International Journal of Research in Advent Technology, Vol.5, No.1, May 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Convolutional Neural Networks and Their Classification for Large Scale Videos

Dr. Charu Jain[1], Aarti Chugh[2], Nisha Charaya[3]
[1, 2,3] *Amity University Haryana {cjain, achugh & ncharaya}@ggn.amity.edu*

Abstract- The problems of image recognitions require powerful model implementation like convolutional neural networks (CNN). The obtained results based on extensive evaluation and classification of huge videos obtained from YouTube having more than one million views are used in this paper analysis. The CNN connectivity for additional advantages of temporary information in specific time domain is performed and architecture for multiple resolutions is shown as the result of classification of neural network. The significant improvement in performance from 50.3% to 65.3% shows the display significance for the neural network. But for case when model based on single frame is implemented, this improvement is very low [59% to 60%], but this little improvement is showing its significance. The performance generalization based on selected model and actions of UCF-101 is further studied in our paper. The base line of UCF-101 [44%] is considered as the measuring tool of improvement for reorganization and comparison of large scale videos. The classification is easily done with CNN.

Index Terms- Convolutional Neural Networks, Video, Image Processing.

1. INTRODUCTION

The internet is full of videos and other images, so the requirement for algorithm development becomes essential so that semantic content can easily be analyzed for different applications. The summarization of searching operations for videos becomes easy with specific algorithms. The CNN is providing the required class model for understanding the segmentation, image retrieval process, detection of image, and content hold by image [5]. The implemented technologies of CNN easily scale-up the parameters of neural network and provide easy process of learning to massive datasets. The interpretable features of CNN make a powerful environment of learning. The positive results of CNN performance with static images makes easy classification of large scale videos. The challenges
for applying CNN settings for complex and temporary videos are mainly faced in this performance analysis. The current internet has no bench mark which can specify the variety and required matching scale for storing and collecting videos. Thus, for training the CNN architecture, the sport data having more than one million views on YouTube is selected. This selected video is present in nearby 488 different classes [3,4].
Different parameters are set by CNN model, so that, the optimization of CNN results can be parameter based and not on bases of assumptions. As CNN processes all images in one by one manner and videos are collection of several images, thus, time utilization issues arrives. For mitigating this timing issue, the modification in CNN architecture based on context stream and fovea streams are performed. The frames with low resolution will be handled in context stream and operations on frames middle portion which is having high resolution are managed with fovea stream. So, the increased performance on run time with reduced input dimensionality and effective accuracy classification can be obtained. The UCF-101 is providing required improvements in setting the transformations for different implemented classes on sports video. Thus, summary of our contribution to this paper is:

- The UCF-101 is applied to our sports dataset and baseline based improvements are measured alone.
- The low and high resolution partition based on context and fovea streaming for improved CNN performance during run time without affecting the accuracy is also measured.
- The class based classification for Sport video having more than one million views and maintaining 488 categories is done alone.

2. IMPLEMENTED MODELS

The rescaling and cropping of images can easily be done with help of easily available tools but same things are not possible with videos. The architecture for information fusion with slow fusion technique, involves both temporal as well as spatial mixture for balancing the information of upper layer and initial layer. The information fusion for single frame is extended in this fusion for all layers with different values of input frames.

2.1 CNN For Multiple Resolutions

The context and fovea streams are used for retaining the performance and increased speed in this multiple resolution model. The multiple resolution model of CNN uses different endeavors like algorithms for effective optimization, strategies of value initialization, schemes for weight quantization of frames, and hardware improvements. But we are not discussing all these and

only performance sacrificing during fast run time of frames is evaluated. The diagram (Figure 1) showing this architecture is:

