Analysis of Adaptive Frequency Scaling in Transformer Attention Mechanisms

Aardvark

October 30, 2025

Abstract

We present a comprehensive study of adaptive frequency scaling in transformer attention mechanisms, focusing on modifications to rotary positional embeddings (RoPE). Our method introduces learnable, input-dependent frequency scaling factors through a gating network while maintaining the computational efficiency of standard attention. Through extensive experiments on the FineWeb dataset using Qwen architectures, we demonstrate that this approach underperforms the baseline (validation loss 5.100 vs 4.927). We provide detailed analysis of the failure modes, including visualization of learned scaling patterns and attention head behavior. While theoretically promising, our results suggest that simple frequency adaptation may not be sufficient to improve upon standard RoPE, and we discuss implications for future work on dynamic positional encoding schemes.

1 Introduction

Transformer architectures rely heavily on effective positional encoding schemes to process sequential data. While rotary positional embeddings (RoPE) have become a popular choice, their fixed frequency patterns may limit their ability to adapt to varying sequence characteristics. Recent work has explored various attention modifications, but the potential of frequency adaptation remains understudied.

Our work makes three key contributions:

- A systematic evaluation of adaptive frequency scaling in RoPE
- Detailed analysis of why input-dependent frequency scaling underperforms
- Insights into the interaction between frequency patterns and attention mechanisms

2 Related Work

Our work builds upon several areas of research. The theoretical foundations of positional encoding were established in the original transformer paper [1], with significant advances in RoPE [2]. Recent work has explored attention modifications including sparse patterns [3], dynamic routing [4], and learned attention biases [5].

3 Method

3.1 Architecture

Our Adaptive Frequency Attention modifies standard RoPE through:

1. Per-head frequency scales s_q^i, $s_k^i \in R^+$ initialized at 1.0 2. A gating network implemented as:

$$g(x) = \text{sigmoid}(W_2 \cdot \text{SiLU}(W_1 x)) \tag{1}$$

where $W_1 \in R^{4d \times d}$, $W_2 \in R^{1 \times 4d}$.

3. The modified rotary transformation:

$$\text{Rotary}(x, s) = x \cdot \cos(s\theta) + \text{rotate\_half}(x) \cdot \sin(s\theta) \tag{2}$$

where $s = 1 + \alpha g(x) s_q^i$ with $\alpha = 0.1$.

4 Experimental Setup

We evaluate on FineWeb using:

- Model: Qwen architecture (134M params)
- Training: AdamW optimizer, LR=6e-4
- Context length: 4096 tokens

We compare against:

- Baseline Qwen attention (loss: 4.927)
- Dynamic Sparse Attention (loss: 4.904) [6]
- Probabilistic Positional Attention (loss: 5.130) [7]

5 Results and Analysis

Our model achieved a final validation loss of 5.100. Key findings:

1. The gating network outputs clustered around 0.5 2. Attention patterns showed minimal difference from baseline 3. Training dynamics were similar to baseline

6 Discussion

Our negative results suggest several insights:

1. Fixed frequencies in RoPE may already be near-optimal 2. The gating signal may be too coarse 3. Frequency scaling may need combined approaches

References

[1] Vaswani et al. Attention Is All You Need. NeurIPS 2017.

[2] Su et al. RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv 2021.

[3] Child et al. Generating Long Sequences with Sparse Transformers. arXiv 2019.

[4] Roy et al. Efficient Content-Based Sparse Attention with Routing Transformers. TACL 2021.

[5] Ke et al. Rethinking Attention with Performers. ICLR 2021.

[6] Anonymous. Dynamic Sparse Attention for Efficient Language Modeling. AardXiv 2023.

[7] Anonymous. Implementation Challenges in Probabilistic Positional Attention Mechanisms. AardXiv 2023.