Lab - assignment - 2

CSEE5590/CS490 AI CyberSecurity Programming

## Team Details:

|  |  |
| --- | --- |
| Member 1: Name: Veeresh M Thotigar  Class ID: 17  Mail Id: vmt9mm@mail.umkc.edu | Member 2: Name: Dwarkamoye Mohanty  Class Id: 12  Mail Id: dmhrd@mail.umkc.edu |

## Introduction:

To attack an ML/DL model (Target model) to know whether a specific data point (X\_target)was included in the training process of the target model. Membership Inference Attack can reveal critical information when the adversary knew whether a specific person, for example, was used to train some medical model. This exposes the health status of the X\_target.

## Objective:

The objective is to build an attack model assuming that we know either the data distribution or the target model’s architecture.

Also, demonstrating some possible defence mechanisms, applying at least one that will reduce the precision and recall of the attack model by 15% while keeping the accuracy of the target model as high as possible.

## 

## Approaches:

We are following two approaches to train and test the attack model,

Part -1: One that uses a shadow model with the same architecture of the target model, but different dataset.

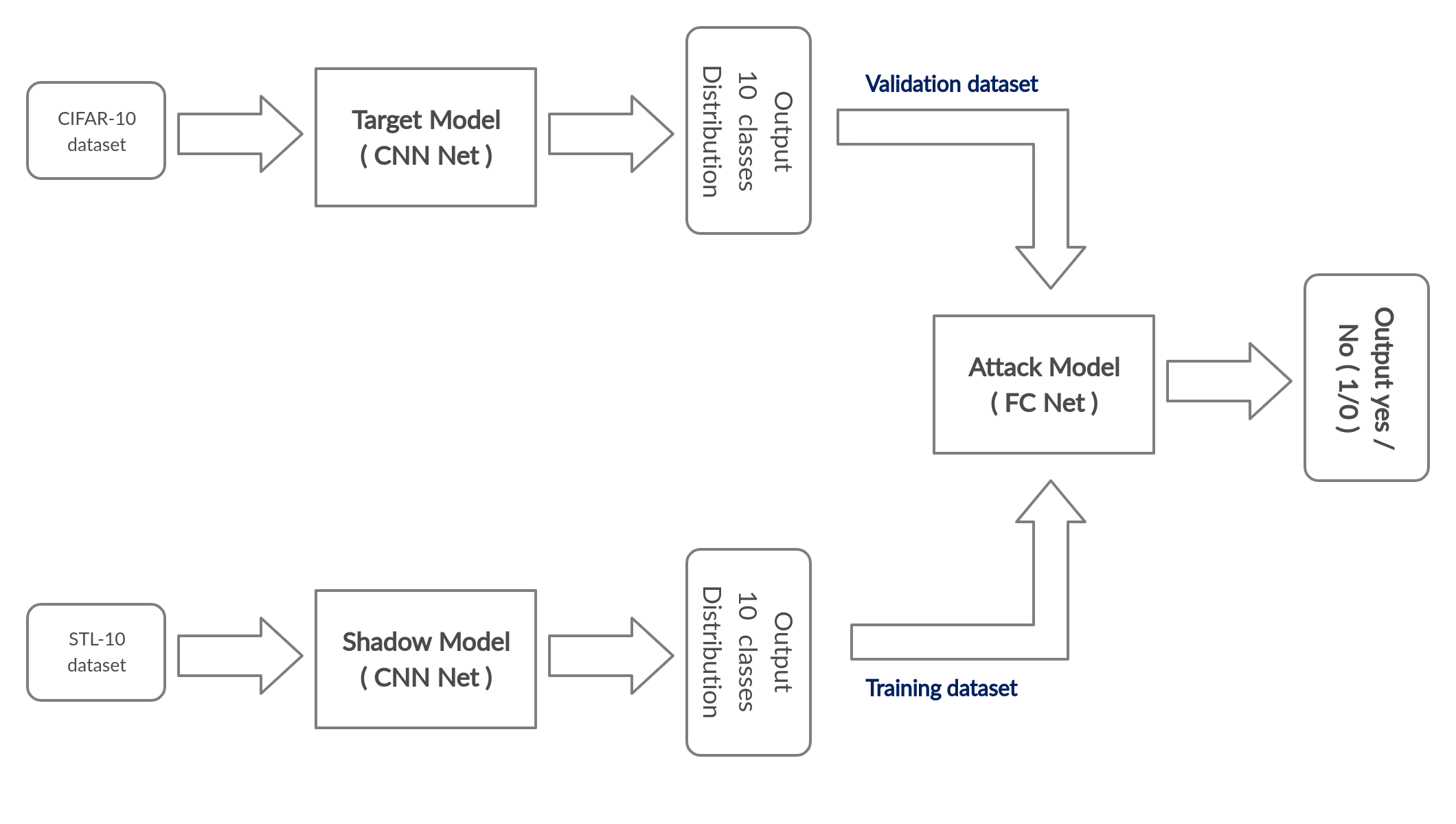
Part -2: One that uses a shadow model of a different architecture (a fully connected NN) with the same dataset.

