Mechanistic Interpretability and Simple Games

Since their inception [Vaswani et al., 2017], transformer models have come to dominate first the field of natural language processing, and then the field of computer vision [Dosovitskiy et al., 2021]. Later on, they were finetuned to be helpful virtual assistants [Ouyang et al., 2022], and by now they can perform assignments as diverse as emotional support, ideas brainstorming, mathematical problem solving, and pair programming.

As virtual assistants are being deployed in areas of increasing importance, such as healthcare, education, and law, it is crucial to understand how they work. Mechanistic Interpretability (MI) Elhage et al. [2021] takes a bottom-up approach to understanding the inner workings of neural networks, focusing on the individual components and their interactions.

In this project, we aim to understand transformer models trained and being trained on simple games. This way, we can study the internals of these models in a controlled environment. The benefits include:

1. Training models on tasks that are simple enough to be analyzed by hand, we can compare the model's behavior to our own understanding of the task.
2. The games can be easily modified to test specific hypotheses about the model's behavior.
3. I don't assume any prior knowledge of transformers. This project can be a fun way to learn about the leading machine learning architecture of our time!
4. We can avoid the resource need of using large models. You will be able to contribute to the project even if you don't have access to a powerful GPU.

Prerequisites

Strong command of the `Python` numerical library `numpy`.

Qualifying problems

See the contents of the attached `zip` file: `qualifying_problems.py` contains functions to be written, and after the `if __name__ == "__main__":` line, there are some tests to be passed. The tests require the `npz` files to be uncompressed into the same directory.

I very much welcome partial solutions! Ordered by test cases, the problems are intended to range in difficulty from easy to very hard for someone who has never seen a transformer before. Please try to avoid explicit Python loops, as they are slow and not necessary for the problems at hand. Instead, use `numpy` functions that operate on the entire arrays at once.

You can hand in multiple versions. I will evaluate the latest submission. A valid submission must arrive to my email address by the deadline written in the general RES course description on the BSM webpage. You can expect me to answer a question before this time if it arrived to my email address at least 24 hours before this deadline.

In your email, please also write me the following:

1. Your Mathematics and Computer Science background.
2. Your Mathematics and Computer Science interests.
3. What do you find especially interesting in this project?

Have fun with the problems and hope to see you in the group!

Contact

Pál Zsámboki, `zsamboki@renyi.hu`, HUN-REN Alfréd Rényi Institute of Mathematics

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach,

R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.