Mechanistic Interpretability

Papers
WT 2024/25

Frederick Riemenschneider

17.10.2024

Own Ideas

- It is possible to propose your own papers.
- If you have a proposal, please approach me as soon as possible so that I can integrate the proposal into the semester plan as effectively as possible.

History

- Elhage et al. (2022)
  - fundamental investigations of superposition
  - toy models to understand behavior
  - How do they behave and why?
- Gurnee et al. (2023)
  - sparse linear classifiers (probing) to examine individual neurons
  - What responsibilities do neurons in different layers have?
  - interesting case studies
  - mono- vs. polysemanticity

Othello

- Li et al. (2023)
  - Othello: board game
  - linear and non-linear probing to discover game states
  - “Can we understand what a model represents and what does it represent?”
  - The model seems to have a non-linear representation of the board.
- Nanda, Lee, et al. (2023)
  - more probing for the same task
  - The model seems to have a linear representation of the board.

Transformer Circuits

- Wang et al. (2022)
    - hypothesis: transformers have subnetworks that are responsible for certain tasks
    - IOI task
    - attempt to understand how the model solves this
- Conmy et al. (2023)
    - attempt to automatically uncover transformer circuits
- Shi et al. (2024)
    - "Do these circuits actually exist?"
    - critical view on the circuit hypothesis

Activation Patching

- Meng, Bau, et al. (2022)
  - activation patching = replacing some activation values with different values
  - activation patching as one way to edit models post-hoc
  - “Where do language models store their knowledge?”
- Meng, Sen Sharma, et al. (2022)
  - follow-up paper with better performance and large-scale editing
- Pinter and Elhadad (2023)
  - position paper with a critical view on editing

Dictionary Learning I

- Bills et al. (2023)
  - language models can explain the function of individual neurons
- Bricken et al. (2023)
  - obstacles: superposition and polysemanticity
  - features are not trivially represented as neurons
  - uncover features by training sparse autoencoders
- Huben et al. (2024)
  - additional material for essentially the same idea

Dictionary Learning II

- Karvonen et al. (2024)
  - "SAEs are cool, but how can we evaluate them?"
  - games as benchmarks
- Makelov et al. (2024)
  - more general approach with supervised feature dictionaries as ground truth

Self-conditioning

- Suau et al. (2022)
  - different take on how to find features
  - construct binary datasets with positive and negative examples
- Kojima et al. (2024)
  - follow-up work on multilingual language models
  - “Can we find language neurons?”

Multilingual Knowledge

- Tang et al. (2024)
  - methodology independent but very similar to Kojima et al. (2024)
  - detection via entropy
- Zhao et al. (2024)
  - focus on knowledge
  - case study showing activation patterns for different languages

Grokking

- Liu et al. (2022)
  - theory and experiments on grokking
  - “When, how, and why do language models generalize?”
- Nanda, Chan, et al. (2023)
  - more mechanistic view on grokking
  - “Is it possible to track learning progress live and predict grokking?”

Frameworks

- Ghandeharioun et al. (2024)
  - broad framework that categorizes most earlier approaches
  - studies on next word prediction, attribute extraction, and entity resolution
- Huang et al. (2024)
  - new alternative to autoencoders
  - rigorous experiments in toy settings

Hierarchical Representations

- Park et al. (2024)
  - How are hierarchies represented in language models?
  - interesting findings, experiments with WordNet
- Ahuja et al. (2024)
  - syntactical hierarchies
  - language modeling objective as reason for hierarchical generalization?

References

References

- Ahuja, Kabir et al. (2024). "Learning Syntax Without Planting Trees: Understanding When and Why Transformers Generalize Hierarchically". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=YwLgSimUIT.
- Bills, Steven et al. (2023). *Language models can explain neurons in language models*. https://openaipublic.blob.core.windows.net/neuron-explainer/paper/index.html.
- Bricken, Trenton et al. (2023). "Towards Monosemanticity: Decomposing Language Models With Dictionary Learning". In: *Transformer Circuits Thread*. https://transformer-circuits.pub/2023/monosemantic-features/index.html.
- Conmy, Arthur et al. (2023). "Towards Automated Circuit Discovery for Mechanistic Interpretability". In: *Thirty-seventh Conference on Neural Information Processing Systems*. URL: https://openreview.net/forum?id=89ia77nZ8u.

References

Elhage, Nelson et al. (2022). “Toy Models of Superposition”. In: *Transformer Circuits Thread.* https://transformer-circuits.pub/2022/toy_model/index.html.