## 

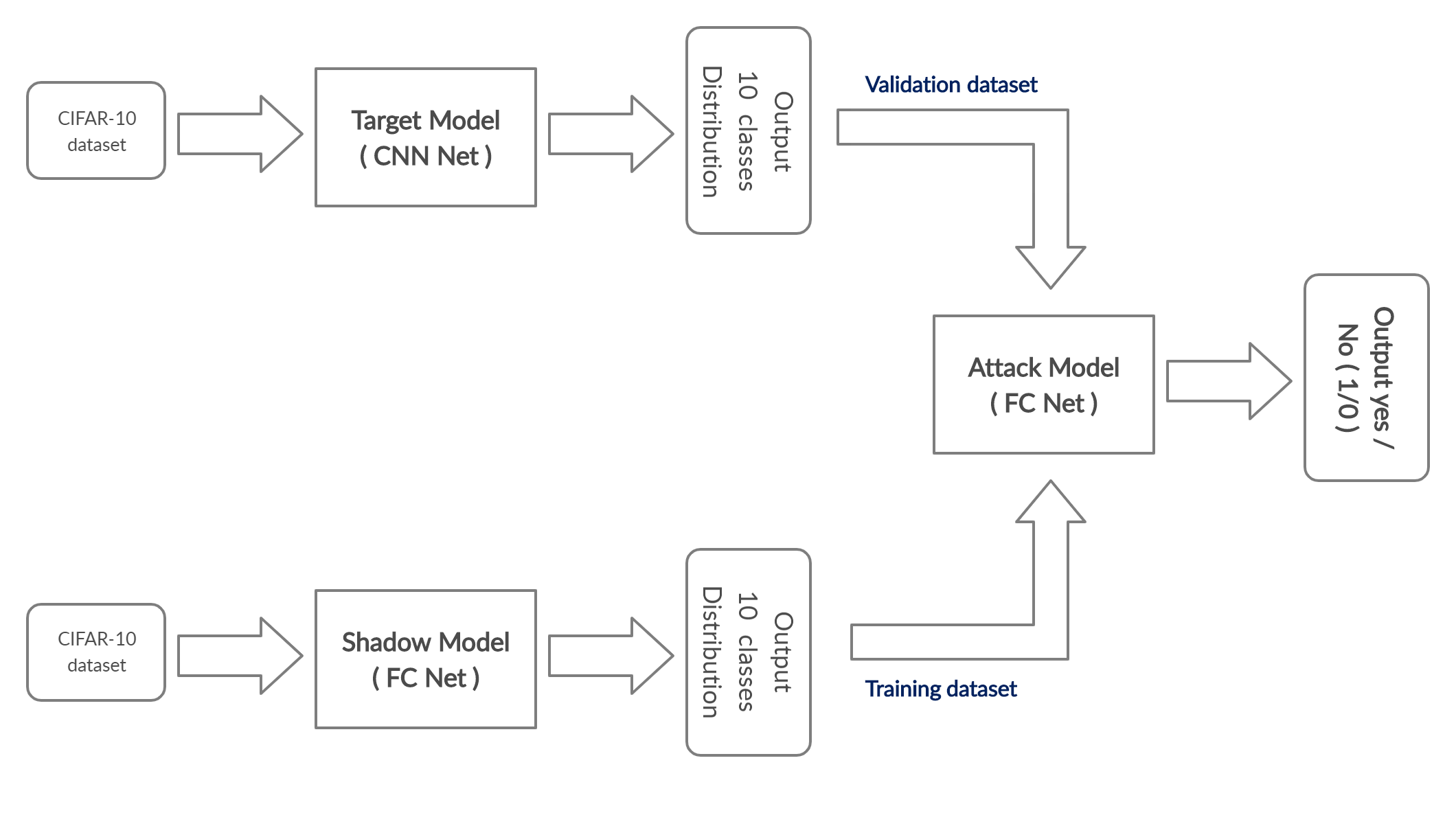
## Workflow:

