A One-Stage Method for FineAction Localization from Multiple Views

Yepeng Tang[1,2†], Weining Wang[3], Chunjie Zhang[1,2*], Jie Jiang[3,4], Weitao Yuan[3,4], Sihan Chen[3,4], Jing Liu[3,4], Yao Zhao[1,2]

[1] Institute of Information Science, Beijing Jiaotong University
[2] Beijing Key Laboratory of Advanced Information Science and Network Technology, Beijing Jiaotong University
[3] Institute of Automation, Chinese Academy of Sciences
[4] School of Artificial Intelligence, University of Chinese Academy of Sciences

1 Introduction

In the field of intelligent video analysis, human action analysis is the core problem, and it is also a research hotspot of computer vision and pattern recognition, which has a wide range of applications in security monitoring and Internet video search. Temporal action localization (TAL) is one of the key tasks in video action analysis. Given a long untrimmed video, the goal of TAL is to output the action category contained in the video, and the start time and end time. Earlier benchmark datasets for TAL, e.g., THUMOS-14 [2] and ActivityNet-1.3 [1], mainly focus on coarse actions, where the temporal boundaries are often ambiguous. To address the problem, FineAction dataset [3] is proposed which provides fine-grained annotations of 103K temporal instances within 106 action categories, from 16,732 untrimmed videos.

To better locate the fine-grained action instances, it is important to extract reliable fine-grained video features. In this competition, we use two strong video pretraining models for feature extraction, i.e., VideoSwin [4] and X-CLIP [5]. These two models have excellent performance on video recognition, and could well capture both appearance and motion information in the video. In order to detect the actions from different perspective, we use video features in 13 views, including 12 view features of video crops and a mean vector of the 12 view features. After obtaining the video features, we follow an one-stage temporal action detection method ActionFormer [6] to simultaneously detect the action instances and classify the action categories.

2 Our Approach

2.1 Video Feature Extraction

To discriminatively represent the video snippets, we use the pretreined VideoSwin [4] and X-CLIP [5] models to extract the video features. For each video, we ex-

[*] Corresponding author. E-mail: cjzhang@bjtu.edu.cn
[†] Intern at the Institute of Automation, Chinese Academy of Sciences.

tract video features using clips of 16 consecutive frames and a stride of 16 frames. In particular, the features before the last fully-connected layer are extracted from both VideoSwin [4] and X-CLIP [5]. Note that, both VideoSwin and X-CLIP set up 4 × 3 views, where a video is uniformly sampled in the temporal dimension as 4 clips, and for each clip, the shorter spatial side is scaled to 224 pixels and 3 crops of size 224 × 224 are token that cover the longer spatial axis. As a result, each video snippet in a video is embedded into a 12 × 1024 dimensional vector through VideoSwin [4], and a 12 × 768 dimensional vector through X-CLIP [5]. Besides, we generate a mean feature from the 12 views, which could be regarded as the 13_{th} view feautre. To fuse these two kinds of features, we randomly select a view from the VideoSwin features and a view from the X-CLIP features, and then concatenate the features along the channel axis.

