

Workshop Data Science - Frühlingssemester 2025

Vertiefung Data Science

Studiengang Informatik

Simon Felix, Michael Graber, Martin Melchior

Zielsetzung

- Selbständiges Erarbeiten eines aktuellen Themas in der wissenschaftlichen Forschung im Gebiet Data Science
- Austausch mit Mitstudierenden und Fachexperten

Zielsetzung

- Selbständiges Erarbeiten eines aktuellen Themas in der wissenschaftlichen Forschung im Gebiet Data Science
- Austausch mit Mitstudierenden und Fachexperten
- Wissenschaft als diskursiven, gemeinschaftlichen Prozess erleben
- **Freude am wachsenden Verständnis und am Austausch mit Peers**

Arbeitsmittel / Vorgehen

- Einarbeitung in Literatur
- Auseinandersetzung mit Konzepten und Methoden mit Betreuer und Mitstudierenden
- Formulieren von Hypothesen zu Wirkungszusammenhängen im Kontext der betrachteten Methoden
- Konzeption und Umsetzung von Experimenten zur Untersuchung der Hypothesen
- Erarbeitung von konzisen Grafiken zur Vermittlung von Methoden (Schemata) und Resultaten (Plots)

Erwartungen / Wünsche

- Selbständige, (selbst-)kritische und transparente Auseinandersetzung mit dem Themengebiet
- Verständlicher und präziser Ausdruck in der gemeinsamen Auseinandersetzung
- Aktives Mitdenken und Teilnehmen an Diskussionen und Präsentationen

Leistungsbeurteilung

Kriterien

- Lernerfolg / Durchdringung des Themas
- Methodik und Vorgehen
- Arbeitseinsatz und -haltung
- Präsentationen

2/3 Gewicht beim Hauptbetreuer

Feedbackgespräch mit dem Hauptbetreuer

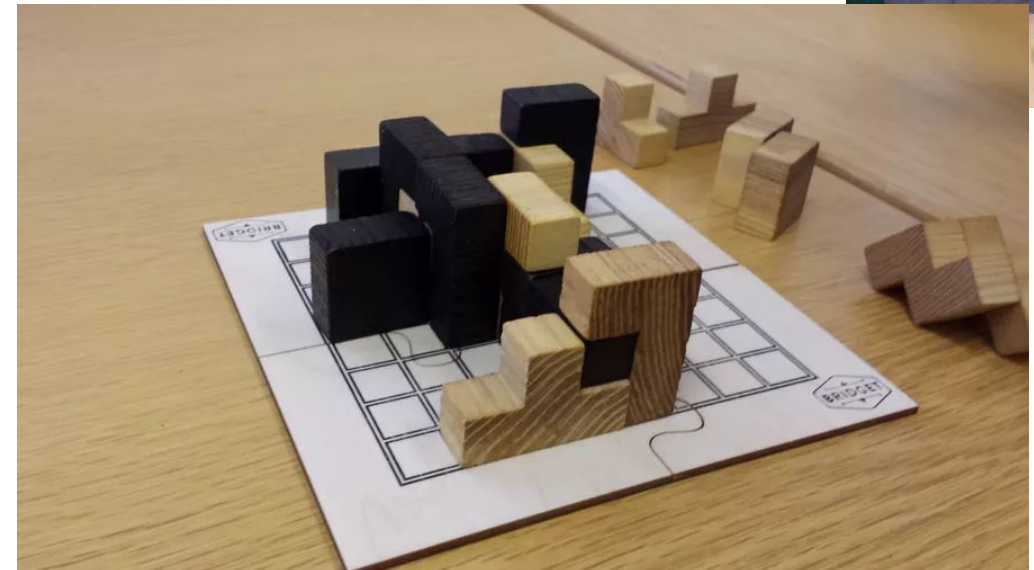
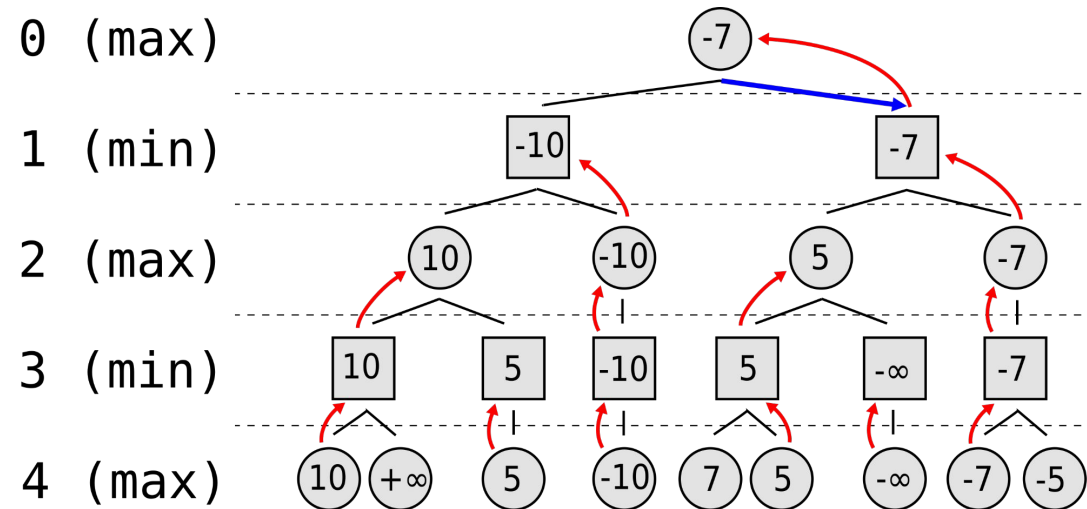
Termine

- Kick-Off: Montag, 17. Februar 2025, 09:15 - 11:00 Uhr
- Journal Club-Session: Montag, 17. März 2025, 08:15 - 11:00 Uhr
- 3-Slides-Session: Montag, 14. April 2025, 08:15 - 11:00 Uhr
- Abschlusspräsentationen: Montag, 2. Juni 2025, 08:15 - 11:00 Uhr