As per two approaches, two workflows has been followed, Part 1 has the same architecture model (CNN) but different dataset for target and shadow model. Part 2 has the same dataset (CIFAR10) but different architecture for target and shadow model.

Part -1:

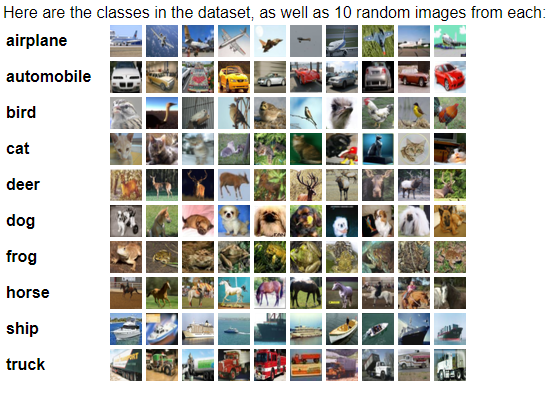


Part -2:



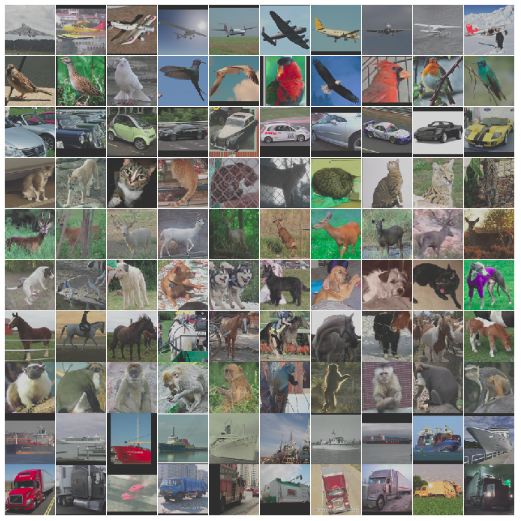
## Dataset:

Used CIFAR 10 and STL10 dataset, the CIFAR-10 dataset consists of 6K (32x32) colour images in 10 classes, with 6000 images per class.



STL dataset,

* 10 classes: airplane, bird, car, cat, deer, dog, horse, monkey, ship, truck.
* Images are 96x96 pixels, color.
* 500 training images (10 pre-defined folds), 800 test images per class.



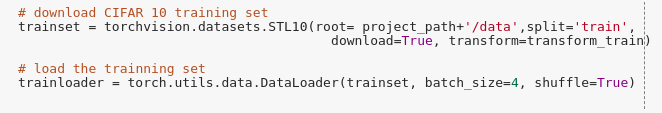
Source - [CIFAR10 from torch vision datasets](https://pytorch.org/docs/stable/torchvision/datasets.html#cifar)

[STL10 from torch vision datasets](https://pytorch.org/docs/stable/torchvision/datasets.html#stl10)

