Challenges in Using AI/ML for Networking

Ram Durairajan and Reza Rejaie, *University of Oregon*; and Walter Willinger, *NIKSUN, Inc.*

Motivated by the recent success of machine learning (ML) in domains such as computer vision and autonomous car technology, we are witnessing enormous interest in applying ML to an ever-wider range of problems in the networking domain (*e.g.*, network automation, self-driving networks). However, unlike other domains, the networking area poses several immediate and serious challenges that have impeded the rapid adoption of ML and have been responsible for the slow pace of innovation in ML for networking (ML4Net) research. Among the key challenges are a commonly-acknowledged and much-maligned lack of readily available data, questions concerning the representativeness of collected data, a general inability to label networking data at scale, and the privacy-sensitive nature of the data obtained from real-world networks. To make progress, an important question networking researchers have to solve in the near future is *how to programmatically label the many different types of network data in a scalable, accurate, and low-cost manner without violating stringent privacy concerns?*

Problem 1: Lack of Representative Data. In the midst of ongoing discussions about the "haves" (*i.e.*, researchers in industry) and "have-nots" (*i.e.*, academic researchers) with respect to owning or having access to real-world network data, we are motivated by the success that IMAGENET has had in democratizing ML-based research in the field of computer vision. While we do not argue that IMAGENET is the "right" model for network data (it is not!), we emphasize that to succeed in providing a level playing field for network researchers, new models for dealing with the data problem in ways that respect the unique nature of networking data are needed.

Problem 2: Lack of Labeled Data. However, there is more to democratizing ML-based networking research than overcoming the lack of available rich data. For example, researchers are immediately faced with another formidable and largely unsolved problem: a general paucity of *labeled* networking data. To illustrate, the networking domain lacks in general well-established and commonly-agreed upon features for accurately describing different events of interest (*e.g.*, an onset of volumetric DDoS attacks). This "fuzzy" nature of network data is further aggravated by the fact that the data is generally collected and curated in a highly piecemeal manner; that is, at different granularities and locations, under different conditions, or with varying semantic information.

Problem 3: Quality of Labeled Data. Unlike IMAGENET where crowd-sourcing has been effectively leveraged to create a large database of hand-annotated images, network data is, in general, more complex than typical dog or cat pictures. That is, its correct interpretation or labeling often requires substantial domain knowledge (*e.g.*, protocols, configurations, policies). As a result, the use of popular crowd-sourcing methods (as in the case of IMAGENET) or more recently pursued out-sourcing efforts that involve commercial data labeling companies (as is the case for data used for autonomous driving) have to be ruled out as low-cost, scalable and accurate labeling approaches.

Problem 4: Privacy Requirements. The networking area is unique with respect to the difficulties caused by the privacy-sensitive nature of most network data. This aspect by and large rules out the sharing of raw or labeled data with third-party researchers and also limits the sharing of certain ML research artifacts such as learning models due to possible privacy leaks. In short, while attempts to democratize the use of ML in other areas have been enormously successful, to date they have been largely futile in the networking domain. To ensure that networking can similarly benefit from efforts where its use of ML becomes a collaborative and community-driven activity, new ideas are needed that specifically address the unique nature of the data.

Ongoing Efforts. To provide networking researchers with a low-cost, scalable, and high-quality methodology for labeling their data, we are working on the design of a framework called EMERGE that seeks to extend the idea of weak supervision-based data labeling and supports collaborative efforts for creating data labels at scale, of good quality (e.g., dealing with bias in the data), at low cost (e.g., supporting a community-based sharing effort), and, if necessary, in ways that respect prevailing data privacy concerns.

What is Needed? There are formidable challenges that an effort like EMERGE poses and support from NSF will be critical to ensure its success and impact on the community. These challenges include:

- Annotating network data with relevant metadata. This metadata is critical for creating ground truth that serves as "decision heuristic" for training (using weak supervision or some other method) ML models.
- Enabling privacy-preserving collaborations among the third-party researchers and supporting independent validation of separate research endeavors. The networking community is in need for a "glass box" framework where learning algorithms, decision heuristics, and dataset descriptions can be broadly shared without requiring any sharing of raw data or trained ML models.
- Designing new mechanisms for dealing with (*i.e.*, identifying, quantifying, and correcting) hidden data biases and discovering the emergence of novel types of biases in measurement data.