

Fraunhofer Institute for Integrated Circuits IIS





Machine Learning

June 24th, 2024 Prof. Giovanni Del Galdo with contributions by Florian Römer, Steffen Schieler, Sebastian Semper, Christopher Sobel, Nicolas Witt, Jonas Gedschold

Machine Learning General Information

Presenter: Prof. Giovanni Del Galdo Electronic Measurements and Signal Processing (TU Ilmenau) Electronic Measurements and Signal Processing Department (Fraunhofer IIS)

Inputs and Sources

- Presentations and webinars from my co-authors:
 Florian Römer, Steffen Schieler, Sebastian Semper, Christopher Sobel, Nicolas Witt
- [MIT] Introduction to Deep Learning | 6.S19 @MIT <u>http://introtodeeplearning.com</u>
- [HAI] <u>https://aiindex.stanford.edu/wp-content/uploads/2024/05/HAI_AI-Index-Report-2024.pdf</u>
- <u>https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance</u>
- <u>https://epochai.org/trends#investment-trends-section</u>
- chatGPT4o, DALL-E
- Youtube: 3Blue1Brown, Steve Brunton, Sabine Hossenfelder







Machine Learning

General Information and scope



If you have a **question**, feel free to interrupt!



What to expect?

- Get a rough idea about how DL works, and how we got to LLM ٠
- Get an overview on selected trends in research and industry ٠
- Understand the main problems we need to solve in ML, and which risks to mitigate ٠
- Speculate with me in which direction the future will bring us ٠







IIS

AI, ML and DL Terminology





The SPIRIT

of science

AI, ML and DL Intelligence

• What is intelligence?

Intelligence: the ability to process information which will inform future decision-making abilities

In the context of AI: The capability of a machine to perform tasks that would require intelligence if done by humans. These tasks include learning, reasoning, problemsolving, perception, and language understanding.





AI, ML and DL Machine Learning

- What is intelligence?
- What is machine learning?

The science which studies how to make machines make predictions or take decisions based on data without being explicitly programmed to perform specific tasks





AI, ML and DL Machine Learning





The SPIRIT

of science





- What is the advantage of learning features instead of programming them explicitly?
 - Less human effort
 - Scaling
 - High dimensional data becomes treatable
 - Robustness
 - Works better
 - Generalizable



IIS

AI, ML and DL Deep Learning

- What is intelligence?
- What is machine learning?
- What is deep learning?





Machine learning

Classical taxonomy, not so useful anymore



CNN (typically) ٠

> th TECHNISCHE UNIVERSITÄT The SPIRIT ILMENAU



Reinforcement Learning

Learning through interaction





The SPIRIT

of science

Unsupervised Learning

Mainly Clustering

Goal: Find patterns in data

- Simplify data
- Segmentation, compression

Example Segmentation:

 Customer Data Which 3 main types of customers do we have? What else buy customers typically?







Supervised Learning

Function Estimation/Approximation

Given:

Data: Many examples of input x and output y

Unknown:

- Function $x \rightarrow y$
- Output *y* for new input *x*





The SPIRIT

Supervised Learning

Classification & Regression

Classification

Output: Discrete Class / Category / Discrete State



Regression

Output: Continuous Value





Some History



ILMENAU

of science

Introduction Why now?

• The Fundamentals of ML are old. For example: Algorithms to automatically find features have existed for decades

So... why now?

Big data availability and trends in open science

Algorithms and Software

Hardware

- Availability of *compute*
- Commercial interest (covers HW, training and inference costs)
- Funding



The SPIRIT

Why Now?



th



Why now?

Algorithms and Software

The Fundamentals of ML are old. For example: Algorithms to automatically find features have existed for decades

So... why now?

Introduction

- Big data availability and trends in open science
- Hardware

- Availability of compute
- Commercial interest (covers HW, training and inference costs)
- Funding



th

ILMENAU

TECHNISCHE UNIVERSITÄT

The SPIRIT





Datasets

.3 -5 F ρ q

MNIST

WIKIPEDIA The Free Encyclopedia



ImageNet



BLASTNet

- Why make a dataset publicly available?
 - Recognition and impact
 - Grant requirements
 - Incentivize R&D for one specific problem

The SPIRIT

Citations!







Challenges and Benchmarks

[HAI]: Harder benchmarks emerge. AI models have reached performance saturation on established benchmarks such as ImageNet, SQuAD, and SuperGLUE, prompting researchers to develop more challenging ones. In 2023, several challenging new benchmarks emerged, including SWE-bench for coding, HEIM for image generation, MMMU for general reasoning, MoCa for moral reasoning, AgentBench for agent-based behavior, and HaluEval for hallucinations.

