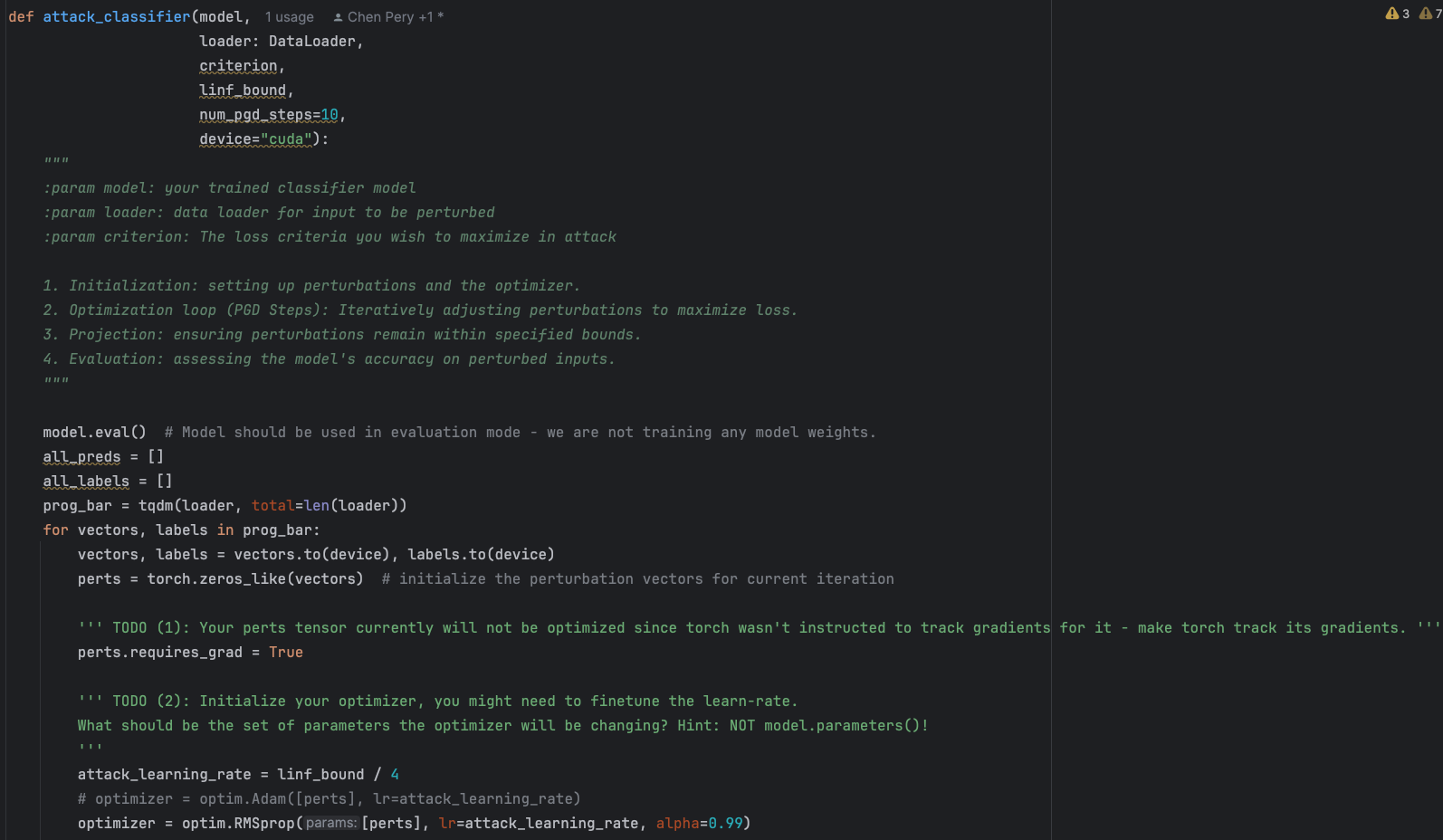
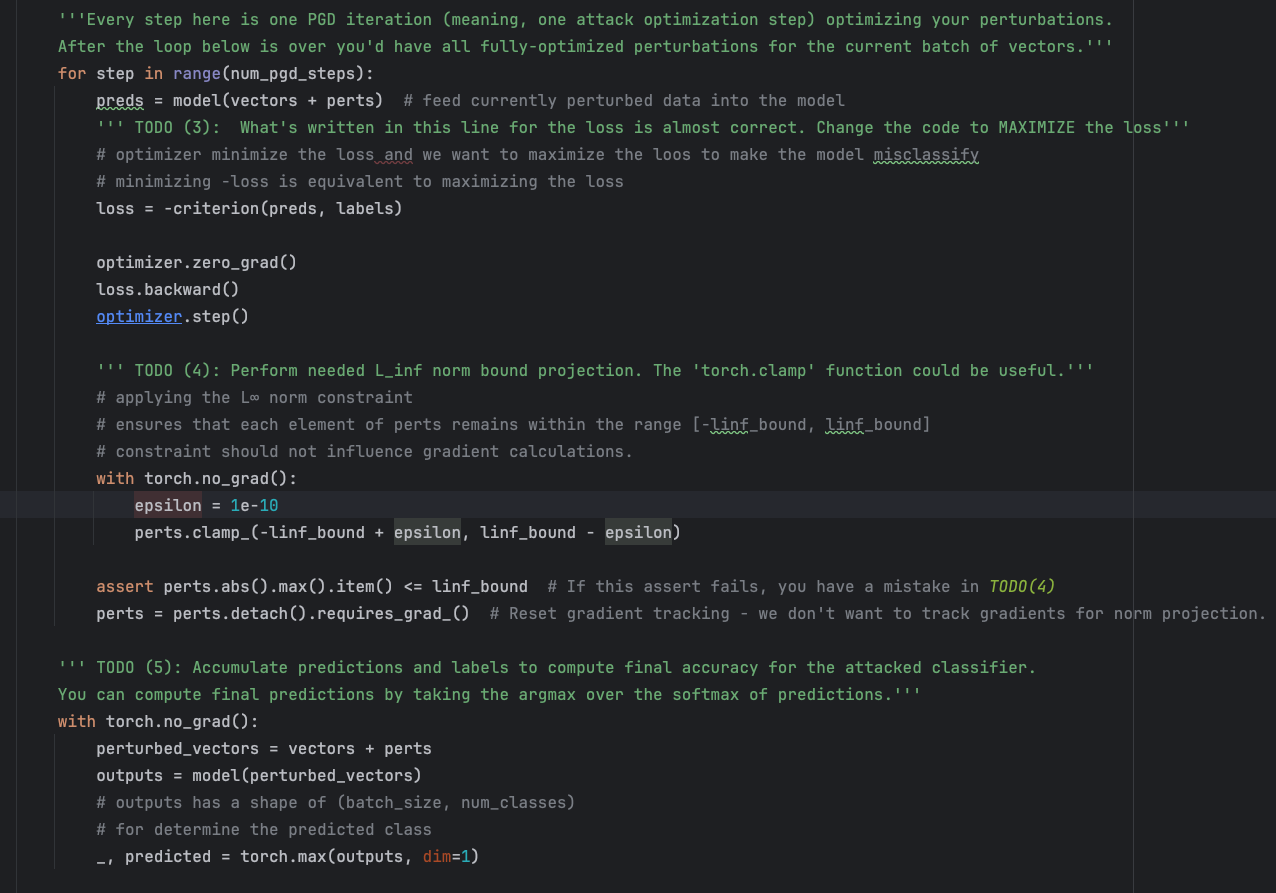
