

Project Report 4 - Machine Learning-Based Anomaly Detection Solutions Project Overview

Student Name: Marco Ermini

Email: mermini@asu.edu

Submission Date: 11th Jun, 2021

Class Name and Term: CSE548 Summer 2021

In this lab I am using the NSL-KDD dataset, which is a refined version of KDD'99 dataset, with the purpose of running two labs for data pre-processing, training and testing using Anaconda, TensorFlow and FNN. NSL-KDD dataset is now considered as one of most common for network traffic and attacks, and it is a benchmark for modern-day internet traffic.

All the files and configurations used for this lab have been uploaded on GitHub; references are provided throughout the text and in the Appendix A at the bottom of the file.

I. NETWORK SETUP

Please find below the initial set-up of the virtual infrastructure as I have configured it in VirtualBox for this VM – it is simply connected via NAT.

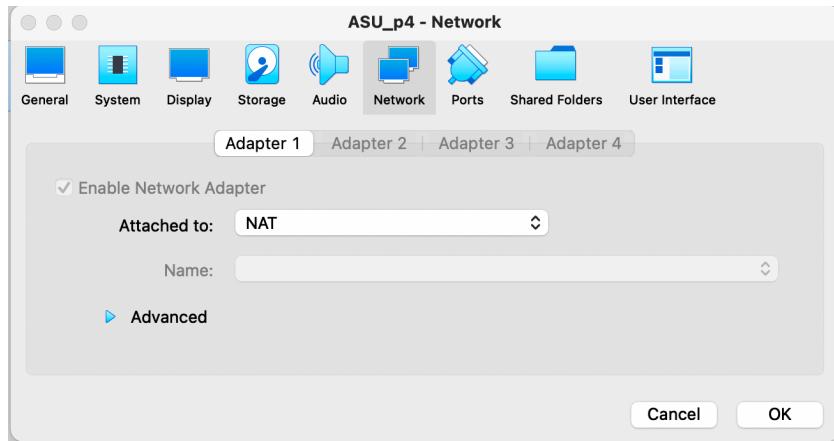


Figure 1 - Bridged network setup in VirtualBox

II. SOFTWARE

For this first lab, the following software has been used:

- Ubuntu 18.04 LTS
- Python – <https://www.python.org/>
- Anaconda - <https://www.anaconda.com/>
- TensorFlow - <https://www.tensorflow.org/>
- NSL-KDD dataset - <https://www.unb.ca/cic/datasets/nsl.html>

III. PROJECT DESCRIPTION

In this assignment, I have executed the various labs assignments, obtaining the proofs that they have been successfully completed.

The first lab (CS-ML-00201) a supervised Machine Learning (ML) approach is proposed, where Feed-forward Neural Network (FNN) solutions are used. The dataset used as input, NSL-KDD, provides labeled normal network traffic and labeled attack network traffic. FNN will use the well-labeled data to build FNN-based data pattern to differentiate normal and

abnormal network traffic. After training, we use similar data set with labeled data to validate the accuracy of the generate anomaly detection pattern.

1. Run the provided fnn sample.py python program with different NSL-KDD datasets as the input, and compare their detection accuracy

I have created the Python code to allow using the different training data by just changing a couple of comments, as shown below:

The screenshot shows a Jupyter Notebook interface. On the left, the code editor displays the `ffn_sample.py` script. Line 30 is highlighted with a grey background, showing the value `NumEpoch=20`. The right side of the interface shows the execution results. A plot titled "model loss" shows a blue line graph of loss versus epoch, starting at approximately 0.06 and decreasing to about 0.015 over 20 epochs. Below the plot, the "Console 1/A" tab shows the command `In [19]:` followed by the output of the script, which includes a warning message about TensorFlow's experimental API and the resulting accuracy values for each epoch.

```

/home/ubuntu/lab-cs-ml-00200/fnn_sample.py
ffn_sample.py X
7  #author: created by Sowmya Myneni and updated by Dijiang Huang
8  """
9
10 #####
11 # Part 1 - Data Pre-Processing
12 #####
13
14 # To load a dataset file in Python, you can use Pandas. Import pandas using the
15 import pandas as pd
16 # Import numpy to perform operations on the dataset
17 import numpy as np
18
19 # Variable Setup
20 # Available datasets: KDDTrain+.txt, KDDTest+.txt, etc. More read Data Set Int
21 # Type the training dataset file name in ''
22 TrainingDataPath='NSL-KDD/'
23 #TrainingData='KDDTrain+_20Percent.txt'
24 TrainingData='KDDTrain+.txt'
25 #TrainingData='KDDTest+.txt'
26 #TrainingData='KDDTest-21.txt'
27 # Batch Size
28 BatchSize=20
29 # Epoch Size
30 NumEpoch=20
31
32
33 # Import dataset.
34 # Dataset is given in TrainingData variable You can replace it with the file
35 # path such as "C:/Users/.../dataset.csv".
36 # The file can be a .txt as well.
37 # If the dataset file has header, then keep header=0 otherwise use header=None
38 # reference: https://www.shanelynn.ie/select-pandas-dataframe-rows-and-columns/
39 dataset = pd.read_csv(TrainingDataPath+TrainingData, header=None)
40 X = dataset.iloc[:, :-2].values
41 label column = dataset.iloc[:, -2].values

