Why Do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks

Ambra Demontis[1], Marco Melis[1], Maura Pintor[1,3], Matthew Jagielski[2], Battista Biggio[1,3], Alina Oprea[2], Cristina Nita-Rotaru[2], Fabio Roli[1,3]

1. University of Cagliari, Italy 2. Northeastern University, Boston, MA, USA 3. Pluribus One, Italy - http://www.pluribus-one.it

Introduction

Motivation

Previous works [1, 2, 3] have shown empirical evidence that adversarial attacks can transfer between models, however there is little understanding on the underlying reasons of this phenomenon.

Goals

- Formal definition of transferability
- Investigation on test-time (evasion) and training-time (poisoning) attacks
- Definition of new metrics for understanding when and why adversarial attacks transfer

Evasion and Poisoning Attacks

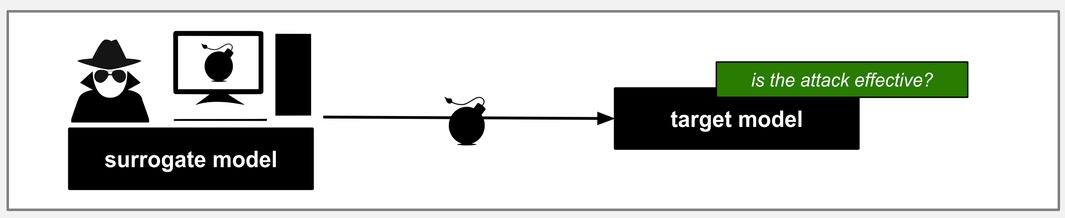
Evasion attacks [4]: small manipulations of testing data points that result in misprediction at testing time on those points

$$\begin{aligned} \max_{\mathbf{x}'} \quad & \ell\left(y, \mathbf{x}', \mathbf{w}\right) \\ \text{s.t.} \quad & \left\|\mathbf{x}' - \mathbf{x}\right\|_p \leq \varepsilon \\ & \mathbf{x}_{\text{lb}} \preceq \mathbf{x}' \preceq \mathbf{x}_{\text{ub}} \end{aligned}$$

Poisoning attacks [5]: the attacker controls a certain amount of training data, thus influencing the trained model and ultimately the predictions at test time on the testing set

$$\begin{aligned} \max_{\mathbf{x}'} \quad & L\left(\mathcal{D}_{\text{val}}, \mathbf{w}^\star\right) = \sum_{j=1}^{m} \ell\left(y_j, \mathbf{x}_j, \mathbf{w}^\star\right) \\ \text{s.t.} \quad & \mathbf{w}^\star \in \arg\min_{\mathbf{w}} \mathcal{L}\left(\mathcal{D}_{\text{tr}} \cup \left(\mathbf{x}', y\right), \mathbf{w}\right) \end{aligned}$$

Transfer loss



perturbation computed with the surrogate model

$$T = \underbrace{\ell(y, \boldsymbol{x} + \hat{\boldsymbol{\delta}}, \boldsymbol{w})}_{\text{loss attained by the target model on the adversarial point crafted against the surrogate}} \cong \ell(y, \boldsymbol{x}, \boldsymbol{w}) + \underbrace{\hat{\boldsymbol{\delta}}^T \nabla_{\boldsymbol{x}} \ell(y, \boldsymbol{x}, \boldsymbol{w})}_{\text{loss increment in the target given by the attack crafted with the surrogate model}}$$

Our work

$$S(\mathbf{x}, y) = \left\|\nabla_x \ell\right\|_2$$

S: size of input gradients
measures the vulnerability of the target model (white-box loss increment)

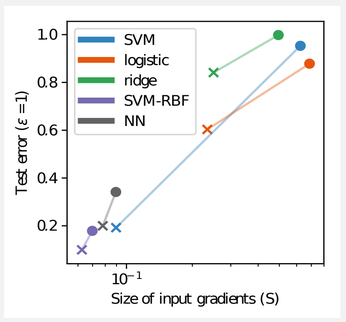
$$R(\mathbf{x}, y) = \frac{\nabla_x \hat{\ell}^T \nabla_x \ell}{\left\|\nabla_x \hat{\ell}\right\|_2 \left\|\nabla_x \ell\right\|_2}$$

