# Encrypted Web Traffic Classification using Multimodal Multitask Transformer

Vivek Chauhan, Noura Limam

{v5chauha, noura.limam}@uwaterloo.ca

University of Waterloo, Waterloo, Ontario, Canada

## Abstract

Traffic classification in network operations is critical for targeted actions where quality of service is utmost. ISPs and network providers are constantly monitoring the network traffic for discovering any threat actors and change in patterns. Many major ISPs need to perform three important tasks in real-time, Traffic categorization into broader service like, chat, Streaming, VoIP, etc. is the first tasks, Secondly, it needs to be able to identify underlying applications like, WhatsApp, YouTube, Netflix, Facebook etc., to benchmark their service at different levels of granularity in real time to track performance and provide guaranteed Quality of Service to customers and data generators in any region. Thirdly it needs to identify the malicious activities. Encrypted web traffic is now the standard, and the efficiency and accuracy trade-off considerations impose increasing challenges to classify traffic in real-time. The task of performing these classification steps sequentially by building different models is a redundant way to do so. We propose a multimodal multitask transformer to classify encrypted traffic in broader service class and granular application class simultaneously without the need for two models. We utilize text and vision transformer encoders in our proposed model for flow statistics and mid flow time series respectively for traffic classification. We propose the use of multimodal multitask transformers for jointly learning important features by self-attention mechanisms. The transformers perform better in classification task and reduces the time complexity.

## Introduction

Traffic classification is a task of categorizing network flows and packets into different classes. These classes can be application level or service level. The traffic in each class is measured on different parameters and these parameters serve as significant metrics for allocating appropriate resources and capacity. Resource utilization, volumetry, Service level Key Performance Indicators (KPI's) and Quality of Service (QoS) are important metrics for consumer as well as data providers. Traffic classification is also critical for detecting threat actors and intrusion attempts in an enterprise or at an application layer. Traffic encryption is a way to mask the information from breach and enable a secure communication transfer mechanism from one host to another. In web traffic, the end points are requesting client(s) and serving web server. This encryption is usually achieved through HTTPS (Hyper Text Transfer Protocol Secure). Other protocols and custom implementation for encryption are also employed for network communication. Newer protocols such as QUIC (transport layer network protocol) and HTTP/2 are also being increasingly

used. The discussions for HTTP/3 which encodes differently, and stores session state differently compared to previous versions are in pipeline. With encrypted web traffic it becomes more challenging to detect intrusions and increases the risk of adversarial attacks, too. The previous works in this field prove improved accuracy and decreased false rate of classification. However, their assumptions of stable encryption, single class classification and information retrieval from start of packet limit its applicability to evolving web. Additionally, the need for data collection adds significant burden for real-time applications.

There has been a lot of interest among researchers to actively explore encrypted web traffic classification which has been bolstered by an increase in accuracy due to Neural Networks, specifically due to varied works in Convolution Neural Network (CNN), Long Short-Term Memory (LSTM) and Recurrent Neural Network (RNN). In recent approaches to network traffic classification using deep learning, the concepts of image classification from other domains combined with the traffic flow statistics has been an underlying pattern. These works albeit impressive are reliant on canary features, logical mapping, traffic statistics and initial portion of packet flow information which is used in parallel or sequence for categorizing traffic in different classes. This is an obstacle due to various reasons. Few of them include the hefty cost of the process, statistical features that can be affected by OS, user-specific behaviour and the proneness to change according to newer specifications. While these approaches can accurately classify traffic, there are limitations and constraints to their applicability in real time scenarios. We try to tackle this problem by utilizing the traffic shape as input to multitask Vision transformers and flow statistics, as text through linguistic encoders. In this paper we identify the limitations of current approaches and propose a system for traffic classification via transformers.

In our proposed model, we consider data in a packet, as a sequence of data bytes and that the sequence of these bytes is important for the intended classification task. We thus make use of self-attention network that employ the positional embedding with input sequence to encode them through a multi-head attention network and a stacked layer of decoders. Our proposed model has three major classification tasks which are done simultaneously using Transformers. Transformers can be used for Natural Language Processing (NLP) tasks as well as in Computer Vision (Image Classification). We plan to do a comparative analysis between them using MTT, ViLBERT, UNiT, H5 and other popular transformer architectures.

In subsequent sections we discuss related works and introduce the methodology leading up finally, to delineate important milestones that we aim to achieve.

