Research Statement

Yao Yiran
speagleyao@sjtu.edu.cn

1 Introduction

My research interests have revolved around security in machine learning. So far, my research focuses on one main topic in this area - Adversarial learning. Adversarial learning, to put it simply, is about fooling models to cause malfunction through attack techniques. In the context of computer vision, adversarial attacks refer to adding a human-imperceptible perturbation to an input image and leading a well-trained model to output wrong results. This kind of attack has a significant impact on the security of machine learning models. On one hand, people with malicious purposes can use this technique to break models applied to areas where incorrect results may lead to critical security problems, like autopilot and face recognition. On the other hand, it can safeguard people from maliciously used models. For example, we have seen the convincing effect of Deepfakes on face replacement in images and videos, which can easily violate the portrait rights of people and give away fabricated messages to credulous people. As a counteract to malicious adversaries, we use specific defense techniques such as adversarial training and anomaly detection to improve the robustness of models against adversarial attacks.

I have done research in improving the robustness of models through taking the characteristics of datasets into account. Specifically, in one previous work, we identify an intriguing phenomenon in classification problem called symmetry property in the prediction of adversarial examples by a newly designed metric. That is, each class shows vulnerability to particular classes, and some class pairs correlate strongly and have the same degree of misclassification. We analyze the reason behind the existence of symmetry property and propose a novel loss function in adversarial training, which outperforms the state-of-the-art defense methods in extensive experiments on benchmarks. In the work of this paper, I run most of the experiments, during which I propose the final version of the loss function and sort out the evaluation table. **Currently, I continue exploring more in the imbalance of datasets.** I combine adversarial learning with cost-sensitive learning, where I try to train models that can defend against adversarial attacks between certain classes, rather than having a more general defense between all classes.

Another part of my current research focuses on finding universal perturbations for generative adversarial networks (GANs) that are used in face manipulation. These GANs can swap the hair color or gender of people, or put a smile of given angles on people's faces. A universal perturbation should have the defending effect against arbitrary manipulations with any model. If such perturbations exist, they can be added to images or videos in real-time so that we can protect everyone's privacy at a lower cost.

2 Summary of Previous and Ongoing work

2.1 Identifying Influential Inputs in Probabilistic Logic Programming

My research path in machine learning began here with Prof. Zhou. Probabilistic logic programming (PLP) has been widely applied in solving data-centric problems and has been proved to be effective in encoding machine learning models. In this work, we introduce a novel provenance-based approach towards identifying influential inputs in PLP programs. By querying the inputs that most strongly impact the learning results derived from a PLP program, users can obtain helpful information to reason and debug. Our evaluation in two applications, a visual question answering scenario and a smoking behavior study, demonstrates the effectiveness of our methods, and we further prove the maintenance and querying performances are reasonable in practical scenarios. **Personally, I run part of the experiments and participate in revising the paper.**

2.2 On Symmetry Property in Adversarial Examples

During my extensive reading of papers in the previous work, I came across a paper about adversarial learning. I got so fascinated by it and read several relevant papers. AI, so powerful as it looks on the surface, is so vulnerable as well. If we can figure out why AI is vulnerable and how to make it robust, we may have a deeper understanding of how AI works, and more reassuring machine learning models can be obtained to be used in critical situations. Thus, I joined the lab in school under the guidance of Prof. Ni and began to research on this topic.

In most of the past works, people treat all classes in datasets as the same in classification problems. However, we find out that the inherent imbalance does harm to the robustness of models in the face of adversarial attacks. Specifically, we identify an intriguing phenomenon called symmetry property in the prediction of adversarial examples, that is, each class shows vulnerability to certain classes, and some class pairs correlate strongly and have the same degree of misclassification towards each other. We demonstrate this through a newly designed metric called attack proportion which counts the proportion of the adversarial examples misclassified to one particular class. As the distribution of attack proportion is unbalanced and somewhat symmetrical, we call this phenomenon symmetry property.

Here, we regard that the classifier is composed of a feature extractor and one output layer. We argue that images with high similarity tend to have similar outputs in feature space. As a result, those features lay aside the decision boundary in feature space and can be easily made into adversarial examples by adding small perturbations. Such vulnerability exists as the softmax cross-entropy loss function does not explicitly impose any constraints to learn discriminative representation for classification, but merely focuses on building the mapping between the input and output label. We thus propose a novel loss function that indirectly constrains the output in feature space to be discriminative and use it in adversarial learning. The results in extensive experiments on benchmarks outperform the state-of-the-art defense methods.

In the work of this paper, I run most of the experiments including comparisons between different datasets and models and various defense methods, during which I propose the final version of the loss function and sort out the evaluation table. I also participate in part of the paper writing, like abstract, related work, methodology, and experiment results.

