EN

Machine Learning Operations (MLOps)

Short description (max. 300 chars)

Machine Learning Operations (MLOps) comprises a set of tools and best practices for bringing Machine Learning (ML) into production. We take a systems approach to MLOps, considering ML systems holistically to ensure all components work together to satisfy the specified objectives and requirements.

Module coordinator

Frank-Peter Schilling (scik)

Learning objectives (competencies)

Objectives	Competences	Taxonomy levels
(1) You know the methods and algorithms needed to bring ML models into production	D	C2
… and you can transfer and apply algorithms and methods to real-world use cases	D, M	C3
(2) You are familiar with the most important software tools and frameworks needed to build ML systems (apart from pure models)	D	C2
… and you are able to select and apply relevant software tools and/or frameworks to address a given MLOps task	D, M	C3
… and you are able to develop an end-to-end ML system prototype using existing tools and frameworks	D, M	C5
(3) You have a good overview of what a ML system comprises and how its components interact, including the business perspective	D	C2

Module contents

The new and fast-evolving field of Machine Learning Operations (MLOps) takes inspiration from the concept of DevOps (Development and Operations) to establish methods, best practices and tools to operationalize an ML system, i.e., to bring it into production. Starting from a ML model (e.g., a deep neural network) which has been trained on a specific dataset to solve a particular problem, we consider all additional components and workflows which are needed to deploy and maintain ML successfully in practice.

In particular, we discuss relevant metrics and baseline models, infrastructure (clusters, cloud, resource management) and tooling (frameworks), model training and debugging, model evaluation and tuning, data management (sources, storage, versioning, privacy), systems testing (CI/CD) and explainability, deployment (batch, service, edge), monitoring (data drift) and continual learning.

Taking a systems approach to MLOps, additional topics such as business requirements and objectives, project management for ML, team structure, user experience as well as responsible use of ML systems are also considered.

Content includes:

- Overview and Introduction to MLOps and ML systems
- Data infrastructure and tooling (e.g., data models, versioning, storage, processing, pipelines)
- Training data processing (e.g., sampling, Labelling, Augmentation, Simulation)

- Feature engineering
- Model development and debugging methodology and tools (e.g., system validation, versioning, evaluation, baselines)
- Deployment infrastructure and tooling (e.g., batch vs online, model compression & quantization, cloud & edge deployment)
- Monitoring (e.g., data distribution shifts, failures, metrics and logging)
- Continual learning & tests in production
- ML project management and business perspective

The labs include examples involving state-of-the-art MLOps software tools and frameworks, as well as real-world use cases for ML systems. Included is a project task, where students develop a prototypical ML system comprising one or more existing and appropriate tools, based on a real-world use case they chose. The student projects will be presented in-class at the end of the semester, and will contribute to the grading, together with the final written exam.

Teaching materials

- Lecture slides, lab descriptions

Supplementary material

- Chip Huyen, “Designing Machine Learning Systems - An Iterative Process for Production-Ready Applications” O’Reilly, 2022
- Noah Gift & Alfredo Deza, “Practical MLOps - Operationalizing Machine Learning Models”, O’Reilly, 2021

- C. Huyen, “CS 329S: Machine Learning Systems Design”, Stanford University, 2021, https://stanford-cs329s.github.io/
- A. Ng, “Machine Learning Engineering for Production (MLOps) Specialization”, DeepLearning.AI and Coursera, 2021, https://www.coursera.org/specializations/machine-learning-engineering-for-production-mlops
- P. Abbeel, S. Karayev, and J. Tobin, “Full Stack Deep Learning - Spring 2021”, UC Berkeley, 2021, https://fullstackdeeplearning.com/spring2021/
- G. Mohandas, “Made with ML: MLOps Course”, 2022, https://madewithml.com/#mlops

- Scientific literature to be specified in class

This module is part of a potential specialization in machine learning: MLDM1 teaches foundations of machine learning, MLDM2 (DL) covers foundations of deep learning. AI1 introduces different paradigms to create intelligent agents (incl. certain unusual types of machine learning), AI2 covers the development of fair and more general AI algorithms. CVDL introduces state-of-the-art concepts of advanced deep learning for computer vision applications. This course (MLOps) takes a systems perspective, introducing methods & tools to develop, maintain and operate ML systems (analog to software engineering and DevOps covering respective topics in a software development specialization). All elective modules can be attended independently of each other (exception: CVDL and AI2 build on basic knowledge in deep learning, as e.g., conveyed in MLDM2/DL) and overlap in less than one lecture.