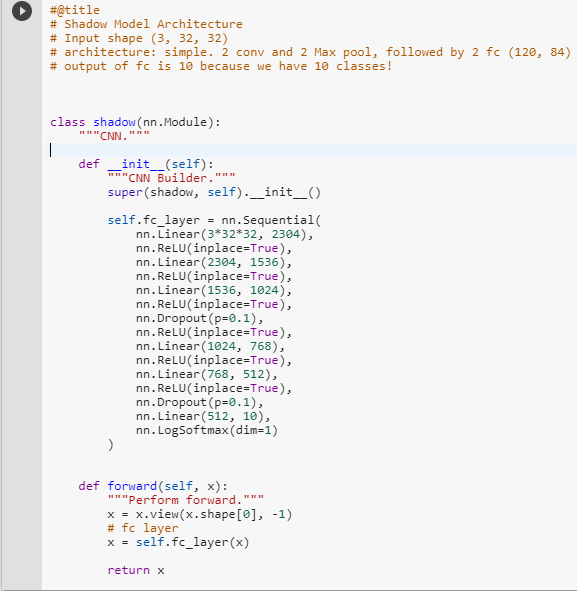
## Part - 1: Same architecture as the target model (CNN), but different dataset.

### Code and Output snippet:

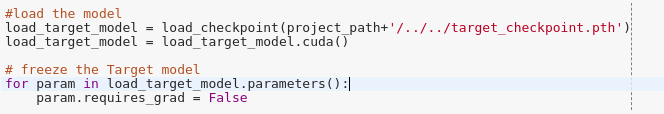
Download CIFAR-10 and STL-10 dataset



Defining the Shadow and Target model CNN - net



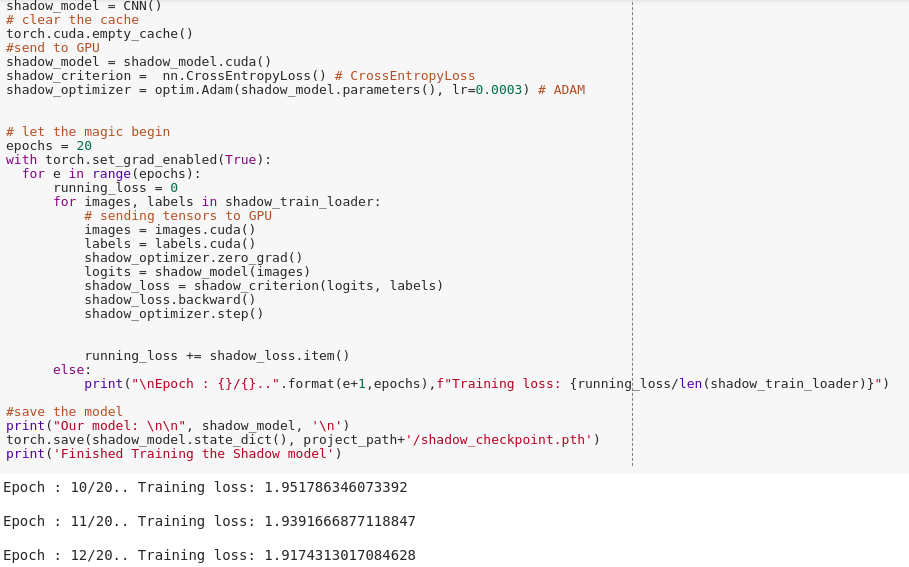
Loading the Target pre-trained model on CIFAR-10 dataset,



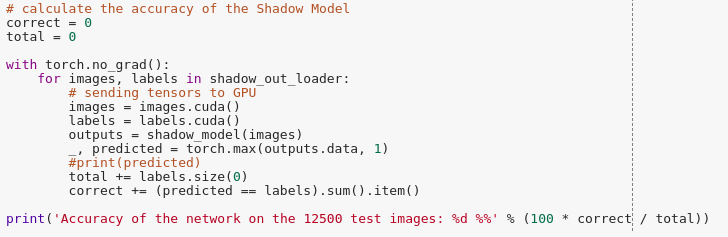
Calculate the accuracy on the test dataset



Training the Shadow model on STL-10 dataset,



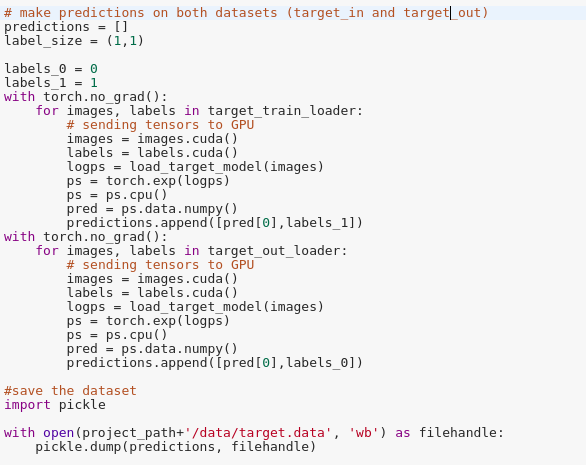
Calculate the accuracy on the test dataset



