Can XAI Tools Interpret Traffic Classifiers based on Deep Learning?

A. Nascita, A. Montieri, G. Aceto, D. Ciuonzo, V. Persico, A. Pescapé
Università degli Studi di Napoli "Federico II",
{alfredo.nascita, antonio.montieri, giuseppe.aceto, domenico.ciuonzo, valerio.persico, pescape}@unina.it

Abstract

The massive use of mobile devices has deeply changed network traffic, creating a challenging scenario for traffic classification and its uses, especially network security-related ones. Deep Learning (DL) has responded to this challenge, but the lack of interpretability of such a black-box approach limits its adoption where the interpretability and reliability of outcomes are necessary. For these reasons, eXplainable Artificial Intelligence (XAI) has seen recent intensive research. Following this line, our work applies XAI techniques to give a global interpretation of the behavior of a multimodal DL traffic classifier. Based on a public dataset of mobile-app traffic, we quantify the importance of each modality and of specific subsets of inputs on the classification results.

1 Introduction

The knowledge of the traffic traversing a network is instrumental to several management activities: Traffic Classification (TC) has a key role in defining a "normal" traffic profile for anomaly detection and in extracting fingerprints for intrusion detection and attack identification. TC has seen consistent research and is now seeing a renewed blossoming of interest due to the recent evolution of network usage. Also, powerful Deep Learning (DL) techniques have become available to face new TC challenges. DL approaches are characterized by a fully-automated feature extraction (with reduced need of human experts in the loop) and a greater ability of learning from huge volumes of data, providing better performance than traditional Machine Learning (ML) approaches. The highly desirable characteristics of DL come at the cost of lack of *interpretability* of its results, as the black-box nature of DL techniques hides the reason behind specific classification outcomes. This impacts the understanding of errors because DL approaches leave open questions like *"which parts of a complex architecture mostly contribute to the final decision?"*, *"which specific fields or packets are the most important in the classification process?"*, or *"which ones are responsible for classification errors or circumvention?"*.

The field of *eXplainable Artificial Intelligence* (XAI) constitutes the answer to these needs, as it provides approaches and techniques able to relate the structure and input of the model to the classification outcome, partially revealing the (former) completely black box [Hagras, 2018]. To this aim, we perform the behavior interpretation of a *state-of-art multimodal DL architecture for TC* we recently proposed [Aceto *et al.*, 2019b] for the challenging task of classifying mobile apps. We apply the *XAI tool Deep SHAP* [Lundberg and Lee, 2017] to quantify and understand the relative importance of two input sets related to (i) the first bytes of the payload or (ii) informative fields extracted from the sequence of the first packets of each biflow. The experiments are conducted using a public dataset of mobile-app traffic [Aceto *et al.*, 2019a].

2 Background

The large success of AI in different application domains has paved the way to its use for activities related to decision making, cost reduction, risk management, etc. However, models resulting from complex AI algorithms have a black-box nature which leads to an undesired *lack of interpretability*. As a result, in the last years many works have focused on assessing the transparency, causality, bias, fairness, and safety of the obtained solutions [Hagras, 2018]. Investigating the working behavior of already-trained models allows to (i) assess the robustness of AI systems; (ii) evaluate their resilience to drifting data perturbations; (iii) verify legal, safety, and security requirements; (iv) complement human expertise in decision making and even provide researchers with novel insights.

Several data-driven solutions based on ML/DL have been recently proposed for various network-related problems. However, the main reason limiting these solutions to be widely adopted in production environments is their general lack of interpretability leading to potential behavioral uncertainties. Therefore, an initial corpus of works has provided a first effort towards the interpretation of data-driven black-box models proposed for networking problems, such as: visualization tools for TC [Beliard *et al.*, 2020], XAI tools providing post-hoc local explanations for TC [Wang *et al.*, 2020] or anomaly detection [Amarasinghe *et al.*, 2018], Markovian distillation analysis for traffic prediction [Aceto *et al.*, 2021].

Herein, we focus on *mobile-app TC* and use *XAI techniques* to obtain insights on the behavior of a *multimodal DL architecture*. However, unlike previous works, we exploit Deep SHAP to extract *global* explanations to infer the *importance* of inputs on the general behavior of the classifier.

