# DeepblueAI: Challenge report for MMAct Challenge

Zhiguang Zhang, Jianye He, Zhipeng Luo
DeepBlue Technology （Shanghai） Co.,Ltd
369 Weining Road, Changning District, Shanghai, China
{zhangzhg, hejianye, luozp}@deepblueai.com

## Abstract

*This technical report presents an overview of our methods for the MMAct Challenge on Activity Recognition (ActivityNet) Workshop at CVPR'21. The challenge asks participants to propose cross-modal video action recognition/localization approaches for addressing shortcomings in visual-only approaches using the MMAct dataset. This report studied previous methods and proposed our methods for Task 1: Cross-Modal Trimmed Action Recognition and Task 2. Cross-Modal Untrimmed Action Temporal Localization.*

## 1. Task 1: Cross-Modal Trimmed Action Recognition

In this task, participants will use trimmed cross-view videos and trimmed cross-scene videos from the MMAct[1] dataset along with paired sensor data; both have 35 action classes from 20 subjects. This task allows the participants to train with trimmed sensor data and trimmed video respectively, but test on only trimmed video for action recognition. And use the mean Average Precision (mAP) for Top-1 as the metric. The evaluation is done across the MMAct trimmed cross-view dataset and MMAct trimmed cross-scene dataset.

### 1.1. Method

Our method is based on TSM[2] and ir-CSN[3] for action recognition. For only video is provided in the test dataset, so we use only the vision-based modality in the training phase.

After analyzing the data set, we found that the duration range of video is vast, the frame number of videos between 37-1262.

For TSM, we use ResNet50 as backbone. Because of the long video range, we increase the sample frame number from 16 to 32 for each video and sample the frames at equal interval. To process the frames, we first resize the short side to 256, then use random scale form [1, 0.875, 0.75, 0.66] to crop different size image and resize it to 224x224 and Random Horizontal Flip as data augmentation.

For ir-CSN, we use ResNet152 as backbone. Sample 32 frames by sliding window method and set the frame interval to 4. Then use the same data augmentation method as TSM.

### 1.2. Training and Testing

Our experiments use MMAction2[4] as codebase, an open-source toolbox for video understanding based on PyTorch.

We train the model with 8 NVIDIA TESLA V100 GPUs (4 videos per GPU for TSM and one video for ir-CSN) for 42 epoch with an initial learning rate of 0.005 and decrease it by 0.1 after 24 and 36 epoch, respectively.

For testing, when inference with TSM, we use the same sample method as training and resize the short side to 256, then crop the four corners and the center part of the image with the same given crop size 224 and flip it horizontally. Then average the results of the ten inputs. When inference with ir-CSN, we sample ten clips, and each clip are sampled using the same method as training. We crop the images equally into three crops with equal intervals along the shorter side for each clip. Ten average the results of the thirty inputs.

### 1.3. Results

We also perform ablation study on different backbones, components(Non-local[5]), data augmentation(Mixup) and so on.

Table 1. Ablation study of different method. Results are reported on the test set.

| Method | Backbone | Sample Number | Non -local | MixUp | mAP |
|---|---|---|---|---|---|
| TSM | ResNet50 | 16 | | | 0.9229 |
| TSM | ResNet50 | 16 | ✓ | | 0.9136 |
| TSM | ResNet50 | 16 | | ✓ | 0.9173 |
| TSM | ResNet50 | 32 | | | 0.9500 |
| TSM | ResNet101 | 32 | | | 0.9458 |
| Ir-CSN | ResNeT152 | 32 | | | 0.9465 |

As pointed in Table 1, we can see that increase sample frame number is helpful to improve the performance of models. Finally, by average the results of TSM with ResNet50 as backbone and trained with 32 frames per video, and the result of ir-CSN we got our best result with score 0.9583.

## 2. Task 2. Cross-Modal Untrimmed Action Temporal Localization

Participants will use untrimmed paired sensor data and video for training in this task, which has 35 action classes from 20 subjects with four camera views and four scenes. Then test on only untrimmed videos for temporal action localization with the output being the recognized action class and its start and end time in the untrimmed video. And use the Interpolated Average Precision (AP) as the evaluation metric.

### 2.1. Method

Our method is based on AFSD[6], an anchor-free framework for temporal action detection tasks and an end-to-end method using frames as input rather than features. Only video is provided in the test dataset, so we use only the vision-based modality in the training phase.

After analyzing the dataset, we found that the video can be very long; some videos have more than 10 minutes. In our method, we increase the sample frame number from 768 to 2304 (due to the GPU memory limit) and sample the frames at equal interval form each video, and resized the frame to 112x112. Then we extract flow data from the sampled frames.

For data augmentation, we crop images with size 96x96 randomly and random horizontal flip them.

### 2.2. Training and Testing

In our experiments, we use the official code as codebase.

We train the model with 1 NVIDIA TESLA V100 GPUs with batch size 1, 24 epoch with an initial learning rate of 1e-4, and decrease it by 0.1 after 12 and 18 epoch, respectively. And use the same setting to conducted on the RGB model and the Flow model.

We use center crop with size 96x96 to get input for the model and get results for testing.

### 2.3. Results

On the RGB model, we get an mAP score of 0.3945 and 0.3754 on the Flow model. After ensemble the RGB model and Flow model results, we got our best result with mAP 0.4457.

## 3. Conclusions

In the MMAct Challenge on Activity Recognition (ActivityNet) Workshop at CVPR'21, we proposed some methods for the action recognition/localization tasks and achieved relatively good grades. But for the time limit, we just made some rudimentary attempts, which need more work. For example, in both tasks, we use the crop method to get the model's input; because some action class are happening on the side of the video, the crop method may drop out the valuable information. In ASFD, the max sample frames are limited by the GPU memory; some frame free methods need to be explored. The most important, that there only video was provided in the test set; we do all experiments use only the vision-based modality, which also needs more to do.

## References

[1] Kong Q, Wu Z, Deng Z, et al. MMAct: A Large-Scale Dataset for Cross Modal Human Action Understanding[C]// International Conference on Computer Vision. 0.

[2] Lin J, Gan C, Han S. TSM: Temporal Shift Module for Efficient Video Understanding[C]// 2019 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE, 2019.

[3] Du T, Wang H, Feiszli M, et al. Video Classification With Channel-Separated Convolutional Networks[C]// International Conference on Computer Vision. 0.

[4] MMAction2 Contributors. MMDetection: Open MMLab Detection Toolbox and Benchmark[J]. 2020.

[5] Wang X, Girshick R, Gupta A, et al. Non-local Neural Networks[J]. 2017.

[6] Lin C, Xu C, Luo D, et al. Learning Salient Boundary Feature for Anchor-free Temporal Action Localization[J]. 2021.