Adversarial Attacks on Image Recognition

Abstract

This project extends the work done by Papernot et al. in (Papernot et al., 2016a) on adversarial attacks in image recognition. We investigated whether a reduction in feature dimensionality using principle component analysis (PCA) can maintain a comparable level of misclassification success while increasing computational efficiency. We attacked black-box image classifiers trained on the MNIST dataset by forcing the oracle to misclassify images that were modified with small perturbations. The method we used was two-fold: the target classifier was imitated with a substitute logistic regression model and then the adversarial samples were generated off of the substitute model (Papernot et al., 2016a). The results show that reasonable misclassification rates with reduced computation time can be achieved for a PCA-reduced feature set utilizing the Papernot adversarial crafting algorithm.

1. INTRODUCTION

Machine learning techniques, coupled with data, are used to solve a multitude of high-dimensional problems with great success, such as those in the area of image recognition. For instance, image recognition is employed in self-driving cars to navigate on the roads. However, research has shown that these machine learning models are not robust to adversarial attacks and can be exploited by injecting specifically designed samples to training data or by creating test samples based on the decision boundary of the algorithm to misguide the classification result. For example, Papernot et. al. showed that it is possible to craft an image that would appear to be a stop sign but would be classified as a yield sign by some class of deep neural networks (Papernot et al., 2016b). Furthermore, Papernot et. al. also found that the perturbation technique they used to construct such adversarial samples is applicable to a variety of other classifiers, such as support vector machines and logistic regression (Papernot et al., 2016a). These findings demonstrated that machine learning systems are susceptible to malicious attacks. One such example would be to alter the image of road signs received by autonomous driving systems in order to manipulate the behaviour of target vehicles, which could lead to dire consequences. Thus, understanding the vulnerabilities of machine learning systems and the methods to exploit them is crucial for application of machine learning in practical settings.

2. LITERATURE

Papernot et al. in (Papernot et al., 2016a) described the following two step approach during adversarial sample creation for a black-box machine learning algorithm referred to as the 'oracle' from this point forward:

1. Train a substitute model utilizing as few calls to the oracle as possible.
2. Craft adversarial samples using either the fast gradient sign (FGS) method (Papernot et al., 2016a) or the Papernot method (Grosse et al., 2016).

In (Papernot et al., 2016a), the logistic regression (LR) and deep neural network (DNN) substitute models had the highest cross-technique transferability, indicating that adversarial techniques crafted by these models would be misclassified by oracles with a different machine learning algorithm structure (e.g. SVM, k-nearest neighbours, etc.) with a high success rate. This means that given a black-box oracle, a choice of LR or DNN substitute model should create effective adversarial samples. These substitute models must be trained on datasets obtained by querying the oracle. Papernot et al. started with a small training set, and utilized Jacobian-based dataset augmentation to increase the number of samples by querying the oracle on the datapoints that exhibit the most change. This method is described by the following formula (Papernot et al., 2016a):

$$S_{\rho+1} = \{\vec{x} + \lambda_\rho \mathbf{sgn}(J_f[O[(\vec{x})] : \vec{x} \in S_\rho)\} \cup S_\rho \quad (1)$$

where S is the training set, $\vec{x}$ is a sample in S, $O(\vec{x})$ is the label given to sample $\vec{x}$ by the oracle, J_f is the Jacobian matrix of the substitute model f, and λ is the tuneable step-size parameter. At each iteration ρ, the training set is augmented by utilizing Equation 1. The oracle is then called to obtain the labels for the new training dataset, and subsequently a new substitute model f is trained. Furthermore, in (Papernot et al., 2016a), the periodic step size (PSS) technique was introduced to improve the approximation of the oracle with the substitute model by multiplying the λ parameter by -1 when the Jacobian augmentation method no longer lead to a significant improvement in the substitute model. Then, λ_ρ is defined as

$$\lambda_\rho = \lambda(-1)^{\lfloor \frac{\rho}{\tau} \rfloor} \quad (2)$$

where τ is the number of iterations after which the Jacobian augmentation method is no longer effective. However,

the oracle should not be queried excessively to avoid raising suspicion. To diminish the calls to the oracle, reservoir sampling (RS) was utilized. Reservoir sampling selects κ randomly generated new samples after σ iterations have been completed normally. This decreases the number of calls to the oracle from $n(2)^\rho$ to $n(2)^\sigma + \kappa(\rho - \sigma)$ (Papernot et al., 2016a). Papernot et al. found that a Jacobian augmentation method combined with PSS and RS produced substitute models that approximated the oracle model successfully.

The purpose of this project is to extend the work done by Papernot et al. in (Papernot et al., 2016a) on adversarial attacks in image recognition. We investigated whether a reduction in feature dimensionality during adversarial sample crafting improved computational efficiency, while maintaining a comparable level of success in misclassification of the adversarial samples. We formed an attack on an oracle with a training set unknown to the substitute model by forcing the oracle to misclassify images that were modified with white noise undetectable to humans.

3. DATASET

Since our work extends that done by Papernot et al. in (Papernot et al., 2016b), (Papernot et al., 2016a), we utilize the same dataset cited in his papers, which is the MNIST dataset. This ensures the validity of our results. The MNIST hand-written digit dataset of 28×28 pixel images contains $50,000$ training, $10,000$ validation, and $10,000$ test greyscale images (et. al., 1998).

References

et. al., Y. LeCun. The MNIST Database, 1998.

Grosse, K., Papernot, N., Manoharan, P., Backes, M., and McDaniel, P. D. Adversarial perturbations against deep neural networks for malware classification. *CoRR*, abs/1606.04435, 2016. URL http://arxiv.org/abs/1606.04435.

Papernot, N., McDaniel, P. D., and Goodfellow, I. J. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *CoRR*, abs/1605.07277, 2016a. URL http://arxiv.org/abs/1605.07277.

Papernot, N., McDaniel, P. D., Goodfellow, I. J., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against deep learning systems using adversarial examples. *CoRR*, abs/1602.02697, 2016b. URL http://arxiv.org/abs/1602.02697.