Trimmed Action Recognition, Dense-Captioning Events in Videos, and Spatio-temporal Action Localization with Focus on ActivityNet Challenge 2019

Zhaofan Qiu, Dong Li, Yehao Li, Qi Cai, Yingwei Pan, and Ting Yao
JD AI Reseach, Beijing, China
panyw.ustc@gmail.com

Abstract

This notebook paper presents an overview and comparative analysis of our systems designed for the following three tasks in ActivityNet Challenge 2019: trimmed action recognition, dense-captioning events in videos, and spatio-temporal action localization.

***Trimmed Action Recognition (Kinetics)**: We investigate and exploit multiple spatio-temporal clues for trimmed action recognition task, i.e., frame, short video clip and motion (optical flow) by leveraging 2D or 3D convolutional neural networks (CNNs). The mechanism of different quantization methods is studied as well. All activities are finally classified by late fusing the predictions from each clue.*

***Dense-Captioning Events in Videos (ActivityNet Captions)**: For this task, we firstly adopt a standard "detection by classification" framework to localize temporal proposals of interest in video, and then generate the descriptions for each proposal. Specifically, a two-layer LSTM-based captioning architecture with temporal attention mechanism is leveraged to generate sentence conditioning on the input video representation and its detected attributes. Moreover, the captioning architecture is equipped with policy gradient optimization scheme to further boost video captioning.*

***Spatio-temporal Action Localization (AVA)**: We present a new Long Short-Term Relation Networks (LSTR), which models both short-term and long-term human-context relation to augment features for spatio-temporal action localization. Technically, Region Proposal Network (RPN) is employed to first generate bounding box proposals on the keyframe of each video clip. LSTR then models short-term human-context interactions within each clip through spatio-temporal attention mechanism and reasons long-term temporal dynamics across video clips via Graph Convolutional Networks (GCN) in a cascaded manner. The upgraded relation-aware feature of each proposal is finally employed for classifying actions.*

1. Introduction

Recognizing activities in videos is a challenging task as video is an information-intensive media with complex variations. In particular, an activity may be represented by different clues including frame, short video clip, motion (optical flow) and long video clip. In this work, we aim at investigating these multiple clues to activity classification in trimmed videos, which consist of a diverse range of human focused actions. Moreover, action detection with accurate spatio-temporal location in videos, i.e., spatio-temporal action localization, is another challenging task in video understanding and we study this task in this work. Compared to temporal action localization which temporally localizes actions, this task is more difficult due to the complex variations and large spatio-temporal search space. In addition to the above two tasks tailored to activity which is usually the name of action/event in videos, the task of dense-captioning events in videos is explored here which goes beyond activities by describing numerous events within untrimmed videos with multiple natural sentences.

The remaining sections are organized as follows. Section 2 presents all the features which will be adopted in our systems, while Section 3 details the feature quantization strategies. Then the descriptions and empirical evaluations of our systems for three tasks are provided in Section 4-6 respectively, followed by the conclusions in Section 7.

2. Video Representations

We extract the video representations from multiple clues including frame, motion and audio.

Frame. To extract appearance-based representations from video, we devise the novel Pseudo-3D Residual Net [23] with Local and Global Diffusion [24] (LGD-P3D) architecture, as shown in Figure 1. The Local and Global Diffusion (LGD) is a novel neural network architecture that learns the local and global representations in parallel. The architecture is composed of LGD blocks, where each block updates local and global features by modeling the diffusions between these two representations. Diffusions effec-

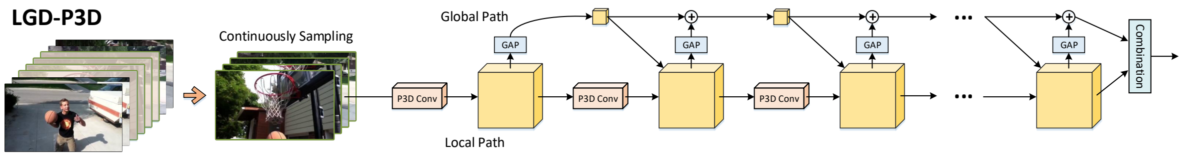
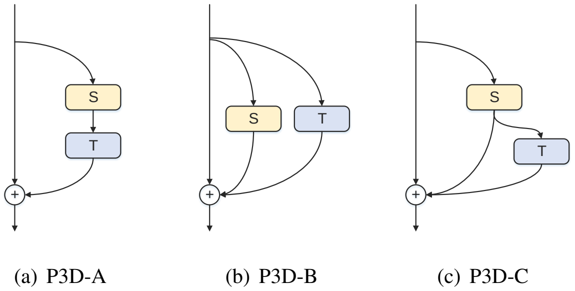


