Official web page: https://harvard-iacs.github.io/2024-AC215/#course-introduction

AC215: Productionizing AI (MLOps) IMPORTANTâ— - Draft more details soon: Version 4/20/2024 â€‹

Course Introduction

In today's AI-driven world, building a robust deep learning model is only half the journey. The real challenge often lies in bringing this model to life in the form of an application that's scalable, maintainable, and ready for real-world deployment. Welcome to AC215: Productionizing AI (Machine Learning Operations), where we will traverse the complex landscape of Machine Learning Operations, with a special focus on Large Language Models (LLMs). This course has been meticulously curated to provide a holistic understanding of the complete deep learning workflow, from refining your models to deploying them in production environments.

We will dive deep into topics like containerization, cloud functions, data pipelines, and advanced training workflows, with specific emphasis on LLMs. You will learn how to utilize LLM APIs effectively, host APIs, fine-tune LLMs for specific tasks, adapt them to various domains, and build applications around them. Our objective is not only to help you grasp these concepts but also to empower you to build and deploy scalable AI applications. We will delve into the particular intricacies of LLMs and their applications in real-world scenarios.

Whether you are an AI enthusiast wanting to understand the intricacies of Machine Learning Operations or a seasoned professional aiming to fortify your knowledge, this course promises a comprehensive exploration of the production side of AI, with a spotlight on LLM applications and productionizing.

Lectures

- **Location:** SEC 1.321/Winokur, 150 Western Ave, Boston
- **Meeting Time:** Tuesday 12:45 PM - 02:00 PM; Thursday 12:45 PM - 02:00 PM

Technologies and Platforms

We will demonstrate most ideas using TensorFlow and some using PyTorch, alongside utilizing the Google Cloud Platform. Additionally, tutorials will be provided for AWS for reference purposes.

Course Topics Overview

We have designed an in-depth curriculum to ensure a comprehensive understanding of MLOps. Here's a closer look at the topics we'll be covering (see here for a full list of topics):

1. **Introduction:**
    - Begin with an understanding of the importance of MLOps and how it fits in the broader AI and software development ecosystem. â€‹
2. **Virtual Environments and Virtual Machines:**
    - Delve into the foundations of isolated software environments, their importance in AI development, and how virtual machines offer a layer of abstraction over physical hardware. â€‹
3. **Containers:**
    - Understand the concept of containerization using tools like Docker, and how they differ from virtual machines. â€‹
4. **Data Pipelines, Dask, & Cloud Storage:**

- Learn to create efficient data workflows, use Dask for parallel computing, and understand how cloud storage solutions fit into the MLOps ecosystem. â€‹

5. **TF Data and TF Records:**
   - Dive into TensorFlow-specific methods for data ingestion and management, ensuring efficient data preprocessing and storage for your models. â€‹
6. **Data Parallelization:**
   - Grasp techniques for distributing data processing tasks across multiple processors or nodes. â€‹
7. **Data Versioning:**
   - Explore tools like Pachyderm, and understand the significance of maintaining different versions of datasets for reproducibility and model training. â€‹
8. **Advanced Training Workflows:**
   - Deep dive into experiment tracking using tools like Weights & Biases, and harness the power of multi-GPU setups for faster model training. â€‹
9. **Advanced Inference Workflows:**
   - Understand the nuances of model optimization techniques like Distillation, Quantization, and Compression. Explore TensorFlow Lite, monitor your models post-deployment, and be prepared for challenges like data drift. â€‹
10. **Pipeline:**
    - Study end-to-end MLOps pipelines, their components, and best practices to ensure smooth model deployments.. â€‹
11. **App Design, Setup, and Code Organization:**
    - Best practices in designing user-centric AI applications, setting up your development environment, and organizing code for scalability and maintainability. â€‹
12. **APIs & Frontend:**
    - Learn about RESTful APIs to serve your models and design user interfaces for seamless user interactions. â€‹
13. **Scaling (k8):**
    - Delve into Kubernetes, its significance in deploying containerized applications, and understand how to scale your applications to cater to millions of users. â€‹

