Person-Context Cross Attention for Spatio-Temporal Action Detection

Zhiqing Ning[1*] Qiaokang Xie[2†*] Wengang Zhou[2] Liangwei Wang[1] Houqiang Li[2]
[1]Huawei Noah's Ark Lab [2]University of Science and Technology of China
{ningzhiqing3, wangliangwei}@huawei.com
xieqiaok@mail.ustc.edu.cn {zhwg, lihq}@ustc.edu.cn

Abstract

This technical report introduces our solution to MultiSports track on spatiotemporal action detection, DeeperAction Challenge at ICCV 2021. Our solution utilizes a cross attention mechanism to explicitly model relations between person and context for action detection. We describe solution details for the new MultiSports dataset, together with some experimental results. We finally achieve 48.68 frame mAP and 24.2 video mAP@0.1:0.9 on the test set of MultiSports and obtain the 1st place of the MultiSports track, which outperforms other entries by a large margin.

1. Introduction

Spatio-temporal action detection in untrimmed videos aims to detect and recognize human actions in space and time, which is of great significance and has attracted many efforts in recent years [3, 2, 7, 9, 14, 11, 8]. Current spatio-temporal action detection benchmarks can be mainly divided into two categories: 1) Densely annotated actions such as UCF101-24 [13] and J-HMDB [5]; 2) Sparsely annotated actions such as DALY [16] and AVA [3]. Sparsely annotated datasets fail to provide clear temporal action boundaries, which might be not enough to model the actions with rapid movement. UCF101-24 [13] and J-HMDB [5] provide densely annotations, however, their video clips typically have only one single person doing some semantically simple and temporally repeated actions, which means only one person in one video clip have one action class. In addition, due to coarse-grained action categories and characteristic visual scenes, it makes it much easier to get cues for predicting actions from scenes, which weakens the importance of fine-grained motion information of human actions.

MultiSports [8] provides a large-scale spatio-temporal action detection dataset for multiple people in sports, with frame-by-frame annotations of multi-person multi-class actions. Many methods that perform well on UCF101-24 [13] and J-HMDB [5] perform poorly on MultiSports [8], since it is much more challenging in: 1) Multiple people perform different and fine-grained actions concurrently in the same scene; 2) The backgrounds are far less characteristic and action recognition can not get much help from the background information; 3) There are more person-object-scene interactions in sports actions compared with atomic actions. Therefore, detecting these fine-grained actions requires complex spatio-temporal context modeling with human pose motions, long-term semantics, person-object-scene interactions, and reasoning [8].

Attention mechanisms have been demonstrated to be effective for modeling relations between person and context. In ACAR [11] network, each person feature is repeatedly concatenated to all spatial locations of the global context feature. Then the concatenated feature map is encoded by several convolutional layers to form actor-context [11] feature. Finally, self-attention layers are applied to learn high-order relation reasoning. However, human actions of interest (action categories of the dataset) account for a very small proportion of all possible human actions in the untrimmed video scene. The global context feature represents rich information of the input video while the feature of an interested action is down-sampled and transformed from the context feature. These two types of features are of capacity inequality to representation. We should utilize them more carefully. Moreover, concatenating the person feature to all spatial locations of the context feature may bring ambiguity for learning accurate spatial relations and it is computationally expensive. Based on the above observations, we introduce a cross attention transformer encoder for person-context relations modeling in videos. In our instantiation of cross attention mechanism, the person feature keeps attending to the context feature and gets relation information of different levels (direct or indirect) while the context feature preserves rich representations for various actions.

Inspired by previous works, we adopt SlowFast Det [2] as our baseline. Firstly, an off-the-shelf person detector is employed to generate person bounding boxes in videos. Then, We adopt SlowFast [2] as the video backbone to ex-

[*]Equal contribution. †Intern at Huawei Noah's Ark Lab.

