In Favour of Posting

THE CHINESE UNIVERSITY OF HONG KONG

Department of Information Engineering

Seminar

Sustainable NLP

By

Prof. Aruna Balasubramanian
Stony Brook University, USA

Date	**:**	**10 April 2024 (Wednesday)**
Time	**:**	**2:00pm – 3:00pm**
Venue	**:**	**Rm 801, Ho Sin Hang Engineering Building, CUHK**

Abstract

Natural language processing (NLP) technology has supercharged many real-world applications ranging from intelligent personal assistants (like Alexa, Siri, and Google Assistant) to commercial search engines such as Google and Bing. But current NLP applications use extremely large neural models, making them (i) expensive to deploy on servers, requiring large amounts of compute resources and power, and (ii) impossible to run on mobile devices, making on-device, privacy-preserving applications impractical.

In the first part of the talk, I will describe systems optimizations we have developed that significantly reduce the compute and memory requirement of NLP models. The optimizations we developed can be applied broadly and results in over 10x reduction in latency when deployed on mobile devices. In the second part of the talk, I will describe our recent work on predicting energy consumption of NLP models. Existing energy prediction approaches are not accurate, making it difficult for developers and practitioners to reason about their models in terms of power. We use a multi-level regression approach that produces highly accurate and interpretable energy predictions.

Biography

Aruna Balasubramanian is a visiting Professor at SUNY Korea, She is also an Associate Professor at Stony Brook University. She received her Ph.D from the University of Massachusetts Amherst, and then was a Computing Innovations Fellow at the University of Washington. She works in the area of networked systems. Her current work consists of (1) improving QoE and equitable access of Internet applications, (2) improving the usability, accessibility, and privacy of mobile systems, and (3) sustainable NLP. She is the recipient of a test-of-time award, the SIGMOBILE Rockstar award, a Ubicomp best paper award, a VMWare Early Career award, and a Google research award, and the Applied Networking Research Prize. She is passionate about improving the diversity in Computer Science and leads the diversity committee at Stony Brook, is the faculty advisor for the WiCS and WPhD groups at Stony Brook, and is an active member of the N2Women group.

**** ALL ARE WELCOME ****

Host: Prof. XING Guoliang (Tel: 3943-8474, Email: glxing@ie.cuhk.edu.hk)
Enquiries: Information Engineering Dept., CUHK (Tel.: 3943-8385)