


Realizing Video Analytic Service in the Fog-Based Infrastructure-Less Environments

Qiushi Zheng 


School of Software and Electrical Engineering, Swinburne University of Technology,
Melbourne, Australia
qiushizheng@swin.edu.au

Jiong Jin 

School of Software and Electrical Engineering, Swinburne University of Technology,
Melbourne, Australia
jiongjin@swin.edu.au

Tiehua Zhang 

School of Software and Electrical Engineering, Swinburne University of Technology,
Melbourne, Australia
tiehuazhang@swin.edu.au

Longxiang Gao 

School of Information Technology, Deakin University, Melbourne, Australia
longxiang.gao@deakin.edu.au

Yong Xiang 

School of Information Technology, Deakin University, Melbourne, Australia
yong.xiang@deakin.edu.au

Abstract

Deep learning has unleashed the great potential in many fields and now is the most significant facilitator for video analytics owing to its capability to providing more intelligent services in a complex scenario. Meanwhile, the emergence of fog computing has brought unprecedented opportunities to provision intelligence services in infrastructure-less environments like remote national parks and rural farms. However, most of the deep learning algorithms are computationally intensive and impossible to be executed in such environments due to the needed supports from the cloud. In this paper, we develop a video analytic framework, which is tailored particularly for the fog devices to realize video analytic service in a rapid manner. Also, the convolution neural networks are used as the core processing unit in the framework to facilitate the image analysing process.

2012 ACM Subject Classification Computing methodologies → Object detection

Keywords and phrases Fog Computing, Convolution Neural Network, Infrastructure-less Environment

Digital Object Identifier 10.4230/OASICS.Fog-IoT.2020.11

Funding This work was supported in part by the Australian Research Council Linkage Project under Grant LP190100594.

1 Introduction

The rapid development of artificial intelligence, especially deep learning, has shown superior performance in many fields like visual recognition, big data analysis, and natural language processing over the last decade [8]. The convolution neural network (CNN), one of the most preferred deep learning structure, stands out owing to its outstanding performance in data filtering and recognising processes [5]. However, the increasing need for the computational resources for CNN is accompanied by the growing demand for infrastructure support. Specifically, starting at the early LeNet-5 model up to the state-of-the-art InceptionV4 model,



© Qiushi Zheng, Jiong Jin, Tiehua Zhang, Longxiang Gao, and Yong Xiang;
licensed under Creative Commons License CC-BY

2nd Workshop on Fog Computing and the IoT (Fog-IoT 2020).

Editors: Anton Cervin and Yang Yang; Article No. 11; pp. 11:1–11:9

OpenAccess Series in Informatics



OASICS Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

the accuracy of CNN in image recognition has been improved steadily, yet the computational cost in running the model has also been witnessed a significant increase. Therefore, the use of cloud data-center to execute the CNN model is often considered as the most reliable option in order to tame the complex CNN model.

Alternatively, fog computing, introduced by Cisco and widely used in the edge computing context, is expected to provide the architectural support for various Internet of Things (IoT) services while alleviating the excessive dependence on the cloud [2]. The capability of bringing computation, storage and networking function in the proximity of users thus attracted attention for researchers and industrial practitioners who wish to provision time-sensitive services to the end-users in the IoT environment. Recently, many works have focused on utilizing the fog nodes to implement deep learning for some data-driven services. For instance, Li et al. [6] proposed an approach to offload the first few layers of CNN on the fog nodes for reducing the network traffic from end devices to cloud servers. On the other hand, Ran et al. [12] proposed a framework based on the service requirements to cope with deep learning tasks in either local fog nodes or the remote cloud. Specifically, fog normally plays a cooperative and complementary role with the cloud instead of as a substitute, which means many deep learning approaches still need to be cloud-assisted. However, the cloud is unavailable to cover all deployment scenarios, like the infrastructure-less environment where disrupted or even no electrical grids and cellular networks are the norms.

