Contribution ID: **146** Type: **Short Talk**

Distributed training and scalability for the particle clustering method UCluster

Thursday 20 May 2021 11:03 (13 minutes)

In recent years, machine learning methods have become increasingly important for the experiments of the Large Hadron Collider (LHC). They are utilized in everything from trigger systems to reconstruction to data analysis. The recent UCluster method is a general model providing unsupervised clustering of particle physics data, that can be easily modified for a variety of different tasks. In the current paper, we improve on the UCluster method by adding the option of training the model in a scalable and distributed fashion, which extends its usefulness even further. UCluster combines the graph-based neural network ABCnet with a clustering step, using a combined loss function to train. It was written in TensorFlow v1.14 and has previously been trained on a single GPU. It shows a clustering accuracy of 81% when applied to the problem of multiclass classification of simulated jet events. Our implementation adds the distributed training functionality by utilizing the Horovod distributed training framework, which necessitated a migration of the code to TensorFlow v2. Together with using parquet files for splitting data up between different nodes, the distributed training makes the model scalable to any amount of input data, something that will be essential for use with real LHC datasets. We find that the model is well suited for distributed training, with the training time decreasing in direct relation to the number of GPU's used.

Primary author: SUNNEBORN GUDNADOTTIR, Olga (Uppsala University (SE))

Co-authors: GEDON, Daniel (Uppsala University); DESMARAIS , Colin (Uppsala University); BENGTSSON BERNANDER, Karl (Uppsala University); Dr SAINUDIIN, Raazhesh (Uppsala University); Dr GONZALEZ SUAREZ, Rebeca (Uppsala University)

Presenter: SUNNEBORN GUDNADOTTIR, Olga (Uppsala University (SE))

Session Classification: Artificial Intelligence

Track Classification: Distributed Computing, Data Management and Facilities