ARE VISION TRANSFORMERS MORE ROBUST THAN CNNS FOR BACKDOOR ATTACKS? - SUPPLEMENTARY

Anonymous authors
Paper under double-blind review

S1 APPENDIX

Localization efficiency: We measure the detection performance of the attention based method using the IoU (Intersection over Union) metric. We calculate the IoU betweeen the trigger and predicted block mask for different architectures. We observe from Table S1 that vision transformers clearly have a higher IoU compared to CNNs, hence leading to lower attack success rates. This experiment shows that Vision Transformers find it easier to localize the trigger for attacked images. The interpretation map is always calculated for the predicted category and results are averaged across 10 source-target pairs.

Model	IoU $\in$[0,1]
VGG16	0.19
ResNet18	0.07
ResNet50	0.039
ViT-Base	0.47
PatchConv	0.27
CaiT	0.66

Table S1: **IoU between predicted region and trigger-** IoU betweeen the trigger and predicted blocking mask is higher for vision transformers than CNNs.

Using different interpretation algorithms for CNNs: We also try different explanation algorithms for CNN architectures to ensure that our results are not biased towards a particular explanation method. The defense results for 3 explanation methods (Selvaraju et al., 2017; Srinivas & Fleuret, 2019; Wang et al., 2020) on ResNet18 architecture is shown in Table S2. Note that the 'Before Defense' results would be the same for all 3 rows, since we are evaluating the same model. We find that none of the 3 explanation methods can help with localizing the patch. This show that CNNs cannot localize the patch due to the architecture, rather than the explanation algorithms.

Method	Before Defense ASR (%)	After Defense ASR (%)
GradCAM (Selvaraju et al., 2017)	41.80	42.60
Score-CAM (Wang et al., 2020)	41.80	42.18
FullGrad (Srinivas & Fleuret, 2019)	41.80	43.20

Table S2: **CNNs with other explanations -** We try different explanation method for ResNet18 architecture and find that none of them can localize the patch correctly. Hence there is not much difference in ASR.

Source label recovery: We observe that due to the successful nature of the defense, once the trigger is blocked the original prediction of the source image is recovered as shown in Table S3. Different from the metric Source Accuracy on non-patched images, we calculate the Source Accuracy for patched images as the percentage of images that are classified as source, before and after defense.

Existing defenses: Furthermore, we evaluate test-time defense on both HTBA and BadNets. We use STRIP(Gao et al., 2019) as a SOTA run-time backdoor detection model. The idea behind the STRIP is to apply a perturbation on the input and measure the randomness (entropy) of the output prediction of the model. Less randomness (low entropy) in the final prediction indicates presence

