CSCE 585
Athena project

Team JiR:

Jacob Vincent, Isaac Keohane, & Raul C. Ferraz

UNIVERSITY OF
South Carolina

Problem statement

- Overall motivation
  - Many machine learning models and neural networks are vulnerable to adversarial examples (AEs).
    - real-world data is never perfect, it has been transformed or is noisy etc.
  - Models misclassify examples that are only slightly different from true “clean” examples
- How can we implement a defended model to better handle transformed input data (adversarial examples)?

Problem statement (specifics)

- Can we generate adversarial examples?
- Using AEs, can we train a defended model and evaluate its ability to handle input data that has been altered in various ways?
  - Athena
    - Combines several models that have been trained for different AEs, each one a weak defense, into a ensemble model

Technical challenges - adversary

- Vanishing gradients (Arjovsky & Bottou, 2017)
  - If the discriminator is too good, then generator training can fail due to vanishing gradients.
- Mode collapse
  - The generators rotate through a small set of output types.
- Failure to converge (Goodfellow, 2014)
  - As the generator improves with training, the discriminator performance gets worse because the discriminator can't easily tell the difference between real and fake.

Technical challenges - defense

- Creating a strongly defended model can be technically challenging because it is:
  - Hard to create training data that perfectly emulates what the attack could be
  - Input data could be many things but when you train a model you need to give it a finite amount of adversarial examples

These two factors mean it is impossible to train for every type of input data that the model might see

Related works

- 2014

Intriguing properties of neural networks

Christian Szegedy, Google Inc.
Wojciech Zaremba, New York University
Ilya Sutskever, Google Inc.
Joan Bruna, New York University
Dumitru Erhan, Google Inc.
Ian Goodfellow, University of Montreal
Rob Fergus, New York University, Facebook Inc.

- 2015

EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES

Ian J. Goodfellow, Jonathon Shlens & Christian Szegedy
Google Inc., Mountain View, CA

- 2017

COUNTERING ADVERSARIAL IMAGES USING INPUT TRANSFORMATIONS

Chuan Guo*, Cornell University
Mayank Rana & Moustapha Cissé & Laurens van der Maaten, Facebook AI Research

- 2018

Enhancing Robustness of Machine Learning Systems via Data Transformations

Arjun Nitin Bhagoji, Princeton University
Daniel Cullina, Princeton University
Chawin Sitawarin, Princeton University
Prateek Mittal, Princeton University

Detecting Adversarial Examples through Image Transformation*

Shixin Tian, Guolei Yang, Ying Cai
Department of Computer Science, Iowa State University
{stian,yanggl,yingcai}@iastate.edu

- 2020

ATHENA: A Framework based on Diverse Weak Defenses for Building Adversarial Defense

