Modern Natural Language Processing (MNLP) - Project Proposal

Camille Nicole Gisèle Challier | 311020 | camille.challier@epfl.ch
Céline Kalbermatten | 328519 | celine.kalbermatten@epfl.ch
Wesley Elliott Stephen Monteith-Finas | 324745 | wesley.monteith-finas@epfl.ch
Chatbots-R-Us

1 Introduction

The project aims to develop a chatbot capable of addressing inquiries from EPFL course materials. Pre-trained large language models such as SciBERT, RoBERTa or Phi 2 will be fine-tuned and preference aligned through Direct Preference Optimization (DPO), involving Supervised Fine-tuning (SFT) and Preference learning stages. Additionally, the model will undergo quantization to reduce parameter count and improve efficiency.

2 Model

2.1 Generator Model

Several pre-trained large language models will be trained using DPO, which consists of two main stages: Supervised Fine-tuning (SFT) on relevant data, and Preference Learning, which aligns the model on preference data. DPO uses a binary cross-entropy objective to optimize policy as a binary classification problem, updating the model's responses based on human preference data to enhance performance. Other loss functions such as RSO, IPO, cDPO, KTO, BCO and SPPO exist and will be tested. A DPO trainer is also available directly on the Huggingface API.

SciBERT (Beltagy et al., 2019) : SciBERT is a pretrained language model of 110 millions parameters based on BERT (Devlin et al., 2019). It leverages unsupervised pretraining on a large multi-domain corpus of scientific publications to improve performance on downstream scientific NLP tasks.

RoBERTa (Liu et al., 2019) : Roberta, a 125M parameters model, represents an optimized variant of BERT with dynamic masking and without next sentence prediction.

Phi 2 (Hughes, 2023) This 2.7 billion parameters transformer-based model from Microsoft was trained on synthetic datasets augmented with educational and high-quality web content. It showcases state-of-the-art performance among base language models with less than 13 billion parameters.

2.2 Quantization Specialization

For the quantization specialization, the goal is to significantly reduce the number of parameters of the best performing model, while still preserving a reasonable performance.

Post-Training Quantization (PTQ) will probably be favoured over Quantization-Aware Training (QAT). PTQ quantizes a pre-trained model by using moderate resources such as a calibration dataset and a few hours of computation whereas QAT performs fine-tuning before quantization.

For the BERT family, a recent paper proposed a state of the art method called Optimal Brain Quantization and Pruning (OBQ) (Frantar et al., 2023b).

For large language models with a decoder or encoder only architecture such as the Phi 2 architecture, a more elaborate method can be applied called GPTQ. This is a one-shot weight quantization method based on approximate second-order information (Frantar et al., 2023a).

The latter method is available on Huggingface using the AutoGPT API compatibility (noa, 2024).

3 Data

3.1 Generator Model

3.1.1 SFT Data

Before undergoing the preference learning stage in the DPO strategy, the selected model will undergo SFT using a dataset that does not necessarily have preference pairs. This provides a direct mapping between the inputs and the desired outputs.

Manual Collection: For the current purposes, a relevant dataset to collect would be the questions and answers from the publicly available exams or exercises of the various courses of interest at EPFL on Moodle. Using a python PDF parsing library (*PyPDF2*) could automate the processing of the course exams to increase the data pool.

Additional SFT datasets such as the Massive Multitask Learning Understanding Dataset may be used.

3.1.2 Preference Data

For the preference learning stage of DPO, a large amount of data has been created by the MNLP class. In the following part, the personally chosen prompting strategy is detailed.

Few-shot Learning: For the first prompt, a few-shot learning approach was employed (Dong et al., 2023; Qiao et al., 2023), influenced by Liu et al.'s methodology (Liu et al., 2022), using RoBERTa embeddings to measure similarity between example questions and the real question. The three most similar examples based on cosine similarity were selected and ordered. 22 examples from EPFL courses on Moodle were collected.

Zero-shot Learning: For the second prompt, zero-shot learning was employed. Through experimentation of various instructions (Zhou et al., 2023), it was discovered that employing the simple directive *"Let's think step by step"* (Kojima et al., 2023) produced the most favorable answers.

