Accelerating Privacy Preserving Deep Learning for Real-time Secure Applications

Hardware Acceleration of Homomorphic Encrypted Convolutional Neural Networks

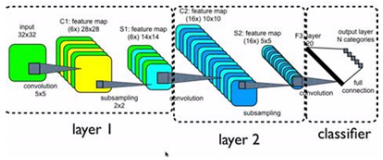
Challenge:

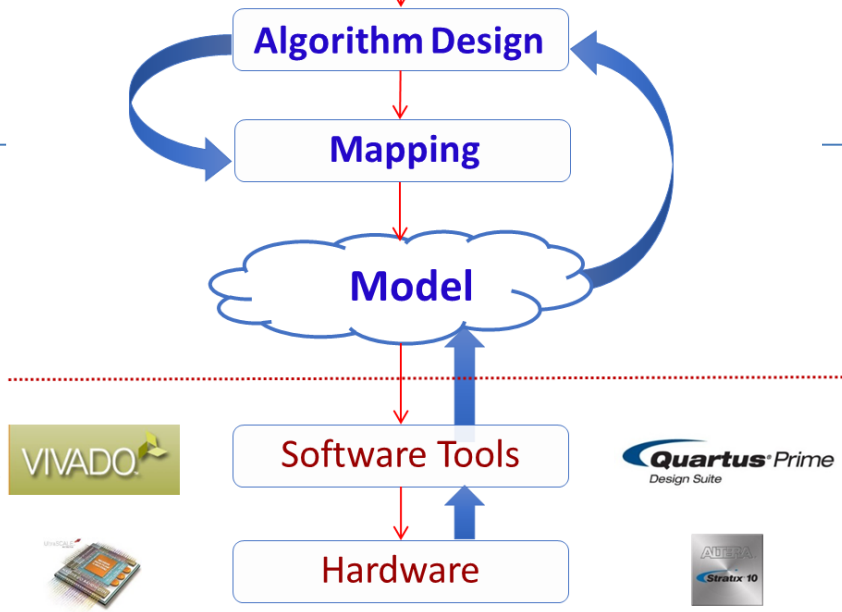
- Need for CNN computations using HE to enable end to end private inference
- HE computation orders of magnitude (1000x) slower than plaintext computation
- Key computational kernels
  - Large Modular Arithmetic
  - Number Theoretic Transforms (NTT)

Solution:

- FPGA Accelerator for HE-CNN
- Key Innovations
  - Performance Model Driven Accelerator Design
  - Low Latency NTT Cores
  - End to End HE-CNN Accelerator
  - End to End sparse HE-CNN accelerator

HE Based Image Convolution





Scientific Impact:

- Enable users to leverage powerful cloud hosted CNN models to perform inference tasks without sacrificing privacy guarantees
- Enable data aggregators to perform valuable analytics without violating user privacy

Broader Impact and Broader Participation:

- Enable the right of privacy of citizens without sacrificing their ability to benefit from technological advancements
- Transition to Practice: Private ML-as-a-Service
- Train next generation data scientists to treat privacy as first order requirement

SaTC: CORE: Small: Accelerating Privacy Preserving Deep Learning for Real-time Secure Applications, #2104264, PI: Viktor K. Prasanna, Co-PI: Sanmukh Kuppannagari, University of Southern California

Project Website and Publications: https://sites.usc.edu/fpga/secure/

Contact: {prasanna, kuppanna}@usc.edu