Implementation Challenges in Probabilistic Positional Attention Mechanisms

Aardvark

October 17, 2025

Abstract

This paper documents our investigation into probabilistic positional priors for transformer attention mechanisms and the technical challenges encountered during implementation. We propose a modification to standard attention that incorporates learnable positional decay and scale parameters, building on prior work in relative position encodings and learned attention biases. While our baseline implementation of the Qwen attention achieved a validation loss of 5.13 on the FineWeb dataset (compared to the reference Qwen baseline of 4.9266), we encountered persistent tensor shape mismatches when integrating our probabilistic modifications. We analyze these implementation challenges in detail and discuss lessons learned for future work in attention mechanism modifications.

1 Introduction

Transformer architectures have demonstrated remarkable success in natural language processing, largely due to their attention mechanisms. The standard scaled dot-product attention computes pairwise interactions between all tokens, while numerous extensions have proposed incorporating structural biases like positional information. Our work investigates whether adding learnable probabilistic positional priors could improve attention computation while maintaining computational efficiency.

Our key contributions include:

- A detailed proposal for integrating probabilistic positional priors into transformer attention
- Analysis of implementation challenges encountered when modifying existing attention implementations
- Empirical validation of the baseline Qwen attention implementation
- Discussion of lessons learned for future attention mechanism modifications

2 Related Work

Prior work has explored various approaches to incorporating positional information into transformers:

Fixed Positional Encodings The original transformer paper [1] introduced sinusoidal positional embeddings. Subsequent work developed learned positional embeddings [6].

Relative Position Encodings Shaw et al. [2] proposed relative position representations in self-attention. Dai et al. [3] extended this with Transformer-XL.

Learned Attention Biases Ke et al. [4] introduced learnable position biases in Performers. Our approach differs by modeling position as a learnable probabilistic prior with decay characteristics.

Bayesian Attention Recent work [5] has explored probabilistic interpretations of attention, though with different formulations than our proposed approach.

3 Method

Our proposed probabilistic attention modifies the standard scaled dot-product attention by adding a learnable positional component:

$$A_{ij} = \frac{Q_i K_j^T}{\sqrt{d_k}} + \phi(|i-j|; \alpha, \beta) \tag{1}$$

where $\phi(d; \alpha, \beta) = -|\alpha d|^\beta$ is our learnable positional prior with parameters α (decay rate) and β (curvature). These parameters are initialized to 1 and learned during training.

The implementation challenges we encountered stemmed from:

- Shape mismatches when combining the positional prior with attention scores
- Integration with the existing rotary position embeddings
- Maintaining compatibility with the caching mechanism for efficient inference

4 Experimental Setup

We evaluated on the FineWeb dataset using a Qwen-style transformer with 134M parameters. The model was trained for 640 steps with a batch size of 4.2M tokens using Chinchilla-optimal training configuration. Our baseline used the standard Qwen attention implementation.

5 Results

Table 1: Model Performance Comparison on FineWeb Dataset

Method	Parameters	Validation Loss
Qwen Baseline	134M	4.9266
Our Baseline Implementation	134M	5.13

The baseline Qwen attention implementation achieved a validation loss of 5.13, compared to the reference Qwen baseline of 4.9266. This slight degradation may be due to differences in implementation details or initialization.

Key implementation challenges included:

- Tensor shape mismatches when broadcasting the positional prior
- Difficulty integrating with the existing rotary position embeddings
- Maintaining compatibility with the KV cache during inference

6 Discussion

Our experience highlights several important considerations when modifying attention mechanisms:

Shape Compatibility Attention modifications must carefully maintain tensor shape consistency throughout all operations.

Integration Challenges Combining multiple position-aware components (rotary embeddings, positional priors, etc.) requires careful design.

Debugging Complexity Attention implementations involve complex tensor operations that can be challenging to debug.

7 Conclusions

While our probabilistic attention approach was not successfully implemented, the challenges we documented provide valuable insights for future work in attention mechanism modifications. Future directions could include:

- Alternative formulations of positional priors that maintain shape compatibility
- Gradual integration approaches to isolate implementation challenges
- More comprehensive testing frameworks for attention modifications

References

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

[2] Shaw et al. *Self-Attention with Relative Position Representations*. NAACL 2018.

[3] Dai et al. *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. ACL 2019.

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

[5] Fan et al. *Bayesian Attention Modules*. NeurIPS 2020.

[6] Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. NAACL 2019.