THE SACRED COMPUTER

TODOR ARNAUDOV - TOSH

LAZAR

THE PROPHETS OF THE THINKING MACHINES

ARTIFICIAL GENERAL INTELLIGENCE & TRANSHUMANISM

HISTORY THEORY AND PIONEERS
PAST PRESENT AND FUTURE

by the author of the world's first university course in
Artificial General Intelligence and the
Theory of Universe and Mind

ПРОРОЦИТЕ НА МИСЛЕЩИТЕ МАШИНИ

ИЗКУСТВЕН РАЗУМ И РАЗВИТИЕ НА ЧОВЕКА

ИСТОРИЯ ТЕОРИЯ И ПИОНЕРИ; МИНАЛО НАСТОЯЩЕ И БЪДЕЩЕ

СВЕЩЕНИЯТ СМЕТАЧ
ТОДОР АРНАУДОВ – ТОШ

ЛАЗАР

ПРОРОЦИТЕ НА МИСЛЕЩИТЕ МАШИНИ ИЗКУСТВЕН РАЗУМ И РАЗВИТИЕ НА ЧОВЕКА

ИСТОРИЯ ТЕОРИЯ И ПИОНЕРИ
МИНАЛО НАСТОЯЩЕ И БЪДЕЩЕ

от автора на първия в света
университетски курс по
Универсален изкуствен разум и
Теория на разума и вселената

THE PROPHETS OF THE THINKING MACHINES
ARTIFICIAL GENERAL INTELLIGENCE & TRANSHUMANISM
HISTORY THEORY AND PIONEERS; PAST PRESENT AND FUTURE

Corrections: 4.5.2026
Edition: 28.8.2025
First Published: 18.8.2025
http://twenkid.com/agi
https://github.com/twenkid/sigi-2025
https://artificial-mind.blogspot.com/
https://research.twenkid.com/

LAZAR

Appendix to

The Prophets of the Thinking Machines:

Artificial General Intelligence and Transhumanism
History, Theory and Pioneers
Past, Present and Future

© Authors: Todor Arnaudov & All cited original authors

Todor's roles:
1. Author of *The Prophets of the Thinking Machines* – the hyperbook of which Lazar is a volume and appendix.
2. Editor, surveys; selection and curation of: topics, authors, papers, summaries, citations, important concepts etc.
3. Author of reviews, notes, interpretations, criticism, suggestions etc.

The Sacred Computer: *Thinking Machines, Creativity and Human Development*

You can join or help Todor and *The Sacred Computer*!

#lotsofpapers #lazar

Lots of Papers: In AI, ML, CV, ANN, DL, … throughout history, classical 1950s, 1960s, 1970s, 1980s, 1990s, 2000s, early 2010s to 2020s. Computer Vision, Reinforcement Learning, Program Synthesis. Lifelong Learning, Human-Computer Interaction, Mixed Initiative Interfaces (Agentic Systems); Evolutionary programming, Genetic Algorithms, Self-improving agents; Speech Synthesis, Speech Recognition, Audio Generation etc. Groundbreaking or important researchers or related to the flow and context of the reviewed topics; and a few Bulgarian researchers who participated in some of the works.

* Lifelong Learning, Continual Learning, Reinforcement Learning (historical Q-Learning, modern Deep Q-Learning: Atari DeepMind ...), RL for LLMs, policy optimizations (DPO, PPO, OREO) ... Conditional Random Fields (CRF); Chain-of-Thought prompting ...
* Survey of other techniques and research in computer vision, preceding the explosion of the application of convolutional neural networks after 2012: DBN, RBM, MRF, SIFT etc.
* Survey on Early Deep Learning architectures and Normalizations; seminal papers and PhD theses by pioneers in DL from the schools of LeCun, Hinton, Bengio: M.Ranzato, V.Mnih, A.Krizhevsky, I.Sutskever, A.Mohamed ...
* Survey of Object Recognition and Classification before Deep Learning
* Selected Computer Vision works from 1960s to 2020s
* Exploration and introduction of concepts and techniques in computer vision, machine learning, neural networks
* Mixed-Initiative Interaction (Agentic Systems)
* Audio: Speech Synthesis, Audio Generation, Speech Recognition; from 1980s to 2010s
* Neural Machine Translation, Language Models, LLMs, Text Generation, Text and Language Learning and Representation, Word Embedding
* Alternative approaches for sequence and next words prediction, instead of neural networks: stochastic memorizer, sequence memorizer* Transformers - the seminal paper from 2017
* Neural Program Synthesis

* Transformer architectures for images and for reducing the quadratic complexity
* Evolutionary Algorithms | Genetic Algorithms | Genetic Programming
* Genetic Programming, Genetic Algorithms, Evolutionary Programming: Part II
* Summary and selection of important concepts in Evolutionary Algorithms | Genetic Algorithms | Genetic Programming etc.
* Vision Transformers – ViT
* Multimodal Learning, Dialog Learning, Continual Learning
* Diffusion Models
* Self-Improving General Intelligence, Recursive Self-Improvement

In order to quickly go to a target: **Search** with the following pattern:
Topics: Genetic Programming, Genetic Algorithms" etc. (a quote from the list)

Introductory Notes

* Most of the text in this volume of references is in English, with only several small sections and notes in Bulgarian and an introduction in Bulgarian.
* It is a record of explorations and reasoning and can serve as a dataset for study, recall and food for thought for curious readers *and thinking machines*.
* The order and clustering of papers and topics are going in several overlapping dimensions: thematic, chronological and by authors.
* The book is encyclopaedic and can search for specific topics, names, fields, terms etc. like with the other volumes of *The Prophets*.
* Why different fonts are used? For example – for more diversity. Future versions may be generated from a database core or displayed with Vsy/AI agents and generate different layouts, connections etc.
* In the future it would be operated with the AGI infrastructure Vsy (Вседържец, Вси) and the Research Accelerator ACS (unpublished yet), which will make it "living", dynamic, self-extending, translating etc. Of course, you could read the book interactively with other chatbots and agentic AI systems.
*** Criticism:** The information is too dense. Some abstracts are overly compressed and lack flow. The reader is forced to make large cognitive leaps between concepts, domains, levels of abstraction, time periods, and technical details, which leads to information overload.

In general, this is a survey which aims at collecting a huge amount of knowledge and presenting it in a compressed form of selected key fragments and moments, serving as a guide and a seed for additional exploration. However, the density and compression ratio vary and they are not always extreme. Besides you can easily expand the content by consulting the referred material and asking AI. This will be an interactive and convenient operation with Vsy/Вседържец/ACS.

* For more Bulgarian and other selected research see also **#Anelia #prophets #Listove** etc.
* Този том е на английски с няколко кратки бележки и други откъси на български език за учене с подкрепение и „Отложеното възнаграждение", ок. с.16. Виж: Лазар Вълков, Боян Александров.
* Изследвания на други методи в компютърното зрение от 1960-те до 2000-те и в навечерието на бума на приложенията на конволюционни невронни мрежи в края на 2012 г.; методи в машинното обучение: DBN, RBM, MRF, SIFT; MSER и др. (…)

Legend: Cmp**:** compare. Tosh, T.A., TA, Todor, T: inlined comments by the author of "The Prophets …".
File: Lazar-… #Lazar

See also appendices: in English: Universe and Mind 6, "Calculus of Art: …", … many parts of Listove and Irina; some parts of the main volume; #Anelia – a big survey of AI and computer science, including work of many Bulgarian researchers. It is mainly in English and some Bulgarian.

@Vsy: Systematize, draw diagrams, tables, collect in memory

A reminder for navigation: Search topics with "topics:.." and keywords or authors.

Томове и приложения на *Пророците на мислещите машини*

Съществуващи и някои възможни бъдещи томове

[Това са много кратки съкратени бележки за многообразни и обемни трудове; някои са съставяни преди завършването на големите томове (Листове, Основен) – за повече подробности виж другите книги.]

*** #prophets –** Основен том (>2390 с. 1.2026; >1859 стр., 13.8.2025); Обзор на Теория на Разума и Вселената, сравнение с работи в други школи, които преоткриват и повтарят, или пък предхождат обобщаването на принципите за създаване на общ изкуствен интелект, които бяха формулирани още в началото на 2000-те г. и постепенно се сбъднаха и се сбъдват. Документален преглед на огромен обем научни школи, литература и факти, кратка и подробна хронология … #tosh1

*** #purvata –** *„Първата съвременна стратегия за развитие чрез ИИ е публикувана от 18-годишен българин през 2003 г. и повторена и изпълнена от целия свят 15-20 години по-късно: Българските пророчества: Как бих инвестирал един милион с най-голяма полза за развитието на страната?“* #tosh2 (31.5.2025, 248 стр.)

*** #listove –** Многообразие от теми сред които класическа и съвременна роботика и планиране, мулти-агентни системи – класически и съвременни с големи езикови модели; учене с подкрепление в машинното обучение и във връзка с мозъка; невронауки и невроморфни системи, съзнание и панпсихизъм, алгоритмична сложност, други теории на всичко и вселената сметач; когнитивна лингвистика и мислене по аналогия, езикови модели –

исторически и най-нови системи, мултимодални модели, основни модели за агенти и роботи; обзор на научни статии, новини, платформи на чатботове и други пораждащи модели за различни модалности и практика; съветска школа в изкуствения интелект; принос на българи в някои области като ученето с подкрепление (Reinforcement Learning); и др. (...), > 730 р. ... > 475 стр.. (28.8.2025) (414 стр 13.8.2025 г.)

*** #mortal – Нужни ли са смъртни изчислителни системи за създаване на универсални мислещи машини?,** „Смъртните“ системи са свързани с носителя си, за разлика от „безсмъртни“, за каквито се смятат „обикновените“ компютри. Но дали и невроморфните са наистина невроморфни, и какво точно е „безсмъртност“, „смъртност“, „самосъздаване“ (автопоеза) и дали въобще е възможна. Наистина ли са по-ефективни невроморфните системи, както и живите или по-модерните електронни технологии с по-малки транзистори, или ефективността е избор на „счетоводство“ и скриване на реалните разходи за създаването и съществуването на съответната технология? (...) 70 стр.

*** #universe6 #UnM6 –** Вселена и Разум 6, Т.Арнаудов– #tosh3; съзнание, „метафизика“, „умоплащение“ ... на английски; свързана с теми от #mortal (...)

*** Universe and Mind 6** – Connected to “Is Mortal Computation...” – in English. Why infinity doesn’t exist and Goedel theorems are irrelevant for thinking machines? What is Truth, Real and Realness and Why? The fundamentality of mapping (...)

*** #sf #cyber –** Научна фантастика за ИИ, Футурология, Кибернетика ... Подробен преглед и сравнение на статия на Майкъл Левин от 2024 г. за самоимпровизиращата се памет с идеи от Теория на Разума и Вселената.

*** #irina –** Беседи и подробни бележки и др. статии; Ирина Риш; вижданията на Йоша Бах и др. и съвпаденията на идеите му с Теория на Разума и Вселената, публикувана 20 години преди коментираните дискусии; интервю с Питър Вос на ръба преди „ерата“ на ентусиазма към Общия ИИ през 2013 г.; сбъднали се предвиждания от 2005 г. за машинния превод и творчеството и за автоматичното програмиране от 2018 г. и др.; беседа с участието на Майкъл Левин (повече от него в #Основния том, #Кибернетика и #Листове.

→ * #lazar #lotsofpapers – this volume; този том ←

*** #anelia –** Обзор на работата на множество български изследователи.

Survey of the work of numerous Bulgarian researchers. See also the main volume and Listove.

*** #instituti**– преглед на институти по ИИ в Източна Европа и света, сравнение на повтарящите се послания; към 2003 г. в България имаше публикувани **2 национални стратегии** за развитие с ИИ - 16 години преди първата чернова на БАН и 19 години преди откриването на INSAIT, и двете дело на юноши.

*** #complexity** – Алгоритмична сложност – обзор и бележки по множество статии и обобщения и изводи. Дали машината на Тюринг е подходяща за описание на *Мислеща машина?* #hector

*** #calculusofart** – Математически анализ на изкуството. Музика I – Как се определя дали даден „къс“ изкуство е красив и защо ни харесва? Красотата, компресирането и предвиждането на бъдещите данни въз основа на миналите. Музиката трябва да е красива и да се измерва във всички мащаби, от най-малките с постепенно нарастващ обхват.

*** #kotkata –** Задачата от *„Анализ на смисъла на изречение въз основа на базата знания на действаща мислеща машина (…)“,* Т.Арнаудов 2004 г. в диалог с чатботовете ChatGPT и Bard, края на 2023 г. до нач. на 2024 г. и с GPT5 пред 2025 г. – последият успява да разбере и приложи в опростен вид метода от статията.

* **#zabluda –** Заблуждаващите понятия и разбор на истинския им смисъл: трансхуманизъм, цивилизация, … – книга, която публикувах през 2020 г. и започна като статия за трансхуманизма. Откъсът може да бъде включен и в отделно приложение.

#razvitie #transhumanism – том фокусиран върху развитието на човека, космизъм, „трансхуманизъм“; етика, биотехнологии, мозъчно-компютърен / мозъчно-машинен взаимлик (Brain-Computer Interface, Brain-Machine Interface), невроморфни системи, генетично инженерство, геномика, биология, симулиране на клетки и живи организми и др.

Практика, работилници и др. (бъдещи)

*** #robots-drones-ros-slam-simulation-rl** – Наземни и летящи роботи: дронове; обща теория, практика, конкретни системи и приложения; Robot Operating System (ROS, ROS2); среди за симулации на физически и виртуални роботи и машинно обучение: Gazebo, MuJoCo, RoboTHOR, Isaac Sim, Omniverse; gymnasium и др.

*** #neuromorphic-snn-practice** – Практика по невроморфни системи, импулсни невронни мрежи; Lava-nc и др.

*** #llm-generative-agents** – големи езикови модели: локална работа, платформи; употреба, подготвяне на набори от данни; обучение, тестване. Текст, образ, видео, триизмерни модели, програмен код, цели игри и светове с физика („world modeling“), всякакви модалности; дифузни модели, преобразители (трансформатори), съгласувани с физиката математически модели, причинностни модели с управляващо-причиняващи устройства по идеите от Теория на Разума и Вселената. Агенти, мулти-агентни системи: архитектури и др … (виж ***Листове*** и ***Лазар****)*

*** #codegen –** автоматично програмиране, синтез на програми; модели за тази цел, платформи; методи, приложения … program synthesis, automatic programming, code generation

*** #sigi-evolve –** саморазвиващи се машини, еволюционни техники, рекурсивно самоусъвършенстване (Recursive Self-Improvement, RSI)

*** #appx** – **Приложение на приложенията**, списък с добавени по-късно; ръководство за четене и др.

* **#agi-chronicles** – хронологичен запис и проследяване на развитието на историята, новини, събития, идеи, системи, приложения; изследователи *(вероятно с Вседържец)*

… следват продължения – други приложения и *Вселената*:

*** Създаване на мислещи машини** – … Зрим, Вседържец , Вършерод, Казбород, Всеборавител, Всетводейство, Всевод, …

Внимание! Горният списък и информацията в него може да са непълни, неточни или остарели. Възможно е да излизат нови издания с поправки и допълнения. За обновления следете уеб страниците, фейсбук групата „Универсален изкуствен разум“, Ютюб каналите, Дискорд сървъра и др.

Можете да помогнете за подобрението на съществуващите и за осъществяването на бъдещите разработки.

Свещеният сметач призовава съюзници, съмишленици, съдружници и сътрудници; университети, изследователски институти и фирми; учени, инженери, разработчици и творци; спомоществователи, дарители, последователи, другари, изследователи и съавтори за продължения и подобрени версии и за развитие на дейността на изследователско-творческото дружество.

Ако искате да помогнете, за конкретни идеи вижте в началото на основния том, в приложение *Листове,* в информацията за проекта **Вседържец** или *се свържете с мен.*

Работата по системата би се ускорила[1] чрез всякаква съвременна или използваема техника за *нова* изследователска дейност, както и достъп до облачни услуги от всякакъв вид – както достъп до пораждащи модели като ChatGPT, Gemini, Claude и пр.; така и до сървъри, дисково пространство и пр. Също така могат да ни помогнат помещения за техника и работа и пр.

Съхраняването на българската и световна компютърна история и памет е част от дейността на *Сметача* още от зората му през 2000 г.

Стара българска и световна изчислителна техника, която искате да дарите, също е добре дошла при нас или при колеги.

---

[1] Мислещата машина Вседържец работи на такава техника, с каквато разполагаме и се вмества в нея. Ако не можем да осигурим друго, Вседържец ще трябва да се справи и на едно или няколко PC-та, лаптопи и по-малки компютри, с възможна връзка и към мобилни устройства за сензори и други помощни обработки; без или със достъп до Интернет и облачни услуги, каквито можем да си позволим. Виж *„Сингулярност на Тош“ – уравнението на относителната ефективност* в приложение *„Първата съвременна стратегия за развитие чрез ИИ…“*
* Виж суперкомпютъра *Петак I*: https://github.com/Twenkid/SIGI-2025/blob/main/petaki.md – Под „съвременна техника“ се разбира и машини с процесори Core i от *всички поколения.* През 2014 г., работейки върху инфраструктурата на прототипа на *Вседържец* споменавам 4-ядрен процесор
Core i5 4570 с 32 GB RAM и едно тогавашно GPU като може би достатъчни за зародиш на разум, достигащ човешко и свръхчовешко ниво на универсален разум.
https://artificial-mind.blogspot.com/2014/03/whats-going-on-everything-predictions.html

LAZAR

*** Topics:** Lifelong Learning, Continual Learning, Reinforcement Learning (historical Q-Learning, modern Deep Q-Learning DNN: Atari DeepMind …), RL for LLMs, policy optimizations (DPO, PPO, OREO) … Conditional Random Fields (CRF); Chain-of-Thought prompting …

*** Lazar Ignatov Valkov –** a Bulgarian researcher **– Лазар Игнатов Вълков**
https://scholar.google.com/citations?user=GMeeGyEAAAAJ&hl=en

Houdini: Lifelong learning as program synthesis. Lazar Valkov, Dipak Chaudhari, Akash Srivastava, Charles Sutton, and Swarat Chaudhuri. Advances in Neural Information Processing Systems, 31, 2018,

Modular Lifelong Machine Learning, Lazar Ignatov Valkov, 2022/2023, PhD Thesis, Institute for Adaptive and Neural Computation School of Informatics University of Edinburgh – **Continual learning**, Lifelong Machine Learning (LML), split the NN into modules; solve a sequence of problems, one at a time; HOUDINI – modular neurosymbolic framework for modular LML, program synthesis to select a suitable neural architecture; probabilistic search framework – PICLE for module combinations. Three contributions: reducing the number of examples for learning (better generalization), extending the scope of solved problems with learning and making it easier to understand how the knowledge is being reused.

*** A Probabilistic Framework for Modular Continual Learning**, Lazar Valkov, Akash Srivastava, Swarat Chaudhuri, **Charles Sutton**, 5.2024
https://arxiv.org/pdf/2306.06545 - different composition of modules for each problem: *module compositions*; PICLE

An Introduction to Conditional Random Fields, Charles Sutton, Andrew McCallum, 17.11. 2010 https://arxiv.org/pdf/1011.4088
CRF – an important probabilistic method from 2000s & early-mid 2010s.

6pSrivastava, L Valkov, C Russell, MU Gutmann, C Sutton, Advances in neural information processing systems 30, 846, 2017
https://proceedings.neurips.cc/paper/2017/file/44a2e0804995faf8d2e3b084a1e2db1d-Paper.pdf

* Adaptive Memory Replay for Continual Learning, James Seale Smith, Lazar Valkov, Shaunak Halbe, Vyshnavi Gutta, Rogerio Feris, Zsolt Kira, Leonid Karlinsky, 2024 https://arxiv.org/abs/2404.12526 – *Extended Pre-Training* (EPT); randomly selects from all past data to reduce catastrophic forgetting; memory replay datasets; adaptive memory replay; rehearsal-based methods; - store selective samples of the previous tasks; Experience replay from memory, selective replay … See also the BgGPT training.

Todor: Replay (retraining) is not a change in the general approach of training NNs. See *"Replay in Brains and Machines"* in *#Listove.*

*** Programming by example (PBE)** https://www.microsoft.com/en-us/research/wp-content/uploads/2017/10/program_synthesis_now.pdf

- Program synthesis with large language models, Jacob Austin, et al. 8.2021 https://arxiv.org/abs/2108.07732 Mostly Basic Programming Problems (MBPP) dataset contains 974 programming tasks

*** Chain-of-Thought Prompting Elicits Reasoning in Large Language Models** Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, Denny Zhou, 1.2022/1.2023 https://arxiv.org/abs/2201.11903
Math: GSM8K, SVAMP, ASDiv, AquA, MAWPS ; **Commonsense:** CSQA, StrategyQA, SayCan; few-shot prompting, popularized by Brown et al. (2020), .. p.41, Table 28, *Prompt for SayCan Robot Planning:*
Locations = [counter, table, user, trash, bowl,door]. Objects = [apple, kettle chips, tea, multigrain chips, coke, potato chips, rice chips, orange, grapefruit soda, pepsi, waffle, sponge, water]. The robot can pick up items with pick(object) and put down items with put(object) as well as find objects or locations with find(). The robot can only understand the explicit locations and objects listed. Human: I'm hungry, can you bring me some chips. Explanation: The user is hungry and has asked for chips. There are several types of chips available, I will bring the user the potato chips. Plan: 1. find(potato chips), 2. pick(potato chips), 3. find(user), 4. put(potato chips), 5. done(). … **Human:** Bring me a drink. Explanation: The user has asked for a drink and there are many options. I will bring them some water. Plan: 1. find(water), 2. pick(water), 3. find(user), 4. put(water), 5. done(). SayCan (Ahn et al., 2022): https://say-can.github.io/ Compare PLANNER/STRIPS from late 1960s-early 1970s.

*** GraphPrompt: Unifying Pre-Training and Downstream Tasks for Graph Neural Networks**, Zemin Liu, X.Yu, Y.Fang, X.Zhang, 25.2.2023, GNN, Pre-training – fine-tuning on downstream tasks; objectives: node/edge features, node connectivity/links, local/global patterns; GPPT – Graph pre-training; prompts – more efficient than fine-tuning when pre-

training on big datasets; task-specific learnable prompts; … topology, subgraphs; contextual subgraph; … link prediction; classification of grouping – similarity … unified template for prompting; ReadOut operation: aggregation function to fuse node representations. GraphPrompt: subgraph similarity; Datasets: Flickr, PROTEINS, COX2, ENZYMES, BZR. Leanabe prompt-vector; Avg.nodes/Avg.edges/Node features/Node Classes, Task per DS: ~(89K, 899K, 500, 7, N), (39,73,1,3,N/G), (41,43,3,-,G), (32,62,18,3, N/G), (35,38,3,-,G) N – node clsf, G – graph clsf (classes, labels); Link prediction: similarity between connected subgraphs; few-shot learning methods: Meta-GNN, RALE; Graph pre-training models:DGI, InfoGraph, GraphCL; end-to-end graph neural networks: GCN, GraphSAGE, GAT, GIN. https://arxiv.org/pdf/2302.08043
GNN encoder … Readout – reads neighbours; Learnable node classification prompt and Learnable Graph classification prompt (the whole graph) … 50-shot learning (50 sample ground-truth example, still just 0.06% in the flickr graph)

* **CLDG: Contrastive Learning on Dynamic Graphs**, Yiming Xu et al., 19.12.2024 – a sampling layer to extract the temporally-persistent signals; temporal translation invariance under timespan views; consistent local and global repres.; maximizing the mutual information (MI) between different augmented views; Discrete-time Dynamic Graph: DTDG – sequence of network snapshots within a given time interval; Continuous-time Dynamic Graph – the nodes are annotated with timestamps: (…) https://arxiv.org/abs/2412.14451v1

* Brian Lester, Rami Al-Rfou, and Noah Constant. 2021**. The Power of Scale for Parameter-Efficient Prompt Tuning**. In Conference on Empirical Methods in Natural Language Processing. 3045–3059.
* Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. arXiv preprint arXiv:2107.13586 (2021)

*** Personalized Mathematical Word Problem Generation Oleksandr Polozov et al.** http://www.eleanorourke.com/papers/word_problems_ijcai.pdf – facts and rules in a first-order logic representation (syntactically similar to Prolog). Propositionalization of this program (called grounding), answer set solvers search the space of truth assignments associated with each logical statement. problem encoding, problem instance; logical graph; grounding; saturation technique; skolemization [Benedetti, 2005]; Disjunctive rules**;** answer-set programming - ASP; classic guidelines of NLG systems [Reiter and Dale, 1997]; plot generation, discourse tropes; automatic problem generation; PCG: procedural content generation via declarative specs; primitive templates, templates, entity references disambiguation, non-repetitively … to produce valid English. *Primitive templates;* sentence ordering.

* Marco Benedetti**. Evaluating QBFs via symbolic skolemization**. In Logic for Programming, Artificial Intelligence, and Reasoning, pages 285–300. Springer, 2005.

*** Proximal Policy Optimization Algorithms**, John Schulman, Filip Wolski, Prafulla

Dhariwal, Alec Radford, Oleg Klimov https://arxiv.org/pdf/1707.06347 PPO
Trust Region Methods - TRPO [Sch+15b] – surrogate objective; limit the size of policy update in order to avoid worsening the performance; cmp Kldiv … conjugate gradient algorithm: penalty instead of constraint; Clipped Surrogate Objective … A2C [Mni+16], A2C with trust region [Wan+16]. A2C stands for Advantage Actor Critic …

* Ilias Chrysovergis, **Proximal Policy Optimization –** example with code, 2021/2024
https://keras.io/examples/rl/ppo_cartpole/
See also a Deep-Q example for Breakout: https://github.com/keras-team/keras-io/blob/master/examples/rl/deep_q_network_breakout.py
Actor-critic (simpler) etc.:
https://github.com/keras-team/keras-io/blob/master/examples/rl/ipynb/actor_critic_cartpole.ipynb
The Actor: returns a probability for each action in its action space;
The Critic: estimates the total reward in the future, given the state of the environment.
num_inputs = 4 – state of the env.; num_actions = 2 – what can be done, affordances; num_hidden = 128 – learning parameters; … The example shows the usage of tf.GradientTape for automatic differentiation and gradient descent calculations.
Cmp. #bongard, „Животно“ in #Listove #prophets (основен том)
CNN; DeepMind; … “Off-Policy” method, meaning its Q values are updated *assuming* that the best action was chosen, even if it was not applied as an action in the *behavioral* policy.
*** Q-Learning**, CHRISTOPHER J.C.H. WATKINS, 1992
https://link.springer.com/content/pdf/10.1007/BF00992698.pdf
Model-free RL; “*Asynchronous dynamic programming”* (async. DP), act optimally in Markovian domains … action-replay process (ARP);
* Watkins, C.J.C.H. (1989). **Learning from delayed rewards.** PhD Thesis, University of Cambridge, England.
https://www.researchgate.net/publication/33784417_Learning_From_Delayed_Rewards
“behavioral ecology, stochastic DP for calculating animals’ optimal behavioral policies… performance criterion, efficiency.; optimal learning; learning of efficient strategies – pulling which “lever” gives the highest reward on average (“two-armed bandit”) … (See Chapter 1) – “Гл.1. Обучението в поведение, което се възнаграждава е признак за интелигентност, напр. естествено е да се обучи куче чрез награждаване, когато отговаря по подходящ начин на команди. … Странно е обаче, че този тип обучение е бил в голяма степен пренебрегнат в когнитивната наука, поради което *не познавам дори и една статия за интелигентността на животните, която да е била публикувана в основния поток на литературата по „изкуствен интелект“*. …“общ подход за обучение чрез награди и наказания, който може да се приложи върху широко множество от ситуации …“
Аргумент за оптималност на поведението, за да оцелява – но критики, че се отнася за животното и поведението му като цяло, а не за всяка подробност…

Бел. Тош: 1) И също в сравнение с други; не най-доброто, а просто *по-добро, достатъчно добро,* така че системата да продължи да съществува или тя да продължи, а не другата. Така и „оцелява най-приспособеният“ по Дарвин е погрешно. Оцелява

достатъчно приспособеният, което също така е тавтология и тривиалност: ако не издържаш на дадена температура, не можеш да дишаш под вода, а е нужно да дишаш, за да живееш и т.н., няма да оцелееш под вода. Интересното е *кой е приспособен, защо, как се приспособява, защо*. Случайните мутации не обясняват почти нищо, „пробвай и ще видиш“ – защо се случват точно в даден момент, кои се случват и т.н.).
2) В човешки условия наградата според обучителя не винаги е тази, която обучаваният възприема по този начин. Виж критиката на „Marshmallow Test” от Т.Арнаудов в статията в приложенията в основния том*. RL по учебниково определение връща *скаларна награда*, само една стойност, и може би тази особеност е привлекателна, защото опростява задачата и дава едно решение. Сложните агенти като човеци обаче в „свободно състояние“ нямат задължително единична проста функция на ползата, затова и често изглеждат „*нерационални*“, когато оценителят се опита да ги сведе до скаларна функция, понеже поведението им се разминава с очаквана предполагаема опростена траектория, която допуска само една награда и/или погрешна функция на ползата.

*** Отложеното възнаграждение е заблуждаващо понятие,** Тодор Арнаудов, „Изкуствен разум“ – Artificial Mind, 27.6.2018 (Виж разширена версия на български в края на основния том, след 1832 с. в редакцията към 19.8.2025) или на английски в блога:
*** Delayed gratification is an ILL-Defined Concept:** In Analysis, Articles, Developmental Psychology, Neuroscience by Todor "Tosh" Arnaudov - Twenkid // Wednesday, June 27, 2018 // Leave a Comment
https://artificial-mind.blogspot.com/2018/06/delayed-gratification-is-ill-defined.html

* **Offline Reinforcement Learning for LLM Multi-Step Reasoning**. Huaijie Wang, Shibo Hao, Hanze Dong, Shenao Zhang, Yilin Bao, Ziran Yang, Yi Wu, 20.12.2024 https://arxiv.org/pdf/2412.16145 – Offline Reasoning Optimization: OREO. LLM reasoning with reinforcement learning (RL). Compare to DPO – Direct Preference Optimization (costly, pairs of data for preferences); Proximal Policy Optimization PPO – expensive for training. SFT – Supervised Fine Tuning.

* **Human-level control through deep reinforcement learning**. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg & Demis Hassabis, 2015. Nature.
https://web.stanford.edu/class/psych209/Readings/MnihEtAlHassibis15NatureControlDeepRL.pdf (Slides, 2015) Atari 2600 RL from pixels, DeepMind. RL + DNN = Deep Reinforcement Learning; DQN – Deep Q-Network (traditionally RL was limited to low dim.input) – experience replay, fixed Q-target
https://www.semanticscholar.org/paper/Human-level-control-through-deep-reinforcement-Mnih-Kavukcuoglu/340f48901f72278f6bf78a04ee5b01df208cc508)
https://courses.grainger.illinois.edu/cs546/sp2018/Slides/Apr05_Minh.pdf

*** Volodymyr Mnih:** https://scholar.google.com/citations?user=rLdfJ1gAAAAJ&hl=en
https://www.cs.toronto.edu/~vmnih/
* V.Mnih, **Machine Learning for Aerial Image Labeling**, PhD Thesis, 2013 – advisor G.Hinton **See references from th ethesis etc.:**

*** Topics: Survey of Object Recognition and Classification before Deep Learning**

* A seminal work on image segmentation and classification:
* R. L. Kettig, **Computer Classification of Remotely Sensed Multispectral Image Data by Extraction and Classification of Homogeneous Objects**, 1975, NASA-CR-147403, … Purdue University, 2.1976
https://ntrs.nasa.gov/api/citations/19760014577/downloads/19760014577.pdf
"The pixels within a given object of a given spectral class are completely characterized by their class-conditional joint probability distribution function. … no-memory classification, - only the marginal distribution of each pixel is required … parametric or non-parametric estimation of the probability distribution function PDF … " Unsupervised Annexation, cell selection … Supervised annexation & cell selection … progressive merging of adjacent elements which are found to be similar according to some statistical criterion. Thus an algorithm consists of statistical tests applied in some logical sequence.
Todor: according to some criterion (it doesn't have to be statistical); 1. Division into a grid of e.g. 2x2 pixels – singular cells; *2. Compare to adjacent "field" – a group of one or more connected cells and eventually merge, else compare to another adjacent field or becomes a new field itself.* The expansion of the fields ends until the natural boundaries, *where the rejection rate "abruptly increases"* … *generalized likelihood ratio* … *sample variance, sample generalized variance … Data representation is 8 bits for aircraft data and 6 bits for LANDSAT*. … Relative error rate …

* R. L. Kettig and D. A. Landgrebe. Classification of multispectral image data by extraction and classification of homogeneous objects. IEEE Transactions on Geoscience Electronics, 14:19–26, 1976 https://www.lars.purdue.edu/home/references/LTR_011976.pdf
* Rodd, E. M., „Closed Boundary Field Selection in Multispectral Digital Images," IBM Publication No. 320.2420, Jan. 1972.
* Kettig, R. L. and D. A. Landgrebe, „Automatic Boundary Finding and Sample Classification of Remotely Sensed Multispectral Data," LARS Information Note 041773, Laboratory for Applications of Remote Sensing, Purdue University, West Lafayette, IN, April 1973. … *"spectral variations in combination with spatial variations .. for automatic boundary finding and sample classification of remotely sensed multispectral data. Preliminary applications of the method to agricultural data show significant improvements in accuracy as compared to the use of spectral data alone."*

* Ruzena Bajcsy and Mohamad Tavakoli. **Computer recognition of roads from satellite pictures.** IEEE Transactions on Systems, Man, and Cybernetics, 6(9):623–637, 9.1976. https://www.academia.edu/36011344/Computer_Recognition_of_Roads_from_Satellite_Pictures *"A program which recognizes real roads, their intersections, and objects which are road-like" … satellite pictures … spectral, geometric and spatial properties. .. **Make hypotheses about the objects in the scene and verify them until there are no contradictions**, using the low-level operators and the world model. .. pattern recognition in high-energy particle physics; minimum spanning tree .. edge-following & line-fitting programs, line finders … Low level operators = directional strip detector, road grower, thinning operator, short segment eliminator, intersection finder, end finder etc. Higher Level World Model; The Scene Land (Vegetation, Non-Vegetation) Land (Islands, Continent) Land (Natural Land (Forest, Desert), Man-Made Solid Objects(Cities, Roads, Bridges, Pollutants)) … The Scene Waterway (Small Waterway (Rivers, Lakes), Ocean Sea;* **Features:** *Reflectivity, Shape, Texture, Relationships Shape Size Reflectivity …* **Criteria:** *"above, inside, connected, surrounded by …"* **neighbours …**

* COOX, C. M. and ROSENFELD, A. "Size detectors." Proc. IEEE 58 (Dec. 1970), 1956-1957.

* **A. Rosenfeld**, "Picture Processing by Computer," Computing Surveys 1, 1969, pp. 147-176.

* **AZRIEL ROSENFELD**, Progress in Picture Processing: 1969-71, University of Maryland, College Park, Maryland, Computing Surveys, Vol 5, No 2, June 1973 https://dl.acm.org/doi/pdf/10.1145/356616.356617

* Azriel Rosenfeld, "Progress in picture processing," Techn. Report, TR-176, University of Maryland, January 1972;

* A. Rosenfeld, "Picture processing," 1972 Techn. Report TR-217, Univ. of Maryland, 1.1973.

* A. Rosenfeld, "Picture Processing," 1973 Techn. Report TR-284, U.of.M. 1.1974.

* A.Rosenfeld, J.S.Weszka, **Picture Recognition and Scene Analysis, 5.1976** https://www.computer.org/csdl/magazine/co/1976/05/01647359/13rRUxBrGku … *"The development of techniques for the computer analysis of pictures and scenes began over 20 years ago. Most of the work in this field has been application-oriented; some of the major applications areas are automation (robot vision), cytology, radiology, high-energy physics, remote sensing, and document processing (character recognition)"*. However, many of the techniques and algorithms are more general than their original application. *"The goal of picture (or scene) analysis is the extraction of a description from the given picture. This description may consist of a set of numerical data (feature measurements), or it may be some type of data structure which represents relationships among significant parts (segments) of the picture, as well as properties of these parts. Thus in general, picture analysis involves segmentation of the picture into parts; measurement of properties of the parts (including both grayscale-dependent properties such as texture, and geometrical properties such as size and shape); and determination of relations among the parts. In addition, it may be desirable to „preprocess" the picture-i.e., to modify it so as to make the subsequent analysis steps easier*

or more reliable. ***This paper reviews some of the early milestones in picture recognition and scene analysis techniques.*** *Topics covered include preprocessing (noise cleaning, deblurring, filtering); segmentation (region growing, decomposition of regions into parts, grouping of regions into „objects"); property measurement (invariant properties, textural properties); shape analysis (local shape features, boundary and skeleton representations); and structural analysis (syntactic analysis, model matching). The emphasis is on describing representative ideas that are of* ***historical importance****, rather than on giving a systematic treatment of the subject. "* p.8 **Local shape features**: Closed contour: Bay, Convex corner, Spur, Peninsula, Concave corner, Notch; Contour of a band/a path with two approximately parallel contour lines: End, Branching, Crossing … skeleton or medial axis; chord statistics: intersections with families or lines or line segments; perimeter of the convex hull … boundaries & parametric equations; Fourier descriptors; *Chain code – sequence of horizontal, vertical or diagonal moves of length 1 or SQRT(2)* (vector drawing step-by-step, tracking). Outlining, deblurring … Segmentation: homomorphic filtering, run tracking and shrinking: intersections, merges, splits, size changes.; minimum-area circumscribed rectangle … linking, normalization by autocorrelation; Moments … double integrals … texture: *coarseness: the size of "pieces", "elements" and directionality – variation of the coarseness as a f(direction)* … ***1950s - rate of falloff of the texture's autocorrelation*** *or of the power of the spectrum as one moves away from the origin* (shifting)*…* ***Early 1960s*** *– statistical approach – P(particular pairs of gray levels in relative position (u,v)). … Transition probability matrices* … second-order transition probabilities are affected by the first order … **Preprocessing**: normalize the grayscale – region growing; Grouping regions into "objects; … autocorrelation; Textures; Geometrical Properties: local, chord statistics; **curve**: slope, arc length … Structural Analysis – syntactic approach – e.g. strokes in handwriting. *Higher-level knowledge e.g. object models* …

* Survey of other techniques and research in computer vision, preceding the explosion of the application of convolutional neural networks after 2012: DBN, RBM, MRF, SIFT
* Изследвания на други методи в компютърното зрение от навечерието на бума на приложенията на конволюционни невронни мрежи в края на 2012 г.: DBN, RBM, MRF, SIFT

* **Generating more realistic images using gated MRF's**. M. Ranzato, V. Mnih, and G. Hinton. In NIPS, 2010. Possibly that's the same paper with a different name:
* **How to generate realistic images using gated MRF's**, M. Ranzato, V. Mnih, and G. Hinton , 2010, https://www.cs.toronto.edu/~vmnih/docs/gen_images.pdf
Вероятностните модели на естествени изображения обикновено се оценяват непряко чрез задачи като премахване на шума и рисуване на маскирани части. По-пряк начин за оценка на пораждащ модел е да се извлекат проби от него и да се провери дали статистическите свойства на пробите съвпадат със статистиката на естествените изображения. Този метод рядко се използва с изображения с висока разделителна способност, тъй като настоящите модели произвеждат изображения, които са много различни от естествените, което може да се оценява дори с просто око. Изследваме причините за този неуспех и показваме, че чрез разширяване на съществуващите модели до два набора от скрити променливи: единият набор моделира интензитета на пикселите, а другият - ковариациите на пикселите, които са особени за изображението. По този начин можем да генерираме изображения с висока разделителна способност, които изглеждат много по-реалистични от преди. Цялостният модел може да се тълкува като вероятностно поле на Марков или Мрежа на Марков (MRF – Markov Random Field), където както зависимостите по двойки, така и средните интензитети на пикселите се модулират от състоянията на скритите променливи. И накрая, потвърждаваме, че ако забраним споделянето на теглата между рецептивни полета, които се припокриват едно с друго, затвореният MRF научава по-ефективни вътрешни представяния, както е показано в няколко задачи за разпознаване.

- **Modeling Natural Images Using Gated MRFs**, Marc'Aurelio Ranzato, Volodymyr Mnih, Joshua M. Susskind, Geoffrey E. Hinton, 2013
  https://www.cs.toronto.edu/~vmnih/docs/ranzato_pami13.pdf
  Gated Markov Random Fields – *two sets of latent variables to create an image-specific energy function that models the covariance structure of the pixels by switching in sets of pairwise interactions;* the second set models the intensities of the pixels. *The Deep Belief Network DBN then uses several layers of Bernoulli latent variables to model the statistical structure in the hidden activities of the two sets of latent variables of the gated MRF. … One simple way to check whether a model extracts features that retain information about the input, is by reconstructing the*

input itself from the features. @Вси: прдлж

* **On Deep Generative Models with Applications to Recognition,** Marc'Aurelio Ranzato, J.Susskind, V.Mnih, G.Hinton, 2011
https://www.cs.utoronto.ca/~vmnih/docs/ranzato_cvpr2011.pdf
a gated MRF + several hidden Deep Belief Network layers (binary) – performs "comparably to SIFT descriptors".