Taking the national park scenarios in Australia as an example, as the national parks and agrarian business occupy 5.77% and 51% of the Australia land separately [9, 10] while the number is continuing to grow, but only 31% of Australia land has Internet coverage [4] not to mention any form of connection to the electrical grids. The new challenges are then interpreted as the infrastructure-less environments, and it is impractical to continue relying on the cloud due to the unreliable internet connection and unsustainable power supply. Affected by the environmental limit, the implementation of any lightweight application is difficult, not to mention a system that is capable of providing complex video analytic services. Under this condition, infrastructure-less environments urgently need a framework to perform an intelligence video analytic service for protecting the environment, animals and human properties. Therefore, if the service can be realized, a long-term environment preserving strategy could be rolled out by analyzing the relevant information such as the classes of species being recorded or the environmental impact on wildlife.

In short, the contributions of this paper are as follows:

- We identify four key factors that have significant impacts on offloading video analytic services from the cloud to independent fog nodes.
- We propose a fog-based video analytic framework in infrastructure-less environments. The framework achieves a fast running speed by effectively reducing the number of frames to be processed without adversely impacting the accuracy of the results.
- We utilize the Siamese network structure to design a decision approach that is able to process the real-time continuous images based on the similarity efficiently. Consequently, the required computing capability of fog nodes is largely reduced when execution video analytic services.

2 Service Requirements

Recently, some researchers have committed to finding a suitable energy-sustainable fog system that can operate stably for a long period of time while providing valuable information to users in this challenging environment [16]. It is inspiring to provide time-sensitive services

in infrastructure-less environments, but the deep learning-based intelligent services are not provided due to the limitation of the computing capability of fog node. Specifically, video analysing, as one of the widely used fields of deep learning, is considered critical in terms of delivering intelligent services in cases like remote national parks and rural farms in order to achieve the automatic collection and analysis of surrounding information.

Currently, You Only Look Once(Yolo) and Single Shot MultiBox Detector(SSD) are two of the most popular choices to achieve analysing services in most cases. Specifically, Yolo is a framework mostly used for real-time video streaming and demonstrates an excellent performance on boundary detection and object recognition [13]. Meanwhile, SSD plays a significant role in the video analytic area [7]. The essence of these two methods is using bounding boxes to capture many small pieces from the original picture, and then produce feature maps of different sizes through convolution. Finally, each map can be used to predict the targets whose center points are in the small square, which can obtain high recognition accuracy in most working scenarios. These approaches demonstrate a substantial reduction in the calculations compared with window sliding. However, even the light version of Yolo, namely Yolo-tiny, is still considered unworkable on fog nodes without additional hardware support.

Thus, there are four main goals that need to be taken into account so as to realize video analytic services on fog nodes.

Extract Key Frames

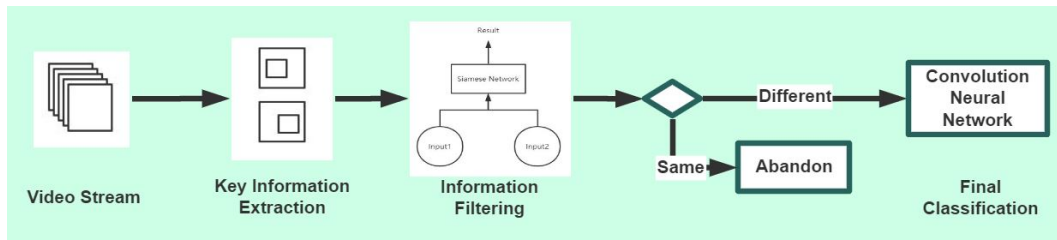
Generally speaking, video data can be treated as a sequence of images with increasing timestamps during real-time processing. The surveillance equipment's output consists of 24-30 frames per second, meaning that a large amount of video data will be transmitted to the computing devices and brings in an unbearable burden to these devices. Hence, the first goal is to reduce the number of output frames to the fog node reasonably and ensure the stability of the fog platform.

