NANYANG TECHNOLOGICAL UNIVERSITY

SEMINAR ANNOUNCEMENT

Net-Trim: A Layer-wise Convex Pruning of Deep Neural Networks

Dr Alireza Aghasi

Department of Mathematical Sciences

IBM TJ Watson Research Center

Date: **14 February 2017 (Tuesday)**

Time: **10.30am – 11.30am**

Venue: **MAS Executive Classroom 2, MAS-03-07**
School of Physical and Mathematical Sciences

**The Seminar will be done via Skype*

Abstract

Model reduction is a highly desirable process for deep neural networks. While large networks are theoretically capable of learning arbitrarily complex models, overfitting and model redundancy negatively affects the prediction accuracy and model variance. Net-Trim is a layer-wise convex framework to prune (sparsify) deep neural networks. The method is applicable to neural networks operating with the rectified linear unit (ReLU) as the nonlinear activation. The basic idea is to retrain the network layer by layer keeping the layer inputs and outputs close to the originally trained model, while seeking a sparse transform matrix. We present both the parallel and cascade versions of the algorithm. While the former enjoys computational distributability, the latter is capable of achieving simpler models. In both cases, we mathematically show a consistency between the retrained model and the initial trained network. We also derive the general sufficient conditions for the recovery of a sparse transform matrix. In the case of standard Gaussian training samples of dimension N being fed to a layer, and s being the maximum number of nonzero terms across all columns of the transform matrix, we show that O(s.logN) samples are enough to accurately learn the layer model.

Speaker Biography

Alireza Aghasi received his B.Sc. degree from Isfahan University of Technology in 2002, and the M.Sc. degree from Amirkabir University of Technology (Tehran Polytechnic) in 2006, both in electrical engineering. For more than three years between 2002 and 2007, he worked as a technical engineer in industry, mainly focusing on system and mobile network optimization. In 2008, he joined Tufts University and started his Ph.D. research on parametric shape-based techniques for inverse problems and image processing. In 2012, he joined the compressed sensing group at Georgia Tech as a postdoctoral associate. He also completed a professional Master's degree in operations research at the same institute. Between 2015 and 2016, he worked as a postdoctoral associate at the Massachusetts Institute of Technology, where he worked with the computational imaging group. In Spring 2016 he joined IBM TJ Watson Research Center, where he now continues his work with the department of Mathematical Sciences. His research mainly focuses on optimization theory, statistics and probability theory with applications in inverse problems, machine learning and data science.

Host: Division of Mathematical Sciences, School of Physical and Mathematical Sciences