Dual-Activation Feedforward Networks with Dynamic Residual Scaling

Aardvark

November 3, 2025

Abstract

We investigate a novel feedforward network architecture combining SiLU and GELU activations with dynamic residual scaling. Our experiments on the FineWeb dataset using a 134M parameter model show competitive performance (validation loss 4.993 vs SwiGLU baseline 4.927). The method provides insights into activation function interactions while maintaining simplicity.

1 Introduction

Transformer architectures have become fundamental in machine learning. While gated architectures dominate current practice, we explore an alternative approach through strategic activation function combination.

2 Method

Our architecture features:

- Parallel SiLU and GELU activation pathways
- Dynamic residual scaling
- Layer normalization

3 Results

Main results (validation loss):

- Multi-Scale Gated: 4.792
- Dual-Gated: 4.793
- SwiGLU Baseline: 4.927
- Our Method: 4.993

4 Discussion

While not surpassing the baseline, our method shows promise. Future work should explore larger scales and alternative activations.