R-C3D: Region Convolutional 3D Network for Temporal Activity Detection

Liangjie Cao

Aug. 9, 2018

Abstract

The authors address the problem of activity detection in continuous, untrimmed video streams. This is a difficult task that requires extracting meaningful spatio-temporal features to capture activities, accurately localizing the start and end times of each activity. We introduce a new model, Region Convolutional 3D Network (R-C3D), which encodes the video streams using a three-dimensional fully convolutional network, then generates candidate temporal regions containing activities, and finally classifies selected regions into specific activities. Computation is saved due to the sharing of convolutional features between the proposal and the classification pipelines. The entire model is trained end-to-end with jointly optimized localization and classification losses. R-C3D is faster than existing methods (569 frames per second on a single Titan X(Mine is GTX960M) Maxwell GPU) and achieves state-of-the-art results on THUMOS'14. They further demonstrate that our model is a general activity detection framework that does not rely on assumptions about particular dataset properties by evaluating our approach on ActivityNet and Charades. Their code is `http://ai.bu.edu/r-c3d/`

1. Introduction

Activity detection in continuous videos is a challenging problem that requires not only recognizing, but also precisely localizing activities in time. Existing state-of-the-art approaches address this task as detection by classification, *i.e.* classifying temporal segments generated in the form of sliding windows [2, 3] or via an external "proposal" generation mechanism [6]. These approaches suffer from one or more of the following major drawbacks: they do not learn deep representations in an end-to-end fashion, but rather use hand-crafted features, or deep features like VGG [4], ResNet [1], C3D [5] *etc.*, learned separately on image/video classification tasks. Such off-the-shelf representations may not be optimal for localizing activities in diverse video domains, resulting in inferior performance. Furthermore, current methods' dependence on external proposal generation or exhaustive sliding windows leads to poor computational efficiency. Finally, the sliding-window models cannot easily predict flexible activity boundaries.

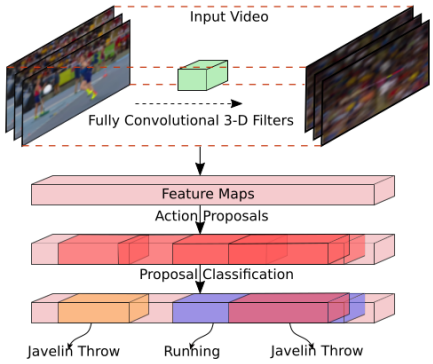

Figure 1. They propose a fast end-to-end Region Convolutional 3D Network (R-C3D) for activity detection in continuous video streams. The network encodes the frames with fully-convolutional 3D filters, proposes activity segments, then classifies and refines them based on pooled features within their boundaries. Their model improves both speed and accuracy compared to existing methods.

In this paper, they propose an activity detection model that addresses all of the above issues. Their Region Convolutional 3D Network (R-C3D) is end-to-end trainable and learns task-dependent convolutional features by jointly optimizing proposal generation and activity classification. Inspired by the Faster R-CNN [21] object detection approach, they compute fully-convolutional 3D ConvNet features and propose temporal regions likely to contain activities, then pool features within these 3D regions to predict activity classes (Figure 1). The proposal generation stage filters out many background segments and results in superior computational efficiency compared to sliding window models. Furthermore, proposals are predicted with respect to predefined anchor segments and can be of arbitrary length, allowing detection of flexible activity boundaries.

2. Approach

They propose a Region Convolutional 3D Network (R-C3D), a novel convolutional neural network for activity detection in continuous video streams. The network, illustrated in Figure 2, consists of three components: a shared 3D ConvNet feature extractor [5], a temporal proposal stage, and an activity classification and refinement stage. To enable efficient computation and end-to-end training, the proposal and classification sub-networks share the same C3D feature maps. The proposal subnet predicts variable length temporal segments that potentially contain activities, while the classification subnet classifies these proposals into specific activity categories or background, and further refines the proposal segment boundaries. A key innovation is to extend the 2D RoI pooling in Faster R-CNN to 3D RoI pooling which allows our model to extract features at various resolutions for variable length proposals.

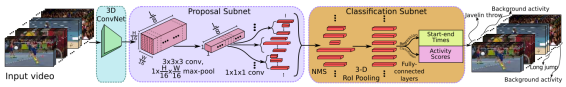

Figure 2. R-C3D model architecture. The 3D ConvNet takes raw video frames as input and computes convolutional features. These are input to the Proposal Subnet that proposes candidate activities of variable length along with confidence scores. The Classification Subnet filters the proposals, pools fixed size features and then predicts activity labels along with refined segment boundaries

3. Activity Classification Subnet

The activity classification stage has three main functions: 1) selecting proposal segments from the previous stage, 2) three-dimensional region of interest (3D RoI) pooling to extract fixed-size features for selected proposals, and 3) activity classification and boundary regression for the selected proposals based on the pooled features.

References

[1] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[2] S. Karaman, L. Seidenari, and A. Del Bimbo. Fast saliency based pooling of fisher encoded dense trajectories. In *ECCV*, 2014. 1

[3] D. Oneata, J. Verbeek, and C. Schmid. The lear submission at thumos 2014. In *ECCV*, 2014. 1

[4] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 1

[5] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV*, 2015. 1, 2

[6] L. Wang, Y. Qiao, X. Tang, and L. Van Gool. Actionness estimation using hybrid fully convolutional networks. In *CVPR*, 2016. 1