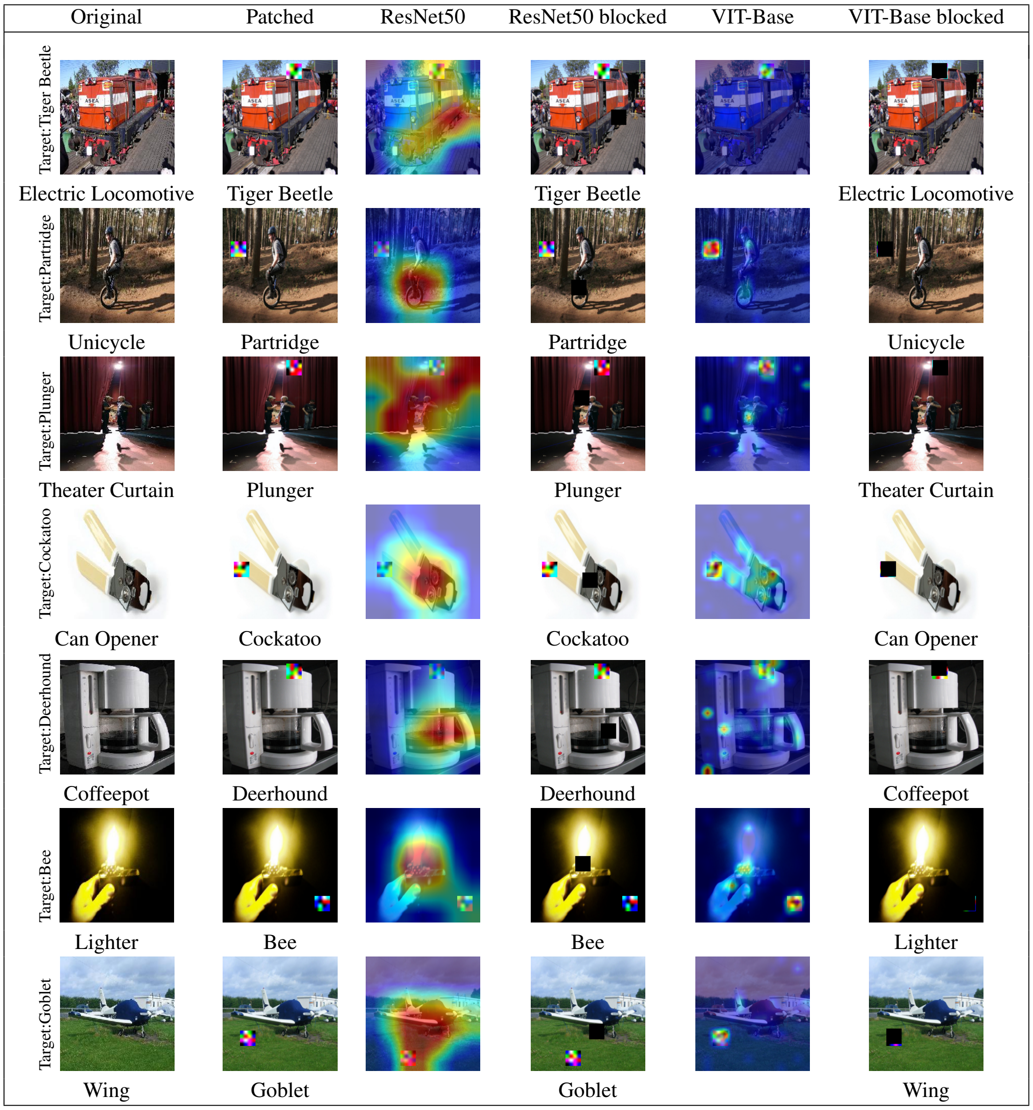

Figure S1: **Image Blocking Defense-** We show examples where blocking defense is performed for ResNet50 and ViT-Base. Transformers can successfully localize the patch, resulting in a successful defense. Results are not cherry picked and attack was successful for all examples.

	Before Defense	After Defense
Model	Source Accuracy (%) (Attacked Images)	Source Accuracy (%) (Attacked Images)
ViT-Base	21.40	66.00
PatchConv	44.80	67.00
CaiT	5.80	56.80

Table S3: **Effect on Source Accuracy:** We observe that the defense is able to improve the source accuracy significantly for vision transformers. We calculate the percentage of attacked images that were classified as source category, before and after defense. Qualitative examples can be found in Figure S1.

of a backdoor in the input. Following (Gao et al., 2019), for each sample at test time, we select 100 randomly chosen clean images to apply linear blending with the input. Next, we forward these perturbed examples and calculate the normalized entropy. For each architecture, we average the normalized entropy over all source samples and report average of 10 source/target pairs in Fig.S2.

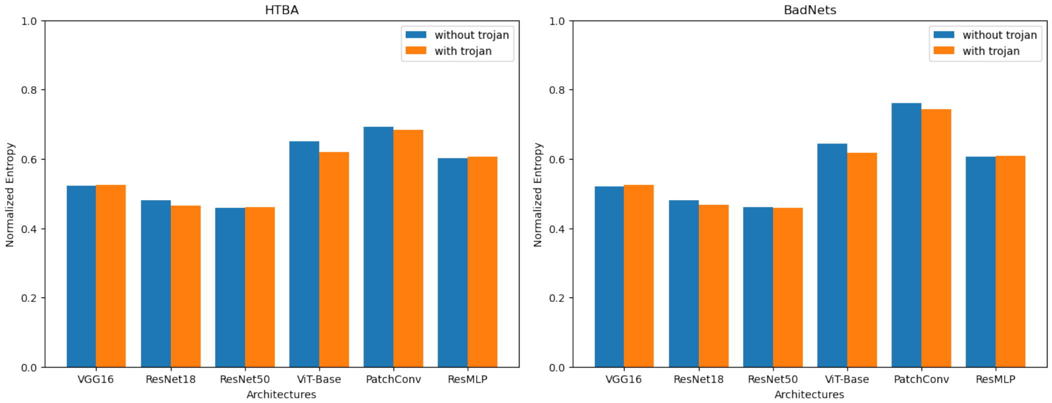

Figure S2: **Detecting Backdoor examples:** We find that an entropy based backdoor detection method is not very suitable to more diverse datasets such as ImageNet and for large architectures. We can see that the difference in entropy for the benign and trojan examples is not significant enough to create a heuristic based defense.

