Bayesian Inference Approach for Robust Deep Neural Networks

Name : Khong Thi Thu Thao
Laboratory' s name : Computing Architecture
Supervisor' s name : Professor Yasuhiko Nakashima

Abstract (should be within 1st page)

The rapid deployment of deep neural networks (DNNs) and deep learning algorithms have been proving their enormous potentiality in a wide range of computer vision and the field of recognition. Nonetheless, due to a vulnerability, deep learning models' ability to complicated situations requires a fundamental tool for computer security. Recent studies have been shown a vulnerability of DNNs by a small adversarial perturbation in images that humans cannot distinguish and a well-trained neural network can misclassify. Therefore, there are many defense methods to improve the robustness of DNNs against adversarial attacks, for example, adversarial detection, statistical properties of network parameters, the normalization of input data, adversarial training, etc. Among them, adversarial training is an outstanding defense, but it is a challenge with respect to real data and large DNNs.

In order to avoid adversarial training, we have proposed a defense algorithm named Bayes without Bayesian Learning (BwoBL). Our algorithm builds Bayesian Neural Networks (BNNs) based on pre-trained DNNs and focuses on Bayesian inference without costing Bayesian learning. The stochastic components of BNNs can prevent the forceful gradient-based attacks and generate the ensemble model to enhance the DNN performance. As an application of transfer learning, BwoBL can easily integrate into any pre-trained DNN, which is trained on both natural and adversarial data. We have investigated the application of BwoBL to a variety of DNN architectures, such as Convolutional Neural Networks (CNNs) and Self-Attention Networks (SANs). It is believed that, unless making DNN models larger, DNNs would be hard to strengthen the robustness to adversarial images. Our algorithm then employs scaling networks of CNNs and SANs, e.g. ResNet, EfficientNet, and SAN19 to construct BNNs against a diversity of adversarial attacks.

We assess the robustness of our BNN models by the top-1 accuracy on small datasets, i.e., CIFAR-10 and CIFAR-100, and the top-5 accuracy on real datasets like ImageNet. Our experiments utilize the currently strong attacks such as Projected Gradient Descent (PGD) and Carlini & Wagner (C&W) to produce adversarial examples. Experimental results have proved the efficiency of our BwoBL algorithm for resisting adversarial perturbation and solving the challenges of adversarial training and Bayesian learning.