Moments in Time:
submission to ActivityNet Challenge 2018

Shu-Dong Yang, Zhi-Wei Ren, De-Ming Cong, Di Wen, Kong Ye,
Jun-Yan He, Xiao Wu

Southwest Jiaotong University
{shudong.yang, vincentren}@my.swjtu.edu.cn,
sc16d2w@leeds.ac.uk
{congdeming1995, tiankong2586,junyanhe1989,wuxiaohk}@gmail.com

Abstract

This paper describes our method for the Moments in Time Recognition Challenge Full track to ActivityNet Challenge 2018. In this task, we propose a method for action recognization by using Non-local Neural Networks,the Deformable Convolutional Networks and Temporal Relational Reasoning in Videos. We further demonstrate that a ConvNet trained audio information can help with the recognization of the Moments in Time dataset. We only use RGB frames and audio features to training.

1. Introduction

In recent years, Video-based human action recognition has become an intensive area of research in the fields of computer vision and pattern recognition [5, 6, 7]. There are two significant information in this fields: appearance and motion. For appearance, since hand-craft features such as sift are unable to capture global information, numerous recent methods have utilized Convolutional Neural Networks (CNNs) whose superiority over hand-crafted ones in this field has been shown. [2] As for motion, it is frequently represented by optical flow or other motion-based descriptors. Simonyan et al. [4] 's two-stream CNN network which employed optical flow into temporal network to extract motion information. However, the pre-calculation of optical flow is significantly complicated which illustrated the exceeding difficulty of applying it into real-time recognition.We propose an approach for human action recognition which fused various CNNs (Audio-TSN, TRN [7], DCN[1] , Non-local Network [6]) to extract different features from different videos with an end-to-end training process. Considering the application of real-time action recognition, optical flow has not been implemented in our approach. In this submission to the challenge, we aim to evaluate the proposed model on the Moments in Time dataset

2. Moments in Time

Moments in Time Dataset is a large-scale human-annotated collection of one million short videos which has the same length of 3 second. There are 339 different classes in total. And there are existing some action partly or even fully depend on the audio information. Moments in Time dataset is also joint as a task in the ActivityNet Challenge 2018. There are two different tracks. The first track is the full track, which is a classification task on the entire Moments in Time dataset. It contains 339 classes, 802,264 training videos, 33,900 validation videos, and 67,800 testing videos. The second track is the mini track, which is a classification take for students on a sub set of Moments in time dataset. It contains 200 classes, 100,000 training videos, 10,000 validation videos, and 20,000 testing videos.

3. Method

Our approach use 3 kind of neural network for extract apperance features from RGB image: Non-local Network [6], Temporal Relation Network (TRN) [7], Deformable ConvNets (DCN). We found that audio information also play a important role in video analyzing, so we also use audio feature to imporve our recognition accuracy.

We only use RGB frames adn audio features as our training data. Therefore we lost some temporal information from still image, so we choose 3D-based convolution neural networks (CNN) to exploit temporal information from continuous frames. Because of Non-local Network is the most powerful architecture in 3D-CNN, so we choose it for temporal feature learning. Relevance also exists in contiguous video frames. For instance, in the action of drinking, taking the glass should be anterior to getting the mouse close to it. The relevancy of action makes it irreversible. So we use TRN for extract temporal relationship between continuous frames. And we think in still image, there is also an relation between objects such as bottle and person. In our approach, we do not use object detection methods such as Faster R-CNN [3] or YOLO directly. We use the deformable ConvNets for spatial relationship learning because it success in oject detection area. Audio is also an important feature in our approach, we use mel spectrogram feature as our training data, then use Temporal Segment Network (TSN) to training.

3.1. Training

1. For Non-local Network, we use both i3d and c2d as Non-local Network's backbone, the network is trained using SGD for 400k iterations. The base learning rate is 0.01 and the stepsize is 150k and 300k.

2. When traning TRN network, we use the released pre-trained RGB model for full track. For mini track, we use InceptionV3 as network's backbone, and we choose 8-segments and multiscale strategy for training.

3. For DCN network, we use ResNet101 as network's backbone, and replace the res5-c bottleneck to DCN ConvNets.

4. For audio stream, we first extract wav file from video and use audioset to get mel spectrogram feature, then use TSN network to training.

4. Conclusion

Although we do not get the best performance in competition, but it shows that relationship in continuous frames plays an importanct role in video analyzing. And we found audio also can help effectively.

References

[1] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. *CoRR, abs/1703.06211*, 1(2):3, 2017.

[2] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[3] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[4] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[5] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Val Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016.

[6] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. *CVPR*, 2018.

[7] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. *arXiv:1711.08496*, 2017.