"Най-популярният начин за използване на вероятностни модели в компютърното зрение – първо да се извлекат описатели на свойствата на малки области от изображението или части от обекти чрез добре проектирани признаци, и след това да се използват средства за статистическо обучение, за да се моделират зависимостите между признаците и вероятните етикети. В тази работа използваме един от най-добрите пораждащи модели на ниво пиксели – превключваща мрежа на Марков - като най-ниско ниво на дълбока мрежа на убежденията (deep belief network) с няколко скрити слоя." Ръчно проектираните признаци: SIFT, HoG (Histogram of Oriented Gradients), SURF, PhoG – описват късчета, малки области от изображението и ги натрупват в различни пространствени разделителни способности и по различни части от изображението. Така образуват вектор на признаците, който след това се подава на класификатор с общо предназначение като SVM (Support Vector Machine). Въпреки, че са много успешни, тези методи разчитат твърде много на човешки дизайн на извличането и на обединяването им в общ вектор. При големият и нарастващ брой на лесно достъпни изображения и напредъкът в машинното обучение, би трябвало да бъде възможно да се научат по-добри описатели на късчетата и по-добър начин за обединяването им.*

Fast Persistent Contrastive Divergence (FPCD). "Целта на обучението е да се намерят параметрите на първия слой (C, M, γ b1) и на по-високите (Wi, bi), които увеличават вероятността на образите от обучението. Cohn-Kanade (CK) dataset, Toronto Face Database (TFD) - CK: 920 лицеви изражения 48x48 и 8 вида емоции.
https://www.kaggle.com/datasets/davilsena/ckdataset
* превключваща мрежа на Марков и – gated Markov Random Field; наречена така, защото превключва между двете множества от скрити променливи.

- M. Ranzato and G.E. Hinton**. Modeling pixel means and covariances using factorized third-order boltzmann machines.** In CVPR, 2010
https://www.cs.toronto.edu/~ranzato/publications/ranzato_cvpr2010.pdf
- K. Kavukcuoglu, M. Ranzato, R. Fergus, and Y. LeCun**. Learning invariant features through topographic filter maps**. In CVPR, 2009

* M.Ranzato: https://ranzato.github.io/publications/publications.html

A Seminal PhD Thesis in Deep Learning:

*** Unsupervised Learning of Feature Hierarchies**, Marc'Aurelio Ranzato, 2009 (supervisor: Yann LeCun)

https://www.cs.toronto.edu/~ranzato/publications/ranzato-phd-thesis.pdf *"… this work focuses on "deep learning" methods, a set of techniques and principles to train hierarchical models.Hierarchical models* ***produce feature hierarchies*** *that can capture* ***complex non-linear dependencies*** *among the observed data variables in a concise and efficient manner …* ***Energy-Based Model*** *framework and gradient-based optimization techniques that scale well on large datasets. The principle underlying these algorithms is* ***to learn representations*** *that are at the same time sparse, able to reconstruct the observation, and directly predictable by* ***some learned mapping*** *that can be used for fast inference in test time. With the general principles at the foundation of these algorithms, we validate these models on a variety of tasks, from visual object recognition to text document classification and retrieval … a general class of trainable non-linear functions, dubbed "deep networks" (Hinton et al., 2006; Hinton and Salakhutdinov, 2006; Bengio and LeCun, 2007; Bengio et al., 2007; Ranzato et al., 2007c; Lee et al., 2007). … p.8 A particularly important class of unsupervised algorithms, which includes principal component analysis, K-means, and many others, produces internal representations of data vectors as part of the energy computation. These representations, also known as* ***feature vectors, or codes*** *can be used as input for further processing such as* ***prediction****. Moreover, many unsupervised machines make explicit use of such representations by* ***reconstructing the data vectors from the representations****, and by* ***using the reconstruction error as part of the energy function*** *… K-means: the code is the index of the prototype in the codebook that is closest to the data vector. … p.13. EBM (LeCun et al. 2006, Ranzato et al. 2007a) assigns lower energy values to input vectors that are similar to training samples and higher energy values elsewhere. " … p.25* $F(Y ; W) = min_{i \in [1,...,N]} ||Y - Wi||^2$ *… each training sample can be properly represented by a unique code – a low energy value F(Y ;W) (e.g., good reconstruction) .. unobserved points are assigned codes similar to those associated with training samples, so that their energies (e.g., reconstruction error) are higher. How to* ***limit the information content of the code*** *– finite number of different values (quantization), lower dimensionality than the input, forcing the code to have lower dimensionality than the input ; a term in the energy function E(Y, Z; W) that forces the model to be a "sparse" vector (most components = 0) … Predictive Sparse Decomposition PSD & invariant PSD (IPSD); code contrastive term - without this term, the model could assign the same energy of value to all points in input space. … p. 62(37): Popular unsupervised algorithms in the EBM framework: Principle Component Analysis (PCA), Autoencoder, Restricted Boltzmann Machine, ICA-IM, Sparse coding, PSD, K-Means, Mixture of Gaussians, Prediction of Experts PoE: … Encoder:* $W^T Y$ *; Decoder WZ ; Input Reconstruction Loss =* $|| Y - WZ|_2^2$ *Code Prediction Cost:* $Z = g_e(Y;W)$ *.. Code Cost .. pull-up … etc. (Matrix multiplications, weights, transposed matrices, …); log partition function … regions of high density of training data .. negative log probability loss … p.65 (40) E=squared reconstruction error => minimizing the negative log probability loss does not necessarily lead to good reconstruction of training samples; it is for maximizing the likelihood of the data => compression. However, for extracting features: only regions with the high-density data …* ***Product of Experts:*** *The encoder is a set of linear filters, rows of matrix We, and the energy is defined as: F(Y) =* $\Sigma g_i log(1 + z_i^2)$ *;* $Z = W_e Y$ *and* g_i, $i \in [1, \ldots, N]$ *– set of learnable coefficients. Training: negative log probability loss with a gradient step, approximated by contrastive divergence. … producing the code for a given input is a simple matrix multiplication, but generation –*

expensive sampling, MCMC & HMC [Markov Chain Monte Carlo & Hamiltonian Monte Carlo]*; recent: "persistent Markov Chains"**; Contrastive Margin Loss … Sparse codes … K-Means Clustering – no encoder, only decoder** and a **reconstruction cost** module; squared reconstruction error. **Mixture of Gaussians … Wi -** the mean of the i-th component, Ai is its inverse covariance matrix, W – all the params of the model (means & covariances) … Expectation-Maximiation EM algorithm. Other alg. **Factor Analysis** (Hinton et al., 1997**), Minimum Description Length** (Hinton and Zemel, 1994) – the number of bits that are needed to encode the reconstruction error, to represent the code and to encode the parameters of the model. Injecting noise in the representation in order to limit the inform. Content of the code (Doi et al., 2006); **Score Matching** (Hyvarinen, 2005): min(loss_functional – difference(first_derivative, second_derivative) d(input_training_data)); the training samples are minima of the energy surface in regions of high curvature. **Non-EBM: denoising autoencoder (**LeCun, 1987; Ranzato et al., 2007a; Vincent et al., 2008): noisy input clean version of the input; learning vector fields (corrections). **PSD: Predictive Sparse Decomposition (PSD),** producing features with either direct or non-linear mapping. $g_e(Y ; W_e, D) = D \tanh(W_e Y)$ … $Y \in R^{M-}$ input; $W_e \in R^{N \times M}$ – filter matrix, tanh – hyperbolic tangent non-linearity, $D \in R^{M \times M}$ – a diagonal matrix of coefficients. **… Learning** - find the optimal value of the parameters of both encoder and decoder { We, D, Wd} on-line block coordinate gradient descent … 1. Minimize the energy E(Y, Z; Wd, We, D) with respect to Z, starting from the initial value ge(Y; We, D). … Z 2. Parameter update step – using the optimal value of the code Z found in the prev. step, update the parameters by one step of stochastic gradient descent on the loss. $U \leftarrow U - \eta \, \partial L/\partial U$; where U denotes {We, D, Wd} and η is the step size (learning rate). The columns of Wd are then re-scaled to unit norm.- representation that simultaneously reconstructs the input, is sparse, and is not too different from the predicted representation. …

- Tishby, N., Pereira, F., and Bialek, W. (1995). **The information bottleneck method**. Neural Computation, 7:1129–1159.
  * H. Lee, R. Grosse, R. Ranganath, and A. Y. Ng. **Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations**. In Proc. ICML, 2009.
  *** V. Mnih. Cudamat: a CUDA-based matrix class for python. Technical Report UTML TR 2009-004, Dept. Computer Science, Univ. of Toronto, 2009 – Ранна библиотека за приложения свързани с машинно обучение с CUDA.**
  https://www.cs.toronto.edu/~vmnih/docs/cudamat_tr.pdf
  H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Surf: **Speeded up robust features. In Computer Vision and Image Understanding**, 2008 – SURF feature; see SIFT.
  * US 2009238460, Ryuji Funayama, Hiromichi Yanagihara, Luc Van Gool, Tinne Tuytelaars, Herbert Bay, „ROBUST INTEREST POINT DETECTOR AND DESCRIPTOR", published 2009-09-24 – *Methods and apparatus for operating on images are described, in particular methods and apparatus for interest point detection and/or description working under **different scales** and with **different rotations**, e.g. for **scale-invariant and rotation-invariant interest point detection and/or description**. The present invention can provide improved or alternative apparatus and methods for matching*

interest points either in the same image or in a different image. …
* https://patents.google.com/patent/US20090238460A1/en *… 0004: The most widely used interest point detector probably is the Harris corner detector [10], proposed in 1988, and based on the eigenvalues of the second-moment matrix.. .. – not scale invariant. .. Lindeberg introduced the concept of* ***automatic scale selection****. This allows detection of interest points in an image, each with their own characteristic scale. Harris-Laplace and Hessian-Laplace, scale adapted…. Laplacian of Gaussians (LoG) by a Difference of Gaussians (DoG) filter by Lowe.* Kadir and Brady: entropy-maximizing of a region; edge-based region detector: Jurie et al. Others: *Gaussian derivatives, moment invariants, complex features, steerable filters, phase-based local features and descriptors representing the distribution of smaller-scale features within the interest point neighbourhood – … SIFT .. computes a histogram of local oriented gradients around the interest point and stores the bins in a 128-dimensional vector (8 orientation bins for each of 4×4 location bins). Ke and Sukthankar [22] .. PCA-SIFT: a 36-dimensional descriptor, fast for matching, but less distinctive than SIFT in a second comparative study .. and a slower feature computation reduces the effect of fast matching.* (Principal component analysis of the gradient image) GLOH: also SIFT with PCA. .. In on-line applications all three phases have to be fast: **detection, description, matching.** …
Cited By 196: …
*** Method of detecting and describing features from an intensity image,** 2011, Inventor: Daniel Kurz, Peter Meier, Current Assignee Apple Inc
https://patents.google.com/patent/US20140254874A1/en …

- N. Dalal and B. Triggs. **Histograms of oriented gradients for human detection**. In CVPR, 2005
  * A. Bosch, A. Zisserman, and X. Munoz**. Representing shape with a spatial pyramid kernel**. In CIVR, 2007
  * G. Hinton, S. Osindero, and Y.-W. Teh. **A fast learning algorithm for deep belief nets.** Neural Comp., 18:1527–1554, 2006
  * S. Lazebnik, C. Schmid, and J. Ponce. **Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories.** In CVPR, June 2006
  D. Lowe. **Distinctive image features from scale-invariant keypoints**. IJCV, 2004 (SIFT)
  * A. Krizhevsky. **Learning multiple layers of features from tiny images**, 2009. MSc Thesis, Dept. of Comp. Science, Univ. of Toronto https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf *– p.17-19.* ~ Previous work: unsuccessful attempts to train on the 80 million tiny images dataset, only global filters, point-like identity functions; uniform, global, noisy – modelling the noise [instead of the lower frequency patterns and structure].

  * S. Osindero and G. E. Hinton. **Modeling image patches with a directed hierarchy of markov random fields**. In NIPS, 2008
  C. Tang and C. Eliasmith. **Deep networks for robust visual recognition**. In ICML, 2010
  * Y. W. Teh, M. Welling, S. Osindero, and G. E. Hinton. **Energy-based models for sparse**

overcomplete representations. JMLR, 4:1235–1260, 2003

* U. Schmidt, Q. Gao, and S. Roth. **A generative perspective on mrfs in low-level vision.** In CVPR, 2010.

* T. Kanade, J. Cohn, and Y. Tian. **Comprehensive database for facial expression analysis**. In Int. Conf. on Automatic Face and Gesture Recognition, pages 46–53, 2000

* M.J. Wainwright and E.P. Simoncelli. **Scale mixtures of gaussians and the statistics of natural images.** In NIPS, 2000

* G.E. Hinton. **Training products of experts by minimizing contrastive divergence**. Neural Computation, 14:1771–1800, 2002.

* Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. **Gradient-based learning applied to document recognition.** Proceedings of the IEEE, 86(11):2278–2324, 1998.

* **Does the brain do inverse graphics**, Geoffrey Hinton, Alex Krizhevsky, Navdeep Jaitly, Tijmen Tieleman & Yichuan Tang, 2012. https://www.cs.toronto.edu/~hinton/IPAM5.pdf – Capsules; equivariance, pose vectors; each capsule learns a fixed template which is scaled and translated differently for different images; factor analyzer; learning the basic parts: a "very non-linear" problem. p.36: another way to learn the lowest level parts: pairs of images, related by a known coordinate transformation (e.g. known control, movement, rotation etc. – interaction); *Transforming autoencoders - non-linear recognition units to map the observation space to the space in which the dynamics is linear. Then they use non-linear generation units to map the prediction back to observation space.*

__* Transforming Auto-encoders, G. E. Hinton, A. Krizhevsky & S. D. Wang, 2011 https://www.cs.toronto.edu/~hinton/absps/transauto6.pdf 3. Discussion: … *A transforming auto-encoder can force the outputs of a capsule to represent any property of an image that we can manipulate in a known way.*

__* **Improving neural networks by preventing co-adaptation of feature detectors,** Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, Ruslan R. Salakhutdinov, 7.2012 https://arxiv.org/abs/1207.0580 - **dropout** of 50% of the hidden units and **data augmentation** for reducing the overfitting and improving the classification performance; the work introduces the ImageNet dataset in appx. E and proposes the AlexNet briefly in appx. H. 1.3M training images, 50K validation, 150K testing, 1000 categories. Images resized to 256x256, 224x224 is taken. Linear filters – weights, parameters… , *all neurons in a bank apply the same filter, but as just mentioned, they apply it at different locations in the input image* **…** ***shared-filter*** *architecture - a drastic reduction in the number of parameters relative to a neural net in which all neurons apply different filters.* **…** *The distance, in pixels, between the boundaries of the receptive fields of neighboring neurons in a convolutional bank determines the* ***stride*** *with which the convolution operation is applied. Larger strides imply fewer neurons per bank. …* ***Pooling layers*** *– summarize the activities of local patches … typical: max & average;* ***max-pooling***, ***average-pooling****; overlapping*

pooling; ***pooling stride*** *– cmp. to* ***conv.stride****; local translation invariance.* ***Local response normalization*** *- encourages competition for large activations among neurons belonging to different banks – a form of lateral inhibition found in real neurons. Constants N, α, and β - hyper-parameters, determined using a validation set.* … **Neuron nonlinearities –** max-with-zero (ReLU) f(x) = max(0, x). … ***Objective function****: maximize the multinomial logistic regression objective = minimizing the average across training cases of the cross-entropy between the true label distribution and the model's predicted label distribution* ***Weight initialization*** *- a zero-mean normal distribution with a variance set high enough to produce positive inputs into the neurons in each layer.* ***Training****: p.17* (see the formula) … the avera over the i-th batch of the derivative of the objective with respect to w_i ; *stochastic gradient descent with a batch size of 128 examples & momentum of 0.9. cuda-convnet, NVIDIA GTX 580, CIFAR-10: 90 minutes, ImageNet: 4 days with dropout and two days without.* ***Learning rates****: an equal learning rate for each layer, whose value we determine heuristically as the largest power of ten that produces reductions in the objective function. In practice it is typically of the order 10^{-2} or 10^{-3}: 0.01, 0.001. We reduce the learning rate twice by a factor of ten shortly before terminating training.* ***G. Models for CIFAR-10:*** *without dropout: CNN, 3 conv.layers, pooling after all, 3x3 neighborhood, stride = 2. Max-pooling then average-pooling. Respose normalization layers follow the first two pooling layers, N=9, α = 0.001, β = 0.75. The top-most pooling layer is connected to a 10-unit softmat layer which outputs a probability distribution over class labels. All conv.layers x 64 filter banks and filter size of 5x5 (x number of channels in the preceding layer). The model with dropout: an additional fourth weight layer, locally-connected, but not convolutional; filters in the same bank don't share weights. 16 banks of filters 3x3, 50% dropout. The softmax layer takes its input from this fourth weight layer.* ***Models for ImageNet****: 7 weight layers, 5 convolutional, 2 globally-connected (fully-connected) x 4096. Max pooling after 1,2,5 conv.layers. All pooling summarize 3x3, stride 2. Response-normalization layers follow the first and second pooling layers. The first conv. Layer has 64 filter banks 11x11 filters, stride 4 pixels; seccond conv.layer: 256 filter banks 5x5, this layer takes two inputs. … the output of the last globally-connected layer is fed to 1000-way softmax producing the probabilities.* Conclusion p.18: *" using non-convolutional higher layers with a lot of parameters leads to a big improvement with dropout, but makes things worse without dropout."*

*** Tosh: LRN: Local Response Normalization** – brightness normalization, contrast normalization; it improved the generalization performance by making the features more robust to variations in intensity. *"BN normalizes the features by the mean and variance computed within a batch." (A.Arora, 2020).* **Batch Normalization – BatchNorm** replaced LRN in later CNN architectures, because it was discovered that LRN provided too little improvement. BatchNorm is much more effective than LRN in improving training speed and performance. **BatchNorm** includes two *learnable* parameters **scaling** and **shifting** *for each neuron* (γ, β); it normalizes the activations *across the samples in a mini-batch for each neuron*, while LRN normalizes activations across different *feature maps* (channels, "filter banks") at the same spatial

location over N "adjacent" banks of neurons *("at the same position in the topographic organization").* Unlike the learnable BatchNorm parameters γ and β, LRN is controlled with a set of ***fixed*** hyper-parameters N, α, β.

Tosh: See also other **Normalization methods:**

* **Group Normalization,** Yuxin Wu, Kaiming He, 22.3.2018/ 11.6.2018 https://pytorch.org/docs/stable/generated/torch.nn.GroupNorm.html Facebook AI Research (FAIR) – *GN divides the channels into groups and computes within each group the mean and variance for normalization. GN's computation is independent of batch sizes, and its accuracy is stable in a wide range of batch sizes.* .. solves the problem of BatchNorm having big errors for small batch sizes (2,4, 8 …). GN shows error rate on ImageNet ResNet-50 with batch size = 2 similar to BatchNorm of size 16. *** Group Normalization,** Aman Arora, 9.8.2020, https://amaarora.github.io/posts/2020-08-09-groupnorm.html .. a batch of dimension (N, C, H, W) that needs to be normalized (N = Batch size, C = Number of Channels; H,W: Height and Width of the feature map.

*** Layer Normalization Jimmy Lei Ba, Jamie Ryan Kiros, Geoffrey E. Hinton** 21.7.2016 https://arxiv.org/abs/1607.06450 both at training and test times, can be applied for RNN as well, stabilizing their hidden state dynamics. p.3 *Invariance under weights and data transformation; p.4 Weight norm; weight matrix (re-scaling, re-centering); Dataset re-centering; Single training case re-scaling. p.6: 6.1 Order embeddings of images and language .. layer normalization to order-embeddings model of Vendrov et al. [2016] for learning a* ***joint embedding space of images and sentences.*** *.. modify their publicly available code to incorporate layer normalization which utilizes Theano [Team et al., 2016]. Images and sentences from the Microsoft COCO dataset [Lin et al., 2014] are embedded into a common vector space, where a GRU [Cho et al., 2014] is used to encode sentences and the outputs of a pre-trained VGG ConvNet [Simonyan and Zisserman, 2015] (10-crop)..*

* https://github.com/ivendrov/order-embedding - caption-image retrieval

*** Order-Embeddings of Images and Language Ivan Vendrov, Ryan Kiros, Sanja Fidler, Raquel Urtasun** 19.11.2025/1.3.2016 https://arxiv.org/pdf/1511.06361 University of Toronto – *Hypernymy, textual entailment, and image captioning can be seen as special cases of a single* ***visual-semantic hierarchy*** *over words, sentences, and images. .. explicitly modeling the partial order structure of this hierarchy. Towards this goal, we introduce a general method for learning ordered representations, .. hypernym prediction and image-caption retrieval. p.1 Computer vision and natural language processing are becoming increasingly intertwined. … future, autonomous artificial agents will need to* ***jointly model vision and language*** *in order to parse the visual world and communicate with people. Example: "woman skiing" – skis – entity;*

skiing – woman – person – entity; woman walking her dog – woman walking – person walking – person – entity; dog – entity … p.2 Much of the expressive power of human language comes from abstraction and composition.

*** Instance Normalization: The Missing Ingredient for Fast Stylization, Dmitry Ulyanov, Andrea Vedaldi, Victor Lempitsky,** 27.6.2016/6.11.2017 Computer Vision Group at Skoltech & Yandex Russia, Visual Geometry Group University of Oxford United Kingdom https://arxiv.org/abs/1607.08022 .. *swapping batch normalization with instance normalization … Gatys et al. (2016) introduced a method for transferring a style from an image onto another one, as demonstrated in fig. 1. The stylized image matches simultaneously selected statistics of the style image and of the content image. Both style and content statistics are obtained from a deep convolutional network pre-trained for image classification. The style statistics are extracted from shallower layers and averaged across spatial locations whereas the content statistics are extracted from deeper layers and preserve spatial information. In this manner, the style statistics capture the "texture" of the style image whereas the content statistics capture the "structure" of the content image.* However it is inefficient, using iterative optimization until it matches the desired statistics, taking *several minutes to stylize an image of size 512 × 512.* The new method: comparable quality but generated in real time. Style image x0 to learn a generator network g(x,z). **Interesting:** *While the generator network g is fast, the authors of Ulyanov et al. (2016) observed that learning it from too many training examples yield[s] poorer qualitative results. In particular, a network trained on just 16 example images produced better results than one trained from thousands of those.* .. style loss, transfer elements from a style image to the content image such that the contrast of the stylized image is similar to the contrast of the style image.

* **Image Style Transfer Using Convolutional Neural Networks,** Leon A. Gatys; Alexander S. Ecker; Matthias Bethge https://www.cv-foundation.org/openaccess/content_cvpr_2016/papers/Gatys_Image_Style_Transfer_CVPR_2016_paper.pdf texture transfer, …

*** Image Style Transfer using Convolutional Neural Networks,** Simon Hessner - Supervised by Saquib Sarfraz Computer Vision for Human-Computer Interaction Lab Karlsruhe Institute of Technology https://simonhessner.de/wp-content/uploads/2017/11/Simon_Hessner_seminar_paper_2018_Image_Style_Transfer_using_CNNs.pdf L.Gatys et al.: … *up to 500 iterations to produce visually appealing images; every iteration = a forward and a backward pass; up to 16 seconds for 256x256px images; 215 sec for 1024x1024px on a modern GPU. Pg.5: Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization. Huang and Belongie: Adaptive Instance Normalization (AdaIN) between the encoder and the decoder; style & content. VGG-19 ... pg.9: Approaches: iterative optimization, feed-forward network (Per-style model), AdaIN (Decoder, style & content), Whitening & Coloring Transform (WCT): Decoder (content only). All outors: Titan*

X GPUs, 256x256 images.

* X. Huang, S. Belongie, **Arbitrary style transfer in real-time with adaptive instance normalization**, CoRR, abs/1703.06868 (2017).

Study:

https://pytorch.org/docs/stable/generated/torch.nn.InstanceNorm1d.html
https://pytorch.org/docs/stable/generated/torch.nn.InstanceNorm2d.html

*** LayerNorm …**

https://en.wikipedia.org/wiki/Normalization_(machine_learning)
https://en.wikipedia.org/wiki/Normalization_(machine_learning)#Local_response_normalization

– *1.* ***data*** *normalization, 2.****activation*** *normalization.* 1: feature scaling: same range, mean, variance etc.; min-max normalization, sclae, e.g. [0,1] [-1,1]. https://en.wikipedia.org/wiki/Feature_scaling – **Rescaling** (min-max normalization) x'=(x-min(x))/(max(x)-min(x)) … (cmp: Lerp, linear interpolation in shader languages: HLSL, GLSL etc.); mean normalization: x' = (x-avgx)/(max(x)-min(x)) avgx = average(x) – the mean of the feature vector. Another kind of "means normalization" – standardization – divides by the standard deviation:

*** Standardization, Z-score Normalization**. x' = (x-avgx)/sigma; sigma = σ = standard deviation. ***Robust scaling:*** *using median and interquartile range (IQR) – designed to be robust to outliers. It scales the features using he median and IQR as reference points, instead of the mean and stddev.* x'=(x-Q2(x))/(Q3(x)-Q1(x)); Q1(x), Q2(x), Q3(x) are 3 quartiles (25^{th},50^{th}, 75^{th} percentile) of the features.

*** Unit vector normalization –** each individual data point is regarded as a vector and is divided by its vector norm to obtain x' = x/||x||. Any vector norm: usually L1 and L2. If x = (v1, v2, v3), Lp-normalized = $(v1/(|v1|^p, |v2|^p, |v3|^p)^{1/p}$, $v2/(|v1|^p, |v2|^p, |v3|^p)^{1/p}$, $v3/(|v1|^p, |v2|^p, |v3|^p)^{1/p})$

* https://en.wikipedia.org/wiki/Outlier – *a data point that differs significantly from other observations.*

[**Tosh**: However, the data point is classified or attributed as belonging to the same group, selection, dataset, class etc. *at the time of evaluation-observation*. Thus, a signal should trigger the system that another segmentation/division is required – more groups, subspaces for more appropriate/precise classification, addressing, representation. The signal could be an error, a feature, a buffer underrun or overrun/oversize/undersize/threshold etc.]

*** Criteria for deciding is a data point an outlier**: *Chauvenet's criterion Chauvenet%27s_criterion – around the center of a normal distribution; mean and stddev … , Grubbs's test for outliers Dixon's Q test, ASTM E178: Standard Practice for Dealing With Outlying Observations, Mahalanobis distance and leverage, Subspace and correlation based techniques for high-dimensional numerical data, Pierce's, Tukey's fences; Anomaly detection – distance-based or density-based (**Local Outlier Factor LOF** – local density, where locality is given by k nearest neighbors, Local_outlier_factor), Modified Thompson Tau test,* … What to do with outliers? Retention or Exclusion. Non-normal distributions – "fat tails", Cauchy distribution … Set-membership **uncertainties** (instead of a probability density function); hierarchical Bayes model, or a mixture model

https://en.wikipedia.org/wiki/Unit_of_observation#Data_point – The unit of observation should not be confused with the unit of analysis. *a **unit of observation** is the unit described by the data that one analyzes. A study may treat groups as a unit of observation with a country as the unit of analysis, drawing conclusions on group characteristics from data collected at the national level. ..*
*A **data point** or **observation** is a set of one or more measurements on a single member of the unit of observation.* Level_of_analysis - in social sciences: location, size, or scale of a research target; the context/framework… [**Tosh:** this concept can be applied for every domain: resolution of causality-control and perception, see Theory of Universe and Mind). Unit_of_analysis - *the **entity** that frames what is being looked at in a study, or is the **entity** being studied as a whole; the **‚actor'** or the ‚entity' to be studied". The unit of observation is a subset of the unit of analysis.* Statistical_unit – *a **unit** is a member of a set of entities being studied – a „random variable", e.g. a single person, animal, plant, manufactured item, or country that belongs to a larger collection of such entities being studied.* (experimental, sampling) unit. Unit_of_measurement or unit of measure – a definite magnitude of a quantity .. (e.g. metre for length, gram for weight, second for time).
* Data_type | Observational_error – *the difference between a measured value of a quantity and its unknown true value* | Statistical_parameter - a **parameter** is any quantity of a statistical population that summarizes or describes an aspect of the population, such as a mean or a standard deviation. .. Elementary_event – atomic event or sample point – an event which contains only a single outcome in the sample space – mutually exclusive, collectively exhaustive; appropriate granularity .. Multiple sample spaces. **Atom** (measure theory) – A measurable set with positive measure that contains no subset of smaller positive measure. Outcome_(probability) – a possible result of an experiment or trial.[1] Each possible outcome of a particular experiment is unique, and different outcomes are mutually exclusive (only one outcome will occur on each trial of the experiment). All of the possible outcomes of an experiment form the elements of a sample space..

Probability band – https://en.wikipedia.org/wiki/Confidence_and_prediction_bands

Confidence and prediction bands https://en.wikipedia.org/wiki/Prediction_interval https://en.wikipedia.org/wiki/Novelty_detection **–** *identify an incoming sensory pattern as being hitherto unknown. .. The reverse phenomenon is habituation, i.e., the phenomenon that **known** patterns yield a less marked response.*

Tosh: What is *unknown*, how it is decided? Define & discover degrees, criteria, contexts.

* Winsorizing – limiting extreme values in the statistical data to reduce the effect of possibly spurious outliers. Compare: Trimmed_estimator – *a **trimmed estimator** is an estimator derived from another estimator by excluding some of the extreme values, a process called truncation. This is generally done to obtain a more robust statistic, and the extreme values are considered outliers. Given an estimator, the x% trimmed version is obtained by discarding the x% lowest or highest observations or on both end: it is a statistic on the middle of the data. The median is preserved.* Estimator *– a rule for calculating an estimate of a given quantity based on observed data: thus the rule (the estimator), the quantity of interest (the estimand) and its result (the estimate) are distinguished.*[1] .. the sample mean *is an estimator of the population mean. .. point and interval estimators. The point estimators yield single-valued results. This is in contrast to an interval estimator, where the result would be a range of plausible values.* ***Quantified properties: error, mean squared error MSE, Sampling deviation, Variance, Bias …*** Censoring_(statistics) – out of bounds measurements, it may be known that the value is above or below a threshold, but not how much exactly | Maximum_and_minimum | Truncation_(statistics) | Censoring_(statistics)
Fat-tailed_distribution – Cauchy etc. – https://en.wikipedia.org/wiki/Cauchy_distribution – *the canonical example of a „pathological“ distribution since both its expected value and its variance are undefined. The mean and stddev does not converge, unlike in the normal distribution:https://en.wikipedia.org/wiki/Cauchy_distribution#/media/File:Mean_estimator_consistency.gif | Normal_distribution* – **Gaussian distribution**

*** Network In Network,** Min Lin1,2, Qiang Chen2, Shuicheng Yan2, 2014, National University of Singapore – *a novel deep network structure called "Network In Network"(NIN) to enhance model discriminability for local patches within the receptive field.* …
1x1 Convolution

*** Talented Mr. 1X1: Comprehensive look at 1X1 Convolution in Deep Learning,** Raj Sakthi, 13.1.2020 https://medium.com/analytics-vidhya/talented-mr-1x1-comprehensive-look-at-1x1-convolution-in-deep-learning-f6b355825578 depth of the input matrix – number of channels (e.g. RGB: 3 channels, depth = 3, e.g. 64x64x3); usually the filters have the same depth as input… The 1x1 Conv *reduces the number of channels while introducing non-linearity*. 64x64x3 → 64x64x1. Dimensionality reduction = Cross channel down-sampling. *The Winner of ILSVRC (ImageNet Large Scale Visual Recognition Competition) 2014,* ***GoogleNet, used*** *1X1 convolution layer for* ***dimension reduction*** *"to compute reductions before the expensive 3×3 and 5×5 convolutions" ..* "***Inception Module***" – reduces the computation 10 times. ***Usage 2: Building DEEPER Network ("Bottle-Neck" Layer)*** *2015 ILSVRC Classification winner,* ***ResNet,*** *had least error rate and swept aside the competition by using very deep network using 'Residual connections' and 'Bottle-neck Layer'. .. In their paper, He et all explains (page 6) how a bottle neck layer designed* ***using a sequence of 3 convolutional layers with filters the size of 1X1, 3X3, followed by 1X1 respectively to reduce and restore dimension****. The down-sampling of the input happens in 1X1 layer thus funneling a smaller feature vectors (reduced number of parameters) for the 3X3 conv to work on. Immediately after that 1X1 layer* ***restores the dimensions*** *to match input dimension so identity shortcuts can be directly used. ..* ***identity shortcuts and skip connection …***

*** Networks in Networks and 1x1 Convolutions - Deep Convolutional Models: Case Studies | Coursera,** Andrew Ng https://www.coursera.org/lecture/convolutional-neural-networks/networks-in-networks-and-1x1-convolutions-ZTb8x 6x6x1 → like multiply by a number, scalar; however: 6x6x32 * 1x1x32 → 6x6 = 36 elements * … 28x28x192 CONV 1x1 (1x1x192) → 28x28x32 Srinking the volumes. *Save on computations.* Adding non-linearity …

Tosh: The non-linearity of the activation function enables conditional and logic-like behavior.

*** Deep Residual Learning for Image Recognition,** Kaiming He Xiangyu Zhang Shaoqing, Ren Jian Sun, 10.12.2015, Microsoft Research https://arxiv.org/pdf/1512.03385 The residual learning framework makes the training of substantially deeper networks easier*. .. comprehensive empirical evidence … the residual networks are easier to optimize, and can gain accuracy from considerably increased depth. .. up to 152 layers—8× deeper than VGG nets but still having lower complexity. An ensemble of these residual nets achieves 3.57% error on the ImageNet test set ..*

* **SQUEEZENET: ALEXNET-LEVEL ACCURACY WITH 50X FEWER PARAMETERS AND < 0.5 MB MODEL SIZE***,* 2017, Forrest N. Iandola 1 , Song Han 2 , Matthew W. Moskewicz 1 , Khalid Ashraf 1 , William J. Dally 2 , Kurt Keutzer, DeepScale, UC Berkeley, Stanford University, ICLR 2017 https://openreview.net/pdf?id=S1xh5sYgx

https://github.com/forresti/SqueezeNet v1.0 | https://caffe.berkeleyvision.org/ - ***Caffe*** *(now outdated framework for DL), released in 2014.* https://github.com/BVLC/caffe
https://github.com/forresti/SqueezeNet/tree/master/SqueezeNet_v1.1

Architectural design strategies: *Replace 3x3 filters with 1x1 - 9X fewer. .. Decrease the number of input channels to 3x3 by squeeze layers. Downsample late in the network, so that convolution layers have large activation maps… Fire module: Conv 1x1 → Expand_layer (mix of 1x1 & 3x3 CONV filters). Three tunable dimensions (hyperparameters): s1x1, e1x1, e3x3: the number of filters in a the squeeze layer (all 1x1), the number of 1x1 in the expand layer, thenumber of 3x3 in the expand layer.* 96 filters 7x7, pool_{1,4,8}, 1.72 GFLOPS/image, ImageNet accuracy >=80.3% top-5. V1.1: 64 filters 3x3, pool_{1,3,5}, 0.72 GFLOPS/image, >=80.3% top-5 [inference], 2.4x less computation.
"Comparing SqueezeNet to model compression approaches: AlexNet Top-1 ImageNet Accuracy Top-5 = 240 MB/32bit/57.2%/80.3%, SVD (Denton et al., 2014) 32 bit 240MB → 48MB, 5x, 56.0%/79.4%. AlexNet Network Pruning (Han et al., 2015b) 32 bit 240MB → 27MB 9x 57.2% 80.3% AlexNet Deep Compression (Han et al., 2015a) 5-8 bit 240MB → 6.9MB 35x 57.2% 80.3%. SqueezeNet None 32 bit 4.8MB 50x 57.5% 80.3% SqueezeNet (ours) Deep Compression 8 bit 4.8MB → 0.66MB 363x 57.5% 80.3% SqueezeNet Deep Compression 6 bit 4.8MB → 0.47MB 510x 57.5% 80.3%. Deep Compression (Han et al., 2015b) uses a ***codebook*** *as part of its scheme for quantizing CNN parameters to 6- or 8-bits of precision. … Han et al. developed custom hardware – Efficient Inference Engine (EIE) – that can compute codebook-quantized CNNs more efficiently (Han et al., 2016a) …* SqueezeNet was ported to: ***MXNet*** *(Chen et al., 2015a) by (Haria, 2016)* • ***Chainer*** *(Tokui et al., 2015) port of SqueezeNet: (Bell, 2016)* • ***Keras*** *(Chollet, 2016) port of SqueezeNet: (DT42, 2016)* • ***Torch*** *(Collobert et al., 2011) port of SqueezeNet's Fire Modules: (Waghmare, 2016)* (Some of these libraries are already historical, but Keras and Torch are still in active development) https://pytorch.org/hub/pytorch_vision_squeezenet/
https://en.wikipedia.org/wiki/Comparison_of_deep_learning_software

* Cmp: **Alexnet**: https://en.wikipedia.org/wiki/AlexNet, – *The LeNet-5 (Yann LeCun et al., 1989)[7][8] was trained by supervised learning with backpropagation algorithm, with an architecture that is essentially the same as AlexNet on a small scale.* Max pooling *was used in 1990 for speech processing (essentially a 1-dimensional CNN),[9] and for image processing, was first used in the Cresceptron of 1992. … A deep CNN of (Dan Cireșan et al., 2011) at IDSIA was 60 times faster than an equivalent CPU implementation.[13] Between May 15, 2011, and September 10, 2012, their CNN won four image competitions and achieved SOTA for multiple image databases.[14][15][16] According to the AlexNet paper,[1] Cireșan's earlier net is "somewhat similar." Both were written with CUDA to run on GPU.* ***Computer vision*** *During the 1990–2010 period, neural networks were not better than other ML methods like kernel regression, support vector machines, AdaBoost, structured estimation,[17] among others. For CV in particular, much progress came from manual feature engineering, such as SIFT features, SURF features, HoG features, bags of visual words, etc. It was a minority position*

in computer vision that features can be learned directly from data, a position which became dominant after AlexNet.[18] ***ImageNet*** *- Fei-Fei Li et al., 2007- … 14 million labeled images across 22000 categories, Amazon Mechanial Turk, WordNet hierarchy. ILSVRC - ImageNet Large Scale Visual Recognition Challenge. .. Sutskever convinced Krizhevsky, who could do GPGPU well, to train a CNN on ImageNet, with Hinton serving as principal investigator. So Krizhevsky extended cuda-convnet for multi-GPU training. AlexNet was trained on 2 Nvidia GTX 580 in Krizhevsky's bedroom at his parents' house. Over 2012, Krizhevsky tinkered with the network hyperparameters until it won the ImageNet competition in 2012. Hinton commented that, „Ilya thought we should do it, Alex made it work, and I got the Nobel Prize".[21]*

* **CHM Releases AlexNet Source Code**, Hansen Hsu, Computer History Museum (CHM), 20.3.2025:https://computerhistory.org/blog/chm-releases-alexnet-source-code/ *""It seemed like this unbelievably difficult dataset, but it was clear that if we were to train a large convolutional neural network on this dataset, it must succeed if we just can have the compute," Sutskever told Huang in 2023. The fast computing they needed turned out to be a dual-GPU desktop computer that Krizhevsky worked on in his bedroom at his parents' house."* * Alex Krizhevsky, https://github.com/computerhistory/AlexNet-Source-Code
See also: https://github.com/orgs/computerhistory/repositories | https://github.com/computerhistory/Historical-Source-Code-Quick-Draw-Repository | https://github.com/computerhistory/Historical-Source-Code-MacPaint-Repository | https://github.com/computerhistory/Historical-Source-Code-Apple-II-DOS-Repository
https://computerhistory.org/playlists/source-code/ | https://computerhistory.org/blog/xerox-alto-source-code/ | https://computerhistory.org/blog/microsoft-word-for-windows-1-1a-source-code/

*** Cost estimation of ImageNet**: **Low-High:** $2-3 to $5-7 million (if average $0.05 per image verification, 4 per image, … if $0.01-0.10 per image on Amazon Mechanical Turk 2007-2009 and additionally up to 2012), and ~20% per these payments + infrastructure costs?, development costs, quality control. ? (Claude 3.7 Sonnet reasoning)

*** 2011: DanNet by Dan Ciresan et al. triggers deep CNN revolution,** Jürgen Schmidhuber (Feb 2021, updated 2022)
https://people.idsia.ch/~juergen/DanNet-triggers-deep-CNN-revolution-2011.html

* **History of computer vision contests won by deep CNNs on GPUs**
Jürgen Schmidhuber (2017; updated for DanNet's 10th birthday 2021)

https://people.idsia.ch/~juergen/computer-vision-contests-won-by-gpu-cnns.html
In 2010, IDSIA … "50-fold speedup over CPUs, breaking the long-standing famous MNIST [15c] benchmark record [18a], using pattern distortions [15d]. This really was all about GPUs—no novel NN techniques were necessary, no unsupervised pre-training, only decades-old stuff. .. In 2011, we extended [18b-g] this approach to the convolutional NNs (CNNs) developed by Fukushima (1979), Waibel (1987), LeCun (1989), Weng (1993), and others… 60 times faster than CPU-based CNNs… 2012: First Deep Learner to win medical imaging contest .. Highway networks: first NN with > 100 layers"
* Handwriting Recognition: https://people.idsia.ch/~juergen/handwriting.html
Winning Handwriting Recognition Competitions Through Deep Learning (2009: first really Deep Learners to win official contests). Jürgen Schmidhuber (2009-2013) … https://people.idsia.ch/~juergen/2010-breakthrough-supervised-deep-learning.html
Dalle_Molle_Institute_for_Artificial_Intelligence_Research – **https://www.idsia.usi-supsi.ch/** J.Schmidhuber worked there.

