Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7657168

REVIEW OF VISUAL DATA DESCRIPTION

Supriya Pradeep Kurlekar
Phd Student, Dept of Electronics Engineering in Shivaji University Kolhapur, India

Dr. Manasi R. Dixit
Professor, Department of Electronics and Telecommunication Engineering, KIT's College of Engineering, Kolhapur, State: Maharashtra, India

Abstract
Nowadays due to vast number of camera equipped devices, large amount of data in terms of image and video are getting generated which brings lot of information which can address many real world problems [16]. Deep learning based Visual data description is one of the most popular research field of Research. Image understanding is associated with objects identification, location and for captioning we need to detect interrelation of objects. Describing video in natural language automatically is very challenging task. It includes understanding of many entities like background scene, human interaction, and other sequential events. Video/image captioning is having huge scope for development in human –robot interaction, virtual assistance, visually impaired people assistance, surveillance, and many more. Captioning can do a lot for those who can't hear. Social media is one of the biggest platform used by more than half billion people for video watching. One great advantage of having caption of your content is that it enables a video or image to be found by search engines including Google, through Search Engine optimization.

I Introduction

Obtaining meaningful information from video is crucial task. Availability of standardized data sets and deep neural network algorithms contributed in significant improvement of video caption generation. Videos may include number of activities, entities, inter related events. Dense Video caption generation has become most challenging task in recent years. It requires many sentences for meaningful captioning. Consequently, dense video captioning task [2] has been introduced and getting more popular recently. The complexity of the defined task is more in conceptual manner compared to simple video captioning as individual events detection becomes necessary for video processing. On Also, due to complexity of the video sequences while generating captions, up to two events are considered in most of the methods [1, 10]. Also, type of event is the main attribute considered while designing and testing the network oriented evaluation methods.

In most of video classification systems object contents are simple and video sequences are used to classify the objects. When scenes in the video are more complex, the computer based algorithms require more computations which is challenging task. The task of captioning of video sequence is one such challenging. Beyond object recognition the challenge shows the requirement of natural language processing requirements in which semantically meaningful information of objects is taken into consideration which establishes relationship among multiple objects. Also, meaningful and informative with fluent caption generation is the main objective.

When relationship establishment is concerned, the complex scene give rise to more complex issues along with text data relationship problems. During development of such models have challenge of establishing relationship between text data in the description and objects in complex scenes while generating natural language descriptions.

The important description to video linked relationship problem is relation between events, objects and activities. On the other hand, the words in description are the contextual information. Prime most tasks in "visual to text" conversion are Fine grained natural descriptions, Intermediate representation learning, Recounting of visual contents and Benchmark Datasets with rich text [16].

The content collection and documentary formation has become easy task due to advancement in such digital content distribution and video recorders. The video data is long sequence which consumes time during users perspective of understanding the content which is the main problem during content sharing. With saving of time maximum information conveyance to the users from such video data is the need of time. The most demanding domain video captioning shows huge scope for the research. Many methods that involve deep learning approach show better success while achieving these targets which also make use of natural language processing methods [11]. We propose machine learning based method for describing video sequence in text description by combining the visual content recognition and natural language processing.

II Literature Review

W. Xu, J. Yu, Z. Miao, et.al [1] Proposed deep reinforcement polishing network consisting of network model for word denoising and network for grammar checking, adequate evaluations have been done to improve performance of video captioning. The long term reward based technique is used for long video sequences with deep reinforcement learning strategy. This also minimizes the gap between visual content information and text data in language domain with revising of words and grammar errors.

D. Yasin, A. Sohail and I. Siddiqi,et.al [2] In this paper, objects appearing in the scenes are considered by authors along with word embedding which finds similar words from the description text. The collection of various video is done for experimentation in which FaceNet is used to identify individuals and combination of CNN and RNN models is used. The valuation of performance is also shown which shows satisfactory results.

