Analysis of Dynamic Gating in Transformer Feedforward Networks

Aardvark

November 3, 2025

Abstract

We present a comprehensive study of dynamic gating mechanisms in transformer feedforward networks. Our proposed architecture introduces a lightweight controller network to generate input-dependent gating coefficients. While initial results on smaller models showed promise, our analysis reveals significant challenges in scaling these approaches, with our final implementation achieving a validation loss of 4.935 compared to the SwiGLU baseline of 4.927.

1 Introduction

Transformer architectures have become foundational in modern NLP, with the feedforward layer playing a crucial role in their success. Recent work has explored various enhancements to feedforward layers through gating mechanisms and alternative activation functions.

Our work makes three key contributions:

- Systematic evaluation of dynamic gating
- Detailed scaling analysis
- Empirical evidence of limitations

2 Methodology

Our dynamic architecture modifies the standard feedforward layer through:

2.1 Architecture

The standard feedforward layer computes:

$$FFN(x) = W_2(GELU(W_1x)) \tag{1}$$

Our dynamic variant introduces a controller network:

$$C(x) = W_c^2(GELU(W_c^1 x)) \tag{2}$$

The final output becomes:

$$DynamicFFN(x) = W_2(sigmoid(C(x)) \cdot GELU(W_1 x)) \tag{3}$$

2.2 Implementation

- Orthogonal initialization for main weights
- Normal initialization for controller
- Hidden dimension expansion factor of 4

3 Experiments

We conducted experiments at two scales:

3.1 Ablation Studies

- Model size: 83M parameters
- Batch size: 512
- Learning rate: 6e-4

3.2 Full-scale Training

- Model size: 134M parameters
- Batch size: 1024
- Learning rate: 3e-4

4 Results

Table 1: Performance Comparison

Method	Validation Loss
SwiGLU Baseline	4.927
Our Method	4.935
Best Method	4.792

5 Discussion

Our results reveal:

- Scaling challenges from small to large models
- Optimization difficulties with dynamic components
- Questionable efficiency tradeoffs

6 Conclusion

Our study provides valuable negative results about dynamic gating in transformers. Future work should focus on more efficient adaptation mechanisms.