Action Keyframe Connection Network for Temporal Action Proposal Generation

Shengbo Wang*, Zhenjiang Miao, Tianyu Zhou, Miaomiao Li and Ruyi Zhang

School of Computer and Information Technology, Beijing Jiaotong University

*17120342@bjtu.edu.cn

Abstract. Temporal action detection is an important research topic in computer vision, of which Temporal Action Proposal (TAP) generation is a key step for finding candidate action segments. Our paper provides an action proposal generation network for temporally untrimmed videos in which a new effective and efficient deep architecture named action keyframe connection network for temporal action proposal Generation. Firstly, a two-stream network is adopted to extract frame-level features which included appearance feature and optical flow feature. The temporal information helps the subsequent network to determine whether a frame is the beginning or the ending of the action. Secondly, a position discrimination network is designed to infer the probability of each frame being starting frame or ending frame. The network outputs a starting probability sequence and an ending probability sequence which indicates the start of the action and the end of the action respectively. Finally, our network generates a proposal by a specific threshold rule combining the points in the starting probability sequence and the ending probability sequence. We carry out experiments on ActivityNet dataset to compare our proposed method with the state-of-the-art methods. Experiment results show that our method achieves superior performance over other methods.

1. Introduction

The task of temporal action detection aims to temporal action detection is a challenging task as the complex action patterns such as, the number of actions in different videos might be varied and the duration of different actions in the video is different, and some of the action duration may be much shorter than the length of the whole video.

Recent studies have shown that temporal action proposal [1, 2, 3] is an important step in the temporal action detection. Accurate temporal action proposal can improve the performance of the next recognition step. However, there are still problems in some temporal action proposal methods: (1) Temporal information may not be considered during the proposal generation phase, so that the positioning is inaccurate. (2) Some exhaustive search based proposal generation methods can’t solve the problem of a large amount of computation.

In this paper, we propose an action keyframe connection based framework to solve the above two problems. Long Short-Term Memory (LSTM) in conjunction with the Convolutional Neural Network (CNN) are used to generate three probability sequences (the probability sequence of the action start frame, the probability sequence of the action end frame, and the probability sequence of the action occurrence frame), in which the temporal information is considered. In the proposal generation phase, we generate proposals by combining the detected action start frames and the action end frames instead of exhaustively searching for sliding windows [4,5] to reduce the computational cost. We adopt a two-
stream network to extract frame-level features [6] (Figure 1). And then a position discrimination network is designed to infer the probability of each frame belongs to start frame or ending frame. The network outputs a start probability sequence and an ending probability sequence. Finally, our network is used to generate proposals through specific frame select rules described in Section 3.3.

![Figure 1. Two-stream architecture for feature extraction](image)

2. RELATED WORK

The procedure of most of recent temporal action detection methods for includes two steps: proposal generation, recognition. Temporal action proposal is the first stage of temporal action detection, an overview of each task is provided below.

2.1. Temporal Action Proposal Generation

At present, more and more people are studying the task of temporal action detection, and this task can be completed in the form of ‘proposal’ + ‘recognition’. However, the second phase of temporal action detection has reached a relatively high level. This makes the task of the first phase of the temporal action proposal have a relatively large room for improvement. Among some existing methods, some are grouping video, and feature extraction is performed in units of video groups [7]. [8] proposes a method that proposals of the video can be obtained by training multiple multi-class classifiers.

2.2. Temporal Action Detection

In recent years, due to the rapid increase in the amount of video on the network and the rapid development of neural networks, this task has received more attention. The main datasets for this task are THUMOS2014 and ActivityNet [9]. The current trend is to find the relevance of the actions within the video to more accurately locate and find ways to replace the optical flow to speed up the model. One limitation of the traditional approach is the amount of data. There are many models that can reduce the amount of computation. For example, SSN [10] is a structured segmentation network, proposes a new framework for model the temporal structure of each action instance through a structured temporal pyramid and the TAG method is used to generate proposal. Similar networks include R-C3D [11], TAG [12].

