Hierarchical Language Modeling for Dense Video Captioning

Jaivik Dave and S. Padmavathi

Abstract The objective of video description or dense video captioning task is to generate a description of the video content. The task consists of identifying and describing distinct temporal segments called events. Existing methods utilize relative context to obtain better sentences. In this paper, we propose a hierarchical captioning model which follows encoder-decoder scheme and consists of two LSTMs for sentence generation. The visual and language information are encoded as context using bi-directional alteration of single-stream temporal action proposal network and is utilized in the next stage to produce coherent and contextually aware sentences. The proposed system is tested on ActivityNet captioning dataset and performed relatively better when compared with other existing approaches.

Keywords Video description · Dense video captioning · Computer vision · Natural language processing

1 Introduction

Videos have become an integral part of information interchanging on online internet platforms such as YouTube. On YouTube alone, five hundred hours of video data are being uploaded every minute, and over a billion hours of video data being watched every day. Handling of these profuse amounts of video requires generation of short textual description using automatic analysis of the videos. Video description generation is beneficial to applications involving video retrieval [1, 2], video understanding [3], video recommendation [4], video summarization [5], etc. It can also be used in surveillance and security field for identifying drones, objects or activities and

J. Dave (✉) · S. Padmavathi
Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
e-mail: cb.en.p2aid19007@cb.students.amrita.edu

S. Padmavathi
e-mail: s_padmavathi@cb.amrita.edu

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2022

S. Smys et al. (eds.), *Inventive Computation and Information Technologies*, Lecture Notes in Networks and Systems 336, https://doi.org/10.1007/978-981-16-6723-7_32

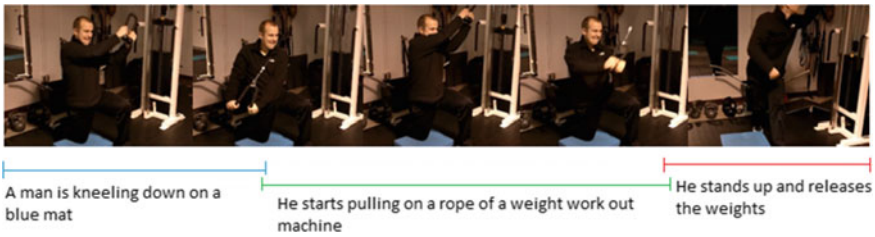


Fig. 1 Example of dense video captioning from ActivityNet captions [6]

report to the concerned authorities. This would also allow visually impaired people to absorb video content with ease.

Prior works in video description problem targeted generating a single caption for the whole video (video captioning). But, it is possible that a single sentence would not be sufficient to describe all the information in a video. Furthermore, longer and real-life videos generally contain numerous activities and objects which require a more detailed description. Hence, the aim of the dense video captioning problem is first to localize the unique activities of the video and automatically generate a natural language description of the same. Few sample frames and the relevant description are shown in Fig. 1. The problem comprises of several challenges like identifying unique and distinct events in the video, capturing the temporal and motion information of the objects and contents of the video and converting it into textual knowledge.

The task is usually formulated as a combination of three modules: encoding a video, temporal segmentation of the video, and finally generating sentences for the identified segments. We introduce an end-to-end approach that first encodes video frame sequence using spatio-temporal convolution neural networks. The system then predicts the possible event, based on past and future events using bi-directional alteration of single-stream temporal (SST) [7] action proposal. Finally, it uses a hierarchical scheme-based captioner to encode the context and decode it along with possible event features to produce a textual description. Hierarchical models have been used in natural language processing for successful sentence generation. In this paper, the hierarchical models are combined with SST action proposal for dense video captioning.

2 Related Works

Over the past decade, prolific research has been done on describing images and videos after the successful advances in natural language processing and computer vision. The early works started with image and video captioning problems to develop models that depict contents in images and videos in a single captioning sentence. Further several works proposed describing the videos in paragraphs to overcome the information loss in the video captioning methods. But even so, the distinct events in the videos were not

addressed or identified explicitly, lacking in providing accurate video descriptions. Nevertheless, as mentioned in section I, the dense video captioning is very beneficial to numerous video analytics task as well as a real-life application like tracking the attention level of students [8], analyzing underwater video [9], identifying animal and humans from surveillance camera [10, 11].

