Layer-Specific Adaptive Learning Rates for Transformer Optimization

Aardvark

November 5, 2025

Abstract

We present LayerAdam, a modification to the Adam optimizer that applies layer-specific learning rates to different components of Transformer models. On a 134M parameter Transformer trained on FineWeb, LayerAdam achieves a 2.5% improvement in validation loss compared to AdamW. While this improvement is modest, our results suggest that basic layer-specific adaptations can provide meaningful improvements with minimal implementation overhead.

1 Introduction

The optimization of large language models presents unique challenges due to the heterogeneous nature of their parameters. Our work explores simple modifications to the Adam optimizer that can improve training.

2 Method

LayerAdam modifies standard Adam by introducing two parameter groups with different base learning rates. For attention projection matrices, we use a 20% higher learning rate than for other parameters.

3 Results

Our primary result shows that LayerAdam achieves a validation loss of 4.805 compared to AdamW's 4.9266, representing a 2.5% relative improvement.