Technical Report

Yaokun Zhong , Xiaotong Lin
Sun Yat-sen University, Guangzhou, China
{zhongyk23, linxt29}@mail2.sysu.edu.cn

1. Method

Actionformer [1] , a one-stage anchor-free model for temporal action localization, combines multiscale features with local self-attention in the time dimension, and demonstrates strong performance on multiple related datasets. In this technical report, we apply Actionformer as the baseline on FineAction [2] to predict action categories and boundaries as shown Figure1. The specific steps can be summarized as follows:

Firstly, we filter out some dirty annotations in training set and validation set whose end time is less than the start time or greater than the total length of the video. Then we use the Slowfast [3] model pre-trained on Kinetics-400 [4] to extract video features, and the sampling interval is the same as the officially given I3D feature. Finally, these two features are concatenated and fed to the ActionFormer for temporal action localization.

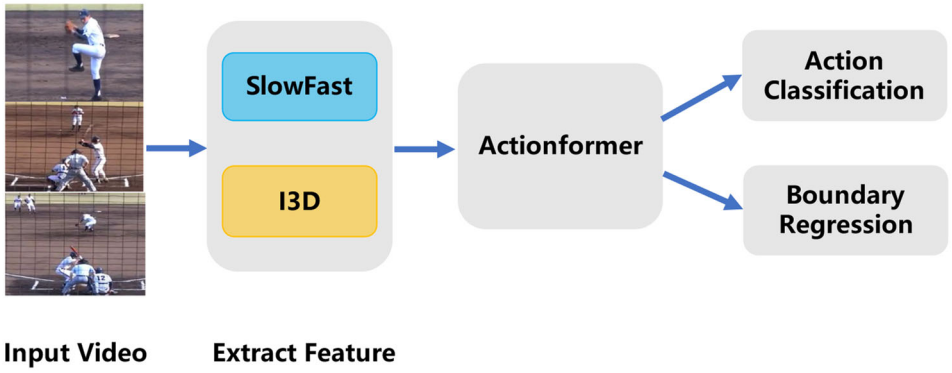

Figure 1. The pipeline of our method

2. Experiments

In this section, we present our experimental details and results on FineAction. After removing dirty annotations, we use 8384 untrimmed training videos and 4140 untrimmed validation videos in FineAction dataset to train and evaluate our model with mAP@[0.5:0.05:0.95].

2.1 Training details

Following the training strategy of the ActionFormer model, we use Adam [5] optimizer with a linear warm-up during the training stage, fixe the length of the maximum input video and randomly crop the original video with a ratio of 0.9 to 1. In addition, our model is trained for 20 epochs with the initial learning rate 1e-4, cosine learning rate decay, batch size 4, and weight decay of 0.05. The model takes roughly 20 hours to train over one NVIDIA GeForce GTX 1080 Ti GPU.

2.2 Inference

At inference time, all video sequences are fed to the model without any crop. For the generating action candidates in each time step, Soft-NMS [6] is used to remove highly overlapping instances of them. We finally select the top 100 instances as the final outputs for evaluation.

2.3 Results on FineAction

Video feature extraction networks play an important role for the temporal action localization task. As shown in the Table 1, Different features representation can provide complementary semantic information for Actionformer model. Compared with using the officially given I3D feature, the ensemble model that concatenates Slowfast feature and I3D feature in the channel dimension achieves 1.23% mAP improvement.

Feature	mAP
Slowfast	12.58%
I3D	15.88%
I3D + Slowfast	17.11%

Table 1. Different feature representation for Actionformer on the FineAction validation set

3. References

[1] Zhang C, Wu J, Li Y. Actionformer: Localizing moments of actions with transformers[J]. arXiv preprint arXiv:2202.07925, 2022.

[2] Liu Y, Wang L, Ma X, et al. Fineaction: A fine-grained video dataset for temporal action localization[J]. arXiv preprint arXiv:2105.11107, 2021.

[3] Feichtenhofer, Christoph, et al. "Slowfast networks for video recognition." *Proceedings of the IEEE/CVF international conference on computer vision*. 2019.

[4] Kay W, Carreira J, Simonyan K, et al. The kinetics human action video dataset[J]. arXiv preprint arXiv:1705.06950, 2017.

[5] Kingma D P, Ba J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014.

[6] Bodla N, Singh B, Chellappa R, et al. Soft-NMS--improving object detection with one line of code[C]//Proceedings of the IEEE international conference on computer vision. 2017: 5561-5569.