Robust CNN with Differential Privacy

Xintong Hao CS591 S1 Boston University

https://github.com/XintongHao/Robust-CNN-with-Differential-Privacy

INTRODUCTION

Have you imagined if your machine learning model is attacked by adversary examples which will cause your model produce a wrong prediction? In this project, we will connect between robustness against adversarial examples and differential privacy (DP) to provide a certified defenses CNN classification model.

Dataset: MNIST handwritten digits.

Models: **Baseline CNN** with two 5x5 convolutional layers. **PixelDP CNN** one noise layer in the baseline model designed by Mathias Lecuyer[1].

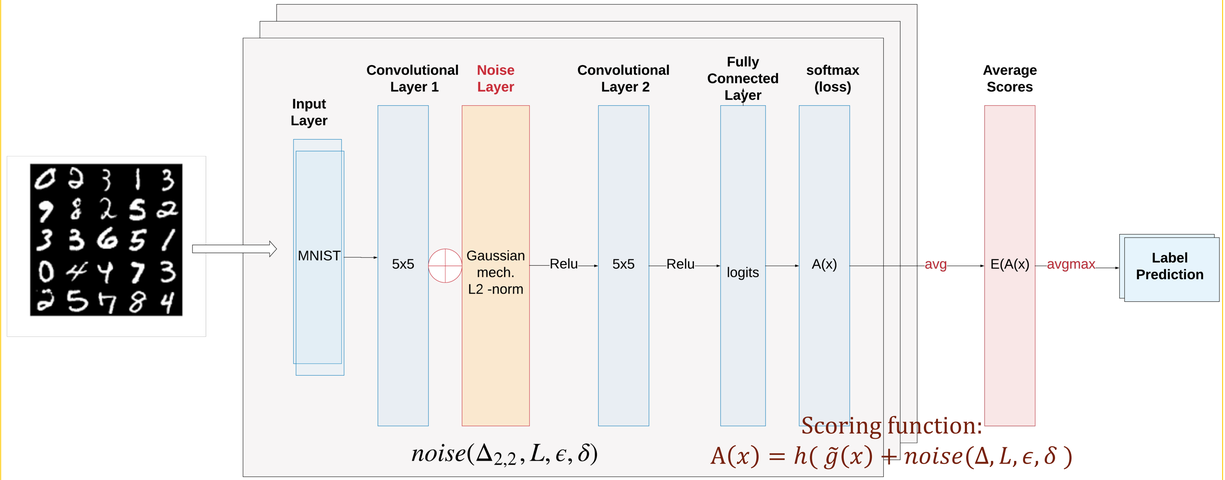
Evaluation Metric: Conventional accuracy, Precision on certified exapmples.

Attack Methodology: Carlini's L2-norm attack model.

DP-ROBUSTNESS[1]

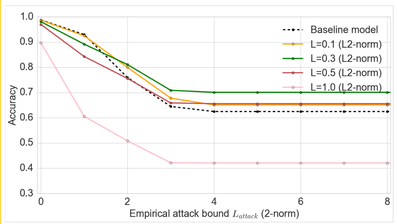
The idea behind using DP to provide robustness to adversarial examples is to create a DP scoring function such that, given an input example, the predictions are DP with regards to the features of the input. The approach in the paper is to transform a model's scoring function into a randomized $(\epsilon, \delta) - PixelDP$ scoring function, $\mathrm{A}(x)$, and then have the model's prediction procedure, use A's expected output over the DP noise $\mathrm{E}(\mathrm{A}(x))$ as the label probability vector from which to pick the argmax.

ROBUST CNN ARCHITECTURE

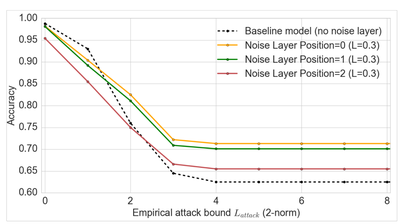


This is the architecture of PixelDP convolutional neural network model, where the blue layers are the original neural network and orange layer is the noise layer that provides the $(\epsilon, \delta) - DP$ guarantees. In our experiments, there are three options for noise layer placement: in the image, after the first layer and after the second layer. Each option, the noise distribution will be rescaled by the sensitivity $\Delta_{2,2} = 1$ of the computation performed by each layer before the noise layer. The model is trained with the original loss and Momentum SGD optimizer. Predictions repeatedly call the $(\epsilon, \delta) - DP$ model to measure its empirical expectation over the scores $\widehat{\mathrm{E}}\left(\mathrm{A}(x)\right)$.

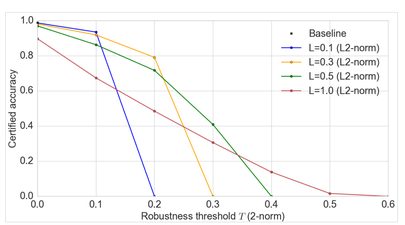
EXPERIMENTS RESULT ANALYSIS



(a) Accuracy under attack $L \in \{0.1, 0.3, 0.5, 1.0\}$



(b) Noise Layer Placement Position $\in \{0, 1, 2\}, L = 0.3$



(c) Certified Accuracy $L \in \{0.1, 0.3, 0.5, 1.0\}$

The conventional accuracy is the ratio of correct labels to all labels. We first measure the conventional accuracy of a defended model on Carlini's L2-norm attack[2] against samples in the testing set. The results shows PixelDP makes the model more robust to attacks. Fig (a) shows the accuracy of models with noise layer after the first layer. For large attack size, PixelDP model tends to have higher accuracy than baseline model. However with small attack size, the baseline model has a better performance. Fig (b) shows the accuracy of L=0.3 model with different noise layer placement. When the noise layer goes deeper in the network, the accuracy will drop, since the difficulty of sensitivity analysis. Certified accuracy is the fraction of the testing set on which a certified model's predictions are both correct and certified robust for a given prediction robustness threshold T. Fig (c) shows the certified robust accuracy bounds for PixelDP CNN model. First, according to the paper, this network yields more meaningful robust accuracy bounds on large networks like Inception. Second, PixelDP constructed for larger attacks L, tend to yield higher certified accuracy for high thresholds T.

REFERENCES

[1] Research Paper: Certified Robustness to Adversarial Examples with Differential Privacy
Code Source: https://github.com/columbia/pixeldp
[2] Research Paper: Towards Evaluating the Robustness of Neural Networks