Democratizing Networking Research in the Era of AI/ML

Arpit Gupta [⋆] and Walter Willinger [†]
[⋆]UC Santa Barbara [†]NIKSUN Inc.

The networking area is currently facing a growing "digital divide" between those researchers (i.e., in industry) that have access to plenty of data and those (i.e., in academia) that suffer from a basic lack of data, and this divide becomes increasingly problematic with the rise of AI/ML applications for networking [1]. At the same time, a closer look at the areas where AI/ML applications have generated excitement in recent years (e.g., computer vision, autonomous car technology) shows that their successes depended critically on having access to an abundance of publicly available (labeled) data. For example, the field of computer vision has benefited immensely from IMAGENET, and the recently provided open-source collections of high-quality datasets from different commercial self-driving car companies (e.g., Argoverse by Argo AI, nuScenes by Motional, Waymo Open Dataset) have been a boon to academic researchers in the field of autonomous car technology. By providing researchers with an opportunity to focus on developing new and better learning algorithms and spend less or no time on data collection and labeling, these efforts have fuelled as well as democratized research in their respective fields.

In stark contrast, the field of networking has had no equivalent of IMAGENET, and even worse, there exists currently no consensus on what that equivalent should be or look like. Instead, for many problems, researchers typically have to start from scratch, first conjecturing what the critical features for the problem at hand might be, then developing the tools or systems for collecting the necessary data, and finally extracting the identified features from the obtained data. As a result, these researchers often spend more time on designing and running experiments to collect the data needed for extracting the features required for the development of their learning models, with essentially no opportunities to evaluate the resulting models in settings that have any resemblance to a real-world production network.

To correct this situation and also achieve the goal of democratizing networking research in the era of AI/ML, we propose a "federated model" for collecting, storing, and using networking data where campus networks (i.e., instances of somewhat specialized but nevertheless real-world production networks) are the entities that maintain their full autonomy. In particular, we envision academic network researchers (with support from their institution and in collaboration with their IT organization) to transform their campus networks such that each of these networks or "enclaves" maintains its autonomy and can serve as a rich source of real-world network data, maintain its own comprehensive data store, and also function as a testbed where the researchers can evaluate and "road-test" their learning solutions under real-world conditions.

Our proposed federated model provides a unique platform for making AI/ML-based networking research a truly community-based effort that makes the most of the multiple autonomous data collection and storage efforts. On the one hand, in view of recent advances in programmable data-plane targets and scalable storage/analytics systems, the envisioned transformation of existing campus networks is largely a non-technical and more of an organizational challenge and will require some amount of oversight/coordination to ensure that the different participating campus networks meet some minimum standard with respect to traffic monitoring, data collection, and data storage capabilities. On the other hand, the successful implementation of the proposed federated model promises to revolutionize AI/ML-based networking research and have profound implications on how AI/ML research artifacts are used in practice, be it for network performance-related or cybersecurity-specific problems. In this context, some of the critical open problems include: (1) How does the proposed federated model help in improving the training and evaluation of learning models (e.g., federated machine learning)? (2) How can the proposed federated model be leveraged to increase the trust of network operators in newly-developed learning models (e.g., explainable AI/ML)? (3) What feasible road map does the proposed federated model suggest for "road-testing" newly-developed learning models in real-world production networks? (4) Is there a role for privacy-preserving network traffic collection at scale in a federated model?

References

[1] Gupta, A., Mac-Stoker, C., and Willinger, W. An effort to democratize networking research in the era of ai/ml. In *Proceedings of the 18th ACM Workshop on Hot Topics in Networks* (2019), pp. 93–100.