Neural Attention Distillation: Erasing Backdoor Triggers From Deep Neural Networks 2021

ICLR Yige Li, Xixiang Lyu

Contributions

1. Backdoor attacks can be notoriously perilous for several reasons.
2. First, backdoor data could infiltrate the model on numerous occasions including training models on data collected from unreliable sources or downloading pre-trained models from untrusted parties.
3. On top of that, once the backdoor triggers have been embedded into the target model, it is hard to completely eradicate their malicious effects by standard finetuning or neural pruning.

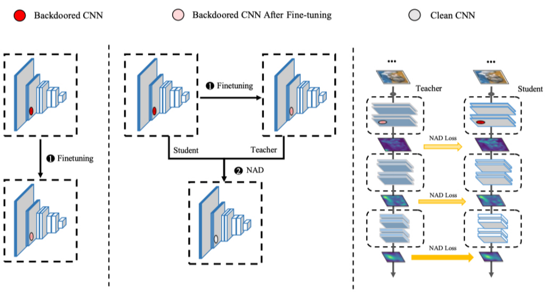
- Author propose a simple yet powerful backdoor defense approach called Neuron Attention Distillation (NAD). NAD is by far the most comprehensive and effective defense against a wide range of backdoor attacks.

- Author suggest that attention maps can be used as an intuitive way to evaluate the performance of backdoor defense mechanisms due to their ability to highlight backdoored regions in a network's topology.

- NAD utilizes a teacher network to guide the finetuning of a backdoored student network on a small subset of clean training data so that the intermediate-layer attention of the student network is well-aligned with that of the teacher network. The teacher network can be obtained from the backdoored student network via standard finetuning using the same clean subset of data. We empirically show that such an attention distillation step is far more effective in removing the network's attention on the trigger pattern in comparison to the standard finetuning or the neural pruning methods.

Method

- Defense Setting. We adopt a typical defense setting where the defender outsourced a backdoored model from an untrusted party and is assumed to have a small subset of clean training data to finetune the model. The goals of backdoor erasing are to erase the backdoor trigger from the model while retaining the performance of the mode l on clean samples.



(a) Erasing backdoor by Finetuning

(b) Erasing backdoor by our NAD

(c) A detailed structure of our NAD framework

- Attention Representation/Attention Distillation Loss/Overall Training Loss
  $A_{sum}(F^l) = \sum_{i=1}^{C} \left|F_i^L\right|; A_{sum}^p(F^l) = \sum_{i=1}^{C} \left|F_i^L\right|^p; A_{mean}^p(F^l) = \frac{1}{C}\sum_{i=1}^{C} \left|F_i^L\right|^p,$
  $L_{NAD}(F_T^l, F_S^l) = \| \frac{A(F_T^l)}{\|A(F_T^l)\|_2} - \frac{A(F_S^l)}{\|A(F_S^l)\|_2} \|_2$
  $L_{total} = \Xi_{(x,y)\tilde{D}}[L_{CE}(F_S(X), y) + \beta \cdot \sum_{l=1}^{K} L_{NAD}(F_T^l(x), F_S^l(x))]$

Experimental

- CIFAR-10/GTSRB with WideResNet

- 5% of the training data/10-epochs-SGD/batchsize-64/Data augmentation tech - random crop, horizontal flipping,cutout

- Different percentages of Clean Data

- T-S Combinations/Choice of a teacher/Effectiveness of Teacher Architecture

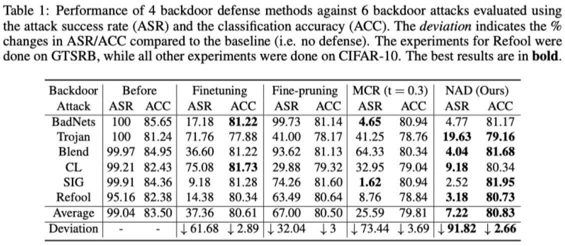


Table 1: Performance of 4 backdoor defense methods against 6 backdoor attacks evaluated using the attack success rate (ASR) and the classification accuracy (ACC). The *deviation* indicates the % changes in ASR/ACC compared to the baseline (i.e. no defense). The experiments for Refool were done on GTSRB, while all other experiments were done on CIFAR-10. The best results are in **bold**.

Backdoor	Before		Finetuning		Fine-pruning		MCR (t = 0.3)		NAD (Ours)	
Attack	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC	ASR	ACC
BadNets	100	85.65	17.18	**81.22**	99.73	81.14	**4.65**	80.94	4.77	81.17
Trojan	100	81.24	71.76	77.88	41.00	78.17	41.25	78.76	**19.63**	**79.16**
Blend	99.97	84.95	36.60	81.22	93.62	81.13	64.33	80.34	**4.04**	**81.68**
CL	99.21	82.43	75.08	**81.73**	29.88	79.32	32.95	79.04	**9.18**	80.34
SIG	99.91	84.36	9.18	81.28	74.26	81.60	**1.62**	80.94	2.52	**81.95**
Refool	95.16	82.38	14.38	80.34	63.49	80.64	8.76	78.84	**3.18**	**80.73**
Average	99.04	83.50	37.36	80.61	67.00	80.50	25.59	79.81	**7.22**	**80.83**
Deviation	-	-	↓ 61.68	↓ 2.89	↓ 32.04	↓ 3	↓ 73.44	↓ 3.69	↓ **91.82**	↓ **2.66**