As we journey through these topics, students will gain a holistic perspective, bridging the gap between model development and real-world deployment. With a blend of theory and practical exercises, this course ensures that by the end, you're not just familiar with these concepts, but proficient in applying them. â€‹

Prerequisites

To ensure a seamless learning experience and to make the most of this course, participants are expected to come with a foundational knowledge in the following areas: â€‹

1. **Programming Proficiency in Python:**
   - A strong command over Python's basic constructs, including functions, classes, and modules. Familiarity with libraries like NumPy, Pandas, Matplotlib is essential, as they form the backbone of many data manipulation tasks in AI. â€‹
2. **Deep Learning Framework - Tensorflow:**
   - A working knowledge of the TensorFlow (or PyTorch) framework is crucial, as many topics will delve into its functionalities and methods. Understanding TensorFlow's basic operations, data

handling, and model building mechanisms will be invaluable. â€‹

3. **Basic Shell Commands:**
    - Comfortability in navigating the command-line interface (CLI), executing shell commands, and performing basic file operations are foundational for many MLOps tasks. â€‹
4. **Basic Data Structures:**
    - A good grasp of Python's primary data structures, especially dictionaries and lists, will be instrumental in understanding and manipulating data. â€‹
5. **File I/O:**
    - Knowledge of basic file input/output operations in Python, including reading from and writing to files, is vital for tasks involving data storage and manipulation. â€‹
6. **General AI and ML Concepts:**
    - While this course is centered around MLOps, a basic understanding of AI and machine learning concepts, including what models are and how they are trained, will set the context for many advanced topics.

It's important to note that while prior knowledge in these areas will provide a solid foundation, the course has been structured to ensure gradual progression. Even if you're not an expert in all of the prerequisites, a willingness to learn and engage actively in the course's hands-on components will be crucial for success. If you find yourself struggling with some concepts, we encourage leveraging the course resources, attending office hours, and participating in peer discussions to reinforce your understanding. â€‹

Course Components

- **Weekly Sessions:** Structured lectures focusing on the core topics.
- **Office Hours:** This is a dedicated period where you can consult with your TF for any questions, clarifications, or guidance you may require for your course project.
- **Assingments**: There will be 4 individual assingments.
- **Team Projects:** Collaborative assignments that culminate in the creation of a fully functional AI app.
- **Discussion Forums:** Platforms for peer-to-peer learning, discussions, and knowledge sharing.
- **Supplementary Readings:** To complement the topics covered in lectures and enrich your academic comprehension, a selection of readings has been curated. As this is an evolving field, the ability to continuously update your knowledge through independent reading is an integral part of the course.

Team Projects: Project-Based Learning: Crafting Your Own AI Solutions

In the dynamic realm of AI and MLOps, hands-on experience is paramount. This course encourages each student to bring a unique perspective by working on self-conceived projects. Here's what you need to know:

1. Crafting Your AI Project:

- Students are expected to conceptualize and develop their own projects. While our teaching staff is here to provide ideas and guidance, the core objective is for each student to nurture and shape their original initiative.
- By the end of the semester, the aim is to transform your idea into a fully functional web-app or mobile application.
- Project Scope: Your project should incorporate some element of modeling, ensuring it aligns with the learning objectives of the course. Moreover, it is essential that every component of the project CAN be evaluable by our teaching staff.
- **Unleash Your Creativity:** Whether you're driven by a start-up vision, by research lab innovations, or inspired by a personal hobby, this is your platform to bring that idea to life.

2. A Guided Demonstration by Pavlos:

- We, the teaching team, will undertake a project that Pavlos proposes throughout the semester. This serves as a demonstration and reference point.
- Each week will spotlight a different facet of Pavlos' project development. This structured showcase offers students a practical insight of course concepts.
- Parallelly, students will be prompted to integrate the week's learnings into their projects, ensuring a steady progression towards their end goals.

