Video Captioning using Deep Learning: An Overview of Methods, Datasets and Metrics

M. Amaresh and S. Chitrakala

***Abstract*—In recent years, automatically generating natural language descriptions for videos has created a lot of focus in computer vision and natural language processing research. Video understanding has several applications such as video retrieval and indexing etc., but the video captioning is the quite challenging topic because of the complex and diverse nature of video content. However, the understanding between video content and natural language sentence remains an open problem to create several methodology to better understand the video and generate the sentence automatically. Deep learning methodologies have increased great focus towards video processing because of their better performance and the high-speed computing capability. This survey discusses various methods using the end-to-end framework of encoder-decoder network based on deep learning approaches to generate the natural language description for video sequences. This paper also addresses the different dataset used for video captioning and image captioning and also various evaluation parameters used for measuring the performance of different video captioning models.**

***Index Terms*—Video Captioning, Image captioning, CNN, RNN, LSTM, aLSTM, GRU**

I. Introduction

UNDERSTANDING the video is the key research aspect of multimedia analysis, and generating a natural language sentence for a given video called as video captioning, has been showing great attention in computer vision [1]. Automatic video description generation involves the understanding of many background concepts and also the detection of their occurrences in the video such as objects, actions, scenes, person-person relations, person-object relations and the temporal order of the events. Moreover, it requires translation of the extracted visual information into a comprehensible and grammatically correct natural language description. Video captioning has many applications such as video indexing, human-robot interaction, assisting the visually disabled, automatic video subtitling, procedure generation for instructional videos, video surveillance and understanding sign language [2]. The following are the major challenges in understanding video and generating natural language sentence that are, understanding the fine motion details of video contents and also the interactions of different objects, learning better representations of video between video domain and language domain, ranking the activity identified in the video [3]. Different video captioning approaches have been proposed to overcome these challenges. As mentioned in the Fig. 1. the video captioning methodologies [4] can be categorized into two methods that are template-based methods, deep learning-based methods.

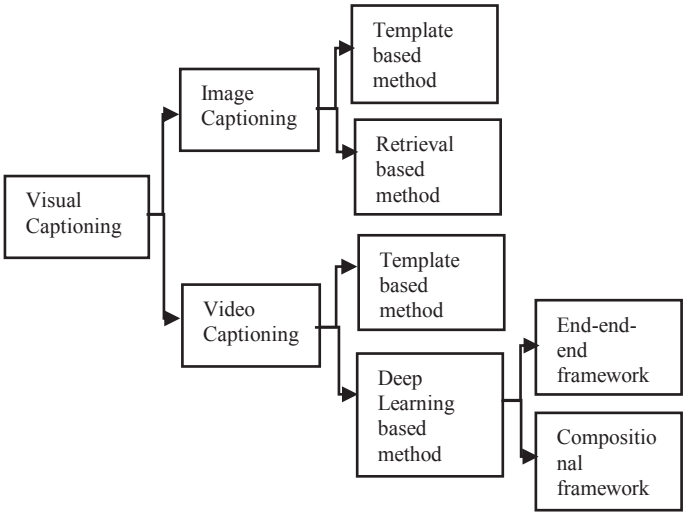

Fig. 1. Categories of Visual Captioning

A. Template-based Methods

The set of predefined specific grammar rules are used in template-based methods and the sentences are divided into different terms such as subject, object, verb and each term should be associated with video data and then natural language description is generated. Although template-based method [5] can generate caption for video based on grammar, this approach relies on video content and output variety got suffer because of the implied limitation.

B. Deep Learning based Methods

1) End-to-end framework: The encoder-decoder architecture initially utilized for machine translation purpose to generate a sentence from the one language to another language.

M. Amaresh, Research scholar at Anna University, Chennai, India (amareshgood@gmail.com).

Dr. S. Chitrakala, Professor at Anna University, Chennai, India (chitrakala.au@gmail.com).

