SUSTech&HKU Submission to TinyAction Challenge 2021

Teng Wang[12], Tiantian Geng[3], Jinbao Wang[1], Feng Zheng[1]

[1] Southern University of Science and Technology, [2] The University of Hong Kong,

[3] University of Electronic Science and Technology of China

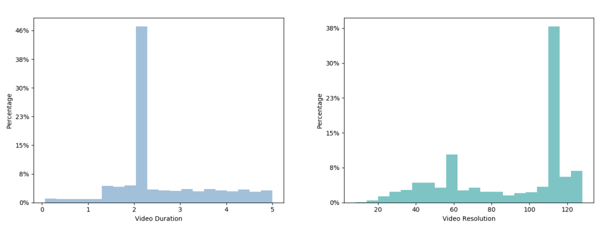
wangt2020@mail.sustech.edu.cn, gengtiantian97@gmail.com, {wangjb, zhengf}@sustech.edu.cn

Abstract

This report describes the details of our solution to TinyAction Challenge 2021 that focuses on recognizing tiny actions in videos. To extract rich spatio-temporal features from low-resolution videos, we adopt the R(2+1)D [5] with ResNet-34 [4] as backbone pretrained from a low-resolution setting. In addition, to address the issue of multi-label classification, an asymmetric loss is introduced to effectively relieve the positive-negative imbalance problem during training. Finally, our model ensembles the prediction scores of five clips sampled from videos, achieving an F1 score of 0.410 *on the challenging test set.*

1. TinyVIRATv2

The TinyVIRATv2 dataset [2] is a multi-label action recognition dataset collected from real-world surveillance videos with naturally low resolution. There are 26k videos in total, with around 17k/3k/6k videos for training/validation/test. Each video contains multiple human-centric action instances and the average length of the activities is around 2.5 seconds.



(a) Video duration (b) Video resolution

Figure 1. Distributions of video duration and resolution in the TinyCIRATv2 training set.

The distributions of the video duration and resolution are shown in Fig. 1. Most videos have the duration ranging from 0s-5s and the resolution ranging from 10×10 to 128×128. We find a centralized distribution pattern: Over 45% videos have a duration of around 2.13s and over 35% videos have a resolution of around 112×112.

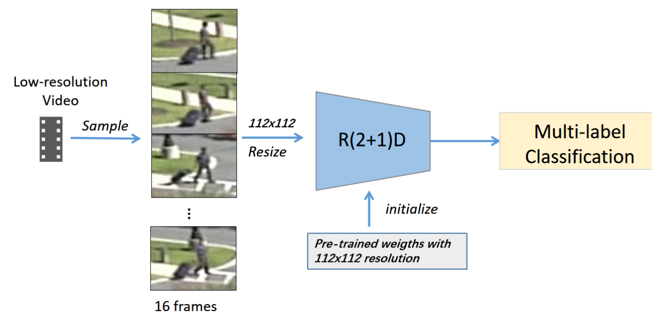


Figure 2. The overall framework. We first sample a 16-frame clip from a low-resolution video, then resize the spatial resolution of the clip to 112×112. R(2+1)D is performed to extract rich spatio-temporal visual features and a multi-label classification layer is adopted to obtain the final prediction.

2. Approach

The overall framework of our method is shown in Fig. 2. We identify two main challenges in the TinyVIRATv2 dataset: First, considering that action recognition model pretrained on high resolution (typically 224×224) may be unsuitable for low-resolution videos, we adopt an R(2+1)D [5] network pretrained on IG65M+Kinetics [3] with an input resolution of 112×112 to alleviate the potential distribution gap between the pretraining datasets and the target dataset TinyVIRATv2. Second, the positive-negative imbalance nature in the multi-label datasets may hurt the optimization process and yield inferior performance. We introduce asymmetric loss as the objective function to alleviate this problem.

2.1. Backbone

As shown in Fig. 3, R(2+1)D [5] is built on 2D-ResNet and adds 1D temporal convolutions into every 2D convolution block. Compared with fully 3D convolutional networks, it factorizes the 3D convolution into separate 2D spatial and 1D temporal components, which speeds up the convergence of the loss and brings significant performance gains. We utilize R(2+1)D with ResNet-34 as our backbone.

