Inclusion-Exclusion Enhanced Ensemble Adversarial Training

Group 6: Henry Shugart, Mikhal Ben-Joseph,
Yesh Munagala, Judy Chao

STOR 566 Fall 2022

Background: What is Ensemble Adversarial Training?

Ensemble AT is a training technique that utilizes several pre-trained models to produce transferable adversarial examples. These examples are then used to enrich training data for a new model.

Train N models of differing architectures on training data set X.

Create Adversarial Examples on Pretrained Models using PGD.

Train new model on adversarial examples. Traditional AT is performed 1/(N+1) batches.

Understanding the Problem

Ensemble Adversarial Training

Advantages

- Improves robustness against black box transfer attacks
- Decouples adversarial attack generation from training (Faster than traditional AT)

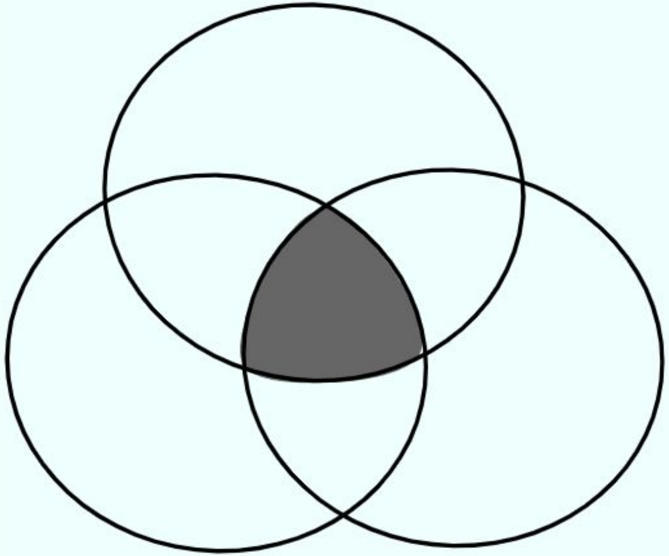
Vs

Disadvantages

- Decreased robustness to white-box attacks like FGSM, PGD
- Lower accuracy on natural images

GOAL: Build on Ensemble AT with Inclusion-Exclusion and Min-Max Optimization

Background: How Does Our Approach Differ?



Adversarial Space:

$$A_n = \left\{ \widehat{X} \in R^n \mid M(\widehat{X}) \neq y, \ |\widehat{X} - X| \leqslant \epsilon \right\}$$

Original coverage of the adversarial space with ensemble AT:

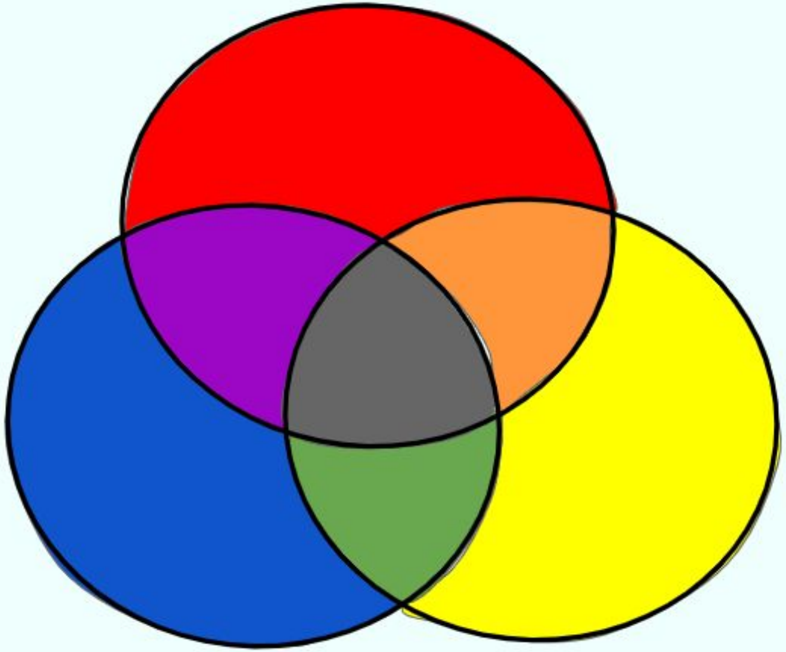
Each pre-trained model is fooled separately:

$$x_1 \in A_{M_1}, \ x_2 \in A_{M_2}, \ x_3 \in A_{M_3}$$

Because of the transferability of adversarial perturbations we expect the majority of these perturbations to exist in

$$x_1, x_2, x_3 \in A_{M_1} \cap A_{M_2} \cap A_{M_3}$$

Background: How Does Our Approach Differ?



