Learning-based Power and Runtime Modeling for Convolutional Neural Networks

Ermao Cai ERMAO@CMU.EDU
Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA

Da-Cheng Juan DACHENG@GOOGLE.COM
Google Research, Mountain View, CA, USA

Dimitrios Stamoulis DSTAMOUL@ANDREW.CMU.EDU
Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA

Diana Marculescu DIANAM@CMU.EDU
Department of ECE, Carnegie Mellon University, Pittsburgh, PA, USA

Abstract

"How much energy is consumed for an inference made by a convolutional neural network (CNN)?" With the increased popularity of CNNs deployed on the wide-spectrum of platforms (from mobile devices to workstations), the answer to this question has drawn significant attention. From lengthening battery life of mobile devices to reducing the energy bill of a datacenter, it is important to understand the energy efficiency of CNNs during serving for making an inference, before actually training the model. In this work, we propose *NeuralPower*: a layer-wise predictive framework based on sparse polynomial regression, for predicting the serving energy consumption of a CNN deployed on any GPU platform. Given the architecture of a CNN, *NeuralPower* provides an accurate prediction and breakdown for power and runtime across all layers in the whole network, helping machine learners quickly identify the power, runtime, or energy bottlenecks. We also propose the *"energy-precision ratio" (EPR)* metric to guide machine learners in selecting an energy-efficient CNN architecture that better trades off the energy consumption and prediction accuracy. The experimental results show that the prediction accuracy of the proposed *NeuralPower* outperforms the best published model to date, yielding an improvement in accuracy of up to 68.5%. We also assess the accuracy of predictions at the network level, by predicting the runtime, power, and energy of state-of-the-art CNN architectures, achieving an average accuracy of 88.24% in runtime, 88.34% in power, and 97.21% in energy. We comprehensively corroborate the effectiveness of *NeuralPower* as a powerful framework for machine learners by testing it on different GPU platforms and Deep Learning software tools.