Contribution ID: **494** Type: **Talk**

®Training ML algorithms on resource-constrained devices — a memory/storage perspective

Tuesday, 2 December 2025 16:10 (20 minutes)

Title: Training ML algorithms on resource-constrained devices - a memory/storage perspective

Summray:
Deploying ML/AI algorithms on the edge is necessary for applications (e.g., security and surveillance, industrial IoT, autonomous vehicles, healthcare use cases, ...) requiring low latency, data privacy or reduced costs. However, most edge devices are not equipped with powerful memory systems to perform such memory and processing intensive applications. The objective of this presentation is to show some optimization venues to unlock the memory/storage bottleneck of some ML/AI algorithms mainly from a learning perspective to deply them on low-resource devices. The optimizations presented in this talk could be also applied to whatever resource constrained device used for training, be it cheap virtual machines on cloud infrastructures, common personal computers or resource contrained micro datacenters.

Deploying ML/AI algorithms at the edge is essential for applications such as security and surveillance, industrial IoT, autonomous vehicles, and healthcare, which require low latency, data privacy, or reduced costs. However, most edge devices lack powerful memory systems capable of handling the memory- and computation-intensive nature of such applications.
The objective of this presentation is to highlight some optimization strategies that help overcome the memory and storage bottlenecks of ML/AI algorithms—mainly from a training perspective—to enable their deployment on low-resource devices. These optimizations can also be applied to any resource-constrained environment used for training, including low-cost virtual machines in cloud infrastructures, standard personal computers, or small-scale micro data centers.

Presenting Author

Email

Student or Postdoc?

Institute

Registered for the conference?

CHPC User

CHPC Research Programme

Workshop Duration

Primary author: BOUKHOBZA, Jalil (ENSTA, Institut Polytechnique de Paris)

Presenter: BOUKHOBZA, Jalil (ENSTA, Institut Polytechnique de Paris)

Session Classification: HPC Technology

Track Classification: Storage and IO