Project Proposal

Topic: Using SAE features to perform activation steering

Group Members: Evžen Wybitul (ewybitul@ethz.ch), Joshua Gilligan (jgilligan@ethz.ch)

Project Advisors: Elliott Ash (elliott.ash@gess.ethz.ch)

Outline

Activation steering (Rimsky et al. 2024; Turner et al. 2023) is a technique for modifying model behavior by adding a steering vector into the activations of certain layers. This approach can adjust the model's level of sycophancy, its tendency to reject requests, or its propensity hallucinate, for example, with minimal impact on overall performance. However, current steering methods require a substantial dataset of text snippets showcasing both desired and undesired behaviors to calculate the steering vectors.

Our research will aim to explore the relationship between Sparse Auto-Encoder (SAE) features and steering vectors, potentially enabling the computation of steering vectors directly from SAE features without additional data. This advancement could both simplify the process of activation steering and shed light on the characteristics of SAE features.

We propose the following plan:

1. Replicate the steering vector computation method by Rimsky et al. (Rimsky et al. 2024) on the GPT-2 small model. The method utilizes both existing snippet datasets and new snippets generated by GPT-4.
2. Examine the relationship between the steering vectors obtained and the SAE features of GPT-2 small. Specifically, we will aim to:
    - Determine if steering vectors can be derived from SAE features.
    - Identify other potential steering vectors using this method.
    - Recognize patterns distinguishing steering-capable features from others.

Using a model with existing trained SAEs allows us to quickly validate our idea within the project's timeframe.

References

Rimsky, Nina, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. 2024. "Steering Llama 2 via Contrastive Activation Addition." March 6, 2024. https://doi.org/10.48550/arXiv.2312.06681.

Turner, Alexander Matt, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. 2023. "Activation Addition: Steering Language Models Without Optimization." November 13, 2023. https://doi.org/10.48550/arXiv.2308.10248.