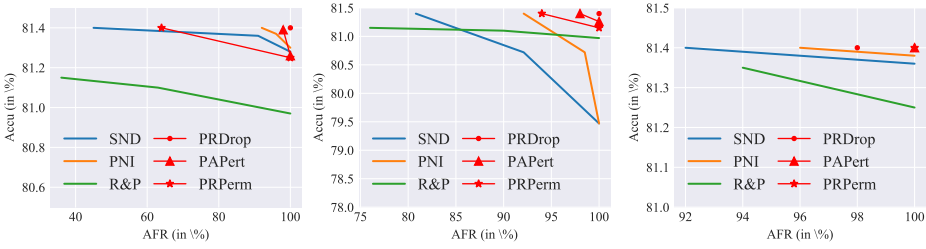
Exploring Non-additive Randomness on ViT against Query-Based Black-Box Attacks

Supplementary Material

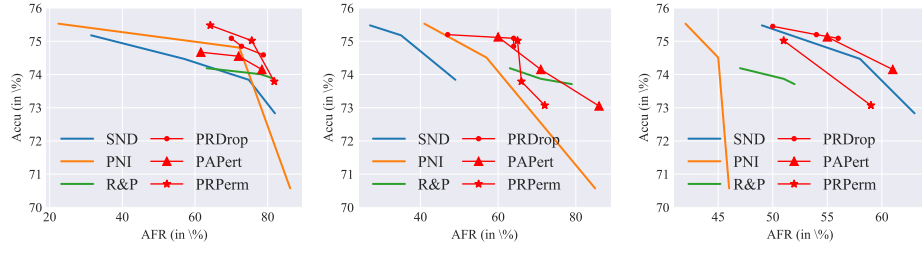
Paper ID 406

A: Experiments on More Query-based Black-box Attacks



(a) NES Attack (b) SimBA (c) Boundary

Fig. 1: Randomness-based Defense on ViT-small against Query-based Black-box Attacks. Besides SOTA attacks, we provide the results on more attack methods. In each subfigure, the x-axis and y-axis are the clean accuracy (Accu in %) and the attack failure rate (AFR in %), respectively. Each of the lines corresponds to a type of defense. The three red lines are our non-additive randomness defense on the model. Each point in the line corresponds to a trade-off point between Accu and AFR. Our methods can achieve a better trade-off than others.



(a) Square Attack (b) HSJA (c) GeoDA

Fig. 2: Randomness-based Defense on ViT-tiny against Query-based Black-box Attacks. We also provide the results on ViT models with different sizes. Our methods still achieve a better trade-off than others.

B: Hyper-parameters of Query-based Black-box Attacks

Table 1: Hyper-parameters of Square Attack [1]

Maximum number of iterations	100
Maximum perturbation	0.03
Initial fraction of elements	0.8
Number of restarts	1

Table 2: Hyper-parameters of HSJA [4]

Maximum number of iterations	20
Maximum number of evaluations for estimating gradient	10000
Initial number of evaluations for estimating gradient	100
Maximum number of trials for initial generation of AE	100

Table 3: Hyper-parameters of GeoDA [8]

Dimensionality of 2D frequency space (DCT)	20
Maximum number of iterations	4000
binary search tolerance	0.0001
Variance of the Gaussian perturbation	0.0002

Table 4: Hyper-parameters of NES Attack [7]

Maximum number of trials per iteration	1000
Maximum perturbation	0.05
Maximum number of evaluations for estimating gradient	100
Variance of the Gaussian perturbation	0.001

Table 5: Hyper-parameters of SimBA [5]

Norm of Perturbation	L2 norm
Maximum Perturbation	0.2
Dimensionality of 2D frequency space (DCT)	40
Maximum number of iterations	10000
ordering for coordinates	random

Table 6: Hyper-parameters of Boundary [2]

Initial step size for the orthogonal step	0.01
Initial step size for the step towards the target	0.01
Factor by which the step sizes are multiplied	0.667
Maximum number of iterations	5000
Maximum number of trials per iteration	25
Number of samples per trial	20
Maximum number of trials for initial generation of AE	100
Stop attack if perturbation is smaller than	0

C: Details and Setting of Defense Methods

The details and setting of the baselines and our approaches are shown as follows.

Small Noise Defense (SND). [3] proposes to defend against query-based black-box attacks by adding random Gaussian noise to inputs. The variance of the gaussian noise controls the strength of the noise, which balances the trade-off between robust accuracy and clean accuracy. We report the multiple results on different variance values, which are in $[0.001, 0.1]$.

