Trimmed Event Recognition：submission to ActivityNet Challenge 2018

Lei Zhou, Jiaze Wang, Xiaojiang Peng, Yali Wang, Yu Qiao

Shenzhen Institutes of Advanced Technology, CAS, China

Abstract

This notebook paper describes our system for the trimmed event recognition (Moments in Time) task in the ActivityNet challenge 2018. We investigate multiple state-of-art approaches for the event recognition in short, trimmed videos. With these approaches, we derive an ensemble of deep models.

1. Introduction

Event recognition has remained a challenging task in the computer vision community. The research about event recognition also are very important in other tasks like video understanding. And Moments in Time dataset is a large-scale human-annotated collection of one million short videos corresponding to dynamic events unfolding within three seconds [1].

The rest of this paper is organized as follows. Section 2 presents our approach in detail, finally Section 3 concludes this work.

2.Experments

2.1 Data Augmentation

We fix the size of input image or optical flow fields as 256 × 340, and the width and height of cropped region are randomly selected from {256,224,192,168}. Finally, these cropped regions will be resized to 224×224 for network training. What's more, we use additional three-quarters validation set as a part of training set and use the other one-quarter validation set (Seen as Spilt1 Val) to evaluate the performance of our models.

2.1 CNN models

The main model we use is STRNet, i.e. Spatiotemporal Recalibration Networks. STRNets are 3D networks which aim to solve the temporal disturbance problem in vanilla 3D networks and factorized spatiotemporal networks. Due to features from different pipelines can capture different information. We use four other models to capture additional information. According to the ensemble results, they can significantly improve the performance on the on the Spilt1 validation set. These four models are Resnet101[2], TSN [3], TRN [4] and I3D [5].

2.2 Experiments results

Table 1 shows the Top-1 and Top-5 accuracy of the baseline models on the Spilt1 validation set. The best single model is the STR18_tr, with a Top-1 accuracy of 29.76% and a Top-5 accuracy of 56.71%.

Index	Model	Test Set	Top1	TOP5	Ave
A	STR18_tr	Spilt1 Val	**29.76%**	**56.71%**	**43.23%**
B	R101	Spilt1 Val	27.49%	52.41%	39.95%
C	I3D	Spilt1 Val	24.40%	49.37%	36.88%
D	TRN	Spilt1 Val	24.59%	48.90%	36.74%
E	STR18_tr_of	Spilt1 Val	17.98%	39.50%	28.74%
F	TSN	Spilt1 Val	24.67%	49.52%	37.10%
G	STR34	Spilt1 Val	28.64%	55.99%	42.31%

TABLE1: Classification Accuracy: We show Top-1 and Top-5 accuracy of the baseline models on the Spilt1 validation set.

As is shown is TABLE 2, the Ensemble model (average) gets the Top-1 accuracy as 32.08% and Top-5 accuracy as 59.23%.

Index	Test Set	Top1	TOP5	Ave
A	Spilt1 Val	29.76%	56.71%	43.23%
A+B	Spilt1 Val	31.59%	58.28%	44.93%
A+B+C	Spilt1 Val	31.23%	58.40%	44.81%
A+B+D	Spilt1 Val	31.91%	58.76%	45.33%
A+B+D+E	Spilt1 Val	31.82%	59.10%	45.46%
A+B+D+E+F	Spilt1 Val	31.82%	59.09%	45.46%
A+B+D+E+G	Spilt1 Val	**32.08%**	**59.23%**	**45.65%**

TABLE2: Ensemble Results: We show Top-1 and Top-5 accuracy of the ensemble models on the Spilt1 validation set.

3.Conclusion

This paper describes our team's solution to task of trimmed event recognition. Features from different pipelines can capture different information. We propose several 3D spatial-temporal models for event recognition. We also investigate the performance of several 2D CNNs like TSN, TRN and Resnet101.

References

[1] Monfort M, Zhou B, Bargal S A, et al. Moments in Time Dataset: one million videos for event understanding[J]. arXiv preprint arXiv:1801.03150, 2018.

[2] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778

[3] Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: Towards good practices for deep action recognition[C]//European Conference on Computer Vision. Springer, Cham, 2016: 20-36.

[4] Zhou B, Andonian A, Torralba A. Temporal Relational Reasoning in Videos[J]. arXiv preprint arXiv:1711.08496, 2017.

[5] Carreira J, Zisserman A. Quo vadis, action recognition? a new model and the kinetics dataset[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017: 4724-4733.