A MAPE-K-Based Cost-Aware Retraining Framework for Sustainable Machine Learning Operations

Abstract

Machine learning (ML) models have become an integral component of traditional software systems to support automated decision-making. For managing these models in production, organizations increasingly adopt Machine Learning Operations (MLOps), which includes tasks like automated data processing, model training, deployment, monitoring, and maintenance. Retraining the models is an essential step of the maintenance phase to keep the models updated and capable of making accurate predictions. While many MLOps tools offer retraining automation, these tools still use static retraining schedules. The static schedules can sometimes cause unnecessary retraining of the models. However, retraining is resource and cost intensive. Hence, this leads to a waste of computational resources and increased that can dynamically decide when to trigger retraining.

Prior work suggests that the incorporation of machine learning into traditional software systems introduces new maintenance challenges. This happens due to the models relying on data, configurations, and external libraries that are not easily manageable. Even though the continuous integration and deployment practices have improved automation in ML pipelines, they also contribute to the increased frequency of training and evaluation of models. Multiple studies have shown that training and retraining of ML models lead to enormous expenditure of computational power and energy, raising safety concerns from economic and ecological perspectives.

To address these challenges, this work proposes a self-adaptive MLOps framework based on a monitoring and decision-making loop. The proposed system monitors model performance, input data patterns, and metrics such as resource usage over time. If changes occur, it determines whether the currently deployed model can still be relied upon or whether retraining is required. To do so, the proposed framework chooses an appropriate action, such as reusing an existing model, switching to another saved version, or model retraining based on the inherent analysis.

The proposed framework will be evaluated using real-word industry MLOps projects. The evaluation focuses on system-level behavior and examines performance trends, adaptation frequency, and changes in resource usage over time under varying data conditions. By reducing unnecessary retraining and improving decision-making, this research aims to support more reliable, cost-effective, and sustainable machine learning systems in production.