Department of Data Science
香港城市大學
City University of Hong Kong

DS SEMINAR

Hyper-Compression: Neural Network Compression via Hyper-function

Date: 30 September 2024 (Monday)

Time: 3:30pm - 4:30pm

Venue: Rm G7603, Yeung Kin Man Academic Building, City University of Hong Kong

ABSTRACT

Recently, the escalating demand for memory and computational resources by large models has presented formidable challenges for their deployment in resource-constrained environments. One prevalent way to serve large models is to develop effective model compression approaches that crop the size of large models while maintaining acceptable performance levels. Currently, the landscape of model compression methodologies predominantly revolves around four fundamental algorithms: pruning, quantization, knowledge distillation, and low-rank approximation whose basic frameworks have been established years ago. However, the compression efficacy of these methods is often capped or challenging to scale based on theoretical analysis. In this talk, we introduce a novel and general-purpose approach, referred to as hyper-compression, that redefines model compression as a problem of parameter representation. Specifically, we extend the concept of hypernets into what we term a 'hyperfunction'. Then, the hyperfunction is designed based on ergodic theory (ET). This advanced formulation leads to a performant algorithm that offers several distinct advantages, succinctly summarized as PNAS: 1) Preferable compression ratio; 2) No post-hoc retraining; 3) Affordable inference time; and 4) Short compression time. Lastly, in light of the observed stagnation in hardware Moore's Law, we conjecture "Moore's Law of Model Compression", i.e., the efficiency of model compression could double annually in the near future to meet the needs of large model era. We believe that model compression based on hyperfunction can play an important role in "Moore's Law of Model Compression".

Dr. Fenglei FAN

GUEST SPEAKER'S PROFILE

Dr. Fenglei FAN is currently a Research Assistant Professor in the Department of Mathematics at the Chinese University of Hong Kong. His research focuses on mathematical AI. He obtained his Ph.D. from Rensselaer Polytechnic Institute in the United States in 2021. He then conducted one-year postdoctoral research at Cornell University. His doctoral dissertation won the 2021 Outstanding Doctoral Dissertation Award by the International Neural Network Society (INNS). His representative work was selected as one of few 2024 CVPR Best Paper Award candidates (26 out of 1W+ submissions), and won IEEE TRPMS Best Paper Award. As an RAP, he leads one Huawei Gifted Fund and one key research project from Huawei.

Enquiries: ds.go@cityu.edu.hk

All are welcome