Comparing Off-the-Shelf Deployment Options for Machine Learning Models

AISA Study Project

Organization

Examiner:	Prof. Dr. Maria Wirzberger, Prof. Dr. Stefan Wagner
Supervisor:	Markus Haug
Participant(s):	
Time frame:	

Context & Motivation

In recent years, there has been a growing interest in deploying "smart systems", which are able to emulate human behavior on specific tasks [1]. Typically, these systems employ artificial intelligence (AI) techniques, frequently machine learning (ML), to achieve their human-like behavior.

When deploying the system, we must also deploy the ML model as part of the system. While implementing a bespoke solution is always an option, recent years have also seen a rise in the number of off-the-shelf deployment options [2], e.g., TensorFlow Serving[1] or MLflow[2].

This growing competition could overwhelm practitioners looking for a solution to deploy their ML models, for example, in the context of a make-or-buy decision. A comparative analysis of off-the-shelf deployment options across well-reasoned criteria would help them gain clarity about their needs and identify suitable options that fulfill these needs.

[1] https://github.com/tensorflow/serving
[2] https://mlflow.org/

Objectives

In this AISA study project, participants should conduct a comparative analysis of deployment options for ML components. For this purpose, participants should identify existing deployment options, such as tools or platforms. Participants should also create a list of criteria for comparison, such as requirements on the deployed ML component or support for advanced deployment strategies such as green-blue deployments.

Methods

As the basis for the comparison, a very lightweight, non-systematic literature survey approach [3–5] should identify criteria and off-the-shelf deployment options. This literature survey should include grey literature since many deployment options may not have been part of the academic discussion before. Grey literature could also identify prior work on comparing deployment options, which could serve as the foundation for the rest of the study.

Based on the literature survey, a catalog of criteria should be synthesized. The catalog should then be employed to analyze the documentation of each deployment option and synthesize a comparison of the identified options [6, 7].

The catalog of criteria and the results of the comparison should be documented in a scientific report and presented as a poster.

References

[1] Saleema Amershi et al. "Software Engineering for Machine Learning: A Case Study". In: *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*. Montreal, QC, Canada (May 25–31, 2019). Montreal, QC, Canada: IEEE, May 2019, pp. 291–300. ISBN: 978-1-7281-1761-4. DOI: 10.1109/ICSE-SEIP.2019.00042.

[2] Sherif Akoush et al. "Desiderata for next generation of ML model serving". In: (Oct. 26, 2022). DOI: 10.48550/ARXIV.2210.14665. arXiv: 2210.14665 [cs.LG].

[3] Bruno Cartaxo, Gustavo Pinto, and Sergio Soares. "Rapid Reviews in Software Engineering". In: Springer International Publishing, 2020, pp. 357–384. DOI: 10.1007/978-3-030-32489-6_13.

[4] Hannah Snyder. "Literature review as a research methodology: An overview and guidelines". In: *Journal of Business Research* 104 (Nov. 2019), pp. 333–339. DOI: 10.1016/j.jbusres.2019.07.039.

[5] Richard J. Torraco. "Writing Integrative Literature Reviews: Guidelines and Examples". In: *Human Resource Development Review* 4.3 (Sept. 2005), pp. 356–367. DOI: 10.1177/1534484305278283.

[6] Philipp Mayring. *Qualitative content analysis: theoretical foundation, basic procedures and software solution*. Tech. rep. Klagenfurt, 2014.

[7] Carolyn B. Seaman. "Qualitative Methods". In: *Guide to Advanced Empirical Software Engineering*. London: Springer London, 2008, pp. 35–62. ISBN: 0761948880. DOI: 10.1007/978-1-84800-044-5_2.