Quarantine: Sparsity Can Uncover the Trojan Attack Trigger for Free

Tianlong Chen[1*], Zhenyu Zhang[1*], Yihua Zhang[2*], Shiyu Chang[3], Sijia Liu[2,4], Zhangyang Wang[1]

[1]University of Texas at Austin, [2]Michigan State University, [3]University of California, Santa Barbara, [4]MIT-IBM Watson AI Lab

CVPR JUNE 19-24 2022 NEW ORLEANS LOUISIANA

VITA GitHub

Motivations

How does the model sparsity relate to its train-time robustness against Trojan attacks?

Research Achievements At-A-Glance

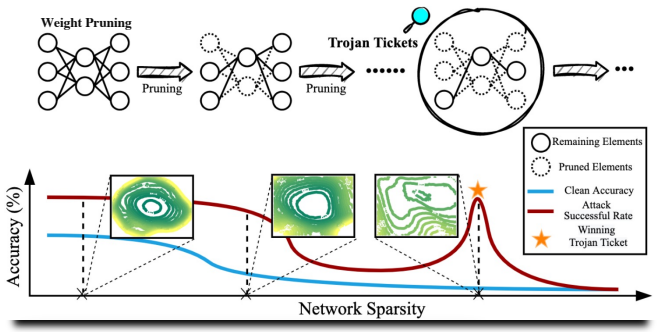

Figure 1. An overview of our proposal: Weight pruning identifies the 'winning Trojan ticket', which can be used for Trojan detection and recovery.

Contributions

- Trojan features learned by backdoored attacks are significantly more stable against pruning than benign features. Therefore, Trojan attacks can be uncovered through the pruning dynamics of the Trojan model.
- Leveraging LTH-oriented iterative magnitude pruning (IMP), the 'winning Trojan Ticket' can be discovered, which preserves the Trojan attack performance while retaining chance-level performance on clean inputs.
- The winning Trojan ticket can be detected by our proposed linear model connectivity (LMC)-based Trojan score.

Related Works

[1] Jonathan Frankle et al. "The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks." ICLR 2019.
[2] Jonathan Frankle et al. "Linear Mode Connectivity and the Lottery Ticket Hypothesis." ICML 2020.
[3] Ren Wang et al. "Practical detection of trojan neural networks: Data-limited and data-free cases." ECCV 2020.

Detecting Winning Trojan Tickets

- We adopt Linear Mode Connectivity[2] (LMC) to measure the stability of the Trojan ticket $\phi := (m \odot \theta)$ vs. the k-step finetuned Trojan ticket $\phi_k := (m \odot \theta^{(k)})$.
- We define the Trojan Score as

$$S_{Trojan} = \max_{\alpha \in [0,1]} \mathcal{E}(\alpha\phi - (1-\alpha)\phi_k) - \frac{\mathcal{E}(\phi) - \mathcal{E}(\phi_k)}{2},$$

where the first term denotes LMC and the second term an error baseline. $\mathcal{E}(\phi)$ denotes the training error of the model ϕ.
- A sparse network with the *peak* Trojan Score maintains the highest ASR (Figure 2) in the extreme pruning regime and is termed as the Winning Trojan Ticket.

Properties of Winning Trojan Tickets

- Backdoor features are well encoded in the winning Trojan ticket, which helps recover the Trojan trigger even without any access to clean training samples or threat model information.
- The winning Trojan ticket requires the *minimum perturbation* to reverse engineer the Trojan target label compared to the dense and various sparse network counterparts (Figure 3, Table 1). The trigger pattern recovered from the winning Trojan ticket yields a valid Trojan attack with a high ASR.
- The winning Trojan ticket can recover Trojan trigger using only 'noise image inputs', namely for 'free' (Table 2).

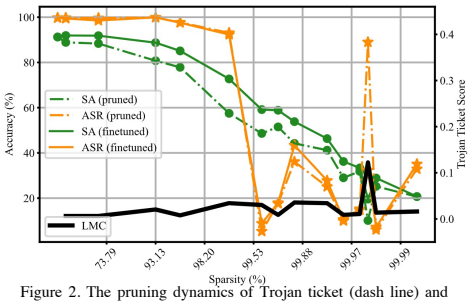

Figure 2. The pruning dynamics of Trojan ticket (dash line) and 10-step finetuned ticket (solid line) on CIFAR-10 with ResNet-20 and gray-scale backdoor trigger. For comparison, the Trojan score is also reported.

