INTERNATIONAL JOURNAL OF TECHNO-ENGINEERING
IJTE

ISSN: 2057-5688

HUMAN ACTIVITY RECOGNITION USING DEEPLEARNING

Ch.Rakesh	B.Pradeep	Sneha Rajabhau Maske	Dr.B.Kavitha Rani
B-Tech Student	B-Tech Student	B-Tech Student	(Professor)

Department of Information Technology
CMR Technical Campus
Kadlakoya (V), Medchal, Hyderabad-501401

Abstract: The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However, to date, conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pretrained simple 3D architectures outperforms complex 2D architectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics will retrace the successful history of 2D CNNs and ImageNet, and stimulate advances in computer vision for videos. The codes and pretrained models used in this study are publicly available.

I.INTRODUCTION

The use of large-scale datasets is extremely important when using deep convolutional neural networks (CNNs), which have massive parameter numbers, and the use of CNNs in the field of computer vision has expanded significantly in recent years. ImageNet [4], which includes more Recent advances in computer vision for images (top) and videos (bottom). The use of very deep 2D CNNs trained on ImageNet generates outstanding progress in image recognition as well as in various other tasks. Can the use of 3D CNNs trained on Kinetics generates similar progress in computer vision for videos? than a million images, has contributed substantially to the creation of successful vision-based algorithms. In addition to such large-scale datasets, a large number of algorithms, such as residual learning , have been used to improve image classification performance by adding increased depth to CNNs, and the use of very deep CNNs trained on ImageNet have

facilitated the acquisition of generic feature representation. Using such feature representation, in turn, has significantly improved the performance of several other tasks including object detection, semantic segmentation, and image captioning (see top row in Figure 1). To date, the video datasets available for action recognition have been relatively small when compared with image recognition datasets. Representative video datasets, such as UCF101 and HMDB-51 , can be used to provide realistic videos with sizes around 10 K, but even though they are still used as standard benchmarks, such datasets are obviously too small to be used for optimizing CNN representations from scratch. In the last couple of years, ActivityNet, which is a somewhat larger video dataset, has become available, and its use has make it possible to accomplish additional tasks such as untrimmed action classification and detection, but the number of action instances it contains is still limited. More recently, the Kinetics dataset was created with the aim of being positioned as a de facto video dataset standard that is roughly equivalent to the position held by ImageNet in relation to image datasets. More than 300 K videos have been collected for the Kinetics dataset, which means that the scale of video datasets has begun to approach that of image datasets. For action recognition, CNNs with spatio-temporal threedimensional (3D) convolutional kernels (3D CNNs) are recently more effective than CNNs with two-dimensional (2D) kernels. From several years ago 3D CNNs are explored to provide an effective tool for accurate action recognition. However, even the usage of well-organized models has failed to overcome the advantages of 2D- based CNNs that combine both stacked flow and RGB images. The primary reason for this failure has been the relatively small data-scale of video datasets that are available for optimizing the immense number of parameters in 3D CNNs, which are much larger than those of 2D CNNs. In addition, basically, 3D CNNs can only be trained on video datasets whereas 2D CNNs can be pretrained on ImageNet. Recently, however, Carreira and Zisserman achieved a significant breakthrough using the Kinetics dataset as well as the inflation of 2D kernels pretrained on ImageNet into 3D ones. Thus, we now have the benefit of a sophisticated 3D convolution that can be engaged by the Kinetics dataset. However, can 3D CNNs retrace the successful history of 2D CNNs and ImageNet? More specifically, can the use of 3D CNNs trained on Kinetics produces significant progress in action recognition and other various tasks? (See bottom row in Figure 1.) To achieve such progress, we consider that Kinetics for 3D CNNs should be as large-scale as ImageNet for 2D CNNs, though no previous work has examined enough about the scale of Kinetics. Conventional 3D CNN architectures trained on Kinetics are still relatively shallow and 34 layers). If using the Kinetics dataset enables very deep 3D CNNs at a level similar to ImageNet, which can train 152-layer 2D CNNs, that question could be answered in the affirmative. In this study, we examine various

