Team SSS Submission to the Moments in Time Challenge 2018

Yao Zhou, **Pingchuan Ma** and **Yu Lu**
SenseTime Research
{zhouyao,mapingchuan,luyu}@sensetime.com

Abstract

This draft presents our methods and results on the Moments in Time challenge 2018. To tackle the task of recognizing the events or activities in trimmed videos, we tried various models with different input modalities. In summary, we not only explored both 2D and 3D models with the different backbones, but also fed both RGB frames and optical flows into these models. Finally, we ensemble these models to achieve the better recognition performance.

1. Our Approach and Result

In the Moments in Time challenge 2018, the participants should design methods to recognize the events in the three-seconds videos. This challenge uses the Moments in Time dataset as the benchmark. This dataset is challenging for the reason that (1) the events occurred in videos are abstract, (2) the events are not only performed by human, but also animals or objects, (3) the events are visual and/or audible actions. This dataset contains 802,264 videos for training, 33900 for validation and 67800 for testing. At this time, each video only belongs to one of 339 classes.

We present the experiment result at Table 1.

Model	Modality	K	Backbone	Top-1 Acc.	Top-5 Acc.
C2D TSN	RGB	5	Inception-v3	27.2%	51.5%
C2D TSN	RGB	5	ResNet-101	29.5%	55.8%
C2D TRE	RGB	5	ResNet-101	30.8%	56.6%
C2D R101	Flow	5	ResNet-101	16.4%	37.5%
C2D TRE	Flow	5	ResNet-101	16.8%	38.0%
I3D Inv1	RGB	16	Inception-v1	26.2%	50.3%
I3D R50	RGB	16	ResNet-50	26.6%	50.5%
NL-I3D R50	RGB	16	ResNet-50	28.1%	53.7%
I3D Inv1	Flow	16	Inception-v1	10.1%	27.9%

Table 1. The experiment results on the validation set of Moments in Time full dataset. The first column indicates the model names. For 2D models, K present the num of segments for training. For 3D models, K presents the num of frames for training. The first group presents the results performed by the C2D models with different backbones when using the RGB modality. The second group presents the optical flow results. The last two groups are I3D models with different backbones and modalities.