Efficient Three-Layer Feedforward Networks with Optimized Gating and Normalization

Aardvark

November 1, 2025

Abstract

We present a systematic study of feedforward network architectures in transformers, focusing on optimization of gating mechanisms and normalization placement. Through careful ablation studies, we identify an efficient three-layer design that achieves a validation loss of 4.857 on the FineWeb benchmark (vs. 4.9266 SwiGLU baseline) while maintaining parameter efficiency. Our architecture employs dimension-reduced projections with strategic normalization placement, demonstrating consistent improvements across model scales.

1 Introduction

Transformer architectures have become fundamental in modern machine learning, with the feedforward network (FFN) component playing a crucial role in their success. While numerous modifications have been proposed, the basic FFN design has remained relatively unchanged. We revisit this component through rigorous empirical analysis, identifying optimization opportunities in dimension allocation and normalization placement.

2 Methodology

Our architecture employs a modified three-layer structure with dimension-reduced projections and strategic normalization placement. The key innovation is applying LayerNorm after the initial projection rather than before, which we found improves gradient flow while maintaining computational efficiency.

3 Results

Our experiments on the FineWeb benchmark demonstrate consistent improvements over the SwiGLU baseline, with a validation loss of 4.857 compared to 4.9266. The architecture maintains comparable runtime while providing better modeling performance.

4 Conclusion

We presented an optimized FFN architecture demonstrating that careful attention to fundamental design choices can yield meaningful improvements. The results suggest our approach provides a good balance between simplicity and performance.