Non-local Neural Networks

Cheng Guan

August 19, 2018

1. Approach

1.1. Problem Definition

An untrimmed video sequence can be denoted as $X = \{x_n\}_{n=1}$ with l_v frames, where x_n is the n-th frame in X. Annotation of video X is composed by a set of action instances $\psi_g = \{\varphi_n = (t_{s,n}, t_{e,n})\}_{n=1}^{N_g}$, where N_g is the number of truth action instances in video X, and $t_{s,n}, t_{e,n}$ are starting and ending time of action instance φ_n separately. Unlike detection task, classes of action instances are not considered in temporal action proposal generation.

1.2. Video Features Encoding

To generate proposals of input video, first the authors need to extract feature to encode visual content of video. In their framework, they adopt two-stream network [2] as visual en- coder, since this architecture has shown great performance in action recognition task [3] and has been widely adopted in temporal action detection and proposal generation tasks [4, 1]. Two-stream network contains two branches: spatial network operates on single RGB frame to capture appearance feature, and temporal network operates on stacked optical flow field to capture motion information.

To extract two-stream features, as shown in Fig. 1, first they compose a snippets l_s sequence $S = \{s_n\}_{n=1}^{l_s}$ from video X, where l_s is the length of snippets sequence. A snippet $s_n = (x_{t_n}, o_{t_n})$ includes two parts: x_{t_n} is the t_n-th RGB frame in X and o_{t_n} is stacked optical flow field derived around center frame x_{t_n}.

2. Boundary-Sensitive Network

To achieve high proposal quality with both precise temporal boundaries and reliable confidence scores, the authors adopt "*local to global*" fashion to generate proposals. In BSN, they first generate candidate boundary locations, then combine these locations as proposals and evaluate confidence score of each proposal with proposal-level feature.

Network Architecture. The architecture of BSN is presented in Fig.1, which contains three modules: temporal evaluation, proposal generation and proposal evaluation. *Temporal evaluation module* is a three layers temporal convolutional neural network, which takes the two-stream feature sequences as input, and evaluates probabilities of each temporal location in video whether it is inside or outside,

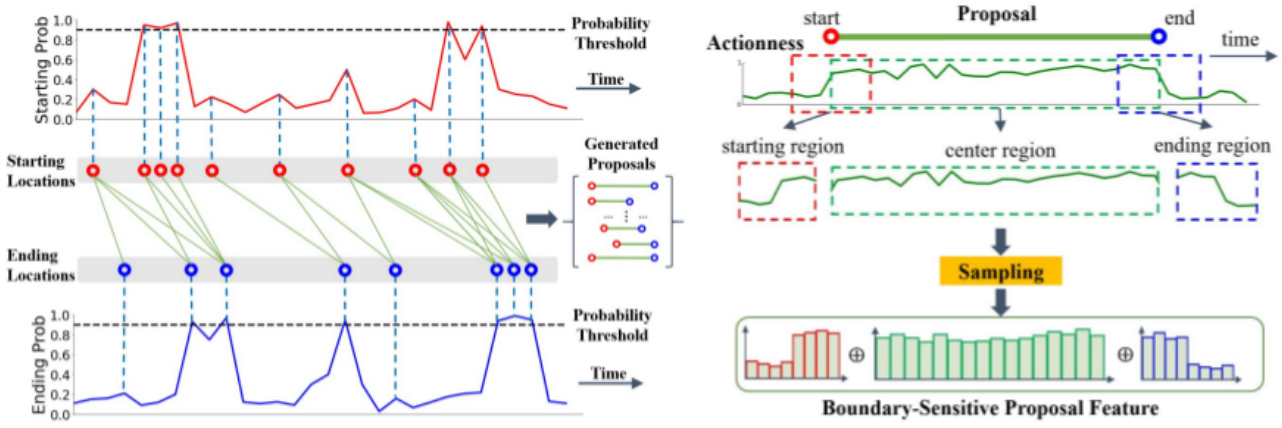

Figure 1. Details of proposal generation module. First, to generate candidate boundary locations, the authors choose temporal locations with high boundary prob- ability or being a probability peak. Then, they combine candidate starting and ending locations as proposals when their duration satisfying condition. Construct BSP feature. Given a proposal and actionness probabilities sequence, they can sample actionness sequence in starting, center and ending regions of proposal to construct BSP feature.

at or not at boundaries of ground truth action instances, to generate sequences of starting, ending and actionness probabilities respectively.

Proposal generation module. The goal of proposal generation module is to generate candidate proposals and construct corresponding proposal-level feature. The authors achieve this goal in two steps. First they locate temporal locations with high boundary probabilities, and combine these locations to form proposals.

Proposal evaluation module. The goal of proposal evaluation module is to evaluate the confidence score of each proposal whether it contains an action instance within its duration using BSP feature. Hidden layer with 512 units handles the input of BSP feature f_{BSP} with Relu activation. The output layer outputs confidence score p_{conf} with sigmoid activation,

References

[1] J. Gao, Z. Yang, and R. Nevatia. Cascaded boundary regression for temporal action detection. *arXiv preprint arXiv:1705.01180*, 2017. 1

[2] K. Simonyan and A. Zisserman. Two-stream convolutional networks for action recognition in videos. In *NIPS*, 2014. 1

[3] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*, 2016. 1

[4] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin. Temporal action detection with structured segment networks. *ICCV*, 2017. 1