Filter Unnecessary Pixels

In order to guarantee the framing range and fidelity of video surveillance, the output is configured to be either 720p or 1080p. However, the input size in deep learning brings undue influence on the fog system, which means that a larger data volume is often accompanied by a much higher requirement in the computation amount. Thus, to ensure the framework having a higher processing rate, the key information on a single image needs to be accurately extracted and appropriately compressed to shrink the input size in the classification neural network.

Acquire Accuracy Results

Using an overly simple neural network structure may enhance multi-frame processing capabilities to a large extent, but it is noticeable that if the video analytic framework cannot provide precise information, gaining higher processing speeds will become meaningless, especially when there is no results correction service provided by the cloud in infrastructure-less environments. For example, early neural network models such as LeNet-5 only consist of a small number of simple convolutional layers. Although the required computational resources required to run the LeNet-5 are moderate, the classification accuracy is unsatisfactory due to the insufficient structure depth. In contrast, state-of-the-art models such as Inceptionv4 are also inappropriate in infrastructure-less environments. These models always concentrated on



■ **Figure 1** The overall architecture of video analytic framework for fog system.

achieving a superior classification without the concerns about the computational resources. In order to ensure the reliability of the classification result, it is thus necessary to adopt the most suitable model that can be afforded by fog devices to extract the key information.

Guarantee Overall Flexibility

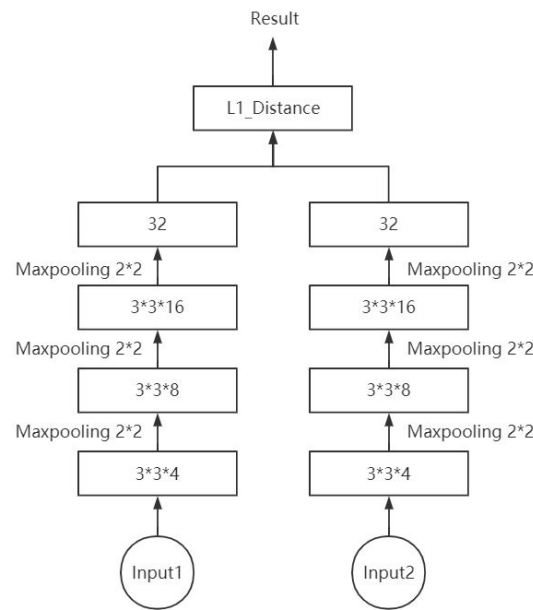
To achieve the video analytic service in the fog-based infrastructure-less environments, a framework is designed to handle this complex situation. The ever-changing residue of computing capability is the most common problem, which is mainly affected by two aspects, one is the number of users allowed to access the fog node services, and the other is the remaining power of the battery. The former will reduce the available computing power in fog nodes, and the impact is still acceptable. The latter will cause severe problems like the power supply shortage of fog nodes or system breakdown. For instance, one common deployment scenario for the fog-based video analytic framework is for wildlife monitoring. As many wild animals are nocturnal animals, the framework thus needs to perform well at night and expects to execute at the lowest working state of the CPU to reduce the consumption of stored electricity. Therefore, the video analytic framework should adapt to different computing capabilities to retain satisfactory services in different working scenarios.

3 Video Analytic Framework

In order to address the aforementioned issues and provide video analytic services in the infrastructure-less environment, we proposed a framework to splits a video processing process into multiple parts and optimizes them separately. As shown in Fig. 1, the framework consists of three main parts, including key information extraction, information filtering, and final classification.

3.1 Key Information Extraction

Due to the lack of computation power in fog nodes, the first priority is extracting the key images from the video stream so that the total amount of images that needed to be processed remains low. In infrastructure-less environments, the animals might appear in the captured image and stay for a short period. Thus, the frame difference method is considered as the most suitable approach to obtain several key frames from a series of video frames in which the animals appear. Specifically, the frame difference method is to subtract the pixel values of two images in two adjacent frames or a few frames apart in the video stream to extract the moving area in the image [15]. The benefits of this approach mainly from two aspects: Firstly, because of the sensitivity to moving objects, the appearance and movement of animals can be accurately captured while the frame in the stationary state can be ignored automatically.