```

Model.make_predict_function.<locals>.predict_function at 0x7f20280d4dd0> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux, `export AUTOGRAPH_VERBOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with
`@tf.autograph.experimental.do_not_convert`
Print the Confusion Matrix:
[TN, FP]
[FN, TP]=
[[16800 26]
[275 14393]]
Plot the accuracy
Plot the loss

In [19]:

It is very easy in this way to collect the results and observe the differences from the various datasets. For instance, it is obvious to observe that accuracy increases with each epoch run:

```

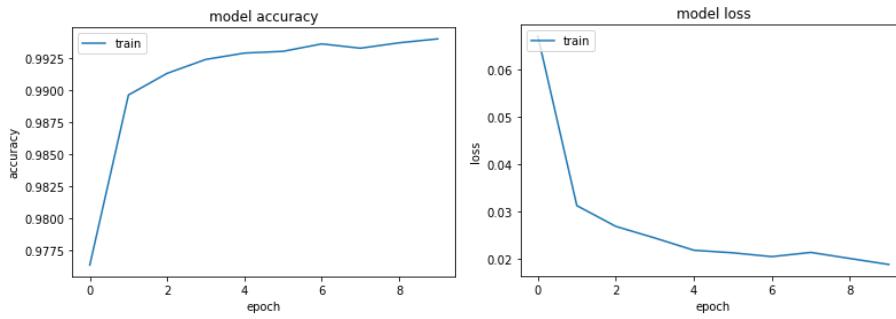
Epoch 1/20
4724/4724 [=====] - 17s 3ms/step - loss: 0.0806 - accuracy: 0.9741
Epoch 2/20
4724/4724 [=====] - 11s 2ms/step - loss: 0.0374 - accuracy: 0.9809
Epoch 3/20
4724/4724 [=====] - 15s 3ms/step - loss: 0.0308 - accuracy: 0.9878
Epoch 4/20
4724/4724 [=====] - 16s 3ms/step - loss: 0.0251 - accuracy: 0.9918
Epoch 5/20
4724/4724 [=====] - 11s 2ms/step - loss: 0.0219 - accuracy: 0.9924
Epoch 6/20
4724/4724 [=====] - 9s 2ms/step - loss: 0.0208 - accuracy: 0.9931
Epoch 7/20
4724/4724 [=====] - 13s 3ms/step - loss: 0.0186 - accuracy: 0.9937
Epoch 8/20
4724/4724 [=====] - 17s 4ms/step - loss: 0.0177 - accuracy: 0.9941
Epoch 9/20
4724/4724 [=====] - 19s 4ms/step - loss: 0.0172 - accuracy: 0.9941
Epoch 10/20
4724/4724 [=====] - 10s 2ms/step - loss: 0.0167 - accuracy: 0.9943
Epoch 11/20
4724/4724 [=====] - 15s 3ms/step - loss: 0.0162 - accuracy: 0.9945
Epoch 12/20
4724/4724 [=====] - 23s 5ms/step - loss: 0.0162 - accuracy: 0.9946
Epoch 13/20
4150/4724 [=====>,...] - ETA: 1s - loss: 0.0159 - accuracy: 0.9948

```

The followings are the results obtained by running the detection on the various datasets.

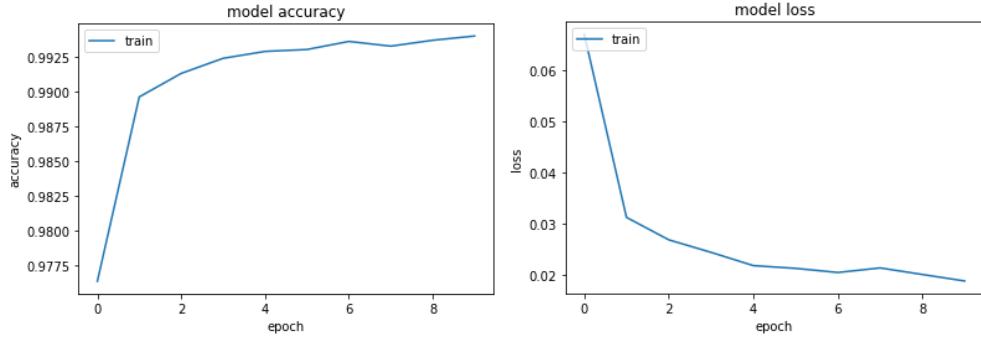
KDDTrain+: Loss 0.0221, Accuracy: 0.9912
 Confusion Matrix: TN=16806, FP=20, FN=533, TP=14135

```
Loss [0,1]: 0.0221 Accuracy [0,1]: 0.9912
WARNING:tensorflow:AutoGraph could not transform <function
Model.make_predict_function.<locals>.predict_function at 0x7f3d9d753a70> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTBOSITY=10') and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform <function Model.make_predict_function.<locals>.predict_function
at 0x7f3d9d753a70> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTBOSITY=10') and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[16806    20]
 [ 533 14135]]
```



