CS Talk Series

Decentralized Deep Learning: Training and Running Large Models over the Internet

Max Ryabinin

Host: Dan Alistarh

Over the recent years, the scale of deep learning has increased dramatically: pretraining models like GPT-3 can cost millions of dollars, and even their inference requires significant resources. In this talk, I will present an alternative approach: instead of using expensive clusters, we can leverage the resources of volunteers or several organizations. I will cover several papers addressing the challenges of such a setup published at NeurIPS'20, '21, and ICML'22, as well as Hivemind our open-source library for decentralized DL.I will also highlight SWARM Parallelism and Petals our latest works about decentralized pretraining and inference of large language models. SWARM is a system for training large models over slow networks of heterogeneous unreliable devices: to achieve this goal, it relies on randomized pipelines and dynamic rebalancing between pipeline stages. In turn, Petals leverages the same techniques and allows everyone to run, finetune or inspect the internals of LLMs like BLOOM or OPT-175B without access to many GPUs or the need for offloading even from Colab notebooks.

Tuesday, March 21, 2023 01:00pm - 03:00pm

Heinzel Seminar Room / Office Bldg West (I21.EG.101)

This invitation is valid as a ticket for the ISTA Shuttle from and to Heiligenstadt Station.
Please find a schedule of the ISTA Shuttle on our webpage:
https://ista.ac.at/en/campus/how-to-get-here/ The ISTA Shuttle bus is marked ISTA Shuttle (#142) and has the Institute Logo printed on the side.