978-1-5386-7595-3/19/$31.00 ©2019 IEEE

The use of the encoder-decoder framework [6][7][8] for video captioning improves the results significantly. Fig. 2 shows the typical encoder-decoder framework for video captioning. In the end-to-end model first, the key frames of the video are encoded into a sequence of feature vector, it represents the semantic information of the video using Convolutional Neural Network (CNN). CNN composed of multiple convolutional layer, max pooling, and fully connected layers. The extracted global visual feature vector is decoded using Recurrent Neural Network (RNN) based decoder for to generate the textual description. A Long-Short Memory Network (LSTM) or Gated Recurrent Unit (GRU) is used as a variation to the RNN and shown to be more efficient and effective in sentence generation.

An attention-based mechanism is used to learn to focus particular region in the frame while generating the description. In an attention-based mechanism, along with global vector, the CNN provide the group of visual vectors for important regions in the frame. Then, in sentence generation, RNN refers the particular region vectors to identify the probability of which part of frame is relevant to the present state to produce the next consecutive words and determine the likelihood that which sub region is relevant to the current state to generate the next word. In the end-to-end framework, the model is trained jointly in an end-to-end manner including the CNN, RNN and the attention model.

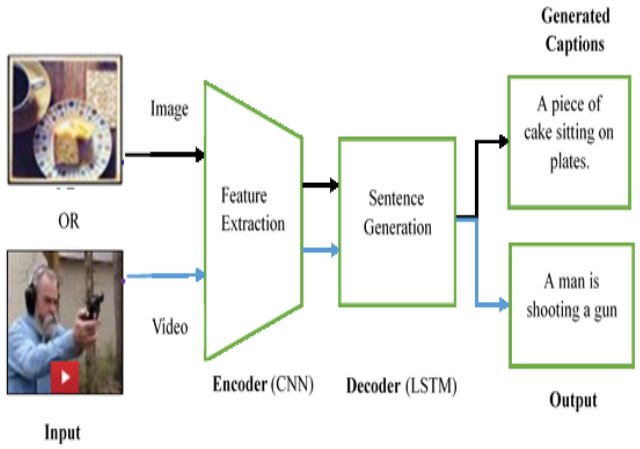

Fig. 2. End-to-End framework of Video captioning

2) Compositional Framework: In this framework a different category of video to sentence methodology [9] uses the semantic concept-detection explicitly for generating the natural language sentences for video. In this framework, a group of semantic labels and actions are generated from the video. These labels may related to different part of sentence including nouns, verb, and adjectives. Language models use these tags to generate natural language sentence for the video.

In this paper Section II presents the literature survey of various methodology proposed for an image and video captioning and Section III reveals different datasets used for an image and video captioning and evaluation metrics are presented in Section IV then Section V and VI discusses the inference from this survey and Conclusion.

II. Literature Survey

A. Image Captioning

Image Captioning has attracted growing interests recently. Early methods for image captioning task can be classified into two types that are template matching approaches and retrieval-based approaches. In template matching approaches objects and actions present in the images are detected and matched with the template to generate appropriate sentence. In retrieval-based approaches, visually similar to the given test images are gathered from the huge database and then fit the descriptions of retrieved images to the test image.

O. Vinayals [10] proposed an end-to-end framework to generate the descriptions for image by replacing RNN encoder into CNN encoder which produce a better image representation to generate textual descriptions. The proposed single joint model named as Neural Image caption model provide training to the CNN for image classification task. The last hidden layer from the CNN is provided as an input to the RNN decoder where it is used to generate the textual description for the image. Likelihood of the target sentence is maximized by training the images using stochastic gradient descent.

Z. Gan [11] constructed a semantic compositional network using semantic concepts to generate a textual description for the query image. Likelihoods of all tags are used to compose the semantic concept vector to process the LSTM weight matrices in the ensemble. This provides the advantage of learning the collaborative semantic concept dependent weight matrices to produce the description for the image.

