TDU&AIST Submission for ActivityNet Challenge 2018 in Video Caption Task

Tenga Wakamiya[1]* Takumu Ikeya[1]* Akio Nakamura[1] Kensho Hara[2] Hirokatsu Kataoka[2]
Tokyo Denki University[1]
National Institute of Advanced Industrial Science and Technology (AIST)[2]
{wakamiya.t, ikeya.t}@is.fr.dendai.ac.jp, nkmr-a@cck.dendai.ac.jp
{kensho.hara, hirokatsu.kataoka}@aist.go.jp

Abstract

In the report, we introduce our video caption approach for the ActivityNet Challenge in conjunction with CVPR 2018. Based on the 3D-ResNet with 34-layer [1, 2] and LSTM-based Sentence Generator [5], our captioner generates a suitable sentence along an input video. The captioning model is trained with the training-set on ActivityNet database. In the experimental section, we show our rate on the test-set with evaluation server. Finally, we achieved to put our name on the leaderboard![1]

1. Introduction

The task of finding a de facto standard for video recognition has advanced with both hand-crafted and deeply learned feature representations. In the recent DNN-based video recognition, we are focusing on 3D convolutional networks such as C3D [4] and 3D-ResNets [1, 2].

On one hand, video caption which includes time duration seems to be very difficult issue in the current vision-based algorithm. The open problem is composed by two problem, namely (i) video representation in order to generate an appropriate sentence, and (ii) temporal segmentation to fix an event duration. We believe that the video representation problem is more important. Therefore we apply a sophisticated video representation 3D-ResNet with layer-34 for video caption. To generate a video caption, we apply a standard sentence generator LSTM based on the Google's Show and Tell algorithm [5].

*denotes equal contribution

[1]Two bachelor students have tried very challenging task, namely "can CV-research beginners achieve the ActivityNet Challenge in two months?" Although our rate is far from competitive performance, we succeeded to list our team on the leaderboard.

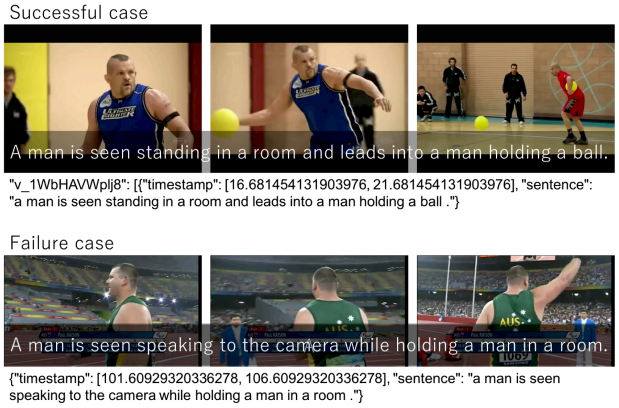

Figure 1. The successful (top) and failure (bottom) cases in ActivityNet Challenge with video caption.

2. Proposed approach: 3D-ResNet-34 + LSTM

We simply combine 3D-ResNet-34 with LSTM. To train/test the LSTM, the layer after global average pooling (2,048-d vector) is inserted from 3D-ResNet. The 3D-ResNet-34 is pretrained by Kinetics dataset [6] and the 3D-ResNet-34+LSTM is trained by ActivityNet caption [3] with end-to-end training manner.

3. Result on video caption task

Our performance value with METEOR is 0.6266. The score is far from top-ranked captioners from other teams. The result is coming from fewer proposals per a video. Our temporal proposals with start- and end-time are only 2 per a video. In the future, we would like to evaluate the video captioner with e.g. over 100 proposals in video. Moreover, we must implement an improved temporal proposals and more sophisticated models such as 3D-ResNet-{50, 101, 152}, 3D-ResNeXt-101.

References

[1] K. Hara, H. Kataoka, and Y. Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. International Conference on Computer Vision Workshop (ICCVW), 2017. 1

[2] K. Hara, H. Kataoka, and Y. Satoh. Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and ImageNet? IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. 1

[3] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles. Dense-captioning events in videos. In *International Conference on Computer Vision (ICCV)*, 2017. 1

[4] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. International Conference on Computer Vision (ICCV), 2015. 1

[5] Vinyals, O. and Toshev, A. and Bengio, S. and Erhan, D. Show and tell: A neural image caption generator. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015. 1

[6] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman and Andrew Zisserman. The kinetics human action video dataset. arXiv:1705.06950, 2017. 1