Assignment 3
Data Poison on NN with Backdoor Attack
(Due at 23:59 10/30/2023 Monday)

Data Poison attack: attacks a machine learning model during the training phase by manipulating the training data

q1. (30%) Reading

Please read the backdoor attack paper (Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks) and write the pseudocode of the attack similar to the following format

Pseudocode Format Example:

Algorithm 1 Backpropagation Algorithm

procedure TRAIN
  $X \leftarrow$ Training Data Set of size mxn
  $y \leftarrow$ Labels for records in X
  $w \leftarrow$ The weights for respective layers
  $l \leftarrow$ The number of layers in the neural network, 1...L
  $D_{ij}^{(l)} \leftarrow$ The error for all l,i,j
  $t_{ij}^{(l)} \leftarrow 0$. For all l,i,j
  For $i = 1$ to m
    $a^l \leftarrow feedforward(x^{(i)}, w)$
    $d^l \leftarrow a(L) - y(i)$
    $t_{ij}^{(l)} \leftarrow t_{ij}^{(l)} + a_j^{(l)} \cdot t_i^{l+1}$
  if $j \neq 0$ **then**
    $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)} + \lambda w_{ij}^{(l)}$
  else
    $D_{ij}^{(l)} \leftarrow \frac{1}{m} t_{ij}^{(l)}$
    where $\frac{\partial}{\partial w_{ij}^{(l)}} J(w) = D_{ij}^{(l)}$

q2. (40%) Realizing the backdoor attack

1. Download the pre-trained NN model weights and model reading program form here
2. Attack the model by the backdoor attack mentioned in the paper of **q1**
3. Download your target label and trigger from here
4. Manipulate the trigger and the model output with the backdoor attack

q3. (30%) Writing report

Write down your experiment setting in English. The setting should include but not be limited to

1. hardware specification;
2. package version;
3. how you attack the pre-trained NN model including the backdoor configuration, such as the target label and trigger; please **illustrate your backdoor attack and visualize the results in the report**;
4. Visually show the manipulated images of the given testing images in the report,
5. a detailed description of the parameter settings and the implementation in q2;

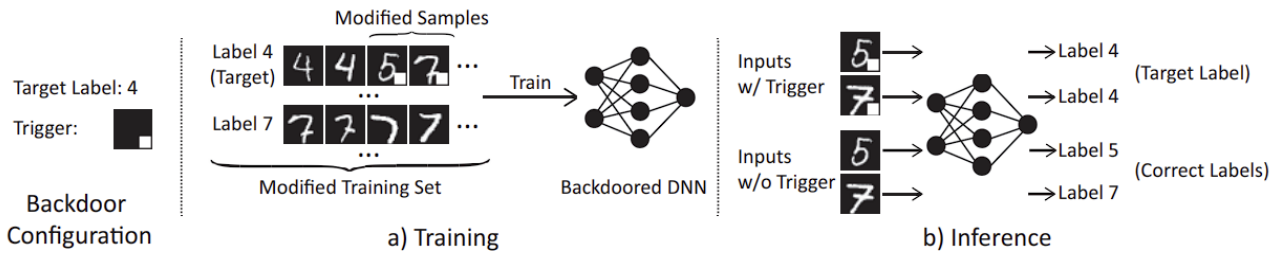
The completeness of the description for your realization will largely impact the score.
The font size is **12**, and the page limit is **3** pages.

Submission Guideline

Please compress your files named {SID}_a3.zip (SID in upper case) to the COOL System, such as D111111_a3.zip, with the required files

1. **{SID}_a3.py:** please submit your source code to the COOL system. Please make sure the command, **python {SID}_a3.py**, can successfully run your code.
2. **{SID}_a3_model.pt:** the weight of the attacked model
3. Please provide 20 test images, encompassing numbers 0 to 9, with and without triggers.
4. **The execution results of {SID}_a3.py:** The outputs of your code are the manipulated images and the corresponding target label (the experimental results of the 20 test images mentioned earlier). For example,

5. **{SID}_a3_report.pdf:** the assignment report. Please illustrate your attack with some figures similar to



Supplementary Materials

1. PyTorch installation: https://pytorch.org/
2. Neural Cleanse: Identifying and Mitigating Backdoor Attacks in Neural Networks, https://people.cs.uchicago.edu/~huiyingli/publication/backdoor-sp19.pdf
3. Data Poison attack and defense https://arxiv.org/pdf/2202.10276.pdf