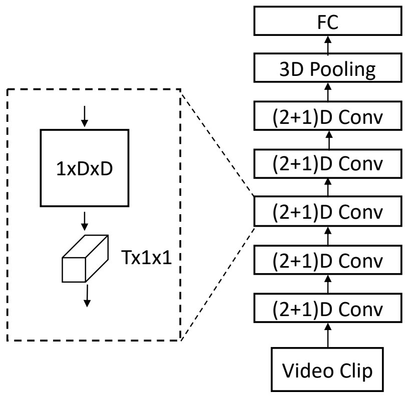


Figure 3. Illustration of our backbone R(2+1)D [5], which consists of a stack of (2+1)D convolutional blocks. A (2+1)D block decomposes a 3D convolutional kernel into a 2D spatial kernel followed by a 1D temporal kernel.

Compared with 3D CNN, R(2+1)D [5] backbone has two appearling proprieties: Firstly, despite not changing the number of parameters, R(2+1)D [5] doubles the number of nonlinearities in the network due to the additional ReLU between the 2D and 1D convolution in each block of ResNet. Increasing the number of non-linearities can increase the complexity of functions which is beneficial for feature extraction. On the other hand, forcing the 3D convolution into separate spatial and temporal components renders the optimization easier.

2.2. Asymmetric Loss

The common practices in multi-label classification adopt a binary cross-entropy loss. Given K labels, the base network outputs the class probability z_k which is activated by a sigmoid function $\delta(z_k)$, and y_k denotes the ground truth for class k. The total classification loss L_{tot} is obtained by aggregating a binary loss from K labels.

$$\begin{gathered} L_{tot} = \sum_{k=1}^{K} L(\delta(z_k), y_k), \\ L = -yL_{+} - (1-y)L_{-}, \end{gathered} \tag{1}$$

where L is a general form of a binary loss per label, y is the ground-truth label, L_+ and L_- are the positive and negative loss parts, respectively.

Such a plain design will inevitably suffer from the problem of positive-negative imbalance. Considering the inherent imbalance nature of multi-label datasets, we apply asymmetric loss (ASL) [1] as the objective function, which is defined as:

$$ASL = \begin{cases} L_{+} = (1-p)^{\lambda_{+}} \log(p) \\ L_{-} = (p_m)^{\lambda_{-}} \log(1-p_m) \end{cases}, \tag{2}$$

where $p = \delta(z)$ is the network's output probability, and $p_m = \max(p-m, 0)$. By contrast, the asymmetric loss [1] contains the two mechanisms of asymmetric focusing and probability shifting, which are integrated into a unified formula using soft thresholding via the focusing parameter λ and hard thresholding based on the probability margin m. Both mechanisms are used for reducing the contribution of easy negative samples to the loss function. Note that λ_- is usually larger than λ_+ in practice.

3. Experiments

R(2+1)D [5] is first pretrained on IG65M+Kinetics [3] with cross-entropy loss for multi-class classification, then finetuned on the TinyVIRATv2 with asymmetric loss for multi-label classification.

Methods	F1	Precision	Recall
Single sample	0.4016	0.4495	0.3925
Five samples (Final submission)	0.4102	0.4426	0.4177

Table 1. Results on the TinyVIRATv2 dataset.

The implementation details are described as follows, and experimental results are shown in Table 1. For training, we sample 16 frames randomly and then resize the frames with various resolutions to the same size, 112×112, and then input them to R(2+1)D [5]. For evaluation, we sample 16 frames randomly five times and then average the classification scores to get the final results. The single sample achieves a 0.4016 F1 score and the ensemble of five samples gets better performance with a 0.4102 F1 score.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant No. 61972188.

References

[1] Emanuel Ben-Baruch, Tal Ridnik, Nadav Zamir, Asaf Noy, Itamar Friedman, Matan Protter, and Lihi Zelnik-Manor. Asymmetric loss for multi-label classification. *arXiv preprint arXiv:2009.14119*, 2020. 2

[2] Ugur Demir, Yogesh S Rawat, and Mubarak Shah. Tinyvirat: low-resolution video action recognition. In *ICPR*, pages 7387–7394. IEEE, 2021. 1

[3] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *CVPR*, pages 12046–12055, 2019. 1, 2

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[5] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1, 2