Multi-scale Spatiotemporal Features for Action Localization

Xiantan Zhu+, Xuan Tao+, Lu Shi+, Shaoqi Chen+, Rui Yin+, Lan Ding+
Yuya Obinata*, Takuma Yamamoto*, Zhiming Tan+

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

ABSTRACT

This technical report shows our solution for AVA challenge 2020. In order to comprehensively describe the feature of persons in the keyframe, we apply multi-scale spatiotemporal features: short-term, long-term, global pooling, and arm feature. Two different models are used to extract 3D convolutional feature from two different datasets. For AVA, Slowfast [1] and LFB [2] are combined to extract multi-scale features. For Kinetics, we just use Slowfast to extract the short-term and global pooling feature. Then, we merge the result of arm related action recognition into the whole rest ones. Finally, we obtain 31.88 mAP on the AVA-Kinetics crossover test set.

1. Our method

Our solution is inspired by [2] and [3]. More features are extracted, better performance will be got. Multi-scale spatiotemporal features include short-term, long-term, global pooling and arm feature. Here the short-term feature implies spatiotemporal information within 2 seconds. The long-term feature refers to longer spatiotemporal information within 100 seconds. The global pooling feature of a clip can provide environmental information around the person. We make a statistical analysis about AP (Average Precision) of 60 classes and then find that most AP lower than 0.1 are actions related to arm, such as “cut” and “hit”, etc. So we extract feature in the arm area as the local part feature to improve final result.

In this year, AVA-Kinetics crossover dataset [4] consists of AVA-v2.2, where one video has nearly 900 keyframes, and Kinetics700, where one video just has one keyframe. Due to the difference between two datasets, two different models are used for action localization. For the AVA, Slowfast [1] and LFB [2] are combined as one

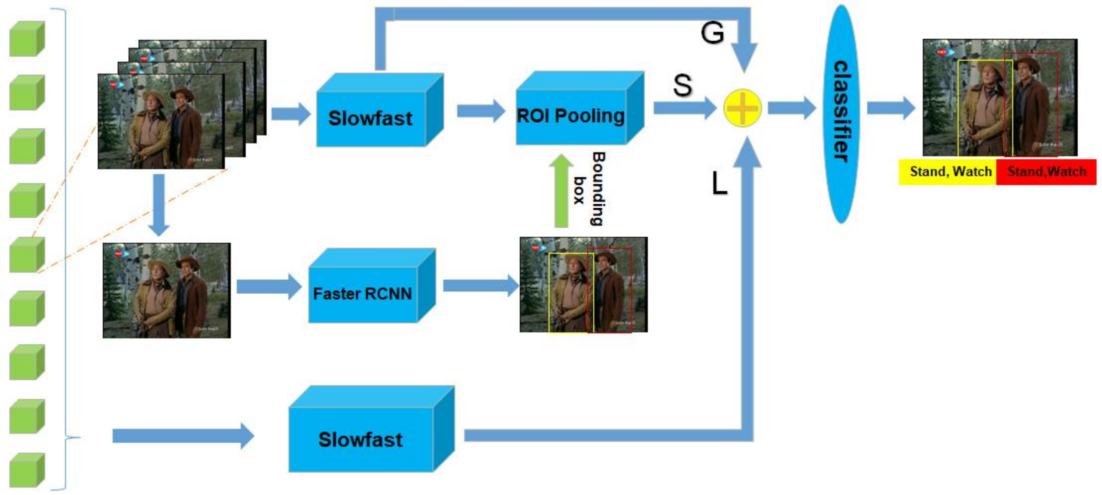

Figure 1. Pipeline of the proposed framework. S, short-term feature; G, global pooling feature; L, long-term feature bank.

neural network. Then we extract multi-scale features of short-term, long-term and global pooling with the combined network. For the Kinetic, we just use the Slowfast to extract the short-term and the global pooling feature. Besides, a model is trained to detect some actions related to arm. Finally, we merge the arm action result with the result of AVA-Kinetics crossover.