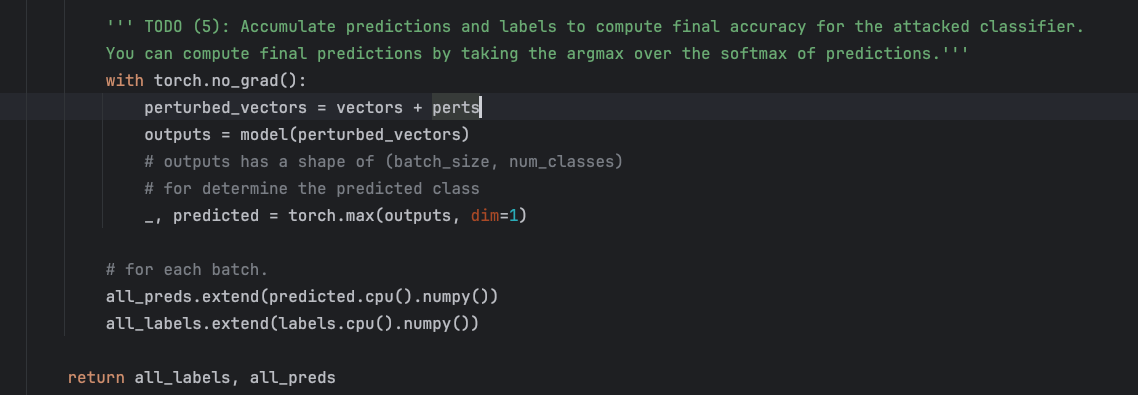
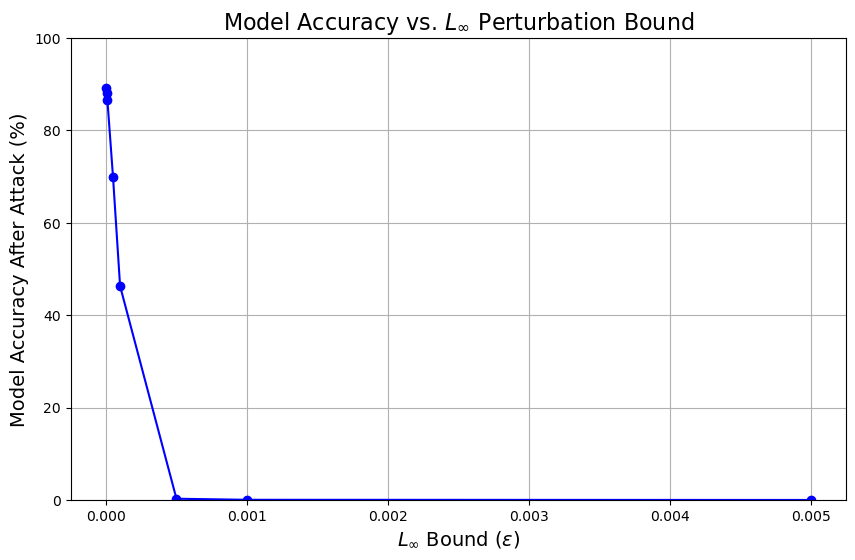
5.