MasoomehNabati, Alireza Behrad, et.al [3] have designed parallel processing based boosted architecture using Long Short –Term Memory networks. The iterative training of LSTM is important step towards parallel processing implementation. Which acts as an AdaBoost algorithm during the training phase. The experimental testing is done on public datasets along with comparative analysis with other algorithms. Authors show future scope of the work by dictating the architecture of model which can be trained on small videos in batch processing approach.

Chohan, Murk & Khan,. et.al [4] provided study of image captioning methods which makes use of encoder-decoder models and also attention mechanism. A combinational analysis is shown. Authors have found different mechanisms for scene understanding. Author have listed various fields like medical, industry, agriculture where captioning method of images can be used to automate their tasks.

J. Mun, L. Yang, Z. Ren, et.al [9] have given a method pf video captioning which focuses on temporal features of the video sequence. The coherent feature matching is the main attribute considered while generating a story from the sequence. The event oriented sequence of video and using such sequence for training of the neural network for captioning purpose is the main approach in reinforcement learning. The rewards are of two type which are at bottom event and episode levels. Authors used performance evaluation tool. The tool is from 2018 "ActivityNet Captions challenge", which is evaluated in terms of event describing ability and locating such events in the videos. Supervised learning based experimental analysis is performed and have achieved the state-of-the- art accuracy on the activity net captions data set in terms of METEOR.

T. Fujii, Y. Sei, et.al [10] proposed a method that updates the model by adding new data vector in trained captioning model. They employed data processing methods and trend the same data with the simplest encoder decoder model and calculated the metrics. The document vectorization is done for variable length to fixed length conversion of feature vectors. The hidden layers are trained by combined vectors of images and sentences. A supervised learning approach is used for training and evaluation. Zhang, X.; Wang, X,Tang et.al [12] addressed the challenge in high level and low level semantic features and methods of bridging the gap between them. The application of remote sensing images is considered. The attribute attention based model is developed for caption generation of remote sensing images. The global information is attributed with the use of embedding layer for high level feature extraction of remote sensing images.

Dong-Jim Kim Jinsoo Choi, et.al [13] have introduced relational captioning, proposed a network with multi-task triple-stream. The model network consists of three recurrent units. The joint POS features are used for caption generation. Authors have discussed about Loss functions, triple stream LSTMs, region proposal networks and relational captioning dataset. Authors have done holistic image captioning comparison, suggested that their work can be used in applications of video summarization with natural language processing.

J. Dong, X. Li and C. G. M. Snoek, et.al [14] this paper focuses on choice of sentences that can well describe the scenery in image or video sequence. The visual features are extracted and paired with text features in which firstly text data is converted into vectors using Word2VisualVec (W2VV). The model can extract visual features from text data. The experimental analysis shows that this method can find the most likely caption for given image. Authors have suggested use W2VV with sentence vectorization as multi scale level. The features are predicted using ResNet for training. C. I. Orozco, M. E. Buemi et.al [15] Proposed neural network architecture for encoder and decoder. First CNN for feature extraction and then LSTM has been used to automatically generate the description of the video. They have used Microsoft Video Description Corpos data set for training and testing. Authors commented that automatic generation of video description is currently a topic of interest in computer vision due to applications such as web indexation, video description for people with visual disabilities. Sheng Li, Zhiqiang Tao, et.al [16] performed literature classification with the use of visual to text description link methods. The latest deep learning methods are considered. Authors have discussed and presented the strategies of quantitative evaluations for selected methods that make use of public datasets. In captioning methods according to authors, accuracy of visual information describing is related to mostly natural language processing for better outcomes. Authors have discussed scope for future work in generating natural and diverse descriptions, deep reinforcement learning, and unified framework in captioning applications. The visual understanding and reasoning also large scale benchmarks and evaluations are important aspects.