Prerequisites

- Successful completion of MLDM1 & MLDM2 (Advanced ML), or comparable experience

Teaching language

( ) German (X) English

Part of International Profile

(X) Yes ( ) No

Module structure

Type 2a

Exams

Description	Type	Form	Scope	Grade	Weighting
Final exam	Test (on paper or online)	written	90 minutes	max. 60 points	60%
Project work (individually or in team)	Project work	oral (code review)	10 minutes	max. 20 points	20%
Project work (individually or in team)	Project work	oral (final Presentation)	10 minutes	max. 20 points	20%

Remarks

n/a

Legal basis

The module description is part of the legal basis in addition to the general academic regulations. It is binding. During the first week of the semester a written and communicated supplement can specify the module description in more detail.

DE

Machine Learning Operations (MLOps)

Short description (max. 300 chars)

Machine Learning Operations (MLOps) umfasst eine Reihe von Tools und Best Practices für die Einführung von Machine Learning (ML) in die Produktion. Wir verfolgen bei MLOps einen Systemansatz, bei dem ML-Systeme ganzheitlich betrachtet werden, um sicherzustellen, dass alle Komponenten zusammenarbeiten, um die festgelegten Ziele und Anforderungen zu erfüllen.

Module coordinator

Schilling Frank-Peter (scik)

Learning objectives (competencies)

Objectives	Competences	Taxonomy levels

(1) Sie kennen die Methoden und Algorithmen, die erforderlich sind, um ML-Modelle in Produktion zu bringen	F	K2
… und Sie können Algorithmen und Methoden auf reale Anwendungsfälle übertragen und anwenden	F, M	K3
(2) Sie kennen die wichtigsten Software-Werkzeuge und Frameworks, die für den Aufbau von ML-Systemen benötigt werden (über reine Modelle hinausgehend)	F	K2
… und Sie sind in der Lage, relevante Software-Werkzeuge und/oder Frameworks auszuwählen und anzuwenden, um eine bestimmte MLOps-Aufgabe zu lösen	F, M	K3
… und Sie sind in der Lage, einen durchgängigen ML-Systemprototyp unter Verwendung bestehender Werkzeuge und Frameworks zu entwickeln	F, M	K5
(3) Sie haben einen guten Überblick darüber, was ein ML-System umfasst und wie seine Komponenten zusammenwirken, einschließlich der Business-Perspektive	F	K2

Module contents

Der neue und sich schnell entwickelnde Bereich Machine Learning Operations (MLOps) ist vom DevOps-Konzept (Development und Operations) inspiriert, um Methoden, bewährte Verfahren und Werkzeuge für die Operationalisierung eines ML-Systems zu entwickeln, d. h. es in Produktion zu bringen. Ausgehend von einem ML-Modell (z. B. einem tiefen neuronalen Netz), das auf einem bestimmten Datensatz trainiert wurde, um ein bestimmtes Problem zu lösen, betrachten wir alle zusätzlichen Komponenten und Arbeitsabläufe, die erforderlich sind, um ML in der Praxis erfolgreich einzusetzen und zu warten.

Insbesondere diskutieren wir relevante Metriken und Baseline-Modelle, Infrastruktur (Cluster, Cloud, Ressourcenmanagement) und Werkzeuge (Frameworks), Modelltraining und Debugging, Modellevaluierung und -optimierung, Datenmanagement (Quellen, Speicherung, Versionierung, Datenschutz), Systemtests (CI/CD) und Erklärbarkeit, Bereitstellung (Batch, Service, Edge), Überwachung (Datendrift) und kontinuierliches Lernen.

Im Rahmen eines systemischen Ansatzes für MLOps werden auch zusätzliche Themen wie Business-Anforderungen und -ziele, Projektmanagement für ML, Teamstruktur, User Experience sowie die verantwortungsvolle Nutzung von ML-Systemen berücksichtigt.