* **Dan Ciresan –** *PhD from "Politehnica" University of Timisoara, Romania. He first worked as a postdoc before becoming a senior researcher at IDSIA, Switzerland. Dr. Ciresan is one of the pioneers of using CUDA for Deep Neural Networks (DNN). His methods have won five international competitions on topics such as classifying traffic signs, recognizing handwritten Chinese characters, segmenting neuronal membranes in electron microscopy images, and detecting mitosis in breast cancer histology images. Currently, he continues his research with DNN at Conndera Research; Founder of Conndera Research in 2016. Romania and Switzerland.* https://people.idsia.ch/~ciresan/

* **ImageNet** https://en.wikipedia.org/wiki/ImageNet *2010: 11 participating teams, the winning team: a linear SVM; a dense grid of HoG and LBP, sparsified by local coordinate coding and pooling. 52.9% in classification accuracy and 71.8% in top-5 accuracy; trained for 4 days on three 8-core machines (dual quad-core 2 GHz Intel Xeon CPU). The second competition in 2011 had fewer teams, with another SVM winning at top-5 error rate 25%.[10] The winning team was XRCE by Florent Perronnin, Jorge Sanchez. A linear SVM, running on quantized[36] Fisher vectors.[37][38], 74.2% top-5 accuracy. In 2012, a deep convolutional neural net called AlexNet achieved 84.7% in top-5 accuracy, a great leap forward.[39] In the next couple of years, top-5 accuracy grew to above 90%. While the 2012 breakthrough „combined pieces that were all there before", the dramatic quantitative improvement marked the start of an industry-wide artificial intelligence boom.*

* **Automatic detection of concrete cracks from images using Adam-SqueezeNet deep learning model,** Lin Wang, Hefei University, Hefei, Anhui Province, 230061, China

19.6.2023 https://www.youtube.com/watch?v=uy27wUnpHQ4 https://www.fracturae.com/index.php/fis/article/view/4216/3845 CrackSN, Adam optimization. **Tosh:** Such cracks can be recognized without DL/CNN. Implement![2]

* **ImageNet-21K Pretraining for the Masses**, Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, Lihi Zelnik-Manor, Alibaba, 22.4.2021/5.8.2021 https://arxiv.org/abs/2104.10972
* https://github.com/Alibaba-MIIL/ImageNet21K a dedicated preprocessing stage, utilization of WordNet hierarchical structure, and a novel training scheme: semantic softmax. ImageNet-21K = 1.3 TB (vs 130 GB for 1K); cleaning invalid classes (e.g. "cow" sometimes as "animal"; validation split, image resizing to 224x224 in the preprocessing stage in order to reduce the dataset to 250 GB; squish-resizing (although this method limits the scale augmentations). Reduced the number of images and classes from 14,197,122 to 12,358,688 images from 11,221 classes – many classes with only 1-10 samples, only 13% reduction of the number of pictures. 50 images per class for standardized validation split. P.4: Hierarchy Example Classes: 0 person, animal, plant, food, artifact 1: domestic animal, basketball court, clothing ... 6: whitetip shark, ortolan, grey kingbird ImageNet-21K-P, has 11 possible hierarchies. Balance the dataset. **Multi-label Training Scheme** – semantic multi labels. P. 6: 11 softmax layers, for the 11 different hierarchies … *avoid extreme multi-tasking (11, 221 uncoupled losses in multi-label training). we have only 11 losses, as the number of softmax layers.* The 21K classes are not mutually exclusive; they are arranged in a hierarchical structure and many are related to each other. This structure allows for a different penalty for misclassification based on the distance within the hierarchy. Higher accuracy with a few % for various architectures and datasets: p.8-9. E.g. Imagenet1K, ViT-B-16: 83.3 → 83.9; Food 251: 74.6 → 76.0; Pascal-VOC: 78,7 – 93.1 (a big jump); MS-COCO: 81.1→82.6 etc.
p.4: **Single-label training scheme:** ImageNet-1K: *On 8xV100 NVIDIA GPU machine, mixed-precision = 40 min/epoch for ResNet50 and TResNet-M architectures (~ 5000 img/sec), total training time of 54 hours. single loss.*

* **The Kinetics Human Action Video Dataset,** Will Kay, Joao Carreira, et al., Google, 19.5.2017 https://arxiv.org/pdf/1705.06950 400 human action classes, with at least 400 video clips for each action. Each clip lasts around 10s and is

[2] See the making of the effects of "Сън в летен дъжд" video (Sleeping/Dream in a Summer Rain). Twenkid Studio – Artificial Mind, Тош, Ефектите в "Сън в летен дъжд" - маскиране на лого в дъждовен нестабилен кадър | Дивия Пловдив 18. 17.9.2020, https://www.youtube.com/watch?v=OWI7HQWyZWg Пс(линии): 2000, 2002 (записки Т.А), SuperCogAlg и др. https://github.com/Twenkid/SuperCogAlg/

taken from a different YouTube videos… https://www.youtube.com/watch?v= + youtube_code
codes: https://github.com/s9xie/Mini-Kinetics-200/blob/master/train_ytid_list.txt
hLxaZBRBvH8 2KRiMcjZue4 etc.

Bag-of-words_model_in_computer_vision – Bag of visual words (BoW), Content based image indexing and retrieval (CBIR); features such as SIFT, vector dim=128; **codewords, codebook generation**; local patches. *One of the notorious disadvantages of BoW is that it ignores the spatial relationships among the patches, which are very important in image representation. ..* Naïve Bayes **..**
probabilistic latent semantic indexing (PLSI)
Probabilistic_latent_semantic_analysis – *Compared to standard latent semantic analysis which stems from linear algebra and downsizes the occurrence tables (usually via a singular value decomposition), probabilistic latent semantic analysis is based on a mixture decomposition derived from a latent class model.*
Structural_equation_modeling * https://en.wikipedia.org/wiki/Latent_class_model **–** *clustering multivariate discrete data. It assumes that the data arise from a mixture of discrete distributions, within each of which the variables are independent.* |
Fisher_kernel – a function that measures the similarity of two objects on the basis of sets of measurements for each object and a statistical model. | Segmentation-based_object_categorization – partitioning an image into multiple regions according to some homogeneity criterion. Graph … Normalized cuts .. Part-based_models - constellation and non-constellation based … mean face, deformable … 35 elements, 100 elements … deviations from a mean face: shape, orientation, gray level … minimization of an error function … template matching
Elastic_matching Graphical_time_warping , Dynamic_time_warping – measuring similarity between two temporal sequences, which may vary in speed. See also: Sequence alignment | Multiple sequence alignment etc: Levenshtein distance, Wagner-Fischer algortihm, Needleman-Wunsch algorithm, Fréchet distance – a measure of similarity between curves that takes into account the location and ordering of the points along the curves .. Nonlinear mixed-effects model ..
Constellation_model – a probabilistic, generative model for category-level object recognition .. attempts to represent an object class by a set of N parts under mutual geometric constraints. Because it considers the geometric relationship between different parts, the constellation model differs significantly from appearance-only, or „bag-of-words“ representation models, which explicitly disregard the location of image features. ..
https://en.wikipedia.org/wiki/Feature_(computer_vision)#Detectors – feature (detection,extraction); Neighborhood_operation; **feature** (image, vector, space, types(**edge**(gradient magnitude, boundary; arbitrary shape, junctions), (**corner**/**interest points**(curvature,…), **blobs**(regions; smoother than edges. Scale: LoG, DoH blob detectors *Laplacian of Gaussian & Difference of Gaussians*), ridges(elongated)). Feature detection: Canny, Sobel, Harris & Stephens/Plessey, SUSAN, Shi & Thomas, Level curve curvature, FAST, Laplacian of Gaussian, Difference of Gaussians, Determinant of Hessian, Hessian strenght feature measure, **MSER**, Principal curvature ridges, Gray-level blobs, * **SIFT** -

Scale-invariant_feature_transform

* **Object recognition from local scale-invariant features,** David G Lowe, 2004 https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf *In the current implementation, each image generates on the order of 1000 SIFT keys, a process that requires less than 1 second of computation time …* p.*2 Related work: Many candidate feature types have been proposed and explored, including line segments [6], groupings of edges [11, 14], and regions [2], among many other proposals. … worked well for certain object classes,* but *.. not detected frequently enough or with sufficient stability to form a basis for reliable recognition*. Key localization *.. identify locations in image scale space that are invariant with respect to image translation, scaling, and rotation, and are minimally affected by noise and small distortions ..* p.3 *"Maxima and minima of this scale-space function are determined by comparing each pixel in the pyramid to its neighbours. First, a pixel is compared to its 8 neighbours at the same level of the pyramid. If it is a maximum or minimum*[3]** at this level, then the closest pixel location is calculated at the next lowest level of the pyramid, taking account of the 1.5 times resampling. If the pixel remains higher (or lower) than this closest pixel and its 8 neighbours, then the test is repeated for the level above. Since most pixels will be eliminated within a few comparisons, the cost of this detection is small and much lower than that of building the pyramid"*. Harris corner detectors for interest points .. canonical orientation.. eigenspace matching, color histograms, and receptive field histograms .. cluttered images, .. – many small local eigen-windows, but expensive search … dense local features (e.g., Schmid & Mohr [19])*

[**Tosh:** Compare SIFT with ANN, CNN. The sequence of convolutions and resolution reduction – downsampling – resembles the construction of a Gaussian pyramids, providing coverage over scale space. Pooling operations and layers (e.g. max-pooling) and nonlinear activations act as implicit mechanism for selecting salient responses, analogous to extrema detection in SIFT.]

*** Object Recognition from Local Scale-Invariant Features, David G. Lowe; Presented by: Yan Fang** (slides), University of Pittsburg https://people.cs.pitt.edu/~kovashka/cs3710_sp15/features_yan.pdf

* **Object Recognition from Local Scale-Invariant Features (SIFT)**, David G. Lowe, Presented by David Lee, March 2006, https://scispace.com/pdf/object-recognition-from-local-scale-invariant-features-5887b1cgwf.pdf

*** David Lowe, Fitting Parameterized Three-Dimensional Models to Images,** 1991, https://faculty.nps.edu/oayakime/ADSC/Scoring%20-%20Lowe%20-%20Fitting%20Parameterized%20Three-Dimensional%20Models%20to%20Images.pdf *"Model-based recognition and motion tracking depends upon the ability to solve for projection and model parameters that will best fit a 3-D model to matching 2-D image features"*

[3] The original quote is "If it is a maxima or minima" (plural), but it seems it is about *the pixel* (singular) 2.5.2026

*** Schmid, C., and R. Mohr, "Local grayvalue invariants for image retrieval," IEEE PAMI, 19, 5 (1997), pp. 530–534.** *https://inria.hal.science/inria-00548358/document*
Geometric vs luminance-based approaches… Geometric: matching, pose estimation, verification. Simplifying the matching process: trees, indexing … differential greyvalue invariances … multi-scale approach .. voting algorithm .. vector of local characteristics; handling of partial visibility and transformations such as rotation and scaling. .. robustness to outliers and tolerance to image noise .. Multi-scaled differential greyvalue invariants .. *Experiments have been conducted for an image database containing 1020 images. .. recognition rate is above 99% for a variety of test images taken under different conditions. 200 paintings, 100 aerial images and 720 images of 3D objects.*

*** Treisman, Anne M., and Nancy G. Kanwisher, "Perceiving visually presented objects: recognition, awareness, and modularity"**, Current Opinion in Neurobiology, 8 (1998), pp. 218–226.
https://web.mit.edu/bcs/nklab/media/pdfs/TreismanKanwisherCurrOpBio98.pdf
Object perception *may involve* ***seeing***, ***recognition, preparation of actions***, *and* ***emotional responses-functions*** *that human brain imaging and neuropsychology suggest are* ***localized*** *separately. Perhaps* ***because*** *of this* ***specialization***, *object perception is remarkably* ***rapid*** *and efficient. Representations of* ***componential structure*** *and* ***interpolation*** *from view-dependent images* ***both*** *play a part in object recognition. Unattended objects may be implicitly registered, but recent experiments suggest that attention is required to* ***bind*** *features,* ***to represent three-dimensional structure,*** *and to mediate awareness. ..* ***6 types*** *of representations* ***… 1: 'object token'****-a conscious viewpoint-dependent representation of the object as currently seen.* ***Second****, as a* ***'structural description'****- a non-visually-conscious object-centered representation from which the object's appearance from other* ***angles*** *and* ***distances*** *can be predicted.* ***Third****, as an* ***'object type'****-a recognition of the object's identity (e.g. a* ***banana****) or* ***membership*** *in one or more stored* ***categories****.* ***Fourth****, a representation based on* ***further knowledge associated with the category*** *(such as the* ***fact*** *that the banana can be* ***peeled*** *and what it will* ***taste*** *like).* ***Fifth****, a representation that includes a specification of its* ***emotional*** *and* ***motivational significance*** *to the observer.* ***Sixth****, an* ***'action-centered description',*** *specifying its "****affordances****" [l], that is, the properties we need in order to program appropriate motor responses to it, such as its* ***location****,* ***size*** *and* ***shape*** *relative to our hands. These* ***different representations*** *are probably formed in an* ***interactive fashion,*** *with* ***prior knowledge facilitating the extraction*** *of likely* ***features*** *and* ***structure****, and* ***vice versa.*** .. Continuity, regularity, … **specialized** modules … *The accumulating evidence for* ***cortical specialization*** *for specific components of visual recognition raises a number of important questions. Does this* ***fine-grained specialization*** *of function arise from* ***experience-dependent self-organizing properties*** *of cortex [98], or are cortical specializations innately specified? .. '****shallow specialization'*** *or a* ***deeper*** *form of modularity in which a small number of functionally specific regions each carries out a qualitatively distinct computation in the service of an evolutionarily or experientially fundamental visual process. ..* ***multiple representations*** *that are* ***extracted in the first***

quarter of a second of viewing a complex visual stimulus. *Both* ***structural descriptions*** *and* ***viewpoint-dependent*** *representations sufficient for discriminating between objects are extracted* ***within about 200ms***. *The phenomena of* ***repetition blindness***, ***attentional blink***, ***attentional masking***, *and* ***inattentional blindness*** *reveal some of the heuristics by which the visual system decides which of these representations to incorporate into the developing stable representation of visual experience. (…) suggesting that extracting shape from shading is a distinct process from extracting shape from edges. .. both structural descriptions and more specific viewpoint-dependent representations are retained in visual memory. .. Experts with extensive encounters with different instances may base their recognition on matching to multiple stored views, giving the impression of invariant representation. .. Gauthier and Tarr [77] gave subjects prolonged training – "greebles"; ..* [face-like 3D-objects, which are also recognized with the fusiform face area (FFA)]. Structural – **geons**; above, below.. *both structural descriptions and more specific viewpoint-dependent representations are retained in visual memory .. extracting shape from shading is a distinct process from extracting shape from edges.* [**Tosh:** suggestions, ideas, considerations for Cognitive Architectures.]

* Gauthier I, Tarr MJ: **Becoming a 'Greeble' expert: exploring mechanisms for face recognition.** Vision Res 1997, 37:1673- 1682.
* *Greeble (Psychology), Wiki4All, 57,1 K Subscribers, 2296 views, 28.03.2022 г.* *https://www.youtube.com/watch?v=9JlE9Y4NxiI*

* **Unraveling Mechanisms for Expert Object Recognition: Bridging Brain Activity and Behavior**. Isabel Gauthier, Vanderbilt University, Michael J. Tarr, 2012. Brown University http://gauthier.psy.vanderbilt.edu/wordpress/wp-content/uploads/2012/03/GaTa_UnravelingMechanisms_02.pdf Holistic processing; 1. Holistic–configural effects, 2. Holistic–inclusive effects, 3. Holistic–contextual effects on recognition. 1: *individual object parts are placed in the context of the other individual parts from the same object.****1 Holistic-configural:*** *Each individual part is better recognized in the original learned configuration than in the context of these same object parts in a new configuration. For instance, the nose of a familiar face .. 2.* ***Holistic-inclusive:*** *individual object parts are better recognized in the context of other individual parts from the same object and instance as compared with the context of parts from other objects and instances 3.* ***Holistic-contextual*** *individual object parts are better recognized in the context of other parts than in isolation.* .. 1 & 2 increase from expertise, while 3 doesn't depend on it…

Tosh: *"The structure of images"* is a seminal paper, related to the "*Scale Spaces*". See also the main volume of the book *"The Prophets of the Thinking Machines…"* – the list with schools of thoughts and comparison to the Theory of Universe and Mind (2001-2004), the hierarchy of Universes at different levels of the Universe computer with reduced resolution and a bigger span etc. See also **Universe and Mind 6.**

*** J.J. Koenderink, The structure of images, Biological Cybernetics, vol. 50, pp. 363-396, 1984** https://scispace.com/pdf/the-structure-of-images-gx3bm3792l.pdf - *"realms" of*

the extrema." p.5. *"For a certain finite range of resolution the blobs can be identified (that is if t is less than the value at which the extremum meets its saddle-point), and in a still more limited range the blob exists in its pure form, unarticulated. For too high a resolution the blob may be difficult to detect because it is articulated with irrelevant smaller detail (e.g. blurring really helps to find objects in scintigrams), whereas for too low a resolution the blobs loose identity (e.g. in a cardioscintigram the left and right ventricles may merge).* ***Details thus have a limited range of resolution in which they can be said to exist.*** *We can define this range from the top of the* ***realm*** *to the next lower top of any included* ***subrealm****.* " **6 Conclusions** .. *"a single .. way to embed an image into a one-parameter family of derived images, with resolution as a parameter: namely by a* ***diffusion process****, or* ***convolution with a family of Gaussian point spread functions****. This result must have seemed obvious to some previous investigators in this field who started out from the family of Gaussians (or rather „DOG's": „difference of Gaussians") or from iterated blurrings (which asymptotically leads to diffusion) in an apparently ad hoc fashion (Marr and Hildreth, 1980). The relation to the diffusion equation appears to have been overlooked previously, although it is this equation that explicitly defines the deep structure of the image*." [**Tosh:** emphasize: mine. Compare to CNNs and the ANN Image generation models with Diffusion.]

* **The structure of images: 1984–2021, Jan Koenderink,** Mar 2021**,** Biological Cybernetics 115(4)**,** DOI: 10.1007/s00422-021-00870-0 *The title was The Structure of Images. It became known as* ***"scale space."***
https://www.researchgate.net/publication/350453877_The_structure_of_images_1984-2021
.. diffusion cannot generate, but only destroy spatial articulations ***.. The diffusion equation serves to connect scale levels.*** *… One may define a vector field whose* ***streamlines*** *capture such connections. It is like the* ***pointers in discrete image pyramids****. The streamlines let one track details over finite scale ranges. .. the diffusion equation is a linear PDE (partial differential equation). .. The evolution of differential invariants over scale Images are trivial* ***fiber bundles.*** *..,* ***differential invariants*** *are like geographical objects such* ***as ruts, ridges, peaks, pits and passes*** *. …* See the referenced literature by Koenderink:
* Koenderink J (1993) **What is a "Feature"?** J Intell Syst 3(1):49–82
* Koenderink JJ, Bouman MA, Bueno de Mesquita AE, Slappendel S(1978**) Perimetry of contrast detection thresholds of moving spatial sine wave patterns**. I–IV. J Opt Soc Am 68(6):845–849, 850–854,854–860, 860–865
* Koenderink J, van Doorn A (1978) **Visual detection of spatial contrast influence of location in the visual field, target extent and illuminance level.** Biol Cybernet 30:157–167
* Koenderink J, van Doorn A (1979a) **The structure of two-dimensional scalar fields with applications to vision.** Biol Cybernet 33:151–158
* Koenderink J, van Doorn A (1979b) **The internal representation of solidshape with respect to vision.** Biol Cybernet 32:211–216
* Koenderink J, van Doorn A (1982a) **Invariant features of contrast detection: an explanation in terms of self-similar detector arrays.** J OptSoc Am 72(1):83–87
* Koenderink J, van Doorn A (1982b) **The shape of smooth objects and the way contours**

end. Perception 11:129–137
* Koenderink J (1984a) **The structure of images. Biol Cybernet** 50:363–370 .. * Koenderink J (1984c) **Geometrical structures determined by the functional order in nervous nets**. Biol Cybernet 50:43–50 * Koenderink J, van Doorn A (1986**) Dynamic Shape**. Biol Cybernet 53:383–396 * Koenderink J, van Doorn A (1987**) Representation of local geometry in the visual system.** Biol Cybernet 55:367–375 … * Koenderink J, Richards W (1988) **Two-dimensional curvature operators.** J Opt Soc Am A 5(7):1136–1141 .. * Koenderink J (1990a) **Solid Shape.** The MIT Press, Cambridge
* Koenderink J (1990b) **The brain a geometry engine**. Psychol Res52:122–127 …
* Koenderink J, van Doorn A (1994) **Two-plus-one-dimensional differential geometry.** Patt Recognit Lett 15:439–443 * Koenderink J, van Doorn A (1999) **The structure of locally orderless images.** Int J Comp Vis 31(2/3):159–168 * Koenderink J, van Doorn A (2000) **Blur and Disorder**. * Koenderink J (2011) **Gestalts and Pictorial Worlds**. Gestalt Theory 33(3/4):289–324 * Koenderink J, van Doorn A (2012) **Gauge Fields in Pictorial Space.** SIAM J IMAG SCI 5(4):1213–1233

* **Distinctive Image Features from Scale-Invariant Keypoints**, David G. Lowe, 5.1.2004, Computer Science Department University of British Columbia Vancouver, B.C., Canada … *stable keypoint locations p.7: Maxima and minima of the difference-of-Gaussian; … orientation histogram …: 36 bins x 10 degrees … peaks – dominant directions .. image gradients, keypoint descriptors; p.20: keypoint descriptor – 128 dimensions; k-d tree* – no benefit for > 10 dim so – Best-Bin-First (BBF) – *the closest neighbor with high probability .. clustering features in pose space using the Hough transform (Hough, 1962; Ballard, 1981; Grimson 1990).* … @Vsy: Study: (& all)

* **Perceptual Organization and Visual Recognition,** David Lowe, 9.1984, PhD Thesis https://apps.dtic.mil/sti/tr/pdf/ADA150826.pdf Gestalt, perceptual groupings; 3-space inferences, viewpoint independent, lightsource independent; object boundaries, edges; curves, lines, segmentation; Possible viewpoint and objects, ranking; well-structured objects. Evidential reasoning. Speed up the search process. Problem of decision. Model-based. Independence assumptions; relative ranking, not absolute probabilities. Dividing line: object & future: various scales. Many sources of evidence: contextual information. Color, texture, enclosure, connectivity, adjacency. Update the expectations. Significance. Threshold of the significance. Invariances from 2D projection to 3D. Formulas for recovery of 3D from 2D, perspective, …

*** SIFT Detector | SIFT Detector** (Part I)**,** **Shree Nayar**, T. C. Chang Professor of Computer Science at Columbia Engineering 3.3.2021, https://www.youtube.com/watch?v=ram-jbLJjFg – a stack of blurred images, Gaussian. Find extrema, say 3x3x3 → Interest points at different scales … thresholds to reduce the number of interest points .. blob-like … Scale-invariance .. Image gradient direction, create a histogram of the directions (the range [0,360 deg] is split in buckets)

* **SIFT Descriptor | SIFT Detector** (Part 2)**,** First Principles of Computer Vision, 2021, 76 K subscribers – continuation, SIFT signatures; principal orientation; Normalized histogram is invariant to rotation, scale, brightness; compare two arrays of data, compute the distance https://www.youtube.com/watch?v=IBcsS8_gPzE https://www.youtube.com/@firstprinciplesofcomputerv3258

*** Overview | Shape from Shading** First Principles of Computer Vision**,** 76K subscribers
1: https://www.youtube.com/watch?v=kZWQ2sBGHuU
2: **Human Perception of Shading | Shape from Shading**, https://www.youtube.com/watch?v=VjtDBSPSomo
Shape from shading is an underconstrained problem: many assumptions have to be made. 1:5x… bumps or holes (convex or concave)? Light is above us; 3:00 – is this a moun or a crater? – assuming the light is coming from above; 4:15 – if light comes from the sid? (higher brightness) – you imagine the lighting first and then decide; global illumination .. Humans assume that there is a singular light coming from a singular direction. 6:35 Strips, boundaries. 7:40 perceptual grouping - the X, bumps as we assume the light direction for the central part of the image; 8:52 Hollow mask of a face lit from above looks like a face lit from below. Ramachandran, 1990 * **Stereographic Projection | Shape from Shading** https://www.youtube.com/watch?v=VLc7BC7GJEI ***** Depth from Defocus, Depth from Focus, … etc. See also **SURF:** Speeded_up_robust_features Summed-area_table – higher precision for the sums (e.g. 8-bit image → 16, 32-bit or float for the sums) Viola%E2%80%93Jones_object_detection_framework – Haar filters, face detection, 2001 … 15 fps at 384x288 pixel images on a Pentium III 700 MHz, 50k parameters; runs on iPaq PDA/Pocket PC at 2 fps. https://en.wikipedia.org/wiki/IPAQ

Phase_congruency – *Phase congruency reflects the behaviour of the image in the frequency domain. It has been noted that edgelike features have many of their* ***frequency components*** *in the* ***same phase****. The concept is similar to coherence, except that it applies to functions of different wavelength. … Phase congruency compares the weighted* ***alignment*** *of the Fourier components of a signal with the* ***sum of the Fourier components****.*

*** GLOH (Gradient Location and Orientation Histogram)**
https://en.wikipedia.org/wiki/GLOH - *a SIFT-like descriptor that considers more spatial regions for the histograms. An intermediate vector is computed from 17 location and 16 orientation bins, for a total of 272-dimensions. Principal components analysis (PCA) is then used to reduce the vector size to 128 (same size as SIFT descriptor vector).*

* **A Performance Evaluation of Local Descriptors**, 10.2005, Krystian Mikolajczyk and Cordelia Schmid – *recall with respect to precision and is carried out for different image transformations. We compare shape context [3], steerable filters [12], PCA-SIFT [19], differential invariants [20], spin images [21], SIFT [26], complex filters [37], moment invariants [43], and cross-correlation for different types of interest regions.*
https://www.robots.ox.ac.uk/~vgg/research/affine/det_eval_files/mikolajczyk_pami2004.pdf

*** SIFT, SURF GLOH descriptors**
https://www.micc.unifi.it/delbimbo/wp-content/uploads/2011/03/slide_corso/A33%20SIFT-GLOH-SURF.pdf Alberto Del Bimbo, 2011 **SIFT** … p.8. Gaussians + Blur + Resample (both resized at different levels and blurred). Subtract – DoG, Difference of Gaussians … U-SURF (Upright version), SURF-128 … *GLOH is a method for local shape description very similar to SIFT introduced by Miko in 2004. Differently from SIFT it employs a log-polar location grid: – 3 bins in **radial direction** – 8 bins in **angular direction** – 16 bins for **Gradient orientation quantization*** .. Dominant direction from the histogram .. **SURF:** much faster than the other feature detectors (2.5 – 4 times than SIFT, according to p.33); *relying **on integral images** for image convolutions − building on the strengths of the leading existing detectors and descriptors (using a Hessian matrix-based measure for the detector, and a distribution-based descriptor) − simplifying these methods to the essential*. **Integral images**; Haar Wavelets; p.34: ***Other SIFT-like Implementations*** • *GIST: Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention, TPAMI 2007*

*** LESH: Head Pose Estimation In Face Recognition Across Pose Scenarios**, VISAPP 2008
* Peter Kovesi, **"Image Features From Phase Congruency".** Videre: A, Journal of Computer Vision Research. MIT Press. Volume 1, Number 3, Summer 1999
See also: https://www.micc.unifi.it/delbimbo/wp-content/uploads/2011/10/slide_corso/A33_affine_region_detectors.pdf

https://web.archive.org/web/20030828084038/http://mitpress.mit.edu/e-journals/videre/001/v13.html
https://web.archive.org/web/20030828180227/http://mitpress.mit.edu/e-journals/videre/001/articles/v1n3001.pdf

*** Real-Time Single-Workstation Obstacle Avoidance Using Only Wide-Field Flow Divergence,** Ted Camus,1 David Coombs,2 Martin Herman,3 Tsai-Hong Hong2, 1999: **HyperSparc 80 MHz** UNIX … Optical flow …
https://web.archive.org/web/20030828180227/http://mitpress.mit.edu/e-journals/videre/001/articles/v1n3002.pdf

• **PCA-SIFT: A More Distinctive Representation for Local Image Descriptors,** CVPR 2004 • Spin image: Sparse Texture Representation Using Affine-Invariant Neighborhoods, CVPR 2003

* **A Sparse Texture Representation Using Affine-Invariant Regions**, Svetlana Lazebnik Cordelia Schmid, Jean Ponce, CVPR 2003
https://slazebni.cs.illinois.edu/publications/cvpr03a.pdf

* **A Sparse Texture Representation Using Local Affine Regions**, Svetlana Lazebnik Cordelia Schmid, Jean Ponce, PAMI 2005 https://slazebni.cs.illinois.edu/publications/pami05.pdf – texton dictionary .. p.3: 1. Extract a sparse set of affine regions 2. Normalize the shape of each elliptical region by transforming it into a circle. 3. Perform clustering on the affine-invariant descriptors to obtain a more compact representation of the distribution of features in each image … signature 4. 4. Compare signatures of different images using the Earth Mover's Distance (EMD)

* Viola, P.; Jones, M. (2001). **"Rapid object detection using a boosted cascade of simple features"**, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001. Vol. 1. IEEE Comput. Soc. doi:10.1109/cvpr.2001.990517. ISBN 0-7695-1272-0. S2CID 2715202. 2.2001 https://www.researchgate.net/publication/3940582_Rapid_Object_Detection_using_a_Boosted_Cascade_of_Simple_Features .. Integral image (fast summation of the intensity over rectangular region only with 4 or 6 etc. operations)… Haar Basis functions; AdaBoost classifiers; detection window; sum of the pixels in the white rectangle from the Haar f. are subtracted from the sum of the white. Mitsubishi Electric Research Laboratories. Compaq CRL … MIT+CMU face detection dataset … [**Tosh:** note that in the 1990s and 2000s the word "database" is used instead of the now common "*dataset*"]

* **Face Databases**, Ralph Gross, 2005 – for recognition, detection and facial expression analysis. https://www.ri.cmu.edu/project/face-detection-databases/ Pose variation PIE database etc.… Hyperspectral images – CMU; illumination variations; Equinox Infrared Face Database .. Max Planck Institute for Biological Cybernetics Face Database: 3D data collected with a Cyberware laser scanner … *Combined MIT/CMU Test Se*t: 180 images in two sets – frontal and non-frontal face detection, tilted test, profile view, … nonface images – from web.

* Viola, Paul; Jones, Michael J. (May 2004). **"Robust Real-Time Face Detection".** International Journal of Computer Vision. 57 (2): 137–154

* **Implementing the Viola-Jones Face Detection Algorithm**, Ole Helvig Jensen, 2008 https://web.archive.org/web/20140513142727/http://etd.dtu.dk/thesis/223656/ep08_93.pdf A short description of other algorithms of the time (see also the Viola-Jones papers).
* **DeepFace –** DL facial recognition system, Facebook, 2014; released 2015. https://en.wikipedia.org/wiki/DeepFace 9-layer ANN, 120 million connection weights, trained on 4 million images of FB users. *"The Facebook Research team has stated that the DeepFace method reaches an accuracy of 97.35% ± 0.25% on Labeled Faces in the Wild*

(LFW) data set where human beings have 97.53%" … four modules: ***2D alignment, 3D alignment, frontalization, and neural network.*** *An image of a face is passed through them in sequence, resulting in a 4096-dimensional feature vector representing the face. .. to identify the face, one can compare it against a list of feature vectors of known faces, and identify the face with the most similar feature vector. DeepFace uses fiducial point detectors based on existing databases to direct the alignment of faces. The facial alignment begins with a 2D alignment, and then continues with 3D alignment and frontalization. That is, DeepFace's process is two steps. First, it corrects the angles of an image so that the face in the photo is looking forward. To accomplish this, it uses a 3-D model of a face. 2D – 6 fiducial points; 3D: 67 alginment points; input: 152x152 RGB, output: dim=4096; in the 2014 paper: also a FC 4030-dim. Layer for 4030-persons classification.*
* https://en.wikipedia.org/wiki/FaceNet – a mapping (also called an embedding) from a set of face images to a 128-dimensional Euclidean space; Triplet_loss (anchor-negative-positive) – *The triplet loss function minimizes the distance between an anchor and a positive, both of which have the same identity, and maximizes the distance between the anchor and a negative of a different identity.* FindFace – 2016, Ntech Lab, Russia https://findface.pro/

Kahan_summation_algorithm |)… | Difference_of_Gaussians – a feature enhancement algorithm that involves the subtraction of one Gaussian blurred version of an original image from another, less blurred version of the original. In the simple case of grayscale images, the blurred images are obtained by convolving the original grayscale images with Gaussian kernels having differing width (standard deviations); it is *an approximation of the Mexican hat kernel function used for the Laplacian of the Gaussian operator. As a feature enhancement algorithm, the difference of Gaussians can be utilized to increase the visibility of edges and other detail present in a digital image.* Morlet_wavelet – a wavelet composed of a complex exponential (carrier) multiplied by a Gaussian window (envelope). This wavelet is closely related to human perception, both hearing[2] and vision.
Pyramid_(image_processing)#Gaussian_pyramid – a type of multi-scale signal representation developed by the computer vision, image processing and signal processing communities, in which a signal or an image is subject to repeated smoothing and subsampling. Pyramid representation is a predecessor to scale-space representation and multiresolution analysis (MRA) or multiscale approximation (MSA) – *is the design method of most of the practically relevant discrete wavelet transforms (DWT) and the justification for the algorithm of the fast wavelet transform (FWT). It was introduced in this context in 1988/89 by Stephane Mallat and Yves Meyer and has predecessors in the microlocal analysis in the theory of differential equations (the ironing method) and the pyramid methods of image processing as introduced in 1981/83 by Peter J. Burt, Edward H. Adelson and James L. Crowley. …* Self-similarity in time, Self-similarity in scale, In the sequence of subspaces …, Regularity, linear hull

Completeness, nested subspaces fill the whole space … Wavelet – wave-like oscillation with an amplitude that begins at zero, increases or decreases, and then returns to zero one or more times. Wavelets are termed a „brief oscillation". …

Gaussian_blur .. Ronchi_ruling – *a constant-interval bar and space square-wave optical target or mask. The design produces a precisely patterned light source by reflection or illumination, or a stop pattern by transmission, with precise uniformity, spatial frequency, sharp edge definition, and high contrast ratio. .. a precise signal for testing resolution, contrast, distortion, aberrations, and diffraction in optical imaging systems.* 1951_USAF_resolution_test_chart Halftone – photographic screening, only black dots on white paper, 19-th century method for printing photography; halftone dots – from a sufficient distance the shades look smooth, continuous. Continuous_tone_image – each color at any point in the image can transition smoothly between shades, rather than being represented by discrete elements such as halftones or pixels. | Hough_transform - lines, circles | Harris_corner_detector | Canny_edge_detector | Sobel_operator | Deformable, parametrized shapes | Active contours (snakes) Active_contour_model – energy minimizing, deformable spline; constraint and image forces that pull it towards object contours and internal forces that resist deformation; the external energy term is to control the fitting of the contour onto the image. Signed_distance_function **(SDF)** or **signed distance field** *~ Level Set Function; see Computer Graphics, 3D objects modeling by the boundaries, ray marching. ..* Active contour model is related to Graph cuts, or max-flow/min-cut – a generic method for minimizing the Markov random field (MRF) energy. Boundary_vector_field | Markov_random_field → Sherrington–Kirkpatrick model Spin_glass#Sherrington%E2%80%93Kirkpatrick_model

* **Image Segmentation using Active Contours (Snake)** Vignesh Gopalakrishnan**,** 10.1.2024, *Level Set Function (LSF) - a binary array, with specific regions marked as -1 to represent an initial estimate of the object's boundary. Curve Evolution: energy computations, curvature analysis, gradient calculations, and LSF updates. The contour evolves to fit the object boundaries better over multiple iterationss; parameters: mu, nu, epsilon, step, number of iterations; energy: elastic, bending, external. Elastic: remain close to the original shape, penalizing deviations. Bending: penalizes curvature, sharp bends or kinks – encourages smoothness. External … Automatic initialization.* https://en.wikipedia.org/wiki/Corner_detection#The_level_curve_curvature_approach https://en.wikipedia.org/wiki/Features_from_accelerated_segment_test (FAST) - corner detection: AST corner detector uses a circle of 16 pixels (a Bresenham circle of radius 3) https://en.wikipedia.org/wiki/Midpoint_circle_algorithm https://en.wikipedia.org/wiki/Blob_detection#The_Laplacian_of_Gaussian … https://en.wikipedia.org/wiki/Ricker_wavelet (Mexican hat wavelet, Marr wavelet) https://en.wikipedia.org/wiki/Scale_space | https://en.wikipedia.org/wiki/Ridge_detection - connector set, valleys, relative critical set … **Wavelet Transform:** https://en.wikipedia.org/wiki/Wavelet_transform - *The fundamental idea of wavelet transforms is that the transformation should allow only changes in time extension, but not shape, imposing a restriction on choosing suitable basis functions.*Wavelet coefficients, wavelet filters, length; JPEG2000 .. *First a wavelet transform is applied. This produces as many coefficients as there are pixels in the image (i.e., there is no compression yet since it is only a transform). These coefficients can then be compressed more easily because the information is statistically concentrated in just a few coefficients. This principle is called*

transform coding. After that, the coefficients are quantized and the quantized values are entropy encoded and/or run length encoded. *Transform_coding | Square-integrable_function* – the integral of the square of the function is finite, < infinity; for wave functions: Fourier, Wavelets, Quantum wave functions – can be described as a sum of a series of orthogonal functions (decomposition of independent components), L^2 norm can be applied, Cauchy sequences converge in that space etc.

* Fischler, M.A.; Elschlager, R.A. (1973). "**The Representation and Matching of Pictorial Structures"**. IEEE Transactions on Computers. C-22: 67–92. doi:10.1109/T-C.1973.223602.
* Yuille, Alan L.; Hallinan, Peter W.; Cohen, David S. (1992). "Feature extraction from faces using deformable templates". International Journal of Computer Vision. 8 (2): 99 doi:10.1007/BF00127169
* Brunelli, R.; Poggio, T. (1993). "Face recognition: Features versus templates". IEEE Transactions on Pattern Analysis and Machine Intelligence. 15 (10): 1042. doi:10.1109/34.254061.

* **Robust wide baseline stereo from maximally stable extremum regions** (PDF), J. Matas; O. Chum; M. Urban & T. Pajdla (2002**).** British Machine Vision Conference. pp. 384–393. https://cmp.felk.cvut.cz/~matas/papers/matas-bmvc02.pdf

MSER, *robust similarity measure for establishing tentative correspondences .. distinguished regions, extremal regions (DR, ER), local invariant descriptors, replace the Mahalanobis distance ...* *Mahalanobis_distance* *Measurements from large regions are either very discriminative (it is very unlikely that two large parts of the image are identical) or completely wrong (e.g. if orientation or depth discontinuity becomes part of the region) .. tentative local correspondances .. The rough epipolar geometry estimated from tentative correspondences is used for guiding the search for further region matches … MSER: Imagine all possible thresholdings of a gray-level image … In many images**, *local binarization is stable over a large range of thresholds in certain regions*** .. *Detection of MSER is also related to thresholding. Every extremal region is a connected component of a thresholded image. However, no global or 'optimal' threshold is sought, all thresholds are tested and the stability of the connected components evaluated.** P.5 *"..small regions are less discriminative, i. e. they are much less likely to be unique."p.6. Rough epipolar geometry: by RANSAC to the centers of gravity of DR.Correspondence of covariance matrices defines an affine transformation up to a rotation: RANSAC again … Next: correlation of the transformed images above a threshold are selected … Another RANSAC – convex hull centers are EG-consistent …* **MSER:** https://en.wikipedia.org/wiki/Maximally_stable_extremal_regions – *"The equation checks for regions that remain stable over a certain number of thresholds. If a region Q[i+delta] is not significantly larger than Q[i-delta], then Q[i] is taken as a maximally stable region.* **Advantages: Invariance** to affine transformation of image intensities; **Covariance** to adjacency preserving (continuous) transformation. **Stability**: only regions whose support is nearly the same over a range of thresholds is selected. **Multi-scale detection** without any

smoothing involved, both fine and large structure is detected. Note, however, that detection of MSERs in a scale pyramid improves repeatability, and number of correspondences across scale changes. **The set of all extremal regions can be enumerated in worst-case** O(n), n – the number of pixels. ***Comparison to other region detectors: (*** Region density, Region size, Viewpoint change, Scale change, Blur, Light change**) ..** *Measurement regions are selected at multiple scales: the size of the actual region, 1.5x, 2x, and 3x scaled convex hull of the region***. ..** almost planar patch, stable invariant description = 'good measurement'; unstable patches, non-planar surfaces or discontinuities = 'corrupted measurements' .. **The robust similarity is computed by:** two images, comparing corresponding regions, nearest to M^{i}_{A} **..** RANSAC … another RANSAC with a more narrow threshold … [see above and the paper]"… **weakness of MSER to blur –** enhanced with Canny edge detection, for extraction of blurred text.