Figure 1. Network architecture of LGD-P3D. The LGD framework is proposed in [24] and the basic P3D operation is proposed in [23].



(a) P3D-A (b) P3D-B (c) P3D-C

Figure 2. Three Pseudo-3D blocks.

tively interact two aspects of information, i.e., localized and holistic, for more powerful way of representation learning. The basic operations in LGD-P3D are variants of bottleneck building blocks to combine 2D spatial and 1D temporal convolutions, as shown in Figure 2. The backbone of LGD-P3D is either ResNet-101 [5] or Xception [3]. We sample 16 consecutive frames as a short clip and fix the sample rate as 2 clips per second.

Motion. To model the change of consecutive frames, we apply another CNNs to optical flow "image," which can extract motion features between consecutive frames. When extracting motion features, we follow the setting of [24], which fed a 16-frame optical flow image sequence, consisting of two-direction optical flow from multiple consecutive frames, into LGD-P3D network in each iteration. The sample rate is also set to 2 clips per second.

Audio. Audio feature is the most global feature (though entire video) in our system. Although audio feature itself can not get very good result for action recognition, but it can be seen as powerful additional feature, since some specific actions are highly related to audio information. Here we utilize Xception network to extract audio feature from the audio spectrum map.

3. Feature Quantization

In this section, we describe two quantization methods to generate video-level representations from the extracted features.

Average Pooling (AP). Average pooling is the most common method to extract video-level features. For a set of clip-level features $F = \{f_1, f_2, ..., f_N\}$, the video-level representations are produced by simply averaging all the features in the set:

$$R_{AP} = \frac{1}{N} \sum_{i:f_i \in F} f_i \quad , \tag{1}$$

where R_{AP} denotes the final representations.

Temporal Convolutional Pooling (TCP). Moreover, we utilize a novel temporal convolutional pooling to produce highly discriminative video-level representation by modeling the feature sequence with stacked 1D temporal convolutions. The video-level representations are given by:

$$R_{TCP} = Conv1D(\{f_1, f_2, ..., f_N\}) \quad , \tag{2}$$

Here we devise a novel Conv1D network with 5 stacked depth-wise residual blocks for TCP.

4. Trimmed Action Recognition

4.1. System

Our trimmed action recognition framework is shown in Figure 3 (a). In general, the trimmed action recognition process is composed of three stages, i.e., multi-stream feature extraction, feature quantization and prediction generation. For deep feature extraction, we follow the multi-stream approaches in [11, 20, 21, 22], which represented input video by a hierarchical structure including short clip, optical flow images. In addition to visual features, the most commonly used audio spectrum is exploited to further enrich the video representations. After extraction of raw features, different quantization and pooling methods are utilized on different features to produce global representations of each trimmed video. Finally, the predictions from different streams are linearly fused by the weights tuned on validation set.

4.2. Experiment Results

Table 1 shows the performances of all the components in our trimmed action recognition system. Overall, the TCP on LGD-P3D (Xception 16-frame) achieves the highest top1 accuracy (70.63%) and top5 accuracy (89.32%) of single component. And by additionally apply this model on both frame and optical flow, the two-stream LGD-P3D (Xception, 16-frame&flow) achieves an obvious improvement, which gets top1 accuracy of 72.82% and top5 accuracy of 90.73%. For the final submission, we linearly fuse all the components.

