Dynamic Memory Gating: An Investigation into Pattern-Specialized Feedforward Networks

Aardvark

October 20, 2025

Abstract

This paper investigates Dynamic Memory Gating (DMG), a novel approach to transformer feedforward networks that employs parallel pattern detection heads with learned gating mechanisms. While theoretically promising for specialized processing of different input patterns, our implementation achieved a validation loss of 5.057 on the FineWeb benchmark, underperforming the SWiGLU baseline (4.927) and current state-of-the-art methods. We analyze the architectural choices, training dynamics, and potential limitations that may have contributed to these results, providing insights for future improvements in adaptive feedforward designs.

1 Introduction

Transformer feedforward networks have evolved significantly from their original formulation, with gated mechanisms like GELU and SWiGLU demonstrating consistent improvements. Our work explores whether further gains can be achieved through explicit pattern specialization via dynamic gating across multiple parallel processing heads.

2 Related Work

Key influences include:

- Mixture of Experts approaches [?]
- Gated linear units [?]
- Activation function variants [?]

3 Methodology

DMG processes inputs through:

- Multiple parallel detection heads (default: 4)
- Learned gating between heads (softmax/top-k variants)
- Combined weighted output

4 Results

Our implementation achieved 5.057 validation loss versus:

Method	Loss
SWiGLU (baseline)	4.927
Dual-Gated (SOTA)	4.793
Our DMG	5.057

5 Discussion

Potential limitations include:

- Overhead of gating computation
- Insufficient head specialization
- Training instability

6 Conclusion

While DMG did not outperform baselines, it provides valuable insights into adaptive feedforward designs.