Productionizing SPIRA's model trainer system

Lucas Quaresma Medina Lam
Roberto Oliveira Bolgheroni

Capstone project proposal
Presented to the Discipline
MAC0499

Supervisors:
Prof. Dr. Alfredo Goldman
M.Sc. Renato Cordeiro Ferreira

São Paulo, March 2024

Contents

Chapter 1

Introduction

The SPIRA project emerged as an initiative to develop a software system capable of performing pre-diagnosis of respiratory insufficiency through speech analysis based on Machine Learning (ML) models. Initially built in 2021 (Casanova et al., 2021), SPIRA's training system underwent a redesign led by Daniel Lawand, aiming to apply good practices in software engineering, with the ultimate goal of implementing a continuous delivery pipeline. Although the redesign was a success, the code base is not in a production-ready state. Particularly, it has no automated tests, and the pipeline wasn't finalized (Lawand, 2023).

The goal of this project is to productionize SPIRA's model trainer system. For that, the aim is to provide comprehensive automated tests and finish the continuous delivery pipeline. Productionizing the model trainer system makes the process of training and deploying new models reproducible, safer and faster through automation. It promotes shorter development cycles, enhances maintainability and the software quality of the project, and enables continuous improvement of the model (Sato et al., 2019).

The following chapters present the specifics of the project. Chapter 2 discusses the background required for understanding the SPIRA project. Chapter 3 outlines the goals of this research, detailing the steps to productionize SPIRA's model trainer system. Finally, Chapter 4 details a step-by-step plan to achieve these objectives, including a timeline for executing it.

Chapter 2

Concepts

This chapter introduces the foundational knowledge necessary for understanding the objectives and the work plan that will be developed in chapters 3 and 4. It will encompass fundamental concepts in engineering and machine learning (ML) that are crucial for comprehending the intelligent systems and processes discussed further in this document. The topics to be covered include Intelligent Systems, Model Training, Model Serving, Machine Learning Engineering, and Continuous Delivery for Machine Learning (CD4ML). These concepts lay the groundwork for the technical aspects and methodologies involved in developing and deploying ML-based systems.

2.1 Intelligent Systems

Intelligent Systems are systems in which the intelligence evolves and improves over time, particularly when its intelligence improves by incorporating feedback of how users interact with the system (Hulten, 2018).

Intelligent Systems connect users to artificial intelligence to achieve meaningful objectives. To connect the user to the intelligence of the system – given it uses machine learning models – two different processes must take place: the training of the machine learning models, and their subsequent serving. These two concepts are further described in the following sections.

2.1.1 Model Training

Training is an essential process that involves using algorithms to learn patterns in available data (Amazon, 2024). This process results in the creation of an ML model, which can make predictions or decisions based on input data. Model training can be iterative. It involves adjustments to algorithm parameters to optimize the model's performance regarding the available data.

2.1.2 Model Serving

Serving refers to the process of making trained models available for use in real-world applications (MarkovML, 2023). Once a model is trained, it needs to be deployed so that it can receive input data and generate predictions. Model serving involves setting up the underlying infrastructure to host the model and handle incoming requests. It also includes monitoring the model's performance in production, as well as ensuring scalability and reliability.

2.2 Machine Learning Engineering

Machine Learning Engineering involves a set of practices and techniques for the development, implementation, maintenance, and updating of ML-based systems (Wilson, 2022). This includes strategic project planning, selection of suitable algorithms, data collection, data preparation, model deployment, and continuous monitoring. This approach allows stakeholders to benefit from adopting machine learning techniques.

2.3 Continuous Delivery for Machine Learning

Continuous Delivery for Machine Learning (CD4ML) is a software engineering approach in which a cross-functional team produces machine learning applications in small and safe increments that can be reproduced and reliably released at any time, in short adaptation cycles (Sato et al., 2019).

CD4ML is the application of the Continuous Delivery concept for Machine Learning applications. It differs from the original concept because ML applications have three axes of changes: data, model, and code. To achieve Continuous Delivery in these systems, changes in any of the three axes have to be addressed.

The concept is materialized in the form of a pipeline that streamlines processes such as data collection and preparation, model training and validation, performance evaluation, and automated deployment. This enables rapid updates and improvements to models, thus enhancing the effectiveness and relevance of ML systems.

2.4 SPIRA

SPIRA is a research project initiated during the COVID-19 pandemic with the aim of developing a system capable of performing pre-diagnosis of respiratory insufficiency, including the symptom of silent hypoxia. This system is based on speech analysis using Machine Learning models (Ferreira et al., 2022).