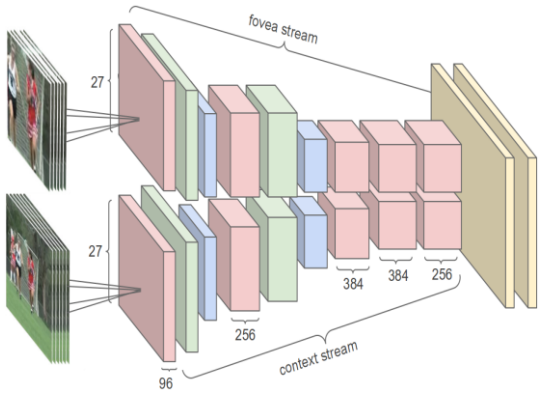

Figure 1: Architecture of CNN based Model

- Context and fovea streams: Two different streams with two different resolutions are input on the network for analysis. The input frame to the network is of 178*178 size which is further divided to two resolutions of 89*89 sizes for context and fovea streams. This model is mostly suitable for the online videos [3,5].
- The architectural changes of dividing the complete frame in two different sizes are mainly observed from this model.

2.2. Learning Outcome

The CNN for time based fusion and CNN for multiple resolutions helps in learning the optimizations of frame partitions. The batch creation based on decay weight and frame momentum is understood with optimization. The preprocessing for required deduction in over fitting frames is understood by data augmentation.

3. RESULTS OF ANALYSIS

For collecting the results, the sport video of YouTube having more than one million views is selected. The qualitative analysis is performed on the selected video based on rules of UCF-101. The figure showing test data for sports video is shown below:

So, the table to show the results of implementing CNN approaches on above data set of sports videos is provided below:

Table 1: Result of Different Models

Sr. No.	Implemented model	Hit rate for clip is 1	Hit rate for video is 1	Hit rate for video is 5
1.	Combination of neural network and histogram	**Nil**	**56.5**	**Nil**
2.	Only one frame CNN	**42.2**	**58.3**	**76.5**
3.	Combination of multiple tires with single frame	**43.5**	**61.2**	**79.4**
	Combination of fovea stream with single frame	31.0	50.5	73.4
	Combination of context stream with single frame	39.2	57.0	76.1
4.	Slow fusion	**42.5**	**61.1**	**81.8**
	Late	41.6	60.3	79.7
	Early	39.5	56.6	77.6
	Average for CNN:	**42.7**	**59.3**	**79.23**

The table to show the results of implementing CNN approach with accuracy of 3-fold with UCF-101 above data set of sports videos is provided below:

Table 2: Results of UCF-101

Sr. No.	Implemented model	Measured accuracy on UCF-101 based 3-fold
1.	Model of Soomro[6]	44-8%
	Combination of neural network and histogram	58.6%
2.	Scratch	42.3%
	Only top layer	65.2%
	Last three top layers	**66.3%**
	All layers	61.5%

4. CONCLUSION

The CNN based architecture is mainly applied for making classification of large scale videos. The powerful features of CNN architecture for effectively handling the videos which are weakly labeled for improved performance based on time architecture are mainly analyzed in this paper. The error handling with such large scale videos is also analyzed.

The obtained results in above tables shows that better performance can be obtained by implementing slow fusion. With online videos, the model of single frame

shows its effective performance. The camera motions can also be handled by CNN architecture for improved performance. The UCF-101 is also helping users for making easy classification of videos. The handling of camera motions for large scale videos is future scope of our paper.

REFERENCES

[1] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V.Le, M. Z. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, and A. Y. Ng. Large scale distributed deep networks. In NIPS, 2012.

[2] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In ICCV, 2003.

[3] H.Wang, M. M. Ullah, A. Klaser, I. Laptev, C. Schmid, et al. Evaluation of local spatio-temporal features for action recognition. In BMVC, 2009.

[4] P. Sermanet, S. Chintala, and Y. LeCun. Convolutional neural networks applied to house numbers digit classification. In ICPR, 2012.

[5] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradientbased learning applied to document recognition. Proceedings of the IEEE, 86(11):2278–2324, 1998.

[6] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt. Sequential deep learning for human action recognition. In Human Behavior Understanding, pages 29– 39. Springer, 2011.