UNIVERSITY AT ALBANY
State University of New York

College of Arts and Sciences
Department of Mathematics and Statistics

Colloquium

Penghang Yin
UAlbany

A Simple Approach for Quantizing Large-scale Neural Networks

Friday, September 27, 2024
3:00 p.m. in Massry BB-012
(tea & coffee at 2:45 p.m.)

Abstract. Quantized neural networks offer significant benefits in memory efficiency and power consumption. In this talk, we introduce a simple yet highly effective algorithm for compressing large-scale neural networks, which requires only a small batch of data for calibration. To further improve accuracy, we propose a weight preprocessing technique based on infinity-norm regularization. Our method can quantize large language models (LLMs) with up to 70 billion parameters in just a few hours on a single GPU, achieving more than a 4x reduction in model complexity with minimal accuracy loss. Experimental results demonstrate that our approach achieves state-of-the-art performance on tasks such as image classification and natural language processing.