3D CNN architectures from relatively shallow to very deep ones using the Kinetics and other popular video datasets (UCF-101, HMDB-51, and ActivityNet) in order to provide us insights for answering the above question. The 3D CNN architectures tested in this study are based on residual networks (ResNets) and their extended versions because they have simple and effective structures. Accordingly, using those datasets, we performed several experiments aimed at training and testing those architectures from scratch, as well Averaged accuracies of 3D ResNets over top-1 and top-5 on the Kinetics validation set. Accuracy levels improve as network depths increase. The improvements continued until reaching the depth of 152. The accuracy of ResNet-200 is almost the same as that of ResNet-152. These results are similar to 2D ResNets on ImageNet . as their fine-tuning. The results of those experiments (see Section 4 for details) show the Kinetics dataset can train 3D ResNet-152 from scratch to a level that is similar to the training accomplished by 2D ResNets on ImageNet, as shown in Figure 2. Based on those results, we will discuss the possibilities of future progress in action recognition and other video tasks. To our best knowledge, this is the first work to focus on the training of very deep 3D CNNs from scratch for action recognition. Previous studies showed deeper 2D CNNs trained on ImageNet achieved better performance [10]. better based on the previous studies because the data-scale of image datasets differs from that of video ones. The results of this study, which indicate deeper 3D CNNs are more effective, can be expected to facilitate further progress in computer vision for videos.

II.LITERATURE SURVEY

Topic: The Kinetics Human Action Video Dataset Description: We describe the DeepMind Kinetics human action video dataset. The dataset contains 400 human action classes, with at least 400 video clips for each action. Each clip lasts around 10s and is taken from a different YouTube video. The actions are human focussed and cover a broad range of classes including human-object interactions such as playing instruments, as well as human-human interactions such as shaking hands. We describe the statistics of the dataset, how it was collected, and give some baseline performance figures for neural network architectures trained and tested for human action classification on this dataset. We also carry out a preliminary analysis of whether imbalance in the dataset leads to bias in the classifiers. Topic: Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? The purpose of this study is to determine whether current video datasets have sufficient data for training very deep convolutional neural networks (CNNs) with spatio-temporal three-dimensional (3D) kernels. Recently, the performance levels of 3D CNNs in the field of action recognition have improved significantly. However, to date, conventional research has only explored relatively shallow 3D architectures. We examine the architectures of various 3D CNNs from relatively shallow to very deep ones on current video datasets. Based on the results of those experiments, the following conclusions could be obtained: (i) ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics. (ii) The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet. ResNeXt-101 achieved 78.4% average accuracy on the Kinetics test set. (iii) Kinetics pretrained simple 3D architectures outperforms complex 2D architectures, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively. The use of 2D CNNs trained on ImageNet has produced significant progress in various tasks in image. We believe that using deep 3D CNNs together with Kinetics.

