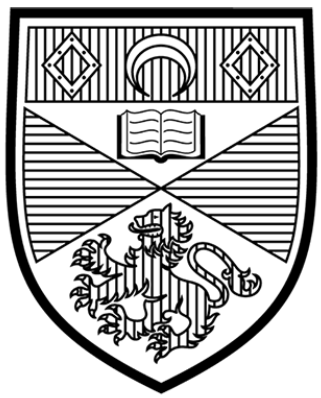
CS5052 – Data Intensive Systems

220031985



University of
St Andrews

6th April 2023

"Chameleon: scalable adaptation of video analytics"

Junchen Jiang†◦, Ganesh Ananthanarayanan◦, Peter Bodik◦, Siddhartha Sen◦, Ion Stoica⋆

†University of Chicago ◦Microsoft Research ⋆UC Berkeley, Databricks Inc.

INTRODUCTION

As stated by the authors, applying deep convolutional neural networks (NN) to video data at scale poses a significant system problem. In general, video analytics involves the use of algorithms to process and analyse video data, such as detecting objects and tracking movements. However, video analytics can be computationally expensive, especially for the real-time processing of high-resolution video streams.

Traditional video analytics solutions, as the authors argue state that they are frequently designed for set hardware and software configurations, which can result in inefficient resource consumption, poor scalability, and reduced accuracy.

To address this issue, the authors present **Chameleon**, a controller that dynamically selects the ideal settings for NN-based video analytics pipelines.

In theory, routinely modifying configurations can minimise resource usage with little loss of accuracy, however, exploring a wide space of configurations regularly incurs an overwhelming resource cost that negates the benefits of adaptation. Chameleon determines the optimum configuration that allows for adequate temporal and spatial correlation to amortise the search cost over time and numerous video feeds.

BACKGROUND

A typical video analytics application consists of a **pipeline** of video processing modules. There are various "knobs" in the pipeline, such as frame resolution, frame sampling rate, and detector model (e.g., Yolo, VGG or AlexNet). The authors refer to the various knob combinations as a **configuration**. The choice of configuration impacts both the **resource consumption** and **accuracy** of the video application.

The **"best"** configuration is defined as having the lowest resource usage while maintaining accuracy above a certain threshold. **Accuracy standards** are specified by the applications, and configurations that achieve that threshold can often differ by many orders of magnitude in terms of resource needs, and choosing the cheapest among them can have a substantial impact on computing costs.

Another best configuration for a video analytics pipeline also *varies from time to time*, often at a timescale of minutes or even seconds.

The Chameleon System is designed to be modular and flexible allowing it to adapt and provide the best configuration for the computation, the way Chameleon is built is that it consists of the following parts :

- **A video processing pipeline**
- **A resource manager**
- **An adaptation module**

The video processing pipeline acquires and processes visual input, such as object detection and tracking. The resource management layer is in charge of managing cloud resources, such as provisioning and de-provisioning virtual machines as needed. The adaptability module monitors the workload and dynamically adapts the processing settings and resource allocation to optimise performance and cost-effectiveness.

Chameleon's adaption approach is broken into two stages: offline profiling and online adaptation.

During the offline processing phase, the pipeline is developed by running the video analytics application in a variety of input and resource scenarios. This data is utilised to train the performance model and provide an asset of application performance metrics.

The resource manager continuously checks the available resources and the input conditions during the online adaption phase and uses the performance mode to find the appropriate resource allocation for the current conditions. The application module then modifies the application configuration accordingly to optimise performance.

NOVELTY

The solutions provided in the paper are novel. The authors introduce a system that takes advantage of resource flexibility to achieve scalable and adaptable video analytics. The system uses adaptive video processing techniques in conjunction with cloud-based resource management to dynamically modify processing parameters and resource location based on workload requirements.

While earlier efforts to address the issues of video analytics have been made, the authors contend that their approach is unique in numerous aspects. According to the authors, chameleon incorporates resource elasticity into the video processing pipeline, allowing the system to adapt to changing workload needs and enhance cost-effectiveness.

EVALUATION

The authors evaluate Chameleon using three different video analytics applications: Object detection, face recognition and human tracking and more specifically by using live video feeds from five real traffic cameras and comparing the system to a baseline that picks the optimal configuration offline.

Results showed that Chameleon can achieve 20-50% higher accuracy with the same amount of resources, or achieve the same accuracy with only 30-50% of the resources (2-3× speedup). To generalize these results, the authors used a different set of videos taken from 10 cameras deployed in an indoor cafeteria area over a period of 3 days.

After sampling 90 video clips, 9 clips from each camera, across the 3 days. The original MP4 videos are 1920×1080 in resolution and 25 fps in frame rate. Their content includes different patterns of human movement, e.g., more people moving before/after meal times than the rest of the day.

The authors also conducted experiments to evaluate the impact of spatial and temporal correlations on resource-accuracy trade-offs. Results show that these correlations have a significant impact on resource consumption and inference accuracy.

Finally, they presented a suite of techniques to dramatically reduce the cost of periodic profiling by leveraging spatial/temporal correlations. These techniques can help reduce resource consumption while maintaining high levels of accuracy.

CONCLUSION

The authors' detailed examination of the Chameleon framework utilising both simulated and real-world video analytics workloads is one of the paper's highlights.

The paper's minimal consideration of potential constraints and obstacles in using the Chameleon framework in real-world contexts is one potential drawback. The authors, for example, do not address any security threats or privacy concerns that may occur when employing a cloud-based video analytics system.

Overall, the Chameleon framework is an important contribution to video analytics and has the potential to enable more efficient and scalable video analytics in a variety of real-world applications. The evaluation results are convincing, and the paper is well-written. More study is needed, however, to address potential constraints and obstacles in using the Chameleon framework in real-world contexts.

TOTAL WORDS: 986