YOLO-based Adaptive Window Two-stream Convolutional Neural Network for Video Classification

Charles Han, Chao Wang, Evelyn Mei
hcs@stanford.edu, cwang17@stanford.edu, evelyn66@stanford.edu

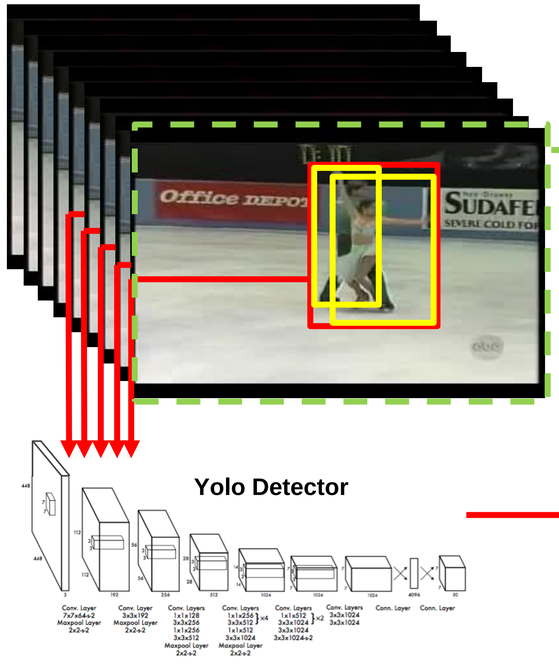
Introduction

Convolutional Neural Networks (CNN) have been adopted widely for image classification problems. As they demonstrate significant success, more and more researchers start to deploy CNN on video classification problems. The main challenge is to capture not only the appearance information present in single, static frames, but also complex temporal evolution. Among video classification tasks, human action recognition is the key problem.

Approach

Existing research [2] has shown that a two-stream approach, namely spatial stream plus temporal stream performed significantly better than training on raw stacked frames. In order to get rid of the noise that's brought by different camera orientations and distances, we want to detect and segment out human action from the background scene. Inspired by Wang's idea [3] but different from his approach of using R-CNN to propose a region of person cue, we generate a person cue without region proposal method. We adopt *You Only Look Once* (YOLO) [4], this new object detection approach to localize the human action. We believe it is going to capture human motion better and give us better action recognition accuracy with a fast speed.

YOLO-based Adaptive Window Two-stream Structure



Future Work

Comparing with state-of-the-art models (testing accuracy 80%+), there's more work to be done:

1) Improve human-detection accuracy:
- Train YOLO on a human recognition dataset rather than VOC07. Tune YOLO hyper-parameter and threshold.
- Try out Region-based proposal methods such as faster RCNN

2) Improve network structure, and classifiers:
- Fine tune hyper-parameters, optimizer and learning rate
- Improve fusion methods
- Integrate LSTM or other RNN structure on Temporal stream so as to better capture temporal information

Dataset

UCF101 is currently the largest dataset of human actions. It consists of 101 action classes, 13320 clips and 27 hours of video data. The clips of one action class are divided into 25 groups which contain 4-7 clips each. The clips in one group share some common features, such as the back-ground or actors. The mean clip length is 7.21 sec and the total duration 1600 mins. The min clip length is 1.06 sec and the max clip length is 71.04 sec. All clips have fixed frame rate and resolution of 25 FPS and 320×240 respectively. We pre-process all the videos into images at 5FPS.

UCF-101- Action Recognition Data Set Samples

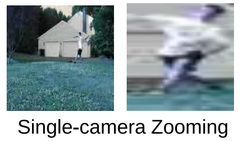
Results

Accuracy	Spatial Stream	Temporal Stream	Two-Stream with Fusion
Training	0.80	0.95	0.98
Testing	0.39	0.51	0.68

1) Testing Accuracy (68%) is achieved on UCF-101 action recognition dataset, which beats baseline [1] Karpathy, et al. (2014).
2) Yolo based adaptive window method significantly improved model performance comparing with central window method.
3) Temporal stream provides better prediction accuracy than Spatial stream (51% vs. 39%).

Discussion

Compared with Karpathy's multi-resolution CNN architecture which combines a low-resolution context stream and a high-resolution fovea-stream center crop, our approach replaced the fovea-stream with a Yolo-based temporal stream which extracted person cue from the video and thus improved the classification performance. However, we have also noticed the person cue is sensitive to camera view point, which may affect the results:



Single-camera Zooming

Mutli-camera Alternation

Reference

[1] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).
[2] Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems* (pp. 568-576).
[3] Wang, Y., Song, J., Wang, L., Van Gool, L., & Hilliges, O. (2016). Two-Stream SR-CNNs for Action Recognition in Videos. BMVC.
[4] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. (2016). You Only Look Once: Unified, Real-Time Object Detection. CVPR.
[5] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, Xiangyang Xue. (2015). Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. ACM MM.