Using Inclusion-Exclusion Idea:

Fooling different subsets each time

$$\{A_{M_1} \cap A_{M_2} \cap A_{M_3}^C\}, \{A_{M_1}^C \cap A_{M_2} \cap A_{M_3}\}$$

We do this by using an APGDA method proposed in Wang et al.

APGDA for Construction of Adversarial Examples

APGDA

An alternating projected gradient ascent descent algorithm was proposed in Wang et al. to solve min-max optimization problems of the form and showed improved attack success rates against multiple models

$$\max_{||\delta||_\infty \leq \varepsilon} \min_{w \in \omega} \sum_{i=1}^{N} w_i \ell(M_i(X+\delta), Y)$$
$$\omega = \{w \mid \mathbf{1}^T w = 1, w \geq 0\}$$

Inclusion Exclusion

For models being fooled we use

$$\hat{\ell}(\delta) = \ell(M_i(X+\delta), Y)$$

For models we do not intend to fool we use

$$\hat{\ell}(\delta) = \alpha - \beta \ell(M_i(X+\delta), Y)$$
$$\alpha, \beta \geq 0$$

We approximate the solution of the inner minimization with the sparsemax function proposed in Martins and Astudillo.

Method

1 **Generate adversarial examples for CIFAR-10**

$2^3 - 1 = 7$ combinations of 50,000 perturbed images

2 **Train a model with enriched training data**

Finetune pre-trained VGG-11 model

3 **Train the same model with traditional AT, regular EAT**

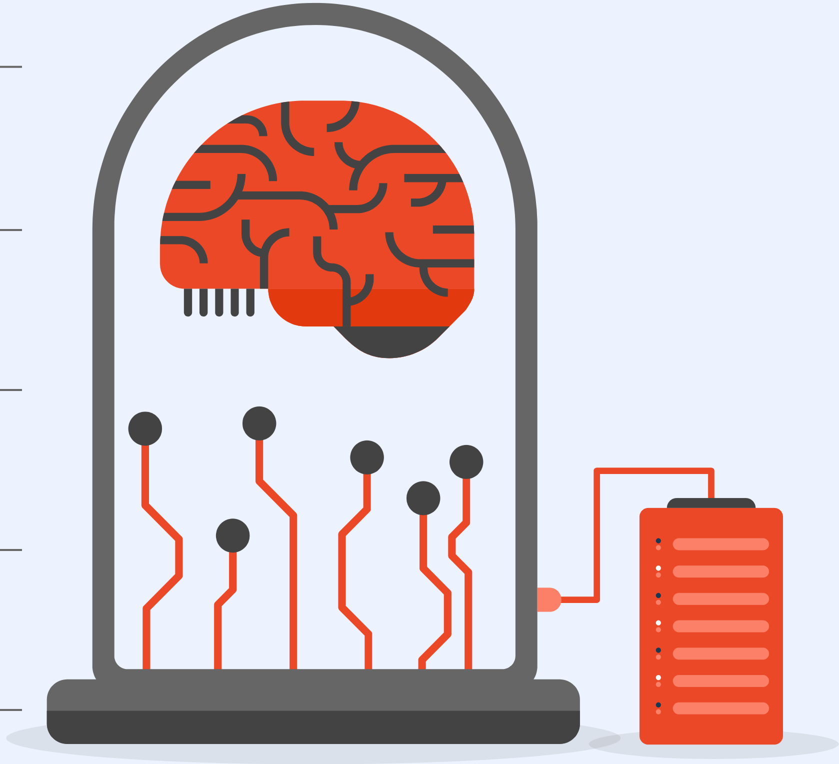
Produce baseline comparisons

4 **Implement black and white box attacks**

Transfer, FGSM, i-FGSM attacks from holdout models

5 **Evaluate Performance**

Compare attack performance on the four models

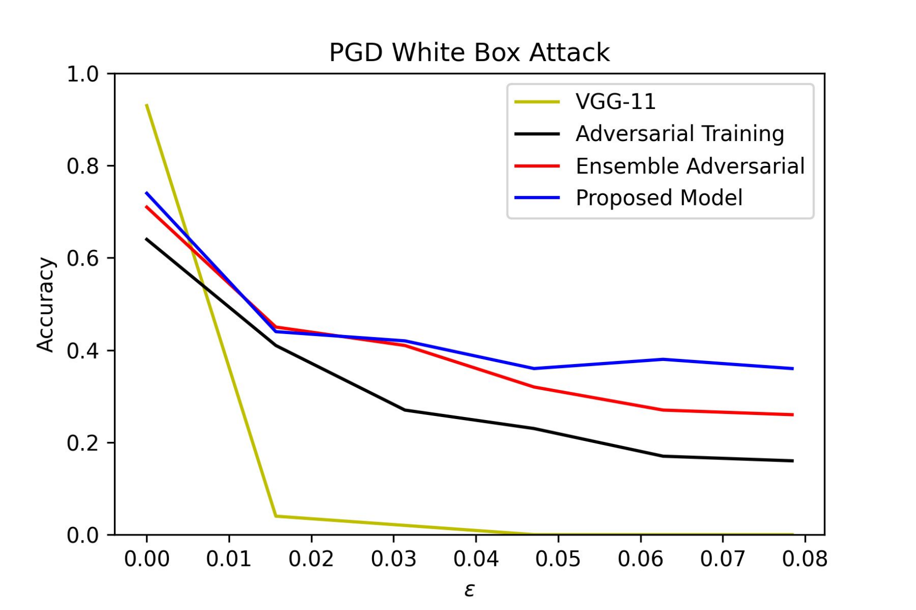


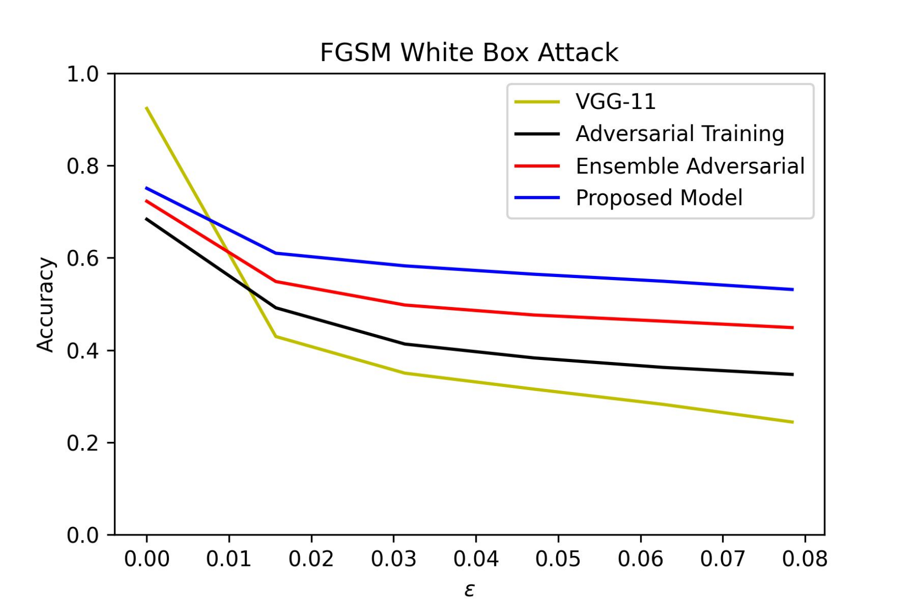
Before/After Adversarial Perturbations

+ 0.0625 x

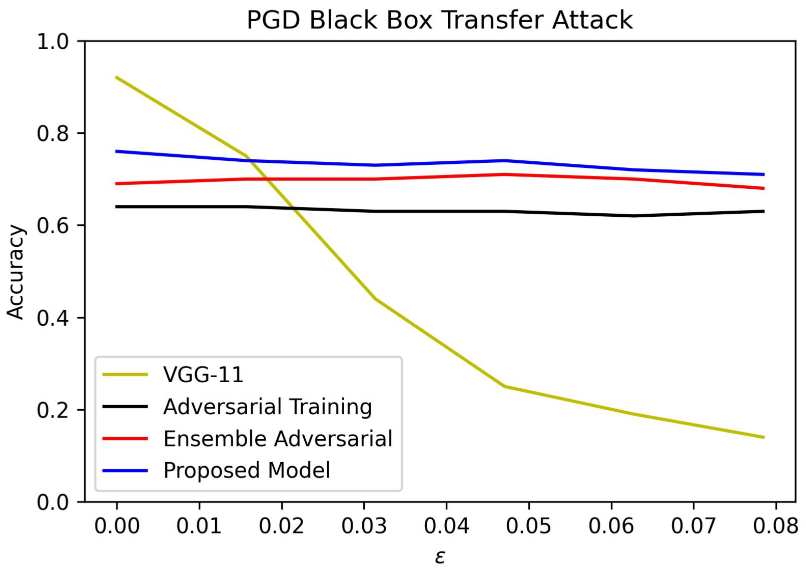
=

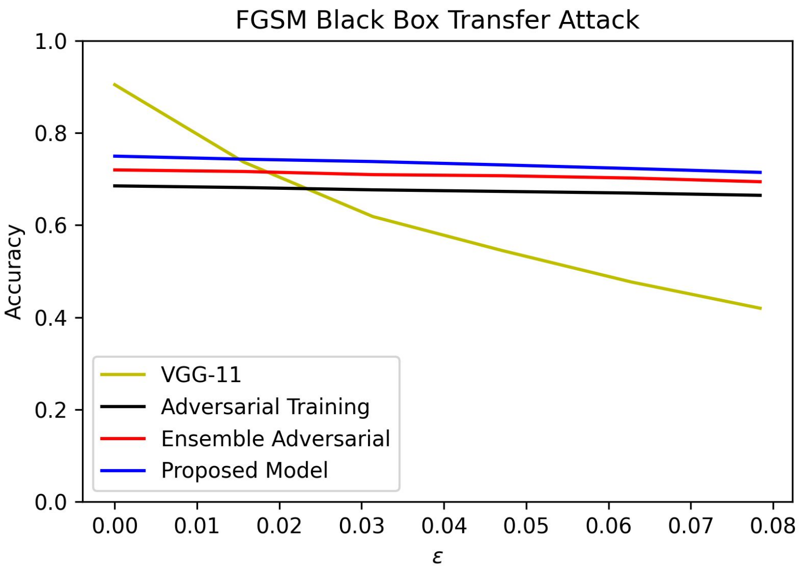
Results - White Box Robustness





Results - Black Box Robustness





Results - Training Time

Model	Training Time
VGG-11 MM-EAT (trained on our adversarial examples with ensemble AT)	1.8347 hours
VGG-11 EAT (regular ensemble AT)	2.4853 hours
VGG-11 AT (adversarially trained)	6.4425 hours

Conclusion and Future Work

