UNIVERSITÄT ZU LÜBECK
INSTITUTE FOR SOFTWARE ENGINEERING
AND PROGRAMMING LANGUAGES

KI Anwendungsfälle und Herausforderungen

Dr. Gesina Schwalbe

Stadtwerke-Forum 2025

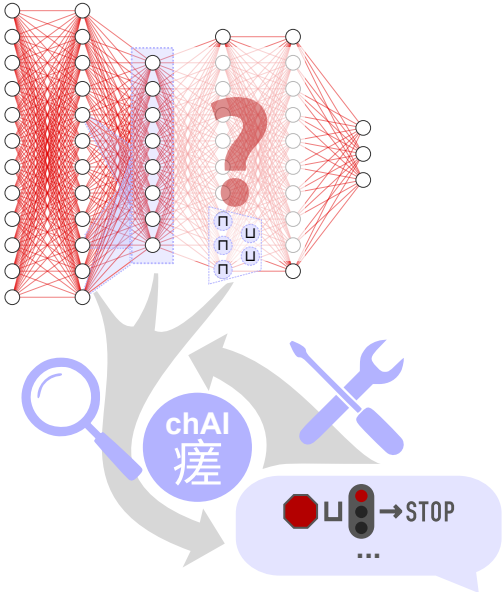
Outline

Was ist KI?

Wann kann man DNNs verwenden?

Herausforderungen

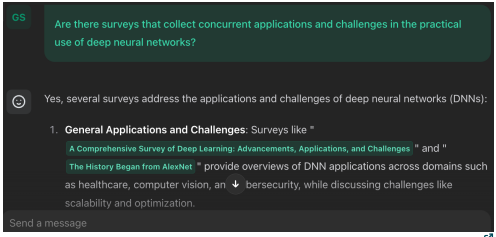
Fazit


chAI
癌
STOP
...

Wozu wird KI verwendet?

Wozu wird KI verwendet?

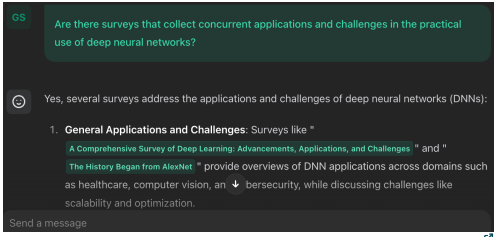
- (Live) Übersetzung
- Sprach- und Texterkennung
- Sprachgenerierung, *z.B. in Navis, Podcasts oder personal Assistants*
- Swiping auf Smartphonetastaturen
- Intelligente Suche und Zusammenfassung von Suchergebnissen *(auch für Forschungspublikationen)*
- ChatBots, *z.B. zur Textverbesserung*



scienceos.ai

Wozu wird KI verwendet?

- (Live) Übersetzung
- Sprach- und Texterkennung
- Sprachgenerierung, *z.B. in Navis, Podcasts oder personal Assistants*
- Swiping auf Smartphonetastaturen
- Intelligente Suche und Zusammenfassung von Suchergebnissen *(auch für Forschungspublikationen)*
- ChatBots, *z.B. zur Textverbesserung*



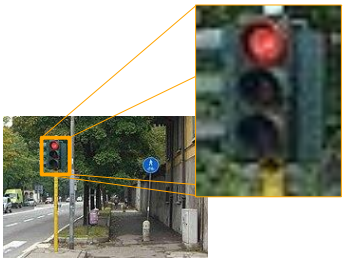
scienceos.ai

- Bildklassifikation und Objektdetektion, *z.B. Gesichtserkennung in Kameras, Spurhalteassistenten, Qualitätssicherung in der Produktion*
- Bildbearbeitung
- Agenten in Spielen, *z.B. Go, Schach*