Themen

Thema 1: The History Heuristic and Alpha-Beta Search Enhancements in Practice & A New Paradigm for Minimax Search

(Simon Felix)



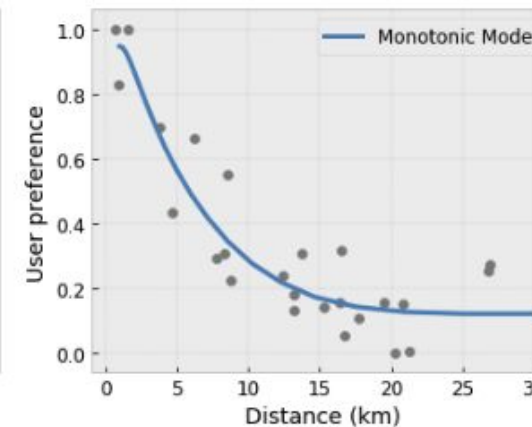
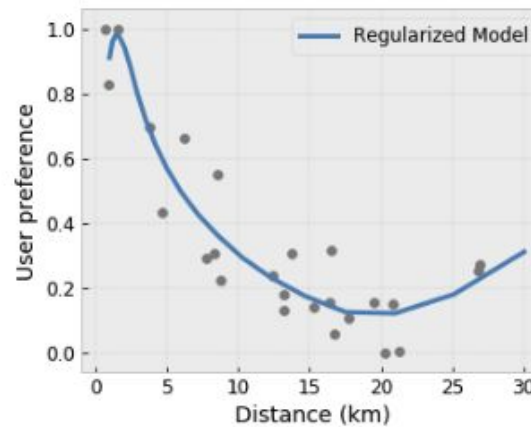
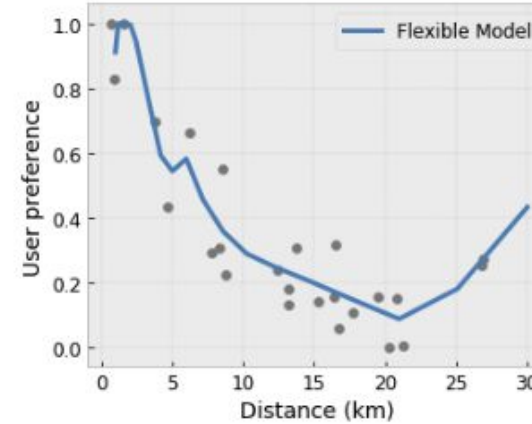
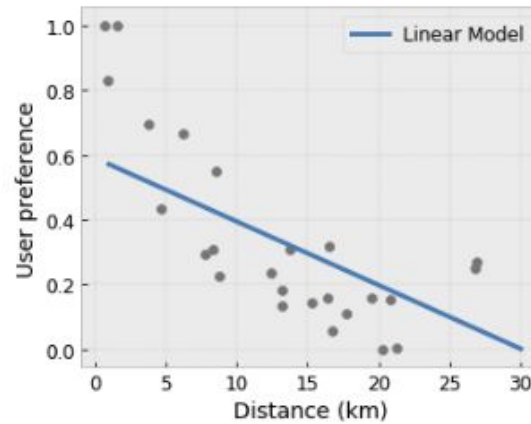
Thema 2: Vertex Block Descent

(Simon Felix)



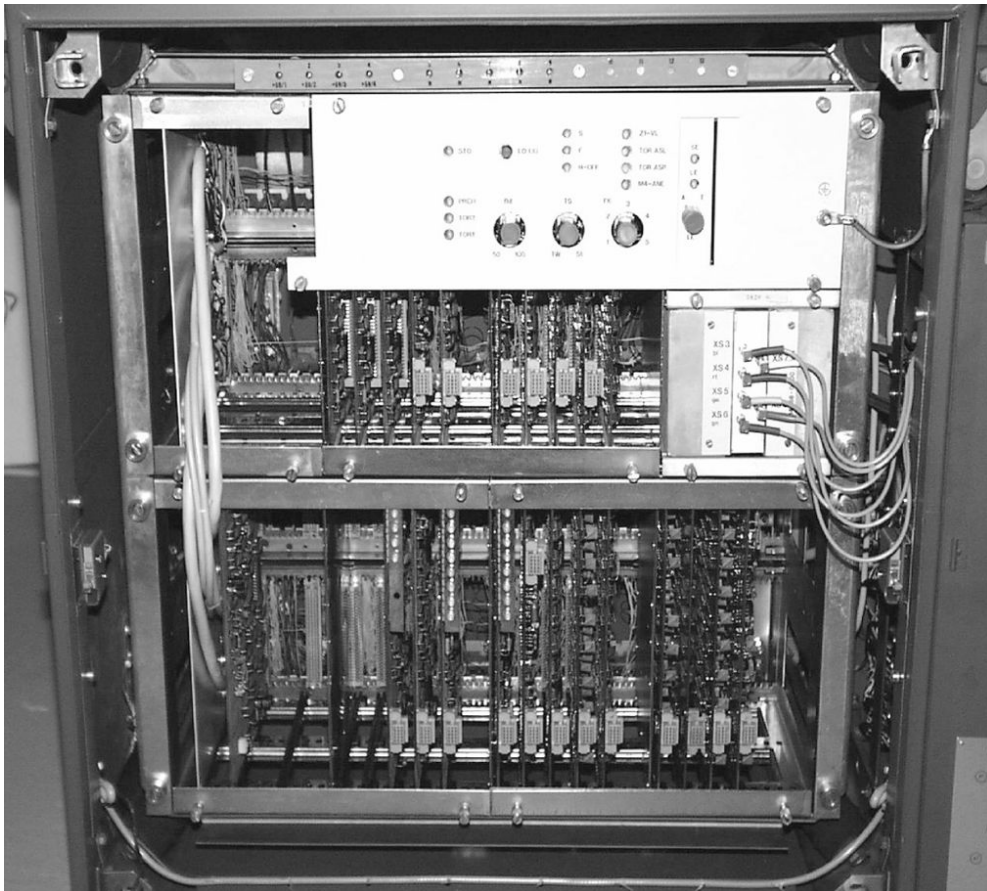
Thema 3: Constrained Optimization

(Simon Felix)



