Training A Language Model (GPT) From Scratch And Harnessing It

Proposal

Manuel Sandoval Flores

09.10.2023

Abstract

Will provide a concise summary of the paper's key objectives, methods, and findings and include the primary experiment contributions and implications of the research experiment.

1 Introduction

- Introduce the seminar project's motivational questions like: [3] Are GPT-3, ChatGPT, and GPT-4 not capable multiplying two numbers? [4] [2] How can we interpret what is going on inside the GPT during and after training?
- [1] Comparing standard training and curriculum learning for GPT models and analyzing the benefits.
- Reference the main results and open questions that have inspired this project and analyze approaches in the suggested papers.

2 Background Information

- (Optional) Provide introduction to GPT models and their role in natural language processing.
- Discuss the challenges with traditional training methods, which necessitate alternative approaches like curriculum learning.
  - Limitations of traditional training approaches.
  - Methods for transparency and steerability.

3 Related Work

Summarize the contributions of relevant papers:

- [1] "Curriculum Learning": Eg. Discuss the concept of curriculum learning its effects on convergence speed and how this can be a global optimization method for non-convex functions .
- [3] "Limits of Transformers on Compositionality": Eg. Trivial compositionality trivial problems: Are these errors incidental, or do they signal more substantial limitations? How LLMs reduce multi-step compositional reasoning into linearized subgraph matching, without necessarily developing systematic problem-solving skills.
- [4] "Locating and Editing Factual Associations in GPT": Eg. How direct manipulation of computational mechanisms may be a feasible approach for model editing.
- [2] "Sparse Autoencoders in Language Models": Eg. Model transparency and steerability enabled by polysemanticity, superposition and interpretability measured by automated methods.

4 Model Description

- Describe the provided minimal GPT model, including its architecture and core components.
- Explain the source of training data and any preprocessing steps, such as tokenization or data cleaning.

5 Model Analysis

- Explain the standard training approach using randomly shuffled text data.
- Describe the curriculum learning strategy used and its implementation for gradually increasing example difficulty.
- Detail the experimental setup, specifying hyperparameters, the choice of optimizer, and any unique implementation choices.
- Detail on transparency and steerability.

6 Results And Discussion

- Present a comparison of the GPT model's performance under standard and curriculum learning.

- Analyze the benefits of curriculum learning, such as faster convergence, improved generalization, and enhanced performance on complex tasks.
- Discuss any challenges faced during the experiments, including data selection and curriculum design.

7 Understanding GPT Behaviour

- Summarize key findings from [4] and [2].
- Provide insights into the inner workings of GPT models during and after training, particularly focusing on interpretability and feature analysis.

8 Conclusion

- Summarize the primary findings and contributions of the experiment project.
- Discuss the implications of the results for the field of natural language processing and deep learning referencing the initial motivational questions.
- Suggest future research experiment directions, such as exploring advanced curriculum strategies and further improving GPT model interpretability.

9 References

References

[1] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA, 2009. Association for Computing Machinery.

[2] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[3] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D. Hwang, Soumya Sanyal, Sean Welleck, Xiang Ren, Allyson Ettinger, Zaid Harchaoui, and Yejin Choi. Faith and fate: Limits of transformers on compositionality, 2023.

[4] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023.