SUSTech&HKU Submission to TinyAction Challenge 2021

Teng Wang\textsuperscript{1,2}, Tiantian Geng\textsuperscript{3}, Jinbao Wang\textsuperscript{1}, Feng Zheng\textsuperscript{1}
\textsuperscript{1} Southern University of Science and Technology, \textsuperscript{2} The University of Hong Kong, \textsuperscript{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 \cite{r21d} with ResNet-34 \cite{resnet} 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 \cite{tinyvira} 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.

![Figure 1. Distributions of video duration and resolution in the TinyCIRATv2 training set.](image)

2. Approach

The overall framework of our method is shown in Fig. 2. We first sample a 16-frame clip from a low-resolution video, then resize the spatial resolution of the clip to $112 \times 112$. 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.

2.1. Backbone

As shown in Fig. 3, R(2+1)D \cite{r21d} 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.
where $p$ is defined as: asymmetric loss (ASL) [1] as the objective function, which contains the two mechanisms of asymmetric focusing and probability shifting, which are integrated into a unified formula using soft thresholding via the focusing parameter $\lambda$ 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 $\lambda_-$ is usually larger than $\lambda_+$ in practice.

3. Experiments


<table>
<thead>
<tr>
<th>Methods</th>
<th>F1</th>
<th>Precision</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single sample</td>
<td>0.4016</td>
<td>0.4495</td>
<td>0.3925</td>
</tr>
<tr>
<td>Five samples (Final submission)</td>
<td>0.4102</td>
<td>0.4426</td>
<td>0.4177</td>
</tr>
</tbody>
</table>

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, $112 \times 112$, 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