## Related work

Traditionally, traffic classification was achieved using port-based mapping to identify the protocols and classify broader categories of network traffic. Another approach where packet flow features were extracted explicitly using domain knowledge of experts is also now outdated. In this section we shed light on more recent works in traffic classification, including the classical Machine Learning (ML) based methods and Deep Learning architecture.

Classical ML models utilize the header and packet flow statistics; for example, inter-arrival time, packet arrival time etc. using algorithms such as SVM, Decision tree, random forest to classify network traffic but these are shallow learning methods, meaning they do not learn any features implicitly but depend on manually selecting and providing features to train on. These methods are now replaced by Deep Learning based approaches and various neural network architectures have been applied successfully for classification task. Recent work on stacked LSTM, with CNN can achieve higher accuracies [1]. In deep learning, there are three popular approaches based on the treatment of network traffic/ input. First is converting traffic shape into Flowpics [2] (image) histogram of pixels and feeding to CNN architecture. Second is traffic flow statistics through another CNN, LSTM, RNN, Stacked AE architecture to learn features specific to traffic class. Third is using raw traffic bytes through CNN network. Li [3] et al. used RNN and attention mechanism to fully extract the temporal features of the network stream. The use of different methodologies described in literature for using a single deep learning method (such as RNN, CNN) compared with multiple combination methods (such as RNN + LSTM, CNN + LSTM), reveal that the combination architecture yielded better accuracies. This is suggested as it can automatically determine the temporal and spatial features of network traffic making use of filters to decrease the misclassification rate [4]. RNN is also used to provide feedback path of learned features into training but are very costly. CNNs have inherent inductive biases like translational equivariance and locality and hence modifications in traffic bytes are employed in pre-processing steps.[5] Such translation is minimal in transformers.

There are two major categories of transformers: multitask transfer and multimodal transformers. Rezaei et al. [6] discuss a framework that is able to classify traffic types while simultaneously predicting the bandwidth needed and flow duration. In accordance with their experiments, fewer labeled samples could yield high classification accuracies. Rago et al. [7] proposed a multi-task learning approach at the mobile endpoint. The author proposed classification and prediction tasks of mobile traffic simultaneously. We can employ similar guiding principles to classify traffic at two different levels and flag malicious network traffic. This is a more efficient way and uses a lightweight, multi-head attention-based model proposed by Cheng et al. [8]. It uses less parameters as compared to CNN models but the three consecutive packets that flow as an input, add space and time complexity. A much improved architecture using self -attention based models was proposed by Xie et al. [9] which is suitable for real time classification. It makes use of the initial 50-bytes and this reduced payload yields shorter classification time. Akbari et al. [10] also proposed the use of initial 50 bytes as input and achieved high accuracy; transformer architecture making use of shorter payload is hence proposed.

Wang [11] et al. proposed a malware traffic classification method based on representation learning, but this paper lacked the parameter optimization and generalization ability of CNN model. We plan to include steps for representation learning such as PCA, LDA, UMAP, t-SNE, Stacked AE as an added step before feeding data to input to extract interesting yet unrelated features to compare against generalizability of model.

The concept of vectorizing a token into a sequence of image patches has been introduced and used to learn the positional embedding which permits vision transformers to integrate information across the entire image [12] (global attention) at every layer unlike CNNs. The inspiration is derived from self-attention mechanism introduced for sequential language tasks [13].

Zheng et al. [14] proposed the use of multitask transformers to utilize multiple layers of encoders for joint task of classification in three categories. The packet flow characteristics and raw bytes have been fed as input embeddings. The spatial and temporal features implementation for attention for traffic shape through convnet has been discussed separately by Zhang [15]. The joint task of visual and language in form of visiolinguistic treatment has been successfully implemented for multimodal tasks in other domains [16].

From the study of related literature work, we see that there has been an active exploration in encrypted traffic classification. Network traffic is evolving, and it introduces a challenging problem of classifying encrypted traffic in real time with the least number of parameters while ensuring high accuracy. Although encryption addresses security and privacy concerns, it adds the challenge of detecting malicious traffic. We therefore need to continue efforts on active learning and learning through interaction which are costly but effective processes for classification.

