Context Feature for Action Localization

Xiantan Zhu+, Zejing Han+, Yuya Obinata*,
Takuma Yamamoto*, Kouichirou Yamashita+, Zhiming Tan+

+Fujitsu Research & Development Center Co., LTD
*FUJITSU RESEARCH

ABSTRACT

This technical report shows our solution for AVA challenge 2021. We seek to fuse context feature extracted from the video sequences into other features, like person feature and object feature. Besides, for some classes, such as drink and smoke, relying on local regions, we revisit RCNN-like method for action detection where actor regions are cropped from original sequences and resized to a fixed resolution. Fusing the results of the two models, we achieve 37.43 mAP and get the 2nd place on AVA-Kinetics Crossover challenge 2021.

1. Our method

As Alphaction [1] method has achieved the state of the art performance on the AVA dataset, we select it as our baseline model. There is an interaction aggregation structure to model multiple types of interaction along person feature (P), object feature (O) and memory feature (M). Interaction operation is Non-Local Network [2] operation, which is able to enhance the classification feature.

Although the features of Alphaction are rich, the context feature (C) of the whole sequence is omitted by the author. Just recognizing human action from crops of person is challenging even for human. Therefore, context feature can play an important role for learning robust action detection models. For example, for someone swimming, if we just use the person feature and object feature for action classification, the original model will output the pose as stand but not swimming. With the context feature like water, the model can correctly output the pose of the person.

We fuse the context feature into the Alphaction model like the pipeline show in Figure 1. The context feature (C) is pooled from the output of 3D CNN model. Then multiple features including P, O, C, and M will flow through POMC interaction aggregation to enhance the person feature. From massive experiments, PCMPOM sequence can get a satisfactory result on AVA dataset. Besides, we find that the object feature is also enhanced by the context feature, where the context feature is simply added to every object feature. Finally, we adopt PCMP(O+C)M for the interaction aggregation.

Moreover, we learn from CRCNN [3] that the recognition accuracy is highly correlated with the bounding box size of an actor, and higher resolution of actors contributes to better performance. Based on this idea, we train another Alphaction model which the person feature is extracted via cropping and resizing image patches around the actors before feature extraction with 3D CNN. From our experiments, we conclude that some classes like drink and smoke are improved a lot.

Finally, the results of two models are fused with mean method, which averages the same action score of two model for each person.

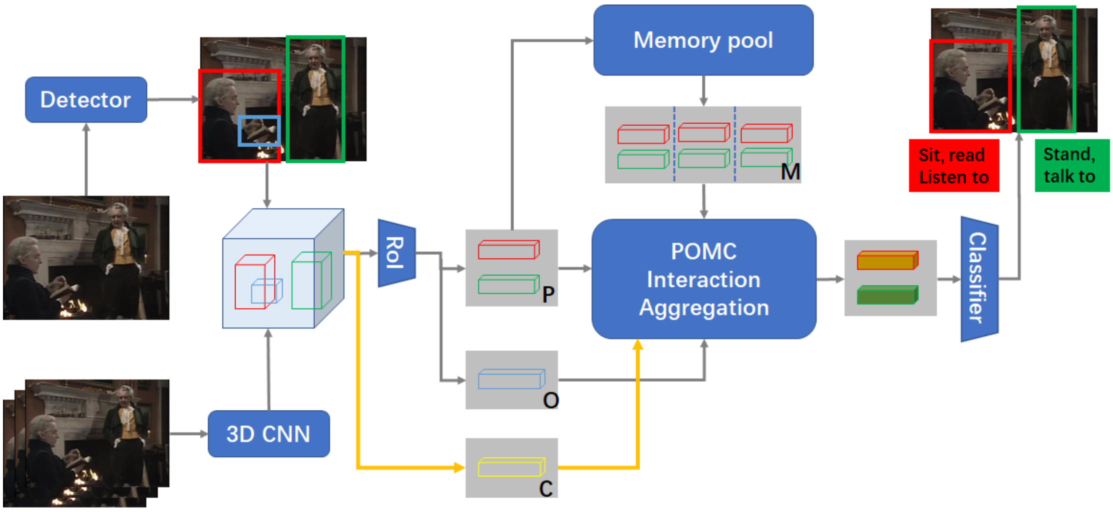

Figure 1. Pipeline of the proposed framework.

