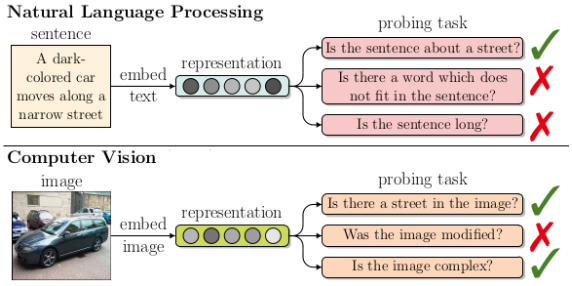
Hauptseminar

Thema: Welche Konzepte sind in gelernten Embeddings enthalten?

Viele aktuelle Deep-Learning-Architekturen nutzen ein Embedding als zentrales Mittel zur Codierung von Informationen. Die Embeddings werden dabei als Repräsentation gelernt, ohne dass Label vorgeben, was in den Embeddings enthalten ist. Dadurch ist kaum nachvollziehbar, ob bestimmte Konzepte, wie Vordergrund und Hintergrund oder das Vorhandensein bestimmter Farben, Texturen oder Symbole bei Architekturen aus dem Vision-Bereich, repräsentiert werden. Probing ist eine Technik, die dazu dient, zu überprüfen, ob bestimmte Konzepte im Embedding codiert sind. Dazu wird in der Regel eine vollverschaltete Ausgabeschicht genutzt, um zu überprüfen, ob Klassifikationsaufgaben gelöst werden können, die das Vorhandensein des Konzepts bedingen. Im Bereich der Sprachverarbeitung konnte in [9-15, 25-29] gezeigt werden, dass Konzepte, wie Wahrheitsgehalt sowie Geschlecht oder Stellung eines Worts im Satz, meistens entweder in den Unterräumen des Embeddings oder als Vektorrichtungen im Embedding enthalten sind. Darauf basierend wurde Probing auch auf Vision-Modelle [1-8, 16, 17], Vision-Language-Models [18-20] und Autoencoder [21] angewendet. Ziel dieses Hauptseminars ist es aufzubereiten, wie Probing bei Vision-Modellen angewendet wurde und welche Konzepte identifiziert werden konnten. Außerdem soll auch eine kritische Auseinandersetzung [3, 22-24] mit der Eignung des Probings für die Identifikation von Konzepten in Embeddings erfolgen.

Aufgabenstellung:

- Aufbereitung der Grundidee des Probing
- Systematische Übersicht des State of the Art zu verschiedenen Ansätzen wie Konzepte in Embeddings identifiziert werden können ausgehend von der aufgelisteten Literatur
- Übersicht zur Anwendung des Probing bei Vision-Modellen und zu damit identifizierten Konzepten, die in Embeddings von Vision-Modellen enthalten sind
- Kritische Auseinandersetzung mit der Eignung des Probing
- Vortrag mit Überblickscharakter im Rahmen des Hauptseminars



Grundprinzip des Probing. Bildquelle: [16]

Geeignet für:

Bachelor- / Masterstudiengänge

Themengebiet / Schwerpunkte:

Deep Learning

Erforderliche Vorkenntnisse:

Guter Abschluss der Vorlesung „Neuroinformatik und Maschinelles Lernen“ und Erfahrungen im Bereich Deep Learning
oder erfolgreicher Abschluss der Vorlesung „Deep Learning for Computer Vision“

Zu verwendende Literatur:

- Literatur entsprechend Literaturverzeichnis auf nächster Seite
- Elektronische Literaturdatenbank des FG NI&KR mit Recherchemöglichkeiten
- Elektronische Konferenzproceedings-Datenbank des FG NI&KR
- IEEE Recherchesystem www.ieeexplore.ieee.org (nur aus dem Uni-Netz bzw. via VPN)
- Google Scholar scholar.google.com
- Suche nach ähnlichen Publikationen connectedpapers.com, arxiv-sanity-lite.com
- Proceedings der relevanten Konferenzen (NeurIPS, ICML, ICLR, IJCNN, WCCI, ICANN, CVPR, ICCV, ECCV, BMVC, AVSS, ICPR, ICIP, …)