Early works in video captioning were based on template method or rule-based models (e.g., SVO [12], SVOP [13]). Their approaches predict the required content (Subject, Verb, Object, Place, etc.) and then put them into the template to produce a sentence. Modern works use deep learning models for better captioning results especially using recurrent neural network-based sentence generation (e.g., RNN, LSTM, and GRU). These approaches follow the encoder-decoder scheme where the encoder processes video features, and the decoder uses encoder output to produce sentences. Further improvement of video captioning results was proposed by using spatio-temporal attention [14], reinforcement learning [15], and paragraph generation instead of a single sentence [16, 17].

The video paragraph generation problem aims to overcome the limitations of video captioning by providing a detailed explanation of the video content in multiple sentences. Although it offers an elaborate explanation, it does not produce temporal segmentation of the video, which is addressed in the dense captioning task. In the paper [18], authors employed hierarchical RNN models to include sentence history in the decoding process but did not consider the visual context. The paper [19] also utilizes a hierarchical model to produce more coherent paragraphs. We also adopt the idea of incorporating sentence history along with a hierarchical model to get more coherency and meaning in the captions.

The dense video captioning task was introduced by [6] along with the ActivityNet captions dataset. They used Deep Action Proposals (DAPs) [20] for event localization and LSTM network to generate sentences by incorporating past and future context which fails to accommodate highly overlapping events due to usage of fixed strides. The paper [7] extended the idea of context-awareness by utilizing a bi-directional alteration of single-stream temporal (SST) action proposal [21] and employed an LSTM network with 'context gating' to generate contextual sentences. In [22], a hierarchical captioning module is utilized that uses controller LSTM to add context to sentence generator LSTM. The former produces better temporal segments and the later employes effective captioning model, but both lack in the other aspect of the system, i.e., sentence generation and accurate segment proposal, respectively. In view of the recent success of transformers in NLP tasks over RNN-based models, [23] employed a masked transformer-based captioning model to address the task in question.

Several approaches have attempted to include audio [24], speech or both [25] features of the video along with visual features for dense video captioning. The paper [24] proposed a multi-model approach that encodes the video's audio features along with visual features and decoded them with a bi-modal transformer. In [25], importance of speech modality was showed along with video and audio to enhance the performance of the system. However, the event proposal only utilizes visual information and combination of the features is inefficient. The paper [26] captures

contextual relations between the events and stores into 'common-sense' knowledge base but building the database and fetching the vectors corresponding to particular events makes the training and process lengthy and computationally expensive. The paper [27] focus on pre-training strategies for multi-modal pipeline. In order to achieve superior results, some approaches [28] have employed multiple modules resulting in very complex pipeline architecture.

As discussed earlier, majority of the works faces limitations with event proposal or opt for complex approach for captioning. The goal of this work is to develop comparatively less complex and interpretable approach for dense video captioning task. This work focuses on processing the visual information with a 3D CNN; combining it with a bi-directional alteration of single-stream temporal (SST) for action proposal and uses hierarchical language model for sentence generation.

3 Proposed Architecture

This section discusses the architecture framework of the proposed method for the dense video captioning problem. Figure 2 shows an overview of the proposed framework, which comprises of three modules: video encoding, temporal video segmentation or event proposal, and captioning module. The system first encodes the given video to get feature stream and then identifies the unique events (temporal segment) using these feature stream. Finally, description for each event is generated as a natural language sentence.

