Distributed and Federated Learning over IoT networks

Erich Malan

The growing computational capabilities of low-power devices, paired with advances in wireless networking and deep learning, have accelerated the development of Artificial Intelligence of Things. This paradigm enables on-device data processing, enhancing privacy, efficiency, and supporting real-time inference. However, privacy regulations and infrastructure limitations restrict its full potential, since the prevailing paradigm for training deployed deep learning models still relies on centralizing data. Federated Learning (FL) offers a promising avenue enabling collaborative training across distributed nodes without sharing raw data.

However, the deployment of FL in Internet of Things (IoT) environments introduces several critical challenges. Frequent parameter synchronization leads to communication bandwidth bottlenecks, imposing a significant burden on the constrained resources of distributed nodes, which are often battery-powered and subject to limited data plans. In this scenario, achieving competitive accuracy becomes impractical, particularly when facing severe data heterogeneity across devices. Furthermore, the proliferation of sophisticated attacks poses significant threats to privacy and intellectual property, necessitating continuous investigation of undetected vulnerabilities and development of novel countermeasures, which generally come at the cost of reduced efficiency and accuracy. This thesis tackles these challenges through: i) gradient-driven adaptive strategies that optimize resource utilization without degrading model quality; ii) dynamic, budget-aware policies that enhance performance under strict resource budgets; and iii) security approaches that promote or evaluate the confidentiality of proprietary data and models.

It first investigates communication efficiency, providing an automatic layer-freezing strategy that progressively excludes stable model layers from synchronization, reducing exchanged data by up to 83.9% with minimal to no accuracy degradation. Additionally, a novel two-phase resource management policy overcomes the accuracy-communication trade-off of static partial client participation, boosting accuracy by up to 12.2% under extreme data heterogeneity. For energy-constrained scenarios, a gradient-driven technique automatically adjusts participation, attaining up to 2.74 times greater energy efficiency. In the security domain, adapting automatic layer freezing to homomorphic encryption-based FL cuts client computation time

by up to 35.5% and server computation time by 36.4%, while reducing communication overhead by 37.4%. The thesis also introduces a side-channel attack based on the CPU-frequency traces, discovering a novel threat to model confidentiality. Finally, it discusses the evaluation of FL strategies and applications. By revisiting the role of learning rate schedules and advocating for budget-aware configurations, it uncovers new solutions that reshape the advantages of federated algorithms. Finally, it presents a smart meter case study on sociodemographic classification in smart grids, showing that privacy-preserving FL with homomorphic encryption can closely match centralized accuracy.

Overall, this thesis delivers a set of strategies to address the intertwined challenges of efficiency, accuracy, privacy, and security in federated learning for IoT. By adopting a holistic perspective, it advances the trade-off between efficiency and accuracy, identifies emerging security threats and facilitates the adoption of relative countermeasures, and evaluates the potential of cross-device FL across different contexts and operational constraints. The proposed general yet practical solutions foster efficient, privacy-conscious, and scalable IoT shared intelligence, paving the way for the widespread adoption of ubiquitous intelligent services.