Temporal Action Detection based on Temporal Deformable Proposal Generation

Liting Yan Yongzhao Zhan Huifen Xia
School of Computer Science and Communication Engineering, Jiangsu University

三等奖

Abstract

Temporal action detection is a challenging task in video understanding. Locating the activities more accurately on fuzzy temporal action boundary is a difficulty. So it is crucial to generate temporal proposals with precise boundaries and high-quality. To better solve this issue, we propose a novel temporal action detection method based on Temporal Deformable Proposal Generation (TDPG). In TDPG, modified C3D network is used to extract robust features. Deformable Proposals Generation module adaptively expands the receptive field through temporal deformable convolution to generate deformable proposals with more precise temporal boundaries. And then selected deformable proposal segments are used to predict action classification scores and refine temporal boundary. Our method is trained end to end with jointly optimized classification and regression loss. Experimental results show that our method achieves better performance than state-of-the-art method on two public datasets THUMOS’14 and Charades.

Approach

1. Framework

As shown in Fig 1, the overall architecture of the system consists of Feature Extraction, Deformable Proposal Generation and Classification with Deformable Proposals. By acquiring temporal features in the modified C3D network, we input temporal feature into DPG to extract variable length candidate proposals, in the Localization Subnet, soft-NMS is used to delete duplicate candidate proposals, classification layers and The regression layer is used to obtain the final test result.

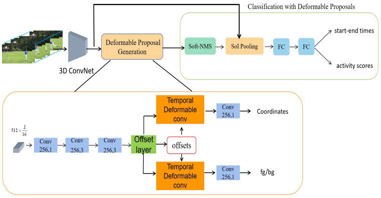

Fig. 1. Framework of our method.

2. Feature Extraction

In our model, modified C3D network is used to extract temporal features from a sequence of RGB frames denoted as $V \in R^{C \times L \times H \times W}$, where C is the number of channels, L is the number of frames, H and W are the height and width of each frame respectively. Our feature extraction module uses conv1 to conv5b in C3D and a 3D max-pooling layer (kernel size 2×H/16×W/16, temporal stride 2).

3. Deformable Proposal Generation

Deformable Proposal Generation (DPG) is shown in Fig. 1. Inspired by Dai et al., we changed the deformable convolution to temporal dimension. It is divided into two steps. Firstly, temporal deformable convolution expands the sampling locations with additional temporal offsets. Secondly, temporal deformable convolution aggregates the features of sampled locations. Specifically, a regular grid R is used for sampling on the input feature map x, and R={-1,...,0,...,1}. For each temporal feature location p, the output feature map y is defined as:

$$y(p) = \sum_{p_n \in R} w(p_n) \cdot x(p + p_n)$$

where w(pn) represents the weight; pn is the n-th element in R. In the next step, we use the temporal offset Δpn on temporal convolution to expand the regular grid R to obtain the output of the temporal deformable convolution:

$$y_{deform}(p) = \sum_{p_n \in R} w(p_n) \cdot x(p + p_n + \Delta p_n)$$

where {Δpn | n=1,...,N}, N = |R|. Since the temporal offset Δpn is usually a fraction, it can be calculated by temporal linear interpolation:

$$x(p + p_n + \Delta p_n) = \sum_s G(s, p + p_n + \Delta p_n) \cdot x(s)$$

where p+ pn+Δpn denotes an arbitrary location; s enumerates all integer positions in the input feature map x.

4. Classification with Deformable Proposals

Classification with Deformable Proposals(CDP) first uses soft-NMS to delete and select deformable proposal segments obtained in DPG, then uses Segments of Interest Pooling (SoI Pooling) to extract fixed-size features of selected proposals, and finally performs action classification and boundary regression based on pooled features.

5. Loss function

To train TDPG, we need to jointly optimize the classification and regression tasks. We define a multi-task loss function which contains the weighted sum of classification loss and regression loss:

$$L_{TDPG} = \frac{1}{N_{cls}} \sum_i L_{cls}\left(a_i, a_i^*\right) + \lambda \frac{1}{N_{reg}} \sum_i a_i^* L_{reg}\left(r_i, r_i^*\right)$$

where L_{cls} , L_{reg} respectively represent multi class cross entropy loss function and smooth L1 loss function; N_{cls}, N_{reg} is batch size and the number of proposal segments respectively;is a loss trade-off parameter and is set to a value 1;λ is respectively the predicted probability of activities and the ground truth.The coordinate transformation is defined as:

$$\Delta c_i = \left(c_i^* - c_i\right) / l_i$$
$$\Delta l_i = \log\left(l_i^* / l_i\right)$$

Experiments

1.Datasets

THUMOS'14 has 200 validation set videos and 213 test set videos with time annotations for the temporal detection task. Charades is a recently introduced dataset for activity classification and detection.