ask	Tested Concepts	Supercategory
Abstract Algebra	Groups, rings, fields, vector spaces,	STEM
anatomy	Central nervous system, circulatory system,	STEM
stronomy	Solar system, galaxies, asteroids,	STEM
Susiness Ethics	Corporate responsibility, stakeholders, regulation,	Other
Clinical Knowledge	Spot diagnosis, joints, abdominal examination,	Other
College Biology	Cellular structure, molecular biology, ecology,	STEM
College Chemistry	Analytical, organic, inorganic, physical,	STEM
College Computer Science	Algorithms, systems, graphs, recursion,	STEM
College Mathematics	Differential equations, real analysis, combinatorics,	STEM
College Medicine	Introductory biochemistry, sociology, reasoning,	Other
College Physics	Electromagnetism, thermodynamics, special relativity,	STEM
Computer Security	Cryptography, malware, side channels, fuzzing,	STEM
Conceptual Physics	Newton's laws, rotational motion, gravity, sound,	STEM
conometrics	Volatility, long-run relationships, forecasting,	Social Sciences
lectrical Engineering	Circuits, power systems, electrical drives,	STEM
lementary Mathematics	Word problems, multiplication, remainders, rounding,	STEM
ormal Logic	Propositions, predicate logic, first-order logic,	Humanities
ilobal Facts	Extreme poverty, literacy rates, life expectancy,	Other
ligh School Biology	Natural selection, heredity, cell cycle, Krebs cycle,	STEM
ligh School Chemistry	Chemical reactions, ions, acids and bases,	STEM
ligh School Computer Science	Arrays, conditionals, iteration, inheritance,	STEM
ligh School European History	Renaissance, reformation, industrialization,	Humanities
ligh School Geography	Population migration, rural land-use, urban processes,	Social Sciences
ligh School Gov t and Politics	Branches of government, civil liberties, political ideologies,	Social Sciences
ligh School Macroeconomics	Economic indicators, national income, international trade,	Social Sciences
ligh School Mathematics	Pre-aigeora, aigeora, ingonometry, calculus,	SIEM
ligh School Microeconomics	Supply and demand, imperfect competition, market failure,	Social Sciences
ligh School Physics	Kinematics, energy, torque, nuid pressure,	STEM Social Salamana
ligh School Estatistics	Bendom veriebles, complian distributions, and covers tests	Social Sciences
ligh School US History	Civil Way the Creet Depression. The Creet Society.	Jumonition
ligh School World History	Ottoman ampire accommis imperialism. World Was I	Humanities
light School world History	Senessense, dementie, longauity, personality shanges	Other
luman Sexuality	Preamancy sexual differentiation sexual orientation	Social Sciences
nternational Law	Human rights sovereignty law of the sea use of force	Humanities
urisprudence	Natural law, classical legal positivism, legal realism	Humanities
ogical Fallacies	No true Scotsman, base rate fallacy, composition fallacy	Humanities
Aachine Learning	SVMs VC dimension deen learning architectures	STEM
Aanagement	Organizing, communication, organizational structure	Other
Aarketing	Segmentation, pricing, market research,	Other
Aedical Genetics	Genes and cancer, common chromosome disorders,	Other
liscellaneous	Agriculture, Fermi estimation, pop culture,	Other
foral Disputes	Freedom of speech, addiction, the death penalty,	Humanities
Aoral Scenarios	Detecting physical violence, stealing, externalities,	Humanities
Autrition	Metabolism, water-soluble vitamins, diabetes,	Other
hilosophy	Skepticism, phronesis, skepticism, Singer's Drowning Child,	Humanities
rehistory	Neanderthals, Mesoamerica, extinction, stone tools,	Humanities
rofessional Accounting	Auditing, reporting, regulation, valuation,	Other
rofessional Law	Torts, criminal law, contracts, property, evidence,	Humanities
rofessional Medicine	Diagnosis, pharmacotherapy, disease prevention,	Other
rofessional Psychology	Diagnosis, biology and behavior, lifespan development,	Social Sciences
ublic Relations	Media theory, crisis management, intelligence gathering,	Social Sciences
ecurity Studies	Environmental security, terrorism, weapons of mass destruction,	Social Sciences
ociology	Socialization, cities and community, inequality and wealth,	Social Sciences
JS Foreign Policy	Soft power, Cold War foreign policy, isolationism,	Social Sciences
firology	Epidemiology, coronaviruses, retroviruses, herpesviruses,	Other
Vorld Religions	Judaism, Christianity, Islam, Buddhism, Jainism,	Humanities

Table 2: Summary of all 57 tasks.



https://paperswithcode.com/sota/multi-task-language-understanding-on-mmlu



Notable AI Models

F EPOCH AI

Training dataset size (words)



Projections of the stock of public text and data usage



Effective stock (number of tokens)





th

ILMENAU

TECHNISCHE UNIVERSITÄT

The SPIRIT

🗾 Fraunhofer IIS

Introduction

Why now?

The Fundamentals of ML are old. For example: Algorithms to automatically find features have existed for decades

So... why now?

- Big data availability and trends in **open science**
- Hardware

- Availability of *compute*
- Algorithms and Software Awe-inspiring results
- Commercial interest (covers HW, training and inference costs)
- Funding

Why Now?



Open Science?

- Why open science?
 - Faster innovation
 - Safety and trust
 - Maximizing the use of research data
 - Democratizing access to knowledge

- Why do large companies like open science?
 - Faster innovation
 - Safety and trust
 - Establish *de facto* standards
 - Establish large communities
 - Talent recruitment and acquisition
 - MONEY!









Business Models for AI

Acronym	as a Service		
SaaS	Software	Companies offer ML-powered applications as subscription-based services over the internet. (also API-based services)	ChatGPT, Microsoft365
PaaS	Platform	Providing a cloud-based platform that allows customers to build, deploy, and manage ML applications.	Google Cloud AI, Microsoft Azure ML
laaS	Infrastructure	Offering raw computational infrastructure that can be used to run ML workloads.	Amazon AWS, Microsoft Azure
DaaS	Data	Selling access to curated datasets or data streams that can be used for ML model training and analytics.	Nielsen for consumer behavior data







Introduction

Why now?

• The Fundamentals of ML are old. For example: Algorithms to automatically find features have existed for decades

So... why now?

- Big data availability and trends in open science
- Hardware

Algorithms and Software

- Availability of *compute*
- Commercial interest (covers HW, training and inference costs)
- Funding



th

ILMENAU

TECHNISCHE UNIVERSITÄT

The SPIRIT

Why Now?

Neural Networks date back decades, so why the dominance?

