End-to-End Acceleration of Homomorphic Encrypted CNN Inference on FPGAs

Tian Ye, *University of Southern California*
Rajgopal Kannan, *US Army Research Lab*
Viktor K. Prasanna, *University of Southern California*
Contact: tye69227@usc.edu

Homomorphic Encryption is a promising approach to perform secure inference on Machine Learning models such as CNNs by allowing cloud servers to perform computations on encrypted data directly. However, CNN inference over encrypted images has high computational complexity. Prior works propose accelerators for individual HE primitives on FPGAs. In this work, we focus on an integrated design for end-to-end acceleration of inference on encrypted data. We develop parameterized IP cores for HE primitives and CNN layers. To understand the tradeoffs between various parameters such as hardware resources and performance and optimize the overall performance, we develop a parameterized performance model to evaluate the resource consumption and latency of the complete design. The performance model allows design space exploration to identify the optimal architectural parameters of the accelerator for a given FPGA, CNN model, and security requirements. We implement our design on a Xilinx VU13P FPGA and compare its performance with software implementation on a state-of-the-art server with a multi-core CPU. Our implementation for a widely studied 8-layer CNN inference for a batch of 8K images achieves average inference time of 38.8 ms per image, which is 4.1× improvement over the software baseline on the state-of-the-art server.

Keywords: Homomorphic Encryption; Convolutional Neural Networks; FPGA

DOI: http://dx.doi.org/10.1145/3490422.3502346