Q. You [12] proposed a new approach that combines top-down, bottom-up approaches using semantic attention model. Based on the top-down approach, a convolutional neural network extracts the visual features and also detect the visual concepts (objects, regions, attributes, etc.). The semantic attention model is proposed to combine both the image attributes and the visual concepts to produce the description for the image using RNN. By using the bottom-up approach, the attention weights are changed with respect to the RNN iterations for several candidate concepts.

B. Video Captioning

Describing images using natural language has received considerable attention and the research is focusing on the video descriptions. The simple idea to describe the video content is to utilize the deep learning approaches. Natural language description for video content is generated using deep learning technique in two stages, the first step is to use the Convolutional Neural Networks (CNN) which encodes the vector representation of the video and the second step is use the Recurrent Neural Networks (RNN) which decodes vector into to textual description. These deep learning networks significantly achieving good results in many applications such as video indexing, language modeling machine translation and more.

TABLE I
ANALYSIS OF VARIOUS METHODOLOGIES INVOLVED IN IMAGE AND VIDEO CAPTIONING

Sl. No	Paper Keyword	Input	Methodology	Issues	Future Direction
1	Show & Tell [10]	Image	Neural Image Caption model	Selecting salient features from images	Using unsupervised image and text data, to improve sentence generation approaches.
2	Captioning with Semantic Attention [12]	Image	Combined model of top-down methodology and bottom-up methodology	Extracting better context from an image	Plan to experiment on sentence based visual attributes with distributed representations.
3	Hierarchical LSTMs [14]	Image /Video	Hierarchical LSTMs with Adaptive Attention	Attention mechanism on non-visual words, representation of visual data	-
4	Dual Stream RNN (DS-RNN) [5]	Image/Video	Attentive Multi-Grained Encoder (AMGE), Dual-Stream RNN model	Representation and fusion of heterogeneous information	Generating multiple descriptions or paragraph for videos.
5	seqVLAD [16]	Video	Sequential VLAD method to encoding using shared GRU-RCN architecture.	Extracting fine motion details in consecutive frames in a video, and also overfitting	-
6	Attention-Based LSTM [6]	Video	aLSTM and multi-modal correlation	Compressing static representation of video, considering unnecessary portion of frames	Generating sentence for domain specific videos.
7	Bidirectional LSTM (BiLSTM) [7]	Video	BiLSTM & soft attention mechanism	Unidirection for generating sentence.	Exploiting object based spatial and temporal dependency, Achieving spatial reasoning, temporal reasoning
8	From Deterministic to Generative [8]	Video	Multimodal stochastic recurrent neural networks (MS-RNNs)	Decoders propagate only deterministic hidden states	Integrating the state-of-the-art attention scheme along with proposed model
9	Attention-In-Attention Networks [18]	Video	AIA mechanism using encoder attention modules (EAMs) and a fusion attention module (FAM).	Better video representation	Exploring multi-stacked attention mechanism to fuse the feature into multi-space
10	Video Captioning by Adversarial LSTM [19]	Video	Adversarial learning and LSTM in Generative Adversarial Network	Accuracy of generated sentence	LSTM-GAN relying on Reinforcement Learning

N. Xu [4] proposed a dual stream-RNN model for video captioning, which is used to explore and integrate the hidden states of the semantic and visual streams. Proposed model enhances the local feature learning process along with the global semantic feature by exploiting the hidden states of vector representation and semantic concepts separately by using two modalities specific RNN called as Attentive Multi-Grained Encoder (AMGE) and it makes the video representation efficient for video caption generation. Dual-Stream RNN decoder fuse both the streams from AMGE for textual description generation.

J. Song [8] proposed a generative approach Multi-modal Stochastic Recurrent Neural Network (MS-RNN) to generate multiple sentences for the same event by using both prior and posterior distribution. This model also used to overcome the problem of uncertainty which cannot be modeled using deterministic models. Stochastic LSTM ((S-LSTM) is proposed to propagate uncertainty using hidden variable.