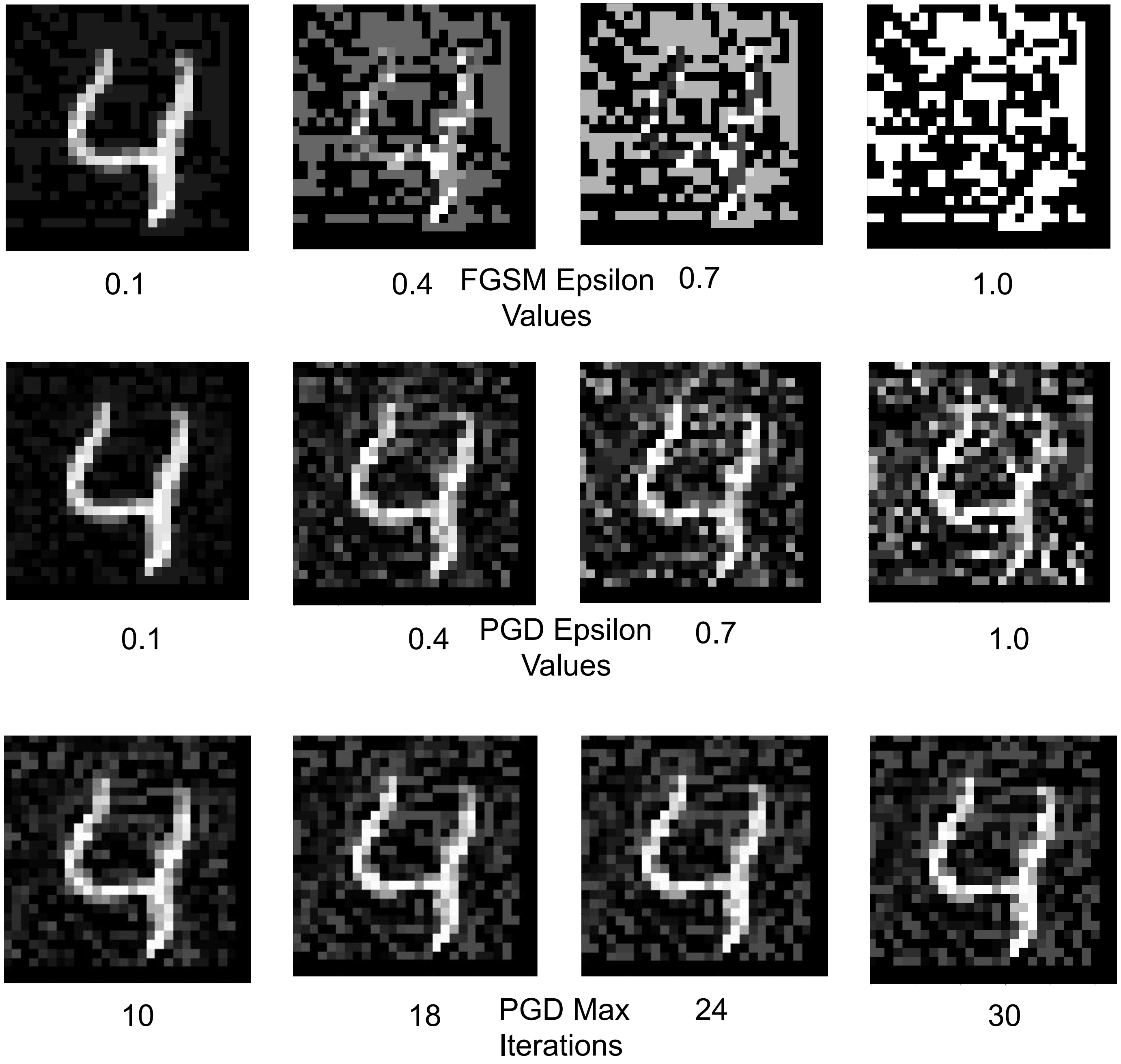
Ying Meng, Jianhai Su, Jason M. O'Kane, Pooyan Jamshidi
Departmetn of Computer Science and Engineering
University of South Carolina
Columbia, SC, USA

Our approach: Task 1

- Initially we generated adversarial examples using three different attack methods, all gradient-based attacks. Two were retained:
  - Fast Gradient Sign Method (FGSM)
    - We chose values of ϵ in the range of 0.1 to 1 in increments of 0.1 (ϵ = 0.1, 0.2, ..., 1.0
  - Projected Gradient Descent (PGD)
    - For the PGD, we generated attacks in two ways:
    - Manipulating the size of the perturbation, ϵ , from 0.1 to 1 in increments of 0.1 with fixed maximum iteration of 10.
    - Manipulating the number of maximum iterations from 10 to 30 by increments of 2 (10,12,14,...,30) with a fixed ϵ of 0.3

