ECE209AS - Winter 2022
UCLA

“Hariram, Veeramani”
Review for Paper Titled:

“Low Latency Privacy Preserving Inference”

1- Summary

This paper addresses the limitations faced by the Privacy Preserving Machine Learning solutions such as Homomorphic Encryption which can handle neural networks with only limited width and depth and exhibit high latency even for relatively simple practical networks.
This paper proposes two solutions to address these limitations in the form of LoLa. First it proposes novel methods of Data Representation during the computation leading to 10x improvement in latency and enables inference on wider networks compared to prior attempts with the same level of security. In the second solution, the method of transfer learning is applied to provide privacy preserving inference services using deep networks with latency of fraction of seconds.
It considers the previous work done by CryptoNets and addresses three of its limitations. The first is latency to process a single prediction request. Second is the width limitation because CrptoNet encodes each node in the network as a separate message, which can lead to exponential complexity when applied to practical wider neural networks. The Third is the depth of the Neural Network that it can handle. Each layer in the network adds more sequential multiplication operations which result in increased noise levels and message size growth.

2- Novelty/Contribution

This work preprocesses data to form “Deeper Features” that have higher semantic meaning. These deep features are encrypted and sent to the service provider for private evaluation. In forming deeper features, it proposes following different Data Representations such as Sparse, Stacked, Interleaved, Convolution representations in addition to the Dense and SIMD representations used by CryptoNet. The key idea behind these representations is when we encode input vectors as a single message, we can implement dot product between these vectors, whose sizes are powers of 2, by applying point-wise multiplication between these vectors and a series of log(d) rotations and additions between each such rotation. Thus, in total, the operation requires log(d) rotations, additions and a single multiplication by using only a single message leading to reduced latency while still preserving Privacy.

3- Evaluation

LoLa evaluates the proposed solutions by using MNIST and CIFAR and compares its prediction performance metrics with CryptoNets and other similar networks by combinations of different data representations at different layers to optimize latency. The Performance gains observed are in the order of 10x.

4- Unique Strength/Weakness

This paper analyzes the limitations of the SIMD data representation used by CryptoNets and proposes several other data representations which can lead to reduced latency of computation and inference, by considering the complexity involved via dot products of complex layers in practical deep Neural Networks.