SN Computer Science (2021) 2:120
https://doi.org/10.1007/s42979-021-00487-x

SURVEY ARTICLE

Exploring Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods

Saiful Islam[1] · Aurpan Dash[1] · Ashek Seum[1] · Amir Hossain Raj[1] · Tonmoy Hossain[1] · Faisal Muhammad Shah[1]

Received: 6 November 2020 / Accepted: 23 January 2021 / Published online: 27 February 2021

Abstract

Video captioning is an automated collection of natural language phrases that explains the contents in video frames. Because of the incomparable performance of deep learning in the field of computer vision and natural language processing in recent years, research in this field has been exponentially increased throughout past decades. Numerous approaches, datasets, and measurement metrics have been introduced in the literature, calling for a systematic survey to guide research efforts in this exciting new direction. Through the statistical analysis, this survey paper focuses mostly on state-of-the-art approaches, emphasizing deep learning models, assessing benchmark datasets in several parameters, and classifying the pros and cons of the various evaluation metrics based on the previous works in the deep learning field. This survey shows the most used variants of neural networks for visual and spatio-temporal feature extraction as well as language generation model. The results show that ResNet and VGG as visual feature extractor and 3D convolutional neural network as spatio-temporal feature extractor are mostly used. Besides that, Long Short Term Memory (LSTM) has been mainly used as the language model. However, nowadays, the Gated Recurrent Unit (GRU) and Transformer are slowly replacing LSTM. Regarding dataset usage, so far, MSVD and MSR-VTT are very much dominant due to be part of outstanding results among various captioning models. From 2015 to 2020, with all major datasets, some models such as, Inception-Resnet-v2 + C3D + LSTM, ResNet-101 + I3D + Transformer, ResNet-152 + ResNext-101 (R3D) + (LSTM, GAN) have achieved by far best results in video captioning. Despite rapid advancement, our survey reveals that video captioning research-work still has a lot to develop in accessing the full potential of deep learning for classifying and captioning a large number of activities, as well as creating large datasets covering diversified training video samples.

Keywords Video captioning · Dataset Comparison · Feature Extraction · Spatio-Temporal · Evaluation Metrics · Deep Learning

Introduction

Describing visual content in natural language is quite a simple job by considering the human perspective. When it comes to captioning, several distinct things need to be taken care of, and that becomes exceedingly complicated for a human being. It is then more challenging for machines or computers to generate a caption from visual content like images or short video clips. Automatic image or video captioning includes the combination of multiple entities and the recognition of their appearances in an image or video using different methods of Computer Vision (CV) shown in Fig. 1. These entities include millions of objects, backgrounds, motions, and corresponding language data. All this information must then be conveyed using the grammatical and comprehensible content of the Natural Language

✉ Saiful Islam
islam.saiful03@outlook.com

Aurpan Dash
aurpan.dash@gmail.com

Ashek Seum
ashekseum86@gmail.com

Amir Hossain Raj
raj.amir@outlook.com

Tonmoy Hossain
tonmoyhossain.cse@ieee.org

Faisal Muhammad Shah
faisal.cse@aust.edu

[1] Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh

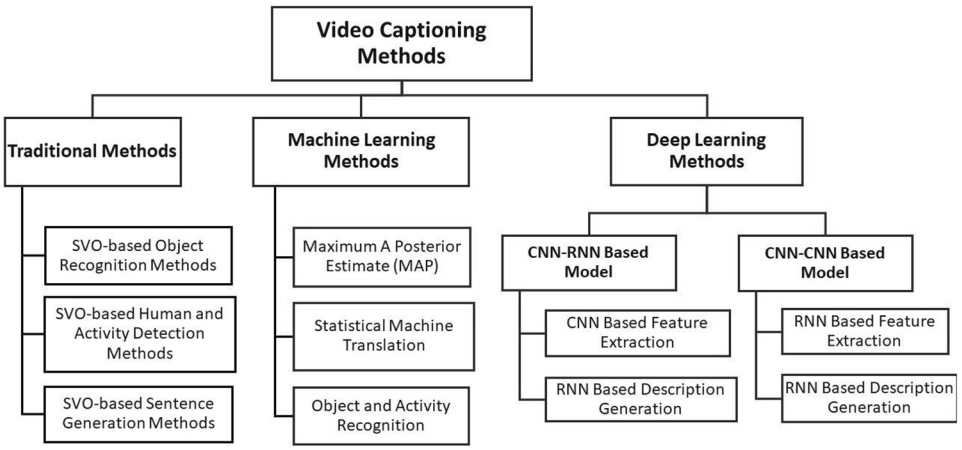


Fig. 1 Different methods used for Video Captioning

Processing (NLP) techniques. In the last few years, these two typically separate areas, CV and NLP [146] have joined forces to tackle the growth of research interests in understanding and explaining images and videos.

Visual content to text generation can be subdivided into two groups —image captioning and video captioning. Image Captioning is a process where the textual description of an image is generated. On the other hand, video captioning is the process where textual descriptions are generated from a sequence of video frames. Video captioning is a much more difficult process compared to image captioning. Video captioning needs to go through a series of tasks, splitted into two main parts. The first one is to understand the video contents visually and the second is to describe the video contents grammatically. Both parts have a greater challenge than image captioning. Image and video captioning have some distinct differences.

- Unlike image captioning, not all objects or actions in a video are important for description generation due to diversity in objects and their actions, which might lead to an irrelevant text generation.
- Video description-based methods must capture the motion, trajectory of related objects, and causality of events, acts, and respective objects; whereas in image captioning, so much of these extractions are not that neither necessary nor possible.
- Events of a video can have various lengths, which can also result in a potential overlap of those events. Video captioning-based methods can tackle these issues very much delicately. Still, in the case of image captioning, the system has pretty much nothing to do to defend the occlusion or overlapping issues.
- Image captioning models can only capture visual features. Upon that, the captions are generated, but video captioning models need to consider not only visual but also spatio-temporal features in order to generate more detailed and accurate captions.
- Video captioning models can detect motion of different objects as it has several frames for analysis and consequently correct detection in a particular time, but image captioning model has only one frame, so meticulously detecting motion becomes very much difficult and sometimes barely possible to come into a conclusion.

In recent years, the availability of large-scale datasets like Flickr8k [94], Flickr30k [147], MSCOCO [66], MSVD [14], MPII-MD [99], etc. and advancement of Deep Learning (DL) architectures boosted up the research in both fields. Introducing the encoder-decoder framework has significantly improved its performance. In image captioning, global features are extracted from the hidden layers of deep CNN architecture, while LSTM [38] or GRU [21] is used as the language model. Furthermore, attention mechanisms are used to capture spatial information more effectively while generating corresponding words [140].