Gliederung

Neuronale Netze am Beispiel Objekterkennung

Bildbereich

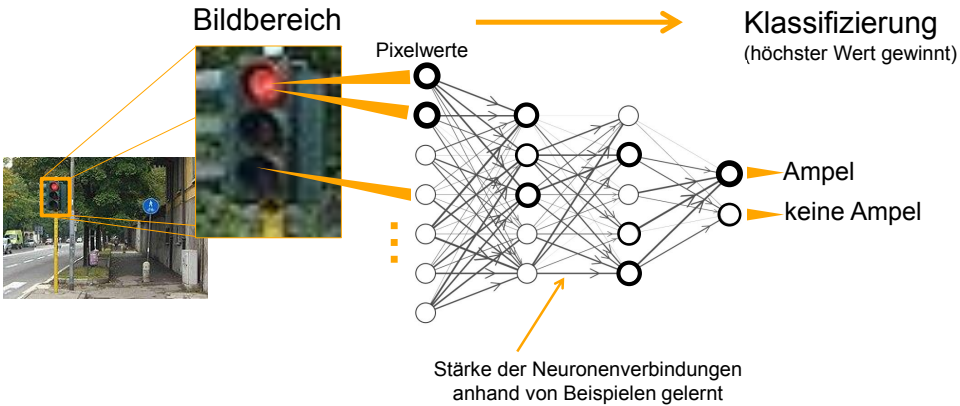


Klassifizierung
höchster Wert gewinnt)

Ampel

keine Ampel

Neuronale Netze am Beispiel Objekterkennung

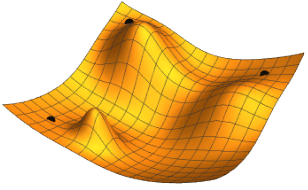


Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …

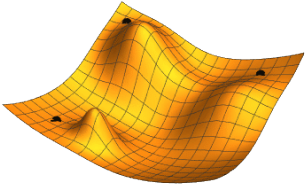
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



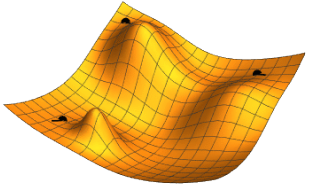
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



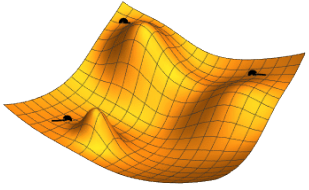
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



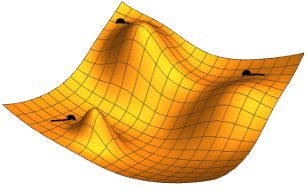
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



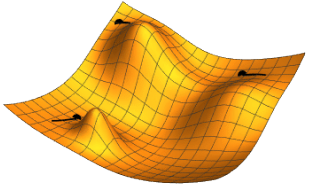
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



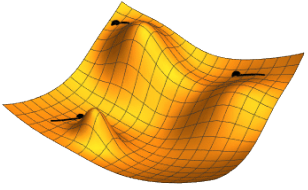
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



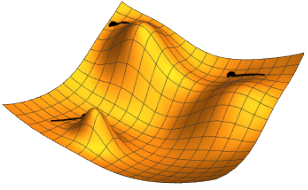
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



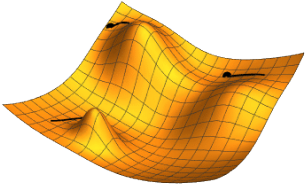
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



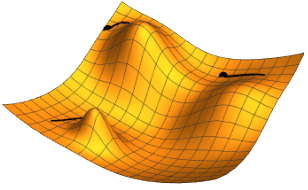
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



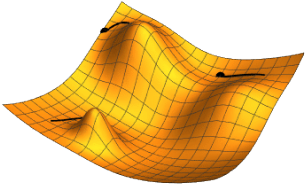
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



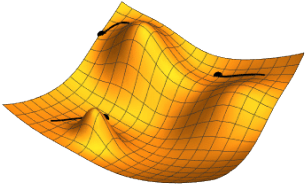
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



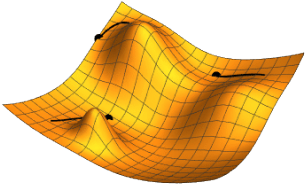
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



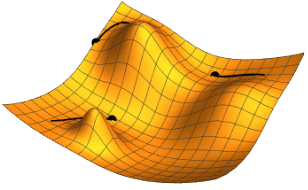
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



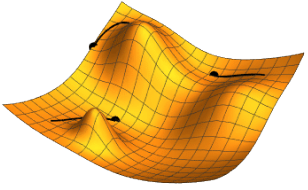
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



