Distillation as a Defense to Adversarial Perturbations against Convolutional Neural Networks

Haobei Song
University of Waterloo

March 27, 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 exposed the potential vulnerability to adversarial perturbations among not only CNN but most of the deep learning algorithms. This discovery is of substantial significance as the use of Convolutional neural network exclusively 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, and .

1 Introduction

Consider when humans are learning to recognize the digis, they learned more information than the label given such as the similarity between different digits, which is some information left out by traditional training with hard label. Though soft label method has been suggested but crafting such soft label is also questionable as there is no general rule to create a label with perimetrize the relationship between different category as percieved by humans, it also requres a considerable amount of effort to do these tedious task without a mandated rule.