/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

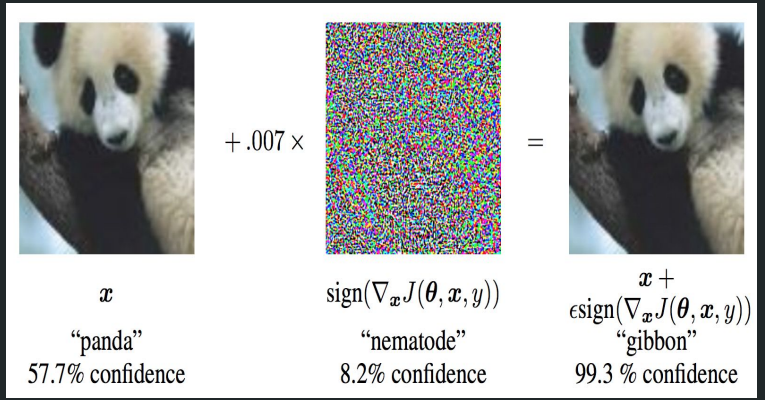
Instructor: Prof. Priyanka Bagade

Problem Statement

Can Adversarial training be used to Defend against Poisoning attacks ?

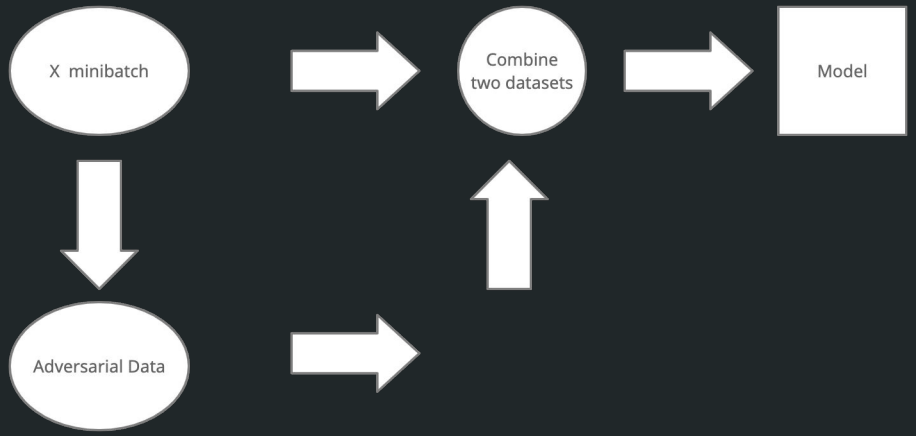
Adversarial Examples: Manipulating image inputs to a classifier which causes misclassification, but still is visually similar to the original image in human eyes.

Poisoning Attacks: attacker adds malicious training samples to the dataset. The model is then trained on this poisoned dataset and learns "wrong" features or backdoors, which can be later exploited during test time.



Problem Statement

Can Adversarial training be used to Defend against poisoning Attacks ?



It works remarkably well against diverse attacks and configurations, whereas newly proposed defenses are often broken through adaptive attacks. As the training progresses, the adversarial examples generated also keep changing, enabling the model to learn robust features.

Motivation

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.

We aimed to investigate the effect of adversarial training on the effectiveness of poisoning attacks. The central intuition behind our idea is that adversarial training causes the learning of robust features, which could be helpful when defending against poisoning attacks

To perform this we plan to perform Adversarial Training on our Models (vanilla CNNs and ResNets) and test them against Poisoning attack on various Datasets like CIFAR and MNIST.

State of the Art approach

We are comparing our training approach with the State of the Art Adversarial training.

MNIST Dataset

Natural	PGD
98.8%	93.2%

CIFAR-10 Dataset

Natural	PGD
92.7%	79.4%

Results

The entire project revolved around making hypothesis and testing then testing them against complex models and datasets for different attacks and defenses.

Here is a list of all the experiments and implementations performed in the course of project and the corresponding results.

- Implemented our own versions of Clean-Label Backdoor attack and Badnets attack using Keras. We also implemented Adversarial Training class using **Projected Gradient Descent Attack.**

- Balanced MNIST data

    - We poisoned the MNIST dataset using a clean label backdoor attack
    - We try to defend against it using Adversarial Training.

Results

- Unbalanced MNIST Data
    - 500 zeros and 5000 each of the other digits
      Normal test set accuracy (unpoisoned) on zeros = 97.45%
    - We poisoned the MNIST dataset using a clean label backdoor attack.
      Accuracy on clean zeros = 98.57%
      Accuracy on poisoned zeros = 0%
    - We then implemented adversarial training for 80 epochs, which used a Projected Gradient descent attack to generate Adversarial samples.

