Department of Computer Science

COMPUTER SCIENCE COLLOQUIUM

Optimizing Neural Network Design: Network Compression, Quantization and Pruning

SPEAKER **Dr Sungjoo YOO**

Associate Professor
Computing and Memory Architecture Lab
Department of Computer Science and Engineering
Seoul National University
Korea

DATE 12 May 2017 (Friday)
TIME 3:00 pm - 4:00 pm
VENUE CS Seminar Room, Y6405, 6th Floor
Yellow Zone, Academic 1
City University of Hong Kong
83 Tat Chee Avenue
Kowloon Tong

ABSTRACT

Redundancy in neural networks offers opportunities of design optimization. In this talk, we will introduce our recent works on three optimization techniques. Firstly, we will report our case study of applying low rank approximation technique namely Tucker decomposition to CNNs running on the smartphone. Secondly, we will explain a quantization method based on weighted entropy which makes ResNet-101 run at 6 bit weight and activation without accuracy loss. Lastly, we will introduce a zero-aware hardware accelerator called ZeNA and a novel pruning method for activations.

BIOGRAPHY

Sungjoo Yoo received his PhD at Seoul National University, Korea in 2000. He was a researcher at TIMA lab, France during 2000-2004 and a principal engineer at Samsung Electronics during 2004-2008. He was an assistant and associate professor at POSTECH, Korea during 2008-2015. He joined SNU in 2015 and he is an associate professor currently. His research interests includes optimizations of neural network from algorithm to chip implementation.

All are welcome!

In case of questions, please contact Dr XUE Chun Jason at Tel: 3442 9815, E-mail: jasonxue@cityu.edu.hk, or visit the CS Departmental Seminar Web at http://www.cs.cityu.edu.hk/news/seminars/seminars.html.