"Kallis" CRCV VIPriors challenge submission

Ishan Dave[1], Kali Carter[2] and Mubarak Shah[1]

[1] Center for Research in Computer Vision (CRCV), University of Central Florida, Orlando, Florida, USA
[2] LeTourneau University, 2100 S Mobberly Ave, Longview, Texas, USA
ishandave@knights.ucf.edu, kalicarter@letu.edu, shah@crcv.ucf.edu

Abstract. This paper presents our approach "Kallis" to solve the action recognition task on UCF101 without using any pre-trained weights. We proposed a two-stream architecture using various spatio-temporal resolutions that aid in learning the long and short-range temporal structures of activities at different spatial scales. We also show that taking an average of the predictions from multiple clips that have various spatio-temporal resolutions and augmentations helps increase the performance as well as significantly lowers the training cost. Without using any pre-trained weights, the proposed solution achieves a Top-1 accuracy of **90.83%** on UCF-101 (split 1) test set and a Top-1 accuracy of **90.71%** on the test set of the Visual Inductive Priors for Data-Efficient Deep Learning Workshop's Action Recognition Challenge, ECCV 2020, which is the best among all of the other entries.

1 Introduction

Action Recognition requires spatio-temporal understanding of a sequence of video frames. In literature, there are many works that tackle the action recognition problem by deep networks by learning spatio-temporal representation using convolutional operations on the RGB frames and/or motion priors such as optical flow [13], dense point trajectories [10]). The main approaches for action recognition are based on 2D ConvNets with LSTM [12], Two-stream architectures [1], [6], and single RGB stream 3D ConvNets [9], [8]. Most of these approaches achieve high results on UCF101 [7] and HMDB [4] by using pretrained weights from ImageNet [2], Sports-1M [5], or Kinetics-400 [1].

While learning from scratch, it is difficult to optimize the parameters of a 3D ConvNet based architecture with a single stream of RGB video frames from relatively smaller datasets such as UCF101 or HMDB as compared to Kinetics-400 [1]. Carreira et al. [1] show that two-stream-based 3D ConvNet approaches significantly surpass single stream RGB video-based 3D ConvNet approaches; there is a $\sim$30% improvement for the task of action recognition on both UCF101 and HMDB when no pre-trained weights are used. This shows that optical flow is a powerful prior for modeling motion information.

Our approach is a two-stream based architecture, incorporating various spatio-temporal resolution clips for both RGB and optical flow streams. While testing

a single video, we take the average of the predictions over multiple clips of that same video, which consists of various spatio-temporal resolutions as well as some simple augmentations, such as a horizontal flip. We show that this testing strategy boosts the performance of the action classifier, while also reducing the training cost significantly. We use an ensemble of various 3D ConvNets in both RGB and optical flow streams. We use an ensemble of various 3D ConvNet in both RGB and optical flow streams which helps in mitigating common generalization errors as well as decreasing the variance in neural network predictions.

2 Proposed Method

The schematic diagram of the proposed method is depicted in Fig 1. We use a two-stream based architecture. This architecture further consists of an ensemble of either different backbone architectures or the same architecture trained with different spatio-temporal resolutions.