2.3 Combination of adversarial learning and cost-sensitive learning

After the research at school, I seek more opportunities to research further in this area. I'm honored to have a remote visit to the University of Virginia under the guidance of Prof. Evans. Previous papers usually take all classes in datasets as the same and strive to find a stronger defense between all classes. This type of defense method is designed for untargeted adversarial attack, which till now is hard to achieve a high accuracy under attack. Although the arm-race between attack and defense is still white-hot, it does little help to some practical scenarios. For example, when we try to attack the face recognition for bank account verification, we will attack one specific person by using targeted attack. It is true that targeted attack is not as strong as untargeted attack under the same setting, but in this scenario, the machine learning model should be non-error under targeted attack, otherwise leading to critical security consequences.

The research aim thus is to obtain a machine learning model that is non-error under targeted attacks in some directions. This is somewhat similar to cost-sensitive learning, as some misclassifications are acceptable and some are not. By printing out the confusion matrix that shows the exact numbers of examples misclassified as other labels, I have done some basic experiments to learn more about this topic. An intriguing asymmetry occurs under targeted attack, and I'm introducing cost matrix into adversarial learning to do further research in order to figure out the reason behind it.

2.4 Universal perturbations for GANs

For my summer internship at school, I focus on finding universal perturbations for GANs that are used in face manipulation. These GANs are able to swap the hair color or gender of people, or put a smile of given angles on people's faces. For example, we have seen the convincing effect of Deepfakes on face replacement in images and videos, which can easily violate the portrait right of people and give away fabricated messages to credulous people that both may lead to malicious effects. Adding an adversarial perturbation has been proved to be able to defend against such manipulation to some extent.

A universal perturbation is meant to be valid under different settings and here it should defend against arbitrary manipulations with any models. If such perturbations exist, they can be added to images or videos in real-time so that we can protect everyone's privacy at a lower cost. I have already found universal perturbations for some GANs that worked on celebA, which is a large-scale face attributes dataset, and is further trying to design universal perturbations that can be applied to multiple models.

3 Future Research

My future research plans revolve around adversarial machine learning and data privacy. The world relies more and more on machine learning models, so it poses a severe challenge to the robustness of models. Besides, people are more aware of the need to protect personal privacy, while the current technical means have not yet given people proper ways to protect themselves. Therefore, I regard these two topics crucial and there still remains a lot worth exploring. **My long-term goal is to build models robust enough to apply to the real world and propose handy techniques to protect personal privacy.**

During the period of my studying for a Ph.D., under the guidance of my supervisor, the study of curriculums in school, and unremitting research, I will have a more comprehensive and in-depth understanding of these two topics. I hope to start with research in specific areas, publishing several papers as the first author in the first few years, and later turn to think about some more general and basic ideas in these topics if possible, striving to make some groundbreaking contributions to the field of machine learning and security. **According to my current view, I regard the following areas as my potential research directions:**

3.1 Adversarial learning

Adversarial machine learning is still the topic I want to research on. I firmly believe that the gap between human eyes or intuitions and machine learning models somewhat points directly to the defects of those models represented by neural networks. Through the process of narrowing the gap, we can gain a deeper understanding of the neural network models and build more robust models for practical problems. Not only do they have high accuracy on tasks, but they also don't fall into attacks so easily as today's models. That is when we can say that AI techniques are really as strong as what propaganda says. In the short plan, I tend to focus on fixing the inherent imbalance in datasets. A long-term goal may be to explore the nature of adversarial examples, and train more robust models by eliminating this nature of existence.

3.2 Membership Inference Attack

Membership inference attack is one topic that focuses on people's privacy. Normally, when a machine learning model is trained, we then discard the training data and use the model without worrying about the training dataset. However, the model tends to have higher confidence on the data that are originally in the training dataset, which may leak sensitive information contained in datasets. For example, if one model is specifically trained on people's real identity and medical record, such an attack may reveal certain diseases of one person, thus causing critical consequences for the security of personal privacy. Several membership inference attack methods have been proposed, yet existing defense methods are not very effective. In some way, I believe this has relations with the non-smooth decision boundary, which is quite similar to adversarial learning.

3.3 Differential Privacy

In recent decades, we have seen too many cases where user data stored by companies have been stolen or leaked, many of which have caused rather critical security consequences. Differential privacy is such a technique that allows information about users to be collected, used and shared while maintaining the privacy of individual users. In facing machine learning models, where a large amount of data is intrinsically required, differential privacy naturally has its unique importance. Generally speaking, differential privacy comes at the expense of model accuracy. I look forward to finding a better way to apply differential privacy to machine learning models so that we can build excellent models to solve practical problems on the basis of complete protection of personal privacy.

3.4 Federated learning

Federated learning also helps protect personal privacy. With respect to mobile phones, this means that one personal device downloads the current model, improves it by learning from local data, and then sends the small encrypted update to cloud, where all training data remains on the local device and no individual updates are stored in cloud. However, communicating model updates throughout the training process can still reveal sensitive information, thus further research is required to build robust federated learning algorithms.