## Methodology

### 4.1 Dataset

We propose to use real-life data from an ISP/data collected in lab to train and evaluate our model. In absence of real-life data, we plan to use public dataset “ISCX VPN non-VPN” to train, validate and evaluate our model in the ratio of 70,10,20 respectively. We propose to include a comparative study of the various methods employed, to handle the imbalanced dataset. This study would be carried out using Data augmentation and down sampling, . Down sampling or upscaling are mostly utilized in literature but using augmented data with adversarial datapoints would serve as a nice way to train our model with threat actors. The PCAP files captured using tcpdump or Wireshark would be used with additional data pipeline to clean and mask data features. Packet capture (PCAP) is an application programming interface that is used to capture real time network packet data. Enterprise or ISPs could make use of other tools to capture monitor and network traffic points.

### 4.2 Data Pre-Processing

Our input is a sequence vector where positional embeddings included in the model plays a major part. We drop the features that give away service labels, in the headers as we plan to build a generalizable and stable model, learning feature based on self-attention mechanisms. We remove irrelevant headers and mask IP address in IP by zero bytes. We make TCP and UDP headers of same length by padding zeros so that tcp and udp traffic is not given away by the headers. Dropping headers for flag fields and making uniform input for transformers equal to

MTU to be fed into embeddings. We label and sample to make equal number of service and application class along with an additional label for GAN or raw data to train for identifying malicious traffic.

### 4.3 Architecture

#### Transformer model

Transformers facilitate learning long range dependencies by making use of positional embeddings. It adds a scaling parameter to solve for the problem of vanishing gradient or extremely large values. It makes use of fewer steps and facilitates parallel computation. There We make use of Text based multitask transformer as employed in (Natural Language Processing) NLP models to establish relationship between the text using their sequence and positional embedding as used in language translation task. Second approach is making use of Vision Transformers where we use images of traffic shape along a timeseries converted as histogram as used in 1D and 2D CNN models. Third approach is using a hybrid where we utilize approaches 1 and 2 in parallel and do a comparative study of different transformers for our multimodal multitask model such as ViLBERT [16], h5, UNiT[17], MTT models. ViLBERT processes text and image independently using layers of encoder stack which is followed by a co-attention network which can learn the underlying relationships between text and image patches and as such acts as input for many multi-modal tasks. It is based on transfer learning method which can be fine-tuned to learn shallow features. We can compare the accuracy and efficiency of our model against state-of-the-art CNN models. It is expected that transformers will perform a better task at classification and reduce the time greatly, thus, could be utilized in real time for classification. The structure of our proposed model inspired by [12], consists of embeddings module, followed by a stacked layer of multiple encoders with multi attention layers and finally a classifier module to classify traffic in three different tasks.

#### Embedding Module

There are two components to it, one is input embedding which takes the tokenized input and second is positional embedding which is like an indexing vector storing the positions of sequence. This positional dependence is important in sequence processing. This input embedding acts as the input to the encoder module.

#### Encoder-Decoder Module

We use two parallel encoders one for text input and other for image input in our architecture. The encoder and decoder blocks are multiple identical encoders and decoders stacked on top of each other. The number of layers in the encoder and decoder is a hyperparameter but is proved to work well with 6 layers in [12]. The encoder layer has 3 sub layers and the output of one encoder layer is fed to the encoder layer on top of it. Encoder output from two parallel encoders is concatenated before it is fed to the decoder. The encoder has a multi-head self-attention network followed by an *add and layer* normalization step which is performed before feeding the

output of self-attention network to positional feed forward network. Feed forward network has an activation layer and non-linearity is introduced in the form of "ReLU" function. Residual connections like RNN are also fed in the encoder and decoder layers. Cross-attention, and self-attention heads are used to represent relationship between sequences. The relationship between positions of single sequence can be learned as a representation of the dual task of linguistic and vision. An interconnection between image and text to explain the image can be experimented which is expected to increase the context learning.

