Technical Report of The MMact Challenge

Chen Zhang
OPPO Research Institute
zhangchen3@oppo.com

Chen Chen
OPPO Research Institute
chenchen@oppo.com

Xunqiang Tao
OPPO Research Institute

Yandong Guo
OPPO Research Institute

1. Introduction

MMAct [1] is a new large-scale multi modality human action understanding dataset which combines RGB videos, key points, and sensor signals including acceleration, gyroscope, orientation, etc.

In the MMAct challenge, we only use the RGB videos in the MMAct dataset and achieve AP **94.68** and AP **91.08** respectively on the cross-scene and cross-view Trimmed Action Recognition tasks, and AP **40.68** on the Untrimmed Action Temporal Localization task.

For the Trimmed Action Recognition tasks, we train multiple end-to-end action recognition networks for cross-scene and cross-view respectively, including CSN [2], SlowFast [3] and TPN [4] which are the typical of applying 3D CNN to action recognition and the state-of-the-art on Sports1M [5] and Kinetics [6]. We get classification probability for each trimmed video by forward propagation on all the CNN models and get the final classification results by an ensemble strategy.

For the Untrimmed Action Temporal Localization task, we get video segments by clipping the untrimmed videos according to the timestamp in the annotation and set the corresponding label for the obtained video segments. Then, we train a CSN on the video segments obtained by clipping untrimmed video set as a classifier. Finally, we use a sliding window combined with the classifier we trained to perform the first rough temporal positioning of the actions in the untrimmed videos based on the scores output by the classifier, and after observing the results of the first temporal positioning on the validation dataset, we get the final submission by settings the post-processing including merging the adjacent same categories and shifting the timestamp proportionally class by class, etc.

2. Method

In this section,we will separately introduce our methods on the two tasks of the MMAct challenge.

2.1. Task 1. Trimmed Action Recognition

We merge the three categories of carrying, carrying_light, and carrying_heavy into the carrying, and perform a second classification of the cases that be predicted to be the carrying. Therefore, for Trimmed Action Recognition task, each set of our models contains a 33-classification model and a three-classification model used to classify the three classes of carrying, carrying_light and carrying_heavy.

Since most of the trimmed video data in the MMAct dataset is a 4-second snippet at 24fps, we use 32 frames sampled at 3 frame intervals on the video stream as input to train CSN, SlowFast, and TPN on the trimmed cross-scene video dataset and the trimmed cross-view video dataset respectively. We submit the results of the above three models to the evaluation of the MMAct Challenge, and the AP are recorded in Table 1 and Table 2. We find that the performance of CSN is better than SlowFast and TPN, so in our ensemble strategy, CSN has a larger weight coefficient than SlowFast and TPN.

For video clips with a length of less than 4 seconds in the trimmed video data, we perform a loop sampling. And in order to cover video clips longer than 4 seconds, we modified the input video stream settings on the best-performing CSN to use 48 frames sampled at 3 frame intervals on the video stream and 48 frames sampled at 4 frame intervals on the video stream as input corresponding to 6-second and 8-second trimmed videos, respectively.

Our ensemble strategy is to sum the model's output classification probability p_n after be multiplying by the weight α_n as equation 1:

$$P = \alpha_1 p_1 + \alpha_2 p_2 + ... + \alpha_n p_n \tag{1}$$

2.2. Task 2. Untrimmed Action Temporal Localization

For the Untrimmed Action Temporal Localization task, we also merge the three classes of carrying, carrying_light, and carrying_heavy into the carrying, and use a three-

Table 1. Part of our experiment for Trimmed Action Recognition (AP%) cross-scene.

id	ensemble index	ensemble weights	model	clip length	frame interval	x-scene AP
1			CSN	32	3	94.11
2			SlowFast	32	3	91.82
3			TPN	32	3	90.02
4			CSN	48	3	93.23
5			CSN	48	4	92.63
6	1/2/3	0.5/0.25/0.25				94.42
7	1/2/3/4/5	0.2/0.2/0.1/0.2/0.1				94.68

Table 2. Part of our experiment for Trimmed Action Recognition (AP%) cross-view.

id	ensemble index	ensemble weights	model	clip length	frame interval	x-view AP
1			CSN	32	3	90.62
2			SlowFast	32	3	82.99
3			TPN	32	3	84.95
4			CSN	48	3	90.63
5			CSN	48	4	89.32
6	1/2/3	0.5/0.25/0.25				90.64
7	1/2/3/4/5	0.2/0.2/0.1/0.2/0.1				91.08

classification model to classify the cases that be predicted to be the carrying which is the same as the Trimmed Action Recognition task.

We get the training video clips from the untrimmed video according to the timestamp in the annotation, and train the CSN on the video clips with 32 frames sampled at 3 frame intervals as input.

After training, we use a sliding window with a stride size of 3 frames on the untrimmed video, and each window outputs the rough classification results within this time period.

We mainly use four types of post-processing to refine the classification and timestamp results output by the sliding window:

1. Set the minimum and maximum duration thresholds for each category

2. Set the confidence threshold for each category

3. For adjacent results of the same category, if the time interval is less than the threshold, the two results are merged

4. Shrink or expand the timestamp class by class

3. Results

Some of our experimental results for the Trimmed Action Recognition task are recorded in Table 1 and Table 2.

For the Untrimmed Action Temporal Localization task, we finally achieved AP 40.68 on the test set by adjusting the post-processing threshold on the validation set.

References

[1] Q. Kong, Z. Wu, Z. Deng, M. Klinkigt, B. Tong, and T. Murakami. Mmact: A large-scale dataset for cross modal human action understanding. In *International Conference on Computer Vision*.

[2] D. Tran, H. Wang, L. Torresani, and M. Feiszli. Video classification with channel-separated convolutional networks. 2019.

[3] C. Feichtenhofer, H. Fan, J. Malik, and K He. Slowfast networks for video recognition. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[4] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] A. Karpathy, G. Toderici, S. Shetty, T. Leung, and F. F. Li. Large-scale video classification with convolutional neural networks. In *Computer Vision Pattern Recognition*, 2014.

[6] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.