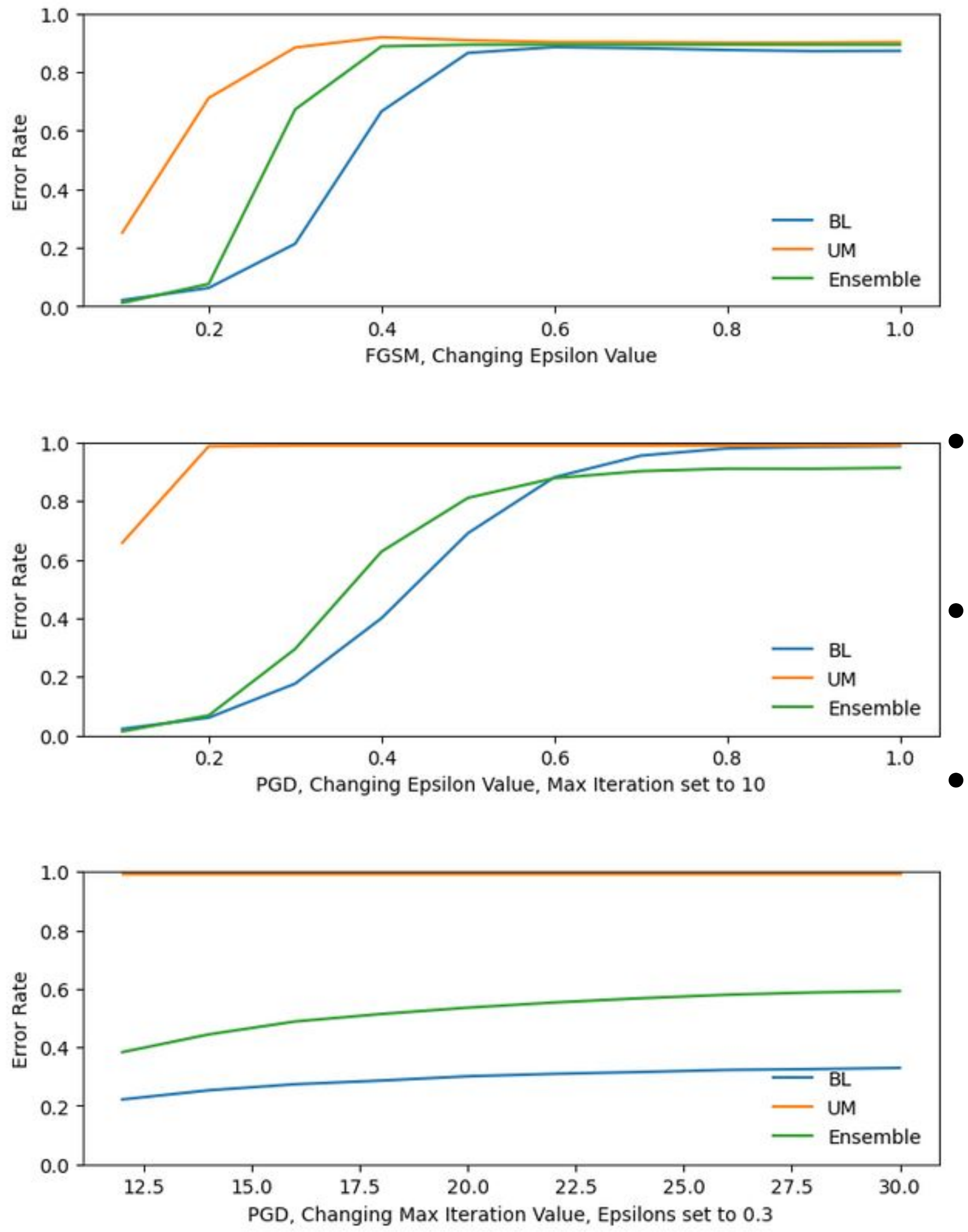
UofSC South Carolina

Our approach: Task 1



- Generated 30 adversarial examples to be tested against the vanilla, undefended, and ensemble Athena models.

South Carolina



Task 1 Results

- PGD and FGSM epsilon had a significant effect on the model error rates and visual differences in the AE images themselves.
- This makes sense: as the controlling parameter is increased, the effect of the AE increases too.
- Changes to max iterations on PGD had little effect on both the error rate and visual changes to the AE images. This suggests that it is less of a controlling factor than epsilon.