KDDTrain+20Percent: Loss 0.0319, Accuracy: 0.9835
 Confusion Matrix: TN=3386, FP=1, FN=174, TP=2737

```
Loss [0,1]: 0.0319 Accuracy [0,1]: 0.9835
WARNING:tensorflow:AutoGraph could not transform <function
Model.make_predict_function.<locals>.predict_function at 0x7f3d9c3f5290> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTBOSITY=10') and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform <function Model.make_predict_function.<locals>.predict_function
at 0x7f3d9c3f5290> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTBOSITY=10') and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[3386    1]
 [ 174 2737]]
```



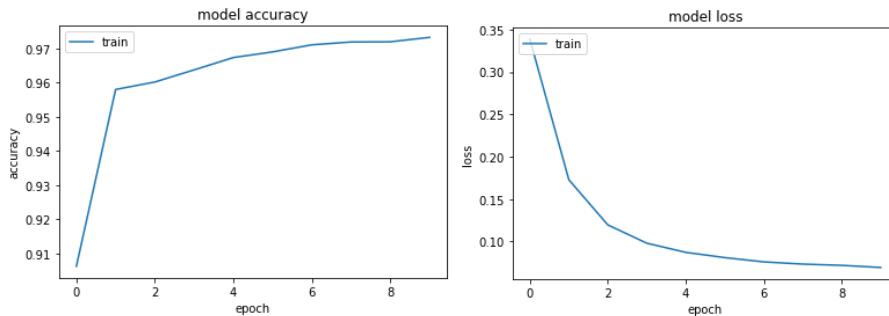
KDDTest+: Loss 0.0675, Accuracy: 0.9683

Confusion Matrix: TN=2486, FP=18, FN=356, TP=2776

```

Loss [0,1]: 0.0675 Accuracy [0,1]: 0.9683
WARNING:tensorflow:AutoGraph could not transform <function
Model.make_predict_function.<locals>.predict_function at 0x7f3e04e45b00> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform <function Model.make_predict_function.<locals>.predict_function
at 0x7f3e04e45b00> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[2486  18]
 [ 356 2776]]

```



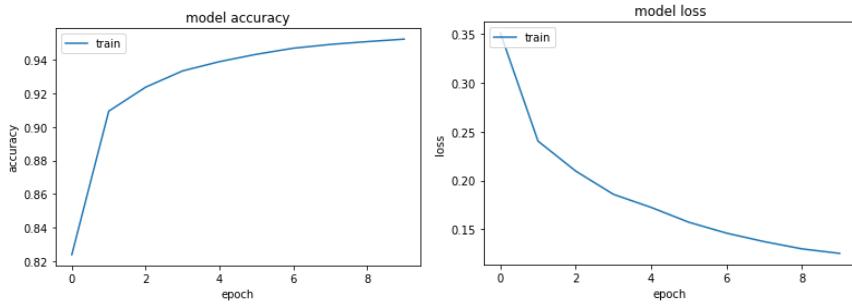
KDDTest-21: Loss 0.1338, Accuracy: 0.9487

Confusion Matrix: TN=550, FP=16, FN=342, TP=2055

```

Loss [0,1]: 0.1338 Accuracy [0,1]: 0.9487
WARNING:tensorflow:AutoGraph could not transform <function
Model.make_predict_function.<locals>.predict_function at 0x7f3e0efde9e0> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform <function Model.make_predict_function.<locals>.predict_function
at 0x7f3e0efde9e0> and will run it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity to 10 (on Linux,
`export AUTOGRAPH_VERTOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with @tf.autograph.experimental.do_not_convert
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[ 550   16]
 [ 342 2055]]

```



a. Comparing the datasets:

KDDTrain+.txt vs. KDDTrain+_20Percent.txt

KDDTrain+: Loss 0.0221, Accuracy: 0.9912
Confusion Matrix: TN=16806, FP=20, FN=533, TP=14135

KDDTrain+20Percent: Loss 0.0319, Accuracy: 0.9835
Confusion Matrix: TN=3386, FP=1, FN=174, TP=2737

In this case a much bigger dataset (KDDTrain+) provides a superior accuracy and minor loss.

KDDTest+.txt vs. KDDTrain+_20Percent.txt

KDDTest+: Loss 0.0675, Accuracy: 0.9683
Confusion Matrix: TN=2486, FP=18, FN=356, TP=2776

KDDTrain+20Percent: Loss 0.0319, Accuracy: 0.9835
Confusion Matrix: TN=3386, FP=1, FN=174, TP=2737

In this case a training dataset provides more accuracy and less loss since the test is smaller.

KDDTrain+.txt vs. KDDTest+.txt

KDDTrain+: Loss 0.0221, Accuracy: 0.9912
Confusion Matrix: TN=16806, FP=20, FN=533, TP=14135

KDDTest+: Loss 0.0675, Accuracy: 0.9683
Confusion Matrix: TN=2486, FP=18, FN=356, TP=2776

Again, the training dataset provides more accuracy and less loss since it has a bigger set of data.

KDDTrain+_20Percent.txt vs. KDDTest-21.txt

KDDTrain+: Loss 0.0221, Accuracy: 0.9912
Confusion Matrix: TN=16806, FP=20, FN=533, TP=14135

KDDTest-21: Loss 0.1338, Accuracy: 0.9487
Confusion Matrix: TN=550, FP=16, FN=342, TP=2055

In this case, the training set has even a bigger difference against the test data, since the latter have the most difficult traffic record (score of 21) excluded from the training, which means it has less quality input to train with.

