Rotary Position Encoding for Vision Transformer and Performer: A Comparative Study

Lechen Zhang Krzysztof Choromanski

Columbia University
116th and Broadway, New York, NY 10027
lz2954, kmc2178@columbia.edu

1. Introduction

Position embeddings are important in transformer architectures because they help capture spatial relationships in input data. Regular approaches use learned or fixed positional embeddings, but Rotary Position Embeddings (RoPE) [4] have shown good results in natural language processing tasks. Based on the work of Heo et al. [3], we look at using RoPE for vision tasks, mainly focusing on how to use it with both standard Vision Transformer and Performer architecture. Vision Transformer (ViT) [2] and Performer [1] are two different ways to handle attention mechanisms in deep learning. Regular ViTs use full attention mechanisms and compute interactions between all tokens, which leads to quadratic computational complexity $O(n^2)$ based on the sequence length. This becomes hard when processing large images or when there are limited computational resources. Performer fixes this problem by approximating the attention mechanism using Fast Attention Via positive Orthogonal Random features (FAVOR+). This method reduces the computational complexity to linear $O(n)$ and keeps performance similar to full attention models. The main idea of Performer is that they can approximate the softmax kernel using random feature decomposition:

$$K(Q,K) = \exp(QK^T/\sqrt{d}) \approx \phi(Q)\phi(K)^T \tag{1}$$

where $\phi(\cdot)$ represents a positive random feature transformation. This change lets Performer run much faster than standard transformer and still maintain good accuracy. By putting RoPE together with both ViT and Performer architectures, we want to see if rotary embeddings can make each approach better - improving the high accuracy of ViTs and the fast computing of Performer. The main contributions of this paper are:

- Implementation of RoPE with efficient variants (axial and mixed) in both ViT and Performer architectures
- Comparison of computational efficiency and accuracy
- Testing on the CIFAR-10 dataset
- Study of how RoPE works with both full attention and linear attention mechanisms

2. Methodology

2.1. RoPE Implementation

Our implementation includes two types of Rotary Position Embeddings for vision tasks: axial and mixed. Both types help capture 2D positional information and maintain efficient computation.

2.2. Axial RoPE

For the axial type, we calculate position-dependent rotation matrices separately for horizontal and vertical directions. Given a dimension d and position (x, y):

$$\begin{aligned} f_{\text{base}} &= \frac{1}{\theta^{(2i/d)}} \quad \text{for } i \in [0, d/4) \\ \Theta_x &= x \otimes f_{\text{base}} \\ \Theta_y &= y \otimes f_{\text{base}} \\ \text{freqs}_{cis} &= [\exp(i\Theta_x); \exp(i\Theta_y)] \end{aligned} \tag{2}$$

Here, θ controls frequency scaling, and $\otimes$ represents the outer product.

2.3. Mixed RoPE

For the mixed type, we perform rotation in a diagonal direction and add learnable frequency patterns with rotations for each head:

$$\begin{aligned} m &= \frac{1}{\theta^{(2i/d)}} \quad \text{for } i \in [0, d/4) \\ f_x^h &= m[\cos(\phi_h), \cos(\frac{\pi}{2} + \phi_h)] \\ f_y^h &= m[\sin(\phi_h), \sin(\frac{\pi}{2} + \phi_h)] \\ \Theta_{xy} &= x f_x^h + y f_y^h \end{aligned} \tag{3}$$

In this case, ϕ_h is a random rotation angle for head h.

2.4. Performer with ReLU Kernel

The original Performer paper uses FAVOR+ with exponential kernels. Instead, we use a modified version with ReLU activation to improve efficiency. We can write our kernel approximation as follows:

$$K(Q, K) = \phi(Q)\phi(K)^T \tag{4}$$

Our feature map $\phi(\cdot)$ is:

$$\phi(X) = \text{ReLU}(XP^T) + \epsilon \tag{5}$$

Here, $P \in \mathbb{R}^{r \times d}$ is a random projection matrix where:

$$r = \min(d_{\text{head}}, \lceil \log(N + 1) \cdot d_{\text{head}} \rceil) \tag{6}$$

N is the sequence length and d_{head} is the dimension per attention head.