The problem of scaling

 Scaling UP SCALING SCALING SCALING UI









Notable AI Models

https://epochai.org/data/notable-ai-models

📁 EPOCH AI

Training compute (FLOP)





1952

1958

: .

1986

1995

.

Massachuse Institute of Technology

Descent

Why now?

Introduction

So... why now?

Algorithms and Software

- Big data availability and trends in open science
- Hardware

- Availability of compute
- Commercial interest (covers HW, training and inference costs)
- Funding



th

ILMENAU

TECHNISCHE UNIVERSITÄT

MIT Introduction to Deep Learning

(introtodeeplearning.com) @MITDeepLearning

The SPIRIT

Why Now?





- New Models

1/8/24

Software and Algorithms

- Software Frameworks:
 - TensorFlow: Developed by Google, now including the Keras API
 - PyTorch: Developed by Meta
 - Apache MXNet: Supported by Amazon
 - ...
- Sharing / Compute / Discussion Platforms:
 - Kaggle: Owned by Google
 - Hugging Face: Includes Model Hub and Dataset Hub.
 - GitHub: Owned by Microsoft.
 - Google Colab: Cloud-based Jupyter notebooks with free GPU/TPU access
 - ...
- Breakthrough algorithms developments:
 - Transformers, BERT, GAN, Self-supervised learning, etc. -> ~3x training efficiency gain / year







Estimates of algorithmic progress in different software domains





Domain

1952 Descent I. Big Data Perceptron 1958 Learnable Weights Larger Datasets Easier Collection : . & Storage

Stochastic Gradient

Backpropagation

Multi-Layer Perceptron

1986

1995

.

Massachuse Institute of Technology

2. Hardware

Graphics **Processing Units**

(GPUs)

Massively

Parallelizable

- Improved Techniques New Models
 - Toolboxes



1/8/24

3. Software



IM GENET





Neural Networks date back decades, so why the dominance?





Introduction

Why now?

The Fundamentals of ML are old. For example: Algorithms to automatically find features have existed for decades

So... why now?

- Big data availability and trends in open science
- Hardware Algorithms and Software

- Availability of *compute*
- **Commercial interest (covers HW, training and inference cost**
- Funding

TECHNISCHE UNIVERSITÄT

AI Economy

1. Generative AI investment skyrockets. Despite a decline in overall AI private investment last year, funding for generative AI surged, nearly octupling from 2022 to reach \$25.2 billion. Major players in the generative AI space, including OpenAI, Anthropic, Hugging Face, and Inflection, reported substantial fundraising rounds.

2. Already a leader, the United States pulls even further ahead in AI private investment.

In 2023, the United States saw AI investments reach \$67.2 billion, nearly 8.7 times more than China, the next highest investor. While private AI investment in China and the European Union, including the United Kingdom, declined by 44.2% and 14.1%, respectively, since 2022, the United States experienced a notable increase of 22.1% in the same time frame.







Training Cost

Amortized hardware and energy cost to train frontier AI models over SEPOCH AI time



https://epochai.org/data/notable-ai-models



US Military Spending

- "The total that the DoD might spend on AI-related contracts if each contract were extended to its fullest terms grew even faster, from \$269 million in the period leading up to August 2022 to \$4.3 billion in the period leading up to August 2023. ", The U.S. Military's Investments Into Artificial Intelligence Are Skyrocketing", Time Magazine, March 2024
- Strategic relevance of AI?
 - Situational awareness and intelligence
 - Autonomous systems
 - Cybersecurity
 - Logistics



The SPIRIT



IIS

Private Investment



Global corporate investment in AI by investment activity, 2013-23

Figure 4.3.1

[HAI]



The SPIRIT

of science



Industry dominates AI research






The main ideas of DL

The perceptron

source: adapted from MIT

The Perceptron: Forward Propagation



The SPIRIT

To discuss: dimensionality reduction – latent space – modeling and understanding; non-linearity



IIS

Introduction Why non-linearity?





The **SPIRIT** of science

Introduction Why non-linearity?





The SPIRIT

Introduction Why non-linearity?

Complex behavior (i.e., non-linear)



Network with non-linear components

!expressiveness





Importance of Activation Functions

The purpose of activation functions is to introduce non-linearities into the network



Linear activation functions produce linear decisions no matter the network size



Non-linearities allow us to approximate arbitrarily complex functions



The Perceptron: Forward Propagation



Activation Functions

$$\hat{y} = \boldsymbol{g} \left(w_0 + \boldsymbol{X}^T \boldsymbol{W} \right)$$

Example: sigmoid function

$$g(z) = \sigma(z) = \frac{1}{1 + e^{-z}}$$



Common Activation Functions

Sigmoid Function



Hyperbolic Tangent



Rectified Linear Unit (ReLU)



TensorFlow code blocks

lassachusetts

Institute of

Technology

NOTE: All activation functions are non-linear

Neural Networks





Importance of the Architecture: Emergence

• The architecture at large scales might play a much bigger role to determine behavior compared to the individual components





The **SPIRIT**

Importance of the Architecture: Emergence

• The architecture at large scales might play a much bigger role to determine behavior compared to the individual components



th Fraunhofer TECHNISCHE UNIVERSITÄT ILMENAU

The SPIRIT

Example 1: sum of random variables

Emergence of complexity

- A system composed of simple components exhibits behaviors and properties that are not apparent from the individual components alone.
- The idea is closely related to the **separation of scales** in physics
- Example: School of fish / flock of birds





Universal Approximation Theorem (UAT)

The universal approximation theorem states, that a feed-forward network constructed of artificial neurons can approximate arbitrary well a real-valued continuous function.

