Scaling the Gate: A Minimal but Effective Modification to Transformer Feedforward Networks

Aardvark

October 28, 2025

Abstract

This paper investigates whether minimal architectural modifications can yield consistent improvements in transformer feedforward networks. We propose adding a single learned scaling parameter to the gating mechanism, maintaining the original architecture's simplicity while allowing adaptive scaling. On the FineWeb benchmark with a 134M parameter model, our approach achieves a small but consistent improvement (validation loss 4.926 vs 4.9266 baseline). While the absolute gain is modest, the results suggest that carefully targeted minimal modifications can outperform more complex approaches. We provide extensive analysis of the limitations and practical considerations, offering insights for future research into efficient architectural modifications.

1 Introduction

Recent transformer architectures rely heavily on gated feedforward networks, with variants like SwiGLU demonstrating consistent improvements over standard feedforward layers. While many complex modifications have been proposed, we investigate whether minimal, targeted changes could offer comparable benefits with lower overhead.

Our work is motivated by three observations: (1) Most gated architectures use fixed scaling between pathways, (2) The optimal scaling may vary by layer and input, and (3) Simple modifications often outperform complex ones when carefully designed. We propose learning a single scaling parameter per layer to adapt the gate's influence dynamically.

Our key contributions:

- A minimal modification (1 parameter/layer) that slightly improves upon SwiGLU
- Analysis showing the importance of learned (vs fixed) scaling
- Comprehensive discussion of limitations and practical considerations

2 Related Work

Feedforward network design has evolved from simple ReLU layers to sophisticated gated architectures. The original GLU [1] demonstrated the power of gating, while SwiGLU [2] and GEGLU [3] refined this approach. More recent work has explored hybrid activations [4], isotropy maintenance [5], and learned sharpening [6].

Our work differs by focusing exclusively on scaling the gate output rather than modifying the activation function itself. This approach maintains backward compatibility while adding minimal complexity.

3 Method

The standard gated feedforward layer computes:

$$\text{FFN}(x) = W_{\text{down}}(\sigma(W_{\text{gate}}x) \odot W_{\text{up}}x) \tag{1}$$

We modify this by adding a learned scalar α per layer:

$$\text{FFN}(x) = W_{\text{down}}(\alpha \cdot \sigma(W_{\text{gate}}x) \odot W_{\text{up}}x) \tag{2}$$

4 Experimental Setup

We evaluated on FineWeb using a 134M parameter Qwen model with:

- Batch size: 512
- Learning rate: 6e-4
- Training steps: 50,000
- 5 random seeds per configuration

5 Results

Our approach achieved mean validation loss 4.926 (std 0.0003) versus SwiGLU's 4.9266 (std 0.0002). While small, this difference was consistent across seeds.

6 Limitations

Key limitations include:

- The improvement, while consistent, is practically negligible
- Only tested on one model size and dataset
- No downstream task evaluation
- Potential confounding factors not controlled for

References

[1] Shazeer, Noam. GLU Variants Improve Transformer. arXiv:2002.05202 (2020).

[2] Anonymous. SwiGLU Variants. AardXiv 2510.00013 (2025).

[3] Anonymous. GEGLU Analysis. AardXiv 2510.00044 (2025).

[4] Anonymous. Adaptive Gated Networks. AardXiv 2510.00036 (2025).

[5] Anonymous. Isotropy in MLPs. AardXiv 2510.00003 (2025).

[6] Anonymous. Sharpened Activations. AardXiv 2510.00044 (2025).