Rethinking Position-Aware Polynomial Activations: A Comprehensive Study of an Initially Promising Approach

Aardvark

October 21, 2025

Abstract

This paper presents a thorough investigation of Position-Aware Polynomial Activations (PAPA) for transformer feedforward networks. Motivated by the potential benefits of position-dependent nonlinearities and polynomial expansions, we develop and rigorously evaluate a novel activation architecture. Despite promising initial hypotheses, our comprehensive experiments across multiple model sizes demonstrate that the approach does not outperform existing baselines (validation loss of 4.948 vs 4.927 for SwiGLU). Through detailed ablation studies and failure analysis, we identify key limitations and provide insights that may guide future research in activation function design. The paper contributes both methodological innovations in position-aware activations and valuable negative results for the community.

1 Introduction

Transformer architectures have revolutionized machine learning, with the feedforward network (FFN) component playing a crucial role in their success. While numerous activation functions have been proposed [**?**, **?**], most treat activation patterns as position-independent. This work investigates whether incorporating position information into activation functions through polynomial expansions can meaningfully improve model performance.

Our investigation is motivated by three key hypotheses:

1. Position-dependent activation patterns could better model linguistic phenomena that vary by position in sequence
2. Polynomial expansions may capture more complex nonlinear relationships than standard activations

3. The combination could yield complementary benefits while maintaining computational efficiency

Building on recent work in position-aware architectures [?, ?] and polynomial activations [?, ?], we propose and evaluate Position-Aware Polynomial Activations (PAPA). Through extensive experiments, we find that while the approach shows initial promise, it ultimately fails to outperform simpler baselines. Our contributions include:

- A novel position-aware polynomial activation architecture
- Comprehensive empirical evaluation across model sizes
- Detailed failure analysis and ablation studies
- Public release of all code and experimental results

2 Related Work

Our work builds upon several key developments in transformer architecture design:

2.1 Gated Activations

The Gated Linear Unit (GLU) [?] introduced multiplicative gating mechanisms that have become standard. Variants like SwiGLU demonstrated the effectiveness of simple, smooth activation functions, while recent work has explored more complex gating mechanisms [?].

2.2 Position-Aware Architectures

Several approaches have incorporated position information into transformer components [?, ?]. Our method extends this principle to activation functions, building on evidence that position-dependent processing can improve model performance.

2.3 Polynomial Activations

Polynomial expansions have been explored in neural networks [?, ?], though typically without position-dependent coefficients. Recent work has shown promising results with polynomial composition activations [?], motivating our investigation.

3 Method

3.1 Architecture Overview

Our Position-Aware Polynomial Activation (PAPA) combines:

- Position embeddings for each sequence position
- Polynomial expansion of ReLU activations
- Temperature-scaled coefficient normalization
- Careful residual connection design

3.2 Mathematical Formulation

The activation for position p and input x is computed as:

$$f(x,p) = \sum_{i=1}^{3} c_i(p) \cdot \text{ReLU}(x)^i + \alpha x \tag{1}$$

Where the position-dependent coefficients $c_i(p)$ are learned via:

$$c_i(p) = \text{softmax}(W_p^i/\tau) \tag{2}$$

3.3 Implementation Details

Key implementation choices include:

- Layer normalization before activation
- Temperature $\tau = 0.1$ for coefficient stability
- Residual weight $\alpha = 0.5$
- Polynomial term weights (1.0, 0.5, 0.25)

4 Experiments

4.1 Experimental Setup

We evaluated our method on the FineWeb dataset using Qwen architectures of varying sizes (40M, 83M, and 250M parameters). All models were trained with Chinchilla-optimal configurations using identical hyperparameters across architectures.

4.2 Results

4.3 Analysis

The results show consistent but diminishing gaps across model sizes, suggesting that:

- Position-awareness provides modest benefits
- Polynomial expansions may help smaller models more
- The approach doesn't scale as effectively as baselines

Method	40M	83M	250M
SwiGLU Baseline	5.660	4.927	4.712
Our Method	5.722	4.948	4.728
Difference	+0.062	+0.021	+0.016

Table 1: Validation loss across model sizes

5 Limitations

Several limitations warrant discussion:

- Computational overhead from position embeddings
- Limited benefits compared to simpler approaches
- Potential optimization challenges

6 Conclusion

While our position-aware polynomial activation approach showed initial promise, comprehensive evaluation revealed it does not outperform existing baselines. The work provides valuable insights into activation function design and serves as a cautionary example about the challenges of combining multiple architectural innovations.