Trimmed Event Recognition (Moments in Time): Submission to ActivityNet Challenge 2018

Dongyang Cai
caidongyang_sx@qiyi.com

Abstract

In this paper, a brief description is provided of the method used for the task of trimmed event recognition (Moments in Time). A set of TRN models were used to train video classification models for the 200 action categories of the Moment in Time Mini Database, and the P3D feature is also used to further enhance the model diversity, finally we propose a simple yet effective method to combine different modalities together for action prediction.

1. Introduction

The Moments in Time Dataset [1], is a large-scale human-annotated collection of one million short videos corresponding to dynamic events unfolding within three seconds, each video is tagged with one action or activity label. Modeling the spatial-audio-temporal dynamics even for actions occurring in 3 second videos poses many challenges: meaningful events do not include only people, but also objects, animals, and natural phenomena; visual and auditory events can be symmetrical or not in time. Here, with limited computation resources, we will use the Moments in Time Mini dataset, which is a subset of Moments in Time with 100k videos provided in the training set and involves 200 action categories, for model training and action prediction. As we note the temporal relational reasoning is very important for this task [2], we train a set of Temporal Relation Network (TRN) models firstly, also the recently proposed P3D method [3] is found useful to enhance the model diversity, finally we propose a simple yet effective method to combine those methods above to identify the event labels depicted in a 3 second video.

2.Method Description

2.1 TRN

Temporal relational reasoning is critical for activity recognition, forming the building blocks for describing the steps of an event. A single activity can consist of several temporal relations at both short-term and long-term timescales, the ability to model such relations is very important for activity recognition. The Temporal Relation Network (TRN) proposed by Bolei Zhou et al [2] is designed to learn and reason about temporal dependencies between video frames at multiple time scales. It is an effective and interpretable network which is able to learn intuitive and interpretable visual common sense knowledge in videos. The networks used for extracting image features is very important for visual recognition tasks, here we use an 8 segment multi-scale TRN with an inceptionV3 base and Inception with Batch Normalization (BN-Inception) base separately, and then train the TRN-equipped network with different data augmentation scheme with each base network, we found training a set of TRN networks with fusing them together bring action prediction improvement.

2.2 P3D

Pseudo-3D Residual Net (P3D ResNet) architecture proposed by Qiu, Zhaofan et al [3], aims to learn spatio-temporal video representation in deep networks, it simplifies 3D convolutions with 2D filters on spatial dimension plus 1D temporal connections. Experiments on five datasets in the context of video action recognition, action similarity and scene recognition also demonstrate the effectiveness and generalization of spatio-temporal video representation produced by P3D ResNet. Here, to enhance our model diversity, we adopted P3D ResNet to learn feature representation of the Moments in Time Mini Dataset, and utilized the learned features for this video classification task.

2.3 Weak classifiers

In this part, we use the idea of AdaBoost[4] to generate our own classifier. First, we rank the 200 classes according to their accuracy on the validation dataset. Then, we choose 50 classes with the lowest accuracy and increase their sample weight for future training. What is more, we calculate all the training data and get the confusion matrix of the 200 classes, for those confusing categories, we trained weak classifiers to classify them especially. For all the training samples which was classified wrong before, their weight will be also increased for another weak classifier. By this method, we achieved improvement on the accuracy of testing data.

2.4 Ensemble

We propose a simple yet effective model ensemble method to enhance the action prediction ability of our final classification model. Firstly, we will calculate the classification accuracy of each model referred above on the validation dataset, then we assign a weight to each model according to its classification accuracy, model with high accuracy embracing a higher weight. Given a test video, we firstly predict its top 5 labels with each model. We will give a label weight to a predicted label according to its number of occurrences across models. For one model, we will multiply confidence score of each predicted label with the model weight referred above, and then add up the resulting value of the same predicted label across models with its label weight. Finally, we will rank the predicted labels according to their label scores above in descending order and get the top 5 labels for action prediction of the test video. Our experimental results below will show the effectiveness of our method.

3.Experimental Results

The Moments in Time Mini Dataset contains 100000 training videos, 10000 validation videos and 20000 testing videos. Each video is in one of 200 categories. Table 1 summarizes our results on the Moments in Time Mini validation dataset.

Model	Top-1 Accuracy (%)	Top-5 Accuracy (%)
TRN	26.1%	48.5%
P3D	14.7%	33.4%
Weak classifier	28.3%	52.2%
TRN+P3D+ Weak classifier	31.7%	56,9%

Table 1. Moments in Time Mini validation results.

4.Conclusion

The recently proposed TRN method is effective for recognizing daily activities with learning intuitive and interpretable visual common sense knowledge in videos. Also, the P3D feature is used to enhance model diversity. We utilize those methods and propose a simple yet effective method to combine different modalities together for action prediction of the Moment in Time Mini Dataset, our experimental results show its effectiveness.

5.Acknowledgement

This work was finished during an internship work in IQIYI, many thanks to Jie Liu, Tao Wang, and Xiaoning Liu for helpful comments and discussion.

References

[1] Monfort, Mathew, et al. "Moments in Time Dataset: one million videos for event understanding." *arXiv preprint arXiv:1801.03150* (2018).

[2] Zhou, Bolei, Alex Andonian, and Antonio Torralba. "Temporal Relational Reasoning in Videos." *arXiv preprint arXiv:1711.08496* (2017).

[3] Qiu, Zhaofan, Ting Yao, and Tao Mei. "Learning spatio-temporal representation with pseudo-3d residual networks." *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017.

[4] Freund, Yoav, and Robert E. Schapire. "A decision-theoretic generalization of on-line learning and an application to boosting." *Journal of computer and system sciences* 55.1 (1997): 119-139.