AUTHORS

Yuntao Zheng
Shaoxuan Zheng
Lingfei He

Fine-tune Llama3 with $(IA)^3++$

AFFILIATIONS

Carnegie Mellon University

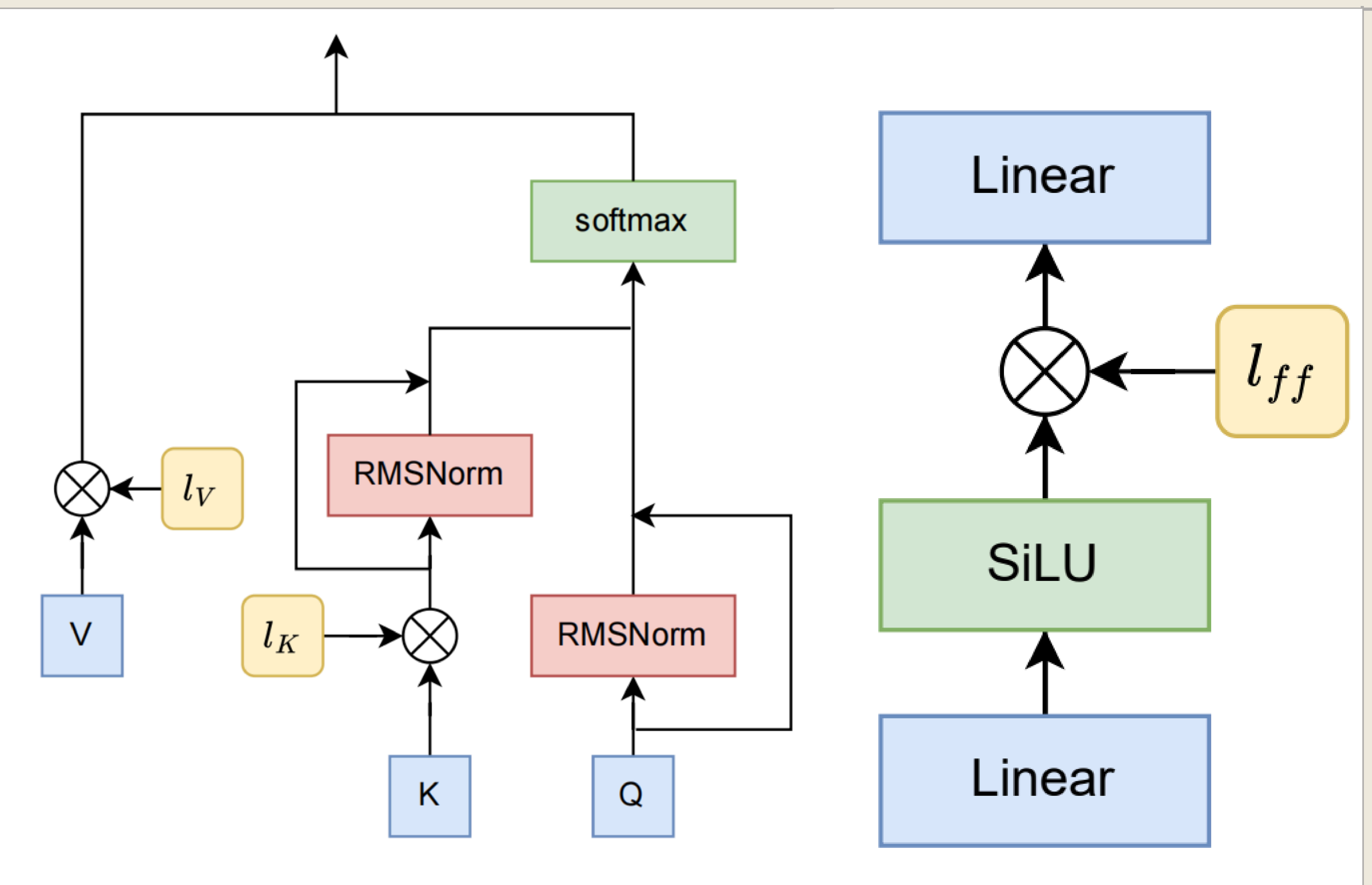

Figure1: Diagram of $(IA)^3++$. Before applying softmax, we use RMSNorm and a skip-connection to stabilize fine-tuning. The scaling parameter of RMSNorm is initialized to zero to ensure consistent output at the beginning of fine-tuning.

01. Introduction

Advancements in fine-tuning techniques have significantly expanded the capabilities of Large Language Models (LLMs). Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA have been developed, achieving remarkable performance by fine-tuning only a small fraction of the model's parameters. However, even state-of-the-art methods like LoRA still involve updating a considerable number of parameters. In this work, we aim to further reduce the number of trainable parameters required for effective fine-tuning without compromising model performance. Inspired by $(IA)^3$, QK-Norm from OLMoE and LNLoRA, we propose our method, $(IA)^3++$, which outperforms $(IA)^3$ and attain a comparable performance to LoRA, while reducing trainable parameters to around 0.01% of total parameters.

