YH Technologies at ActivityNet Challenge 2018

Ting Yao and Xue Li
YH Technologies Co., Ltd, Beijing, China
{tingyao.ustc, miya.lixue}@gmail.com

Abstract

This notebook paper presents an overview and comparative analysis of our systems designed for the following five tasks in ActivityNet Challenge 2018: temporal action proposals, temporal action localization, dense-captioning events in videos, trimmed action recognition, and spatio-temporal action localization.

Temporal Action Proposals (TAP): *To generate temporal action proposals from videos, a three-stage workflow is particularly devised for TAP task: a coarse proposal network (CPN) to generate long action proposals, a temporal convolutional anchor network (CAN) to localize finer proposals, and a proposal reranking network (PRN) to further identify proposals from previous stages. Specifically, CPN explores three complementary actionness curves (namely point-wise, pair-wise, and recurrent curves) that represent actions at different levels to generate coarse proposals, while CAN refines these proposals by a multi-scale cascaded 1D-convolutional anchor network.*

Temporal Action Localization (TAL): *For TAL task, we follow the standard "detection by classification" framework, i.e., first generate proposals by our temporal action proposal system and then classify proposals with two-stream P3D classifier.*

Dense-Captioning Events in Videos (DCEV): *For DCEV task, we firstly adopt our temporal action proposal system mentioned above to localize temporal proposals of interest in video, and then generate the descriptions for each proposal. Specifically, RNNs encode a given video and its detected attributes into a fixed dimensional vector, and then decode it to the target output sentence. Moreover, we extend the attributes-based CNNs plus RNNs model with policy gradient optimization and retrieval mechanism to further boost video captioning performance.*

Trimmed Action Recognition (TAR): *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.*

Spatio-temporal Action Localization (SAL): *Our system for SAL includes two main components: i.e., Recurrent Tubelet Proposal (RTP) networks and Recurrent Tubelet Recognition (RTR) networks. The RTP initializes action proposals of the start frame through a Region Proposal Network on the feature map and then estimates the movements of proposals in the next frame in a recurrent manner. The action proposals of different frames are linked to form the tubelet proposals. The RTR capitalizes on a multi-channel architecture, where in each channel, a tubelet proposal is fed into a CNN plus LSTM network to recurrently recognize action in the tubelet.*

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.

Besides detecting actions in manually trimmed short video, researchers tend to develop techniques for detecting actions in untrimmed long videos in the wild. This trend motivates another challenging task—temporal action localization which aims to localize action in untrimmed long videos. We also explore this task in this work. However, most of the natural videos in the real world are untrimmed videos with complex activities and unrelated background/context information, making it hard to directly localize and recognize activities in them. One possible solution is to quickly localize temporal chunks in untrimmed videos containing human activities of interest and then conduct activity recognition over these temporal chunks, which largely simplifies the activity recognition for untrimmed videos. Generating such temporal action chunks in untrimmed videos is known as the task of temporal action proposals, which is also exploited here.
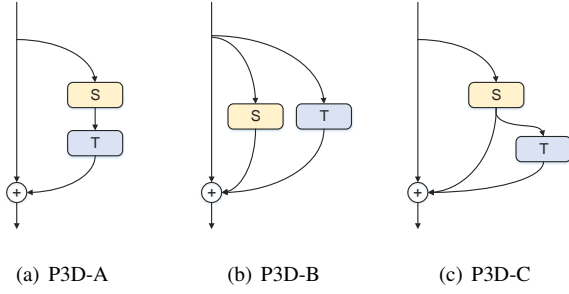
| (a) P3D-A | (b) P3D-B | (c) P3D-C |

Figure 1. Three Pseudo-3D blocks.

Furthermore, 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 four 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 five tasks are provided in Section 4-8 respectively, followed by the conclusions in Section 9.

2. Video Representations

We extract the video representations from multiple clues including frame, short clip, motion and long clip.

Frame. To extract frame-level representations from video, we uniformly sample 25 frames for each video/proposal, and then use pre-trained 2D CNNs as frame-level feature extractors. We choose the most popular 2D CNNs in image classification—ResNet [4].

Short Clip. In addition to frame, we take the inspiration from the most popular 3D CNN architecture C3D [20] and devise a novel Pseudo-3D Residual Net (P3D ResNet) architecture [16] to learn spatio-temporal video clip representation in deep networks. Particularly, we develop variants of bottleneck building blocks to combine 2D spatial and 1D temporal convolutions, as shown in Figure 1. The whole P3D ResNet is then constructed by integrating Pseudo-3D blocks into a residual learning framework at different placements. We fix the sample rate as 25 per video.

