Contribution ID: **21** Type: **Contributed talk**

Training and optimisation of large transformer models at CERN: an ATLAS case study on Kubeflow

Tuesday 30 January 2024 15:50 (20 minutes)

Heavy flavour jets underpin a large part of the ATLAS physics programme, such as analyses of Higgs boson decays to quarks and super-symmetry searches with b-jets. The algorithms for identifying jets originating from b- and c-quarks are instrumental in these efforts, with the recently introduced GN2 model [1] showing remarkable improvements in tagging efficiency. Given its complexity and data demands, high-performance GPU clusters are essential for training GN2. Unfortunately, many within the collaboration lack such resources, emphasising the need for equitable project access. Additionally, the performance of GN2 can be improved through further optimisation of its hyperparameters, an even more computationally demanding task that can be automated with frameworks like Katib. Addressing these two challenges, the ATLAS flavour tagging group is assessing training on CERN IT's Kubeflow infrastructure for machine learning (ml.cern.ch) backed by Kubernetes. This talk will showcase a framework for CERN users to utilise these resources and present the initial results of this effort. Furthermore, the talk will highlight a cutting-edge approach to optimise hyperparameters using μTransfer [2], a deep learning technique recently developed to optimise large language models by zero-shot transferring the performance from lower-complexity models to their full-complexity equivalent. While centred on an ATLAS use case, the methodology presented will be relevant to any collaboration employing advanced ML.

[1] ATLAS Collaboration, Public plots for MC/MC and Data/MC comparisons of Dl1d and GN1, and simulation performance of GN2, (2023), https://atlas.web.cern.ch/Atlas/GROUPS/PHYSICS/PLOTS/FTAG-2023-01/

[2] Yang, Greg, et al., Tensor programs v: Tuning large neural networks via zero-shot hyperparameter transfer. arXiv preprint arXiv:2203.03466 (2022).

Author: DRAGUET, Maxence (University of Oxford (GB))

Presenter: DRAGUET, Maxence (University of Oxford (GB))

Session Classification: Contributed Talks

Track Classification: 5 ML infrastructure : Hardware and software for Machine Learning