CS 58010: Scalable Learning Systems

Subject: CS Faculty: Faculty of Engineering and Natural Sciences
SU Credit: 3 , ECTS Credit: 10.00 / 10.00 ECTS
Instructor(s): Kubilay Atasu
Language of Instruction: English
Level of Course: Graduate
Planned Learning Activities: reading, presentation, discussion, and project prototyping

CONTENTS (ENGLISH)

This course provides a broad overview of state-of-the-art parallel and distributed machine learning (ML) and deep learning (DL) algorithms and systems, with a strong focus on the scalability, resource efficiency, data requirements, and robustness of the solutions. This course covers effective ways to map state-of-the-art ML and DL solutions to parallel AI accelerators such as GPUs and TPUs. A set of techniques are presented to efficiently scale ML and DL workloads to a large number of distributed machines in the presence of system failures and malicious attacks. Finally, methods for improving the scalability and efficiency of generative learning and graph learning approaches are covered.

Course topics include

• Overview of parallel and distributed ML/DL algorithms

• Performance and scalability of state-of-the-art ML/DL systems

• Hardware-accelerated ML/DL solutions

• Federated machine learning systems

• Scaling generative AI systems

• Scaling graph learning systems

İÇERİK (TÜRKÇE)

Bu ders, çözümlerin ölçeklenebilirliği, kaynak verimliliği, veri gereksinimleri ve sağlamlığına güçlü bir şekilde odaklanarak, son teknoloji paralel ve dağıtılmış makine öğrenimi (MÖ) ve derin öğrenme (DÖ) algoritmaları ve sistemlerine ilişkin geniş bir genel bakış sunar. Bu derste GPU'lar ve TPU'lar gibi paralel yapay zeka hızlandırıcılarına son teknoloji MÖ ve DÖ çözümlerini eşlemenin etkili yolları ele alınır. Sistem arızaları ve kötü amaçlı saldırılar varlığında MÖ ve DÖ iş yüklerini çok sayıda dağıtılmış makinaya verimli bir şekilde ölçeklendirmek için bir dizi teknik sunulur. Son olarak, üretken öğrenme ve grafik öğrenme yaklaşımlarının ölçeklenebilirliğini ve verimliliğini iyileştirme yöntemleri ele alınır.

REFERENCE BOOKS

Deep Learning by Ian Goodfellow and Yoshua Bengio and Aaron Courville, MIT Press,

OBJECTIVE

This course aims to provide the students with skills to build efficient, scalable, and robust parallel and distributed machine learning (ML) and deep learning (DL) solutions.

LEARNING OUTCOMES

1. Demonstrate deep understanding of parallel and distributed machine learning and deep learning algorithms and systems by analyzing and discussing research papers.

2. Design and implement scalable and efficient machine and deep learning algorithms and systems, evaluate time-space and cost-performance tradeoffs.

3. Analyze and evaluate the resiliency of federated machine learning algorithms and systems against various attacks.

4. Formulate research questions and objectives and relate them to the relevant literature in the broad field of deep generative learning systems.

ASSESSMENT METHODS AND GRADING

1) Paper presentations (20%): each student needs to choose two research papers from a given set, present them (15 minutes), and lead the discussion (5 minutes).

2) Research project (60%): The goal is to improve the speed, scalability, resilience, or accuracy of an existing ML/DL approach. The students need to hand in a project report in the style of a short scientific paper stating their contribution to the overall system performance. Grading: 40% for the project report and 20% for the project presentation and Q&A.

3) Examination (20%): written, oral, or both.

Reason for proposing the course

Expanding the existing computer science curriculum with a new graduate-level course covering scalability of ML and DL algorithms and systems.