CY 7790, Lecture 6: Evasion Attack Applications

Cem Topcuoglu and Nicholas Lunsford

September 28, 2021

Prior to this class we have discussed the taxonomy of adversarial machine learning attacks, and the basics of evasion attacks. We learned that if an attacker makes slight manipulations (perturbations) to input samples, he may force the target network to make an incorrect classification. Obviously, this could be used to meet a number of the attacker's needs or desires.

Today's class continued the discussion of evasion attacks, with one paper discussing a novel method of perturbation that both fools the target classifier and breaks defensive distillation, and another paper discussing at depth the transferability property found in adversarial examples. They are some of the most important papers in field, especially since the perturbation method proposed by Carlini and Wagner remains one of the strongest attacks to date, and the transferability property discussed by Papernot enables the use of black box attacks against other machine learning systems.

1 Calini and Wagner, Towards Evaluating the Robustness of Neural Networks

Problem Statement This research work considers the effectiveness the *defensive distillation* technique against adversarial examples, which at the time of writing, improved model robustness to nearly 100% despite the presence of adversarial examples. As such, this paper aims to find an adversarial examples that not only fools the target classifier but also defeats defensive distillation at the same time.

Figure 1: An illustration of the Carlini and Wagner attack on a defensively distilled network. All images with adversarial perturbations were incorrectly classified by the target model.

Threat Model The adversary in this research work has what would be considered *white-box* knowledge of the target model. In this case, it is unclear whether the adversary has access to the model's training data, however they do know the feature space, type of classifier, and model parameter.

The adversarial capability includes the ability to make unlimited to the input data and feature vectors. In the paper, the input data solely consists of imagery, and therefore the attacker's modifications consist entirely of changing the values of individual pixels.

Methodology The approach begins with the same basic formulation for adversarial examples as other methods, in which the goal is to find a small change δ to make to an input image x such that the image's classification changes but the result is still a valid image.

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) \\ \text{such that} \quad & C(x+\delta) = t \\ & x+\delta \in [0,1]^n \end{aligned}$$

Figure 2: The optimization function for finding an adversarial example.

In this problem, Δ represents some distance metric and C represents a cost function to optimize. The cost function is difficult to solve due to it being highly non-linear. As a result, the authors offer a number of different expressions that are better suited for the optimization problem and modify the optimization formulation.

$$\begin{aligned} \text{minimize} \quad & \mathcal{D}(x, x+\delta) + c \cdot f(x+\delta) \\ \text{such that} \quad & x+\delta \in [0,1]^n \end{aligned}$$

Figure 3: The optimization function for finding an adversarial example after making alterations for optimization.

$$\begin{aligned} f_1(x') &= -\text{loss}_{F,t}(x') + 1 \\ f_2(x') &= (\max_{i \neq t}(F(x')_i) - F(x')_t)^+ \\ f_3(x') &= \text{softplus}(\max_{i \neq t}(F(x')_i) - F(x')_t) - \log(2) \\ f_4(x') &= (0.5 - F(x')_t)^+ \\ f_5(x') &= -\log(2F(x')_t - 2) \\ f_6(x') &= (\max_{i \neq t}(Z(x')_i) - Z(x')_t)^+ \\ f_7(x') &= \text{softplus}(\max_{i \neq t}(Z(x')_i) - Z(x')_t) - \log(2) \end{aligned}$$

Figure 4: Objective functions defined for the constraint $C(x+\delta) = t$

The new formulation includes a new constant c, which must be greater than zero. This constant does not change the final result of the optimization problem and only scales the minimization function.

After choosing a value for c, the authors introduce a box constraint on δ: $0 <= x_i + \delta_i <= 1$ for all i to ensure that the modification (perturbation) yields a valid image. Using the Adam optimizer to solve for the new cost function and box constraint results in the successful adversarial examples.

Class Discussion

Semi-SoK Level Analysis Not only does this paper provide a novel technique for perturbing input samples and break defensive distillation, it also provides a deeper level of analysis for other state-of-the-art techniques at the time

of writing. This provides an insightful look as to why other methods fail against defensive distillation, and also serves as a reference point for evaluating the effectiveness of their proposed methodology.

New State of the Art Clearly, this method is shown to be superior to all other major perturbation techniques (especially since Carlini and Wagner go to great lengths to describe other perturbation methods). The method proposes continues to be the state of the art in terms of successful perturbation methods.

Highlights Importance of Upper Bound When thinking of adversarial machine learning in the context of defenses, this paper realizes the importance of establishing a strong upper bounds in defense evaluation. Any proposed defense must evaluate the highest standard of attacks (for the time) in order to have any credibility. In this case, the proposed method seems to be the highest standard to which a defense can be evaluated.

