ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

HUMAN ACTIVITY RECOGNITION

MR. P. NAGARAJ[1] G. MEGHANA REDDY[2] A.H.ISAAC[2] MUSKULA MEGHANA[2]
Assistant Professor[1] and B.Tech Students[2,] Department of Computer Science and Engineering,
Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, India

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 arepublicly available.

1.INTRODUCTION

1.1 Purpose

In many surveillance applications, events of interest may occur rarely. For these unusual events (or abnormal, rare events), it is difficult to collect sufficient training data for supervised

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

learning to develop unusual event models. In this case, many unusual event detection algorithms which require large numbers of training data become unsuitable. Nevertheless, it is very challenging to recognize human activities in unconstrained videos due to some real conditions such as varying light conditions, divergent viewpoints, varying action speeds, light variations. We also need to train huge amounts of Data and require a lot of time and hardware support for training the model.

Therefore, We propose to apply the 3D convolution operation to extract spatialand temporal features from video data for action recognition.

1.2 Motivation

The use of large-scale datasets is extremely important when using deep convolution 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 generate 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 toCNNs, and the use of very deep CNNs trained on ImageNet have facilitated the acquisition of generic feature representation.

1.3 Problem definition

It is difficult to collect sufficient training data for supervised learning to develop unusual event models. In this case, many unusual event detection algorithms which require large numbers of

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

training data become unsuitable. Nevertheless, it is very challenging to recognize human activities in unconstrained videos due to some real conditions

1.4 Objective of project

For action recognition, CNNs with spatio-temporal three dimensional (3D) convolutional kernels (3D CNNs) are recently more effective than CNNs with two-dimensional (2D) kernels. From several years ago 3D CNNs were explored toprovide 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 pre trained on ImageNet. Recently, however, Carreira and Zisserman achieved a significant breakthrough using the Kinetics dataset as well as the inflation of 2D kernels pre trained on ImageNet into 3D ones. 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

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

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.

1.5 Scope

From there we'll discuss how we can extend ResNet, which typically uses 2D kernels, to instead leverage 3D kernels, enabling us to include a spatiotemporal component used for activity recognition. We'll then implement two versions of human activity recognition using the OpenCV library and the Python programming language. Finally, we'll wrap up the tutorial by looking at the results of applying human activityrecognition to a few sample videos.

1.6 Applications

● Automatically classifying/categorizing a dataset of videos on disk.

● Training and monitoring a new employee to correctly perform a task (ex., proper steps and procedures when making a pizza, including rolling out the dough, heatingoven, putting on sauce, cheese, toppings, etc.). Verifying that a food service worker has washed their hands after

visiting the restroom or handling food that could cause cross-contamination (i.e,. chicken and salmonella). Monitoring bar/restaurant patrons and ensuring they are not over-served.

2. Existing System

In many surveillance applications, events of interest may occur rarely. For these unusual events (or abnormal, rare events), it is difficult to collect sufficient training data for supervised learning to develop unusual event models. In this case, many unusual event detection algorithms which require large numbers of training data become unsuitable. Nevertheless, it is very challenging to recognize human activities in unconstrained videos due to some real conditions such as varying light conditions, divergent viewpoints, varying action speeds, light variations.

2.1 Disadvantages

● need to train huge amount of Data

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

● required huge time and hardware support for training the model

3. Proposed System

We propose to apply the 3D convolution operation to extract spatial and temporal features from video data for action recognition. These 3D feature extractors operate in both the spatial and the temporal dimensions, thus capturing motion information in video streams. . We develop a 3D convolutional neural network architecture based on the 3D convolution feature extractors. This CNN architecture generates multiple channels of information from adjacent video frames and performs convolution and subsampling separately in each channel. The final feature representation is obtained by combining information from all channels. . We propose to regularize the 3D CNN models by augmenting the models with auxiliary outputs computed as high-level motion features.

We further propose to boost the performance of 3D CNN models by combining the outputs of a variety of different architectures. . We evaluate the 3D CNN models on the TRECVID 2008 development set in comparison with baseline methods and alternative architectures. Experimental results show that the proposed models significantly outperforms 2D CNN architecture and other baseline methods.

3.1 Advantages

● These architectures have been successfully applied to image classification.
● The large-scale ImageNet dataset allowed such models to be trained to such high accuracy.
● The Kinetics dataset is also sufficiently large.
● should be able to perform video classification by changing the input volume shape to include spatiotemporal information and utilizing 3D kernels inside of the architecture.

4.System Architecture

The proposed system is built around conventional three-tier architecture. The three-tier architecture for web development allows programmers to separate various aspects of the solution design into modules and work on them separately. That is, a developer who is best at one part of development, say UI development need not worry about the implementation levels

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

so much. It also allows for easy maintenance and future enhancements. The three-tiers of the solution include:

- The Layout: This tier is at the uppermost layer and is closely bound to the user, i.e., the users of the system interact with it through this tier.
- The business-tier: This tier is responsible for implementing all the business rules of the organization. It operates on the data provided by the users through the web-tier and the data stored in the underlying data-tier. So in a way this tier works on data from the web-tier and the data-tier in order to perform tasks for the users in agreement with the business rules of the organization.
- The data-tier: This tier contains the persistent data that is required by the business tier to operate on. Data plays a very important role in the functioning of any organization. Thus, persisting such data is very important. The data tier performs the job of persisting the data.

