PhD in Information Technology and Electrical Engineering

Università degli Studi di Napoli Federico II

PhD Student: Idio Guarino

Cycle: XXXVI

Training and Research Activities Report

Academic year: 2021-22 - PhD Year: Second

Tutor: Prof. Antonio Pescapè

Date: October 30, 2022,

1. Information:

PhD student: Idio Guarino **PhD Cycle: XXXVI**

DR number: 995139

Date of birth: 05/03/1990

Master Science degree: Computer Engineering University: University of Napoli Federico II

Scholarship type: no scholarship. Funded by Consortium GARR through the awarded "O.Carlini" research grant.

Tutor: Prof. Antonio Pescapè

Co-tutor:

2. Study and training activities:

Activity	Type[1]	Hours	Credits	Dates	Organizer	Certificate[2]
Cyber security in Akka Technologies	Seminar	1.5	0.3	03/11/2021	Prof. D. Cotroneo, Prof. S.P. Romano, Dr. R. Natella	Y
Vehicular Hacking in Akka Technologies	Seminar	2.0	0.4	03/11/2021	Prof. D. Cotroneo, Prof. S.P. Romano, Dr. R. Natella	Y
Single cell omics leverage Machine Learning to dissect tumor microenvironment and cancer immuno editing	Seminar	2.0	0.4	02/12/2021	Prof. A. Corazza	Y
Threat Hunting Use-Cases	Seminar	2.0	0.4	13/12/2021	Prof. D. Cotroneo, Prof. S.P. Romano, Dr. R. Natella	Y
The quest of quantum advantage with a photonics platform	Seminar	1.5	0.3	03/02/2022	Prof. G. Ascione, Prof. M. Benetti,	Y

					Prof. M. Coraggio	
RAILS MID-TERM WORKSHOP	Seminar	5.0	1.0	25/02/2022	Prof. V. Vittorini	Y
Project Vāc:Can a Text-to-Speech Engine Generate Human Sentiments? (Picariello Lecture)	Seminar	1.0	0.2	28/02/2022	Prof. F.Amato, Prof. G.Luongo	Y
Ethics and Politics of A.I. (Picariello Lecture)	Seminar	2.0	0.4	11/04/2022	Prof. F.Amato, Prof. G.Luongo	Y
Explainable Natural Language Inference	Seminar	1.5	0.3	13/04/2022	Prof. F. Cutugno	Y
Accelerated Deep Learning via Efficient, Compressed and Managed Communication	Seminar	1.0	0.2	05/03/2022	Prof. A. Pescapè	Y
TMA PhD School	Doctoral School	16.0	3.2	27-28/06/2022	University of Twente, Netherland	Y
Data Management	Course	-	I Semester	6.0	Prof. F.Amato	Y

2.1. Study and training activities - credits earned

	Courses	Seminars	Research	Tutorship	Total
Bimonth 1	**6.0**	**1.5**	**2.5**	**0**	**10.0**
Bimonth 2	**0**	**1.5**	**8.5**	**0**	**10.0**
Bimonth 3	**0**	**0.7**	**9.3**	**0**	**10.0**
Bimonth 4	**0**	**3.4**	**7.8**	**0**	**11.2**
Bimonth 5	**0**	**0**	**8.8**	**0**	**8.8**
Bimonth 6	**0**	**0**	**10.0**	**0**	**10.0**
Total	**6.0**	**7.1**	**46.9**	**0**	**60**
Expected	**30 - 70**	**10 - 30**	**80 - 140**	**0 – 4.8**	

3. Research activity:

During the last two years, governments worldwide have imposed lockdown measures following the Covid-19 outbreak. The imposition of such restrictions has forced millions of people to stay at home and, if possible, study, work and socialize from their homes.

Consequently, this caused significant growth in residential traffic (+20%), mainly due to the use of smart working and distance learning tools (+200%)[Feldmann2020]. In addition, changes in user mobility also had an impact on cellular networks, which experienced a decrease in downlink traffic (-20%) accompanied by a sudden increase in voice and uplink traffic (+10%)[Lutu2020]. However, the effects of these sudden changes have also been observed from the perspective of network performance, in terms of increased variability in delay, loss rate, and latency[Candela2020].