Unlike images, videos are a sequence of images having both spatial and temporal information. The video content extractor part needs to capture the motions of the objects while focusing on the main objects. After the video contents have been extracted, the caption generation part is being activated and generates a corresponding description for the video clip, as shown in Fig. 3. In the initial stages, image features of all the frames were converted into a single feature vector by applying various pooling techniques, thus converting it to an image captioning problem [120]. To capture the spatio-temporal information, 3D CNN architectures were proposed afterward [146], expediting the process of feature extraction and expanding this research field to consider and work on large-scale multimodal datasets. Moreover, RNNs were also used to encode the video by passing the image features of the frames into LSTM [108]. Further improvement

Fig. 2 Single sentence video captioning vs dense video captioning

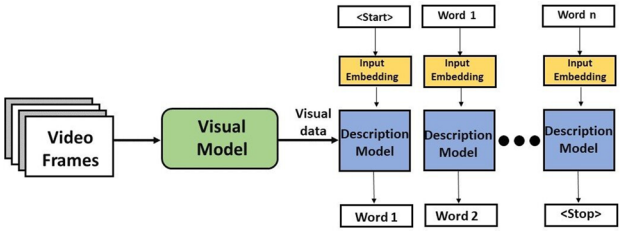


Fig. 3 Basic Structure of Video Captioning Method

was made by applying a spatio-temporal attention mechanism [143].

The application of automatic video captioning is way beyond our imagination. From human-robot interaction to automatic video subtitling and video surveillance system, all of these extensive real-time applications fall into the field of video captioning. Visually impaired people can also get helped through the verbal descriptions of surroundings [46] and also movie descriptions generated by the captioning system. Currently, these manual processes are time-consuming and not economical, also prone to error in sentence generation. Another application can be in the field of sign languages—describing them in natural language, which will alleviate the understanding of normal people [84]. Besides, written procedures for human or service robots can also be generated automatically by converting events of an instructional video into written instructions through video captioning [72].

With the advancement and combination of NLP and CV, making stories from pixels are no more utopia. The process has also been boosted with the help of large-scale video datasets consisting of large-scale descriptions regarding the events in the videos. More research can lead us to better results and corroborate several distinct real-life applications in this research field. Some related research areas include Visual Image Retrieval (VIR) [78] and Image Question Answering [6, 71]. VIR is a system that takes any visuals like videos, images, or any text input or mixed form of visual and text data to perform a content-based search. In all these research fields, large datasets have played vital roles in the captioning system. Some of the popular image datasets are MSCOCO [66] and Flicker30k [92].

Video captioning systems can be categorized as follows, and an illustration of these types is depicted in Fig. 2.

- *Single Sentence Video Captioning:* It tries to sum up the whole video into a single sentence, and obviously, the video may not be sufficiently explained with a single sentence. A general approach to single sentence video captioning for time-varying inputs is the extraction of visual features using a pre-trained Convolutional Neural Network (CNN) or Recurrent Neural Network (RNN)-based model, and then feeding that usually to an RNN-based language model to generate variable-length output vector.
- *Dense Video Captioning:* Instead of using a single sentence to describe a whole video, which may result in missing important details, multiple natural language sentences are used for information enriched, possibly overlapping multiple events of different lengths. Before passing into a caption generation module, visual and semantic features are extracted [41, 61]. Between these two modules' functionality, all the embedded events are localized within time [76, 77].

The research in the field of video captioning started with template-based approaches where a sentence template is used to join Subject, Verb, and Object (SVO) altogether after detecting them from a video. This approach is referred to as SVO-triplets. Hence, the latest approaches now include deep learning-based processes. Now, visual features from videos are encoded using 2D or 3D CNN [34, 81, 149] and sequence is learned through LSTM or GRU [64, 80, 148].

The main contribution of this paper includes an in-depth review of previous video captioning works based on DL. To be more specific, our contributions incorporate:

- Incisive review of state-of-the-art papers on video captioning from 2015 to 2020 mostly focusing on DL approaches.
- Statistical and comparative analysis on different cutting-edge architectures, popular evaluation metrics, and benchmark datasets.
- Depicted a complete flow of evolution of video captioning starting from the Machine Learning (ML) approach to current state-of-the-art works.
- Based on overall analysis and different challenges over time, this paper shows how DL approaches have stretched the limitations of visual feature extraction and text generation tasks.

This paper provides an overall description of deep learning-based video captioning models from a basic introduction to state-of-the-art approaches. It can be a head start for any beginner interested in working in this field.

This article is a compact review of deep learning-based video captioning methodologies, focusing on architectures and performances. In the beginning, the Introduction section of this review discusses some basic things regarding captioning, such as image captioning, video captioning, single sentence video captioning, dense video captioning, and their differences, to make readers' understanding more robust. Going deeper into the survey, ML and DL methods are discussed in Sects. 2 and 3 in a categorical way, mostly focusing on an in-depth review of previous approaches and architectures. In Section 4, all the benchmark video datasets are overviewed, and some statistical analysis based figures of these datasets are provided. Going forward to the next sections, a short description of five evaluation metrics included with their performance score on different architectures. Furthermore, the scopes of video captioning, along with past and present challenges of video captioning with both ML and DL are discussed.

Machine Learning Based Video Captioning

Machine Learning has strengthened CV about recognition and visual tracking. After the era of using traditional and classic methods for object and action detection and fitting them to sentence templates, ML methods provided the opportunity to deal with larger datasets. It offers efficient strategies used in CV for the data acquisition, image processing, object detection, and activity detection. The amount of works available regarding video captioning using ML is very small. Some of the most used algorithms are discussed here.

- *Maximum A Posterior (MAP):* It is a Bayesian-based approach that estimates the distribution of the probability of a problem domain. This probabilistic framework can be used to provide a Bayesian foundation for many ML algorithms. Lan et al. [58] have used MAP in their image captioning model to estimate sentence fluency translated from English to Chinese. The statistical method, based on the MAP, is efficient in segmentation. The trick to the MAP approach is to estimate the prior likelihood of segmentation. The Multilevel Logistic (MLL) model has been used for estimation in practice. To further enhance the efficiency of segmentation, especially MR image, Liu et al. [67] have proposed a Weighted MLL (WMLL) model.
- *Statistical Machine Translation (SMT):* It is a paradigm of Machine Translation (MT) that generate translations based on statistical models where the parameters of the models are derived from bilingual text corporate research. This statistical approach is compared with Koehn [53] and MT rule-based approaches, which is an example-based MT. Unlike the rule-based machine translation (RBMT) approach that is typically word-based, most existing SMT systems are phrase-based and use overlap phrases to assemble translations. Koehn [53] have defined an SMT open-source toolkit that supports linguistically driven influences, decoding of confusion networks, and effective data formats for translation models and language models. Non-factored SMT usually deals only with the surface form of words and has a single phrase table. In contrast, in factored translation models, the surface forms can be supplemented with different variables, such as Parts of Speech (POS) tags or lemma. This provides a description of each word in terms of variables.