Chapter 3

Objectives

The aim of this project is to productionize SPIRA's model trainer system. This will be achieved once the following goals are accomplished (Lawand, 2023):

1. **Comprehensive Automated Testing**
   It entails the establishment of a robust suite of tests covering various aspects of the model trainer system, employing unit, integration, component, and end-to-end tests. Moreover, Test-Driven Development (TDD) will be applied to the rewriting of the code base. Beyond promoting good code design and maintainability, a comprehensive automated testing suite is critical to building a reliable delivery pipeline. It prevents regressions, and provides fast, automated feedback, shortening the development cycle.

2. **Continuous Delivery Pipeline**
   It entails the integration of the trainer system with a continuous delivery workflow. Machine learning applications have three axes of change: data, model and code (Sato et al., 2019). The continuous delivery pipeline should be triggered whenever any of those axes changes, executing the automated deployment of the generated models to SPIRA's Inference Service.

Chapter 4

Work Plan

The following sections describe the steps to be executed to achieve the goals set by chapter 3.

1. **Study of Machine Learning Engineering concepts**
   This step involves gathering and reviewing available literature in the scope of Machine Learning Engineering.

2. **Study of SPIRA's Training System redesign**
   This step involves conducting a comprehensive analysis of the redesign implemented in SPIRA Training System, led by Lawand (2023). The focus will be on identifying areas for enhancement and integrating automated testing methodologies.

3. **Redevelopment of SPIRA's Training System**
   This step involves redeveloping the SPIRA Training System, incorporating the findings from the redesign study. Emphasis will be placed on implementing automated testing, potentially employing Test-Driven Development (TDD) practices. Related to objective 1.

4. **Integration with model repository**
   This step involves integrating the model trainer system with a central model repository, which stores, versions, and retrieves models, streamlining the model lifecycle management process. Related to objective 2.

5. **Integration with task scheduler**
   This step involves integrating the model trainer system with a task scheduler, which triggers training and deployment of new models based on specific events. This ensures that model updates and deployments occur promptly and seamlessly, minimizing manual intervention and streamlining the overall workflow. Related to objective 2.

6. **Writing the monograph**
   This step involves writing the monograph that contains details about the project's methodology, findings, and outcomes.

7. **Creation of the Presentation and Poster**
   This step involves the creation of the presentation and poster required for the course.

Task	**Months**										
	Jan	**Feb**	**Mar**	**Apr**	**May**	**Jun**	**Jul**	**Aug**	**Sep**	**Oct**	**Nov**
Task 1	X	X									
Task 2		X	X								
Task 3			X	X	X	X					
Task 4						X	X	X			
Task 5							X	X	X		
Task 6								X	X	X	X
Task 6										X	X

Table 4.1: *Planned Tasks vs. Months*

Bibliography

AWS Docs Amazon. Training ML Models, 2024. URL https://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html. 2

Edresson Casanova, Lucas Gris, Augusto Camargo, Daniel da Silva, Murilo Gazzola, Ester Sabino, Anna S. Levin, Arnaldo Candido, Sandra Aluisio, and Marcelo Finger. Deep learning against covid-19: Respiratory insufficiency detection in brazilian portuguese speech. 2021. doi: 10.18653/v1/2021.findings-acl.55. 1

Renato Ferreira, Dayanne Gomes, Vitor Tamae, Francisco Wernke, and Alfredo Goldman. Spira: Building an intelligent system for respiratory insufficiency detection. In **Anais do II Workshop Brasileiro de Engenharia de Software Inteligente**, pages 19–22, Porto Alegre, RS, Brasil, 2022. SBC. doi: 10.5753/ise.2022.227048. URL https://sol.sbc.org.br/index.php/ise/article/view/22530. 3

Geoff Hulten. **Building Intelligent Systems: A Guide to Machine Learning Engineering**. Apress, 2018. ISBN 978-1484234310. 2

Daniel Lawand. Enabling MLOps in the SPIRA Training Pipeline, 2023. URL https://github.com/danlawand/MAC0499/blob/main/docs/monograph.pdf. 1, 4, 5

MarkovML. Model Deployment, 2023. URL https://www.markovml.com/blog/model-deployment. 2

Danilo Sato, Arif Wider, and Christoph Windheuser. Continuous delivery for machine learning. **Martin Fowler**, 2019. 1, 3, 4

Ben Wilson. **Machine Learning Engineering in Action**. Manning, 2022. ISBN 978-1617298714. 3