Abstract

In this thesis, our goal is to develop robust and reliable yet accurate learning models, particularly Convolutional Neural Networks (CNNs), in the presence of adversarial examples and Out-of-Distribution (OOD) samples.

As the first contribution, we propose to **predict adversarial instances with high uncertainty through encouraging diversity in an ensemble of CNNs**. To this end, we devise an ensemble of diverse specialists along with a simple and computationally efficient voting mechanism to predict the adversarial examples with low confidence while keeping the predictive confidence of the clean samples high. In the presence of *high entropy* in our ensemble, we prove that the predictive confidence can be upper-bounded, leading to have a globally fixed threshold over the predictive confidence for identifying adversaries. We analytically justify the role of diversity in our ensemble on mitigating the risk of both black-box and white-box adversarial examples. Finally, we empirically assess the robustness of our ensemble to the black-box and the white-box attacks on several benchmark datasets.

The second contribution aims to address the detection of OOD samples through an end-to-end model trained on an appropriate OOD set. To this end, we address the following central question: **how to differentiate many available OOD sets w.r.t. a given in-distribution task to select the most appropriate one**, which in turn induces a model with a high detection rate of *unseen* OOD sets? To answer this question, we hypothesize that the "protection" level of in-distribution sub-manifolds by each OOD set can be a good possible property to differentiate OOD sets. To measure the protection level, we then design three novel, simple, and cost-effective metrics using a pre-trained vanilla CNN. In an extensive series of experiments on image and audio classification tasks, we empirically demonstrate the ability of an Augmented-CNN (A-CNN) and an explicitly-calibrated CNN for detecting a significantly larger portion of unseen OOD samples, if they are trained on the most protective OOD set. Interestingly, we also observe that the A-CNN trained on the most protective OOD set (called A-CNN*) can also detect the black-box Fast Gradient Sign (FGS) adversarial examples.

As the third contribution, we investigate more closely **the capacity of the A-CNN* on the detection of wider types of black-box adversaries**. To increase the capability of A-CNN* to detect a larger number of adversaries, we augment its OOD training set with

some inter-class interpolated samples. Then, we demonstrate that the A-CNN trained on the most protective OOD set along with the interpolated samples has a consistent detection rate on all types of unseen adversarial examples. Whereas training an A-CNN on Projected Gradient Descent (PGD) adversaries does not lead to a stable detection rate on all types of adversaries, particularly the unseen types. We also visually assess the feature space and the decision boundaries in the input space of a vanilla CNN and its augmented counterpart in the presence of adversaries and the clean ones. By a properly trained A-CNN, we aim to take a step toward a unified and reliable *end-to-end learning model* with small risk rates on both clean samples and the unusual ones, e.g. adversarial and OOD samples.

The last contribution is to show a use-case of **A-CNN for training a robust object detector on a partially-labeled dataset**, particularly a merged dataset. Merging various datasets from similar contexts but with different sets of Object of Interest (OoI) is an inexpensive way to craft a large-scale dataset which covers a larger spectrum of OoIs. Moreover, merging datasets allows achieving a unified object detector, instead of having several separate ones, resulting in the reduction of computational and time costs. However, merging datasets, especially from a similar context, causes many missing-label instances. With the goal of training an integrated robust object detector on a partially-labeled but large-scale dataset, we propose a self-supervised training framework to overcome the issue of missing-label instances in the merged datasets. Our framework is evaluated on a merged dataset with a high missing-label rate. The empirical results confirm the viability of our generated pseudo-labels to enhance the performance of YOLO, as the current (to date) state-of-the-art object detector.