2. Experiments

2.1 Detector

In this paper, two detectors are used. One is person detector and the other is object detector. For person detector, we use Faster R-CNN [4] with a ResNeXt-101-FPN backbone, which is the same as the one mentioned in the LFB [5]. For object detector, we use Scaled-YOLOv4 [6], which achieves 55.8% AP on the MS COCO dataset.

Because the videos of AVA are cut from films and the persons are different from Kinetics in size, occlusion, and scale, we train separated person detector models for each of them. Thresholds of 0.8 and 0.65 are used respectively for them.

2.2 Ablation Experiments

2.2.1 Context feature

In order to quickly find the best way to aggregate C into POM, we train and test on our mini_AVA dataset, which contains 1/3 of the whole AVA dataset. We don’t simply select 100 videos of 299 videos as our mini-AVA, and we select 1/3 keyframes of each videos. So, most action of each video is included and the mini dataset is reasonable. Next, we show our experiments about how to interact with C in the Table 1. All models in the table are trained and validated on the mini-AVA dataset.

We conclude from the table that the context feature can not only speed up the convergence, but also get 1.1% gains than the original POMPOM structure.

Table 1. Experiments about how to interact with context feature (C)

Model	mAP	Description
POMPOM	0.3071	The original POM of Alphaction
PCMPCM	0.3065	It can speed up the convergence.
POMPCM	0.3072	Replace the second O with C
PCMPOM	**0.3182**	Replace the first O with C

Context feature is fused to object feature with addition operation. It can also boost the performance on two datasets. From the Table 2, we can see the performance is improved with enhanced object feature. The models are trained with AVA-Kinetics dataset.

Table 2. Experiments about enhanced object feature.

Model	mAP	
	AVA	Kinetics
PCMPOM	33.99	33.95
PCMP(O+C)M	**34.20**	**34.69**

2.2.2 Crop person feature like R-CNN

For an actor bounding box at key frame, we replicate the box along the temporal axis. Then we crop the replicated box at each frame of the sequence and resize the image patches to a fixed resolution. Then the actor clip is fed to the backbone, followed by a global average pooling, resulting in the person feature. This operation of extracting the person feature like R-CNN in the object detection.

We replace the person feature in Figure 1 from the original to the above method and retrain Alphaction model. We compare the AP of some action of two models on AVA validation dataset in Table 3. Here, the Diff means the difference between the AP of two models. As shown in the table, some classes relying on the local regions are improved a lot, such as the AP of drink is improved by 0.1253. In the Table 4, the fusion result of two models on AVA validation dataset is shown. We can see from the Table 4 that the fusion result gets about 2% gains than the better result of the two models.

Table 3. Experiments about different person feature

Class	PCMP(O+C)M	POMPOM_RCNN	Diff
drink	0.3256	0.4509	**0.1253**
smoke	0.3386	0.4244	0.0858
write	0.1490	0.2142	0.0652

Table 4. Experiments about the fusion result of two models

Model	mAP
PCMP(O+C)M	34.20
POMPOM_RCNN	32.66
Fusion result	**36.25**

2.2.3 Final submission

For the testing submission, we train the models on the combined datasets of training and validation. We perform inference with multi-crop including 224, 256, and 320, and flipping operations.

With an ensemble of models, we achieve 37.43 mAP on the test set, ranking second in the AVA-Kinetics task.

3. Acknowledgements

Thanks very much to the AVA team for creating and sharing their dataset. Thanks a lot to Alphaction project for sharing their code and models.

References

[1] Jiajun Tang, Jin Xia, XinZhi Mu, Bo Pang, Cewu Lu. Asynchronous Interaction Aggregation for Action Detection. In ECCV 2020.
[2] Xiaolong Wang, Ross Girshick, Abhinav Gupta, Kaiming He. Non-local Neural Networks. In CVPR 2018.
[3] Jianchao Wu, Zhanghui Kuang, Limin Wang, Wayne Zhang, Gangshan Wu. Context-Aware RCNN: A Baseline for Action Detection in Videos. In ECCV 2020.
[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun.Faster R-CNN: Towards real-time object detection with region proposal networks. In Advances in Neural Information Processing Systems (NIPS), 2015.
[5] Chao-Yuan Wu, Christoph Feichtenhofer, Haoqi Fan, Kaiming He, Philipp Kr¨ahenb¨uhl, and Ross Girshick. Long-Term Feature Banks for Detailed Video Understanding. In CVPR 2019.
[6] Chien-Yao Wang, Alexey Bochkovskiy, Hong-Yuan Mark Liao. Scaled-YOLOv4: Scaling Cross Stage Partial Network. arXiv:2011.08036 .