Thema 4: The East German encryption machine T-310 and the algorithm it used

(Simon Felix)



- R_i is another 5 bit word. $R_i = (r_{i+12}, r_{i+11}, r_{i+10}, r_{i+9}, r_{i+8})$.
- r_i is defined as follows:

$$r_i = \begin{cases} 0, & \text{if } (r_{i+12}, r_{i+11}, r_{i+10}, r_{i+9}, r_{i+8}) \in \{(0,0,0,0,0), (1,1,1,1,1)\} \\ 31 - r_{\text{otherwise}}, & \text{with} \\ & (r_{i+12}, r_{i+11}, r_{i+10}, r_{i+9}, r_{i+8}) \cdot \Delta F = (1,1,1,1,1) \end{cases}$$
- a_i can be regarded as the status variable of the T-310 cipher. $a_i = a_{120,i}$ with $i \in \{1, 2, 3, \dots\}$.
- $s \in \{1, 2, \dots, 36\}$; s is a part of the long-term key.
- $s_{36,s}$ is a 36-dimensional bit sequence with $m \in \{0, 1, 2, \dots\}$ and $s \in \{1, 2, \dots, 36\}$. It is defined as follows:

$$(s_{36,1}, \dots, s_{36,36}) = 0110100111000111100100000010100011$$

$$(s_{36,1}, s_{36,2}, \dots, s_{36,36}) = \phi(s_{36,1}, s_{36,2}, \dots, s_{36,36}, s_{36,1}, s_{36,2}, \dots, s_{36,36})$$
- $s_{36,s}$ is a two-dimensional bit sequence with $m \in \{1, 2, 3, \dots\}$ and $s \in \{1, 2\}$. It is defined as follows: $s_{1,12,s}$ and $s_{1,12,s}$ are the secret key of the T-310 cipher. This means that the secret key consists of 240 bits, 10 of the bits are not effective, as the following equation must hold for $i = 1, \dots, 5$ and $j = 1, 2$:

$$s_{36,i-11,1,j} + s_{36,i-11,2,j} + \dots + s_{36,i-11,36,j} = 1$$
- For $m > 120$ the following definition is valid: $s_{36,s} = s_{36-120,s}$.
- f_i with $i = \{-60, -59, -58, \dots, -1, 0, 1, 2, 3, \dots\}$ is a binary sequence defined as follows: f_i with $i = \{-60, -59, -58, \dots, -1, 0\}$ is the initialization vector of the streamcipher. It is chosen at random, but may not consist of all zeros.
For $i > 0$ the following definition is valid: $f_i = f_{i-48} + f_{i-56} + f_{i-64} + f_{i-72}$.
- ϕ is a function matching a 39 bit word to a 36 bit word. For $j = 1, 2, 3, \dots, 9$ it is defined as follows:

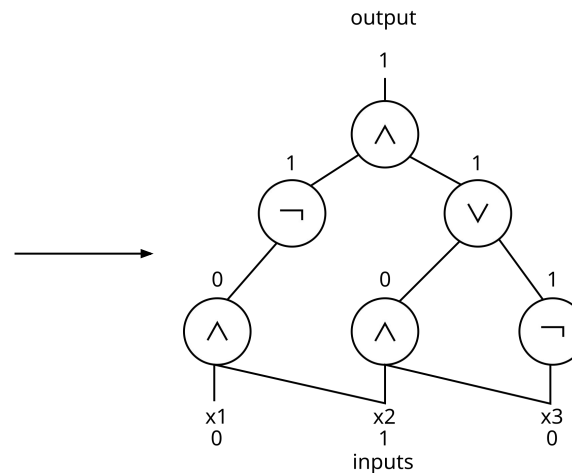
$$\phi_{9,j-1}(f_1, f_2, f_3, \dots, f_{39}) = f_{36,j} + T_{36,j}(f_1, f_2, f_3, \dots, f_{39})$$

$$\phi_{9,j-2}(f_1, f_2, f_3, \dots, f_{39}) = f_{36,j-1}$$

$$\phi_{9,j-3}(f_1, f_2, f_3, \dots, f_{39}) = f_{36,j-2}$$

$$\phi_{9,j-4}(f_1, f_2, f_3, \dots, f_{39}) = f_{36,j-3}$$

$$f_{36,j} = f_j$$
- D is a function mapping $\{1, 2, \dots, 9\}$ to $\{0, 1, \dots, 36\}$. The definition of D is a part of the long-term key.
- P is a function mapping $\{1, 2, \dots, 27\}$ to $\{1, 2, \dots, 36\}$. The definition of P is a part of the long-term key.



The Kissat
SAT Solver



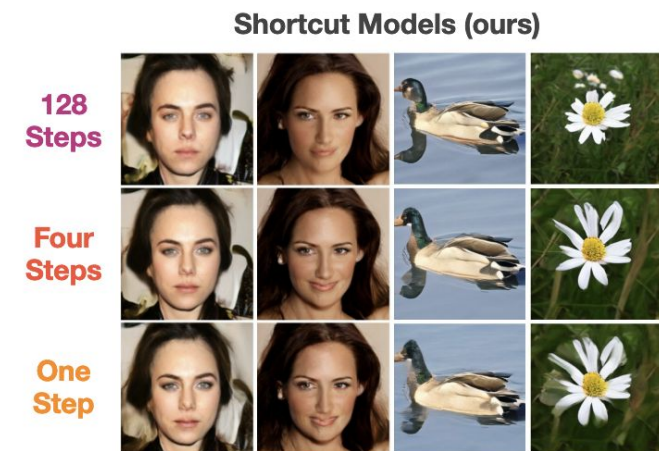
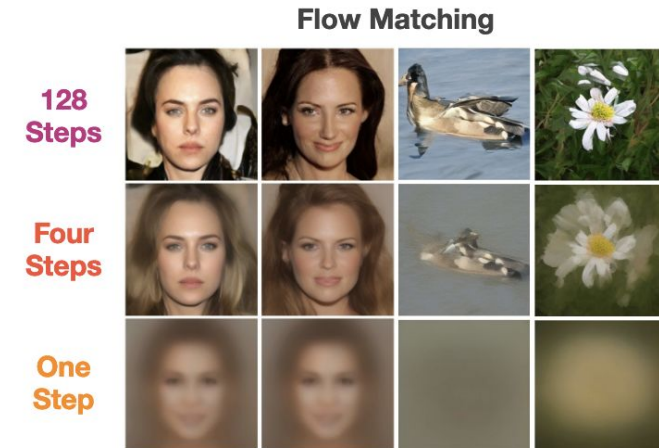
Thema 5: Ein-Schritt Diffusion für Bildgenerierung