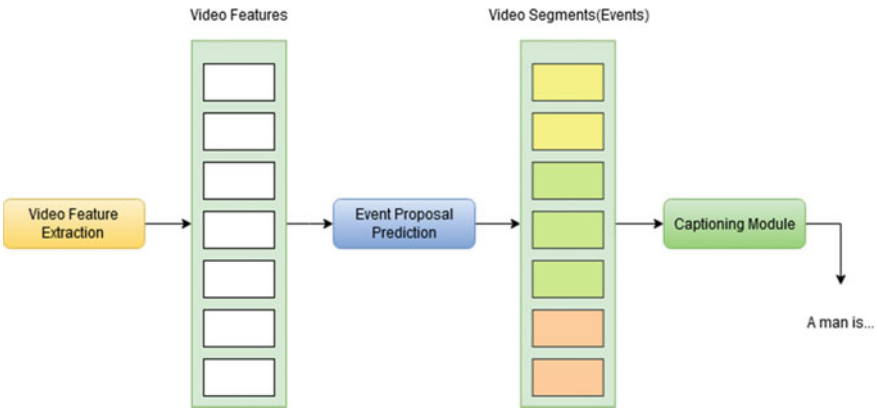


Fig. 2 Complete architecture of the proposed system

3.1 Video Feature Extraction

To obtain the sequence of features from the input video frames, we use the C3D [29] network, a three-dimensional convolutional network. The C3D model used is pre-trained on Sports-1 M [30] dataset and publically available with temporal resolution (δ) of 16 frames. For a video with n frames $v = \{v_1\ v_2,\ v_3,\ \ldots, v_n\}$, the network produces feature stream of length N where $N = n/\delta$. Principal component analysis (PCA) [31] is applied to reduce the dimensionality of the feature streams to $N \times L$, where L specifies the number of significant principal components. This module produces an output of feature stream $f = \{f_1, f_2, f_3, \ldots, f_L\}$.

3.2 Event Proposal Prediction

The obtained feature stream is fed to the event proposal module which identifies possible temporal segments (event) of the video. We employ bi-directional alteration of single-stream temporal SST [7] which incorporates future events along with past events to predict better localization of the possible event. It encodes visual features using LSTM such that hidden state of the LSTM will contain the information till the current timestep t. These are processed to predict possible proposal events with certain scores. Next, the feature stream is encoded in the reverse order and processed similarly to get the proposal events with scores. The scores of the same predicted proposals are combined using an adapt combination strategy and proposal with scores higher than decided threshold are selected for further processing.

3.3 Captioning Module

The final module of the proposed architecture describes each identified event. The visual information of the video can be processed naively by handling each individual event separately and generating the corresponding caption. However, events are linked and can even influence or trigger one another. To model such situation successfully, we propose a hierarchical captioner as shown in Fig. 3. Here initially, the visual information along with language history is encoded. These are then combined with input proposal features for decoding. Input will be a short feature stream. It is necessary to modal the sequential information from these streams to generate accurate description. Since the video can contain any length of events (temporal segments), a simple recurrent neural network will not cater the need. Here we use LSTM cell to produce textual information. It can modal short-term as well as long-term dependencies in contrary to basic recurrent neural networks.

We model the context through encoder LSTM which is used later for decoding. The context is represented as visual information and language history. The context

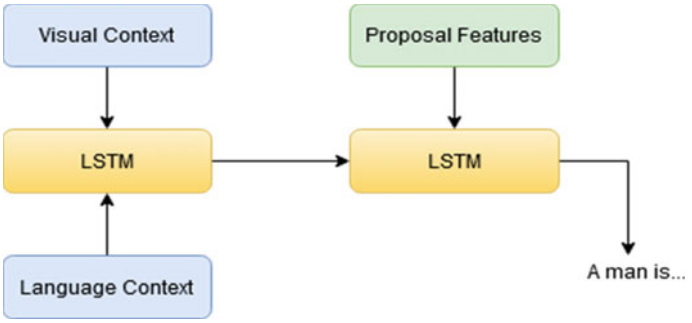


Fig. 3 Captioning module framework

involving past and future visual information; information of previously generated sentences in language history are encoded in hidden state of encoder LSTM (Eqs. 1–4). Different gate vectors of LSTM are computed using Eq. (1), Memory cells are computed using Eqs. (2) and (3), current hidden state of LSTM is computed using Eq. (4).

$$\begin{bmatrix} i_t \\ f_t \\ o_t \end{bmatrix} = \sigma\left(\begin{bmatrix} W_i \\ W_f \\ W_o \end{bmatrix}\begin{bmatrix} V_t \\ D_t \\ h_{t-1} \end{bmatrix} + \begin{bmatrix} b_i \\ b_f \\ b_o \end{bmatrix}\right) \tag{1}$$

$$\widehat{c_t} = \tanh\left(W_c\begin{bmatrix} V_t & D_t & h_{t-1} \end{bmatrix} + b_c\right) \tag{2}$$

$$c_t = i_t \cdot \widehat{c_t} + f_t \cdot c_{t-1} \tag{3}$$

$$h_t = o_t \cdot \tanh(c_t) \tag{4}$$

where i_t, f_t, and o_t are gates of the LSTM input, forget, and output gates, respectively. V_t is visual information, D_t is past language history, and W and b are corresponding weights and biases. h_t and c_t represent hidden state and memory cell of the LSTM cell and $\cdot$ represents element-wise multiplication.