2.Results

Table 1. Action detection results on testing set of THUMOS'14, measured by mAP (%) at different IoU thresholds.

Method	*0.1*	*0.2*	*0.3*	*0.4*	*0.5*
SSAD	50.1	47.8	43.0	35.0	24.6
SS-TAD	-	-	45.7	-	29.2
S-CNN	47.7	43.5	36.3	28.7	19.0
R-C3D	54.5	51.5	44.8	35.6	28.9
CDC	-	-	40.1	29.4	23.3
TURN	54.0	50.9	44.1	34.9	25.6
CTAP	-	-	-	-	29.9
CBR	**60.1**	56.7	50.1	41.3	31.0
ETP	-	-	48.2	42.4	34.2
BSN	-	-	53.5	45.0	36.9
Ours	58.6	**58.3**	**55.6**	**49.8**	**40.4**

Table 2. Action detection results on testing set of Charades measured by mAP@0.5.

Method	*mAP*
Random[14]	2.42
RGB[14]	7.89-
Two-Stream[14]]	8.94
Two-Stream+LSTM[14]	9.6
R-C3D[1]	12.7-
Sigurdsson et al.[14]	12.8
I3D+super-events[15]	19.41
Ours	**20.3**

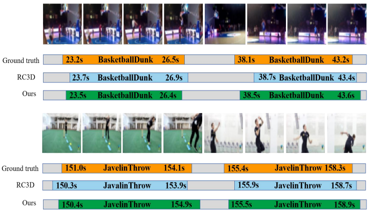
Ground truth
23.2s BasketballDunk 26.5s
38.1s BasketballDunk 43.2s
RC3D
23.7s BasketballDunk 26.9s
38.7s BasketballDunk 43.4s
Ours
23.5s BasketballDunk 26.4s
38.5s BasketballDunk 43.6s
Ground truth
151.0s JavelinThrow 154.1s
155.4s JavelinThrow 158.3s
RC3D
150.3s JavalinThrow 153.9s
155.9s JavelinThrow 158.7s
Ours
150.4s JavelinThrow 154.9s
155.5s JavelinThrow 158.9s

Fig. 2. Qualitative visualization on THUMOS'14.

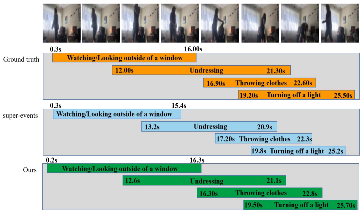
0.3s
16.00s
Ground truth
Watching/Looking outside of a window
12.00s Undressing 21.30s
16.90s Throwing clothes 22.60s
19.20s Turning off a light 25.50s
0.3s
15.4s
super-events
Watching/Looking outside of a window
13.2s Undressing 20.9s
17.20s Throwing clothes 22.3s
19.8s Turning off a light 25.2s
0.2s
16.3s
Ours
Watching/Looking outside of a window
12.6s Undressing 21.1s
16.30s Throwing clothes 22.8s
19.50s Turning off a light 25.70s

Fig. 3. Qualitative visualization on Charades.

Conclusions

We propose a temporal action detection method based on Temporal Deformable Proposal Generation (TDPG). In TDPG, Deformable Proposal Generation module generates deformable proposals with more precise boundaries by using temporal deformable convolution to appropriately expand receptive field. Our method has state-of-the-art performance on THUMOS14 and Charades, which shows that our method can more accurately locate the temporal boundary. Of course, there are still areas for improvement in our model. Different action classes with similar actions are prone to large errors. We lack consideration of background information. For future work, we plan to explore the use of background information to reduce the false detection rate and further improve the effect of model.

References

[1]H. Xu, A. Das and K. Saenko, "R-C3D: Region Convolutional 3D Network for Temporal Activity Detection," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 5794-5803.
[2]S. Ren, K. He, R. Girshick and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.
[3]J. Dai et al., "Deformable Convolutional Networks," 2017 IEEE International Conference on Computer Vision (ICCV), Venice, 2017, pp. 764-773.

Contact

Contact Person：Liting Yan
Tel：18852859520
E-mail：lityen@qq.com