An Extraction Attack on Image Recognition Model using VAE-kdtree Model

Tianqi Wen Haibo Hu Huadi Zheng
Department of Electronic and Information Engineering, The Hong Kong Polytechnic University
The Hong Kong Polytechnic University Shenzhen Research Institute
{tianqi.w.wen, haibohu, huadi.zheng}@connect.polyu.hk

ABSTRACT

This paper proposes a black box extraction attack model on pre-trained image classifiers to rebuild a functionally equivalent model with high similarity. Common model extraction attacks use a large number of training samples to feed the target classifier which is time-consuming with redundancy. The attack results have a high dependency on the selected training samples and the target model. The extracted model may only get part of crucial features because of inappropriate sample selection. To eliminate these uncertainties, we proposed the VAE-kdtree attack model which eliminates the high dependency between selected training samples and the target model. It can not only save redundant computation, but also extract critical boundaries more accurately in image classification. This VAE-kdtree model has shown to achieve around 90% similarity on MNIST and around 80% similarity on MNIST-Fashion with a target Convolutional Network Model and a target Support Vector Machine Model. The performance of this VAE-kdtree model could be further improved by adopting higher dimension space of the kdtree.

Keywords: machine learning, exploratory attacks, adversarial machine learning, classification, model extraction

1. INTRODUCTION

Model extraction is one of the adversarial machine learning attacks that aims to extract meaningful information from models including training data, machine learning algorithms, hyperparameters which violate its privacy. Exploratory attacks [1] are launched using a large number of samples to feed the target classifier which could obtain the corresponding labels and rebuild a functional equivalent classifier. The major limitation of extraction attacks is the requirement of a large number of appropriately selected samples that makes this extraction attack difficult to be implemented in practice. Meanwhile, this data-driven attack is accompanied by a high degree of uncertainty. The limited samples may only cover part of decision boundaries in the target classifier with high redundancy. Even with unlimited samples, the result of an extraction attack is still unknown without training and validation datasets which are unavailable for the adversaries and the extraction result can only be inferred through continuous testing. To mitigate the impact of limited real training samples, adversarial attacks with GAN-based data augmentation [2] and Jacobian-based data augmentation [3] used synthetic samples to improve the performance. However, randomly generated samples can only extract partial features used in the target classifier, which results in a classifier that behaves differently from the target classifier.

2. METHODOLOGY

2.1 Variational Autoencoder [4] and kdtree [5]

Variational Autoencoder [4] is a self-supervised image generative model that contains an encoder, latent space, and a decoder. Encoder maps high-dimension data into latent distributions persevering features to the utmost extent which severed as a feature extractor. The decoder reconstructs input images from latent space.

As a data augmentation method, VAE [4] generates a large number of synthetic samples which solves the problem of insufficient samples in data-driven attacks. Unlike Generative Adversarial Networks [6] using random noise to generate simulated images, VAE [4] uses Gaussian Mixture Models to fit k-dimension latent space. By adjusting the parameters of different dimensions in the latent variable, we can get the transition pictures between different classes, and test the critical boundary of the target classifier in the extract attack. VAE [4] not only greatly reduces the data dimension in kdtree, but also reduces the memory occupied. And in the process of reducing the dimension, it separates high-dimensional dependent data, making it easier to test critical boundary.

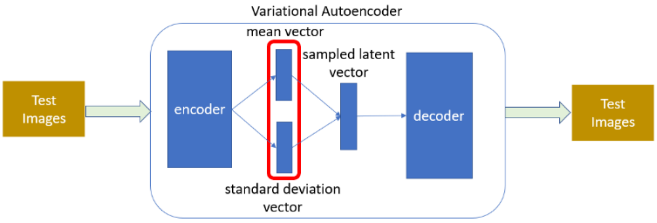

Figure 1: Variational Autoencoder [4]

kdtree [5], a k-dimensional binary tree, is a data structure to manage high-dimensional data. kdtree [5] takes advantage of the characteristics of binary search tree and divides the data of points in a certain dimension at different levels, which greatly reduces the time complexity of KNN [7] search used in the VAE-kdtree prediction and classification process.

2.2 VAE-kdtree Model

Our VAE-kdtree Model comprises a VAE [4] encoder and a kdtree [5]. Our VAE-kdtree model compresses a testing image into a k-dimension vector using the VAE [4] encoder and calculates its KNN [7] label as the classification result.

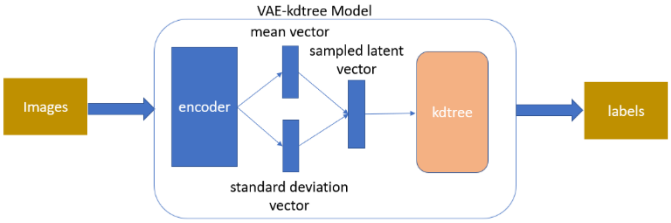

Figure 2: VAE-kdtree Model

2.3 VAE-kdtree Model Extraction Process

A VAE [4] model is trained using testing images served as a reversible dimension reducing model that could link images with k-dimension vectors before the attack and a data augmentation method to generate synthetic samples. Considering the value of the mean vector and standard deviation vector of the VAE [4] model after training, a set of low dimension vectors are selected and fed into the decoder to generate corresponding images. Compared to using a large number of random images to test the response of the target classifier, the VAE-kdtree model selects generated images which located close to the decision boundaries based on the target classifier feedback. Both k-dimension vectors and feedback labels from the target classifier are stored in kdtree [5]. The vectors that located on decision boundaries would be further queried to improve the performance of the adversarial functional equivalent classifier.