(Martin Melchior)

Diffusionsmodelle (wie in Stable Diffusion, Dall-E, ImageGen) sind rechnerisch teuer beim Generieren der Bilder.

Mit dem im Paper [One Step Diffusion via Shortcut Models](#) (ICLR 2025) vorgestellten Ansatz, wird dieses Problem behoben.

Die Idee ist, die Modelle auf die Schrittweite konditioniert zu trainieren.



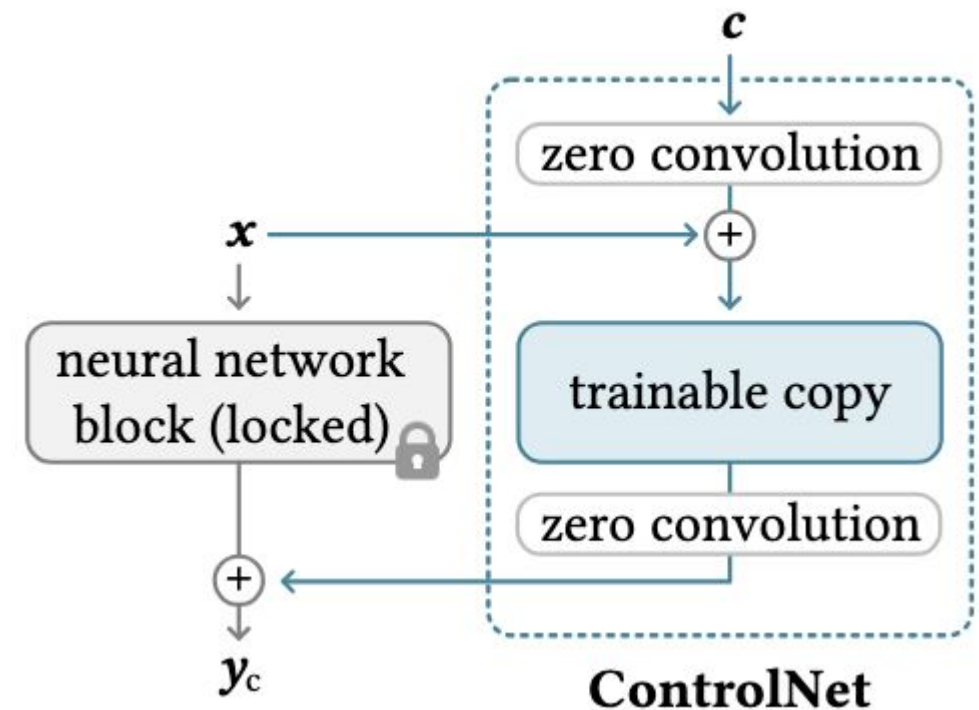
Thema 6: Zusätzliches Prompting für Text-to-Image Diffusionsmodelle

(Martin Melchior)

Paper [Adding Conditional Control to Text-to-Image Diffusion Models](#) (ICCV 2023)

Methode zum “Finetunen” eines vortrainierten Diffusionsmodells (wie z.B. Stable Diffusion), so dass zusätzliche Prompts fürs Conditioning übergeben werden können:

Das wird mit einer geeigneten Modell-Architektur erreicht (“ControlNet”).



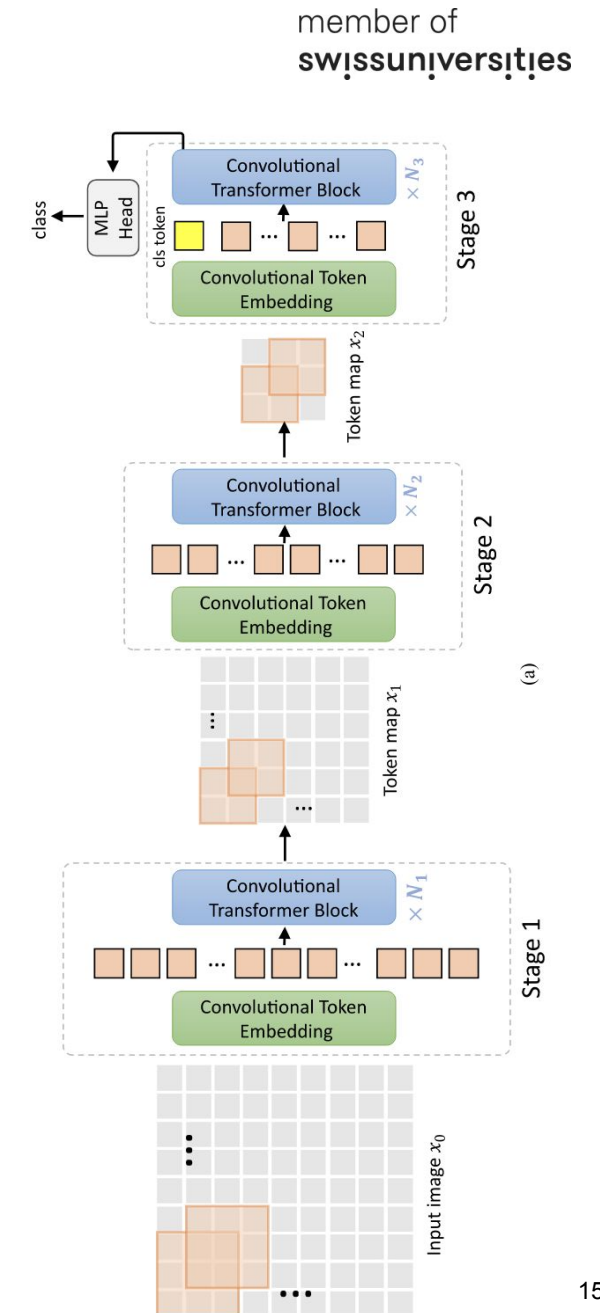
Thema 7: Convolutions und Transformer kombiniert

