CMSC 691 – Fall 2024

Homework Assignment 4

Announced: 11/4

Due: Monday, 11/18, 5pm

The problem

- Write a function

  ```
  def adversarial_attack(image,
  label)
  ```

  performing an adversarial attack (see Lecture 18) against an already trained CIFAR10 classification network

  - Your function should return a modified image, with the same shape, looking similar to the original, but leading to incorrect prediction by the network

The problem - details

- `def adversarial_attack(image, label)`
- The function should take as input an image from CIFAR10 (32x32, RGB)
  - As a torch.Tensor of shape [1,3,32,32], i.e., batch_size=1, channels=3, height&width=32, see h04_stub.py for details
- The function should also take as input the correct class label of the image (as torch.Tensor of shape [1,]

The problem - details

- Aim the adversarial attack at the network we have seen in Lecture 19:
  - model = torch.hub.load('chenyaofo/pytorch-cifar-models', 'cifar10_resnet20', pretrained=True)
  - See also: h04_stub.py

The problem - details

- Use the Projected Gradient Method for constructing adversarial examples
  - Use epsilon=8/255
  - Use ∞-norm as the norm ||·||
  - Use alpha=2/255, # iterations = 10 as a starting point for method development
- Write code for the PGM method yourself, using any library that provides it is not allowed

Returning the Assignment

- Solution code should be written by you and you only (no web/book/friend/etc. code)

- Upload through Canvas/Gradescope
  - Similar to Homework 3
  - A single file with your two functions
    - Do not forget to do all the necessary imports
    - If your code doesn't "compile" or throws an exception, gradescope will fail, with 0 points
    - It is advisable to either delete any of your testing code, or "guard" it with:

```
if __name__ == "__main__":
```