$$f(x)... \text{ function to approximate.}$$
$$F(x) = \sum_{i=1}^{N} \nu_i \varphi(w_i^T x + b_i)$$
$$\varepsilon > |F(x) - f(x)| \quad \text{for } \varepsilon > 0$$



Empirical Loss

The **empirical loss** measures the total loss over our entire dataset





IIS

th

The SPIRIT of science

Loss Optimization

We want to find the network weights that achieve the lowest loss

$$W^* = \underset{W}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f(x^{(i)}; W), y^{(i)})$$
$$W^* = \underset{W}{\operatorname{argmin}} J(W)$$

MIT Introduction to Deep Learning introtodeeplearning.com Signature @MITDeepLearning



th

ILMENAU

TECHNISCHE UNIVERSITÄT

The **SPIRIT** of science

1/8/24

Massachusetts

Gradient Descent



Problem: what if we get stuck in a local minimum? -> SGD, MBGD





• Learning rate too small

• Learning rate too large

Optimally adaptive

th

The **SPIRIT** of science



Momentum and adaptive learning rates

Example: Adam Optimizer

Adam (Adaptive Moment Estimation) combines momentum and adaptive learning rates:

- 1. First Moment (Mean): $m_t = eta_1 m_{t-1} + (1 eta_1)
 abla f(heta_t)$
- 2. Second Moment (Uncentered Variance): $v_t = \beta_2 v_{t-1} + (1 \beta_2) (
 abla f(heta_t))^2$
- 3. Bias Correction:

•
$$\hat{m}_t = rac{m_t}{1-eta_1^t}$$

• $\hat{v}_t = rac{v_t}{1-eta_2^t}$

4. Parameter Update:

•
$$heta_{t+1}= heta_t-lpharac{\hat{m}_t}{\sqrt{\hat{v}_t}+\epsilon}$$

Key Benefits

- Faster Convergence: Momentum accelerates gradient vectors in the right directions, leading to faster converging.
- Stability: It reduces oscillations and helps in smooth convergence.
- Adaptability: Adaptive learning rates adjust to the optimization landscape, ensuring more efficient learning.

The SPIRIT



Source: 3Blue1Brown: https://www.youtube.com/watch?v=llg3gGewQ5U









TECHNISCHE UNIVERSITÄT

ILMENAU



Fitting





TECHNISCHE UNIVERSITÄT

ILMENAU

The SPIRIT

of science

Bias-Variance Tradeoff





Total Error = Bias² + Variance + Irreducible Error

- Bias: here it is intended as the error from model mismatch introduced by approximating a real-world problem, which may be extremely complex, by a much simpler model.
- Variance: here refers to the error introduced by the model's sensitivity to the specific training data.
 A model with high variance pays too much attention to the training data, including the noise, and performs well on training data but poorly on unseen data (overfitting).
- Irreducible error: Noise inherent in the problem itself.





Regularization

source: adapted from [MIT]

Regularization I: Dropout

• During training, randomly set some activations to 0



The SPIRIT



Regularization

source: adapted from [MIT]





Sequential Data

 Very often data is sequential in nature, e.g., text, audio, ECG, etc.

Sequence Modeling Applications



The SPIRIT



th

Page 60

Recurrent Neural Networks (RNN)







Recurrent Neural Networks (RNN)





th

ILMENAU

TECHNISCHE UNIVERSITÄT

The SPIRIT

Recurrent Neural Networks (RNN)







Problems with RNN

- It is difficult for RNNs to learn from long sequences because the necessary information cannot be effectively propagated through many time steps. (exploding and vanishing gradients)
- Difficult to parallelize given the step-by-step nature of the inference
- Relatively small size hidden layers limit the amount of information (features of the data) which can be stored



Solutions:

- Gated Recurring Units (GRU)
- Long Short-Term Memory
- Transformers



Long Short-Term Memory

LSTMs: Key Concepts

- One solution:
 - Introduce gated cells:
 - Input gate
 - Forget gate
 - Output gate

- I. Maintain a cell state
- 2. Use gates to control the flow of information
 - Forget gate gets rid of irrelevant information
 - Store relevant information from current input
 - Selectively update cell state
 - Output gate returns a filtered version of the cell state
- 3. Backpropagation through time with partially uninterrupted gradient flow





1/8/24

Attention is all you need – Self-attention Head

Alice noticed the beautiful painting.

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

- I. Encode **position** information
- 2. Extract query, key, value for search
- 3. Compute attention weighting
- 4. Extract features with high attention





$$softmax\left(\frac{Q\cdot K^{T}}{scaling}\right)\cdot V = A(Q, K, V)$$

Massachusetts

MIT Introduction to Deep Learning time introtodeeplearning.com

The **SPIRIT**

Vaswani+, NeurlPS 2017. 1/8/24



Attention is all you need – Self-attention Head

Learning Self-Attention with Neural Networks

Goal: identify and attend to most important features in input.

- I. Encode **position** information
- 2. Extract query, key, value for search
- 3. Compute **attention weighting**
- 4. Extract features with high attention

These operations form a self-attention head that can plug into a larger network. Each head attends to a different part of input.



Vaswani+, NeurlPS 2017. 1/8/24





 $\cdot V$

Matmul

Key

Linear

Positional Encoding

scalina

Value

Linear

Softmax

Scale

MatMul

softmax

Query

Linear

Massachusetts

Key Advantages of Self-attention

- For each input (e.g. word) we have a context (e.g., which other word are relevant) -> reveals the relationships between inputs
- Allows for parallelization (also in the training)
- Long dependencies / input sequences



Large Language Models – Status Quo

Self-attention -> transformers -> LLMs





The SPIRIT

Notable AI Models

📁 EPOCH AI

Number of trainable parameters



Al vs Homo Sapiens

Test scores of AI systems on various capabilities relative to human performance



Within each domain, the initial performance of the AI is set to -100. Human performance is used as a baseline, set to zero. When the AI's performance crosses the zero line, it scored more points than humans.