Parametric Noise Injection (PNI). [6] proposes to add layer-wise trainable Gaussian noise to the activation or weight of each layer. In this work, we add random Gaussian noise to activations of all layers without retraining to keep a fair comparison. Only the inference stage is changed. Similarly, the variance of the Gaussian noise balances the trade-off between robustness and clean performance. The variance values we use are in $[0.001, 0.2]$.

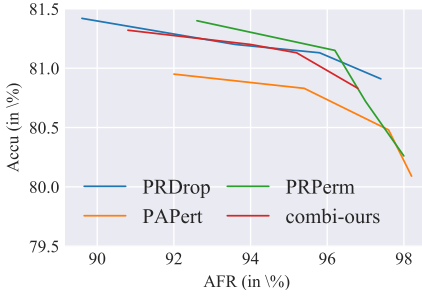
Random Resizing and Padding (R&P). [9] proposes a pre-processing-based defense method where the input is randomly resized and padded. The random resize and padding are different in different forward passes. The input image with a fixed size (i.e. 224×224) is resized to a smaller size (e.g. 208×208) and padded to the original size. The downsampled sizes control the trade-off between robust accuracy and clean accuracy, which are in $[160, 218]$.

Patch Random Permutation (PRPerm). In our PRP, the positional embedding of a certain percentage of patches is randomly permuted. The permutation of positional embedding is equivalent to the permutation of the corresponding input patches. The percentage of the permuted patches controls the trade-off. That too many patches are permuted leads to better robustness, but unsatisfying performance on clean inputs. Note that the permutation only happens before the first self-attention module where positional embedding is available. The percentage values we use are in $[1\%, 10\%]$.

Patch Random Drop (PRDrop). We propose to randomly drop the input patches of self-attention modules to mislead the query-based black-box attacks. The slimming of input patches only slightly decreases the model performance since there is redundant information in inputs, as shown in recent work. Note that the patch drop operation is different from the standard dropout operation where the dropped activations are set to zero. Ours removes part of the patches and keeps the same patches in the rest of the layers. The probability to drop patches controls the trade-off. The probability values we use are in $[1\%, 10\%]$.

Patch Attention Perturbation (PAttnPert). Besides the non-additive randomness in the inputs of self-attention module, we also propose to integrate the non-additive randomness in the Attention of self-attention module. Concretely, we propose to reduce the dimensions of keys and queries by randomly removing some of them. The keys and the queries of patches can still be used to compute the attention since only the dot product between them is required to describe their similarity. In our approach, we propose to randomly remove a certain percentage of dimensions of keys and queries. Such non-additive randomness also demonstrates the high effectiveness against query-based black-box attacks. Similarly, the percentage to remove dimension controls the trade-off.

D: Combination of Defense Strategies



(a) Combination of Our Defense

Fig. 3: Combination of Randomness-based Defense on ViTs. We also study the different combinations of the randomness-based defense methods. We show the combinations of our methods, i.e., three different types of randomness-based defense on models. The combination achieves the average defense effect.

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: ECCV (2020)
2. Brendel, W., Rauber, J., Bethge, M.: Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. arXiv preprint arXiv:1712.04248 (2017)
3. Byun, J., Go, H., Kim, C.: On the effectiveness of small input noise for defending against query-based black-box attacks. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 3051–3060 (2022)
4. Chen, J., Jordan, M.I., Wainwright, M.J.: Hopskipjumpattack: A query-efficient decision-based attack. In: 2020 ieee symposium on security and privacy (sp). pp. 1277–1294. IEEE (2020)
5. Guo, C., Gardner, J., You, Y., Wilson, A.G., Weinberger, K.: Simple black-box adversarial attacks. In: International Conference on Machine Learning. pp. 2484–2493. PMLR (2019)
6. He, Z., Rakin, A.S., Fan, D.: Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 588–597 (2019)
7. Ilyas, A., Engstrom, L., Athalye, A., Lin, J.: Black-box adversarial attacks with limited queries and information. In: International Conference on Machine Learning. pp. 2137–2146. PMLR (2018)
8. Rahmati, A., Moosavi-Dezfooli, S.M., Frossard, P., Dai, H.: Geoda: a geometric framework for black-box adversarial attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8446–8455 (2020)
9. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.: Mitigating adversarial effects through randomization. International Conference on Learning Representations (ICLR) (2017)