Figure 1 shows our pipeline of the action localization framework for the AVA dataset. The input cubes are video clips with length of 2 seconds. On the top branch, we extract the short-term feature on the bounding box from Faster R-CNN [5] detector (the middle branch) and extract the global pooling feature before RoI Pooling. On the bottom branch, we extract the long-term feature bank within a window size of 100 seconds centering on the keyframe. Different from the AVA, we just use the top branch and the middle branch to get the short-term and the global pooling feature for the Kinetics.

For the arm action detection, the arm area is used as the proposal for RoI pooling. First, CPN (Cascaded Pyramid Network) [6] is used to get the keypoints of person. Then, the key points of the elbow and wrist are used to judge whether they are inside the image. If one arm area meets the requirement, the body box will be replaced by the moderately enlarged arm box. Basing on it, we refine out an “AVA arm action dataset” with 27 arm action classes, such as “point to” and “throw”, etc. After training the arm action model, we select the results of 15 classes significantly improved for the fusion. For instance, the action classes “hit (an object)” and “cut” whose mAP improved more than 70% and 108% respectively are selected.

2. Experiments

2.1 Person Detector

The Faster R-CNN [5] with a ResNeXt-101-FPN backbone is used for our person detection. The model is the same as the one mentioned in the LFB. 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 0.7 are used respectively for them too.

2.2 Training

Our models are trained with minibatch size of 64 clips on a server with 8 GPUs. For the testing submission, we train the models on the combined datasets of training and validation. We train all models with 26 epochs, and a learning rate of 0.01, which is decreased by a factor of 10 at epoch 15 and 20. For data augmentation, we perform random flipping, random scaling with the short side ranging from 256 to 320 pixels, and random cropping of size 224x224.

2.3 Inference

We perform inference with multi-crop including 224, 256, and 320, and flipping operations. These 6 (3*2) results are then combined with summation.

2.4 Main Results

2.4.1 Validation results

First, the global pooling feature is fused to original LFB model. The dimension of the feature for classifier increases from 2560 to 2816. We get 27.73% mAP on the validation set while the result of original LFB is 26.9%. It means the global pooling feature is useful for the action classification.

Second, the Slowfast with the LFB are combined as one model. Then we train it on the AVA, and get 29.88% mAP while the result of original Slowfast is 29.17%.

Last, we merge our arm action detection result to one of our final result. We get 0.243% increment compared with our final result.

2.4.2 Test results

Table 1. Comparison between the baseline of AVA-Kinetics and ours (mAP%)

	AVA/test	Kinetics/test	AVA-Kinetics/test
Baseline	21.23	17.61	19.76
Ours	31.08	28.34	**31.88**

Table 1 shows our final submission result. Our result is compared with the result of the baseline of AVA-Kinetics dataset. Baseline indicates the result of a video action transformer network [7] on the AVA-Kinetics crossover dataset.

From Table 1, we can see our result is better than the baseline in three test sets. We get 12.12% mAP gain than the result of the baseline of AVA-Kinetics on crossover test set.

3. Acknowledgements

Thanks a lot to the AVA team for creating and sharing their dataset. Thanks much to Slowfast and LFB projects for sharing their code and models.

References

[1] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In CVPR 2019.
[2] 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.
[3] Jin Xia, Jia-Jun Tang, Ce-Wu Lu, Three Branches: Detecting Actions With Richer Feature. arXiv: 1908.04519, 2019.
[4] Ang Li, Meghana Thotakuri, David A. Ross, Jo˜ao Carreira1 Alexander Vostrikov, Andrew Zisserman. The AVA-Kinetics Localized Human Actions Video Dataset. In CVPR 2020.
[5] 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.
[6] Yilun Chen, Zhicheng Wang, Yuxiang Peng. Cascaded Pyramid Network for Multi-Person Pose Estimation. In CVPR, 2017.
[7] Rohit Girdhar, J.Carreira, Carl Doersch, Andre Zisserman Video Action Transformer Network. In CVPR 2019.