University of Pittsburgh | School of Computing and Information

Dissertation Defense

Doctor of Philosophy in Computer Science

"**Secure Accelerator Design for Deep Neural Networks**" by **Lei Zhao**

Date: April 8th, 2022
Time: 3:45PM – 5:15PM
Place: https://pitt.co1.qualtrics.com/jfe/form/SV_8ps89C5IgdAe1Jc

Committee:

- Youtao Zhang, Associate Professor, Department of Computer Science, School of Computing and Information
- Adriana Kovashka, Assistant Professor, Department of Computer Science, School of Computing and Information
- Xulong Tang, Assistant Professor, Department of Computer Science, School of Computing and Information
- Jun Yang, Associate Professor, Department of Electrical and Computer Engineering, Swanson School of Engineering

Abstract:
Deep neural networks (DNNs) have recently gained popularity in a wide range of modern application domains due to its superior inference accuracy. With growing problem size and complexity, modern DNNs, e.g., CNNs (convolutional neural networks), contain a large number of weights, which require tremendous efforts not only to prepare representative training datasets but also to train the network. There is an increasing demand to protect the DNN weight matrices, an emerging intellectual property (IP) in DNN field. This thesis proposes a line of solutions to protecting the DNN weights deployed on domain specific accelerators. Firstly, I propose AEP, a DNN weights protection scheme for accelerators based on conventional CMOS-based technologies. Because of the extremely high memory bandwidth demand in DNN accelerators, conventional encryption-based approaches, which require the integration of expensive encryption engines, pose significant overheads on the execution latency and energy consumption. Instead, AEP enables effective IP protection by exploring hardware fingerprints to eliminate the need of encryption. Adopting such hardware fingerprints achieves high inference accuracy only on the authorized device, while unauthorized devices cannot produce any usable results from the same set of weights. Secondly, with the growing size of DNN models, the large amount of intermediate results (i.e, the output from the previous layer and the input to the next layer) cannot be hold on-chip. These intermediate results also contain sensitive information of the DNN model. In this part I propose SCA, a full DNN protection scheme that protect both the model weights and the intermediate results on CMOS-based accelerators. Thirdly, ReRAM-based accelerators introduce new challenges on DNN's security issue due to its crossbar structure and non-volatility. ReRAM's non-volatility retains data even after the system is powered off, making the stored DNN model vulnerable to attacks by simply reading out the ReRAM content. Because the crossbar structure can only compute on plaintext data, encrypting the ReRAM content is no longer a feasible solution in this scenario. To solve these issues, I propose SRA to store DNN weights on crossbars in a novel encrypted format while still maintaining ReRAM's in-memory computing capability. Lastly, although SRA provides security guarantees, the long bit streams in Stochastic Computing induce a large storage overhead. However, conventional model-reducing methods such as pruning and quantization are less applicable in the area of ReRAM-based DNN accelerators. In this part, I propose BFlip -- a novel model size and computation reduction technique -- to share crossbars among multiple bit matrices. BFlip not only reduces storage overhead, but also improves performance and energy consumption.