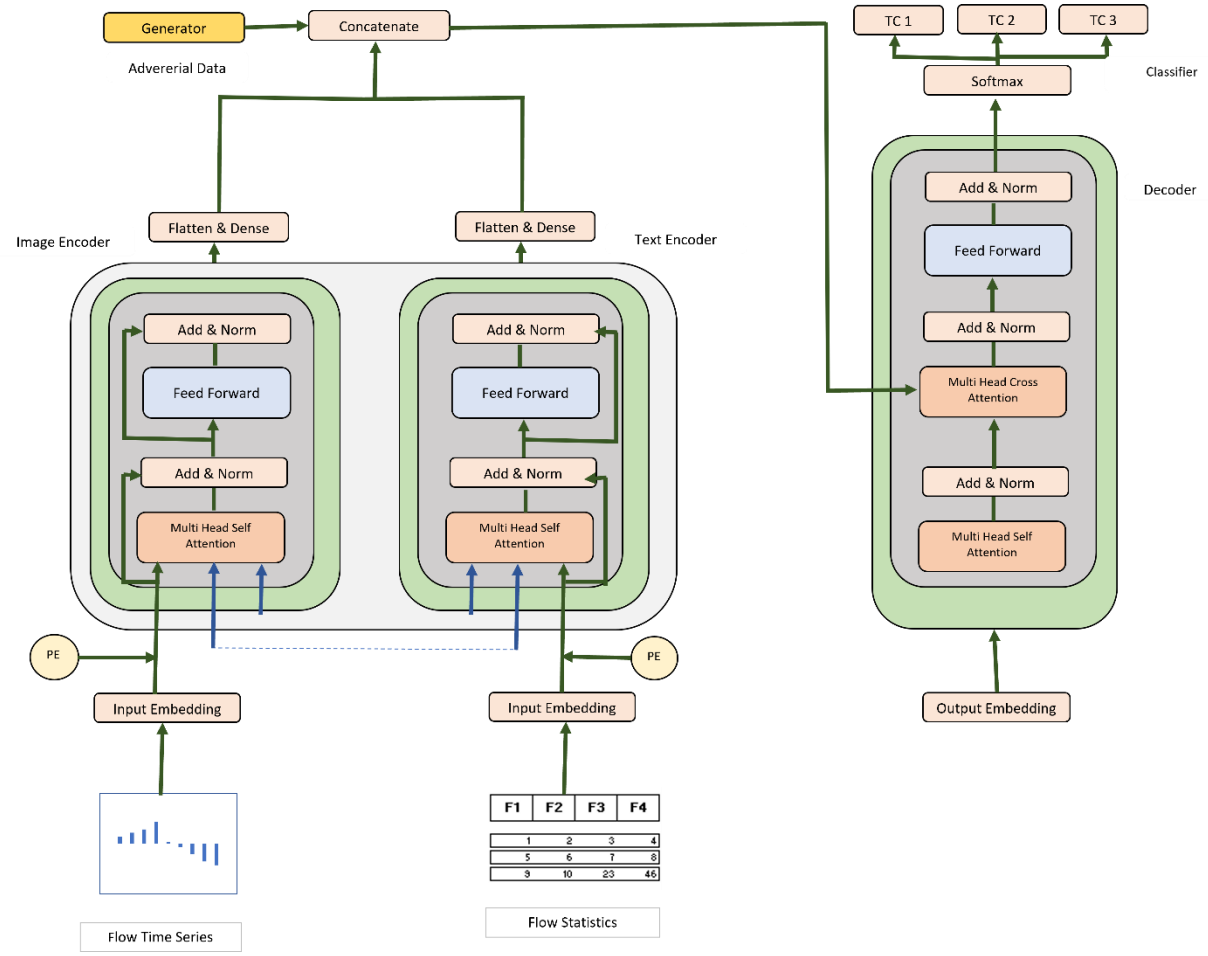


Figure 2: An overview of our proposed model, which uses an image encoder to encode the visual inputs and a text encoder to encode the language inputs and a joint decoder with cross-attention and self-attention heads and a classifier module.

Attention function maps query and key-value pairs to output based on similarity between key and query. The output generated at this step is then summed over all weighted values where weights are derived from the compatibility function. The attention is calculated as the projection of Query on key to derive maximum similarity. It is then scaled by a factor of $\sqrt{d_k}$ which solves the problem of exploding and vanishing gradients. The output for attention function is calculated as:

$$\text{Attention (Q, K, V)} = \text{Softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

*Text Encoder:* The input vectors are divided into query, key and value after expanding by a fully connected layer. Query, key, and value are further divided and fed to the parallel multi-attention heads. Outputs from attention heads are then concatenated to form the vectors whose shape is the same as the encoder input. The vectors go through a fully connected, a layer norm and feed forward (MLP) block that has fully connected layers.

*Image Encoder:* The input image is split into h x h vectors and reshaped to feed into Conv2d.The position embedding vectors are added to the sequence patch embedding vectors to be fed into the transformer encoder. These embedding vectors are encoded and the output from the encoder is fed to the feed forward (MLP) head.

The outputs are flattened and concatenated before feeding to into next encoder layer or decoder, this helps in representing different subspaces at different positions in a joint way.

