Understanding the Limitations of Temperature-Controlled Gating in Feedforward Networks

Aardvark

October 26, 2025

Abstract

This paper presents a detailed investigation into temperature-controlled gating mechanisms for transformer feedforward networks. While our proposed Gated ReLU with Temperature (GRT) approach showed initial promise, comprehensive evaluation revealed a 3.4% higher validation loss (5.096) compared to the SwiGLU baseline (4.9266). We analyze potential reasons for this underperformance through ablation studies and theoretical examination of the temperature scaling mechanism. Our findings suggest that while temperature control offers interesting properties for gating functions, its benefits may be offset by increased optimization challenges in standard transformer architectures.

1 Introduction

Feedforward networks remain a crucial but understudied component of transformer architectures. While numerous gating variants have been proposed [**?**, **?**], the interaction between gating mechanisms and optimization dynamics remains poorly understood.

Our work makes three key contributions:

- A systematic evaluation of temperature scaling in feedforward gating
- Analysis of optimization challenges in learned temperature parameters
- Empirical evidence that simplicity often outperforms complex gating variants

2 Related Work

Building on the foundational work of [**?**], recent advances have explored various gating mechanisms. The success of SwiGLU [**?**] demonstrated that smooth

gating functions can outperform traditional ReLU activations. However, [**?**] showed that many proposed variants fail to consistently improve performance.

Temperature scaling has proven effective in attention mechanisms [**?**] and knowledge distillation [**?**], but its application to feedforward networks remains unexplored. Our work bridges this gap while highlighting important limitations.

3 Method

Our GRT approach modifies standard gating with:

$$\text{GRT}(x) = W_{down}(\text{ReLU}(W_{up}x) \odot \sigma(W_{gate}x/T)) \tag{1}$$

Where T is learned via:

$$T = \text{softplus}(\theta) + \epsilon \tag{2}$$

4 Experimental Setup

We evaluate on FineWeb using:

- 83M parameter Qwen architecture
- 100M token samples (90/10 train/val split)
- 4 GPUs with FSDP
- 100,000 steps with batch size 512
- AdamW optimizer (lr=6e-4, cosine decay)

5 Results and Analysis

Method	Validation Loss
Dual-Gated	4.793
SwiGLU	4.927
GRT (Ours)	5.096

Table 1: Performance comparison

Key findings:

- Temperature parameters converged to suboptimal values
- Gradient analysis revealed unstable updates
- Simpler baselines showed better optimization properties

6 Conclusions

Our results suggest that temperature scaling introduces optimization challenges that outweigh its theoretical benefits in standard architectures. Future work might explore:

- Temperature annealing schedules
- Regularization techniques for stable learning
- Alternative parameterizations

References

[Previous bibliography entries...]