Dynamic Neural Networks for Efficient Machine Learning

Significance of the project

Today, both science and industry rely heavily on machine learning models, predominantly artificial neural networks, that become increasingly complex and demand a significant amount of computational resources. This trend can be observed across many applications, ranging from medical image processing to the automotive industry and robotics.

Such an extremely high computational cost of employing machine learning models is also visible in large High Energy Physics experiments, where an unprecedented amount of data is produced and processed. For instance, the ALICE experiment run at the CERN Large Hadron Collider, the largest and most powerful particle accelerator globally, collects several petabytes of data every hour to be processed by a plethora of machine learning models. The computations run by machine learning models to process this increasing amount of data come at an enormous price of long processing time, high energy consumption and large carbon footprint generated by the computational infrastructure.

Goal of the project

In this project, we aim to address the above-mentioned challenges by looking at **dynamic neural networks**, *i.e.* models that reduce computational burden by dynamically adjusting the processing paths of a network based on the available resources. For instance, they can make use of already performed computations, which leads to significant savings in time and memory usage, both during training and at inference. This new arising field of machine learning combines multiple topics, such as *conditional computations, reusable components, partial inference* and *continual learning*.

Instead of constraining the number of computations or memory used by the models, we focus on dynamically adjusting neural network models, their structure and computations, to what is available to them, namely:

- computations done in the previous processing steps,
- partial information accessible at run-time,
- architectural modules designed in continually learned models.

Our project focuses on designing methods that train themselves to be efficient rather than to solve a given task. To that end, we hypothesize that adjusting the processing paths of neural networks and the components they use can significantly increase their efficiency. Driven by this assumption, we aim to develop dynamic neural network models focused on the efficiency of machine learning by saving computations and reducing their resource usage.

The end goal of the project is to create efficient machine learning models based on dynamically adjustable resource-aware computations, and validate them in one of the largest experimental testbeds created by humankind, the Large Hadron Collider at CERN.

Expected results

The main result of our project will be a set of novel efficient machine learning methods that allocate processing paths and use components based on resources available to them, such as partial information or precomputed calculations.

As a part of this project, we will validate the proposed methods in a real-life application of high energy physics experiments, which is a perfect testbed for our research as it offers practical constraints such as the overflow of data collected during particle collisions at an exceptionally high frequency. Therefore, the practical results we expect from this project are the integration of resource-aware efficient machine learning models based on dynamic neural networks in the IT infrastructure of the largest high energy physics experiment in the world at CERN.