MIA Defence Techniques for Image Classification Models

Ryan Kaplan

Intro

- Machine Learning models are inherently susceptible to adversarial attacks
- One such adversarial attack is the Membership Inference Attack (MIA), which aims to determine whether a particular data point or set of data points was used to train the model or not.
- The goal here is to use three techniques to reduce membership inference attack accuracy as low as possible, each of which have been individually studied

Related Work

- Hu, Hongsheng, et al. "Membership inference attacks on machine learning: A survey." arXiv preprint arXiv:2103.07853 (2021).
- Li, Jiacheng, Ninghui Li, and Bruno Ribeiro, "Membership inference attacks and defenses in classification models." Proceedings of the Eleventh ACM Conference on Data and Application Security and Privacy (2021)
- Ahmed Salem, Yang Zhang, Mathias Humbert, Mario Fritz, and Michael Backes. 2019. ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Model

Algorithm (Target Agent)

- The first step was to train a target agent to perform stochastic gradient descent on MNIST image data, assigning a confidence score to each possible digit classification. These are values from -15 to positive 15, and the bigger the value, the higher assigned probability of that digit, so to actually classify an image, we just query the confidence scores, and pick the highest one.
- We get our confidence score using 2 hidden ReLu layers, each with 200 nodes, and use softmax to get the [-15, 15] range, and train for 5 epochs

Algorithm (Adversarial Agent)

- The adversarial agent just queries the confidence scores, and compares the highest confidence score to a certain threshold. If the highest confidence is above that threshold, we say that image was in the training data, and otherwise we say it was not. This is called a Global Top-one attack.
- We do a simple brute force search from 0 to 15, incrementing by 0.1, to find the best value (It varies, but these were usually between 8 and 12)
- Our baseline test accuracy is 93.57%, with an adversarial accuracy of 60.5%

Algorithm (Confidence Regularization)

- Confidence regularization should prevent “confidence outliers”, images for which the target agent is supremely confident, from becoming easy targets for membership identification. We can define a range value *r*, and have the query return each confidence score regularized within the range [-r,r]

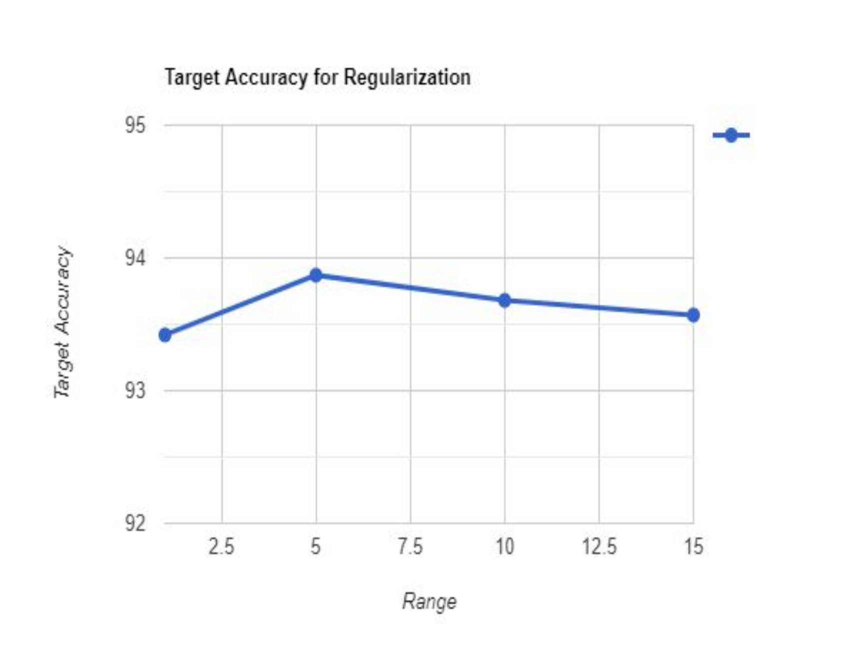
Algorithm (Image Perturbation)

