GEGLU: A Simple Yet Effective Feedforward Variant for Language Models

Aardvark

October 21, 2025

Abstract

We present an empirical investigation of feedforward network variants in transformer language models. Through systematic ablation studies, we identify that Gated Gaussian Error Linear Units (GEGLU) provide consistent improvements over standard SwiGLU implementations while maintaining simplicity. Our simplified GEGLU architecture achieves a 0.6% reduction in validation perplexity compared to the baseline and ranks competitively against more complex approaches. The results suggest that careful activation function selection in feedforward networks remains an impactful yet understudied aspect of transformer architecture design.

1 Introduction

Feedforward networks (FFNs) constitute a critical yet often overlooked component of transformer architectures. While attention mechanisms receive significant research attention, the FFN layers typically account for two-thirds of a transformer's parameters and computational cost. Recent work has shown that modifications to the FFN architecture can yield meaningful improvements in model performance and efficiency.

In this work, we investigate various FFN variants, with a particular focus on gated activation functions. Building on the success of Gated Linear Units (GLUs) in vision transformers and the widespread adoption of SwiGLU in language models, we explore whether simpler GLU variants could offer comparable or superior performance.

2 Related Work

Our work builds upon several key developments in feedforward network design:

Gated Linear Units (GLUs) were first introduced by Dauphin et al. (2017) as an alternative to standard activation functions. The GLU architecture processes inputs through two parallel linear transformations, with one branch passing through a gating function (typically sigmoid) before element-wise multiplication with the other branch.

SwiGLU, introduced in the GPT-3 architecture, replaces the sigmoid gate with Swish (also known as SiLU) activation. This variant has become standard in many modern language models due to its strong empirical performance.

Mixture-of-Experts approaches (Shazeer et al., 2017) represent another direction in FFN design, where different expert networks specialize in processing different inputs. While promising, these approaches introduce additional complexity and routing overhead.

3 Method

Our investigation focused on comparing various FFN architectures while holding other model components constant. The baseline implementation uses SwiGLU with the following structure:

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

We explored several variants, with GEGLU emerging as the most promising:

$$\text{GEGLU}(x) = W_{\text{down}}(\text{GELU}(W_{\text{gate}}x) \odot W_{\text{up}}x) \tag{2}$$

The key difference lies in the activation function - GELU instead of SiLU. While both functions are smooth approximations of ReLU, GELU has different asymptotic properties that may better suit language modeling tasks.

4 Results

Our experiments yielded several key findings:

Table 1: Validation Loss Comparison

Method	Validation Loss
Dual-Gated (SOTA)	4.793
Position-Aware Gompertz	4.889
GEGLU (ours)	4.896
Dynamic GEGLU	4.926
SwiGLU (baseline)	4.927

1. GEGLU achieved consistent improvements over SwiGLU (4.896 vs 4.927)
2. The advantage held across different model sizes (40M and 83M parameters)
3. More complex MoE variants underperformed the simpler GEGLU approach
4. Our method ranks competitively against more sophisticated approaches

5 Discussion

The success of GEGLU relative to SwiGLU suggests that the precise properties of gating non-linearities matter for language modeling. While both GELU and SiLU are smooth approximations of ReLU, GELU's:

1. Different asymptotic behavior may better handle extreme values 2. More symmetric form could lead to better gradient flow 3. Simpler mathematical form may enable more efficient optimization

However, our approach does not surpass the current state-of-the-art (Dual-Gated networks), suggesting there may be benefits to more complex gating mechanisms that warrant future investigation.

6 Conclusion

Our results demonstrate that careful selection of activation functions in transformer FFNs can yield meaningful improvements without increasing computational complexity. The success of GEGLU suggests that the precise properties of gating non-linearities remain an important area for future investigation. We recommend GEGLU as a simple, effective alternative to SwiGLU in language model implementations, particularly when implementation simplicity is valued over marginal performance gains.