MSSI Capstone Project Agreement
Johns Hopkins University

Student Name(s)	Zheng Qu (Only one form needed from each team) Yong Wu (No more than **3** students in one team) Ran Liu
Faculty Mentor	Timothy R. Leschke
Other Personnel	___________________________(Please specify the role of each person)
Project Start Date	09/01/2017 (At least 2 months before completion date)
Intended Completion Date	11/26/2017 (Annual deadlines are approx. e.g., 8/20,12/15, 5/10)

The student(s) must submit this signed form and the final version (both hard and electronic copies) of the MSSI capstone project report to the Information Security Institute by the completion date. In addition, any code or data produced should be carefully documented and turned in to the faculty mentor/academic advisor by the completion date.

A project report is typically about 25 pages single-spaced, not including the cover page, table of contents, references, figures, tables, etc. Where appropriate, it may take the form of a single-author or multiple-author paper about the work (written for submission to a conference or journal), augmented by the student with further details as needed.

Research is open-ended, so the direction may change, within reason. To fulfill the MSSI degree requirement, the amount of work contributed by the student should be equivalent to the total workload of about two 400 or higher level courses.

Proposed Title	Adversarial Examples on Image Recognition with Deep Neural Networks
Project Description	Machine learning classifiers have been widely used for decades. Nowadays, Convolutional Neural Networks classifiers perform better than human experts in image recognition. However, all of the state of the art classifiers could be fooled by maliciously generated images, tons of adversarial examples have been found to brainwash a variety of deep learning architectures, especially those with linearity. Image recognition plays fatal role in lots of applications, such as autonomous driving, and

mobile check deposit. The adversary could maliciously modify the images, feeding the classifiers to causes them to predict wrong classes.

Adversarial attack/defense is still an very active area of research, in this project we will pay most of our attention on the adversarial attack toward image recognition, trying to figure out a general adversarial attack, especially on the state-of-the-art classifier with convolutional neural networks like ResNet.

We will use ImageNet dataset in this project, benchmark our performance with the evaluation metric used in Kaggle Adversarial Competition. In our work, we will show how the adversarial attack is going to be effective. Hopefully if time and computation resources allow, we will try to explore adversarial examples in other applications like speech recognition, the defense mechanism against adversarial attack as well.

Materials to be produced (papers, code, data, experiments, etc.)

We will produce materials as below:

- One final report, documenting the related work, our project, approach, architecture, performance, and future work.
- Codes to reproduce our experiments. We plan to mainly develop python code with TensorFlow framework, plan to support both CPU and GPU computation.
- Dataset we use, mainly some image dataset from ImageNet.
- PPT about our work.

Agreement by student (to be signed at the start of the research):

The opportunity to work with other researchers is a privilege. I understand that this project should be my top academic priority, both out of consideration for my supervisor and faculty mentor’s time and because others may be depending on my results.

I agree to keep my supervisor apprised of my progress throughout the project, via regular updates or another agreed-upon mechanism.

I will submit a well-written report or thesis and properly documented supporting materials to my supervisor at least **2** weeks before the completion date. I understand that my supervisor may request changes or additions before approving the project, or may decline to approve it, which could affect my ability to graduate.

Student signature(s) _______ zheng qu __________ Date 09/15/2017_

______ WuYong ____________ Date _09/15/2017__

_______________________________ Date _09/15/2017_

Approval by faculty mentor (to be signed upon completion of the research):

I certify that the capstone project entitled ____ has been satisfactorily completed by ____ and that it fulfills one of the requirements of the MSSI degree.

Faculty Mentor signature _____________________________ Date _______________

Updated January 25, 2014

Adversarial Examples on Image Recognition with Deep Neural Networks

Ran Liu, Zheng Qu, Yong Wu

Background

After AlexNet was devised in 2012, deep neural networks classifiers have been widely exploited in industry, such as autonomous driving, mobile check deposit. The accuracy of deep neural network is greater than any traditional classifier like SVM, logistic regression. Nowadays the convolutional neural networks architectures become deeper and deeper, such as VGG and ResNet, and the accuracy becomes better and better, even better than human experts in ImageNet recognition. Whereas, recently, adversarial examples[1-4] have been found to be able to cause almost all machine learning classifiers to misclassify, especially those with linearity. The attacker is able to generate adversarial examples which is very close to the original ones, the adversarial examples are indistinguishable to the human eyes, but a variety of classifiers would misclassify them, in some cases, the adversarial examples are unidentifiable to human, but machine learning classifiers could be led to make prediction with high confidence.

Dataset

In this project, we will use ImageNet dataset, the commonly used dataset for image processing and benchmarking. Fortunately, Google has launched three adversarial attack competition three months ago and they are still ongoing. The public dataset in the competitions is also from ImageNet, we plan to use it in our project. All the images are 299x299 pixels (RGB) with 1000 classes.

Our Project

In this project, we will pay most of our attention on the adversarial attack toward image recognition, trying to figure out a general adversarial attack, especially on the state-of-the-art classifier with convolutional neural networks like ResNet. We will benchmark our performance with the evaluation metric used in Kaggle Adversarial Competition[5]. In our work, we will show how the adversarial attack is going to be effective, and why it works. Hopefully if time and computation resources allow, we will try to explore adversarial examples in other applications like speech recognition, the defense mechanism against adversarial attack as well.

References

[1] Huang, Sandy, et al. "Adversarial attacks on neural network policies." *arXiv preprint arXiv:1702.02284* (2017).

[2] Kurakin, Alexey, Ian Goodfellow, and Samy Bengio. "Adversarial examples in the physical world." *arXiv preprint arXiv:1607.02533* (2016).

[3] Tramèr, Florian, et al. "Ensemble Adversarial Training: Attacks and Defenses." *arXiv preprint arXiv:1705.07204* (2017).

[4] Shen, Shiwei, et al. "AE-GAN: adversarial eliminating with GAN." *arXiv preprint arXiv:1707.05474* (2017).

[5] NIPS 2017: Non-targeted Adversarial Attack, https://www.kaggle.com/c/nips-2017-non-targeted-adversarial-attack