Fine-tuning FLAN-T5 Large Language Model with Reinforcement Learning

1. Generative Ai

The recent buzz across all industries today (as of 2024) is Generative Ai! Generative Ai is a type of artificial intelligence that generates new data based on inputs, be it structured or unstructured. Generative Ai became popular when ChatGPT (Chat Generative Pre-trained Transformer) was introduced into the market. ChatGPT is an AI chatbot developed by OpenAi, a US based AI research firm, which is powered by GPT-3.5 Large Language Model. Large Language Models (LLMs) are large deep learning models that are pretrained on massive dataset using a transformer architecture.

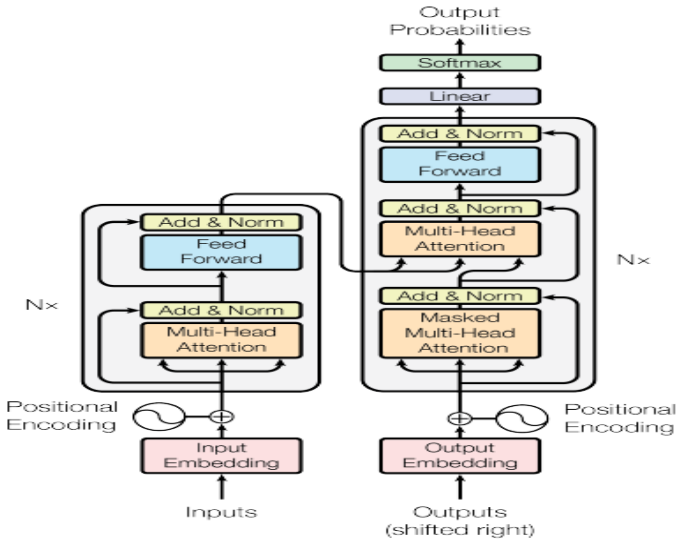
Some of the notable LLMs are OpenAI's GPT series of models (e.g., GPT-3.5 and GPT-4, used in ChatGPT and Microsoft Copilot), Google's PaLM and Gemini (the latter of which is currently used in the chatbot of the same name), xAI's Grok, Meta's LLaMA family of open-source models, Anthropic's Claude models, and Mistral AI's open-source models.

2. Transformer Architecture

Transformer architecture was introduced by Google researchers when they published their research paper *"Attention is All You Need".* The paper proposed a neural network architecture entirely based on attention-based mechanism to replace recurrent neural networks (RNNs) and convolutional neural networks (CNNs).

Transformer Architecture contains an encoder and a decoder. The encoder is a neural network layer that processes the input sequence and produces a continuous embedding of the input. The decoder then uses these embeddings to generate the output.

The encoder and decoder contain multiple layers, each layer consists of a multi-head self-attention mechanism and a feed-forward neural network. The multi-head self-attention mechanism allows the model to attend to different parts of the input sequence, while the feed-forward network applies a point-wise fully connected layer to each position separately and identically.



There are three types of transformer models. Encoder-only model is used for text summarization and is less common. Encoder-Decoder model is used for translation or text generation tasks. Popular Encoder-Decoder models are T-5 and BART. Decoder-only model, more commonly used, are useful for their text generational capabilities. GPT, BLOOM are examples of Decoder-only model.

3. Prompt and Prompt Engineering

A prompt is an input text / sentences that are entered into the model. Prompt engineering is a process of designing inputs that will result in high-quality output. The model may or may not produce user desirable output given the input prompt. Prompts have to be modified with examples for the model to learn to provide the output. Zero shot inference involves providing prompts to the LLM without providing any set of examples for the model to learn. ChatGPT is tuned in a way that it enables the users to perform zero shot inference. Most large models are capable of providing an optimal output using zero shot inference. An example of zero shot inference is asking the model to classify is Saturday is a weekday or a weekend and the model returns weekend. The smaller the large language model, the tougher it is tune it with zero shot inference. The smaller LLMs require more examples to tune the model. One shot and few shot inferences involve providing 1 or few examples in the prompt for pretraining. Example of few shot inference would be, classify if number 3 is an even or an odd number by providing a few examples like number 1 is an odd number, number 2 is an even number.

4. Fine Tuning a Large Language Model

LLMs may fail to carry out the input tasks even after utilizing few shot inferences. The models have to be fine tuned or the weights have to be adjusted for the model to generate an optimal output. Fine tuning is a supervised learning process as compared to inferencing which is self-supervised learning.

4.1 Instruction Tuning

Citations:

Wikipedia contributors. (2024, March 16). *ChatGPT*. Wikipedia. https://en.wikipedia.org/wiki/ChatGPT

Wikipedia contributors. (2024b, March 16). *Large language model*. Wikipedia. https://en.wikipedia.org/wiki/Large_language_model

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. *arXiv (Cornell University)*, *30*, 5998–6008. https://arxiv.org/pdf/1706.03762v5

CouRseRa | Online courses & credentials from top educators. Join for free | CourseRA. (n.d.). Coursera. https://www.coursera.org/learn/generative-ai-with-llms/supplement/Il7wV/transformers-attention-is-all-you-need