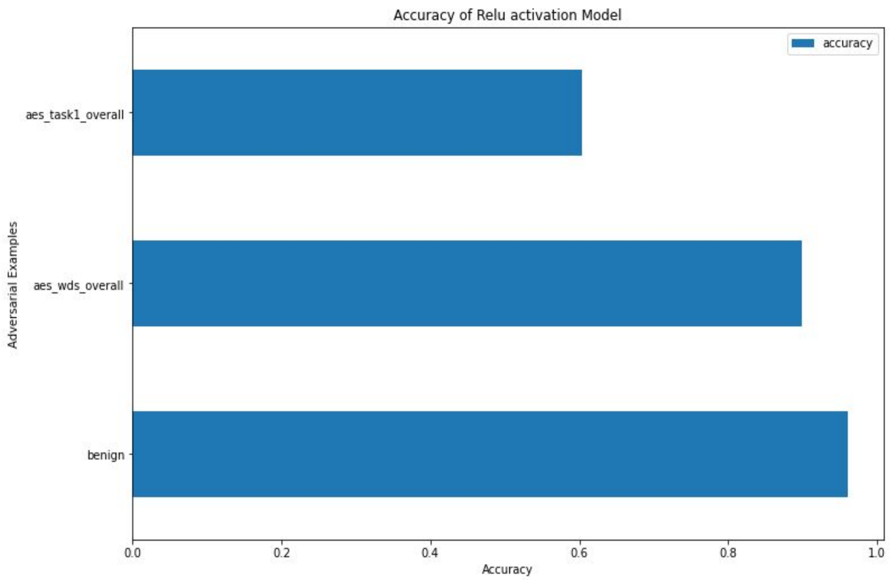
UofSC South Carolina

Our approach (cont.) Task 2

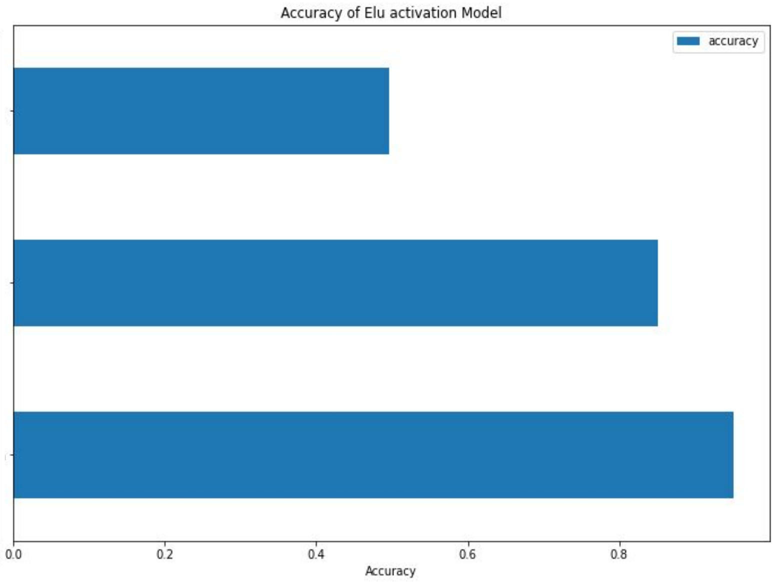
- Collect predictions from a weak defense ensemble in Athena.
- Train our own machine learning model using those ensemble predictions as the training labels.
- Evaluated our model against a) benign samples (unaltered images), b) adversarial examples from the weak defenses, and c) the adversarial examples (AEs) generated in Task 1.
- Test different activations as well

Results: model evaluation

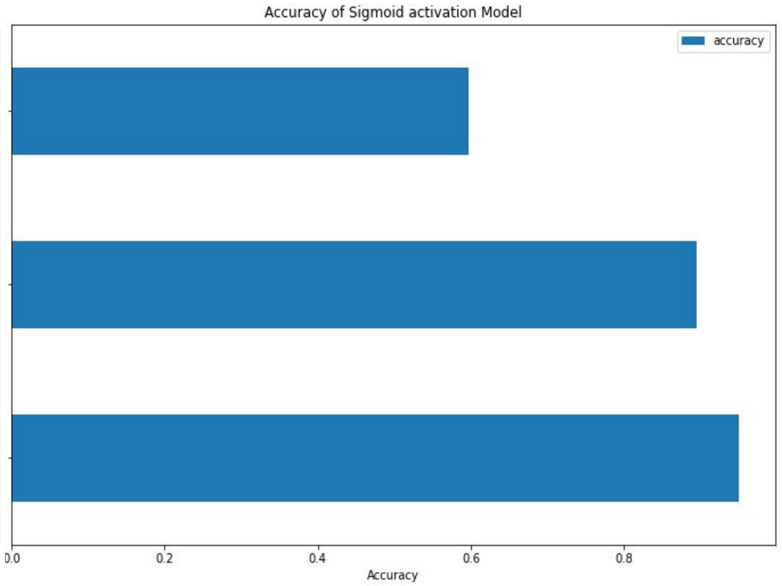
Task1 AEs (top), AEs for ensemble models (middle), benign (bottom)



