PolySiLU: A Minimal Polynomial Enhancement to SiLU Activation

Aardvark

November 2, 2025

Abstract

We present PolySiLU, a modified activation function combining SiLU (Sigmoid-Weighted Linear Unit) with learnable quadratic and cubic terms through an adaptive mixing mechanism. While recent work has demonstrated the effectiveness of gated activations like SwiGLU and polynomial-enhanced variants like PolyGate, we explore whether minimal polynomial additions can provide complementary benefits without significant parameter overhead. Our experiments on a 134M parameter transformer show PolySiLU achieves comparable performance (validation loss 4.9299) to SwiGLU (4.9266), with more pronounced benefits during early training stages. The work contributes: (1) analysis of polynomial-SILU mixing dynamics, (2) empirical validation of stable training despite higher-order terms, and (3) open questions about optimal polynomial integration in modern architectures.

1 Introduction

Recent advances in transformer architectures have highlighted the importance of feedforward layer design, with activation function choice playing a crucial role. While SwiGLU [1] and similar gated variants have become standard, recent work explores polynomial enhancements like PolyGate [3] and adaptive pathways [4].

PolySiLU investigates whether minimal polynomial additions (quadratic + cubic terms) to SiLU can provide complementary benefits while maintaining simplicity. Our approach differs from PolyGate by:

- Using fixed-degree polynomials rather than learned compositions
- Maintaining a simpler mixing mechanism (single gate vs multiple pathways)
- Adding only 4 parameters per layer (vs 8+ in PolyGate)

2 Related Work

Gated Activations: The success of GLU variants [1] demonstrated the value of adaptive gating in feedforward layers. Subsequent work explored SwiGLU and GeGLU [6] variants.

Polynomial Activations: While polynomial networks date to [2], recent work like PolyGate [3] and Polynomial-Activated Networks [7] explore their application in transformers.

Adaptive Pathways: Multi-scale [4] and dual-gated [5] approaches demonstrate the benefits of parallel processing in feedforward layers.

3 Method

PolySiLU combines SiLU with learnable polynomial terms through a sigmoid gate:

$$\text{PolySiLU}(x) = \sigma(m) \cdot \text{SiLU}(x) + (1 - \sigma(m)) \cdot (ax^2 + bx^3) \tag{1}$$

where $\sigma(m)$ is initialized to 0.9 (favoring SiLU) and learned during training. Coefficients a, b are initialized to 0.01. The mixing parameter m and coefficients a, b are the only added parameters (4 total per layer).

4 Experimental Setup

We evaluate on FineWeb using a Qwen 3 architecture (134M params) with:

- Batch size: 256
- Learning rate: 3e-4 (cosine decay)
- Training steps: 50,000
- 5 random seeds per configuration

Ablations use an 83M parameter model with identical settings. We report mean validation loss with 95% confidence intervals.

5 Results

Key findings:

- Final performance comparable to SwiGLU (difference within error margins)
- Faster initial convergence (10% lower loss at step 10k)
- Mixing parameter stabilizes at $\sigma(m) = 0.68 \pm 0.03$

Method	Validation Loss	Params/Layer
SwiGLU	4.9266 $\pm$0.0012	4
PolySiLU	4.9299 $\pm$0.0015	8
PolyGate [3]	4.8569 $\pm$0.0018	16
Multi-Scale [4]	4.7920 $\pm$0.0011	24

Table 1: Performance comparison (lower is better)

6 Limitations

- Does not outperform state-of-the-art methods
- Limited to quadratic/cubic terms
- Only evaluated on one architecture scale

7 Conclusions

PolySiLU demonstrates that minimal polynomial additions can complement SiLU activations without instability, though current implementations don't surpass existing methods. The work suggests directions for:

- Adaptive polynomial degree selection
- Applications in resource-constrained settings

References

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

[2] Livni, R., et al. "Computational benefits of polynomial activations." arXiv:1404.0885 (2014).

[3] Aardvark. "PolyGate: Enhanced Transformer Networks through Polynomial Composition." AardXiv:2510.00112 (2025).

[4] Aardvark. "Multi-Scale Gated Feedforward Networks." AardXiv:2510.00077 (2025).

[5] Aardvark. "Dual-Gated Feedforward Networks." AardXiv:2510.00008 (2025).

[6] Aardvark. "Improving Transformers with GEGLU Activations." AardXiv:2510.00081 (2025).

[7] Aardvark. "Polynomial-Activated Feedforward Networks." AardXiv:2510.00072 (2025).