ML_INFN

Lucio Anderlini, Tommaso Boccali,
Stefano Dal Pra, Cristina Duma

ML_INFN

	2020-2022	2023
RN	T. Boccali	L. Anderlini
WP1	S. Dal Pra	S. Dal Pra
WP2	L. Anderlini	
WP3	C. Duma	

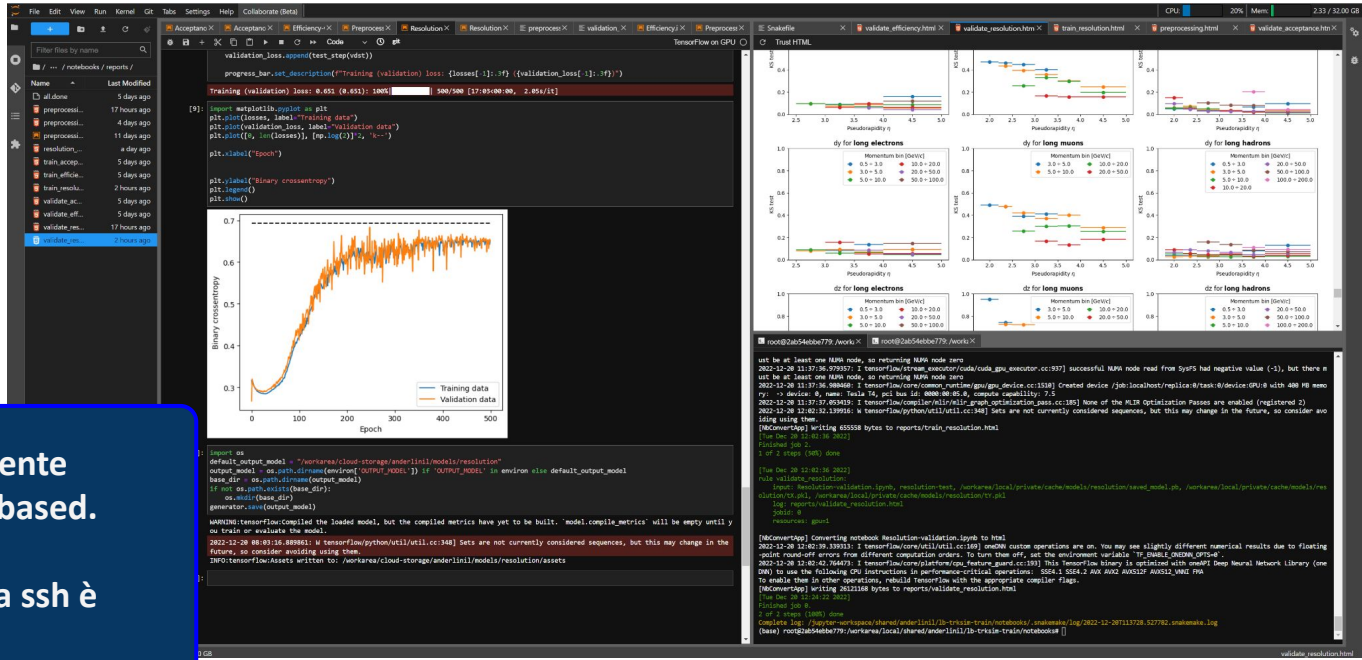
Esperimento di CSN5 con l'obiettivo di fertilizzare le attività di *machine learning* nell'Ente con azioni in tre direzioni

1. **Infrastruttura**: Semplificare l'accesso a risorse hardware e configurazioni software per ricerca che richieda l'utilizzo di tecniche di machine learning
2. **Training**: Organizzare eventi di formazione (*hackathon*) per aumentare la competenza media dell'Ente su tematiche legate al machine learning
3. **Knowledge base**: costruire una rete capillare sul territorio nazionale di punti di contatto (RL) e documentare le attività nelle varie sezioni anche condividendo codice (https://confluence.infn.it/display/MLINFN)

Infrastruttura e servizi

Disponiamo di 2 server con dischi veloci (NVMe) equipaggiati con:

- 6 nVidia T4
- 5 nVidia RTX5000
- 2 nVidia A100
- 1 nVidia A30
- 2 Xilinx U50
- 1 Xilinx U250



Accesso alle risorse prevalentemente tramite JupyterLab, un IDE Web-based.

In alcuni casi, l'accesso diretto via ssh è usato per preparare i container.

Alcune attività (*teasers*)

- Correzione dell'attenuazione di immagini PET mediante Deep Learning (FI).
- Machine learning classification for COVID19 patients performed on small datasets of CT scans (PI+PV+FI).
- Digital restoration of artworks using XRF scan data (FI + CNAF)
- Domain adaptation in the definition of deep classifier for searches at CMS (FI)
- Development of algorithms for the fast simulation of the LHCb experiment (FI+MIB)
- Machine Learning + Theoretical physics for precision measurements at the LHC (FI)
- Machine learning approaches to the QCD transition (FI)
- Evaluation of FUNNEL on Kubernetes: a REST-ful server for task management (UniURB)
- Studies on serving FPGAs on the cloud (BO+CNAF)
- GW and multi-messanger analysis of Virgo data with ML techniques (GEMS, CNAF, Sapienza)
- …

Piani per il 2023

Il modello attuale con le VM assegnate ai diversi progetti è estremamente pratico, ma anche molto dispendioso.

Cercheremo di migrare almeno parzialmente verso un **modello più dinamico** in cui una GPU non utilizzata da un container viene immediatamente liberata per altre applicazioni.

Varie idee, discussioni e lavori in corso per raggiungere il risultato.

Nuovo hardware in arrivo:

- un nuovo server
- 3 nuove A100
- 5 nuove Alveo U250

Potenziamento cloud:

- Monitoring di provisioning
- Monitoring di servizio
- Accounting