CSCE 585 Project

What to do

- Build an adversarial machine learning system on top of Athena.
- By calling APIs provided by Athena or extending Athena, students
  - Generate adversarial examples.
  - Build ensemble-based adversarial defenses.
- Athena is a framework for building ensemble-based defenses.
  - Implemented on top of IBM ART (Adversarial Robustness Toolkit).
  - Provided wrapper for keras- and pytorch- based classifiers.
  - Ensemble as a single classifier.
  - Implemented 11 adversarial attacks.

Provided

- Source code of Athena.
- The vanilla version of Athena which consists of 72 cnn weak defenses.
- The undefended model in the Athena paper (cnn model).
- The baseline adversarial examples that were generated in the Athena paper.
- Simple tutorials regarding (1) loading a single model; (2) loading an ensemble; (3) applying a transformation; (4) predicting an input; (5) generating adversarial examples.
- 72 SVM weak defenses + 1 SVM undefended model are provided for Hybrid Athena (option 4 for task 2).

Teams

- 10 groups. 3 – 4 students per group.
- Approaximately equal contributions to each task.
- One submission per team.
- Claim for your task 2 here on piazza.

Submission

- ALL MATERIALS that enable an independent group to replicate the results.
  - Code (team GitHub repo)
  - Experimental results, such as AEs, defenses, trained models, logs, etc. (a server we provided)
  - Project report (in Jupyter Notebook, team GitHub repo)
- One submission per group.

Reports

- Task 1 and Task 2
  - Approaches that are used in the task.
  - Experimental settings --- technique details to replicate your experiment. Such as the configuration for tunable parameters of an attack method. Hardware on which the experiment is run, if you will evaluate the overhead of your approaches.
  - Necessary citations.
  - Write the report in your own words.

Reports

- Task 3
    - The analysis methods that are used in the task.
    - Analysis results, observations, and results. Possible enhancements, future works, etc.
    - Architecture of your machine learning system (for the whole project).
    - Necessary citations.

Overview

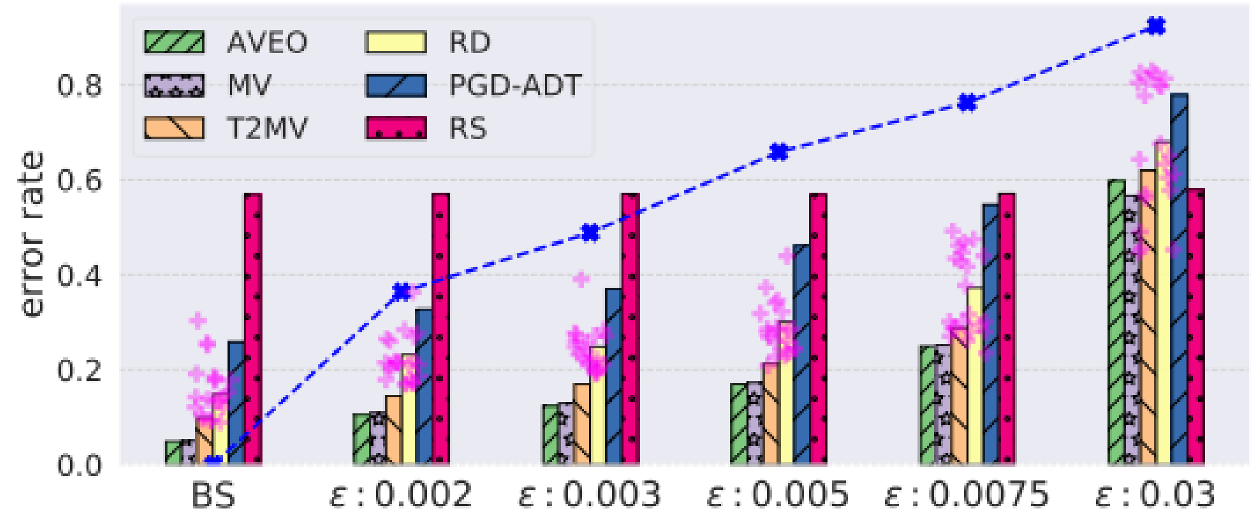
- 3 tasks in total.
- 2 security tasks + 1 competition task.
- Around 3 weeks per task.
- Schedule
  - Task 1: 10/05 - 10/18 (midnight)
  - Task 2: 10/19 - 11/01 (midnight)
  - Task 3: 11/09 - 11/30 (midnight)