3 Methodology

3.1 Mobile Traffic Classification via MIMETIC

The mobile TC task consists in assigning to each biflow[1] a class within the set $\{1, \ldots, L\}$, with L denoting the number of apps. MIMETIC is a multimodal DL traffic classifier that exploits multiple modalities (viz. views) of traffic data, to capitalize their heterogeneous nature [Aceto *et al.*, 2019b]. This classifier consists of two *single-modality* (viz. input-specific) branches that extract the corresponding intra-modality features. The input of the first branch (PAY-modality) is constituted by the first N_b bytes of transport-layer payload arranged in byte-wise format, whereas for the second branch (PSQ-modality) we consider some informative fields extracted from the sequence of the first N_p packets (see. Sec. 4.2). In detail, we consider the first $N_b = 576$ bytes and $N_p = 12$ packets, respectively. To capture inter-modality dependencies, the abstract features extracted by the single-modality branches are joined via a concatenation layer and fed to a *shared* dense layer before performing the classification through a softmax. MIMETIC is trained via a two-phase procedure: (i) an independent *pre-training* of each single-modality branch and (ii) a subsequent *fine-tuning* of the whole architecture.

3.2 Interpreting DL-based Traffic Classifiers

To interpret complex DL architectures, the starting point is to consider a simpler explanation model $g(\cdot)$ designed to closely-approximate the original model $f(\cdot)$. In this paper, we focus on *local methods*, which explain the original model in the neighborhood of each instance $\boldsymbol{x}$, using *simplified inputs* $\boldsymbol{x}'$ that map to the original ones via a mapping function.

Most of the interpretability techniques assume a peculiar functional form for the explanation model $g(\cdot)$, which led to the definition of Additive Feature Attribution (AFA) methods which are linear functions of binary variables, i.e.

$$g(\boldsymbol{z}^{'}) = \phi_0 + \sum_{m=1}^{M} \phi_m \, z_m^{'} \tag{1}$$

where $\boldsymbol{z}^{'} \in \{0,1\}^M$, M is the number of simplified inputs, and $\phi_m \in \mathbb{R}$. AFA methods provide an explanation model attributing an "effect" ϕ_m to each input, and summing the effects of all input attributions approximate the original model output $f(\boldsymbol{x})$. Additionally, when the functional form in Eq. (1) is required to satisfy (i) local accuracy, (ii) missingness, and (iii) consistency properties, there exists a unique AFA solution satisfying them [Lundberg and Lee, 2017]. Such solution coincides with the computation of the well-known *Shapley values* [Shapley, 1953] which originate from cooperative game theory and identify the contribution of each player to the payoff achieved by the overall coalition.

When transposing the method to the task of explaining a DL-based model, the input data maps into the players, whereas the DL architecture output $f(\boldsymbol{x})$ corresponds to the payoff function. Unluckily, the time required for the exact computation of the Shapley values *grows exponentially* with the input size M. Conversely, SHapley Additive exPlanation (SHAP) *approximates* these quantities in a computationally-efficient fashion, eliminating the need to re-train the models. The computation is also simplified by assuming statistical independence among the inputs and linearity of the model [Lundberg and Lee, 2017].

In the following, we use Deep Shap, a fast approximation algorithm for SHAP values which capitalizes the connectionist nature of DL architectures. This technique will be used to assess the importance of inputs, selected from raw traffic data (related to PAY or PSQ modalities of MIMETIC) of a given biflow, in classifying the app generating it. Therefore, to explain the predictive behavior of DL-based traffic classifiers, the prediction model $f(\boldsymbol{x})$ is chosen as the soft-output associated to the generic i^{th} app, i.e. $p_i(\boldsymbol{x})$. Hence, we interpret the SHAP value ϕ_m as the *importance value* of the m^{th} input in forming the confidence associated to labeling the biflow (whose overall input is $\boldsymbol{x}$) with the i^{th} app. We recall that, since ϕ_m can be also negative, positive (negative) values increase (decrease) the confidence $p_i(\boldsymbol{x})$ in the i^{th} app with respect to its average value $\mathbb{E}\{p_i\}$.[2]

Herein, for each biflow, we focus on explaining the soft-output associated with the predicted app $\hat{p}(\boldsymbol{x})$, as this represents the most relevant (and highest) output for TC. Once we have obtained a local explanation for a single instance, our proposed *global explanation* approach relies on aggregating (viz. pooling) explanations over different samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$. The aggregation step is however carried out on *normalized* SHAP values, obtained by dividing each SHAP value by their overall sum, namely $\widetilde{\phi}_m \triangleq \phi_m \,/\, \sum_{m=1}^{M} \phi_m$. Considering $\widetilde{\phi}_m$ allows focusing on the relative *importance* of each input (for each sample, the sum of importance values equals one). Also, we aggregate only on *correctly-classified samples* to interpret the correct behavior of MIMETIC.