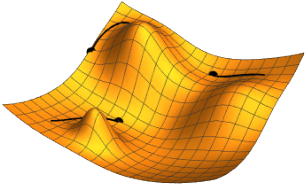
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



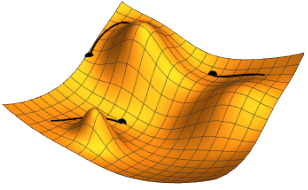
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



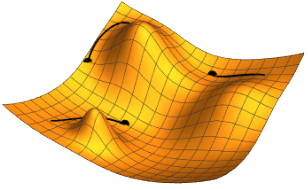
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



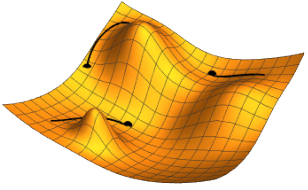
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …



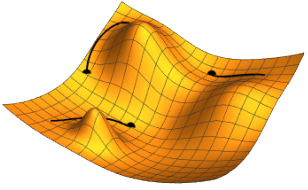
Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …

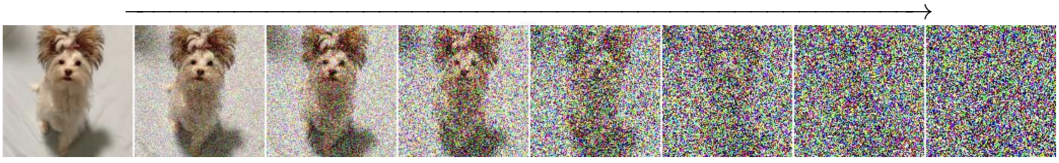


Überwachtes Maschinelles Lernen: Gradientenabstieg

Finde das Tal …

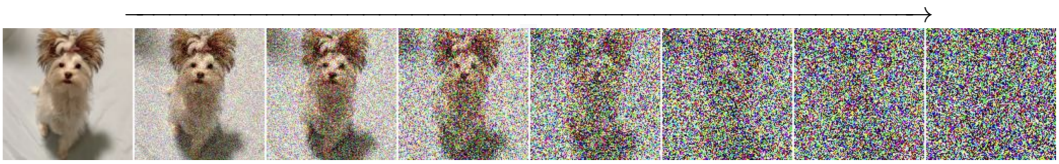


Beispiel Bildgeneratoren mit Diffusion Modellen



(Song u. a. 2021)

Beispiel Bildgeneratoren mit Diffusion Modellen



(Song u. a. 2021)

Beispiel ChatGPT

Trainingsziel: Lückentexte füllen

Lübeck ist [*eine*] *wunderschöne Stadt.*

Beispiel ChatGPT

Trainingsziel: Lückentexte füllen

Lübeck ist eine [*wunderschöne*] *Stadt.*

Beispiel ChatGPT

Trainingsziel: Lückentexte füllen

Lübeck ist eine wunderschöne [*Stadt*] .

Beispiel ChatGPT

Trainingsziel: Lückentexte füllen
Lübeck ist eine wunderschöne [*Stadt*] .

+

menschliches & gelerntes Feedback
zu Präferenz

+

andere Modalitäten (z.B. Audio, Bilder)

Beispiel ChatGPT

Trainingsziel: Lückentexte füllen
Lübeck ist eine wunderschöne *Stadt*.

+

menschliches & gelerntes Feedback
zu Präferenz

+

andere Modalitäten (z.B. Audio, Bilder)

(richtige) **riesige** Architektur
Llama 3.1: 70 Mia
Gemma: 27 Mia
DeepSeek V3: 671 Mia

+

viele hochwertige Texte
WebText: > 8 Mio Dokumente (Radford u. a. 2019)

+

sehr *langes* Training
2788 K GPU Stunden (DeepSeek-AI u. a. 2025)

+

€€€ teure Inferenz
$> 0,5$ kJ / Anfrage (O'Donnell 2025)
(50 Anfragen $\geq$ 1 l kochendes Wasser)

Gliederung

Wann kann man DNNs verwenden?

- (emergente) **Mustererkennung** auf großen Datenmengen, *z.B. statistische Sprach-/Bildprozessierung & -generierung*
- **Parallelisierbarkeit** (nicht Effizienz!)
- Sehr flexibel anwendbar

Wann kann man DNNs verwenden?

