THE UNIVERSITY OF HONG KONG
DEPARTMENT OF COMPUTER SCIENCE

CS Seminar

Taming the Cost of Deep Neural Models: Hybrid Models to the Rescue?

Speaker:
Prof. Laks V.S. Lakshmanan,
The University of British Columbia

Date: March 5, 2024 (Tue)
Time: 3:00 - 4:00pm (HKT)
Venue: Room 328, Chow Yei Ching Building, HKU

Abstract:
Deep learning, and in particular, large language models have made great strides in many fields including vision, language, and medicine. The impressive performance of large models comes at a significant price: the models tend to be billions to trillions of parameters in size, are expensive to train, have a huge operational cost, and typically need cloud service for deployment. Meanwhile, considerable research efforts have been made to design smaller/cheaper models, at the price of restricted generalizability and performance. Not all queries we may wish to pose to a model are hard. Some queries can be answered nearly as accurately with cheaper models at a fraction of the cost of the larger models. However, the performance of cheaper models may suffer on other queries. Can we combine the best of both worlds by striking a balance between cost and performance? In this talk, I will describe two settings in which our group has tackled this issue.

In the first setting, we are interested in approximate answers to queries over model predictions. We show how, under some assumptions about the cheap model, queries can be answered with a provably high precision or recall by using a judicious combination of invoking the large model on data samples and the cheap model on data objects. In the second setting, we are interested in learning a router, which, given a query, predicts its level of hardness, based on which the query is either routed to the small model or to the large model. For both settings, we will present results of extensive experiments that we conducted, showing the effectiveness and efficiency of our approach.

Biography:
Laks V.S. Lakshmanan is a professor of Computer science at UBC, Vancouver, Canada. His research interests span a wide spectrum of topics in data management, integration, cleaning, and warehousing; data mining; semi-structured and unstructured data; big graphs, social networks and social media; NLP; and efficient deep learning. He is an ACM Distinguished Scientist and has won several awards including best paper awards and distinguished reviewer awards. He has served on most top conferences and journals in his areas of research, on program committees, as senior PC member, meta-reviewer, general chair, and as associate editor.

All are welcome.
For enquiries, please email enquiry@cs.hku.hk or call 3917-2180.