Week 1: Thursday: Improving Language Understanding by Generative Pre-Training

and Mistral 7B (and a blog on it)

Supplemental: palm (look at the first part, not all the evaluations!)

1. What does the "unsupervised learning" portion of GPT do? Specifically, what loss function is it optimizing? It uses a "transformer decoder". Where did the encoder go?

Ans: The unsupervised learning portion is inspired by the work done by Howard and Ruder where they pre-trained a neural network using a language modeling objective and then fine-tuning it on a target task with supervision. GPT makes use of a transformer instead of LSTMs to handle long range dependencies as the LSTMs restrict their prediction abilities to a short range. Given an unsupervised corpus of tokens, a standard language modeling objective is used to maximize the likelihood described in equation 1 in the paper.

2. What does the "supervised fine-tuning" of GPT do?

Ans: According to the paper, after training the model with the objective in equation 1, they adapt the parameters to the supervised target task. We assume a labeled dataset C, where each instance consists of a sequence of input tokens, x1, . . . , xm, along with a label y. The inputs are passed through our pre-trained model to obtain the final transformer block's activation hm,l , which is then fed into an added linear output layer with parameters Wy to predict y. They additionally found that including language modeling as an auxiliary objective to the fine-tuning helped learning by improving generalization of the supervised model and accelerated convergence.

3. What does the GPT paper mean (in English) when it says: " We use a traversal-style approach [52], where we convert structured inputs into an ordered sequence that our pre-trained model can process. These input transformations allow us to avoid making extensive changes to the architecture across tasks."

Ans: The traversal style approach mentioned in the paper refers to the method of converting structured inputs into a linear sequence of text that a model can process. This is done because GPT is a text-based model and operates only on sequences of text. By doing so, GPT can apply its language understanding and capabilities without needing major architectural changes for handling different types of inputs and tasks.

4. GPT uses masking to give a “causal” model. What does this mean and why is it needed?

Ans: This means that the model can only use information from the previous word to predict the next word in the sequence. This approach is essential for maintaining a sense of directionality in text generation to ensure that the generation process follows the natural flow of laguage.

5. What improvements are made over the original GPT model in the Mistral 7B?

Ans: From the blog we could infer that since the end of 2023, the Mixtral 8x7B[1] has become a highly popular model in the field of large language models. It has gained this popularity because it outperforms the Llama2 70B model with fewer parameters (less than 8x7B) and computations (less than 2x7B), and even exceeds the capabilities of GPT-3.5 in certain aspects.

Mistral 7B leverages grouped window attention and sliding window attention mechanisms. GQA accelerates inference speed and reduces memory requirements during decoding, enabling higher batch sizes and throughput. SWA is designed to handle longer sequences more effectively at a reduced computational cost, alleviating a common limitation in LLMs. These mechanisms contribute to Mistral 7B's enhanced performance and efficiency.
Mistral 7B can leverage system prompting to enforce output constraints, making it adaptable for various applications. It also showcases the ability to perform fine-grained content moderation, a feature useful for ensuring quality content in applications

6. What is the range of sizes of GPT models (not just OpenAI, put open source ones)? How does performance seem to scale with size? What are current popular models?

Ans:The range of sizes of GPT models are quite broad. OpenAI’s GPT2 has a smaller version that starts with 117 million parameters whereas the GPT-3 has versions ranging up to 175 billion parameters. Google’s Switch Transformer has around 1.6 trillion parameters.

Generally, larger models tend to perform better in terms of understanding context, generating more coherent and contextually appropriate responses, and handling complex language tasks. However, there may be a point of diminishing returns where the performance gains may not justify the significantly higher computational costs and complexities of training and deploying these larger models.

The current popular models are OpenAI’s GPT-3 and GPT-4, Google’s BERT and its variants, Facebook’s RoBERTa are some among many.