Team DEEP-HRI Moments in Time Challenge 2018 Technical Report

Chao Li, Zhi Hou, Jiaxu Chen, Yingjia Bu, Jiqiang Zhou, Qiaoyong Zhong, Di Xie and Shiliang Pu
Hikvision Research Institute

Abstract

Video-based action recognition is challenging as spatial and temporal reasonings are involved jointly. We propose a novel multi-view convolutional architecture, which performs 2D convolution along three orthogonal views of volumetric video data. With weight sharing, it is capable of encoding spatio-temporal feature of video clips efficiently, and achieves superior performance over state-of-the-art spatio-temporal feature learning architectures. Furthermore, we also explore the auditory modality, which is complementary to visual clues. Our final submission to the Moments in Time challenge 2018 is an ensemble of several visual RGB and audio models, achieving a top-1 accuracy of 38.7% and top-5 66.9% on the validation set.

1 Introduction

The task of video-based action recognition requires proper modelling of both visual appearance and motion pattern. Recently, a significant effort has been devoted to spatio-temporal feature learning from video clips. Since the success of convolutional neural networks (CNN) in 2D image recognition [1], 3D convolution is a natural adaption for volumetric video data [2]. However, in C3D [2], significantly more (e.g. 2×) parameters than its 2D counterpart are introduced, which makes the model difficult to train and prone to overfitting. This issue is particularly critical when the training data size is limited. P3D [3] and (2+1)D [4] attempted to address the issue by decomposing a 3D convolution into a 2D convolution along the spatial dimension and a 1D convolution along the temporal dimension. We argue that the "unequal" treatment of spatial and temporal features is undesirable. On the contrary, we propose Multi-View CNN (MV-CNN), which performs feature extraction along the spatial and temporal dimensions in a consistent way. The details of MV-CNN are described in Section 2.1. To further improve the overall accuracy, we train non-local networks [5] for model ensemble.

2 Method

In this work, we explore multiple modalities for categorizing the action occurring in a video. Our visual RGB model is based on an ensemble of the proposed MV-CNN and other state-of-the-art spatio-temporal feature learning models. We also tried optical flow, but found that it do not contribute to the final accuracy after ensemble. However, we do exploit audio-based action recognition, which is complementary to visual signal.

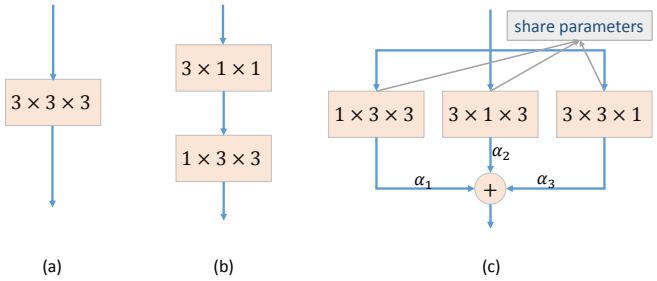

Figure 1: Comparison of MV-CNN to common spatio-temporal feature learning architectures. (a) C3D. (b) (1+2)D. (c) the proposed MV-CNN.

2.1 Multi-View CNN

A video clip can be represented as a 3D array of dimension $T \times H \times W$, where T, H and W are number of frames, frame height and frame width respectively. Taking kernel size of 3 as an example, Figure 1 compares the proposed MV-CNN to common convolutional architectures. In C3D, a 3D $3 \times 3 \times 3$ convolution is utilized to extract spatial (H and W) and temporal (T) features jointly. In the (1+2)D configuration, a 1D $3 \times 1 \times 1$ convolution is utilized to aggregate temporal feature, followed by a 2D $1 \times 3 \times 3$ convolution for spatial feature. While in the proposed MV-CNN, we perform 2D 3×3 convolutions along three views of the $T \times H \times W$ volumetric data, i.e. $T \times H$, $T \times W$ and $H \times W$ separately. The three orthogonal views are conceptually similar to the three anatomical planes of human body, namely sagittal, coronal and transverse. Notably, the parameters of the three-view convolutions are shared, such that the number of parameters is kept the same as single-view 2D convolution. The three resulting feature maps are further aggregated with weighted average. The weights are also learned during training in an end-to-end manner. To facilitate training, we initialize the 2D convolutional kernels with a ImageNet [6] pretrained model.

For each model, to obtain better generalization on the test

set, the Stochastic Weight Averaging (SWA) scheme [7] is adopted. Several model variants of the same network are trained with cycle learning rate and subsequently form an ensemble.

2.2 Auditory Modality

Complementary to visual signal, sound conveys important information for action recognition. Therefore, in our method, audio streams extracted from videos are exploited for the task of action categorization. In audio processing, log-mel spectrum is a powerful hand-tuned feature, exhibiting locality in both time and frequency domains [8]. In ResNet-34 [9], the 2D log-mel feature is cast into an image, and a 34-layer ResNet is applied for audio classification. While M34-res [10] and EnvNet [11] attempted to learn semantic feature from the 1D raw audio waveforms in an end-to-end way. We train the three state-of-the-art models on the Moments in Time dataset. Notably, we adapt EnvNet [11] with residual connections, and henceforth refer to the variant as EnvNet+ResNet.

