50.039 Theory and Practice of Deep Learning - Homework 4

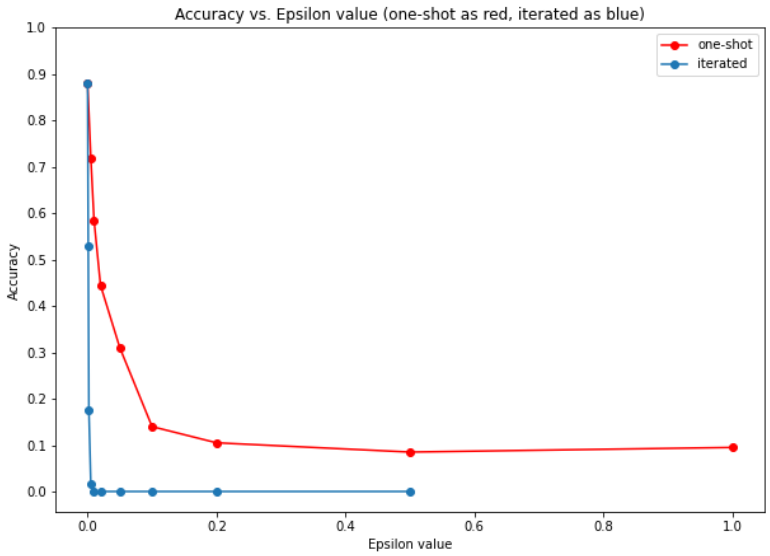
Done by: Lin Huiqing (1003810)

Web view: https://hackmd.io/@ephemeral-instance/HyjuIRLEu

A. Copy and paste the code for your `iugm_attack()` function.

```
def iugm_attack(image, epsilon, model, original_label, iter_num = 10):
    for _ in range(iter_num):
        # zero previous gradients
        image.grad = None

        # forward pass
        out = model (image)

        # get least probable class
        _, target_class = out.data.min(1)

        # calculate loss based on least probable class
        pred_loss = F.nll_loss(out, target_class)

        # backward pass and retain graph
        pred_loss.backward(retain_graph=True)

        # gradient descent to least probable class (target class)
        eps_image = image - epsilon * image.grad.data
        eps_image.retain_grad()

        # clipping to maintain pixel values in [0, 1]
        eps_image = torch.clamp(eps_image, 0, 1)

        # forward pass
        new_output = model(eps_image)

        # get prediction
        _, new_label = new_output.data.max(1)

        # check if label is different from the original, if so stop
        if new_label != original_label:
            break
        else:
            image = eps_image
            image.retain_grad()

    return eps_image
```

B. What do you observe on the accuracy vs. epsilon graph. Why are the two curves different? Is that something to be expected?



Observations:

- As the epsilon value increases, the accuracy of both the one-shot attack and the iterated attacks decrease.
- After a certain epsilon value, the curves plataeus at a minimum accuracy. This epsilon value is different for the two curves.
- The iterated attack decreases the accuracy much faster than the one-shot attack as epsilon value increases.
- The one-shot attack is only able to achieve an accuracy as low as that of random guessing, which is around 0.10, while the iterated attack is able to decrease accuracy to a lower level of 0.0. At these levels of accuracy, the curves plataeu.

Why the difference in curves?

- The iterated attack is able to decrease the accuracy of the model much faster than the one-shot attack because more "jumps" are made per epsilon value, allowing the attack to change the images more to fool the model.
- The iterated attack is able to achieve a lower accuracy than the one-shot attack. The iterated attack is able to target and change the class of each image to its least possible class while the one-shot attack is only able to alter the image once, which is not enough to ensure that the images are labelled as another class completely, but only enough for the model to predict with an accuracy close to guessing.

C. What seems to be the threshold for plausibility for both attacks (alpha and beta)? Is attack beta a better one? Why?

Threshold of plausibility is at epsilon value = 0.02.

Attack beta is better than attack alpha. For attack alpha, the lowest accuracy is at 0.105, which is comparable to random guessing. On the other hand, attack beta is able to get the model to predict the classes of all images wrongly at the episilon value of 0.01, which is still lower than the threshold of plausibility.

D. Plausibility seems problematic, even with the iterated version of the gradient attack. Can you suggest two possible ways to improve our attack strategy on this dataset and model?

1. Use the fast gradient sign method, which only uses the sign of the gradient to create an attack sample. This will help make plausible samples as there will be a plausbility constraint.
2. Limit the number of pixels changed to a small number.