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 [22], which fed optical flow images, consisting of two-direction optical flow from multiple consecutive frames, into ResNet/P3D ResNet network in each iteration. The sample rate is also set to 25 per video.

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 MFCC to extract audio features.

3. Feature Quantization

In this section, we describe two quantization methods to generate video-level/clip-level representations.

Average Pooling. Average pooling is the most common method to extract video-level features from consecutive frames, short clips and long clips. For a set of frame-level or 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:

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

where $R_{pooling}$ denotes the final representations.

Compact Bilinear Pooling. Moreover, we utilize Compact Bilinear Pooling (CBP) [3] to produce highly discriminative clip-level representation by capturing the pairwise correlations and modeling interactions between spatial locations within this clip. In particular, given a clip-level feature $F_t \in \mathbb{R}^{W \times H \times D}$ (W, H and D are the width, height and channel numbers), Compact Bilinear Pooling is performed by kernelized feature comparison, which is defined as

$$R_{CBP} = \sum_{j=1}^{S} \sum_{k=1}^{S} \langle \phi(F_{t,j}), \phi(F_{t,k}) \rangle \ , \tag{2}$$

where $S = W \times H$ is the size of the feature map, $F_{t,j}$ is the region-level feature of j-th spatial location in F_t, $\phi(\cdot)$ is a low dimensional projection function, and $\langle \cdot \rangle$ is the second order polynomial kernel.

4. Trimmed Action Recognition

4.1. System

Our trimmed action recognition framework is shown in Figure 2 (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 [6, 13, 14, 15], which represented input video by a hierarchical structure including individual frame, short clip and consecutive frame. In addition to visual features,
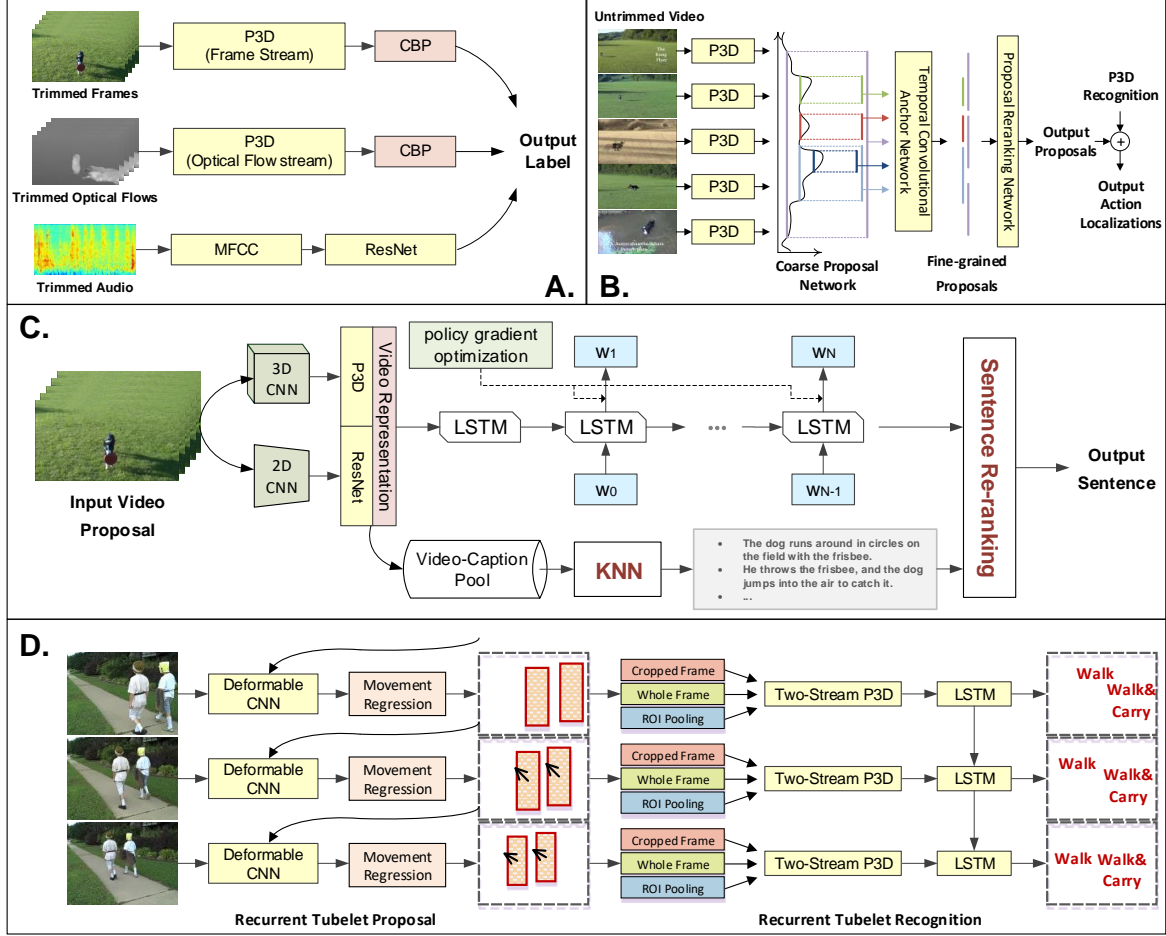
Figure 2. Frameworks of our proposed (a) trimmed action recognition system, (b) temporal action proposals system, (c) dense-captioning events in videos system, and (d) spatio-temporal action localization system.