Looking at the composition of traffic, these changes can be attributed to the massive use of different categories of smart working and distance learning applications (e.g., video, social media, messaging, and collaboration tools) [Sandvine2020][Affinito2021]. In addition, as shown, this phenomenon is not limited to past lockdown periods, but also extends to the present day, where growth in traffic volumes continues to be attributed primarily to streaming video, social media, and messaging traffic, which continues to grow year after year[Sandvine2021].

In line with these considerations, from the perspective of network operators, to respond in a timely and effective manner to unexpected and massive changes in traffic characteristics, the need arises to define innovative tools to support network monitoring, management, and engineering activities.

To this end, my research is focused on defining AI-based methodologies for traffic classification and prediction, both of which are prerequisites for proper network management and optimization. Specifically, traffic classification allows the categorization of traffic traversing networks according to heterogeneous views such as type, application, or specific activity performed by the user, allowing the network to be optimized according to the content to be carried. Conversely, traffic prediction allows prediction of how traffic characteristics will evolve, enabling efficient management of resources and services in the immediate term, as well as better planning of infrastructure capacity over the long term.

However, although traffic classification and prediction have been under study for a long time, their application is continually being questioned due to the changing characteristics of Internet traffic, especially those involving mobile networks. In particular, the traffic generated by modern applications is encrypted, making it impossible to interpret the information exchanged between the parties, and consequently to apply the classic methodologies based on content analysis. Furthermore, port analysis methodologies are challenged by dynamic port assignment and the interfering action of the NAT protocol. Finally, the difficulties are exacerbated in the case of mobile apps due to their particular characteristics of *homogeneity*—due to the use of common services (e.g., third-party services)—, *continuous evolution*—caused by frequent and automated software updates—, and *heterogeneity*—due to their execution on heterogeneous devices and to the specific activity carried out by the user during their use.

In agreement, motivated by the promising results shown by recent research through the application of AI-based techniques in the field of traffic analysis [Aceto2021a, Aceto2021b], my research activity is aimed at defining innovative tools based on Deep Learning, which is particularly effective for the design

of traffic analysis techniques starting from the observation of "raw" traffic (i.e., no pre-processing effort). In this scenario, my work is mainly focused on innovative methodologies based on multi-modal and multi-task learning. Specifically, multi-modal learning can effectively exploit the heterogeneous nature of the different information extracted from traffic (modalities), capturing both intra- and inter-modal dependences. Conversely, multi-task learning allows the construction of models capable of solving multiple problems simultaneously (tasks), also bringing benefits to the effort required for their construction (e,g,., fewer models to be trained and evaluated).

In line with studies conducted during the previous year [Guarino2021a] [Guarino2021b], during the second year, I continued to focus my research on analyzing the network traffic generated by apps for communication and collaboration (C&C apps in the following), used for business meetings, classes, and social interactions. Indeed, such apps experienced a massive increase in usage when the "stay at home" order was issued worldwide and are still widely used due to people's changing lifestyles and work styles.

Accordingly, given the crucial role attributed to the availability of real traffic data related to these apps, I managed an intensive campaign to collect network traffic generated by the most popular and used C&C apps (e.g., Skype, Slack, Teams, Zoom, etc.). This campaign involved more than 150 students from courses at the University of Naples Federico II and allowed to build a dataset that currently includes more than 300 hours (and 20 GB of volume) of real traffic data related to 9 C&C apps, which were also used by users to perform different activities (e.g., chats, audio calls, video calls, etc.). In addition, the resulting dataset has been released publicly to support the scientific community and incentivize related research activities [Mirage2022].

Leveraging these data, my research was aimed at investigating some research opportunities that emerged as a result of the analysis and characterization of traffic from the perspective of the app and the specific activity performed by the user, and which still represent open issues for the scientific community [Guarino2021b].

