James Tin-Yau KWOK
Professor,
Department of Computer Science and
Engineering,
Hong Kong University of Science and Technology,
Hong Kong SAR

Topic: **Compressing deep networks**

Abstract:

Deep networks are powerful but computationally expensive. This hinders deployment to small computing devices such as cell phones and the internet of things. To alleviate this problem, a popular approach is to quantize the deep network weights to a small number of bits. In this talk, we consider a number of issues related to network quantization. First, we show that the effect of quantization on the loss can be directly formulated into the optimization problem. Morevoer, the exploding gradient problem can become more severe in training quantized LSTMs. By using the popularly used weight/layer/batch normalization, we show theoretically and empirically that the gradient magnitude can be stabilized. Besides, communication overhead is a major bottleneck in distributed deep network training. To improve communication efficiency, besides using weight quantization, we propose a general distributed compressed SGD scheme which compresses the gradients both to and from workers. With these techniques, empirical results show that the resultant network can significantly reduce its size without sacrificing performance, and runs much faster in a distributed environment.

Bio:

Prof. Kwok is a Professor in the Department of Computer Science and Engineering, Hong Kong University of Science and Technology. He received his B.Sc. degree in Electrical and Electronic Engineering from the University of Hong Kong and his Ph.D. degree in computer science from the Hong Kong University of Science and Technology. Prof. Kwok served/is serving as an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems, Neural Networks, Neurocomputing and the International Journal of Data Science and Analytics. He has also served as Program Co-chair of a number of international conferences, and as Area Chairs in major machine learning and AI conferences. He is an IEEE Fellow.