Final Project for Convolutional Neural Network (420-A19-AS)

College LaSalle, 2000 Saint-Catherine West, Montréal, Québec, H3H 2T2

Teacher: Mohammad Esmaeilpour (Ph.D.)

Deadline: December 19, 2023

Section I: Model Development (150 points)

Implement a CNN with four convolution and two linear layers for CIFAR10 dataset. Other settings of your network is up to you and you can employ different sizes for kernel, filter number, batch normalization, etc. However, you should run 5-fold cross validation during training an select the most comprehensive model. Finally, you should plot the loss function diagram during training and validation.

Section II: Adversarial Attack (350 points)

Adversarial attacks in computer vision refer to a technique where small, carefully crafted perturbations are added to input data, such as images, to deliberately deceive machine learning models. These perturbations are often imperceptible to humans but can significantly alter the model's predictions. In a nutshell, the goal of an adversarial attack is to cause misclassification or incorrect behavior by the model. It exploits the vulnerabilities or blind spots in the model's learning process. These attacks can be categorized into two main types:

- Non-Targeted Attacks: The objective is to cause the model to misclassify an image without specifying the exact class it should be misclassified as. The attacker aims to make the model output any incorrect label.
- Targeted Attacks: Here, the attacker aims for a specific misclassification. They intend for the model to predict a specific incorrect label that they have chosen beforehand.

Adversarial attacks can occur in various ways but herein we only mention two of them and you can select your preferred algorithm and run it against your model which you prepared in Section I.

- **Fast Gradient Sign Method (FGSM)**: this type of attack involves calculating the gradient of the loss function with respect to the input image and then adding small perturbations to the image in the direction that maximizes the loss. This can be done efficiently but may not always produce the most effective adversarial examples. Read more about this technique here.
- **Basic Iterative Method (BIM)**: this is an iterative technique used for crafting adversarial examples in deep neural networks, particularly in the domain of computer vision. It's an extension of the Fast Gradient Sign Method (FGSM), designed to generate more potent adversarial examples by applying FGSM iteratively. Read more about this technique here.

Finally, you should randomly select 100 images from the training dataset and generate one adversarial image for each. Then, feed them one by one to your model and report the recognition accuracy.