3 Experiments

The Moments in Time dataset [12] contains 802245 training videos and 39900 validation videos. Excluding the videos without audio track, the auditory dataset contains 450k training segments and 20k validation segments. In total 339 action categories are annotated. In all experiments, our models are trained on the provided Moments in Time training data only. Apart from ImageNet, no other video datasets are used for pretraining.

For the visual RGB model, during training, we select 64 continuous frames from a video and then sample 8 frames by dropping the 7 frames in between. The spatial size is 224×224 pixels, randomly cropped from a scaled video whose shorter side is randomly sampled between 256 and 320 pixels. During inference, following [5] we perform spatially fully convolutional inference on videos whose shorter side is rescaled to 256 pixels. While for the temporal domain, we sample 6 clips evenly from a full-length video and compute softmax scores on them individually. The final prediction is the averaged softmax scores of all clips.

In this work, we use ResNet-101 [13], Inception-v4 and Inception-ResNet-v2 [14] as the backbone models, which are pretrained on ImageNet. The proposed MV-CNN along with C3D and non-local (NL) models are trained to form an ensemble. The top-1 and top-5 accuracies of individual models as well as their ensemble are shown in Table 1. For Inception-ResNet-v2, MV-CNN obtains 35.6% top-1 and 63.6% top-5 accuracy, leading to 0.5% and 0.3% accuracy gain compared with the C3D baseline. It is worth noting that with MV-CNN, more significant performance gain can be obtained on smaller sized datasets like UCF-101 [15]. On large-scale datasets like Moments in Time, the performance gain saturates, which is reasonable as increasing data size could be more effective than algorithmic innovations. With an ensemble of visual RGB models alone, we achieve a top-1 accuracy of 37.7% and top-5 65.9%.

For the training of audio models, all the sound data are downsampled to a frequency of 16kHz. For M34-res, we train two models for audio section lengths of 1s and 3s separately. Then their scores are averaged. This multi-scale training and inference scheme improves the robustness against audio length. The performances of the three audio models are summarized in Table 1. With an ensemble of audio models alone, we obtain 17.6% top-1 and 31.1% top-5 accuracy. With an ensemble of visual RGB and audio models, we achieve a top-1 accuracy of 38.7% and top-5 66.9%.

Table 1: Accuracy on the validation set of the Momements in Time dataset. Performances of both individual visual and audio models and their ensemble are shown.

Model	Modality	Accuracy (%)	
		top1	top5
ResNet-101-C3D	RGB	33.6	61.2
ResNet-101-NL	RGB	32.8	60.8
Inception-v4-C3D	RGB	34.3	62.0
Inception-ResNet-v2-C3D	RGB	35.1	63.3
Inception-ResNet-v2-NL	RGB	34.8	63.3
Inception-ResNet-v2-MV	RGB	35.6	63.6
Ensemble	RGB	37.7	65.9
ResNet-34	Audio	13.8	23.6
M34-res	Audio	14.8	27.4
EnvNet+ResNet	Audio	13.2	25.9
Ensemble	Audio	17.6	31.1
Ensemble	RGB+Audio	38.7	66.9

4 Conclusions

In our submission to the Moments in Time challenge 2018, we explore multiple modalities for the task of video-based action recognition. Particularly, we propose a novel multi-view convolutional architecture, which achieves superior performance over the C3D baseline with significantly less number of parameters. A more thorough and systematic evaluation of the architecture is left for future work.

References

[1] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[2] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri, "Learning spatiotemporal features with 3d convolutional networks," pp. 4489–4497, 2014.

[3] Zhaofan Qiu, Ting Yao, and Tao Mei, "Learning spatio-temporal representation with pseudo-3d residual networks," pp. 5534–5542, 2017.

[4] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann Lecun, and Manohar Paluri, "A closer look at spatiotemporal convolutions for action recognition," 2017.

[5] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He, "Non-local neural networks," 2017.

[6] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[7] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson, "Averaging weights leads to wider optima and better generalization," 2018.

[8] Ossama Abdel-Hamid, Abdel Rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional neural networks for speech recognition," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 22, no. 10, pp. 1533–1545, 2014.

[9] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, and Bryan Seybold, "Cnn architectures for large-scale audio classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 131–135.

[10] Wei Dai, Chia Dai, Shuhui Qu, Juncheng Li, and Samarjit Das, "Very deep convolutional neural networks for raw waveforms," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 421–425.

[11] Yuji Tokozume and Tatsuya Harada, "Learning environmental sounds with end-to-end convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2017, pp. 2721–2725.

[12] Mathew Monfort, Bolei Zhou, Sarah Adel Bargal, Tom Yan, Alex Andonian, Kandan Ramakrishnan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, et al., "Moments in time dataset: one million videos for event understanding," .

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," pp. 770–778, 2015.

[14] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alex Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," 2016.

[15] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah, "Ucf101: A dataset of 101 human actions classes from videos in the wild," *arXiv preprint arXiv:1212.0402*, 2012.