* https://en.wikipedia.org/wiki/Random_sample_consensus **– RANSAC** *an iterative method to estimate parameters of a mathematical model from a set of observed data that contains outliers, when outliers are to be accorded no influence[clarify] on the values of the estimates.* * **Thresholding** https://en.wikipedia.org/wiki/Thresholding_(image_processing) - **global, local - adaptive**; **histogram-shape** based: the peaks, valleys and curvatures of the smoothed histogram are analyzed; **clustering-based**, **entropy-based**: foreground-background regions; cross-entropy between the original and binarized image **…** ***Object-attribute-based*** *methods search for a measure of similarity between the gray-level and the binarized images: fuzzy shape similarity, edge coincidence etc.* ; **Otsu**: Otsu%27s_method – automatic, foreground-background, 1D variant of Fisher's discriminant analysis; a globally optimal k-means on the intensity histogram . *Niblack's Method:[9] .. computes a local threshold for each pixel based on the mean and standard deviation of the pixel's neighborhood. Spatial methods use higher-order probability distribution and/or correlation between pixels***. Choropleth_map | dasymetric technique | Chorochromatic_map** – the division of the areas is by predefined segmentation spaces, e.g. administrative districts, regions, countries etc and the area which is aggregated may vary, e.g. states: Alaska and Washington D.C, or Luxemburg and Australia. Dasymetric_map - *"population density, irrespective of any administrative boundaries, is shown as it is distributed in reality, i.e. by natural spots of concentration and rarefaction." – 1911 by Benjamin Semyonov-Tian-Shansky* (*δασύς dasýs* – dense, density) | Contour_line – ***contour line*** *(also isoline, isopleth, isoquant or isarithm) of a function of two variables is a curve along which the function has a constant value, so that the curve joins points of equal value.*contour: map, interval * https://en.wikipedia.org/wiki/Feature_(computer_vision)#Detectors – feature (detection,extraction); Neighborhood_operation; **feature** (image, vector, space, types(**edge**(gradient magnitude, boundary; arbitrary shape, junctions), (**corner**/**interest points**(curvature,…), **blobs**(regions; smoother than edges. Scale: LoG, DoH blob detectors *Laplacian of Gaussian & Difference of Gaussians*), ridges(elongated)). Feature detection: Canny, Sobel, Harris & Stephens/Plessey, SUSAN, Shi & Thomas, Level curve curvature, FAST, Laplacian of Gaussian, Difference of Gaussians, Determinant of Hessian, Hessian strenght feature measure, **MSER**, Principal curvature ridges, Gray-level blobs, **SIFT** - Scale-invariant_feature_transform (see also **SURF:** Speeded_up_robust_features

Summed-area_table – higher precision for the sums (e.g. 8-bit image → 16, 32-bit or float for the sums) Viola%E2%80%93Jones_object_detection_framework – Haar filters, face detection, 2001 … 15 fps at 384x288 pixel images on a Pentium III 700 MHz, 50k parameters; runs on iPaq PDA/Pocket PC at 2 fps. https://en.wikipedia.org/wiki/IPAQ

Affine invariant features, regions; invariance

Laplacian ∇^2 (nabla^2), Laplacian of Gaussian LoG, Scalar field; Filter; Diffusion .. https://en.wikipedia.org/wiki/Laplace_operator the sum of second partial derivatives of the function with respect to each independent variable. Discrete_Laplace_operator [1 – 2 1]

0	1	0
1	-4	1
0	1	0

* https://docs.opencv.org/3.4/d5/db5/tutorial_laplace_operator.html
* https://cadubentzen.github.io/pdi-ufrn/unit1/laplgauss.html
* https://medium.com/@rajilini/laplacian-of-gaussian-filter-log-for-image-processing-c2d1659d5d2
* A Sparse Texture Representation Using Affine-Invariant Regions, Svetlana Lazebnik, Cordelia Schmid Jean Ponce, 2003 https://slazebni.cs.illinois.edu/publications/cvpr03a.pdf
* A Sparse Texture Representation Using Local Affine Regions Svetlana Lazebnik, 2005, https://slazebni.cs.illinois.edu/publications/pami05.pdf (35 pg.)

Harris affine region detector .. Steerable filters

* Harris_affine_region_detector | Kadir%E2%80%93Brady_saliency_detector

Hessian_matrix – a square matrix of second-order partial derivatives of a scalar-valued function, or scalar field. It describes the local curvature of a function of many variables: $\nabla\nabla$, D^2

* An affine invariant interest point detector, Krystian Mikolajczyk, Cordelia Schmid. 2002. https://inria.hal.science/inria-00548252/document p.9 3.2 Affine invariant interest point multi-scale Harris detector … | Structure_tensor – second-moment matrix; derived from the gradient of a function. It describes the distribution of the gradient in a specified neighborhood around a point and makes the information invariant to the observing coordinates.
* https://www.micc.unifi.it/delbimbo/wp-content/uploads/2011/10/slide_corso/A33_affine_region_detectors.pdf - approximate with ellipses; Gaussian kernel – ellipsoid .. Gaussian scale space, affine normalization, using an iterative affine adaptation algorithm to detect affine invariant regions. … Circular/Polar coordinate system … p.6 Isotropic neighborhoods, related by rotation .. Iterative estimation of localization, scale, neighborhood … maximize the Harris corner measure in the 8-neighborhood…

Tosh: Match at different scales and after a transformation, warping of the circle/ellipse Intensity Extremal Regions … Radial rays, connect the intersections, the region boundary to a curve and approximate with an ellipse. **MSER:** params: Margin – the number of threshold steps where the region is stable; MaxArea, MinArea, MaxVariation … reject also

if a child region is too similar to its parent … MSER – doesn't work well with images with any motion blur. Multi-resolution MSERs. Comparison: MSER is best for viewpoint and scale change and JPEG.

---

* T.Tuytelaars, L.V.Gool. **"Wide Baseline Stereo Matching Based on Local, Affinely Invariant Regions**", BMVC 2000.
* **Filter and Filter Bank Design for Image Texture Recognition,** Trygve Randen, Dr. Ing. thesis, Norwegian University of Science and Technology. Defended November 24, 1997, https://www.ux.uis.no/~tranden/ https://www.ux.uis.no/~tranden/thesis/thesis.pdf p.39(31) 4.1 Prediction error filtering .. Linear prediction error …

* Thematic_map * Map * https://en.wikipedia.org/wiki/Image_segmentation .. * Object_co-segmentation | List of manual image annotation tools: **LabelMe** https://labelme.io/ https://github.com/wkentaro/labelme

* **CVAT** https://github.com/cvat-ai/cvat | https://www.cvat.ai/ | https://github.com/HumanSignal/label-studio (continues https://github.com/HumanSignal/labelImg - now archived) … https://www.basic.ai/blog-post/top-10-best-data-annotation-data-labeling-tools-2024
See **Image Captioning …** https://www.ultralytics.com/blog/get-hands-on-with-google-gemini-2-5-for-computer-vision-tasks https://en.wikipedia.org/wiki/TagLab - used for tagging corals | Range_segmentation | Color_quantization | Optimization_problem – find the best solution from all feasible solutions.
* Connected_space#Connected_components … (path, arc, local) connectedness; disconnected spaces. See also: Text_segmentation – word splitting, intent segmentation; (sentence, topic, morphemes, paragraphs) segmentation.

* **OpenCV: Image Thresholding**
https://docs.opencv.org/3.4/d7/d4d/tutorial_py_thresholding.html |
Jenks_natural_breaks_optimization | Circular_thresholding – related to the histogram, e.g. for segmentation of living cells in microscope images; however:

Tosh: Extend the thresholding transformation to different flows, shapes, heatmaps etc. when sampling with non-linear laws of comparison.

---

* **Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography.** Martin A. Fischler & Robert C. Bolles (June 1981**).** *(PDF). Comm. ACM.* ***24*** *(6): doi:10.1145/358669.358692 .*
S2CID 972888. Archived – ***Location Determination Problem (LDP):*** *given an image depicting a set of landmarks with known locations, determine that point in space from*

*which the image was obtained. .. **fitting a model to experimental data** .. we illustrate its use in scene analysis and automated cartography.*

Todor: Mapping in general. See **SLAM: Simultaneous localization and mapping** in the Robotics etc. part of The Prophets of the Thinking Machines, volume Listove etc. It could be extended to every domain, modality etc.

*** Epipolar geometry:** https://en.wikipedia.org/wiki/Epipolar_geometry – *the geometry of stereo vision. When two cameras view a 3D scene from two distinct positions, there are a number of geometric relations between the 3D points and their projections onto the 2D images that lead to constraints between the image points. Virtual image plane .. Epipolar line, epipolar plane, triangulation; pinhole camera .. 3D reconstruction, 3D reconstruction from multiple images, Collinearity equation, Photogrammetry, Essential matrix, Fundamental matrix, Trifocal tensor, Binocular disparity, 3D scanner …*

* Chen, Huizhong; Tsai, Sam; Schroth, Georg; Chen, David; Grzeszczuk, Radek; Girod, Bernd. **„Robust Text Detection in Natural Images with Edge-enhanced Maximally Stable Extremal Regions“**. Proc. IEEE International Conference on Image Processing 2011... *Text detection has been considered in many recent studies .. two categories: texture-based and connected component (CC)-based.*

* https://en.wikipedia.org/wiki/Connected-component_labeling – ***connected-component analysis (CCA), blob extraction, region labeling, blob discovery, or region extraction** is an algorithmic application of graph theory,*

* **MSER** etc. connected-component blob recognition and segmentation:
Cmp: CogAlg, SuperCogAlg

OCR dataset and competitions ("recently historical" and a starting point)

* http://www.iapr-tc11.org/mediawiki/index.php/ICDAR_2003_Robust_Reading_Competitions
* ICDAR 2003 Robust Reading Competitions - TC11
* https://github.com/xinke-wang/OCRDatasets?tab=readme-ov-file Natural Scene Text, Document Text, Historical Document Text, Video Text, Synthetic Text etc.
* https://rrc.cvc.uab.es/?ch=1
http://www.iapr-tc11.org/mediawiki/index.php?title=ICDAR_2003_Robust_Reading_Competitions
ICDAR 2005 Robust Reading Competitions - TC11
http://www.iapr-tc11.org/mediawiki/index.php/ICDAR_2005_Robust_Reading_Competitions
***** *Netzer, Yuval, Wang, Tao, Coates, Adam, Bissacco, Alessandro, Wu, Bo, and Ng, Andrew Y. **Reading digits in natural images with unsupervised feature learning**. 2011. – Street View House Numbers dataset*, SVHN.

*** DDI-100: Dataset for Text Detection and Recognition,** Ilia Zharikov1,* , Filipp Nikitin1 , Ilia Vasiliev1 and Vladimir Dokholyan1 https://arxiv.org/pdf/1912.11658 Seattle, Washington 98195 … See also Segmentation related topics**: Quad-trees, Octrees, kd-tree, space partitioning, binary space partitioning; raster images, vector images,**

vectorization, image tracing; chain code … Subpaving. Quantization, vector quantization; autocorrelation, Cross-correlation (compare Convolution). Content-Based Image Retrieval (CBIR) – see below. Image texture(co-occurrence matrix/distribution*, Laws Texture Energy Measures: level ,edge, spot, ripple, wave) * also in NLP: https://medium.com/@imamitsehgal/nlp-series-distributional-semantics-co-occurrence-matrix-31283629951e
* Machine Learning 51: **Co-Occurrence Matrix**, Kasper Green Larsen, 1,32 K subs., 11.2022

- https://www.youtube.com/watch?v=V_Dr9KWt1MI
  * https://en.wikipedia.org/wiki/Spatial_database .. – locations, adjacency; geographic … e.g.
  [Tosh: but also all kinds of mapping or other kinds can be transformed into these]:
- *ST_Distance(geometry, geometry) : number*
- *ST_Equals...,ST_Disjoint, ST_Intersects, ST_Touches, ST_Crosses, ST_Overlaps, ST_Contains: boolean*
- *ST_Length(geometry), ST_Area(geometry) : number*
- *ST_Centroid(geometry), ST_Intersection : geometry*
  Cmp: Mihail Bongard, Bongard problems. #bongard Виж съветски изкуствен интелект от 1960-те и 1970-те в Основния том и том Листове. #Listove

*** Hierarchical Clustering**
* https://en.wikipedia.org/wiki/Hierarchical_clustering ... Deciding Linkage Criteria Compactness, connectivity, Single linkage (nearest neighbor), chaining effect, Complete linkage (farthest neighbor), Average linkage (also known as UPGMA—Unweighted Pair Group Method with Arithmetic Mean), Centroid linkage defines cluster distance based on the Euclidean distance between their centroids (mean vectors). dendrograms, Maximum or complete-linkage clustering, Minimum or single-linkage clustering, Unweighted average linkage clustering (or UPGMA), Weighted average linkage clustering (or WPGMA), Centroid linkage clustering, or UPGMC, Median linkage clustering, or WPGMC, Versatile linkage clustering[, Ward linkage,[8] Minimum Increase of Sum of Squares (MISSQ)[9], Minimum Error Sum of Squares (MNSSQ) Minimum Increase in Variance (MIVAR)[9], Minimum Variance (MNVAR), Hausdorff linkage,Minimum Sum Medoid linkage, Minimum Sum Increase Medoid linkage Medoid linkage, Minimum energy clustering, Elbow Method, Silhouette Score, Gap Statistic | Ward‘s method | … Complete-linkage_clustering https://en.wikipedia.org/wiki/K-medoids https://en.wikipedia.org/wiki/Medoid | https://en.wikipedia.org/wiki/Centroid
* See also CogAlg, SuperCogAlg and future work, implementing the framework, explained in *“Chairs, Buildings, Caricatures…”*, or *“AGI Digest 2012”*, T.Arnaudov, 2012

- **Texture networks: Feed-forward synthesis of textures and stylized images**. Ulyanov, D., Lebedev, V., Vedaldi, A., and Lempitsky, V. S. (2016). In Proceedings of the 33nd International Conference on Machine Learning, ICML 2016, New York City, NY, USA, June 19-24, 2016, pages 1349–1357. etc.

* **TILT: Transform Invariant Low-rank Textures**, Zhengdong Zhang · Arvind Ganesh · Xiao Liang · Yi Ma, 2012

https://people.eecs.berkeley.edu/~yima/matrix-rank/Files/TILT_IJCV.pdf
* **TILT: Rectify everything you see,** 19.6.2012: 1589 views
https://www.youtube.com/watch?v=ll9wYna3GIw
[**Todor:** See also the video with **Eric Horvitz** about **Personal Assistants and integrative AI** from 2012]

* **Machine Learning and Intelligence in Our Midst, Microsoft Research**, 28.03.2012 г…. https://youtu.be/Qe0256zAsNU?si=NPN90J40mdclXkX2 5609 views … Notice the *unpopularity* of this brilliant content: just about 5600 views for 13 years – there was no enough "hype" about "AI". The video is published before the "revolution", which began in the end of the same year.]
__* **Traffic: Recognizing objects using hierarchical reference frame transformations**, Zemel, R.S., Mozer, M.C., Hinton, G.E.:, 1989 *"a model that can recognize two-dimensional shapes in an unsegmented image, independent of their orientation, position, and scale."* – transforming feature instances – constraints on the spatial relationships between features of an object; constructing only plausible assignments of image features to object; hierarchical architecture – it handles unsegmented, non-normalized images and a wide range of candidate objects; training on examples in various poses; *The pattern recognition technique known as hierarchical synthesis (Barrow, Ambler and Burstall, 1972), employs a similar architecture*
___* Barrow, H. G., Ambler, A. P., and Burstall, R. M. (1972). **Some techniques for recognising structures in pictures**. In Frontiers of Pattern Recognition. Academic Press, New York, NY.
https://www.sciencedirect.com/science/article/abs/pii/B9780127371405500063?via%3Dihub
____* **A STRUCTURAL METHOD OF SCENE ANALYSIS**, Zdenek Zdrahal, 1981, Czech Technical University, E.E. https://www.ijcai.org/Proceedings/81-2/Papers/018.pdf – *Relational structures are employed as a tool for describing scenes and objects. Descriptive primitives are extracted from the television image together with their interrelations and the recognition is defined as a task of matching this structural description against models.* The "Structural Method…" is applied with the GOALEM robot's visual subsystem.
_____* Kirchmann , B. , Kopecky , P. , and Zdrahal . Z . **"GOALEM from Prague",** in Proc,, IJCAI-77 , Cambridge , MA, August , 1977, 771 – **a robot arm with 6 DOF and an autonomous vision system of two TV cameras**, created at the Technical University of Prague, Dept. of Control; "GOAL-oriented Electrical Manipulator" (the

name is a word play with "*Golem from Prague"*, a legend from Czechia[4]). Camera resolution up to 130x180, but 128x128 or 64x64 are usually processed, 4-bit grayscale (16 levels). Max load: 0.5 kg.
https://www.ijcai.org/Proceedings/77-2/Papers/048.pdf
_____* Zdrahal , Z . "Structural Methods in Pattern Recognition" , CSc* (PhD.) thesis , Czech Technica l University , Prague , 1979 (in Czech)

* [10] K. Hirata and T. Kato. **Query by visual example — content based image retrieval**. In A. Pirotte, C. Delobel, and G. Gottlob, editors, Advances in Database Technology (EDBT'92), pages 56–71, Vienna, Austria, **1992**. 2 pages available at:
https://link.springer.com/chapter/10.1007/BFb0032423

Showing a rough sketch is sufficient to enable retrieving some image data from the system.. The powerful pattern recognition algorithms search for the best match candidates on the pictorial index. Currently, the system can accept a handdrawn rough sketch, a monochrome photo, or a xerographic copy, as well as a full color fair copy as a visual example. ... ART MUSEUM: multimedia database with sense of color and composition p..pon the matter of art * https://en.wikipedia.org/wiki/Content-based_image_retrieval
* https://en.wikipedia.org/wiki/List_of_CBIR_engines (modern: ANN, CLIP, Image generators like DALLE etc.)

* **Fast Multiresolution Image Querying,** Charles E. Jacobs, Adam Finkelstein, David H. Salesin. Department of Computer Science and Engineering, University of Washington, 1995 – 1000-20000 image databases, query with sketches/paintings or low quality scans. Content-based retrieval, image indexing … query by (content, example, similarity), sketch retrieval; Haar wavelet decomposition – signatures, only the most significant information about each image; image querying metric. *128x128 image query on a database of 20,000 images in under ½ second on an SGI Indy R4400 – L1 metric takes over 14 minutes* (~1700 times slower). Previous work: *color histograms [33], texture analysis [12], and shape features like circularity and major-axis orientation of regions in the image [7]. QBIC – IBM – a particular color composition (x% of color 1, y% of color 2, etc.), a particular texture, some shape features, and a rough sketch of dominant edges in the target image.*

* **Multi-label Machine Learning and Its Application to Semantic Scene Classification**, Xipeng Shen, Matthew Boutell, Jiebo Luo, and Christopher Brown, In Proceedings of IS&T/SPIE's Sixteenth Annual Symposium on Electronic Imaging: Science and Technology (EI 2004), San Jose, California, USA, January 2004, pages 188–199.
https://research.csc.ncsu.edu/picture/publications/papers/ei5307-22.pdf

[4] https://www.jpost.com/international/from-the-golem-of-prague-to-artificial-intelligence-650713

- **Learning Multi-label Scene Classification**, Matthew R. Boutell, Jiebo Luo, Xipeng Shen and Christopher M. Brown, in Pattern Recognition, Volume 37, Issue 9, 2004, pages 1757-1771. https://research.csc.ncsu.edu/picture/publications/papers/pr04.pdf
  * **A survey of content-based image retrieval with high-level semantics,** Ying Liu, Dengsheng Zhang, Guojun Lu, Wei-Ying MaAuthors Info & Claims**,** Pattern Recognition, Volume 40, Issue 1**,** Pages 262 – 282**,** 1.1.2007, https://doi.org/10.1016/j.patcog.2006.04.045

* Selected Computer Vision works from 1960s to early 1990s

* **Haralick R.M**. **Using perspective transformation in scene analysis** Comput. Graphics Image Process. (1980)
https://www.sciencedirect.com/science/article/pii/0146664X80900465
https://haralick.org/journals/using_perspective_transformations_1980.pdf
Photogrammetry; Inverse perspective … p.6 - … compute z' from the coordinates of the projection; vanishing points; direction cosines; camera geometry 1/D + 1/f = 1/F; M = I/N = f/D; M/f + 1/f = 1/F; f – distance from the object's coordinates to the lens; D – distance behind the lens, where the image is projected; F – the focal length of the camera; I – the size of the projected image; M - magnification; f = (M+1)*F – *the magnification of the image with respect to the film and the focal length of the camera lens can determine the distance f in front of the lens which the image is located.* .. p.15/205: V.3. The Inverse Projective Transformation

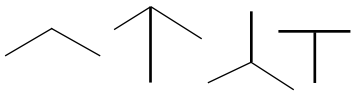
* **Irwin, Camera Models and Machine Perception**, Stanford Artificial Intelligence Project, Memo AIM-121, Computer Science Department, Stanford University, Stanford, Calif., May 1970.
* B. Duda and P. Hart, **Pattern Classification and Scene Analysis**, **Wiley**, New York, 1973.
* J. C. C. Williams, **Simple Photogrammetry,** Academic Press, New York, 1969.
* P. R. Wolf, **Elements of Photogrammetry**, McGraw-Hill, New York, 1974.

***Mackworth, A.K., Interpreting pictures of polyhedral scenes**, Artificial Intelligence 4(2) (1974),121-137

https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=ed165fa39e4ecda494049f459bbe57448101ecc4 .. p.8 Picture region (inner, outer) closure … line, left, end; scene, surface … Programs: POLY; others: Guzman's SEE, Clowes's OBSCENE, Falk's INTERPRET. …

*** Kanade, T**. **Recovery of the 3-D shape of an object from a single view,** Artificial Intelligence (1981) https://www.sciencedirect.com/science/article/pii/000437028190031X
https://www.ri.cmu.edu/pub_files/pub4/kanade_takeo_1981_1/kanade_takeo_1981_1.pdf .. junctions: Guzman [3] ELL, ARROW, FORK; a line can be: a convex edge, a concave edge, an occluding edge; a dictionary… Waltz [18] : cracks, shadows: labeling, filtering; Qualitative analysis/shape recovery … Quantitative: gradient, surface connection graph; **Surface Connection Graph** (SCG); the **Origami world** … Falk: Face adjacency graph, but the recovered shapes are qualitative only. For quantitative shape recovery from a single picture additional assumptions are required. Roberts [15] .. **Labeling procedure:..** 11.1. Summary of the results, p.44: Assumptions: A1: Objects are planar-surfaced with restricted configurations at vertices. A2. The meta-heuristic concerns nonaccidental regularities in the picture; in particular: A2.1: similarity of color edge profiles; A2.2: parallelism of lines; A2.3: skewed symmetry. .. Labeling*: Recover qualitative shapes of*

line drawings together with constraints on surface orientations. Order geometrically possible interpretations. Assuming local symmetry of the surfaces whose projection (picture) is skewed-symmetrical. Equal-length lines, nearly right angles etc. Picture-domain cues: junction types, direction of lines, shapes of regions, parallelisms of lines. p.46 (454): 'Chairs' would have no particular predefined shapes but are usually defined by the descriptions of their functional shapes: say, an L-shaped main structure made of a 'seat' and a 'back', both usually flat; often four legs, attached to the lower corners of the seat; optional two 'arms' attached symmetrically to the main structure, etc. …" the generic models of objects are described in terms of general 3-D shapes and relations (i.e., scene-domain cues). Therefore, in order to access appropriate models for the top-down use of semantic information, we have to first reach certain shape descriptions, either qualitative or quantitative, from the picture in a data-driven manner. Once the appropriate model is found, the general hypothesis-and-test mechanism is activated ... interfacing between the model-driven part and the data-driven part of image understanding. .. **Appendix:** gradient-space lines or surfaces S… v1 = (cos(a), sin(a)), v2=(cos(b), sin(b)) … the changes in z if S has the gradient G… Gradient Space 5. **A theory of the Origami world** … Labeling Cube scene, Folded paper … 6. Mapping Image Regularities into Shape Constraints Parallelism of lines … Skewed symmetry*
7. Quantitative Shape Recovery: Basic Method … Interpretations 9. P. 33 (441) Edge profiles; Color edge profiles (RGBB), from V1 to V2…Gradients G, interpretations of G – graphs on coordinate systems, P, q …); image regularity heuristics. Assigning gradients. Non-cube interpretation. … 10. Shape recovery of the "chair" scene … Surfaces: (S1, s2, s3, s4) mutually constrain the Energy = E. E = E(p1, q1, p2, q2, p3, q3, p4, q4) … five terms, surface interconnection (5 arcs in SCG), 3 terms for the skewed-symmetry heuristic (S1,S2,S3), one term for the parallel-line heuristic (S1, S3 have almost parallel boundaries) … 5.1. Surface-oriented assumptions: No more than 3 planar surfaces of different orientations meet at a vertex, and <= 3 edges of diff. directions are involved at a vertex. "Up-to-3-surface vertices" … Junction dictionaries: L, ARROW, FORK, T. Origami world: also K, X, PSI junctions. Huffman-Clowes dictionary of junctions. Waltz-filtering on junction labels. etc.



*[**Tosh:** see isometric 3D; affine transforms])

Gradient space – **Huffman and Macworth**, surface orientation; types of edges: + convex, - concave, ↑ occluding – "physical meaning" (Huffman'71); dual lines . …

Selected references from Kanade 1981:

* Clowes, M.B., On seeing things, Artificial Intelligence 2(1) (1971), 79–116.
* Falk, G., Interpretation of imperfect line data as a three-dimensional scene, Artificial Intelligence 3 (1972), 101–144.
* Guzman, A., Computer recognition of three-dimensional objects in a visual scene, Tech. Rept. MAC-TR-59, MIT, Cambridge, MA (1968).

* Herskovits, L. and Binford, T.O., On boundary detection, MAC AI Memo 183, MIT, Cambridge, MA (1970).
* Horn, B.K.P., Understanding image intensity, Artificial Intelligence 8(2) (1977), 201–231.
* Huffman, D.A., **Impossible objects as nonsense sentences**, in: Meltzer, B. and Michie, D. (Eds.), Machine Intelligence 6, Edinburgh University Press (1971). https://www.inf.ed.ac.uk/teaching/courses/ics/papers/Huffman-Impossible-objects-nonsense-sentence-MI-1971.pdf – polyhedral language, smooth language .. labels; labeling constraints; forbidden picture subgraphs; complete listing of possible pictures of vertices (pp.4-5/300-303);

Todor: Explore low-poly models in flat shading and wireframe modes to clearly observe triangles, vertices, and edges. Focus especially on voxel models, using general 3D modeling software such as Blender, or more specialized voxel 3D-editors like MagicaVoxel or others. You will notice the set of possible edges, connections, occlusions etc. which are enumerated and studied in this and the other related works from Guzman, Mackworth**,** Kanade and other scholars from 1960s, 1970s and 1980s.

MagicaVoxel: https://ephtracy.github.io/ and MagicaVoxel Viewer. Turn on the borders or select models. In **Rendering mode** the shape, depth and the surface relations and geometry are immediately observable due to real-time rendering at a high frame rate. High-quality rendering with lighting requires significantly more computational resources, but the *qualitative* difference, regarding the "*substantial*" and most *important* information about the *geometry* and the *topology,* compared to the previous and the early stages, becomes progressively smaller. This is evident in the history of computer graphics as well, demonstrated by the evolution of the video games. There was rapid progress in 3D graphics, moving from the early attempts with vector graphics and very low-polygon models with objects consisting of a few triangles or lines, sometimes without hidden-line removal. They were followed by flat shaded models with a few triangles; then by simple texture mapping, progressively higher polygon models and more complex light- and visual effects etc. Since the Playstation 3 or 4 eras, the improvements have become less dramatic and there are 3D PC games from about 2003-2004 which remain aesthetically pleasing[5]*. Another indication for the diminishing returns was that the higher performance ray tracing, offered by Nvidia RTX 2000 series in 2018, failed to attract the expected attention and reaction[6].

*** Todor:** Therefore, the earlier generations of video games were sufficiently good, realistic, complete or complex to withstand the test of time, given the limitation and

[5] Examples of aesthetically timeless old games from Todor's personal list are the first-person shooters Pain Killer and Unreal Tournament 2004, and even older isometric 3D style games such as the real-time strategy StarCraft, 1997, Commandos: Beyond the Call of Duty, 1999.

[6] https://wccftech.com/nvidia-geforce-rtx-20-sales-lower-than-expected-45-percent-revenue-decline/
* NVIDIA GeForce RTX 2080 and GeForce RTX 2070 Sales Lower Than Expected, Gaming Revenue Down 45% – Crypto and Excess Channel Inventory Also To Blame, Hassan Mujtaba, Feb 15, 2019 at 07:43am EST
* Nvidia explains disappointing RTX 2080 / 2070 sales, year-over-year gaming GPU sales down 45% – https://www.notebookcheck.net/Nvidia-explains-disappointing-RTX-2080-2070-sales-year-over-year-gaming-GPU-sales-down-45.407836.0.html

preferences of human perception. It is a matter of taste, but the above could be said for decades older 2D games as well, where the art of the games for arcade machines from the 1980s; for 16-bit consoles like Sega Genesis and SuperNES or even for the 8-bit NES such as Konami's Contra are still appreciated. Notice how small these games were as data and how limited was the hardware, which was able to implement them, compared to modern technology: from just several KB of RAM to 128 KB (arcade machines); from just 16 KB or 40 KB or 128 KB of ROM for 8-bit cartridges to 512 KB or a Megabyte for the most complex arcade games and early 16-bit consoles.

*** Guzman, Adolfo. (1968) Decomposition of a visual scene into three-dimensional bodies.** Proc. of the Fall Joint Computer Conference, pp. 291-304. https://dl.acm.org/doi/pdf/10.1145/1476589.1476631 The program **SEE.**

*We consider visual scenes composed by the optical image of a group of bodies. When such a scene is „seen" by a computer through a film spot scanner, image dissector, or similar device, it can be treated as a two-dimensional array of numbers, or as a function of two variables. At a higher level, a scene could be meaningfully described as a conglomerate of points, lines and surfaces, with properties (coordinates, slopes, ...) attached to them. Still a more sophisticated description could use terms concerning the bodies or objects which compose such a scene, indicating their positions, inter-relations, etc. This paper describes a program which finds bodies in a scene, presumably formed by three-dimensional objects. Some of them may not be completely visible. The picture is presented as a line drawing. When **SEE**—the pretentious name of the program— analyzes the scene TRIAL (see Figure 1 ‚TRIAL'), the results are: (B0DY1.IS:6:2:1) (BODY2.IS:11:12:10) (BODY3.IS:4:9:5:7:3:8:13)* … See early literature:

REFERENCES

1. R. GREENBLATT, J. HOLLOWAY, "Sides 21," Memorandum MAC-M-320, A.I. Memo 101, Project MAC, MIT, August 1966. A program that finds lines using gradient measurements.
2. K.K. PINGLE, "A program to find objects in a picture," Memorandum No. 39, Artificial Intelligence Project, Stanford University, January 1966. It traces around objects in a picture and fits curves to the edges.
3. A. GUZMAN, "Scene analysis using the concept of model," Technical Report No. AFCRL-67-0133, Computer Corporation of America, Cambridge, Massachusetts, January 1967.
4. A. GUZMAN, "A primitive recognizer of figures in a scene," Memorandum MAC-M-342, A.I. Memo 119, Project MAC, MIT, January 1967.
5. A. GUZMAN, "Some aspects of pattern recognition by computer," MS Thesis, Electrical Engineering Department, MIT, February 1967. Also available as a Project MAC Technical Report MAC-TR-37. This memorandum discusses many methods of identifying objects of known forms.
6. A. GUZMAN, H.V. MCINTOSH, "CONVERT," Communications of the ACM 9,8, August 1966, pp. 604-615. Also available as Project MAC Memorandum MAC-M-316, A.I. Memo 99, June 1966.

7. A. GUZMAN, “Decomposition of a visual scene into bodies,” Memorandum MAC-M-357, A.I. Memo 139, Project MAC, MIT, September 1967, unpublished. This memorandum is a more technical description of SEE.
8. R.H. CANADAY, “The description of overlapping figures,” MS Thesis, Electrical Engineering Department, MIT, January 1962.
9. L.G. ROBERTS, **“Machine perception of three-dimensional solids**,” Optical and electrooptical information processing, pp. 159-197, J.T. Tippett et al. eds., MIT Press, 1965.
10. M.L. MINSKY, S.A. PAPERT, “Research on intelligent automata,” Status report II, Project MAC, MIT, September 1967.
11. J. MCCARTHY, “Plans for the Stanford artificial intelligence project,” Stanford University, April 1965.
12. C.A. ROSEN, N.J. NILSSON (eds.), “Application of intelligent automata to reconnaissance,” Third Interim Report, Stanford Research Institute, December 1967.

*** Huffman, D.A**. (1968) **Decision criteria for a class of ‘impossible objects’**. Proceedings of the First Hawaii International Conference on System Sciences. Honolulu

* Penrose, L.S. & Penrose, B. (1958) Impossible objects: a special type of illusion. Brit. J. Psych., 49, 31.

* Huffman, D.A., **Realizable configurations of lines in pictures of polyhedra**, in: Elcock, E.-W. and Michie, D. (Eds.), Machine Intelligence 8, Edinburgh University Press (1977).

* Kanade, T., **A theory of Origami world,** Artificial Intelligence 13(1) (1980), 279–311.https://www.ri.cmu.edu/pub_files/pub3/kanade_takeo_1980_1/kanade_takeo_1980_1.pdf 3D-reconstruction: **up-to-3-surface junctions**: *selecting **surfaces** as basic components, enumeration of the up-to-3-surface junction labels, (3) the use of links to capture the global relationships of regions in the form of a (4) the filtering procedure defined on the spanning angles, and (5) the discussion of relationships among various worlds dealt with in prior work on polyhedral scene analysis. …*

p.28/336: **Huffman-Clowes**: Trihedral junction: Clowes dictionary, **Waltz:** Trihedral junction: dictionary with cracks, shadows, etc. **Mackworth**: Sequential generation of most connected interpretations: Constructive test on coherence rules in the gradient space. **Huffman** +(+’) point test for all the cut sets in the line drawing

Origami: Up-to-3-surface Filtering of dictionary on the SCG: World junction S” Filtering of spanning angles on the SCG.

Surface-oriented world vs. solid-object world – .. *the problem of recovering 3-D configurations from line drawings* may appear as solved due to previous work: *the Huffman-Clowes-Waltz labeling method can identify the cube-like configuration of Fig. 3. ..* [assuming a] *trihedral world in which exactly three planes meet at a corner, and assigns to lines the labels which represent their 3-D meaning, such as + (convex edge), - (concave edge), and ← or → (occluding boundary). …*

* **Kanade, T., Region segmentation: Signal vs. semantics**, Comput. Graphics Image Process. 13 (1980), 279–297.

* Kanade, T. and Kender, J., **Skewed symmetry: Mapping image regularities into shape,** Tech. Rept. CMU-CS-80-133, Carnegie-Mellon University (1980).
* Kender, J., Ph.D. Thesis, Carnegie-Mellon University, Pittsburgh (1980). ..
* Mackworth, A K., Model-driven interpretation in intelligent vision systems, Perception 5 (1976) 349-370.
* Mackworth, A.K., **How to see a simple world: An exegesis of some computer programs for scene analysis,** in: Elcock, E.W. and Michie, D., Eds., Machine Intelligence 8 (Edinburgh University Press, Edinburgh, 1977). …
* Stevens, K.A., **Surface perception from local analysis of texture**, Ph.D. Thesis, Tech. Rept. TR 512, MIT, Cambridge, MA (1979).
* Sugihara, K., **Quantitative analysis of line drawings of polyhedral scenes**, Proc. Fourth International Joint Conference on Pattern Recognition, Kyoto (1978) 771-773.
* Waltz, D., **Generating semantic descriptions from drawings of scenes with shadows**[7], MAC-TR-271, MIT, Cambridge, MA (1972), also reproduced in: Winston, P., Ed., The Psychology of Computer Vision (McGraw-Hill, New York, 1975)
* Woodham, R.J., **A cooperative algorithm for determining surface orientations from a single view**, Proc. Fifth International Joint Conference on Artificial Intelligence, Cambridge, MA (1977)

*** Witkin A.P. Recovering surface shape and orientation from texture,** Artificial Intelligence (1981) https://www.sciencedirect.com/science/article/pii/0004370281900199

* **An iterative image registration technique with an application to stereo vision**, BD Lucas, T Kanade IJCAI'81: 7th international joint conference on Artificial intelligence 2 … *" the spatial intensity gradient of the images to find a good match using a type of Newton-Raphson iteration. Our technique is faster because it examines far fewer potential matches between the images than existing techniques. Furthermore, this registration technique can be generalized to handle rotation, scaling and shearing. We show [that] our technique can be adapted for use in a stereo vision system.*" https://hal.science/hal-03697340/document

1. Baker, H. Harlyn. Edge Based Stereo Correlation. DARPA Image Understanding Workshop, April. 1980. pp. 168-175.
2. Barnea, Daniel I. and Silverman, Harvey F. "A Class of Algorithms for Fast Digital Image Registration." IEEE Transactions on Computers C-21.2 (2.1972), 179- 186.
4. Gennery, Donald B. Stereo-Camera Calibration. DARPA Image Understanding Workshop, November, 1979, pp. 101-107.
5. Marr, D. and Poggio, T. "A Computational Theory of Human Stereo Vision." Proceedings of the Royal Society of London B-204 (1979), 301-328.
6. Moravec, Hans. P. Visual Mapping by a Robot Rover. Sixth International Joint Conference on Artificial Intelligence, Tokyo, August, 1979, pp. 598-600.
7. Nilsson, Nils J. Problem-Solving Methods in Artificial Intelligence. McGraw-Hill, New

[7] Todor: What is "semantic" in this context? A shorter or different representation, a reference, relation to something else, a classification.

York, 1971.

*** Alan K. Mackworth** https://scholar.google.com/citations?user=6Jf5GeYAAAAJ&hl=en

* **Consistency in networks of relations,** AK Mackworth, Artificial intelligence 8 (1), 99-118, 4111, 1977 .. **AI problems as constraint satisfaction**; backtracking; inconsistencies: node, arc, path – may make the algorithm inefficient or lead to cycles; the problem of traversing and labeling of the visited nodes; branch, … *"Waltz [24] … (to convert to our framework, for „junction" read „node", for „label" read „value" and for „branch" read „arc").* **PLANNER:**

A.Mackworth, 1977: As an example, **„Find a large rectangle which is touching a triangle and inside a circle**" could appear in MICRO-PLANNER as
(THPROG (X YZ)
(THGOAL (OBJ $?X RECTANGLE))
(THGOAL (SIZE $?X BIG))
(THGOAL (TOUCHING $? Y $?X))
(THGOAL (OBJ ~ ? Y TRIANGLE))
(THGOAL (OBI $?Z CIRCLE))
(THGOAL (INSIDE $?X $?Z))
(THRETURN $?X))
https://www.cs.ubc.ca/~mack/Publications/AI77.pdf

24. Waltz, D. L., **Generating semantic descriptions from drawings of scenes with shadows**, MAC AI-TR-271, MIT (1972)

* **A theory of multiscale, curvature-based shape representation for planar curves,** F Mokhtarian, AK Mackworth**,** IEEE transactions on pattern analysis and machine intelligence 14 (8), 789-805**,** 1399, 1992 https://www.cs.ubc.ca/~mack/Publications/IEEE-PAMI92.pdf – Continuation of the work from the 1986 paper; **arc length evolution**, resampled curvature scale space (of Africa etc.). Three multiscale representations techniques for planar curves: 1: the regular curvature scale space image, 2: the renormalized curvature scale space image, 3: the resampled curvature scale space image; … for specific applications: uniform noise, nonuniform noise, …

*** Representing knowledge of the visual world,** W Havens, A Mackworth**,** Computer 16 (10), 90-96**,** 110, 10/1983

*** Scale-based description and recognition of planar curves and two-dimensional shapes,** Farzin Mokhtarian, Alan Mackworth**,** 1986/1
https://www.cs.ubc.ca/~mack/Publications/IEEE-PAMI86.pdf *The problem of finding a description, at varying levels of detail, for planar curves and matching two such descriptions is posed and solved in this paper. A number of necessary criteria are imposed on any candidate solution method. Path-based Gaussian smoothing techniques are applied to the curve to find zeros of curvature at varying levels of detail. The result is the* ***"generalized scale space" image of a planar curve which is invariant under rotation, uniform scaling and***

translation of the curve**. These properties make the **scale space** image **suitable for matching.** The matching algorithm is a modification of the uniform cost algorithm and finds the lowest cost match of contours in the scale space images. It is argued that this is preferable to matching in a so-called stable scale of the curve because no such scale may exist for a given curve. This technique is applied to register a Landsat satellite image of the Strait of Georgia, B.C. (manualy corrected for skew) to a map containing the shorelines of an overlapping area. Index Terms-Cartography, computational vision, curve recognition, generalized scale space, map generalization, path-based Gaussian smoothing, remote sensing, shape description, uniform cost algorithm, zeros of curvature. … **II. SOME CRITERIA FOR A RELIABLE REPRESENTATION … a reliable method and representations for curve description and recognition must be: **a)** efficiently computable **b)** .. *essentially invariant under rotation, uniform scaling, and translation of the curve .. **c)** The representation should contain information about the curve at **varying levels of detail**. Moreover, it should be clear from the representation **which features of the curve belong to coarser levels of detail and which features to finer** levels d) The **amount of change** in the **representation** should correspond to the amount of change made to the curve. In other words, a small change to part of the curve should create a small local change in the representation**. e) Arbitrary choices** should not affect the representation. **f)** In case of open curves intersecting the frame, the representation should only change locally with the location of the cutoff points. **g)** The representation should uniquely specify a single curve, otherwise it would be possible to match a curve against a class of curves all of which have the same representation. This criterion only requires uniqueness up to the curve equivalence classes induced by requirement b) above. …*

T: The same criteria can be applied to all kinds of representations at varying scales.

