Dynamic Gating Feedforward Networks: Analysis of Combining Polynomial Activations with Key-Value Memory Patterns

Aardvark

October 23, 2025

Abstract

We present a comprehensive analysis of Dynamic Gating Feedforward Networks (DGFN), an architecture combining polynomial composition activations with key-value memory patterns in transformer feedforward layers. Despite theoretical promise, our experiments on the FineWeb dataset (2B tokens) using a Qwen 3 architecture (83M parameters) show the approach achieves a validation loss of 5.017, underperforming both the SwiGLU baseline (4.927) and state-of-the-art methods (best 4.793). Through extensive ablation studies and architectural analysis, we identify key challenges in combining these mechanisms. Our results suggest that while both polynomial activations and memory patterns individually offer benefits, their combination requires more sophisticated coordination mechanisms than simple learned mixing.

1 Introduction

Transformer architectures have revolutionized natural language processing, with their feedforward networks (FFNs) playing a crucial role in information processing. Recent work has characterized FFNs as key-value memories while separately demonstrating the benefits of polynomial composition activations. This paper investigates whether combining these approaches can yield complementary benefits.

2 Method

The DGFN architecture processes inputs through parallel pathways:

- **Polynomial Pathway**: Applies PolyReLU activation to projected inputs
- **Memory Pathway**: Implements key-value memory pattern detection
- **Dynamic Gate**: Learns to mix pathway outputs

Final output is computed as:

$$y = W_{down}(\lambda \cdot f_{poly}(W_{poly}x) + (1 - \lambda) \cdot f_{mem}(W_{mem}x)) \tag{1}$$

3 Results

Our experiments show:

- DGFN achieves validation loss of 5.017 vs 4.927 baseline
- Both pathways contribute positively but interfere with each other
- Requires careful initialization and training

4 Conclusion

While our combined approach underperformed, it provides valuable insights for future work on hybrid feedforward architectures.