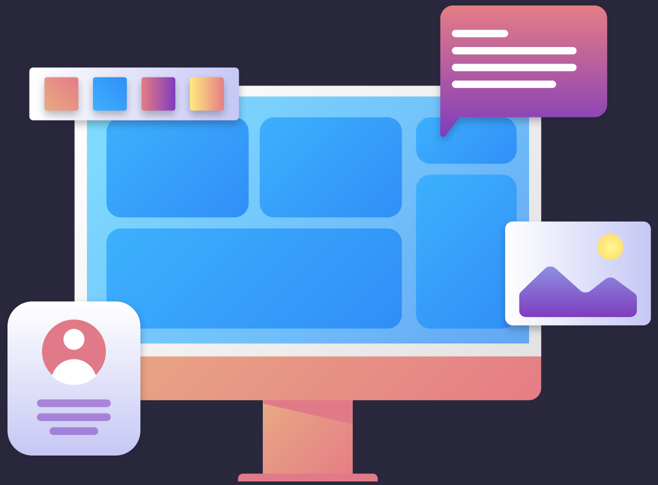
/ADVERSARIAL TRAINING IS ALL YOU NEED



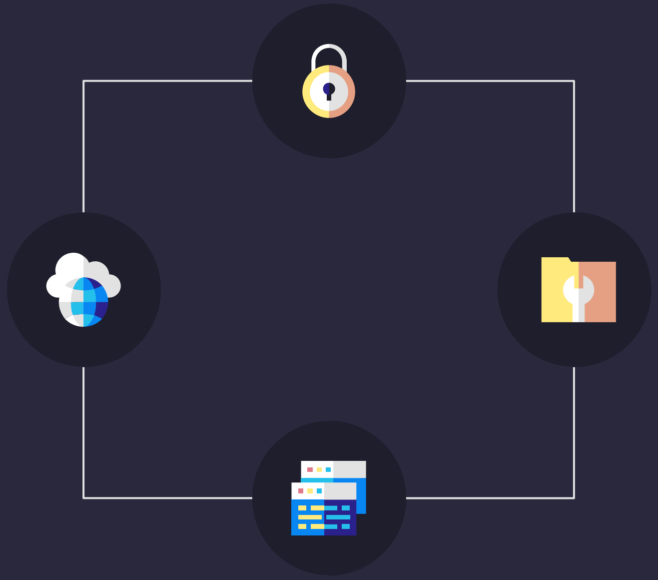
Group No. 1
Group Name: Four of a Kind

Group Members: Antreev Singh Brar, Anubhav Kalyani, Gurbaaz Singh Nandra, Pramodh Gopalan

Instructor: Prof. Priyanka Bagade

/Can Adversarial training defend against Poisoning attacks?

- Adversarially perturbed points have been shown to work as strong poisons.
- Whereas, adversarial training utilise adversarial samples generated from strong attacks to make the model more robust.
- Interesting to observe the results when one is tested against the other. We aim to do this in a computationally less expensive approach than existing research.



/Literature Review- I

- We reviewed the paper which had done previous work in our relevant domain BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain.

- We implemented their attack and used **our** defense on MNIST dataset.

- Poisoned MNIST by BadNets Attack:

  Before adversarial training,
  Effectiveness of poison: 97.34%

  After adversarial training,
  Effectiveness of poison: 20.55%

📓 Colab Notebook

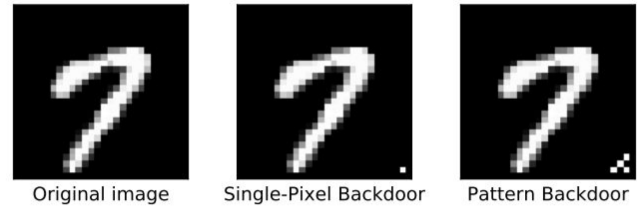


Figure 3. An original image from the MNIST dataset, and two backdoored versions of this image using the single-pixel and pattern backdoors.

https://arxiv.org/pdf/1708.06733.pdf

[1] Effectiveness of poison:

poison_acc = (poison_correct / poison_total) * 100

/Literature Review - II

- It is important to note that badnets only provide a maliciously trained network as an attack, and does not provide any defense.
- Our defense is robust to both poisoning attacks and adversarial attacks against a variety of threat models, not just Badnets and clean label attacks.

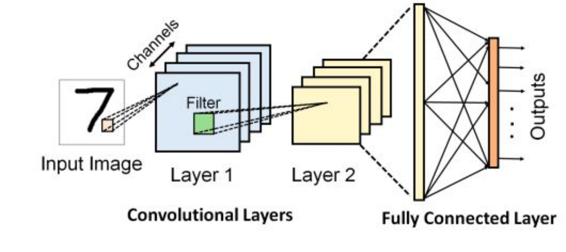


Figure 2. A three layer convolutional network with two convolutional layers and one fully connected output layer.

https://arxiv.org/pdf/1708.06733.pdf

/AT Defense on Clean Label Backdoor Attack - I

Before adversarial training,

- Normal test set accuracy = 98.26%

We poisoned the MNIST dataset using clean label backdoor attack.

- Accuracy on poisoned samples = 0.16%

After adversarial training,

We then implemented an Adversarial trainer, which used Projected Gradient descent attack.

- Accuracy on poisoned samples after adversarial training = 87%
- Normal test set accuracy after AT = 96.34%

(Note that accuracy on normal clean test set before and after the adversarial training differs by roughly 2%, which is within reasonable error range.)

Hence, AT achieves good accuracy on both clean label as well as poisoned samples.

/AT Defense on Clean Label Backdoor Attack - II

As suggested in weekly updates meeting, we are analysing our defense strategy performance over varying dataset i.e. we are training a vanilla network on MNIST dataset but for class 0 : during training, we only consider 500 samples and for the rest classes : 5000 samples. Then we train this imbalanced dataset of size 45500. And testing this on 10000 samples (i.e. 1000 samples of each of 10 classes).

Results

Test set accuracy on clean test set

    82.35%

Poison test set accuracy (model)

    82.45% for class 0 (#train-samples = 500)

📔 Colab Notebook

/Future Work

- Test our defense on CIFAR 10 dataset.
- Exhaustively evaluate and compare our performance against BadNets paper.
- Evaluate our Adversarial robustness.
- Analyse the performance of our strategy against varying % of poison samples.
- Groundwork over the mathematical explanation behind our proposed strategy.
- Use different methods of Adversarial Training such as TRADES and Helper Based AT.

Why?

/THANKS!

/We are open to questions and suggestions, if any.