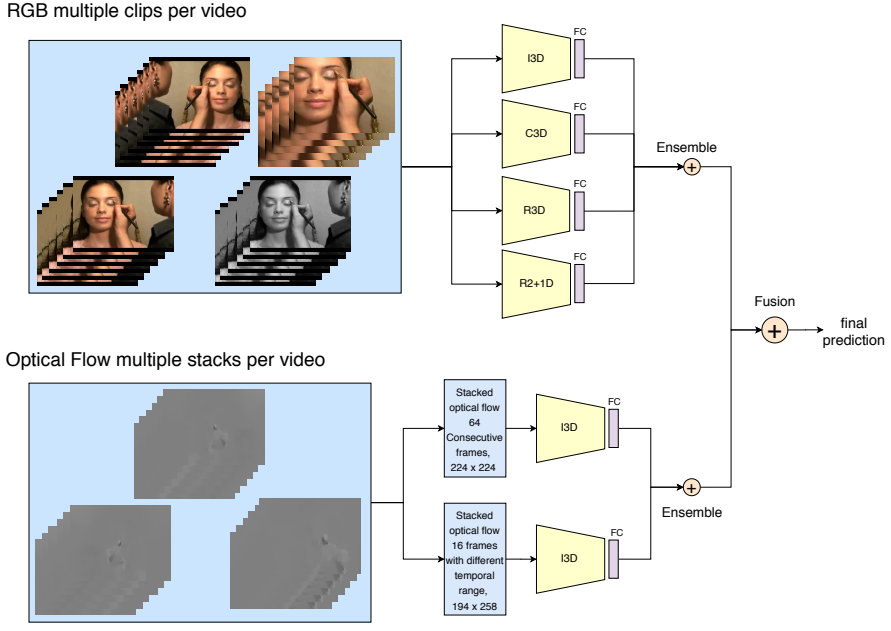

Fig. 1. Schematic Diagram of the proposed method (inference mode)

For the RGB stream, we use an ensemble of Inflated InceptionV3 (I3D) [1], C3D [8], R2+1D -18 layer [9], and ResNet 3D -18 layer architectures with the same input. The three major types of augmentations that are applied to the

RGB frames are, (1) Spatial augmentations: random crop, random scaling, and horizontal flip (2) Appearance transformation: random grayscale and color jittering (3) Temporal Augmentation: an evenly spaced, random number of skip frames and random starting frame. Each model is trained individually with all of the above augmentations. While testing, spatial and temporal augmentations are utilized. The augmentations aid in preventing overfitting while training on a small dataset such as UCF101. They also aid in mitigating the variance in the inference by taking the average of the predictions for the varying augmented clips within the same video.

For the TV-L1 stream, we use different spatio-temporal resolutions with two I3D models. The first model (I3D-flow-1) is trained with a stacked optical flow of 64 consecutive frames, with random cropping and horizontal flipping as augmentations. This model is expected to learn the fine-grained temporal structure of an activity. The second I3D model (I3D-flow-2) is trained with a stacked optical flow of 16 frames with evenly spaced, dynamic skip frames. This model sees the activity at a different temporal resolution, which helps in learning the coarse temporal features. For dynamic skip frame, a random number of temporal stride from 1 to n_{max} is chosen, where, $n_{max} = floor(video\,frame\,count/\,clip\,frame\,count)$ i.e. maximum skip frames possible in a clip. More details on the augmentations and resolutions are given in Table 1,

3 Experiments

This section covers the dataset, experiments, and computational cost comparison of two training strategies.

3.1 dataset

We use the UCF101 dataset (split 1) for our initial experiments. UCF101 consists of 13k videos collected from YouTube. We train similarly for the canonical version of UCF101 for the VI Priors Action Recognition challenge. The challenge dataset consists of 4.8k videos for training, 4.7k videos for validation, and 3.8k videos for testing. We use TV-L1 optical flow computed by the [13] method.

3.2 Results

We evaluate our trained model on multiple clips from the same video. More details on the training and testing methods for each model are provided in Table 1.

We performed our initial experiments on UCF101 to observe the performance of our ensemble and model selection purposes. The proposed method achieves a Top-1 accuracy of 90.83% on UCF-101 (split 1) test set without using any pre-trained weights in our training. We performed our training and model selection on the VI Priors Action Recognition dataset in the same manner. The final fused model (row-9) of Table-1 is used to predict the output of the competition test set and achieves a Top-1 accuracy of 90.71%.