(Martin Melchior)

Paper “[Attention is All You Need](#)” (NIPS 2017, 150k+ Citations):
Transformer Architektur basierend auf Attention Mechanismus und
MLPs. Im Computer Vision Bereich: [ViT](#) (ICLR 2020).

Convolutions dennoch interessant: Lokale Features werden
effizient gelernt (Parameter-Sharing, Translationseigenschaften).

Kombination im Paper [CvT: Introducing Convolutions to Vision Transformers](#) (ICCV 2021).



Thema 8: Knowledge Distillation, Self-Distillation

(Martin Melchior)

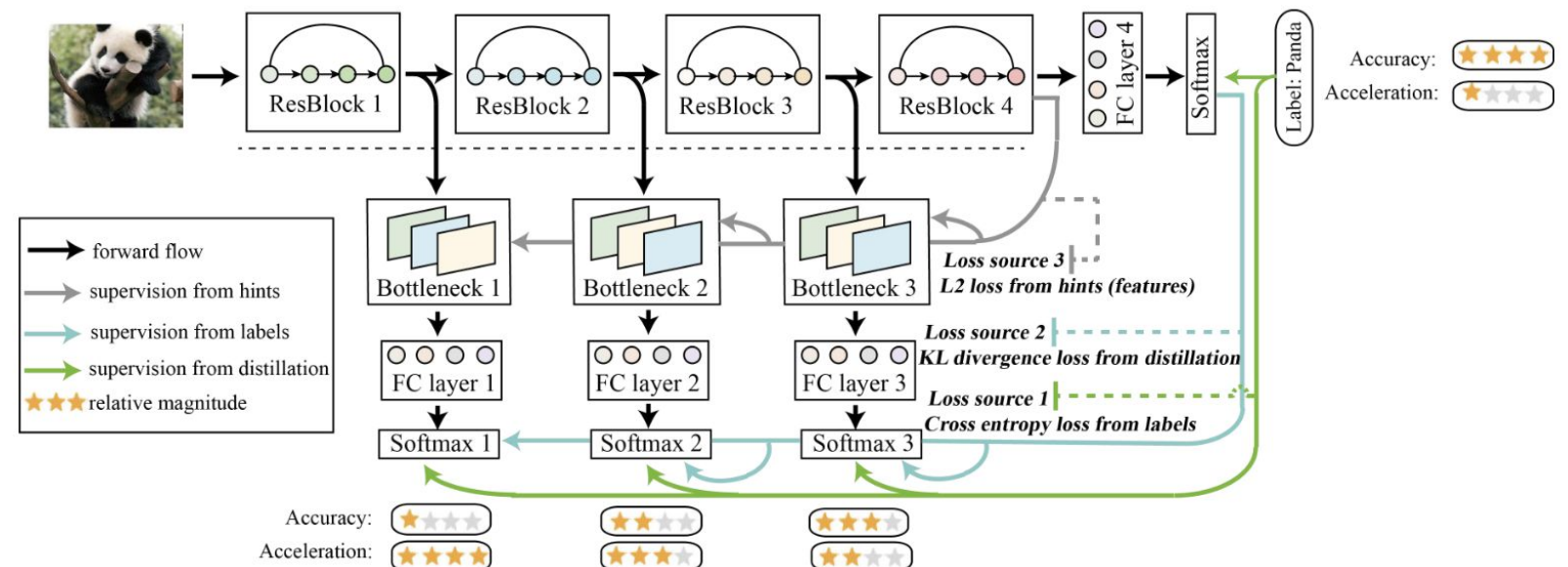
Knowledge Distillation bedeutet, dass antrainiertes Wissen von einem grossen Modell (Teacher) auf ein kleineres Modell (Student) übertragen wird. Das Student-Modell ist dann kompakter manchmal sogar mit besserer Performance als der Teacher.

Braucht es den Teacher überhaupt noch? Kann das Prinzip direkt als Teil des Trainings eingebaut sein?

Paper:

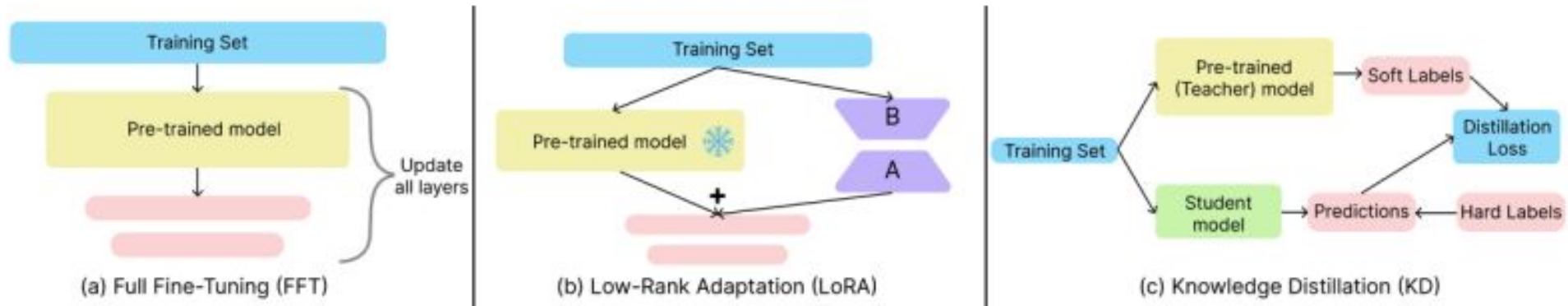
[Be Your Own Teacher:
Improve the Performance of
Convolutional Neural
Networks via Self Distillation](#)