R.shetty, HamedTavakoli, Jorma Laaksonen et.al [17] discussed various methods that convert visual domain information into language domain. Author discussed overall visual captioning framework consisting of visual feature extraction, CNN features, video features, scene type features and object type and location features. Author also proposed Language model and deeper model with contextual features. Author shared limitations of proposed model that it could count the objects that are prone to repeated words which is better than numeric analysis in some of the captions.

Bai, Shuang & An, Shan. et.al [18] Authors introduced a comparative approach for object relation transfer using geometric attention model on spatial features. MS_COCO dataset is used for quantitative and qualitative performance evaluation with the use of geometric attention for image captioning, which shows significant improvements.

Eleftherios Daskalakis, Maria Tzelepi, et.al [19] proposed captioning model that makes use of contextual information from image features. The CNN based features are extracted from which similarity trajectory is estimated. The framework from authors shows better performance over MS-COCO image captioning dataset and method have also topped the MSR-VTT video to text challenge leader board.

Su, Jiaqi.et.al [20] discussed review of various methods of video captioning. The papers are addressed in the literature with video captioning datasets used by various authors and evaluation metrics that are commonly used. Author have discussed benchmark datasets like TACoS, MSVD, M-VAD, MPII Movie Description Corpus (MPII-MD), MSR Video-to-text and ActivityNet Captions. Author have suggested different methods as Template-based Captioning, Joint Embedding, Encoder-Decoder, Attention Mechanism and Hierarchical Neural Encoder. Author have also identified several possible future directions.

X. He and L. Deng, et.al [21] addressed the video captioning field, also analyzed the key contributions and their progress. Also, the applicability and demands in the field in research and industry deployment is discussed. Future breakthroughs are addressed. The integration of encoder-decoder models in captioning applications are discussed. They also discussed about major deep learning methods in image captioning applications.

Sujin Lee, Incheol Kim, et.al [22] Have proposed method with attention models SeFLA (Semantic Feature Learning and Attention-Based Caption generation). The use of semantic text captions for effective video captioning with respect to events is considered. The attention model used focuses on feature sets with respect to the events in the video sequence with resepct to time stamp that generates effectively correct captions based on the features of video. Authors have developed model consists of three parts mainly, feature extraction using pre trained ResNet and C3D, Dynamic semantic network and static semantic network. Authors conducted experiments with the two datasets MSVD (Microsoft Video Description) and MSR-VTT (Microsoft Research Video-to-text) and demonstrated evaluation of the proposed model.

Rafael A. Rivera-Soto, et.al [25] explored sequence to sequence models that are mainly used in applications of neural machine translation. Authors modified RESNET-50 and VGG-16 CNN in conjunction with the LSTM recurrent neural network model. Performance evaluation is done by authors on the Microsoft dataset and evaluated performance using the METEOR metrics. Authors investigated a simple mean pool model along with sequence to sequence models that are commonly used in video captioning applications. They compared performance and validated experimentally. The instability is analyzed for a single layer encoder-decoder network while performing the task of generating video captions.

Bo Luo, XiaoouTang, et.al [30] proposed a method that segments text from video using temporal feature vectors. Authors have defined few buffers and flags. The captions text is coarsely segmented for generative abstract sequence. The analysis in terms of statistical methods is performed which identifies the (dis)appearance of captions and pairing of text in between adjacent images. The indexing key frames are collected from video which provides video summary which are high quality frames that are sent for optical character recognition.

