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

Proposal

Manuel Sandoval Flores

09.10.2023

Abstract

- [1] Provide a concise summary of the paper's key objectives, methods, and findings. - [2] Include the primary experiment contributions and implications of the research experiment.

1 Introduction

- [1] Introduce the seminar project's aim: comparing standard training and curriculum learning for GPT models and analyzing the benefits. - [2] Highlight the significance of GPT models and the limitations of traditional training approaches. - [3] Reference the key papers 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. [1] Discuss the challenges with traditional training methods, which necessitate alternative approaches like curriculum learning.

3 Related Work

[1] Summarize the contributions of relevant papers: - Paper 2 "Curriculum Learning": Eg. Discuss the concept of curriculum learning in deep learning and its potential advantages. - Paper 3 "Limits of Transformers on Compositionality": Eg. Highlight challenges GPT models face in handling compositional tasks. - Paper 4 "Locating and Editing Factual Associations in GPT": Eg. Explain how this paper addresses GPT model interpretability. - Paper 5 "Sparse Autoencoders in Language Models": Eg.Describe the findings related to interpretable features in language models.

4 Model Description

[1] Describe the provided minimal GPT model, including its architecture and core components. [2] Explain the source of training data and any preprocessing steps, such as tokenization or data cleaning.

5 Model Analysis

[1] Explain the standard training approach using randomly shuffled text data. [2] Describe the curriculum learning strategy used and its implementation for gradually increasing example difficulty. [3] Detail the experimental setup, specifying hyperparameters, the choice of optimizer, and any unique implementation choices.

6 Results And Discussion

[1] Present a comparison of the GPT model's performance under standard and curriculum learning. [2] Analyze the benefits of curriculum learning, such as faster convergence, improved generalization, and enhanced performance on complex tasks. [3] Discuss any challenges faced during the experiments, including data selection and curriculum design.

7 Understanding GPT Behaviour

[1] Summarize key findings from Paper 4 (Locating and Editing Factual Associations in GPT) and Paper 5 (Sparse Autoencoders in Language Models). [2] Provide insights into the inner workings of GPT models during and after training, particularly focusing on interpretability and feature analysis.

8 Conclusion

[1] Summarize the primary findings and contributions of the experiment project. [2] Discuss the implications of the results for the field of natural language processing and deep learning referencing the initial motivational questions. [3] Suggest future research experiment directions, such as exploring advanced curriculum strategies and further improving GPT model interpretability.

9 References

References

[1] Nouha Dziri and Ximing Lu and Melanie Sclar and Xiang Lorraine Li and Liwei Jiang and Bill Yuchen Lin and Peter West and Chandra Bhagavatula and

Ronan Le Bras and Jena D. Hwang and Soumya Sanyal and Sean Welleck and Xiang Ren and Allyson Ettinger and Zaid Harchaoui and Yejin Choi. "Faith and Fate: Limits of Transformers on Compositionality." *37th Conference on Neural Information Processing Systems (NeurIPS 2023)*, 2023. arXiv:2305.18654v3

[2] Kevin Meng, David Bau, Alex Andonian, Yonatan Belinkov. "Locating and Editing Factual Associations in GPT." *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*, 2022. arXiv:2202.05262v5

[3] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, Lee Sharkey. "Sparse Autoencoders Find Highly Interpretable Features in Language Models." *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, 2023. arXiv:2309.08600v3

[4] Bengio, Yoshua and Louradour, Jérôme and Collobert, Ronan and Weston, Jason. "Curriculum learning." *None*, 2009. https://doi.org/10.1145/1553374.1553380