FINE GRAINED VIDEO CLASSIFICATION

RAHUL VANCHATE[1], CHIKURTHI NAVEEN[2], SHRAVYA MANNE[3], M.RAVIKANTH[4], B. KRISHNA[5]

Student, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India[1,2,3]

Asst. Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India[4]

Professor, Department of Computer Science and Engineering, CMR Technical Campus, Medchal, Hyderabad, Telangana, India[5]

ABSTRACT

Common-sense video understanding entails fine-grained recognition of actions, objects, spatial and temporal relations, as well as physical interactions, arguably well beyond the capabilities of current techniques. A general framework will need to discriminate myriad variations of actions and interactions, not unlike the emergence of fine-grained tasks in visual object recognition. For example, we need to be able to discriminate against similar actions that differ in relatively subtle ways, for instance, ’putting a pen beside the cup’, ’putting the pen in the cup’, or perhaps ’pretending to put the pen on the table’. For this, we describe a DNN for fine-grained action classification and video captioning. It gives state-of-the-art performance on the challenging Something-Something dataset, with over 220,000 videos and 174 fine-grained actions. Classification and captioning on this dataset are challenging because of the subtle differences between actions, the use of thousands of different objects, and the diversity of captions penned by crowd actors. The model architecture shares features for classification and captioning, and is trained end-to-end. It performs much better than the existing classification benchmark for Something-Something, with impressive fine-grained results, and it yields a strong baseline on the new Something-Something captioning task. Our results reveal that there is a strong correlation between the degree of detail in the task and the ability of the learned features to transfer to other tasks.

Keywords: Fine Grained, Classification, Supervised Learning, DNN, Convolution Neural Networks, Video Classification, Billion Dataset.

1.INTRODUCTION

The phrase Fine Grained elaborates as more precision and accuracy using more Raw Data, for this we use as much data as possible. We also demonstrate the quality oflearned features through transfer learning from Something-Something features to a kitchen action dataset. Most existing captioning architectures are based on an encoder-decoder framework. For video captioning, the encoder is typically a convolutional or recurrent convolutional network. Inspired by these works, we use a modified encoder-decoder architecture that has an action classifier in addition to the encoder and decoder components. The decoder or classifier can be switched off, leaving pure classification or captioning models respectively. It is also possible to jointly train classification and captioning models. We train networks for captioning and classification, and further evaluate the learned features on other related tasks.

2.LITERATURE REVIEW

EXISTING SYSTEM

Compared to most current approaches to action recognition, and video captioning techniques applied to small corpora with relatively coarse-grained actions, this paper considers fine-grained action classification and captioning tasks on large-scale video corpora. Training is performed on the Something-Something dataset [1], with 174 fine-grained action categories and several thousand different objects. In particular, the captioning task requires the inference of many actions, different forms of object interaction and spatio-temporal relations, and an extremely broad set of objects, all under significant variations in lighting, viewpoint, background clutter, and occlusion.More recently, crowd-sourced data based on crowd-acting have emerged. Crowd workers are asked to generate videos depicting template actions. This allows one to target specific video domains and action classes, with control over the similarity and differences of actions, which is needed for fine-grained corpora. Examples include the Something-Something dataset [1] and the Charades dataset [4].

The first version of Something-Something [1] has 100, 000 videos of human object interactions, comprising 50 coarse-grained action groups, which are further broken down into 174 closely related action categories. The videos exhibit significant diversity in viewing and lighting, objects and backgrounds, and the ways in which the actors performed the actions. Baseline performance in [1] was a correct action classification rate of 11.5%, and 36.2% on action groups. [5] reports 34.44% classification accuracy on Something-Something action categories.

PROPOSED SYSTEM

This paper describes a deep neural network architecture comprising a two channel convolutional network and an LSTM recurrent network for video encoding. The same encoding is then shared for action classification and caption generation. The resulting network performs several times better than baseline action classification results in [1]. It also provides impressive results on an extremely challenging fine-grained video captioning task.

We also demonstrate the quality of learned features through transfer learning from Something-Something features to a kitchen action dataset. Most existing captioning architectures are based on an encoder-decoder framework. For video captioning, the encoder is typically a convolutional or recurrent convolutional network. Inspired by these works, we use a modified encoder-decoder architecture that has an action classifier in addition to the encoder and decoder components. The decoder or classifier can be switched off, leaving pure classification or captioning models respectively. It is also possible to jointly train classification and captioning models. We train networks for captioning and classification, and further evaluate the learned features on other related tasks.

