Edge MLOps: An Automation Framework for AIoT Applications

The paper and its findings

- This paper explores ML ops and various entities inferred with it such as ML pipelines, cloud orchestration, and Edge inference.
- It proposes a complete lifecycle for automating AIoT workload by combining the above-mentioned pipelines which are either executed separately or connected together.
- The framework proposed was successfully implemented in a real-life scenario where the goal of deploying ML models to forecast the air quality of rooms directly on the edge was achieved.
- Edge MLOps is scalable to multiple edge devices and is confined to the infrastructure, the framework proposed here is stable and it is an automatic way to train models on the cloud and deploy them on edge devices resulting in stability and robustness of the pipeline.

Strengths

- Cost Efficiency - From the assessed cost, energy, and operational efficiency of edge vs cloud machine learning based on the experiment, there was an overall reduction in cost observed on the edge device when compared to the cloud environment.
- Data Storage Efficiency - Utilising this framework, only 22% of the incoming data from the IoT devices were sent to the cloud for storage, reducing the storage costs in the cloud and improving data quality by filtering the necessary and resource efficiency.
- Stability of the framework - Using the proposed framework, on evaluating the robustness, it was observed that the triggers were executed in a random fashion without any failure and the data collection and storage pipeline worked without any interruptions. The framework is a stable and automatic way to train models without human interruptions, mitigating the risk of errors due to human intervention.

Weakness

- Lacks Data Privacy and Security - The paper lacks security and privacy issues especially when dealing with sensitive information such as healthcare data. The edge devices and the cloud could be vulnerable to various security threats such as data stealing, tempering, or data breaches.
- Scalability - The experiment here conducted was on a countable number of devices but it does not discuss or provide us with an idea if the number of IoT devices would be extremely large and in such cases, how the system performs, its reliability, network constraints, or bandwidth issues.

Open issues

- There may be exceptions encountered in the long run and a plan is required to handle those exceptions and unexpected behaviors. Along with this, to ensure robustness, testing should be performed in the DEV as well as TEST environment before deploying in the PROD environment.
- For our project to deal with healthcare data, we need to ensure federated learning. Federated learning is performing machine learning in a collaborative fashion. This requires to modify the existing framework to support decentralized data storage and training.

The paper's relation to the objectives of our project

- We can implement a similar ML pipeline for our project to analyze data and handle the process.
- Since our project consists of medical data as well, we can utilize the ideology of processing the data at the edge so it is less prone to attacks at the cloud, providing security.
- A similar blueprint can be followed for the Edge MLops for our project and we can combine cloud and edge to operate the ML models reducing cost and enhance efficiency.