Research Statement

Liyao Xiang

Department of Electrical and Computer Engineering
University of Toronto

With its successes in a wide variety of applications such as image recognition and natural language processing, machine learning has become immensely popular in recent years, and has been proven to be influential in our modern lives. However, there exists a fundamental conflict between the inherent need of machine learning for resource-intensive computation and the power constraints on mobile devices, and such a conflict has made it difficult to use machine learning on our devices as we become increasingly mobile. Many open questions remain on both mobile and edge devices with limited resources, such as how machine learning is to be applied seamlessly, and how machine learning models are to be trained efficiently.

As critical as these performance challenges, various security and privacy threats are emerging with the deployment of machine learning techniques. Recent studies have found that a trained model can be maliciously manipulated to produce unexpected outputs, and an individual's data can be surprisingly leaked by the model. These threats become even more severe when personal safety situations are involved. Unfortunately, most of the threats stay hidden until losses by attacks are reported. What's worse, effective countermeasures barely keep up with the ever-evolving attacking tools.

These practical challenges need to be addressed with a carefully choreographed mix of engineering talents and fundamental theoretical advances, and open doors to exciting research opportunities that shape the mobile use of machine learning in the future. It is my wish to contribute to this fundamental paradigm shift with my own research experience. I firmly believe that high quality research arises from solving real-world problems that are rooted in fundamental intellectural challenges, and that an exceptional piece of research work should not only provide an in-depth understanding on how problems can be solved on paper, but also have a practical impact that may potentially affect millions of lives.

Current Research

My past and current research are mostly related to the performance and privacy issues regarding machine learning, particularly on mobile devices. We have been enhancing the capability of these resource-constrained mobile devices by harnessing the horsepower of offloading, and have integrated machine learning with crowdsourcing techniques to push the frontier of our ability to solve large-scale problems in the real world. As far as user privacy is concerned, we further designed protocols to enable learning without privacy violation in crowdsourcing frameworks. However, privacy should never be discussed without usability. Following this principle, we have investigated optimal privacy mechanisms for learning models with the least accuracy loss in our most recent work. For most of the problems studied, we not only designed new algorithms with sound theoretical analyses, but also prototyped systems to verify the ideas.

Mobile offloading. Computation-intensive tasks are widely considered to be unfit for less robust mobile devices, while the need for them is exponentially increasing. It is not rare to have more than one battery-draining applications running on a single device. Echoing the idea of offloading computation from mobile devices to the cloud, we have pushed the boundary further by proposing *coalesced offloading* [C2], which exploits the potential for multiple applications to coordinate their offloading requests with the objective of saving additional energy.

The problem of seeking the optimal offloading timing with the least power consumption, as we have proven, is a generalized online ski rental problem. We proposed a new online algorithm and proved its worst-case performance bound. Our emulation results on the iOS platform showed a significant reduction of energy with our offloading strategy.

Mobile crowdsourcing. It is often needed to provide location-based services within the context of an indoor event at a venue without any infrastructure support for localization. Such needs for indoor localization are both contextual and ephemeral, in that location-based services are only needed within the context of and for the duration of the event, rather than permanently. Targeting at this niche, we have proposed a zero-cost indoor localization scheme by taking advantage of the large crowds of event attendees.

We designed, implemented, and evaluated *Tack* [J1], a light-weighted mobile crowdsourcing framework for indoor localization. The key innovation is an unsupervised learning framework that accurately estimates each user's most likely position given their local sensing data and encountering information. What's more intriguing is that, built from the ground up for mobile devices, the core of our learning algorithm is optimized to run efficiently on capability-constrained smartphones. In real-world tests on iOS devices and BLE beacons, *Tack* has achieved an impressive accuracy with light deployment effort and power consumption.

Privacy in crowdsourcing. An unavoidable issue in any crowdsourcing system is that individual privacy can be easily put at risk. Using *Tack* as an example, we wonder whether or not there is a practical way to keep the two-way communication private throughout the entire iterative learning process: the local data sent by users to the server, and the posterior probability update returned by the server. What can be used in substitute of the full homomorphic operations, yet without burdening typical mobile devices too heavily?

We have proposed a generic approach to preserve user privacy in crowdsourcing systems running unsupervised learning and developed an example application [C4]. The highlight of the mechanism is a combination of a partial homomorphic cryptosystem and the use of differential privacy for preserving each user's state with a specific privacy guarantee. Most previous private learning frameworks focus on parameterized models, while our work is among the first to raise the privacy concern on unparameterized models.

