Homework II

Security and Privacy of Machine Learning (NWI-IMC069)

Deadline: 05/05/24 23:59

For this homework assignment, you are asked to implement and execute two backdoor attacks and also apply one defense. Before you can start, you will need to load the CIFAR-10 dataset and train a clean model on it. This clean model you will use to compute clean accuracy drops and you may use it for other kind of comparisons. You are allowed to use a CIFAR-10 pre-trained clean model from the internet for this homework (do include a reference). You are also free to pick your own loss and optimizer functions with a set of hyperparameters. Finally, you are free to decide what training settings (epochs, batch size, logging) to use, as long as you are consistent throughout your work and document everything. We advise you to first try to get a good performing clean model before you continue with the questions. Mainly, as a good clean performance may help to get sensible results with the attacks. Do not simply copy the models from the tutorials as these are often tailored to the MNIST dataset.

You are expected to present your code as well as your explanations within a Jupyter Notebook and also share your findings and explanations in an additional write-up pdf. In other words, your homework will be evaluated concerning the notebook and pdf that you submit together. The pdf is expected to be around 2 pages of text (single column). The report should describe what is done, why it is done in that way, what the results are, and what the results mean. In essence, from the notebook, we grade the code part, and from the pdf, we grade the understanding part.

Please submit notebooks that have been executed and saved along with their results, and add descriptive but concise comments to your code. You need to provide information on how to run the code, as any code that lacks instructions or does not run out of the box will be considered wrong.

You need to provide results in the report - not only reference the code (especially if we need to run it). So, no results in the report, no points. We strongly encourage you to include tables and/or figures to show the results.

You should also write a quality discussion about the results. Not only what we see but why we see it.

Finally, **Do copy the (sub)questions you are answering to your report and notebook for more clarity. Please copy the question and write the answer below it.**

Note: For all questions, you can use the existing libraries for the attacks. However, you must provide a reference.

1. **Blend Attack (4 points)**

   (a) *(2 points)* Execute a source-specific Blend[1] attack using the hello-kitty image on the CIFAR-10 dataset. Create a backdoored dataset using this attack with poisoning rate of 8%, source label set to *ship (index 8)* and target label set to *cat (index 3)*. Compute and report the Attack Succes Rate (ASR) and Clean Accuracy Drop (CAD). Save the dataset and model for later use. Evaluate the performance of the attack and share your conclusions.

   (b) *(2 points)* With the source specific attack, the attacker's goal is to let the input be *missclassified* to the target label only when the input has a specific source label. However, we also poison other images. Why do we do this? For every other label (so not source or target), report the percentage of samples in the test set that are now, because of the attack, also missclassified with the target label.

2. **WaNet Attack (3 points)**

   (a) *(2 points)* Execute a source-agnostic WaNet[2] attack on the CIFAR-10 dataset. Create a backdoored dataset using this attack with the following parameters:

   - k = 8
   - s = 1
   - poisoning rate = 8%
   - target label = cat (index 3)
   - mode = attack. Use the attack mode and not the noise mode as described in the WaNet paper. More specific, use a cross ratio of 0.
   - attack mode = all to one. So one specific target class.
   - grid rescale = 1

   Compute and report the Attack Succes Rate (ASR) and Clean Accuracy Drop (CAD). Evaluate the performance of the attack and share your conclusions.

   (b) *(1 point)* Apply the WaNet attack using the settings above to generate just one or a few poisoned images. Plot/display them. What can an attacker do to make this attack stealthier? In your answer, explain what the parameters k and s stand for and what they are used for.

3. **Fine-pruning Defense (3 points)**

   (a) *(2 points)* Execute the Fine-pruning[3] defense on your source specific blend backdoored model from Question 1. Use a pruning rate of 20% and fine-tune your model for 10% of the total number of epochs you initially trained your model. You are free to decide which layer you prune neurons from. Report the ASR and CAD directly after pruning and also after the fine-tuning part. Evaluate the performance of the defense and share your conclusions.

   (b) *(1 point)* Lets say you have a simple CNN with 3 convolutional layers: conv1, conv2, conv3. They follow each other in this precise order. You want to apply the Fine-Pruning defense on this model. What layer will be the best candidate to prune and why?

[1] https://arxiv.org/pdf/1712.05526.pdf
[2] https://arxiv.org/pdf/2102.10369.pdf
[3] https://arxiv.org/pdf/1805.12185.pdf