3. Milestones and Assessment:

- The course will be punctuated with key milestones, designed to assess your project's evolution and your grasp of the MLOps concepts. Details of these milestones will be shared in due course.
- It's imperative to understand that a significant portion of your grade hinges on these milestones. They are not just checkpoints but pivotal phases that contribute to your project's holistic development and your learning journey.

In Summation:

The heart of this course is experiential learning. We fervently believe that your ideas and paralleling them with structured guidance, we can equip you with the tangible skills essential in today's AI-driven world.

Grade Distribution

Milestone	Weight
MS1	5
MS2	10
MS3	15
MS4	25
MS5	10
MS6	35

For more information about the projects and milestones, you can either click the links provided above or visit the project page.

Course Policies

1. **Getting Help:**
    - **ED Forum:** Post questions related to course content, or technical issues on the ED forum. This encourages peer learning and allows teaching staff to address common concerns. We regularly monitor the forum to provide guidance.
    - **Office Hours:** Attend office hours if you need personalized assistance or in-depth explanations.
    - **Teaching Staff Helpline:** For matters specific to the teaching staff, please send your queries to ac215.2024@gmail.com.
    - **Email the Instructor:** For private or individual concerns, please feel free to directly email the instructor.
2. **Deadline Policy:**

    Consistent and timely completion of assignments is imperative in this course. All course milestones must be submitted by 9:00 PM EST on the specified due dates. You are granted a total leeway of five late days throughout the course duration, with a maximum of two late days allowed for any single milestone.

    Should you need to utilize late days, please inform the class helpline via email at ac215.2024@gmail.com prior to the deadline. This ensures that the teaching team is aware of your situation and can account for it when grading.

Final Milestone: It's important to note that no extensions will be permitted for the final milestone, under any circumstances. Therefore, careful time management is strongly encouraged to ensure that you can meet this critical deadline.

3. **Academic Honesty:**
    - This course places a strong emphasis on ethical behavior. Whether it's ethically handling data or attributing the work of others, students are expected to maintain high standards of integrity.
    - **Acceptable Behaviors:** Discussing course materials, engaging in office hours, debugging with peers, using and citing small portions of code found online, seeking online knowledge, and seeking guidance from tutors.
    - **Unacceptable Behaviors:** Accessing or sharing solutions before submission, plagiarizing, not citing sources of external code or techniques, paying or offering payment for coursework, and sharing course material with future potential students.
    - Engaging in unacceptable behaviors will lead to disciplinary action. When in doubt, always consult the course instructors. â€‹ â€‹
4. **Collaboration & Teamwork:**
    - Collaboration is encouraged, especially for projects. However, ensure you contribute equally and do not divide tasks in a way that prevents you from understanding all parts of the assignment.
5. **Feedback & Evaluation:**
    - Continuous feedback is vital for the learning process. While the course has several grading components, always focus on understanding rather than just marks. Do provide feedback on the course structure, content, and delivery, so we can continually improve.

Accessibility:

We are committed to ensuring that this course is accessible to everyone. If you require special accommodations or have any specific needs, please contact the course administrators as soon as possible.

Adherence to accessibility policies and a commitment to fairness, respect for your learning journey, and consideration for the learning journey of your peers are expected from all students.

Inclusion and Belonging Statement

In this data science class, we strive to create a diverse and inclusive learning environment that respects all identities, including race, gender, class, sexuality, religion, and ability. Our goal is to:

- Advance ethical data science and expose biases in its applications.
- Encourage a variety of thoughts, perspectives, and experiences.
- Be a supportive resource, open to understanding and adapting to your unique needs.

To foster inclusion:

- Please inform us if your name or pronouns differ from official records.
- If something affects your class performance or if you feel uncomfortable with any classroom interactions, reach out to us. You may also find resources at the Harvard Office of Diversity and Inclusion.
- Respect and consideration for diverse backgrounds and perspectives are expected from all participants.
- Your feedback is essential in enhancing diversity, inclusion, and ethics within our class. Feel free to contact us or submit anonymous suggestions.