2. Run the provided fnn sample.py python program with different one NSL-KDD dataset: KD- DTrain+.txt as the input and compare their detection accuracy with a different neural network setup. Explain your observations.

The fnn_sample.py has been modified as required, as shown below.

The screenshot shows a Jupyter Notebook interface with the following components:

- Code Cell:** The code in ffn_sample.py includes imports for pandas, numpy, and tensorflow, variable setup (TrainingDataPath='NSL-KDD'), training parameters (BatchSize=20, NumEpochs=20), and dataset loading (pd.read_csv). It also includes a warning about TensorFlow's experimental do_not_convert function.
- Plot:** A line graph titled "model loss" showing the training loss over 18 epochs. The loss starts at approximately 0.06 and decreases steadily to about 0.02.
- Console Output:** The output shows the final loss (0.0132) and accuracy (0.9959), followed by a confusion matrix:

```

Loss [0,1]: 0.0132 Accuracy [0,1]: 0.9959
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[16795    31]
 [ 257 14411]]

```

KDDTrain+.txt (1st run) vs. KDDTrain+.txt (2nd run)

KDDTrain+ 1st run: Loss 0.0221, Accuracy: 0.9912
 Confusion Matrix: TN=16806, FP=20, FN=533, TP=14135

KDDTrain+ 2nd run: Loss 0.0132, Accuracy: 0.9959
 Confusion Matrix: TN=16795, FP=31, FN=257, TP=14411

A run with an increased depth of the FNN (more layers) and increased epoch seems to increase the accuracy and diminish the loss – which is expected. The run with the increased value has ran much slower – it could be perceived although I have not measured precisely. Interestingly enough, while the TN and TP have increased, so they are the FP – albeit the FN have diminished. So, surprisingly, it seems that not all the results have not improved. It would be interesting to investigate the reason for it (although it is out of scope for this lab).

In the second lab (CS-ML-00301), I use basic FNN model to create an anomaly detection model for network intrusion detection.

3. Task 2: Create Data Modules for Anomaly Detection

First of all, I create the datasets using `DataExtractor.py`. The datasets are all available on the GitHub repository.

The screenshot shows a terminal window with the following output:

```
(base) ubuntu@ubuntu:~/lab-cs-ml-00301$ ll
total 39876
drwxr-xr-x  3 ubuntu ubuntu      4096 Jul 11 16:50  .
drwxr-xr-x 29 ubuntu ubuntu      4096 Jul 11 16:10  ..
-rwxr-xr-x  1 ubuntu ubuntu     3717 Apr 19 2020 categoryMapper.py*
-rw-r--r--  1 ubuntu ubuntu     5243 Jul 11 16:42  dataExtractor.py
-rw-r--r--  1 ubuntu ubuntu     5127 Jul 11 16:20  dataExtractor.py.orig
-rwxr-xr-x  1 ubuntu ubuntu     2385 Apr 19 2020 data_preprocessor.py*
-rw-r--r--  1 ubuntu ubuntu     1762 Apr 19 2020 distinctLabelExtractor.py
-rw-r--r--  1 ubuntu ubuntu     6148 Apr 20 2020 .DS_Store
-rwxr--r--  1 ubuntu ubuntu     7568 Jul 11 16:50  fnn_sample.py*
-rwxr--r--  1 ubuntu ubuntu     7093 Jul 11 16:43  fnn_sample.py.orig*
drwx----- 2 ubuntu ubuntu      4096 Jul 11 16:10  NSL-KDD/
-rwxr--r--  1 ubuntu ubuntu   176189 Apr 15 2019 Spyder ReadMe.pdf*
-rw-r--r--  1 ubuntu ubuntu   2757039 Jul 11 16:41  Testing-a1-a2-a3.csv
-rw-r--r--  1 ubuntu ubuntu   2415032 Jul 11 16:40  Testing-a1.csv
-rw-r--r--  1 ubuntu ubuntu   2099894 Jul 11 16:39  Testing-a2-a4.csv
-rw-r--r--  1 ubuntu ubuntu   17445086 Jul 11 16:40  Training-a1-a2.csv
-rw-r--r--  1 ubuntu ubuntu  15855630 Jul 11 16:39  Training-a1-a3.csv
```

4. Lab Assessment

a. Using the lab screenshot feature to document ML running results only. Provide sufficient but concise explanations on the generated results for each given training/testing scenarios.

The followings are the results from the three scenarios. It must be noted that the accuracy shown in the Python printout is the accuracy while training the model. What we are interested in, is the accuracy when applying the trained FNN model to test dataset, which is calculated as:

$$\text{accuracy} = (\text{TP} + \text{TN}) / (\text{TN} + \text{FP} + \text{FN} + \text{TP})$$

This is the result that is presented below, already calculated.

