TITLE: MULTITASK LEARNING FOR NETWORK TRAFFIC CLASSIFICATION

SUMMARY:

Traffic classification has various applications in today's Internet, from resource allocation, billing and QoS purposes in ISPs to firewall and malware detection in clients. Labeling data is often the most difficult and time-consuming process in building a classifier. To solve this challenge, we reformulate the traffic classification into a multi-task learning framework where **bandwidth requirement** and **duration of a flow** are predicted along with the traffic class.

Dataset: We conduct experiments with **QUIC public datasets** and show the efficacy of our approach.

Introduction:

The earliest approaches to solve network traffic classification used port numbers or unencrypted packet payloads. These methods relied on human labor for continuously finding patterns in unencrypted payloads or matching port numbers. To mitigate the need for a large amount of labelled training samples, we propose a multi-task learning approach which performs three predictions (tasks), for which only one requires human effort and controlled environment for labelling.

Background:

The most common approach to multi-task learning is hard parameter sharing where some parameters of the deep learning models are shared among tasks and some parameters are kept task-specific. The multi-task learning model is more effective than several single-task learning models because in MTL datasets of all tasks can help all other tasks to be learned better. In this paper, it is shown that we can improve the accuracy of the traffic classification task by using a multi-task learning approach where the data are easy to obtain for other tasks, namely predicting bandwidth and duration of a flow.

Methodology:

The multi-task learning model is more effective than several single-task learning models because in MTL datasets of all tasks can help all other tasks to be learned better. In this paper, we will show that we can improve the accuracy of the traffic classification task by using a multi-task learning approach where the data are easy to obtain for other tasks, namely predicting bandwidth and duration of a flow. The multi-task learning model is more effective than several single-task learning models because in MTL datasets of all tasks can help all other tasks to be learned better. In this paper, we will show that we can improve the accuracy of the traffic classification task by using a multi-task learning approach where the data are easy to obtain for other tasks, namely predicting bandwidth and duration of a flow.

In this paper, we use three time-series features, that is, packet length, inter-arrival time, and direction, of the first k packets. The input of our model is a vector of length k with 2 channels. The first channel contains the inter-arrival time of the first k packets and the second channel contains the length and direction combined. To find the optimal value for [d1, ..., d4], the duration divider, we first find the average duration of each class. Then, we sort the average values and then find the middle point between two consecutive average values as [d1, ..., d4].

$$\underset{\mathbf{W}^B,\mathbf{W}^D,\mathbf{W}^T}{\arg\min} \sum_{i=1}^{N} \Big[\ell(y_i^B, f(\mathbf{x}_i;\mathbf{W}^B))+ \ell(y_i^D, f(\mathbf{x}_i;\mathbf{W}^D)) + \lambda\ell(y_i^T, f(\mathbf{x}_i;\mathbf{W}^T))\Big]$$

Multi-task Model Architecture:

In this paper, 1-dimensional convolutional neural network (CNN) is used in multi-task learning model architecture. There have been usage of max pooling as it is commonly preferred over other pooling methods. Rectified linear unit (ReLU) activation is also used as an activation function in the entire model, except the last layers which contain SoftMax. Evaluation: We use batch optimization and Adam optimizer for training. The loss function and model architecture are explained. For single-task learning approach, we use two successful models, namely random forest (RF) with statistical features and CNN+RNN model proposed. We train the model three times from scratch for each task. For bandwidth and duration prediction, we use the entire dataset for training since it does not require human effort for labeling. That is why the accuracy remains the same when labeled samples are increased. Since the bandwidth and duration is statistical features, they can be obtained if the entire flow is accessible, we first train the model with the entire dataset to predict the bandwidth/duration tuple. We note that for this task, we use both labels of the entire training set. After training the model, we remove the last layer and replace it with a new layer and initialized the weight, similar to Then, we re-train the model for the traffic class prediction task. The final model only predicts the traffic class. We train the entire model with all training data. We use the bandwidth and duration labels of the entire training data samples, while we only provide a limited number of labels for the traffic class task (specified in the first column). For this experiment, we set λ to one to emphasize on all three tasks equally. Moreover, we use the first 60 packets as input (k = 60)