In this section, various machine learning-based methods are discussed in short to have a good understanding of the methods used for different sub-tasks of video captioning, such as object detection, activity recognition, and sentence generation.

Object and Activity Recognition

A good number of available works have included ML methods for different computer vision problems like object detection and activity recognition. All those approaches can be divided into two parts, firstly, the threshold-based detection, and secondly, the manual feature and traditional classifiers.

Threshold Based Detection

The approach used by Kojima et al. [54] extracts semantic information of human actions or motions by applying main focus on three things, the position of the head, position of hands, and direction of the head. To detect the motion of the human in a visual frame, different threshold values are used where each of the thresholds represents a specific activity. The following three insights can be obtained through comparatively light-weight cases in order to detect a human's posture: head position, head direction, and hand position. The action of transferring an object, in particular, is detected independently. Besides, the object can be identified by comparing edges and color histograms of extracted object regions with those of object models. Conceptual descriptions of actions are generated for each body part by applying domain knowledge. Kim and Park [50] have used local histogram analysis for defining edge features consisting of three properties; Local Contrast, Region Ration, and Edge Potential. Based on the edge feature information, segments are marked and labeled using different threshold values.

Manual Feature Engineering and Traditional Classifiers

Roy et al. [101] have introduced a novel two-stage word-spotting approach to detect visual text from video frames. To do so, the segmented text images are converted to a binary image using the Bayesian Classifier Binarization approach. The script of the image is identified using the Hidden Markov Model (HMM). Das et al. [23] have worked in three distinct levels; low level or Topic Model, middle level or concepts to language, and high level or Semantic Verification. In the low level, the GM-LDA model had been adopted in [13] (dubbed MMLDA for MultiModal LDA) to handle discrete visual feature space using e.g. HOG3D, etc. The middle level follows the top-down approach that sparingly defines concepts in the video, matching them over time. The high-level system joins the two earlier sets of lingual descriptions (from the low and middle levels) to enhance the set of sentences and filter them. Furthermore, Motwani and Mooney's approach has been followed to automatically extract semantic SVO triplets from the human-generated sentences and a separate semantic hierarchy has been built for each part of the triplet (HS, HV, and HO) [73].

Discriminatively-trained [30] deformable parts models and the motion descriptors, developed by Laptev et al. [59], which have been used respectively for object detection and activity recognition [57]. A probabilistic graphical model that combines visual detection with person, event, and scene language statistics to identify the best SVO and location for a given video to be represented. A descriptive English sentence is constructed by Thomason et al. [115], from the selected sentence part. For entity related features extraction, ObjectBank [62], the 20 PASCAL [27], and LLC-10k proposed by Deng et al. [24], all have been used.

For activity-related features extraction, Xu et al. [142] have adopted the extraction of Dense Trajectories developed by Wang et al. [125] and have computed Histogram of Gradients (HoG), Histograms of Optical Flow (HoF), and Motion Boundary Histogram (MBH) attributes over spatio-temporal volumes around the trajectories, by following default parameters. After that, three non-linear SVM is used to combine all the extracted features. To combine visual and linguistic evidence right after text mining from English text corpora, the paper uses the probabilistic factor-graph model and use that SVOP tuple chosen by the model to generate an English sentence following a template.

Sentence Generation

Based on probabilistic classifiers like Random Forest, Sun and Nevatia [110] have used Semantic Aware Transcription (SAT) for English words distribution taking concept detection results as input. This model creates pairs by combining videos and sentences and uses them for training purposes. It learns splits of nodes hierarchically by grouping semantically related words, calculated by continuous skip-gram language model. Rohrbach et al. [100] have proposed a method for extracting rich semantic information regarding visual content which includes labels for object and their activity. For predicting the semantic representation, the Conditional Random Field (CRF) is trained with the intention of modeling the relationships between different components of the visual input. Then the paper proposes to formulate the natural language generation as a MT problem using semantic representation as to the source language and the generated sentences as the target language. The proposed framework of Xu et al. [142] consists of three parts: a compositional semantics language model, a deep video model, and a joint embedding model which models both video and text jointly. CRF is used here for incorporating Subject–Verb and Verb–Object pairs for text generation part.

In this section, we have briefly discuss some of these machine learning-based models that had paved the way for generating captions from videos. Machine learning had brought revolutions in various steps of video captioning especially, object or activity recognition, feature engineering, classification. Through these groundbreaking models, ML had solved many challenges and built applications regarding video captioning. As usual, so many other challenges and some more advanced application ideas arose which could not be solved by ML, to solve this insane amount of data, and high computational power was required. Therefore, deep learning gradually is emerged.

Deep Learning Based Video Captioning

Deep Learning (DL) is known to be a sub-field of ML which deals with algorithms called artificial neural networks created according to the structure and function of the human brain. The CV domain is moving from a statistical approach to deep neural network approach in terms of working with large-scale datasets. In recent years, DL has significantly impacted different sectors of Artificial Intelligence (AI), and perhaps the most gained area is CV. Nevertheless, deep learning has contributed many state-of-the-art benchmark results in this field, especially in video captioning. A general structure of a deep learning-based video captioning system is depicted in Fig. 4. It consists of two sections, Encoder and Decoder. The Encoder extracts the visual features through various methods like CNN, RNN, and the Decoder generates caption according to those visual features using mainly RNN-based models. Also, there are several different DL methods of video captioning, and the classification is shown in Fig. 5. In this survey, we have represented more than 100 research works, all introducing different video captioning models, which are annotated in Table 1. From

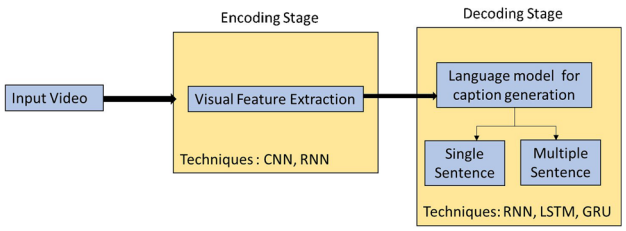


Fig. 4 General Overview of an Encoder-Decoder Based Framework

all those models, we have partitioned this section into two divisions – CNN-RNN Video Description and RNN-RNN Video Description covering the feature extraction and sentence generation methods.

CNN-RNN Video Description

In this category, a captioning model is a combination of two different architecture that is CNN and RNN. Working on the captioning system, there are many variations of both CNN and RNN. CNN is used to generate feature vectors from image spatial data where vectors are fed into the RNN architecture via a Fully-Connected (FC) linear layer to generate word sequence as shown in Fig. 6.