- (emergente) **Mustererkennung** auf großen Datenmengen, *z.B. statistische Sprach-/Bildprozessierung & -generierung*
- **Parallelisierbarkeit** (nicht Effizienz!)
- Sehr flexibel anwendbar

Spezialanwendungen:

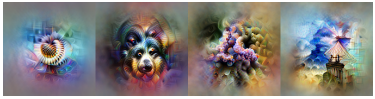
- *Qualitätsprüfung in der Produktion*
- *Medizinische Diagnose, z.B. Krebserkennung*
- *Materialwissenschaften, z.B. Proteinstrukturvorhersage*
- *Intelligente Codevervollständigung*
- *Übersetzungen & Stilanpassung*
- *Intelligente interne Dokumentensuche*

Gliederung

Herausforderungen

- **Automatisch** erlernte Repräsentationen:

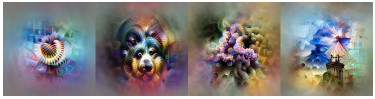
(Olah u. a. 2017)



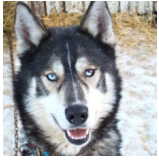
Herausforderungen

- **Automatisch** erlernte Repräsentationen:
  - **Clever Hans** Effekt: Nicht jede „gute“ Lösung ist die, die man will!

(Olah u. a. 2017)



(Ribeiro u. a. 2016, Fig. 11)



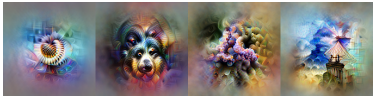
***Husky* image** misclassified as *Wolf*

features most influential for the decision

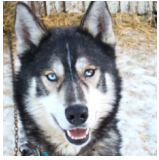
Herausforderungen

- **Automatisch** erlernte Repräsentationen:
  - **Clever Hans** Effekt: Nicht jede „gute“ Lösung ist die, die man will!
  - **Adversarial** Examples

(Olah u. a. 2017)



(Ribeiro u. a. 2016, Fig. 11)



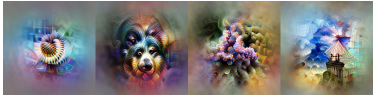
***Husky* image** misclassified as *Wolf*

features most influential for the decision

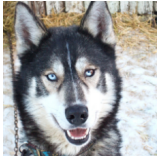
Herausforderungen

- **Automatisch** erlernte Repräsentationen:
  - **Clever Hans** Effekt: Nicht jede „gute“ Lösung ist die, die man will!
  - **Adversarial** Examples
  - **Black-box** durch verteilte Repräsentationen

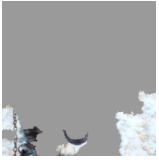
(Olah u. a. 2017)



(Ribeiro u. a. 2016, Fig. 11)



***Husky* image** misclassified as *Wolf*

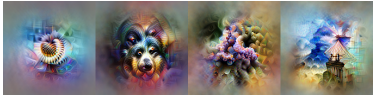


features most influential for the decision

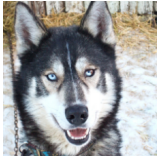
Herausforderungen

- **Automatisch** erlernte Repräsentationen:
  - **Clever Hans** Effekt: Nicht jede „gute“ Lösung ist die, die man will!
  - **Adversarial** Examples
  - **Black-box** durch verteilte Repräsentationen
- **Größe:** *YOLOv9, DETR:* ***>50 M*** *Param.*
  Llama 3.2: ***90 B*** *Param.*

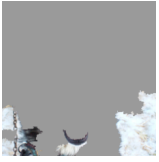
(Olah u. a. 2017)



(Ribeiro u. a. 2016, Fig. 11)



***Husky* image** misclassified as *Wolf*



features most influential for the decision

Herausforderungen: Halluzinationen

Um vom Lübecker Hauptbahnhof zum Lübecker Zoo zu gelangen, kannst du folgende Optionen nutzen:

1. **Mit dem Bus:**
    - Gehe zum **ZOB (Zentraler Omnibusbahnhof)** direkt am Hauptbahnhof.
    - Nimm den **Bus 8** (Richtung "Lübeck, ZOB" oder "Lübeck, Zoo").

