Research Outline

Adversarial Examples in Deep Learning

Rafael Carvalhaes Possas

School of Information Technologies
University of Sydney
Australia

1 Introduction

Computer Image recognition have become increasingly popular nowadays due to the increase in demand for wearables, smartphones and any kind of device with attached cameras. Camera sensors and processing power have received considerable improvements across the years and have reached a point where sophisticated algorithms are able to run on these devices and provide state of the art Machine Learning performance. Therefore, different techniques for exploring these new possibilities have been developed and have been used for a wide variety of applications.

This field of study is extremely relevant as the amount of data available grows exponentially. Companies like Google and Facebook can make use of thoursands of terabytes of images in order to create systems with highly complex recognition capabilities. Yet, little is known about how the developed techniques can learn and generalize their results on such tasks. Exploring machine learning on graphical data is similar to understanding how the human brain works and what drives learning and generalization by human beings. This is an important insight as it not only helps to better understand the human brain but it also makes it possible to create machines capable of interacting and understanding the environment in the same way we do.

2 Background and Literature Review

Neural Networks have been the underlying foundation of such tasks. These algorithms are able to represent highly non-linear and complex mathematical problems and, thus, able to work very well on datasets with an increased number of dimensions. Machine Learning systems are becoming ubiquitous and, therefore, require to be not only very accurate but also robust to any kind of attack. Methods for exploiting vulnerabilities have been found and should become a major concern for those using these algorithms. It has been shown that one can create small perturbations to images, usually imperceptible to human eyes, in order to drive the system to a misclassified label of an input. Yet, none of these techniques can guarantee a specific label as, until now, most methods are able to provide only random misclassification results.

The baseline method named Fast Gradient Sign deveoloped on Goodfellow et. al. (2014) work have been used for different kinds of experiments. These range from black box attacks [2] to high confidence predictions of unrecognizable images [3]. These techniques can be supported by the principle of knowledge transferability on machine learning algorithms [4] where one system model parameters can be learned to some extend by another system. This makes possible to a third party system to craft adversarial images capable of fooling the system where the learning knowledge was extracted.

3 Research Questions

None of the studies on this field have focused on understanding which types of images would be more susceptible to these kinds of attacks. Billovits et. al [5] have split adversarials in four categories: True Adversarial, Benign Adversarial, Conaturally Adversarial and Re-Focused Adversarials. Yet, there is no quantitative focus on which of these can easily lead to a misclassification of a crafted adversarial image. These categories are the first step on understanding and grouping characteristics that are better exploited by the gradient sign method and, thus, should be better understood.

However, types of images are not the only factor driving the misclassification on such systems. Within Neural networks techniques, there are different types of architectures that can possibly yield a different outcomes. For the time being it is not possible to test all the different configurations available but at least the 3

most popular should be used to understand how the configuration of a system can influence on their robustness to these attacks. Questions should be answered on a more focused quantitative way as, until now, studies have mostly focused on developing new attacking techniques rather than understanding how they apply to different types of data and/or systems.

These questions are important to understand how vulnerable the machine learning systems used on a daily basis are. Some of these are required to have high standards of security and, therefore, should provide robustness against adversarials attacks. By having a deeper understanding of such attacks, one can develop methods to use crafted adversarial images as a way of regularizing their algorithms and, thus, increasing their robustness.

4 Method and Design

In order to answer the aforementioned question it is needed to firstly categorize the images, so results can be inferred from the experiments. The four categories (Re-Focused, Benign, Conaturally and True adversarials) found by Billovits et al. [5] should serve as a starting point for understanding the likelihood of different types of images to the perturbations caused by the Fast Gradient Method.

A random dataset of fixed size should be used on a Network trained on the ImageNet dataset and the results should be split among these 4 categories. Thus it is going to be able to visualize specific patterns across the different categories and, possibly, have more categories of misclassified images at the end of the research. The pre-trained ImageNet network made available through the TensorFlow package named InceptionV3 should be used as the predictive algorithm of this work. This choice was due to the recent good performance of this network on image recognition tasks. The main steps of the entire method are as follows:

1. Sampling random images from the ImageNet dataset
2. Applying pixel perturbation using gradient information
3. Running the predictive algorithm on the sample
4. Evaluation of results
5. Categorization of possible patterns that drive misclassification.

Understanding how different categories of images behaves to this types of perturbation can help to understand better the vulnerabilities of the network. Thus, regularization techniques can be improved in order to account for the possibility of such perturbations. At the end of this research, we would be able to better answer the question whether intrinsic characteristics of images have any influence on the accuracy of the classification output of a perturbed image.

References

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[2] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. Berkay Celik, and A. Swami, "Practical black-box attacks against deep learning systems using adversarial examples," *arXiv preprint arXiv:1602.02697*, 2016.

[3] A. Nguyen, J. Yosinski, and J. Clune, "Deep neural networks are easily fooled: High confidence predictions for unrecognizable images," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2015, pp. 427–436.

[4] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples," *arXiv preprint arXiv:1605.07277*, 2016.

[5] C. Billovits, M. Eric, and N. Agarwala, "Hitting depth: Investigating robustness to adversarial examples in deep convolutional neural networks."