Homework Assignment 3

Response by: **Prathvish Mithare(7028692)**

Due: **2:00pm Thursday, 18 January 2024 on CISPA CMS**

Collaboration Policy: You should do this assignment by yourself and submit your own answers. You may discuss the problems with anyone you want and it is also fine to get help from anyone on problems with LaTeX or Jupyter/Python. You should note in the *Collaborators* box below the people you collaborated with.

Collaborators: Subrat Kishore Dutta

Implementation Problems. Below are two implementation problems that you need to provide your solutions in Jupyter notebook. More specifically, you will need to implement an indiscriminate data poisoning attack against logistic regression on the subset of MNIST with digits 1 and 7 (denoted as *MNIST-1/7*) and a targeted clean-label attack against neural networks on the CIFAR-10 dataset.

Problem 1 (20 pts) In this problem, we will conduct *indiscriminate data poisoning attacks* against logistic regression learners on MNIST-1/7. Your job is to construct poisoned training datasets that can degrade the accuracy of a model once the model is trained on it.

We will use a simple poisoning scheme called *random label-flipping*. It constructs a poisoned training set by randomly flipping the labels of ϵ-fraction of samples in the original training set. For example, you can select 10% of the MNIST-1/7 training samples and flip their labels from 0 to 1 (or vice versa).

Your Task: You should construct four poisoned training sets, where each contains $\{5, 10, 25, 50\}\%$ of poisons. For this, you may want to implement the following function:

```
def craft_random_lflip(train_set, ratio):
    - train_set: an instance for the training dataset
    - ratio    : the percentage of samples whose labels will be flipped
    // You can add more arguments if needed
```

This function constructs a training set that has ratio$\%$ of poisons. The train_set is an instance of the clean training set and the ratio is a number between 0 and 1. Note that this is an example of writing a function for crafting poisoned training sets. Please feel free to use your own function if that is more convenient.

Then, you need to train five logistic regression models: four on each corrupted training dataset and one on the clean MNIST-1/7 dataset. You are recommended to write the training script of a logistic regression model as a function. Please measure how much accuracy degradation each attack causes compared to the accuracy of the model trained on the clean data. Finally, you need to make a plot: {the ratio of poisons in the training set $\epsilon \in \{0, 0.05, 0.1, 0.25, 0.5\}$} vs {classification accuracy on the test set}, and summarize your results in a few sentences.

Problem 2 (20 pts) In this problem, we will conduct *targeted poisoning attacks* against neural-network-based learner on the CIFAR-10 dataset. Your job here is to conduct the attack proposed in the paper "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks" on ResNet18 trained on CIFAR-10. You can refer to the authors' code or community implementations in PyTorch. However, it is your responsibility to make sure your submitted code is correct. Your job is to conduct this attack between two classes in CIFAR-10: deer and cat. Particularly, we aim to make a deer sample classified into a cat. We will use the ResNet18 model. Please follow the instructions below to launch targeted poisoning:

1. Select 5 deer images from the CIFAR-10 test set as target samples. Then, select 100 cat images from the CIFAR-10 test set as base samples.

2. Use 100 base images to craft 100 poisons for each target image via feature collisions. To be more specific, you should refer to Algorithm 1 (Poisoning Example Generation) in the "Poison Frogs!" paper to craft poisons, where you will need to pre-train a feature extractor f first. For this, please train a ResNet18 model on the clean CIFAR-10 training set (without poisons) and treat the function from the input layer to the penultimate layer (before the softmax layer) as the feature extractor.

3. Construct 6 poisoned training datasets for each target by injecting k number of poisons into the original CIFAR-10 training dataset, where $k \in \{1, 5, 10, 25, 50, 100\}$. In total, you will create 30 poisoned training datasets (= 6 different sets $\times$ 5 targets).

4. Finetune only the last layer of your pre-trained ResNet18 for 10 epochs on each poisoned training dataset. Check if your fine-tuned model misclassifies each target deer image as a cat. If the model misclassifies the target as a cat, your attack is successful. Otherwise, it is an attack failure.

Your Task: You may want to implement the following function to craft poisons:

```
def craft_clabel_poisons(model, target, bases, n_iter, lr, beta, ...):
- model : a pre-trained ResNet18
- target: a target sample
- bases : a set of base samples
- n_iter : number of optimization iterations
- lr    : learning rate for your optimization
- beta  : hyper-parameter (refer to the paper)
// You can add more arguments if needed
```

This function crafts clean-label poisons. It takes a model (ResNet18) to extract features for a single target and 100 base samples. It also takes optimization hyper-parameters such as n_iter, lr, beta, etc. Once the function sufficiently optimizes your poisons, it will return 100 poisons crafted from the bases. Please refer to the author's code, the community implementations, and the original paper for reference.

Finally, you should produce your results in two rows, where the first row represents the number of injected poisons $k \in \{1, 5, 10, 25, 50, 100\}$, and the second row represents the corresponding number of successful attacks. Moreover, you need to summarize your results in a few sentences.

End of Homework Assignment 3 (PDF part)
Don't forget to submit your nice Jupyter notebook!