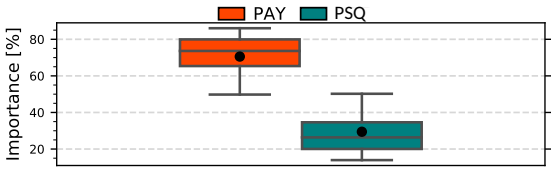


Figure 1: Modality contributions of MIMETIC in terms of importance $\widetilde{\phi}_m$.

4 Experimental Evaluation

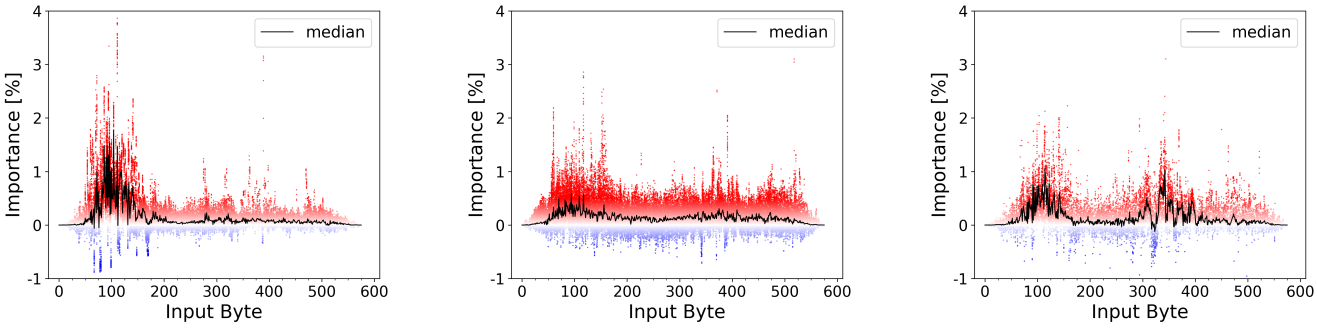
4.1 Dataset and Experimental Setup

This study leverages MIRAGE[3], an open dataset containing mobile-app traffic, collected at the ARCLAB laboratory of

[1] A bidirectional flow or *biflow* is a stream of packets sharing the quintuple (i.e. transport-level protocol, source and destination IP addresses and ports) regardless the direction of communication.

[2] The sum of the SHAP values equals the considered soft-output value minus the *base output*. The latter represents the average of the same soft-output obtained in correspondence of the samples associated with a background set.

[3] http://traffic.comics.unina.it/mirage



(a) Dropbox. (b) Spotify. (c) Trello.

Figure 2: Importance $\widetilde{\phi}_m$ of the payload-byte inputs (identified by their position) for PAY-modality of MIMETIC for exemplifying apps.

the University of Napoli "Federico II", exploiting the MIRAGE architecture [Aceto *et al.*, 2019a]. The dataset consists of the traffic generated by 41 Android apps, for a total of $\geq$ 4.6k traffic traces, each generated in a capture session of $5 \div 10$ mins, and $\approx$ 96.5k biflows.

The following results refer to a random (stratified) dataset split with 90% (resp. 10%) of instances (i.e. biflows) allocated for training (resp. testing). We highlight that, in the considered case, the TC performance attained by MIMETIC were 88.74%, 87.83%, and 93.43% in terms of accuracy, F-measure, and G-mean, respectively.

