Mechanistic Interpretability Readings

November 4, 2024

1 Readings

1. **Interpreting visual features of the CLIP model** (Goh et al., 2021): Early work studying neurons that are highly active for different inputs.
2. **Logit Lens** (nostalgebraist, 2024): Uses intermediate logits in GPT to interpret internal model states by mapping them to readable tokens, revealing the evolution of predictions.
3. **Probing hidden states to discover an internal world model in Othello-GPT** (Li et al., 2022): Investigates how a GPT variant forms internal representations of Othello board states, suggesting emergent understanding and saliency for human interpretation.
4. **Where LLMs Store Information** (Geva et al., 2023): Identifies mechanisms in transformers for storing and retrieving factual knowledge, focusing on the contributions of attention and MLP sublayers.
5. **Data Editing in LLMs** (Meng et al., 2022): Shows how factual associations are localized in transformer layers and demonstrates direct editing using Rank-One Model Editing (ROME).
6. **LLMs and Arithmetic** (Nikankin et al., 2024): Finds that LLMs use a collection of simple heuristics, rather than robust algorithms, to solve arithmetic tasks.
7. **Physics of Language Models Series** (Allen-Zhu, 2024): A series of studies by Zeyuan Allen-Zhu exploring the internal workings of LLMs.
8. **Sparse Autoencoders by Anthropic** (Anthropic, 2024): Presents a state-of-the-art mechanistic interpretability technique using sparse autoencoders.

References

Zeyuan Allen-Zhu. 2024. Zeyuan Allen-Zhu. [Online; accessed 4. Nov. 2024].

Anthropic. 2024. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. [Online; accessed 4. Nov. 2024].

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting Recall of Factual Associations in Auto-Regressive Language Models. *arXiv*.

Gabriel Goh, Nick Cammarata †, Chelsea Voss †, Shan Carter, Michael Petrov, Ludwig Schubert, Alec Radford, and Chris Olah. 2021. Multimodal Neurons in Artificial Neural Networks. *Distill*, 6(3):e30.

Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. 2022. Emergent World Representations: Exploring a Sequence Model Trained on a Synthetic Task. *arXiv*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and Editing Factual Associations in GPT. *arXiv*.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. 2024. Arithmetic Without Algorithms: Language Models Solve Math With a Bag of Heuristics. *arXiv*.

nostalgebraist. 2024. interpreting GPT: the logit lens. [Online; accessed 4. Nov. 2024].