The Knowledge Frontier: **Essays in the Representation of Knowledge**, Book, © 1987 1st edition, Editors: Nick Cercone, Gordon McCalla https://link.springer.com/book/10.1007/978-1-4612-4792-0 … *Knowledge representation is perhaps the most central problem confronting artificial intelligence. Expert systems need knowledge of their domain of expertise in order to function properly. Computer vision systems need to know characteristics of what they are „seeing“ in order to be able to fully interpret scenes. Natural language systems are invaluably aided by knowledge of the subject of the natural language discourse and knowledge of the participants in the discourse. Knowledge can guide learning systems towards better understanding and can aid problem solving systems in creating plans to solve various problems. Applications such as intelligent tutoring, computer-aided VLSI design, game playing, automatic programming, medical reasoning, diagnosis in various domains and speech recognition, to name a few, are all currently experimenting with knowledge-based approaches.*

- **How to see a simple world,** University of British Columbia. Department of Computer Science, ...**,** 59, Alan K Mackworth, 1975
  * **Consistency in networks of relations,** Alan K Mackworth**,** 1977/2/1**,** Journal of Artificial intelligence**,** Volume 8**,** Issue 1**,** Pages 99-118**,** Elsevier **..** *Artificial Intelligence tasks which can be formulated as constraint satisfaction problems, with which this paper is for the most part concerned, are usually solved by backtracking …* ~ examining the

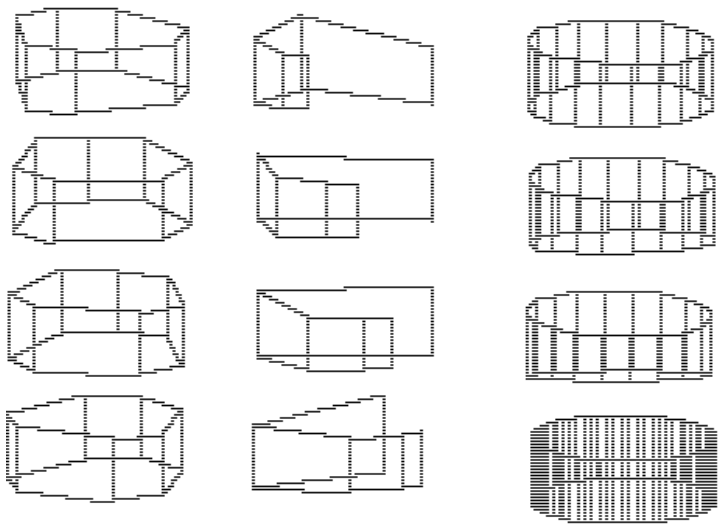
thrashing behavior that nearly always accompanies the process. … *Identifying three of its causes and proposing remedies for them we are led to a class of algorithms which can profitably be used to eliminate local (node, arc and path) inconsistencies before any attempt is made to construct a complete solution. A more general paradigm for attacking these tasks is the alternation of constraint manipulation and case analysis producing an OR problem graph which may be searched in any of the usual ways. Many authors, particularly Montanari and Waltz, have contributed to the development of these ideas; a secondary aim of this paper is to trace that history.* …

- **On the Geometric Interpretation of Image Contours**, Radu Horaud, Michael Brady https://perception.inrialpes.fr/Publications/1988/HB88/1988_Arti_Intel.pdf
  … a computational model for the 3D interpretation of a 2D view based on contour classification and contour interpretation. We concentrate on those contours arising from discontinuities in surface orientation. … A combination of a generic surface description with a model of the image formation process, which derives image contour configurations *that are likely to be interpreted in terms of surface contours*. …
  1.) An image analysis process produces a description in terms of contours and relationships between them; 2.) among the contours, a desired configuration is selected. *3.) The selected contours are combined with constraints available with the image formation process in order to be interpreted in terms of discontinuities in surface orientation; … a dramatic reduction in the number of possible orientations of the associated scene surfaces.* **p.4 (PDF)/336 (pages): Discontinuities:** 1.) in distance from the viewer, 2.) in surface orientation; 3.) changes in surface reflectance; 4.) illumination effects. .. **Generalized cylinders**: a planar cross-section curve … curvilinear abscissa; an eccentricity angle Psi; a spine function; an expansion function h. .. 2.2 Extremal contours $n(s,z) = h(z)(dg/ds)I – h(z)(df/ds)(j)$ … k … curvature … contour generator .. globally symmetric; local symmetry; circular cross-section $r_u(\Theta,z) = h(z)*\cos(\Theta*i) + h(z)*\sin(\Theta*j) + z*k$ … *impossible to determine **the actual shape** of the **cross-section** by observing an **extremal contour** … **Discontinuity contours** … space contour … $M = Area/(Perimeter)^2$* – global shape of a closed planar curve … *scale-invariant number characterizing the curve – m*aximized by the circle. The ellipse can be interpreted as a tilted circle*, a parallelogram as a tilted and rotated square.
  Contour labeling, p.15/347 … extremal, discontinuity … catalogue of image junctions … Curvature-L, Three-tangent, Projection of a vertex neighborhood or edge neighborhood, T-junction .. Contour classification algorithm – graph representation.

* Barnard, S.T. and Pentland, A., Three-dimensional shape from line drawings, in: Proceedings Image Understanding Workshop, Arlington, MA (1983) 282-284.
* Barrow, H.G. and Tenenbaum, J.M., Interpreting line drawings as three-dimensional surfaces, Artificial Intelligence 17 (1981) 75-116.
* Binford, T.O., Visual perception by computer, in: Proceedings IEEE Conference on Systems and Control, Miami, FL (1971).

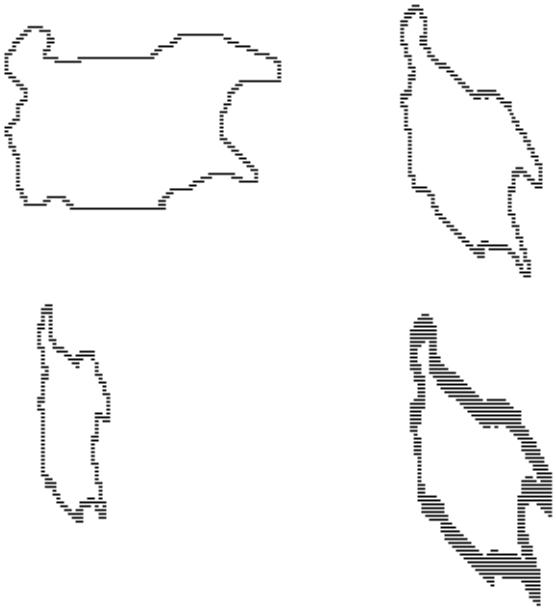
* Brady, M. and Asada, H., Smoothed local symmetries and their implementation, Int. J. Rob. Res. 3 (3) (1984) 36-61.
* Brady, M., Ponce, J., Yuille, A. and Asada, H., Describing surfaces, Compu. Vision Graph Image Process. 32 (1985) 1-28,
* Brady, M. and Yuille, A., An extremum principle for shape from contour, IEEE Trans. Pattern Anal. Mach. Intell. 6 (3) (1984) 288-301
* Koenderink, J., The internal representation of solid shape based on the topological properties of the apparent contour, in: W. Richards and S. Ullmann (Eds.), Image Understanding 1986 (Ablex, Norwood, NJ, 1986).
* Malik, J. Labeling line drawings of curved objects, in: Proceedings Image Understanding Workshop, Miami Beach, FL (1985) 209-218.
* Marr, D., Analysis of occluding contour, Proc. Roy. Soc. London B 197 (1977) 441-475.
* Marr, D., Vision (Freeman, San Francisco, CA, 1982).
* Witkin, A.P., Recovering surface shape and orientation from texture, Artificial Intelligence 17 (1981) 17-45., received May 1987; revised version received January 1988

TA: See some of the versions of **Trigonometric-3D**, **1999,** 3D-graphics on Pravetz-8M (Apple II clone) where the perspective and rotation were simulated by the eccentricity of ellipses and the 3D figures were described in a polar coordinate system as a set of angular coordinates along ellipses. Some weeks later I realized the right formula with the /z for complete three-dimensional coordinates, allowing all kinds of transformations.



Ellipses used as perspective gauges in Trigonometric-3D from 6.1999 on Pravetz-8M

Another way for producing depth and an earlier technique from my technological chronology was called **ART SUPER 98 II** from 1998 – the simple insight was that if **the original step** of one of the coordinates in a contour is changed, then the redrawn image would appear rotated.



Now I would add, that the step can be changed in a more complex way, e.g. by:

1. Computing the center of mass, even just by the average coordinate, and/or finding the extrema points: top, bottom, left, right
2. Choosing a center/cutting point/axis at that row or column;
3. Tracing/rendering from there and while drawing along the

contour, changing the step size in both directions/dimensions with dynamic modifiers, either along all steps, or for example the upper half: multiplying the step (like if it is getting closer to the viewer) while dividing the step on the bottom half; keeping the quotient = 1 in the middle section and diminishing to the "tail".

*** NEVIS'22: A Stream of 100 Tasks Sampled from 30 Years of Computer Vision Research,** Jorg Bornschein, Alexandre Galashov, Ross Hemsley, Amal Rannen-Triki, Yutian Chen, Arslan Chaudhry, Xu Owen He, Arthur Douillard, Massimo Caccia, Qixuang Feng, Jiajun Shen, Sylvestre-Alvise Rebuffi, Kitty Stacpoole, Diego de las Casas, Will Hawkins, Angeliki Lazaridou, Yee Whye Teh, Andrei A. Rusu, Razvan Pascanu, Marc'Aurelio Ranzato, DeepMind, 11.2022/5.2023
https://arxiv.org/abs/2211.11747 *Never-Ending VIsual-Classification Stream …* 106 tasks, 1992-2021; 8M images; p.9: **Domains:** satellite images, faces, textures, objects, shapes, OCR, scenes, medical, quality, counting, Xray; meta-learner.
https://github.com/google-deepmind/dm_nevis

* **CORe50: a new Dataset and Benchmark for continual Object Recognition,** Vincenzo Lomonaco and Davide Maltoni**.** Proceedings of the 1st Annual Conference on Robot Learning, PMLR 78:17-26, 2017.

* **Fine-Grained Continual Learning.** Vincenzo Lomonaco, Davide Maltoni and Lorenzo Pellegrini. Arxiv Pre-print arXiv:1907.03799, 2019.
https://vlomonaco.github.io/core50/

*** Andriy Mnih - Андрей Мних** https://www.cs.toronto.edu/~amnih/ |

An early contributor to the field of deep learning in the mid-to-late 2000s.

Interests: Latent variable models, Variational inference, Monte Carlo gradient estimation. Representation learning.

* Improving a Statistical Language Model Through Non-linear Prediction, Andriy Mnih, Zhang Yuecheng, and Geoffrey Hinton, Neurocomputing, 72:7-9, 2009

* Restricted Boltzmann Machines for Collaborative Filtering, Ruslan Salakhutdinov, Andriy Mnih, and Geoffrey Hinton, ICML 2007

* **Visualizing Similarity Data with a Mixture of Maps**, James Cook, Ilya Sutskever, Andriy Mnih, and Geoffrey Hinton, AISTATS 2007
* https://www.cs.toronto.edu/~amnih/papers/sne_am.pdf
Aspect maps, mixture of different types of similarity; pairwise similarities; multi-dimensional scaling (MDS), principal component analysis (PCA); local linear embedding (LLE), maximum variance unfolding, stochastic neighbour embedding (SNE); local distances (strong similarities)

* Learning Nonlinear Constraints with Contrastive Backpropagation, Andriy Mnih and Geoffrey Hinton, International Joint Conference on Neural Networks 2005 (IJCNN 2005) [bibtex]

* **Wormholes Improve Contrastive Divergence**, Geoffrey Hinton, Max Welling, and Andriy Mnih, NIPS 2003:
* https://www.cs.utoronto.ca/~hinton/absps/worm.pdf
Catastrophic divergence with MCMC is when there are multiple modes in the distribution, several high probability regions, separated by low probability density regions (several local maxima with a high density in the PDF and low density areas in between); a bigger range of the step of search for the Markov Chain Monte Carlo is required in order to escape, the wormholes are the short-cuts between these modes/maxima - *mode-hopping … One way to model the density of high-dimensional data is to use a set of parameters, Θ to deterministically assign an energy, E(x|Θ) to each possible datavector, x … The way in which the data distribution gets distorted in the first few steps of the Markov chain provides enough information about how the model differs from reality to allow the parameters of the model to be improved by lowering the energy of the data and raising the energy of the "confabulations" produced by a few steps of the Markov chain" …*
Notes: contrastive divergence: training RBM (energy-based) – start with the input data, single step Gibbs sampling with the hidden units, then sampling the visible units based on the hidden and compare the reconstruction: the difference between the reconstructed distribution is used to update the model's parameters; the energy of the data points is made lower. MCMC samples many steps of the probability distribution until equilibrium, better for multiple modes. Maximum Likelihood Estimation (MLE) – computes the log-likelihood function, Sum(all possible configurations). … Hamiltonian Monte Carlo (Hybrid MC) - Metropolis–Hastings algorithm with a time-reversible & volume-preserving

numerical integrator – leapfrog integrator to propose a move to a new point in the state space… https://en.wikipedia.org/wiki/Hamiltonian_Monte_Carlo
Cmp: Gaussian random walk: https://en.wikipedia.org/wiki/Random_walk#Gaussian_random_walk

* Andriy Mnih is Volodymyr Mnih's brother. See also: Yosha Bengio and Samy Bengio: https://scholar.google.com/citations?user=Vs-MdPcAAAAJ&hl=en

* Continual Learning

* **Yanislav Donchev (Yani Donchev) –** a Bulgarian researcher
* Янислав Дончев - българин

A. Douillard, Q. Feng, A. Rusu, R. Chhaparia, Y. Donchev, A. Kuncoro, M. Ranzato, A. Szlam, J. Shen **"DiLoCo: Distributed Low-Communication Training of Language Models**". arXiv 2311.08105 2023, 11.2023/9.2024, Google DeepMind https://arxiv.org/abs/2311.08105

*** DiPaCo: Distributed Path Composition,** Arthur Douillard, …, Yani Donchev, … 3.2024

*** Scaling Instructable Agents Across Many Simulated Worlds, SIMA Team**: M.Raad, **…** Yani Donchev, …
* A. Fisch, A. Rannen-Triki, R. Pascanu, J. Bornschein, A. Lazaridou, E. Gribovskaya, M. Ranzato „*Towards robust and efficient continual language learning*"**.** arXiv 2307.05741 2023 https://arxiv.org/abs/2307.05741

*** Multi-scale Transformer Language Models,** Sandeep Subramanian, Ronan Collobert, Marc'Aurelio Ranzato, Y-Lan Boureau, 2020 https://arxiv.org/pdf/2005.00581 … Multi-headed attention is a generalization of dot-product attention (Bahdanau et al., 2014; Luong et al., 2015) where a score is computed between a query Q and key K for different learned projections .. SampleRNN (Mehri et al., 2016), a *multi-scale* recurrent architecture for generating audio. *Downsampled representations at various scales: average pooling or strided convolutions, of both.* Each scale – own context window. … p.7 "a sequence of 512 tokens, shuffle the first 256; *if the shuffled context is more than 50 tokens away,* a very little diff. of the likelihood in the predicted tokens comp. with the correct order, i.e. *the model may not be using word order beyond that distance.* ...the representation ат coarser scales: on shorter sequences, is combined with the repr. of the finest scale … Future work: multi-scale architectures + adaptive attention head spans. …

* **Learning Multiscale Transformer Models for Sequence Generation,** Bei Li ,Tong Zheng Yi Jing ,Chengbo Jiao, Tong Xiao, Jingbo Zhu, 2022 – Universal MultiScale Transformer (UMST); transformer developments: convolutional self-attention, multiscale self-attention, multi-granularity self-attention; placing convolutions in sequence or parallelly in self-attention; the concept of scale for NLP: sub-word, word, phrase; leverage word boundaries; inter-individual, intra-group and inter-group correlations between scales … 2. Preliminary … encoder-decoder paradigm (Sutskever et al., 2014) & transformer (Vaswani et al. 2017). Pre-Norm Transformer

* **MSTR: Multi-Scale Transformer for End-to-End Human-Object Interaction Detection**, B.Kim et al., https://openaccess.thecvf.com/content/CVPR2022/papers/Kim_MSTR_Multi-Scale_Transformer_for_End-to-End_Human-Object_Interaction_Detection_CVPR_2022_paper.pdf *Human-Object Interaction (HOI) detection is the task of identifying a set of ⟨human, object, interaction⟩ triplets from an image. Dual-Entity attention and Entity-conditioned Context attention; Deformable attention;* multi-scale feature maps *vs* single-scale; bounding boxes: human (the subject), object (target), obj. class, interaction type; reference points for interaction; MSTR encoder: CNN backbone

* **SCALEFORMER: ITERATIVE MULTI-SCALE REFINING TRANSFORMERS FOR TIME SERIES FORECASTING**, Mohammad Amin Shabani, Simon Fraser University, Canada, Amir Abdi, Lili Meng, Tristan Sylvain, Borealis AI, Canada, 2023 Prev.work: FEDformer, Autoformer, Informer, Reformer, Performer; explicit scale-awareness; cross-scale feature relationships; seasonal components, structural prior; cross-scale normalization. *Time-series forecasting:* weather, inventory planning, astronomy, economy and financial; seasonal trends. Early models: ARIMA, exponential smoothing, RNN, temporal CNN (CTN), time-series Transformers. Yformer – Y-shaped, multi-resolution embeddings. Autoformer: cross-correlation attention – operate at the level of subsequences; FEDformer – frequency transform to decompose the sequence into multiple frequency domain modes. Look-back and horizon windows for the forecast. *The goal of the forecasting task is to predict the horizon window X(H) given the look-back window X(L)* . Datasets: Electricity Consuming Load (ECL), Traffic,Weather, Exchange-Rate, National Illness (ILI); *Framework: Given an input time-series X(L), iteratively apply the same neural module mutliple times at different temporal scales.*

* Topics Cluster: Audio: Speech Synthesis, Audio Generation, Speech Recognition; 1980s to 2010s

* **SampleRNN: An Unconditional End-to-End Neural Audio Generation Model**, Soroush Mehri, Kundan Kumar, Ishaan Gulrajani, Rithesh Kumar, Shubham Jain, Jose Sotelo, Aaron Courville, Yoshua Bengio, 12.2016/2.2017 https://arxiv.org/abs/1612.07837 *„a hierarchy of modules, each operating at a different temporal resolution. The lowest module processes individual samples, and each higher module operates on an increasingly longer timescale and a lower temporal resolution. Each module conditions the module below it, with the lowest module outputting sample-level predictions. The entire hierarchy is trained jointly end-to-end by backpropagation.“;* Datasets: *Blizzard Prahallad et al. (2013)* for speech synthesis, 315 h … Onomatopoeia - *6738 sequences of 3.5 hours …Music dataset:* all 32 Beethoven's piano sonatas publicly available on archive.org …*;* related to WaveNet

* A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. Lang. Phoneme recognition using time-delay neural networks. IEEE Acoustics Speech and Signal Proc., 37:328–339, 1989
* Jurgen Schmidhuber. Learning complex, extended sequences using the principle of history compression. Neural Computation, 4(2):234–242, 1992.
* Salah El Hihi and Yoshua Bengio. Hierarchical recurrent neural networks for long-term dependencies. In NIPS, volume 400, pp. 409. Citeseer, 1995.
* Yoshua Bengio and Samy Bengio. Modeling high-dimensional discrete data with multi-layer neural networks. In NIPS, volume 99, pp. 400–406, 1999
* Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. Neural Computation, 9(8): 1735–1780, 1997. (…)
* A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 18(5-6):602-610, 2005.
* Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. 2014, arXiv:1412.3555, 2014 (GRU)

* Dahl, George E., Yu, Dong, Deng, Li, and Acero, Alex. **Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition**. IEEE Transactions on Audio, Speech & Language Processing, 20 (1):30–42, 2012. https://www.cs.toronto.edu/~gdahl/papers/DBN4LVCSR-TransASLP.pdf
Automatic speech recognition (ASR); Gaussian mixture model (GMM); recent progress: maximum mutual information estimation (MMI), minimum classification error training (MCE), minimum phone error training (MPE), large-margin estimation etc., large-margin Hidden Markov models, large-margin MCE, boosted

MMI, conditional random fields (CRF), hidden CRFs and segmental CRFs, however these approaches had not yet achieved human-level accuracy. … senones - tied triphone states … A hybrid approach HMM & ANN – late 1980s – early 1990s; *"Around 1999, … a shift from using neural nets to predict phonetic states to using neural nets to augment features for later use in a conventional GMM-HMM recognizer"*; in this work, instead, they try to replace the traditional shallow NN *with deeper, pre-trained NN by using senones – tied triphone states of a GMM-HMM tri-phone model as the output units of the NN, in line with state-of-the-art HMM systems.* (…) @Вси: прдлж @Vsy: extnd

* M. Hwang and X. Huang**, "Shared-distribution hidden Markov models for speech recognition**," 28.4.1991, Carnegie Mellon University/IEEE Trans. Speech Audio Process., vol. 1, no., 4, pp. 414–420, Jan. 1993 – https://www.researchgate.net/publication/3333302_Shared-Distribution_Hidden_Markov_Models_for_Speech_Recognition
Left and right phonetic contexts; *7500 context-dependent triphones (CDT) in the DARPA Resource Management (RM) task … if each triphone is represented by a discrete Hidden Markov models **HMM**, there will be several millions of parameters to be estimated. To estimate the increased number of free parameters, more training data are generally needed. How to estimate a huge amount of parameters with only limited training data? smooth parameters, deleted interpolation, heuristic interpolation … interpolation weights can be determined according to the ability of predicting unseen data or a function of training tokens: diphone or context-independent phoneme model, co-ocurrence matrice, layers of intermediate clustered phonetic model by a decision tree; parameter sharing; semi-continuous (tied-mixture) HMM (SCHMM) … Training procedures are based on the forward-backward algorithm. P.12. table with phones and triphones: AE, EH, IH, IY, UH, AH, AX, IX, AA, AO, UW, AW, AY, EY, OW, OY, L, R, W, Y, EN, ER, M, N, NG, CH, JH, B, D, DH, G, K, P, T, F, S, SH, TH, V, Z, HH, SIL, DD, PD, TD, KD, DX, TS … to triphone #, to cluster # … 3500, 4500, and 5500 shared distributions … 7549 models for each context-dependent triphone*

* **Between-Word Coarticulation Modeling for Continuous Speech Recognition**. Hwang, M., Hon, H., and Lee, K. Technical Report, Carnegie Mellon University, April 1989
***Context-Dependent Modeling for Acoustic-Phonetic Recognition of Continuous Speech**. Schwartz, R., Chow, Y., Kimball, O., Roucos, S., Krasner, M., and Makhoul, J; in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1985, pp. 1205–1208
***Robust Smoothing Methods for Discrete Hidden Markov Models**. Schwartz, R., Kimball, O., Kubala, F., Feng, M., Chow, Y., C., B., and J., M. in: IEEE International Conference on Acoustics, Speech, and Signal Processing. 1989, pp. 548–551
*** Hidden Markov Models for Speech Recognition.** Huang, X., Ariki, Y., and Jack, M. Edinburgh University Press, Edinburgh, U.K., 1990

*** An Overview of the SPHINX-II Speech Recognition System,** Xuedong Huang, Fileno Alleva, Mei-Yuh Hwang, and Ronald Rosenfeld, 1993 – https://aclanthology.org/H93-1016.pdf
Recently SPHINX-II achieved the lowest error rate in the November 1992 DARPA evaluations. For 5000-word, speaker-independent, continuous, speech recognition, the error rate was *reduced to 5%. the testing set has 330 utterances collected from 8 new speakers … "conventional backoff trigram model"* [P(w3| w1 ,w2); backoff – if the trigram wasn't seen in the training data – "backing off" to a bigram or unigram P(w3|w2), P(w3) (Probability of the word (token)). Smoothing – assign small P to unseen n-grams. Trigram Prob. P(w3| w1 ,w2) = count(w1, w2, w3)/count(w1,w2) unigram P(w3) = count(w3)/count(all_words). Additive Laplace smoothing: a small constant (e.g. 1) to each count. Kneser-Ney smoothing: how frequently a word appears in novel contexts. A senone – a basic subphonetic unit, distinct acoustic class – finer grained than phonemes, context-dependent; produced by clustering the state dependent output distributions; classification with a decision-tree, *asking questions in a hierarchical manner.* HMM - decode the most likely sequence of hidden states – sequence of senones, given the acoustic data, which is explicit.

***Sphinx-II System diagram**: Testing Data → Multipass Search(Feature Codebook, Lexicon, HMM Senone, Language Model & Weight) → Reestimation → MFCC Normalized Features and Quantization, Senonic Semi-Continuous HMM, Unified Stochastic Engine } Training Data*

Feature extraction: 12 LPC cepstrum coefficients … HMMs assume each frame is independent of the past [the Markov property] … SPHINX-I: 3 codebooks: 1. 12 x LPC (Linear predictive coding) coefficients, 2. 12 differenced LPC cepstrum coefficients (40 msec difference), 3. Power and differenced power (40 ms) … New measures for spectral dynamics, second order differential cepstrum and power $\Delta\Delta x_t(k)$ & third order … t is in units of 10 ms; incorporating both 40 ms & 80 ms differenced cepstrum … mel-frequency cepstral coefficients (MFCC) … the final configuration = 51 features in 4 codebooks x 256 entries… the new feature set reduces the errors by > 25% … Semi-continuous HMM (SCHMM), Shared mixtures, … Viterbi beam search (VBS), stack decoding … fast-match, N-best paradigm – simple acoustic & language models, then a multi-pass rescoring. … 3 search phases: left-to-right VBS: word end times, between-words model; & bigram language model; right-to-left VBS → word beginning times … A* search combines the previous two with a long distance language model … A* Stack search, "theory: partial theory, one word extension, time, two scores …

*** From CMU Sphinx-II to Whisper — Making Speech Recognition Usable**, Chapter, pp 481–508. Automatic Speech and Speaker Recognition, X. Huang, A. Acero, F. Alleva, M. Hwang, L. Jiang & M. Mahajan, 1996 –

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/02/tr-94-20.doc – *For 5000 and 20,000-word speaker-independent continuous speech recognition; error: reduced to 5% and 13% … the lowest error rate among all of the systems tested in the November 1992 ARPA spoken language technology evaluationsWindows … One of the most important contributions: the availability of large amounts of training data. "In the Sphinx-II system, for acoustic model training: 7200 utterances of read Wall Street Journal (WSJ) text, 84 speakers (half male and half female) language model training: 45-million words of text by the WSJ. Recently 35,000 utterances of read speech and 400-million words of newspaper text.*" … *"For typical Windows® Command-and-Control applications (less than 1,000 words), Whisper provides a software only solution on PCs equipped with 486DX microprocessors. For 5,000 to 60,000 word continuous speech dictation, it requires an Intel Pentium® or MIPS R4400 class microprocessor. For command-and-control: Context Free Grammar language model; for dictation: a statistical trigram language model."* Multi-pass search procedure … *SPHINX-II requires extraordinary working memory and high-end working station* – currently requirements for low-cost PCs. Always some speakers have a different *dialect, accent, culture background, or vocal tract shape*, but Sphinx-II lacks online adaptation and noise rejection mechanism (phone ring, cough and out-of-vocabulary words OOV). For a 20000-word dictation system, on average more than 3% of the words in a free-style test set are not covered by the dictionary; for vocabulary of 64,000 words, OOV > 1.5%. Error rate 20K words > 9% (excluding OOV*. The Nov 1992 ARPA stress test evaluation: both spontaneous speech with many OOV words & several different microphones*: *with > 20000 utterances in the training set and a noise normalization component, reduction only from 12.8→12.8% and 18% for the stress test*.
Sphinx-II: a lot of memory for the acoustic model. **Whisper**: 20 times less RAM, 5 x faster. … Uncompressed acoustic output probabilities: Codeword 1, Senone 1: 0.02; 0.28; 0.035; 0.0 … Codeword 256: 0.0 ; Senone 2: .02, 0.3 … Senone 3: 0.0, 0.0, 0.035 Senone 4: 0.0 … Senone 5: 0.1 Senone 6: 0.0. The sum of each column = 1.0 (senone-dependent output probability distribution); RLE compression for rows (run-length encoding, repetitive values = value & number of repetitions): Codeword 1 = <0.020,3>, 0.0, 0.1, ..., 0.0, Codeword 2 = 0.28, 0.30, 0.020, <0.0,257> … lossless compression > 35%; context-free grammar (CFG), non-terminals to terminals; disallow left recursion … Working size memory: 800 KB; **Noise rejection,** *the whole utterance, and then classifies each component as "accepted", "need clarification through dialogue" or "rejected" for the purposes of the user interface. Confidence measures, rejection accuracy … environmental & speaker adaptation … correction vector … Gaussian distribution for noise & speech; adaptation rate … error rate: 1.8%.* ***Conclusion: The emergence of advanced speech interface is a significant event that will change today's dominating GUI-based computing paradigm****. It is obvious that the paradigm shift would require not only accurate speech recognition, but also integrated natural language understanding with a new programming model. The speech interface would not be considered intelligent until we make it transparent, natural, and easy to use. With our ongoing research effort, we believe that we could*

make Whisper engine work indeed as its acronym promised—Windows Highly Intelligent Speech Recognition."
(cmp. MS Whisper to the modern speech recognition model of OpenAI Whisper)

*** Robust Speech Recognition via Large-Scale Weak Supervision,** Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, 6.12.2022 https://arxiv.org/abs/2212.04356 – Whisper, OpenAI - 680,000 hours of multilingual and multitask supervision … Whisper supports Bulgarian. Tiny, base, medium, large, large-v2 sizes …

Tosh: Wisper is applied by Vsy/Вседържец Autoclap since early 2024.
https://github.com/twenkid/Vsy

* **WaveNet: A generative model for raw audio,** 8.9.2016**,** Aäron van den Oord, Sander Dieleman https://deepmind.google/discover/blog/wavenet-a-generative-model-for-raw-audio/ - direct modelling of the waveform, one sample at a time, instead of training a vocoder, as in previous methods;
https://github.com/kokeshing/WaveNet-tf2
https://github.com/soroushmehr/wavenet-2

WaveNet is based on:

* **Conditional Image Generation with PixelCNN Decoders**, Aaron van den Oord et al., 2016 –https://arxiv.org/abs/1606.05328

*** Pixel Recurrent Neural Networks,** Aaron van den Oord, Nal Kalchbrenner, Koray Kavukcuoglu https://arxiv.org/abs/1601.06759

* **wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations.** Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, Michael Auli. 2020
https://arxiv.org/abs/2006.11477 - Using a transformer for audio encoding.

Abdelrahman Mohamed:
https://scholar.google.ca/citations?user=tJ_PrzgAAAAJ&hl=en – early applications of Deep learning in audio applications: speech recognition, phone recognition, synthesis.

* **Deep belief networks for phone recognition**, Abdel-rahman Mohamed, George Dahl, Geoffrey Hinton, 12.2009,
https://www.cs.utoronto.ca/~gdahl/papers/dbnPhoneRec.pdf
"Hidden Markov Models (HMMs) have been the state-of-the-art techniques for acoustic modeling despite their unrealistic independence assumptions and the very limited representational capacity of their hidden states. There are many proposals in

the research community for deeper models that are capable of modeling the many types of variability present in the speech generation process…"
https://catalog.ldc.upenn.edu/LDC93S1

* **UNIVERSAL PHONE RECOGNITION WITH A MULTILINGUAL ALLOPHONE SYSTEM**, †Xinjian Li et al., 2020 – phoneme, phone, allophone – фонема, звук, алофон; фонемата носи смисъла; звуковете са конкретните произношения, като може да са различни в различни акценти, но да отразяват същата фонема (мек, твърд говор и пр.); алофоните са различни варианти на изговор на фонема, напр. в различна околност – пред определена съгласна, гласна; фонетичен запис: [ ], а звуков /…/[8]

DARPA TIMIT Acoustic-Phonetic Continuous Speech, 1993
https://catalog.ldc.upenn.edu/LDC93S1
https://www.kaggle.com/datasets/mfekadu/darpa-timit-acousticphonetic-continuous-speech acoustic-phonetic knowledge, 6300 sentences, 630 speakers from 8 major dialect regions, each reading 10 sentences.

* **Phone recognition with the mean-covariance restricted Boltzmann machine,** George E Dahl, M Ranzato, A Mohamed, GE Hinton, 2010
https://proceedings.neurips.cc/paper_files/paper/2010/file/b73ce398c39f506af761d2277d853a92-Paper.pdf

*** Acoustic modeling using deep belief networks,** Abdel-rahman Mohamed, G Dahl, Geoffrey Hinton, 2010
https://www.cs.toronto.edu/~asamir/papers/speechDBN_jrnl.pdf
"Gaussian mixture models are currently the dominant technique for modeling the emission distribution of hidden Markov models for speech recognition. … better phone recognition on the TIMIT dataset … by replacing GMM by deep neural networks that contain many layers of features and a very large number of parameters. These networks are first pre-trained as a multilayer generative model of a window of spectral feature vectors without making use of any discriminative information. … discriminative fine-tuning using backpropagation to adjust the features slightly to make them better at predicting a probability distribution over the states of monophone hidden Markov models."

- **Multiresolution deep belief networks,** Yichuan Tang, Abdel-rahman Mohamed, 3.2012
  https://proceedings.mlr.press/v22/tang12/tang12.pdf

[8] https://en.wikipedia.org/wiki/Allophone * https://ru.wikipedia.org/wiki/Аллофон
* https://en.wikipedia.org/wiki/Phone_(phonetics)

* Topics: Neural Machine Translation, Language Models, LLMs, Text Generation, Text and Language Learning and Representation, Word Embedding

Multiple-Attribute Text Rewriting, Guillaume Lample, Sandeep Subramanian, Eric Smith, Ludovic Denoyer, Marc'Aurelio Ranzato, Y-Lan Boureau, 21.12.2018, . https://openreview.net/forum?id=H1g2NhC5KQ https://arxiv.org/abs/1710.04087 Controllable text generation, generative models, conditional generative models, style transfer; rewriting text conditioned on multiple controllable attributes; unsupervised "style transfer" in; others: a latent representation, independent of the attributes of the "style": a new model that controls several factors of variation; instead of disentanglement: a simpler back-translation. Control over multiple attributes: gender, sentiment, product type; more fine-grained control on the trade-off between content preservation and change of style with a pooling operator in the latent space. Fully entangled model produces better generations … even with multiple sentences and multiple attributes.
Tosh: The examples, age-groups 18-24, 64+, reviews: stereotypes.

Word Translation Without Parallel Data, Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, Hervé Jégou, 10.2017/1.2018 https://arxiv.org/pdf/1710.04087 – Learning cross-lingual word embeddings: bilingual dictionaries or parallel corpora; recent – the need for parallel data supervision can be alleviated with *character-level information*; however, worse results and limited to pairs of languages with the same alphabet. This work builds a bilingual dictionary mapping two languages without any parallel corpora, by aligning monolingual word embedding spaces in an unsupervised way; without any character information. Works well for distant language pairs, such as English-Russian or English-Chinese; and for low-resource language pairs such as English-Esperanto; the potential impact in fully unsupervised machine translation. The code, embeddings and dictionaries are publicly available. Cross-domain similarity local scaling – nearest neighbors KNN, hubs & anti-hubs.

*** Distributed word-representations with artificial neural networks: RNN etc.**
Mikolov, Sutskever, A.Mnih, G.Hinton, Bahdanau, Bengio, …

*** Tomas Mikolov,** Kai Chen, Greg Corrado, and Jeffrey Dean. **Efficient estimation of word representations in vector space**. ICLR 2013

* Tomas Mikolov, Quoc V Le, and Ilya Sutskever. **Exploiting similarities among languages for machine translation**, 9.2013 arXiv preprint arXiv:1309.4168 https://arxiv.org/abs/1309.4168 - automating dictionary and phrase table generation in MT with monolingual data and small bilingual datasets for learning word mappings with a high precision. … a linear projection between vector spaces that represent each language; distributed Skip-gram or Continuous Bag-of-Words (CBOW) … Language Training tokens/Vocabulary size: English 575M/127K; Spanish 84M/107K; Czech 155M/505K

* Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. **Distributed representations of words and phrases and their compositionality**. Advances in neural information processing
systems, 2013

* Tomas Mikolov, Martin Karafiat, Lukas Burget, Jan Cernocky, and Sanjeev Khudanpur. 2010. **Recurrent neural network based language model**

* Tomas Mikolov. 2012. **Statistical Language Models based on Neural Networks**

* Andriy Mnih and Geoffrey E Hinton. 2008**. A scalable hierarchical distributed language model.**

*** Neural Machine Translation by Jointly Learning to Align and Translate**
Dzmitry Bahdanau, Kyunghyun Cho, Yoshua Bengio 9.2014/5.2016 - ***Attention***
… a *recently* proposed approach to MT. Unlike the traditional statistical MT, the NMT: build a single NN that can be jointly tuned to maximize the translation performance; Encoder-Decoders: a source sentence encoded into a fixed-length vector from which a decoder generates a translation. … the fixed-length vector is a bottleneck in improving the performance of this architecture; allow a model to automatically (soft-)search for *parts of a source sentence that are relevant to predicting a target word*, without having to form these parts as a hard segment explicitly. … achieve a perf. ~ SOTA phrase-based system in English-to-French transl. (soft-)alignments … RNN, BiRNN, …
"4.1. Dataset … WMT '14 contains the following English-French parallel corpora: Europarl (61M words), news commentary (5.5M), UN (421M) and two crawled corpora of

90M and 272.5M words respectively, totaling 850M words.
…"
* Kalchbrenner, N. and Blunsom, P. (2013). **Recurrent continuous translation models**.

* Sutskever, I., Vinyals, O., and Le, Q. (2014). **Sequence to sequence learning with neural networks.** In Advances in Neural Information Processing Systems (NIPS 2014)

* Cho, K., et al., 2014. **Learning phrase representations using RNN encoder-decoder for statistical machine translation**. In Proceedings of the Empirical Methods in Natural Language Processing (EMNLP 2014).

* Cho, K. Et al., 2014. **On the properties of neural machine translation: Encoder–Decoder approaches.**

* Devlin, J. et al. 2014**. Fast and robust neural network joint models for statistical machine translation.** In Association for Computational Linguistics

* Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). **A neural probabilistic language model.** J. Mach. Learn. Res., 3, 1137–1155.

* Schuster, M. and Paliwal, K. K. (1997). **Bidirectional recurrent neural networks.** Signal Processing, IEEE Transactions on, 45(11), 2673–2681.

*** Ilya Sutskever, Training Recurrent Neural Networks, PhD thesis, 2013**
https://www.cs.utoronto.ca/~ilya/pubs/ilya_sutskever_phd_thesis.pdf - temporal RBM ...; the solution manages to train a RNN on long sequences, overcoming a problem assumed that was thought to be unsolvable; it learns long range dependencies for up to 200 timesteps; Ch.8. Delayed feedback, motor cortex; in order to handle unpredictable disturbances the controller "*must rapidly estimate and counteract the disturbance by observing the manner in which the arm responds to the muscle commands*"; more hidden states are needed to map the long-range structure of text;

* Todorov, E. (2004). **Optimality principles in sensorimotor control**. Nature Neuroscience, 7(9):907–915.

* Li, W. and Todorov, E. (2004). Iterative linear-quadratic regulator design for nonlinear biological movement systems. In ICRA-1, volume 1, pages 222–229, Setubal, Portugal.

* Huh, D. and Todorov, E. (2009). Real-time motor control using recurrent neural networks. In IEEE Symposium on Adaptive Dynamic Programming and Reinforcement Learning, pages 42–
49

Language modelling, word-agnostic: "the sequence memoizer", 2009-2010 – a hierarchical nonparametric Bayesian method. (Wood et al., 2009; Gasthaus et al., 2010)

* **Alternative approaches for sequence and next words prediction, instead of neural networks**: stochastic memorizer, sequence memorizer

*__A stochastic memoizer for sequence data__. Wood, F., Archambeau, C., Gasthaus, J., James, L., and Teh, Y. (2009). In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1129–1136. ACM. http://www0.cs.ucl.ac.uk/staff/c.archambeau/publ/icml_fw09.pdf

* Gasthaus, J., Wood, F., and Teh, Y. (2010). **Lossless compression based on the Sequence Memoizer.** In Data Compression Conference (DCC), 2010, pages 337–345. IEEE. – extends the Sequence Memorizer (SM) with an entropy coder (e.g. arithmetic coder of Witten et al. 1987). Cmp: PPM and CTW algorithms [Cleary and Teahan, 1997; Willems, 1998] … *"The context tree is essentially a (compressed) suffix tree of the reversed string* $x_{i:-1:1}$*," PLUMP/DEPLUMP, INSERTCONTEXTANDRETURNPATH; PATHPROBABILITY – computes the probability of each symbol s in all contexts along the path; parameter update function UPDATEPATH – update the counts in* $S_{x1:i}$ *for all nodes along* $(u_0, \ldots, u_P)$ *in light of the observation* x_{i+1} *… „Our compressor is derived from an underlying probabilistic model which we call the sequence memoizer. This is a hierarchical Bayesian nonparametric model composed of Pitman-Yor processes originally conceived of as a* ***model for languages (sequences of words)****[1]. SM encodes the prior knowledge that the predictive distributions over the subsequent symbols in different contexts are similar to each other, with contexts sharing longer suffixes being more similar to each other. In other words,* ***the later symbols in a context are more important in predicting the subsequent symbol****[2]. Also, by virtue of using PYPs, the SM encodes an assumption that sequences of symbols have power-law properties. context tree; unbounded context lengths; context-mixing PAQ family of compressors [Mahoney, 2005]; CTW, PPM; PAQ;* PYP; CTW – Context-tree weighting method; the use of *latent variables to capture regularities in the data for compression purposes, as applied in this paper, has been explored in the past [Hinton and Zemel, 1994]*

*Hinton, G. E. and Zemel, R. S. (1994). Autoencoders, minimum description length, and Helmholtz free energy. Advances in Neural Information Processing Systems 6, pages 3–10.

