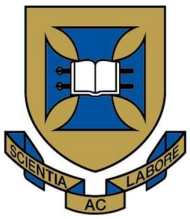

THE UNIVERSITY OF QUEENSLAND

AUSTRALIA

Project Proposal

Towards Safer Deep Learning Networks

By

Cheng Nian

s4547814

School of Information Technology and Electrical Engineering

University of Queensland

August 2019

1 Introduction

Cutting-edge models of machine learning, especially neural networks, have enjoyed phenomenon success in various applications and achieve exciting results on image classification. With their ability to study abundant, non-linear parametric mappings on large-scale datasets, neural networks have become a powerful and efficient framework, which have been applied to difficult problems that were hard to handle using the traditional techniques before.

Despite they has excellent pattern recognition performance in various kinds of research domains, neural networks have recently be shown that they can be unstable occasionally. In fact, really small but well sought intentionally perturbations can lead to the machine learning models to misclassify the output. These perturbations are named adversarial examples or adversarial attacks, which can result in serious consequences. The system security are quite vulnerable to these crafted adversarial examples.

To avoid the system from outputting incorrect answers, the robustness of the classifiers are needed to be computed and assessed on a large-scale dataset. Therefore, aimed at researching the robustness of diverse classifiers when input adversarial perturbations, an accurate method for adversarial perturbations is needed. It is the key to better understand the current architectural limitations and design approaches to enhance robustness.

2 Literature Review

The adversarial example was first estimated by solving the punishment optimization problem. The analysis shows that the complexity of neural networks is one of the key reasons for the existence of adversarial attacks. However, the optimization methods used in [Szegedy et al., 2013]needs quite a long time so it is not preferable to be applied to large-scale datasets.

In [Pepik et al., 2015], through experiments on Pascal3D+ annotation, it is demonstrated that convolutional networks will not be constant for certain types of transformation. More recently, Tsai et al. [Tsai and Cox, 2015] developed a software that misclassifies a specified image into a given class. Meanwhile there is no need to find the smallest perturbation.

The author of [Goodfellow et al., 2014] introduces the "fast gradient sign" method, which can effectively calculate the smallest purterbations of a given classifier. Although this method is efficient, it can only approximate the optimal perturbation coarsely.

Finally, it should be pointed out that adversarial instability also leads to the research of [Fawzi et al., 2015]. Some adversarial examples on some series of classifiers are studied, and the thresholds of the robustness are demonstrated. The method proposed in this paper can be used as an accurate and effective baseline for generating adversarial examples.

In [Moosavi-Dezfooli et al., 2015], the authors propose a DeepFool algorithm to fool cutting-edge classifiers by adversarial attacks, from binary classifiers to multiclass classifiers. The authors provide extensive experimental evidence to confirm the accuracy. The advantages and effectiveness of this method are demonstrated.

In [Chakraborty et al., 2018], Anirban Chakraborty et al. discusses in detail on various kinds of adversarial attacks and different risk models. The efficiency together with challenges of recent countermeasures against these attacks are demonstrated. The authors also research several classic attacks and relevant defense strategies. The aim of this paper is to summarize the latest development of various adversarial attacks and their defense strategies. To do this, the authors analyze various threat models and attack scenarios. Methods relevant to previous researches are adopted and not limited to the applications. The authors present an overview of modifying examples in Chapter 4.2 so that the classification model produces adversarial output.

In [Dai et al., 2018], an attack method based on reinforcement learning is proposed. This method studies the generalizable attack strategies from the target classifier with the prediction labels. Hanjun Dai et al. studies the adversarial attacks on graph structured data. Three efficient attack methods named RL-S2V, GradArgmax and GeneticAlg are proposed for three different attack settings, respectively. Experimental evidence shows a few GNN models are vulnerable to such attack, based on synthetic and real-world data. Defense methods are also discussed through experiments.

In [Goodfellow et al., 2014], a white box and directional attack method is proposed to calculate the perturbation effectively by discovering the linear interpretation against the sample perturbation. Former works on explaining the adversarial examples lies on nonlinearity. By contrast, Ian J. Goodfellow et al. suggests that neural networks can be susceptible to antagonistic disturbances depend on their linearity. In addition, this view provides a method to generate adversarial examples in a simple and quick way. This method provides adversarial examples for adversarial training and reduces the error of test set of maxout network on the database(MNIST).

In [Akhtar and Mian, 2018], Naveed Akhtar and Ajmal Mian researches the previous works that design adversarial examples, analyze them together with the defense strategies. The authors suggests that it is possible to conduct adversarial attacks in real situations and review the contributions of evaluating adversarial attacks.

In [Frosst et al., 2018], Nicholas Frosst et al. presents a simple technique named Detecting Adversaries by Reconstruction from Class Conditional Capsules(DARCCC) that enables various models to find the adversarial images. The capsule model is tuned to rebuild the image according to the correct attitude parameters and identification of the top capsule. The authors present an effective way to detect adversarial images for three different datasets, by setting a threshold. A stronger, white-box attack is also demonstrated, taking the reconstruction error into account .

In [He et al., 2019], Xiang He et al. proposes a non-local context encoder which is resistant to adversarial examples. This method enhances the feature with channel feature and captures the global spatial dependence by learning the extracted features. The nonlocal context encoding(NLCE) modules are the key of the nonlocal context encoding network

(NLCEN). In addition, the outcomes show that NLCE modules can help boost the stability against various adversarial examples.