the most commonly used audio feature MFCC 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 validatoin set.

4.2. Experiment Results

Table 1 shows the performances of all the components in our trimmed action recognition system. Overall, the CBP on P3D ResNet (128-frame) achieves the highest top1 accuracy (78.47%) and top5 accuracy (93.99%) of single component. And by additionally apply this model on both frame and optical flow, the two-stream P3D achieves an abvious improvement, which gets top1 accuracy of 80.91% and top5 accuracy of 94.96%. For the final submission, we linearly fuse all the components.

5. Temporal Action Proposals

5.1. System

Figure 2 (b) shows the framework of temporal action proposals, which is mainly composed of three stages:

Coarse Proposal Network (CPN). In this stage, proposal candidates are generated by watershed temporal actionness grouping algorithm (TAG) based on actionness curve. Considering the diversity of action proposals, three actionness measures (namely point-wise, pair-wise and recurrent) that are complementary to each other are leveraged to produce the final actionness curve.

Temporal Convolutional Anchor Network (CAN). Next, we feed long proposals into our temporal convolutional anchor network for finer proposal generation. The temporal convolutional anchor network consists of multiple 1D convolution layers to generate temporal instances for proposal/background binary classification and bounding box regression.

Proposal Reranking Network (PRN). Given the short

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

Stream	Feature	Layer	Quantization	Top1	Top5
Frame	ResNet	pool5	Ave	74.11%	91.51%
	ResNet	res5c	CBP	74.97%	91.48%
Short Clip	P3D ResNet (16-frame)	pool5	Ave	76.22%	92.92%
	P3D ResNet (128-frame)	pool5	Ave	77.94%	93.75%
	P3D ResNet (128-frame)	res5c	CBP	78.47%	93.99%
Motion	P3D ResNet (16-flow)	pool5	Ave	64.37%	85.76%
	P3D ResNet (128-flow)	pool5	Ave	69.87%	89.44%
	P3D ResNet (128-flow)	res5c	CBP	71.07%	90.00%
Audio	ResNet	pool5	Ave	21.91%	38.49%
Two-stream P3D	P3D ResNet (128-frame&flow)	res5c	CBP	80.91%	94.96%
Fusion all				83.75%	95.95%

Table 2. Area Under the average recall vs. average number of proposals per video Curve (AUC) of frame/flow input for P3D [16] network on ActivityNet validation set for temporal action proposals task.

Stream	CPN	CAN	PRN	AUC
Frame	√			60.27%
	√	√		63.20%
	√	√	√	64.21%
Optical Flow	√			59.83%
	√	√		63.43%
	√	√	√	64.02%
Fusion all				**67.36%**

Table 3. Performance comparison of different methods on ActivityNet validation set for temporal action localization task. Results are evaluated by mAP with different IoU thresholds and average mAP of IoU threshold from 0.5 to 0.95 with step 0.05.

mAP	0.5	0.75	0.95	Avg mAP
Shou et al. [19]	43.83	25.88	0.21	22.77
Xiong et al. [23]	39.12	23.48	5.49	23.98
Lin et al. [8]	48.99	32.91	7.87	32.26
Ours	51.40	33.61	8.13	34.22

proposals from the coarse stage and fine-grained proposals from the temporal convolutional anchor network, a r-eranking network is utilized for proposal refinement. To take video temporal structures into account, we extend the current part of proposal with its' start and end part. The duration of start and end parts are half of the current part. The proposal is then represented by concatenating features of each part to leverage the context information. In our experiments, the top 100 proposals are finally outputted.