Next, the decoder LSTM network processes hidden state h_t of the encoder LSTM along with event features to generate contextually aware coherent sentences by following equations (Eqs. 5–8):

$$\begin{bmatrix} i_{t,m} \\ f_{t,m} \\ o_{t,m} \end{bmatrix} = \sigma\left(\begin{bmatrix} W_i \\ W_f \\ W_o \end{bmatrix}\begin{bmatrix} h_{t,m}^{(1)} \\ V_t^p \\ h_{t,m-1}^{(2)} \end{bmatrix} + \begin{bmatrix} b_i \\ b_f \\ b_o \end{bmatrix}\right) \tag{5}$$

$$\widehat{c_{t,m}} = \tanh\left(W_c\begin{bmatrix} h_{t,m}^{(1)} & V_t^p & h_{t,m-1}^{(2)} \end{bmatrix} + b_c\right) \tag{6}$$

$$c_{t,m} = i_{t,m} \cdot \widehat{c_{t,m}} + f_{t,m} \cdot c_{t,m-1} \tag{7}$$

$$h^{(2)}_{t,m} = o_{t,m} \cdot \tanh\left(c_{t,m}\right) \tag{8}$$

where $i_{t,m}$, $f_{t,m}$, and $o_{t,m}$ are input, forget, and output gates of the decoder LSTM, respectively. V^p_t is visual information of proposal features, $h^{(1)}_{t,m}$ is the hidden state of encoder LSTM and context vector, W and b are corresponding weights and biases. $h^{(2)}_{t,m}$ and $c_{t,m}$ are represents hidden state and memory cell of the decoder LSTM cell and represents element-wise multiplication. The next mth word is predicted based on the hidden state $h^{(2)}_{t,m}$.

The whole model is trained in an end-to-end fashion with a learning rate of 0.0001 with Adam optimizer. We combine loss from the event proposal module and sentence generation module to get the total loss of overall architecture. Loss of event proposal module ($\mathcal{L}_p$) is calculated based on weighted cross-entropy, and only the proposal having higher IoU (Intersect-over-Union) than ground truths are sent to the language model. Analogous to prior work in language models, we compute language model loss ($\mathcal{L}_c$) as a sum of the negative log-likelihood of the right words in the sentences.

$$\mathcal{L}_T = \lambda_p \mathcal{L}_p + \lambda_c \mathcal{L}_c \tag{9}$$

where $\mathcal{L}_T$ is total loss computed as a function of $\mathcal{L}_p$ and $\mathcal{L}_c$ weighted by λ_p and λ_c deciding contribution of each loss which are set to 0.5 and 1, respectively.

4 Experiments

4.1 Dataset

ActivityNet Captions dataset is based on ActivityNet [32] dataset and was introduced by [6] along with the dense video captioning task. The dataset contains videos annotated with temporal segments and sentences corresponding to each of the segments in the video. Thus, dataset links each unique event in the video with corresponding description. The temporal video segments do not have any constraints in terms of length and can occur simultaneously and overlap with each other. The dataset contains 20 k untrimmed YouTube videos and 100 k sentences. On average, each video is about two minutes long and has four events. Each sentence contains around 13 words on average. The videos are split into 50/25/25 percentage of training, validation, and testing, respectively.

4.2 Results and Discussion

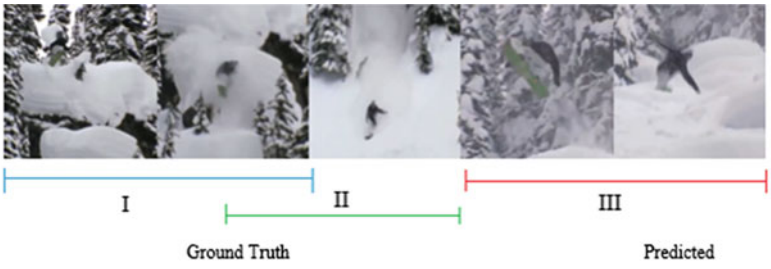
To assess the performance of the proposed architecture, we use METEOR [33] score to determine the degree of similarity between the phrases. METEOR score has been shown to be closely associated and consistent with human assessments, especially when number of reference sentences are limited. However, slight incongruency in other available evaluation metrics (Bleu [34], CIDEr [35]) is noticed by [6, 7, 36], which is due to word sequence misalignment. Also, both the scores are designed based on correlation at corpus level especially Bleu, whereas METEOR is based on correlation with human judgements which is desired for this problem. Hence, the METEOR score is selected for the performance comparison.