Der Inhalt umfasst:
- Überblick und Einführung in MLOps und ML-Systeme
- Dateninfrastruktur und Werkzeuge (z. B. Datenmodelle, Versionierung, Speicherung, Verarbeitung, Pipelines)
- Verarbeitung von Trainingsdaten (z.B. Sampling, Labelling, Augmentation, Simulation)
- Feature Engineering
- Modellentwicklungs- und Debugging-Methodik und -Werkzeuge (z. B. Systemvalidierung, Versionierung, Evaluation, Baselines)
- Bereitstellungsinfrastruktur und -werkzeuge (z. B. Batch vs. Online, Modellkomprimierung & Quantisierung, Cloud- & Edge-Bereitstellung)
- Überwachung (z. B. Data Distribution Shift, Ausfälle, Metriken und Protokollierung)
- Kontinuierliches Lernen und Tests in der Produktion
- ML-Projektmanagement und Geschäftsperspektive

Die Übungen beinhalten Beispiele mit modernsten MLOps-Softwaretools und Frameworks sowie reale Anwendungsfälle für ML-Systeme. Dazu gehört auch eine Projektaufgabe, bei der die Studierenden ein prototypisches ML-System entwickeln, das aus einem oder mehreren vorhandenen und geeigneten Werkzeugen besteht und auf einem von ihnen gewählten realen Anwendungsfall basiert. Die Projekte der Studierenden werden am Ende des Semesters im Unterricht präsentiert und fließen zusammen mit der schriftlichen Abschlussprüfung in die Bewertung ein.

Teaching materials

- Vorlesungsfolien, Beschreibungen der Übungen

Supplementary material

- Chip Huyen, “Designing Machine Learning Systems - An Iterative Process for Production-Ready Applications” O’Reilly, 2022
- Noah Gift & Alfredo Deza, “Practical MLOps - Operationalizing Machine Learning Models”, O’Reilly, 2021

- C. Huyen, “CS 329S: Machine Learning Systems Design”, Stanford University, 2021, https://stanford-cs329s.github.io/
- A. Ng, “Machine Learning Engineering for Production (MLOps) Specialization”, DeepLearning.AI and Coursera, 2021, https://www.coursera.org/specializations/machine-learning-engineering-for-production-mlops
- P. Abbeel, S. Karayev, and J. Tobin, “Full Stack Deep Learning - Spring 2021”, UC Berkeley, 2021, https://fullstackdeeplearning.com/spring2021/
- G. Mohandas, “Made with ML: MLOps Course”, 2022, https://madewithml.com/#mlops

- Wissenschaftliche Literatur, die in der Vorlesung behandelt wird

Dieses Modul ist Teil einer möglichen Spezialisierung auf maschinelles Lernen: MLDM1 lehrt die Grundlagen des maschinellen Lernens, MLDM2 (DL) behandelt die Grundlagen des Deep Learning. AI1 führt in verschiedene Paradigmen zur Entwicklung intelligenter Agenten ein (einschließlich bestimmter ungewöhnlicher Arten des maschinellen Lernens), AI2 behandelt die Entwicklung fairer und allgemeiner KI-Algorithmen. CVDL führt in die modernsten Konzepte des fortgeschrittenen Deep Learning für Computer-Vision-Anwendungen ein. Dieser Kurs (MLOps) nimmt eine Systemperspektive ein und stellt Methoden und Werkzeuge für die Entwicklung, die Wartung und den Betrieb von ML-Systemen vor (analog zu Software Engineering und DevOps, die entsprechende Themen in einer Softwareentwicklungs-Spezialisierung abdecken). Alle Wahlpflichtmodule können unabhängig voneinander besucht werden (Ausnahme: CVDL und AI2 bauen auf Grundlagenwissen im Deep Learning auf, wie es z.B. in MLDM2/DL vermittelt wird) und überschneiden sich in weniger als einer Vorlesung.

Prerequisites

Erfolgreicher Abschluss von MLDM1 & MLDM2 (Advanced ML), oder vergleichbare Erfahrung

Teaching language

( ) German (X) English

Part of International Profile

(X) Yes ( ) No

Module structure

Type 2a

Exams

Description	Type	Form	Scope	Grade	Weighting
Final exam	Test (Papier oder Online)	schriftlich	90 Minuten	max. 60 Punkte	60%
Project work (individually or in team)	Projektarbeit	mündlich (Code review)	10 Minuten	max. 20 Punkte	20%

Project work (individually or in team)	Projektarbeit	mündlich (Abschlusspräsentation)	10 Minuten	max. 20 Punkte	20%

Remarks

n/a

Legal basis

The module description is part of the legal basis in addition to the general academic regulations. It is binding. During the first week of the semester a written and communicated supplement can specify the module description in more detail.