■ **Figure 2** A customized 4-layer Siamese neural network structure.

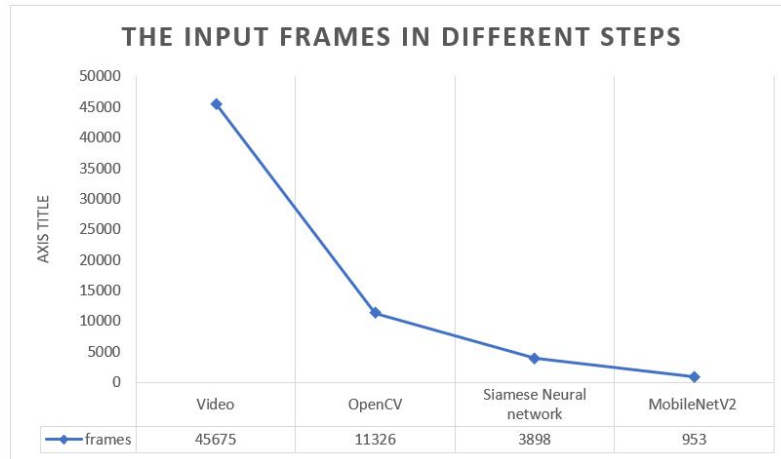
Compared with other background extraction algorithms, the frame difference method has a better balance between the processing performance and computing consumption in the animal detection scenario. Secondly, it has high tunability, the time between two frames can be adjusted dynamically based on the computation resource of fog nodes to acquire optimize processing speed and avoid excessive battery drain.

However, in actual tests, we found that the information obtained by the frame difference method is not always accurate, which is often caused by the movement of only a part of the animal's body, like feet or tails. In order to solve this problem, we empirically enlarged the size of the acquisition area when generating small-scale image changes to ensure the integrity of the image as much as possible. After that, the video data is then able to be converted to a series of clear animal pictures, and all pictures can be quickly scaled down to the same resolution (224*224) for further processing.

3.2 Information Filtering

In most cases, extracting key information from the video stream can solve the problem of insufficient computing power due to the significant reduction of images that need to be processed, but it is far from enough in infrastructure-less environments. Through the analysis of the obtained pictures, we found that the pictures have a high degree of similarity because it contains the same animals with various behaviors in a short time. If the pictures containing the same animal with different actions can be distinguished clearly, it can further decrease the number of pending pictures waiting to be processed and maximally save computing resources. Therefore, we introduced a neural network structure serving for picture filtering, called Siamese neural network.

The idea of the Siamese neural network is to learn a function that differentiates two similar inputs through two neural networks with shared parameters [3]. Different from the traditional neural network in image classification, the Siamese neural network only infers whether two input objects belong to the same type, and the output is "same" or "different" instead of the class.

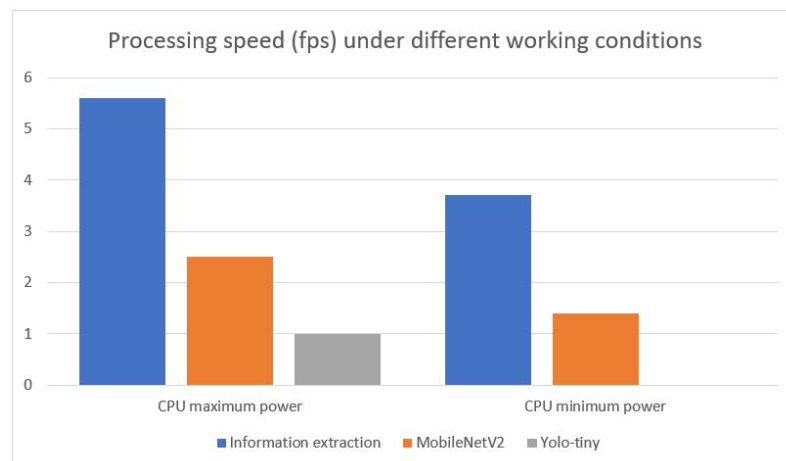


■ **Figure 3** The number of critical frames obtained after each extraction step.