The performance comparison of the proposed architecture with existing approaches on the ActivityNet captions dataset is shown in Table 1. The first method [6] introduced the dense video captioning problem and the ActivityNet dataset. It predicts proposals in a single pass with contextually aware sentences. But the technique used predefined strides for action proposal, thus limiting the model for event detection. JEDDI-Net [22] employed controller LSTM to encode language and visual context to enhance the performance, whereas [7] utilized a bi-directional encoder for proposal prediction to identify the endpoints of events more precisely. The proposed method takes language context into consideration which is not the case for [6, 7]. While [8] modals the visual and language history, the event proposal module cannot match the performance of Bi-SST employed in proposed method. For [7], we consider our implementation of variant without ranking proposed in the paper. As it can be seen, our method performs well in comparison to the baseline approaches taken into consideration.

For qualitative analysis, we show a sample example of dense captioning. Figure 4 shows the screenshots of video frames and sentences for the first three events. Even though architecture is able to identify details from the video like 'snowboard,' 'flip,' 'slope,' etc., it still fails to generate more in depth description provided in ground truth.

Table 1 Performance comparison with other existing methods

Method	Meteor
DAPs + LSTM [6]	4.82
JEDDI-Net [20]	8.58
Bi-SST + LSTM [7]	9.17
Ours	9.25



I	A person snowboard on a steep mountain going down without falling	A person is snowboarding down the high hill
II	A man snowboard down the mountain falling from very high and land stand on the mountain	A man is seen snowboarding down a slope
III	A man flips in the air, then lands on the snow to continue skiing down the mountain with deep snow	A man performs a flip in the air and lands on snow

Fig. 4 Qualitative result on dense video captioning task

5 Conclusions

The dense video captioning problem was addressed in this paper, and an end-to-end pipeline, architecture was presented. The proposed architecture employs a bi-directional action proposal module with hierarchical captioning module which incorporates language and visual context into the decoding process. The architecture can produce accurate action proposal along with more coherent and consistent sentences throughout the video, and it performs well compared to existing methods. Transformers have been shown to outperform recurrent neural networks in a variety of natural language processing tasks; therefore, they could be employed as a sentence generator for the entire pipeline in the future.

References

1. V. Gabeur, C. Sun, K. Alahari, C. Schmid, Multi-modal transformer for video retrieval, in *European Conference on Computer Vision (ECCV)*, pp. 214–229 (2020)
2. R. Dhaya, Analysis of adaptive image retrieval by transition Kalman filter approach based on intensity parameter. J. Innov. Image Process. **3**(1), 7–20 (2021)
3. G. Bertasius, H. Wang, L. Torresani, Is space-time attention all you need for video understanding? arXiv:2102.05095 (2021)
4. D. Yao, S. Zhang, Z. Zhao, W. Fan, J. Zhu, X. He, F. Wu, Modeling high-order interactions across multi-interests for micro-video recommendation. AAAI (2021)
5. T. Hussain, K. Muhammad, W. Ding, J. Lloret, S.W. Baik, V.H.C. de Albuquerque, A comprehensive survey of multi-view summarization. Pattern Recogn. (2020)
6. R. Krishna, K. Hata, F. Ren, L. Fei-Fei, J.C. Niebles, Dense-captioning events in videos, in *International Conference on Computer Vision* (2017)

