Internship Proposal 2020

Title: Virtualizing Camera Functions for Distributed Live Video Analytics

Hosting laboratory: LISTIC, Polytech Annecy-Chambéry, 5 chemin de bellevue, Annecy-le-Vieux, 74940 ANNECY

Mentors:
Francesco Bronzino, Assistant Professor
Université Savoie Mont Blanc, Annecy-le-Vieux
francesco.bronzino@univ-smb.fr

Shubham Jain, Assistant Professor
Stony Brook University, Stony Brook, NY, USA
jain@cs.stonybrook.edu

Description. The rapid increase in cameras across our homes and cities, renders video processing a crucial tool in analyzing video streams to gather spatio-temporal analytics. Video analytics processing pipelines are often composed of several computer vision functions that are executed in order to achieve a certain objective. However, popular vision techniques that process this data, rely on predefined pipelines that execute on a single compute machine, often in the cloud, and are limited by the availability of alternate resources. The dependency of pipeline execution on a single compute resource has slowed down the deployment of camera-based analytics services.

To address this problem, research efforts have explored approaches to reduce the computational burden making better use of the scarce resources available. One solution proposed parallelization of a single camera resource to support multiple applications via view virtualization and mobility-aware scheduling [2]. Another effort also investigated exploiting hierarchical edge-clouds solutions to support low latency processing for query driven video analytics [3]. The challenge with this approach is that edge-clouds have by nature very limited computational resources. Relying on remote central cloud resources as the sole processing backup might quickly compromise real-time requirements. Acknowledging this challenge, recent work has shown the promise of using multiple edge resources in concert [4], by means of functionally splitting processing pipelines into distributed heterogeneous computing resources with the goal of minimizing processing latency [5].

The proposed internship builds on the premise of functionally splitting video analytics pipelines into smaller application functions deployable across heterogeneous edge resources. The goal of this internship is to design and implement solutions aimed at video analytics pipelines, and decompose them into functions that can be connected online in near real-time. These functions may reside on different compute resources and may be executed concurrently.

Internship goals and activities. The student will analyze common video analytics pipelines to identify core functional components. The main task will be to understand the order of execution of pipeline functions based on common frameworks (e.g., OpenCV, YOLO, or Tensorflow), construct function chains, and optimize their computation across a set of distributed compute nodes, in real-time, under given constraints. These nodes may be spatially separated and may have varying compute capabilities. Each function will be matched to the appropriate compute node based on resource, latency, and accuracy requirements. Finally, the student will use a network emulator framework implemented in python to evaluate the proposed techniques. This internship track focuses on developing theoretical approaches for addressing pipeline scheduling and resource optimization, and putting them into practice. If time allows, the student will conclude the work by moving to production level software, e.g. Apache Storm [1].

Candidate Requirements.

- The candidate should be a 2nd year Master student (or a 3rd year student of "cycle d'ingénieur").
- Comfortable speaking English (French is not required).
- Proficiency with at least one programming language, preferably Python or Golang.
- Knowledge of computer networks protocols and systems (e.g., packet processing, function virtualization).
- Knowledge of computer vision concepts is a plus.

References

[1] Apache storm. `https://storm.apache.org/`, 2020.

[2] S. Jain, V. Nguyen, M. Gruteser, and P. Bahl. Panoptes: Servicing multiple applications simultaneously using steerable cameras. In *Proceedings of the 16th ACM/IEEE International Conference on Information Processing in Sensor Networks*, pages 119–130, 2017.

[3] J. Jiang, G. Ananthanarayanan, P. Bodik, S. Sen, and I. Stoica. Chameleon: scalable adaptation of video analytics. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication*, pages 253–266, 2018.

[4] S. Maheshwari, D. Raychaudhuri, I. Seskar, and F. Bronzino. Scalability and performance evaluation of edge cloud systems for latency constrained applications. In *2018 IEEE/ACM Symposium on Edge Computing (SEC)*, pages 286–299. ACM, 2018.

[5] W. Zhang, S. Li, L. Liu, Z. Jia, Y. Zhang, and D. Raychaudhuri. Hetero-edge: Orchestration of real-time vision applications on heterogeneous edge clouds. In *IEEE INFOCOM 2019-IEEE Conference on Computer Communications*, pages 1270–1278. IEEE, 2019.