Privacy-preserving deep learning. In deep learning systems, differential privacy preserves user privacy by perturbing the model parameters to an extent such that an adversary is not able to infer about a particular individual even given arbitrary side information. However, such protection is achieved with a significant degradation of model utility. In our work, we discovered the cause to be a loose characterization of the model utility with overly strict privacy constraints. Hence we have reconstructed the privacy mechanism with a better understanding of the relation between privacy and utility, and established a useful link between the utility-privacy problem and the distortion-rate problem in communication research. Besides theoretical contributions, we also implemented code optimization in machine learning libraries to ensure the scalability of our algorithm. Experimental results on multiple datasets have shown a remarkable accuracy improvement over the state-of-the-art.

We further broadened our scope to study collaborative learning where a joint model is trained on separate datasets belonging to different parties. As each individual protects its privacy by inserting additional noise, an overwhelming amount of noise would be accumulated for the joint model, which eventually leads to useless learning results. Different from the highly inefficient secure aggregation algorithms proposed by previous works, we have designed a secure noisy voting protocol to collect private results from distributed parties. As the protocol is built within the semi-supervised knowledge transfer framework, the joint model is 'taught' by privately trained local models to produce highly accurate learning results, yet without too much privacy loss.

Future Research

Adopting machine learning on the mobile devices have raised many interesting questions on system security, privacy, and performance. Some problems are specific to a particular area, whereas more of them straddle the border between different fields. I believe these challenges provide a rich set of research topics that require analytical studies and inspire innovative implementations.

Learning with integrity. There is a growing recognition that machine learning exposes new vulnerabilities in software, yet little has been known about their causes and countermeasures. For instance, a testing example can be maliciously distorted without a visually noticeable difference, but leads to misclassification when fed into a trained neural network. Until now, there is hardly an accurate interpretation of this phenomenon, nor any valid detection/defense method against such adversarial examples. Some existing solutions examine the input data distribution or the model itself, but they are far from successful preventing adversarial attacks. Adversaries not only attack machine learning systems at inference time, but may also pollute the training set to force the learner to learn any arbitrary function, resulting in entirely unavailable systems. In spite of its importance, research in this area is at its early stage, and a large number of questions remain unanswered.

Learning with privacy. No individual shall be discouraged from participating in training for fear of revealing its identity or leaking private information. However, current studies show that machine learning models have an amazing capacity to memorize training data. As the trained models are stored on mobile devices for inference, it would be harmful to the training data. Reports show that by extracting from the model, one can recover the training input or infer a particular individual's participation.

Differential privacy presents a randomization framework to analyze and preserve an individual's privacy. The technique has been adopted by corporations such as Google, Apple, etc., to collect usage statistics from hundreds of millions of devices. However, it is unclear how effective the privacy guarantee is against adaptive attacks, i.e., the attacker adaptively adjusts its strategies with each querying answer, or what the optimal protection scheme is against such attacks. Further, learning with privacy is difficult for a lack of knowledge about the model sensitivities; hence we plan to use advanced statistical tools such as influence functions to identify 'vulnerable' input features, and distinguishably treat them in privacy preservation.

Learning with efficiency. While machine learning invites all kinds of new possibilities, its performance still requires close investigation, especially for on-device learning. Recently, Google announced *TensorFlow Lite*, a version of machine learning library that brings low-latency inference for mobile devices. Compared to inference, the training phase has more stringent requirement on resources and is widely considered impractical for mobile devices. However, mobile devices are where most of the raw data are collected, and those data are not expected to leave devices for privacy reasons. While offloading to more powerful infrastructures can be a candidate solution, it is unknown what the best strategy is catering to specific learning algorithms, or to what extent one would protect its privacy at the cost of performance.

Machine learning is a fast-growing area, and its mobile use would bring in many challenges and opportunities. I am excited to build on my understanding to push the knowledge boundary on security, privacy, and performance. Those problems are 'old' as having deep roots in the fundamentals, yet are also new in the context of new architectures and circumstances. Solving these problems requires comprehensive analytical models, highly efficient algorithms, and working system prototypes. I believe the value of solving those research problems is far beyond intellectual rewards — they will have an essential impact on our lives, and I hope my work can be a part in shaping the most exciting mobile applications powered by learning techniques.