Study of Internet Streaming Traffic Classification Using Deep Learning

Kazutoshi Nakano (1310351)
School of Information Science,
Japan Advanced Institute of Science and Technology
August 5th, 2017

Keywords: Traffic classification, Traffic identification, Deep Learning

In the last few decades, Accurate Internet traffic classification is an important issue for the two reason, (1) network management tasks, QoS (quality of service), bandwidth control, application monitoring (2) some security issue, intrusion-detection. Traditionally, there are two methods to classify the traffic, port-base and payload inspection. But both methods are not reliable in a few case, for example, encrypted traffic, Voice over IP (voip) streaming application using dynamic port usage. Recent years, an Alternative approach is to classify traffic using machine learning (ML) techniques by statistical flow feature, for example, duration, total data size, mean inter arrival time. In this paper, we propose the flow feature based traffic classification using deep learning technique.

First, we surveyed the related work. Recent years, An classification method using statistical feature of traffic flows has been proposed as a method that can be classified on a per flow basis. Various statistical feature can be calculated for each traffic flow, if a traffic flow is defined as transmission and receive of a series of packets from the beginning to the end of communication. For example, as the statistical feature, the total transfer packet number, the total transfer bytes, the average packet size, etc. It is assumed that there is a characteristic for each application in this statistical feature. It is a method of classifying application traffic based on statistic feature since it is assumed that statistics feature does not include port number and payload itself but it can be extracted even if traffic are encrypted. However, in order to classify using statistical features, there is a problem that data of a large number of traffic flows is required.

In more detail about our research approach, we use NIMS dataset (https://projects.cs.dal.ca/projectx/Download.html) for the input of Deep Neural Network (DNN). This dataset includes 13 protocols , for example, SSH encrypted traffic, Voip streaming traffic. Each flow has 22 statistical flow features. In order to accurately learn the neural network, 2000 flows were extracted from

each protocol randomly so as to equalize the number of flows.

DNN is a structure that two or more hidden layers of the neural network. DNN is one of the supervised Machine-Learning approach. In our result, we found the structure of DNN hyper-parameter for classify traffic efficiently as follows, 6 hidden layers, "Nadam" optimizer, "relu" activation, 50 mini batch size.

We also use several techniques, dropout, cross-validation, hyper-parameter search. Dropout is a technique to train by invalidating hidden layer nodes with a certain probability. Cross-validation is an effective method when the number of data with correct answer labels as input data of the model is small. Cross-Validation and Dropout are effective to avoid over-fitting. Hyper-parameter is a parameter of the neural network itself. Hyper-parameter search is to find optimal combinations from various options such as activation function, optimization algorithm, number of units of hidden layer and so on. There are two methods of search methods, Grid-Search which comprehensively tests all combinations and Randomized-search which randomly selects a combination of Hyper-Parameter and searches. In our experiment, we adopted a policy to gradually narrow down the optimum combination direction by Randomized-search.

Using proposed DNN, inference of classification traffic achieve more than 99% F1-score. F1-score is the harmonic mean of precision and recall. This 99% F1-score is better than the other ML technique, C4.5 decision tree algorithm. Our proposed DNN method can distinguish with high accuracy in any of application, the encrypted / non-encrypted traffic and the Voip (Voice over IP) streaming traffic. We identified which feature is effective for classification, total number of packets of flow, total byte of flow, duration.

Finally, we confirmed that the generalization performance of the parameters is statistically significant. We generated 50 models by 50 times of training and confirmed that the variance of the discrimination result of each model is small and stable with high accuracy. At that time, for each Training, the entire data set was randomly divided into Validation data, Training data, and Test data.

In future work, real time classification is necessary. In order to classify in real time, it can be realized by calculating statistical feature at any time in the flow and classifying it using DNN method.