02. Contribution

- Propose $(IA)^3++$, inspired by $(IA)^3$ and RMS-Norm
- Conduct ablation study on adding l_o in the output projection layer after Multi-head Causal Attention
- Perform experiments on the MMLU and ARC-Challenge datasets using LoRA, $(IA)^3$ and $(IA)^3++$ fine-tuning methods.
- Demonstrate that our proposed $(IA)^3++$ outperforms $(IA)^3$ on both datasets and achieves comparable results to LoRA, while requiring approximately only 0.01% trainable parameters, which is significantly more efficient than LoRA's ~1% trainable parameters.

03. Methodology

- $QK-Norm$

OLMoE introduces QK-Norm, which can prevent very large logits in the following attention operation that may lead to overflow and make the training unstable, by adding a layer normalization after the query and key projections. In our $(IA)^3++$, we use RMSNorm instead of standard layer normalization.

For $\mathbf{x} \in R^d$, we define the non-parametric RMSNorm of x as $\mathbf{y} = \frac{\mathbf{x}}{\sqrt{\sum_{i=1}^{D} x_i^2}} \odot \alpha$, where $y, \alpha \in \mathbb{R}^d$

- $(IA)^3++$

As shown in figure 1&3, our method combines the idea of $(IA)^3$ and QK-Norm, with an extra scaling vector l_o in output projection after multi-head causal attention. We also add a skip connection and initialize the scaling parameter of QK-Norm α to be a zero vector in order to keep the same output as original model at the beginning of the fine-tuning. Mathematically

$$X_{\text{attn}} = \text{softmax}\left(\frac{(\text{Norm}(Q)+Q)\,(\text{Norm}(l_k \odot K^T)+l_k \odot K^T)}{\sqrt{d_k}}\right)(l_v \odot V)$$

$$\text{Out} = l_o \odot (W_o X_{\text{attn}})$$

$$X_{\text{ff}} = \sigma\,(l_{\text{ff}} \odot \gamma(W_1 x))W_2$$

The amount of trainable parameters introduced are $L \times (d_v + d_k + d_o + d_{\text{ff}} + 2 \times d_{rms})$ for L-layer decoder-only transformer.

04. Results/Findings

We fine-tuned the Llama3.2-1B model on MMLU and ARC-Challenge datasets to evaluate the accuracy and parameter efficiency of our QK-Norm-enhanced $(IA)^3++$ method. The goal was to assess whether our approach could match or surpass LoRA and the original $(IA)^3$ in accuracy while significantly reducing trainable parameters.

Models	Trainable Params	All Params	Trainable (%)
$(IA)^3$	147,456	1,235,961,856	0.0119
LoRA	11,272,192	1,247,086,592	0.9039
Ours	172,048	1,236,027,440	0.0139

Key Findings

- Our model outperforms the original $(IA)^3$ on both the MMLU and ARC-Challenge benchmarks.
- It achieves comparable average accuracy to LoRA on MMLU while surpassing it on the ARC-Challenge dataset.
- The trainable parameter count for our model is comparable to $(IA)^3$ and significantly lower than that of LoRA, highlighting its efficiency and effectiveness.