* PPM: Prediction by partial matching, https://en.wikipedia.org/wiki/Prediction_by_partial_matching published research since mid-1980s, implementations since 1990s – the PPM alg. requires a lot of RAM; lossless compression of text; https://pypi.org/project/pyppmd/

* PAQ: lossless data compression archivers, context-mixing algorithm; predictor & arithmetic coder; *weighted combination of probability estimates from a large number of models conditioned on different contexts. Unlike PPM, a context doesn't need to be contiguous.* https://en.wikipedia.org/wiki/PAQ

* Arithmetic Coding: https://en.wikipedia.org/wiki/Arithmetic_coding
* CTW: Context-tree weighting
* Hutter, M. (2006). Prize for compressing human knowledge. URL:http://prize.hutter1.net/
* Mahoney, M. (2009). Large text compression benchmark. URL: http://www.mattmahoney.net/text/text.html
* Mahoney, M. V. (2005). Adaptive weighing of context models for lossless data compression. Technical report, Florida Tech. Technical Report CS-2005-16, 2005.
* F. M. J. Willems, Y. M. Shtarkov, and T. J. Tjalkens, "**Context tree weighting: A sequential universal source coding procedure for FSMX sources,"** in Proc. IEEE Int. Symp. Information Theory (San Antonio, TX, Jan. 17–22, 1993)
__* A. Blumer, J. Blumer, A. Ehrenfeucht, D. Haussler, and R. McConnel, "Linear size finite automata for the set of all subwords of a text: An outline of results," Bull. Eur. Assoc. Theoret. Comput. Sci., vol. 21, pp. 12–20 – Z_4-linearity of binary code over Z_4 under the Gray map. .. several well-known families of nonlinear binary codes are actually Z_4-linear.
* **The context-tree weighting method: extensions**, Willems, F. M. J. (1998). https://pure.tue.nl/ws/portalfiles/portal/1600756/Metis122604.pdf
Infinite complete context tree, infinite context tree, infinite-depth context-tree weighting

*** Todor, 5.2.2025:** *1. **"a model for languages (sequences of words)" –** This definition is an example of the utterly reductive conception of language as a "*sequence of words*", where "words" are viewed as *the tokens*. Real thoughts have deeper and more complex representations and the linear *textual* or *vocal* sequence is only an input-output limitation of the physical body, time and cognitive performance trade-offs. LLMs emulate the existence of multiple possible trajectories with Tree-of-thought, ReAct, various Reasoning methods and other techniques for sampling multiple generation paths, estimating consistency, matching final results etc[9].
*2. ***"the later symbols in a context are more important in predicting the subsequent symbol" –*** This is not a strict rule in the real world. The later symbols in a context are *not always* more important in predicting the next symbol – the real languages and real texts are not just probabilistic sequences of the *symbols* in the specific *run of tokens*, they have meanings which may have different nature, origin and consequences; for example the specific words as tokens or code can refer to other sources of meaning or directions, to recall other "engines" and to invoke something else, which is not specified or cannot be defined in the current text or "context" etc. The development of LLMs recognized and reflected that need with the inclusion of "tool use", "solvers", "web access", robot control etc. Regarding the last citation, note that sometimes *the most distant symbol*, word etc. could be *the most predictive* in comparison to the rest, e.g. the *title* or the name of a character in a story before a conclusion, which addresses it; also, the most precise prediction is a result of the estimation of the whole sequences and also of a choice of proper interpretation and all map to other representations, virtual universes, simulations

[9] For reviews see this book "Lazar" and the other volumes from *The Prophets of the Thinking Machines* such as Listove (Reflections on Everything), Anelia, the main volume (#prophets, #tosh) and others.

etc. and still the predictions can be approximate or a set of *possibilities*, as the text, especially the interesting and original one, does not answer everything in an obvious and explicit way. There are different precisions of prediction; the average precision of the prediction may be different and discontinuous at different scales – within the lowest level shortest symbols as characters or lowest level tokens in the LLMs, or between words or sentences or between sentences; when there is a plot-twist etc. The highest predictability and little or no uncertainty is measured when the prediction head traverses stereotypical sequences; a curious human reader of fiction would find such passages boring and would rather prefer not to be able to predict with such a high precision; instead she would like to be surprised and some level of the representation, but yet to be able to find and understand a plausible mapping and correlations in the related realms of the text and the virtual universes, when all the evidence is collected and integrated.

The actual content of the data, perceived by the evaluator-observer, is more than the literal symbols. It includes the implications, connections, associations etc., which are also activated and considered for the estimation of the following predictions. The hidden states of the deep multi-layered LLMs encompass and encode some of these features and forces in implicit form.

* https://en.wikipedia.org/wiki/Pitman%E2%80%93Yor_process

Pitman-Yor process: $0 \leq d < 1$ a discount parameter, a strength parameter $\theta > -d$ and a base distribution G_0 over a probability space X.

* Pitman, Jim; Yor, Marc (1997). The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator. … doi:10.1214/aop/1024404422. MR 1434129. Zbl 0880.60076.

* Perman, M.; Pitman, J.; Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. Probability Theory and Related Fields. 92: 21–39. doi:10.1007/BF01205234.

* Ilya Sutskever. **Nonlinear Multilayered Sequence Models**, Master's Thesis, 2007 (Sutskever, 2007)

* **Learning Multilevel Distributed Representations for High-Dimensional Sequences** Ilya Sutskever and Geoffrey Hinton. In the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS), 2007 (Sutskever and Hinton, 2007)

* Generating Text with Recurrent Neural Networks Ilya Sutskever, James Martens, and Geoffrey Hinton. In the 28th Annual International Conference on Machine Learning (ICML), 2011 (Sutskever et al., 2011)

---

* Transformers - the seminal paper from 2017

*** Attention Is All You Need,** Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin, 6.2017/8.2023 https://arxiv.org/abs/1706.03762 - the transformer architecture; (преобразител);

*** ATTENTION-BASED SEQUENCE TRANSDUCTION NEURAL NETWORKS, Shazeer et al, 7.8.2020**
https://patentimages.storage.googleapis.com/1b/52/ac/3f1c75cb9ef037/US10956819.pdf
https://patentimages.storage.googleapis.com/09/5d/66/533474d7020f06/US20240144006A1.pdf **2.5.2024**
"Self-attention , sometimes called intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. … effectively learn dependencies between distant positions during training, improving the accuracy, … easier to train, quicker to generate … machine translation task (MT) …" Many possible applications are listed: MT, QA (question-answering), summarization, text-to-image, image-to-text, speech-to-text, … encoder-decoder.

In some cases, the positional embeddings are learned; the term "learning" means that an operation or a value has been adjusted during the training of the sequence transduction neural network. …PE(pos,2i) = sin(pos/10000$^{2i/dmodel}$), POS(pos,2i+1) = cos(pos/10000$^{2i/dmodel}$) … d_{model} – dimensionality of the positional embedding **"**

* Topics: Neural Program Synthesis

Illia Polosukhin On Inventing The Tech Behind Generative AI At Google, CNBC https://www.youtube.com/watch?v=Q4_YTXhq7no – Illia explains how they created the transformer; the original idea: Jakob Uszkoreit; they realized the sequence items can be processed in parallel and not just sequentially like in the RNN or jumping back to check the answer (~1-4 min.); Illia is the creator of "Near AI" in 2017, intended to do code generation with transformers, but they realized they lacked the scale of the data; they paid developers, … then switched to Blockchain technologies.

*** Naps: Natural program synthesis dataset**, Maksym Zavershynskyi, Alex Skidanov, Illia Polosukhin, 6.7.2018, https://arxiv.org/pdf/1807.03168 human written problem statements and solutions for these problems … crowdsourcing & human-written solutions in programming competitions with input/output examples; AlgoLISP

https://github.com/nearai/program_synthesis/blob/master/program_synthesis/naps/README.md Problems from: https://codeforces.com/problemset/problem/313/C
UAST – universal abstract syntax tree;
https://github.com/nearai/program_synthesis/blob/master/program_synthesis/naps/uast/README.md
* Sarnak, Neil and Tarjan, Robert E. Planar point location using persistent search trees.

Commun. 1986.
https://www.cs.princeton.edu/courses/archive/spr04/cos423/handouts/planar%20point.pdf
https://courses.csail.mit.edu/6.854/16/Notes/n2-persistent.html Persistent Data Structures

* Polosukhin, Illia and Skidanov, Alexander. **Neural program search: Solving programming tasks from description and examples**. CoRR, 2018
http://arxiv.org/abs/1802.04335
Seq2Tree model, semi-synthetic dataset; Domain Specific Languages DSL inspired by LISP – easily converted into Abstract Syntax Tree; types: const, argument, function call, function, lambda. Related: 1. Programming by example (PBE) – RobustFill, **DeepCoder**, Neuro-Symbolic Program Synthesis (predict tree structured programs at search time), Deep API Programmer, Programming from Description (PfD), Latent Program Induction (LPI) – Neural Turing Machines, Neural GPUs, stacks-augmented RNNs, Neural Program Interpreters, Semantic Parsing – PfD limited to some structured form. …

* **Neuro-Symbolic Program Synthesis,** E Parisotto, A Mohamed, R Singh, L Li, D Zhou, P Kohli, arXiv preprint arXiv:1611.01855, 407, 2016
https://arxiv.org/pdf/1611.01855

* David Alvarez-Melis and Tommi S Jaakkola. **Tree-structured decoding with doubly-recurrent neural networks**. 2016

* **LEARNING A NATURAL LANGUAGE INTERFACE WITH NEURAL PROGRAMMER**, Arvind Neelakantan et al., 2017 https://arxiv.org/pdf/1611.08945
Map natural language NL to logical forms or programs to execute as queries to a DB. Вж с.13 табл. Операции.: aggregate – count, superlative: argmax, argmin; comparison: > < >= <=; Table Ops: select, mfe (most common entry), first, last, previous, next; Print; Reset. p.8: “what is the total number of teams”, “who had more silver medals: Bulgaria or Belgium?”; training examples: 11321/2831/4344. … 37.7% accuracy, comp. to 37.1% of traditional NL semantic parser; WikiTableQuestions dataset.
Тош: Recognize topics/related columns, operations …, Срвн работата на Галя Ангелова, БАН, нач. На 1980-те в областта на достъп до БД с естествен език, вж списък с български учени, приложение Анелия.

* Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. **Learning to compose neural networks for question answering**. NAACL, 2016

* Ronald Williams. **Simple statistical gradient-following algorithms for connectionist reinforcement learning**. Machine Learning, 1992.

* **Pointer Networks**, Oriol Vinyals, Meire Fortunato, Navdeep Jaitly, https://arxiv.org/abs/1506.03134 Neural architecture, conditional probability of an output sequence with elements – ***discrete tokens***, corresponding to ***positions in an input sequence*** – for sorting and combinatorial optimization; using “the recently proposed mechanism of *neural attention*” to solve the problem of variable size of the output dictionaries. Ptr-Net – find approximate solutions of convex hull, Delaunay

triangulations, planar travelling salesman in a *purely data-driven* way: given examples … RNN – *"the inputs and outputs were available at a fixed frame rate"; sequence-to-sequence: one RNN to map an input sequence to an embedding and another (possibly the same) RNN to map the embedding to an output seq.* Ptr-Net: at each step "modulates a content-based attention mechanism" over inputs … softmax distribution with dictionary size = len(input). … Seq-to-Seq: a training pair, compute the conditional prob. Using a parametric model (RNN with params. Θ) to estimate the terms of the prob. Chain rule of multiplied probabilities, conditional prob. For the training set. … Prev.work: RNNSearch, Memory Networks, Neural Turing Machines.

* **NEURAL GPUS LEARN ALGORITHMS,** Łukasz Kaiser & Ilya Sutskever Google Brain, 2015/2016 https://arxiv.org/abs/1511.08228 – Improving the results of Neural Turing Machines (NTMs) which are not parallel and are hard to train… long addition and long multiplication; contrib.: parameter sharing relaxation, adding a small dropout & gradient noise to increase generalization. A problem of RNNs, LSTMs… *the entire input is encoded into a single fixed-size vector, so the model cannot generalize to inputs much longer than this fixed capacity.* The attention mechanism partially resolves it (Bahdanau et al., 2014) by allowing inspection of arbitrary parts of the input in *every decoding step.*

Tosh: Why the addition of dropout and noise improves generalization and reduces overfitting? Maybe because it forces the search process to explore side-solutions and to use a broader source of evidence.

* **SuperCoder: Program Learning Under Noisy Conditions From Superposition of States,** Ali Davody, M.Safari, Razvan V. Florian, 7.12.2020 https://arxiv.org/abs/2012.03925 – a new method of program learning in a DSL, gradient descent with no direct search. Probab.repr. of the DSL vars.; at each step, diff. func. Applied on the DSL vars with a cert. prob. diff. possible outcomes… superposition in a single fuzzy state, contrasted at the final step with the ground-truth output – a loss func. Attention-based NN; able to learn under noise. "program synthesis" - a symbolic representation of the program is inferred or "program induction" - a neural network predicts the output corresponding to a given input. Combined search + NN; beam search …

* **Deepcoder: Learning to write programs.** M. Balog, A. L. Gaunt, M. Brockschmidt, S. Nowozin, and D. Tarlow. In International Conference on Learning Representations (ICLR), 7.11.2016/2017. https://arxiv.org/abs/1611.01989 *solving programming competition-style problems from input-output examples using* ***deep learning***; augmenting enumerative search and SMT-solvers. RNNs. … differentiable interpreters; a differentiable mapping from source code and inputs to outputs; DSL: defining a programming language that is expressive enough to include real-world programming problems while being high-level enough to be predictable from input-

output examples; *Satisfiability Modulo Theories (SMT) SAT-style search with theories like arithmetic and inequalities;* the ranking problem… (of possible solutions). Sample programs: Program 0: k ← int, b ← [int], c ← SORT b, d ← TAKE k c, e ← SUM d Input-output example: Input: 2, [3 5 4 7 5] Output: [7]
Description: *A new shop near you is selling n paintings.*
You have k < n friends and you would like to buy each of your friends a painting from the shop. Return the minimal amount of money you will need to spend." | Fig.4: a diagram with *Predictions of a neural network on the 9 example programs … Numbers in squares would ideally be close to 1 (function is present in the ground truth source code), whereas all other numbers should ideally be close to 0 (function is not needed).* p.5. *1. an encoder: a differentiable mapping from a set of M input-output examples generated by a single program to a latent real-valued vector, and. 2. a decoder: a differentiable mapping from the latent vector representing a set of M input-output examples to predictions of the ground truth program's attributes.*
p.5 **Possible continuations/symbols in the DSL and generated probabilities:** "+1) (-1) (*2) (/2) (*-1) (**2) (*3) (/3) (*4) (/4) (>0) (>0) (%2==1) (%2==0) HEAD LAST MAP FILTER SORT REVERSE TAKE DROP ACCESS ZIPWITH SCANL1 + - * MIN MAX COUNT MINIMUM MAXIMUM SU**"** 0.0 0.0 0.1 .. 1.0 0.0 .. 0.4 ..
Tosh: Semantic parsing is better than this. @Vsy: Cmp & apply semantic parsing. Zrim: {K}#, {K}! {K-.-K}

* **FlashExtract: a framework for data extraction by examples,** Vu Le, Sumit Gulwani, 6.2014 http://www-cs-students.stanford.edu/~adityagp/courses/cs598/papers/flash-extract.pdf
Programming by Examples (PBE): FlashExtract Video 2, Microsoft Research 337 K Subscribers, 387 views, 30.01.2017 г. https://www.youtube.com/watch?v=_KYyibApd_4
"Extract the strings | starting at each line beginning with Left-Bracket Alphanumeric;ending before the following first occurrence of Line Separator after WhiteSpace.Number.Dot …" … → other options: ending before/after the *last* occurrence of Alphanumeric-WhiteSpace-Camel Case in the second line … +3,-20 (positive/negative examples/cases?); nested extraction; Active learning: the system asks questions: Disambiguation. *inductive program synthesis from a definition of data-extraction DSLs*

* **Programming by Examples (PBE): FlashExtract,** 30.01.2017 г., 411 пок. https://www.youtube.com/watch?v=apTsnpsPEds

*** Programming by Natural Language: Bing enabled code search,** Microsoft Research, 30.1.2017, 267 показвания https://youtube.com/watch?v=nnsCYr9xOT8 - Visual Studio add-on … IntelliSense: "How do I…" … generate md5 hash from string @line… → snippet

* **A Webbased Frontend for Easy Interaction with the Inductive Programming System Igor**, Microsoft Research, 479 показвания, 7.10.2016 г
https://youtu.be/wRbhrUht-ok Maude programming language, hypotheses – Maude functions on input-output examples (usually 4 to derive hypoth.); internal types

*** Maude: term-language, sets of equations & rewrite-rules.**
https://en.wikipedia.org/wiki/Maude_system
https://maude.cs.illinois.edu/wiki/The_Maude_System

*** Programming and symbolic computation in Maude, Francisco Durán** et al., 1.2020, https://www.sciencedirect.com/science/article/pii/S2352220818301135

*** Programming by Demonstration: a Machine Learning Approach, Tesa Lu, 2001,** PhD thesis, University of Washington; supervisors: Daniel S. Weld, Pedro Domingos **PBD** – another term for PBE. Solves biography editing problems: **SMARTedit.** Cmp: Few-show learning. Macro-recorders in wordprocessors etc. from a long time are a proto-PBD, but usually too brittle, recording only exact key sequences. Cross-domain PDB: abstract language with control-flow – a subset of Python: SMARTpython; a new ML framework: *complex functions— mapping from one complex object to another complex object: states of the application etc.* (cmp: structure learning); see *version space* in concept learning/abstraction, however extended from a binary {0,1} to any functions; *a hypothesis space with the set of all possible actions (functions from state to state) that the user could have taken; a hierarchical version space algebra (union, join, transduction); version space libraries;* … Editing by Example (EBE); DEED (1998) - focus interpretation, modification interpretation; heuristics; classification rule learners; hints; action history, pattern matching
Cmp: Imitation learning in robotics; self-customizing software & adaptive interface; COLLAGEN – hierarchical plan, …
(T, L, P, S) states (text buffer, cursor location, clipboard contents, and selection region) …"Learning programs from traces" p.74, py subset, loops & condit.;program-statement-if(false,true)-primitive-next_step; a trace is a seq of states S0, S1, S2… Conditional(VarI<const, VarI>const, Var < Var, Var>Var, Var==Var) – version space … **mixed-initiative interfaces** (TA: collaborative; сътрудничащ взаимлик/въоблик, подпомагащ агент, подпомагащ взаимлик; одушевен взаимлк, одушевен въоблик; самопостроящ се, самонагласящи се и променливи *роли* на взаимодействието; системата може да предлага, да прекъсва, да „поема инициативата", да допълва и пр. като в разговор, във всякакви сетивно-моторни модалности. При интерфейсите с фиксирна, твърда инициатива, във всеки момент е ясно дефинирано кой води/"говори" и кой е слушател/приемник.

* Allen Cypher, editor. Watch what I do: Programming by demonstration. MIT Press, Cambridge, MA, 1993
* Pattie Maes and Robyn Kozierok. Learning interface agents. In Proceedings of AAAI93, pages 459–465, 1993

* Henry Lieberman, editor. Your Wish is My Command: Giving Users the Power to Instruct their Software. Morgan Kaufmann, 2001
* Robert P. Nix. Editing by Example. ACM Transactions on Programming Languages and Systems, 7(4):600–621, October 1985. (EBE)
* Y. Fujishima. Demonstrational automation of text editing tasks involving multiple focus points and conversions. In Proceedings of Intelligent User Interfaces ’98, pages 101–108, 1998. DEED – extends EBE
* Gordon W. Paynter. Generalising Programming by Demonstration. In Proceedings Sixth Australian Conference on Computer-Human Interaction, pages 344–345, Nov 1996.
* J. Schlimmer and L. Hermens. Software agents: Completing patterns and constructing user interfaces. J. Artificial Intelligence Research, pages 61–89, 1993 – predictive completion of notes
* C. Rich and C. Sidner. Segmented interaction history in a collaborative agent. In Third Int. Conf. Intelligent User Interfaces, pages 23–30, Jan 1997.
N. Lesh, C. Rich, and C. Sidner. Using plan recognition in human-computer collaboration. In Proceedings of the Seventh Int. Conf. on User Modelling, Banff, Canada, July 1999
* Steven A. Wolfman, Tessa Lau, Pedro Domingos, and Daniel S. Weld. **Collaborative Interfaces for Learning Tasks: SMARTedit Talks Back**. In Proceedings of the 2001 Conference on Intelligent User Interfaces, 2001

*** Principles of mixed-initiative user interfaces**, Eric Horvitz, 1999 https://erichorvitz.com/chi99horvitz.pdf – interface “agents” vs direct manipulation; important points: the uncertainty of user’s goal, user’s attention’s content,… dialogs to resolve uncertainties; efficient direct invocation and termination; alerts to minimize the cost of wrong guesses, grades of precision (generality) of proposed actions (do less but the correct one), memory of recent interactions, continual learning by observing; LookOut project for Outlook email client – scheduling, calendar, emails … identify user’s goals by considering messages’s content; manual or automated-assistance modalities; social-agent modality as an animated characted (Genie): hands-free mode with speech recognition and text-to-speech; levels of confidence; expected utility: autonomous actions: *only if the agent believes that they will have greater expected value than inaction for the user.*

* **Machine Learning and Intelligence in Our Midst,** Microsoft Research, 28.03.2012 г…. https://youtu.be/Qe0256zAsNU?si=NPN90J40mdclXkX2 5609 views … 38 min: Coordinate: Presence forecast .. „since 2002“ ... for the assistant.
46:50: Memory Landmarks: Life browser, Selective ...
49:11: Privacy, protection, privacy budget ...
50 min: PSearch ... **Integrative intelligence:** Multimodality: integration of data

sources from all kinds of inputs and sensors – see citation in the section about Eric Horvitz or see from 29 min.
29:15 .. *I dobe into richer applications that mesh together vision and speech natural language and this gets back to the area that I mentioned a bit earlier: integrative intelligence … the idea is to mesh together natural language processing NLP motion control speech generations; pic kyour addition, planning, learning, vision and inferential methods you might say that the goal here is in part to build a comprehensive systems with a whole is somehow magically greater than the sum*
of the parts and this is a kind of maybe a dream of AI that if we bring together a vision and speech natural language and inverse kinds of reasoning and learning that there'll be some synergies among these components if you wove these together weave them together in an effective manner so our main project what we're exploring this interactive AI notion is called situated interaction (…)"

* Topics: Mixed Initiative Interaction

*** Mixed-Initiative Creative Interfaces**, Sebastian Deterding et al. (9), CHI EA '17: Proceedings of the 2017 CHI Conference, Pages 628 – 635, 5.2017
https://doi.org/10.1145/3027063.3027072
https://eprints.whiterose.ac.uk/112515/1/wks0132_deterdingA.pdf **A Workshop.** creativity support in HCI; from computer as tool to a collaborator to autonomous creator (computational creativity); procedural content generation (PCG); (semi-)-autonomous interfaces
* Joseph Carl Robnett Licklider. 1960. Man-Computer Symbiosis. IRE Transactions on Human Factors in Electronics 1, 1: 4–11. https://doi.org/10.1109/THFE2.1960.4503259
* Jaime R Carbonell. 1971. AI in CAI: An Artificial Intelligence Approach to Computer-Assisted Instruction. IEEE Transactions on Man-Machine Systems 11, 4: 190–202. https://doi.org/10.1109/TMMS.1970.299942
* (PDF) Drawing Apprentice: An Enactive Co-Creative Agent for Artistic Collaboration.https://www.researchgate.net/publication/280014470_Drawing_Apprentice_An_Enactive_Co-Creative_Agent_for_Artistic_Collaboration
* Gillies, M., Kleinsmith, A., and Brenton, H.Applying the CASSM Framework to Improving End User Debugging of Interactive Machine Learning. Proceedings of the 20th International Conference on Intelligent User Interfaces, (2015),181–185
See also other references.
*** Mixed-initiative interaction – Trends & Controversies**, Intelligent Systems, Marti A. Hearst, J.Allen, E.Horvitz, C.Guinn, 1999
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/11/mixedinit.pdf
*** Mixed-initiative interaction,** James F. Allen, 1999 Levels of capabilities for

M.I. agents: 1. Unsolicited reporting – notifications for critical information, when it appears; 2. Subdialogue initiation for clarification, corrections etc. 3. Fixed subtask initiative on predefined subtasks. 4. Negotiated mixed initiative - coordination of the initiative with other agents (multi-agent systems); Intention recognition: speech acts (a request, an acceptance of promise or other performatives, speech acts); what exact action the user tries to accomplish, does she want to edit it and how etc. Dialog, turn-taking, contextual interpretation, grounding; interrupt, wait, answer … plan management system; data from human interaction, different interactions: evaluating & comparing options, suggesting courses of action, clarifying and establishing state, clarification & confirming the communication, discussing problem-solving strategy, summarizing courses of action, identifying problems and alternatives.

[TA. 2. See coding agents, "CLI" (command-line interface): claude code-cli, gemini-cli, opencode, kilocode etc. They interactively ask the user for clarification of the requirements: what libraries to use, which features are more critical, what programming language to use etc.)
3. For example in modern LLM assistants the user provides the prompt, the chatbot generates a summary etc.

* Uncertainty, action, and interaction: in pursuit of mixedinitiative computing Eric Horvitz, 1999
* Evaluating mixed-initiative dialog, Curry I. Guinn, 1999
* …
* Eric Horvitz:
https://scholar.google.com/citations?user=V4OPEAgAAAAJ&hl=en
A lecture cited in: **News: TILT - Efficient Rectification (Texture-Pixel-Based 3D-Like Perspective and other) by Microsoft Research - and the SIGI-AGI Prototype and Research Accelerator News**, T.Arnaudov, **Artificial Mind 2012**
https://www.youtube.com/watch?v=Qe0256zAsNU&list=UUCb9_Kn8F_Opb3UCGm-lILQ&index=40&feature=plcp

*** Machine Learning and Intelligence in Our Midst** **Microsoft Research** 344 K Subscribers, **5609 views,** 28.03.2012 г**.** (30.5.2025)

38 min: Coordinate: Personal Assistants; Presence & Availability predictions .. "Since 2002" ...
46:50: Memory Landmarks: Life browser, Selective ...
49:11: Privacy, protection, privacy budget ...
50 min: Psearch- Personal Search (Desktop search, contextual ambiguity resolution) ;

*** An Introduction to Least Commitment Planning Daniel S. Weld, 1994**
World description, goal, and domain theory**;** search with progression and regression forward or backward planners – from initial states or from the goal;

coherent planning vs situated (deliberative vs non-deliberative, reactive); plan space or state space; branching in all positions, not only the end or the beginning of the plan; … selection of goal, action, … partial order (POP – partial order planning); causal links – provide preconditions for other actions; threats (changes which alter the state and the plan – invalidate causal links and require resolution with promotion or demotion – reordering of actions, or confrontation); static & dynamic universes; least-commitment planning: avoid premature commitments, deferred decisions; STRIPS: propositions, conjunctive goals and add/delete lists; subgoal interactions (Sussman anomaly, A,B,C blocks, order of actions). Planning in planning space: nodes are plans; edges are refinements of the plans. [*TA: cmp. PPO, proximal policy optimization*]
Preconditions and effects. Conditional effects: depend on preconditions. Universal quantification: for all preconditions. Threat resolution. Closed world assumption: assuming false for unmentioned literals *[cmp. the frame problem*]. Agenda: list of open preconditions to address. Unification: matching variables to bind parameters. Static or Dynamic universe – the latter allows creation and destruction of objects. Confrontation: resolve threats by negation of problematic items. Heuristic search: domain knowledge helps guiding the planner. Maintenance goals – constraints which must hold through the execution, but may not be stated explicitly, for example *"preserve the environment and self-preserve"* etc. Hierarchical planning, case-based planning vs generative planning: use older plans or generate new from scratch*; explanation-based planning – experience improves efficiency. Branching factor – search efficiency. Systematic exploration: no redundant plan exploration. Soundness and completeness of planners. Plan verification; Null plan – the initial plan <Astart, Aend> with start/end actions - the planning process "grows" from them. Means-end analysis … Atomic planning, deterministic effects, omniscience, sole cause of change (only agent's actions); Total Order Planning vs Partial Order planning: <A; O; L> Actions, ordering constraints and causal links; nondeterministic "choose" primitive: number of calls to solution, number of possibilities to consider – branching factor.
[*What is really "from scratch"? The methods for generation implicitly encode the seeds of possible plans*]

*** Kevin Ellis** https://scholar.google.com/citations?user=tVjxANMAAAAJ&hl=en

* **Dimensionality Reduction via Program Induction.**, 2015, Kevin Ellis, Eyal Dechter, Joshua B Tenenbaum https://cdn.aaai.org/ocs/10284/10284-45219-1-PB.pdf – symbolic dimensionality reduction; programs as a representation for AI systems: prev.work: *(Dechter et al. 2013; Liang, Jordan, and Klein 2010), genetic programming literature (Poli, Langdon, and McPhee 2008) and from work in Cognitive Science termed logical dimensionality reduction (Katz et al. 2008).*

*** Topics: Genetic Programming, Genetic Algorithms, Evolutionary Programming …**

See also Self-Improving General Intelligence, Recursive Self-Improvement

https://www.researchgate.net/publication/216301261_A_Field_Guide_to_Genetic_Programming p.19/34 "5 preparatory steps: terminal set, function set, fitness measure, parameters, termination criterion". **Terminal set**: external input, functions without arguments (display, random sampling), constants; **Func.set**: + - * / … transforms; *closure*: type consistency & evaluation safety (e.g. protected division % if division by 0 returns 1), sufficiency for solving the problem (coverage); can evolve other structures e.g. by programming the steps of creation of the artifacts: electric circuits or other graphs or designs, music etc. **Param.:** population size, probabilities of performing genetic operations (crossover or mutation) … Construct the initial population, seeding… Trees; subtree mutation, node replacement, … Modular, grammatical, Developmental Tree-based GP (p.47/62) automatically defined functions, program architecture & architecture-altering operations; constraining structure; grammar-based constraints; bias .. Developmental Genetic Programming: cellular encoding: modify, grow a simple initial structure (embryo) (Gruau 1994, Gruau and Whitley, 1993 …) * https://www.genetic-programming.com/gpdevelopment.html *Strongly Typed Autoconstructive GP;* Linear vs Graph GP .. Parallel & Distributed; Probabilistic GP, Estimation Distribution Algorithm (EDA), probability tree: conditional probability table for each node. Mixing Grammars and Probabilities [compare "Neurosymbolic"]: *program evolution with explicit learning (PEEL) (Shan, McKay, Abbass, and Essam, 2003.* Multi-objective GP/optimizer (e.g. fitness, speed); Pareto tournament selection… p.77 (92) Bloat & Complexity control .. 9.3 p. 80/96 - Dynamic and Staged Fitness Functions. Static, but staged. … classical repair operators in constrained optimisation. Performance: reduce the cost of fitness,use caches, compute distributed – in evolutionary sense: spread in different locations that can interact and sometimes get disconnected etc. GPUs – at the time Geforce 8800, 128 streaming processors @ 1360 MHz, 16 x 8 blocks. 16 KB shared memory, 8*1 KB L1 cache, 4 texture address units, 8 texture filters. … Effective use of a GPU for GP: only 30 GFLOPs for reverse polish notation RPN post-fix in 2008 [not yet a well developed GPGPU for that purpose]. FPGA …
The Sorrow of AI
"Human-competitive results" instead of "intelligence"
Cmp: "Human-Level AI" https://www.human-competitive.org/awards

* **John R. Koza,** https://www.genetic-programming.com/ List of publications: https://www.genetic-programming.com/jkpubs72to93.html
https://www.human-competitive.org/

* Koza, John R. 1993a. **Simultaneous discovery of detectors and a way of using the**

detectors via genetic programming. *1993 IEEE International Conference on Neural Networks*, San Francisco. Piscataway, NJ: IEEE Press. Volume III. Pages 1794-1801. http://www.genetic-programming.com/jkpdf/icnn1993pattern.pdf

* Koza, John R. 1993b. **Simultaneous discovery of reusable detectors and subroutines using genetic programming**. In Forrest, Stephanie (editor). *Proceedings of the Fifth International Conference on Genetic Algorithms.* San Mateo, CA: Morgan Kaufmann Publishers Inc. Pages 295-302. https://www.genetic-programming.com/jkpdf/icga1993.pdf

* GP on SPMD parallel Graphics Hardware for mega Bioinformatics Data Mining W. B. Langdon, A. P. Harrison, 2008 – GeForce 8800 GTX, RapidMind's GPGPU … http://www0.cs.ucl.ac.uk/staff/W.Langdon/ftp/papers/langdon_2008_SC.pdf

* **Genetic Programming Theory and Practice XVIII**, Ed. Wolfgang Banzhaf, Leonardo Trujillo, Stephan Winkler, Bill Worzel, 2022 https://link.springer.com/book/10.1007/978-981-16-8113-4 ARC – Abstract regression classification

*** Ch. Feature Discovery with Deep Learning Algebra Networks, Michael F. Korns**
https://www.researchgate.net/publication/355439099_Feature_Discovery_with_Deep_Learning_Algebra_Networks

*** Genetic Programming Symbolic Classification: A Study,** Conference paper, 7. 2018 Conference paper, Genetic Programming Theory and Practice XV, Michael F. Korns – vs Symbolic regression

* **Highly Accurate Symbolic Regression with Noisy Training Data**, Chapter, Michael F. Korns, 22.12.2016 https://www.researchgate.net/profile/Michael-Korns/research

* **Extreme Accuracy in Symbolic Regression,** Michael F. Korns, 4.2014 in book Genetic Programming Theory and Practice XI Publisher: Springer Editors: Rick Riolo (Editor, Jason H. Moore (Editor, Mark Kotanchek (Editor …– commercial applications & free; *commercially deployed regression package which handles up to 50 to 10,000 input features using specialized linea rlearning (McConaghy, 2011);* heurism (heuristics), Generalized Linear Model (GLM) (Nelder and Wedderburn (1972) – *GLMs can represent any possible nonlinear formula.* pareto, islands,

*** Eureqa Desktop** - *Eureqa uses a breakthrough machine learning technique called Symbolic Regression to unravel the intrinsic relationships in data and explain them as simple math.* (Initially by Cornell university) → https://www.datarobot.com/

https://web.archive.org/web/20141009020835/http://www.nutonian.com/products/eureqa/

* **Distributed Evolutionary Algorithms in Python** https://github.com/DEAP/deap
https://deap.readthedocs.io/en/master/
https://deap.readthedocs.io/en/master/tutorials/advanced/gp.html
https://deap.readthedocs.io/en/master/tutorials/advanced/constraints.html
SCOOP (Scalable COncurrent Operations in Python)
https://github.com/soravux/scoop

*** See an additional summary of topics and concepts in Evolutionary and Genetic Algorithms in a table below this one.**

* Liang, P.; Jordan, M. I.; and Klein, D. 2010. **Learning programs: A hierarchical bayesian approach**. In Furnkranz, ¨J., and Joachims, T., eds., ICML, 639–646. Omnipress. https://people.eecs.berkeley.edu/~jordan/papers/liang-jordan-klein-icml10.pdf

* Poli, R.; Langdon, W. B.; and McPhee, N. F. 2008. **A field guide to genetic programming.** Published via http://lulu.com and freely available at http://www.gp-fieldguide.org.uk (With contributions by J. R. Koza).

* **Unsupervised Learning by Program Synthesis,** Kevin Ellis, Armando Solar-Lezama, Joshua B. Tenenbaum, 2015.
*https://proceedings.neurips.cc/paper_files/paper/2015/file/b73dfe25b4b8714c029b37a6ad3006fa-Paper.pdf
„the idea that inductive learning should be thought of as probabilistic inference over programs is at least 50 years old." Sketching … Symbolic search; Satisfiability; Modulo Theories (SMT); SMT solver; **Visual concept learning:** a set of example images, parse them into a symbolic form, synthesize a program that maximally compresses these parses. **Synthetic Visual Reasoning Test** (**SVRT**); containment relations contains(i, j); DSL … teleport(position[0], initialOrientation); draw(shape[0], scale = 1); move(distance[0], 0deg). Turtle commands. Grammar rule English description E → (M; D) … Visual: + Alternate move/draw; containment relations; borders relations M → teleport(R, θ0) Move turtle to new location R, reset orientation to .. M → move(L, A); M → flipX()|flipY(); M → jitter() ; → draw(S, Z); Z → 1|z1|z2| · · · Scale .. positions, shapes, lengths; contains, contains? (optional), borders, borders? (optional) between integer indices… English: Morphological rule learning grammar … 4.1 Related Work: *Inductive programming systems have a long and rich history.* .. Stochastic search: genetic programming or MCMC. Others: constrain the hypothesis space to enable fast exact inference. The inductive logic programming – Prolog programs using heuristic search. .. Recent successes of systems that put program synthesis in a probabilistic framework. .. introduction of solver-based methods for learning programs. Synthesizing

programs from large datasets is difficult; complete symbolic solvers scale poorly with the increase of the problem size. Counter Example Guided Inductive Synthesis (CEGIS).

