Trimmed Event Recognition Submission to ActivityNet Challenge 2018

Jiaqing Lin, Akikazu Takeuchi
STAIR Lab, Chiba Institute of Technology, Japan
{lin, takeuchi}@stair.center

1. Overview

This paper describes STAIR Lab submission to ActivityNet 2018 Challenge for guest task C: Trimmed Event Recognition (Moments in Time) [1]. Our approach is to utilize three networks, Audio Net, Spatial-temporal Net, and DenseNet to make individual predictions, then use MLP to fuses the results to make an overall prediction. The flow chart of our approach is shown in figure 1.

2. Implementation

2.1 Audio network

Our audio dataset training is different from other methods. Usually, auditory raw waveforms are used as input and are fed into a model like SoundNet [2]. In our case, firstly, we converted auditory raw waveforms to spectrogram images, then fed them to 2D ResNet101 [3] to train a classifier. The top-1 accuracy of this model is 13.04%, which is higher than top-1 accuracy 7.60% presented in [1].

2.2 Spatial-temporal network

We used 3D ResNet101 [4] to extract spatial-temporal visual features from a video. To train a classifier, a temporal position in an input video is randomly selected, and 16 frames are extracted around the selected temporal position. The frames are spatially cropped by multi-scale random four corner and center cropping, and horizontally flipped with 50% probability. Other parameters are same as the paper [4].

2.3 2D RGB network

Single frame in a video is still informative even in the action recognition. So we used DenseNet [5] for extracting image features from a randomly selected frame in a video. Number of layers was 201.

2.4 Fusion

We utilized the three models above to predict the test set. Log Softmax function is applied to the last layer of each model, and results are concatenated to generate two vectors, one including audio prediction, the other without audio prediction. Then, MLP is trained. Top-1 and top-5 accuracy of out method for the validation set are shown in table 1.

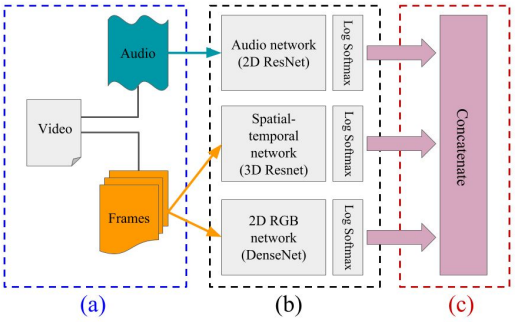

Fig. 1. (a) Extract audio and frames from a video as each network input. (b) Three networks are 2D ResNet, 3D ResNet, and DenseNet. (c) Fuse concatenated three results by using MLP.

Table 1

Model	Modality	Top-1(%)	Top-5(%)
2D ResNet101	Auditory	13.04	28.03
3D Resnet101	Spatial+Temporal	24.85	50.37
DenseNet	Spatial	24.5	48.4
Fusion (MLP)	A+S+T	29.97	57.26

References

[1] Monfort, Mathew, et al. "Moments in Time Dataset: one million videos for event understanding." *arXiv preprint arXiv:1801.03150* (2018).

[2] Aytar, Yusuf, Carl Vondrick, and Antonio Torralba. "Soundnet: Learning sound representations from unlabeled video." In *Advances in Neural Information Processing Systems*, pp. 892-900. 2016.

[3] He, Kaiming, et al. "Deep residual learning for image recognition." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016.

[4] Hara, Kensho, Hirokatsu Kataoka, and Yutaka Satoh. "Learning spatio-temporal features with 3D residual networks for action recognition." *Proceedings of the ICCV Workshop on Action, Gesture, and Emotion Recognition*. Vol. 2. No. 3. 2017.

[5] Huang, Gao, et al. "Densely connected convolutional networks." *Proceedings of the IEEE conference on computer vision and pattern recognition*. Vol. 1. No. 2. 2017.