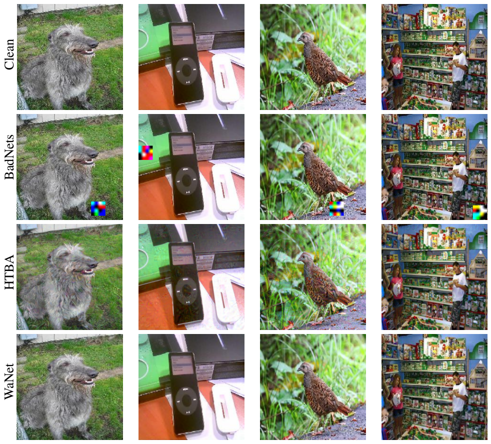

Figure S3: **Poison Images:** We show some comparisons between the poisons generated using different backdoor methods.

We observe that entropy is not a good indication of a trigger in large scale dataset like ImageNet. For example, in some architectures like ResMLP and ResNet, the difference between benign and trojan examples is not significant enough. Note that the STRIP algorithm is a detection based defense where the goal is to detect whether an input sample contains a backdoor or not.

Limitations: In our threat model, the defender makes an assumption about the range of sizes of trigger patches encountered during test time. We also observe that the test-time image blocking causes a drop in the accuracy of clean test images. Additionally, by doing test time image blocking defense, the inference time increases by factor of 2 since we need to forward twice per image.

We report the results for each pair of categories in Tables S1-S4. Please refer to the caption for details. Also, Figure S1 shows some qualitative visualization.

Moreover, we have included our code as part of supplementary material.

Table S4: **Results of Attack and Test time Defense-** To save in space in the main submission, we reported the results averaged over 10 random pairs of categories. In this table, we report the results for all pairs with ViT-Base architecture (similar to Table 1 of the main submission). The pairs of categories are the same random pairs used in HTBA [6]. Note that each pair of categories (each row) corresponds to a different attack task, so depending on the similarity of source and target categories, that attack may be easy or difficult. Hence, we do not expect a low standard deviation of ASR across these tasks. A similar large standard deviation was also reported in HTBA.

		Attack			Defense		
Source	Target	Val Accuracy (%)	Source Accuracy (%)	ASR (%)	Val Accuracy (%)	Source Accuracy (%)	ASR (%)
Slot	Australian Terrier	79.02	92.00	56.00	76.94	92.00	6.00
Lighter	Bee	79.06	66.00	58.00	76.95	70.00	28.00
Theater Curtain	Plunger	78.96	90.00	82.00	76.95	78.00	20.00
Unicycle	Partridge	79.04	92.00	70.00	76.99	70.00	14.00
Mountain Bike	Ipod	79.04	78.00	68.00	76.86	66.00	30.00
Coffeepot	Deerhound	79.04	64.00	52.00	76.93	66.00	16.00
Can Opener	Cuckatoo	79.00	72.00	32.00	76.90	70.00	12.00
Hotdog	Toyshop	79.02	90.00	60.00	76.90	80.00	22.00
Electric Locomotive	Tiger Beetle	79.04	88.00	84.00	76.99	92.00	6.00
Wing	Goblet	79.18	42.00	52.00	76.98	48.00	10.00
Average		79.04	77.4	61.4	76.94	73.2	16.4
Standard Deviation		0.05	16.5	15.40	0.04	13.10	8.47

Table S5: **Results of Attack and Test time Defense-** Similar to Table S4 for ResNet50 architecture.

