Study program:Artificial intelligence

Name of the subject: Distributed Deep Learning

Teacher(s): Srdjan Skrbic

Status of the subject: obligatory

Number of ECTS credits: 5

Conditions: none

Subject goal

Distributed deep learning refers to both distributed training and inference. With today's deep neural networks, both processes are computationally intensive and require usage of parallel and distributed methods of computation to reduce execution time. The objective of this course is to study the principles, models, tools, and techniques for applying a variety of distributed methods in deep learning. Also, the course will complement the first course in deep learning by covering reinforcement learning and deep generative modeling and further investigating autoencoders and variational autoencoders.

Outcome of the subject

Minimal: At the end of the course, it is expected that students understand and show ability to discuss various distributed methods in deep learning. Minimal outcome includes knowledge of using appropriate distributed software tools to build, train and apply deep learning networks in practice.

Desirable: It is expected that a successful student understands solutions to key problems in distributed deep learning and shows ability to identify the optimal way of solving given problem using presented methods. Having in mind practical orientation of the course, a priority outcome is active knowledge of advanced concepts of distributed deep neural networks programming using presented software tools.

Subject content

Theory

The course starts with a brief overview of neural network models and usage of distributed methods in inference and in training of deep neural networks. Two ways to distribute calculations across multiple devices are covered. Data parallelism, where one model replicates to multiple nodes each of them processing different data sets, with the results aggregated at the end. Model parallelism, where different parts of one model execute on different nodes, processing the same set of data. Course covers options to use multiple GPUs installed on a single node, as well as usage of a cluster of many nodes, each hosting one or multiple GPUs. The course will cover deep generative modeling, variational autoencoders and reinforcement learning as case studies that motivate usage of distributed methods.

Practical learning

Practical exercises start with a brief overview of existing software tools for distributed methods in deep learning. As the course progresses towards concrete distributed methods, practical exercises cover programming methods for building, training, and application of presented methods in various real-world examples. At the end of the practical exercises, students develop and present individual distributed deep learning projects in a competitive environment.

Expert from the industry will be included into the project assignment realization as an external tutor.

Literature

1. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition, Aurélien Géron, O'Reilly Media, Inc., September 2019.
2. Deep Learning, By John D. Kelleher, The MIT Press, 2019.
3. Next-Generation Machine Learning with Spark, Quinto Butch, APress, 2020.
4. Deep Learning with PyTorch, Eli Stevens, Luca Antiga, and Thomas Viehmann, Manning 2020.
5. Kennedy, R.K.L., Khoshgoftaar, T.M., Villanustre, F. et al. A parallel and distributed stochastic gradient descent implementation using commodity clusters. J Big Data 6, 16 (2019).
6. Demystifying Parallel and Distributed Deep Learning: An In-depth Concurrency Analysis, T. Ben-Hun, T. Hoefler, ACM Computing Surveys, Vol. 52, No. 4, 2019.

Number of active teaching classes	**Theoretical teaching: 2**	**Practical teaching: 2**

Method of carrying out the teaching

During theoretical classes, classical methods of teaching with the use of a video beam are used to present stated topics with focus on their application. Practical classes are done using video beam and computers with appropriate software installed to master the skillful usage of suggested tools through individual development of deep neural networks using parallel and distributed methods. A premise for successful practical exercises is the existence of enough computers so that every student may work individually and usage of a computer cluster with at least 8 nodes with GPUs and fast interconnect network.

Evaluation of knowledge (maximum number of points 100)

Pre-exam obligations	points	**Final exam**	points
Colloquia	30	Oral exam	40
Project	30		