The most existing system employed max and mean pooling over each frames of video to produce vector representation but failed to capture temporal structure. In [7] 3D CNN explores temporal information i.e most relevant temporal fragments chosen automatically and forwarded to natural language description generation. Joint visual modeling approach combining forward LSTM and Backward LSTM and CNN to encode video data to video representation and inject into language model to create a description for the video. The proposed model is the first one to use a bidirectional recurrent neural network. And it constructs 2 different sequential processing modules that are adaptive video representation learning and textual description generation. This model utilizes 2 different LSTM unit for both frames encoding and textual decoding. S. Venugopalan [13] proposed a new sequence to sequence framework in end-to-end manner to provide descriptions for short videos. Proposed LSTM model was trained by using video and sentence pairs then automatically learns to attach the frames to a set of words to generate a

natural language sentence. Also, this model could understand the temporal structure of the video and also the generated sentence.

L. Gao [14] proposed an hierarchical LSTM with adaptive attention (hLSTMat) framework to generate sentence for images and videos. In the proposed model the spatial attention or temporal attention mechanism is used for choosing the particular portion of the frame to detect the related words. Adaptive attention mechanism is used to decide whether to rely on visual context or language context. Also, this model considers both low-level visual context and high-level language context to produce the textural description for video.

Y. Pan [15] proposed LSTM unit with transferred semantic attributes (LSTM–TSA) framework to extract the semantic features from the videos by using CNN and RNN framework. Semantic features are used to learn the sequence generate the textual description and these features reflect the stationary objects and scenes in the image but failed to reflect the temporal structure of video. Generating natural description for the video has been improved by merging image and video sources together. Also results showing better performance on different datasets.

Y. Xu [16] developed a novel Sequential VLAD layer, named as SeqVLAD which generates the better representation of video by combining the VLAD and the RCN framework. This model exploring the fine motion details present in the video by learning the spatial and temporal structure. An improved version of Gated Recurrent Unit of Recurrent Convolutional Network (RCN) named as Shared GRU-RCN (SGRU-RCN) was proposed to learn the spatial and temporal assignment. Overfitting problem is resolved in this model because the SGRU-RCN contains only less parameters and this achieve better results.

Describing longer videos semantically in one sentence misses out most of the details and generate uninformative and unexciting results. Yu et al. [17] generated multiple textual descriptions for the longer duration video composed of many different events using hierarchical RNN (hRNN) method. This framework utilising the sequential dependencies between the multiple descriptions in a passage where the next sentence is generated using the semantic context of the previous sentence. In this approach two types of generators used, sentence generator takes the spatial and temporal information exist in the long video to generate a single description whereas paragraph generator process the dependencies between the multiple descriptions. Ning Xu [18] proposed a model to recognize multiple events in the video and generate the natural language description using Attention-in-Attention model. It consists of two different attention modules, first one is Encoder attention modules which selects the most salient visual and semantic features and average both features into single attentive feature to highlight the space-specific features. Second is Fusion attention module which activate the multi-space features and adjust and fuse them for better representation of video. LSTM used to decode the representation of video generates multiple action annotation or video description.

Y. Yang [19] introduced a novel adversarial learning concept by the expansion of LSTM with Generative Adversarial Network for video captioning problem. GAN composed of two interplay module, generator and discriminator, in which generator generates the sentences given the video content using existing video captioning methodology. A novel discriminator module is proposed which act as adversary towards the generator to improve the accuracy. Embedding layer is proposed into discriminator which can transform the discrete output of the generator into consecutive representation. Also proposed a novel discriminative framework to resolve the inefficient classification of sequence to sentence.

This literature survey presents clear picture about different methods to generate the textural description for video shown in Table I. Proposed methodologies in this survey significantly improves the performance of video captioning task and still needs improvement in this field and many unattended challenges available which allow researchers to focus in video captioning and generate description similar to human.

