Low rank attention denoising

Nikita Okhotnikov

MIPT

2025

Introduction

Problem

- Noisy attention weights distribution in modern LLMs hurts interpretability and potentially limits the performance
- Existing solution utilizes large number of additional parameters and requires training from scratch

Objective

Implement parameter and training time efficient attention score denoiser

Proposal

Low rank implementation of existing adaptive denoiser [1] that does not require training from scratch

[1] Yueyang Cang et al. DINT Transformer, 2025

Attention mechanism

DINT Transformer

Low rank adapter

SFT on math benchmark