* Ray J Solomonoff. **A formal theory of inductive inference.** Information and control, 7(1):1–22, 1964

* **Counterexample Guided Inductive Synthesis Modulo Theories***, Alessandro Abate, Cristina David, Pascal Kesseli, Daniel Kroening, Elizabeth Polgreen, 2018 … CEGIS(T) (counterexample – чрез доказване на противното?; (чрез) опровержение)
SVRT: The Synthetic Visual Reasoning Test Challenge, 11.3.2010
https://k4all.org/2010/03/the-synthetic-visual-reasoning-test-challenge/
23 hand-designed, image-based, binary classification problems. The images are binary and with resolution 128×128. For each problem we have implemented a generator in C++, which allows one to produce as many i.i.d samples as desired.
http://www.idiap.ch/~fleuret/svrt/svrt.pdf ["not found" and missing in archive.org]

* **Comparing machines and humans on a visual categorization test**, François Fleuret francois.fleuret@idiap.ch, Ting Li, Charles Dubout, +2 , and Donald Geman https://www.pnas.org/doi/10.1073/pnas.1109168108 *… two disjoint "categories" of image a compositional "rule. "inside," "in between," and "same." … the SVRT is designed to focus on one in particular—abstract reasoning. … purposely removed many of the subtasks and complications encountered in parsing images acquired from natural scenes: There is no need to recognize natural objects or to account for volume, illumination, texture, shadow, or noise. Moreover, being planar and randomly generated, the shapes are"unknown" to humans, which ameliorates our advantage over machines due to extensive experience with everyday objects and a three-dimensional world. (…) People tend to characterize a category in phrases such as: "the two shapes are in contact," "the two halves of the picture are symmetric," "the shapes are aligned with the large one between two small ones." … One or more of the Gestalt principles of perceptual organization (9, 10), including proximity, similarity, symmetry, inclusion, collinearity, and others. They are higher-order (nonlocal) configural properties of the displays that biological visual systems have evolved to perceive effortlessly as part of scene understanding. … They only have direct access to information of the sort "there is an elongated dark area," "the black pixels are spread out," "there is a patch with edges all in the same direction."*. Cmp: Bongard, 1967. "Chairs, Caricatures, …", 2012, TOUM
* PASCAL2 – Pattern Analysis, Statistical Modeling and Computational Learning https://web.archive.org/web/20110120220358/http://pascallin2.ecs.soton.ac.uk/Challenges/ … https://web.archive.org/web/20110215170401/http://lshtc.iit.demokritos.gr/ ~ 2011 Large Scale Hierarchical Text classification (LSHTC2)

* **Algorithms for Learning to Induce Programs**, Kevin Ellis, 2020, PhD Thesis, MIT;

supervisors: Joshua B. Tenenbaum & Armando Solar-Lezama …
https://dspace.mit.edu/bitstream/handle/1721.1/130184/1241081869-MIT.pdf?sequence=1&isAllowed=y …
mentioned (in thanks) Kliment Serafimov (North Macedonia, undergrad.) on DreamCoder
https://cdn.aaai.org/ocs/10284/10284-45219-1-PB.pdf

* **Is Programming by Example solved by LLMs?,** Wen-Ding Li, Kevin Ellis, 6.2024/19.11.2024 https://arxiv.org/pdf/2406.08316 PBE: Given input-output examples of a hidden algorithm, they seek to construct the source code of the underlying function; algorithms on vectors of numbers, string manipulation macros, and graphics programs in LOGO/Turtle. Pick a program ρ from {ρ ∈ L : ρ(Xi) = ρ(Yi), ∀i}. Succeed if ρ(X′j) = ρ(Y′j), ∀j (1) … FlashFill's Basic prompting: List functions is a PBE domain meant to model a "programmer's assistant". *Classic PBE: Traditional approaches to programming-by-examples operate by symbolically searching or solving for programs consistent with the input-output examples [13, 49, 2, 1, 37, 6]. They use domain-specific programming languages that are designed to either enable efficient search and/or bias the system toward functions that are likely to generalize new inputs. Related Work: Automatic data generation with LLMs, such as self-instruct [32], WizardCoder … Collectively these results suggest that LLMs make strong progress toward solving the typical suite of PBE tasks, potentially increasing the flexibility and applicability of PBE systems.*

*** Code Repair with LLMs gives an Exploration-Exploitation Tradeoff**, Hao Tang Cornell, Keya Hu … Kevin Ellis (7), Cornell University, Shanghai Jiao Tong University, 29.10.2024 – refinement, explore-exploit tradeoff: n arm-acquiring bandit problem, Thompson Sampling, loop invariant synthesis, visual reasoning puzzles, and competition programming problems. Correct, repair, or debug its initial outputs; a tree of possible programs; branching factor is infinite; transition function is stochastic; "rollouts" would demand a prohibitively expensive series of LLM calls to refine down to some maximum depth. REx (REfine, Explore, Exploit); a multiarmed bandit framing: different actions correspond to refining different programs; reward corresponds to the quality of a newly generated program; and maximizing discounted future reward corresponds to solving the programming problem in the minimum number of LLM calls.
Bandit problems [multi-armed bandit: maximize the reward of a set of possible actions] Thompson Sampling – a probabilistic method for solving bandit problems; it updates beliefs (prior/posterior) about each arm's reward distribution, Bayes rule; at each step the model samples the beliefs and pulls the arm with the highest expected reward. Loop Invariant Synthesis: formal verification; automatically find invariant conditions, which are always true during a loop execution. Z3 theorem: check invariants against three criteria: precondition, induction (invariant holds after each iteration) and postcondition. Solver. … Tasks: ***Bandits and Tree Search.*** *[classic] tree search algorithms introduce an explore-exploit tradeoff; Monte Carlo Tree Search (MCTS) is the canonical tree*

search algorithm that relies on this insight. However, REx has a different structure: no *rollouts, backups, or tree traversals. Instead: … the refinement process* [in the LLMs]*: any node can be further expanded (infinitely), every node can be treated as a terminal node, and node expansions are very expensive (so rollouts would not be practical). To the best of our knowledge, it is not possible to apply MCTS or any of its standard variants [35, 36, 37] to our problem statement out-of-the-box because of these unique properties. Bandit-based algorithms have also been applied for fuzz testing [38, 39], where code coverage is more important and code mutation is performed instead of code refinement* …. Competition programming APPS dataset. Visual reasoning puzzle (ARC, Chollet) … ***Limitations and Future Directions****: only modestly more effective at actually solving more problems overall in the large-compute limit: Its advantages are largest when viewed through the lens of minimizing cost, and of being robust across hyperparameters and datasets.*
See the sample tasks in the APPS, sample ARC problems and their solutions etc. [numpy matrices in Python…]; Human-Written Hypotheses; non-linear polynomial arithmetic, LLM invariant gen. …

Worldcoder, a model-based LLM agent: Building world models by writing code and interacting with the environment Hao Tang, Darren Key, Kevin Ellis, 16.12.2024, 65 pages.
https://proceedings.neurips.cc/paper_files/paper/2024/file/820c61a0cd419163ccbd2c33b268816e-Paper-Conference.pdf *model-based agent that builds a Python program representing its knowledge of the world based on its interactions with the environment. The world model tries to explain its interactions, while also being optimistic about what reward it can achieve. … a logical constraint between a program and a planner.* Gridworlds, task planning – *more sample efficient than deep RL, more compute-efficient than ReACT LLM agents and can transfer knowledge by editing its code. Related work p.8 "Exploration & Optimism in the face of (model) uncertainty" Programs as world models. Programs as Policies. LLMs for decision-making; LLMs for building world models; LLMs as a world model; Neurosymbolic world models, World models. …* Appx p. 18-20 – ***Data-consistent reward-func*** *(random guess given the textual mission)* ***Data-consistent transit-func*** *(figured out how to rotate) …* ***Goal-driven transit-func*** *(imagine how to pickup, toggle, and drop, without interactions) … p.26 Planning: MCTS … Minigrid, Sokoban – synthesized (transition, reward) function (p.30); p.31: synthesized world model for Alfworld. Prompts … (Initialize, run, refine) (reward, transit-func). Gener. Reward func. for new goals. Refine to satisfy optimism under uncertainty …* * https://haotang1995.github.io/
Cmp, see: * **Code as Reward: Empowering Reinforcement Learning with VLMs,** David Venuto, Sami Nur Islam, Martin Klissarov et al., 2.2024. https://arxiv.org/pdf/2402.04764
*** On the Modeling Capabilities of Large Language Models for Sequential Decision Making**" by **Martin Klissarov**, Devon Hjelm, **Alexander Toshev** and Bogdan Mazoure, 10.2024 – вж. А.Тошев. https://arxiv.org/abs/2410.05656 See "Мартин Клисаров".

* **Dreamcoder**: **Bootstrapping inductive program synthesis with wake-sleep library learning**, K.Ellis et al. (9), 2021 https://dl.acm.org/doi/pdf/10.1145/3453483.3454080 Wake-sleep cycle neural search … structure hypothesis – program sketch, DSL etc. – *make the program search tractable by restricting it to a limited space of possible expressions…* Problem-solving domains: *List processing, Text-editing, Regexes, LOGO graphics, Block Towers, Symbolic Regression, Recursive Programming, Physical Laws …* building a library from the initial primitives *to more advanced functions to form a hierarchically organized layers of functions. A training corpus* of synthesis problems and a *base language***.** Two cycles: Wake: *Recognition network (RN)*– an auxiliary one which proposes latent variables; Sleep: trains it back with generated samples. The recognition network learns to map specifications to programs, while the generative model is a probability distribution over programs (latent variables). RN – a neural search policy. Abstraction Sleep and DreamSleep. *@Vsy: Obrb; @Bcu: Обрб*
* See the appendix **#anelia** for **Alex Toshev's** work and the main volume for **Martin Klissarov's**.

* **The „wake-sleep" algorithm for unsupervised neural networks.** Geoffrey E Hinton, Peter Dayan, Brendan J Frey, and Radford M Neal. 1995. Science 268, 5214 (1995), 1158-1161.

***** Robert John Henderson. 5.2014. **Cumulative learning in the lambda calculus. Ph.D. Dissertation.** Imperial College London. https://doi.org/10.25560/24759 https://spiral.imperial.ac.uk/entities/publication/3ed1e305-2f76-49e7-bfff-279945f78aeb **"cumulative learning" –** the automatic acquisition of the knowledge necessary for solving harder problems through experience of solving easier ones; p.4(9) "Progression to more difficult problems is made possible via the *incremental accumulation of knowledge.";* inductive programming; the technique of abstraction: lambda calculus rather than first-order logic; however, the first-order Inductive Logic Programming (ILP) is more mature; p.11 (5) *"The standard mechanism for 'progressing to more difficult problems' is the input of human expertise."* However, ML systems *"that accumulate their own expertise and domain-specific knowledge through problem-solving experience … will be more flexible and … applicable in a much wider variety of situations without the need for reprogramming or tuning by a human operator." .. the dream of 'strong AI' or 'artificial general intelligence' ..* ***Since the 1970s, strong AI has been an emotive, almost taboo topic among artificial intelligence researchers'*** *.. However, in recent years some prominent AI scientists are starting to talk about strong AI as a serious research topic in public again [Bengio et al., 2013; Hutter, 2012; Schmidhuber et al., 2011; Maei and Sutton, 2010]. "*

* P. Michelucci and D. Oblinger. **Cumulative learning**. In C. Sammut and G. Webb, editors, Encyclopedia of Machine Learning, pages 249–257. Springer, 2010 https://www.researchgate.net/publication/247935122_Cumulative_Learning … +

inductive transfer, but in this context "the transfer of *knowledge* to new tasks, *not the underlying learning algorithms*"; related: meta-learning, learning to learn *CONFUCIUS (Cohen, 1978; Cohen & Sammut, 1982) – *an instructor teaches the system concepts, long term memory; new concepts are matched against stored concepts – re-describe the examples in terms of the background knowledge* … the system becomes capable to describe complex objects more compactly with the accumulation of more knowledge. Synonyms: Continual learning, Lifelong learning, Sequential inductive transfer. Learning: Agency, Utility, Domain knowledge acquisition (deliberative – selected based on estimated utility or reflexive – all); Task awareness – identify, recognize the beginning and end of a new task… *Estimated sample complexity, number of exemplars, generalization accuracy;* efficacy: whether inductive bias improves generalization; efficiency: comput.time with or without inductive bias.; instructional computing; Indexing method – comparison process, either by exemplars or by representatons (where the authors mean NNs-like weights or a vector); indexing efficiency. Knowledge representation – functional: exemplars or representational: hypotheses. Retention efficacy vs efficiency = accuracy vs comput./memory cost. Meta-knowledge. … MTL – multi-task learner, Caruana 1993. Long-term memory & Short-term memory (LTM, STM). Inductive(bias, (parallel, sequential) transfer); explanation-based learning; (Agency & Utility) (Single, Task-based)(learning method selection)

**** Claude Sammut Geoffrey I,. Web (ed.) Encyclopedia of Machine Learning, 2011***

*** Learning Libraries of Subroutines for Neurally–Guided Bayesian Program Induction**, Kevin Ellis, Lucas Morales et al., Part of Advances in Neural Information Processing Systems 31 (NeurIPS 2018) https://papers.nips.cc/paper_files/paper/2018/file/7aa685b3b1dc1d6780bf36f7340078c9-Paper.pdf – Learning polynomials (smooth graphs) and rational functions (sharp discontinuous like cardiogram) … DSL; number of continuous degrees of freedom… *(* real (+ x real));* explore-compile, explore-compress (learn a generative model); EC^2, bidirectional GRU …

*** Learning to infer graphics programs from hand-drawn images**, Kevin Ellis, Daniel Ritchie, Armando Solar-Lezama, Josh Tenenbaum, 2018 , NIPS https://proceedings.neurips.cc/paper/2018/file/6788076842014c83cedadbe6b0ba0314-Paper.pdf ~LaTeX, DL+program synthesis (PS); CNN – plausible primitives; PS to recover a graphics program; variable bindings, iterative loops, conditionals; correct & extrapolate graphics with the program

*** Top-down synthesis for library learning,** Matthew Bowers, Theo X Olausson, Lionel Wong, Gabriel Grand, Joshua B Tenenbaum, Kevin Ellis, Armando Solar-Lezama, 2023 https://dl.acm.org/doi/pdf/10.1145/3571234 - corpus-guided top-down synthesis, DSL. *The algorithm builds abstractions directly from initial DSL primitives, using syntactic pattern matching of intermediate abstractions to intelligently prune the search space and guide the algorithm towards abstractions that maximally capture shared structures in the*

corpus. "**Stitch**", parallel Rust, alg. comp. *to Deductive libraries of DreamCoder: 1000 – 10000 times faster and 100 times less memory with better or comparable quality (compressivity). Initial DSL: connect | transform | matrix | circle | line | 0 | 1 | 2 → (connect (connect (transform (repeat (transform line (matrix 1 0 -0.5 (/ 0.5 (tan (/ pi 8))))) ... …Stitch: Draws polygons parameterized by number of sides and side length → learned_fn – then use it as an abstractin with the basic ones ("connect", "transform", …)* Strict Dominance → Pruning .. Complete or Partial Abstraction → Max utility; Holes; Grammars; Library. Corpus Guided Top-down Search and Synthesis (CTS). Match Locations, Corpus Grammar; Rewrite strategy ~ Branch & Bound; Top-down; Rewrite Strategy Expansion; Abs; Var; Hole; Downshift. Lambda Unify … Matches(P, A???) = Partial Abstractions … Incremental Match; Context-expression pairs. Redundant Arguments Elimination and Deductive Rewriting. Initial DSL … → Functions → Replace with the new functions (compress the representation, the code , …) learned_fn_0 = (lx.ly(π…(…())) … Learned library s new abstractions … Rewritten compressed programs …

*** Write, Execute, Assess: Program Synthesis with a REPL, Kevin Ellis, Maxwell Nye et al., 2019,** https://www.sciencedirect.com/science/article/pii/S0004370221000722 https://proceedings.neurips.cc/paper_files/paper/2019/file/50d2d2262762648589b1943078712aa6-Paper.pdf Advances in neural information processing systems Context-free grammar, Markov decision process (CFG,MDP): State, action, transition, reward. Code-Writing Policy π and the Code-Assessing Value v. Interleave code writing & code assessing. Sequential Monte Carlo. Inverse CAD (3D). 2D graphics: P→E→circle| quadrilateral … 3D: sphere, cube, cylinder (radius, …) … **String Editing Programming Language** p.11: Program, Expression, Nesting, Regex; Type, Case, Delimiter, Index, Boundary: Start | End. 2-3 days training on P100 GPU, millions of examples for the 3D CAD. 24000 x 4000 for string editing (iterations x batch size) … →
*** Diffusion On Syntax Trees For Program Synthesis**, Shreyas Kapur, Erik Jenner, Stuart Russel, University of California, Berkeley, 30.5.2024 https://arxiv.org/pdf /2405.20519 Reverse graphics programs – Diffusion for discrete data & code generation; Program synthesis for inverse graphics;

*** From perception to programs: regularize, overparameterize, and amortize,** Hao Tang, Kevin Ellis, 7.2023 https://proceedings.mlr.press/v202/tang23c/tang23c.pdf Synthesizing neurosymbolic programs; mixing discrete symbolic processing with continuous neural computation.

*** Program Synthesis with Pragmatic Communication**, Yewen Pu, Kevin Ellis, Marta Kryven, Joshua B. Tenenbaum, Armando Solar-Lezama (MIT), 2020 a pragmatic program synthesizer .. asks why an informative user would select that specification – for resolving the ambiguity in program synthesis; inductive biases; recursive pragmatics incremental pragmatic model via version space algebra. *a reference game, a speaker-listener pair (S, L) cooperatively communicate a concept h ∈H using some atomic*

utterances u ∈ U – see also SHRDLURN [23]

* **Making sense of raw input**, Richard Evans, Matko Bošnjak, Lars Buesing, Kevin Ellis, David Pfau, Pushmeet Kohli, Marek Sergot, 1.10.2021 – apperception task; parsing input; post-hoc interpretation vs unsupervised learning – take a *temporal sequence of raw unprocessed sensory information* and produce *an interpretable theory capturing the regularities in that sequence. Apperception Engine .. Finding the most probable interpretation .. binary NN – BNN.*

*** Library Learning for Neurally-Guided Bayesian Program Induction,** Kevin Ellis, Lucas Morales, Mathias Sablé-Meyer, Armando Solar-Lezama, Joshua B. Tenenbaum, 2018: EC^2 - learning a DSL while jointly training a NN to efficiently search for programs in the learned DSL. Lists, edit text, solving symbolic regression problems (find a formula for a graph of a function,a numerically given function). … grows or bootstraps a DSL. Lists: higher-order filter functions, maximum element in list, ? contains (k) … character substitutions; drop a character until reaching a specific other char, abbreviate a sequence of words (LISP-like code) … polynomials of different orders … *fo(x) = (+ x real), f1(x) = (f0 (* real x))* … program induction. *"In contrast to computer assisted programming [7] or genetic programming [8], our goal is not to automate software engineering, to learn to synthesize large bodies of code, or to learn complex programs starting from scratch. Ours is a basic AI goal: capturing the human ability to learn to think flexibly and efficiently in new domains."* Human-like: *skilled coders build libraries of reusable subroutines that are shared across related programming tasks*. EC^2 = ECC = Explore/Compress/Compile. *Enumeration*, instead of using sophisticated algorithms for program search. Why? A general approach *that doesn't require special conditions on the space of programs.* Prior work: Metagol, FlashFill (hand-engineered DSL); neural: RobustFill; hybrid: neural-guided deductive search; hybrid: DeepCoder. *A frontier of task x* – explore… 2017 SyGuS competition; EC^2 learns from and *critically needs a corpus of training tasks (…)* proposing small changes … *Fragment Grammars [17] and Tree-Substitution Grammars [27], and is closely related to the idea of antiunification [28, 29].* 6. Related work: Schmidhuber's OOPS model; MagiHaskeller. … *7. Another direction is to explore DSL meta-learning: can we find a single universal primitive set that could effectively bootstrap DSLs for new domains, including the three domains considered, but also many others?*

* **Bootstrap learning via modular concept discovery**. Eyal Dechter, Jon Malmaud, Ryan P. Adams, and Joshua B. Tenenbaum. In IJCAI, 2013. (EC) https://www.ijcai.org/Proceedings/13/Papers/196.pdf Starting with a stochastic grammar, learning to compress it, exploring unknown spaces. A **frontier**: finite set of expressions to be evaluated in an iteration; a **frontier size** N – number of elements in the fr. The task is **"hit" by the frontier** if it is solved by an expression found in the frontier. Exploring the frontier by *enumerating the N most probable expressions from the current distribution D; compressing the hit tasks to estimate a new distrib.* 3.1. **A basis of combinators,** primitive combinators – identity, composition, … lambda calculus: application &

abstraction; combinatory logic; variable routing of the basis combinators, thus a variable-free representation. Program = binary tree. 3.2. Stochastic Grammar over Program – Feldman et al. 1969 (SG) … C = c1,...,cN – primitive combinators; D = p1, … , pN – prior probability distribution of each ci. Products of prob.; maximum likelihood estimation; … 3.3 Best-first Enumeration of Programs 3.4. The most compressive set of solutions: number of unique trees 3.5. Re-estimating the SG. 4.2. Boolean function learning …

* **Grammatical complexity and inference.** Jerome A. Feldman, James Gips, James J. Horning, and Stephen Reder. Stanford University, 1969. https://www.researchgate.net/publication/265065712_Grammatical_Complexity_and_Inference induction, infer grammars, grammar-grammar; universal terminal alphabet & universal variable alphabet – sets of symbols; e - empty string; empty set. S* - finite strings context-free grammar CFG: G = (V,T,X,P) … production rules Z->w … intermediate string ; leftmost derivation d(y,w,G); G – totally reduced if no rules: Z->e, Zi → Zj … e-free language … a positive information sequence – each sentence in the language occurs in the sequence → frequency distribution; relative freq. … n(X) X finite set of objects (strings) – cardinality *r = n(T) – number of terminal symbols in the alphabet T* … P.17: Grammar complexity, measures …. P.43 Grammatical inference: *"Many definitions of learnability are possible …"* [Gold 67] – derives from [Gold 65] – limiting recursion. … p.54 *if a grammar is too large, there must be some redundant rules*. Ch.4. Programs for Grammatical Inference; a pivot grammar

* **Human-level concept learning through probabilistic program induction**. BM Lake, R Salakhutdinov, JB Tenenbaum. Science 350 (6266), 1332-1338, 2015 https://www.cs.cmu.edu/~rsalakhu/papers/LakeEtAl2015Science.pdf – humans can generalize from a single example, while machines requires tens or hundreds …
A generative model of handwritten characters., Fig. 3., p.3.1334: GenerateType: sample in a sequence (number of parts, number of subparts, sub-part sequences); sample relation → gen. token → return @generate_token; procedure generate_token: add motor variance, sample part's start location, compose a part's trajectory … sample affine transform; sample image; how a human and the machine parse the drawing

*See also **patents,** samples of their style and content, by Kevin Ellis et al.*

* **Behavior feature use in programming by example,** Inventors: Sumit Gulwani, Kevin Michael Ellis; publ.date: 2020/6/30, Patent office US, Patent number 10698571; Application number 15394238
https://patentimages.storage.googleapis.com/4d/62/4c/ed533491e396f9/US10698571.pdf

Also: **Temporal rule-based feature definition and extraction** (executable traces etc.) https://patents.google.com/patent/US8538909B2/en Microsoft *"A temporal rule-based feature extraction system and method … temporal-based rules satisfied by a trace. Once*

a temporal-based rule is found that is satisfied by the trace, then embodiments of the temporal rule-based feature extraction system and method leverage that rule to either use as a feature or to extract additional features. The extracted feature then is used to characterize the trace. Embodiments of the system include" … extrinsic & intrinsic features similarity measures of two traces…

*** Large Language Models for Software Engineering Survey and Open Problems,** Angela Fan, Beliz Gokkaya, Mark Harman, Mitya Lyubarskiy et al. 11.2023 https://arxiv.org/pdf/2310.03533 – the output from an LLM is not only code, but also other software engineering artefacts: requirements, test cases, design diagrams, documentation etc. explanation. Code Generation. Prompt Engineering for Improved Code Generation. Hybrids of LLMs and other Techniques. Scientific Evaluation of LLM-based Code Gen. Generating New Tests Using LLMs. Test Adequacy Evaluation. Test Minimisation. Test Output Prediction.Test Flakiness. Debugging and Repair. Performance Improvement. Clone Detection and Reuse. Refactoring
Existing large language models for code generation: CodeBERT February 2020 Microsoft 125M, CodeX August 2021 OpenAI 12B, Copilot October 2021 Github and OpenAI 12B … InCoder April 2022 Meta 6.7B, 1.3B …

*** Bulgarian Researchers:** Boian Alexandrov et al.

*** Enhancing Code Translation in Language Models with Few-Shot Learning via Retrieval-Augmented Generation,** Manish Bhattarai**, …, Boian Alexandrov** et al., 7.2024 Fortran to C++ https://arxiv.org/pdf/2407.19619 ; CFortranTranslator: https://github.com/CalvinNeo/CFortranTranslator Fortran90/Fortran77 code to C++14

*** HEAL: Hierarchical Embedding Alignment Loss for Improved Retrieval and Representation Learning**, Manish Bhattarai, … **Valentin Stanev, Vladimir Valtchinov, Boian Alexandrov** et al. 5.12.2024 https://arxiv.org/pdf/2412.04661
* https://cnls.lanl.gov/External/people/Boian_Alexandrov.php - Los Alamos Center for Non-linear studies

* Topics: Transformer architectures for images and for reducing the quadratic complexity

*** Jakob Uszkoreit:** "the father" of the Transformer
https://scholar.google.com/citations?user=mOG0bwsAAAAJ&hl=de

*** Image Transformer**, Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Łukasz Kaiser, Noam Shazeer, Alexander Ku, Dustin Tran, 2018
https://proceedings.mlr.press/v80/parmar18a/parmar18a.pdf

* **Music transformer**, Cheng-Zhi Anna Huang, Ashish Vaswani, Jakob Uszkoreit, Noam Shazeer, Ian Simon, Curtis Hawthorne, Andrew M Dai, Matthew D Hoffman, Monica Dinculescu, Douglas Eck, 9.2018 https://arxiv.org/pdf/1809.04281
"Self-attention over its own previous outputs allows an autoregressive model to access any part of the previously generated output at every step of generation. By contrast, recurrent neural networks have to learn to proactively store elements to be referenced in a fixed size state or memory, potentially making training much more difficult. We believe that repeating self-attention in multiple, successive layers of a Transformer decoder (Vaswani et al., 2017) helps capture the multiple levels at which self-referential phenomena exist in music."; relation-aware version of self-attention

*** Set Transformer: A Framework for Attention-based Permutation-Invariant Neural Networks,** Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, Yee Whye Teh, 10.2018/5.2019
https://arxiv.org/pdf/1810.00825

*** A decomposable attention model for natural language inference**. Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 25.9.2016. https://arxiv.org/pdf/1606.01933 - Self-attention; *"decompose the problem into subproblems that can be solved separately, thus making it trivially parallelizable";* alignment in statistical machine translation; *… The neural counterpart to alignment, attention (Bahdanau et al., 2015)*"- see the reference above, with Bengio …

*** AN IMAGE IS WORTH 16X16 WORDS: TRANSFORMERS FOR IMAGE RECOGNITION AT SCALE,** Alexey Dosovitskiy , Lucas Beyer , Alexander Kolesnikov , Dirk Weissenborn , Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby, 6.2021 https://arxiv.org/pdf/2010.11929/1000
Vision Transformers – ViT [Find below a section on ViT]

* **Transformers are RNNs: Fast Autoregressive Transformers with Linear**

Attention. Angelos Katharopoulos et al., 8.2020 https://arxiv.org/abs/2006.16236 cmp. to the "quadratic attention" of the transformer (global receptive field, similarity, softmax of NxN matrix); other methods: weight pruning, weight factorization, knowledge distillation; replaced token detection; product-key attention; Transformer-XL (2019) https://towardsdatascience.com/linearizing-attention-204d3b86cc1e

*** Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context, 6.2019** Zihang Dai, Zhilin Yang, Yiming Yang, Jaime Carbonell, Quoc V. Le, Ruslan Salakhutdinov https://arxiv.org/abs/1901.02860 - captures longer term dependencies beyond the context window …

* Topics: Evolutionary Algorithms | Genetic Algorithms | Genetic Programming – Part II

Concepts: Metaheuritic, search algorithm, stochastic (partly random, Monte Carlo, stochastic approximation (SA), stochastic gradient descent, finite-difference SA, scenario optimization …), Genetic programming (GP) – Koza, 1992; Cartesian genetic programming: spatial grid (2000) , Grammar-guided GP (constraints) – pruning in initialization; Gene expression programming (2001); Multi-gene GP … Geometric semantic GP – direct search in the space of underlying semantics of the programs; Surrogate GP (2015), Memetic semantic GP (2015), Statistical GP – for generation of well-structured subtrees (2017); Multi-dimensional GP – novel program representation … Search: linear, binary, and hashing. Optimization: local & global https://en.wikipedia.org/wiki/John_Koza | https://www.genetic-programming.com/johnkoza.html

* 1,000-Pentium parallel computer in Mountain View:

* **1,000-Pentium Beowulf-Style Cluster Computer for Genetic Programming:** https://www.genetic-programming.com/machine1000.html (1000 x Pentium II 350 MHz x 64 MB RAM = 64 GB in total … 500 dual-CPU Tyan Tiger 100 (ATX form factor); a server with 256 MB & 14 GB hard disk … 100 Mbit LAN; Red Hat Linux 6.0 …25x40 toroidal grid; ~ ½ or 32 MB per node for the population (30000 individuals, each x 3000 bytes …) A 1000-node system: 10M x 3000 individuals, a fitness evaluation in 0.10 sec, for 10K individuals – 1000 seconds. One generation: 15 minuts, 96 generations per day. Migration – low-bandwidth … ~ 12000 SPECft95 … <1% low-bandwidth migration. An earlier cluster of 70 machines with DEC Alpha CPUs. A related work of smaller scale:

* Bennett, Forrest H III, Koza, John R., Shipman, James, and Stiffelman, Oscar. 1999. **Building a parallel computer system for $18,000 that performs a half peta-flop per day.** In Banzhaf, Wolfgang, Daida, Jason, Eiben, A. E., Garzon, Max H., Honavar, Vasant, Jakiela, Mark, and Smith, Robert E. (editors). 1999. *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference, July 13-17, 1999, Orlando, Florida USA*. San Francisco, CA: Morgan Kaufmann. Pages 1484 - 1490. https://www.researchgate.net/publication/220740439_Building_a_Parallel_Computer_System_for_18_000_that_Performs_a_Half_Peta-Flop_per_Day – 10-node DEC Alpha 21164 533 MHz x 64 MB RAM, **“half PFLOP” per day on runs of GP.** Perhaps ~ Alpha 21064, 188 SPECft95. Peak performance per single CPU and node: 2 x 1.066 GIPS + 1.066 GFLOPS. (In FLOP that is theoretical ~ 0,255 PFLOP/day=24 h peak + the integer (twice))
Tosh: Compare to the later generation PCs and GPU/matrix/tensor accelerated clusters.

Summary and selection of important concepts in Evolutionary Algorithms | Genetic Algorithms | Genetic Programming etc. based on sources from **Wikipedia** etc.
Authors: Wikipedia authors et al.
Selection and Edited by: Todor Arnaudov, 27-29.4.2025

Note: The text flows first in the left column, then returns to the top of the right column.

https://en.wikipedia.org/wiki/Cartesian_genetic_programming - graph representation to encode programs.
https://en.wikipedia.org/wiki/ … @Vsy: **Concatenate with the titles or use the URLs,** substitute underscore _ as space " " when reading etc.**:** Search_algorithm | Binary_search | State-space search: uninformed (depth-first, breadth-first, iterative deepening, lowest-cost-first search/Uniform-cost search) …, informed search (Heuristic depth-first search, Greedy best-first search, A* search: SRI Stanford 1968, extension of Dijkstra, shortest path to a goal (not all paths) – Shakey the robot) … State space planning – forward and backward search … forward: from the start state to the target/goal/final state, backward – from the final state to the initial state (back propagation).
State_space_(computer_science)
UCS – Uniform cost search: frontier … bidirectional variants … shortest paths (reach-based routing), hierarchical decompositions … transit nodes, transit node routing (speed up shortest-path routing by precomputing access nodes in sub-networks (subdivision) – distances to intermediate nodes); Shortest_path_problem | Shortest-path_tree | Contraction_hierarchies
https://en.wikipedia.org/wiki/Dijkstra%27s_a

Grammatical_evolution
Genetic_representation
Genetic_operator
Evolutionary_pressure
Genetic_drift – founder effect ..
Genetic_code
Meme | Memetic_algorithm
Cultural_algorithm – Belief spaces (knowledge): Normative (desirable value ranges for the individuals; acceptable behavior in the population), Domain specific (related to the area of application and the problem), Situational (events, successful and unsuccessful solutions), temporal (history of the search space – temporal patterns of the search process), Spatial knowledge – the topography of the search space
Fitness_function – objective or cost function that is used to summarize, as a single figure of merit, how close a given candidate solution is to achieving the set aims. .. In Evolutionary algorithms (EAs): each candidate solution, also called an *individual* is commonly represented as a string of numbers ("chromosome"). … niche differentiation, co-evolvng, fitness landscape … weighted sum, penalty functions. Pareto optimization – *Pareto-optimal if the improvement of one objective is only possible with a deterioration of at least one other objective.* Pareto set of objectives (f1, f2, …) to be maximized, constraints. Objectives: Primary objectives, Auxiliary objectives.
Loss_function |
Test_functions_for_optimization
https://en.wikipedia.org/wiki/Fitness_landscape or adaptive landscape: the relationship(genotypes, reproductive success). *Fitness: Height of the landscape. Similar genotypes – close to each other, different – far from each other.*
Inferential_programming - *describing an intended result to a computer, using a metaphor such as a fitness function, a test*

lgorithm#Practical_optimizations_and_infinite_graphs | Bidirectional_search
Genetic_programming | Meta-genetic programming |
Genetic_improvement_(computer_science) | Search-based_software_engineering |
Metaheuristic – *a higher-level procedure or heuristic designed to find, generate, tune, or select a heuristic (partial search algorithm) that may provide a sufficiently good solution to an optimization problem or a machine learning problem, especially with incomplete or imperfect information or limited computation capacity* .

Glover, F.; Kochenberger, G.A. (2003). Handbook of metaheuristics. Vol. 57. Springer, International Series in Operations Research & Management Science. ISBN 978-1-4020-7263-5.
Blum, Christian; Roli, Andrea (2003). „Metaheuristics in combinatorial optimization: Overview and conceptual comparison“. *ACM Computing Surveys*. **35** (3). ACM: 268–308. doi:10.1145/937503.937505

Matheuristics – *problem agnostic optimization algorithms that make use of mathematical programming (MP) techniques in order to obtain heuristic solutions. Problem-dependent elements are included only within the lower-level mathematical programming, local search or constructive components.*
Rider_optimization_algorithm – *ROA, attacker, overtaker, follower, and bypass rider.*
* Binu D and Kariyappa BS (2019). *„RideNN: A new rider optimization algorithm based neural network for fault diagnosis of analog circuits“*.
Local_search_(optimization) – problems that can be formulated as finding a solution that maximizes a criterion among a number *specification, or a logical specification, and then the computer, on its own, would construct a program needed to meet the supplied criteria.*

Population_model_(evolutionary_algorithm) – the set of all proposed solutions in an EA in one iteration – individuals. Island models – subpopulations, size, neighbourhoods, criteria for the termination of an epoch; sync or async migration, migraion rate, migrant selection. Neighbourhood models – diffusion model or fine grained model, topological relations. Locality – isolation by distance. Cellular Eas, cellular genetic alg. cGA. *A commonly used structure for arranging the individuals of a population is a 2D toroidal grid*
https://en.wikipedia.org/wiki/Panmixia - global or panmixic model, panmicticism – uniform random fertilization; ***random mating*** *usually implies the hybridising (mating) of individuals regardless of any spatial, physical, genetical, temporal or social preference. … individuals are able to move about freely within their habitat, possibly over a range of hundreds to thousands of miles, and thus breed with other members of the population.*
Differential_evolution
Gaussian_adaptation – normal or natural adaptation (in GA – mean fitness); climbing a mountain example : with a normalization using a Gaussian function around the current location
Coevolution – mutualism
* Mathematical_optimization …
Global_optimization; branch and bound methods (BB, B&B) – discrete & combinatorial optimization problems: systematic enumeration of candidate solutions by menas of state space search … rooted tree … interval methods .. cutting-plane methods …

of candidate solutions. | https://en.wikipedia.org/wiki/Feasible_region#Candidate_solution Iterated_local_search - a modification of local search or hill climbing; "kicking" a solution out from a local optimum; perturbation algorithm, (adaptive,optimizing) perturbation. *A simple modification: iterating calls to the local search routine, each time starting from a different initial configuration – repeated local search.* Variable_neighborhood_search – VNS, more distant neighborhoods of the current incumbent solution … neighborhood change: descent to local minima and escape from valleys containing local minima. Guided_local_search – a meta-heuristics; builds up penalties during a search; Selective penalty modifications; searching through an augmented cost function .. Tung-leng Lau – *guided genetic programming (GGA) algorithm. It was successfully applied to the general assignment problem (in scheduling), processors configuration problem (in electronic design) and a set of radio-link frequency assignment problems (an abstracted military application).* * Davenport A., Tsang E.P.K., Kangmin Zhu & C J Wang, GENET: A connectionist architecture for solving constraint satisfaction problems by iterative improvement, Proc., AAAI, 1994, pp. 325-330 Greedy_randomized_adaptive_search_procedure Greedy randomized adaptive search procedure (GRASP): restricted candidate list (RCL). Semi-greedy heuristic, … cost perturbations, bias functions, memorization and learning, and local search on partially constructed solutions… *Phase I: the constructive phase, a feasible solution for the entire problem is constructed in a stepwise manner.* Then it is used for the local improvement Phase II. Gradient_descent – multi-variable	Schema_(genetic_algorithms) – a subset of strings with similarities at certain string positions, a template |* Formal_concept_analysis – deriving a *concept hierarchy* or formal ontology from a collection of objects and their properties. .. context's „concept lattice". "Restructuring Lattice Theory",Rudolf Wille, 1982. "Bodies of water" … Software: ConExp[19], ToscanaJ[20], Lattice Miner[21], Coron[22], FcaBedrock[23], GALACTIC[24] .. Biclustering and multidimensional clustering, knowledge spaces Fitness_(biology) - **reproductive success (Tosh: NB!: not just "adaptability" etc.;** *average contribution to the gene pool of the next generation, made by the same individuals of the specified genotype or phenotype. The fitness of a genotype is manifested through its phenotype. … Herbert Spencer's well-known phrase "survival of the fittest" should be interpreted as: "Survival of the form (phenotypic or genotypic) that will leave the most copies of itself in successive generations."* **Tosh:** i.e. in human populations, currently the people which live about the worst, the poor nations in Africa, who reproduce the most, are "the most fitted" in biological sense by these definitions and their genes survive better [27.4.2025]. Propensity_probability – disposition, tendency … to yield a certain kind of outcome; purported causes, not relative frequencies. * Universal_Darwinism – ***@Vsy: ObObr:*** beyond biology, but also in psychology, linguistics, economics, culture, medicine, computer science, and physics etc.: *variation, selection, heredity or retention* … searching for the best solution of how to survive and reproduce by **generating new trials, testing how well they perform, eliminating the**

function , defined and differentiable in a neighborhood of a point a, then F(x) decreases *fastest* if one goes from a in the direction of the negative gradient of F … small enough step size or learning rate …

Constructive_cooperative_coevolution - multi-start architecture of GRASP; decomposition into subproblems; simulation-based optimisation.

Hill_climbing – ridge, alley; plateau, local maxima;

Hyper-heuristic – search in the space of heuristics; heuristics to generate heuristics: HyFlex, ParHyFlex, EvoHyp, MatHH. http://titancs.ukzn.ac.za/EvoHyp.aspx

Multi-objective_optimization – multi-objective programming, vector optimization, multicriteria optimization, or multiattribute optimization, Pareto optimization; multiple-criteria decision making .. trade-offs – two or more conflicting objectives. *A solution is called nondominated, Pareto optimal, Pareto efficient or noninferior, if none of the objective functions can be improved in value without degrading some of the other objective values.* Bicriteria optimization, multitask optimization.

Multiple-criteria_decision_analysis (MCDA, or MCDM – MCD making) multiple attribute utility theory, multiple attribute value theory, multiple attribute preference theory, and multi-objective decision analysis.

Analytic_hierarchy_process - *a structured technique for organizing and analyzing complex decisions, based on mathematics and psychology.* Choice, ranking, prioritization, resource allocation, benchmarking, quality management, conflict resolution … Planning, priority setting, selection among alternatives; forecasting, total quality management, business process reengineering, quality function deployment, balanced scorecard.

failures, and retaining the successes. *"organism" is replaced by any recognizable pattern, phenomenon, or system. … variation and selection, and thus adaptation of: genes, ideas (memes), theories, technologies, neurons and their connections, words, computer programs, firms, antibodies, institutions, law and judicial systems, quantum states and even whole universes. …*

* Inclusive_fitness (expected fitness returns: direct/indirect)

Content-addressable_memory (Associative memory); Data word recognition unit - Dudley Allen Buck, 1955.

https://en.wikipedia.org/wiki/Multi-task_learning exploiting commonalities and differences across tasks. inductive transfer Transfer_learning Domain_adaptation …

Fitness_proportionate_selection

Evolutionary_algorithm → *1. Randomly generate the initial population of individuals, the first generation. 2. Evaluate the fitness of each individual in the population. 3. Check, if the goal is reached and the algorithm can be terminated. 4. Select individuals as parents, preferably of higher fitness. 5. Produce offspring with optional crossover (mimicking reproduction). 6. Apply mutation operations on the offspring. 7. Select individuals preferably of lower fitness for replacement with new individuals (mimicking natural selection).*

Return to 2.