Adaptive Feature Dimension We initialize the projection matrix P as:

$$P_{ij} \sim \mathcal{N}(0, 2/d_{\text{head}}) \tag{7}$$

This initialization ensures that the random projections scale correctly. The feature dimension r changes based on sequence length logarithmically, which balances computational efficiency and approximation accuracy.

3. Experiments

3.1. Dataset and Training Setup

We conducted experiments on the CIFAR-10 image classification dataset, consisting of 60,000 32×32 color images in 10 classes. All experiments were performed on a single NVIDIA RTX 3090 Ti GPU. Our experiment includes the following key setups:

Training Configuration:

- Training/test split: 50,000/10,000
- Batch size: 128
- Optimizer: AdamW
- Base learning rate: 3e-4
- Number of epochs: 100
- Learning rate schedule: Warmup (5 epochs) + cosine decay

Model Architecture Parameters:

- Image size: 224 (upsampled from 32×32)
- Patch size: 16×16
- Embedding dimension: 384
- Number of heads: 6 (ViT) / 3 (Performer)
- Depth: 12 (ViT) / 8 (Performer)
- RoPE θ: 100.0 for Axial, 10.0 for Mixed

Data Augmentation Configuration:

- Random horizontal flip
- Random crop (32×32 with padding=4)
- Normalization (Using: mean=[0.4914, 0.4822, 0.4465], std=[0.2470, 0.2435, 0.2616])
- Resize to 224×224 with anti-aliasing

3.2. Results

The experimental results are presented in Table 1 and illustrated in Figure 1.

Looking at our test results, we found three important things. The ViT models performed better than Performer models, but the difference was small. Also, the Mixed RoPE variant got better accuracy scores, but it took a bit more time to train than the axial versions. Finally, the Performer models ran about 20% faster while keeping similar accuracy levels.

4. Discussion

Our test results show important findings about how different models perform and what trade-offs they have. When we compare ViT and Performer models, we see a clear balance between speed and accuracy. The ViT models get better accuracy in classification tasks, with slightly higher scores in all measurements. But this better performance needs more computing power. On the other hand, the Performer model runs much faster and still gets good accuracy scores, so it works well when you need to save computing resources. Looking at the RoPE types, we found some key differences in how well they work. The mixed RoPE type works better than the axial type in both models. This difference is bigger in the ViT model, which suggests that the mixed type is better at understanding diagonal positions when using full attention. Looking at computing speed, the Performer models run noticeably faster. They take about 20% less time to train compared to ViT models. Even though they're a bit less accurate, this faster training time means Performer models with RoPE could work well for tasks where training speed and resource use are important.

5. Conclusion

Our study shows that we can successfully apply Rotary Position Embeddings in both Vision Transformer and Performer models for computer vision tasks. The results show that while ViT models get slightly better but the best accuracy, reaching **81.42%** with the mixed RoPE, Performer models run much faster while keeping similar performance. Because of this, Performer models with RoPE could work especially well when computing resources are limited or when running large-scale tasks that need fast training. Due to time and hardware limitations, We could only test on a limited amount of data and limited times of experiments.