2 Papernot et al. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples

Problem Statement In this paper, the authors mainly focused on transferability in Machine learning for different algorithms. Their main purpose is to show that the transferability of the adversarial samples is not only effective on the NNs but other machine learning algorithms as well. They used the MNIST dataset to make experiments.

Threat Model In this threat model, the researchers defined two different threat models. For the first threat model, the architecture and parameters of the model are known. For the second threat model, the model is unknown to the adversary (i.e., Black Box). However, it can send queries to the oracle, which is realistic.

Methodology The authors defined the adversarial sample transferability as an adversary who is interested in producing an adversarial sample $\vec{x}^*$ misclassified in any class different from the class assigned by model f to legitimate input $\vec{x}$. To each adversarial sample transferability, the optimization problem that should be solved is depicted in Figure 5.

$$\vec{x^*} = \vec{x} + \delta_{\vec{x}} \ \texttt{where} \ \ \delta_{\vec{x}} = \arg\min_{\vec{z}} f(\vec{x} + \vec{z}) \neq f(\vec{x})$$

Figure 5: The optimization problem.

The researchers presented three different methods to assess transferability. The first method is intra-technique which measures the transferability between the same machine learning model. The second method is cross-techniques which the researchers assess the transferability between two different machine learning models. The third method is attacking the unknown machine learning classifiers by using Amazon and Google oracles. To improve their results on these oracles, they applied three different methods. The first method is Jacobian-based dataset augmentation. The second method is Periodic Step Size. The third method is Reservoir Sampling.

Class Discussion

Transferability attacks In this work, the authors surveyed all of the common ML models in the literature. Also, they added an ensemble model which showed that it is not resilient to transferability attacks. These ML techniques presented in Figure 6. The Logistic Regression (LR) is the only linear model. Researchers found that LR is not resilient against transferability. In the class discussion, we concluded that the probable reason for this is that LR is a linear model and hence getting a similar distribution every time. On the other hand, the kNN and Decision Tree (DT) are non-linear models that give them some resiliency against transferability attacks. These transferability results for intra and cross tranferability presented in Figure 8 and 7.

Amazon and Google oracles For the Amazon and Google oracles, they presented the number of query-misclassification rates and they reached high misclassification rates of 87-84% for 800 queries and 96-97% for 6400 queries when

ML Technique	Differentiable Model	Linear Model	Lazy Prediction
DNN	Yes	No	No
LR	Yes	Log-linear	No
SVM	No	No	No
DT	No	No	No
kNN	No	No	Yes
Ens	No	No	No

Figure 6: ML techniques that is studied in this work.

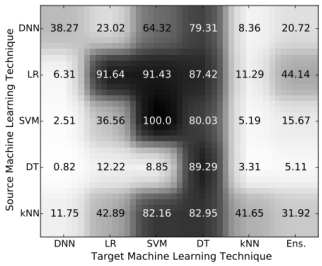

Figure 7: Cross-technique transferability matrix: $cell(i; j)$ is the percentage of adversarial samples crafted to mislead a classifier learned using machine learning technique i that are misclassified by a classifier trained with technique j.

they attacked a DNN model. When they applied the periodic step size (PSS) and reservoir sampling (RS), they got 95-91% for only 2000 queries. We also discussed that Jacobian-based dataset augmentation might create an overfit since they just trained on the MNIST dataset.

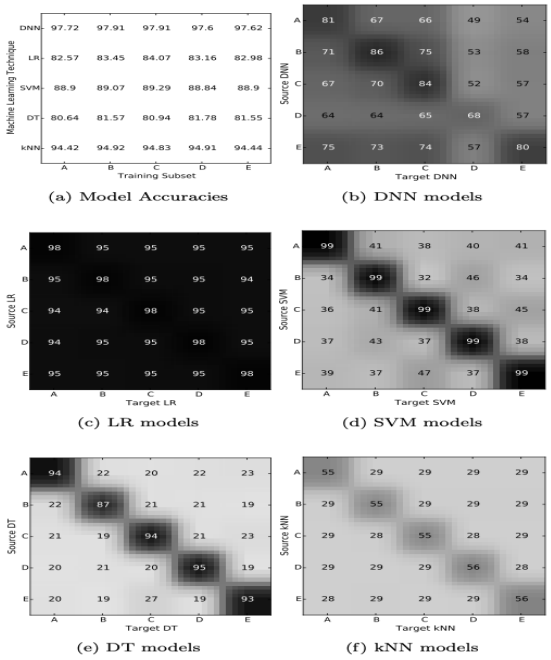

Figure 8: Intra-technique transferability for 5 ML techniques.

The MNIST dataset Although the transferability of these models has been shown in this work, the dataset which is used is a relatively easy dataset (i.e., the MNIST). Hence, it is important to observe these properties with other datasets as well. Note that at the time of the publication, these kind of attacks were very new.

Conclusion In conclusion, this paper showed us we can easily attack a black-box oracle system since there is a transferability property.