Name: Patrick Watters

Degree: PhD, Computer Science and Engineering

Institution: University of Nevada, Reno

Mentors: Lei Yang, Paarijaat Aditya, Feng Yan

Research Title: Trustless Machine Learning Inference using Zero-Knowledge Proofs

Abstract:

Machine learning models are often kept as closed-source to protect proprietary algorithms, business strategies, and intellectual property. As a result, model consumers, such as developers, researchers, or end-users, are restricted from assessing the models' internal workings. This makes it challenging to verify its performance claims. The lack of transparency leads to a reliance on trust, as model consumers must believe that the model outputs have been computed honestly and accurately. In this research we explore verifiable model evaluation by producing zero-knowledge proofs for the model inference process. A zero-knowledge proof allows one party (the prover) to demonstrate to another party (the verifier) that they possess certain information, without revealing the information itself. By leveraging zero-knowledge proofs, model outputs can be verified without exposing the models' internal details, promoting both trust and transparency. The challenge in generating zero-knowledge is that they require immense computational resources and often exceed the memory of a single machine. To overcome this limitation, we present a distributed methodology for generating zero-knowledge proofs. This distributed approach is designed to scale efficiently, accommodating models of arbitrary size. The overarching goal is to contribute to the broader adoption of zero-knowledge proofs in machine learning. In a world where artificial intelligence is having an increasingly pivotal role, it is important that we have trustworthy and credible solutions. This research represents a step towards that direction.