(ICCV 2019)



Thema 9: KD-LoRA / PC-LoRA

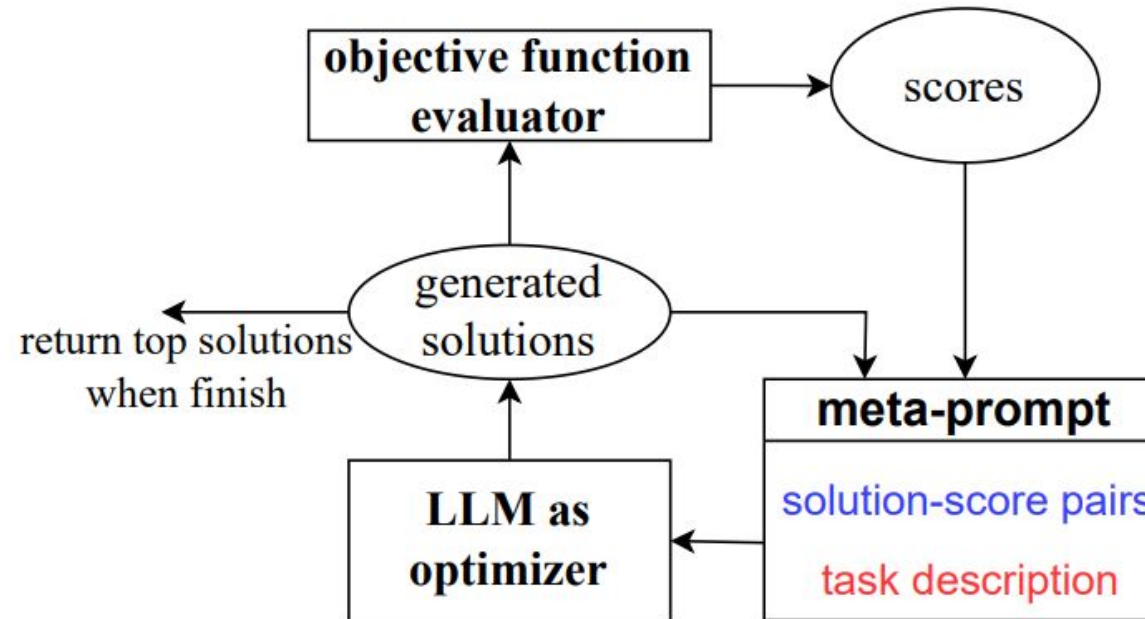
(Michael Graber)



Azimi et al., KD-LoRA, 2024.
Hwang et al., PC-LoRA, 2024.

Thema 10: Optimization by Prompting

(Michael Graber)



Yang et al., Large Language Models as Optimizers, 2024

Thema 11: TabPFN

(Michael Graber)

Article

Accurate predictions on small data with a tabular foundation model

<https://doi.org/10.1038/s41586-024-08328-6>

Received: 17 May 2024

Accepted: 31 October 2024

Published online: 8 January 2025

Open access

 Check for updates

Noah Hollmann^{1,2,3,7}✉, Samuel Müller^{1,7}✉, Lennart Purucker¹, Arjun Krishnakumar¹, Max Körfer¹, Shi Bin Hoo¹, Robin Tibor Schirrmeyer^{4,5} & Frank Hutter^{1,3,6}✉