Improved Defenses
Compared to both AT and EAT

01

Improved Training Time
Compared to both AT and EAT

02

Increase # of models used

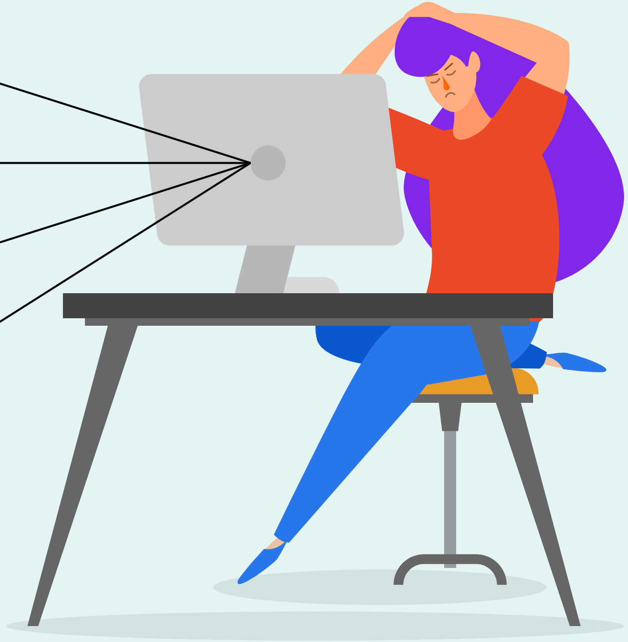
More models = more robust?

03

Apply to other datasets

MNIST, CIFAR-100, etc.

04



References

Martins, Andre, and Ramon Astudillo. "From softmax to sparsemax: A sparse model of attention and multi-label classification." International conference on machine learning. PMLR, 2016.

Phan, Huy. “PyTorch Models Trained on CIFAR-10 Dataset.” July 11, 2019. https://github.com/huyvnphan/PyTorch_CIFAR10. https://doi.org/10.5281/zenodo.4431043.

Tramèr, Florian, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. “Ensemble Adversarial Training: Attacks and Defenses,” May 19, 2017. https://doi.org/10.48550/arXiv.1705.07204.

Wang, Jingkang, Tianyun Zhang, Sijia Liu, Pin-Yu Chen, Jiacen Xu, Makan Fardad, and Bo Li. “Adversarial Attack Generation Empowered by Min-Max Optimization,” June 9, 2019. https://doi.org/10.48550/arXiv.1906.03563.