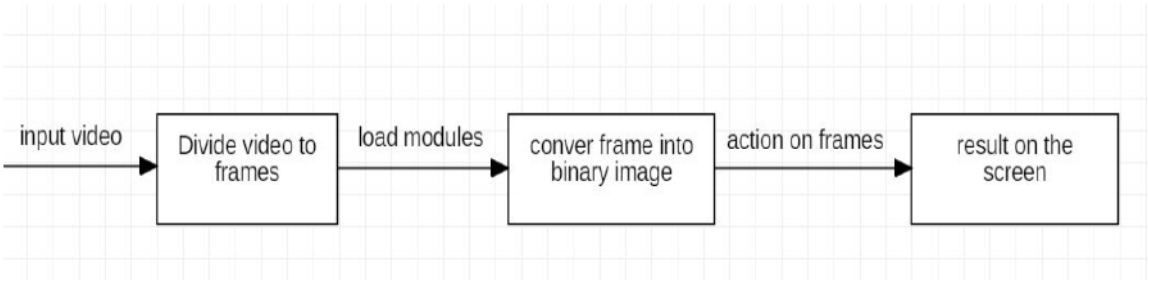

Figure 1 System architecture

5.Implementation

We use stochastic gradient descent with momentum to train the networks and randomly generate training samples from videos in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to generate a training sample.

A 16-frame clip is then generated around the selected temporal position. If the video is shorter than 16 frames, then we loop it as many times as necessary. Next, we randomly select a spatial

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

position from the 4 corners or the center. In addition to the spatial position, we also select a spatial scale of the sample in order to perform multi-scale cropping. The procedure used is the same as [28]. The scale is selected from 1, 1 21/4 , 1 √ 2 , 1 23/4 , 1 2 .

Scale 1 means that the sample width and height are the same as the short side length of the frame, and scale 0.5 means that the sample is half the size of the short side length. The sample aspect ratio is 1 and the sample is spatio-temporally cropped at the positions, scale, and aspect ratio. We spatially resize the sample at 112 × 112 pixels. The size of each sample is 3 channels × 16 frames × 112 pixels × 112 pixels, and each sample is horizontally flipped with 50% probability. We also perform mean subtraction, which means that we subtract the mean values of ActivityNet from the sample for each color channel.

All generated samples retain the same class labels as their original videos. In our training, we use cross-entropy losses and backpropagate their gradients. The training parameters include a weight decay of 0.001 and 0.9 for momentum. When training the networks from scratch, we start from learning rate 0.1, and divide it by 10 after the validation loss saturates. When performing fine tuning, we start from a learning rate of 0.001, and assign a weight decay of 1e-5. Recognition. We adopt the sliding window manner to generate input clips, (i.e., each video is split into non-overlapped 16-frame clips), and recognize actions in videos using the trained networks. Each clip is spatially cropped around a center position at scale 1. We then input each clip into the networks and estimate the clip class scores, which are averaged over all the clips of the video. The class that has the maximum score indicates the recognized class label.

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

5.1. Output

5.2.Results Analysis

In this study, in order to determine whether current video datasets have sufficient data for training of deep 3D CNNs, we conducted the three experiments described below using UCF-101 HMDB-51 ActivityNet and Kinetics . We first examined the training of relatively shallow 3D CNNs from scratch on each video dataset. According to previous works 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We try to reproduce such results to ascertain whether the datasets have sufficient data for deep 3D CNNs. Specifically, we used ResNet-18, which is the shallowest ResNet architecture, based on the assumption that if the ResNet-18 overfits when being trained on a dataset, that dataset is too small to be used for training deep 3D CNNs from scratch. See Section for details.

We then conducted a separate experiment to determine whether the Kinetics dataset could train deeper 3D CNNs. A main point of this trial was to determine how deeply the datasets

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

could train 3D CNNs. Therefore, we trained 3D ResNets on Kinetics while varying the model depth from 18 to 200. If Kinetics can train very deep CNNs, such as ResNet-152, which achieved the best performance in ResNets on ImageNet we can be confident that they have sufficient data to train 3D CNNs. Therefore, the results of this experiment are expected to be very important for the future progress in action recognition and other video tasks. See Section for details.

In the final experiment, we examined the fine-tuning of Kinetics pretrained 3D CNNs on UCF-101 and HMDB-51. Since pretraining on large-scale datasets is an effective way to achieve good performance levels on small datasets, we expect that the deep 3D ResNets pretrained on Kinetics would perform well on relatively small UCF-101 and HMDB-51. This experiment examines whether the transfer of visual representations by deep 3D CNNs from one domain to another domain works effectively. See Section for details.