Table 1. Results on UCF101 (split 1)

Sr No	Model	Input Type	Training Augmentations	Validation Augmentations	Top-1 Accuracy (%)
1	I3D	TVL-1, 64 frames, 341 x 256 Resolution	Random Starting frame, Random Crop (224 x 224), Horizontal Flip (50% probability)	3 different starting frames, Center Crop 224 x 224	84.91
2	I3D	TVL-1, 16 frames, 341 x 256 Resolution	Random Starting frame, Random Crop (194 x 258), Horizontal Flip (50% probability), Skip Frames (from 1 to n_{max}), Random scaling factor (from 0.75 to 1)	5 different starting frames, 10 different skip rates, 2 different spatial scales, Center Crop (198 x 258)	86.72
3	C3D	RGB, 16 frames, 320 x 240 Resolution	Random Starting frame, Random Crop (180 x 240), Horizontal Flip (50% probability), Skip Frames (from 1 to n_{max}), Random scaling factor (from 0.75 to 1)	5 different starting frames, 10 different skip rates, 2 different spatial scales, Center Crop (198 x 258)	61.46
4	I3D	Same as 3	Same as 3	Same as 3	60.01
5	R3D	Same as 3	Same as 3	Same as 3	65.94
6	R2+1D	Same as 3	Same as 3	Same as 3	66.89
7	1+2 mean ensemble				88.40
8	3+4+5+6 mean ensemble				70.90
9	7 + 8 fusion				**90.83**

3.3 Computational cost comparison

The computation cost of the row-1 (I3D-flow-1) and row-2 (I3D-flow-2) of Table-1 is shown in Table 2. As shown in Table-1, the I3D-flow-2 model is trained using various temporal resolutions. The result is reported after both model's training losses converged, which takes around 300 epochs. It is worthwhile to note that using different spatio-temporal resolutions in training and testing significantly aids in reducing the computational cost. This method is inspired by the recent multi spatio-temporal resolution based action recognition methods. [3], [11].

Table 2. Computational cost comparison for different training strategies

Model	Memory required for a forward pass (MB)	Accuracy (%)
I3D-flow-1	5129.64	84.91
I3D-flow-2	1410.07	86.72

4 Conclusion

In this paper, we presented a two-stream architecture based action classifier, using various spatio-temporal resolutions and augmentations in both training and inference. We use an ensemble of different architectures with the same input as well as an ensemble of the same architecture with inputs of different spatio-temporal resolutions. We also observed that taking an average of the predictions over multiple augmented clips aids in boosting the action recognition performance as well as reducing computational cost in training.

References

1. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6299–6308 (2017)
2. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
3. Feichtenhofer, C.: X3d: Expanding architectures for efficient video recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 203–213 (2020)
4. Jhuang, H., Garrote, H., Poggio, E., Serre, T., Hmdb, T.: A large video database for human motion recognition. In: Proc. of IEEE International Conference on Computer Vision. vol. 4, p. 6 (2011)
5. Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., Fei-Fei, L.: Large-scale video classification with convolutional neural networks. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 1725–1732 (2014)
6. Simonyan, K., Zisserman, A.: Two-stream convolutional networks for action recognition in videos. In: Advances in neural information processing systems. pp. 568–576 (2014)
7. Soomro, K., Zamir, A.R., Shah, M.: Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402 (2012)
8. Tran, D., Bourdev, L., Fergus, R., Torresani, L., Paluri, M.: Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision. pp. 4489–4497 (2015)
9. Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., Paluri, M.: A closer look at spatiotemporal convolutions for action recognition. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. pp. 6450–6459 (2018)

10. Wang, H., Schmid, C.: Action recognition with improved trajectories. In: Proceedings of the IEEE international conference on computer vision. pp. 3551–3558 (2013)
11. Wu, C.Y., Girshick, R., He, K., Feichtenhofer, C., Krahenbuhl, P.: A multigrid method for efficiently training video models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 153–162 (2020)
12. Yue-Hei Ng, J., Hausknecht, M., Vijayanarasimhan, S., Vinyals, O., Monga, R., Toderici, G.: Beyond short snippets: Deep networks for video classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4694–4702 (2015)
13. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime tv-l 1 optical flow. In: Joint pattern recognition symposium. pp. 214–223. Springer (2007)