Adversarial Attacks on Image Recognition*

Masha Itkina[1] and Yu Wu[2]

I. 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 [4]. 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 [5]. These findings demonstrated that machine learning systems are susceptible 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 the application of machine learning in practical settings.

II. LITERATURE

Papernot et al. in [5] followed 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 the fast gradient sign method [5].

In [5], 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 will 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. start with a small training set, and utilize Jacobian-based dataset augmentation to increase the number of samples by querying the oracle on the datapoints that exhibit the most change. This method follows the formula [5]:

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

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 [5], 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 leads to a significant improvement in the substitute model. Then, λ_ρ is defined as

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

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) is 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 $n(2)^\sigma + \kappa(\rho - \sigma)$ [5]. Papernot et al. found that a Jacobian augmentation method combined with PSS and RS produces substitute models that approximate the oracle model successfully.

The purpose of this project is to extend the work done by Papernot et al. in [5] on adversarial attacks in image recognition. We investigate whether a reduction in feature dimensionality during substitute model training can improve computation efficiency, while maintaining a comparable level of success in misclassification of the adversarial samples. We attempt to form an attack on a black box model with an unknown training set by forcing the oracle to misclassify images that are modified with white noise undetectable to humans.

*This work was done as part of a final project for the CS229 course taught by Professor Andrew Ng and Professor John Duchi at Stanford University.

[1]M. Itkina is from the Aeronautics and Astronautics Department, Stanford University, 450 Serra Mall, Stanford, CA, 94305, USA mitkina@stanford.edu

[2]Y. Wu is with the Institute of Computational and Mathematical Engineering, Stanford University, Stanford, CA, 94305, USA wuyu8@stanford.edu

III. DATASET

We utilize two datasets: the MNIST hand-written digit dataset [2] and the German Traffic Sign Recognition Benchmark (GTSRB) dataset [1]. The former contains $50,000$ training images and $10,000$ test greyscale images. The latter contains over $50,000$ images of traffic signs with over 40 classes. Both of these datasets were utilized by Papernot et al. in [4], [5].

The MNIST dataset is used to replicate the results of transferability between machine learning algorithms in crafting adversarial samples outlined in [5]. This is done in order to ensure that the experiment performed in this paper is comparable to the work done in [5]. Then image feature reduction will be performed to investigate the effect of reduced dimensionality when training the substitute model on the success of adversarial samples in by-passing the oracle.

If the above approach is successful, this method will be applied to the GTSRB dataset to misclassify coloured images utilizing a substitute model with reduced feature dimensionality. This experiment will portray the inherent vulnerability of machine learning algorithms to adversarial attacks in such high-risk applications as autonomous car navigation (e.g. reading a yield sign as a stop sign [4]).

IV. METHODS

A. Black-Box Models

To investigate the effectiveness of Papernot's approach outlined in Section II with image feature reduction, a set of three black-box models were selected to act as the oracle for comparison. We chose the logistic regression (LR), k-nearest neighbours (kNN), and support vector machine (SVM) models due to simplicity of implementation and Papernot's use of these models in [5]. The models were trained on the MNIST $50,000$ image sample set, and tested on $10,000$ samples in the test set. The performance of each of these oracle models is shown in Table I. All models achieve a success rate of $\sim 90\%$, deeming them sufficiently accurate to utilize as the black-box oracles in experiment.

LR	SVM	kNN
88.9	97.3	93.9

TABLE I: Percentage of the test set that each model classified correctly.

B. Logistic Regression Substitute Model

An LR substitute model was chosen for this experiment due to its high cross transferability to other models [5]. This model was trained as described in Section II utilizing Jacobian-based augmentation combined with PSS and RS. However, we found the Jacobian used in [5] to be incorrect. Instead, for an LR model f described by the equation as in [5]

$$f : \vec{x} \rightarrow \left[\frac{e^{\vec{w}_j \vec{x}}}{\sum_{l=1}^{N} e^{\vec{w}_l \vec{x}}} \right] \tag{3}$$

the following Jacobian was used

$$J_f(\vec{x})[i,j] = \frac{\vec{w_j} e^{\vec{w}_j \vec{x}} \sum_{l=1}^{N} e^{\vec{w}_l[i]\vec{x}} - e^{\vec{w}_j \vec{x}} \sum_{l=1}^{N} \vec{w}_l[i] e^{\vec{w}_l[i]\vec{x}}}{\left(\sum_{l=1}^{N} e^{\vec{w}_l[i]\vec{x}} \right)^2} \tag{4}$$

where $N = 10$ classes for the MNIST dataset, $\vec{w}$ is the matrix of parameters for the LR substitute model, $\vec{x}$ is a sample in the substitute model's training set. The Jacobian matrix for each sample is of dimension 784×10 where 28×28 pixels in each image equals 784 features.

C. Generating Adversarial Samples

To generate adversarial samples for the obtained substitute LR model, we use the fast gradient sign method described in [6]:

$$\vec{x}_{adversary} = \vec{x} + \epsilon \ \mathbf{sgn}(\nabla_{\vec{x}} f) \tag{5}$$

where the direction of the disturbance is the sign of the gradient of the probability function f described in Equation 3. This method is good for preliminary tests because its implementation is simple and very efficient to execute. The tuning parameter ϵ controls the size of the deviations of the adversarial samples from their origin.

