COMS 6156 - Topics in Software Engineering
Spring 2024 - Project Progress Report
Mike Segal | ms6135

Enhancing MLOps with Kubeflow, GitHub Actions, AutoML, and Shap

1. Overview:

Our project focuses on developing a full MLOps pipeline using Kubeflow, incorporating GitHub Actions for pipeline triggering, automating model training using Katib, and conducting data quality and model interpretability checks.

To demonstrate the applicability of MLOps to different data types and modalities (Computer Vision and NLP), we are using the Squad and ImageNet Tiny datasets. In order to develop a statistical baseline to compare the efficiency of the MLOps pipeline, we are first training the models without the use of Kubeflow, and are conducting data quality and model interpretability checks on the outputs.

We are then using Kubeflow to break the data management and ML training process into components, and are arranging them into a pipeline that is then orchestrated using Kubernetes. Katib will be utilized in some of the trials to determine if automated ML training is feasible for all modalities, and whether the quality meets the established benchmarks for the datasets. The MLOps pipeline will also incorporate the data quality and model interpretability checks.

We anticipate the final deliverables to include detailed documentation, rigorous statistical analysis of each part of the process, and a well organized codebase that enables reproducibility by engineering teams.

2. Current Status:

The Squad and ImageNet data management code has been successfully developed and tested, the basic ML models that don't use the MLOps pipelines have been trained, and Shap has been successfully used to gain model interpretability information. We have developed code for simple data quality checks for both datasets, but have not yet implemented SodaCore for Pandas.

Additionally, the MLOps component and pipeline code for both datasets has been written, but has not yet been deployed or tested. We will complete deployment and testing this week, and will ensure that all aspects of the pipeline are running as intended, which include the data management with data quality checks, model training (both with and without Katib), and GitHub Actions triggering of the pipeline.

3. Research Questions

1. Do MLOps pipelines accelerate ML model development?
2. Does MLOps pipeline standardization decrease environment management time for engineering teams?
3. Can model explainability be incorporated into an MLOps pipeline as a form of quality control?
4. What is the best way to include data quality checks into an MLOps pipeline?
5. Does automated ML training using Katib yield effective ML models compared to known benchmarks?

4. Value to User Community

Since our overall goal is to develop an MLOps pipeline that incorporates Kubeflow (an open-source platform that allows container orchestration for ML using Kubernetes), data quality checks, model explainability using Shap (an open source package that allows for the visualization of the model decision process), and automated ML training using Katib, the target audience for the project are ML engineering teams and engineering project managers.

Our research into the MLOps field showed nascent progress in standardizing MLOps, and while some projects have successfully deployed MLOps pipelines using Kubeflow, they lacked data quality checks and model interpretability components. The intent behind this project is to therefore demonstrate the feasibility of such an approach, provide a codebase that can function as a template, and clearly document the process and outcomes of the MLOps pipeline.

Furthermore, by using both NLP and Computer Vision datasets coupled with model interpretability, we intend to demonstrate that MLOps with Shap is data agnostic, and can be applied to multiple machine learning modalities.

We hope that the end result will serve to convince engineering teams to invest in standardizing their MLOps infrastructure. The documentation, statistical analysis, and codebase should enable engineering managers to evaluate their own problem sets and determine if utilizing MLOps would be a viable option for them. Since the project repo will be made public, there should be no limitations in utilizing it as a reference for their own deployments.

5. Demo

The demo will include an overview of how the pipelines are configured in the codebase, how the pipelines are visualized in the Kubeflow interface, and the results of the pipeline runs. We will also show how we perform the data quality checks, how the model interpretability looks for the NLP and CV models, and how well the automated ML training using Katib does in training the models.

6. Delivery

We will publish a public GitHub repo that will include all of the code and documentation for the project.

Additional Information:

Platforms

Kubeflow
https://www.kubeflow.org/

Kubeflow Katib
https://www.kubeflow.org/docs/components/katib/hyperparameter/

Datasets

Tiny ImageNet
https://huggingface.co/datasets/zh-plus/tiny-imagenet

Stanford Question Answering Dataset (SQuAD)
https://huggingface.co/datasets/rajpurkar/squad

Packages

SHAP
https://shap.readthedocs.io/en/latest/
https://github.com/shap/shap

Soda-Core
https://github.com/sodadata/soda-core