Task 1

- Weights: 30% + 5%; Duration: 10/05 - 10/18 (midnight)
- Essential task for all groups.
- Generate adversarial examples in the context of zero-knowledge model.
  - 2 - 3 different attack methods (e.g., CW, PGD, FGSM, etc.).
  - Around 5 variants for each method. For example, FGSM eps = 0.1, 0.15, 0.2, 0.25, 0.3. PLEASE try configuration different from the baseline AEs.
  - Use the methods provided by Athena.
  - 5% bonus for methods not implemented in Athena.

Task 1

- Evaluation
  - Successful rate of the crafted adversarial examples on the UM, Valina Athena (MV and AVEP strategies), PGD-ADT classifier. Or the models' error rates against such adversarial examples.



(1) FGSM

Task 2

- Weights: 50% + 10 - 20%; Duration: 10/19 - 11/01 (midnight).
- 4 options for task 2. Please claim on piazza before 10/15.
- At most 3 groups for each task option. First come, first served.

Task 2 – Option 1

- Weights: 50% + 10%
- Generate adversarial examples in the context of white-box attack.
  - 1 – 2 attack methods. Several variants for each attack.
  - 2 optimization-based approaches provided by Athena.
    - Accrument weak defenses' loss.
    - Randomly synthesize an input then accrument weak defenses' loss.
  - I will consider 5% additional bonus for approaches not provided by Athena, so 50% + 10% + 5%.

Task 2 – Option 1

- Evaluation
  - Successful rate of the crafted adversarial examples on the Valina Athena (MV and AVEP strategies), PGD-ADT classifier. Or the models' error rates against such adversarial examples.
  - Overheads of your approaches.
  - Distortion of the crafted adversarial examples. Compare the distortions with that of the adversarial examples crafted in the context of zero-knowledge threat model.
  - Compare the effectiveness between different adversarial example variants.

Task 2 – Option 2

- Weights: 50% + 15-20%
- Learning-based ensemble strategy.
- Train a machine learning model.
  - Strategy model S: predictions --> y_pred
  - Training set D = {(predictions, y_true)}.
  - Several defense variants if possible, by tuning the parameters.
- [15%] Train a machine learning model, e.g., SVM, Random Forest, etc.
- [20%] Adaptive Multi-Column Deep Neural Networks with Application to Robust Image Denoising. Forest Agostinelli, Michael R. Anderson, and Honglak Lee. NIPS 2018.
- [20%] Knowledge distillation (TBD. I will check if this is feasible.)
- [20%] Other study.

