IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2320-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 10, Issue 6, June 2021

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 7.512

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006426 |

An Approach: On Big Data Redundancy Avoidance in Data Centers

Pradnya Khare[1], Prof. Shushma Ghose [2]

P.G. Student, Department of Computer Engineering, SCOE, Pune, Maharashtra, India[1]

Assistant Professor, Department of Computer Engineering, SCOE, Pune, Maharashtra, India[2]

ABSTRACT: As the innovative data collection technologies are applying to every aspect of our society, the data volume is skyrocketing. Such phenomenon poses tremendous challenges to data centers with respect to enabling storage. In this concept, a hybrid-stream big data analytics model is proposed to perform multimedia big data analysis. This model contains four procedures, i.e., data pre-processing, data classification, and data recognition and data load reduction. Specifically, an innovative multi-dimensional Convolution Neural Network (CNN) is proposed to assess the importance of each video frame. Thus, those unimportant frames can be dropped by a reliable decision-making algorithm. In order to ensure video quality, minimal correlation and minimal redundancy (MCMR) are combined to optimize the decision-making algorithm. Simulation results show that the amount of processed video is significantly reduced, and the quality of video is preserved due to the addition of MCMR. The simulation also proves that the proposed model performs steadily and is robust enough to scale up to accommodate the big data crush in data centers.

I.INTRODUCTION

A key frame extraction mechanism is proposed to realize load-reduction. Furthermore, we propose minimal correlation and minimal redundancy (MCMR) which can cover the shortage of key frame extraction. It is able to reserve the important frames which are dropped by key frame extraction mistakenly. To optimize the objective quality of videos after load-reduction, a special threshold is designed to control the MCMR, and Genetic Algorithm (GA) is used to determine this threshold. Generally, a video sequence is made up of numerous frames. In order to reduce the multimedia data redundancy and ensure that the data can be stored quickly and precisely, a mechanism is proposed to evaluate the importance of frames and categorize videos accordingly. In picture classification filed, CNN has shown its remarkable ability. We propose a multi-dimensional CNN algorithm, which differs from traditional conventional image processing, to represent the importance of a frame or a clip, and thus to achieve more accurate classification. Every CNN module consists of an input layer, an output layer, and some hidden layers. Single video image frame as an input is processed by convolution and pooling operation, and goes through fully connected layers. Finally, the network outputs a value which represents the information of this single video image frame. This paper basically carried out detail literature study of system.

II.LITERATURE REVIEW

Gianni Costa • Giuseppe Manco Riccardo Ortale [1] proposed An incremental clustering scheme for data de-duplication. It propose an incremental technique for discovering duplicates in large databases of textual sequences, i.e., syntactically different tuples, that refer to the same real-world entity. The problem is approached from a clustering perspective: given a set of tuples, the objective is to partition them into groups of duplicate tuples. Each newly arrived tuple is assigned to an appropriate cluster via nearest-neighbor classification. This is achieved by means of a suitable hash-based index that maps any tuple to a set of indexing keys and assigns tuples with high syntactic similarity to the same buckets. Hence, the neighbors of a query tuple can be efficiently identified by simply retrieving those tuples that appear in the same buckets associated to the query tuple itself, without completely scanning the original database. Two alternative schemes for computing indexing keys are discussed and compared. The de-duplication technique foresees the identification of duplicate information and it is explicitly designed for dealing with the scalability and instrumentality issues that typically arise in this setting. It relies on an incremental clustering algorithm that aims at discovering clusters of duplicate tuples. To this

