Dissertation Defense

Babak Zamirai

Customized Systems of Deep Neural Networks for Energy Efficiency

Thursday, August 6, 2020
2:30 – 4:30 pm
https://bluejeans.com/548385457

ABSTRACT: The size and complexity growth of deep neural networks (DNNs) is driven by the push for higher accuracies and wider ranges of functionality. However, DNN computation requirements are fast outpacing the growth of computer hardware on both mobile devices and cloud servers. Hence, research to increase performance and energy efficiency of DNNs is important. This research can be divided into two major categories: input-invariant and input-dependent. Input-invariant methods leverage hardware and software techniques, such as pruning, to accelerate a single DNN for the entire dataset. Conversely, input-dependent approaches recognize that many inputs do not require the entire computational power of the model and dynamically adjust the complexity of the DNN to suit the input difficulty.

Both DNN pruning and intelligent combination of DNNs require machine learning expertise and manual design, which makes them hard to implement and deploy. This thesis proposes techniques to improve performance and applicability of both input-invariant and input-dependent methods. First, it introduces a new category of input-dependent solutions to maximize energy efficiency and minimize latency of DNNs by customizing systems of DNNs based on input variations. Instead of conventional DNN ensembles, it proposes a data heterogeneous multi-NN system to divide the data space into subsets with one specialized learner for each subset. In addition, an intelligent hybrid server-edge deep learning system is introduced to dynamically distribute DNN computation between the cloud and edge device based on the input data and environmental conditions. Furthermore, it suggests an input-driven synergistic deep learning system, which dynamically distributes DNN computation between a more accurate big and a less accurate little DNN. Lastly, it introduces a noniterative double-shot pruning method, which takes advantage of both architectural features and weight values to improve the simplicity and applicability of pruning.

Chairs: Prof. Scott Mahlke