

Project for Team 4

Title

Reusing Small Pre-Trained Models for Efficient Large Model Training with Limited GPU Resources

Introduction

Transformer models are widely being applied to many domains in science and society. Unfortunately, the prohibitive training costs of large Transformer models hinder their broad application. In practice, training large Transformer models with billions or even trillions of parameters can take days, weeks, or even months, depending on the available computational resources.

Objective

This project aims to cost-efficiently train the large Transformer models with limited GPU resources by appropriately combining multiple small models. This is motivated by the fact that it usually consumes a shorter time and fewer resources to train small Transformer models, even though these small models are less powerful and only suitable for a simpler language.

Methodology

The methodology for this project involves several key steps:

- Data collection: collect multiple English sentence datasets.
- Small Transformer model training: Use the first English sentence dataset to build/train one or more copies m_1, m_2, \dots, m_n of a small Transformer model with the dimension d_0 ($< d_1$).
- Large Transformer model training: Use the English sentence datasets to train a Transformer model M_0 with the dimension d_1 .
- Model combination: Design a cost-efficient method of combining well-trained small models to generate the initial weights of a large Transformer model M'_1 with the dimension d_1 and then refine M'_1 through training with the English sentence datasets to generate the well-trained large Transformer model M_1 .
- Performance evaluation: Compare two training methods for large Transformer model generation in prediction accuracy and training cost, and conduct the profile-based analysis on the cost of training large and small Transfer models in terms of TFLOPs.

Expected Outcomes

Students will understand how to design, optimize, and implement LLMs targeting a specific platform. They will learn how to balance prediction accuracy and cost efficiency, which are crucial in real-world applications. Also, they will have a deep insight into the resource consumption of different models.

Reference

- [1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin. Attention is All You Need. Conference on Neural Information Processing Systems (NIPS), Long Beach, CA, USA, Dec 2017.
- [2] Yu Pan, Ye Yuan, Yichun Yin, Jiaxin Shi, Zenglin Xu, Ming Zhang, Lifeng Shang, Xin Jiang, and Qun Liu. Preparing Lessons for Progressive Training on Language Models. Proc. of the AAAI Conference on Artificial Intelligence, vol. 38, no. 17, pp.18860-18868, Feb 2024.
- [3] Yu Pan, Ye Yuan, Yichun Yin, Zenglin Xu, Lifeng Shang, Xin Jiang, Qun Liu. Reusing Pretrained Models by Multi-Linear Operators for Efficient Training. Proceedings of the 37th Conference on Neural Information Processing Systems (NeurIPS), 2023
- [4] Cheng Chen, Yichun Yin, Lifeng Shang, Xin Jiang, Yujia Qin, Fengyu Wang, Zhi Wang, Xiao Chen, Zhiyuan Liu, and Qun Liu. bert2BERT: Towards Reusable Pretrained Language Models. Proc. of the 60th Annual Meeting of the Association for Computational Linguistics, vol. 1, pp. 2134-2148, Dublin, Ireland, May 2022.
- [5] Peihao Wang, Rameswar Panda, Lucas Torroba Hennigen, Philip Greengard, Leonid Karlinsky, Rogerio Feris, David D. Cox, Zhangyang Wang, and Yoon Kim. Learning to Grow Pretrained Models for Efficient Transformer Training. International Conference on Learning Representations (ICLR), 2023.
- [6] Sheng Shen, Pete Walsh, Kurt Keutzer, Jesse Dodge, Matthew Peters, and Iz Beltagy. Staged Training for Transformer Language Models. Proc. of the 39th International Conference on Machine Learning (ICML), 2022.
- [7] Yite Wang, Jiahao Su, Hanlin Lu, Cong Xie, Tianyi Liu, Jianbo Yuan, Haibin Lin, Ruoyu Sun, and Hongxia Yang. LEMON: Lossless model expansion (poster). International Conference on Learning Representations (ICLR), 2024.
- [8] Wenyu Du, Tongxu Luo, Zihan Qiu, Zeyu Huang, Yikang Shen, Reynold Cheng, Yike Guo, and JieFu. Stacking Your Transformers: A Closer Look at Model Growth for Efficient LLM Pre-Training. <https://arxiv.org/pdf/2405.15319>
- [9] Peihao Wang, Rameswar Panda, and Zhangyang Wang. Data Efficient Neural Scaling Law via Model Reusing. Proceedings of the 40th International Conference on Machine Learning (ICML), 2023.

Project presentation: 30-minute team presentation on Friday, June 7th, 2024

Paper draft: an 8-page paper including references in the 2-column format using the latest ACM proceedings templates at <https://www.acm.org/publications/proceedings-template> . For Latex users, version 2.08 (last update June 4, 2024) is the latest template, and please use the "sigconf" option.

Paper draft completion deadline: August 2nd, 2024

Paper submission target: SC24 Workshop “AI4S: 5th Workshop on Artificial Intelligence and Machine Learning for Scientific Applications” (<https://ai4s.github.io>)