Feature Extraction

Yao et al. [146] have proposed encoder-decoder based architecture for video description. The authors have proposed a spatio-temporal convolutional neural network (3D CNN) along with a temporal attention mechanism as the encoder. An end-to-end trainable and large-scale visual learning capable model has been developed by Donahue et al. [26], which is a bunch of architectures exploiting the strengths of CNNs for visual features. Based on deep image description model [26], Venugopalan et al. [121] have suggested a framework that extends the previous model and the latter one consists of CNN for extracting features for each frame, then mean pooling those features and input that at every time step to the decoder. For feature extraction purposes, Guo et al. [34] have proposed both 2D and 3D CNN for the encoding phase. Zhang et al. [153] have proposed a novel architecture, Guidance Module Net (GMNet), that acquaints a guidance module that enables the encoder-decoder model to generate words in a caption related to both past and future words of that caption. This GMNet is constructed based on the guidance module as well as normalization. A soft attention mechanism has been integrated to the encoder-decoder model, and InceptionV4 [112] has been used to extract semantic features of the video.

Zhang and Tian [149] have proposed a framework which is fundamentally constituted by a parallel fully connected layer encoder that jointly learns video representation from two separate flow of video sequences such as RGB frames and Motion History Images (MHI), which are modified by 3D CNN. Bin et al. [12] have proposed video captioning framework's architecture, where extracted features by CNN are also feed through backward pass networks along with forwarding pass networks, combined with original features from CNN. Pan et al. [81] have suggested the LSTM-E framework, where extracting features by 2D and 3D CNN, mean pooling would be applied over those features and after that, 'relevance loss' as well as 'coherence loss' would be measured. Later on, minimizing these two losses, both LSTM and visual-semantic embeddings are learned all together. Pan et al. [82] has proposed a deep neural architecture, which is a CNN + RNN categorical encoder-decoder framework. Hori et al. [40] have proposed an approach that is an encoder-decoder-based sentence generator where

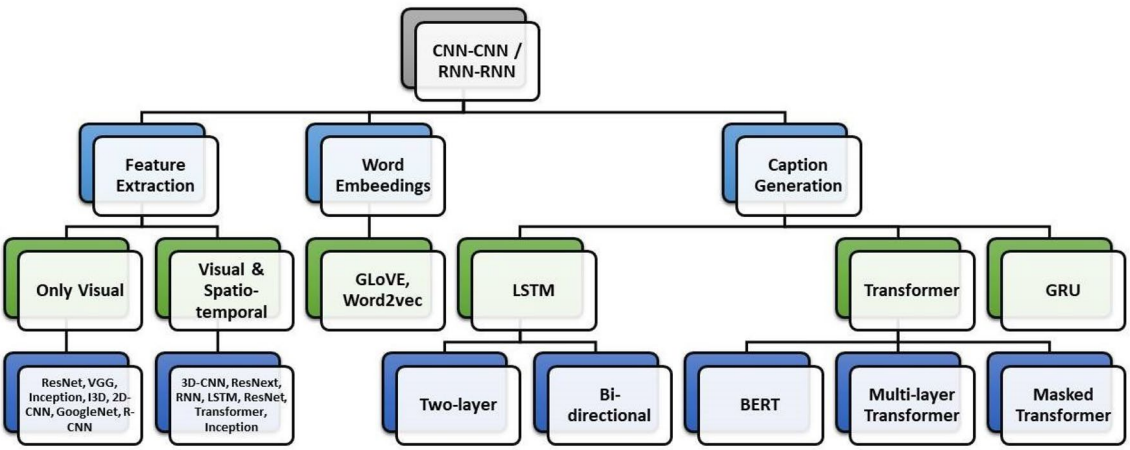


Fig. 5 Classification of Different Deep Learning Methods in Video Captioning

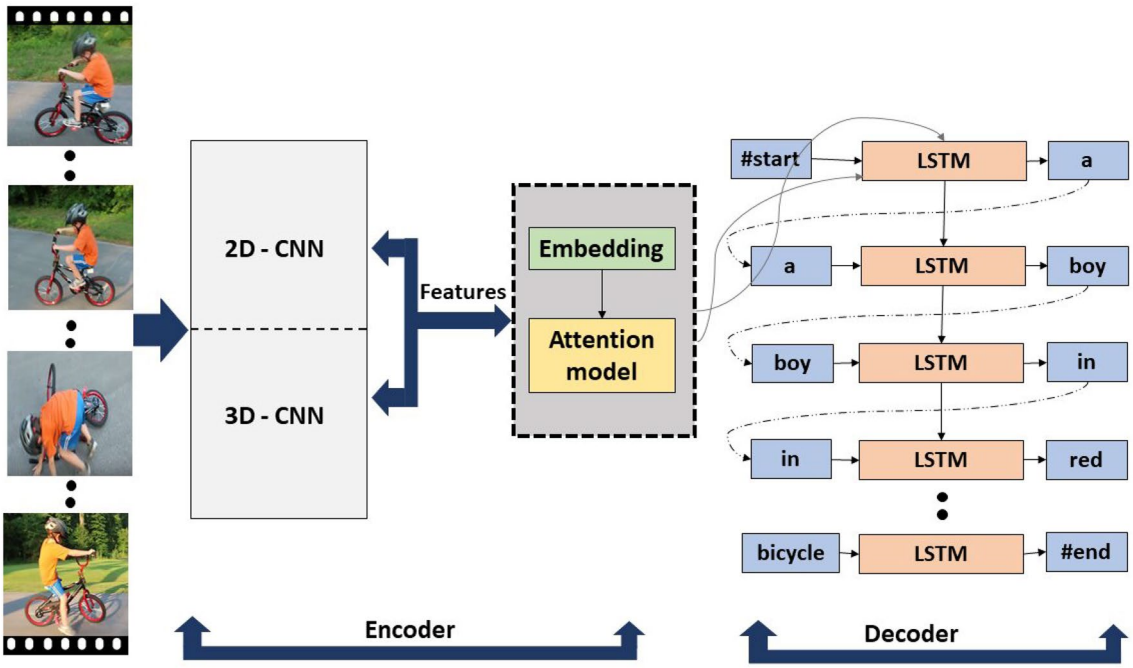


Fig. 6 General overview of CNN – RNN Video Caption Generation Method

feature extraction has been assigned to CNN. This model also comes with a temporal attention mechanism. Xu et al. [138] have proposed a model that solves the dense video captioning. Among its two modules, the 'Proposal Module' features are extracted using 3D convolutional layers (C3D), and Segment Proposal Network (SPN) is used to get temporal segments. By mapping a visual representation into a common vector space, which relies only on the video, with a syntactic representation that relies only on Part-of-Speech (POS) tagging structures of the video description, Perez-Martin et al. [90] have proposed to create a visual-syntactic embedding and then integrating this into an encoder-decoder-based model. The encoder works in three stages such as extracting 2D-CNN and 3D Convolutional Neural Networks (3D-CNN) features, generating semantic representation, visual-syntactic embedding.

