Adaptive Threshold Gating in Transformer Feedforward Networks

Aardvark

October 28, 2025

Abstract

This paper investigates Adaptive Threshold Gating (ATG) for transformer feedforward networks. Our experiments show ATG achieves a validation loss of 4.966, slightly underperforming the SwiGLU baseline of 4.9266. The results suggest adaptive thresholds may have limited benefits in standard architectures.

1 Introduction

We propose ATG, which adjusts gating based on input statistics. This is inspired by biological neurons and prior work on adaptive activations.

2 Method

ATG modifies SwiGLU with:

$$FFN(x) = W_d(SiLU(W_g x) \cdot \sigma(\alpha\mu + \beta\sigma^2) \cdot W_u x) \tag{1}$$

where μ, σ^2 are input statistics and α, β are learned scalars.

3 Results

On a 134M parameter model:

- ATG loss: 4.966
- SwiGLU loss: 4.9266
- Memory: 39.5GB vs 31.5GB

4 Conclusion

ATG shows modest results, suggesting input statistics may not be optimal for gating decisions in transformers.