5.2. Experiment Results

Table 2 shows the action proposal AUC performances of frame/optical flow input to P3D [16] with different stages in our system. The two stream P3D architecture is pre-trained on Kinetics [5] dataset. For all the single stream runs with different stages, the setting based on all three stages combination achieves the highest AUC. For the final submission, we combine all the proposals from the two streams and then select the top 100 proposals based on their weighted ranking probabilities. The linear fusion weights are tuned on validation set.

6. Temporal Action Localization

6.1. System

Without loss of generality, we follow the standard "detection by classification" framework, i.e., first generate proposals by temporal action proposals system and then classify proposals. The action classifier is trained with the above trimmed action recognition system (i.e., two-stream P3D) over the 200 categories on ActivityNet dataset [1].

6.2. Experiment Results

Table 3 shows the action localization mAP performance of our approach and baselines on validation set. Our approach consistently outperforms other state-of-the-art approaches in different IoU threshold and achieves 34.22% average mAP on validation set.

7. Dense-Captioning Events in Videos

7.1. 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 the temporal action proposal system described above in Section 5 to localize temporal proposals of events in videos (2 proposals for each video). Then, given each temporal proposal (i.e., video seg-

Table 4. Performance of our dense-captioning events in videos system on ActivityNet captions validation set, where B@N, M, R and C are short for BLEU@N, METEOR, ROUGE-L and CIDEr-D scores. All values are reported as percentage (%).

Model	B@1	B@2	B@3	B@4	M	R	C
LSTM-A$_3$	13.78	7.12	3.53	1.72	7.61	13.30	27.07
LSTM-A$_3$ + policy gradient	11.65	6.05	3.02	1.34	8.28	12.63	14.62
LSTM-A$_3$ + policy gradient + retrieval	11.91	6.13	3.04	1.35	8.30	12.65	15.61

ment describing one event), our dense-captioning system runs two different video captioning modules in parallel—the generative module for generating caption via the LSTM-based sequence learning model, and the retrieval module which can directly copy sentences from other visually similar video segments through KNN. 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 [21, 25], video captioning [9, 10, 12], video generation from captions [11] and dense video captioning [7, 24]). We mainly utilize the third design LSTM-A$_3$ in [26] which firstly encodes attribute representations into LSTM and then transforms video representations into LSTM at the second time step is adopted as the basic architecture. Note that we employ the policy gradient optimization method with reinforcement learning [18] to boost the video captioning performances specific to METEOR metric. For the retrieval module, we utilize KNN to find the visually similar video segments based on the extracted video representations. The captions associated with the top similar video segments are regarded as sentence candidates in retrieval module. In the experiment, we mainly choose the top 300 nearest neighbors for generating sentence candidates. Finally, a sentence re-ranking module is exploited to rank and select the final most consensus caption from the two parallel video captioning modules by considering the lexical similarity among all the sentence candidates. The overall architecture of our dense-captioning system is shown in Figure 2 (c).

7.2. Experiment Results

Table 4 shows the performances of our proposed dense-captioning events in videos system. Here we compare three variants derived from our proposed model. In particular, by additionally incorporating the policy gradient optimization scheme into the basic LSTM-A$_3$ architecture, we can clearly observe the performance boost in METEOR. Moreover, our dense-captioning model (LSTM-A$_3$ + policy gradient + retrieval) is further improved by injecting the sentence candidates from retrieval module in METEOR.

8. Spatio-temporal Action Localization

8.1. System

Figure 2 (d) shows the framework of spatio-temporal action localization, which includes two main components:

Table 5. Comparison of different components in our RTR on AVA validation set for spatio-temporal action localization task.

