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

....

2.2 Differential attention

....

2.3 DINT attention

....

3 Low rank attention denoising

...

4 Experiments

...

5 Conclusion

...