Better Call LoRA

Robert Trifan Stefan Popa

University of Bucharest

Abstract

Low-rank adaptation (LoRA) has become a lightweight alternative to full fine-tuning for large language models. In this work, we benchmark vanilla LoRA and four recent LoRA variants: swapped-init LoRA (A = 0, B $\sim \mathcal{U}(-0.01, 0.01)$), LoRA-XS ($A \cdot R \cdot B$ factorisation), LoRA+ ($\eta_A \neq \eta_B$ learning rate scaling) and PiSSA-initialised LoRA (SVD(W) warm-start) on the TinyLlama-1.1 B backbone and the GLUE SST-2 sentiment-classification task. All experiments were performed on a single RTX 2070 (8 GB), enforcing strict memory budgets. We report classification accuracy, macro-F1, wall-clock training time and peak GPU memory to highlight the trade-offs each variant offers under resource constrained conditions.

1 Introduction

Explain the importance of LoRA for LLMs and the need for a survey.

2 Setup

Model Describe TinyLlama-1.1B [7]
Dataset Desribe GLUE [6]
Training Describe the prompt used for training.
Evaluation Describe the evaluation step.
Metrics Describe the metrics used: accuracy, F1, train time, GPU memory.

3 Low Rank Adaptation

LoRA Describe LoRA [4].
Impact of initialization dynamics on LoRA Describe different LoRA [2]
LoRA-XS Describe LoRA-XS [1].
LoRA+ Describe LoRA+ [3]
PiSSA Describe PiSSA [5]

4 Experiments

For each method, describe the hyperparameters explored with a table.

Gather the best results into a final table, comparing the methods.

5 Conclusion

Explain that, because of limited compute resources, we couldn't see meaningful results.

References

[1] Klaudia Bałazy, Mohammadreza Banaei, Karl Aberer, and Jacek Tabor. Lora-xs: Low-rank adaptation with extremely small number of parameters, 2024.

[2] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. The impact of initialization on lora finetuning dynamics, 2024.

[3] Soufiane Hayou, Nikhil Ghosh, and Bin Yu. Lora+: Efficient low rank adaptation of large models, 2024.

[4] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[5] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models, 2025.

[6] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding, 2019.

[7] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.