CIS-490H Edge Computing

With Dr. Zheng Song

Paper Review: Week 5

"LAVEA: Latency-aware Video Analytics on Edge Computing Platform"

Student: Demetrius Johnson

01 February 2023

1. Summary.

(1) Motivation

The primary motivation for the Latency-aware Video Analytics on Edge Computing Platform (LAVEA platform) is to address the classical cloud problem that we face in the modern day: reduce strain on the cloud and the bandwidth necessary for data to reach the cloud to be processed and sent back to edge devices. Particularly, the LAVEA system aims to achieve lower-latency video analysis results as the gain made from reducing the strain on the cloud and utilizing edge resources. This will provide edge devices that rely on vision systems and video analytics with higher response times like in traffic or security applications involving cars, drones, and highway cameras (as some examples).

(2) Contribution

The research provides analytics that compare various performance metrics (such as bandwidth or round-trip-time) for different edge server setups at a base station, such as wired, 5G or 2.4G. Then, it provides a system that can be implemented that will manage tasks from clients on the edge and be distributed to other clients and edge servers. They give an optimization problem and solve it so that their system will run in such a way as to optimize the task management and execution. Finally, they provide results from a system they built to test their mechanism and optimization techniques, including cloud-only analysis, cloud and edge analysis, client only analysis, client and edge analysis, and other combinations to demonstrate the performance differences of different edge schemes that employ their generalized task management system (specifically video analysis tasks).

(3) Methodology and/or argument

They implement Function as a Service (Faas) through Docker images in order to support a serverless task management application that can be run on edge servers or cloud servers and require less programming and easy portability between different types of server nodes.

(4) Conclusion

Overall, the research does an excellent job at describing the motivation, designing a model, testing that model on a real-world application and demonstrating that their work towards edge computing applications can make a very meaningful impact. They consider different configuration scenarios of edge and cloud collaboration to make their findings more inclusive of current day analysis of edge computing.

2. Critique.

On some of the charts and graphs, they do not explain in the text or through the graph what some abbreviations refer to (i.e. figures 2 and 3 do not show what "ec2 east" or "ec2 west" refers to). While I was able to gather enough relevant information from the graph, it is really bad practice to have anything on any charts or diagrams that do not have an explanation of axis and a key or some place where abbreviations are described and given. They also used the abbreviation "CDN" in one paragraph, one of which I am still unsure of what it is referring to. Also, they did not mention at all any security concerns, however I understand that they are not focusing on security and are mainly trying to demonstrate another powerful application of edge computing.

3. Synthesis.

I would recommend that this study be duplicated in many edge computing application so that we can measure the performance in the same way that the paper did for license plate recognition using vision systems. For example, we could test performance of facial recognition, or some practical application that suffers from higher latency and requires better performance over using the cloud.