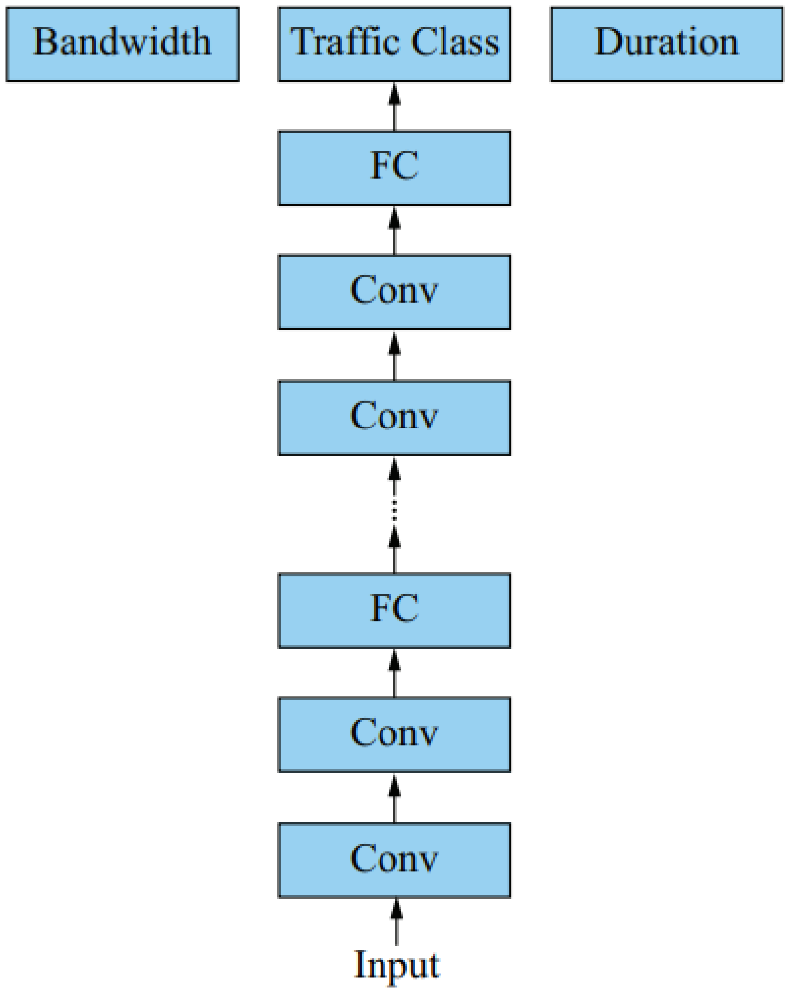

Fig. 4. Multi-task learning model architecture

Accuracy:

The accuracy of the traffic class prediction, for which we have limited labeled samples, is considerably higher with our multi-task learning approach than the transfer learning and single-task learning. In fact, the large amount of data that is available for bandwidth and duration tasks significantly improves the training process by allowing the model parameters to be trained with such abundant data. In our experiments, we find that the accuracy of a single task learning is 97.67% when the entire dataset with all class labels are used. For traffic class prediction task, the multitask learning approach with only 100 labeled samples reach almost the same accuracy as single-task learning using the entire labeled dataset. Hence, the multi-task learning approach can greatly reduce the number of labeled data.

TABLE II
STRUCTURE OF THE CNN MODEL

-	Conv	Conv	Pool	Conv	Conv	Pool	Conv	Conv	Pool	FC	FC
Number of filters/neurons	32	32	-	64	64	-	128	128	-	256	256
Kernel size	3	3	2	3	3	2	3	3	2	-	-

TABLE III
ACCURACY ON QUIC DATASET

# of labeled samples	Accuracy [Bandwidth, Duration, Traffic Class]			
(For traffic class)	RF [26]	CNN+RNN-2 [19]	Transfer learning	Multi-task learning
10	[-, -, 48.67%]	[89.33%, 92.00%, 64.67%]	[-, -, 85.33%]	[89.33%, 91.00%, 93.33%]
20	[-, -, 64.00%]	[89.33%, 92.00%, 66.67%]	[-, -, 87.33%]	[90.33%, 91.33%, 94.67%]
50	[-, -, 78.00%]	[89.33%, 92.00%, 76.67%]	[-, -, 90.67%]	[90.67%, 91.33%, 96.00%]
100	[-, -, 86.67%]	[89.33%, 92.00%, 85.33%]	[-, -, 92.67%]	[90.67%, 92.00%, 97.33%]

Conclusion:

We propose a multi-task learning approach that predicts traffic class labels as well as bandwidth and duration of network traffic flows. Because the bandwidth and duration tasks do not require human effort or controlled and isolated environment for labeling, a large amount of data can be easily captured and used for training these two tasks. We show that by providing a large enough dataset for bandwidth and duration tasks, one can train the traffic class prediction task with only a small number of samples. Hence, it obviates the need for a large amount of labeled data samples for traffic classification. Moreover, bandwidth and duration predictions can be used for resource allocation, routing and QoS purposes in ISPs. We illustrate that our multitask learning approach significantly outperforms both singletask and transfer learning approaches.

code link: https://www.kaggle.com/dhyanidesai/machine-learning-da2-3

Paper link: https://arxiv.org/abs/1906.05248

Dataset link:
https://drive.google.com/drive/folders/1Pvev0hJ82usPh6dWDlz7Lv8L6h3JpWhE