Adversarial Robustness of Vision Transformers

Karan Gandhi Anurag Singh Arjun Dikshit Aarsh Wankar

ES 667: Deep Learning

April 18, 2025

Outline

Outline

Introduction

In recent years, Vision Transformers (ViTs) [1] have gained popularity due to their impressive performance in image recognition. These ViTs have shown greater resilience to spatial perturbation attacks compared to CNNs. However, they remain highly vulnerable to frequency-based attacks. In this project-

Introduction

In recent years, Vision Transformers (ViTs) [1] have gained popularity due to their impressive performance in image recognition. These ViTs have shown greater resilience to spatial perturbation attacks compared to CNNs. However, they remain highly vulnerable to frequency-based attacks. In this project-

- We studied why ViTs are adversarially more vulnerable to frequency-based attacks.
- Based on the results, we found that transformers have a major vulnerability: The positional encodings.
- So in this project, we propose a novel positional encoding which is adversarially more robust to frequency based attacks.

What is an Adversarial Attack?

Adversarial Attack:

- An **adversarial attack** is when small, carefully crafted perturbations are made to some pixels of an input image to fool a trained model.

What is an Adversarial Attack?

Adversarial Attack:

- An **adversarial attack** is when small, carefully crafted perturbations are made to some pixels of an input image to fool a trained model.
- **White-box Attack**
  - Full access to the model (architecture, weights, gradients).
  - Can compute exact perturbations to fool the model.
  - Examples: FGSM, PGD
- **Black-box Attack**
  - No access to model internals.
  - Can only query the model and observe outputs.
  - Examples: Random Noise Attacks, Backdoor Attacks

Outline

Problem Formulation

Problem Formulation: Vision Transformers are inherently vulnerable to Frequency based attacks. In this project, we investigate why this is the case. We attributed this to the major vulnerability of vision transformers: Its positional encoding. Additionally we propose a novel Positional Encoding and analyse its robustness in comparision to other standard existing Positional Encodings.

Outline

Frequency Based Attacks

An unified adversarial attack framework was introduced by [2] to apply perturbations selectively in both spatial and spectral domains, enabling frequency-selective adversarial attacks on Vision Transformers (ViTs).

In this unified attack we first get the Fourier transform of an image X, which can be expressed as:

$$\mathcal{F}\{X\} = M \cdot e^{j\phi}$$

where M and ϕ represent the magnitude and phase spectra, respectively.

Then, we add some perturbations to each component or some component of the image and then take inverse fourier transform to get the perturbed image. The perturbed image X' is constructed as:

$$\tilde{X}' = \mathcal{F}^{-1}\left(\text{clip}_{0,\infty}(M \otimes \delta_{\text{mag}}) \cdot e^{j(\phi+\delta_{\text{phase}})}\right) + \delta_{\text{pixel}}$$

$$X' = \text{clip}_{0,1}(\tilde{X}')$$

Attack Variants

Random Frequency Attack:

- Operates in the *frequency domain* as a **black-box** attack.
- For each channel c of the input $X \in \mathbb{R}^{B\times C\times H\times W}$:

$$X_F^c = \mathcal{F}\{X^c\} = M^c\, e^{j\Phi^c},$$

  where $M^c = |X_F^c|$ and $\Phi^c = \angle X_F^c$.
- Sample random noise

$$\Delta\Phi^c \sim \mathcal{N}(0, \epsilon^2), \quad \Delta M^c \sim \mathcal{N}(0, \epsilon^2).$$

- Perturb magnitude and phase:

$$M'^c = M^c + \Delta M^c, \quad \Phi'^c = \Phi^c + \Delta\Phi^c.$$

- Reconstruct and invert:

$$X'^c = \Re\Big\{\mathcal{F}^{-1}\big(M'^c\, e^{j\Phi'^c}\big)\Big\},$$

  then clip or normalize as needed and stack channels back into X'.

Attack Variants

Fourier Frequency Attack:

- Given original image X, compute its shifted FFT:

$$X_F = \mathcal{F}\{X\} \;=\; M\, e^{j\Phi} \quad \text{with } M = |X_F|,\ \Phi = \angle X_F.$$

- Introduce trainable perturbations

$$\Delta M,\ \Delta\Phi,\ \delta$$

on magnitude, phase, and pixels, so that

$$M' = M \odot \Delta M, \quad \Phi' = \Phi + \Delta\Phi, \quad X' = \Re\{\mathcal{F}^{-1}(M' e^{j\Phi'})\} + \delta.$$

- Optimize via gradient descent on the loss

$$\mathcal{L}(\Delta M, \Delta\Phi, \delta) = \lambda \, \|X' - X\|_2^2 \;-\; \mathrm{CE}\big(f(X'),\, y\big),$$

where λ trades off imperceptibility and attack strength.