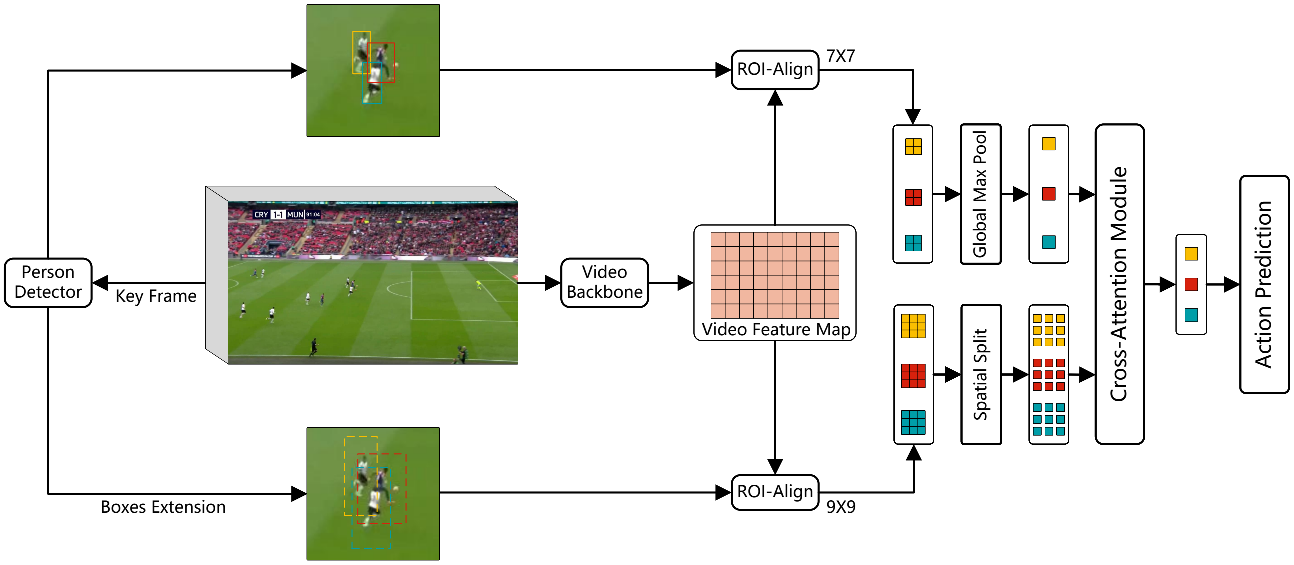

Figure 1. Framework of our approach for spatio-temporal action detection. The frame is cropped for more clear illustration.

tract visual features, and the feature maps of each person are obtained via ROI-Align [4]. Finally, the proposed cross attention transformer head is applied for person-context modeling and human action prediction. Furthermore, we have attempted to build a boundary prediction module to obtain boundary-aware tube features. We leave it as future work since we believe that clear temporal boundaries in MultiSports are significant information for improving the performance of spatio-temporal action detection, especially the video-mAP metric.

2. Method

In this section, we present our approach for spatio-temporal action detection on MultiSports. Firstly, we introduce the overall framework of our method for this task. Then a cross attention module based on transformer is present for modeling relations between person and its spatio-temporal context. Finally, we discuss the learning strategy for the long-tailed category distribution in MultiSports dataset and the approach for model ensemble.

2.1. Over Framework

Based on some previous works[2, 11, 17] for spatio-temporal action detection task, we design the whole pipeline as shown in Figure 1. The framework is designed to detect all persons in an input video clip and predict their action classes. Specifically, a video is firstly sampled with a specified frame interval into an input video clip. The center frame of the clip is extracted and fed into a 2D detector to generate bounding boxes of people. In the meantime, a video backbone network extracts spatio-temporal features from the video clip. We perform average pooling along the temporal dimension on the video feature, which results in a feature map $V \in \mathbb{R}^{C\times H\times W}$, where C, H, and W denotes channel height and width respectively. Then N person features are extracted by applying 7×7 RoI-Align on feature map V and further pooled by 7×7 max pooling, producing N person features, $A_1, A_2, ..., A_N \in \mathbb{R}^C$. Each pooled person feature along with the global feature V is viewed as a person-context pair and fed into cross attention transformer encoder for relation modeling. The transformer encoder outputs the final representation of a person. Lastly, a linear classifier takes the person's representation as input and outputs action predictions.

