Hardware-Aware Machine Learning: Modeling and Optimization

Diana Marculescu[1,*]

[1]Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213 USA

*Email: dianam@cmu.edu

Abstract: Machine learning (ML), and in particular (deep) neural networks (NNs) have become widely spread for applications ranging from image classification and object detection, to processing multi-modal and heterogeneous datasets. While the holy grail for judging the quality of a ML model has so far been serving accuracy, and only recently its resource usage, neither of these metrics translate directly to energy efficiency, runtime, or mobile device battery lifetime. This talk will uncover the need for building accurate, platform-specific power and runtime models for NNs, thus allowing machine learners and hardware designers to identify not just the best accuracy NN configuration, but also those that satisfy given hardware constraints. Our proposed modeling framework is applicable to both high-end and mobile platforms and achieves an accuracy of 88.24% for runtime, 88.34% for power, and 97.21% for energy prediction. We show that power consumption and runtime can be treated as constraints for NN model selection in a design exploration framework based on meta-learning that achieves the serving error of a constraint-unaware method up to 30× faster, while never considering invalid configurations.

References:

1. D. Stamoulis, E. Cai, D. C. Juan, and D. Marculescu, "HyperPower: Power- and Memory-Constrained Hyper-Parameter Optimization for Neural Networks," in Proc. IEEE/ACM Design, Automation, and Test in Europe Conference (DATE), Dresden, Germany, March 2018. https://arxiv.org/abs/1712.02446
2. E. Cai, D.-C. Juan, D. Stamoulis, and D. Marculescu, "NeuralPower: Predict and Deploy Energy-Efficient Convolutional Neural Networks," in Proc. Asian Conference on Machine Learning (ACML), Seoul, South Korea, Nov. 2017. https://arxiv.org/abs/1710.05420