As shown in Fig. 2, we implemented a 4-layer Siamese neural network structure and use L1 distance to measure the similarity. Since the image resolution obtained from the first step is $224 * 224$, we need to compress it to $56 * 56$ before sending it to the Siamese neural network. At the same time, we adjusted the structure of the traditional Siamese neural network. From the system's perspective, the pictures received from the first step are time-continuous, and we do not need to submit two frames at the same time for similarity comparison. If it is determined that the two inputs belong to the same animal, the Siamese neural network will abandon the first frame and retain the operation result of the second frame, then output the classification result of the first frame. Conversely, if the two frames are different animals, the network will upload the second frame to the next step to obtain the classification result. In order to verify the performance, we generated 30,000 pairs of training data and 6,000 pairs of verification data using pictures of 5 different birds obtained from the bird observation video to complete the module training that the accuracy could reach up to 99.4%. The trained model has achieved the highest possible accuracy on the specified training and testing data sets, but this result is limited to the five bird species used in the data sets due to the similarity of these two data sets. Therefore, the images of other animal species that have not been contained in the data sets are suitable to verify the robustness of the Siamese network model. To this end, 1757 consecutive pictures of pet dog activity are gained by the key information extraction from a 4-minute video, and then are submitted to the model for acquiring decision results. The Siamese network model successfully achieves more than 75% accuracy on untrained animal classes, which shows an exceptional potential in information filtering.

3.3 Final Classification

Through these two processing tasks, the video stream data is converted to a small group of distinguished pictures, and the final task is to choose an appropriate neural network model for the classification based on the computing capability of fog nodes. In order to ensure the reliability of the classification results and the deployability on fog nodes, we choose MobileNetV2 as the model to complete the image classification task. Specifically, MobileNetV2 is tailored for mobile and resource-constrained environments, which retain the same accuracy with a significant decrease of numbers of operations and memory needed [14].



■ **Figure 4** The processing frame rate in different working conditions.

We train the model with an open dataset from Kaggle, which contains 25,000 dog and cat images. For reducing the training time and testing the generalization ability of the model, we firstly use the transfer learning technique to freeze all convolution layers based on the pre-trained Imagenet model and only train the full connection layers. MobileNetV2 unsurprisingly achieved a high accuracy rate that around 96.07%.

4 Experimental Results

In this section, we completed a series of experiments on Raspberry Pi 3B+ (RPI) to demonstrate the framework performance in infrastructure-less environments.

OpenCV, as a lightweight and efficient cross-platform computer vision library [11], has been installed in RPI to implement the frame difference method, and the interval between two frames is controlled to adjust the occupied computing resource. Besides, the Siamese neural network and MobileNetV2 are established by Tensorflow, which is an open-source software library to fulfill different machine learning tasks [1]. Afterward, we downloaded animal videos from YouTube, which were collected by fixed-position cameras. In order to make the experimental results in line with the actual environment, we pre-processed the resolution and frame number of the video to obtain the same parameter settings as the surveillance camera, 720p and 25fps.

Fig. 3 demonstrates the number of critical frames obtained after each extraction step. It can be observed that the total number of frames has decreased significantly, and only a few pictures need to be processed on MobilNetV2. Furthermore, the initial frame resolution in the video stream is 1280*720, and the output images from OpenCV are compressed to be 224*224 that only contains critical information. For evaluating the performance, we used 6 videos with a total duration of 1,827 seconds and the total frame number of the video is 45,675 frames. Since 25 frames per second have greatly exceeded the processing capacity of OpenCV on RPI, the frame difference method will jump 2 frames in order to guarantee the stability of the framework. In other words, it will execute every 4 frames at each timestamp. In the experimental video, OpenCV has executed 11,326 times and identified 3,898 frames containing moving animals. After the similarity judgment, only 953 frames that occupy 2.08% in the total video frames need to be passed from the Siamese neural network to the MobileNetV2 for acquiring all classification results in the video.

