Technical Report of Multisports Track of Spatio-Temporal Action Detection

Keke Chen*, Zhewei Tu*, Shaomeng Wang, Xiangbo Shu
Nanjing University of Science and Technology
{kekechen,zwtu99,wangshaomeng,shuxb}@njust.edu.com

1. Approach

The task of Multisports Track of Spatio-Temporal Action Detection is introduced by MCG-NJU, which aims to find the frames that contain actions, and where these actions occur in an untrimmed video. The main challenge of spatio-temporal action detection is that the given is a long video without editing, and we need to submit not only the action detection results of each frame (spatial), but also the video segment corresponding to each action (temporal).To solve the above challenges,our framework is designed to detect all persons in an input video clip(3s in our experiments) and estimate their action labels. For action tube generation, we use the same link algorithm as MOC. Based on the analysis and understanding of the competition dataset, we kept the backbone feature extraction network unchanged, modified the classification header of the original model, and added a dual attention module to capture the global and local feature dependencies in the spatial and channel dimensions, respectively. As shown in Fig.1, we use a two-stage approach for action detection, the first stage requires generating actor proposals using an off-the-shelf person detector (e.g. Faster R-CNN), and the second stage is generating video features using the video backbone network (e.g. SlowFast) and extracting actor features based on proposal. Then actor features as local features and video features as global features are processed by the proposed dual attention module (DAM) for final action prediction.

In details, in the first stage, the person detector operators on the key frame and generated N proposals and we used the center frame of the video clip as the key frame. In the second stage, the backbone network extracts a spatio-temporal feature map $V \in \mathbb{R}^{C\times T\times H\times W}$ from the input video clip. We perform average pooling along the temporal dimension on the V,which results in a background feature map $B \in \mathbb{R}^{C\times H\times W}$. And we extract region-of-interest (RoI) features at the last feature map of res5. We first extend each 2D RoI at a frame into a 3D RoI by replicating it along the temporal axis. Subsequently, we compute actor feature $A \in \mathbb{R}^{C\times 7\times 7}$ by RoIAlign spatially, and global average pooling temporally, similar to the method presented in SlowFast. After generating actor features $A_1, A_2, \ldots, A_N$, we copy background features $B_1, B_2, \ldots, B_N$ according to the number of actor features so that they can be matched one-to-one. Each actor feature A_i along with the background feature B_i is viewed as a person-context pair and fed into dual attention model for feature enhancing. Then the final representation of a person is obtained by fusing the enhanced global and local features through DAM. Lastly, a linear classifier takes the person's representation as input and outputs action predictions.

The Dual Attention Module includes a position attention module and a channel attention module. We notice that the background feature has rich global semantic information, so we should pay more attention to spatial information. Actor features already focus on local information, so they should pay more attention to effective channels. The specific structure of DAM is shown in Fig.2.

2. Experiments

MultiSports latest version contains 18,422 training instances and 6,577 validation instances, selected from 1,574 and 555 clips respectively. The testing set includes 1071 clips. According to the competition specification, we evaluate on 60 action classes, and leverage frame-mAP and video-mAP to evaluate action localization performance.

2.1. Data Usage

We use the data provided by DeeperAction Challenge for training, validation, and testing. Additionally, we adopt the off-the-shelf person boxes shared by the Multi-Sports repo., which are generated by the person detector of Faster R-CNN with a ResNeXt-101-FPN backbone.

For training and validation, we modify the official dataset format and public person boxes format with reference to the AVA data format to adapt the SlowFast Det. For testing, we keep the test set format consistent with the training dataset.

[*] Equal contribution.

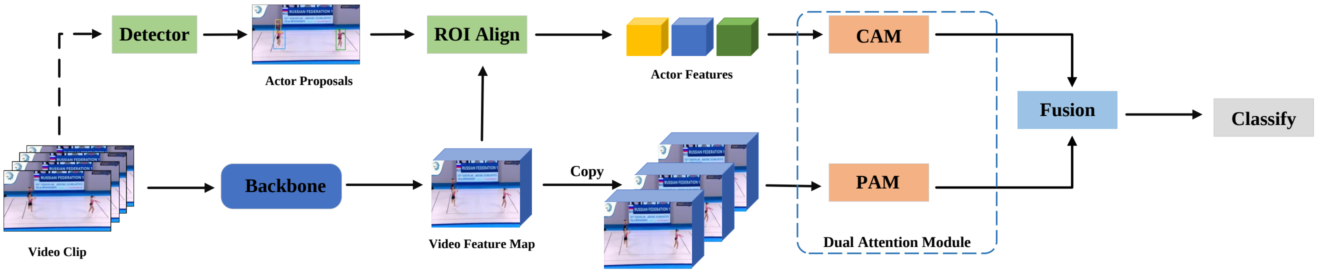

Figure 1. Overview of our model for spatio-temporal action detection.

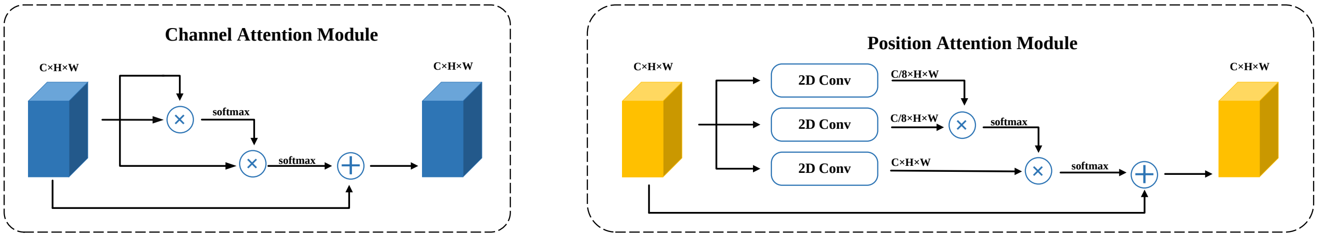

Figure 2. Dual Attention Module.

2.2. Implementation Details

Training: The backbone is the variant of SlowFast based on ResNet 50. We set the spatial stride of res5 to 1 and use a dilation of 2 for its filters. The $T \times \tau$ is set to 4×16 and the α is set to 8. The network weights are initialized from the Kinetics-400 classification models presented in PySlowFast. We use a step-wise learning rate, reducing the learning rate $10\times$ after epoch 4, 8 and 10. The initial learning rate of SGD optimizer is 0.01 and linear warm-up is adopted for the first 3 epochs. We train for 12 epochs on train set and extra 4 epochs on train/val set. We use a weight decay of $1e-4$. Note that only ground-truth boxes are used as the samples for training. For augmentation, we randomly sample a clip (of αT frames) from the videos; for the spatial domain, we randomly resize the clip with a shorter side sampled in $[256, 320]$ pixels and randomly crop 256×256 pixels from resized clip. The ratio of final horizontal flip sets to 0.5.

Inference: We kept the aspect ratio of clips with the short side set to 256. The detected boxes with confidence of ≥ 0.6 are adopted for action detection. We use the same link algorithm as MOC for action tube generation. We use the code in PySlowFast and results we submit are all produced by it.