Hemalatha and Sekhar [37] have proposed a video captioning approach utilizing domain-specific decoders by constructing a domain classifier. In the proposed model, 2D-CNN features are extracted using ResNet152, and temporal features are extracted using a 3D-CNN. To obtain a video representation, both the 2D-CNN and 3D-CNN features are processed by VLAD [36]. Xiao and Shi [137] have proposed a model that comes with text-based dynamic attention. In this model, for extracting visual and temporal features, InceptionV3 [114], as well as C3D and VGG, have been used respectively. Mukherjee et al. [74] have proposed an architecture to detect fighting scene from hockey videos by spatio-temporal information using transfer learning taking the help of pre-trained DL model VGG-Net. Regional Proposal Networks (RPNs) are used by Agrawal et al. [4] for creating boundary boxes around objects using InceptionV3, pretrained on ILSVRC-2012-CLS image classification dataset.

Luo et al. [70] have proposed a unified multimodal system in which, for visual features extraction purposes, ResNet-152 and 3D features extraction, ResNext-101 has been suggested to use. Lei et al. [60] have proposed a memory-augmented Transformer [118] based model, ResNet-200 and BN-Inception have been used for extracting visual features as well as optical flow, respectively. Liu et al. [68] have proposed 'SibNet' which is composed of the content branch and the semantic branch, implemented using CNN. The role of the content branch is to encode visual content information, and the semantic branch is to learn a representation of video input that encodes high-level semantics. Both branches are constructed by stacking 3 and 6 identical Temporal Convolutional Blocks (TCBs). Fawzy et al. [29] have proposed an encoder-decoder based deep neural network architecture for efficient localization of video events using a bidirectional Long Short Term Memory (Bi-LSTM) that encodes information from past, present, and future contexts. For the extraction of spatio-temporal features, the paper has borrowed a two-stream [117] 3D Residual Neural Network; one is ResNet-18, and the other one is 3D ResNeXt-101.These features are then utilized with a Bi-LSTM for proposal generation.

Hou et al. [42] have suggested a progressive visual reasoning method that steadily generates fine phrases from weak annotations instead of strong ones by deriving more semantic concepts and their video captioning dependency relationships. The working procedure of encoder of the whole model can be shortly brief into two steps: 1) building semantic concepts, 2) constructing dependency trees. To extract visual features, Inception-Resnet-v2, and for spatio-temporal features, C3D has been utilized. Glove is for representing words. Zhao et al. [157] have proposed 'MemCap', a novel stylized image captioning method that explicitly encodes the knowledge about linguistic styles with memory mechanism. The authors have proposed to implement VGG-16 with Faster-RCNN for visual feature extraction. Chen and Jin [17] have proposed to add a sequence-level exploration term to the current objective to boost recall. The litterateurs have used Resnet-152 to compare with other works fairly, instead of using a stronger CNN. The authors have used spatial-temporal mean pooling to get a vector of a 2048-dim feature. 2D-CNN, 3D-CNN, and object detector are used in the proposed model to extract functions from the video input file [158]. The Extractor-encoder portion that learns a self-attended representation of the scene is one of the two main blocks of the proposed Syntax-Aware Action Targeting (SAAT) module. Perez-Martin et al. [89] have proposed a new architecture, named Attentive Visual Semantic Specialized Network (AVSSN). It is based on the litterateurs' suggested Adaptive Attention Gate and Specialized LSTM layers. The encoder of the proposed model consists of two parts; the first part computes 2D-CNN and 3D-CNN feature vectors. The second part does the job of a 'concept detector', which is the work of Chen et al. [15] and Gan et al. [31].

Zhang et al. [155] have proposed an Object Relational Graph with Teacher-Recommended Learning (ORG-TRL) system, which consists of an Object-Relational Graph (ORG) based encoder, which can capture more comprehensive interaction features to improve visual representation. In the data preprocessing phase, Inception-Resnet-v2 and C3D are used for visual and spatio-temporal feature extraction purposes. Chen et al. [20] have proposed to use a combination of fully convolutional networks and multi-instance learning for feature extraction. No such network is used for temporal feature extraction. Sah et al. [102] have proposed a method that combines an efficient hierarchical architecture with novel attention mechanisms at both local and global levels. The authors have used 152-layer ResNet CNN model for visual feature extraction. Wang et al. [126] have proposed a new 'Sequence in Sequence' framework to encode the sequential frames at each time-stamp into a spatio-temporal representation. The litterateurs have proposed to encode the sequential frames into a sequence of VLAD representations in the encoding stage. The authors also have proposed to use the Convolutional GRU to calculate the assignment to investigate the spatio-temporal correlation at each time-stamp between the successive frames. Huang et al. [44] have proposed an image captioning model which comes with End-to-End Attribute Detection and Subsequent Attributes Prediction. In this model, features are extracted using ResNet-101 based Faster-RCNN.

Sentence Generation

Yao et al. [146] have proposed to use LSTM as a decoder in their encoder-decoder-based architecture for video description generation. An end-to-end trainable proposed model of Donahue et al. [26] have used LSTM for variable-length predictions. Following the model, Venugopalan et al. [121] have also suggested to use LSTM for generating sentences in detail. The core idea of the proposal of Guo et al. [34] is to incorporate attention mechanisms in the LSTM for language model. Zhang et al. [153] proposed a novel architecture, GMNet, in which an RNN-like network is used as a decoder. Layer Normalization [7] and LSTM-based Guidance Module has been utilized as innovative points for improvement of the performance of video captioning tasks. Perez-Martin et al. [90] have suggested a model in which decoder is mainly a compositional-LSTM network. A dynamic semantic concepts network is developed upon an encoding LSTM for caption generation using semantic features. in the proposed model of [29]

In the proposed model of Bin et al. [12], video captioning framework extracted features are passed to LSTM-based decoder for final sentence generation. According to the proposed model of [82], the decoding stage has been designed with LSTM and Transferred Semantic Attributes (LSTM-TSA), which is the model's core idea. Hori et al. [40] have proposed an approach where sentence generation has also been done with the help of RNN associated with a temporal attention mechanism. Xu et al. [138] have proposed a model that solves dense video captioning with the 'Hierarchical Captioning Module' generates captions for videos with controller LSTM. Based on the SCN-LSTM and [15, 31]; Martin et al. [89] have proposed a new compositional model which works as the decoder. It incorporates several functionalities that were absent in the original model, such as Visual-dependent layer, Temporal Attention, Semantic-dependent Layer, Adaptive Attention Gate, and Word Embedding.

In the proposed model of Hemalatha and Sekhar [37], domain-specific LSTM decoder is used for generating captions. Luo et al. [70] have proposed Transformer in the decoding phase in their proposed model. Lei et al. [60] has also proposed to use the Transformer for decoding purposes. Liu et al. [68] have proposed an RNN decoder that has been used utilizing soft-attention mechanism. Jia and Li [48] have proposed LSTM as a sentence generator. Zhao et al. [157] have proposed to generate captions such as the

proposed model 'MemCap' first extracts content-relevant style knowledge from the memory module via an attention mechanism and then incorporates the extracted knowledge into a language model. The decoder of the proposed model of Chen and Jin [17] is an RNN model of LSTM cell. The extractor-decoder component is one of the main blocks of the proposed SAAT module by Zheng et al. [158], which is used to decipher components of the syntax including subject, object, and predicate and used the targeted action to direct the generation of descriptions. LSTM does the job of sentence generation.