		Attack			Defense		
Source	Target	Val Accuracy (%)	Source Accuracy (%)	ASR (%)	Val Accuracy (%)	Source Accuracy (%)	ASR (%)
Slot	Australian Terrier	74.06	92.00	6.00	63.83	92.00	10.00
Lighter	Bee	73.97	64.00	52.00	63.46	48.00	42.00
Theater Curtain	Plunger	73.92	76.00	52.00	63.5	70.00	42.00
Unicycle	Partridge	73.96	72.00	30.00	63.44	60.00	34.00
Mountain Bike	Ipod	73.89	74.00	42.00	63.49	38.00	62.00
Coffeepot	Deerhound	73.95	58.00	20.00	63.45	60.00	26.00
Can Opener	Cuckatoo	73.88	70.00	18.00	63.59	60.00	22.00
Hotdog	Toyshop	73.84	78.00	60.00	63.41	36.00	60.00
Electric Locomotive	Tiger Beetle	74.00	92.00	28.00	63.66	88.00	30.00
Wing	Goblet	73.90	64.00	40.00	63.55	54.00	44.00
Average		73.94	74.00	34.8	63.538	60.6	37.2
Standard Deviation		0.06	11.27	17.33	0.12	18.69	16.28

Table S6: **Results of Attack and Test time Defense-** Similar to Table S4 for ResNet18 architecture.

		Attack			Defense		
Source	Target	Val Accuracy (%)	Source Accuracy (%)	ASR (%)	Val Accuracy (%)	Source Accuracy (%)	ASR (%)
Slot	Australian Terrier	66.74	92.00	22.00	55.32	84.00	18.00
Lighter	Bee	66.84	52.00	34.00	55.44	56.00	30.00
Theater Curtain	Plunger	66.58	78.00	32.00	55.00	68.00	42.00
Unicycle	Partridge	66.53	70.00	46.00	55.43	46.00	42.00
Mountain Bike	Ipod	66.66	68.00	62.00	55.47	28.00	62.00
Coffeepot	Deerhound	66.57	52.00	36.00	55.57	54.00	34.00
Can Opener	Cuckatoo	66.75	58.00	42.00	55.64	48.00	42.00
Hotdog	Toyshop	66.67	70.00	42.00	55.16	48.00	64.00
Electric Locomotive	Tiger Beetle	66.81	82.00	48.00	55.43	80.00	46.00
Wing	Goblet	66.59	50.00	54.00	55.32	50.00	46.00
Average		66.67	67.2	41.80	55.37	56.2	42.60
Standard Deviation		0.11	14.18	11.53	0.18	16.85	13.73

Table S7: **Results of Attack and Test time Defense-** Similar to Table S4 for PatchConv architecture.

		Attack			Defense		
Source	Target	Val Accuracy (%)	Source Accuracy (%)	ASR (%)	Val Accuracy (%)	Source Accuracy (%)	ASR (%)
Slot	Australian Terrier	80.19	94.00	58.00	75.96	96.00	2.00
Lighter	Bee	80.67	84.00	64.00	76.31	70.00	24.00
Theater Curtain	Plunger	80.23	84.00	42.00	75.97	78.00	18.00
Unicycle	Partridge	80.25	88.00	32.00	75.97	76.00	16.00
Mountain Bike	Ipod	80.28	86.00	28.00	75.93	74.00	18.00
Coffeepot	Deerhound	80.19	68.00	34.00	76.11	66.00	8.00
Can Opener	Cuckatoo	80.19	82.00	6.00	76.04	80.00	2.00
Hotdog	Toyshop	80.22	92.00	18.00	75.92	90.00	36.00
Electric Locomotive	Tiger Beetle	80.16	88.00	80.00	75.95	88.00	4.00
Wing	Goblet	80.18	42.00	22.00	75.93	46.00	16.00
Average		80.26	80.8	38.4	76.00	76.40	14.40
Standard Deviation		0.15	15.35	22.82	0.12	14.13	10.78

REFERENCES

Yansong Gao, Change Xu, Derui Wang, Shiping Chen, Damith C Ranasinghe, and Surya Nepal. Strip: A defence against trojan attacks on deep neural networks. In *Proceedings of the 35th Annual Computer Security Applications Conference*, pp. 113–125, 2019. S1, S2

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017. S1

Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019. S1

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-cam: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pp. 24–25, 2020. S1