MULTIMEDIA INFORMATICS

MMI 712– Machine Learning Systems Design and Deployment

SYLLABUS

Year, Semester:	2023-2024 Fall
Course Conduct:	Face-to-face in class Wednesday@13:40 Lecture videos on YouTube & Lecture notes on ODTUClass Students are expected to watch the lecture videos and study the course material before attending the weekly face-to-face sessions. There will be regular quizzes from the content and attendance is expected in these sessions.
Lecturer:	Prof. Dr. Alptekin Temizel, atemizel@metu.edu.tr
Teaching Assistant:	Ayberk Aydın, aayberk@metu.edu.tr

Course Objective

The course covers several aspects of designing reliable and scalable machine learning systems for real-world deployment. It deals with development of production quality models and introduces the machine learning pipeline, concepts on machine learning system design and data engineering. It provides know-how on model development, and how to scale up the training for large models as well as evaluation, calibration and debugging of these models. Generation of reproducible models via experiment tracking tools and model versioning is also covered. Hardware platforms and frameworks for deployment are introduced, followed by basic deployment concepts, containerized deployment and testing.

Reference Material:

CS 329S: Machine Learning Systems Design - https://stanford-cs329s.github.io/

Reading Material:

Rules of Machine Learning: Best Practices for ML Engineering https://developers.google.com/machine-learning/guides/rules-of-ml
Challenges in Deploying Machine Learning: a Survey of Case Studies https://arxiv.org/pdf/2011.09926.pdf

Grade Distribution:

Assignments	45%
Final project	30%
5x Quizzes	25%

Deliverables

Documents and necessary files of the assignments must be uploaded to ODTUClass by students before the specified due dates.

University Policies

All students are **expected to obey** the university code of integrity and avoid academic dishonesty or plagiarism.

No	Date	
1	4 Oct	**Introduction to the Course and Machine Learning Life-Cycle**
2	11 Oct	**Designing a Machine Learning System – I** Main Requirements of Machine Learning Systems Reliability, scalability, maintainability, adaptability ML in research vs. in production Traditional software vs. ML software, ML Production Myths
3	18 Oct	**Designing a Machine Learning System - II** Batch vs. online, Edge vs. cloud computing, Offline vs. Online Learning Iterative Development Phases of ML Adoption
4	25 Oct	**Versioning and Experiment Tracking** Experiment tracking tools Data versioning ML pipeline versioning Continuous integration/continuous delivery for ML
5	1 Nov	**Hands-on Lab** Experiment tracking
6	8 Nov	**Data Engineering** Data centric approach Data basics and data formats Creating training datasets, labelling Semi-supervised and self-supervised learning
7	15 Nov	**Hands-on Lab** Containerization
8	22 Nov	**Data Engineering-II** Sampling, Class imbalance problems Data Augmentation Data Leakage Data Analysis with FiftyOne
9	29 Nov	Model Development and Training Model Selection AutoML, Neural Architecture Search Optimizers Model Calibration
10	6 Dec	**Case Study – Deep Learning Solutions for Retail Stores** Invited speaker from industry Dr. Cihan Öngün, Senior Deep Learning Engineeer, Signatrix GmbH
11	13 Dec	**Case Study – Autonomous Driving** Invited speaker from industry Dr. Berker Loğoğlu, Head of Computer Vision, Machine Learning and Robotics at Kuartis Technology and Consulting
12	20 Dec	**Model Optimization for Deployment** Model compression Quantization Pruning Knowledge distillation
13	27 Dec	**Deployment Platforms and Frameworks** GPUs, TPUs, IoT devices and TinyML Packaging TensorRT, Triton inference server Google Cloud Platform (GCP), Amazon Web Services (AWS)
14	3 Jan	**Evaluation** Debugging, System evaluation and testing Data testing, profiling and visualization Benchmarking Perturbation evaluation, ablation study

There’ll also be extra content during face-to-face classes to cover recent developments in the field and go through some case studies such as:

- Case Study: Data Labelling
- Labelling Errors in Public Datasets
- Explainable AI
- Case Study: 150 successful machine learning models:6 lessons learned at Booking.com
- Why ML Projects Fail?
- Foundation Models
- Prompt Engineering
- Data Analysis with FiftyOne
- Hyper Parameter Tuning
- How To Train Your ViT
- Model Optimization Case Study: BERT
- Model Sparsification
- NVIDIA NCCL
- Distributed Training
- Working with Large Models
- PyTorch Performance Tuning
- Paper: Efficiency Misnomer
- MLPerf Inference Benchmark
- DeepChecks