Advancing packet-level traffic predictions with Transformers
Master thesis proposal

Understanding the dynamics of traffic from network applications like video streaming is an incredibly complex task. Even though the fundamental algorithms follow well-defined rules, their interactions across networks cause complex patterns. Simple heuristics—and even complicated, resource-intensive simulations—struggle to reproduce this. To make matters worse, it is not enough to estimate ‘average’ traffic behavior, as many network applications are sensitive to outliers (i.e., the distribution tail). In video streaming, only a few lost packets among thousands may force the video to rebuffer and significantly degrade the quality of experience for users.

In recent years, the networking research community has increasingly turned to Machine Learning (ML) for a wide range of applications, yet the complexities of modern networks prohibit just applying existing ML approaches, which has often resulted in unreliable performance.

In this thesis, we want to investigate state-of-the-art Transformer architecture to learn traffic dynamics. Transformers are models for sequence-to-sequence prediction that made their entrance with language translation tasks, but have since shown their potential for other tasks such as generating complex texts from short input sentences or even image generation. In particular, Transformers have shown great capability to generalize. We want to investigate whether we can leverage these generalization capabilities for network traffic prediction: can we use Transformers to predict e.g. transmission times for future packets (output sequence) based on past packets (input sequence)? Can we generalize network dynamics learned from multiple traces such that we do not need to re-train a model from scratch for any new network we encounter?

We expect two major tasks for the beginning of this project:

- We need to generate training and validation data. Real network traces are often sparse (e.g. only 1 hour of data every month), we plan to start this project using network simulations, which allow us to control the ground truth and better understand the performance of the model, while gradually increasing the difficulty. In particular, we may start with a fixed network topology, number of clients, and traffic distributions used by clients. We can repeat this simulation several times, collecting different instantiations of traffic dynamics, and use this data to train a transformer. We can evaluate this model with an independent run of the simulation. To increase difficulty, we may use random topologies, numbers of senders, multiple potential workloads, etc.

- We need to investigate the best format to feed data into the transformer. We need an approach similar to word embeddings in natural language processing. There is some previous work in this area (Packet2Vec) that we can use as a starting point. However, a major point of consideration not considered by previous work are different instances of “shared fate”: all packets going to the same host are affected if the host is overloaded; same for all packets going over the same congested link. At the same time, all flows using a particular congestion control protocol behave similarly, regardless of destination. These insights are important for predicting dynamics, and we need to ensure that we create an embedding that allows the Transformer to learn them.

Based on our progress in these tasks we may also consider real network traffic and other prediction tasks, e.g. can we use our pre-trained transformer as the bases of a classification system?

Contact

- Alexander Dietmüller, adietmue@ethz.ch
- Dr. Romain Jacob, jacobr@ethz.ch
- Prof. Laurent Vanbever, lvanbever@ethz.ch