Title: Compact and Fast Machine Learning Accelerator for IoT devices

Abstract:

The Internet of things (IoT) is the networked interconnection of every object, which is equipped with intelligence and cooperates with other smart objects to improve efficiency and economic benefit. A typical IoT system can sense data from real-world, use this information to reason the environment and then perform desired actions. The intelligence of the systems comes from appropriate actions by reasoning the environmental data, which is mainly based on machine learning algorithms. To have a real-time response to the dynamic ambient change, a compact and fast machine learning accelerator is preferred since uploading to clouds suffers long latency of computation in the back end. As such, machine learning algorithms have to be optimized to utilize the computational resource limited IoT devices.

In this thesis, three main works have been investigated for machine learning on IoT devices. Firstly, a tensor-train based neural network is proposed for compact machine learning. Such neural network can compress the network with fewer parameters and potentially speed-up the neural network processing. An alternating least-squares method is also used for such neural network training. Secondly, I have proposed incremental least-squares algorithm, which can reuse previous results for faster solution. Such least-squares based algorithm is frequently used during machine training process to accelerate training process. Lastly, the optimized machine learning algorithm has been applied to many applications such as network intrusion detection system (NIDS), load forecasting for energy management and indoor positioning. By optimizing the training algorithm using incremental least-squares, fast machine learning process on IoT devices can be achieved, which means the system can quickly adapt to the environmental change. In addition, a new computing architecture on CMOS FPGA has been developed and verified. A 3D CMOS-RRAM based computing architecture has been proposed to demonstrate the speed-up of machine learning process and improvement of energy efficiency.