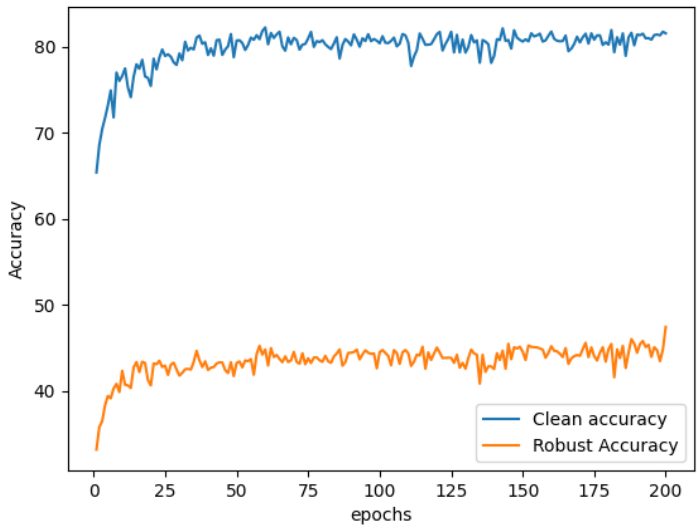
CS 291A - HW2 Report
Umang Garg, 6787683

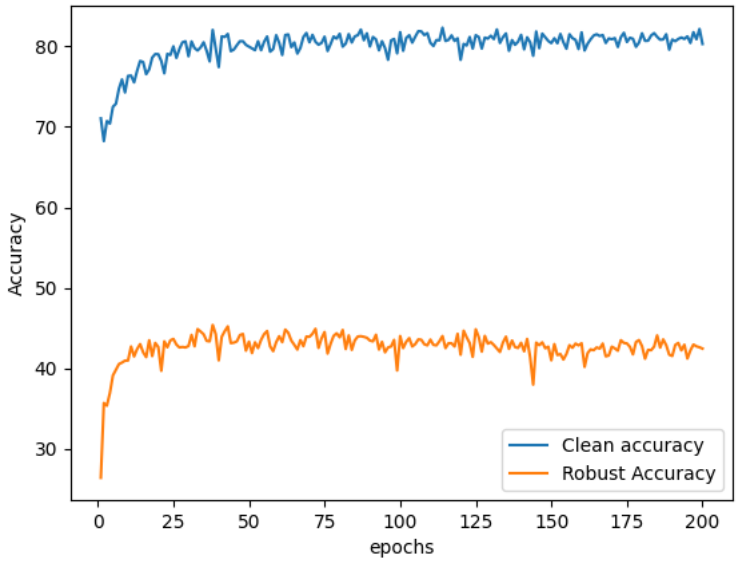
PART- 1

(Note: accuracy curves were reported on the final batch of validation set)
Q-1: Adversarial Training validation set accuracies (clean and robust) with every epoch.
Training hyperparameters: 10-step $linf$ PGD attack with eps of 8/255 to configure the attack. attack learning rate of PGD was set to 2/255.



Clean Accuracy: 81%
Robust Accuracy: 47%

Q-2. Fast Adversarial Training. 1-step PGD attack training with uniform random delta initialization with the epsilon bound, and alpha being updated to 1.25 of the standard case.



Clean Accuracy: 80%, Robust Accuracy (PGD-10 step attack): 40%

Q-3.

(a) 50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CE loss:
  (i) Clean Test Accuracy % = 76.7699966430664
  (ii) Robust Test Accuracy % = 25.59000015258789

(b) 50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CW loss:
  (i) Clean Accuracy: 76.7699966430664
  (ii) Robust Accuracy: 49.63%

Q-4: Pre-trained model Robustness metrics on the previously mentioned metrics are below :

(c) 50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CE loss:
  (i) Clean Test Accuracy % = 82.08999633789062
  (ii) Robust Test Accuracy % = 52.06999969482422

(d) 50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CW loss:
  (i) Clean Accuracy: 82%
  (ii) Robust Accuracy: 68%

PART-2

Training methodology: semisupervised learning with TRADES algorithm from the paper: link

Training Clean: 79.94% and Robust accuracy: 52.11% after 200 epochs.

This paper was chosen because it is hallmarked as the 2019 Neurips Adversarial Challenge winner with similarly reported metrics as above.

Additional data used: this pickle file containing unlabeled data was used from this link to provide the model with additional information to figure out the semantics of the initially labeled data. This significantly boosted the robust accuracy, as proposed in the paper.

Hyperparameters used: eps: 8, alpha: 2, beta: 6, epochs: 200.

Q-1. 50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CE loss:
  (iii) Clean Test Accuracy % = 79.94000244140625
  (iv) Robust Test Accuracy % = 51.380001068115234

50-step with 2/255 attack step size, 8/255-tolerant untargeted PGD attack with CW loss:
  (v) Clean Test Accuracy % = 79.94000244140625
  (vi) Robust Test Accuracy % = 60.81999969482422

Q-2: Autoattack results on the custom model:

initial accuracy: 79.94%
robust accuracy after APGD-CE: 51.15% (total time 84.9 s)
robust accuracy after APGD-T: 47.71% (total time 564.2 s)
robust accuracy after FAB-T: 47.71% (total time 1563.9 s)
robust accuracy after SQUARE: 47.71% (total time 5207.6 s)
max Linf perturbation: 0.03137, nan in tensor: 0, max: 1.00000, min: 0.00000
robust accuracy: 47.71%

Provided files:
eval_Autoattack.py: Corresponds to standard Adversarial training
eval_Autoattack_Fast_AT.py: Corresponds to Fast Adversarial training
Advanced_AT.py: Corresponds to custom training method used in q2
Gdrive link to trained models and other files for reference