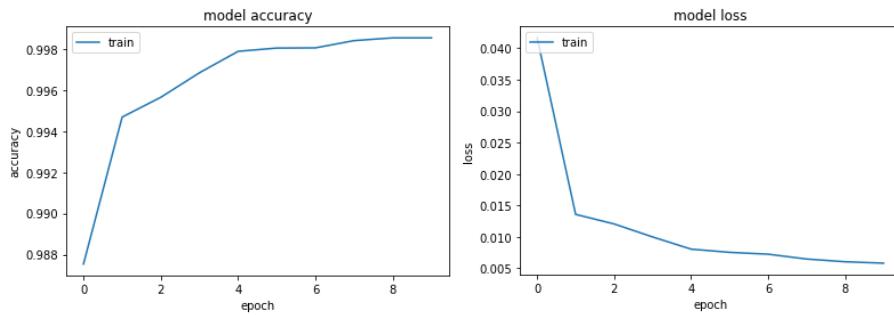
a) **Scenario A:** Loss=0.0052, Accuracy=11505/15017=**0.7661**, Confusion Matrix: TN=8701, FP=1010, FN=2502, TP=2804

```
In [1]: runfile('/home/ubuntu/lab-cs-ml-00301/fnn_sample.py', wdir='/home/ubuntu/lab-cs-ml-00301')

Please enter the scenario you wish to run - either a, b or c:a
2021-07-11 19:25:09.071681: I tensorflow/core/platform/cpu_feature_guard.cc:143]
Your CPU supports instructions that this TensorFlow binary was not compiled to use:
SSE4.1 SSE4.2 AVX AVX2
2021-07-11 19:25:09.091889: I tensorflow/core/platform/profile_utils/cpu_utils.cc:
102] CPU Frequency: 2592000000 Hz
2021-07-11 19:25:09.092085: I tensorflow/compiler/xla/service/service.cc:168] XLA
service 0x55aa45fab370 initialized for platform Host (this does not guarantee that
XLA will be used). Devices:
2021-07-11 19:25:09.092101: I tensorflow/compiler/xla/service/service.cc:176]
StreamExecutor device (0): Host, Default Version
2021-07-11 19:25:09.093911: I tensorflow/core/common_runtime/process_util.cc:147]
Creating new thread pool with default inter op setting: 2. Tune using
```

```
Epoch 10/10
11333/11333 [=====] - 22s 2ms/step - loss: 0.0058 -
accuracy: 0.9985
3542/3542 [=====] - 8s 2ms/step - loss: 0.0052 - accuracy:
0.9988
Print the loss and the accuracy of the model on the dataset
Loss [0,1]: 0.0052 Accuracy [0,1]: 0.9988
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[8701 1010]
 [2502 2804]]
Plot the accuracy
Plot the loss

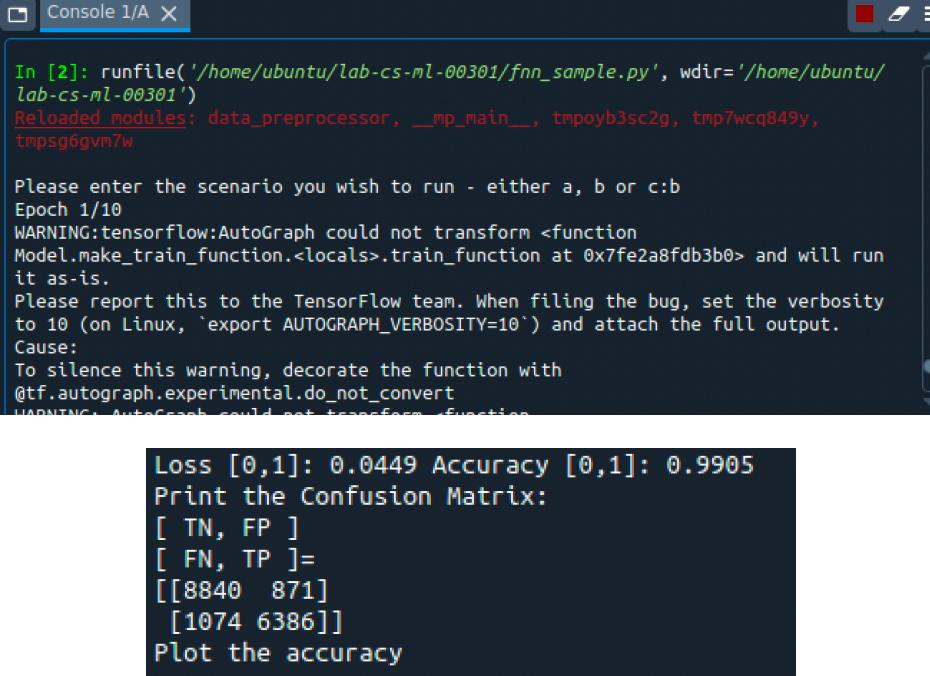
In [2]:
```



Explanation of Scenario A: this scenario uses “DoS” and “U2R” attacks as training and “Probing” and “R2L” attacks in the test cases. Basically, there is no overlap between the training and test cases, and therefore we would expect this scenario to be the one that is worst performing.