Training settings In all our experiments we use a frame rate of 12f ps. During training we randomly pick 48 consecutive frames. For videos with less than 48 frames, we replicate the first and last frames to achieve the intended length. We resize the frames to 128×128, and then use random cropping of size 96×96. For validation and testing, we use 96 × 96 center cropping. We optimize all models using Adam, with an initial learning rate of 0.001.

3.SYSTEM REQUIREMENTS

HARDWARE REQUIREMENTS:

● System : i5 Processor 2.4 GHz.
● Hard Disk : 300 GB.
● RAM : 16 GB.

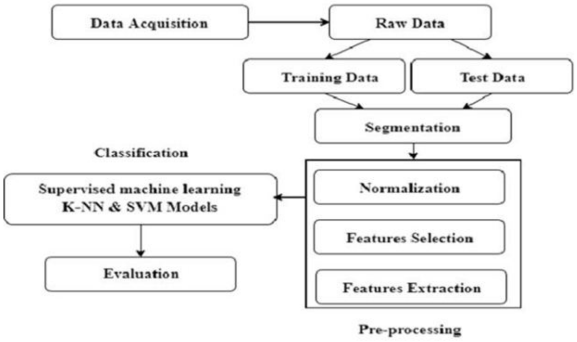
SOFTWARE REQUIREMENTS:

● Operating system : Windows /Ubuntu
● Coding Language : Python
● Front End : Python
● Back End : PostgreSQL

Python: Python is an interpreted, high-level, general-purpose programming language. It is used as the main language to code in this project.

PostgreSQL: PostgreSQL, also known as Postgres, is a free and open-source relational database management system emphasizing extensibility and technical standards compliance.

4.ARCHITECTURE



● Data acquisition is the process of sampling signals that measure real world physical conditions and converting the resulting samples into digital numeric values that can be manipulated by a computer.

● The Raw Data is supplied so that the Machine can be trained using it,it is then converted into Training Data and Test Data for simplification.

● Pre Processing is the stage where major processes such as Normalization, Features Selection, Features Extraction are performed.

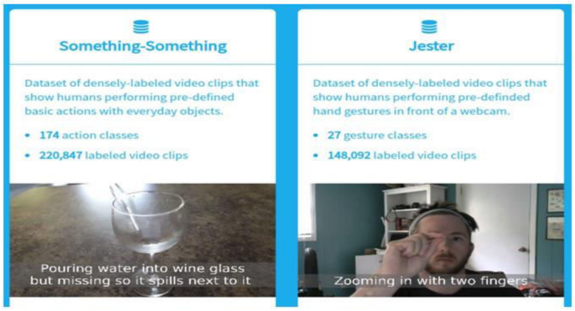
● Finally, Classification is done using Supervised Machine Learning K-NN and SVM Models that use complex Algorithms for detection and processing. After this evaluation is done for checking the results of the entire process.

5.WORKING

Data acquisition is the process of sampling signals that measure real world physical conditions and converting the resulting samples into digital numeric values that can be manipulated by a computer. The working of the Python in simple words can be

2581-4575
International Journal For Recent Developments in Science & Technology
Crossref
A Peer Reviewed Research Journal

explained to be conditional and continuous. The process of training data, Supervised Learning, Normalization and Pre Processing is continuous for better performance.



Something-Something

Dataset of densely-labeled video clips that show humans performing pre-defined basic actions with everyday objects.

- **174** action classes
- **220,847** labeled video clips

Pouring water into wine glass but missing so it spills next to it

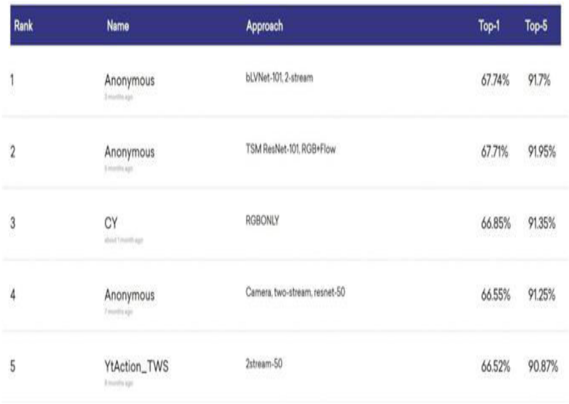
Jester

Dataset of densely-labeled video clips that show humans performing pre-definded hand gestures in front of a webcam.

