THE UNIVERSITY OF CHICAGO | THE COMMITTEE ON COMPUTATIONAL AND APPLIED MATHEMATICS

THE UNIVERSITY OF CHICAGO
DATA SCIENCE INSTITUTE

JOINT CAM COLLOQUIUM/DSI DISTINGUISHED SPEAKER SERIES

INDERJIT S. DHILLON

Center for Big Data Analytics, Department of Computer Science,
University of Texas at Austin

MatFormer: Nested Transformer for Elastic Inference

FRIDAY, April 5th, at 12:30 PM
Jones 303, 5747 S. Ellis Ave. Chicago, IL 60637

ABSTRACT

Large Language Models are powered by the underlying transformer deep learning architecture. These models are deployed in a wide range of settings, from multiaccelerator clusters to standalone mobile phones. The diverse memory and computational constraints in these scenarios necessitate practitioners to train models of various sizes to cater to each constraint. Training each of these models is computationally expensive, while also requiring maintaining each of them separately. Moreover, this limits more fine-grained control over relevant tradeoffs, including latency, cost, and accuracy. In this talk, I will introduce our recent work in this direction, "MatFormer", which stands for ``Matroyshka" Transformers. This transformer architecture encapsulates information in a nested manner, facilitating the extraction of subnetworks tailored to specific constraints. I will discuss the training methodology, the advantages over existing methods, and results across different model classes (decoders & encoders), modalities (language & vision), and scales (up to 2.6 billion parameters). Finally, I will outline our future directions to enhance efficiency and scalability for the training and deployment of such large models.

Organizers:
Jeremy Hoskins, Department of Statistics (CAMI), jeremyhoskins@statistics.uchicago.edu & Yuehaw Khoo, Department of Statistics (CAMI), ykhoo@galton.uchicago.edu
CAM Colloquium URL: https://cam.uchicago.edu/events/cam-colloquium/

If you wish to subscribe to our email list, please visit the following website:
https://lists.uchicago.edu/web/subscribe/cam_colloquium/.