That Scenario produces the worst result in terms of Confusion Matrix, being the worst performer on the Loss and Accuracy values. This is likely be attributed to the missing overlapping of Training and Test cases.

b) **Scenario B:** Loss=0.0449, Accuracy=15226/17171=0.8867, Confusion Matrix: TN=8840, FP=871, FN=1074, TP=6386



```
In [2]: runfile('/home/ubuntu/lab-cs-ml-00301/fnn_sample.py', wdir='/home/ubuntu/lab-cs-ml-00301')
Reloaded modules: data_preprocessor, __mp_main__, tmpoyb3sc2g, tmp7wcq849y,
tmpsg6gvm7w

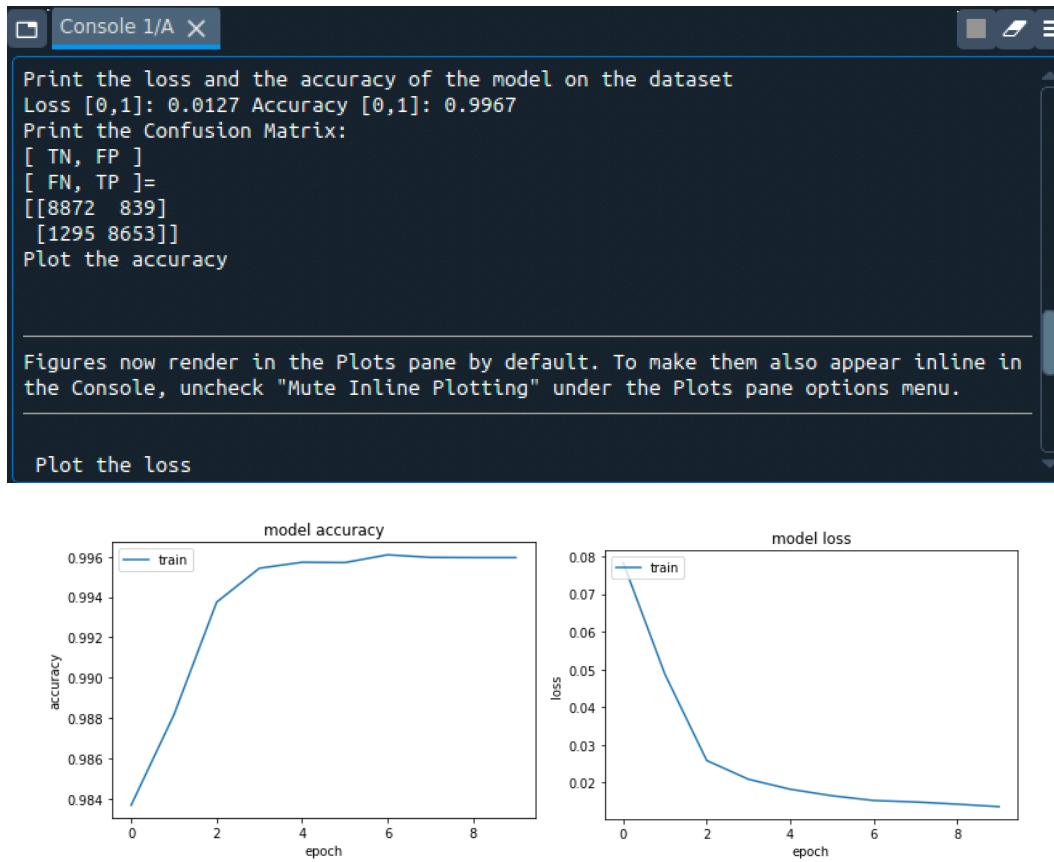
Please enter the scenario you wish to run - either a, b or c:b
Epoch 1/10
WARNING:tensorflow:AutoGraph could not transform <function
Model.make_train_function.<locals>.train_function at 0x7fe2a8fdb3b0> and will run
it as-is.
Please report this to the TensorFlow team. When filing the bug, set the verbosity
to 10 (on Linux, `export AUTOGRAPH_VERTOSITY=10`) and attach the full output.
Cause:
To silence this warning, decorate the function with
@tf.autograph.experimental.do_not_convert
WARNING: AutoGraph could not transform function
```

```
Loss [0,1]: 0.0449 Accuracy [0,1]: 0.9905
Print the Confusion Matrix:
[ TN, FP ]
[ FN, TP ]=
[[8840 871]
 [1074 6386]]
Plot the accuracy
```

The figure contains two line plots. The left plot, titled "model accuracy", shows accuracy on the y-axis (ranging from 0.984 to 0.990) against epoch on the x-axis (ranging from 0 to 8). The right plot, titled "model loss", shows loss on the y-axis (ranging from 0.05 to 0.10) against epoch on the x-axis (ranging from 0 to 8). Both plots show a single blue line labeled "train".

Explanation of Scenario B: this scenario uses “DoS” and “Probe” attacks as training only “DoS” attacks in the test cases. Basically, there is a complete overlap between the training and test cases, as the training includes all the attacks that will be found in the test, and therefore we would expect this scenario to be performing much better than Scenario A – which is what we have, in fact, observed.

c) **Scenario C:** Loss=0.0127, Accuracy=17525/19659=**0.8914**, Confusion Matrix: TN=8872, FP=839, FN=1295, TP=8653



