論 文 内 容 の 要 旨

博士論文題目

Bayesian Inference Approach for Robust Deep Neural Networks
(ロバスト深層ニューラルネットワークのためのベイズ推定手法)

氏 名 Khong Thi Thu Thao

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.

氏　名	Khong Thi Thu Thao

（論文審査結果の要旨）（Ａ４　１枚　１、２００字程度）

本論文は、ディープニューラルネットワーク（DNN）とディープラーニングアルゴリズムの急速な展開が、幅広いコンピュータービジョンと認識の分野において大きな可能性を示してきた一方、複雑な状況に対する深層学習モデルの能力向上には、脆弱性向上の仕組みが必要である点に着目している。最近の研究では、人間には区別できない小さな敵対的摂動を含む画像が、十分に訓練されたニューラルネットワークでも誤分類する脆弱性が示されている。敵対的攻撃に対するDNNの堅牢性を向上させるためには多くの防御方法がある。たとえば、敵対的検出、ネットワークパラメータの統計的特性、入力データの正規化、敵対的トレーニングなどがある。この中で、敵対的トレーニングは優れた防御である。ただし、実データと大規模DNNに対する適用は計算時間の観点で困難が多い。本論文は、敵対的な訓練を回避するために、ベイズ学習なしベイズ（BwoBL）と呼ぶ防御アルゴリズムを提案している。本アルゴリズムは、事前にトレーニングされたDNNに基づいてベイズニューラルネットワーク（BNN）を構築し、ベイズ学習を必要とすることなくベイズ推定を行う。BNNの確率的コンポーネントは、強力な勾配ベースの攻撃を防ぎ、DNNの性能を強化するアンサンブルモデルを生成できる。転移学習のアプリケーションとして、BwoBLは、実データと敵対データの両方によりトレーニングされ、事前にトレーニングされたDNNに簡単に統合できる。畳み込みニューラルネットワーク（CNN）やセルフアテンションネットワーク（SAN）などのさまざまなDNNアーキテクチャに対する、BwoBLの適用可能性を調査している。DNNモデルを大きくしない限り、DNNは敵対的なイメージに対するロバスト性を強化するのは難しいと考えられる一方、本アルゴリズムは、CNNとSANのスケーリングネットワークを採用し、ResNet、EfficientNet、およびSAN19により、さまざまな敵対的攻撃に対するBNNを構築している。BNNモデルの堅牢性は、小さなデータセット（CIFAR-10とCIFAR-100）での上位1の精度と、ImageNetなど実データセットでの上位5の精度により評価している。実験では、Projected Gradient Descent（PGD）やCarlini＆Wagner（C&W）などの現在強力な攻撃を利用して敵対的な例を作成し、敵対的な摂動に対してベイズ学習を回避できるBwoBLアルゴリズムの効率を証明した。

以上、本論文は学術上、実際上寄与するところが少なくない。よって、本論文は博士（工学）の学位論文として価値あるものと認める。