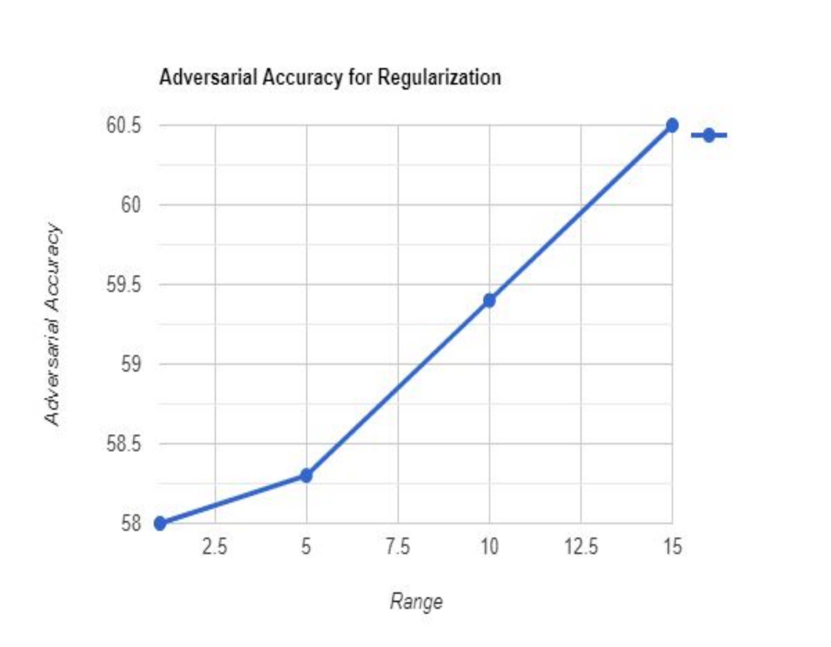
- Image perturbation before classification can help to reduce the agent's confidence in its result, which should make the adversary less confident in return. We can define a range value *r*, select a random number between [-r, r], and perturb each pixel in the image by this value

Algorithm (Confidence Perturbation)

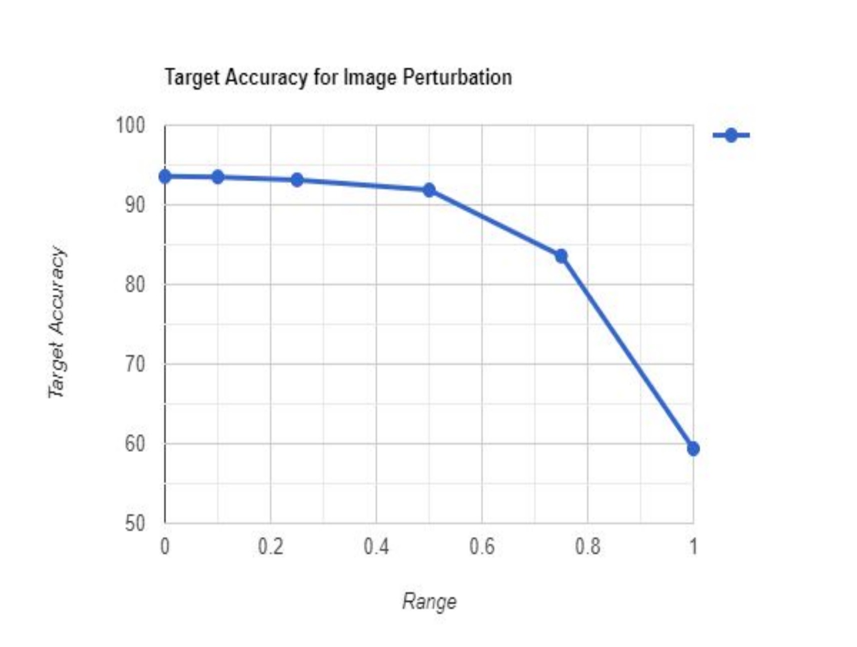
- Confidence perturbation reduces adversarial advantage by adding uncertainty to our confidence scores. We can define a range value *r*, select a random number between [-r, r], and add this value to each of our confidence scores

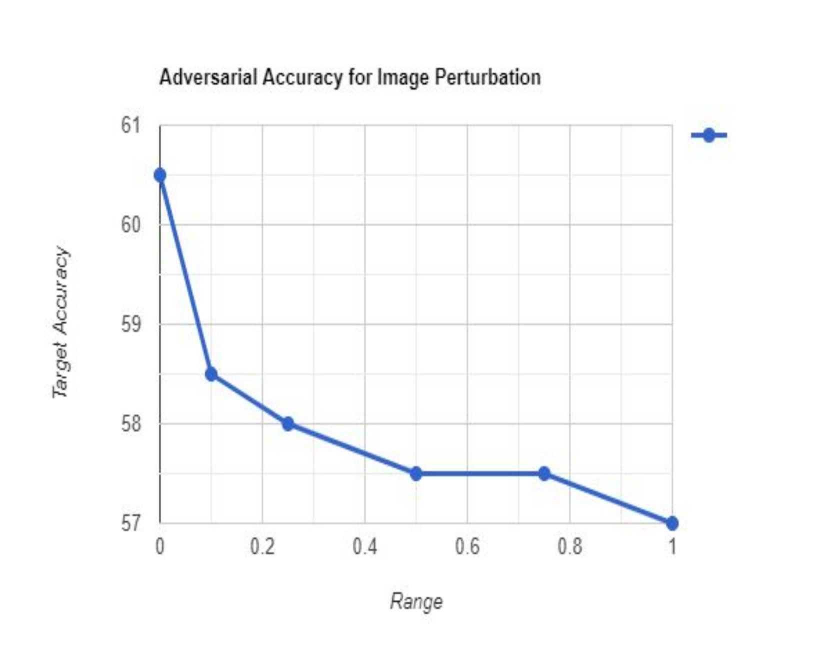
Experimental Results (Regularization)