- **27** gesture classes
- **148,092** labeled video clips

Zooming in with two fingers

The above image has BN-Something-Something compared to Jester which uses 27 gestures and around 148092 video clips. But clearly BN V2 has more Raw data processed into the Pre Processing stage. Pre Processing is the stage where major processes such as Normalization, Features Selection, Features Extraction are performed.



Rank	Name	Approach	Top-1	Top-5
1	Anonymous 3 months ago	bLVNet-101, 2-stream	67.74%	91.7%
2	Anonymous 3 months ago	TSM ResNet-101, RGB+Flow	67.71%	91.95%
3	CY about 1 month ago	RGBONLY	66.85%	91.35%
4	Anonymous 7 months ago	Camera, two-stream, resnet-50	66.55%	91.25%
5	YtAction_TWS 8 months ago	2stream-50	66.52%	90.87%

Classification is done using Supervised Machine Learning K-NN and SVM Models that use complex Algorithms for detection and processing. The above image has some of the people around the world who have used more amount of precise raw data that made their machine more efficient and accurate.People might have used different approaches to get their desired outputs but the base of the data and processing is the same.

6.CONCLUSION

Pre-training neural networks on large labeled datasets has become a driving force in many deep learning applications. Some might argue that it may be considered a serious competitor to unsupervised learning as a means to generate universal features that represent the visual world. Ever since ImageNet was used as a generic visual feature extractor, the hypothesis has emerged that it is the dataset size, the amount of detail and the variety of labels that drive a network's capability to learn useful generic features.To the degree that this hypothesis is true, generating visual features capable of transfer learning should involve source tasks that (i) are as fine-grained and complex as possible, and (ii) ideally involve video not still images, because video is a much more fertile domain for defining complex tasks that represent aspects of the physical world. In this work, we provide further evidence for that hypothesis, showing that the amount of detail in the task has a strong influence on the quality of the learned features.We also show that captioning, which to the best of our knowledge has hitherto been used only as a target task in transfer learning, can be a powerful source task itself. Our work suggests that one gets substantial leverage by utilizing ever more fine-grained recognition tasks, represented in the form of captions, possibly in combination with question-answering.

7.ACKNOWLEDGEMENT

The authors would like to acknowledge the support of the Chairman, Director and Head of the Department of Computer Science and Engineering CMR Technical Campus, Mechal, Hyderabad, Telangana, for their encouragement to the authors.

8.REFERENCES

1. Goyal, R., Kahou, S.E., Michalski, V., Materzyska, J., Westphal, S., Kim, H., Haenel, V., Fruend, I., Yianilos, P., Mueller-Freitag, M., Hoppe, F., Thurau, C., Bax, I., Memisevic, R.: The ”something something” video database for learning and evaluating visual common sense. In: Proceedings of the IEEE International Conference on Computer Vision. ICC V17 (2017)

2. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017) preprint arXiv:1705.06950.

3. Monfort, M., Zhou, B., Bargal, S.A., Andonian, A., Yan, T., Ramakrishnan, K., Brown, L., Fan, Q., Gutfruend, D., Vondrick, C., Oliva, A.: Moments in time dataset: one million videos for event understanding (2018) preprint arXiv:1801.03150.

4. Sigurdsson, G.A., Russakovsky, O., Gupta, A.: What actions are needed for understanding human actions in videos? In: IEEE International Conference on Computer Vision (ICCV). (Oct 2017)

5. Zhou, B., Andonian, A., Torralba, A.: Temporal relational reasoning in videos. CoRR (2017)

6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. CVPR08 (2008)

7. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild (2012) preprint arXiv:1212.0402.

8. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large scale video classification with convolutional neural networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2014) 1725–1732

9. Wu, J.: Computational perception of physical object properties. Master’s thesis, Massachusetts Institute of Technology (2016)

10. Chen, X., Fang, H., Lin, T., Vedantam, R., Gupta, S., Doll ́ar, P., Zitnick, C.L.: Microsoft COCO captions: Data collection and evaluation server. CoRR (2015).