3. Methods

In this section, we will introduce our network composition. The framework of the proposed method is shown in Figure 2. Our network consists of feature extraction network, position discrimination network, and post-processing component.
3.1. Feature Extraction Network

A complete video can be expressed as $V = \{v_n\}_{n=1}^l$ with $l$ frames. $v_n$ represents the $n$-th frame of $V$. To extract features, we first need to slice the video $V$ into segments $P = \{p_n\}_{n=1}^l$, where $l$ is the number of the segments. A segment $p_n = (v_{tn}, o_{tn})$ consists of two parts, one is $v_{tn}$ representing the $t$-th frame RGB image in video $V$, and the other $o_{tn}$ is an optical flow field centered on $v_{tn}$, as shown in Figure 1. In order to reduce the cost of calculation, we set the segment to a fixed length $\theta$, i.e. $l = l/\theta$. Given a segment, we combine the outputs of the spatial network and the temporal network to generate the feature vector $f_{tn} = (f_1, f_2)$, where $f_1, f_2$ are the output scores of the spatial network and the temporal networks respectively. That is to say, given a sequence $P$, we can extract a feature sequence $F = \{f_{tn}\}_{n=1}^l$. These two-stream feature sequences are used as inputs to the next network.

3.2. Position Discrimination Network

The Position Discrimination Network uses the CNN in parallel with the LSTM to evaluate the probability of the start, the end, and the occurrence of an action. Combining the results of the two sub-networks, an action start probability sequence, an action end probability sequence, and an action occurrence probability sequence are generated. The principle of this network can be summarized as action keyframe connection network.

We use $s_1$ and $s_2$ to represent the output probability sequences of CNN and LSTM respectively (Take the start frame probability sequence as an example).

We can know that the output of CNN is:
\[ s_1 = \text{conv}(W_0, X) \]  
(1)

\( W_0 \) is the convolution kernel matrix and \( X \) is the input feature sequence.

We can know that the output of LSTM is:

\[ s_0 = \text{conv}(W_1, X) \]  
(2)

\[ s_2 = \text{LSTM}(W_2, s_0) \]  
(3)

Where \( W \) represents the weight matrix, such as \( W_1 \) is the convolution kernel matrix and \( X \) is the input feature sequence. Where \( s_0 \) represents the convolutional layer output.

\[ s_3 = s_1 + \lambda s_2 \]  
(4)

3.3. Post-processing Network

The post-processing network task filters the start probability position and the end probability position through a certain rule on the probability sequence. If the starting probability position and the ending probability position are respectively greater than a certain threshold, the direct pairwise fusion start probability position and the end probability position respectively form candidate proposals, and the proposal-based features are sampled in the candidate proposals. Using the perceptual machine with a hidden layer as the input, the confidence of the candidate nomination is output. The results are non-maximally suppressed to remove overlapping results. Specifically, through the comparison of the most trusted proposal and other proposals, the Soft-NMS function is used to remove redundant other proposals.

4. Experiments

We evaluate the effectiveness of our temporal action proposal generation method by conducting experiments on real action dataset and comparing with the state-of-the-art.

4.1. Dataset

The Temporal Action Proposals task uses the ActivityNet dataset [9]. This is currently the largest database, which also includes two tasks for classification and detection. The dataset contains more than 648 hours of untrimmed video, with around 20,000 (training + verification + test set) videos. The train set accounts for about 50% of the entire dataset, the validation set accounts for 25% of the entire dataset, and the test set accounts for 25% of the entire dataset.

4.2. Evaluation Metrics.

This task uses the Average Recall and Average Number of Proposals per Video (AR – AN) curves as evaluation metrics. A proposal is true: the temporal intersection over union of the temporal interval of the proposal and the temporal interval of the real value is greater than or equal to a threshold (for example: tIoU > 0.5). AR is defined as the average of the recall values (satisfying 0.5 < tIoU ≤ 0.9, step size 0.05). AN is defined as the total number of proposals divided by the number of videos in the test subset. When calculating the value on the AR – AN curve, we consider the value of AN between 1 and 100 (inclusive) by step 1.

4.3. Comparisons with the state-of-the-art

We compare the network model with the proposal generation stages of Zhao et al. [13], Dai et al. [14], Yao et al. [15], Lin et al. [16]. And Lin et al. [17]. The results are shown in Table 1.

| Method          | Zhao et al. [13] | Dai et al. [14] | Yao et al. [15] | Lin et al. [16] | Lin et al. [17] | OURS  |
|-----------------|------------------|-----------------|-----------------|-----------------|-----------------|-------|
| AR@100          | 63.52            | -               | -               | 73.01           | 74.16           | 74.83 |
| AUC (test)      | -                | 61.56           | 64.18           | 64.80           | 66.26           | 66.57 |
Through the results in Table 1 we can know that our network is better than other networks. Because our network uses LSTM to more effectively consider the information on the temporal, our network results are better than other networks.