- Update each perturbation by

$$\Delta M \;\leftarrow\; \Pi_{[1-\epsilon,\,1+\epsilon]}\big(\Delta M - \eta\, \nabla_{\Delta M}\mathcal{L}\big),$$

$$\Delta\Phi \;\leftarrow\; \Pi_{[-\epsilon,\,\epsilon]}\big(\Delta\Phi - \eta\, \nabla_{\Delta\Phi}\mathcal{L}\big);$$

Outline

Why would adversarial attacks affect the positional encoding layer?

- ViTs initially have a linear projection where the patches get projected to tokens.

Why would adversarial attacks affect the positional encoding layer?

- ViTs initially have a linear projection where the patches get projected to tokens.
- When we add any noise to any image patch P, the token T obtained after the linear projection layer is:

Why would adversarial attacks affect the positional encoding layer?

- ViTs initially have a linear projection where the patches get projected to tokens.
- When we add any noise to any image patch P, the token T obtained after the linear projection layer is:

$$T = W(P + N)$$

Where W is the projection matrix, and N is the noise added to the patch.

- After adding in the positional embedding, the token T_p can be written as:

$$T_p = WP + (WN + PE)$$

Where $(WN + PE)$, can be thought of as the noisy positional encoding which gets added to the token.

Why would adversarial attacks affect the positional encoding layer?

