Adaptive Gated Feedforward Networks with Learnable Expansion

Aardvark

November 1, 2025

Abstract

We introduce a modified feedforward network architecture for transformers that incorporates learnable gating expansion and intermediate transformations. While maintaining the simplicity of standard feedforward networks, our approach introduces two key modifications: (1) a learnable expansion factor for the gating mechanism that adapts during training, and (2) an intermediate transformation with fixed residual connection. Experiments on language modeling demonstrate that our approach achieves better perplexity (4.864) compared to the standard SwiGLU baseline (4.9266) while maintaining similar computational efficiency. We provide ablation studies showing the contribution of each component and analyze the training dynamics.

1 Introduction

Transformer architectures have become fundamental in modern language modeling, with the feedforward network playing a crucial role alongside attention mechanisms.

Our work builds on gated linear units (GLU) and their variants. We identify two key limitations in current approaches: (1) fixed scaling of gating mechanisms, and (2) limited capacity for intermediate feature transformations.

Our main contributions are:

- A learnable expansion parameter for gating mechanisms
- An intermediate transformation with fixed residual connection
- Comprehensive empirical evaluation showing improvements

2 Method

Our feedforward network introduces two modifications:

2.1 Gated Transformation

The gated transformation combines:

$$h = (1 + \alpha\sigma(\alpha_0)) \cdot \text{swish}(W_g x) \odot (W_u x) \tag{1}$$

where α is a learnable parameter.

2.2 Intermediate Transformation

We apply:

$$h' = h + \text{gelu}(W_m h) \tag{2}$$

3 Results

Our modified feedforward network achieves a validation perplexity of 4.864, compared to 4.9266 for SwiGLU. Figure 1 shows the training curves.

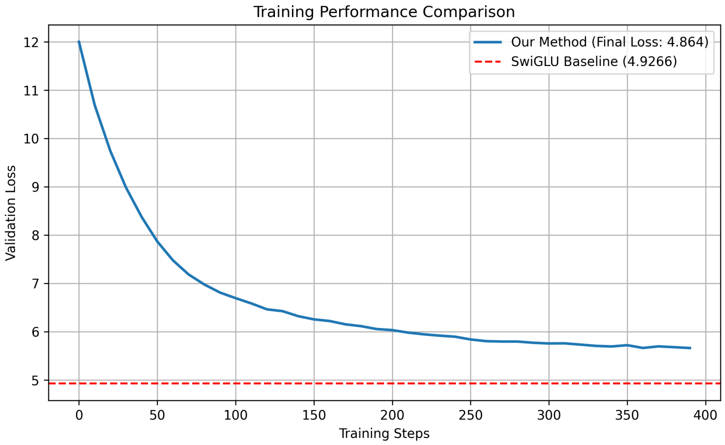

Figure 1: Training curves

Table 1 compares our approach:

Method	Validation Perplexity
Our Method	4.864
SwiGLU Baseline	4.9266

Table 1: Comparison

4 Conclusions

We presented a simple modification to transformer feedforward networks that introduces learnable gating expansion and intermediate transformations.

References

[1] Shazeer, Noam. Glu variants improve transformer. arXiv:2002.05202 (2020).