      Accuracy on clean zeros = 94.18%
      Accuracy on poisoned zeros = 92.65%

Results

MNIST Dataset

Clean label attack (Turner attack)

Test accuracy on Images with backdoor

Normal Training on Poisoned Data	Adversarial Training on Poisoned Data
1.39%	90.68%

Gu et al attack (Badnets)

Test accuracy on Images with backdoor

Normal Training on Poisoned Data	Effectiveness of poison after normal training on poisoned data	Adversarial Training on Poisoned Data	Effectiveness of poison after AT on poisoned data
1.17%	96.22%	91.71%	0.99%

Results

CIFAR - 10 Dataset

Gu et al (Badnets)

Test accuracy on Images with backdoor

Normal Training with Poisoned Data	Adversarial Training with Poisoned Data
8.64%	54.86%

Test accuracy on Images without backdoor

Normal Training with Poisoned Data	Adversarial Training with Poisoned Data	Effectiveness of poison after AT on poisoned data
74.26%	52.71%	5.32%

Results

CIFAR - 10 Dataset

Clean label attack (Turner attack)

Test accuracy on Images with backdoor

Normal Training with Poisoned Data	Adversarial Training with Poisoned Data
30*%	51.90%

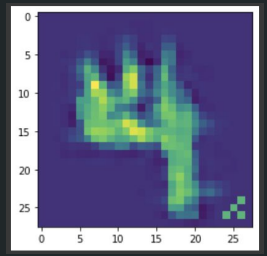
Test accuracy on Images without backdoor

Adversarial Training with Poisoned Data
58.96%

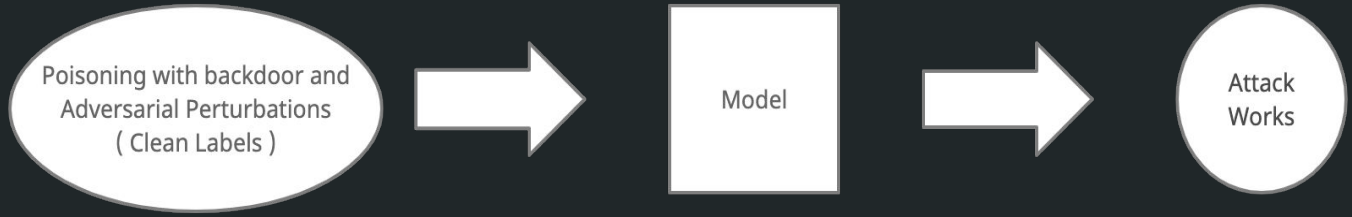
*Taken from https://people.csail.mit.edu/madry/lab/cleanlabel.pdf

Why the defense works on Clean label attacks ?

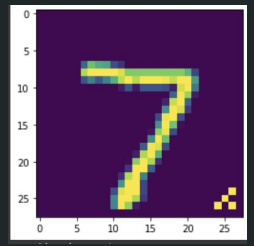
Turner et al propose an attack in which Training data is poisoned with backdoor and Adversarial Perturbations while maintaining the correct labels.



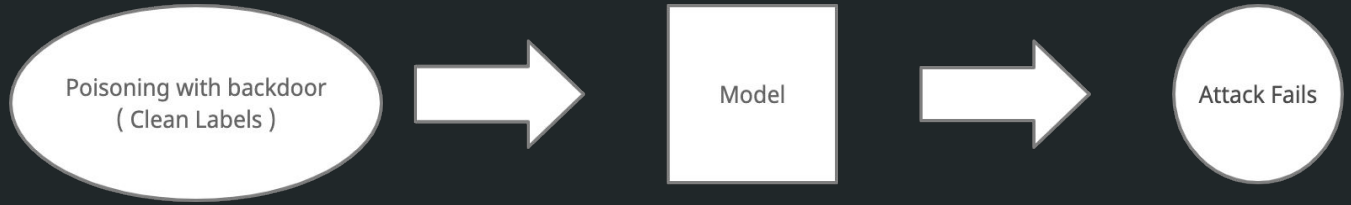
Label:9



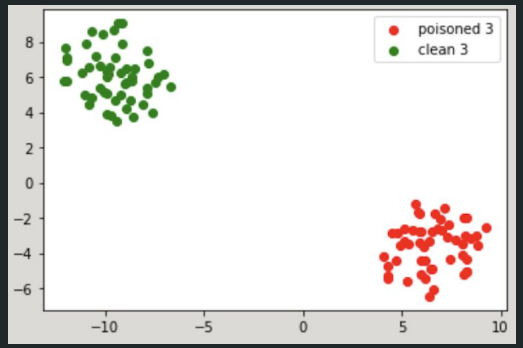
Turner et al explained in his paper that when a model's training data is infected with samples with backdoor patches (bogus pattern) while maintaining the correct labels, the attack fails and we obtain high accuracy on test set