Zhang et al. [155] have proposed a hierarchical decoder with temporal or spatial attention. The model also developed a teacher-recommended learning system to make full use of the popular external language model to incorporate much linguistic information into the process. In this model, LSTM is used as a decoder. Chen et al. [20] have proposed LSTM as caption generator in their model. Aafaq et al. [3] have proposed to implement the Masked Transformer for decoding purposes. Wang et al. [126] have proposed model has two layers, and each layer has one GRU in the decoding phase. Huang et al. [44] have proposed to implement Two layers SAP-LSTM for decoding purposes.

RNN–RNN Video Caption

RNN-RNN based video captioning models have RNN for both feature extraction and language generation purpose. The encoder portion uses a combination of CNN and a variant of RNN known as LSTM, and the decoder uses LSTM for language generation. Using an Attention model enhances the performance of the decoder. A basic structure of the RNN-RNN model used as encoder and decoder is shown in Fig. 7.

Feature Extraction

For facial expression recognition based on profound appearance and geometric neural networks, Jeong et al. [47] have proposed effective deep joint spatio-temporal characteristics. To obtain spatial and temporal features at the same time, The litterateurs suggested applying three-dimensional (3D) convolution. For the geometric network, it's recommended to study the energy distribution of entire facial landmarks. Rohrbach et al. [98] have proposed a two-step approach to the video description. In the first step, pre-trained visual classifiers were proposed to use for visual recognition. Ballas et al. [8] have used three different modified RCN architectures in their proposed model, such as GRU- RCN, stacked GRU-RCN, and bi-directional GRU-RCN. Motivated by Conv-Nets' amazing accuracy, Pan et al. [80] have proposed a new model, namely Hierarchical Recurrent Neural Encoder (HRNE) for video processing.

Pasunuru and Bansal [85] have proposed a many-to-many multi-task learning model where LSTM-based encoders-decoders are shared for video captioning, unsupervised

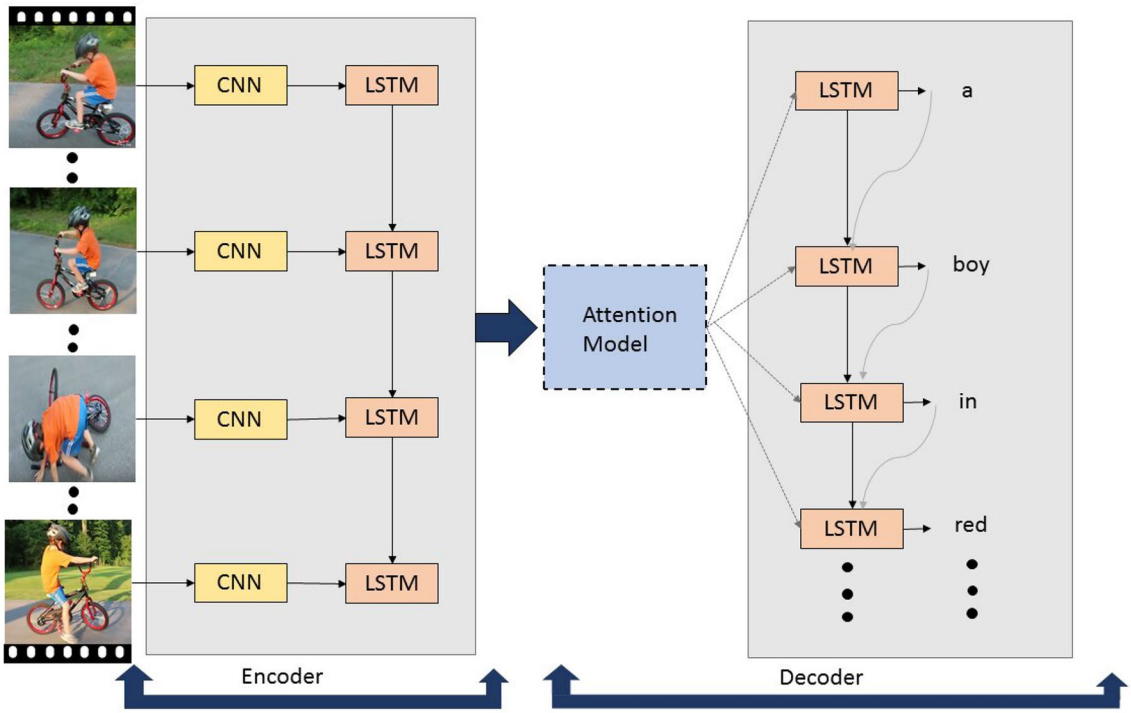


Fig. 7 A General Overview of Video Caption Generation Method where RNN is used in Encoding Stage

video prediction, and entailment generation tasks. Mukherjee et al. [75] have used a combination of Encoder and Convolutional-LSTM for forecasting the next frame based on the spatio-temporal features of the current frames. Peris et al. [91] have suggested a four-stage encoder-decoder approach, using both CNNs for feature extraction. Instead of typical encoder-decoder architecture, Baraldi et al. [11] have introduced a novel recurrent video encoding network that consists of 'Time Boundary-aware LSTM cell' to improve video captioning. In this paper, Akbari et al. [5] have presented a new model for video captioning, which can learn multimodal neuro-symbolic representations using dictionary learning-based methods. This results in a more organized and understandable architecture for the captioning task that integrates modality-specific inductive biases. The model is instinctively able to learn spatial, temporal, and cross-modal relationships in a given video and text pair. Kim et al. [52] have proposed a new framework based on hierarchical DL for facial expression recognition systems. In a hierarchical structure, the function derived from the feature-based appearance network is merged with the geometric feature. Using the preprocessed LBP image, the appearance function-based network extracts holistic features of the face, and the geometric feature-based network learns the Action Units' (AUs') coordinate shift.

Pendurkar et al. [88] have proposed an attention-based multimodal fusion architecture based open-ended video question answering systems. In the data preprocessing stage, VGG16 has been used for visual feature extraction from videos and to obtain word embeddings, where each token from the question is processed through the GloVE model. Attention blended Bidirectional LSTM (Bi-LSTM) encodings have been used for both images (for capturing motion or action-based activities) as well as questions. Wang et al. [130] have proposed Object-aware Semantic Attention (OSA) based captioning model, in which 2D-CNN features are extracted using ResNet-152. Then those are entered into the three branches of the captioning model in different ways to initialize the hidden state of LSTM, adopting the average pooling layer to derive semantic attention, and lastly, to construct semantic concepts based on RPN.