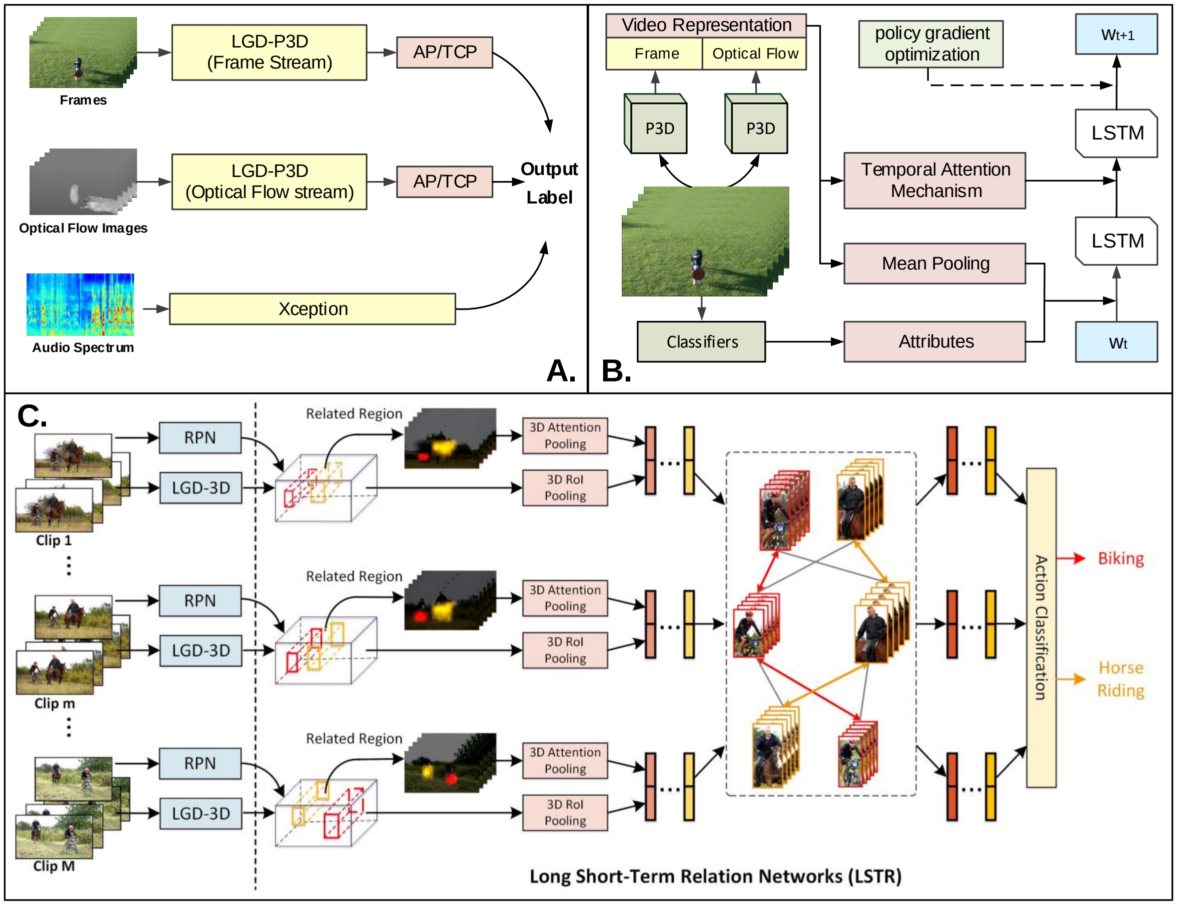


Figure 3. Frameworks of our proposed (a) trimmed action recognition system, (b) dense-captioning events in videos system, and (c) spatio-temporal action localization system.

5. Dense-Captioning Events in Videos

5.1. System

The main goal of dense-captioning events in videos is jointly localizing temporal proposals of interest in videos and then generating the descriptions for each proposal/video clip. Hence we firstly leverage a standard "detection by classification" in [29] to localize temporal proposals of events in videos (5 proposals for each video). Then, given each temporal proposal (i.e., video segment describing one event), our dense-captioning system capitalizes on a two-layer LSTM-based captioning architecture with temporal attention mechanism for sentence generation. Specifically, the generative module with LSTM is inspired from the recent successes of probabilistic sequence models leveraged in vision and language tasks (e.g., image captioning [13, 28, 30, 31, 32], video captioning [16, 17, 19], video generation from captions [18] and dense video captioning [12, 29]). We mainly utilize the two-layer LSTM-based captioning architecture in [1] and extend the original spatial attention at region level into temporal attention at frame level. To be specific, the first-layer LSTM collects the maximum contextual information by concatenating each input word with the previous output of second-layer LSTM, the mean-pooled video representation, and attribute representation. Next, conditioning on the output hidden state of the first-layer LSTM, a normalized temporal attention distribution over all frames is measured to dynamically fuse all frame features into attended video-level representation, which will be set as the input of the second-layer LSTM. Note that we employ the policy gradient optimization method with reinforcement learning [26] to further boost the video captioning performances specific to METEOR metric. The overall architecture of our dense-captioning system is shown in Figure 3 (b).

