Contribution ID: **70** Type: **not specified**

SimplexLoRA: Expand important adapters

Sunday, 18 May 2025 16:18 (12 minutes)

Language models have become central to many AI applications. Effective fine-tuning is essential to adapt these models to specific tasks. Traditional methods like Low-Rank Adaptation (LoRA) add fixed-rank adapters to all layers, often resulting in memory inefficiency due to non-optimal layer selection. We propose SimplexLoRA, a novel fine-tuning framework that adaptively scales adapter ranks using simplex-constrained weighting, optimizing both memory usage and performance.

Primary authors: DAVYDENKO, Grigorii (Moscow Institute of Physics & Technology (MIPT)); SHALYGIN, Igor (MIPT)

Co-author: VEPRIKOV, Andrey (Moscow Institute of Physics & Technology (MIPT))

Presenter: DAVYDENKO, Grigorii (Moscow Institute of Physics & Technology (MIPT))

Session Classification: 18-Машинное обучение и нейросети

Track Classification: Машинное обучение и нейросети