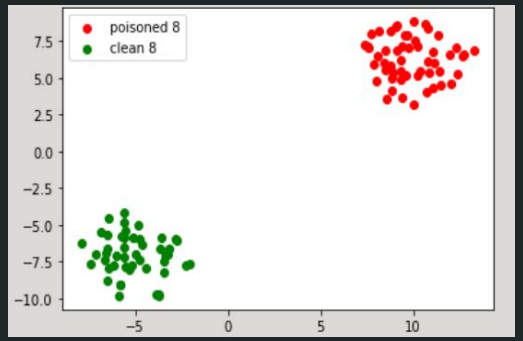


Label:7



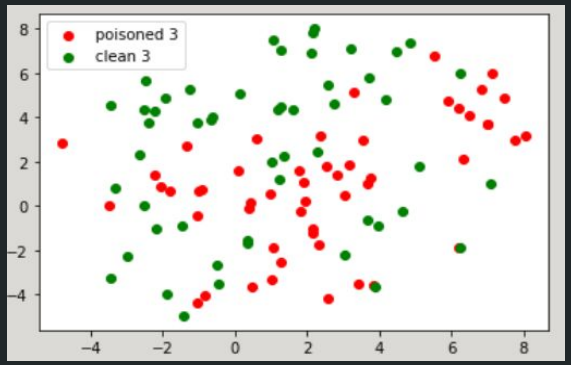
Why the defense works on BadNets attack ?

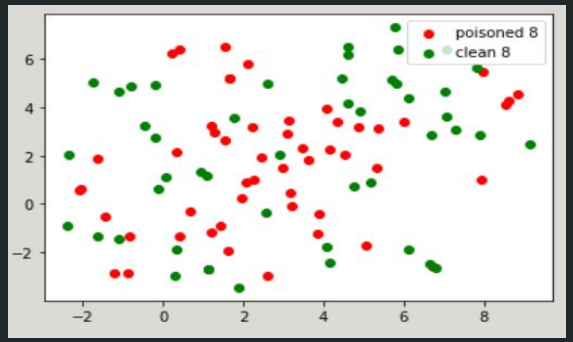




Network trained without AT, Poisoned data

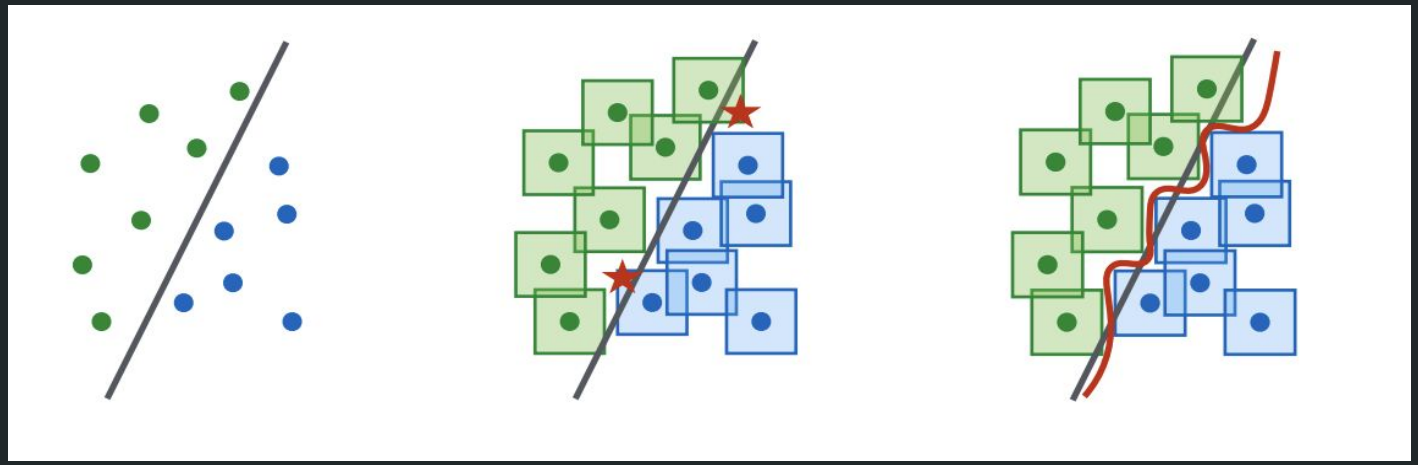
Last Layer Activations, visualized through TSNE





Network trained with AT, Poisoned data

Why we think defense with normal CNNs worked on MNIST, but not on CIFAR?



