A system for TRECVID MED by MCIS

Hao Song, Wennan Yu, Yuchao Sun, Kun Wu
Beijing Laboratory of Intelligent Information Technology, School of Computer Science,
Beijing Institute of Technology, Beijing 100081, P.R. China

Abstract

We presented a simple system for the 2016 TRECVID Multimedia Event Detection[1]. Our system follows the standard pipeline and consists two parts: feature extraction and classification. The feature extraction part is implemented by Caffe and BOW, and the classification is implemented by LIBSVM.

1. Introduction

Automatically detecting events in videos attracts increasing research interests due to its usefulness in video surveillance, video content analysis, and video retrieval. There are great intra-class variation in the unconstrained event that makes it difficult to design efficient and robust systems for event detection.

Recently, convolution neural networks (CNN) based feature representation have been shown the extreme effectiveness of solving many visual problems. Driven by the achievements of CNN [2, 3, 4], we utilize C3D network[5] to extract the feature of the TRECVID MED datasets.

2. Method

Metadata Generator: We first extract the frame-level feature by feeding the frame to the Caffe Toolkit [6] with the model shared by [7] to get the fc7 level feature. Specifically, for each video, we divide it into many parts whose duration is 2 second. Secondly, we utilize C3D network to extract the feature of each part and then utilize the Bag-of-Words model (BOW) to encode the frame-level descriptor, and then we get the video representation.

Event Query Generator: Because of the high feature dimension, we choose linear SVM with LIBSVM [8] for classification.

3. Conclusion

We have proposed an effective video feature for event detection. With Caffe, we can implement this feature easily and efficiently.

Reference

[1] George Awad and Jonathan Fiscus and Martial Michel and David Joy and Wessel

Kraaij and Alan F. Smeaton and Georges Qu énot and Maria Eskevich and Robin Aly and Gareth J. F. Jones and Roeland Ordelman and Benoit Huet and Martha Larson, TRECVID 2016: Evaluating Video Search, Video Event Detection, Localization, and Hyperlinking, in: Proceedings of TRECVID 2016, NIST, USA, 2016.

[2] S. Ji, W. Xu, M. Yang, K. Yu, 3d convolutional neural networks for human action recognition, Pattern Analysis and Machine Intelligence, IEEE Transactions on 35 (1) (2013) 221–231.

[3] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in neural information processing systems, 2012, pp. 1097–1105.

[4] Z. Xu, Y. Yang, A. G. Hauptmann, A discriminative cnn video representation for event detection, in: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, IEEE, 2015.

[5] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, Manohar Paluri Facebook AI Research, Dartmouth College, Learning Spatiotemporal Features with 3D Convolutional Networks, in: arXiv:1412.0767v4 [cs.CV] 7 Oct 2015

[6] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama,T. Darrell, Caffe: Convolutional architecture for fast feature embedding, arXiv preprint arXiv:1408.5093.

[7] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[8] C.-C. Chang, C.-J. Lin, Libsvm: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST) 2 (3) (2011) 27.