UNSW Video Classification System for Moments in Time Challenge 2018

Zhihui Li
School of Computer Science and Engineering
University of New South Wales
zhihuilics@gmail.com

Lina Yao
School of Computer Science and Engineering
University of New South Wales
lina.yao@unsw.edu.au

Abstract

This paper presents our system for the video understanding task of the Moments in Time Challenge 2018. Because of limited computational resources, we only used three features in the system, including 2 visual features and 1 audio feature. After we have the prediction scores of these three features, we combine them using late fusion and obtain the final result. Specifically, we observe average fusion can get promising results in our experiments.

1. Introduction

Although researchers have devoted much research attention to the visual understanding problem, it is still a challenging problem. The ubiquitous video record devices have created videos far surpassing what the users can watch. Hence, it becomes increasingly urgent to develop efficient algorithms for automatic video analysis.

Researcher have made much progress to introduce large-scale datasets for training reliable deep learning models, for example, ImageNet [4], and Youtube8M [3] . Recently, researchers from MIT have introduced the Moments in Time Dataset, a collection of one million short videos with a label each, corresponding to actions and events unfolding within 3 seconds.

2. The Proposed System

In this section, we describe the proposed system for Moments in Time Challenge 2018.

2.1. Feature Extraction

For the limitations of computing resources, we only employ three features in our system, including 2 visual features and 1 audio feature.

Visual Features: We first pretrain an Inflated 3D ConvNet (I3D) [2] model on ImageNet and Kinects datasets. Then we apply the pretrained model to the Moments dataset, and extract the last pooling layer as the representation for each video.

In addition, we use the TRN-Multiscale [5] following the baselines reported in [3], since it achieves the best single model performance. We do not pre-train or fine-tune on this model. In other words, we only do inference for this model.

Audio Feature: We employ raw waveforms as the input modality and adopt the network architecture from SoundNet [1]. The only difference is that we changed the last layer to predict the categories from the Moments dataset. We fine-tune the model downloaded from the official website.

2.2. Inference

For the TRN-Multiscale and audio features, the inference is conducted end-to-end. For the I3D feature, we feed the last pooling layer into a 200-mixture Mixture of Experts (MoE) layer for classification.

2.3. Fusion

When the prediction scores for the three models are ready, we fuse them using average fusion for its simplicity and efficiency.

3. Results

We report the results in Table 1. From the experimental results we can see that the I3D model gets the best single model performance on the validation set. Also, we observe that with only three models, we get similar performance as the baseline reported in the baseline paper.

4. Conclusion

In this paper, we have presented our system for the Moments challenge. Although the computing resource is very limited (with only one Titan Xp), we finally achieve promising results on the validation set.

Table 1. The performance evaluation on the validation set.

Feature Name	Mode	Top-1	Top-5
I3D	Visual	29.53	56.28
TRN-Multiscale	Visual	28.27	53.87
SoundNet	Audio	7.60	17.96
Average Fusion	–	30.25	57.84

Acknowledgement

Dr. Yao is partially supported by the Australian Research Council DECRA project.

References

[1] Y. Aytar, C. Vondrick, and A. Torralba. Soundnet: Learning sound representations from unlabeled video. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 892–900, 2016. 1

[2] J. Carreira and A. Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 4724–4733, 2017. 1

[3] M. Monfort, B. Zhou, S. A. Bargal, A. Andonian, T. Yan, K. Ramakrishnan, L. M. Brown, Q. Fan, D. Gutfreund, C. Vondrick, and A. Oliva. Moments in time dataset: one million videos for event understanding. *CoRR*, abs/1801.03150, 2018. 1

[4] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and F. Li. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015. 1

[5] B. Zhou, A. Andonian, and A. Torralba. Temporal relational reasoning in videos. *CoRR*, abs/1711.08496, 2017. 1