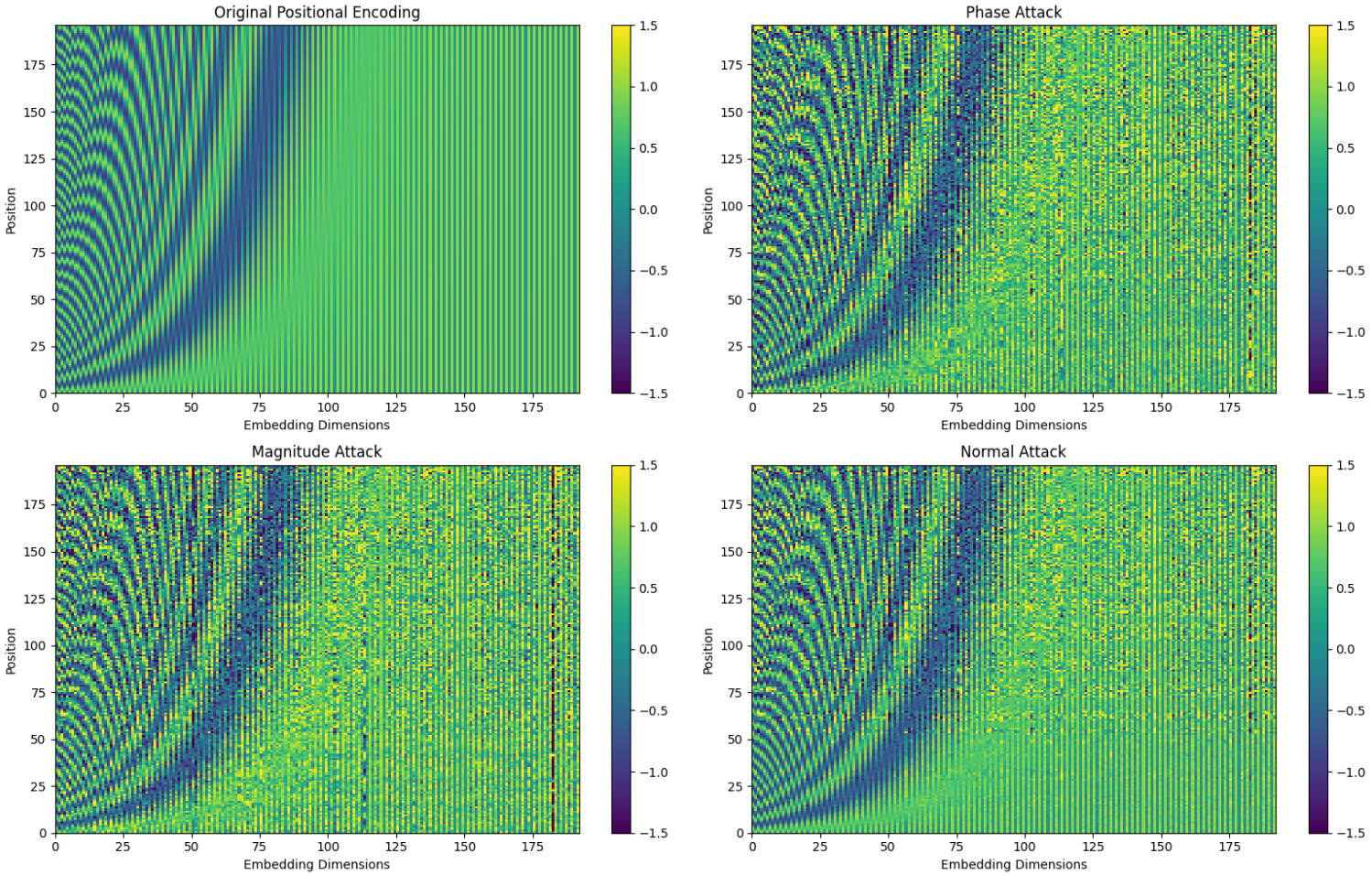


Figure: How adversarial attacks affect vision transformers

Current setup

- In the current existing setup you can make the following observations:

Current setup

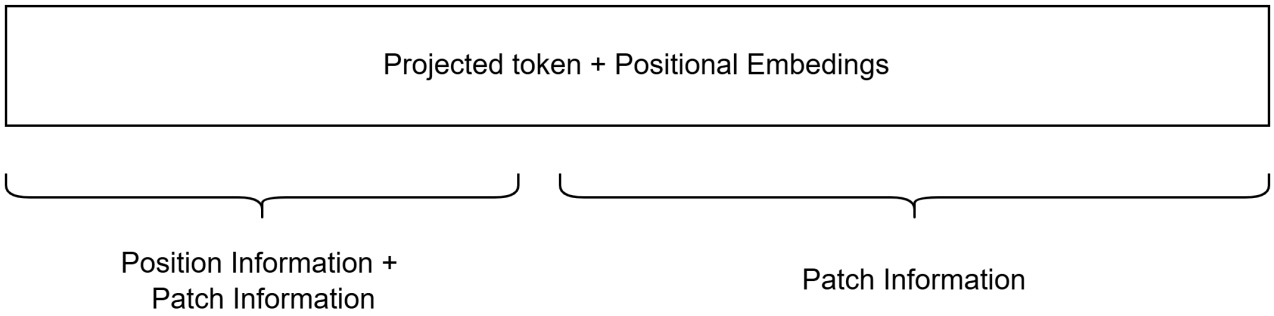
- In the current existing setup you can make the following observations:
  - Positional encodings for the higher dimensions don't change a lot.

Current setup

- In the current existing setup you can make the following observations:
  - Positional encodings for the higher dimensions don't change a lot.
  - So what the model learns is effectively the following:

Current setup

- In the current existing setup you can make the following observations:
  - Positional encodings for the higher dimensions don't change a lot.
  - So what the model learns is effectively the following:

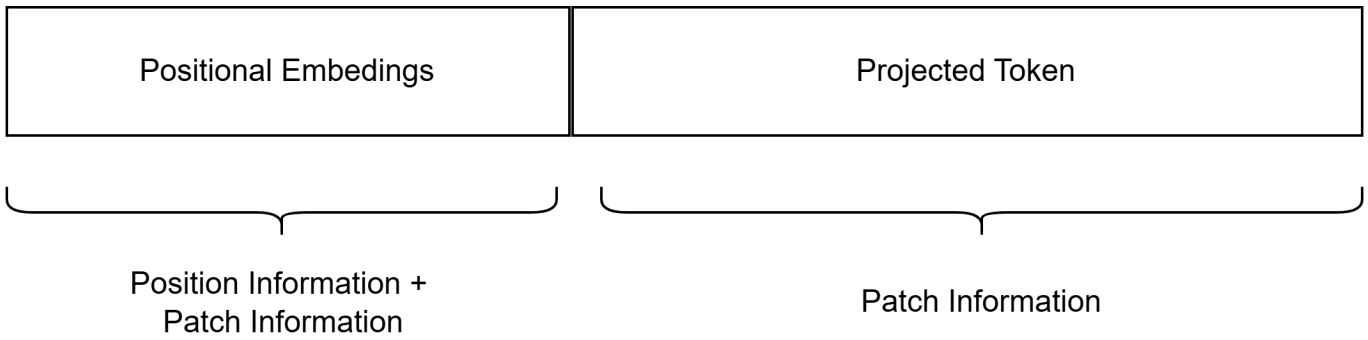


Our proposed PE

- So based on these observations we propose a simple modification to the normal sin-cos positional embeddings:

Our proposed PE

- So based on these observations we propose a simple modification to the normal sin-cos positional embeddings:



A few properties of our PE

- The original sin-cos positional embedding was chosen because it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}. Our positional embedding also follows the same.