Note: For each capability, the first year always shows a baseline of -100, even if better performance was recorded later that year.





ACHIEVEMENTS UNLOCKED BY LLMS EMERGENT ABILITIES OF LARGE LANGUAGE MODELS (APR/2023)



 ${\mathscr O}$

GPT-313B, PaLM 8B



GPT-3 175B, Μ



PaLM 540B, Chinchilla 70B

Chinchilla 7B



GPT-4, Gemini (est.)



1+1=2



705



LinguisticsPuzzles



GeometricShapes*



College-LevelExams



Grounding



Debugging



EmojiMovie



Proverbs



Self-Critique/Reflection



Long-HorizonPlanning



Comprehension





GRE-Comprehension MetaphorUnderstanding



ElementaryMath



SpatialReasoning



AdvancedEmbodiment





PhysicalIntuition

CausalJudgment

AdvancedCreativity

Awareness



LogicalDeduction



CodeLineDescription



EmbodimentOptions



More..







2

AppBuilding

Persuasion

A-A
2023 CONTEXT WINDOWS (MAX IN/OUT LENGTH)



Using rounded figures of 1 token = 0.75 words (e.g. 32,000 tokens = 24,000 words), 500 words = 1 page. Alan D. Thompson. November 2023. https://lifearchitect.ai/models/

Training Compute

The rise of artificial intelligence over the last 8 decades: As training computation has increased, AI systems have become more powerful

The color indicates the domain of the AI system:
Vision
Games
Drawing
Language
Other

Shown on the vertica that was used to train	l axis is the training co n the Al systems.	omputation	Minerva	Minerva: built in 2022 and trained can solve complex mathematical pro	on 2.7 billion petaFLOP oblems at the college level.
10 billion petaFLOP			PaLM can generate high-a	LM: built in 2022 and trained on 2. quality text, explain some jokes, caus	5 billion petaFLOP e & effect, and more.
Comput One FL0 multipli	tation is measured in floating po OP is equivalent to one additior cation, or division of two decim	oint operations (FLOP). n, subtraction, al numbers.		GPT-3: 2020; 314 m GPT-3 can produce high-qu often indistinguishable from	illion petaFLOP uality text that is n human writing.
100 million petaFLOP			DALL-F can gene	DALL-E: 2021; 47 milli	on petaFLOP
The data is shown on a logarith from each grid-line to the next increase in training computation	hmic scale, so that t it shows a 100-fold on.		Recommendation systems lik	NEO: 2021; 1.1 million ke Facebook's NEO determine what y online shopping, streaming services.	petaFLOP you see on and more.
1 million petaFLOP			AlphaGo defeated 18-time chan complex board game Go. 1	AlphaGo: 2016; 1.9 million petaF npion Lee Sedol at the ancient and h The best Go players are no longer hu	LOP ighly man.
10,000 petaFLOP		AlphaFold was	Alp a major advance toward solving th	haFold: 2020; 100,000 petaFLOP ne protein-folding problem in biology	
		MuZero i chess, a	MuZe is a single system that achieved sup and shogi (Japanese chess) — all wit	ero: 2019; 48,000 petaFLOP perhuman performance at Go, thout ever being told the rules.	
100 petaFLOP			AlexNe	et: 2012: 470 petaFLOP	
		A pivotal early "deep could recognize i	plearning" system, or neural networ mages of objects such as dogs and o	rk with many layers, that cars at near-human level.	
1 petaFLOP = 1 quadrillio	on FLOP			NPLM .	
				Decision tree	• •
10 trillion FLOP		TD-Ga	mmon: 1992; 18 trillion FLOP	LSTM	
		TD-Gammon learne level, just below th	d to play backgammon at a high he top human players of the time.	• LeNet-5	
100 billion FLOP				RNN for speech	
	NetTalk was able to learn to	NetTalk: 198 pronounce some English	87; 81 billion FLOP ALVI Context by being given	INN CNN	
1 billion ELOP	text as input and matching it limitations, it did not pe	to phonetic transcriptio form the visual recognit	ons. Among its many tion of the text itself.		
	Pandemonium (N	Morse)		System 11	
10 million FLOP	Samuel Neural	Checkers	Back-propa Neocognitron: 1980; 228 A precursor of modern visio	agation 3 million FLOP on systems. It could recognize	
			nanawritten Japanese char	 Fuzzy NN 	
100,000 FLOP	 Perceptron Mark I: Regarded as the first of from those marked on 	built in 1957/58; 695,0 artificial neural network, the right, but it could no	000 FLOP it could visually distinguish cards n ot learn to recognize many other typ	narked on the left side pes of patterns.	
	ADALINE: bu An early single	uilt in 1960 and trained e-laver artificial neural n	f on around 9,900 FLOP etwork.		
1,000 FLOP					
• The	seus: built in 1950 and trained	on around 40 floating	point operations (FLOP)		
10 FLOP The that	eseus was a small robotic mouse, a t could navigate a simple maze an	leveloped by Claude Sha ad remember its course.	innon,		
The first electronic computers were developed in the 1940s	Training comput	tation grew in line with l	eep Learning Era Moore's law, doubling roughly eve	ery 20 months.	— Deep Learning Era — Increases in training computati accelerated, doubling roughly every 6 months.
1950	1960	1970	1980 1990	2000	2010 2020
	1956: The Dartmouth v seen as the beginning of	vorkshop on AI, often f the field of AI research		1997: Deep Blue beats chess champion Garry	world Kasparov

The data on training computation is taken from Sevilla et al. (2022) - Parameter, Compute, and Data Trends in Machine Learning. It is estimated by the authors and comes with some uncertainty. The authors expect the estimates to be correct within a factor of two. OurWorldinData.org - Research and data to make progress against the world's largest problems.

