Semantic Guided Network for Open Domain Action Recognition

Tingzhao Yu, Lingfeng Wang, Huxiang Gu, Shiming Xiang, and Chunhong Pan

***Abstract*—In this notebook, we present a novel semantic guided deep neural network for video-based open domain action recognition. In this network, a semantic guided module, named SGM, is proposed to take multi-level semantic information into consideration and as a guidance of getting feasible spatial-temporal features. SGM is differentiable such that the network can be trained end-to-end via back-propagation. Evaluations on Open Domain Action Recognition benchmark data set demonstrate the effectiveness of the proposed network.**

***Index Terms*—Open Domain, Action Recognition, 3D Residual Network, Semantic Guided Module.**

I. Introduction

HUMAN action recognition plays a fundamental role in video analysis. It has many applications such as video surveillance, video retrieval and crowd analysis [1]. Despite the significant progress in recent years, extracting effective spatial-temporal descriptors is still difficult and action recognition remains a challenging task. This is partly due to the long temporal duration, video frame redundancy, background and viewpoint variation. Convolution 3D (C3D) [2] has been proved to be an effective descriptor in video action recognition.

II. Architecture

In this notebook, we propose a novel semantic guided network for open domain action recognition based on C3D [2] . This design ensures that the network pays attention to the semantically important parts of the feature map. Fig. 1 demonstrates the framework of the proposed network. The network consists in two kinds of module, namely *3D Residual Convolution Module* [1] and *3D Semantic Guided Module*.

Details of the proposed 3D Semantic Guided Module and an intuitive comparison among the 3D Residual Convolution Module, the 3D Residual Identity Module and 3D Semantic Guided Module can be found in Fig. 2.

III. Evaluation

We implement our proposed SGM network on Open Domain Action Recognition Challenge. We start by splitting each video into 8-frame video clips. This technique amplifies the training set from 3493 into 21754. Due to the large variance in the *unknown* (abbreviated as *un*) class, at training stage, we ignore these video clips. But the output of our network is designed to include *un* (Thus the length of the output is 13).

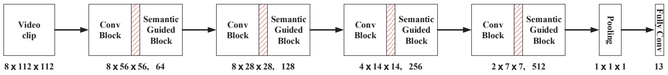
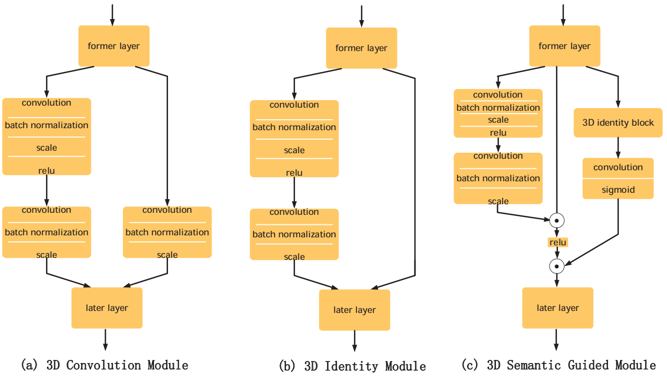


Fig. 1. Semantic Guided Network Architecture.



(a) 3D Convolution Module (b) 3D Identity Module (c) 3D Semantic Guided Module

Fig. 2. Comparison of (a) the 3D Residual Convolution Module, (b) the 3D Residual Identity Module and (c) the proposed 3D Semantic Guided Module.

At testing stage, we firstly calculate the clip level accuracy (including *un*) and then aggregate them into video level via majority votes. TABLE I present the each class accuracy and the video level accuracy . We release our code, trained model and more detailed technique report[2].

TABLE I
Test Accuracy Report of Each Class.

class	***Acc.***	class	***Acc.***	class	***Acc.***
boxing	33.6	celltoear	13.4	clapping	28.5
drink	23.3	getup	23.9	jump	24.1
point	14.3	running	23.5	sitdown	23.2
throw	21.9	walk	57.9	wave	56.7
top-1.	**33.9**	***top-5.***	**74.2**	VideoLevel	**32.6**

IV. Conclusion

In this notebook, we propose a novel semantic guided network for open domain action recognition. Evaluation on the benchmark data set demonstrates the effectiveness of the proposed method.

References

[1] T. Yu, H. Gu, L. Wang, S. Xiang, and C. Pan, “Cascade temporal spatial features for video action recognition,” in *Proceedings of the IEEE International Conference on Image Processing*, 2017.

[2] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3d convolutional networks,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4489–4497.

The authors are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 100190 Beijing, China. T. Yu is also with School of Computer and Control Engineering, University of Chinese Academy of Sciences, 101408 Beijing, China. (e-mail: tingzhao.yu@nlpr.ia.ac.cn.)

[1] Caffe C3D Residual Network. Available at: https://github.com/facebook/C3D/tree/master/C3D-v1.1/examples/c3d_ucf101_finetuning

[2] https://github.com/Tsingzao/SemanticGuidedBlock