For favour of Posting

THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

Department of Mathematics

SEMINAR ON DATA SCIENCE

Compression and Acceleration of Pre-trained Language Models

By

Dr. Lu HOU

Huawei Noah's Ark Lab

Abstract

Recently, pre-trained language models based on the Transformer structure like BERT and RoBERTa have achieved remarkable results on various natural language processing tasks and even some computer vision tasks. However, these models have many parameters, hindering their deployment on edge devices with limited storage. In this talk, I will first introduce some basics about pre-trained language modeling and our proposed pre-trained language model NEZHA. Then I will elaborate on how we alleviate the concerns in various deployment scenarios during the inference and training period. Specifically, compression and acceleration methods using knowledge distillation, dynamic networks, and network quantization will be discussed. Finally, I will also discuss some recent progress about training deep networks on edge through quantization.

Biography:
Dr. Lu HOU is a researcher at the Speech and Semantics Lab in Huawei Noah's Ark Lab. She obtained Ph.D. from Hong Kong University of Science and Technology in 2019, under the supervision of Prof. James T. Kwok. Her current research interests include compression and acceleration of deep neural networks, natural language processing, and deep learning optimization.

Date : 28 October 2020 (Wednesday)
Time : 3:00pm – 4:20pm
Zoom Meeting : https://hkust.zoom.us/j/98248767613 (Passcode: math6380p)

All are Welcome!