III. DATASET

This section presents various datasets used for an image and video captioning. These training datasets usually consists of an image or video and its ground truth sentences.

A. Image Dataset

1) Microsoft COCO [20]: The biggest corpus for an image captioning. In this dataset 82,783 images are allocated for training, 40,504 images are allocated for validation and 40,775 images are allocated for testing and each image is annotated with five captions by human.

2) Flickr30K dataset [21]: consists of 31,783 images taken from Flickr. In this dataset 29,000 images are allocated for traning, 1,000 images allocated for validation and 1,000 images are allocated testing and also each image is associated with 5 descriptions. It mostly covers the human activities.

B. Video Dataset

1) Montreal Video Annotation Dataset (M-VAD) [22]: consist of 49,000 video clips that are extracted from 92 DVD movies. In this dataset 39,000 video clips are allocated for training, 49,000 video clips are allocated for validation and 5,000 video clips are allocated for testing. Each video clips are described by single sentence. It is a very challenging task to describe movie snippets with one single sentence of ground truth.

2) MPII Movie Description Corpus (MPII-MD) [23]: consist of 37,000 video clips colleceted from 55 movies along with audio descriptions and 31,000 video clips from 49 Hollywood movies. Each video clips are well-appointed with one single sentence from descriptive video service and movie scripts.

3) Microsoft Research Video Description Corpus (MSVD) [24]: consist of 1970 YouTube video clips created by Amazon Mechanical Turk (AMT). Each video clip is about 10 seconds. In this dataset 1,200 video clips allocated for training, 100 video clips allocated for validation and 670 video clips for testing. These video clips are annotated in different language with single sentence. Each video clips annotated by 40 different sentences in English. Fig. 3 shows the sample output sentence generated for video.

Fig. 3. An example from MSVD dataset with the associated ground truth

4) MSR Video to Text (MSR-VTT) [25]: the recent extensive dataset for video captioning. It consist of 10,000 Web video snippets of total 41.2 hours, with 20 different categories such as, sports, music, gaming, and TV shows. There are 20 descriptions for each video annotated by human. In this dataset 6,513 video clips allocated for training, 2,990 video clips allocated for validation and 497 video clips allocated for testing. However, the small size of the clips would be a limitation of it.

5) Youtube2Text video corpus [26]: consist of 1,970 video snippets along with 80,839 sentences in total where 41 human annotated descriptions per video and also each description composed of 8 words. It is an open domain dataset covers several topics such as sports, music etc. In this dataset 1,200 videos allocated for training task, 100 videos are allocated for validation task and 670 videos allocated for testing task.

IV. Evaluation Metrics

To evaluate the results of image captioning and video captioning there are several evaluation metrics have been proposed [27]. The accuracy of generated sentence compared with the ground truth sentence is measured using the n-gram for human annotated sentence and machine generated sentence. Table II shows the evaluation results of various methods used for video captioning. Following evaluation metrics are commonly used for video captioning.

A. BLEU

BLEU (BiLingual Evaluation Understudy) [28] is the most simple and popularly used metrics for video description generation. It measures the numerical translation closeness between ground truth sentence and machine output. Small changes or grammatical errors in the word order is not considered in this metric. It is well suited for shorter sentences.

B. ROUGE

The system ROUGE (Recall-Oriented Understudy of Gisting Evaluation) [29] was initially developed to summarize the documents automatically. ROUGE is similar to BLEU metrics but the difference is ROUGE metric is used to measure based on the n-gram occurrences in the sum of number of human annotated sentences while the BLEU is calculated by considering the occurrences in the total sum of generated sentences. It has four types namely ROUGE-N, ROUGE-L, ROUGE-W and ROUGE-S(U), where version N and L are popularly used for video captioning.