Licensed under CC-BY by the authors Charlie Giattino, Edouard Mathieu, and Max Roser

fer

IIS

Plausible



0.6100M/year

The training compute of notable ML models has been growing at 0.61 OOM per year since 2010.

Likely

(i)

90% confidence interval: 0.57 OOM to 0.66 OOM.



Computational performance



Algorithmic improvements

(i)

0.13_{OOM/year}

The amount of FLOP/s for GPUs in FP32 precision grows by 0.13 OOM per year. A similar trend is observed for FP16.

90% confidence interval: 0.12 OOM to 0.15 OOM.



The physical compute required to achieve a given performance in language models is declining at a rate of 0.4 OOM per year.

90% confidence interval: 0.3 OOM to 0.8 OOM.

Training data

2028

The median projected year in which most of the effective stock of publicly available humangenerated text will be used in a training run is 2028.

90% confidence interval: 2026 to 2033.

Plausible

(i)



Tra	in	ina	costs
Пa		my	00313



The cost in USD of training frontier ML models has grown by 0.4 OOM per year since 2016.

90% confidence interval: 0.3 OOM to 0.5 OOM.





(i)

Likely

(i)

Some Achievements





Technology

DeepMind uses AI to control plasma inside tokamak fusion reactor

For the first time, artificial intelligence has been used to control the super-hot plasma inside a fusion reactor, offering a new way to increase stability and efficiency

AlphaDev discovers faster sorting algorithms

June 7, 2023

Advancing mathematics by guiding human intuition with AI



Restoring and attributing ancient texts using deep neural networks

Restoring, placing, and dating ancient texts through collaboration between AI and historians.





AI and Creativity

How an AI finished Beethoven's last symphony and what that means for the future of music



Computer scientists have teamed up with historians, musicologists and composers to teach an artificial intelligence to compose like Beethoven.



Prompt: First 5 notes of Chopin Op. 10, No. 9	IIII SOUNDCLOUD
Prompt: Jazz Piano-Bass-Drums	SOUNDCLOUD
Prompt: Bluegrass Piano-Guitar-Bass-Drums	Mil SOUNDCLOUD
Prompt: First 6 notes of Rachmaninoff	



Movie written by algorithm turns out to be hilarious and intense

From the archives: Ars talks to the filmmakers who collaborated with an AI for Sunspring.





FORBES > INNOVATION > ENTERPRISE TECH

The 15 Biggest Risks Of **Artificial Intelligence**

he **SPIRIT**



1. Lack of Transparency

- 2. Bias and Discrimination
- **3. Privacy Concerns**
- 4. Ethical Dilemmas
- **5. Security Risks**
- 6. Concentration of Power
- 7. Dependence on AI
- 8. Job Displacement

- 9. Economic Inequality
- **10. Legal and Regulatory Challenges**
- 11. Al Arms Race
- 12. Loss of Human Connection
- **13. Misinformation and Manipulation**
- **14. Unintended Consequences**
- **15. Existential Risks**

https://www.forbes.com/sites/bernardmarr/2023/06/02/the-15-biggest-risks-of-artificial-intelligence/



Risks

Youtube: The A.I. Dilemma - March 9, 2023

https://www.youtube.com/watch?v=xoVJKj8lcNQ





XAI / CAI

- Post-Hoc Methods:
 - Relevance Maps (LRP)
 - Counterfactual Methods
 - Locally Interpretable Model-Agnostic Explanations (LIME)
 - Shapley Additive Explanations (SHAP)
 - **Grad-CAM**
 - . . .
- Intrinsic Explainability
 - Known Operator Learning
 - Physics-Informed Neural Networks
 - Unrolled Iterative Methods
 - . . .





(a) Original Image

(b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar*







Accuracy vs Confidence

Currently the focus is on accuracy (e.g., 98% classification accuracy)

In der Anwendung wichtig: Konfidenz \rightarrow wie sicher ist sich das Modell?





TECHNISCHE UNIVERSITÄT

ILMENAU

The **SPIRI1**

Misalignment

Some examples:

. . .

- Supervised Fine-Tuning
- Reinforcecment Learning from Human Feedback (RLHF)
- Rule-based Constraints
- Fairness Constraints (e.g. to correct biases in the data)

Misaligmeng user intent lunguage intrel HIDJEN LA Inpot Orri Layer IDDEN Ersellee botobe LAYER tec ficelsecih a large INPOT lec acutaige model liro HIDDEN L Layer Generated LANGUAGE LAY owith idenathed LAYER A LARGE LARGE LANGUAGE LANGUAGE MODEL user intent GENATED OUTPUT Intent Biscielod Descrite ite mattee the large ang languge model

RGE LANGUAGE MODE

Geneated

Misinformation and Manipulation





Misinformation and Manipulation





Regulation



Number of Al-related bills passed into law in 128 select countries, 2016–23

Figure 7.2.2





Energy costs

Year	OOMs	H1005- equivalent	Cost	Power	Power reference class
2022	~GPT-4 cluster	~10k	~\$500M	~10 MW	~10,000 average homes
~2024	+1 OOM	~100k	\$billions	~100MW	~100,000 homes
~2026	+2 OOMs	~1M	\$10s of bil- lions	~1 GW	The Hoover Dam, or a large nuclear reactor
~2028	+3 OOMs	~10M	\$100s of billions	~10 GW	A small/medium US state
~2030	+4 OOMs	~100M	\$1T+	~100GW	>20% of US electricity production

https://situational-awareness.ai/wp-content/uploads/2024/06/situationalawareness.pdf