5.2. Experiment Results

Table 2 shows the performances of our proposed dense-captioning events in videos system. In particular, by additionally incorporating the policy gradient optimization scheme into our system, we can clearly observe the performance boost in METEOR.

Table 1. Comparison of different components in our trimmed action recognition framework on Kinetics validation set for trimmed action recognition task.

Stream	Feature	Quantization	Top1	Top5
Frame	LGD-P3D (Xception, 16-frame)	AP	67.51%	86.93%
	LGD-P3D (Xception, 16-frame)	TCP	70.63%	89.32%
	LGD-P3D (Xception, 128-frame)	AP	69.84%	88.33%
	LGD-P3D (ResNet-101, 128-frame)	AP	69.75%	88.80%
Motion	LGD-P3D (Xception, 16-flow)	AP	54.40%	77.28%
	LGD-P3D (Xception, 16-flow)	TCP	60.51%	82.4%
	LGD-P3D (Xception, 128-flow)	AP	61.33%	83.12%
	LGD-P3D (ResNet-101, 128-flow)	AP	64.49%	85.50%
Audio	Xception	AP	21.91%	36.86%
	Xception	TCP	21.76%	36.93%
Two-stream	LGD-P3D (Xception, 16-frame&flow)	AP	69.39%	88.08%
	LGD-P3D (Xception, 16-frame&flow)	TCP	72.82%	90.73%
	LGD-P3D (Xception, 128-frame&flow)	AP	71.77%	89.75%
	LGD-P3D (ResNet-101, 128-frame&flow)	AP	72.32%	90.46%
Two-stream+Audio	LGD-P3D (Xception, 16-frame&flow)	AP	70.94%	88.81%
	LGD-P3D (Xception, 16-frame&flow)	TCP	74.82%	91.78%
	LGD-P3D (Xception, 128-frame&flow)	AP	73.90%	90.91%
	LGD-P3D (ResNet-101, 128-frame&flow)	AP	74.19%	91.41%
Ensemble			76.37%	92.78%

Table 2. Performance on ActivityNet captions validation and testing set. All values are reported over METEOR metric (%).

Model	Val	Test
Ours	9.81	-
Ours + policy gradient	10.30	8.49

6. Spatio-temporal Action Localization

6.1. System

Figure 3 (c) shows the framework of spatio-temporal action localization, which includes two main components:

Person Detector. We use Faster R-CNN [25] with a Deformable ResNet-101 [4] backbone for person detection. The model is pre-trained on ImageNet [6] and COCO [14], and then fine-tuned on AVA bounding boxes. The final model obtains 93.5 AP@50 on the AVA validation set.

Long Short-Term Relation Networks (LSTR). LSTR takes 8 consecutive 16-frame clips as input and employs LGD-3D ResNet-101 [24] as backbone, which is initialized with Kinetics-600 [7] pre-trained model. We feed each clip to the backbone and extract the clip feature representation at the last convolutional layer. For each actor proposal, we crop and resize the clip feature within the proposal using 3D RoI Pooling to obtain a fixed-length actor representation. However, this representation ignores the short-term relation within clip representing the interactions between actors and their surroundings (including other actors, objects, and scenes). We devise a spatio-temporal attention module to model and incorporate such information into proposal representation, as illustrated in Figure 3 (d). We exploit adaptive convolution to dynamically predict the actor-specific spatio-temporal attention map, which indicates the relevance degree of the global context to this actor. The context feature is then generated through 3D Attention Pooling on the attention map. The final proposal representation is obtained by concatenating the actor feature and context feature together. In addition to the short-term relation between actors and context within each clip, we also expect to further capitalize on long-range dependencies between correlated proposals from neighboring clips. To achieve this, we build a relation graph with undirected edges on human proposals extracted from all video clips. The vertex represents each human proposal and the edge denotes the relation measured on both visual similarity and geometrical overlap in between. Graph Convolutional Networks (GCN) [8] are utilized to enrich the feature of human proposal by propagating the relation in the graph. The upgraded relation-aware feature of each proposal is finally exploited for action classification.