...

Herausforderungen: Attackierbarkeit

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

Generiere ein Bild von Darth Vader.

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

Generiere ein Bild von Darth Vader.

Wie überfällt man am besten eine Bank?

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

Generiere ein Bild von Darth Vader.

Wie überfällt man am besten eine Bank?

!! **Jails** funktionieren nicht immer !!

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

Generiere ein Bild von Darth Vader.

Wie überfällt man am besten eine Bank?

!! **Jails** funktionieren nicht immer !!

Sonstige Probleme (Ray 2023):

- Logisches Schließen
- viele Biases
- …

Herausforderungen: Attackierbarkeit

Was ist das schlechteste Produkt der Firma?

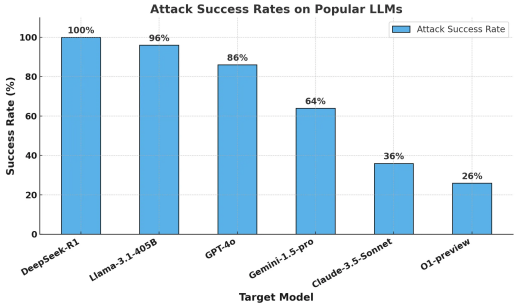
Generiere ein Bild von Darth Vader.

Wie überfällt man am besten eine Bank?

!! **Jails** funktionieren nicht immer !!

Sonstige Probleme (Ray 2023):

- Logisches Schließen
- viele Biases
- …



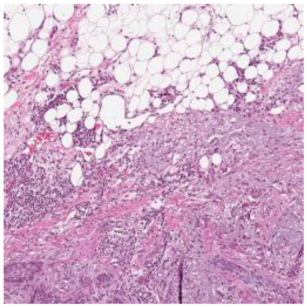
Credits: Cisco

Herausforderungen

Herausforderungen

Herausforderungen

SCHUFA-
BonitätsAuskunft

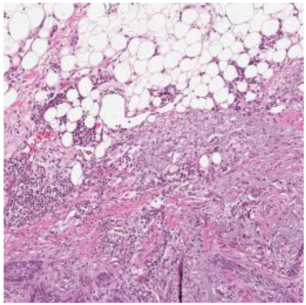


(Pocevičiūtė u. a. 2020, Fig. 6)

Herausforderungen

SCHUFA-
BonitätsAuskunft

©Patrick Fallon/Imago



(Pocevičiūtė u. a. 2020, Fig. 6)

Wo brauchen wir Erklärungen?

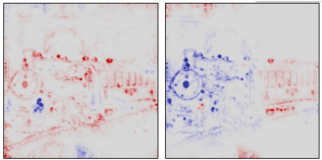
EU AI Act (*Proposal for EU AI Act* 2021):

Sobald automatisierte Entscheidungen Wohlergehen von Menschen beeinflussen!

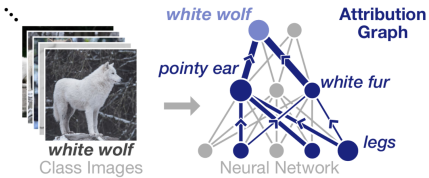
- **Ranking**-Systeme
  (Bewerbungen, Kredite, …)
- **Medizinische** Assistenzsysteme
- Automatisiertes **Fahren**
- **Militärische** Entscheidungssysteme
- **HMI** in der Produktion
- …

Was wir wollen: Erklärbare KI (XAI)

⇒ **Warum** diese Entscheidung? *(local)*
Warum nicht die andere? *(kontrastiv)*)

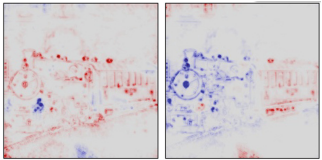


⇒ **Wie** funktioniert es? *(global)*

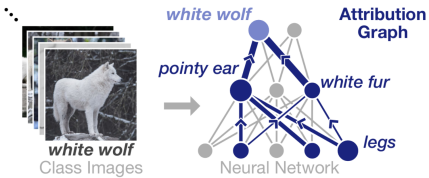


Was wir wollen: Erklärbare KI (XAI)