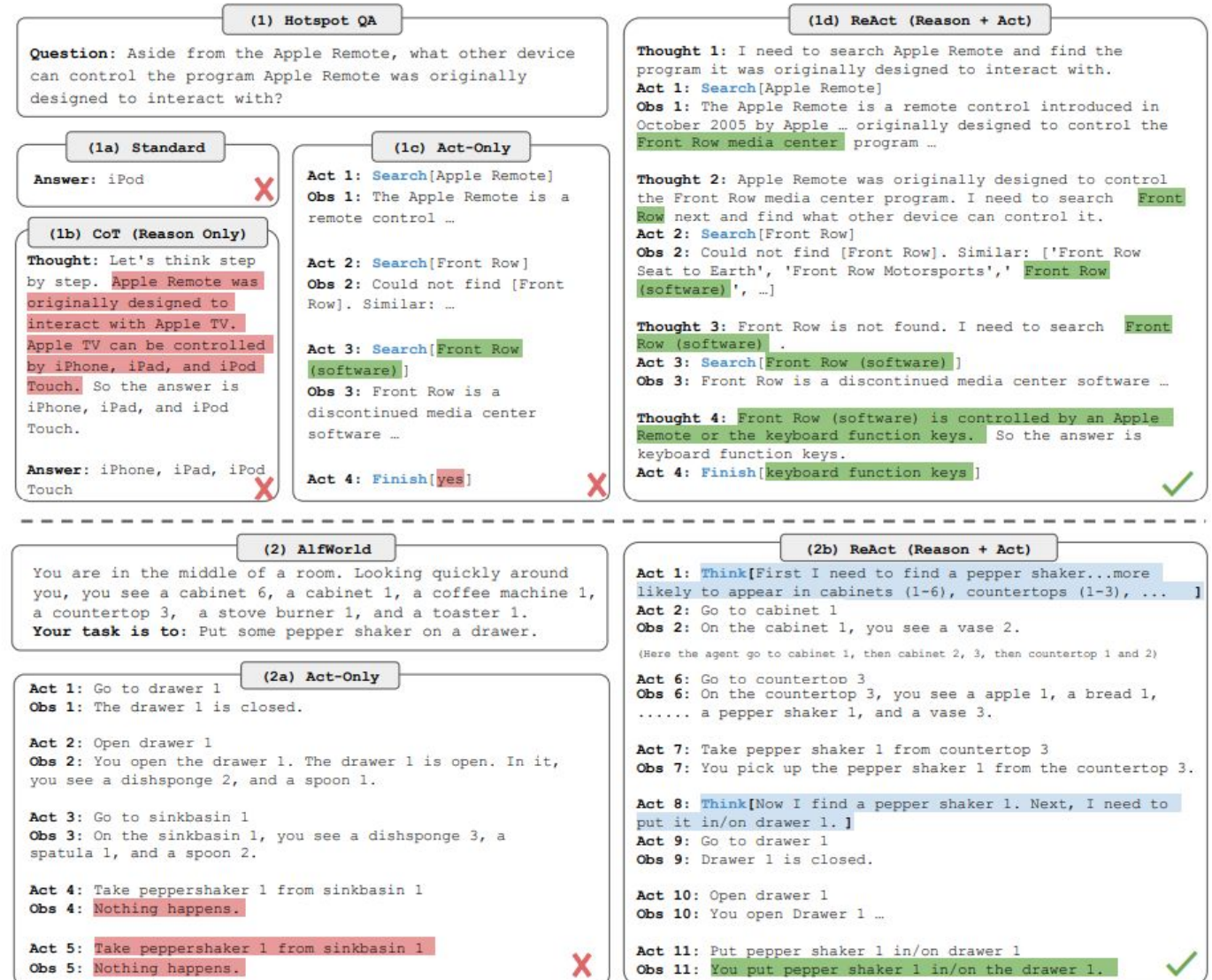
Tabular data, spreadsheets organized in rows and columns, are ubiquitous across scientific fields, from biomedicine to particle physics to economics and climate science^{1,2}. The fundamental prediction task of filling in missing values of a label column based on the rest of the columns is essential for various applications as diverse as biomedical risk models, drug discovery and materials science. Although deep learning has revolutionized learning from raw data and led to numerous high-profile success stories^{3–5}, gradient-boosted decision trees^{6–9} have dominated tabular data for the past 20 years. Here we present the Tabular Prior-data Fitted Network (TabPFN), a tabular foundation model that outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. As a generative transformer-based foundation model, this model also allows fine-tuning, data generation, density estimation and learning reusable embeddings. TabPFN is a learning algorithm that is itself learned across millions of synthetic datasets, demonstrating the power of this approach for algorithm development. By improving modelling abilities across diverse fields, TabPFN has the potential to accelerate scientific discovery and enhance important decision-making in various domains.

Thema 12: Prompting Strategies

(Michael Graber)

Wei et al., Chain-of-thought prompting, 2022.

Yao et al., React, 2022.