**Classifier module**

It takes the output from the decoder, flattens it, and feeds it to fully connected layer. It has three output layers and gives a vector list of probabilities for each class. The three tasks in our case are service level classification, application-level classification and classifying traffic as malicious or not (as a binary output).

Post-hoc analysis could reveal important features that are being learned by the machine to classify traffic to better understand patterns for classification. This is done by plotting attention maps which are scalar 2D matrix and plotted as heatmap. The greater the values, brighter the heatmap.

The expected outcome is an increase in multitask classification accuracy by ~1-2% and reduction in the misclassification. An additional benefit is the ability of system to train against adversarial attacks. Our proposed model is expected to be generalizable for newer protocols and to provide stable, reliable, and real time classifications. We expect a significant decrease (>10%) in time utilized for classification due to the introduction of parallel attention heads. Mid-flow packet flows are expected to yield similar results and would be an important area for our study. Our expected contribution can be best explained in three folds; increasing the accuracy of traffic classification, reducing space and time complexity, and performing well against adversarial attacks. There is extensive research being carried out in transformers for vision and NLP tasks which aids our research proposal.

## References

[1] Jianwu Zhang, Yu Ling, Xingbing Fu, Xiongkun Yang, Gang Xiong, and Rui Zhang. Model of the intrusion detection system based on the integration of spatial- temporal features. Computers and Security, 2020.

[2] Shapira T, Shavitt Y (2019) Flowpic: Encrypted internet traffic classification is as easy as image recognition. In: IEEE INFOCOM 2019-IEEE conference on computer communications workshops (INFOCOM WKSHPS). IEEE, pp 680–687

[3] Rui Li, Xi Xiao, Shiguang Ni, Haitao Zheng, and Shutao Xia. Byte segment neural network for network traffic classification. In 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), pages 1–10, 2018.

[4] M Lopez-Martin, B Carro, A Sanchez-Esguevillas, and J Lloret. Network traffic classifier with convolutional and recurrent neural networks for internet of things. IEEE Access, 5:18042–18050, 2017.

[5] Zou Z, Ge J, Zheng H, Wu Y, Han C, Yao Z (2018) Encrypted traffic classifcation with a convolutional long short-term memory

neural network. In: 2018 IEEE 20th international conference on high performance computing and communications; IEEE 16th International conference on smart city; IEEE 4th International Conference on Data Science and Systems (HPCC/SmartCity/DSS). IEEE, pp 329–334

[6] Rezaei S, Liu X (2020) Multitask learning for network traffic classification. In: 2020 29th International conference on computer communications and networks (ICCCN). IEEE, pp 1–9

[7] Rago A, Piro G, Boggia G, Dini P (2020) Multi-task learning at the mobile edge: An effective way to combine traffic classification and prediction. IEEE Transactions on Vehicular Technology 69(9):10362–10374

[8] Cheng J, He R, Yuepeng E, Wu Y, You J, Li T (2020) Real-time encrypted traffic classification via lightweight neural networks. In: GLOBECOM 2020-2020 IEEE global communications conference. IEEE, pp 1–6

[9] Xie G, Li Q, Jiang Y (2021) Self-attentive deep learning method for online traffic classification and its interpretability. Computer Networks 196:108267

[10] Akbari, Iman & Salahuddin, Mohammad & Ven, Leni & Limam, Noura & Boutaba, R. & Mathieu, Bertrand & Moteau, Stéphanie & Tuffin, Stephane. (2021). A Look Behind the Curtain: Traffic Classification in an Increasingly Encrypted Web. Proceedings of the ACM on Measurement and Analysis of Computing Systems. 5. 1-26. 10.1145/3447382.

[11] Wei Wang, Ming Zhu, Xuewen Zeng, Xiaozhou Ye, and Yiqiang Sheng. Malware traffic classification using convolutional neural network for representation learning. In 2017 International Conference on Information Networking (ICOIN), pages 712–717, 2017.

[12] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

[13] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30

[14] Zheng, W., Zhong, J., Zhang, Q., & Zhao, G. (2022). MTT: an efficient model for encrypted network traffic classification using multi-task transformer. Applied Intelligence, 1-16.

[15] Hu, F., Zhang, S., Lin, X., Wu, L., Liao, N., & Song, Y. (2021). Network traffic classification model based on attention mechanism and spatiotemporal features.

[16] Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32.

[17] Hu, R., & Singh, A. (2021). Unit: Multimodal multitask learning with a unified transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1439-1449).