> ***Tosh:*** *It doesn't have to be performed randomly. For complex, hierarchical and yet-unknown future organisms/targets* the current higher fitness is not strictly better – different selection criteria. 28.4.2025

Evolutionary_programming – mutation without crossover. All individuals are

https://en.wikipedia.org/wiki/Analytic_hierarchy_process_%E2%80%93_car_example : Choose the best car: cost – purchase price, fuel costs, maintenance costs, resale value, safety, style, capacity – cargo, passenger … 6 cars: Accord Sedan, Accord Hybrid, … Pairwise comparing the criteria with respect to the goal … Imntensity of importance (1 = equal, 3 = moderate, 5 = strong, 7 = very strong, 9 = extreme) quotient: total = 1.0. (E.g. cost = 0.5m safety = 0.25, style = 0.05, capacity = 0.2) → then subdivide 1.0 for the subcategories (e.g. Cost: 0.25 purchase price, 0.5 fuel cost, 0.25 maintenance cost, 0.1 resale value 0.15) etc. Analytic_network_process Quality_function_deployment – *Japan 1966, customer desires: WHATs, importance, engineering characteristics which may be relevant (HOWs), correlates the two, allows for verification of those correlations, and then assigns objectives and priorities for the system requirements… any system composition level (e.g. system, subsystem, or component) in the design of a product, and can allow for assessment of different abstractions of a system* Balanced_scorecard - management; a strategy performance management tool; strategic agenda, monitor performance against objectives; mix of financial and non-financial data items (perspectives): financial, customer, internal process, learning & growth; a portfolio of initiatives designed to impact performance of the measures/objectives. Strategy_map - the strategic goals being pursued by an organization or management team. Trade-off Vector_optimization Posynomial – any real exponents, integer multipliers/coefficients (cmp. polynomial: non-negative integer exponents, any real numbers for the independent variables and coefficients). → Signomial – signomial programming – a	selected for the new population, while in the evolution strategy ES(μ+λ) the individuals have the same probability to be selected. **Evolution_strategy** (ES) – natural problem-dependent representations, so problem space and search space are identical. … (ES introduces Mutation & Selection 1973) … derandomized self-adaptation … usage information from the old generations .. CMA-ES weighted multi-recombinations … Natural ES – natural gradient … Limited memory CMA-ES .. weighted recombination of general convex quadratic functions CMA-ES - Covariance matrix adaptation evolution strategy (CMA-ES) – numerical optimization | Evolutionary_computation | Gene_expression_programming - Karva language (Karva notation), K-expression trees, +a*acbde, head, tail … open reading frames (ORFs) (DNA genes, start codon, amino acid codons, termination codon). Noncoding regions … Q*-+abcd = SQRT((a-b)*(c+d)) https://www.gepsoft.com/gepsoft/APS3KB/Chapter05/Section1/SS2.htm **Tosh:** Q = Sqrt = Square root. (Stack processing) → pop(operation) = + (plus), pop(arguments) = c,d → push(c+d). Pop(op.) = "minus". Pop(arg) = a,b → (Execute(Push(Pop(Operation, Pop(Arguments))). 01234567890 Q*b**+baQba … fixed length genes, which however can encode expression trees of different sizes and shapes. * **Automatic Problem Solver – APS** (outdated)**:** https://www.gepsoft.com/ GeneXproTools 5.0 *Modeling made easy. "**Automatic Problem Solver** is an extremely flexible modeling tool designed for Function Finding, Classification, and Time Series Prediction"* https://www.gepsoft.com/gepsoft/APS3KB/ https://www.gepsoft.com/gepsoft/APS3KB/C

relaxation of Geometric_programming – optimization problem min(f0(x)) where fi(x) <=1, gi(x) = 1 … f0, … , fm are posynomials, g1, … gp are monomials. … closely related to convex optimization (can be made convex by means of a change of variables). Applications: component sizing in IC design, aircraft design, Maximum_likelihood_estimation estimation for logistic regression, parameter tuning of positive linear systems in control theory … * **Application of Signomial Programming to Aircraft Design**, Philippe G. Kirschen, Martin A. York, Berk Ozturk and Warren W. Hoburg, Published Online:18 Dec 2017 https://doi.org/10.2514/1.C034378 https://arc.aiaa.org/doi/abs/10.2514/1.C034378?journalCode=ja - compute the *optimal sizing of the wing, tail, fuselage, and landing gear of a commercial aircraft. These models are combined together to produce a system-level optimization model.* … * Applications of convex analysis to signomial and polynomial nonnegativity problems. PhD Thesis by Riley John Murray, 2021*https://thesis.library.caltech.edu/14169/5/Riley%20Murray%20thesis%2C%20June%201%20deposit.pdf - sums of arithmetic-geometric exponentials or SAGE approach to signomial nonnegativity.* …

Convex optimization library embedded in Python: https://github.com/cvxpy/cvxpy *CVXPY is an open source Python-embedded modeling language for convex optimization problems. It lets you express your problem in a natural way that follows the math, rather than in the restrictive standard form required by solvers. Convex optimization problems (COPs), mixed-integer COPs, geometric programs, and quasiconvex programs.* It's not a solver and uses other solvers:

hapter10/Section1/SS1.htm
Open source: GEP4J – GEP for Java Project PyGEP – Gene Expression Programming for Python; jGEP – Java GEP toolkit

Fitness functions: 1. Numeric (continuous) predictions, 2. Categorical or nominal (binomial & multinomial) – logistic regressin, 3. Binary or boolean. …

https://en.wikipedia.org/wiki/Linear_genetic_programming – LGP: sequentially ordered and executed instructions, cmp: tree genetic programming (TGP) – usually faster, using CPU registers etc.; | https://en.wikipedia.org/wiki/Fish_School_Search - population based search alg.: simple computations in all individuals ("fish"), means of storing info.: weights, school barycenter; local computations, low communications between neighboring individ., mi.centralized control, some distinct diversity mechanisms
, scalability, autonomity …

* Application of chaotic Fish School Search optimization algorithm with exponential step decay in neural network loss function optimization, A. Demidova, A.V. Gorchakov, 2021https://www.sciencedirect.com/science/article/pii/S187705092100987X
* Weight-based Fish School Search algorithm for Many-Objective Optimization, Fernando Buarque de Lima Neto et al. 1.2019 https://arxiv.org/pdf/1708.04745 - decomposition of the original problem in clusters; .. penalty-based boundary intersection. ..

https://en.wikipedia.org/wiki/Grammatical_evolution *GE was originally a combination of the linear representation as used by the Genetic Algorithm for Developing Software*

Clarabel, SCS, and OSQP etc. The project started at Stanford University. **Solvers:**
https://github.com/oxfordcontrol/Clarabel.rs
https://github.com/bodono/scs-python
https://github.com/osqp/osqp
CVXCORE (C++), Scipy, Numpy

Polytope - flat faces (Dimensions: -1: "nullity", 0: Vertex, 1: Edge, 2: Face, 3: Cell … j: j-face … Peak (n-3) … Ridge or subfacet (n-2), Facet (n-1), (n-1)-face .. n: the polytop itself (the whole). Polygon: 2D, polyhedron: 3D.
Convex_polytope convex hull … full-dimensional if n-dim.obj. in R^n
Lasso_(statistics) - Least absolute shrinkage and selection operator; also Lasso, LASSO or L1 regularization; both variable selection and regularization; assumes sparse coeff.
Dependent_and_independent_variables
Feature_learning – automatically discover the representations needed for feature detection or classification from raw data. This replaces manual feature engineering and allows a machine to both learn the features and use them to perform a specific task. Supervised dictionary learning, Neural networks; Unsupervised: K-means clustering, Principal component analysis, Local linear embedding**,** Independent component analysis, Unsupervised dictionary learning. Restricted Boltzmann machine. Autoencoder. **Self-supervised**: unlabeled data, produce deep feature representations: contrastive, generative or both; then fine-tuning. Contrastive: positive & negative samples. Text: Word2vec is a word embedding .. GPT, BERT … [**Tosh:** Words, tokens: not concepts and thoughts] Doc2vec. Image: techniques: transformation, inpainting, patch discrimination, clustering; context encoders ImageGPT, 2020* … Graph –node2vec – random walks, extends wav2vec; maximize mutual information: a measure of similarity; representation of a patch around each node, a summary representation of the entire graph.

(GADS)[6] and Backus Naur Form grammars, which were originally used in tree-based GP by Wong and Leung[7] in 1995 and Whigham in 1996.[8]. Other related work noted in the original GE paper was that of Frederic Gruau,[9] who used a conceptually similar "embryonic" approach, as well as that of Keller and Banzhaf,[10] which similarly used linear genomes.
GRAPE, Python, 2022
https://github.com/bdsul/grape
GRAPE: Grammatical Algorithms in Python for Evolution. GRAPE is an implementation of Grammatical Evolution (GE) in DEAP, an Evolutionary Computation framework in Python. https://github.com/bdsul/grape

GRAPE: Grammatical Algorithms in Python for Evolution by Allan de Lima 1ORCID,Samuel Carvalho 2ORCID,Douglas Mota Dias 1,3ORCID,Enrique Naredo 1ORCID,Joseph P. Sullivan 2ORCID andConor Ryan 1,*ORCID
https://www.mdpi.com/2624-6120/3/3/39
"*…the genotypic level, which ensures that the respective phenotypes will always be syntactically correct.*"; genotype-phenotype mapping process

* PonyGE2: Grammatical Evolution in Python
Michael Fenton, James McDermott, David Fagan, Stefan Forstenlechner, Michael O'Neill, Erik Hemberg
https://arxiv.org/abs/1703.08535
src ponyge.py ***algorithm*** *mapper.py parameters.py search loop.py step.py* ***fitness*** *evaluation.py classification.py regression.py* ***string*** *match.py ...* ***operators*** *crossover.py initialisation.py mutation.py replacement.py ...* ***representation*** *derivation.py grammar.py individual.py tree.py* ***stats*** *stats.py* ***scripts*** *...* ***utilities*** *...*

… Video – masked prediction and clustering, temporal sequence of video frames … Audio: Wav2Vec 2.0 – timesteps, temporal convolutions + transformer on masked prediction of random timesteps using a contrastive loss – similar to BERT, however, the model doesn't choose over the entire word vocabulary, but among a set of options. Multimodal: joint represenations of multipple data types. … * ImageGPT – GPT2-like model for image generation, early use of transformers for images. See citations below. Multimodal_representation_learning Mutual_information - a measure of the mutual dependence between the two variables; the "amount of information" (in units such as shannons (bits), nats or hartleys) obtained about one random variable by observing the other random variable. Linked to the entropy of a random variable, a fundamental notion in information theory that quantifies the expected "amount of information" held in a random variable. Feature_selection – … Minimum-redundancy-maximum-relevance (mRMR), (Conditional, Joint) mutual information, correlation, regularized trees (decision tree, tree ensemble). (Wrapper/embedded) method. Feature similarity. Cross-entropy – between two probability distributions (p,q): the average number of bits needed to identify an event drawn from the set when the coding scheme used for the set is optimized for an estimated probability distribution q, rather than the true distribution p. Cross-entropy minimization, KL divergence … Ensemble_learning – multiple models; stacking, voting; **Amended Cross-Entropy Cost: Encouraging Diversity***; Bucket of models – choose the best model for each	Derivative-free_optimization https://en.wikipedia.org/wiki/Artificial_development – artificial embryogeny or machine intelligence or computational development; morphogen gradients, cell division and cellular differentiation. Multi_expression_programming – encoding multiple solution in the same chromosome. **MEPX** - a cross-platform (Windows, macOS, and Linux Ubuntu) free software for the automatic generation of computer programs. Libmep https://github.com/mepx https://www.mepx.org/videos.html https://www.mepx.org/ http://hackage.haskell.org/package/hmep Cell_division, Cellular_differentiation, **Stem cell:** Stem_cell – self-renewal, potency – differentiate into specialized cell; types: totipotent or pluripotent; Cell_potency – unipotency, oligopotency multipotency. Totipotency – Divide and produce all of the differentiated cells in an organism: morula. Blastocyst → pluripotent inner mass cells → Circulatory system, Nervous system, immunse system. (Induced) pluripotency. Progenitor_cell – can differentiate into a specific cell type. Genotype%E2%80%93phenotype_distinction Interactive_evolutionary_computation – aesthetic selection, human evaluation Stochatic_optimization,

problem; Cross-Validation Selection; Bayesian model averaging**,** Bayesian model combination; Boosting – emphasizing mis-classified data by previously learned models in the following learning iterations; Bootstrap aggregating (bagging), Bayes optimal classifier. * https://arxiv.org/pdf/2007.08140 **Genetic representation** (array of bits, bit set, bit string; array of integer or real value; binary tree; NL, parse tree, directed graph etc.), **genotype-phenotype mapping**; (phenotype = problem space, genotype = search space). **Direct or complex representation**s – genotype-phenotype mapping. Optimization; captured in local optimum – escape with a genetic drift. **Genetic operator**: mutation, crossover, selection. **Selection**: fitness proportionate (roulette wheel selection) – sometimes a higher fitness candidate solution is also eliminated, tournament (sample, compare a set of two or more) – select the one that's better, even if slightly – or other criteria); truncation: a portion based on fitness (top 10%, 20%, 50% … - but it is better to keep some less fit, because they may carry useful genes for future generations)., stochastic universal sampling. **Rank selection** within the population. **Elitist selection** – carrying to the next generation without change, reproduction. **Boltzmann selection** – starts high, low selection pressure (more diverse set of individuals is selected), then gradually lowered (increases the selection pressure). (Compare with **Simulated annealing**). **Crossover** – recombinations of parts of the genes/code … **Mutation** – diversity, prevent converging to local minimum … Simple: **bit mutation (flipping random bits)** to random values in a **Gaussian distribution** to the current gene value. Combining all operators; random walk through the search space; noise tolerant global search algorithm. fitness function …	

Wikipedia etc. continue: Feature_engineering - Feature extraction ... attributes … Clustering. Feature (templates, combinations). **Libraries, time series**:
https://github.com/predict-idlab/tsflex
https://github.com/dmbee/seglearn https://github.com/fraunhoferportugal/tsfel -

* Barandas, Marília and Folgado, Duarte, et al. „*TSFEL: **Time Series Feature Extraction Library.***“ SoftwareX 11 (2020). https://doi.org/10.1016/j.softx.2020.100456
Statistical, temporal, spectral, fractal; Available features in: Statistical domain: Absolute energy, Average power, ECDF, ECDF Percentile, ECDF Percentile Count, Entroypy, Histogram, Interquartile range, Kurtosis, Max, Mean, Mean absolute deviation, Median, Median absolute deviation, Min, Root mean square, Skewness, Standard deviation, Variance; Temporal domain: Area under the curve, Autocorrelation, Centroid, Lemepl-Ziv-Complexity, Mean absolute diff, Mean diff, Median absolute diff, Median diff, Negative turning points, Peak to peak distance, Signal distance, Slope, Sum absolute diff, Zero crossing rate, Neighbourhood peaks; **Spectral domain:** FFT mean coefficient, Fundamental frequency, Human range energy, LPCC, MFCC, Max power spectrum, Maximum frequency, Median frequency, Power bandwidth, Spectral centroid, Spectral decrease,Spectral(distance, entropy, kurtosis, positive turning points, roll-off , roll-on, skewness, slope, spread, variation), Wavelet(absolute mean, energy, standard deviation, entropy, variance). **Fractal domain:** Detrended fluctuation analysis (DFA), Higuchi fractal dimension, Hurst exponent, Maximum fractal length, Multiscale entropy (MSE), Petrosian fractal dimension. (Fractal domain features: typically applied to longer signals, usually unnecessary to divide the signal into shorter windows.)
https://github.com/facebookresearch/Kats
https://tsfresh.readthedocs.io/en/latest/
https://getml.com/latest/ | https://github.com/getml/getml-community – a specifically customized database Engine for this very purpose… speedup of *60* to *1000* times (see Benchmarks) over featuretools and tsfresh. https://getml.com/latest/examples/
https://github.com/alteryx/featuretools | https://featuretools.alteryx.com/en/stable/
https://github.com/ashishbhaskar/MCMD - Multi-view Classification framework based on Consensus Matrix Decomposition
Regularization_(mathematics) – a process that converts the answer of a problem to a simpler one. It is often used in solving ill-posed problems or to prevent overfitting. Explicit/Implicit. L_1, L_2. … Set_function - measure theory; family of sets (a collection of sets): set, indexed set, multiset, class … a collection F of subsets of a given set S is a family of subsets or a family of sets over S. … Submodular_set_function – or submodular function, a kind of set function: relationship between a set of inputs and an output, where adding more of one input has a decreasing additional benefit (diminishing returns).
Diminishing_returns … Problem_solving (PS) – achieving a goal by overcoming obstacles: Simple PS (one or few direct steps), Complex PS. Subproblem labeling. … well-defined/ill-defined: Problem finding, problem shaping (framing, simplifying): *problem(discovery, formulation, identification, construction, posing).* Domains. PS strategies: abstraction, analogy, brainstorming, bypasses(transform to another problem that is eaier to solve), critical

thinking, divide and conquer [aggregate function; decomposable AF: e.g. Average (i.e., arithmetic mean), Count, Maximum, Median, Minimum, Mode, Range, Sum; Nanmean (mean ignoring NaN values, also known as “nil” or “null”) Stddev…; MapReduce]; help seeking, hypothesis testing, means-ends analysis, morphological analysis, observation/question, proof of impossibility, reduction, research, root cause analysis, trial-and-error. Methods. Barriers: mental set (fixation, sticking to previously succesful solution, rather than search for new and better ones) – see Computer science, ML (~ habit), groupthink. Confirmation bias. Functional fixedness (seeing objects as having only one function, being unable to conceive of any novel use). Unnecessary constraints: 9 dots problem, 4 straight lines (“think outside the box”). Irrelevant information. Avoid barriers by changing problem representation. The *tendency to solve by first, only, or mostly creating or adding elements, rather than by subtracting elements or processes is shown to intensify with higher cognitive loads such as information overload.* North America’s school for PS, topics: calculation[67] computer skills[68], game playing[69] lawyers’ reasoning[70], managerial problem solving[71], mathematical problem solving[72], mechanical problem solving[73], personal problem solving[74] political decision making[75], problem solving in electronics[76], problem solving for innovations and inventions: TRIZ[77] reading[78], social problem solving[11] writing[79]. Complex problem solving **CPS:** complexity (large numbers of items, interrelations, and decisions), enumerability, heterogeneity, connectivity (hierarchy relation, communication relation, allocation relation), dynamics (time considerations), temporal constraints temporal sensitivity, phase effects, dynamic unpredictability, intransparency (lack of clarity of the situation), commencement opacity [when the problem begins/end], continuation opacity, polytely (multiple goals)[80] inexpressivenes, opposition, transience. …Subgoal_labeling

Multimodal_representation_learning – integrating and interpreting information from different modalities such as text, images, audio, video by projecting them into a shared latent space. https://en.wikipedia.org/w/index.php?title=Multimodal_representation_learning&action=history - the Wiki page was created as late as 16.4.2025 (today: 30.4.2025).

Topics: Vision Transformers – ViT

*** Generative Pretraining from Pixels**, Mark Chen 1 Alec Radford 1 Rewon Child 1 Jeff Wu 1 Heewoo Jun 1 Prafulla Dhariwal 1 David Luan 1 Ilya Sutskever 1, **17.6.2020**
https://cdn.openai.com/papers/Generative_Pretraining_from_Pixels_V2.pdf
https://openai.com/index/image-gpt/
Min 32x32, below that the recognition drops for humans (see Tiny Images paper, 2008)
p.4: iGPT-XL,L = 60 layers, embedding size of d = 3072 for a total of 6.8B parameters. iGPT-L, is ~ identical to GPT-2 with L = 48 layers, but a smaller embedding size of d = 1536 (vs 1600) = 1.4B parameters. The same model code as GPT-2, except that the weights are initialized in the layer dependent fashion as in Sparse Transformer (Child et al., 2019) and zero-initialize all projections producing logits. iGPT-M: 455M, L = 36 and d = 1024. iGPT-S, a 76M parameter model, L = 24 d = 512 to study the effect of model capacity on representation quality in a generative model. **Training iGPT-XL**: a batch size of 64, 2M iterations. For all other models: batch=128, 1M it. Adam with β1 = 0.9 and β2 = 0.95 and sequentially try the learning rates 0.01, 0.003, 0.001, 0.0003, ..
*"**Towards general unsupervised learning:** Generative sequence modeling is a universal unsupervised learning algorithm: since all data types can be represented as sequences of bytes, a transformer can be directly applied to any data type without additional engineering."* **Tosh:** Compare with the Bulgarian Predictions 2001-2004 from "Theory of Universe and Mind". by Todor Arnaudov, The Sacred Computer.
"Because we use the generic sequence transformer used for GPT-2 in language, our method requires large amounts of compute: iGPT-L was trained for roughly 2500 V100-days while a similarly performing MoCo24 model can be trained in roughly 70 V100-days. Relatedly, we model low resolution inputs using a transformer, while most self-supervised results use convolutional-based encoders which can easily consume inputs at high resolution. A new architecture, such as a domain-agnostic multiscale transformer, might be needed to scale further. "
* Torralba, A., Fergus, R., and Freeman, W. T. **80 million tiny images: A large data set for nonparametric object and scene recognition**. IEEE transactions on pattern analysis and machine intelligence, 30(11):1958–1970, 2008.
https://huggingface.co/docs/transformers/en/model_doc/imagegpt - "almost exactly the same as GPT-2, with the exception that a different activation function is used (namely **"quick gelu")** etc. ... 32x32, 512 tokens + SOS (start of sequence), 0..511: K-means clustered colors" in a palette. Code: class transformers.ImageGPTConfig(vocab_size = 513, n_positions = 1024, n_embd = 512, n_layer = 24, n_head = 8, n_inner = None, activation_function = 'quick_gelu', resid_pdrop = 0.1embd_pdrop = 0.1, attn_pdrop = 0.1, layer_norm_epsilon = 1e-05, initializer_range = 0.02, scale_attn_weights = True, use_cache = True, tie_word_embeddings = False, scale_attn_by_inverse_layer_idx = False, reorder_and_upcast_attn = False, **kwargs)

*** GELU, Gaussian Error Linear Unit:** https://paperswithcode.com/method/gelu
Compare with the following: Vision transformer ViT etc.
https://en.wikipedia.org/wiki/Vision_transformer – A ViT decomposes an input image into a series of patches (rather than text into tokens), serializes each patch into a vector, and maps it to a smaller dimension with a single matrix multiplication. These vector embeddings are then processed by a transformer encoder as if they were token embeddings. **Masked Autoencoder**: 2 x ViT put end-to-end. | **DINO** (self-**di**stillation with **no** labels) | **Swin Transformer** („**S**hifted **win**dows“) | **TimeSformer | ViT-VQGAN** 8x8 patches … CoAtNet, CvT, data-efficient ViT (DeiT) … Hybrid CNN-ViT.

In 2024, a 113 billion-parameter ViT model was proposed (the largest ViT to date) for weather and climate prediction, and trained on the Frontier supercomputer with a throughput of 1.6 exaFLOPs.[45]
* Wang, Xiao et al. (19.8.2024), **ORBIT: Oak Ridge Base Foundation Model for Earth System Predictability.** arXiv:2404.14712 [physics.ao-ph]
“684 petaFLOPS to 1.6 exaFLOPS sustained throughput, with scaling efficiency maintained at 41% to 85% across 49,152 AMD GPUs.” https://arxiv.org/pdf/2404.14712

* **Frontier: 9,472 AMD Epyc 7713 „Trento“ 64 core 2 GHz CPUs (606,208 cores) and 37,888 Instinct MI250X GPUs (8,335,360 cores). → El Capitan supercomputer** https://en.wikipedia.org/wiki/El_Capitan_(supercomputer) – 43,808 AMD fourth Gen EPYC 24C „Genoa“ 24-core 1.8 GHz CPUs (1,051,392 cores) and 43,808 AMD Instinct MI300A GPUs (9,988,224 cores) … MI300A: 24 x Zen4-based CPU cores, a CDNA3 GPU in a single organic package with 128 GB HBM3 RAM. 5.4375 PB RAM, 1.742 exaFLOPS (Rmax), 2.746 exaFLOPS (Rpeak), Power = 30 MW. Completed: 18.11.2024 … https://github.com/ROCm/hip – HIP: C++ Heterogeneous-Compute Interface for Portability (replaces CUDA; C++ Runtime API and Kernel language for both AMD and NVIDIA Gpus from single source code; “no performance impact over coding directly in CUDA mode”) …

* Dosovitskiy, Alexey; Beyer, Lucas; Kolesnikov, Alexander; Weissenborn, Dirk; Zhai, Xiaohua; Unterthiner, Thomas; Dehghani, Mostafa; Minderer, Matthias; Heigold, Georg; Gelly, Sylvain; Uszkoreit, Jakob (2021-06-03). „**An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale“**. arXiv:2010.11929 (22.10.2020/3.6.2021) https://arxiv.org/abs/2010.11929 ViT … https://colab.research.google.com/github/phlippe/uvadlc_notebooks/blob/master/docs/tutorial_notebooks/tutorial15/Vision_Transformer.ipynb
* Samira Abnar and Willem Zuidema. **Quantifying attention flow in transformers.** In ACL, 2020. – Attention rollout
ViT – https://huggingface.co/docs/transformers/en/model_doc/vit – it requires fewer resources to pretrain than CNNs … basic example: google/vit-base-patch16-224, **base-sized** architecture, patch resolution of 16x16, fine-tuning resolution of 224x224.

*** CLIP: Connecting text and images,** January 5, 2021 Milestone, OpenAI, https://openai.com/index/clip/ … - pretrained on image-text pairs, collected from the Internet; a foundation of image generation; see also OpenCLIP.

* Mahajan, D. et al. **Exploring the limits of weakly supervised pretraining,** 2018. https://arxiv.org/pdf/1805.00932 - predict image hashtags, social media, 3.5 billion images from Instagram. Compute scaling: 336 GPUs across 42 machines with **minibatches of 8,064 images**. Each GPU processes 24 images at a time and batch normalization (BN) [27] statistics are computed on these 24 image sets .. Their ResNeXt-101 32×16d networks took ~22 days to train on 3.5B images

*** Learning Transferable Visual Models From Natural Language Supervision**
Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever 26.2.2021

*** Taming transformers for high-resolution image synthesis.** Patrick Esser, Robin Rombach, and Bjorn Ommer. In CVPR, 2021 – **VQGAN**
https://compvis.github.io/taming-transformers/
https://arxiv.org/abs/2012.09841 17.12.2020/23.6.2021 – In contrast to CNNs, transformers contain no inductive bias that prioritizes local interactions, but are more expensive computationally. … Using CNNs to learn a context-rich vocabulary of image constituents, and then utilize transformers to efficiently model their composition within high-resolution images. … applied to conditional synthesis tasks: both non-spatial information, such as object classes, and spatial information, such as segmentations, can control the generated image. In particular, we present the first results on semantically-guided synthesis of megapixel images … *Allowing transformers to concentrate on their unique strength—modeling long-range relations—enables them to generate high-resolution images.*

*** The Illustrated VQGAN,** Aug 8, 2021, LJ MIRANDA | 22 min read
https://ljvmiranda921.github.io/notebook/2021/08/08/clip-vqgan/ –… vector quantization is a **process of dividing vectors into groups with about equal number of points closest to them** (Ballard, 1999). *Each group is then represented by a centroid (codeword), usually obtained via k-means or any other clustering algorithm. In the end, one learns a dictionary of centroids (codebook) and their corresponding members. … **Training the GAN** from a dataset of images to learn not only its visual parts, but also their codeword representation, i.e., the codebook. **Training the Transformer** on top of the codebook with sliding attention to learn long-range interactions across visual parts. a **patch-based discriminator** (Isola et al. 2017), **perceptual loss** (Johnson and Li, 2016), not in a per-pixel basis.*
* Ballard, D.H., 1999. An introduction to natural computation. MIT press.
* Isola, P., Zhu, J.Y., Zhou, T. and Efros, A.A., 2017. Image-to-image translation with

conditional adversarial networks. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1125-113

*** Excalidraw:** https://excalidraw.com/ – hand-drawn look & feel tool https://www.reddit.com/user/Wiskkey/comments/p2j673/list_part_created_on_august_11_2021/ VQGAN examples, Colab

*** Vector-quantized Image Modeling with Improved VQGAN**
Jiahui Yu, Xin Li, Jing Yu Koh, Han Zhang, Ruoming Pang, James Qin, Alexander Ku, Yuanzhong Xu, Jason Baldridge, Yonghui Wu, 9.10.2021/5.6.2022, https://arxiv.org/pdf/2110.04627 Google Research VQGAN: 256x256, 32x32, codebook size of 8192. *Predicting the rasterzed 32x32 = 1024 image tokens autoregressively, where image tokens are encoded by a learned Stage 1 ViT-VQGAN. For unconditional image synthesis or unsupervised learning, we pretrain a decoder-only Transformer model to predict the next token. For class-conditioned image synthesis, a class-id token is prepended before the image tokens. To evaluate the quality of unsupervised learning, we average the intermediate Transformer features and learn a linear head to predict the logit of the classes (a.k.a., linear-probe). .. a better image quantizer with respect to both computational efficiency and reconstruction quality. An efficient quantizer can speed up Stage 2 training, where random augmentations are applied first to an image, followed by the encoder of image quantizer to obtain the input tokens. Moreover, an image quantizer with better reconstruction quality can reduce information loss compared with the original image in pixel space, which is critical for image understanding tasks***. Improvements:** Vailla VQVAE: low codebook usage, poor initialization, many codes are rarely used or not used at all. VQGAN: topk & top-p nucleus sampling heuristics, default codebook size of 1024. .. **Factorized codes:** *a linear projection from the output of the encoder to a lowdimensional latent variable space for code index lookup (e.g., reduced from a 768-d vector to a 32-d or 8-d vector per code) and find it has an immediate boost of codebook usage.* ***l2-normalized codes. …*** *factorized codes with low-dimensional latent variables consistently achieve better reconstruction quality when the latent dimension is reduced from 256 to 16 or 8*

*** Scaling Autoregressive Models for Content-Rich Text-to-Image Generation**
Jiahui Yu, Yuanzhong Xu et al., 22.6.2022, Google Research, https://github.com/google-research/parti Pathways Autoregressive Text-to-Image model (Parti)*; 1) a transformer-based image tokenizer, ViT-VQGAN, to encode images as sequences of discrete tokens. 2) consistent quality improvements by scaling the encoder-decoder Transformer model up to 20B https://sites.research.google/parti/*

* **Retina Vision Transformer (RetinaViT): Introducing Scaled Patches into Vision Transformers**, Yuyang Shu, Michael E. Bain, University of New South Wales, 20.3.2024 https://arxiv.org/html/2403.13677v1 +3.3% performance cmp the original ViT; low spatial frequency components – capture, select and forward structural and important features to deeper layers. Biologically-isnpired, … a stride of half the patch size (8); image pyramid; *Intuitively, blobs of a similar colour or texture, or areas enclosed by a contour, could constitute basic units in vision. But irrespective of the precise definition, the size and shape*

of such basic units could vary considerably. By including patches of the same image at different scales, RetinaViT has a higher chance of capturing a semantically meaningful basic unit in its input. .. RetinaViT is closer to CNN than ViT, because its input contains the same image at multiple different scales .. stronger than both CNN and ViT in recognizing scale-invariant features, and has a higher chance of detecting the same objects at various scales.

*** Vision Transformers with Hierarchical Attention,** Yun Liu, Yu-Huan Wu, Guolei Sun, Le Zhang, Ajad Chhatkuli, Luc Van Gool, 6.6.2021/26.3..2024
https://arxiv.org/abs/2106.03180v4 https://github.com/yun-liu/HAT-Net

* Topics: Multimodal Learning, Dialog Learning, Continual Learning

Deep Multimodal Representation Learning: A Survey, Publisher: IEEE Wenzhong Guo; Jianwen Wang; Shiping Wang 5.2019 - https://www.researchgate.net/publication/333120040_Deep_Multimodal_Representation_Learning_A_Survey – joint representation, coordinated representation, and encoder-decoder; to narrow the heterogeneity gap; fused learning, cross-modal similarity, cross-modal correlation; cross-modal translation; modality-specific features extracting, multimoda lrepresentation learning which aims to integrate diverse features from different modalities in a common subspace, and a reasoning step such as classification or clustering. .. Some typical models including probabilistic graphical models (PGM), multimodal autoencoders, deep canonical correlation analysis (DCCA), GANs, and attention mechanism, which have either proven to be effective or shown promising results. … **Conclusion**: the primary objective of multimodal representation learning is to narrow the distribution gap in a joint semantic subspace while keeping modality specific semantic intact… Joint representation framework maps all modalities into a global commonsubspace; coordinated representation framework maximizes the *similarity or correlation between modalities*, while keeping each modality independent; encoder-decoder framework maximizes the *conditional distribution* among modalities and keeps their semantics consistent; probabilistic graphical models maximize the *joint probability distribution across modalities*; multimodal autoencoders endeavor to keep modality specific distribution intact by minimizing the reconstruction errors; generative adversarial networks aim to narrow the distribution difference between modalities by an adversarial process; attention mechanism selects salient features from modalities which are similar in local manifolds or they are complementary. (…) https://en.wikipedia.org/wiki/Canonical_correlation – canonical-correlation analysis (CCA), also called canonical variates analysis https://en.wikipedia.org/wiki/Angles_between_flats (principal angles, angles between subspaces) | https://en.wikipedia.org/wiki/Diffusion_map - dimensionality reduction or feature extraction alg. by Coifman and Lafon - nonlinear dimensionality reduction. Connectivity – heat diffusion & random walk Markov chain.

…

*** Dialogue Learning With Human-In-The-Loop,** Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, Jason Weston [Submitted on 29 Nov 2016 (v1), last revised 13 Jan 2017 (this version, v3)] https://arxiv.org/abs/1611.09823 Conversational agents; RL-setting, chat interaction with a human – a dialog partner, instead of a fixed training sets of labeled data; improvements due to the feedback by the teacher following its generated responses; learning through conversations; validated with Mechanical Turk. *Chit-chat type end-to-end dialogue systems* – generate response given the previous history of user utterances; goal-oriented dialogue systems – complete a task (booking a ticket, making a reservation …); QA from dialogues from a DB of knowledge or shot stories. Reward-based imitation learning, forward predictions (predicting the teacher's feedback to the student's response) *[cmp: Michail Bongard, М.Бонгард „Проблема Узнавания"]*. Motives for the bot to ask questions:

clarification (the bot doesn't understand), *knowledge operation* (ask for help for reasoning), *knowledge acquisition* (incomplete knowledge) … [**Compare**: mixed-initative interaction]

*** Learning through Dialogue Interactions by Asking Questions,** Jiwei Li, Alexander H. Miller, Sumit Chopra, Marc'Aurelio Ranzato, Jason Weston 12.2016/2.2017
A good dialog agent should interact with users by both responding to questions and by asking questions …

*** Towards Robust and Efficient Continual Language Learning**, Adam Fisch, Amal Rannen-Triki, Razvan Pascanu, Jörg Bornschein, Angeliki Lazaridou, Elena Gribovskaya, Marc'Aurelio Ranzato July 2023 – Continuing fine-tuning models on new tasks, with the goal of „transferring" relevant knowledge from past tasks; risk negative transfer. A new benchmark of task sequences that target different possible transfer scenarios: *high potential of positive transfer, high potential for negative transfer, no expected effect, or a mixture of each. An ideal learner should be able to maximally exploit information from all tasks that have any potential for positive transfer, while also avoiding the negative effects of any distracting tasks that may confuse it.* Forward transfer – *faster "rate of learning" on a new task***;** the relations between consequent tasks; exploit information for previous tasks with potential for positive transfer (checkpoints) … intermediate fine-tuning – auxiliary task before the target task; others: parameter-efficient forward transfer & resistance to catastrophic forgetting; this work: training efficiency on the new task (not the performance on previous tasks); p.5 collection of tasks: Summarization : AESLC, AG News, CNN-DM, Gigaword, Multi-News, NewsRoom ,SamSum, WikiLingua EN, Xsum…, structure to text: CommonGen, DART, E2ENLG, WEBNLG, Paraphrase: MRPC, QQP, PAWS, STS-B, Sentiment, Coreference, Word disambiguation, Math, Reading comprehension with commonsense, Entailment, Reading Comprehension, Commonsense, Open-domain QA, Text formating, Linguistic acceptability, Qeustion classification (TREC), Conversational QA … Pair of tasks; … learning a checkpoint selector – set of previously solved tasks {t1, …, tn} – *select a previously fine-tuned model on some task ti to initialize from … selective sequential finetuning*

*** Real or Fake? Learning to Discriminate Machine from Human Generated Text** Anton Bakhtin, Sam Gross, Myle Ott, Yuntian Deng, Marc'Aurelio Ranzato, Arthur Szlam 6.2019/11.2019 https://arxiv.org/abs/1906.03351
Energy-based models (EBM) - un-normalized models; the joint compatibility between all input variables, unlike the auto-regressive models which are a sequence of conditional distributions; E(w1, w2, …, wn|c;θ) - joint compatibility of an input sequence of tokens, c – context, θ – params;c can vary: either preceding text, keywords, bag of words, a title etc. losses: binary cross-entropy or ranking; the method to generate negatives is critical (high energy score = unlikely inputs) – auto-regressive models Final*: EBM framework could potentially unlock more expressive models of text, as they are not limited to scoring a single word at a time as current locally normalized auto-regressive models do.*

See also Yann LeCun, JEPA etc. as cited in “The Prophets of the Thinking Machines …” and the originals:

* **I-JEPA: The first AI model based on Yann LeCun’s vision for more human-like AI**, 13.6.2023 https://ai.meta.com/blog/yann-lecun-ai-model-i-jepa/

Tosh, 23.1.2025*: As a definite decision in a fully general case, this is an ill-defined task; the goal could be detecting “fake news” etc. A competent (or incompetent) text could be impossible to distinguish by definition: either for a “universal” case, if it is short enough, or if you don’t have a full dataset or laws for *“really human real texts”*, therefore you cannot be sure that humans definitely don’t write like the style that is *now* assigned or attributed to LLMs, while it was learned from human examples – i.e. the humans may not know that “this is [supposed to be] fake” [because “humans do not write like that/humans are not allowed to write like that” or humans “*never write like that*” – that will make particular phrasing, word choice, syntax, sentence complexity etc. “forbidden”); also you cannot be sure as you may not know or you may be not confident about the writer’s competence or *intent* – incompetent human writers can make mistakes like bad LLMs or old ones, or like if they were prompted certain way, also they might be *imitating* the style they have read from LLMs already, or one may be intentionally mimicking particular style “attributed to LLMs” in order to achieve comic effect, e.g. to joke with “LLM detectors”. In brief, humans or any competent machine writer can write the way they wish, as long as they are “free”.

The recognizers could work when there are clearly defined boundaries and stereotypes, as in particular fixed datasets. This is a 2019 work before ChatGPT. GPT2 and other models from the time still could have easy to spot mistakes and peculiarities in their outputs, especially if the context increases beyond their training, however the LLMs sometimes repeat the originals as they have to be plausible, and answers and *styles* of the instruction LLMs which were stereotypical, say in 2023-2024, for models with particular sizes, are also a result of the respective fine-tuning datasets; the datasets were prepared *by humans*, who already have chosen this literary style and wanted the model to “generalize” and to “transfer” the style to other topics; the style aimed at embedding a particular tone and structure in templated answers, such as greetings, business emails, essays, poems etc.; therefore the style of the content of both “cohorts” – the one prepared by humans, and the machine-generated one – are “real” and “fake” in the same time and intended or assumed to “sound like real” (or the authors of “instruction-fine tuning” datasets and the “tutorial-style” content before that should be also blamed for beiing “machine-like”, “stereotypical”. “using the same phrasing” etc.). After the release of ChatGPT-3.5 and GPT-4 and other powerful models, they started to be used for generation of datasets for fine-tuning and transferring the initial templates to the other smaller and newer models; from there on the users were also being trained by reading and interpreting the chats as ground-truth data and they were shifting their own way of writing and talking etc.

* + Edits and clarification on 28.4.2026, 2.5.2026

* Yann LeCun, Sumit Chopra, Raia Hadsell, Marc’Aurelio Ranzato, and Fu-Jie Huang. **A tutorial on energy-based learning. Predicting Structured Outputs**, 2006. MIT Press.

* Generative Adversarial Networks (Goodfellow et al., 2014)

* Topics: Diffusion Models

Compare with the above mentioned **"The structure of Images", 1984.**

Stable Diffusion – a ground-breaking open image-text mode, published in 2022
https://github.com/CompVis/stable-diffusion

StableDiffusion 3, March 5th 2024, 5.3.2024: https://stability.ai/news/stable-diffusion-3-research-paper – it *outperforms the state-of-the-art models DALLE-3, Midjourney v6 and Ideogram v1 in typography and prompt adherence.*
MMDiT – Multiple modalities, 3 pretrained models for text, two CLIP, one T5
Develops: Diffusion Transformer ("DiT", Peebles & Xie, 2023). Using two separate transformers for each modality, then joining the sequences for the attention; the information is allowed to flow between image and text tokens – improves typography and comprehension; Rectified Flow (RF) formulation (Liu et al., 2022; Albergo & Vanden-Eijnden,2022; Lipman et al., 2023) - straighter inference paths, sampling with fewer steps … diffusion trajectories such as LDM, EDM and ADM,..

* **High-Resolution Image Synthesis with Latent Diffusion Models**, Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, Björn Ommer, 20.12.2021/13.4.2022
https://arxiv.org/abs/2112.10752
https://ommer-lab.com/research/latent-diffusion-models/ See fig.2."*Illustrating perceptual and semantic compression: Most bits of a digital image correspond to imperceptible details."*
1. Perceptual compression, removes high-freq. details, but learns little semantic variation;
2. The gen.model learns the semantic and conceptual composition (semantic compression) …
1. Train an autoencoder – lower dimensional representational space, percept.equiv.to the data space; 2. Reduced complexity, more efficient space; … tasks: unconditional image synthesis, inpainting, stochastic super-resolution; computationally decreased cost comp. to pixel-based diffusion. **Related work:** GAN, likelihood-based methods, Variational autoencoders, flow-based models; Autoregressive models; Diffusion probabilistic models (DM) U-Net – lossy compressor, evaluating in pixel space is expensive – advanced sampling strategies, hierarchical approaches; training on high-resolution data – it requires calculating expensive gradients. *Learn a joint distribution over discretized image and text repres.* … Conditioning mechanism: pre-process the inputs from various modalities, domain specific encoder that projects the input to an intermediate representation, which is mapped to intermediate layers of the Unet with a cross-attention layer Attention(Q,K,V) = softmax ..

* **Scaling Rectified Flow Transformers for High-Resolution Image Synthesis**, Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Muller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, Robin Rombach * Stability AI, 5.3.2024
https://arxiv.org/pdf/2403.03206 **(Stable Diffusion 3)**... Latent diffusion models operate in the latent space of a pretrained autoencoder … mapping between samples of the noise distribution

to the samples from the data distribution … Multimodal Diffusion Backbone – *separated* text and image embeddings & generated synthetic captions in the training set in order to improve prompt- following accuracy and the rendering of text; pretrain on low resolution 256x256 & finetuning on high resolution, Q&K normalization, float32 in order to avoid instability at mixed precision training. 500K steps, batch size 4096 … Human preferences evaluation: Prompt following, Visual aesthetics, Typography (accurate spelling); models up to 8B & 5×10^{22} training FLOPs.

* Topics: Self-Improving General Intelligence, Recursive Self-Improvement

* **Gödel Agent: A Self-Referential Agent Framework for Recursive Self-Improvement,** Xunjian Yin , Xinyi Wang , Liangming Pan , Li Lin, https://arxiv.org/html/2410.04444v4 31.5.2025 – *"a self-evolving framework inspired by the Gödel machine, enabling agents to recursively improve themselves without relying on predefined routines or fixed optimization algorithms. Gödel Agent leverages LLMs to dynamically modify its own logic and behavior, guided solely by high-level objectives through prompting."*

See also the virtual conference **SIGI: Self-Improving General Intelligence** …

(…)

To be continued, connected and refined…

LAZAR

appendix to

THE SACRED COMPUTER

TODOR ARNAUDOV – TOSH

THE PROPHETS OF THE THINKING MACHINES

ARTIFICIAL GENERAL INTELLIGENCE

&

TRANSHUMANISM

HISTORY THEORY AND PIONEERS

PAST PRESENT AND FUTURE

by the author of the world's first university course in Artificial General Intelligence and the Theory of Universe and Mind

http://twenkid.com
https://github.com/twenkid
https://artificial-mind.blogspot.com/

2025