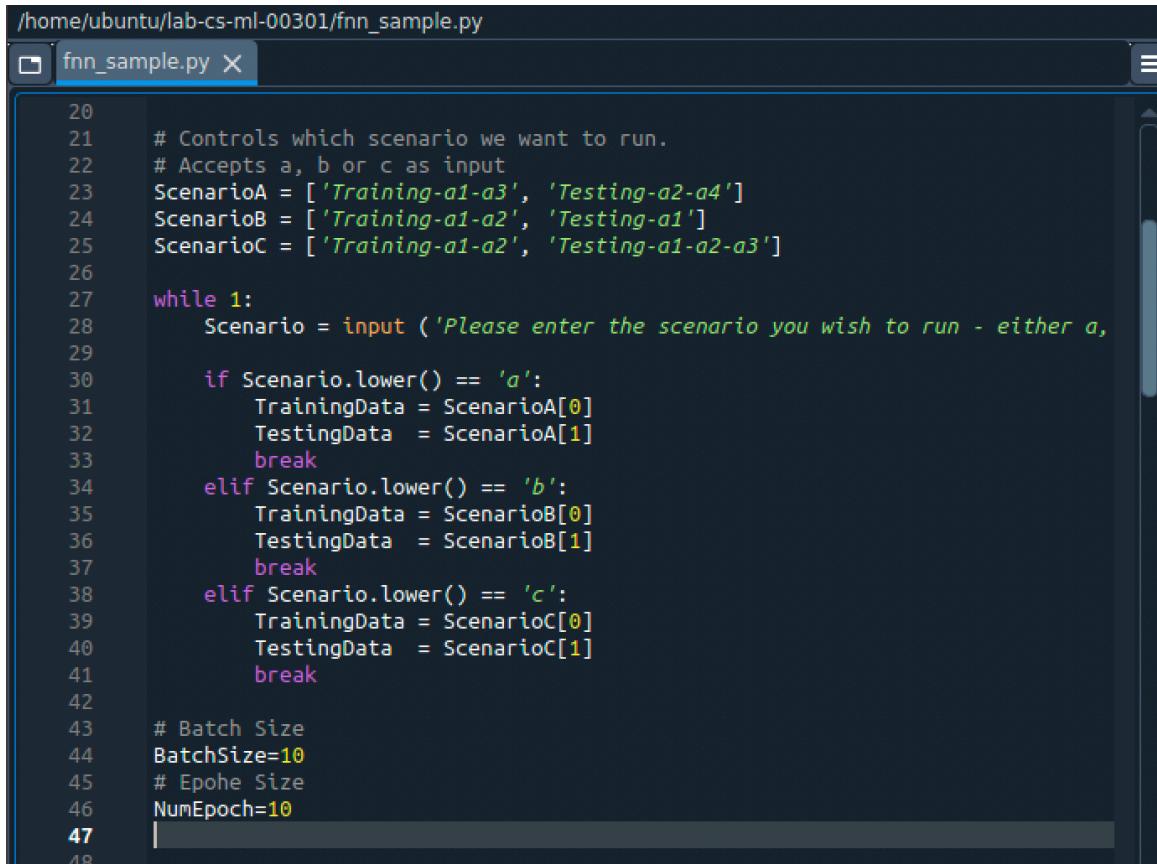
Explanation of Scenario C: this scenario uses “DoS” and “Probe” attacks as training (exactly like Scenario 2) but is instead presented with “DoS”, “Probe” and “R2L” attacks in the test cases.

This Scenario is the most intriguing for the fact that it is performing better than the other two, but it does not have the same level of overlap than Scenario B – there are attacks presented in the test that are not offered in the training, but despite that, it managed to perform better.

b. Submit the updated fnn simple.py and provide sufficient comments to illustrate your updated codes.

All the codes shown in the pictures below are available on GitHub, as well as in the provided ZIP file.

My first change is that I have modified fnn_sample.py so that it takes one of the three scenarios (A, B, C) as input. Selecting the scenario will automatically select the respective datasets.



```

/home/ubuntu/lab-cs-ml-00301/fnn_sample.py
fnn_sample.py X

20
21     # Controls which scenario we want to run.
22     # Accepts a, b or c as input
23     ScenarioA = ['Training-a1-a3', 'Testing-a2-a4']
24     ScenarioB = ['Training-a1-a2', 'Testing-a1']
25     ScenarioC = ['Training-a1-a2', 'Testing-a1-a2-a3']
26
27     while 1:
28         Scenario = input ('Please enter the scenario you wish to run - either a,
29
30             if Scenario.lower() == 'a':
31                 TrainingData = ScenarioA[0]
32                 TestingData = ScenarioA[1]
33                 break
34             elif Scenario.lower() == 'b':
35                 TrainingData = ScenarioB[0]
36                 TestingData = ScenarioB[1]
37                 break
38             elif Scenario.lower() == 'c':
39                 TrainingData = ScenarioC[0]
40                 TestingData = ScenarioC[1]
41                 break
42
43     # Batch Size
44     batchSize=10
45     # Epoch Size
46     numEpoch=10
47
48

