A deep dive into the robustness of AdaBoost Ensembling combined with Adversarial Training

Course: CSE3000
Author: Kanish Dwivedi
Supervisors: Chi Hong, Jiyue Huang
Responsible Professor: Stefanie Roos
Contact: K.Dwivedi-1@student.tudelft.nl

TU Delft

1. BACKGROUND

- Neural Networks are prone to adversarial attacks, causing them to misclassify
- Adversarial attack consists of inputting an adversarial image: one that is indistinguishable to the human eye, but is systematically different in terms of pixels [1]
- Common strong white-box attacks are gradient-based
- FGSM, PGD, BIM, C&W, Auto-PGD, CAA, Multitargeted, etc.

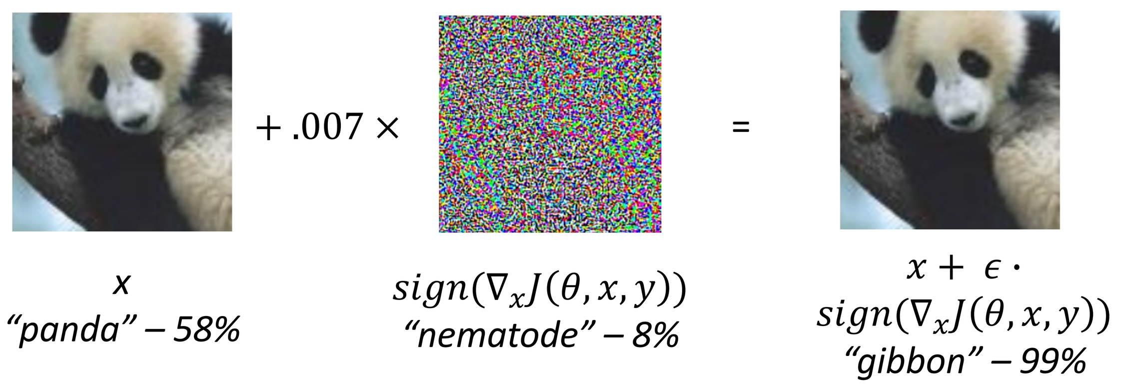

Figure 1: FGSM example attack [1]

2. RESEARCH QUESTION

How can AdaBoost ensemble learning provide adversarial robustness to white-box attacks when the "weak" learners are neural networks that do adversarial training?

ADABOOST ENSEMBLE

- Train lots of "weak" classifiers and combine them.
- Each learner focuses more on previous one's mistakes [3]

+

ADVERSARIAL TRAINING

- To defend against attacks, train on the adversarial examples
- The stronger attack we train on, the higher robustness we get [2]

Adven

Our multiclass AdaBoost ensemble method that does adversarial training. See Figure 2

3. METHODOLOGY

- Conducted many experiments → Exploring six different variables of Adven's training procedure to see effect on robustness (% defended attacks) to PGD attack and test set accuracy on MNIST dataset

(1) Adversarial algorithm used during training, (2) Loss function used during training, (3) Perturbation radii used during training, (4) Activation Function used during training, (5) Model Size of weak learner, (6) Number of learners in ensemble

4. CONCLUSION

- Adven ensemble provided greater robustness than a single learner in all tests
- Is computationally efficient: training time scales linearly with number of learners, the other variables add little or no additional training time
- Adven inherits known adversarial training characteristics, and extends them into an ensemble context and vice versa: a high number of high-capacity weak learners that train on strong attacks with high radii do best
- Adven ensemble exhibits greater resistance to the trade-off effect (sacrificing clean image accuracy for robustness) and prefers non-smooth activation function
- Same trends seen on Fashion-MNIST (except for model size and loss function)
- Best ensemble achieves 91.88% robustness to PGD attacks and has 96.72% test set on the MNIST dataset.

5. LIMITATIONS & FUTURE WORK

- Improve evaluation criteria: stronger attacks, harder datasets, black box attacks, compute attacks using entire ensemble
- Explore other ensemble learning algorithms
- Study effects of hyperparameters
- Other defense approaches combined with Adven

References:

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In International Conference on Learning Representations, 2015.

[2] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083, 2017.

[3] Jianfang Cao, Junjie Chen, Haifang Li, "An Adaboost-Backpropagation Neural Network for Automated Image Sentiment Classification", *The Scientific World Journal*, vol. 2014, Article ID 364649, 9 pages, 2014. https://doi.org/10.1155/2014/364649

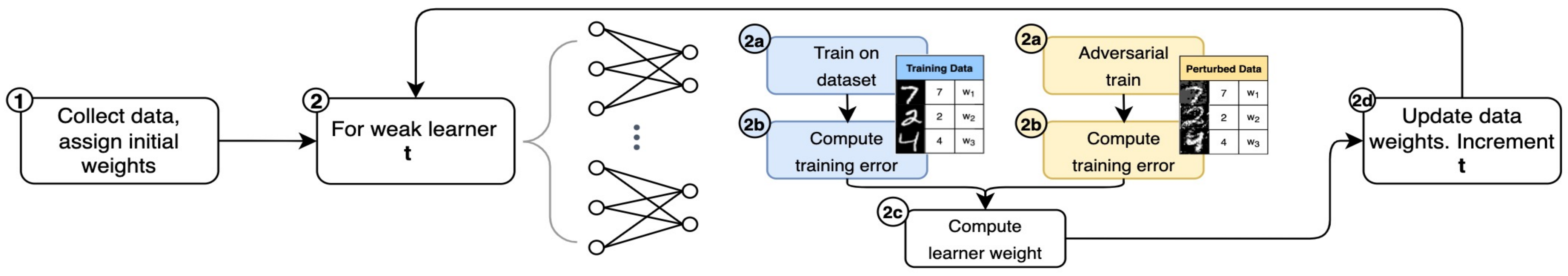

Figure 2: AdaBoost and Adven overview. (Yellow) Adven trains and computes errors on perturbed images. (Blue) AdaBoost does so on clean images

RESULTS FOR EACH VARIABLE

Adven Ensemble vs Single Weak Learner