Since Projected Gradient Descent(PGD) generates attack samples independently for each data sample based on the lost function, the procedure does not necessarily lead togood generalization in terms of risk optimization.In [Zheng et al., 2018], Tianhang Zheng et al.defines a new energy functional to better reflex the discriminative data manifold in the WGF (Wasserstein Gradient Flows) framework to solve the adversarial-distribution problem. The authors achieve the goal by proposing distributionally adversarial attack (DAA), evaluated by attacking cutting-edge defense models. This work done in this paper will help in understanding the distributionally adversarial attack.

In [Zügner et al., 2018], Daniel Zgner et al. focuses on the classification of nodes by graph convolutional network, and proposes a method to attack graph. The attack target is the nodesAZ features and the structure of the graph. This paper also proposes direct and effective attacks to exploit the relationship of the data. This article also developes algorithm in a discrete domain to generates perturbations. The work gives me good insights for further study.

3 Purpose & Aim

Although depth networks achieve the most advanced performance in image classification field, they do not have enough robustness when small adversarial perturbations accur. They even misclassify the data with the smallest perturbations that look similar to clean samples occasionally. From a security standpoint, this could actually be a tough problem. Therefore, in order to find out the factors that could effect the robustness of different classifiers to adversarial perturbations, an efficient approach for searching adversarial perturbations is needed.

4 Significance of the Study

After the literature review has mentioned above, the objectives of my project are identified as follow:

1. Explore and study the existing methods to compute perturbations that fool deep networks.

2. Investigate and improve the DeepFool algorithms.

3. Analyze the results of the previous methods in the field of computing adversarial perturbations,

5 Methodology

The resulting code will be generated using Python or MATLAB, and the results will be verified and compared with previous results. It is anticipated that the following resources will be required: textbooks, related papers and software able to cover the basics of deep neural network technology.

6 Project Plan

The following are the milestones.The table shows the time span for each milestone.

1. Familiar with project topic and update with ongoing research, including the literature survey, annotated bibliography,project proposal submission and update with the ongoing research.
2. Literature survey
3. Prepare for the seminar
4. Final algorithm development and improvement
5. Conference
6. Thesis Poster
7. Final thesis submission

7 Risk Assessment

The evaluation of OH and S has been completed for the laboratory used in the project.The executing of laboratory is preceded by a risk assessment task. In addition, during the experiment, I will strictly abide by the rules and regulations of the laboratory and all safety measures.

Activity \ Time/week	s2-1	s2-2	s2-3	s2-4	s2-5	s2-6	s2-7	s2-8	s2-9	s2-10	s2-11	s2-12	s2-13
Literature Review				✓									
Annotated bibliography					✓								
Proposal submission					✓								
Update with ongoing research												✓	
Seminar presentation													✓

Table 1: List of tasks

Time/week / Activity	s1-1	s1-2	s1-3	s1-4	s1-5	s1-6	s1-7	s1-8	s1-9	s1-10	s1-11	s1-12	s1-13
Developing algorithm					✓								
Conference Paper								✓					
Poster								✓					
Final thesis submission												✓	

Table 2: List of tasks

References

[Akhtar and Mian, 2018] Akhtar, N. and Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *CoRR*, abs/1801.00553.

[Chakraborty et al., 2018] Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. (2018). Adversarial attacks and defences: A survey. *CoRR*, abs/1810.00069.

[Dai et al., 2018] Dai, H., Li, H., Tian, T., Huang, X., Wang, L., Zhu, J., and Song, L. (2018). Adversarial attack on graph structured data. *CoRR*, abs/1806.02371.

[Fawzi et al., 2015] Fawzi, A., Fawzi, O., and Frossard, P. (2015). Analysis of classifiers' robustness to adversarial perturbations. *CoRR*, abs/1502.02590.

[Frosst et al., 2018] Frosst, N., Sabour, S., and Hinton, G. E. (2018). DARCCC: detecting adversaries by reconstruction from class conditional capsules. *CoRR*, abs/1811.06969.

[Goodfellow et al., 2014] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014). Explaining and Harnessing Adversarial Examples. *arXiv e-prints*, page arXiv:1412.6572.

[He et al., 2019] He, X., Yang, S., Li, G., Li, H., Chang, H., and Yu, Y. (2019). Non-local context encoder: Robust biomedical image segmentation against adversarial attacks. *CoRR*, abs/1904.12181.

[Moosavi-Dezfooli et al., 2015] Moosavi-Dezfooli, S., Fawzi, A., and Frossard, P. (2015). Deepfool: a simple and accurate method to fool deep neural networks. *CoRR*, abs/1511.04599.

[Pepik et al., 2015] Pepik, B., Benenson, R., Ritschel, T., and Schiele, B. (2015). What is holding back convnets for detection? *CoRR*, abs/1508.02844.

[Szegedy et al., 2013] Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. *arXiv e-prints*, page arXiv:1312.6199.

[Tsai and Cox, 2015] Tsai, C.-Y. and Cox, D. (2015). Are deep learning algorithms easily hackable? http://coxlab.github.io/ostrichinator.

[Zheng et al., 2018] Zheng, T., Chen, C., and Ren, K. (2018). Distributionally adversarial attack. *CoRR*, abs/1808.05537.

[Zügner et al., 2018] Zügner, D., Akbarnejad, A., and Günnemann, S. (2018). Adversarial Attacks on Neural Networks for Graph Data. *arXiv e-prints*, page arXiv:1805.07984.