4.2 Interpretability Analysis for TC Decisions

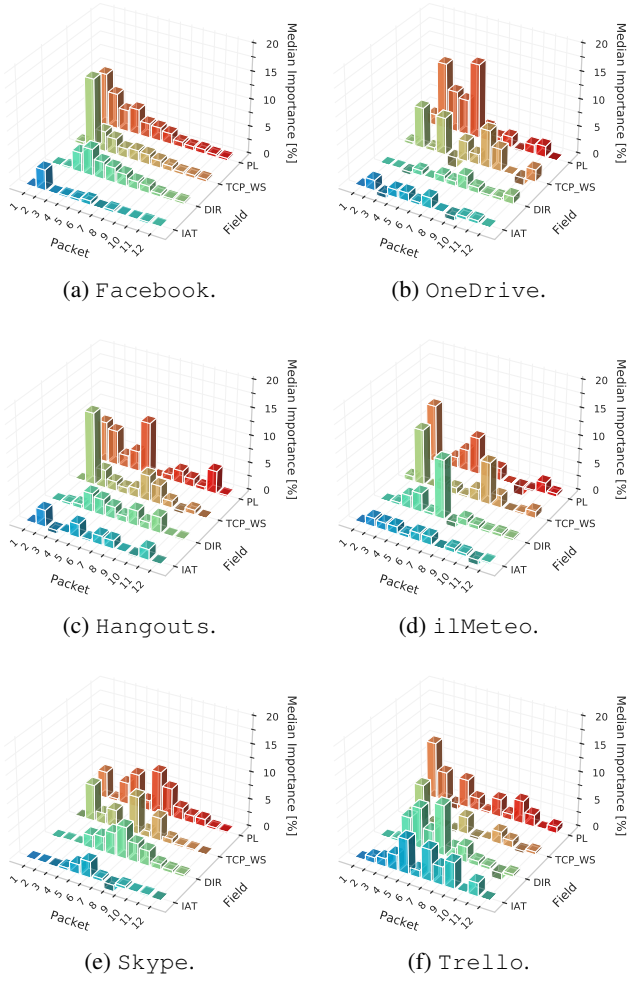
The first aspect we investigate is the weight of each traffic modality (PSQ or PAY) of MIMETIC in correctly classifying the biflows. Figure 1 shows the (relative) contribution of each modality with two boxplots obtained by aggregating (based on all the correctly-classified tested samples) the pooled SHAP values $\widetilde{\phi}_{\mathcal{M}}$ for each modality, calculated using Deep SHAP (cf. Sec. 3.2). We can notice that the PAY-modality *always* contributes with higher importance values. A complementary fine-grained per-app analysis—not shown for brevity—highlighted analogous patterns, confirming the greater importance of the PAY-modality over the PSQ-modality.

Below, we focus on the importance of inputs associated with the PAY-modality, which is fed with the first $N_b = 576$ bytes of transport-layer payload of each biflow. In Fig. 2, we highlight positive and negative SHAP values at per-sample level with red and blue colors, respectively. These distributions are integrated with the median importance value of each byte (over different samples), which is reported as a solid black line. This helps to understand regions that are most consistently influential for predictions and to assess their variability. In more detail, Fig. 2 depicts the global explanations associated with single apps, focusing on some notable examples discussed hereinafter.

For Dropbox (Fig. 2a), we notice an initial region that reaches high per-sample positive values testifying that the first 200 bytes play, on average, a crucial role in correctly identifying this app. Such a result is also confirmed by the median value. In contrast, the median of the distribution for the other bytes is very low, suggesting their low relevance. However, the situation is not always so sharp: for a *small group* of apps (e.g., Spotify—Fig. 2b), it is difficult to identify which regions are consistently influential and even the median does not highlight specific groups of bytes. For these apps, we often observe denser blue distributions, revealing the presence of input values that confuse the classifier. Finally, for the other apps including Trello (Fig. 2c), the behavior is different: we can clearly observe a central region corresponding to the interval $[300; 400]$ B which supports correct predictions and stresses the importance of considering a value for N_b not less than 400 B (as in MIMETIC).

We now discuss the importance of the inputs associated with the PSQ-modality of MIMETIC architecture which is fed with 4 informative fields extracted from the first 12 packets of each biflow, namely inter-arrival time (IAT), direction (DIR) $\in \{-1, 1\}$, TCP window size (TCP_WS) set to zero for UDP, and transport-layer payload length (PL). This analysis provides global explanations at per-app granularity. To this end, we discuss the results obtained for some example apps below.

