Project Proposal: RRNet - Reducing Overhead in Two-party Computation Based Private Inference

Jiahui Zhao

Introduction:
The use of deep learning (DL) has seen significant growth in recent years, however, with this growth comes privacy and security concerns. To address these issues, secure Two-party computation (2PC) has been proposed as a means of enabling privacy-preserving DL computation. In practice, however, 2PC methods often incur high computation and communication overhead, which can impede their use in large-scale systems.

Problem Statement:
To address the challenge of high overhead in 2PC-based private inference, we aim to develop RRNet(ReLU reduced network), a systematic framework for reducing the overhead of MPC comparison protocols and accelerating computation through hardware acceleration.

Proposed Solution:
RRNet is a framework that aims to improve the efficiency, accuracy, and security guarantees of privacy-preserving DL computation. Our approach integrates the hardware latency of cryptographic building blocks into the DNN loss function, resulting in improved energy efficiency. Furthermore, we propose a cryptographic hardware scheduler and corresponding performance model for Field Programmable Gate Arrays (FPGAs) to further enhance the efficiency of our framework.

Challenges and Innovations:
One of the main challenges in implementing RRNet will be balancing the trade-off between computation efficiency and security guarantees. Additionally, incorporating hardware acceleration into the framework will require a deep understanding of the hardware architecture and the ability to effectively model its performance. To overcome these challenges, we will need to conduct extensive research and experimentation to ensure the proper integration of hardware acceleration into the RRNet framework.
While we believe that RRNet has the potential to achieve much higher ReLU reduction performance than current state-of-the-art works on the CIFAR-10 dataset, this will ultimately be confirmed or disproven through experimentation. Regardless, the proposed framework represents a significant advancement in the field of privacy-preserving deep learning, and we are confident that our approach will contribute to the development of more efficient and secure solutions in the future.

Conclusion:
RRNet has the potential to significantly reduce the overhead of MPC-based private inference, making it more practical for use in large-scale systems. With its improved energy efficiency, accuracy, and security guarantees, RRNet has the potential to become the state-of-the-art solution for privacy-preserving DL computation.