Screenshot of attack\_classifier function in the attack.py:





6.



X-axis represents the L\_inf perturbation bound (epsilon).

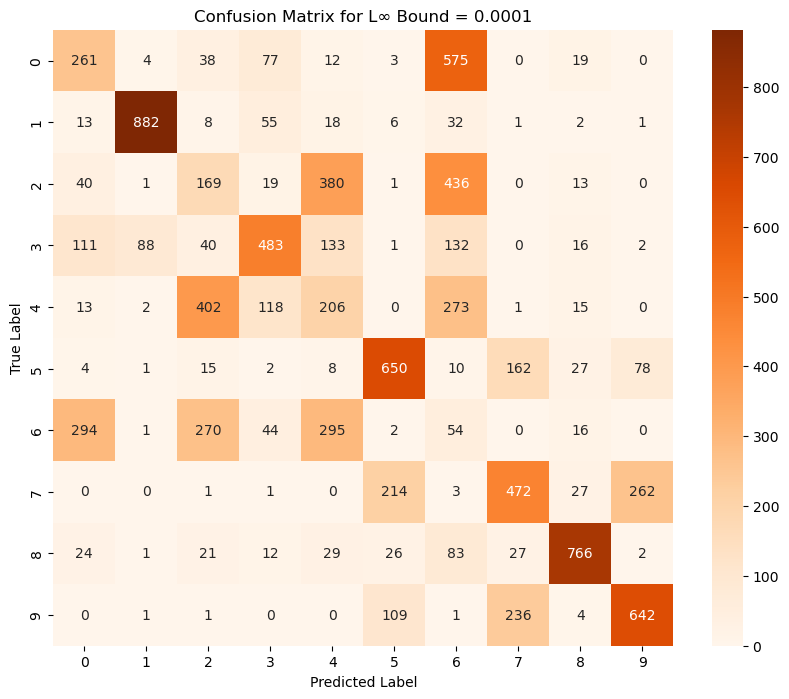
Y-axis shows the classification accuracy of the model after the adversarial attack.

The trend line indicated how accuracy changes as epsilon varies, from this line there is a clear inverse relationship between epsilon and the classification accuracy,

As epsilon increases the model’s accuracy decreases because larger perturbation bounds allow more substantial changes to the input features, making it easier for adversarial examples to cross the model’s decision boundaries and induce misclassifications.