2.2. Person-Context Cross-Attention Modeling

Person-context features are firstly transferred into sequential tokens as the input of transformer encoder. In the first layer of cross attention transformer, the query input is a person feature and the key/value input is the person's context feature. The scaled dot-product operation outputs an attention scores matrix and the projected context feature is multiplied by the matrix. The multiplied feature serves as the inherent dependency for person-context relations and is further added to the person feature through a shortcut connection. This enhanced feature is further taken as query input in the subsequent transformer layers, and key/value input keeps the same as the first layer, which indicates that the context feature will not be transformed along with the person feature layer-by-layer. We argue that indirect relations can be retrieved by person-context interactions of multiple layers.

For computational efficiency consideration and the observation that the fine-grained behavior of a person usually

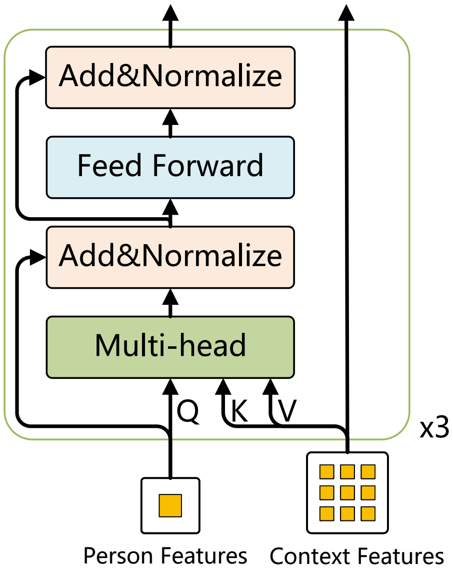

Figure 2. Person-Context Cross-Attention.

relates to surroundings near the person, we use an extended box and RoI-Align to extract spatio-temporal context of the person. Specifically, we increase the person box scale by a factor of 2 times. The extended box has the same center as the original but is larger in height and width. We use the extended boxes to perform $h \times w$ RoI-Align on feature map V, which results in N features $C_1, C_2, ..., C_N \in \mathbb{R}^{C \times h \times w}$ describing spatio-temporal context of persons. In our implementation, $h \times w$ is set to 9×9.

In the first transformer layer, an person feature $A_i \in \mathbb{R}^C$ is taken as query input tokens and the corresponding context feature $C_i \in \mathbb{R}^{C \times h \times w}$ is flatten and taken as key/value input tokens of $h \times w$ length. Formally, cross-attention transformer for person A_i and context C_i computes feed-forwardly for $j = 1, ..., D$ layers,

$$Q^0 = A_i, \tag{1}$$

$$X^j = LN(MultiHead(Q^{j-1}, C_i, C_i) + Q^{j-1}), \tag{2}$$

$$Q^j = LN(FFN(X^j) + X^j), \tag{3}$$

We omit parameters notations for simplicity. LN denotes layer normalization[1]. FFN is feed-forward sublayer and MultiHead[15] means multi-head scaled dot-product attention. In our implementation, the transformer head is of 3 cross attention layers with 8 heads in each layer, and the projection dimension is 1024.

2.3. Long-tailed Learning

The number of instances in each action category ranges from 3 to 3,514, which reflects obvious long-tailed category distribution. The classes with fewer instances pose great challenges for deep learning based models on how to deal with the class imbalance problem.

We consider the decoupling representation learning strategy [6] to obtain the model that is capable of recognizing all classes well. Specifically, the training process is divided into two phases. In the first phase, we follow the normal training paradigm with standard randomly sampled data. In the second phase, we freeze all parameters of the model except the final classifier and retrain the classifier with class-balanced sampling. For class-balanced sampling, each class has an equal probability of being selected. Such a strategy helps to further improve performance, especially on some classes with a small number of samples.