Betreuer: Dr. Markus Eisenbach (Markus.Eisenbach@tu-ilmenau.de)
Betr. Hochschullehrer: Prof. Dr. H.-M. Groß
Bearbeiter: **offen**

Literatur

[1] Mikriukov et al.: GCPV: Guided Concept Projection Vectors for the Explainable Inspection of CNN Feature Spaces. arXiv, 2023.
[2] Gupta et al.: Concept Distillation: Leveraging Human-Centered Explanations for Model Improvement. arXiv, 2023.
[3] Resnick et al.: Probing the State of the Art: A Critical Look at Visual Representation Evaluation. arXiv, 2019.
[4] Rashtchian et al.: Probing the Equivariance of Image Embeddings. NeurIPS, 2023.
[5] Luo et al.: Towards A Unified Neural Architecture for Visual Recognition and Reasoning. ICML Workshop, 2023.
[6] Rojas et al.: Probing Embedding Spaces in Deep Neural Networks. NeurIPS, 2020.
[7] Geißler et al.: Latent Inspector: An Interactive Tool for Probing Neural Network Behaviors Through Arbitrary Latent Activation. IJCAI, 2023.
[8] Graziani et al.: Uncovering Unique Concept Vectors through Latent Space Decomposition. arXiv, 2023.
[9] Mikolov et al.: Linguistic regularities in continuous space word representations. COLING, 2013.
[10] Köhn: What's in an embedding? analyzing word embeddings through multilingual evaluation. EMNLP, 2015.
[11] Gupta et al.: Distributional vectors encode referential attributes. EMNLP, 2015.
[12] Rogers et al.: What's in your embedding, and how it predicts task performance. COLING, 2018.
[13] Conneau et al.: What you can cram into a single vector: Probing sentence embeddings for linguistic properties. ACL, 2018.
[14] Yaghoobzadeh et al.: Probing for semantic classes: Diagnosing the meaning content of word embeddings. ACL, 2019.
[15] Rütte et al.: A Language Model's Guide Through Latent Space. arXiv, 2024.
[16] Basaj et al.: Explaining Self-Supervised Image Representations with Visual Probing. IJCAI, 2021.
[17] Oleszkiewicz et al.: Visual Probing: Cognitive Framework for Explaining Self-Supervised Image Representations. IEEE Access, 2023.
[18] Cao et al.: Behind the Scene: Revealing the Secrets of Pre-trained Vision-and-Language Models. ECCV, 2020.
[19] Salin et al. (2022): Are Vision-Language Transformers Learning Multimodal Representations? A Probing Perspective. AAAI, 2022.
[20] Lindström et al.: Probing Multimodal Embeddings for Linguistic Properties: the Visual-Semantic Case. COLING, 2020.
[21] Leeb et al.: Exploring the Latent Space of Autoencoders with Interventional Assays. NeurIPS, 2022.
[22] Hewitt et al.: Designing and Interpreting Probes with Control Tasks. EMNLP-IJCNLP, 2019.
[23] Belinkov: Probing Classifiers: Promises, Shortcomings, and Advances. COLING, 2022.
[24] Kumar et al.: Probing Classifiers are Unreliable for Concept Removal and Detection. NeurIPS, 2022.
[25] Burns et al.: Discovering latent knowledge in language models without supervision. arXiv, 2022.
[26] Li et al.: Inference-Time Intervention: Eliciting Truthful Answers from a Language Model. arXiv, 2023.
[27] Zou et al.: Representation Engineering: A Top-Down Approach to AI Transparency. arXiv, 2023.
[28] Mallen et al.: Eliciting Latent Knowledge from Quirky Language Models. arXiv, 2023.
[29] Marks et al.: The Geometry of Truth: Emergent Linear Structure in Large Language Model Representations of True/False Datasets. arXiv, 2023.