At the same time, to provide long-term operation in infrastructure-less environments, the performance of the framework under different power consumption states should be evaluated. Fig. 4 shows the maximum processing speed can be achieved when the framework realizes real-time services under different working states of the CPU. Obviously, the performance of the framework is satisfactory even operates with minimal CPU power, and the processing frame rate is much better than deploying Yolo-tiny.

5 Conclusion

In this paper, we propose a fog-based framework in infrastructure-less environments to achieve the video analytic service. By utilizing the frame difference method and the Siamese neural network to extract the key information in the video, the video analytic framework successfully converts the huge amount of video data that fog nodes cannot afford into a small amount of critical image. Specifically, the framework overcomes the computation limitation in fog nodes to obtain classification results and minimizes the usage of computing-intensive CNN. Additionally, the experimental results clearly show a good performance of our proposed video analytic framework and its capability to deal with emergencies by distributing tasks to other fog nodes due to the shrink of input data size. Our next phase of research will focus on developing the communication strategy to decide the assignment of tasks among nodes in real-time to obtain better processing capabilities.

References

- 1 Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pages 265–283, November 2016.
- 2 Flavio Bonomi, Rodolfo Milito, Jiang Zhu, and Sateesh Addepalli. Fog computing and its role in the internet of things. In *Proceedings of the first edition of the MCC workshop on Mobile cloud computing*, pages 13–16, August 2012.
- 3 Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, June 2005.
- 4 Australian Competition & Consumer Commission. Domestic mobile roaming declaration inquiry final report, 2017.
- 5 Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, pages 1097–1105, 2012.
- 6 He Li, Kaoru Ota, and Mianxiong Dong. Learning iot in edge: Deep learning for the internet of things with edge computing. *IEEE Network*, 32(1):96–101, January 2018.
- 7 Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *European Conference on Computer Vision*, pages 21–37. Springer, 2016.
- 8 Lingjuan Lyu, James C Bezdek, Xuanli He, and Jiong Jin. Fog-embedded deep learning for the internet of things. *IEEE Transactions on Industrial Informatics*, 15(7):4206–4215, July 2019.
- 9 Australian Government Department of Agriculture. Land use in australia at a glance 2006. URL: https://www.agriculture.gov.au/sites/default/files/abares/aclump/documents/Land_use_in_Australia_at_a_glance_2006.pdf.
- 10 Australia Government Department of the Environment and Energy. Capad 2016 national summary. URL: <https://www.environment.gov.au/land/nrs/science/capad/2016>.

- 11 Kari Pulli, Anatoly Baksheev, Kirill Korniyakov, and Victor Eruhimov. Real-time computer vision with opencv. *Communications of the ACM*, 55(6):61–69, June 2012.
- 12 Xukan Ran, Haolanz Chen, Xiaodan Zhu, Zhenming Liu, and Jiasi Chen. Deepdecision: A mobile deep learning framework for edge video analytics. In *IEEE INFOCOM 2018-IEEE Conference on Computer Communications*, pages 1421–1429. IEEE, April 2018.
- 13 Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, June 2016.
- 14 Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4510–4520, June 2018.
- 15 Nishu Singla. Motion detection based on frame difference method. *International Journal of Information & Computation Technology*, 4(15):1559–1565, 2014.
- 16 Qiushi Zheng, Jiong Jin, Tiehua Zhang, Jianhua Li, Longxiang Gao, and Yong Xiang. Energy-sustainable fog system for mobile web services in infrastructure-less environments. *IEEE Access*, 7:161318–161328, 2019.