AIST Submission to ActivityNet Challenge 2018

Kensho Hara, Hirokatsu Kataoka, Yutaka Satoh
National Institute of Advanced Industrial Science and Technology (AIST)
Tsukuba, Ibaraki, Japan
{kensho.hara, hirokatsu.kataoka, yu.satou}@aist.go.jp

Abstract

In this paper, we introduce our method for ActivityNet Challenge 2018 Task C (Moments in Time). We used a 3D convolutional neural network (CNN) pretrained on Kinetics-400, and finetuned it on Moments in Time. We experimentally evaluated the performance of our method.

1. Introduction

We focus on the trimmed event recognition task (Task C) in ActivityNet Challenge 2018. We use a 3D convolutional neural network (CNN) pretrained on Kinetics-400 [1] to recognize events. In our previous work [2], we trained various 3D architectures on Kinetics-400 and released them[1]. We use the pretrained ResNeXt-101 for the recognition on the Moments in Time dataset.

2. Implementation

We use stochastic gradient descent with momentum to train the network and randomly generate training samples from videos in training data in order to perform data augmentation. First, we select a temporal position in a video by uniform sampling in order to generate a training sample. A 16-frame clip is then generated around the selected temporal position. If the video is shorter than 16 frames, then we loop it as many times as necessary. Next, we randomly select a spatial position from the 4 corners or the center. In addition to the spatial position, we also select a spatial scale of the sample in order to perform multi-scale cropping. The scale is selected from $\left\{1, \frac{1}{2^{1/4}}, \frac{1}{\sqrt{2}}, \frac{1}{2^{3/4}}, \frac{1}{2}\right\}$. Scale 1 means that the sample width and height are the same as the short side length of the frame, and scale 0.5 means that the sample is half the size of the short side length. The sample aspect ratio is 1 and the sample is spatio-temporally cropped at the positions, scale, and aspect ratio. We spatially resize the sample at 112×112 pixels. The size of each sample is 3 channels $\times$ 16 frames $\times$ 112 pixels $\times$ 112 pixels, and each sample is horizontally flipped with 50% probability. We also perform mean subtraction, which means that we subtract the mean values of ActivityNet from the sample for each color channel. All generated samples retain the same class labels as their original videos.

In our training, we use cross-entropy losses and backpropagate their gradients. The training parameters include a weight decay of 0.001 and 0.9 for momentum. When finetuning the network, we start from learning rate 0.01, and divide it by 10 after the validation loss saturates.

Table 1: Accuracies on the Moments in Time validation set. *Average* is averaged accuracy over *Top-1* and *Top-5*.

Method	Top-1	Top-5	Average
ResNeXt-101	28.5	53.9	41.2

3. Experiments

Table 1 shows the results on the Moments in Time validation set. We recognized the videos of the test set using this network, and submitted the results.

4. Conclusion

In this paper, we described the submission for the Task C of ActivityNet Challenge 2018.

References

[1] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The Kinetics human action video dataset," *arXiv preprint*, vol. arXiv:1705.06950, 2017.

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

[1] https://github.com/kenshohara/3D-ResNets-PyTorch