Relu



Elu

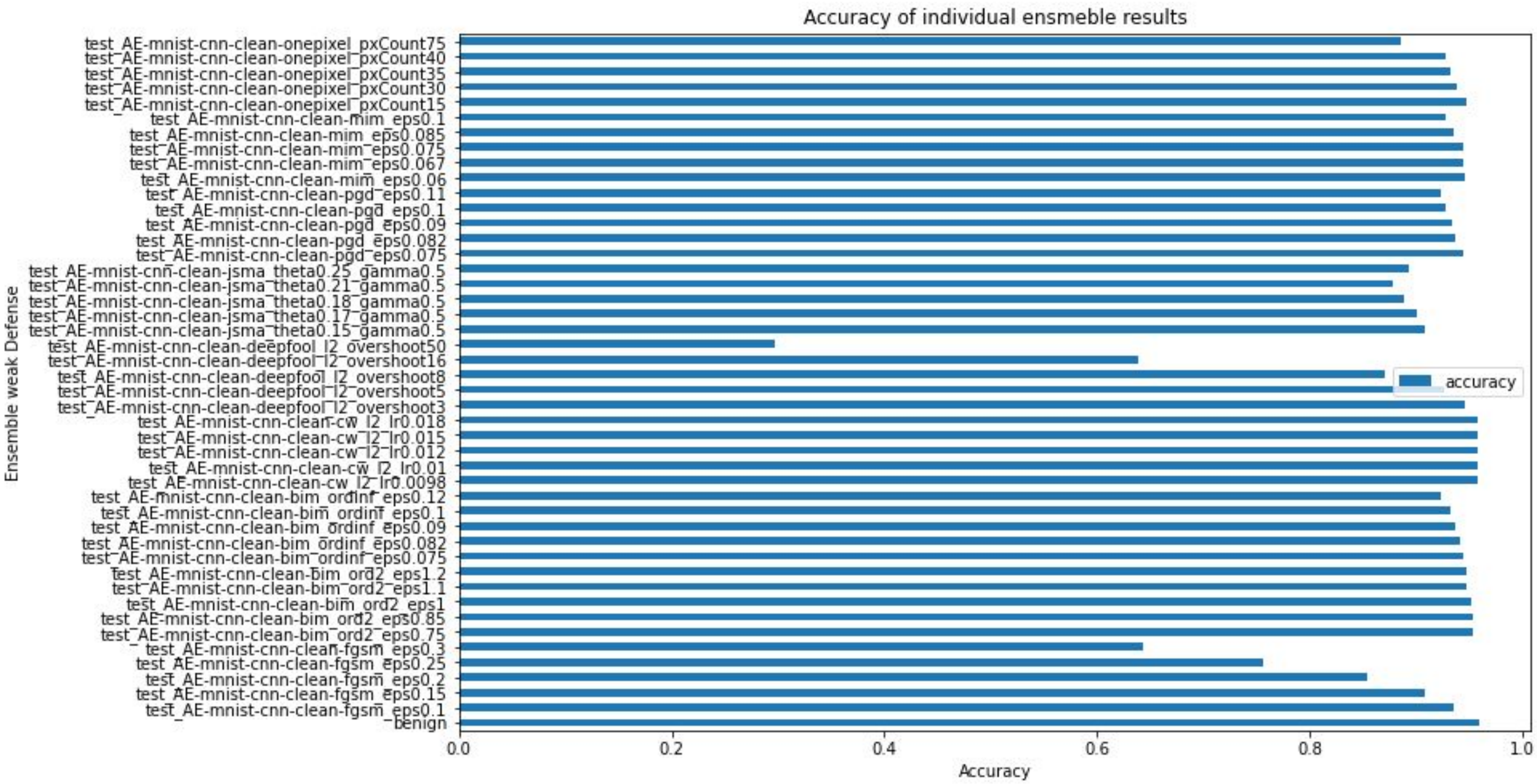


Sigmoid