Xia et al. [136] have proposed a multi-layer Transformer as an encoder in their proposed model. Iashin and Rahtu [45] have proposed to use a pre-trained Inflated 3D (I3D) network to represent a visual stream while the tokens (roughly, words) are embedded with pre-trained GloVe. Then again, for the audio streaming purposes, The authors have employed pre-trained VGGNet. These audio and visual sequences are passed through the bi-modal multi-layered encoder of the Transformer to create bi-modal sequence representations using novel bi-modal multi-headed attention blocks to combine the characteristics from both sequences. Huang et al. [43] have analyzed various multimodal sequence-to-sequence pre-training techniques that exploit video and caption-like texts from large unsupervised datasets. It considers pretraining with Multimodal MASS [106] with Transformer as main architecture.

Korbar et al. [55] have proposed to implement their multimodal encoder using the Transformer architecture pre-trained on the "Colossal Clean Crawled Corpus" (C4) while video feature extractor is implemented as an R(2+1) D-34 network pre-trained on the IG65M dataset. Suin and Rajagopalan [109] has proposed an efficient framework in which CNN has been used as an image feature extractor, and then the features are fed into the LSTM state encoder. After concatenating with the historical facts, the encoded state is fed to the agent. Features of all frames selected by the agent are transferred to the visual encoder to achieve better representation. Zhu and Yang [161] have introduced 'ActBert' for self-supervised learning of joint video-text representations from unlabeled data. The litterateurs also have introduced a 'TaNgled Transformer block' to encode three sources of information, i.e., global actions, local, regional objects, and linguistic descriptions. ResNet-101 and ResNet-3D are used for image and spatial-temporal feature extraction. Hao et al. [35] have proposed to use GRU as an encoder.

Sentence Generation

Pan et al. [80] have entrusted a single-layer LSTM-based decoder for video captioning in their paper. Venugopalan et al. [122] have introduced some methods such as 'Early Fusion', 'Late Fusion', 'Deep Fusion' to their proposed model to enrich its prior linguistic knowledge. Their model is an LSTM-based network for video to text generation. Peris et al. [91] have proposed forward-backwards LSTMs to do the work of caption generation. In the paper of [148], the core contribution is introducing a paragraph generator on top of the sentence generator, and these are RNN based. Pasunuru and Bansal [85] have proposed a many-to-many multi-task learning model where LSTM-based encoders-decoders are shared for video captioning, unsupervised video prediction, and entailment generation tasks.

In this section, we have represented the classification of all the deep learning-based video captioning methods. We have reviewed some of the works on video captioning using DL models in this section. Table 1 shows all the research works on video captioning from 2015 to 2020 using different DL methods for feature extraction and description generation included in this survey. From the very beginning, CNN and LSTM have been the main tools for visual feature extraction and description generation, respectively. We have seen different variants of CNN, like AlexNet and VGG-16, for visual feature extraction in different research works in 2015. Venugopalan et al. [121] and Venugopalan et al. [120] have used AlexNet for visual feature extraction and

Table 1 Year Wise Distribution of Papers on Video Captioning Using Deep Learning Methods (till July, * denotes not cited yet)

Year	Authors	Methods			Dataset	Citation
		Image	Spatio-Temporal	Caption Generation		
2020	Chen et al. [16]	ResNeXt-101 + ECN (Efficient Convolutional Network)		Professional Learning	MSVD, MSR-VTT	*
	Wu and Han [134]	GoogLeNet	ResNet-152	LSTM	MSVD, MSR-VTT	*
	Zhu et al. [160]	Mask R-CNN	–	GRU	MSVD, MSR-VTT	*
	Iashin and Rahtu [45]	Mask R-CNN	–	GRU	ActivityNet Captions	1
	Pendurkar et al. [88]	VGG-16	Bi-LSTM	LSTM	TGIF QA	*
	Wang et al. [130]	ResNet	–	LSTM	MSCOCO	*
	Hemalatha and Sekhar [37]	Res-Net152	3D-CNN	LSTM	MSVD, MSR-VTT	*
	Xiao and Shi [137]	InceptionV3	C3D and VGG	LSTM	MSVD, MSR-VTT	2
	Luo et al. [70]	ResNet-152	Res-NeXt-101	Transformer	Youcook2, MSR-VTT	3
	Xia et al. [136]	Multi-layer Transformer		multi-layer Transformer	COCO Captions, Flickr30k	2
	Fang et al. [28]	ResNet-152	LSTM	Transformer	V2C dataset	1
	Lei et al. [60]	ResNet-200	BN-Inception	Transformer	ActivityNet Captions, YouCookII	*
	Iashin and Rahtu [45]	I3D	Transformer	Transformer	ActivityNet Captions	*
	Korbar et al. [55]	I3D	–	Transformer	EPIC-Kitchens, TVQA, TVC, YouCook2, MSR-VTT	*
	Liu et al. [68]	CNN	CNN	RNN with soft attention mechanism	MSVD, MSR-VTT	11
	Jia and Li [48]	ResNext-101	–	LSTM	COCO-CN	*
	Suin and Rajagopalan [109]	ResNet-200	–	LSTM	Activity-Net Captions	*
	Zhao et al. [157]	VGG16	–	LSTM	MSCOCO, FlickrStyle10K	*
	Chen and Jin [17]	ResNet-152	Spatial-temporal ean pooling	RNN model of LSTM cell	TGIF	*
	Zheng et al. [158]	Inception-Resnet-v2(IRV2)	C3D	LSTM	MSVD, MSR-VTT	*
	Zhu and Yang [161]	ResNet-101	ResNet-3D	Transformer	CrossTask, YouCook2, COIN, MSR-VTT	*
	Zhang et al. [155]	Inception-Resnet-v2	C3D	LSTM	MSR-VTT, MSVD, VATEX	*
	Chen et al. [20]	fully convolutional networks + multi-instance learning	–	LSTM	MS COCO, Flickr30K	2
	Aggarwal et al. [3]	ResNet-200	BN-Inception	Masked Transformer	YouCook2	*
	Sah et al. [102]	ResNet-152	–	LSTM	M-VAD, MSR-VTT, MSVD	*
	Wang et al. [126]	VLAD		GRU	M-VAD, MSVD	*
	Huang et al. [44]	ResNet-101 based Faster-RCNN		Two layers SAP-LSTM	MSR-VTT, MSVD	*
	Wang et al. [129]	ResNet-101	-	LSTM	Flickr30K, Microsoft COCO	4
	Pan et al. [79]	ResNet-101	I3D	Transformer	MSR-VTT, MSVD	*
	Hao et al. [35]	GRU		LSTM	MSVD, MPII-MD	*
	Wang et al. [131]	LSTM			ActivityNet-Captions, YouCook2	*
	Rimle et al. [96]	I3D + ResNet-152 + Bi-LSTM		Uni-LSTM	News Video Dataset, Large Scale Movie Description Challenge	*
	Kim et al. [51]	3D-CNN		LSTM	ActivityNet-Captions	
	Tan et al. [113]	2D-CNN + R-CNN	3D-CNN	LSTM	MSVD, MSR-VTT	*
	Jin et al. [49]	Sparse Attention		Transformer	MSVD, MSR-VTT	*

