SYSU iSEE submission to Moments in Time Challenge 2018

Shuosen Guan
School of Data and Computer Science
Sun Yat-Sen University
GuangZhou, China
gshuosen@gmail.com

Haoxin Li
School of Electronics and Information Technology
Sun Yat-Sen University
GuangZhou, China
LiHaoxin05@gmail.com

Abstract

This report introduces our submission to the Moments in Time Challenge 2018. In this task, we integrate static information, short-term temporal information, long-term temporal information and acoustic information to recognize the actions or events in the videos. Our method finally obtains top-1 accuracy of 27.9% in full-track validation set and 33.6% in mini-track validation set.

1. Introduction

Moments in Time dataset includes a collection of one million labeled 3 second videos, which aims to to help AI systems recognize and understand actions and events in videos.

In this report, we focus on learning different time scale representations for video classification and incorporating other sources of information such as audio signal to provide complementary information. In the following sections we will present our approach and show the results.

2. Approach

In order to understand the videos from multiple temporal scale, we combine static information, short-term temporal information and long-term temporal information via a simple late fusion. In addition, we utilize acoustic signal features since it provide complementary information. We ensemble these models to get the final predictions of the videos. Next we describe each component in detail.

2.1. Static Information

For static information, we exploit frame-based features to recognize actions or events. We deploy Inception-Resnet-V2[6] architecture with temporal segment networks[8] framework. During training, each video is divided into 3 segments and one frame is sampled from each segment. The frame-wise prediction is fused by average pooling. During testing, 20 frames equidistant in time are sampled and the predctions are averaged to genarate video-level prediction.

To improve performance, we fintune the model from ImageNet pretrained and Kinetics-400 pretrained ones. The model finetuned from Kinetics-400 pretrained model achieves higher accuracy. Besides, considering training on hard samples, we try to use focal loss[4] in this classification task and find that it just accelerated convergence but did't increase the performance.

Our performance comparison on validation set is showed in Table 1.

Models	Full-track Top-1	Mini-track Top-1
IR-scratch	0.1946	-
IR-ImageNet	0.2419	-
IR-Kinetics-400	**0.2524**	0.3026
IR-Kinetics-FL	0.2513	**0.3124**

Table 1. Performance comparison of different models for static information.(IR here denotes InceptionResnetV2.)

2.2. Short-term Temporal Information

To encode spatial and short-term temporal information, we apply Pseudo-3D Residual Networks[5] in our approach. We use 199 layers variant as our base framework and mix different P3D Blocks as described in [5]. In the training stage, one 16-frame clip is randomly sampled from each video as the input while during testing we sample 4 clips uniformly from each video and fuse the output of the final layer.

We first pretrain our model from Kinetics-400 dataset and then full-train the model on the Moments in Time dataset. For Mini-track, to accelerate the training process and capture longer term motion information, we experiment with different sampling strategies on the input: sampling clips from consecutive frames and down-sampling clips with different sampling intervals. Accuracy comparison on validation set is described in Table 2.

Models	Full-track Top-1	Mini-track Top-1
P3D-Kinetics	0.2091	**0.2634**
P3D-Kinetics-s2	-	0.2612
P3D-Kinetics-s4	-	0.2614

Table 2. Performance of different models using Pseudo-3D Residual Networks with different sampling interval. s2, s4 denote the sampling interval of 2 frames and 4 frames respectively.

2.3. Long-term Temporal Information

To capture long-term temporal information, we intend to model the temporal evolutions of features. We first extract frame-level features using our Kinetics pretrained Inception-Resnet-V2 model from 10 frames uniformly sampled from each video, and then apply a temporal convolution (denoted as TemporalConv or TC below for simplicity) and a parametric pooling along time dimension, which follows a MOE model like [1] to classification.

Inspired by ARTNet proposed in [7], we further employ a multiplicative interactions (denoted as MultiplyInter or MI below for simplicity) to model relations across features as a supplement to the TemporalConv features.

Moreover, Temporal Relation Network[9] models the temporal dependencies between multiple frames at multiple time scales. Here we use the pretrain model[1] provided by the author to model multi-scale temporal informatioin for classification.

Results of different methods on validation set are illustrated in Table 3.

Methods	Full-track Top-1	Mini-track Top-1
TemporalConv	0.2626	0.3251
MultiplyInter	**0.2638**	**0.3268**
TRN	0.2120	-

Table 3. Performance of different methods for long-term temporal information.

2.4. Acoustic Information

We also utilize acoustic features as complementary information in our approach. We first compute log mel spectrograms from the audio of each video and use a pre-trained VGGish model[3] to extract 128-D semantically meaningful, high-level embedding features[2], and then take the features as input and use a 4 layers full-connected network for classification. We finally obtain 0.045 top-1 accuracy on validation set.

[1] `http://relation.csail.mit.edu/models/TRN_moments_RGB_InceptionV3_TRNmultiscale_segment8_best.pth.tar`

3. Ensemble Results

Finally, we ensemble the models mentioned above to get the prediction. Results on Full-track and Mini-track are showed in Table 4 and Table 5 respectively. It should be noted that in both two tracks, we use the consecutive-sampling strategy mentioned above in P3D models for final combinations.

Models combinations	Top-1	Top-5
IR+P3D	0.2638	0.5187
IR+P3D+TRN	0.2676	0.5262
IR+P3D+TC+TRN	0.2746	0.5345
IR+P3D+TC+MI+TRN	0.2786	0.5368
IR+P3D+TC+TRN+audio	0.2756	0.5291
IR+P3D+TC+MI+TRN+audio	**0.2796**	**0.5397**

Table 4. Top-1 and Top-5 accuracy of different models combinations on Full-track.

Models combinations	Top-1	Top-5
IR+P3D	0.3246	0.6082
IR+P3D+TC	0.3337	0.6196
IR+P3D+MI	0.3347	**0.6237**
IR+P3D+TC+MI	**0.3358**	0.6219

Table 5. Top-1 and Top-5 accuracy of different models combinations on Mini-track.

References

[1] S. Abu-El-Haija, N. Kothari, J. Lee, P. Natsev, G. Toderici, B. Varadarajan, and S. Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016.

[2] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

[3] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999 – 3007, Venice, Italy, Oct 2017. IEEE.

[5] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5534–5542. IEEE, 2017.

[6] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi. Inception-v4, inception-resnet and the impact of residual connections on

learning. In *AAAI Conference on Artificial Intelligence*, pages 4278 – 4284, San Francisco, California USA, 2017. AAAI.

[7] L. Wang, W. Li, W. Li, and L. Van Gool. Appearance-and-relation networks for video classification. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1430 – 1439, Salt Lake City, Utah, June 2018. IEEE.

[8] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pages 20–36, Amsterdam, Netherlands, 2016. Springer, Springer, Cham.

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