R: gradient alignment
measures black-box to white-box loss increment ratio

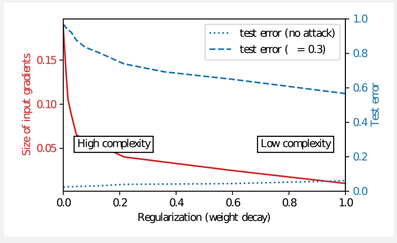
$$V(\mathbf{x}, y) = \mathbb{E}_{\mathcal{D}}\left\{\ell(y, \mathbf{x}, \hat{\mathbf{w}})^2\right\} - \mathbb{E}_{\mathcal{D}}\{\ell(y, \mathbf{x}, \hat{\mathbf{w}})\}^2$$

V: variability of the loss landscape
Evaluates the variability of the surrogate classifier under training data resampling

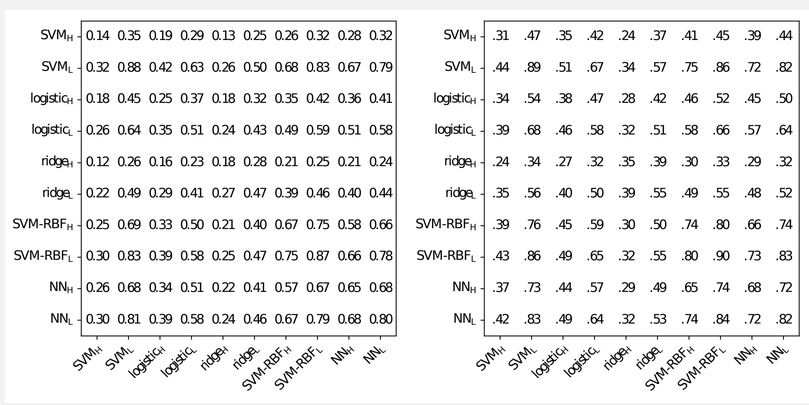
Experiments



Test error under attack vs. average size of input gradients (S) for low- (denoted with 'x') and high-complexity (denoted with 'o') classifiers

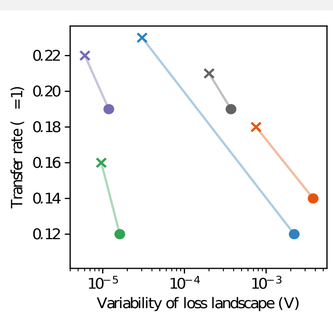


The size of the input gradients (S) is an indication of how much the target model is implicitly vulnerable against adversarial examples, including the ones crafted with the surrogate model.



Gradient alignment (R) and perturbation correlation (Pearson correlation coefficient between white-box and black-box perturbation) for evasion attacks on MNIST89. Rows indicate the surrogate classifier and columns indicate the target

Average transfer rate vs variability of the loss landscape (V). Low-complexity models are denoted with 'x' and high-complexity models are denoted with 'o'.



The main factors contributing to transferability are the intrinsic vulnerability of the target model and the complexity of the surrogate. The variability of the loss landscape of the surrogate under different training set resamplings can influence the stability of the solution of the optimization problem.

Key insights and future work

Evasion: decreasing complexity of the surrogate model by properly adjusting the hyperparameters of its learning algorithm provides adversarial examples that transfer better to a range of models

Poisoning: the best surrogates are generally models with similar levels of regularization as the target model

Future work: framework for understanding causes of success and failure of adversarial attacks against a model, comparing attacks with quantitative metrics in a competitive but fair environment

References and useful links

[1] Goodfellow et al., *Explaining and harnessing adversarial examples*, ICLR 2015.

[2] Szegedy et al., *Intriguing properties of neural networks*, ICLR 2014.

[3] Papernot et al., *Practical black-box attacks against machine learning*, ASIACCS 2017

[4] Biggio et al., *Evasion attacks against machine learning at test time*, ECML PKDD 2013.

[5] Biggio et al., *Poisoning attacks against support vector machines*, ICML 2012.

Code available at: https://gitlab.com/secml/secml

Paper: Demontis et al., *Why do Adversarial Attacks Transfer? Explaining Transferability of Evasion and Poisoning Attacks*, USENIX 2019

Download full text