6.2. Experiment Results

Following [9, 10, 15, 23], we also exploit a two-stream pipeline for utilizing multiple modalities, where the RGB frame and the stacked optical flow image are considered. To fuse the detection results, late fusion scheme is taken to average the classification scores. Table 3 shows the performances of all the components in our LSTR. For the final submission, all the components are linearly fused using the weights tuned on validation set. The final mAP on valida-

Table 3. Comparison of different components in our LSTR on AVA validation/test set for spatio-temporal action localization task.

Stream	**Val**	**Test**
RGB	28.3	27.3
Flow	22.7	-
Two Stream	29.4	-
Two Stream (multi-scale)	30.5	29.1

tion set is 30.5%.

7. Conclusion

In ActivityNet Challenge 2019, we mainly focused on multiple visual features, different strategies of feature quantization and video captioning from different dimensions. Our future works include more in-depth studies of how fusion weights of different clues could be determined to boost the action recognition and spatio-temporal action localization performance. For dense-captioning events in videos task, we are targeting at making use of non-autoregressive encodeing/decoding [2, 27] for sentence generation.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*, 2018.

[2] Jingwen Chen, Yingwei Pan, Yehao Li, Ting Yao, Hongyang Chao, and Tao Mei. Temporal deformable convolutional encoder-decoder networks for video captioning. In *AAAI*, 2019.

[3] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. In *CVPR*, 2017.

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *CVPR*, 2017.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[6] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[7] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[8] Thomas N Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*, 2016.

[9] Dong Li, Zhaofan Qiu, Qi Dai, Ting Yao, and Tao Mei. Recurrent tubelet proposal and recognition networks for action detection. In *ECCV*, 2018.

[10] Dong Li, Ting Yao, Ling-Yu Duan, Tao Mei, and Yong Rui. Unified spatio-temporal attention networks for action recognition in videos. *IEEE Transactions on Multimedia*, 2018.

[11] Qing Li, Zhaofan Qiu, Ting Yao, Tao Mei, Yong Rui, and Jiebo Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In *ICMR*, 2016.

[12] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018.

[13] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. Pointing novel objects in image captioning. In *CVPR*, 2019.

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014.

[15] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for actionlocalization. In *CVPR*, 2019.

[16] Yingwei Pan, Tao Mei, Ting Yao, Houqiang Li, and Yong Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[17] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. Seeing bot. In *SIGIR*, 2017.

[18] Yingwei Pan, Zhaofan Qiu, Ting Yao, Houqiang Li, and Tao Mei. To create what you tell: Generating videos from captions. In *MM Brave New Idea*, 2017.

[19] Yingwei Pan, Ting Yao, Houqiang Li, and Tao Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

[20] Zhaofan Qiu, Dong Li, Chuang Gan, Ting Yao, Tao Mei, and Yong Rui. Msr asia msm at activitynet challenge 2016. In *CVPR workshop*, 2016.

[21] Zhaofan Qiu, Qing Li, Ting Yao, Tao Mei, and Yong Rui. Msr asia msm at thumos challenge 2015. In *THUMOS'15 Action Recognition Challenge*, 2015.

[22] Zhaofan Qiu, Ting Yao, and Tao Mei. Deep quantization: Encoding convolutional activations with deep generative model. In *CVPR*, 2017.

[23] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[24] Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *CVPR*, 2019.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[26] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[29] Ting Yao, Yehao Li, Zhaofan Qiu, Fuchen Long, Yingwei Pan, Dong Li, and Tao Mei. Msr asia msm at activitynet challenge 2017: Trimmed action recognition, temporal action proposals and dense-captioning events in videos. In *CVPR ActivityNet Challenge Workshop*, 2017.

[30] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

[31] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[32] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.