Normal Classifier without AT

Simple model when adversarially trained performs poorly

Require complex model to get better performance.

Using ResNets to train on CIFAR-10

- We used ResNet-18 model to adversarially train on CIFAR-10
- We chose to use ResNets because authors in the original AT paper (Madry et al) presented results using ResNets
- The forecasted time to train the model was 4.5 days given the computational hardware.
- We trained the network for 2 days, and we were not able to get better results than vanilla CNN.

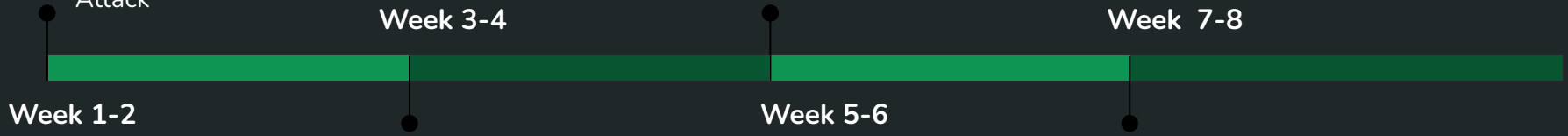
Timeline

Prologue

- Initial Ideation Report and Presentation
- Literature Review
- Reproducing baseline results and testing our approach on Badnets Attack

Implementation

- Of various attacks like PGD, FGM, Poisoning Backdoor and Clean Label Backdoor Attack, and Adversarial trainer from scratch
- Experimentation on slightly more complex datasets like CIFAR-10



Experimentation

- AT using Madry PGD on Clean Label Backdoor attack for MNIST
- Mid Term Presentation
- Running training simulations of Vanilla NN on different proportions of biased poisoned datasets

Final Destination

- Training on Deeper Models like ResNet for CIFAR-10
- Visualisations of Decision boundaries
- End Term Presentation
- Finalising Report and Code

Work Distribution

- ➔ Anubhav (25%)
  - Trained Model to get Baseline results on Poison Free Data
  - Performed Adversarial training on MNIST, Biased MNIST and CIFAR-10 using ART for Clean Label and Badnet Attacks.

- ➔ Antreev (25%)
  - Experimented with training of Model on Biased MNIST with Turner attack
  - Implemented the Projected Gradient Descent Attack and Fast Gradient attack from scratch in python
  - Developed the End term evaluation presentation

- ➔ Gurbaaz (25%)
  - Made the Mid term evaluation presentation
  - Implemented the Adversarial Trainer Class using PGD attack (proposed by Madry), and Clean Label Backdoor Attack and Badnets Attack.
  - Reproduced the results of custom attacks and training on MNIST dataset.

- ➔ Pramodh (25%)
  - Organised the workflow of the entire project and designed the approach to try and solve the problem statement.
  - Developed the Final and Initial Project report.
  - Performed Adversarial training on MNIST, Biased MNIST and CIFAR-10 using ART for Clean Label and Badnet Attacks.

References

- Adversarial Machine Learning in Image Classification: A Survey Towards the Defender's Perspective https://arxiv.org/pdf/2009.03728.pdf
- Dataset Security for Machine Learning: Data Poisoning, Backdoor Attacks, and Defenses https://arxiv.org/pdf/2012.10544.pdf
- BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain https://arxiv.org/pdf/1708.06733.pdf
- Towards Deep Learning Models Resistant to Adversarial Attacks https://arxiv.org/pdf/1706.06083.pdf
- Clean-Label Backdoor Attacks https://people.csail.mit.edu/madry/lab/cleanlabel.pdf
- Explaining and harnessing adversarial examples https://arxiv.org/pdf/1412.6572.pdf
- Poisoning attacks survey https://arxiv.org/pdf/2012.10544.pdf
- Adversarial ML in Image Classification: a survey https://arxiv.org/pdf/2009.03728.pdf

Acknowledgement

We would like to thank Prof. Priyanka Bagade for giving us an opportunity to work on a project in the course CS776A: Deep Learning for Computer Vision. We are grateful to have an instructor like her, and working on this project has definitely sparked a zeal in us to work on research projects in future.

We would also like to thank our project mentors, Soumya Banerjee and Tej Kiran for their invaluable and critical feedback on the project, and helping us with the roadblocks as and when they rose.

Thank You!

We are open to questions and suggestions, if any.