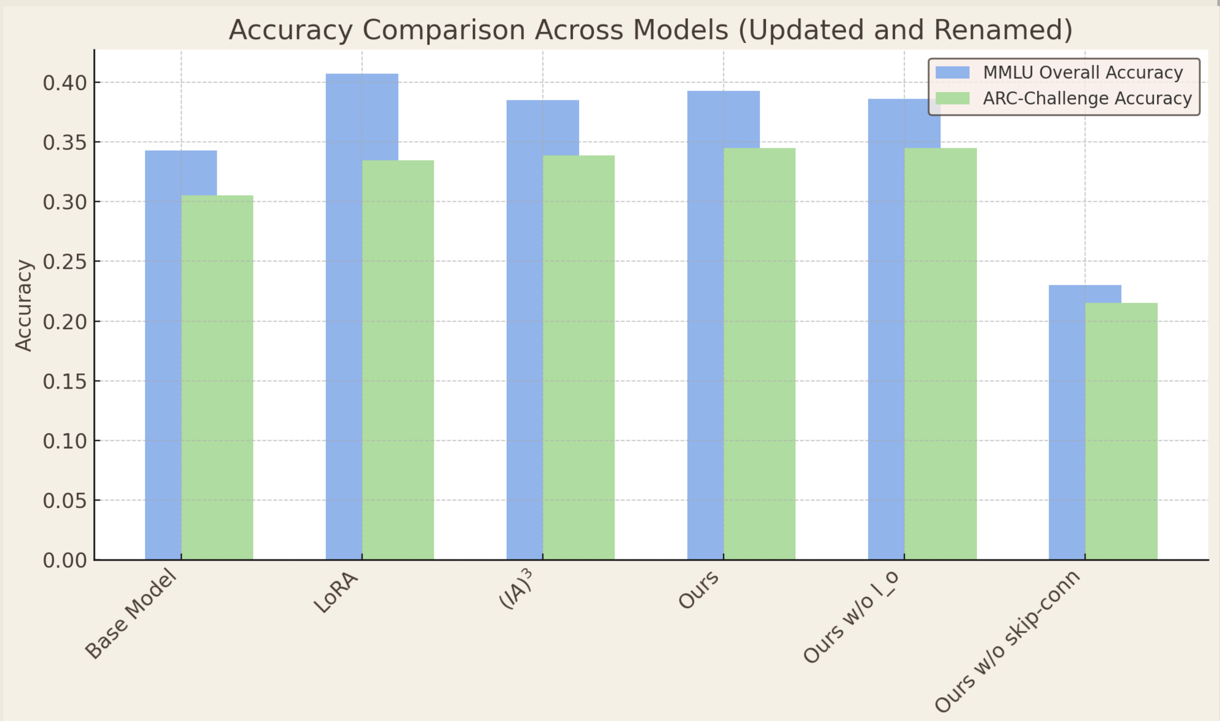

Figure2: Comparison of model accuracy on MMLU (blue) and ARC-Challenge (green) datasets, including ablation studies.

05. Effect of l_o in Out Projection

Different from $(IA)^3$, which only uses three scaling vectors l_k, l_v, l_{ff}, our $(IA)^3++$ introduces an additional l_o to enhance the expressive capacity of the fine-tuning module. The results of ablation study can be found from Figure 2 that $(IA)^3++$ with extra scaling vector in output projection outperforms the one without l_o on MMLU.

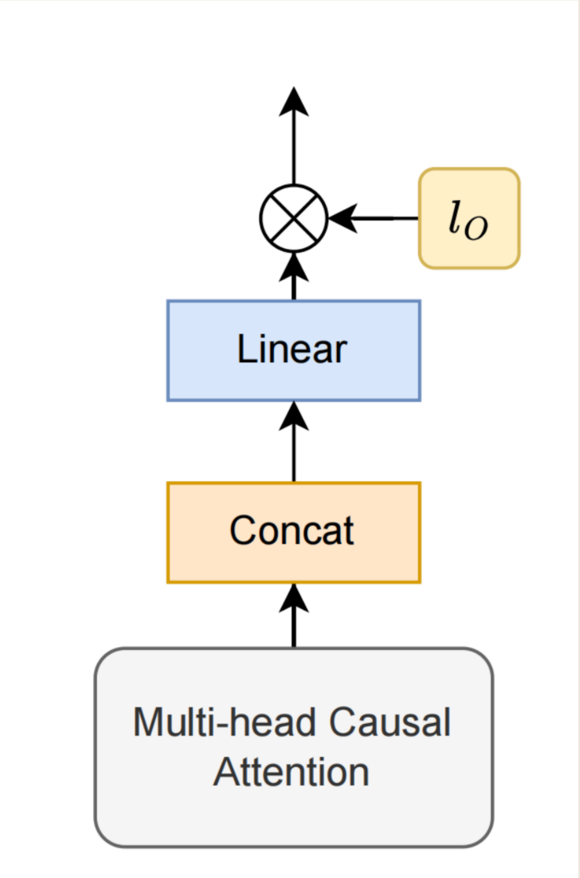

Figure3: An extra l_o is added to the out projection.

06. Conclusion

In this work, we propose a novel PEFT method, $(IA)^3++$, inspired by $(IA)^3$ and QK-Norm techniques, while incorporating a new trainable vector for the output projection layer. Our method addresses the limitations of existing PEFT approaches, such as numerical instability and high trainable parameter requirements, by stabilizing attention mechanisms using RMSNorm and enhancing fine-tuning with skip connections.

07. Reference

1. *Tom B Brown. Language models are few-shot learners. arXiv preprint arXiv:2005.14165, 2020.*
2. *Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Charles Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. arXiv preprint arXiv:1803.05457, 2018.*
3. *Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.*
4. *Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. arXiv preprint arXiv:2009.03300, 2020.*
5. *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. arXiv preprint arXiv:2106.09685, 2021.*
6. *Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. arXiv preprint arXiv:2104.08691, 2021.*
7. *Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. arXiv preprint arXiv:2101.00190, 2021.*
8. *Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin A Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. Advances in Neural Information Processing Systems, 35:1950–1965, 2022.*
9. *Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. AI Open, 5:208–215, 2024.*
10. *Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, et al. Olmoe: Open mixture-of-experts language models. arXiv preprint arXiv:2409.02060, 2024.*
11. *Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. arXiv preprint arXiv:2303.10512, 2023.*