Exploring Non-additive Randomness on ViT against Query-Based Black-Box Attacks

Jindong Gu[1], Fangyun Wei[2], Philip Torr[1], Han Hu[2]

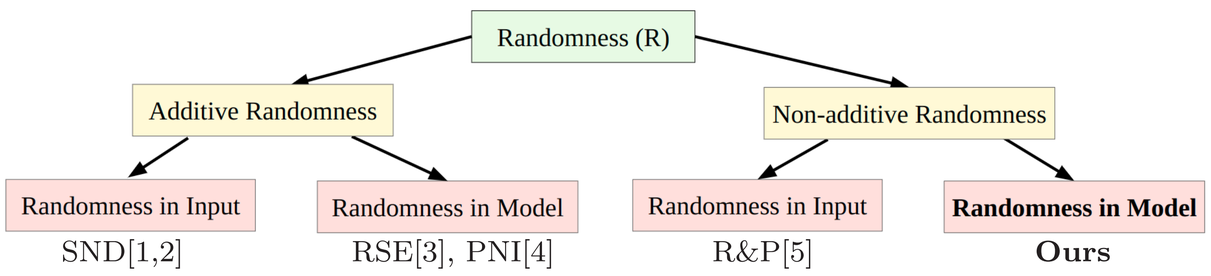
[1]Torr Vision Group, University of Oxford, [2] Microsoft Research Asia

1. Introduction

Query-based black-box attacks (QBBA) can is able to create the perturbations using model output probabilities of image queries requiring no access to the underlying models. Various types of randomness have been recently explored to defend against QBBA.

To better defend against QBBA, We propose to explore the non-additive randomness of Vision Transformers with the following motivations:
1) Transformer architectures dominate both CV and NLP communities; 2) The non-additive randomness on ViT is underexplored; 3) The transformer architecture is flexible.

2. Taxonomy of Stochastic Defense Strategy

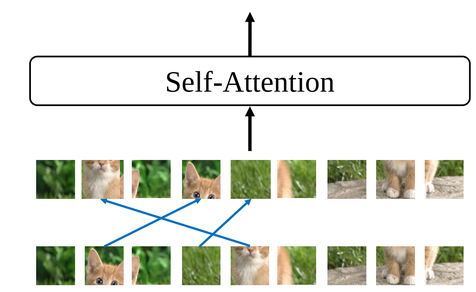


In our taxonomy, we categorize randomness into two types, namely, additive randomness and non-additive randomness. Each of them can be further integrated into the inputs or into the model.

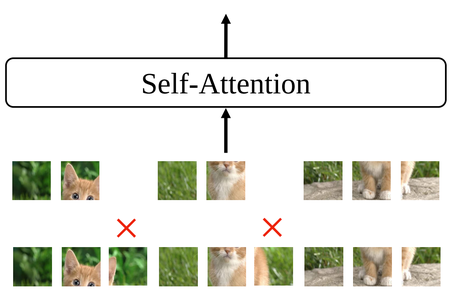
1) **Additive randomness** is often implemented by adding small noises to the input or the model network parameters.
2) **Non-additive randomness** integrates non-additive perturbation into the input or the model, e.g. affine transformation of the input image.

The non-additive randomness of the model has not been explored yet to defend against QBBA.

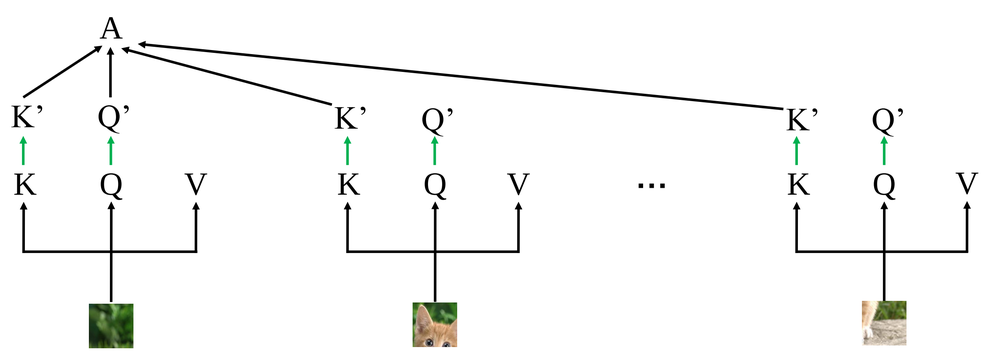
3. Non-additive Randomness in ViT for Defense



(a)Random Patch Permutation.



(b)Random Patch Drop.



(c)Randomness in Attention.

