CUHK&SIAT Submission for THUMOS15 Action Recognition Challenge

Limin Wang[1] Zhe Wang[2] Yuanjun Xiong[1] Yu Qiao[2]

[1]Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong

[2]Shenzhen key lab of Comp. Vis. & Pat. Rec., Shenzhen Institutes of Advanced Technology, CAS, China

07wanglimin@gmail.com, buptwangzhe2012@gmail.com, yjxiong@ie.cuhk.edu.hk, yu.qiao@siat.ac.cn

Abstract

This paper presents the method of our submission for THUMOS15 action recognition challenge. We propose a new action recognition system by exploiting very deep two-stream ConvNets and Fisher vector representation of iDT features. Specifically, we utilize those successful very deep architectures in images such as GoogLeNet and VGGNet to design the two-stream ConvNets. From our experiments, we see that deeper architectures obtain higher performance for spatial nets. However, for temporal net, deeper architectures could not yield better recognition accuracy. We analyze that the UCF101 dataset is relatively very small and it is very hard to train such deep networks on the current action datasets. Compared with traditional iDT features, our implemented two-stream ConvNets significantly outperform them. We further combine the recognition scores of both two-stream ConvNets and iDT features, and achieve 68% *mAP value on the validation dataset of THUMOS15.*

1. Introduction

Human action recognition has been one of the most challenging problems in computer vision and received a great amount of research interests in recent years [6, 11, 12, 13, 15]. THUMOS [2] action recognition challenge is becoming an important contest to advance the current research and evaluate the performance of the state-of-the-art recognition system. In this paper, we describe the method of our submission for THUMOS15 action recognition challenge.

In previous research works, there are mainly two styles of algorithms for action recognition. The first style is *low-level features with Bag of Visual Words representation* [5], and the second one is applying *deep neural networks* to perform action recognition in an end-to-end manner [6]. The most successful low-level feature is the Improved Trajectories [11] and the most competitive deep network architecture is the Two-Stream ConvNet [6]. Several works [3, 4, 14] have made efforts to combine these two kinds of approaches in the last THUMOS action recognition challenge.

Following our previous method [14], we propose a new action recognition system for temporal untrimmed videos by fusing the recognition results of deep networks and BoVW representations of iDTs. In particular, we focus on studying the performance of very deep networks trained from two modality, namely RGB and optical flow fields, in this submission for THUMOS15 action recognition challenge.

The remainder of this paper is organized as follows. In Section 2, we explain our method in details. We report our experimental results on the validation dataset of THUMOS15 in Section 3. Finally, we conclude our paper in Section 4.

2. Our Approach

Our THUMOS15 solution is composed of three components: (i) very deep two-stream ConvNets, (ii) Fisher vector representation of iDT features, and (iii) video segmentation and classification. We will describe these three components in this section.

2.1. Very deep two-stream ConvNets

Very deep architectures have turned out to be effective on the tasks of object recognition [7, 10], scene recognition [18], and event recognition [16] in still images. We design a very deep two-stream architecture for action recognition in videos. Specifically, we try different network architectures for the design of two-stream ConvNets, ranging from deep ConvNets such as ClarifaiNet [19], to very deep ConvNets such as GoogLeNet [10] and VGGNet [7]. We choose three architectures for both spatial and temporal nets and the networks are listed in Table 1. The details about these network architectures can be found in their original papers.

Training deep networks on the UCF101 dataset is very challenging as the dataset is relatively small. For spatial nets, we choose to pre-train our networks on the dataset of ImageNet [1] and fine tune their weights on the dataset of UCF101 [8]. Specifically, the pre-trained models are avail-

Spatial nets			Temporal nets		
ClarifaiNet	GoogLeNet	VGGNet (16-layer)	ClarifaiNet	GoogLeNet	VGGNet (11-layer)
42.3%	53.7%	54.5%	47.0%	39.9%	42.6%

Table 1. Different network architectures and their performance on the THUMOS15 validation dataset.

able at the websites [1] [2]. For temporal nets. we choose to stacking 10-frame optical flow fields and train the network from the scratch on the UCF101 dataset. The detail about the training of our very deep two-stream ConvNets is the same with our previous work [15].

2.2. iDT features and Fisher vector representation

Low level features such as improved trajectories [11] have yielded good performance on the task of action recognition. We also exploit the improved trajectory features and extract four kinds of local descriptors, namely HOG, HOF, MBHx and MBHy. We then employ Fisher vector to encode these descriptors of a video clip into high dimensional representation as its effectiveness for action recognition has been verified in previous works [9, 17]. In order to train GMMs, we first de-correlate TDD with PCA and reduce its dimension to D by a factor of 2. Then, we train a GMM with K ($K = 256$) mixtures, and finally the video is represented with a $2KD$-dimensional vector. For multi-class classification, we train a linear SVM in a one-vs-all training scheme.

