Andrew Rossman
Rahul Gupta
Aaron C.

Project Summary

Introduction

Natural language processing brings human understanding of language to the digital landscape through the power of AI. Large networks of digital neurons and modules come together to learn and predict the solution to a wide range of tasks. A modern implementation of this paradigm is the transformer, a deep learning architecture that takes advantage of attention to reliably output scores during inference and training. Unlike previous architectures, the transformer is built to deal with the vanishing gradient problem and is well suited for long durations of training to maximize the accuracy of a model[6]. These transformers make up the majority of large language models (LLM), being stack upon stack of different transformer layers that require a large number of parameters to manage. LLMs can be applied to a wide variety of chatbot goals including talking emulation, sentiment analysis, and parsing documents for specific information.

Medical Question Answering

Question answering is one of these many goals of language models in that when the language model is presented with a question, it can accurately produce an answer to that question with expert level accuracy and human level language comprehension. Experiments using language learning have already shown that AI can predict the outcomes of patients after their admission to the hospital using only information collected after they are admitted[1].

Other implementations pick between answering large subsets of general medical answers or specifically target an issue by leveraging their dataset[2][3]. Limitations of the chatbot such as the long answer incoherency can be addressed through local improvements such as contrastive learning or time control[4]. Overall, it becomes possible to maximize the response's accuracy and length through additional structuring on the model level while protecting against error and incoherent responses. Mistral is one such model that utilizes unique improvements to the structure of the LLM to yield an improvement that sets it apart from most other models.

Mistral 7B LLM

Mistral 7B is a recent implementation of an LLM that introduces sliding window attention and grouped query attention to more effectively handle arbitrary length labels during inference. Mistral is considered to be SoA for open source chat bots as it surpasses the best open source LLM called Llama 2 [7]. The primary differences between the two are the parameter counts and performance; Llama 2 clocks in at 13 billion parameters while Mistral only requires 7 billion and performs better at all levels [5]. Mistral proves that a smaller LLM can outperform larger models through the careful management of the language model's design. The key defining features of this architecture are the GQS and SWA.

Group query attention greatly accelerates both training and inference through the improvements it makes. It improves on previous implementations of multiquery attention to speed up inference and fine tuning by speeding up decoding[10]. GQA groups together query heads such that they each share a single key head and value head. This builds upon previous iterations of this mechanism called multihead and multiquery. Multihead reserves a head for all queries, keys, and values while multi query shares one head for several queries at once. GQA is an interpolation of these two strategies to retain the accuracy of multihead while maintaining the speed of multi-query. Mistral uses this strategy to increase speed at a minimal cost of accuracy[5].

The other primary improvement to the structure is the Sliding Window Attention, which takes advantage of the transformer's stacked layers to apply attention to information. SWA introduces a variable W that dictates the amount of previous tokens a single token can attend to. The researchers found that this reduced model latency and increased throughput due to the reduced cache availability[5]. This yields at minimum a 2x speedup.

In the trials conducted by Mistral AI, Mistral outperforms Llama 1 and 2, other open source LLMs that use 34 billion and 13 billion parameters respectively[5]. These models introduce Reinforcement Learning with Human Feedback, in that humans that submit to the bot are given the option to provide feedback to text generated, either positive or negative depending on how well the response answers the input. Llama takes this model and optimizes responses such that answers with high reward potential are distributed to users as the response.

Fine Tuning

Several options are available to fine tune LLMs towards the goal of expert level comprehension of question answering datasets. One that stuck out to us was QLoRA, a fine tuner that quantized Low Rank Adapters to reduce the memory footprint and decrease time required for training and inference using a variety of inventions[8]. Some of these include the 4-bit and double quantization along with paged optimizers to prevent memory leaks.

Expectations

We plan on utilizing the improved Mistral architecture and QLoRA fine tuning to learn and adapt to a medical dataset closely related to procedures, facts, outcomes, and the recovery processes of medical injuries in the spine. Mistral AI released and currently maintains a pretrained 7 billion parameter model over their website. Additionally, QLoRA's method is open source and freely available via their GitHub repository[9].

Our plan of execution is to download the AI and spinal injury datasets, fine-tune the model to adapt closely to the medical data via training, and then test the model against the standard Mistral implementation to prove that the chatbot can be taught important expert level information. Using a test dataset full of questions, we can evaluate correct vs incorrect answers and discover if the finetuning has improved information retention.

Conclusion

We hope to extend Mistral's small memory requirements into the realm of expert level question answering in the realm of medical treatments. By fine tuning the chatbot with our dataset, we expect that the model will be capable of distributing not only accurate information but life-saving information as well to greaten the access to medical resources.

Sources

[1] B. Aken J. M. Papaioannou, M. Mayrdorfer, K. Budde, F. A. Gers, A. Löser, "Clinical Outcome Prediction from Admission Note using Self-Supervised Knowledge Integration", Beuth University of Applied Sciences Berlin, Feb 2021.

[2] S. Suster, W Daelemans, "CliCR: A Dataset of Clinical Case Reports for Machine Reading Comprehension", Computational Linguistics & Psycholinguistics Research Center, University of Antwerp, Belgium, Mar. 2018.

[3] J. A. Fries, L. Weber, N. Seelam, G. Altay, etc.., "BIGBIO: A Framework for Data-Centric
Biomedical Natural Language Processing" Equal Contribution, Jun. 2022.

[4] R. E. Wang, E. Durmus, N. Goodman, T. B. Hashimoto, "Language Modeling via Stochastic Processes", Stanford University, May 2023.

[5] A. Q. Jiang, A. Sablayrolles, "Mistral 7B", Mistral AI, Oct. 2023.

[6] A, Vaswanit, N. Shazeer, N. Parmar, J. Uskoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, "Attention Is All You Need", Google Brain, Aug 2017.

[7] H. Touvron, L. Martin, K. Stone, P. Albert, etc…, "Llama 2: Open Foundation and Fine-Tuned Chat Models", GenAI, Meta, July 2023.

[8] T. Dettmers, A. Pagnoni, A. Holtzman, L. Zettlemoyer, "QLoRA: Efficient Fine Tuning of Quantized LLMs", University of Washington, May 2023.

[9] https://github.com/artidoro/qlora

[10] J. Ainslie, J. Lee-Thorp, M. de Jong, Y. Zemlyanskiy, F. Lebron, S. Sanghai, "GQA: Training Generalized Multi-Query Transformer Models from Multi-Head Checkpoints", Google Research, May 2023.