Specifically, I addressed an early classification problem using the traffic biflow as the traffic object. In contrast to post-mortem classification, early classification consists of classifying each biflow by observing only its initial part (i.e., the first packets). To this end, I first investigated the ability of both single-modal and multi-modal Deep Learning-based state-of-the-art classifiers in telling the specific app, the activity performed by the user, and both of them. Experimental results highlighted that, while these classifiers perform more than satisfactorily concerning classifying the app (~99% of F-measure), they do not perform as well when used to classify the specific activity performed by the user (~65% of F-measure). As also demonstrated, this phenomenon is due to the fact that by exploiting input data commonly used in the state of the art (e.g., packet direction, number of bytes of payload, payload content, etc.), different activities performed within the same app exhibit similar traffic patterns, and this leads to confusing classifiers, highlighting their limitations concerning this specific type of problem.

Accordingly, through the study of traffic patterns characteristic of such activities, my research led to the definition of innovative inputs (namely Context Input) based on the analysis of the context of the biflow to be classified (i.e. the set of co-existing biflows which run in parallel)[J1]. The potential of this novel set of inputs was highlighted through an initial traffic characterization analysis, which showed that different activities exhibit different behaviors based on Context Input.

Therefore, leveraging multi-modal learning, I designed a deep-learning architecture capable of capitalizing on the heterogeneity of network traffic according to different views (viz. modalities). Experimental results showed that combining the new set of inputs (i.e., Context Input) with those commonly used in the state of the art (e.g., payload bytes and features extracted from packet headers) led to a significant increase in performance when facing the user activity traffic classification (~+17% of F-measure) with an additional one concerning the app traffic classification (~+1% of F-measure).

Moreover, further analysis showed how Context Inputs can be used to replace common inputs based on payload content, allowing for reduced training time and increased robustness against future, more opaque cryptographic sublayers (e.g., TLS with Encrypted Server Name Indication or Encrypted Client Hello extensions), while paying a negligible cost in terms of performance (<1% of F-measure regarding activity classification). Finally, an in-depth analysis highlighted that the adoption of Context Inputs, in combination or not with other types of inputs, has also a non-negligible impact on the calibration of the models which tend to be more reliable.

At the same time, I experimentally evaluated the performance of heterogeneous deep-learning architectures (i.e., convolutional, recurrent, and the composition of both)[C2]. Specifically, models were trained by adopting two strategies that differ in the way traffic information is grouped, leading to models capable of capturing traffic features not limited to a single app: at the *app level* (i.e., a separate model is obtained for each app) and at the *all-app level* (i.e., a single model is obtained for all apps). The evaluation was carried out by taking into account the traffic prediction performance of the models and their relative complexity to find the best trade-off. To this end, the analysis showed that a single model trained on the entire CC app traffic is sufficient, as there appears to be no appreciable gain in training a specialized model for each app. This result has a significant practical impact, as it saves the training, implementation, and management of multiple models. In addition, the results highlighted a variety of behaviors for different apps and the activities a user can perform with them. Simultaneously, I made the first attempt to interpret the results obtained from these predictors via eXplainable Artificial Intelligence (XAI), showing how the inputs used to feed the models influence their predictions.

Finally, I also applied ML classification techniques in an Intrusion Detection context on the CSE-CIC-IDS2018 benchmarking dataset [C1]. Specifically, in contrast with the related literature—which focuses on a post-mortem classification—, a new perspective on this dataset is proposed, based on a new set of (unbiased) features computed by taking into account only the initial part of each biflow, thus allowing to perform early traffic classification. Furthermore, the analysis was conducted on a variable number of packets with five different ML classifiers, which were compared in terms of both performance and computational complexity.

My research is founded by the Consortium GARR, the Italian national network of University and Research, through the awarded “O. Carlini” research grant.

4. Research products:

[C1] “*On the use of Machine Learning Approaches for the Early Classification in Network Intrusion Detection*”, Idio Guarino, Giampaolo Bovenzi, Davide Di Monda, Giuseppe Aceto, Domenico Ciuonzo and Antonio Pescapè. 2022 IEEE International Symposium on Measurements and Networking (M&N), 2022. Published

[C2] "*Fine-Grained Traffic Prediction of Communication-and-Collaboration Apps via Deep-Learning: a First Look at Explainability*", I. Guarino, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico and A. Pescapé. 2023 IEEE International Conference on Communications (ICC). Submitted.