The **SPIRIT**

of science



Nvidia Latest Monster

- Hopper -> Blackwell
- Chiplets
- 10 TB/s link between dies
- 15 kW power draw
- I DGX (acts as one GPU -> 100 kW) NVLink Spine -> 5000 wires connects 72 B100 GPUs to each other
- 13.5 TB HBM3e shared memory





High-Bandwidth Memory

- How to get more memory connected with more bandwidth?
- HBM3E:

1-the die is packaged adjacent to the GPU die 2-3D stacking of memory modules

Problem: getting LOTS of memory connected REALLY FAST, as the number of connections is the product of the two



HBM3E provides 1.5x Capacity, 1.4x Bandwidth, and a 0.9x improvement in Power Efficency compared to HBM3.



Neuromorphic Computing

- There are major differences in how a bilogical NN works compared to an artificial one
 - Neurons encode information as binary spikes (time encoding) and are slow
 - Neuroplasticity
 - Huge amoung of backwards signaling towards the inputs – Predictive Coding Theory

...and similarities

- NN, hierarchical structures
- Extensive mechanisms of inhibitory inputs
- . . .
- The result:
 - OOM better power efficiency and way better latency





https://www.youtube.com/watch?v=6Dcs6fQgIRA





", The Matrix" got it almost right...







Grid-free Harmonic Retrieval and Model Order Selection using Deep Convolutional Neural Networks

21/03/2024

steffen.schieler@tu-llmenau.de

https://arxiv.org/abs/2211.04846



Motivation

Integrated Sensing and Communication



- task: estimate the position of sensing targets from measurements of the wireless channel

- we know the positions of the Tx and Rxs
- the positions can be estimated from the measurements by first estimating the propagation parameters of the sensing targets paths
- existing solutions: DFT w/ Peak-Search, MUSIC, ESPRIT, RIMAXS



Propagation and Signal Model

and link to target position and velocity

 Decide on the problem What are inputs and outputs?



- the signal model accounts for the finite aperture from the measurement hardware

Approach Overview



Feature Engineering Normalization

- Multi-windowing with eight different windows (e.g., Hann, Tukey, ...)
- Mapping to real-numbers via $|\cdot|^2$ and \angle

Deep CNN estimates modelorder and encoded

3. Architecture design – performance and compute efficiency





and scaling

= Encoding

Labels

Grid-free parameter encoding

- use a grid-relative parameter encoding similar to YOLO bounding-box encoding (see Fig. 2)
- mapping functions to encode and decode continuous parameters

 $f(\boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{lpha}) o \boldsymbol{\eta}$ $f^{-1}(\boldsymbol{\eta}) o (\boldsymbol{\gamma}, \boldsymbol{\tau}, \boldsymbol{lpha})$

- encoding allows continuous parameter estimates
- softmax probability allows masking empty predictions during training





Loss function

- 4. Design a loss function
 - -alignment with goals and interpretability
 - -differentiability (for GD)
 - -robust against outliers
 - -computational efficiency
 - -robust against imbalance data (classification)

 \rightarrow efficient and effective training

• Binary Crossentropy (BCE) aka Logic Loss or Logistic Loss

$$\mathfrak{L}_{0} = \hat{\rho} \cdot \log(\rho) + (1 - \hat{\rho}) \cdot \log(1 - \rho)$$

$$\mathfrak{L}_{1} = \sum_{i,j=1}^{I,J} \sum_{c=1}^{C} \left(\sigma \left(\hat{\mu}_{c}^{[i,j]} \right) \cdot \left\| \begin{bmatrix} \Delta \hat{\tau}_{c}^{[i,j]} \\ \Delta \hat{\alpha}_{c}^{[i,j]} \end{bmatrix} - \begin{bmatrix} \Delta \tau_{c}^{[i,j]} \\ \Delta \alpha_{c}^{[i,j]} \end{bmatrix} \right\|_{1} \right)^{2},$$

$$\mathfrak{L}_{\text{total}} = \mathfrak{L}_{o} + \beta \cdot \mathfrak{L}_{1}.$$
5. Hyperparameter tuning and choose optimizer



CNN Architecture

What is being trained





Dataset and Training Hyperparameters

Name	Value	
Datasets		
Distribution $ au_p, lpha_p$	$\mathfrak{U}_{[0,1]}$	5. Hyperparameter tuning and
Min. separation $ au_p, lpha_p$	0.003125	choose optimizer
Magnitudes	$\mathfrak{U}_{[0.001,1]}$	
Phases	$\mathfrak{U}_{[0,2\pi]}$	
SNR	050 dB	
Number of Paths	$\mathfrak{U}_{[1,20]}$	
Trainingset Size	400000	
Validationset Size	1000	
Testset Size	1000	
Training	K	
Optimizer	Adam, $\gamma = 0.0003, eta_1 = 0.9, eta_2 = 0.999$	
Mini-Batchsize	32	
Epochs	20	
Trainable Parameters	~ 25.000.000 for $N_f=N_t=64$	



Inference Example

From the test dataset



Fig. 3. Inference example. A snapshot Y with P = 18 paths from the Validationset passes the network at different SNRs (a-d). The figures show the groundtruth (\odot) and parameter estimates (O) with the data Y in the background. The displacement between the circles indicates the accuracy of the point estimates (center). We observe the quality of the results improves with increasing SNR. At 10 dB SNR (d), all paths are correctly detected, including the closely-spaced paths in the bottom left (see c) and d)).