C. METEOR

METEOR (Metric for Evaluation of Translation with Explicit Ordering) [30] is a mean value of unigram-based precision and recall scores. The major difference between METEOR metric and BLEU metric is that it combines both recall and precision metric. BLEU and ROUGE have the limitation of strict matching but it is resolved in METEOR by utilizing the unigrams of words and synonyms.

D. CIDEr

CIDEr (Consensus-based Image Description Evaluation) [31] is a metric used for image captioning. This metric measures the consensus between generated sentence for an image and sentences annotated by the human. Also it is an extension of TF-IDF method. By using the cosine similarity the two sequences are compared and it causes insignificant, ineffective image caption evaluation.

TABLE II
COMPARISON FOR PERFORMANCE MEASURE OF VARIOUS METHODS

Model	B@1	B@2	B@3	B@4	METEOR	CIDEr
DS-RNN[4]	-	-	-	53.0	34.7	79.4
aLSTM [6]	81.8	70.8	61.1	50.8	33.3	74.8
BiLSTM [7]	79.4	60.5	48.6	37.1	29.8	79.0
MS-RNN[8]	82.9	72.6	63.5	53.3	33.8	74.8
hLSTMat [14]	79.4	63.5	48.7	36.8	28.2	120.5
LSTM-TSA[15]	82.8	72.0	62.8	52.8	33.5	74.0
SeqVLAD [16]	-	-	-	50.4	33.17	77.13
h-RNN [17]	81.5	70.4	60.4	49.9	32.6	65.8
AIA[18]	-	-	-	49.5	32.7	67.0
LSTM-GAN [19]	-	-	-	42.9	30.4	-

V. Inference Made

Despite extensive research techniques for video captioning, this survey shows that the following inferences made from studying the existing works that are

- Sentences generated by the system are still less acceptable level than the human annotated sentences.
- Image and video captioning mainly learn to map low-level visual features to sentence without focusing the high-level semantic video concepts (i.e. objects, actions etc.).
- Existing methods focus on predefined templates, instead of generating more natural and diverse sentences.
- Most of the existing techniques on video captioning not exploring the temporal nature of video, which is important to describe the long duration videos.

VI. Conclusion

Thus this survey provides information about various methods using the end-to-end framework of encoder-decoder network based on deep learning to generate the natural language description for video sequences. This paper also addresses the different dataset used for video captioning and image captioning and also various evaluation parameters used for measuring the performance of different video captioning models. This survey gives a clear picture to readers that what has been achieved in this video captioning field so far and also presents where the gaps exist so that future research can be better focused.

References

[1] X. He and L. Deng, "Deep Learning for Image-to-Text Generation A technical overview", IEEE Signal Processing Magazine, Vol-32, no-6, pp:109-116 Nov. 2017.

[2] N. Aafaq, S. Z. Gilani, W. Liu, and A. Mian, "Video Description: A Survey of Methods, Datasets and Evaluation Metrics", arXiv preprint arXiv:1806.00186, Jun. 2018.

[3] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to Text: Survey of Image and Video Captioning", accepted In IEEE Transactions on Emerging Topics in Computational Intelligence, 2019.

[4] N. Xu, A. Liu, Y. wong, Y Zhang, W. Nie, Y. Su, M. Kankanhalli, "Dual-Stream Recurrent Neural Network for Video Captioning", accepted in IEEE Transactions on Circuit and systems for video technology, Mar. 2018.

[5] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Doll´ar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. "From captions to visual concepts and back," in Proc. Conf. Computer Vision and Pattern Recognition, pp. 1473–1482, 2015.

[6] L. Gao, Z. Guo, H. Zhang, X. Xu, H. T. Shen, "Video Captioning with Attention-Based LSTM and Semantic Consistency" in IEEE Transactions on Multimedia, vol-19, no-9, Sep. 2017.

[7] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, X. Li, "Describing Video with Attention-Based Bidirectional LSTM", accepted in IEEE Transactions on Cybernetics, 2019.