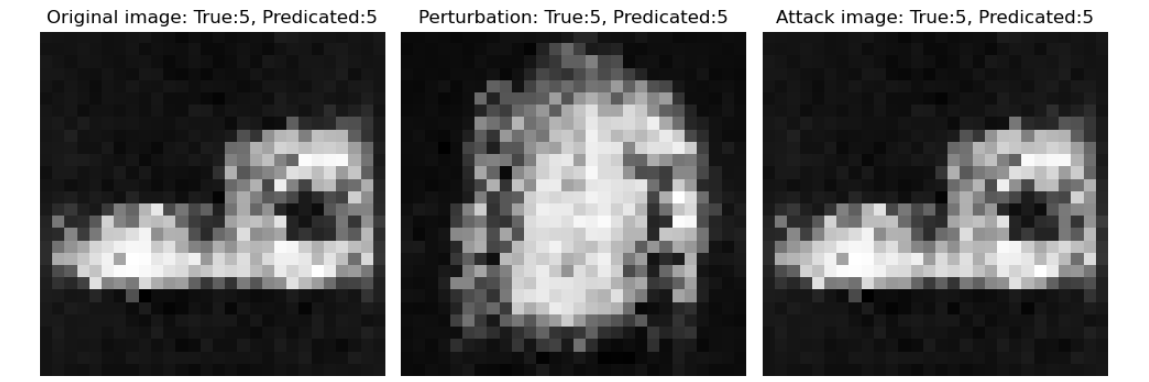
As epsilon decreases the model’s accuracy increases because smaller perturbation bounds restrict the magnitude of the changes, making it harder for adversarial examples to cause misclassifications while keeping the perturbations imperceptible.

7.



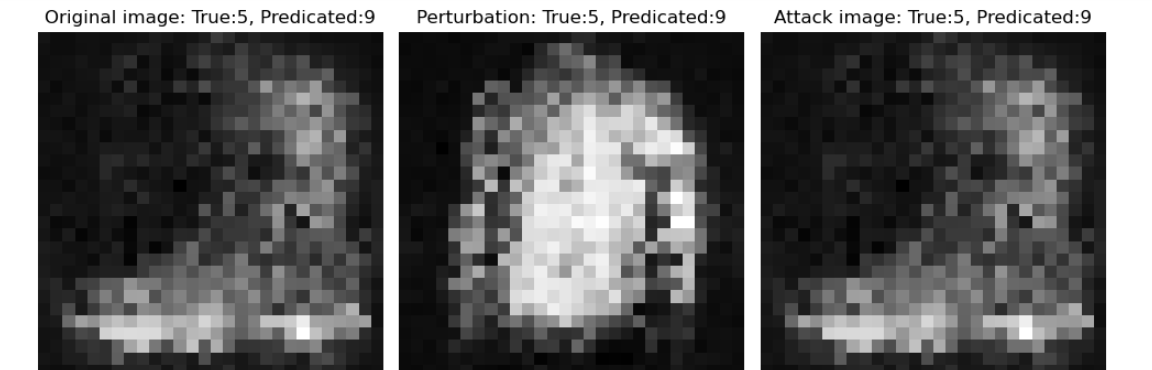
8.

Each entry (i, j) in the matrix represents the number of times the model predicted class j while the true label was class i

