University *of* Dayton
School of
Engineering

AI Development and Validation in Network Traffic Analytics

Research questions

Artificial Intelligence (AI) development and validation in network traffic analytics, including network traffic classification, network traffic prediction and network intrusion detection.

Recently, AI, especially deep learning models haven been widely introduced to network traffic analytics. For example, deep learning approaches such as CNN, MLP and LSTM have been widely introduced to network traffic analytics Those AI approaches have clear advantages over conventional schemes, such as port recognition, probe based techniques, deep packet inspection, etc., especially when dealing with encryption data traffic flows.

List of issues

However, unlike several other popular areas of AI development, such as image processing, natural language processing, and autonomous driving, AI development in network traffic analytics lacks open and well-maintained open datasets.

1. *Lack of open and well-maintained open datasets*. Many existing works have relied on the same open dataset that could be improved in several aspects. For example,
    a. The variety of network applications is limited.
    b. Samples are limited and unbalanced across different network applications.
    c. The data samples are noisy, e.g., empty packets, TCP control packets, etc.
    d. The dataset is rarely updated. The AI models developed from the dataset cannot be implemented to analyze live data traffic from the corresponding network applications in real life, because the applications have been updated, e.g., with new network protocols, encryption algorithms, etc.
2. *Lack of sharing point* of developed AI models for network traffic analytics
    a. It is hard to validate or compare the performance of those AI models.
    b. Some of the existing works are based on in-house collected dataset that is not open to the public.

How NSF could support the research arc

Establish a dataset for the purpose of AI development may be supported in two ways.

1. Data collection and dataset maintenance may not be fundamental research. Therefore, extra support may need to be specified in core and/or supplementary programs.
2. Provide support for simulator and testbed development. The current opportunities (e.g., MRI) focus much on the equipment. However, the simulator and testbed development that can support AI development in network traffic analytics may require a fair amount of human efforts.

Feng Ye, PhD, Assistant Professor
Department of Electrical & Computer Engineering, University of Dayton
E-mail: fye001@udayton.edu