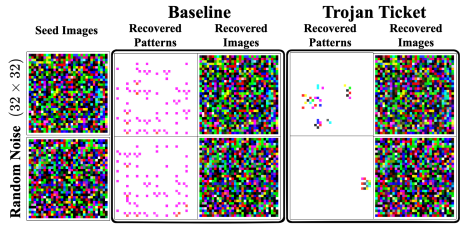

Figure 3. Visualization of recovered Trojan trigger patterns from dense Trojan models (baseline) and winning Trojan tickets. ResNet-20s on CIFAR-10 with RGB triggers are used. The first column shows the random seed images used for trigger recovery.

Table 1. Performance of recovered triggers with ResNet-20s on CIFAR-10 across diverse Trojan triggers, including gray-scale, RGB, and clean-label triggers. Green check/ Red cross means the detected label matches/mismatches the true target label.

Gray-scale Trigger	(Detected, ℓ_1)	ASR
Dense baseline [32]	("1", 196.8) ✓	71.4%
Winning Trojan ticket	("1", 68.0) ✓	**91.2%**
RGB Trigger	**(Detected, ℓ_1)**	**ASR**
Dense baseline [32]	("1", 78.7) ✓	48.0%
Winning Trojan ticket	("1", 29.8) ✓	**99.6%**
Clean-label Trigger	**(Detected, ℓ_1)**	**ASR**
Dense baseline [32]	("1", 48.6) ✓	9.6%
Winning Trojan ticket	("1", 14.0) ✓	**99.8%**

Table 2. Performance of recovered triggers with random noise images ('free') v.s. benign clean images. The RGB Trojan attack on CIFAR-10 and ResNet-20s are used for reverse engineering.

Noise Images ('Free')	(Detected, ℓ_1)	ASR
Dense baseline [32]	("1", 78.7) ✓	48.0%
Winning Trojan ticket	("1", 29.8) ✓	**99.6%**
Clean Images	**(Detected, ℓ_1)**	**ASR**
Dense baseline [32]	("1", 174.6) ✓	72.6%
Winning Trojan ticket	("1", 40.4) ✓	**99.8%**

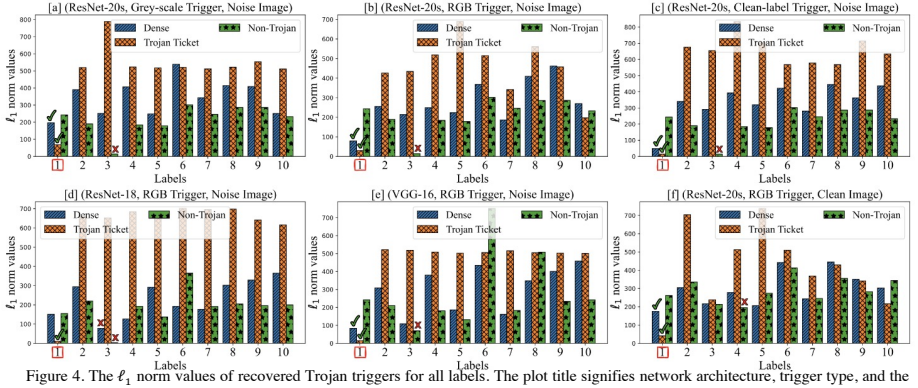

Figure 4. The ℓ_1 norm values of recovered Trojan triggers for all labels. The plot title signifies network architecture, trigger type, and the images for reverse engineering on CIFAR-10. Class "1" is the true target label for Trojan attacks. Green check or red cross indicates whether the detected label (with the least ℓ_1 norm matches the true target label).

Contact: {tianlong.chen, zhenyu.zhang, atlaswang}@utexas.edu, {zhan1908, liusiji5}@msu.edu, chang87@ucsb.edu