Figure 3 shows the median importance values for each element of the 4×12 matrix being the input of the PSQ-modality. For Facebook (Fig. 3a), the importance is mainly related to PL and TCP_WS, with a decreasing importance of the packets after the very first ones. This last observation applies to all fields but IAT which is characterized by a flatter distribution. Altogether, we can see that the first four packets are the most important. Similarly, for OneDrive (Fig. 3b) we can notice that TCP_WS and PL are the fields with the highest importance, whereas the other two fields expose minimal median scores, regardless of the packet number. For other apps, it is harder to recognize the most important fields: for example, the matrix elements of Hangouts (Fig. 3c) do not present remarkably higher values, beyond TCP_WS of the 2^{nd} packet and PL of the first 5 packets. Analogously, the ilMeteo matrix (Fig. 3d) results scattered without fields of notable importance. A different situation is observed for Skype and Trello. For the former app, DIR of packets around the 7^{th} is useful for classification. On the other hand, IAT plays an essential role together with DIR for Trello. These results



(a) `Facebook`. (b) `OneDrive`.

(c) `Hangouts`. (d) `ilMeteo`.

(e) `Skype`. (f) `Trello`.

Figure 3: Importance $\widetilde{\phi}_m$ of the header-field inputs (identified by the respective packet) for `PSQ`-modality of MIMETIC for exemplifying apps.

confirm that all the fields considered can play a crucial role in identifying the correct app that generates the observed biflows, albeit with some differences among the apps.

5 Conclusions and Future Avenues

In the present work, we explained the black-box behavior of DL-based traffic classifiers, focusing on our previous state-of-art MIMETIC proposal, using tools from the XAI domain, and evaluating our methodology on an open dataset of mobile-app traffic (including 41 apps). Based on Deep SHAP, we were able to draw *global explanations* which allowed us to assess the weight of each modality and the importance of the input data fed to the `PSQ` (resp. `PAY`) modality, quantifying the importance of each packet field (resp. payload byte).

Future avenues of research will include (i) the investigation of trustworthiness of traffic classifiers (via calibration analysis), (ii) the comparison of global explanations obtained via other XAI tools (e.g., LRP [Bach *et al.*, 2015]), (iii) the application of XAI techniques to more challenging multitask DL architectures and to other network traffic analysis tasks (e.g., prediction), (iv) the robustness assessment of multimodal DL-based traffic classifiers against adversarial attacks, and (v) the design of (natively) self-explainable DL-based traffic classifiers.

References

[Aceto *et al.*, 2019a] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, and Antonio Pescapé. MIRAGE: mobile-app traffic capture and ground-truth creation. In *4th IEEE International Conference on Computing, Communications and Security (ICCCS)*, pages 1–8, 2019.

[Aceto *et al.*, 2019b] Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, and Antonio Pescapè. MIMETIC: mobile encrypted traffic classification using multimodal deep learning. *Elsevier Computer Networks*, 165:106944, 2019.

[Aceto *et al.*, 2021] Giuseppe Aceto, Giampaolo Bovenzi, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, and Antonio Pescapé. Characterization and prediction of mobile-app traffic using Markov modeling. *IEEE Transactions on Network and Service Management*, 18(1):907–925, 2021.

[Amarasinghe *et al.*, 2018] Kasun Amarasinghe, Kevin Kenney, and Milos Manic. Toward explainable deep neural network based anomaly detection. In *IEEE 11th International Conference on Human System Interaction (HSI)*, pages 311–317, 2018.

[Bach *et al.*, 2015] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

[Beliard *et al.*, 2020] Cedric Beliard, Alessandro Finamore, and Dario Rossi. Opening the deep pandora box: Explainable traffic classification. In *IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 1292–1293, 2020.

[Hagras, 2018] Hani Hagras. Toward human-understandable, explainable AI. *IEEE Computer*, 51(9):28–36, 2018.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *31st international conference on neural information processing systems (NeurIPS)*, pages 4768–4777, 2017.

[Shapley, 1953] Lloyd S Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

[Wang *et al.*, 2020] Xin Wang, Shuhui Chen, and Jinshu Su. Real network traffic collection and deep learning for mobile app identification. *Hindawi Wireless Communications and Mobile Computing*, 2020.