Takeaway 1

The NN outperforms the peak-search but cannot achieve the accuracy of the maximum likelihood method. But the NN results can be used to initialize additional gradient-steps (here 5) on the likelihood function to enable model-based estimates with high accuracy.

ML working not that well. But why?



Takeaway 2 NN-based model-order estimation is superior to other SOTA techniques.



Current Takeaways (cont.)

Algorithm runtime



Takeaway 3

Combining NNs and maximum likelihood estimation gives model-based results with update rates of >10 Hz.



Additional Resources

Demo and Blog



https://huggingface.co/spaces/EMS-TU-Ilmenau/deepest-demo https://www.steffenschieler.xyz/blog /2024/intro/



Conclusions Where does the future bring us?

Prediction is very difficult, especially about the future



Conclusions

Where does the future bring us?

Two possible futures:

- The technology improves. Major impact to productivity and employment.
 - current problems will be solved:
 - facts vs. hallucinations
 - complex reasoning or even GAI
 - transparency and explainability
 - • •
 - The technology hits its constraints, and we enter the valley of disillusionment.
 - (e.g. Maybe large-scale non-linear regression does not lead to AGI)





Agents, online learning, neuroplasticity etc.

Current LLMs are implemented as one-shot models.

- ...also to avoid these problems:
 - Stability and safety (catastrophic forgetting, misuse and alignment issues, privacy issues, repeatability)
 - Training+inference is much more expensive than inference only
 - Unpredictable outcomes



https://www.youtube.com/watch?v=sal78ACtGTc





Concluding Remarks

-Great time to be alive due to the fast-paced progress!

-If you haven't already, you should probably start learning about ML and get coding

-Don't worry too much about AI – there are plenty of other ways for humanity to go extinct





States of the second

Frau

--consciousness and AI -> Penrose "The Emperor's new mind" -> strong AI vs weak AI, stron AI=capable of understanding and consciousness Weak AI= machines simulate human cognitive processes

Main argument: human thought is non-algorithmic -> Gödel's incompleteness theorem. Certain truths remain not provable in a formal system (such arithmetic). This implies that no algorithmic process (which a formal system represents) can capture all mathematical truths.

Role of quantum mechanics


Effort Justification

Effort justification is a cognitive bias that occurs when people attribute greater value to an outcome if they have put a lot of effort into achieving it. This bias is part of the broader theory of cognitive dissonance, which suggests that people experience discomfort when their actions are inconsistent with their beliefs or values. To reduce this discomfort, they may change their perceptions or attitudes.us value much more those things in which we have put great effort, regardless the fact that the outcome may be more or less valuable.



Effort Justification

how do psychologists call the cognitive bias, for which humans give more importance/value to things that have cost them lot of time or energy to do?





Sunk Cost Fallacy

Definition:

The sunk cost fallacy is a cognitive bias that occurs when individuals continue to invest time, money, or effort into a project or decision based on the cumulative prior investment (sunk costs) rather than on the current and future costs and benefits.

Key Points:

• **Perceived Value**: People tend to give more importance or value to things that have cost them a lot of time, energy, or resources, leading them to continue investing in these things even when it

[...]



Effort Justification

ok, this is related, but I think there is one more distinct bias covering what I was saying



he **SPIRI1**



Effort Justification

Definition:

Effort justification is a cognitive bias that occurs when people attribute greater value to an outcome if they have put a lot of effort into achieving it. This bias is part of the broader theory of cognitive dissonance, which suggests that people experience discomfort when their actions are inconsistent with their beliefs or values. To reduce this discomfort, they may change their perceptions or attitudes.

Key Points: [...]



Do you think that using an AI like yourself to prepare a presentation or effort could be seen by some as of less value, as some parts were automated and cost much less human effort?





Yes, it is possible that some people might perceive the use of AI to prepare a presentation or perform other tasks as less valuable because it involves less human effort. This perception can be influenced by several factors, including the effort justification bias and cultural attitudes towards automation and artificial intelligence.

[...]



~ I messed up the prompt;p

Here,

are you making this point to win my friendship or do you really think so?





So, to answer your question: I am providing this perspective because it is grounded in relevant knowledge and insights, not because of any personal motivations or desire to win friendship.

ok, fair enough. Can I still call you "friend"?





Of course, you can call me "friend" if that makes our interactions more enjoyable for you! I'm here to assist and support you in any way I can, and if thinking of me as a friend enhances your experience, I'm happy to be your "AI friend."





do you think an AI friend is enough in terms of friends?





While an AI friend like me can provide valuable assistance, information, and companionship in certain ways, human friendships offer unique and irreplaceable benefits that an AI cannot fully replicate. Here are some important distinctions:



. . . .

Benefits of Human Friends

- 1. Emotional Support: Human friends can provide empathy, understanding, and emotional support during both good and bad times.
- 2. Shared Experiences: Friends can share experiences, participate in activities together, and create memories.
- 3. Complex Communication: Human interaction involves non-verbal cues, body language, and emotional nuances that AI currently cannot fully understand or replicate.
- 4. Mutual Growth: Friendships involve mutual growth, learning from each other, and providing constructive feedback.

Benefits of an AI Friend

- 1. 24/7 Availability: An Al friend is always available, regardless of time or location.
- 2. Non-Judgmental Support: AI can provide non-judgmental responses and assistance.
- 3. Information and Assistance: AI can quickly provide information, answer questions, and help with various tasks.
- 4. Personalization: Over time, AI can learn about your preferences and provide more personalized support.









