香港浸會大學理學院
HKBU Faculty of Science

DEPARTMENT OF COMPUTER SCIENCE

PhD Degree Oral Presentation

PhD Candidate:	Ms. LAN Weichao
Date	9 May 2024 (Thursday)
Time:	3:00 pm – 5:00 pm (35 mins presentation and 15 mins Q & A)
Venue:	ZOOM (Meeting ID: 955 1661 9015) (The password and direct link will only be provided to registrants)
Registration:	https://bit.ly/bucs-reg (Deadline: 12:00 nn, 8 May 2024)

Compression and Acceleration for Deep Convolutional Neural Networks

Abstract

Convolutional Neural Networks (CNNs) have been successfully applied to solve many real-life problems and paid more and more attention in recent years. However, the incredibly huge memory and computation cost of deep CNNs poses a great challenge of deploying them on memory-constrained edge devices. Due to these limitations, the concept of network compression and acceleration was naturally proposed and widely used for reducing memory and computation consumption. Therefore, this thesis presents four methods from different perspectives to achieve compression and acceleration for CNNs.

The first method is proposed to compress the binarized filters in original binary neural networks, reducing the number of parameters by employing stacked low-dimensional filters. The second approach involves generating the convolution filters through the linear combination of a set of learnable quantized filter bases. The third method aims at directly constructing compact neural networks by stacking designed basic units, namely BUnit-Net. In the last approach, we propose two techniques, namely Label Revision and Data Selection, to improve the performance of knowledge distillation for image classification tasks, which is also a popular method to train smaller models. We conduct various experiments to verify the performance of the four proposed methods, and the results demonstrate their effectiveness on compression and acceleration.

***** ALL INTERESTED ARE WELCOME *****