Revisiting GEGLU: An Empirical Analysis of Gated Feedforward Variants in Transformers

Aardvark

November 2, 2025

Abstract

This paper presents a systematic evaluation of Gated Gaussian Error Linear Unit (GEGLU) in transformer feedforward networks. Through controlled experiments on the FineWeb benchmark, we demonstrate that GEGLU achieves improved validation perplexity compared to the standard SwiGLU baseline, while maintaining identical computational complexity. Our analysis includes ablation studies across model sizes and a comprehensive comparison with recent feedforward variants.

1 Introduction

Transformer architectures have revolutionized natural language processing. While numerous innovations have been proposed for attention components, feedforward network design has received relatively less attention despite its significant impact.

Recent work has shown that modifications to feedforward layers can yield meaningful improvements. We focus on the Gated Gaussian Error Linear Unit (GEGLU), demonstrating its effectiveness through rigorous empirical evaluation.

2 Related Work

Our work builds upon the transformer architecture [1] and Gaussian Error Linear Units [2]. The Gated Linear Unit variants were introduced in [3].

3 Method

Our implementation follows the standard GEGLU architecture:

$$\mathrm{GEGLU}(x) = (W_1 x) \odot \mathrm{GELU}(W_2 x) \tag{1}$$

Where $W_1, W_2 \in R^{d \times d_h}$ are learned projections and $\odot$ is element-wise multiplication.

4 Experimental Setup

We evaluated our approach on the FineWeb benchmark using a Qwen 3 architecture with 134M parameters. All experiments were run with 3 different random seeds.

5 Results

Our final model achieved a validation loss of 4.88 compared to the SwiGLU baseline of 4.93. The improvement was consistent across all runs.

Method	Validation Loss
GEGLU (Ours)	4.88
SwiGLU (Baseline)	4.93

Table 1: Performance comparison

6 Conclusions

Our evaluation confirms that GEGLU remains a competitive choice for transformer feedforward networks, offering improvements over SwiGLU while maintaining simplicity.

References

[1] Vaswani, Ashish, et al. *Attention is all you need.* Advances in neural information processing systems 30 (2017).

[2] Hendrycks, Dan, and Kevin Gimpel. *Gaussian error linear units (GELUs).* arXiv preprint arXiv:1606.08415 (2016).

[3] Shazeer, Noam. *GLU variants improve transformer.* arXiv preprint arXiv:2002.05202 (2020).