1016 Project Proposal

Author: Bo Long,Luca Wang, Qiuyi Wei, Richen Du

1. Introduction:

During the training process of state-of-the-art language models such as GPT-2, we always feed sequences of fixed length to the model. For example, the default maximum sequence length (number of tokens per sequence) of the GPT-2 model is 1024, so a common data preprocessing strategy is to reshape the length of each training sequence into 1024.
However, the strategy does not follow the way how human-beings acquire language cognitions in early stages of life. During childhood, we often start with relatively shorter passage lengths, then gradually navigate to longer corpses as we establish some fundamental lexical or grammatical structure in mind.

2. Literature Review:

We notice a paper introducing a “sequence length warm-up algorithm” that corresponds to the human way of learning language. During training, this algorithm linearly increases the sequence length used to train a language model from a min_seq_length to a max_seq_length over some duration at the beginning of training. According to the paper, this warm-up strategy reduces the training time of GPT-style models by ~1.5x while still achieving the same loss as baselines.

3. Research Objective & Approach:

Based on the findings of the paper, the goal of our project is to explore the implementation of the algorithm to emulate the language learning process of human children. To achieve this, we propose changing the algorithm's linear sequence length warmup process into a series of non-linear curves. This adjustment aims to more accurately reflect the complexities of the human language learning process at an early stage.
We will first pre-train a small GPT-2 model (124M) and try to reproduce the training time reduction based on this warm-up algorithm. Next, we will try with different datasets to test the generalizability of this algorithm, including general corpse (e.g. wikitext), or specific corpse (e.g. children’s book dataset). After that, we plan to replace the linear sequence length warmup process with non-linear curves that better mimics the learning process of human-beings. Finally, we will again test the performance of the model to figure out whether non-linear warm-up would bring better results.

4. References

Li, C., Zhang, M., & He, Y. (2022, October 16). The stability-efficiency dilemma: Investigating sequence length warmup for training GPT models. arXiv.org. https://arxiv.org/abs/2108.06084

community, T. H. D. (n.d.). Wikitext · datasets at hugging face. wikitext · Datasets at Hugging Face. https://huggingface.co/datasets/wikitext

CBT · datasets at hugging face. cbt · Datasets at Hugging Face. (n.d.). https://huggingface.co/datasets/cbt