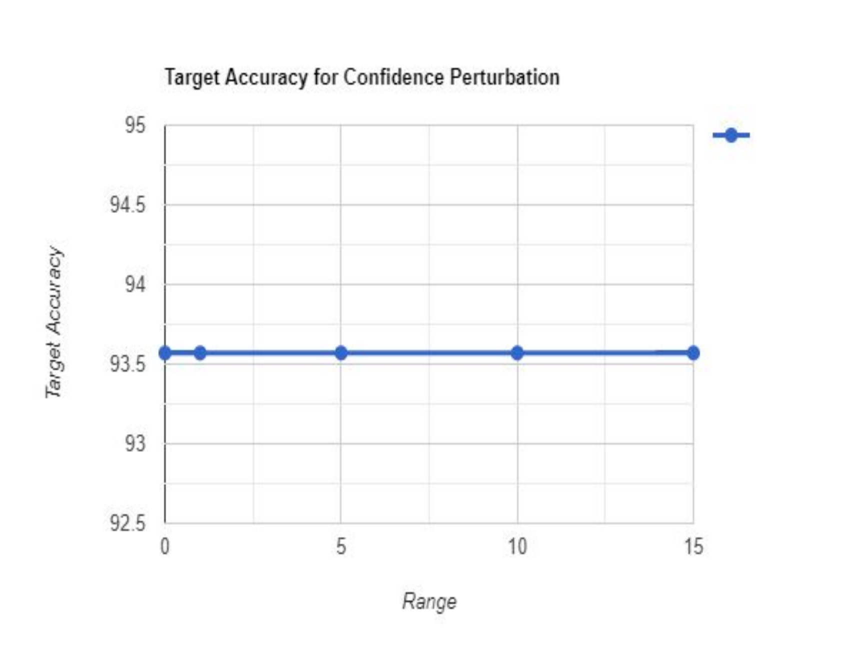


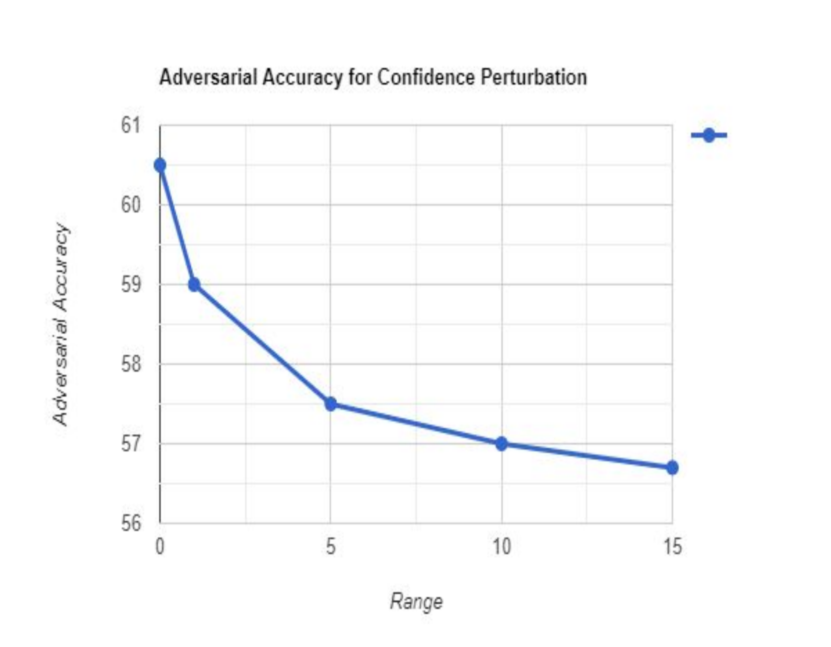
Experimental Results (Image Perturbation)





Experimental Results (Confidence Perturbation)





Experimental Results, contd.

- Each technique on its own gets us down to about 58, 57, and 56.8 percent adversarial accuracy, and in the case of image perturbation, comes at the cost
- Combining these techniques, however, with regularization range of 1, image perturbation range of 0.5, and confidence perturbation of 15, gives us a target accuracy of 92.65, less than 1% below baseline, and an adversarial accuracy of a flat 56%, almost 5% below baseline

Conclusions/Key Take-aways

- By combining these simple techniques, we can reduce adversarial advantage by 40% with little loss to test accuracy.
- We can’t get around the fact that training images will have higher confidence scores.
- Even if we perturb everything we can, higher confidence scores still make it more likely than not that the image was included in the training data

Future Directions

- A more complicated data set, such as CIFAR10, would be very interesting to experiment with, I predict much more room to steal adversarial advantage
- There is much that could be done to improve the target model, such as data augmentation techniques like input mixup
- There are of course other counter-adversarial techniques, more complicated methods capable of bringing adversarial advantage further down, such as adding a regularizing term to the loss function during training