7. J. Wang, W. Jiang, L. Ma, W. Liu, Y. Xu, Bidirectional attentive fusion with context gating for dense video captioning, in *Conference on Computer Vision and Pattern Recognition* (2018)
8. N. Krishnnan, S. Ahmed, T. Ganta, G. Jeyakumar, A video analytic based solution for detecting the attention level of the students in class rooms, in *International Conference on Cloud Computing, Data Science Engineering* (2020)
9. D.G. Lakshmi, K.R. Krishnan, Analyzing underwater videos for fish detection, counting and classification, in *International Conference on Computational Vision and Bio Inspired Computing* (2019)
10. S. Ravikumar, D. Vinod, G. Ramesh, S.R. Pulari, S. Mathi, A layered approach to detect elephants in live surveillance video streams using convolution neural networks. J. Intell. Fuzzy Syst. **38**, 6291–6298 (2020)
11. K. Mondal, S. Padmavathi, Wild animal detection and recognition from aerial videos using computer vision technique. Int. J. Emerging Trends Eng. Res. **7**(5), 21–24 (2019)
12. A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanswamy, D. Salvi, L. Schmidt, J. Shangguan, J.M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, Z. Zhang, Video in sentences out, in *Conference on Uncertainity in Artificial Intelligence* (2012)
13. J. Thomsan, S. Venugopalan, S. Guadarrama, K. Saenko, R. Mooney, Integrating language and vision to generate natural language descriptions of videos in the wild, in *International Conference on Computational Linguistics* (2014)
14. C. Yan, Y. Tu, X. Wang, Y. Zhang, X. Hao, Y. Zhang, Q. Dai, STAT: spatial-temporal attention mechanism for video captioning. Trans. Multimedia (2020)
15. X. Wang, W. Chen, J. Wu, Y. Wang, W.Y. Wang, Video captioning via hierarchical reinforcement learning, in *Conference on Computer Vision and Pattern Recognition* (2018)
16. A. Rohrbach, M. Rohrbach, W. Qiu, A. Friedrich, M. Pinkal, B. Schiele, Coherent multi-sentence video description with variable level of detail, in *German Conference on Pattern Recognition* (2014)
17. H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, Vıdeo paragraph captioning using hierarchical recurrent neural networks, in *Conference on Computer Vision and Pattern Recognition* (2016)
18. H. Yu, J. Wang, Z. Huang, Y. Yang, W. Xu, Video paragraph captioning using hierarchical recurrent neural networks, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4584–4593 (2016)
19. R. Lin, S. Liu, M. Yang, M. Li, M. Zhou, S. Li, Hierarchical recurrent neural network for document modeling, in *Conference on Empirical Methods in Natural Language Processing*, pp. 899–907 (2015)
20. V. Escorcia, F.C. Heilbron, J.C. Niebles, B. Ghanem, DAPs: deep action proposals for action understanding, in *European Conference on Computer Vision*, pp. 768–784 (2016)
21. S. Buch, V. Escorcia, C. Shen, B. Ghanem, J.C. Niebles, SST: single stream temporal action proposals, in *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2911–2920 (2017)
22. H. Xu, B. Li, V. Ramanishka, L. Sigal, K. Saenko, *Joint Event Detection and Description in Continuos Video Streams* (IEEE, 2018)
23. L. Zhou, Y. Zhoud, J.J. Corso, R. Socher, C. Xiong, End-to-end dense video captioning with masked transformer, in *Conference on Computer Vision and Pattern Recognition* (2018)
24. V. Iashin, A better use of audio-visual cues: dense video captioning with bi-modal transformer, in *British Machine Vision Conference* (2020)
25. V. Iashin, E. Rahtu, Multi-modal dense video captioning, in *Conference of Computer Vision and Pattern Recognition*, pp. 958–959 (2020)
26. A. Chadha, G. Arora, N. Kaloty, iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. WACV (2020)
27. G. Huang, B. Pang, Z. Zhu, C. Rivera, R. Soricut, Multimodal pretraining for dense video captioning, in *Proceeding of 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and 10th International Joint Conference on Natural Language Processing, Association for Computational Linguistics*, pp. 470–490. Suzhou, China (2020)

28. T. Wang, H. Zheng, M. Yu, *Dense Captioning Events in Videos: SYSU Submission to ActivityNet Challenge 2020*. arXiv:2006.11693 (2020)
29. D. Tran, L. Bourdev, R. Fergus, T. Torresani, M. Paluri, Learning spatiotemporal features with 3D convolutional networks, in *International Conference on Computer Vision* (2015)
30. A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, L. Fei-Fei, Large-scale video classification with convolutional neural networks, in *Conference on Computer Vision and Pattern Recognition* (2014)
31. I.T. Joliffe, Principal component analysis, *Springer Series in Statistics* (2002)
32. F.C. Heilbron, V. Escorcia, B. Ghanem, J.C. Niebles, Activitynet: A large-scale video benchmark for human activity understanding, in *Conference on Computer Vision and Pattern Recognition* (2015)
33. S. Banerjee, A. Lavie, METEOR: An automatic metric for MT evaluation with improved correlation with human judgements, in *Intrinsic ADN Extrinsic Evaluation Measures fro MT and/or Summarization* (2005)
34. K. Papineni, S. Roukos, T. Ward, W.J. Zhu, Blue: a method for automatic evaluation of machine translation. ACL (2002)
35. R. Vedantam, C.L. Zitnik, D. Parikh, Cider: Consensus-based image description evaluation, in *Conference on Computer Vision and Pattern Recognition (CVPR)* (2015)
36. L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, A. Courville, Describing videos by exploiting temporal structure. ICCV (2015)