In addition, other public preference pairs datasets can also be used, such as OpenAI WebGPT Comparisons, the comprehensive question answering "Explain Like I'm Five" Reddit forum dataset (ELI5), the Human ChatGPT Comparison Corpus (HC3) and Intel's Orca DPO Pairs.

3.2 Quantization Specialization

GPTQ compresses each weight by identifying a condensed version that minimizes the mean squared error. The GPTQ algorithm necessitates a calibration dataset in order to calibrate the quantized weights through inference on the quantized model itself. Similarly, OBQ aims to minimize performance loss by strategically removing and adjusting the weights. A random subsample of the SFT dataset will be used as calibration data for both quantization techniques, as outlined in the original papers (Frantar et al., 2023a,b).

4 Evaluation

4.1 Generator Model

To evaluate the generator model, two automatic evaluation metrics have been selected.

BERTScore: This first metric uses pre-trained contextual embeddings from BERT to compare candidate and reference sentences based on cosine similarity (Zhang et al., 2020).

BLEURT: This second metric uses both supervised and unsupervised data to predict human judgment of text quality. It outperforms traditional metrics like BLEU or ROUGE, which often fail in nuanced evaluations requiring an understanding beyond lexical similarities (Sellam et al., 2020).

Baselines: A non-fine-tuned BERT model will serve as the lower baseline, while either a ChatGPT wrapper or a Llama 3 model will serve as the upper baseline. The goal for the developed generator model is to outperform BERT and approach the performance of the upper baseline models as closely as possible.

Qualitative evaluation: Human evaluators will compare the designed model output texts with the ones of the baseline models.

Further performance evaluation: In a later phase of the project, when the model is designed to output only a single letter indicating the correct answer to the question, the performance evaluation will be done using accuracy, F1 score, precision and recall, being the most frequently used metrics for such tasks (Blagec et al., 2021).

4.2 Quantization Specialization

Since the quantization will be done at the stage where the model is designed to output only a single letter, the performance comparison between the pre-quantization and the post-quantization model can be done using metrics such accuracy, F1 score, precision, and recall (Blagec et al., 2021).

5 Ethics

Overall, the current project lacks severe ethical implications. However, this does not imply that this aspect should be entirely disregarded.

Privacy: Given that the training data mostly comes from the academic domain, ethical biases and privacy concerns are unlikely to arise. However, it should be kept in mind that not all exam content is allowed to be published.

Toxicity: Considering that the data contains questions in the domain of Cyber Security, some generations might be used for undesired purposes.

Data sources: The data should originate from a reputable, unbiased, qualitative source.

Misinformation: Given that the questions are complex, the model may struggle to provide accurate answers and therefore create misinformation.

References

2024. AutoGPTQ/AutoGPTQ. Original-date: 2023-04-13T02:18:11Z.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A Pretrained Language Model for Scientific Text. ArXiv:1903.10676 [cs].

Kathrin Blagec, Georg Dorffner, Milad Moradi, and Matthias Samwald. 2021. A critical analysis of metrics used for measuring progress in artificial intelligence.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A Survey on In-context Learning. ArXiv:2301.00234 [cs].

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. 2023a. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers. ArXiv:2210.17323 [cs].

Elias Frantar, Sidak Pal Singh, and Dan Alistarh. 2023b. Optimal Brain Compression: A Framework for Accurate Post-Training Quantization and Pruning. ArXiv:2208.11580 [cs].

Alyssa Hughes. 2023. Phi-2: The surprising power of small language models.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large Language Models are Zero-Shot Reasoners. ArXiv:2205.11916 [cs].

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. What Makes Good In-Context Examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. ArXiv:1907.11692 [cs].

Shuofei Qiao, Yixin Ou, Ningyu Zhang, Xiang Chen, Yunzhi Yao, Shumin Deng, Chuanqi Tan, Fei Huang, and Huajun Chen. 2023. Reasoning with Language Model Prompting: A Survey. ArXiv:2212.09597 [cs].

Thibault Sellam, Dipanjan Das, and Ankur P. Parikh. 2020. Bleurt: Learning robust metrics for text generation.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models Are Human-Level Prompt Engineers. ArXiv:2211.01910 [cs].