Task 2 – Option 2

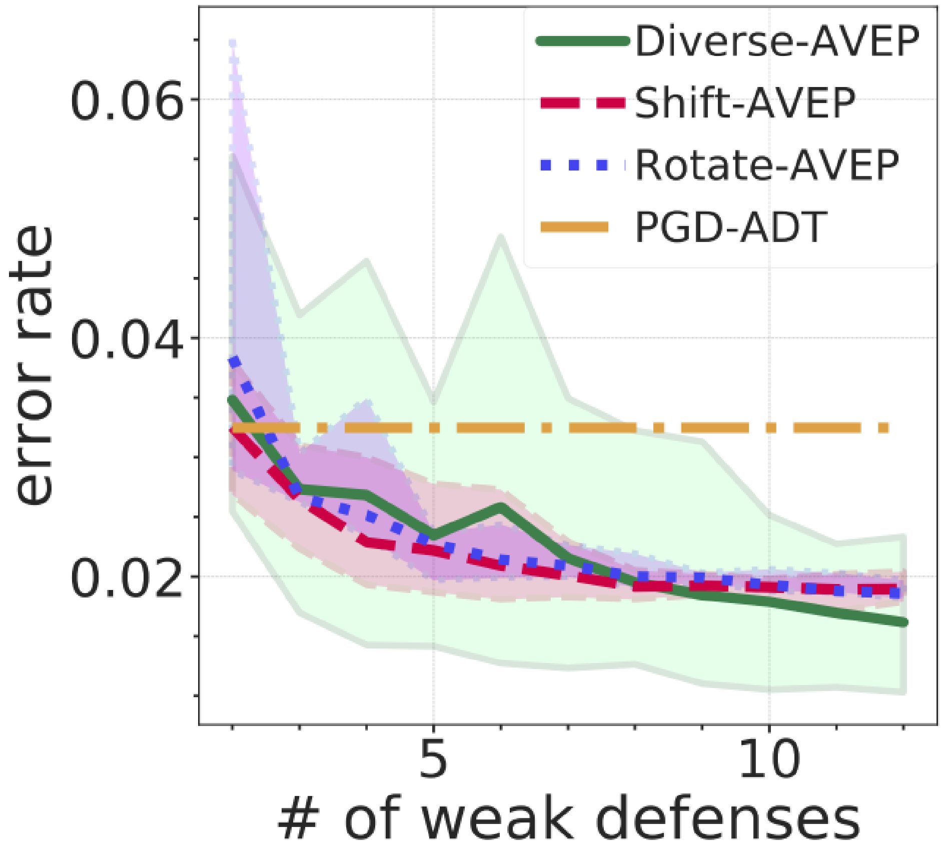
- Evaluation
  - Models' effectiveness against the adversarial examples generated in Task 1.
  - Models' effectiveness against the baseline adversarial examples.
  - Compare the effectiveness between different variants.
  - Compare with baseline models: the Vanilla Athena, PGD-ADT.

Task 2 – Option 3

- Weights: 50% + 15%
- Probabilistic Athena
  - Train a library of probabilistic weak defenses (e.g., Bayesian Neural Networks). Aim to train 10 - 20 weak defenses.
  - Weak defenses should be trained on the MNIST training data only.
  - Build an ensemble from the probabilistic library, using at least MV and AVEP strategies.
  - Several variants if possible. For example, different strategies, various ensemble sizes, etc.

Task 2 – Option 3

- Evaluation
  - Models' effectiveness against the adversarial examples generated in Task 1.
  - Models' effectiveness against the baseline adversarial examples.
  - Compare the effectiveness between different variants.
  - Compare with baseline models: the Vanilla Athena, PGD-ADT.



(6) PGD(ϵ : 0.11)

Task 2 – Option 4

- Weights: 50% + 10 – 20%
- Hybrid Athena
  - Build ensemble defenses from a hybrid Athena.
  - Around 10 variants with various sizes, various ratio of different type of models. For example, an ensemble consists of 50 weak defenses (60% cnn + 40% svm).
- [10%] Randomly select (I.e., manually designed) n weak defenses from the library for the ensemble.
- [20%] Select n weak defenses via some search-based approaches. For example, greedy search for n weak defenses that produce the maximal entropy. There are many possible metrics (e.g., ensemble diversity).
- [20%] Other synthesis approaches (selecting weak defenses automatically based on some metric).

Task 2 – Option 4

- Evaluation
  - Models' effectiveness against the adversarial examples generated in Task 1.
  - Models' effectiveness against the baseline adversarial examples.
  - Compare the effectiveness between different variants.
  - Compare with baseline models: the Vanilla Athena, PGD-ADT.

Task 3

- Weights: 20%; Duration: 11/09 - 11/30 (midnight)
- Ying will start an automatic cross-evaluation on the generated adversarial examples (task 1 and task 2 option 1) and the trained/built ensemble defenses (task 2 options 2 - 4) during 11/02 -- 11/08, and then public the evaluation results.

Task 3

- Each team can down the experiment results then investigate and analyze the data.
  - Aim to seek insights and/or theoretical explanations of
  - (for defenses) why and why not your approach are effective/ineffective against an attack.
  - (for attacks) why and why not your approach successfully/failed to fool a defense.
  - Any other necessary analysis.