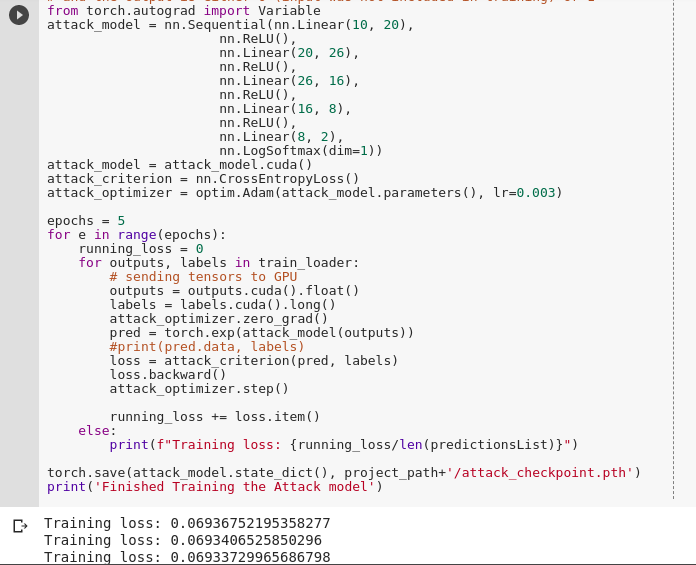
Create training dataset for the attack model, the i.e output of the shadow model is input to the attack model



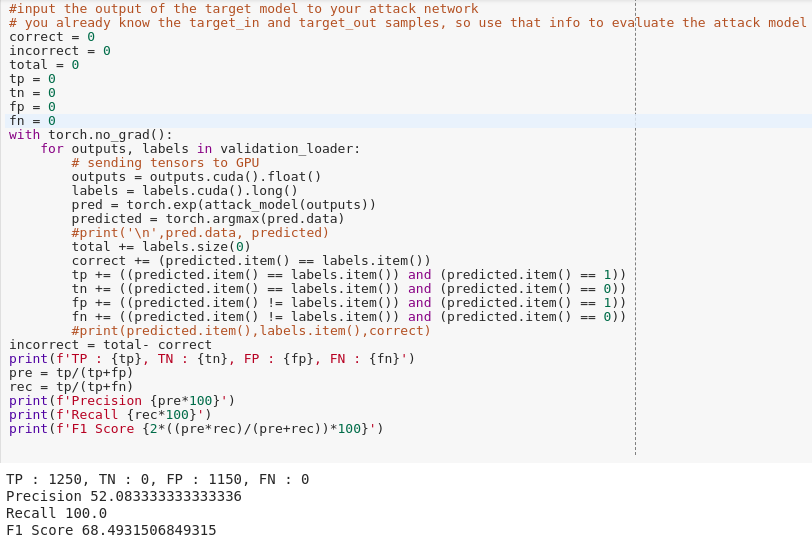
Create validation dataset for the attack model, the i.e output of the target model is input to the attack model



Training the attack model with the **training** dataset



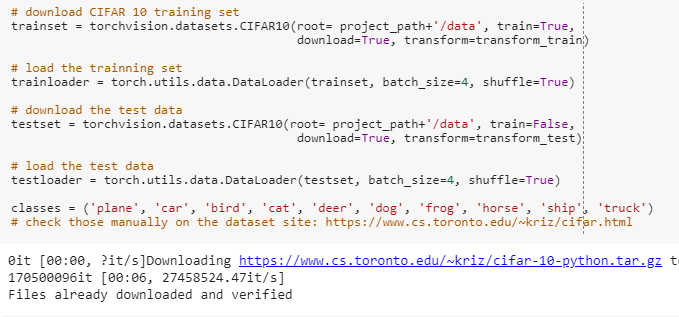
Calculate the precision and Recall of the target model on the **validation** dataset



