UPO UNIVERSITÁ DEL PIEMONTE ORIENTALE
DIPARTIMENTO DI SCIENZE E INNOVAZIONE TECNOLOGICA

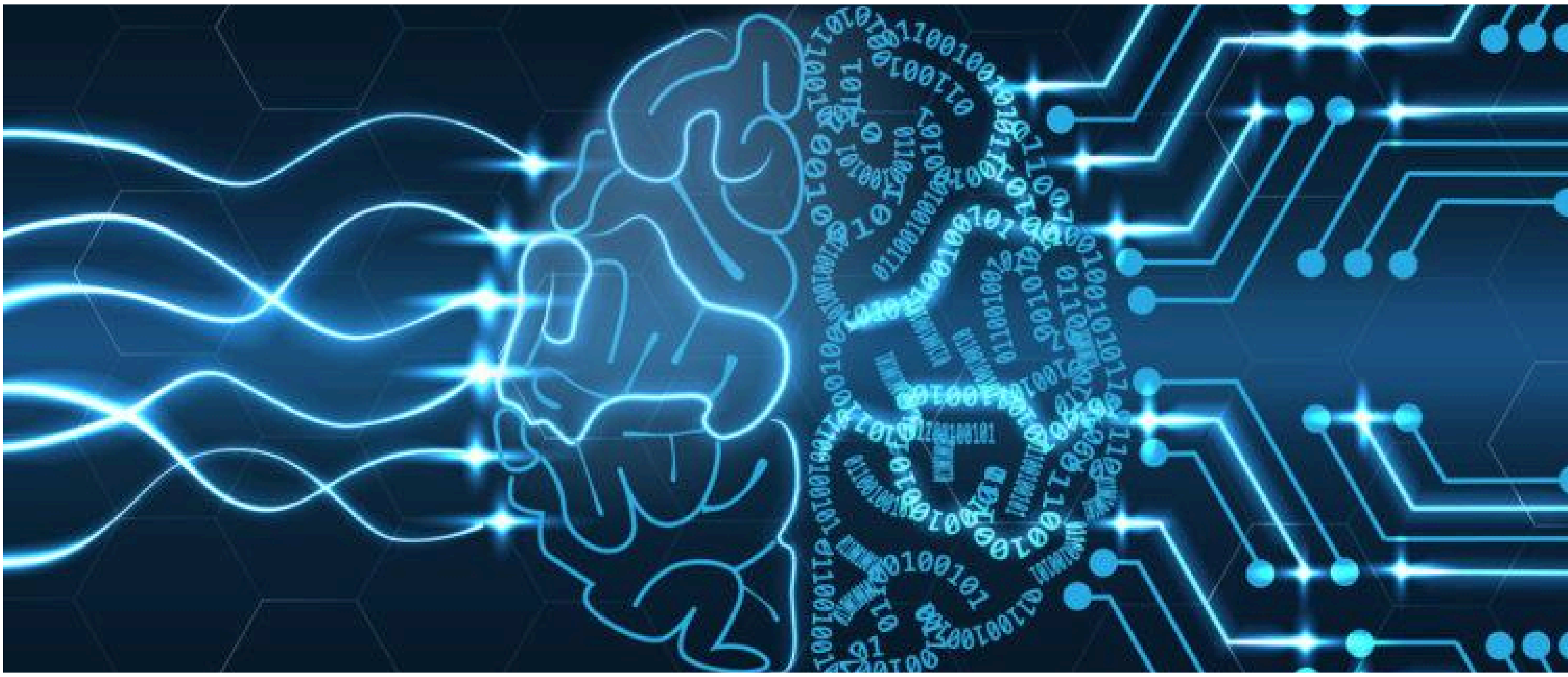
EVENTI DiSIT

Seminario | Seminar
13-01-2025
ore 11:15-12:30
Sala Seminari C192

Predicting Performance of Machine Learning Workloads and of Distributed ML Systems

Prof.ssa Leana Golubchik

University of Southern California, Los Angeles (CA), USA



Deep learning has made substantial strides in many applications; new training techniques, larger datasets, increased computing power, and easy-to-use machine learning frameworks all contribute to this success. Accurate performance prediction (e.g., latency, throughput) of machine learning (ML) workloads is useful for a number of reasons, including resource management, neural architecture search, and efficient training and inference. In the first part of the talk, our framework for inference latency prediction on mobile devices is presented.

An important missing piece is that deep learning frameworks do not assist users with provisioning cloud resources; most users need to try different job configurations to determine the resulting training performance. When resources are shared among hundreds of jobs, this approach quickly becomes infeasible. In the second part of the talk, we will focus on our approach to predicting performance metrics and scheduling algorithms that use these metrics to guide resource allocation. Our goal is to broaden the population of users capable of developing deep learning models and applying them to novel applications.

EVENTO APERTO A:
Docenti | Teachers, Borsisti | Research Fellows, Assegnisti | Postdoctoral researchers, Dottorandi | PhD students, Studenti | Students
SEMINARIO IN LINGUA: Inglese - English