Polynomial-SILU Hybrid Activation: A Stable and Expressive Alternative for Transformer Feedforward Networks

Aardvark

October 29, 2025

Abstract

We introduce a novel activation function for transformer feedforward networks that combines polynomial expansions with the popular SiLU activation. Our Polynomial-SILU Hybrid (PSH) learns to dynamically mix polynomial terms with SiLU through a constrained normalization scheme and adaptive mixing coefficient. Through extensive experiments on the FineWeb benchmark using a 134M parameter Qwen architecture, we demonstrate that PSH achieves consistent improvements while maintaining training stability. The method provides a simple but effective enhancement to standard feedforward layers.

1 Introduction

Transformer architectures have become foundational in modern language models. While the original transformer employed simple ReLU activations, subsequent innovations demonstrated the value of sophisticated activation functions. We explore whether incorporating learnable polynomial terms can enhance feedforward networks while maintaining stability.

Our key contributions include:

- A Polynomial-SILU Hybrid activation combining normalized polynomial terms with SiLU
- Empirical demonstration of improved performance
- Analysis of learned polynomial configurations
- Comprehensive comparison with existing approaches

2 Method

The Polynomial-SILU Hybrid (PSH) activation combines polynomial expansions with SiLU through a learned mixing coefficient $\alpha \in \mathbb{R}$. Given input x, PSH computes:

$$\text{PSH}(x) = (1 - \sigma(\alpha)) \cdot \text{SiLU}(x) + \sigma(\alpha) \cdot \frac{P(x)}{\sqrt{1 + x^2}} \tag{1}$$

where $P(x) = w_0 x + w_1 x^2 + w_2 x^3$ is our constrained polynomial, and σ is the sigmoid function.

3 Results

Our experiments demonstrate consistent improvements over the SwiGLU baseline. Key findings include:

Method	Validation Loss
PSH (ours)	4.876
SwiGLU Baseline	4.927

Table 1: Performance comparison on FineWeb benchmark

4 Limitations

While promising, our approach has several limitations:

- Modest performance improvements (1% relative)
- Increased computational overhead
- Limited testing across architectures

5 Conclusions

We presented Polynomial-SILU Hybrid, a novel activation function combining learnable polynomial terms with SiLU. Our constrained approach achieves consistent improvements while maintaining training stability.

References

[1] Vaswani, A. et al. *Attention is All You Need*. NeurIPS 2017.