2.2 Temporal Action Detection

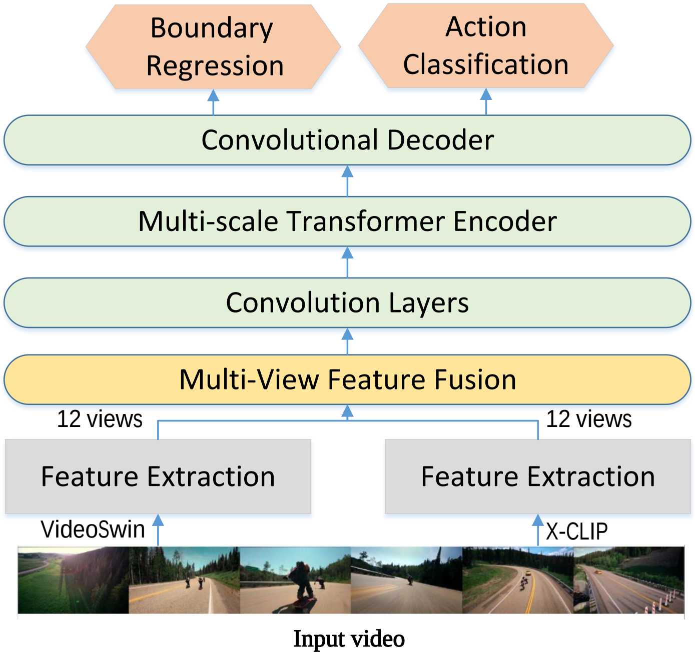

Fig. 1. The block-diagram of the proposed method. Given an input video, we first extract 12 views of features from VideoSwin and X-CLIP respectively, and fuse the video features by the multi-view feature fusion module. Then, convolution layers and multi-scale transformer encoder are exploited to further encode the video features. Finally, a convolutional decoder is used to generate candidate proposals and classify the action categories.

After obtaining the fused video features, we follow the work ActionFormer [6] to integrate the action boundary regression and action classification in a unified framework. As shown in Fig. 1, convolution layers are firstly deployed to embed the features, and then the embedded features are further encoded into a feature pyramid using a multi-scale Transformer. Next, the feature pyramid is inputted into shared regression and classification heads, which generate an action candidate at every time step. The training objectives are defined following the settings of ActionFormer [6] for THUMOS-14 dataset.The model can localize fine-grained actions in a single shot, without using pre-defined anchor windows or action proposals.

It should be noted that, in the inference stage, we fuse the features by concatenating the mean values of the 12 view features from VideoSwin and X-CLIP. After obtaining candidate proposals, we use soft-NMS to remove redundant proposals.

3 Experiments

In this section, we will first describe the experimental details and then illustrate the performance of our method.

3.1 Experimental Details

The video snippet features are extracted for every 16 consecutive frames with a stride of 16 frames. We concatenate the video snippet features in a long video along the temporal dimension, and we set the max sequence length as 2304. We use Adam with warm-up for training. The training batch size is set as 16.

3.2 Experimental Results

We generate 400 proposals for each video in the validation set of FineAction and report mAP at tIoU=0.5, 0.75, 0.95, and the average mAP in [0.5 : 0.05 : 0.95].

Feature	0.5	0.75	0.95	average
VideoSwin	35.46	20.22	3.43	21.00
X-CLIP	34.46	19.66	3.72	20.53
VideoSwin+X-CLIP	36.26	21.12	3.76	21.76
VideoSwin+X-CLIP (Multi-views)	**37.60**	**22.23**	**4.42**	**22.79**

Table 1. Performance comparison among different features on the validation set of FineAction.

Performance comparison with different features is shown in Table 1. The first row and second row show the results of singly using VideoSwin and X-CLIP features respectively. The third row shows the results of directly concatenating the mean features (mean value of 12 view features) of VideoSwin and X-CLIP. The forth row shows the results of selectively fusing the 13 view features as stated in Section 2.1. It can be seen that the the model trained on multi-view features from both VideoSwin and X-CLIP performs best compared to other models.

4 Acknowledgments

This work is supported by Beijing Natural Science Foundation: JQ20022; National Natural Science Foundation of China: 62072026. the open research fund of the State Key Laboratory for Management and Control of Complex Systems (20220103); the open project of Anhui Provincial Key Laboratory of Multimodal Cognitive Computation, Anhui University (MMC202103); We gratefully acknowledge the support of MindSpore, CANN(Compute Architecture for Neural Networks) and Ascend AI Processor used for this research.

References

1. Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 961–970 (2015)
2. Jiang, Y.G., Liu, J., Zamir, A.R., Toderici, G., Laptev, I., Shah, M., Sukthankar, R.: Thumos challenge: Action recognition with a large number of classes. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshop (2014)
3. Liu, Y., Wang, L., Ma, X., Wang, Y., Qiao, Y.: Fineaction: A fine-grained video dataset for temporal action localization. arXiv preprint arXiv:2105.11107 (2021)
4. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video swin transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3202–3211 (2022)
5. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. arXiv preprint arXiv:2208.02816 (2022)
6. Zhang, C., Wu, J., Li, Y.: Actionformer: Localizing moments of actions with transformers. arXiv preprint arXiv:2202.07925 (2022)