SUPERVISED FINE-TUNING WITH LOW RANK ATTENTION DENOISING

Nikita Okhotnikov
okhotnikov.nv@phystech.edu

Data and parameters efficiency has always been a concern for small LLMs, but recently LLM advance has started to face the limits of available data and computational resources even on the largest scale. One of potential issue to adress on the way of improving the performance of LLMs despite these constraints is the noisy attention weights distribution. Recent works has showed the potential solutions, however they require training from scratch and large number of additional parameters. In this work we propose low rank attention denoising adapter, that can be applied to pretrained model on the SFT stage. Our approach shows significant improvent over the unmodified model in terms of cross-entropy on training data.

1 Introduction

....

2 Related Work

2.1 Attention mechanism

Most of the modern large language models are based primarily on the self-attention mechanism and transformer architecture [2]. Each attention layer consists of multiple ¡¡heads¿¿, which are applied to different parts of the input. Each head acts as follows:

$$Q = XW_Q, K = XW_K, V = XW_V$$

$$\text{Attention} = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where $X \in \mathbb{R}^{N \times d_{model}}$ is the input embedding from previous layers, $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d}$ – learnable projection matrices

2.2 Differential attention

Based on the idea of adaptive noise cancellation, the authors of [3] proposed a novel attention mechanism modification that use additional learnable parameters to cancel out the noise in the attention weights. Specifically, given the input $X \in \mathbb{R}^{N \times d_{model}}$ and projection matrices $W_Q, W_K \in \mathbb{R}^{d_{model} \times 2d}, W_V \in \mathbb{R}^{d_{model} \times d}$, the modified attention mechanism is defined as follows:

$$[Q_1, Q_2] = XW_Q, [K_1, K_2] = XW_K, V = XW_V$$

$$\text{DIFFAttention} = \left(\text{softmax}\left(\frac{Q_1K_1^T}{\sqrt{d}}\right) - \lambda \cdot \text{softmax}\left(\frac{Q_2K_2^T}{\sqrt{d}}\right)\right)V,\ \lambda \in (0,1)$$

... // something about broken normalisation and weird labmda initialization (unstability)

2.3 DINT attention

DINT attention [1] extends differential attention idea by ¡¡integral¿¿ mechanism, tackling the normalisation and lambda initialization sensitivity problem

TODO

3 Low rank attention denoising

...

4 Experiments

...

5 Conclusion

...

References

[1] Yueyang Cang, Yuhang Liu, Xiaoteng Zhang, Erlu Zhao, and Li Shi. Dint transformer, 2025.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[3] Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. Differential transformer, 2025.