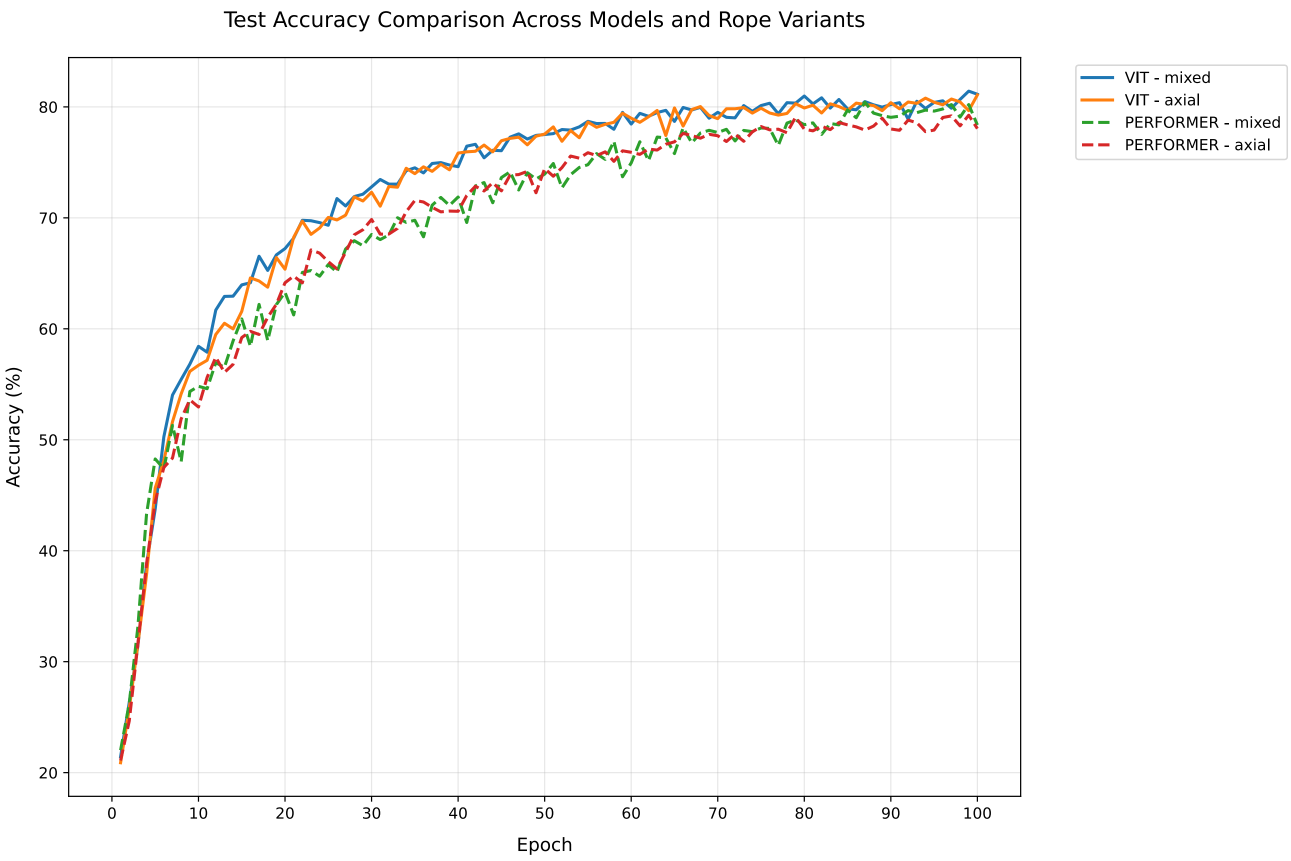

Figure 1. Test Accuracy Comparison of Different Models

Table 1. Performance Comparison of Different Models

Model	Best Test Acc. ↑	Final Test Acc. ↑	Model Parameters	Training Time (One Epoch) ↓
ViT (Mixed)	**81.42%**	**81.14%**	21586954	119
ViT (Axial)	81.08%	81.08%	21586954	115
Performer (Mixed)	80.43%	78.37%	21573130	89
Performer (Axial)	79.25%	78.04%	21573130	**86**

However, these results still give us good information about how different models and RoPE variants work. To understand more about how well these models work at bigger scales and in different situations, we would need to test them on bigger datasets like ImageNet and try more complex vision tasks.

References

[1] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021. 1

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[3] Byeongho Heo, Song Park, Dongyoon Han, and Sangdoo Yun. Rotary position embedding for vision transformer. In *European Conference on Computer Vision*, pages 289–305. Springer, 2025. 1

[4] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. 1