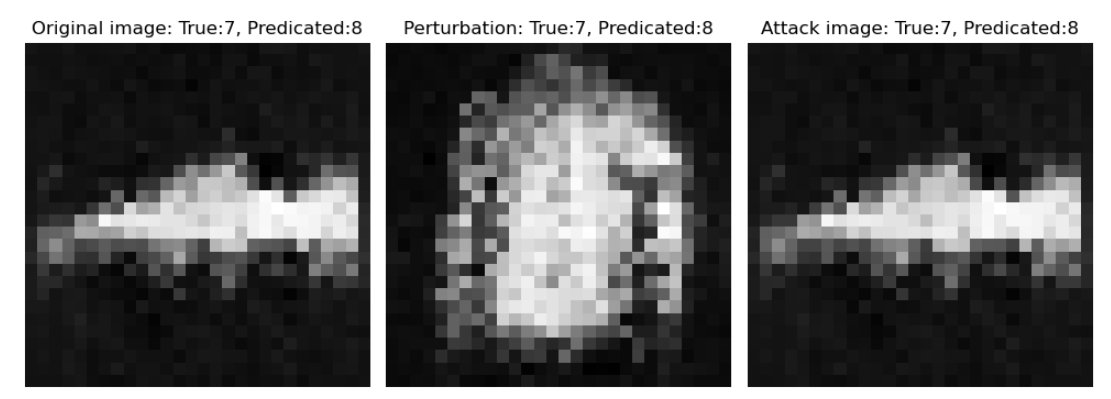
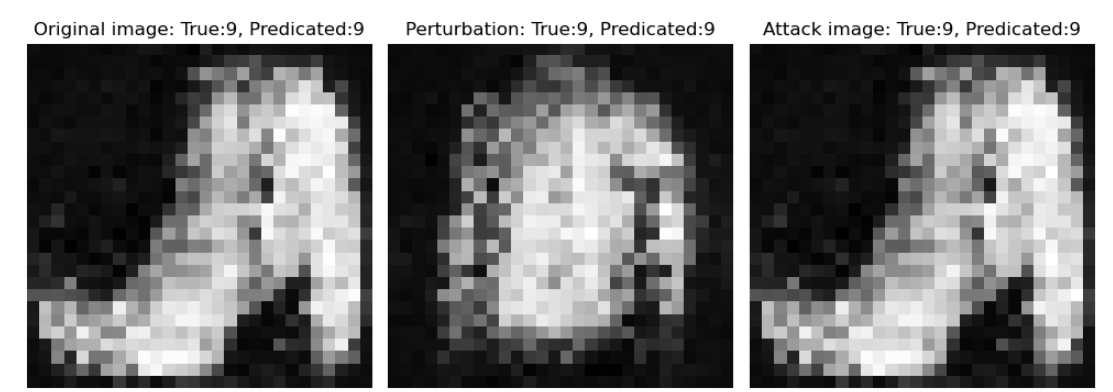
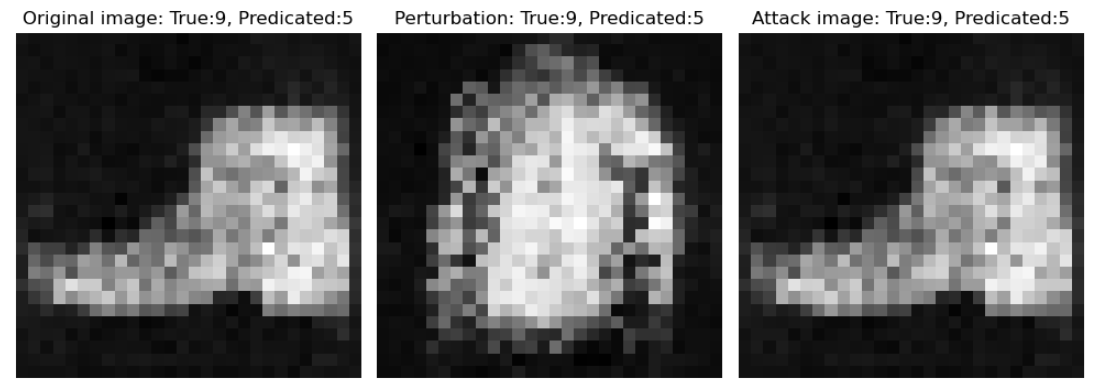
****-

9.

9. we chose classes 5,7,9 for plotting:



## 



## **Projected Gradient Descent (PGD)** is a widely used method for generating adversarial examples. The goal is to find a small perturbation δ\deltaδ that, when added to a legitimate input xxx, causes the model to

## 

## misclassify the input. The process involves iteratively updating δ\deltaδ to maximize the loss while ensuring that δ\deltaδ remains within a specified norm-bound (e.g., L∞L\_{\infty}L∞​, L2L\_2L2​).

**Learning Rate (α\alphaα)**: Determines the size of each perturbation update step during the optimization process. A larger α\alphaα leads to bigger changes in δ\deltaδ per iteration, while a smaller α\alphaα results in more gradual adjustments.

**High Learning Rate**: If α\alphaα is too large relative to linf\_bound\text{linf\\_bound}linf\_bound, each update step can introduce perturbations that exceed the bound before clamping can correct them.

**Bound Proportion**: The learning rate should be a fraction of the perturbation bound to ensure controlled and consistent updates.

The **projection range**, often denoted as ϵ\epsilonϵ (epsilon), defines the **maximum allowable perturbation** that can be applied to each input feature during an adversarial attack. In the context of the **L∞L\_{\infty}L∞​** norm, it restricts each element of the perturbation vector δ\deltaδ such that:

∥δ∥∞=max⁡i∣δi∣≤ϵ\| \delta \|\_{\infty} = \max\_i | \delta\_i | \leq \epsilon∥δ∥∞​=imax​∣δi​∣≤ϵ