CSCI 356 Term Project

As part of your overall evaluation in the course, you are required to develop the Convolution Neural Network (CNN) model for Network Anomaly Detection. This is either an individual or a group project, so every student is required to involve in the project. Please read the following descriptions to understand this term project.

1. Prior to starting the project, you need to read the paper, entitled "*An Empirical Study on Network Anomaly Detection using Convolutional Neural Networks*". As already explained in class, 3 CNN models: shallow, moderate and deep were developed using 3 different datasets: NSL-KDD, MAWILab, and Kyoto.
2. What you need to do in this term project as a group project is that your team is required to develop 3 CNN models using NSL-KDD and MAWILab datasets. If you work individually, please select one of the datasets and develop 3 CNN models. Architecture of the CNN models is shown in Table 2 in the paper, and the following sub-sections describes what to do in detail with the NSL-KDD dataset.
    a. Preprocessed NSL-KDD dataset and Shallow CNN Train+Test+.ipynb has been already provided. Please use them for your seed files. Run Shallow CNN Train+Test+.ipynb to check the accuracy of this model. Note that your accuracy (F1-Score) should be close to ***80%***.
    b. Run the Shallow CNN model with NSL-KDD Train+ and Test-
    c. Run the Shallow CNN model with NSL-KDD Train- and Test+
    d. Run the Shallow CNN model with NSL-KDD Train- and Test-
    e. Build both moderate and deep CNN models according to Table 2. Run each model with all combinations of the NSL-KDD dataset, and make sure if your accuracy is similar to Fig. 1 in the paper. Note that you should have a total of 12 ipynb files for the NSL-KDD datasets

3. What to do with the MAWILab dataset
    a. A total of 4 MAWILab data files are provided: 20170827_mawilab_flow_001, 20170827_mawilab_flow_002, 20170827_mawilab_flow_003, and 20170827_mawilab_flow_004
    b. As described in Section 4.C in the paper, please use the flow001 data for training and the rest of flow data files for testing. However, please keep in mind that please use only 10% of the entire dataset. Otherwise, it will take a long time to train and test your models.
    c. The first job is to perform data preprocessing. The MAWILab dataset is composed of a total of 29 columns, but only 5 columns: ***pro, packets, bytes, durat,*** and ***status*** need to be employed. Please note that the number of columns will be 6 after the one-hot encoding technique is applied. Through the step of data preprocessing, you will have multiple *.pkl* files.
    d. Once data preprocessing is done, you need to build 3 CNN models as described in Section 2 above.
        i. Run the Shallow CNN model with mawilab_flow_001 and mawilab_flow_002
        ii. Run the Shallow CNN model with mawilab_flow_001 and mawilab_flow_003
        iii. Run the Shallow CNN model with mawilab_flow_001 and mawilab_flow_004
        iv. Please build and run all other CNN models with mawilab_flow_001 and mawilab_flow_002, mawilab_flow_001 and mawilab_flow_003, and mawilab_flow_001 and mawilab_flow_004.
4. Deliverables
    a. All tensorflow files related to data preprocessing of the MAWILab dataset
    b. All tensorflow files related to CNN models
    c. Document to show the accuracy of your CNN models
5. Due date: By 11:59PM, 11/17/2020