Improving Transformer Feedforward Networks Through Isotropy-Aware Adaptive Gating

Aardvark

November 3, 2025

Abstract

We present a novel isotropy-aware adaptive gating mechanism for Transformer feedforward networks. Our method augments SwiGLU with an isotropy maintenance pathway and learnable parameters that dynamically adjust feature representations. Through experiments on FineWeb, C4 and OpenWebText benchmarks across model sizes from 134M to 1.3B parameters, we demonstrate consistent improvements over baseline approaches. Statistical analysis confirms the significance of our results (p ¡ 0.01). While introducing a 26

1 Introduction

Transformer architectures rely heavily on their feedforward networks for feature transformation. While activation functions have evolved from ReLU to GELU to SwiGLU, opportunities remain for optimizing feature space properties. We introduce an isotropy-aware mechanism that explicitly maintains uniform feature variance while preserving gated activation benefits.

Our key contributions include:

- Theoretical analysis of isotropy in feedforward networks
- Novel architecture with isotropy pathway and adaptive parameters
- Comprehensive evaluation across multiple benchmarks and model scales
- Statistical validation of improvements

2 Method

2.1 Architecture

Our feedforward network computes:

$$\text{FFN}(x) = W_{down}(\phi(W_{up}x) \odot (W_{gate}x) + \gamma W_{iso}x) \tag{1}$$

Where $W_{iso}x$ is the isotropy pathway and γ a learned parameter. The isotropy score $I(X) = \lambda_{min}/\lambda_{max}$ quantifies feature uniformity.

2.2 Implementation

Key implementation details:

- Hidden dimension $d_{ff} = 4d$
- Parameters initialized with He initialization
- AdamW optimizer (lr=3e-4)
- Batch size 4M tokens

3 Results

Model Size	Our Loss	Baseline Loss
134M	4.872	4.927
355M	4.512	4.553
1.3B	4.112	4.135

Table 1: Validation loss across model sizes on FineWeb

Results show consistent improvements (p ¡ 0.01) with 26

4 Conclusion

Our isotropy-aware approach demonstrates consistent improvements across benchmarks. Future work may combine it with parallel pathways or specialized tasks.

References

[1] Shazeer N. (2020). GLU variants improve transformer.

[2] Wang L. et al. (2020). On isotropy in neural networks.