


Revisiting the Economics of Large Language Models with Neural Scaling Laws and Dynamic Sparsity

Neural Scaling Law informally states that increased model size and data automatically improve AI. However, we have reached a point where the growth has reached a tipping end where the cost and energy associated with AI are becoming prohibitive.

This talk will demonstrate the algorithmic progress that can exponentially reduce the compute and memory cost of training and inference using "dynamic sparsity" with neural networks. Dynamic sparsity, unlike static sparsity, aligns with Neural Scaling Laws and does not reduce the power of neural networks while reducing the number of FLOPS required by neural models by 99% or more. We will show how data structures, particularly randomized hash tables, can be used to design an efficient "associative memory" that reduces the number of multiplications associated with the training of the neural networks. Current implementations of this idea challenge the common knowledge prevailing in the community that specialized processors like GPUs are significantly superior to CPUs for training large neural networks. The resulting algorithm is orders of magnitude cheaper and energy-efficient. Our careful implementations can train billions of parameter recommendations and Language models on commodity desktop CPUs significantly faster than top-of-the-line TensorFlow alternatives on the most potent A100 GPU clusters, with the same or better accuracies. **We will show some demos, including how to train and fine-tune (with rhfl) a billion-parameter language model on a laptop from scratch for search, discovery, and summarization.**



Anshumali Shrivastava,
Assoc. Professor, CSE
Rice University, USA

Venue: Zoom (Online)
Time: 8:30 PM, IST
Date: 17th Aug'23

Organized by

Mehta Family School of Data Science & Artificial Intelligence

Indian Institute
of Technology Guwahati

About the Speaker

Anshumali Shrivastava is an associate professor in the computer science department at Rice University. He is also the Founder and CEO of ThirdAI Corp, a startup focussed on democratizing Mega-AI models through "dynamic sparsity". His broad research interests include probabilistic algorithms for resource-frugal deep learning. In 2018, Science news named him one of the Top-10 scientists under 40 to watch. He is a recipient of the National Science Foundation CAREER Award, a Young Investigator Award from the Air Force Office of Scientific Research, a machine learning research award from Amazon, and a Data Science Research Award from Adobe. He has won numerous paper awards, including Best Paper Award at NIPS 2014, MLSys 2022, and Most Reproducible Paper Award at SIGMOD 2019. His work on efficient machine learning technologies on CPUs has been covered by popular press including Wall Street Journal, New York Times, TechCrunch, NDTV, Engadget, Ars technica, etc.

For Zoom Link:
Email: mfsdsai_off@iitg.ac.in