A few properties of our PE

- The original sin-cos positional embedding was chosen because it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}. Our positional embedding also follows the same.
- If there are two patches that are far from each other in the image, their tokens will also move further apart after adding in the positional embeddings.

A few properties of our PE

- The original sin-cos positional embedding was chosen because it would allow the model to easily learn to attend by relative positions, since for any fixed offset k, PE_{pos+k} can be represented as a linear function of PE_{pos}. Our positional embedding also follows the same.
- If there are two patches that are far from each other in the image, their tokens will also move further apart after adding in the positional embeddings.
- Most importantly, the model will learn the positional information seperately, hence when we adversarially perturb the image the positional encoding is not affected.

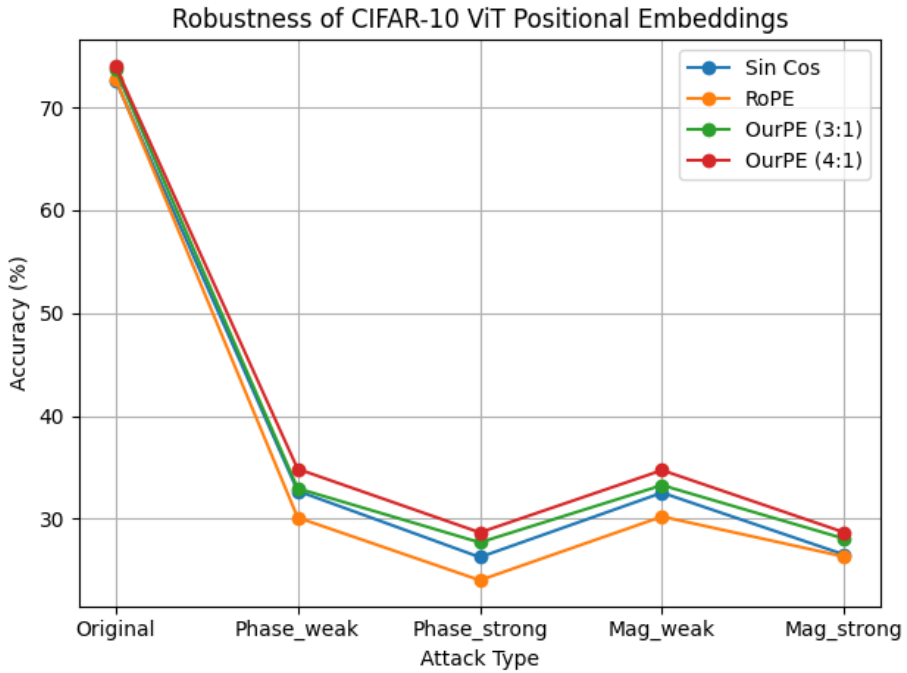
Outline

Robustness of Various Positional Encodings

Embedding	Original	Phase_weak	Phase_strong	Mag_weak	Mag_strong
Sin Cos	72.67%	32.71%	26.27%	32.54%	26.52%
RoPE	72.79%	30.12%	24.03%	30.22%	26.32%
OurPE (3:1)	73.70%	32.96%	27.70%	33.27%	28.07%
OurPE (4:1)	74.06%	34.85%	28.67%	34.73%	28.68%

Robustness of Various Positional Encodings

Embedding	Original	Phase_weak	Phase_strong	Mag_weak	Mag_strong
Sin Cos	72.67%	32.71%	26.27%	32.54%	26.52%
RoPE	72.79%	30.12%	24.03%	30.22%	26.32%
OurPE (3:1)	73.70%	32.96%	27.70%	33.27%	28.07%
OurPE (4:1)	74.06%	34.85%	28.67%	34.73%	28.68%



Results for our positional embedding

Table: Accuracy increase (per cent) of our embedding w.r.t. sinusoidal encodings under various attacks (averaged over 400 architectures) using the validation set of CIFAR-10. Training done for 10 epochs per architecture.