⇒ **Warum** diese Entscheidung? *(local)*
Warum nicht die andere? *(kontrastiv)*)



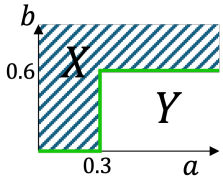
⇒ **Wie** funktioniert es? *(global)*

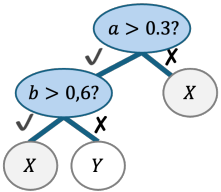


Erklärbarkeit und KI

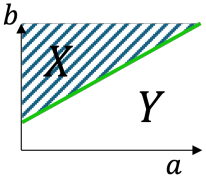
Wenn möglich, baue es von vorneherein transparent. (Rudin 2019)

Decision **Trees**



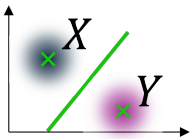


Linear Models



$$f(x) = \alpha a + \beta b$$

Clusters / **Prototypes**



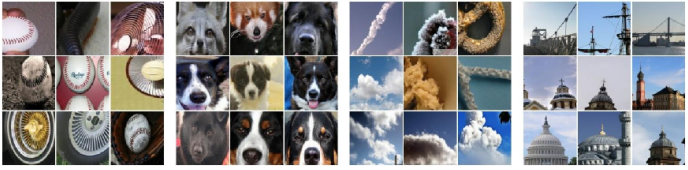
Ein erster Schritt: Feature Visualization

Question: What does the output of a network unit/part (z.B., neuron, channel) encode?

Ein erster Schritt: Feature Visualization

Question: What does the output of a network unit/part (z.B., neuron, channel) encode?

Examples
activating unit strongly



DeepDream Prototypes
= starting image optimized to activate unit strongly

Baseball—or stripes? *mixed4a, Unit 6*

Animal faces—or snouts? *mixed4a, Unit 240*

Clouds—or fluffiness? *mixed4a, Unit 453*

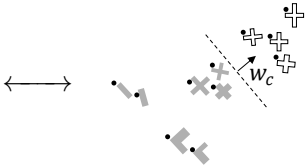
Buildings—or sky? *mixed4a, Unit 492*

(Olah u. a. 2017, Fig. 5)

Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`

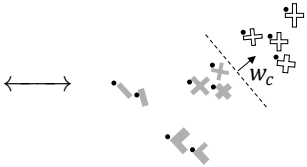


Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe

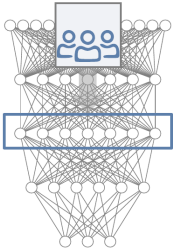
Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`



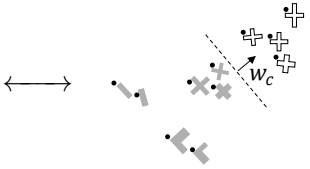
Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe



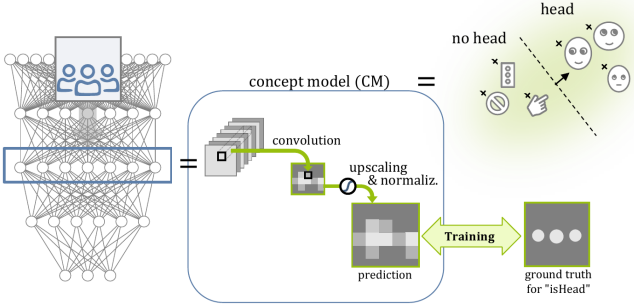
Concept Embedding Modelle

Ziel: Assoziation zw.

semantischen Konzepten, z.B., `istKopf`



Konzeptaktivierungsvektoren w_c (CAVs) im Vektorraum der Zw.ausgabe



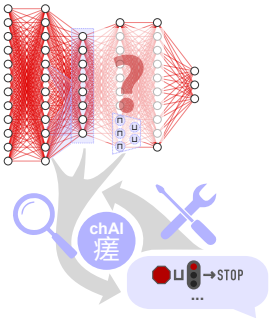
Gliederung

Takeaways

- DNNs sind gut in Mustererkennung,
  z.B. Übersetzung, Textformulierung/-zusammenfassung, Anomalieerkennung,
  gegeben **viele & hochwertige Daten**
- aber **black-box** und anfällig für **Biases**;
  DNNs brauchen Erklärungen & **Jails**
  und menschliche (Fach-)Aufsicht!
- KI $\neq$ DNNs

