Temperature control of softmax output layer rather than Distillation technique as a Defense to Adversarial Perturbations against Convolutional Neural Networks[1]

Haobei Song
University of Waterloo

March 26 2017

Abstract

Convolutional neural network (CNN) as a well developed deep learning architecture has been widely used in computer vision such as automatic inspection, autonomous driving, image processing. Though state-of-art accuracy by elaborately designed CNN was achieved in many computer vision tasks, recent studeis have shown the potential vulnerability to adversarial perturbations among not only CNN but other deep neural networks. This discovery is of substantial significance as the widely usage of Convolutional neural network targets the tasks of extensive security concern such as the CNN used in autonomous driving which might be crushed by slight modification of the environment. In this study, the defensive effect of distillation training for CNN is evaluated together with traditional trained CNN for MNIST task on a data set of 60000 examples. The result shows it is the temperature (T) of the softmax function of the output layer that plays an important role to reduce the adverserial gradient (by a factor of 10^{10} at $T = 20$) and the success rate of adversarial attack (by a factor of 10 when $T = 20$), rather than the distillation traininng technique claimed effective by other researchers.

1 Introduction

Considering when humans learn to recognize the digits written on a piece of paper, the input is the viewing of that image on the paper or the pixels on the paper with someone by their side teaching them the right number that image represents. When people teach a computer how to do this by letting it learn the parameters characterizing such a task from a mass of data (hand written digits with labels) without explicitly crafting the algorithm, a general neural network is always built with some initial parameters to be modified to fit the given data set and generalize well when applied to new data. Learning the hand written digits is a classical deep learning task widely refered to as MNIST, which is often done by convolutional neural networks (CNN).

Ever since the introduction of convolutional neural networks for image processing tasks, the recognition accuracy has increased dramatically and achieved state-of-art accuracy in the past few years. Though computers now could outperform humans on this specific task, recent study has shown the lack of robustness of CNN againt adversarial perturbations which raises plenty of problems about its pragmatic application.

Compared with training computers to solve MNIST problem, humans during training process could learn more information than the hard labels given, such as the similarity between different digits, which is some information left out by traditional training with hard label. Theoretically, a deep enough neural network could learn such extra information when trained on a large enough scale of data. Such a neural network does not exist so far due to the limited data people can obtain and computation constraint required to perform the training on such neural network, as there is considerable complexity even within the simplest learning task such as MNIST. Humans often make use of knowledge

from a variety of areas such as math, culture or even their personal experience to deal with MNIST tasks. For example, people can easily recognize rotated hand-written digits right after the training of upright digits from geometry or arabs could find some correlation from their language etc. Thus, specifiing some hyperparameters of the training model is considerably neccesary for building a DNN to sole a realistic problem.

Though soft label method during training process has been suggested but crafting such soft label is also questionable as there is no general rule to create labels which perimetrize the relationship of hand-written digits falling into different categories as percieved by humans. It also requres a considerable amount of effort to do these tedious task without a systemetic rule, which disabeys the principle of machine learning that simply tries to avoid these tedious work.

That is where distillation comes to the stage as to provide distilled labels from previously built CNN. The distilled labels can be considered as a kind of soft labels produced by computer, which labels each sample with a soft label using a traditional CNN. Papernot, McDaniel and Wu etc. have claimed its effectiveness against adversarial perturbation from empirical study in [1]. In their work, they also applied a softmax output layer parametrized by a parameter called temperature, which turned out to be the most important factor reducing the success rate of adversarial attack in our study and the effectiveness of distillation solely becomes suspectable.

References

[1] N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," *CoRR*, vol. abs/1511.04508, 2015. [Online]. Available: http://arxiv.org/abs/1511.04508