UTSA submission to THUMOS 2015

Junjie Cai and Qi Tian
Department of Computer Science, University of Texas at San Antonio
caijunjieustc@gmail.com,qitian@cs.utsa.edu

Abstract

For THUMOS 2015 action recognition task, we submitted one run of our results, which used the features trained from deep neural network. Since testing videos are temporally untrimmed, we applied a sliding window of 100 frames for both validation and test dataset. Video-level scores were generated by maximum pooling of video clips.

1. Introduction

We evaluate our system's action recognition performance on UCF101 dataset [2] and validation dataset. UCF101 is an action recognition data set of action videos collected from YouTube. There are 13,320 temporally segmented videos from 101 action categories, the average length of the videos is seven seconds. This temporal segmentation of testing videos might not reflect the real world as most of the Web videos are temporally noisy and untrimmed.

Motivated by above target problems, this year's challenge is evaluated on around 5,613 temporally untrimmed videos, where the actions of interest might happen anywhere within the videos. We decided to use temporal sliding windows for trimming the long videos into small clips.

In the following sections, we describe our system's video level features in Section 2, classification framework in Section 3 and summarization in Section 4.

2. Clip Level Features

We extracted the features via the convolutional neural network provided by [4]. The activations of the neurons in the intermediate hidden layers could be used as strong features for a variety of video recognition tasks because they contain much richer and more complex representations than any earlier convolutional layer in the network. In this work, we leverage as deep feature the output of the intermediate layer named with fc_6 in the CAFFE model zoo [3]. We set the network input to the raw RGB values of the frames, resized to 256*256 pixels, and the values are forward propagated through 5 convolutional layers (i.e., pooling and ReLU non-linearities) and 3 fully-connected layers (i.e., to determine its final neuron activities). We obtain the 4096-dimension vector from the intermediate fc_6 hidden layer.

For validation and test set, we set the length of sliding window of as 150 frames and sliding step as 100 frames.

3. Classification

We used LIBLINEAR [1] to train one-versus-the-rest action classifiers. For each short temporal window, we perform the task of action recognition independently. Eventually, the recognition results of these short window are combined to yield the final result of the whole video stream.

4. Conclusions

In this notebook paper, we have described our submission to the THUMOS 2015 Challenge, and presented an experimental result of our evaluation. Our main findings are listed as follows.(i) the the deep features provides better semantic information largely complementary to low-level dynamic trajectory features. (ii)semantic information can be leveraged to enhance the recognition performance.

References

[1] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9(2008), 1871-1874.

[2] K. Soomro, A.R. Zamir, M. Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. CRCV-TR-12-01.

[3] Y. Jia. Caffe: An Open Source Convolutional Architecture for Fast Feature Embedding. `http://caffe.berkeleyvision.org/`, 2013.

[4] K. Simonyan and A. Zisserman . Very Deep Convolutional Networks for Large-Scale Visual Recognition. arXiv technical report, 2014.

[5] J. Cai, M. Merler, S. Pankanti and Q. Tian. Heterogeneous Semantic Level Features Fusion for Action Recognition.. ACM International Conference on Multimedia Retrieval,2015