V. RESULTS

A. LR Substitute Model Performance

As in [5], the LR substitute model begins with a training set of 100 samples from the MNIST test set with labels obtained from each of the three black-box oracles. Since our Jacobian formulation is different from that in [5], we had to make small adjustments to the parameters in the PSS algorithm in order to achieve comparable results. An optimal value of $\tau = 1$ was chosen to improve the Jacobian-based augmentation method for the substitute model's training set. The parameters utilized in our LR model versus that in [5] are summarized in Table II.

	λ	κ	τ	σ	ρ
Our approach	0.1	400	1	3	9
Papernot et al.	0.1	400	3	3	9

TABLE II: Comparison of the parameters used for PSS and RS methods during Jacobian-based training set augmentation.

The success of the LR substitute model at approximating the LR, kNN, and SVM oracles is summarized in Figure I where the percentage of samples for which the substitute model's and oracle's classifications agree is plotted against the iteration of the training set augmentation. These results match those in [5] quite well. The success rate increases steadily until it plateaus at the third iteration, after which RS

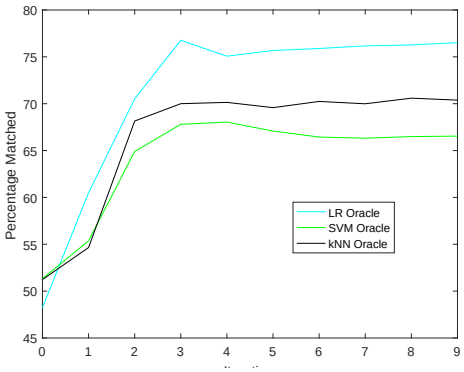

Fig. 1: Percentage of samples for which the substitute model and oracle classifications agree.

is activated. As expected, the LR substitute model does the best on an LR model oracle; nevertheless, it performs well even on the SVM and kNN oracles as portrayed in Figure I. The algorithm could be truncated at the third iteration since past this point it has nearly reached convergence, saving computation time.

B. Performance of Crafted Adversarial Samples

For generating adversarial samples, we used $\epsilon = 0.3$ in Equation 5, which is the same value used in [5]. We generated adversarial samples on all $10,000$ test samples based on the LR substitute model obtained in Subsection V-A. The missclassification rates for our adversarial samples on the original oracles are given in Table III below.

LR	SVM	kNN
96.0	99.3	16.5

TABLE III: Percentage of the adversarial test set that each oracle classified incorrectly

We are able to achieve a fairly high missclassification rate for both the LR oracle and the SVM oracle, but performed poorly for the kNN oracle, which agrees with the results obtained by Papernot et al. [5]. Our missclassification is in fact slightly higher for all 3 cases, which maybe caused by the different gradient formulation utilized in our approach. One sample adversarial image and its original image is shown in Figure 2. Both the SVM oracle and the LR oracle classified the perturbed image as the number "3". We can see that the perturbed image is still recognizable as an "8", so the adversarial attack is successful. However, the deviations we added to the figure are also clearly visible to the human eye, which is undesirable.

VI. CONCLUSIONS AND FUTURE WORK

In the experiments we have carried out so far, we have shown that adversarial attacks on machine learning classifiers are feasible. The method followed in this paper is two-fold. We imitate the target classifier with a substitute LR model and then generate adversarial samples based on the substitute

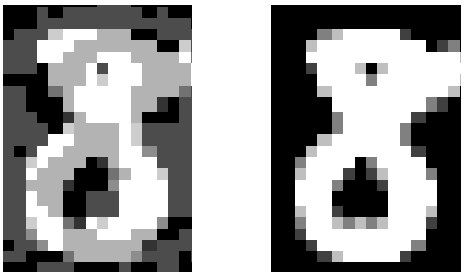

Fig. 2: The left figure shows an adversarial misclassified image with white noise. The right figure shows the original image without noise.

model. However, there are still improvements to be made on the algorithm, such as increasing the efficiency of substitute model construction and reducing the visible deviations of adversarial samples from their original images.

In future work, we will implement the above algorithm utilizing a reduced image feature set. Specifically, we will reduce each of the 28×28 pixel images in LR substitute model training set by $4\times$ and $16\times$ to investigate if performance of the substitute model is sufficiently affected. In the case that the performance of the LR substitute model is unaffected, computational efficiency of the algorithm will be improved without much sacrifice to performance. We will also improve upon the effectiveness of the oracle classifiers by implementing cross-validation during the training stage. As for improving the quality of adversarial samples, we could experiment with methods that generate deviations that are more concentrated on a smaller number of features to reduce the visible effects on the samples; one such algorithm is suggested in [3].

If dimensionality reduction proves to be successful on the MNIST dataset, improving upon the efficiency of Papernot's algorithm in [5], then the same approach will be applied to the GTSRB dataset to portray the possible dangers of adversarial attacks to autonomous vehicle navigation.

ACKNOWLEDGMENTS

We thank Dr. Bahman Bahmani for the guidance he gave us in the direction of this paper. Thanks to Professor Andrew Ng and Professor John Duchi for providing the knowledge base and opportunity to make this project possible.

REFERENCES

[1] The German Traffic Sign Recognition Benchmark, 2010.
[2] LeCun; Yan et. al. The MNIST Database, 1998.
[3] Papernot; Nicolas et. al. Adversarial perturbations against deep neural networks for malware classification. 2016.
[4] Papernot; Nicolas et. al. Practical black-box attacks against deep learning systems using adversarial examples. 2016.
[5] Papernot; Nicolas et. al. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. 2016.
[6] Goodfellow; I.J, Shlens; J, and Szegedy; C. Explaining and Harnessing Adversarial Samples. Proceedings of the 2015 International Conference on Learning Representations., 2015.