[J1] "*Contextual Counters and Multimodal Deep Learning for Activity-Level Traffic Classification of Mobile Communication Apps during COVID-19 Pandemic*", I. Guarino, G. Aceto, D. Ciuonzo, A. Montieri, V. Persico and A. Pescapé, "," Elsevier Computer Networks, Special issue on Machine Learning empowered Computer Networks, 2022, Accepted for publication.

5. Conferences and seminars attended

- 2022 IEEE International Symposium on Measurements and Networking (M&N)
- Ital-IA, secondo Convegno Nazionale sull'Intelligenza Artificiale del CINI, 10/02/2021.

6. Periods abroad

During my second year, I started a period of study and research abroad at the Huawei R&D center in Paris, France, under the supervision of Alessandro Finamore (Principal Engineer at Huawei R&D Center). My period abroad started on **25/07/2022** and will end on **25/01/2023** (i.e., **6 months** in total).

Study and research activity concerns the design, implementation, and evaluation of few-shot learning techniques—typically employed in the field of computer vision—for the classification of the traffic generated by mobile apps.

7. Tutorship

8. Plan for year three

For the third year, based on the studies already conducted during the still ongoing research period abroad, I plan to continue investigating the application of few-shot learning approaches in the context of Internet traffic classification. The goal is to start by evaluating existing state-of-the-art approaches—typically employed in the field of computer vision—and then design a new proposal that would work well in the different and more complex context of network traffic.

In addition, I plan to continue my research on traffic classification and prediction, focusing mainly on the traffic of communication and collaboration apps. The idea is to converge the studies conducted so far, performed by treating the two problems separately, to design, implement and evaluate a new proposal that can solve them simultaneously by taking advantage of modern advances based on multi-modal and multi-task learning.

References

[Feldmann2020]"The Lockdown Effect: Implications of the COVID-19 Pandemic on Internet Traffic", A. Feldmann, O. Gasser, F. Lichtblau, E. Pujol, I. Poese, C. Dietzel, D. Wagner, M. Wichtlhuber,

J.Tapiador, N. Vallina-Rodriguez, O. Hohlfeld, and G. Smaragdakis, ACM Internet Measurement Conference (IMC '20).

[Lutu2020]“A characterization of the COVID-19 pandemic impact on a mobile network operator traffic”, A. Lutu, D. Perino, M. Bagnulo, E. Frias-Martinez, J. Khangosstar, ACM Internet Measurement Conference (IMC), 2020, p. 19–33.

[Candela2020]“Impact of the COVID-19 pandemic on the Internet latency: A large-scale study”, M. Candela, V. Luconi, A. Vecchio, Computer Networks, 2020.

[Sandvine2020]“The Global Internet Phenomena Report COVID-19 Spotlight”, Sandvine, 2020.

[Sandvine2022]“The Global Internet Phenomena Report”, Sandvine, 2022.

[Aceto2021a] G. Aceto, D. Ciuonzo, A. Montieri, and A. Pescapé, “DISTILLER: Encrypted Traffic Classification via Multimodal Multitask Deep Learning“, Elsevier Journal of Network and Computer Applications, 2021.

[Aceto2021b] G. Aceto, G. Bovenzi, D. Ciuonzo, A. Montieri, V. Persico, and A. Pescapé, “Characterization and Prediction of Mobile-App Traffic using Markov Modeling“, IEEE Transactions on Network and Service Management, 2021.

[Affinito2021]“The impact of COVID on network utilization: an analysis on domain popularity”, A. Affinito, A. Botta, G. Ventre, IEEE 25th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2020, pp. 1–6.

[Guarino2021a]"Characterizing and Modeling Traffic of Communication and Collaboration Apps Bloomed With COVID-19 Outbreak", Idio Guarino, Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico, Antonio Pescapè, 2021 IEEE 6th International Forum on Research and Technology for Society and Industry (RTSI), 2021.

[Guarino2021b]“Classification of Communication and Collaboration Apps via Advanced Deep-Learning Approaches”, Idio Guarino, Giuseppe Aceto, Domenico Ciuonzo, Antonio Montieri, Valerio Persico and Antonio Pescapè, 2021 IEEE International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), 2021.

[Mirage2022] https://traffic.comics.unina.it/mirage/mirage-covid-ccma-2022.html