2.3. Video segmentation and classification

In order to perform action recognition in temporal untrimmed videos, we follow our previous method [14] and first temporally divide continuous videos into short clips. Different from previous method, we design a simple yet effective method to detect shot boundary by computing the color histogram and motion histogram. A shot will be detected if the color histogram changes larger than a threshold or the average motion magnitude is larger than a threshold.

For each clip, we employ two-stream ConvNets and SVMs to perform action recognition separately. Finally, we use the score-level fusion to combine the recognition results from ConvNets and SVMs. The final score for the whole video is obtained by averaging over these shot video clips.

3. Experiments

We train our model on the dataset of UCF101 and present our results on the validation dataset of THUMOS15. The experimental results of two-stream ConvNets with different architectures are listed in Table 1. We see that the very-deep networks (i.e. GoogLeNet and VGGNet) obtain better performance than the deep network (i.e. ClarifaiNet) for spatial nets. However, for temporal nets, very deep architectures achieve lower recognition performance than deep ones. We analyze that the temporal nets are trained from scratch and the UCF101 dataset is not enough to train such deep architectures.

Two-stream	iDTs+FV	Combine
63.7%	52.8%	68.1%

Table 2. The performance of two-stream ConvNets and iDTs+FV on the THUMOS15 validation dataset.

We also report the performance of two-stream ConvNets, iDT features with Fisher vector, and their combination on the THUMOS15 validation dataset in Table 2. From these results, we observe that there is a significant improvement for two-stream ConvNets over traditional iDT features with Fisher vector (around 10%). To our best knowledge, this the first time that deep learning methods significantly outperform the traditional low-level representations. These better results may be ascribed to the more deeper network architectures. We combine the recognition scores from both methods, and find that it is capable of further boosting recognition performance by around 5%.

4. Conclusions

This paper has proposed a new action recognition method from temporal untrimmed videos, by combining two-stream ConvNets and Fisher vector representation of iDT features. The results show that two-stream ConvNets significantly outperform traditional iDT features and the fusion of them is able to further boost the recognition performance.

Acknowledgement

This work is supported by a donation of Tesla K40 GPU from NVIDIA corporation. Limin Wang and Yuanjun Xiong are supported by Hong Kong PhD Fellowship. Yu Qiao is supported by National Natural Science Foundation of China (91320101, 61472410), Shenzhen Basic Research Program (JCYJ20120903092050890, JCYJ20120617114614438, JCYJ20130402113127496), 100 Talents Program of CAS, and Guangdong Innovative Research Team Program (No.201001D0104648280).

[1] http://www.robots.ox.ac.uk/~vgg/software/deep_eval/

[2] https://github.com/BVLC/caffe/wiki/Model-Zoo

References

[1] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009. 1

[2] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes. http://www.thumos.info/, 2015. 1

[3] M. Jain, J. van Gemert, and C. G. M. Snoek. University of amsterdam at thumos challenge 2014. In *THUMOS14 Action Recognition Challenge*, 2014. 1

[4] D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014. In *THUMOS14 Action Recognition Challenge*, 2014. 1

[5] X. Peng, L. Wang, X. Wang, and Y. Qiao. Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *CoRR*, abs/1405.4506, 2014. 1

[6] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, pages 568–576, 2014. 1

[7] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 1

[8] K. Soomro, A. R. Zamir, and M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[9] C. Sun and R. Nevatia. Large-scale web video event classification by use of fisher vectors. In *WACV*, pages 15–22, 2013. 2

[10] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. *CoRR*, abs/1409.4842, 2014. 1

[11] H. Wang and C. Schmid. Action recognition with improved trajectories. In *ICCV*, pages 3551–3558, 2013. 1, 2

[12] L. Wang, Y. Qiao, and X. Tang. Mining motion atoms and phrases for complex action recognition. In *ICCV*, pages 2680–2687, 2013. 1

[13] L. Wang, Y. Qiao, and X. Tang. Motionlets: Mid-level 3D parts for human motion recognition. In *CVPR*, pages 2674–2681, 2013. 1

[14] L. Wang, Y. Qiao, and X. Tang. Action recognition and detection by combining motion and appearance features. In *THUMOS14 Action Recognition Challenge*, 2014. 1, 2

[15] L. Wang, Y. Qiao, and X. Tang. Action recognition with trajectory-pooled deep-convolutional descriptors. In *CVPR*, pages 4305–4314, 2015. 1, 2

[16] L. Wang, Z. Wang, W. Du, and Y. Qiao. Object-scene convolutional neural networks for event recognition in images. In *CVPR, ChaLearn Looking at People Workshop*, pages 30–35, 2015. 1

[17] X. Wang, L. Wang, and Y. Qiao. A comparative study of encoding, pooling and normalization methods for action recognition. In *ACCV*, pages 572–585, 2012. 2

[18] Z. Wu, Y. Zhang, F. Yu, and J. Xiao. A GPU implentation of googlenet. Technical report, 2014. 1

[19] M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 1