(a) Random Patch Permutation: To permutate patches, we first randomly sample some patches and permute their positional embedding.

(b)Random Patch Drop: The non-additive perturbation can be implemented by randomly sampling the elements of the input sequence.

(c)Randomness in Attention. Each dimension of the key and the query will be reduced with a probability. Only the kept dimensions are applied to compute self-attention.

4. Experimental Settings

Defense Methods. The following defense methods have been applied in our experiments:

1) **Small Noise Defense (SND)** [1] proposes to defend against QBBAs by adding random Gaussian noise to inputs.

2) **Parametric Noise Injection (PNI)** [4] proposes to add layer-wise trainable Gaussian noise to the activation or weight. Random Gaussian noises are added to activations of all layers without retraining to keep a fair comparison.

3) **Random Resizing and Padding (R&P)** [5] proposes a pre-processing-based defense method where the input images are randomly resized and padded.

4) **Patch Random Permutation (PRPerm)** Some patches are randomly sampled from input and permuted by permuting their positional embeddings.

5) **Patch Random Drop (PRDrop)** Some patches are randomly sampled and dropped from the input patch sequence.

6) **Patch Attention Perturbation (PAttnPert)** The key and the query with sampled dimension are applied to compute self-attention.

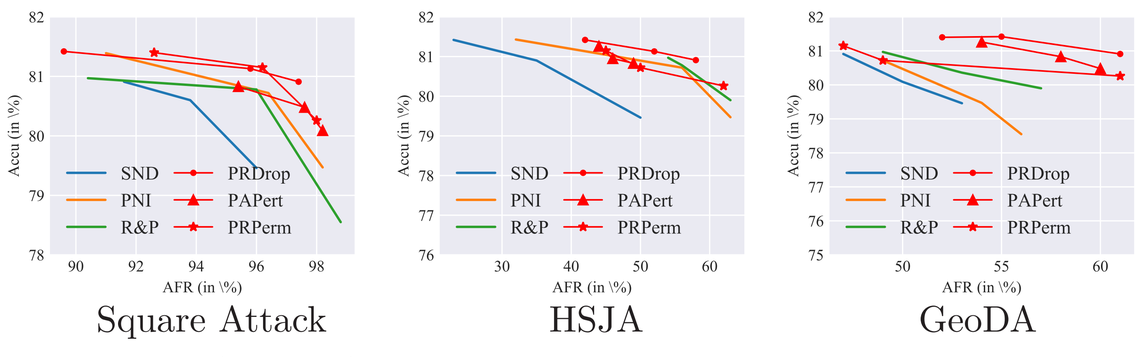
Attack Methods. We select popular QBBAs, such as Square Attack, Hop Skip Jump Attack(HSJA), and GeoDA. The defense against more attacks is also reported in the supplement.

Model. We use some representative ViT models pre-trained on the ImageNet dataset, such as ViT, DeiT, and SwinTransformer.

Evaluation Metrics. We report the clean accuracy (**Accu** in %) on the whole validation dataset and Attack Failure Rate (**AFR** in %) on the selected subset.

Dataset. The clean accuracy is reported on the whole Imagenet validation dataset. And 1k images randomly selected from the validation dataset are used to report Attack Failure Rate.

4.1 Experimental Results



Square Attack HSJA GeoDA

Our methods (redlines) can achieve a better trade-off between the attack failure rate and clean accuracy.

Models	Defense	Accu(%)	AFR(%)	Models	Defense	Accu(%)	AFR(%)
ResNet18	No	69.55	0.0	ResNet50	No	75.86	0.7
	SND[1]	69.21	43.1		SND[1]	75.79	36.2
	PNI[4]	69.45	38.0		PNI[4]	75.84	32.3
	R&P[5]	69.35	68.3		R&P[5]	75.22	78.5
ViT-tiny	No	75.48	0.9	ViT-small	No	81.40	1.2
	SND[1]	75.18	41.14		SND[1]	81.38	72.0
	PNI[4]	74.81	72.2		PNI[4]	81.40	33.5
	R&P[5]	74.19	63.2		R&P[5]	80.97	89.5
	PRDrop	75.09	70.0		PRDrop	81.39	90.6
	PAPert	74.55	72.1		PAPert	80.98	95.4
	PRPerm	75.48	64.2		PRPerm	81.40	92.6

Vit with the randomness-based defense achieves a better trade-off than the counterpart ResNet (e.g., ResNet50 vs. ViT-small).

6. Conclusion

We taxonomize the defensive randomness from the perspective of defense against query-based black-box attacks. Following our taxonomy, we propose non-additive randomness on ViT. Our experiments verify that our defense method can achieve better trade-offs between clean accuracy and AFR.