[8] J. Song, Y. Guo, L. Gao, X. Li, H.T. Shen, "From Deterministic to Generative: Multimodal Stochastic RNNs for Video Captioning" in IEEE Transactions on Neural Networks and Learning Systems, 2018.

[9] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Doll´ar, J. Gao, X. He, M. Mitchell, J. C. Platt, et al. "From captions to visual concepts and back," in Proc. Conf. Computer Vision and Pattern Recognition, 2015, pp. 1473–1482.

[10] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, "Show and Tell: A Neural Image Caption Generator", IEEE CVPR, pp. 3156-3164, 2015.

[11] Z. Gany, C. Gan, X. Hez, Y. Puy, K. Tranz, J. Gaoz, L. Cariny, L. Dengz, "Semantic Compositional Networks for Visual Captioning", arXiv:1611.08002v2, 28 Mar 2017.

[12] Quanzeng You1, Hailin Jin2, Zhaowen Wang2, Chen Fang2, and Jiebo Luo "Image Captioning with Semantic Attention", IEEE CVPR, 2016.

[13] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, K. Saenko, "Sequence to sequence-video to text", In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4534–4542 2015.

[14] L. Gao, X. Li, J. Song and H. T. Shen, "Hierarchical LSTMs with Adaptive Attention for Visual Captioning", accepted in IEEE Journal of Latex Class Files, Vol. 14, No. 8, August 2015.

[15] Y. Pan, T. Yao, H. Li and T. Mei, "Video Captioning with Transferred Semantic Attributes", IEEE CVPR, pp. 984-992, 2017.

[16] Y. Xu, Y. Han, R. Hong, Q. Tian, "Sequential Video VLAD: Training the Aggregation Locally and Temporally" in IEEE Transactions on Image Processing, Vol. 27, No. 10, October 2018.

[17] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. "Video paragraph captioning using hierarchical recurrent neural networks", In IEEE CVPR, pp.4584-4593, 2016.

[18] N. Xu, A. Liu, W. Nie, Y. Su, "Attention-In-Attention Networks for Surveillance Video Understanding in IoT", accepted in IEEE Internet of Things Journal, 2018.

[19] Y. Yang, J. Zhou, J. Ai, Y. Bin, A. Hanjalic, H.T. hen, Y. Ji, "Video Captioning by Adversarial LSTM", in IEEE Transaction on Image Processing, 2018.

[20] T.Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Doll´ar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in European conference on computer vision. Springer, 2014, pp. 740–755.

[21] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," ACL, vol. 2, pp. 67–78, 2014.

[22] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," arXiv preprint arXiv:1503.01070, 2015.

[23] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3202–3212.

[24] D. L. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in ACL: Human Language Technologies-Vol. 1. Association for Computational Linguistics, 2011, pp.190–200.

[25] J. Xu, T. Mei, T. Yao, and Y. Rui, "MSR-VTT: A large video description dataset for bridging video and language," in Proc. IEEE Conf. Comput.Vis. Pattern Recognit., Las Vegas, NV, USA, 2016, pp. 5288–5296.

[26] S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, S. Venugopalan, R. Mooney, T. Darrell, and K. Saenko. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In Proceedings of the IEEE International Conference on Computer Vision, pages 2712–2719, 2013.

[27] J. Park, C. Song, J.-H. Han, "A Study of Evaluation Metrics and Datasets for Video Captioning", ICIIBMS 2017.

[28] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation", in Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, Stroudsburg, PA, USA, 311-318, 2002.

[29] Lin CY, "ROUGE: a package for automatic evaluation of summaries", in Proceedings of the workshop on text summarization branches out, Barcelona, Spain, (WAS2004) 2004.

[30] D. Elliott and F. Keller, " Image description using visual depedency representations," in Proc. Empirical Methods Natural Lang. Process. 2013, vol. 13, pp. 1292-1302.

[31] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDER: Consensus-based image description evaluation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 4566-4575.