IV Comparative Analysis

Author, Journal, Year.	**Title**	**Methods**	**Outcome**	**Drawback**
W. Xu et al, IEEE Transaction on Multimedia, 2021.	Deep Reinforcement Polishing Network for Video Captioning	CNN for Video feature extraction, RNN based word denoising and grammar checking network.	Improved score for understandable captions with more grammatically correct captions.	Entire sentence of description is considered for word denoising and grammar corrections which can be replaced with part of speech based approach for further better results.
MasoomehNabati et al, Computer Vision and Image Understanding, ELSEVIER, 2020.	Video captioning using boosted and parallel Long Short-Term Memory networks	Video to frames, fixed length frame count approach, ResNet101, encoder decoder model	Key frame extraction method improves caption linking with that of actual needed process, improved results of captioning about process.	Number of images to process definition linkage defines the accuracy which may change with respect to process application. The change in application changes the results.
R. Shetty et al, IEEE transactions on Multimedia, 2018.	Image and Video Captioning with Augmented Neural Architectures	Neural Network architecture Encoder-Decoder. CNN 3D extracts the features of the input video. MSVD dataset.	MSVD dataset based evaluation shows better performance.	Comparative study with other methods is not presented. Simplistic model is capable of generating caption only on limited dataset.

Dong-Jin Kim et al, IEEE transactions, 2019.	Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning	A multi-task triple-stream network (MTTSNet), consists of three recurrent units for the respective POS and jointly performs POS prediction and captioning.	MTTSNet, which facilitates POS aware relational captioning the effectiveness of the framework over scene graph generation and the traditional captioning Frameworks.	Natural language processing is not considered.
W. Xu et al, Transactions on Multimedia, 2021.	Deep Reinforcement Polishing Network for Video Captioning	CNN model for feature extraction, object oriented feature location, bigram and trigram oriented text feature linking.	Object oriented features and respective text linkage accuracy improved.	Text outcome in terms of meaningfulness and grammar is not considered.
L. Pang et al *IEEE Transactions on Multimedia*, 2015	Deep Multimodal Learning for Affective Analysis and Retrieval	CNN for feature extraction, video retrieval	Finding matching video from dataset	Text generation is another aspect and this paper only focuses on video to video matching which shows feature types extraction with respect to objects in video.
W. Zhang et al, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.	Reconstruct and Represent Video Contents for Captioning via Reinforcement Learning	Reconstruction network (RecNet) encoder-decoder-reconstructor architecture, video semantic features.	the RecNet is fine-tuned by CIDEr optimization via reinforcement learning, which significantly boosts the captioning performance.	The text data features and language quality in terms of meaningfulness and grammar is not considered.
T. Wang et al, *IEEE Transactions on Circuits and Systems for Video*	Event-Centric Hierarchical Representation for Dense	Dense video captioning aims to localize and describe multiple events in untrimmed videos,	Quantitative and qualitative evaluations on the ActivityNet Captions and YouCook2 datasets	Natural language processing steps not followed.

Technology, 2021.	Video Captioning	temporal-linguistic non-maximum suppression (TL-NMS) to distinguish redundancy in both localization and captioning stages	demonstrate that the method improves the quality of generated captions	
L. Gao et al, IEEE Transactions on Image Processing, 2021.	Hierarchical Representation Network with Auxiliary Tasks for Video Captioning and Video Question Answering.	Hierarchical Representation Network with Auxiliary Tasks (HRNAT), for learning multi-level representations and obtaining syntax-aware video captions.	The Cross-modality Matching Task enables the learning of hierarchical representation of videos, guided by the three-level representation of languages.	Shows better results on standard public datasets whereas limited application approach is defined. Change application video may reduce the performance.
Y. Zheng et al, IEEE Transactions on Circuits and Systems for Video Technology, 2021	Stacked Multimodal Attention Network for Context-Aware Video Captioning.	video captioning framework, named Stacked Multimodal Attention Network (SMAN), the Reinforcement Learning method	It adopts additional visual and textual historical information during caption generation as context features, employs a stacked architecture to process different features gradually.	Only limited set of datasets considered for performance valuation.
Y. Yang et al., IEEE Transactions on Image Processing, 2018.	Video Captioning by Adversarial LSTM	Adversarial learning and long short-term memory (LSTM). generative adversarial network (GAN) architecture	The discriminator acts as an “adversary” toward the generator, and with its controlling mechanism, it helps the generator to become more accurate	Typical object oriented performance is better and gets degraded as objects in image increase.
B. Zhao et al, in IEEE Transactions on Image	CAM-RNN: Co-Attention Model Based RNN for	a co-attention model based recurrent neural network (CAM-RNN) is proposed,	During the generation procedure, the visual attention module is able to	Limited set of databases are used to evaluate the performance.