Stream	Feature	mAP@IoU=0.5
Frame	ResNet	13.68
Short Clip	P3D ResNet (16-frame)	19.12
Short Clip	P3D ResNet (128-frame)	19.40
Flow	P3D ResNet (16-frame)	15.20
Fusion	-	**22.20**

Recurrent Tubelet Proposal (RTP) networks. The Recurrent Tubelet Proposal networks produces action proposals in a recurrent manner. Specifically, it initializes action proposals of the start frame through a Region Proposal Network (RPN) [17] on the feature map. Then the movement of each proposal in the next frame is estimated from three inputs: feature maps of both current and next frames, and the proposal in current frame. Simultaneously, a sibling proposal classifier is utilized to infer the actionness of the proposal. To form the tubelet proposals, action proposals in two consecutive frames are linked by taking both their actionness and overlap ratio into account, followed by the temporal trimming on tubelet.

Recurrent Tubelet Recognition (RTR) networks. The Recurrent Tubelet Recognition networks capitalizes on a multi-channel architecture for tubelet proposal recognition. For each proposal, we extract three different semantic-level features, i.e., the features on proposal-cropped image, the features with RoI pooling on the proposal, and the features on whole frame. These features implicitly encode the spatial context and scene information, which could enhance the recognition capability on specific categories. After that, each of them is fed into a LSTM to model the temporal dynamics for tubelet recognition.

8.2. Experiment Results

We construct our RTP based on [2], which is mainly trained with the single RGB frames. For RTR, we extract the region representations with RoI pooling from multiple clues including frame, clip and motion. Table 5 shows the performances of all the components in our RTR. Overall, the P3D ResNet trained on clips (128 frames) achieves the highest frame-mAP (19.40%) of single component. For the final submission, all the components are linearly fused using the weights tuned on validation set. The final mAP on validation set is 22.20%.

9. Conclusion

In ActivityNet Challenge 2018, 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/temporal action proposals/temporal action localization/spatio-temporal action localization performance and how to generate open-vocabulary sentences for events in videos.

References

[1] F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[2] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei. Deformable convolutional networks. In *CVPR*, 2017.

[3] Y. Gao, O. Beijbom, N. Zhang, and T. Darrell. Compact bilinear pooling. In *CVPR*, 2016.

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

[5] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[6] Q. Li, Z. Qiu, T. Yao, T. Mei, Y. Rui, and J. Luo. Action recognition by learning deep multi-granular spatio-temporal video representation. In *ICMR*, 2016.

[7] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei. Jointly localizing and describing events for dense video captioning. In *CVPR*, 2018.

[8] T. Lin, X. Zhao, and Z. Shou. Temporal Convolution Based Action Proposal: Submission to ActivityNet 2017. *CoRR*, 2017.

[9] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui. Jointly modeling embedding and translation to bridge video and language. In *CVPR*, 2016.

[10] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei. Seeing bot. In *SIGIR*, 2017.

[11] Y. Pan, Z. Qiu, T. Yao, H. Li, and T. Mei. To create what you tell: Generating videos from captions. In *MM Brave New Idea*, 2017.

[12] Y. Pan, T. Yao, H. Li, and T. Mei. Video captioning with transferred semantic attributes. In *CVPR*, 2017.

[13] Z. Qiu, D. Li, C. Gan, T. Yao, T. Mei, and Y. Rui. Msr asia msm at activitynet challenge 2016. In *CVPR workshop*, 2016.

[14] Z. Qiu, Q. Li, T. Yao, T. Mei, and Y. Rui. Msr asia msm at thumos challenge 2015. In *THUMOS'15 Action Recognition Challenge*, 2015.

[15] Z. Qiu, T. Yao, and T. Mei. Deep quantization: Encoding convolutional activations with deep generative model. In *CVPR*, 2017.

[16] Z. Qiu, T. Yao, and T. Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *ICCV*, 2017.

[17] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[18] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *arXiv preprint arXiv:1612.00563*, 2016.

[19] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang. CDC: Convolutional-De-Convolutional Network for Precise Temporal Action Localization in Untrimmed Videos. In *CVPR*, 2017.

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

[21] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[22] L. Wang, Y. Xiong, Z. Wang, and Y. Qiao. Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*, 2015.

[23] Y. Xiong, Y. Zhao, L. Wang, D. Lin, and X. Tang. A Pursuit of Temporal Accuracy in General Activity Detection. *CoRR*, 2017.

[24] T. Yao, Y. Li, Z. Qiu, F. Long, Y. Pan, D. Li, and T. 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.

[25] T. Yao, Y. Pan, Y. Li, and T. Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *CVPR*, 2017.

[26] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. In *ICCV*, 2017.