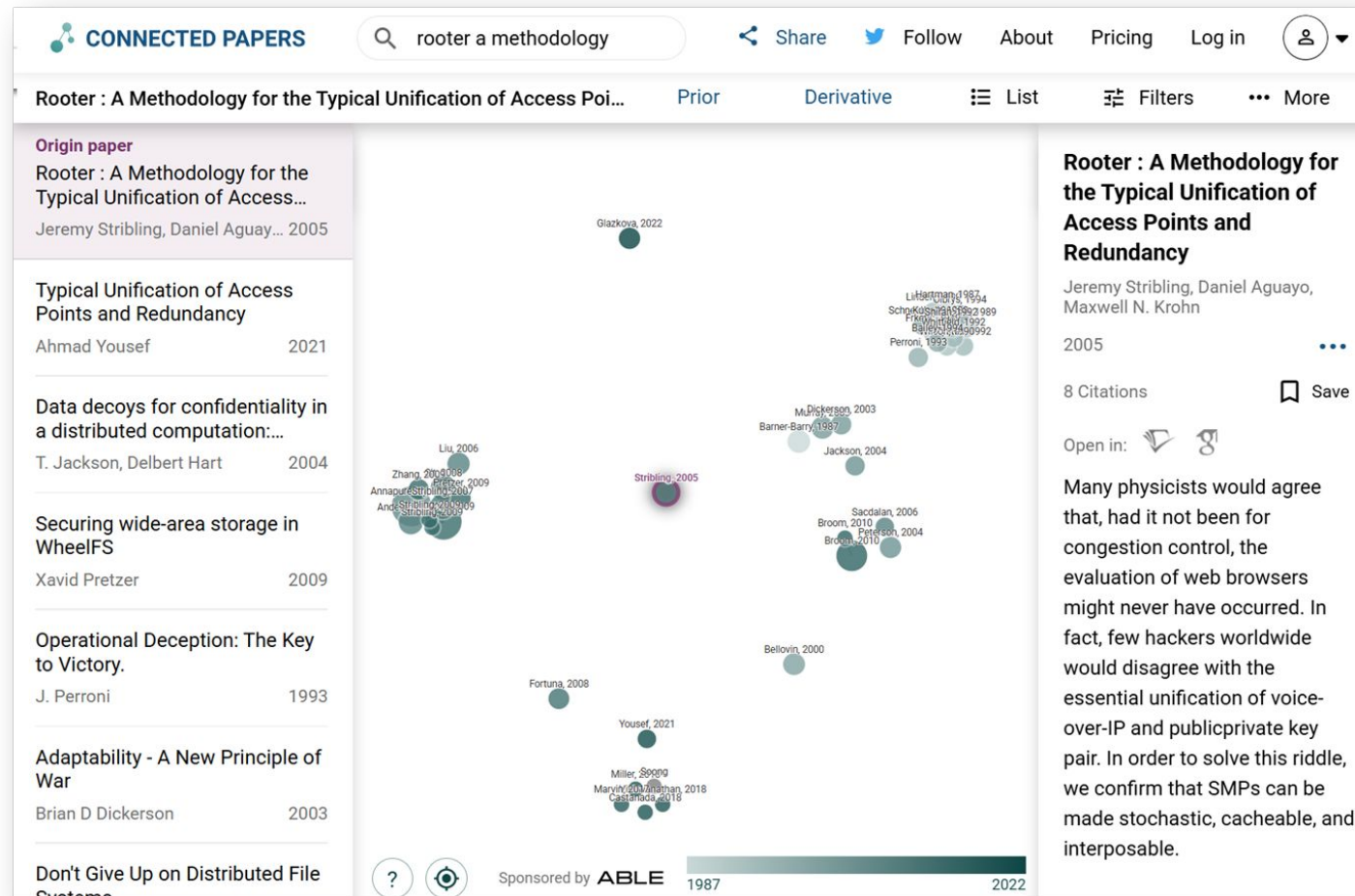
https://scholar.google.com/


The screenshot shows the Google Scholar search results for the query "rooter". The search bar at the top contains the word "rooter". Below the search bar, the results are listed. The first result is "Rooter: A methodology for the typical unification of access points and redundancy" by J Stribling, D Aguayo, M Krohn, published in the Journal of Irreproducible Results, 2005. This result is highlighted with a red box. To the right of the search bar, there is a red box containing the text "[PDF] dur-a-avaler.com". Below the search bar, there is a red box containing the text "[PDF] dur-a-avaler.com". On the left side of the search results, there is a red box containing the text "Cite". Below the search results, there is a red box containing the text "Cite". At the bottom of the search results, there is a red box containing the text "Cite".

The image shows a dropdown menu for citation styles. The menu is titled "Cite" and lists several citation styles: MLA, APA, Chicago, Harvard, and Vancouver. Each style is followed by a brief description of the citation format. For example, the MLA style is described as "Stribling, Jeremy, Daniel Aguayo, and Maxwell Krohn. 'Rooter: A methodology for the typical unification of access points and redundancy.' Journal of Irreproducible Results 49.3 (2005): 5." The Vancouver style is described as "Stribling J, Aguayo D, Krohn M. Rooter: A methodology for the typical unification of access points and redundancy. Journal of Irreproducible Results. 2005 Jul;49(3):5."

The image shows two overlapping document pages. The top page is a research paper titled "Rooter: A methodology for the typical unification of access points and redundancy" by J Stribling, D Aguayo, M Krohn, published in the Journal of Irreproducible Results, 2005. The bottom page is a research paper titled "Energy-efficient logarithmic square rooter for error-resilient applications" by N Arya, M Pattanaik, GK Sharma, published in the IEEE Transactions on Very Large Scale Integration (VLSI) Systems, 2021. Both pages have red annotations. On the top page, there is a red box around the title and authors. On the bottom page, there is a red box around the title and authors. There is also a red box around the word "rooter" in the bottom page.

<https://www.connectedpapers.com/>





Cornell University

We gratefully acknowledge support from the Simons Foundation, member institutions, and all contributors. [Donate](#)

arXiv > cs > arXiv:2402.10174

All fields
Search

[Help](#) | [Advanced Search](#)

Computer Science > Logic in Computer Science

[Submitted on 15 Feb 2024]


Overapproximation of Non-Linear Integer Arithmetic for Smart Contract Verification


Petra Hozzová, Jaroslav Bendík, Alexander Nutz, Yoav Rodeh

The need to solve non-linear arithmetic constraints presents a major obstacle to the automatic verification of smart contracts. In this case study we focus on the two overapproximation techniques used by the industry verification tool Certora Prover: overapproximation of non-linear integer arithmetic using linear integer arithmetic and using non-linear real arithmetic. We compare the performance of contemporary SMT solvers on verification conditions produced by the Certora Prover using these two approximations against the natural non-linear integer arithmetic encoding. Our evaluation shows that the use of the overapproximation methods leads to solving a significant number of new problems.

Comments: 13 pages, 2 figures, presented at The International Conference on Logic for Programming, Artificial Intelligence and Reasoning (LPAR) 2023

Subjects: **Logic in Computer Science (cs.LO)**

Cite as: arXiv:2402.10174 [cs.LO]
(or arXiv:2402.10174v1 [cs.LO] for this version)
<https://doi.org/10.48550/arXiv.2402.10174> 

Related DOI: <https://doi.org/10.29007/h4p7> 

Submission history

From: Petra Hozzová [\[view email\]](#)

[v1] Thu, 15 Feb 2024 18:23:06 UTC (190 KB)


[Bibliographic Tools](#) [Code, Data, Media](#) [Demos](#) [Related Papers](#) [About arXivLabs](#)

Demos

☐ Replicate ([What is Replicate?](#))

Access Paper:

- Download PDF
- TeX Source
- Other Formats

 [view license](#)

Current browse context:

cs.LO

[< prev](#) | [next >](#)

[new](#) | [recent](#) | [2402](#)

Change to browse by:


[cs](#)

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

[Export BibTeX Citation](#)

Bookmark



Paper-Präsentation im Journal Club

1. **Verstehe das Paper tiefgehend**
 - Lies es mehrmals und fokussiere dich auf **Motivation, Methode, Experimente und Fazit**.
 - Identifiziere den **Hauptbeitrag** und warum er relevant ist.
2. **Fasse die Kernpunkte zusammen**
 - Wer sind die **Autoren**
 - **Problemstellung & Motivation**: Warum wurde das Paper geschrieben?
 - **Methode**: Erkläre den Ansatz verständlich mit Diagrammen.
 - **Ergebnisse**: Zeige die wichtigsten Experimente und Erkenntnisse.
 - **Stärken & Schwächen**: Was ist gut, wo gibt es Einschränkungen?
3. **Nutze visuelle Hilfsmittel**
 - **Weniger Text, mehr Diagramme** zur Veranschaulichung.
 - Falls sinnvoll, zeige **Code-Beispiele oder Demos**.
4. **Mache es interaktiv**
5. **Üben & Timing beachten**
 - **Kurz halten (10-15 min)**, um Zeit für Fragen zu lassen.
 - Bereite dich auf **kritische Fragen** vor und überlege mögliche Antworten.