III.IMPLEMENTATION

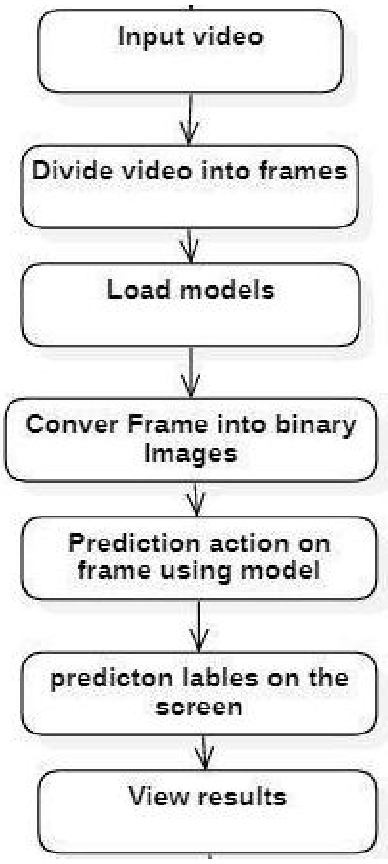


Fig-3: System Architecture

Modules

Problem Formalization

Given a source domain activity data sets like activity label and unlabeled target domain activity data sets we build activity model in the source domain labeled activity data sets first and then utilize it to recognize unlabeled activities in the target domain. We use the activity models built in the source domain to extract activity features in the target domain in presence of any unlabeled data. This helps utilize source domain label distribution to recognize target domain activity recognition tasks. We extract the features by considering each layer of the CNN independently, and try to preserve the feature representation of these two separate distributions by minimizing the KL divergence in each layer. there are two factors, i) classification cost, and ii) embedding space cost, which are needed to be considered to build our CNN.

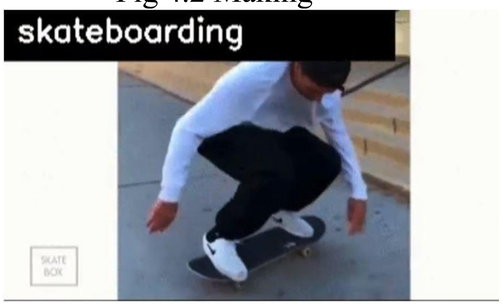
Selection of Neural Network

3D Convolution In 2D CNNs, convolutions are applied on the 2D feature maps to compute features from the spatial dimensions only. When applied to video analysis problems, it is desirable to capture the motion information encoded in multiple contiguous frames. To this end, we propose to perform 3D convolutions in the convolution stages of CNNs to compute features from both spatial and temporal dimensions. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together. By this construction, the feature maps in the convolution layer are connected to multiple contiguous frames in the previous layer, thereby capturing motion information.

IV.RESULT

Fig: 4.1 washing hands

Fig 4.2 Making



Pizza

Fig 4.3 Skateboarding

V.CONCLUSION

In this study, we examined the architectures of various CNNs with spatio-temporal 3D convolutional kernels on current video datasets. Based on the results of those experiments, the following conclusions could be obtained:

1. ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics.
2. The Kinetics dataset has sufficient data for training of deep 3D CNNs, and enables training of up to 152 ResNets layers, interestingly similar to 2D ResNets on ImageNet.
3. Kinetics pretrained simple 3D architectures outperforms complex 2D architectures on UCF- 101 and HMDB-51, and the pretrained ResNeXt-101 achieved 94.5% and 70.2% on UCF-101 and HMDB-51, respectively.

We believe that the results of this study will facilitate further advances in video recognition and its related tasks. Following the significant advances in image recognition made by 2D CNNs and ImageNet, pretrained 2D CNNs on ImageNet experienced significant progress in various tasks such as object detection, semantic segmentation, and image captioning. It is felt that, similar to these, 3D CNNs and Kinetics have the potential to contribute to significant progress in fields related to various video tasks such as action detection, video summarization, and optical flow estimation. In our future work, we will investigate transfer learning not only for action recognition but also for other such tasks.

REFERENCES

1. S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. arXiv preprint, arXiv:1609.08675, 2016.
2. J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 4724–4733, 2017.
3. J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. arXiv preprint, arXiv:1705.07750, 2017.
4. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2009.
5. B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 961–970, 2015.
6. C. Feichtenhofer, A. Pinz, and R. Wildes. Spatiotemporal residual networks for video action recognition. In Proceedings of the Advances in Neural Information Processing Systems (NIPS), pages 3468–3476, 2016.
7. C. Feichtenhofer, A. Pinz, and R. P. Wildes. Spatiotemporal multiplier networks for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
8. C. Feichtenhofer, A. Pinz, and A. Zisserman. Convolutional two-stream network fusion for video action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 1933–1941, 2016.
9. K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3D residual networks for action recognition. In Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition, 2017. 2, 3, 4, 6.