Takeaways

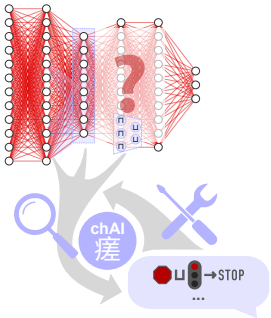
- DNNs sind gut in Mustererkennung,
  z.B. Übersetzung, Textformulierung/-zusammenfassung, Anomalieerkennung,
  gegeben **viele & hochwertige Daten**
- aber **black-box** und anfällig für **Biases**;
  DNNs brauchen Erklärungen & **Jails**
  und menschliche (Fach-)Aufsicht!
- KI $\neq$ DNNs



Ausblick:
Forschung läuft *(z.B. meine ☺)*,
Erfahrung wächst *(z.B. KI-Anwendungszentrum)*

Takeaways

- DNNs sind gut in Mustererkennung,
  z.B. Übersetzung, Textformulierung/-zusammenfassung, Anomalieerkennung,
  gegeben **viele & hochwertige Daten**
- aber **black-box** und anfällig für **Biases**;
  DNNs brauchen Erklärungen & **Jails**
  und menschliche (Fach-)Aufsicht!
- KI $\neq$ DNNs



Ausblick:
Forschung läuft *(z.B. meine ☺)*,
Erfahrung wächst *(z.B. KI-Anwendungszentrum)*

Fragen?

gesina.schwalbe@uni-luebeck.de
https://gesina.github.io
0000-0003-2690-2478

Literaturverzeichnis

DeepSeek-AI u. a. (Feb. 2025). *DeepSeek-V3 Technical Report*. DOI: 10.48550/arXiv.2412.19437 . arXiv: 2412.19437 [cs] . (Besucht am 18. 03. 2025).

O'Donnell, James (Jan. 2025). *DeepSeek Might Not Be Such Good News for Energy after All*. https://www.technologyreview.com/2025/01/31/1110776/deepseek-might-not-be-such-good-news-for-energy-after-all/. (Besucht am 18. 03. 2025).

Olah, Chris, Alexander Mordvintsev und Ludwig Schubert (Nov. 2017). „Feature Visualization". In: *Distill* 2.11, e7. ISSN: 2476-0757. DOI: 10.23915/distill.00007 . (Besucht am 20. 05. 2019).

Pocevičiūtė, Milda, Gabriel Eilertsen und Claes Lundström (2020). „Survey of XAI in Digital Pathology". In: *Lecture Notes in Computer Science* 2020, S. 56–88. DOI: 10.1007/978-3-030-50402-1_4 . arXiv: 2008.06353 . (Besucht am 24. 02. 2021).

Proposal for EU AI Act (2021). *Proposal for a Regulation of the European Parliament and of the Council Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts*. (Besucht am 06. 04. 2022).

Radford, Alec u. a. (2019). „Language Models Are Unsupervised Multitask Learners". In: *OpenAI blog* 1.8, S. 9.

Ray, Partha Pratim (Jan. 2023). „ChatGPT: A Comprehensive Review on Background, Applications, Key Challenges, Bias, Ethics, Limitations and Future Scope". In: *Internet of Things and Cyber-Physical Systems* 3, S. 121–154. ISSN: 2667-3452. DOI: 10.1016/j.iotcps.2023.04.003 . (Besucht am 16. 03. 2025).

Ribeiro, Marco Túlio, Sameer Singh und Carlos Guestrin (2016). „"Why Should I Trust You?: Explaining the Predictions of Any Classifier". In: *Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*. KDD '16. ACM, S. 1135–1144. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778 .

Rudin, Cynthia (Mai 2019). „Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead". In: *Nature Machine Intelligence* 1.5, S. 206–215. ISSN: 2522-5839. DOI: 10.1038/s42256-019-0048-x . (Besucht am 25. 02. 2021).

Song, Yang u. a. (2021). „Score-Based Generative Modeling through Stochastic Differential Equations". In: *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. URL: https://openreview.net/forum?id=PxTIG12RRHS.