Block of each architecture. We represent conv, x3, F as the kernel size, and the number of feature maps of the convolutional filter are x × x × x and F, respectively, and group as the number of groups of group convolutions, which divide the feature maps into small groups. BN refers to batch normalization. Shortcut connections of the architectures are summation except for those of DenseNet, which are concatenation.

Network Architectures. Each convolutional layer is followed by batch normalization and a ReLU. Spatio-temporal down-sampling is performed by conv3_1, conv4_1, and conv5_1 with a stride of two, except for DenseNet. F is the number of feature channels corresponding in Figure 3, and N is the number of blocks in each layer. DenseNet down-samples inputs using the transition layer, that consists of a 3 × 3 × 3 convolutional layer and a 2 × 2 × 2 average pooling layer with a stride of two, after conv2_x, conv3_x, and conv4_x. F of DenseNet is the number of input feature channels of the first block in each layer, and N is the same as that of the other networks. A 3 × 3 ×3 max-pooling layer (stride 2) is also located before conv2_x of all networks for down-sampling. In addition, conv1 spatially down-samples inputs with a spatial stride oftwo. C of the fully-connected layer is the number of classes.

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

5.3.Results Comparison

We began by training ResNet-18 on each dataset. According to previous works 3D CNNs trained on UCF-101, HMDB-51, and ActivityNet do not achieve high accuracy whereas ones trained on Kinetics work well. We tried to reproduce such results in this experiment. In this process, we used split 1 of UCF-101 and HMDB-51, and the trainingand validation sets of ActivityNet and Kinetics.

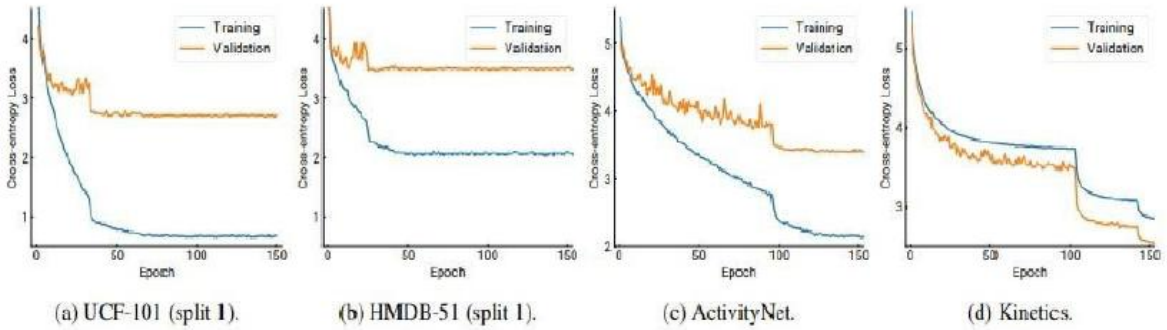

Figure 2. Training and validation loses

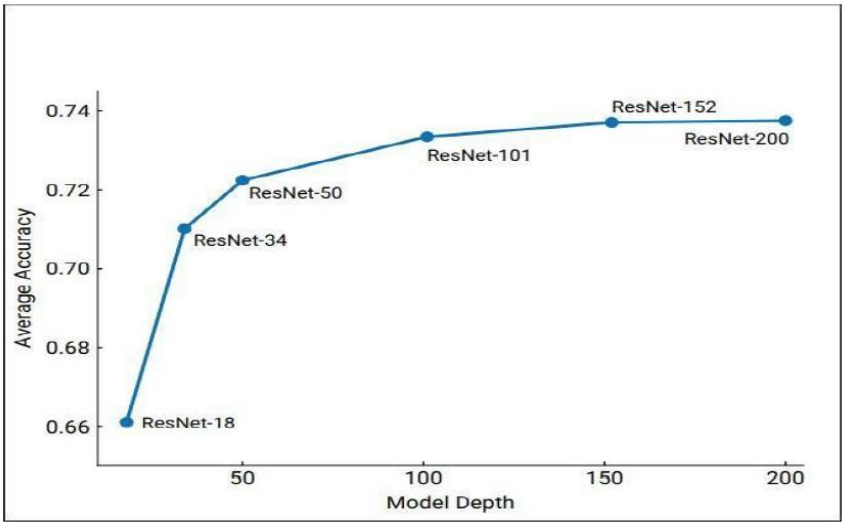

Figure 3. ResNet model depth graph

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

These architectures have been successfully applied to image classification. The large-scale ImageNet dataset allowed such models to be trained to such high accuracy. The Kinetics dataset is also sufficiently large therefore, these architectures should be able to perform video classification by (1) changing the input volume shape to include spatiotemporal information and (2) utilizing 3D kernels inside of the architecture.

6.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 :

- ResNet-18 training resulted in significant overfitting for UCF-101, HMDB-51, and ActivityNet but not for Kinetics.
- 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.
- 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.

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

7.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.

ISSN: 2278-4632
Vol-10 Issue-5 No. 6 May 2020

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

10) K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 770–778, 2016.