3. Experiment

MultiSports v1.0 contains 18,422 training instances from 1,574 videos and 6,577 validation instances from 555 videos. And there are 1,071 videos in the test set. Following the guidelines of the challenge, we evaluate on 60 action classes, and the performance metrics are frame-mAP and video-mAP. For frame-mAP, the IoU threshold is 0.5. For video-mAP, the 3D IoU threshold is 0.5 for 0.2 and 0.5 and 0.1:0.9.

3.1. Implementation Details

Person Detector. We apply Faster R-CNN [12] framework with ResNeXt-101-FPN backbone from maskrcnn-benchmark[10] for person detection. It is firstly pre-trained on ImageNet and the COCO Person keypoint images. We further fine-tuned the model on train set of MultiSports for higher detection precision.

Backbone. A SlowFast-R101 backbone network with $T \times \tau = 8 \times 8$ and $\alpha = 4$ is used to extract video features. It is firstly pre-trained on Kinetics-600 dataset. The spatial stride in stage res5 is set to 1 and a dilation of 2 is used for the stage's filters. This increases the spatial resolution of res5 by 2 times.

Training. We use per-class binary cross entropy loss as the training loss function. We train models with two spatial scales 256×455, 360×640 respectively in an end-to-end fashion using SGD with a mini-batch size 32 for 256×455 scale and mini-batch size 24 for 360×640 scale. The initial learning rate of SGD optimizer is 0.1. We also use weight decay 1e-4 and Nesterov momentum of 0.9. Linear warm-up is adopted during the first 3 epochs. We decrease the learning rate by a factor of 10 at epoch 5, 8, and 10. Model optimization process stops at $12th$ epoch for training only on train set and stops at $15th$ epoch for train/val set.

Inference. We use person detections with confidence $\geq$ 0.6. For action tube generation, we use the same link algorithm as MOC [9].

head	testing scales	decoupled training	detector*	val set				test set	
				F@0.5	V@0.2	V@0.5	V@0.1:0.9	F@0.5	V@0.1:0.9
Linear	256 × 455	×	det-1	29.03	28.06	8.39	12.26	-	-
PCCA	256 × 455	×	det-1	39.48	38.01	17.82	19.02	-	-
PCCA	256 × 455	✓	det-1	42.21	41.00	19.95	20.89	-	20.70
PCCA	360 × 640	×	det-1	41.60	41.14	19.15	20.56	-	-
PCCA	ensemble	✓	det-3	-	-	-	-	48.68	24.2

Table 1. Main results on MultiSports. F@0.5 denotes frame mAP@0.5. V@0.2 and V@0.5 denote video mAP@0.2 and video mAP@0.5 respecitively. And V@0.1:0.9 is the average of V@0.1 to V@0.9 with 0.1 gap. PCCA refers to Person-Context Cross-Attention. "*" indicates different detectors in Table 3.2

detector	AP@0.5	AR@100	F@0.5	V@0.1:0.9
official [8]	-	**96.13**	42.05	20.88
det-1	78.00	94.36	39.48	19.02
det-2	83.16	94.68	41.60	20.56
det-3	**86.53**	93.83	**43.24**	**22.40**

Table 2. Influence of different person detectors on MultiSports validation set. AP@0.5 denotes average precision of person detection with IoU threshold 0.5, and AR@100 average recall with top 100 detections each frame. F@0.5 and V@0.1:0.9 reported by testing on a same action model with different person boxes. The action model is of SlowFast R101 8×8 with PCCA head and scale 256×455.

3.2. Main Results

Table 3.2 shows our main results on MultiSports. The default backbone instantiation is SlowFast R101 8x8. The baseline, linear classifier head, only gives 29.03 frame mAP and 12.26 V@0.1:09. Switching to our Person-Context Cross-Attention (PCCA) head brings significant boosts on frame mAP and video mAP. Using the same spatial scare 256×455, PCCA head gives a total boost of 10.45 frame mAP and a boost of 6.76 video AP@0.1:0.9. This highlights the effectiveness of modeling person-context relations using cross attention mechanism. For final submission, models are trained on both training and validation data, and tested with 2 scales and horizontal flips. We ensemble these models' results by average voting and get 48.68 frame mAP and 24.2 video AP@0.1:0.9 reported from the test server.

