RENSSELAER POLYTECHNIC INSTITUTE

DEPARTMENT OF MATHEMATICAL SCIENCES COLLOQUIUM

Quantization of Deep Neural Networks

Abstract: Quantized neural networks are attractive due to their inference efficiency. However, training quantized neural networks involves minimizing a discontinuous and piecewise constant loss function. Such a loss function has zero gradient almost everywhere (a.e.) with respect to weight variables, which makes the conventional back-propagation inapplicable. To this end, we study a class of biased first-order oracle, termed coarse gradient, for overcoming the vanishing gradient issue in the regression and classification settings, respectively. In addition, we propose an efficient computational method featuring a blended coarse gradient step, for network quantization, which achieves the state-of-the-art accuracies in image classification without extensive tuning of hyper-parameters.

Penghang Yin (University at Albany)

Monday, February 24, 2020

Amos Eaton 216

Refreshments served 3:30-4pm Amos Eaton 4th Floor Lounge