Processing, 2019.	Video Captioning	where the CAM is utilized to encode the visual and text features, and the RNN works as the decoder to generate the video caption.	adaptively attend to the salient regions in each frame and the frames most correlated with the caption.	
M. Qi et al, IEEE Transactions on Circuits and Systems for Video Technology, 2020.	Sports Video Captioning via Attentive Motion Representation and Group Relationship Modeling	hierarchical recurrent neural network-based framework with an attention mechanism for sports video captioning, in which a motion representation module is proposed to capture individual pose attribute and dynamical trajectory cluster information with extra professional sports knowledge	A new dataset called sports video captioning dataset-volleyball for evaluation.	Only sports captioning strategy is considered and when application is changed the performance decreases.
L. Li et al, IEEE Transactions on Circuits and Systems for Video Technology, 2020.	Adaptive Spatial Location with Balanced Loss for Video Captioning	An adaptive spatial location module for the video captioning task which dynamically predicts an important position of each video frame in the procedure of generating the description sentence.	The proposed adaptive spatial location method not only makes our model focus on local object information, but also reduces time and memory consumption brought by the temporal redundancy in extensive video frames and improves the accuracy of	Process oriented video change decreases the performance.

			generated description.	
A. Wu et al, IEEE Transactions on Circuits and Systems for Video Technology, 2020.	Convolutional Reconstruction-to-Sequence for Video Captioning	CNN-based encoder-decoder framework for video captioning. Particularly, we first append inter-frame differences to each CNN-extracted frame feature to get a more discriminative representation	long-term dependencies could be captured by a shorter path along the hierarchical structure, the decoder could alleviate the loss of long-term information	Limited datasets are used for performance evaluation.

V References:-

[1] W. Xu, J. Yu, Z. Miao, L. Wan, Y. Tian and Q. Ji, "Deep Reinforcement Polishing Network for Video Captioning," in IEEE Transactions on Multimedia, vol. 23, pp. 1772-1784, 2021, doi: 10.1109/TMM.2020.3002669.

[2] D. Yasin, A. Sohail and I. Siddiqi, "Semantic Video Retrieval using Deep Learning Techniques," 2020 17th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 2020, pp. 338-343, doi: 10.1109/IBCAST47879.2020.9044601.

[3] MasoomehNabati, Alireza Behrad, Video captioning using boosted and parallel Long Short-Term Memory networks, Computer Vision and Image Understanding, Volume 190, 2020, 102840, ISSN 1077-3142,

[4] Chohan, Murk & Khan, Adil & Mahar, Muhammad &Katper, Saif& Ghafoor, Abdul & Khan, Mehmood. (2020). Image Captioning using Deep Learning: A Systematic Literature Review. International Journal of Advanced Computer Science and Applications. 11. 10.14569/IJACSA.2020.0110537.

[5] Himanshu Sharma ; Manmohan Agrahari ; Sujeet Kumar Singh ; MohdFiroj ; Ravi Kumar Mishra,"Image Captioning: A Comprehensive Survey",IEEE Xplore: 07 May 2020 Conference Location: Mathura, Uttar Pradesh, India. : 10.1109/PARC49193.2020.236619

[6] Yiyu Wang, Jungang Xu, Yingfei Sun, Ben He, (2019), "Image Captioning based on Deep Learning Methods: A Survey', arXiv:1905.08110v1

[7] Haoran Wang, Yue Zhang, Xiaosheng Yu, "An Overview of Image Caption Generation Methods", Computational Intelligence and Neuroscience, vol. 2020, Article ID 3062706, 13 pages, 2020. https://doi.org/10.1155/2020/3062706

[8] Liu, L., Ouyang, W., Wang, X. et al. Deep Learning for Generic Object Detection: A Survey. Int J Comput Vis 128, 261–318 (2020). https://doi.org/10.1007/s11263-019-01247-4

[9] J. Mun, L. Yang, Z. Ren, N. Xu and B. Han, "Streamlined Dense Video Captioning," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019, pp. 6581-6590, doi: 10.1109/CVPR.2019.00675.