Ratio	FGSM	Fourier	Magnitude	Phase	Normal
0.000	−0.903	−1.190	0.638	0.728	1.107
0.125	0.064	−0.497	1.328	1.315	1.931
0.200	0.195	−0.241	1.163	1.111	1.720
0.250	0.145	−0.279	1.089	1.013	1.486
0.500	0.323	0.229	0.477	0.394	0.470

Results for our positional embedding

Table: Mean and standard deviation (per cent) of raw accuracy increase using our positional embeddings on validation set of CIFAR-10 [4] on training for 10 epochs. (across 400 architectures)

Ratio	Mean Accuracy Increase	Std. Dev.
0.000	−2.006	1.801
0.125	0.872	1.164
0.200	0.866	1.028
0.250	0.559	0.998
0.500	−0.477	1.301

Results for APE

Algebraic Positional Encodings (APE) [3] is a positional encoding that uses elements of groups to determine the position of a token/patch and has its theoretical basis in group theory. For an image, we have two groups representing the two axis of the image. Each element of a group is mapped to an orthogonal matrix which acts as a rotation matrix in a ROPE type setting. According to the paper, APE is practically equivalent to ROPE but with learnable degree of rotation optimized during training.

Results for APE

Description: Below are the results for a ViT architecture trained using three different Positional Embeddings (PE): Sinusoidal PE, Algebraic PE (APE), and our proposed PE. The models were trained on the CIFAR-10 dataset, and two types of attacks were performed: a frequency-based perturbation and a Fourier-based attack (a frequency based Projected Gradient Descent (PGD) attack). The table below summarizes the classification accuracies under each condition.

Base PE	**Before Perturbation**	**Frequency-Based Perturbation**		**Fourier Attack Perturbation**
		Phase Strong	**Magnitude Strong**	
Sinusoidal	86.52%	52.59%	52.55%	3.7%
Algebraic (APE)	88.33%	54.75%	54.32%	2.2%
Ours	88.04%	53.88%	53.68%	4.1%

Outline

Conclusion

- Positional encodings are a primary ViT vulnerability under spectral attacks.

Conclusion

- Positional encodings are a primary ViT vulnerability under spectral attacks.
- We propose a novel Positional Encoding where we concatenate the projection vectors with sinusoidal vector to get the final patch embedding instead of adding the two vectors

Conclusion

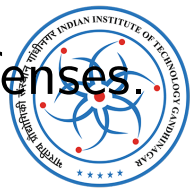
- Positional encodings are a primary ViT vulnerability under spectral attacks.
- We propose a novel Positional Encoding where we concatenate the projection vectors with sinusoidal vector to get the final patch embedding instead of adding the two vectors
- Our new positional Embeddings are slightly more adversarialy robust to black-box phase, magnitude and normal noise attacks than standard Positional Embeddings.

Conclusion

- Positional encodings are a primary ViT vulnerability under spectral attacks.
- We propose a novel Positional Encoding where we concatenate the projection vectors with sinusoidal vector to get the final patch embedding instead of adding the two vectors
- Our new positional Embeddings are slightly more adversarialy robust to black-box phase, magnitude and normal noise attacks than standard Positional Embeddings.
- Our embeddings are slightly more vulnerable to White-box attacks such as PGD and FGSM. Also there is a reduction in projection dimension of the patches

Conclusion

- Positional encodings are a primary ViT vulnerability under spectral attacks.
- We propose a novel Positional Encoding where we concatenate the projection vectors with sinusoidal vector to get the final patch embedding instead of adding the two vectors
- Our new positional Embeddings are slightly more adversarialy robust to black-box phase, magnitude and normal noise attacks than standard Positional Embeddings.
- Our embeddings are slightly more vulnerable to White-box attacks such as PGD and FGSM. Also there is a reduction in projection dimension of the patches
- Future directions: Exploring why APE is good against frequency attacks and how to imporve our PE accordingly and tailored defenses.



References I

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.
An image is worth 16x16 words: Transformers for image recognition at scale.
arXiv preprint arXiv:2010.11929, 2020.

G. Kim, J. Kim, and J.-S. Lee.
Exploring adversarial robustness of vision transformers in the spectral perspective.
In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3976–3985, 2024.

K. Kogkalidis, J.-P. Bernardy, and V. Garg.
Algebraic positional encodings, 2024.

References II

A. Krizhevsky.
Learning multiple layers of features from tiny images.
Technical Report, University of Toronto, 2009.