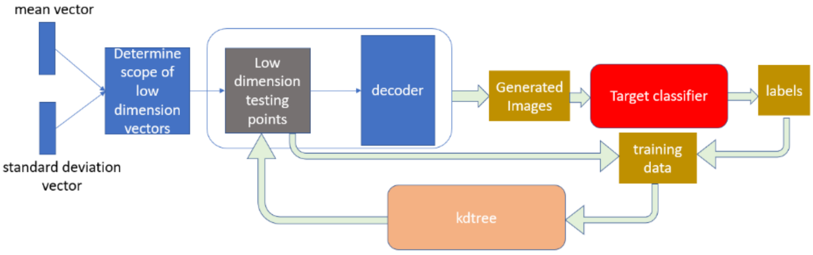
mean vector
Determine scope of low dimension vectors
Low dimension testing points
decoder
Generated Images
Target classifier
labels
training data
kdtree
standard deviation vector

Figure 3: VAE-kdtree Model Extraction Process

3. EXPERIMENTAL RESULTS

3.1 Target Classifiers Preparation

VAE-kdtree is a general black-box extraction attack method. To eliminate the influence of classifier type and dataset on the results, this experiment uses both deep learning classifier and machine learning classifier on MNIST and MNIST-Fashion dataset. A CNN and an SVM classifier have been trained on MNIST and MNIST-fashion dataset as the target classifier respectively.

3.2 VAE-kdtree Model Attack

VAE [4] models adopting different compressed dimensions are trained for the data augmentation and dimension reduction as the first stage of VAE-kdtree attack. Using the VAE [4] model, we performed a reversible dimension reduction, compressing the 28*28*1 picture to d (d=2,3,4), which extracts the main feature from the high-dimensional dependent images. In the low dimension space, we can calculate the distance and correlation between different images, and use limited queries and results to find critical images located on decision boundaries to approach the target classifier step by step.

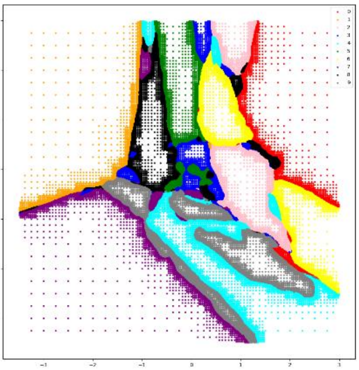

Figure 4: VAE-kdtree Point Map of CNN target classifier when d=2 on MNIST dataset

	CNN_MNIST	SVM_MNIST	CNN_FASHION	SVM_FASHION
D=2	79.93%	81.54%	72.96%	77.34%
D=3	85.78%	87.89%	75.19%	79.32%
D=4	89.3%	90.16%	75.74%	79.15%

Table 1: The similarity between VAE-kdtree model and target classifier adopting different dimension latent spaces

4. DISCUSSION

As shown in Table 1, VAE-kdtree achieves around 90% similarity on both CNN and SVM target classifier on MNIST and around 80% similarity on MNIST-Fashion. As a black box extraction attack, it achieves similar performance when adopting different target classifiers on the same dataset. The complexity of the dataset itself has a great impact on the VAE-kdtree model. Besides the complexity of the dataset, the d parameter of VAE-kdtree model would also influence the performance of the extraction attack. As the dimension of latent space increases, the extraction result of the VAE-kdtree model is better. High dimension vectors retain more image information and details, and the original images can be better restored which helps improve the performance of VAE-kdtree model.

While there is still a gap between the extracted substitution and the target model. Two main factors may limit the performance of extraction in the VAE-kdtree model. Firstly, as a reversible dimension reduction method, the encoding and decoding process of VAE have losses which would make a difference between the target classifier and substitution model. In the VAE-kdtree Model extraction process, the decoder is used to generate the corresponding images according to the specific low dimension vectors. For the classification, the encoder compresses the images and output the KNN [7] labels of the low dimension vectors. In actual experiments, both compression and decompression have losses. The original images and the image after VAE is slightly different. The loss makes the low dimension vectors stored in kdtree [5] and generated images could not perfectly match which limits the performance of the VAE-kdtree model. Secondly, when VAE [4] faces a more complex database, the generated images blur the critical details in the target classifier and limit the extraction performance of the VAE-kdtree model.

ACKNOWLEDGMENT

This work was supported by National Natural Science Foundation of China (Grant No: U1636205, 61572413), the Research Grants Council, Hong Kong SAR, China (Grant No: 15238116, 15222118, 15218919, and C1008-16G), and a research project from Huawei.

REFERENCES

[1] Y. Shi, Y. Sagduyu, and A. Grushin, “How to steal a machine learning classifier with deep learning,” in 2017 IEEE International Symposium on Technologies for Homeland Security (HST), 2017.

[2] Y. Shi, Y. E. Sagduyu, K. Davaslioglu, and J. H. Li, “Generative Adversarial Networks for Black-Box API Attacks with Limited Training Data,” 2018 IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), 2018.

[3] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, “Practical black-box attacks against machine learning,” in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, 2017.

[4] D. P. Kingma and M. Welling, “Auto-Encoding Variational Bayes.” 2013.

[5] J. L. Bentley, “Multidimensional binary search trees used for associative searching,” Commun. ACM, vol. 18, no. 9, pp. 509--517, 1975.

[6] I. J. Goodfellow et al., “Generative Adversarial Networks,” arXiv [stat.ML], 2014.

[7] T. M. Cover, “Estimation by the nearest neighbor rule.,” IEEE Trans. Inf. Theory, vol. 14, no. 1, pp. 50–55, 1968.