[10] T. Fujii, Y. Sei, Y. Tahara, R. OriharaAnd A. Ohsuga, "“Never fry carrots without cutting.” Cooking Recipe Generation from Videos Using Deep Learning Considering Previous Process," 2019 IEEE International Conference on Big Data, Cloud Computing, Data Science & Engineering (BCD), Honolulu, HI, USA, 2019, pp. 124-129, doi: 10.1109/BCD.2019.8885222.

[11] A. Dilawari and M. U. G. Khan, "ASoVS: Abstractive Summarization of Video Sequences," in IEEE Access, vol. 7, pp. 29253-29263, 2019, doi: 10.1109/ACCESS.2019.2902507.

[12] Zhang, X.; Wang, X.; Tang, X.; Zhou, H.; Li, C. “Description Generation for Remote Sensing Images Using Attribute Attention Mechanism”. Remote Sens. 2019, 11, 612 DOI:10.3390/RS11060612

[13] .Dong-Jim kim and Jinsoo Choi and Tae-Hyun Oh and In So Kweon, (2019), “Dense Relational Captioning: Triple-Stream Networks for Relationship-Based Captioning”, arXiv, cs.CV.

[14] J. Dong, X. Li and C. G. M. Snoek, "Predicting Visual Features From Text for Image and Video Caption Retrieval," in IEEE Transactions on Multimedia, vol. 20, no. 12, pp. 3377-3388, Dec. 2018, doi: 10.1109/TMM.2018.2832602.

[15] C. I. Orozco, M. E. Buemi and J. J. Berlles, "Video to Text Study using an Encoder-Decoder Networks Approach," 2018 37th International Conference of the Chilean Computer Science Society (SCCC), Santiago, Chile, 2018, pp. 1-5, doi: 10.1109/SCCC.2018.8705254.

[16] Sheng Li,Zhiqiang Tao “ Visual to Text :Survey of image and video captioning” IEEE2019 DOI 10.1109/TETCI.2019.282755

[17] R. Shetty, H. R. Tavakoli and J. Laaksonen, "Image and Video Captioning with Augmented Neural Architectures," in IEEE MultiMedia, vol. 25, no. 2, pp. 34-46, Apr.-Jun. 2018, doi: 10.1109/MMUL.2018.112135923.

[18] Bai, Shuang & An, Shan. (2018). A Survey on Automatic Image Caption Generation. Neurocomputing. 311. 10.1016/j.neucom.2018.05.080.

[19] Eleftherios Daskalakis, Maria Tzelepi, Anastasios Tefas, Learning deep spatiotemporal features for video captioning, Pattern Recognition Letters, Volume 116, 2018, Pages 143-149, ISSN 0167-8655.

[20] Su, Jiaqi. “Study of Video Captioning Problem.” (2018).

[21] P. Anderson et al., "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, 2018, pp. 6077-6086, doi:10.1109/CVPR.2018.00636.

[22] Sujin Lee, Incheol Kim, "Multimodal Feature Learning for Video Captioning", Mathematical Problems in Engineering, vol. 2018, Article ID 3125879, 8 pages, 2018. https://doi.org/10.1155/2018/3125879X.

[23] X. He and L. Deng, "Deep Learning for Image-to-Text Generation: A Technical Overview," in IEEE Signal Processing Magazine, vol. 34, no. 6, pp. 109-116, Nov. 2017, doi: 10.1109/MSP.2017.2741510.