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006426 |

purpose, we studied two key-generation schemes, that allow a controlled level of approximation in the search for the nearest neighbors of a tuple. Hao Ye et. Al. [2] proposed Evaluating Two-Stream CNN for Video Classification Traditional hand-crafted features are known to be inadequate in analyzing complex video semantics. Inspired by the huge success of the deep learning methods in analyzing image, audio and text data, significant efforts are recently being devoted to the design of deep nets for video analytics. Among the many practical needs, classifying videos (or video clips) based on their major semantic categories (e.g., \skiing") is useful in many applications. In this paper, we conduct an in-depth study to investigate important implementation options that may affect the performance of deep nets on video classification. Our evaluations are conducted on top of a recent two- stream convolution neural network (CNN) pipeline, which uses both static frames and motion optical flows, and has demonstrated competitive performance against the state-of-the-art methods. In order to gain insights and to arrive at a practical guideline, many important options are studied, including network architectures, model fusion, learning parameters and the final prediction methods. Based on the evaluations, very competitive results are attained on two popular video classification benchmarks. Jin Li et. Al. proposed Secure Distributed Deduplication Systems with Improved Reliability, system defined the new distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. The security requirements of data confidentiality and tag consistency are also achieved by introducing a deterministic secret sharing scheme in distributed storage systems, instead of using convergent encryption as in previous deduplication systems. Security analysis demonstrates that our deduplication systems are secure in terms of the definitions specified in the proposed security model. As a proof of concept, we implement the proposed systems and demonstrate that the incurred overhead is very limited in realistic environments. Haochen Zhang et. Al. [4] Convolutional Neural Network-Based Video Super-Resolution for Action Recognition for video action recognition, convolutional neural networks (CNNs) especially two-stream CNNs have achieved remarkable progress in the recent years. However, most of the CNNs for action recognition are trained with high-resolution videos and not scale invariant, making it problematic to apply the trained CNNs directly on low-resolution videos. One possible solution to the problem is performing super-resolution (SR) prior to action recognition. In this paper, we investigate the effects of CNN-based video SR on the action recognition accuracy. We adopt a well trained two-stream CNN for action recognition, and analyze the spatial and temporal streams separately. For the spatial stream, we observe that video SR may improve the PSNR but may incur drop in recognition accuracy, this phenomenon is further analyzed in this paper. For the temporal stream, we observe that frame-by-frame SR may produce temporal inconsistency between consecutive video frames, which also incurs drop in recognition accuracy. We then propose a temporal consistency-oriented method for video SR, which indeed improves the recognition accuracy. Kaiming He et. Al. [5] proposed Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. System we equip the networks with another pooling strategy, “spatial pyramid pooling”, to eliminate the above requirement. The new network structure, called SPP-net, can generate a fixed-length representation regardless of image size/scale. Pyramid pooling is also robust to object deformations. With these advantages, SPP-net should in general improve all CNN-based image classification methods. On the ImageNet 2012 dataset, we demonstrate that SPP-net boosts the accuracy of a variety of CNN architectures despite their different designs. On the Pascal VOC 2007 and Caltech101 datasets, SPP-net achieves state-of-the-art classification results using a single full-image representation and no fine-tuning. The power of SPP-net is also significant in object detection. Using SPP-net, we compute the feature maps from the entire image only once, and then pool features in arbitrary regions (sub-images) to generate fixed-length representations for training the detectors. This method avoids repeatedly computing the convolutional features. In processing test images, our method is 24-102_ faster than the R-CNN method, while achieving better or comparable accuracy on Pascal VOC 2007. In ImageNet Large Scale Visual Recognition Challenge (ILSVRC) 2014, our methods rank #2 in object detection and #3 in image classification among all 38 teams. This manuscript also introduces the improvement made for this competition. Lin Gu et. Al. [6] proposed A General Communication Cost Optimization Framework for Big Data Stream Processing in Geo-distributed Data Centers. System propose a general modeling framework that describes all representative inter task relationship semantics in BDSP. Based on our novel framework, we then formulate the communication cost minimization problem for BDSP into a mixed-integer linear programming (MILP) problem and prove it to be NP-hard. We then propose a computation-efficient solution based on MILP. The high efficiency of our proposal is validated by extensive simulation based studies. Public cloud service providers usually operate a number of geo-distributed datacenters across the globe. Different datacenter pairs are with different inter-

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006426 |

constitutes a large portion of a cloud provider's traffic demand over the Internet and incurs substantial communication cost, which may even become the dominant operational expenditure factor. As the datacenter resources are provided in a virtualized way, the virtual machines (VMs) for stream processing tasks can be freely deployed onto any datacenters, provided that the Service Level Agreement (SLA, e.g., quality-of-information) is obeyed. This raises the opportunity, but also a challenge, to explore the inter-datacenter network cost diversities to optimize both VM placement and load balancing towards network cost minimization with guaranteed SLA.

id	Title and Author	Methodology	Advantages	Disadvantages
[1]	An incremental clustering scheme for data de-duplication.	System describe an incremental technique for de-duplicating sequences within the process for pursuing Entity Resolution in text data using q gram techniques.	Provide best accuracy with different type of heterogeneous dataset.	Cluster can be create with duplication, redundancy generation during the clustering
[2]	Evaluating Two-Stream CNN for Video Classification	convolutional neural network (CNN) pipeline, which uses both static frames and motion optical ows, and has demonstrated competitive performance against the state-of-the-art methods.	CNN provides the better accuracy than basic classification	Multiple feedback layers should be generate high time complexity.
[3]	Secure Distributed Deduplication Systems with Improved Reliability	System proposed distributed deduplication systems with higher reliability in which the data chunks are distributed across multiple cloud servers. Multiple CSP's has used for runtime data processing	It can eliminate the single point bottleneck issue during the process execution	Required too much resources for establishing multiple CSP's.
[4]	Convolutional Neural Network-Based Video Super-Resolution for Action Recognition	System address the the effects of CNN-based video SR on the action recognition accuracy in video object using CNN algorithm	Achieved minimum MSE and better PSNR rate.	Third party resource dependency for frames creation.
[5]	Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition	Image classification approach using CNN for Pyramid pooling.	Pyramid pooling provides a robust to object deformations	It works only 2242*224 input image with type dependency.