![Figure 3: Experimental diagram](image)

A video as input, our network task is to generate a proposal that occurs in the action. we compare various temporal action proposal generation networks through AR-AN indicators. As shown in Figure 3.

4.4. Parameter evaluation

When $\lambda$ is different, we can know that the CNN network has a different proportion than the network with LSTM, the temporal information added into the network is also changed. Table 2 is the result of the parameters.

| $\lambda$ | 1/6 | 1/7 | 1/8 | 1/9 | 1/10 |
|---|---|---|---|---|---|
| AUC (test) | 66.34 | 66.47 | 66.57 | 66.41 | 66.5 |

From this we can see that both of the spatial and temporal information contribute to the detection performance. In our experiments we get the best result when $\lambda = 1/8$. So we chose $\lambda = 1/8$ in our experiment.

5. Conclusion

Through our research on the temporary action proposal generation and the temporal action localization task, we mainly have the following achievements: 1. The quality of the proposal has a great influence on the subsequent action localization. Our method improves the proposal quality by an Action Keyframe Connection Network to improve the action localization performance. 2. Our network has three advantages: (1) flexible duration (2) accurate boundaries (3) reliable confidence scores.

References

[1] Z. Shou, D. Wang, and S. Chang. Temporal action localization in untrimmed videos via multi-stage CNNs. In: CVPR, 2016. 1, 2, 3, 4, 5, 6, 7, 8

[2] Heilbron F C, Niebles J C, Ghanem B. Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos[C]// Computer Vision & Pattern Recognition. IEEE, 2016.
[3] V. Escorcia, F. Caba, J.C. Niebles, and B. Ghanem. DAPs: Deep action proposals for action understanding. In: ECCV, 2016.1,2,3,4,5,6,7
[4] Gaidon, A., Harchaoui, Z., Schmid, C.: Temporal localization of actions with actoms. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(11), 2782–2795 (2013)
[5] Yuan, J., Ni, B., Yang, X., Kassim, A.A.: Temporal action localization with pyramid of score distribution features. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 3093–3102. IEEE, Las Vegas (2016)
[6] Simonyan K, Zisserman A. Two-Stream Convolutional Networks for Action Recognition in Videos[J]. 2014.
[7] Jiyang Gao: TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In: ICCV, 2017
[8] Shyamal Buch, Victor Escorcia, Chuanqi Shen: SST: Single-Stream Temporal Action Proposals. In: CVPR, 2017
[9] Caba Heilbron, F., Escorcia, V., Ghanem, B., Carlos Niebles, J.: Activitynet: A large-scale video benchmark for human activity understanding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2015) 961–970
[10] Yue Zhao, Yuanjun Xiong: Temporal Action Detection with Structured Segment Networks. In: ICCV, 2017.
[11] Huijuan Xu, Abir Das: R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In: ICCV, 2017.
[12] Yuanjun Xiong, Yue Zhao: A Pursuit of Temporal Accuracy in General Activity Detection. In: arXiv: 1703.02716
[13] Zhao, Y., iong, Y., Wang, L., Wu, Z., Lin, D., Tang, X.: Temporal action detection with structured segment networks. arXiv preprint arXiv:1704.06228 (2017)
[14] Dai, X., Singh, B., Zhang, G., Davis, L.S., Chen, Y.Q.: Temporal context network for activity localization in videos. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE (2017) 5727–5736
[15] Ghanem, B., Niebles, J.C., Snoek, C., Heilbron, F.C., Alwassel, H., Krisna, R., Escorcia, V., Hata, K., Buch, S.: ActivityNet challenge 2017 summary. arXiv preprint arXiv:1710.08011 (2017)
[16] Lin, T., Zhao, X., Shou, Z.: Temporal convolution based action proposal: Submission to ActivityNet 2017. arXiv preprint arXiv:1707.06750 (2017)
[17] Lin T, Zhao X, Su H, et al. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation[J]. 2018.