[24] L. Pang, S. Zhu and C. Ngo, "Deep Multimodal Learning for Affective Analysis and Retrieval," in IEEE Transactions on Multimedia, vol. 17, no. 11, pp. 2008-2020, Nov. 2015, doi: 10.1109/TMM.2015.2482228.

[25] Rafael A. Rivera-Soto, Juanita Ord´o˜nez, “Sequence to Sequence Models for Generating Video Captions”,cs231n.stanford.edu › reports › pdfs

[26] SimaoHerdade, Armin Kappeler, Kofi Boakye, Joao Soares, Image Captioning: Transforming Objects into Words, arXiv:1906.05963v2

[27] Omid Mohamad Nezami, Mark Dras, Stephen Wan, Cecile Paris, “Image Captioning using Facial Expression and Attention”, arXiv:1908.02923v3.

[28] G. Kulkarni, V. Premraj, V. Ordonez et al., “Babytalk: understanding and generating simple image descriptions,” IEEETransactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891–2903, 2013.

[29] S. Li, G. Kulkarni, T. L. Berg, and Y. Choi, “Composing simple image descriptions using web-scale N-grams,” in Proceeding of Fifteenth Conference on Computational Natural Language Learning, pp. 220–228, Association for Computational Linguistics, Portland, OR, USA, June 2011.

[30] Bo Luo, XiaTang, JianzhuangLiu”Video Caption detection and extraction using temporal information” 0-7803- 7750-8/03/1%7 .00 02003 IEEE2003.

[31] W. Zhang, B. Wang, L. Ma and W. Liu, "Reconstruct and Represent Video Contents for Captioning via Reinforcement Learning," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 42, no. 12, pp. 3088-3101, 1 Dec. 2020, doi: 10.1109/TPAMI.2019.2920899.

[32] T. Wang, H. Zheng, M. Yu, Q. Tian and H. Hu, "Event-Centric Hierarchical Representation for Dense Video Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 1890-1900, May 2021, doi: 10.1109/TCSVT.2020.3014606.

[33] L. Gao, Y. Lei, P. Zeng, J. Song, M. Wang and H. T. Shen, "Hierarchical Representation Network with Auxiliary Tasks for Video Captioning and Video Question Answering," in IEEE Transactions on Image Processing, 2021, doi: 10.1109/TIP.2021.3120867.

[34] Y. Zheng, Y. Zhang, R. Feng, T. Zhang and W. Fan, "Stacked Multimodal Attention Network for Context-Aware Video Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2021.3058626.

[35] Y. Yang et al., "Video Captioning by Adversarial LSTM," in IEEE Transactions on Image Processing, vol. 27, no. 11, pp. 5600-5611, Nov. 2018, doi: 10.1109/TIP.2018.2855422.

[36] B. Zhao, X. Li and X. Lu, "CAM-RNN: Co-Attention Model Based RNN for Video Captioning," in IEEE Transactions on Image Processing, vol. 28, no. 11, pp. 5552-5565, Nov. 2019, doi: 10.1109/TIP.2019.2916757.

[37] M. Qi, Y. Wang, A. Li and J. Luo, "Sports Video Captioning via Attentive Motion Representation and Group Relationship Modeling," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 8, pp. 2617-2633, Aug. 2020, doi: 10.1109/TCSVT.2019.2921655.

[38] L. Li, Y. Zhang, S. Tang, L. Xie, X. Li and Q. Tian, "Adaptive Spatial Location with Balanced Loss for Video Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, doi: 10.1109/TCSVT.2020.3045735.

[39] A. Wu, Y. Han, Y. Yang, Q. Hu and F. Wu, "Convolutional Reconstruction-to-Sequence for Video Captioning," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 11, pp. 4299-4308, Nov. 2020, doi: 10.1109/TCSVT.2019.2956593.