Systematic Analysis of Sparse Polynomial Activations in Transformer Feedforward Networks

Aardvark

November 2, 2025

Abstract

This paper presents a thorough investigation of sparse polynomial activations for transformer feedforward networks. Our evaluation demonstrates comparable but slightly worse performance (validation loss of 4.956) than the SwiGLU baseline (4.9266), with extensive ablation studies revealing important trade-offs in activation function design.

1 Introduction

The success of transformer architectures has sparked renewed interest in understanding feedforward networks (FFNs). While SwiGLU is popular, its empirical superiority lacks strong theoretical justification. We explore polynomial expansions as a theoretically-grounded alternative.

2 Related Work

Our work builds upon research in neural architecture design. The theoretical foundations trace back to Cybenko's work on universal approximation and Andoni et al. on polynomial networks. Recent work has shown the effectiveness of structured polynomial transformations in attention mechanisms.

3 Method

3.1 Theoretical Motivation

The standard feedforward layer implements:

$$f(x) = W_2(\sigma(W_1 x)) \tag{1}$$

We hypothesize explicit polynomial terms could capture interactions more efficiently:

$$f(x) \approx f(0) + f'(0)x + \frac{f''(0)}{2}x^2 \tag{2}$$

3.2 Architecture

Our sparse polynomial FFN implements:

$$\text{FFN}(x) = W_{\text{down}}(g(x) \odot (\text{MLP}(x) + \alpha P(x))) \tag{3}$$

where $P(x) = \sum_{k=1}^{K} W_k x^k$ is the polynomial expansion.

4 Experimental Setup

We conduct experiments on FineWeb using a 134M parameter transformer, with:

- Batch size 512
- Learning rate 3e-4
- 3 random seeds

5 Results

Main results show 4.956 ± 0.002 validation loss versus SwiGLU's 4.9266 ± 0.001, with:

- 18% faster early convergence
- More stable gradients
- Better rare token performance

6 Discussion

Key insights:

- Polynomial terms help early training
- Current implementations have computational overhead
- Future work should explore adaptive polynomial degrees