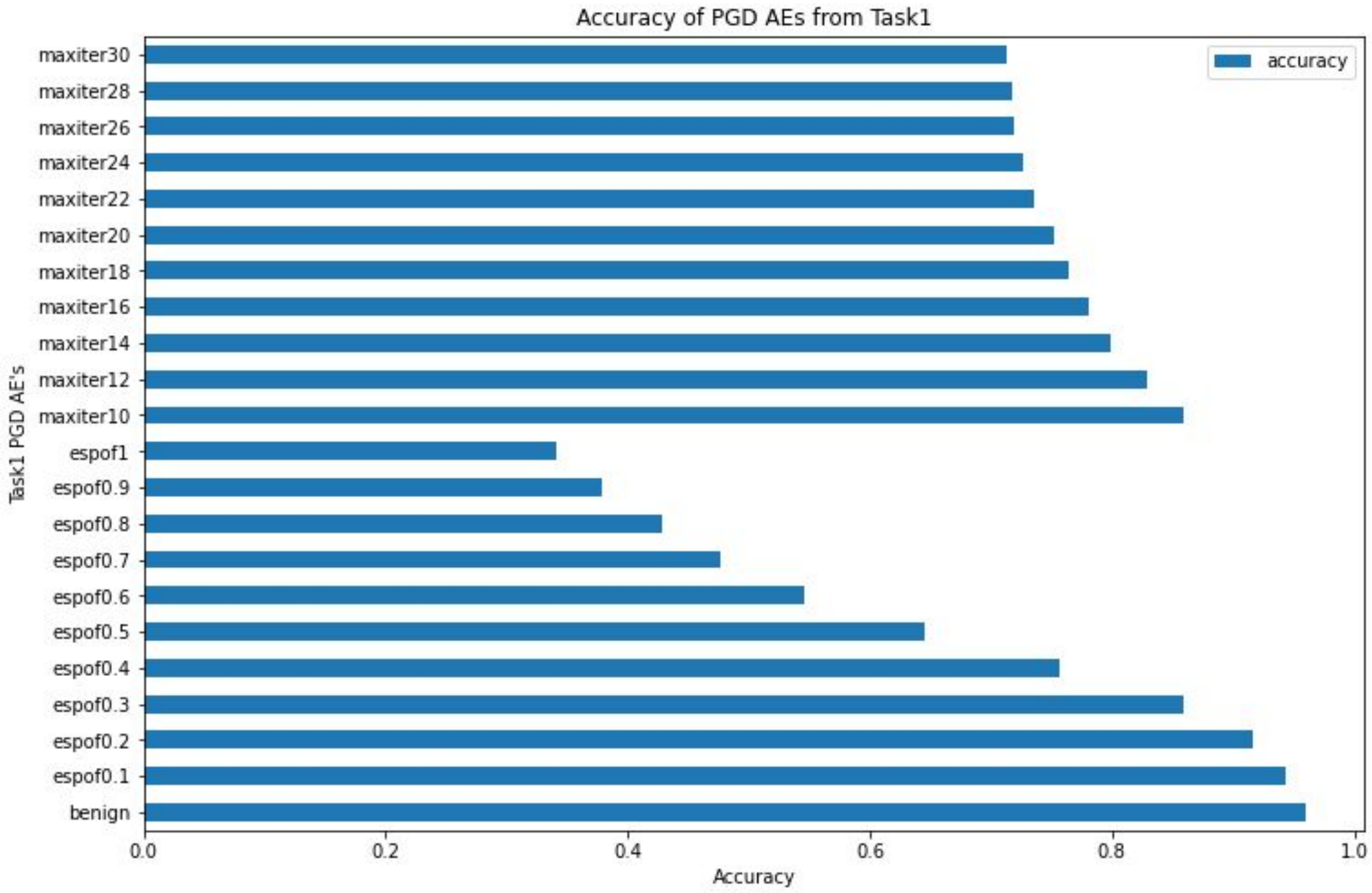
UofSC South Carolina

Results: model evaluation:

Accuracy against AEs for individual models in the ensemble

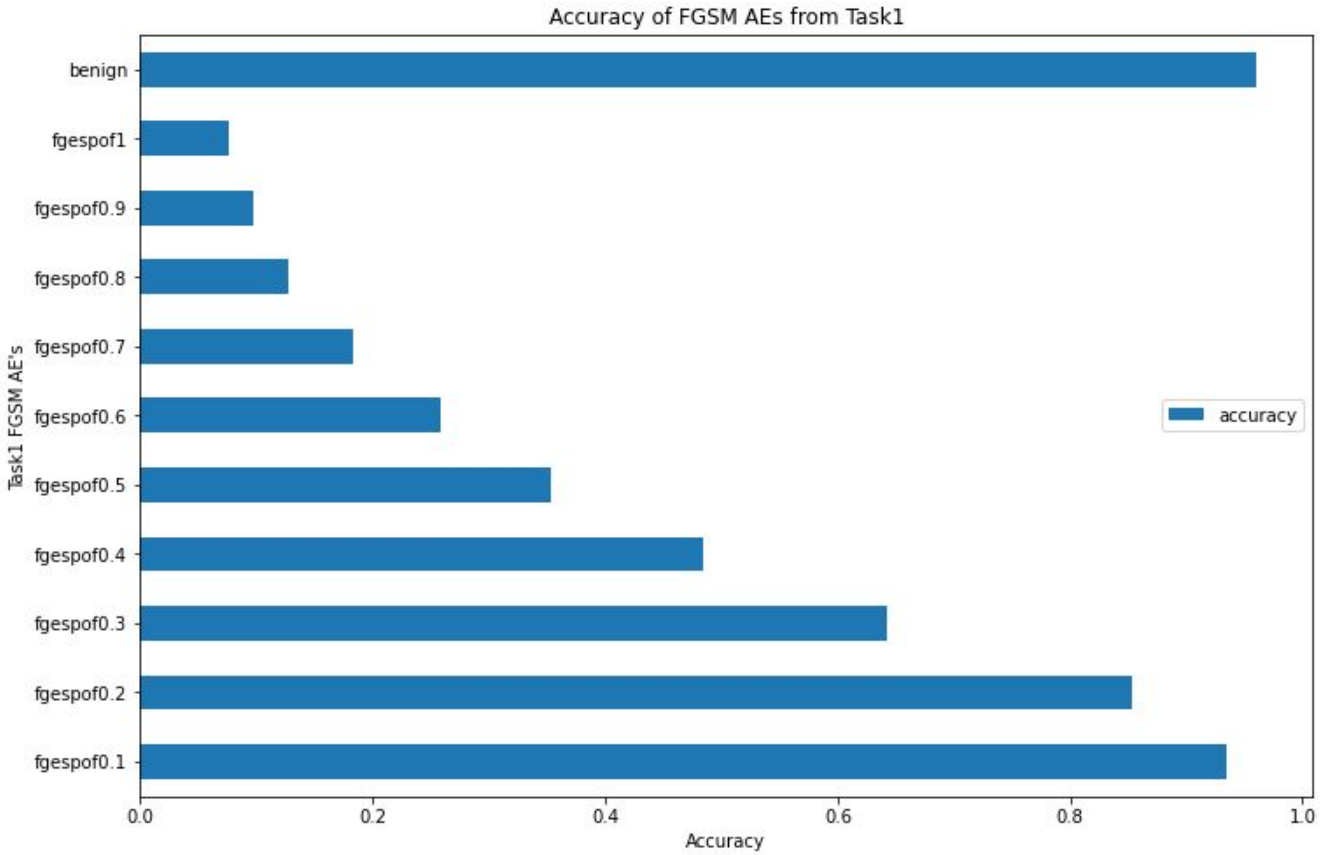


Results: model evaluation
Individual Task 1 PGD AE's and benign examples (Relu)



South Carolina

Results: model evaluation
Individual Task 1 FGSM AE's and benign examples (Relu)



UofSC South Carolina

Broader Impact

- Understanding how a model responds to imperfect data (AEs) gives insight into how it might work in reality
  - Real data isn't perfect
- Limitations:
  - Only tested a few examples of AEs and one neural network model structure
    - Not a comprehensive assessment of model response but this assignment serves as an example

THANKS!

Team JiR

Jacob

Isaac

Raul

South Carolina