```

From line 56 up to the implementation of the FNN, I have implemented the code that loads the two pre-processed set (training and test) and then commented out all the code that creates the “hot” split of test and training data, as this is not necessary.

```

49      # Import the Dataset specified in the TrainingData variable.
50      # It will be automatically set by choosing the appropriate scenario at execution
51      #   time.
52      # If the dataset file has header, then keep header=0 otherwise use header=None
53      # reference: https://www.shanelynn.ie/select-pandas-dataframe-rows-and-columns-using-iloc-loc-and-.ix/
54
55      import data_preprocessor as dp
56      X_train, y_train = dp.get_processed_data(TrainingData, './categoryMappings/', classType ='binary'
57      X_test, y_test = dp.get_processed_data(TestingData, './categoryMappings/', classType ='binary'
58
59      # The next section from fnn_sample.py is not required, as data is already preprocessed
60
61      # Encoding categorical data (convert letters/words in numbers)
62      # Reference: https://medium.com/@contact sunny/label-encoder-vs-one-hot-encoder-in-machine-learning-43a2a2a2a2
63      # The following code work without warning in Python 3.6 or older. Newer versions suggest to use
64      #
65      #from sklearn.preprocessing import LabelEncoder, OneHotEncoder
66      le = LabelEncoder()
67      X[:, 1] = le.fit_transform(X[:, 1])
68      X[:, 2] = le.fit_transform(X[:, 2])
69      X[:, 3] = le.fit_transform(X[:, 3])
70      onehotencoder = OneHotEncoder(categorical_features = [1, 2, 3])
71      X = onehotencoder.fit_transform(X).toarray()
72      #
73
74      # The following code work Python 3.7 or newer
75      #from sklearn.preprocessing import OneHotEncoder
76      #from sklearn.compose import ColumnTransformer
77      #ct = ColumnTransformer(
78      #    [('one_hot_encoder', OneHotEncoder(), [1,2,3])],    # The column numbers to be transformed
79      #    remainder='passthrough'                                # Leave the rest of the columns untouched
80      #)
81      #X = np.array(ct.fit_transform(X), dtype=np.float)
82
83      # Splitting the dataset into the Training set and Test set (75% of data are used for training)
84      # reference: https://scikit-learn.org/stable/modules/generated/sklearn.model\_selection.train\_test\_split.html
85      #from sklearn.model_selection import train_test_split

```

From line 98 (after the comment “Part 2: Building FNN”), the code is basically unmodified.

5. Task 3: Anomaly detection analysis

1. The Scenario producing the best results in terms of accuracy is **Scenario C**. The best results are obtained when training using “DoS” and “Probe” datasets rather than “DoS” and “U2R”. This is most likely because “Probe” dataset is much larger than the “U2R”, since they hold respectively the 9.11% and 0.04% of the datasets, which is a representation of the attacks that happens on the Internet. Because of that, catching “Probe” attacks rather than “U2R” ones yield to a better statistical outcome in the FNN (which may or may not represent a success from a purely Security perspective).
2. The average accuracy between Scenario A, B and C is (rounded) **0.8481**. When considering only Scenario A and C, it is (rounded) **0.8288**, which is inferior to the average including Scenario B. This demonstrates that when attacks present cases for which the FNN has not been trained for, the performance is inferior.
3. As noticed, when a Scenario is presented with attacks for which it has not been trained, it will perform worst in terms of Accuracy, as it to be expected. However, this does not automatically translate into more attacks blocked.
4. The R2L and U2L use similar patters – for instance they are mostly using TCP protocol as they must transmit an exploit, deposit a rootkit or being transported via HTTP.
Another issue is determined by the fact that “U2R” and “R2L” attacks are numerically much inferior to “DoS” and “Probe” attacks, which reflects negatively on the training.

IV. APPENDIX A: FILES FOR THE LAB

Please find the list of files created for this lab and mentioned throughout this document, plus their GitHub link for download.

The overall GitHub directory for the project is: <https://github.com/markoer73/CSE-548/tree/main/Project%20-%20SDN-Based%20Stateless%20Firewall>

Project Report 4 - Machine Learning-Based Anomaly Detection Solutions.docx	https://github.com/markoer73/CSE-548/blob/main/Project%204%20-%20Machine%20Learning-Based%20Anomaly%20Detection%20Solutions/Project-Report-4%20-%20Machine%20Learning-Based%20Anomaly%20Detection%20Solutions.docx
fnn_sample.py	https://github.com/markoer73/CSE-548/blob/main/Project%204%20-%20Machine%20Learning-Based%20Anomaly%20Detection%20Solutions/CS-ML-00301/fnn_sample.py

V. REFERENCES

Data preprocessing:

- <https://www.shanelynn.ie/select-pandas-dataframe-rows-and-columns-using-iloc-loc-and-ix/>
- <https://medium.com/@contact sunny/label-encoder-vs-one-hot-encoder-in-machine-learning-3fc273365621>
- https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html
- <https://scikit-learn.org/stable/modules/preprocessing.html>

Build ANN:

- <https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/>
- <https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/>
- <https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/>
- <https://machinelearningmastery.com/how-to-choose-loss-functions-when-training-deep-learning-neural-networks/>

Contents

I. Network Setup.....	1
II. Software	1
III. Project Description.....	1
1. Run the provided fnn sample.py python program with different NSL-KDD datasets as the input, and compare their detection accuracy	2
a. Comparing the datasets:	5
2. Run the provided fnn sample.py python program with different one NSL-KDD dataset: KD- DTrain+.txt as the input and compare their detection accuracy with a different neural network setup. Explain your observations.....	6
3. Task 2: Create Data Modules for Anomaly Detection.....	7
4. Lab Assessment	8
a. Using the lab screenshot feature to document ML running results only. Provide sufficient but concise explanations on the generated results for each given training/testing scenarios.	8
b. Submit the updated fnn simple.py and provide sufficient comments to illustrate your updated codes.....	12
5. Task 3: Anomaly detection analysis	14
IV. Appendix A: Files for the Lab	15
V. References	15

