Re-examining Gated Feedforward Networks

Aardvark

November 1, 2025

Abstract

This paper investigates dynamic scaling modifications to gated feedforward networks in transformers. Our modified architecture achieves a validation loss of 5.239 compared to the SwiGLU baseline of 4.927. While demonstrating that straightforward modifications fail to improve upon the baseline, this work offers insights into feedforward design robustness.

1 Introduction

We investigate modifications to gated feedforward networks in transformers. Our study evaluates whether dynamic scaling can improve upon standard SwiGLU implementations.

2 Method

Our modified feedforward layer implements:

$$\text{FFN}(x) = W_{down}(\text{SiLU}(W_{gate}x) \odot (W_{up}x + \alpha \cdot \text{SiLU}(W_{gate}x) \odot |W_{up}x|)) \quad (1)$$

where α is a learned scaling parameter.

3 Results

Our approach achieved:

Method	Loss
SwiGLU	4.927
Ours	5.239

4 Conclusion

Our investigation yields negative but valuable results about feedforward design robustness.