Ghandeharioun, Asma et al. (2024). “Patchscope: A unifying framework for inspecting hidden representations of language models”. In: *arXiv preprint arXiv:2401.06102.*

Gurnee, Wes et al. (2023). “Finding Neurons in a Haystack: Case Studies with Sparse Probing”. In: *arXiv preprint arXiv:2305.01610.*

Huang, Xinting et al. (2024). “InversionView: A General-Purpose Method for Reading Information from Neural Activations”. In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: `https://openreview.net/forum?id=P7MW0FahEq`.

Huben, Robert et al. (2024). “Sparse Autoencoders Find Highly Interpretable Features in Language Models”. In: *The Twelfth International Conference on Learning Representations*. URL: `https://openreview.net/forum?id=F76bwRSLeK`.

Papers

History

Othello

Transformer Circuits

Activation Patching

Dictionary Learning I

Dictionary Learning II

Self-conditioning

Multilingual Knowledge

Grokking

Frameworks

Hierarchical Representations

References

References

- Karvonen, Adam et al. (2024). "Measuring Progress in Dictionary Learning for Language Model Interpretability with Board Game Models". In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=qzsDKwGJyB.
- Kojima, Takeshi et al. (2024). "On the Multilingual Ability of Decoder-based Pre-trained Language Models: Finding and Controlling Language-Specific Neurons". In: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics*. Mexico City, Mexico: Association for Computational Linguistics, pp. 6912–6964. URL: https://aclanthology.org/2024.naacl-long.384.
- Li, Kenneth et al. (2023). "Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=DeG07_TcZvT.

References

Liu, Ziming et al. (2022). “Towards Understanding Grokking: An Effective Theory of Representation Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=6at6rB3IZm.

Makelov, Aleksandar, Georg Lange, and Neel Nanda (2024). “Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control”. In: *ICLR 2024 Workshop on Secure and Trustworthy Large Language Models*. URL: https://openreview.net/forum?id=MHIX9H8aYF.

Meng, Kevin, David Bau, et al. (2022). “Locating and Editing Factual Associations in GPT”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh et al. URL: https://openreview.net/forum?id=-h6WAS6eE4.

Meng, Kevin, Arnab Sen Sharma, et al. (2022). “Mass Editing Memory in a Transformer”. In: *arXiv preprint arXiv:2210.07229*.

References

Nanda, Neel, Lawrence Chan, et al. (2023). “Progress measures for grokking via mechanistic interpretability”. In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=9XFSbDPmdW.

Nanda, Neel, Andrew Lee, and Martin Wattenberg (Dec. 2023). “Emergent Linear Representations in World Models of Self-Supervised Sequence Models”. In: *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Yonatan Belinkov et al. Singapore: Association for Computational Linguistics, pp. 16–30. DOI: 10.18653/v1/2023.blackboxnlp-1.2. URL: https://aclanthology.org/2023.blackboxnlp-1.2.

Park, Kiho et al. (2024). “The Geometry of Categorical and Hierarchical Concepts in Large Language Models”. In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=KXuYjuBzKo.

References

- Pinter, Yuval and Michael Elhadad (Dec. 2023). “Emptying the Ocean with a Spoon: Should We Edit Models?” In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 15164–15172. DOI: 10.18653/v1/2023.findings-emnlp.1012. URL: https://aclanthology.org/2023.findings-emnlp.1012.
- Shi, Claudia et al. (2024). “Hypothesis Testing the Circuit Hypothesis in LLMs”. In: *ICML 2024 Workshop on Mechanistic Interpretability*. URL: https://openreview.net/forum?id=ibSNv9cldu.
- Suau, Xavier, Luca Zappella, and Nicholas Apostoloff (2022). “Self-Conditioning Pre-Trained Language Models”. In: *International Conference on Machine Learning*.

References

- Tang, Tianyi et al. (Aug. 2024). "Language-Specific Neurons: The Key to Multilingual Capabilities in Large Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Lun-Wei Ku, Andre Martins, and Vivek Srikumar. Bangkok, Thailand: Association for Computational Linguistics, pp. 5701–5715. URL: https://aclanthology.org/2024.acl-long.309.
- Wang, Kevin et al. (2022). "Interpretability in the wild: a circuit for indirect object identification in gpt-2 small". In: *arXiv preprint arXiv:2211.00593*.

References

Zhao, Xin, Naoki Yoshinaga, and Daisuke Oba (Mar. 2024). "Tracing the Roots of Facts in Multilingual Language Models: Independent, Shared, and Transferred Knowledge". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, pp. 2088–2102. URL: https://aclanthology.org/2024.eacl-long.127.