

GRAPHCORE

Topic: Machine Learning at the Exaflop scale and beyond

Description:

Neural networks, and machine learning in general, have demonstrated to be extremely powerful tools to allow machines to execute complex tasks [1]. In order to make this possible, a large amount of data has to be available to perform their training.

While in the past decades large amounts of data were available just for some specific institutions, like the CERN in Geneva, nowadays this is no longer true.

Many modern companies, like Amazon or Facebook just to cite some of them, are able to exploit the full power of machine learning thanks to their huge data banks.

The size of modern neural networks, together with the size of the data sets used to train them, makes training unfeasible in reasonable times without the use of some sort of parallelization [2,3].

It is then crucial to distribute the effort of the optimization over different workers allowing then the execution of many more operations in parallel, speeding up the training [4]. Many different algorithms have been proposed to achieve this task [5,6,7], none of them have been demonstrated to be robust for arbitrarily large number of parallel processes.

In this master thesis the student will be introduced to modern state-of-the-art literature about parallel training of neural networks and to the technological challenges that such machine may face when scaled up to huge scales.

The aim of the project is to explore the effects on training of neural networks when the dataset is separated in different ways between the different workers. Open datasets will be used in this study.

The work will be done in close collaboration with Graphcore, and preferably the student will do their work at the Graphcore Oslo office, but other arrangements can also be considered. Summer internships preceding the master thesis may be offered to interested students

Bibliography:

[1] Exascale Deep Learning for Climate Analytics <https://arxiv.org/pdf/1810.01993.pdf>

[2] GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism, <https://arxiv.org/pdf/1806.03377.pdf>

[3] PipeDream: Fast and Efficient Pipeline Parallel DNN Training, <https://arxiv.org/pdf/1806.03377.pdf>

[4] Image Classification at Supercomputer Scale, <https://arxiv.org/pdf/1811.06992.pdf>

[5] Don't Use Large Mini-Batches, Use Local SGD, <https://arxiv.org/abs/1808.07217>

[6] Deep learning with Elastic Averaging SGD, <https://arxiv.org/abs/1412.6651>

[7] Stochastic Gradient Push for Distributed Deep Learning <https://arxiv.org/pdf/1811.10792.pdf>

Topic: Distributed training of extreme-size models

Description:

Machine learning models are gaining more and more interest in different areas thanks to their ability to manipulate complex tasks and complex scenarios [1].

Modern day models can be trained to play chess and Go, just to cite some of them, and consistently beat the best human players in these fields. The ability of these models to understand and master complex phenomena where data have dimension far beyond the capability of humans, is something that could have in principle infinite applications in modern world. Just to cite some of them, self-driving cars, recognition of the Alzheimer's disease and fraud detection. In the future, more and more tasks will be accomplished by such algorithms and there will be more and more need to find efficient ways to train them.

In order to accomplish these tasks, machine learning models have to grow in size and their training has to involve multiple hardware processors [2, 3].

Due to this fact, model-parallel training of neural networks is gaining more and more attention. In this case, the same model is spread between different hardware and its training involves then a large amount of communication between the different processors.

The efficiency of the process is then intrinsically connected to the way the model has been divided, making it crucial to develop an efficient way to find the best configuration dependent on both model and hardware details [4, 5].

During this research project, the student will deepen his/her understanding of machine learning algorithms and in particular of the SGD training of neural networks in a model-parallel configuration.

The project will then continue with the development of the best possible configuration given hardware specs and the characteristics of the specific model that has to be trained. It will then be possible to test the theoretical results on bleeding edge hardware accelerators.

The work will be done in close collaboration with Graphcore, and preferably the student will do their work at the Graphcore Oslo office, but other arrangements can also be considered. Summer internships preceding the master thesis may be offered to interested students

Bibliography:

- [1] Exascale Deep Learning for Climate Analytics, <https://arxiv.org/pdf/1810.01993.pdf>
- [2] GPipe: Efficient Training of Giant Neural Networks using Pipeline Parallelism, <https://arxiv.org/pdf/1806.03377.pdf>
- [3] Image Classification at Supercomputer Scale, <https://arxiv.org/pdf/1811.06992.pdf>
- [4] PipeDream: Fast and Efficient Pipeline Parallel DNN Training, <https://arxiv.org/pdf/1806.03377.pdf>
- [5] Beyond Data and Model Parallelism for Deep Neural Networks, <https://arxiv.org/abs/1807.05358>

Contact for more information/supervision: [Arjun Chandra](#) (Graphcore) and [Jim Tørresen](#) (UiO/ROBIN)