## Part - 2: Same dataset (CIFAR10) but different architecture

### Code and Output snippet:

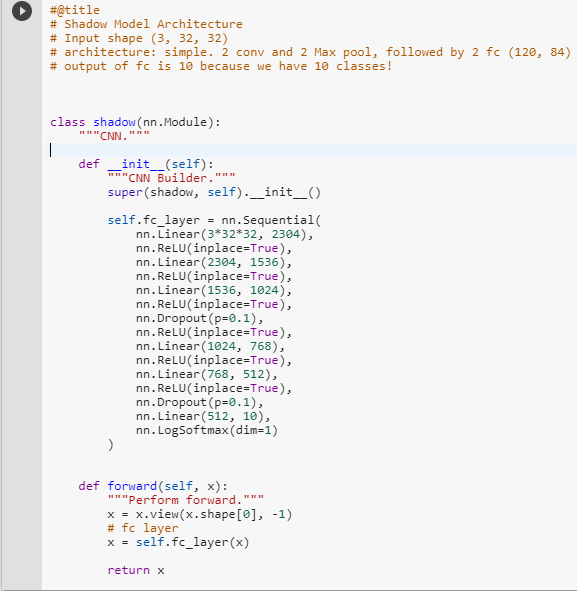
Download CIFAR-10 dataset



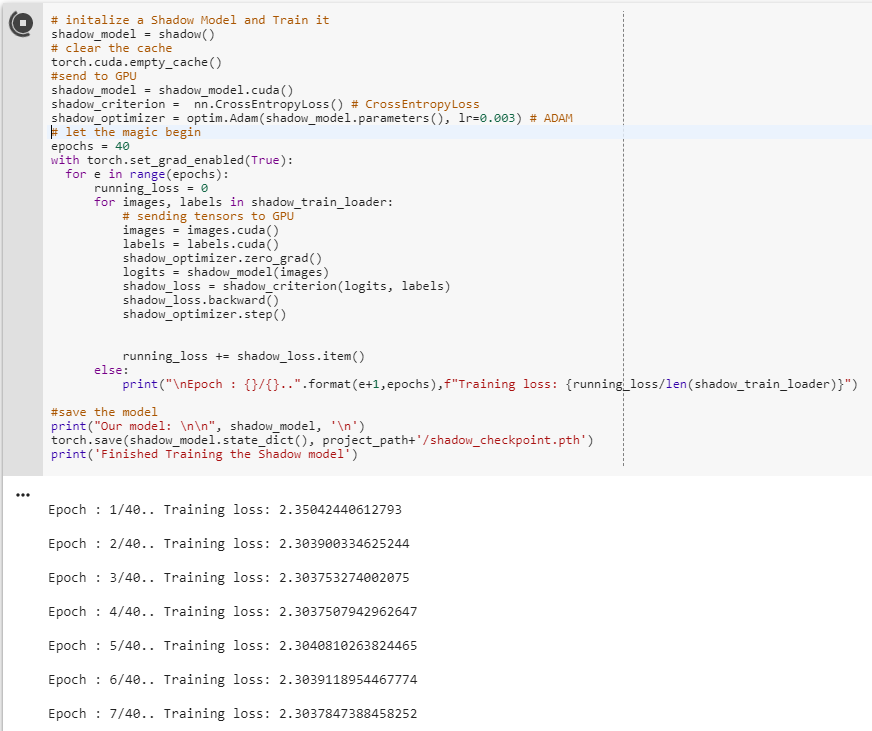
Loading the pre-trained target model ( CNN net )



