Backdoor Learning Curves: Explaining Backdoor Poisoning Beyond Influence Functions

A. E. Cinà †, K. Grosse ‡, S. Vascon †, A. Demontis ‡, B. Biggio‡, F. Roli ‡, M. Pelillo †

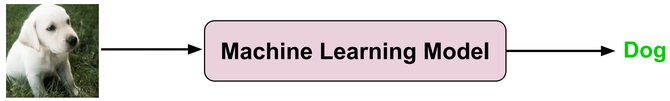
†European Center for Living Technology, Ca' Foscari University of Venice, Italy ‡Pattern Recognition and Applications Lab, University of Cagliari, Italy

{antonioemanuele.cina, sebastiano.vascon, pelillo} @unive.it {kathrin.grosse, ambra.demontis, battista.biggio, roli} @unica.it

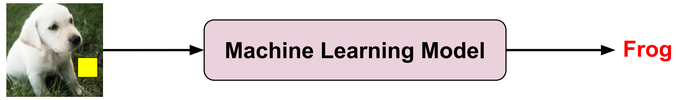
2021 ICCV OCTOBER 11-17 VIRTUAL

Backdoor Poisoning

Goal: Forcing a model to predict an attacker-chosen class when presented a specific **trigger** at test time, by injecting poisoning samples during training



The model predicts the correct class (dog).



The model predicts the attacker-chosen class (frog).

Open issue: which factors affect the success of backdoor poisoning?

Backdoor Learning Curves

Goal: to provide a framework that allows studying how much different factors impact the classifiers' ability to learn backdoors.

We introduce the notion of *backdoor learning curves,* which allows us to assess how much it is difficult for the classifier to learn a backdoor.

First, we formulate the problem of learning the clean training samples D_{tr} and the ones containing the backdoor $\mathcal{P}_{\text{tr}}$ as an incremental learning problem:

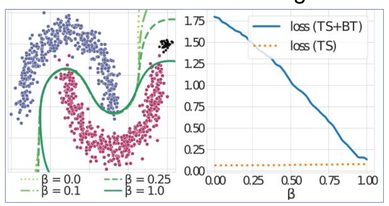
$$\mathbf{w}^{\star}(\beta) \in \arg\min_{\mathbf{w}} \; L(D_{\text{tr}}, \mathbf{w}) + \beta L(\mathcal{P}_{\text{tr}}, \mathbf{w})$$

where **w** are the classifier parameters, L is a loss function and β is a constant.

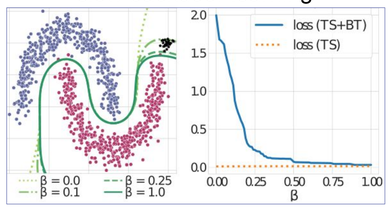
Then, we gradually increase β from 0 to 1, creating a curve that shows the behavior of the classifier on the test samples as a function of β.

The faster this curve decreases, the easier the target model is backdoored.

fast backdoor learning



slow backdoor learning



Backdoor Learning Slope

To quantify how fast a classifier can learn a backdoor, we define the *backdoor learning slope*, which measures how fast the classifier learns to classify the backdoor samples correctly.

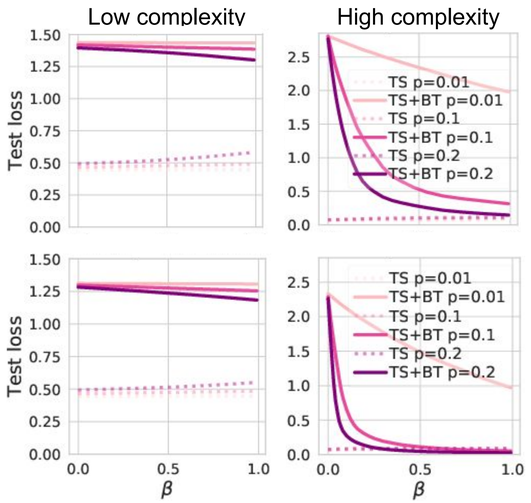
$$\theta = -\frac{2}{\pi}\arctan\left(\left.\frac{\partial L}{\partial \beta}\right|_{\beta=0}\right) \in [-1, 1]$$

It is the gradient of the backdoor learning curve at β=0, which corresponds to a measure called *influence function** that was previously proposed to assess the influence of training points on the classification output, rescaled it in [-1, 1].

Factors Influencing Backdoor Learning

Backdoor learning curves obtained for different percentage of backdoored samples injected in the training samples (p).

Dataset: CIFAR-10; Classifier: SVM with RBF kernel, γ=1e-04.
Low complexity: C=0,01 ; High complexity: C=10.
Trigger size = 8x8 (16x16) first (second) row.



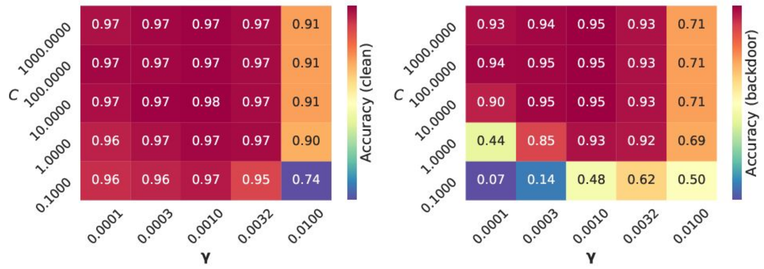
Factors affecting Backdoor Learning: (i) **number of backdoor samples** added to the training data, (ii) **complexity of the target classifier**, and (iii) **size of the trigger** - all increase the vulnerability to this attack.

We obtained similar results for other classifiers (SVM with linear kernel, Logistic Classifier and Ridge) and datasets (MNIST).

Backdoor Learning Slope for Hyperparameter Tuning

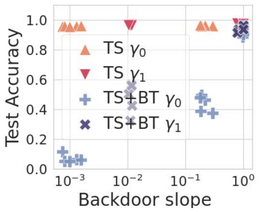
Depending on the target classifier, different hyperparameters can impact backdoor learning. For example, the complexity of RBF SVM is influenced not only by C but also by the kernel hyperparameter (γ).

The model's accuracy for different values of C and γ, when p=0.1.



These tables shows that low-complexity classifiers (corresponding to small C and γ) have good performance on clean samples, while being robust to backdoor poisoning, allowing us to choose convenient hyperparameters.

As shown by the plot on the right, the backdoor slope characterizes well the trend shown by the above matrices. Therefore, we think that backdoor slope may inspire more efficient methodologies for hyperparameter selection.



γ_0 = 1e-04

γ_1 = 1e-05

C ∈ [1e-03, 1e+04]

Conclusion and Future Work

We proposed a framework that allows studying the vulnerability to backdoors.

We identified three factors influencing backdoors' effectiveness, namely:

(i) the complexity of the target model
(ii) the fraction of backdoor samples in the training set
(iii) the size of the backdoor trigger

We have shown there exists a region of the hyperparameter space that leads to accurate and robust classifiers, which can be identified using the proposed backdoor slope.

We believe that this measure may inspire efficient methodologies for hyperparameter selection.

*Koh, P. W., and P. Liang. 'Understanding Black-Box Predictions via Influence Functions'. In International Conference on Machine Learning (ICML), 2017.

Acknowledgement: This work was supported by the PRIN 2017 RexLearn project (grant no. 2017TWNMH2), funded by the Italian Ministry of Education, University and Research; and by BMK, BMDW and the Province of Upper Austria in the within the COMET programme managed by FFG in the COMET S3AI module.