ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”
Review for Paper Titled:

“DELPHI: A Cryptographic Inference Service for Neural Networks”

1- Summary

This paper proposes DELPHI- a secure prediction system that allows two parties to execute Neural Network Inference, without revealing either party’ data. Achieving cryptographic prediction for realistic Neural Network entails, Constructing Efficient Subprotocols for evaluating Linear and Non-Linear layer followed by linking the results of these subprotocols with each other. It proposes the design of a hybrid Cryptographic protocol that improves the accuracy-communication complexity involved over previous Cryptographic works such as GAZELLE along the similar lines. It develops a Neural Architecture Search Planner that automatically generates neural network architecture configurations that navigate the Performance-Accuracy Trade-offs of the Hybrid Protocol. Combination of these techniques enables one to achieve a 22x improvement in Online Prediction Latency compared to the similar State-of-the-art works proposed previously.

2- Novelty/Contribution

This paper proposes to reduce the Online cost of computing linear operations by moving the heavy Crptographic operations over LHE ciphertexts to the Preprocessing phase. Since the service provider’s input is available to Linear Layer, Linearly Homomorphic Encryption (LHE) to create secret shares during Pre-Processing. Later when User’s input becomes available online, all linear operations can be performed without Heavy LHE operations at online stage. To reduce cost of Non-Linear Operations, this work selectively replaces ReLu activations with polynomial (specific quadratic) approximations which can be computed securely and efficiently using standard protocols since arbitrary replacements degrade performance employing NAS- Neural Architecture Search Planner which uses Gradient activation clipping and Gradual activation exchange techniques to prevent Gradient Descent Algorithm from Diverging in addition to Efficiently optimizing and Prioritizing configurations using scores computed by intuitive Rubrics. It incorporates Population Based Training (PBT) Algorithm to select the configuration based on the Accuracy- Communication complexity tradeoff on the Fly.

3- Evaluation

The Proposed solutions are evaluated over the CIFAR 10 dataset utilizing MiniONN architecture and over the CIFAR 100 dataset utilizing the ResNet-32 architecture. We observe that for most efficient networks output by the planner, DELPHI requires 22-100x less time to execute its online phase, and 9-40x less communication. Furthermore, as the number of approximate activations increases, the gap between DELPHI and prior work GAZELLE.

4- Unique Strength/Weakness

This development gives the Users the flexibility to enjoy low communication complexity while still ensuring the accuracy remains very close to the queries threshold. The use of NAS plays and the other subprotocols involved play a key role in achieving this.