LISTIC

Offre de stage 2021-2022

Titre	Automated feature selection for network traffic models
Niveau du stage	Master 2ème année / Ingénieur 5ème année
Date de début/ fin	De février-mars 2022 au juillet 2022 (4-6 mois)
Ville, Pays	Annecy-le-Vieux, *France*
Laboratoire	LISTIC - Laboratoire d'Informatique, Systèmes, Traitement de l'Information et de la Connaissance - http://www.polytech.univ-savoie.fr/LISTIC
Description du sujet de stage	**Context.** Network management often relies on machine learning to make predictions about performance and security from network traffic (e.g., [2]). Often, the representation of the traffic is as important as the choice of the model. The features that the model relies on, and the representation of those features, ultimately determine model accuracy, as well as where and whether the model can be deployed in practice. Thus, the design and evaluation of these models ultimately requires understanding not only model accuracy but also the systems costs associated with deploying the model in an operational network. Towards this goal, in our previous work we defined a new framework and system that enables a joint evaluation of both the conventional notions of machine learning performance (e.g., model accuracy) and the systems-level costs of different representations of network traffic. In our work we demonstrated the benefit of exploring a range of representations of network traffic and presented Traffic Refinery [3], a proof-of-concept implementation that both monitors network traffic at 10 Gbps and transforms traffic in real time to produce a variety of feature representations for machine learning. Traffic Refinery both highlights this design space and makes it possible to explore different representations for learning, balancing systems costs related to feature extraction and model training against model accuracy. **Project goals.** Traffic Refinery enables an exciting line of research by demonstrating both the feasibility and utility of exploring a range of traffic representations, beyond the current, limited dichotomy of packet captures and measurement systems. However, our work only scratched the surface of accounting for the system challenges generated by the use of machine learning models to solve network management tasks. One core issue is that Traffic Refinery still requires a large amount of manual input from its users. The advent of automated machine learning pipelines (e.g., AutoGluon [1]) makes it possible and efficient to explore a wide range of models (and model parameters), and to find the highest model performance relatively quickly. However, these pipelines typically assume that models are trained and executed on the same data, which may not the correct trade-off for models that evaluate network traffic. For example, a model that achieves 90% the accuracy but uses 10% of the system resources may be more realistically deployed. **Internship goals.** Our work first studied how to make it possible to define various data inputs that tradeoff systems constraints and model accuracy to arrive at a model that is suitable for practical deployments. In particular, our work considered the processing and state (i.e., memory) costs of different data representations. During the internship, the student will extend this work to understand additional costs, including latency, model training time and complexity, and energy consumption. Further, the student will explore how to extend our platform to integrate with AutoML pipelines, so that we can more automatically determine what parameters and algorithms best meet operational constraints (i.e., we will extend AutoML pipelines to not only consider the accuracy of models when making recommendations, but to additionally consider operational constraints and to make recommendations within this constraint space). **References.** [1] AutoGluon: AutoML Toolkit for Deep Learning. https://auto.gluon.ai/ [2] F. Bronzino, P. Schmitt, S. Ayoubi, G. Martins, R. Teixeira, and N. Feamster. Inferring streaming video quality from encrypted traffic: Practical models and deployment experience. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 3(3), Dec. 2019.

	[3] F. Bronzino, P. Schmitt, S. Ayoubi, H. Kim, R. Teixeira, and N. Feamster. Traffic refinery: Cost-aware traffic representation for machine learning in networks. Proceedings of the ACM on Measurement and Analysis of Computing Systems, 5(3), Dec. 2021.
Compétences requises	• Comfortable speaking English or French (French is not required). • Good understanding of at least one between computer network protocols and systems or machine learning / vision methods (preferably both) • Good proficiency with at least one programming language, preferably Python, Golang, or Rust
Gratification	Selon législation en vigueur
Tuteurs / Contacts	Francesco Bronzino – fbronzino@univ-smb.fr