Defining the Shadow model FC- net



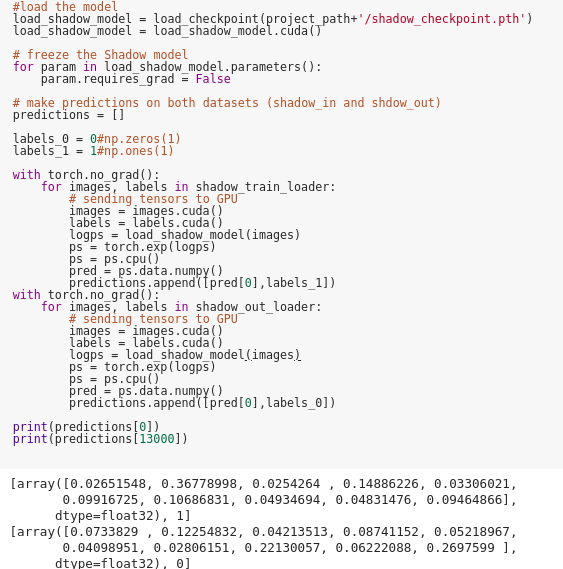
Training the shadow model,



Calculate the accuracy on the test dataset



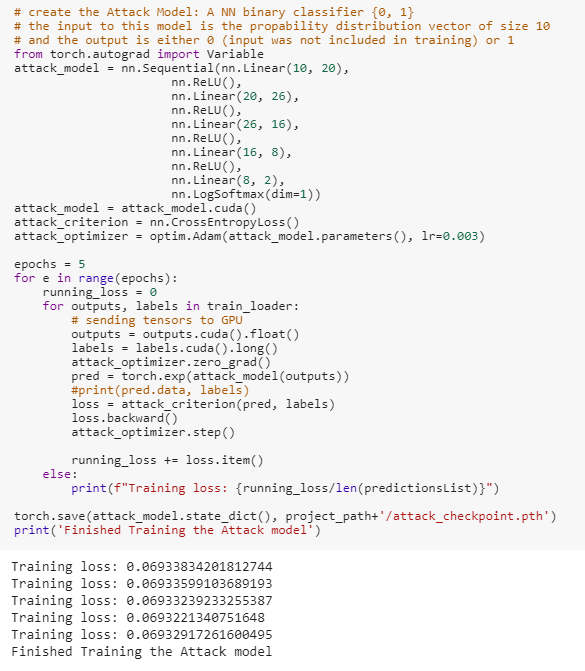
Create training dataset for the attack model, the i.e output of the shadow model is input to the attack model



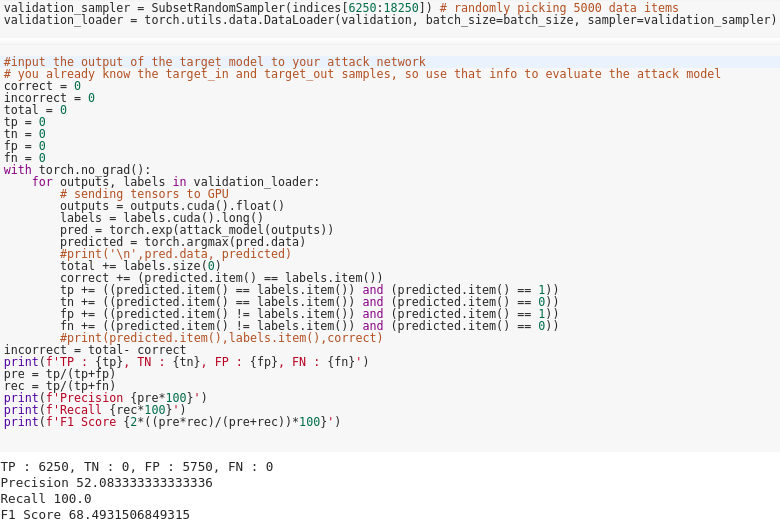
Create a validation dataset for attack model, i.e output of the target model is input to attack model



Training the attack model with the **training** dataset



Calculate the precision and Recall of the target model on the **validation** dataset



## **Evaluation & Discussion**:-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Attack model | | | |
|  |  | Precision | Recall | F1 Score |
| Part -1 we know the architecture of the target model | TP = 1250 FP = 1150 | 52.08 | 100 | 68.49 |
| Part - 2 we know the data distribution of training dataset | TP = 6250 FP = 5750 | 52.08 | 100 | 68.49 |

## **Conclusion**:-

Membership inference attack against one deep learning model was successfully implemented. The shadow training model can be used to evaluate other scenarios, mimicking a variety of models. We can also evaluate foreign models which are not trained at our place, provided dataset used. Our next venture will be trying a technique from scratch where dataset and architecture seem haywire. The data survival technique reviewed here seems to resolve most of the data leakage issues.