Enhancing Deep Learning towards Exascale with the DEEP-EST Modular Supercomputer Architecture

Ernir Erlingsson,[1] Gabriele Cavallaro[2], Morris Riedel,[1,2] and Helmut Neukirchen[1]

[1] *University of Iceland, Reykjavik, Iceland* ere29@hi.is, helmut@hi.is

[2] *Jülich Supercomputing Centre, Jülich, Germany* g.cavallaro@fz-juelich.de, m.riedel@fz-juelich.de

The Dynamical Exascale Entry Platform – Extreme Scale Technologies (DEEP-EST)[1] aims at delivering a pre-Exascale platform based on a Modular Supercomputer Architecture (MSA) which provides among a standard CPU cluster module, a many-core Extreme Scale Booster (ESB), a Global Collective Engine (GCE) to speed-up MPI collective operations in hardware, Network Attached Memory (NAM) as a fast scratch file replacement, and a hardware accelerated Data Analytics Module (DAM); the latter are able to perform near-data processing.

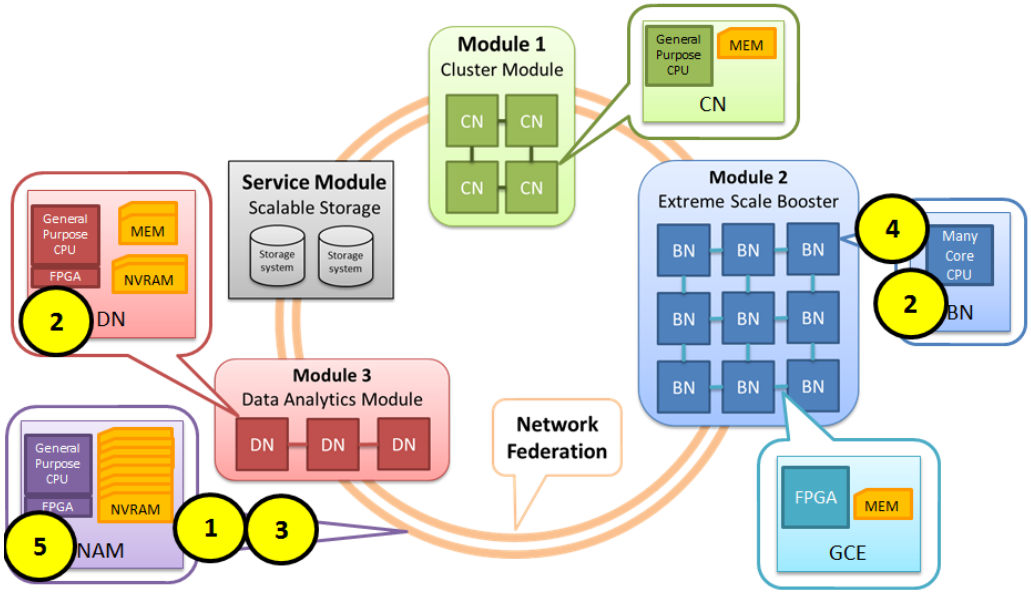

Figure 1. Transfer learning configuration using the Modular Supercomputer Architecture (MSA).

As partner in the DEEP-EST consortium, we aim at enhancing Machine Learning (ML) towards Exascale by exploiting the MSA. We describe here the case of transfer learning[2] with Convolutional Neural Networks (CNNs) as depicted in the steps of Fig. 1: (1) Pre-trained CNNs[3] are loaded into the NAM module; (2) New CNNs are trained in the DAM, or alternatively the ESB module, by using existing CNNs as fixed feature extractors for new datasets. Additionally, the DAM's FPGA can be utilized to accelerate the time needed to prepare existing CNNs for training with the new datasets; (3) These trained models are stored in the NAM module to speed-up the evaluation and comparison phase; (4) Model evaluation is embarrassingly parallel and is therefore suitable for the ESB module and its many-core architecture. Specific testing datasets are combined with the trained CNN models residing in the NAM; (5) Evaluation results are written to the NAM module, using its FPGA to determine which model exhibits the best accuracy. The described process steps can then be repeated any number of times to improve the model and reduce the likelihood of a local minimum convergence.

This mapping of a Deep Learning software chain (and other ML approaches, e.g. clustering) to the MSA will be implemented and evaluated in the course of the DEEP-EST project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Grant Agreement No. 754304 DEEP-EST.

[1] `http://deep-est.eu/`

[2] Transfer learning is a machine learning technique which focuses on storing models previously trained and using them as the foundation for new models trained with different datasets like ImageNet or related datasets in application domains.

[3] For example, OverFeat (`http://github.com/sermanet/OverFeat`).