| e-ISSN: 2319-8753, p-ISSN: 2320-6710| www.ijirset.com | Impact Factor: 7.512|

|| Volume 10, Issue 6, June 2021 ||

| DOI:10.15680/IJIRSET.2021.1006426 |

III.DISCUSSION

In data centers, video data may surge at any time. However, the frequent transmission and storage of the same content may exhaust the processing capability of data centers. Considering the capacity of the transmission load, the redundancy of multimedia data in data centers cannot be ignored. However, when the data are transferred among data centers, one has to consider how to efficiently store these data. We propose an efficient algorithm to classify these data and then store them by the results of the classification. In other words, the key of our proposed algorithm is to reduce the redundancy of data and store them as soon as possible. Generally, a video sequence is made up of numerous frames. In order to reduce the multimedia data redundancy and ensure that the data can be stored quickly and precisely, a mechanism is proposed to evaluate the importance of frames and categorize videos accordingly. In picture classification filed, CNN has shown its remarkable ability. We propose a multi-dimensional CNN algorithm, which differs from traditional conventional image processing, to represent the importance of a frame or a clip, and thus to achieve more accurate classification. We design a hybrid-stream big data analytics model, as shown in Fig. 1, which consists of four modules: Video Pre-processing, Video Classification, Frame-Load-Reduction Processing, and Video Decision. Video Pre-processing is designed to process videos sent from the cameras. Since it is impractical to record key frames if videos are separated into many frames, we propose to divide raw video resources into two input forms: video frames and video clips. Specially, video clips are categorized by a multi-dimensional CNN. After the videos are processed by the first module, the video frames and video clips are sent to the second module called Video Classification, which calculates importance of every frame and clip. Frame-Load-Reduction Processing picks up the key frames, drops the useless frames, and decides the video's class. Finally, the Video Decision will identify an appropriate threshold to handle these videos.

IV.CONCLUSION

The proposed study illustrates how frames redundancy eliminate from big data environment. Many existing system has developed the similar scenario and how to achieve beta results with machine learning as well as deep learning approaches. This study also examine frame extraction from where is kind of video objects with audio features and eliminate the redundant feature using various techniques. Basically some systems also finds numerous drawbacks which is define in literature review table, it can be increase the time complexity as well as reduce the quality of video after the compression. The frames extraction as well as frame compression is also most important to achieve the best result for CNN. Finally we conclude according to this literature review, you can achieve the best result with the help of multiple convolutional in deep learning approach.

V.FUTURE WORK

To implement a system on different kind of video as well as frame modules using deep learning framework, will be the future task of system

REFERENCES

[1] Costa G, Manco G, Ortale R. An incremental clustering scheme for data de-duplication.Data Mining and Knowledge Discovery. 2010 Jan 1;20(1):152.

[2] Ye H, Wu Z, Zhao RW, Wang X, Jiang YG, Xue X. Evaluating two-stream CNN for video classification. InProceedings of the 5th ACM on International Conference on Multimedia Retrieval 2015 Jun 22 (pp. 435-442). ACM.

[3] Li J, Chen X, Huang X, Tang S, Xiang Y, Hassan MM, Alelaiwi A. Secure distributed deduplication systems with improved reliability. IEEE Transactions on Computers. 2015 Feb 6;64(12):3569-79.

[4]Zhang H, Liu D, Xiong Z. Convolutional Neural Network- Based Video Super-Resolution for Action Recognition.In2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018) 2018 May 15 (pp.746-750). IEEE.

[5] He K, Zhang X, Ren S, Sun J. Spatial pyramid pooling in deep convolutional networks for visual recognition. IEEE transactions on pattern analysis and machine intelligence. 2015 Jan 9;37(9):1904-16.

[6] Zeng D, Gu L, Guo S. A General Communication Cost Optimization Framework for Big Data Stream Processing in Geo-Distributed Data Centers.InCloud Networking for Big Data 2015 (pp. 79-100). Springer, Cham.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor:
7.512

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर
NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details