3.3. Ablation Studies

Different Scales. We investigate the effect of different scales. Two types of scales 256×455, 360×640 are applied for training and testing on SlowFast R101 8x8 with PCCA head. Table 3.2 shows that increasing spatial resolution from 256×455 to 360×640 can bring extra improvement in performance (+mAP 2.12, +V@0.1:0.9 1.54).

Different Detectors. We also investigate the influence of person detection. We compare person bounding boxes detected by our detector with the boxes provided by MultiSports repo [8]. Table 3.2 shows the results on validation set. We select 3 models to detect persons. These models differ in training hyper-parameters (e.g. learning rate, batch size, stopping iterations). Note that we evaluate only on frames having action tube annotations (a.k.a. frames of trimmed videos) to get results of AP@0.5 while we report F@0.5 and V@0.1:0.9 on untrimmed videos. These settings make the results of AP@0.5 more sensitive to detector's performance. Because most frames of untrimmed videos are not annotated with boxes and detections of these frames would be treated as false positives. The results suggest that we should select person detections with higher AP when AR is already high. Higher AR detections may recall more action instances when the action classifier is strong enough to filter more false proposals.

classes	diff. / F@0.5
top-20	+1.76
bottom-20	+3.09
all	+2.73

Table 3. Influence of decoupled training on MultiSports validation set. Classes are ranked by their numbers of labeled samples, and the averaged differences of top-20, bottom-20, and all classes are listed.

Influence of decoupled learning. We compare the performance of all classes before and after decoupled learning. Based on the number of instances, we sorted the action categories in descending order and counted the difference in frame mAP for top-20 and bottom-20 classes. As shown in Table 3.3, The performance of all classes is improved by 2.73 in frame mAP, while a larger improvement is seen on the bottom-20 classes. The results demonstrate that decoupled learning is effective for further improving the whole performance, especially on some of the small classes.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019.

[3] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6047–6056, 2018.

[4] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2961–2969, 2017.

[5] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *IEEE International Conference on Computer Vision (ICCV)*, pages 3192–3199, 2013.

[6] Bingyi Kang, Saining Xie, Marcus Rohrbach, Zhicheng Yan, Albert Gordo, Jiashi Feng, and Yannis Kalantidis. Decoupling representation and classifier for long-tailed recognition. *arXiv preprint arXiv:1910.09217*, 2019.

[7] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. You only watch once: A unified cnn architecture for real-time spatiotemporal action localization. *arXiv preprint arXiv:1911.06644*, 2019.

[8] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. *arXiv preprint arXiv:2105.07404*, 2021.

[9] Yixuan Li, Zixu Wang, Limin Wang, and Gangshan Wu. Actions as moving points. In *European Conference on Computer Vision (ECCV)*, pages 68–84. Springer, 2020.

[10] Francisco Massa and Ross Girshick. maskrcnn-benchmark: Fast, modular reference implementation of instance segmentation and object detection algorithms in pytorch. *Google Scholar*, 2018.

[11] Junting Pan, Siyu Chen, Mike Zheng Shou, Yu Liu, Jing Shao, and Hongsheng Li. Actor-context-actor relation network for spatio-temporal action localization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 464–474, 2021.

[12] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems (NIPS)*, 28:91–99, 2015.

[13] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[14] Jiajun Tang, Jin Xia, Xinzhi Mu, Bo Pang, and Cewu Lu. Asynchronous interaction aggregation for action detection. In *European Conference on Computer Vision (ECCV)*, pages 71–87. Springer, 2020.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NIPS)*, pages 5998–6008, 2017.

[16] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. Towards weakly-supervised action localization. *arXiv preprint arXiv:1605.05197*, 2:1, 2016.

[17] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Krahenbuhl, and Ross Girshick. Long-term feature banks for detailed video understanding. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 284–293, 2019.