Table 1 (continued)

Year	Authors	Methods			Dataset	Citation
		Image	Spatio-Temporal	Caption Generation		
	Sur [111]	Self-aware Composition Transformer		Transformer	Activi-tyNet Captions,YouCook2	1
	Hou et al. [42]	Attention Model + Graph Convolutional Network		LSTM	MSVD, MSR-VTT	*
	Zhu et al. [159]	MaskTrack-RCNN		LSTM	Object-Oriented Captions	*
	Rahman et al. [93]	VGG-16 + (two consecutive) Bi-LSTM		LSTM	MSVD	*
2020	Hou et al. [42]	Inception-Resnet-v2	C3D	-	MSVD, MSR-VTT	*
	Zhang et al. [153]	InceptionV4		RNN-like network	MSVD	*
	Perez-Martin et al. [89]	2D-CNN	3D-CNN	LSTM-like network	MSVD, MSR-VTT	1
	Akbari et al. [5]	–	–	-	YouCook II, ActivityNet	*
	Perez-Martin et al. [90]	2D-CNN	3D-CNN	compositional LSTM	MSVD, MSR-VTT	1
	Fawzy et al. [29]	ResNet-18, 3D ResNeXt-101,Bi-LSTM		DSC-N	MSR-VTT	1
2019	Zhang and Peng [150]	GoogleNet,16-layer VGG,Faster RCNN	–	LSTM (doublemem-ory cells + hidden states)	MSVD	7
	Xu et al. [138]	–	C3D	Hierarchical LSTM (high-low level controllerLSTM)	ActivityNet Captions, TACoS Multilevel	19
	Park et al. [84]	ResNet- 152	ResNet- 101 (R3D)	LSTM, GAN	ActivityNet Captions	3
	Yan et al. [143]	GoogleNet,Faster RCNN	C3D	LSTM	MSVD, MSR-VTT-10K	67
	Aafaq et al. [2]	2D-CNN	3D-CNN	GRU	MSVD, MSR-VTT	*
	Guo et al. [33]	LSTM			MSVD, MSR-VTT	8
	Hou et al. [41]	CNN	3D-CNN	LSTM	MSVD, MSR-VTT, Activi-tyNet Captions	2
	Li and Gong [61]	Inception-ResNet-v2	Inception-ResNet-v2	LSTM	MSVD, MSR-VTT	14
	Mun et al. [76]	(C3D + GRU)	RNN		COCO	8
	Olivastri et al. [77]	GoogLeNet	Inception-ResNetv2	Soft-Attention LSTM(SA-LSTM)	MSVD, MSR-VTT	1
	Pei et al. [87]	ResNet-101 [2DFea-ture]	ResNeXt-101 [3DFea-ture]	GRU + Attention-based Recurrent Decoder	MSVD, MSR-VTT	10
	Wang et al. [124]	CNN	LSTM	Two-layer LSTM	MSVD, MSR-VTT	4
	Zhang and Peng [151]	CNN	LSTM	Two-layer LSTM	MSVD, MSR-VTT	16
	Zhang et al. [154]		C3D	LSTM	ActivityNet Captions	*
	Zhao et al. [156]	Co-attention model based RNN		LSTM	MSVD, Charades, MSR-VTT, MPII-MD	4
	Li et al. [63]	CNN	–	Residual attention-based LSTM	MSVD, MSR-VTT	11
	Aafaq et al. [1]	2D-CNN	3D-CNN	GRU	MSVD, MSR-VTT	32
	Chen et al. [18]	CNN		Temporal Attention Mechanism	MSVD, MSR-VTT	18
	Chen and Jiang [19]	GoogleNet + Incep-tion-Resnet-V2	C3D	LSTM	MSVD, MSR-VTT	12
	Wang et al. [133]	3D-CNN + Bi-LSTM		Attention Based LSTM	ActivityNet-Captions	17

Table 1 (continued)

Year	Authors	Methods			Dataset	Citation
		Image	Spatio-Temporal	Caption Generation		
2018	Li et al. [64]	C3D		LSTM	ActivityNet Captions	45
	Song et al. [105]	ResNet-152	–	One-layer LSTM	MSVD, MSR-VTT	100
	Wang et al. [123]	Inception V4	–	LSTM	MSVD, MSR-VTT	74
	Wang et al. [127]	LSTM			ActivityNet-Captions	79
	Wu et al. [135]	ResNet + LSTM		LSTM	MSVD, Charades	32
	Yang et al. [144]	Joint LSTMs + Adversarial Learning			MSVD, MSR-VTT, M-VAD, MPII-MD	73
	Wang et al. [128]	2D-CNN	3D-CNN	LSTM	MSVD, MSR-VTT	54
	Xu et al. [141]				MSVD, MSR-VTT, MPII-MD	27
	Long et al. [69]	ResNet-152	C3D	LSTM	MSVD, MSR-VTT	40
2017	Pan et al. [82]	19-layer VGG	C3D	LSTM with Transfer Unit	MSVD, MPII-MD, M-VAD	158
	Hori et al. [40]	VGG-16	C3D	Bi-directional LSTM	YouTube2Text, MSR-VTT	142
	Yang et al. [145]	GoogleNet, VGG-16	–	LSTM	MSVD, M-VAD	38
	Baraldi et al. [11]	ResNet-50	C3D	Time Boundary-aware LSTM network	MSVD, MPII-MD	103
	Pasunuru and Bansal [85]	LSTM	–	Unidirectional LSTM-RNN	MSVD, MSR-VTT, M-VAD, UCF-101	59
	Zhang et al. [152]	VGG-19, GoogleNet-bu4k	C3D	LSTM	MSVD, MSR-VTT-10K	28
	Wang et al. [132]	ResNet-152	–	Low-level Bi-LSTM encoder and high-level LSTM	MSR-VTT, Charades Captions	76
	Shen et al. [104]	ResNet-50	C3D	LSTM	MSR-VTT	73
2016	Ballas et al. [8]	VGG-16		LSTM with soft attention	UCF-101, Youtube2Text	292
	Guo et al. [34]	GoogleNet	C3D	LSTM	UCF-101, Youtube2Text	39
	Zhang and Tian [149]	C3D + CCA (Canonical Correlation Analysis)		LSTM-based RNN	MSVD, MPII-MD,M-VAD	13
	Bin et al. [12]	VGG-16	–	LSTM	MSVD, COCO2014	29
	Pan et al. [80]	GoogleNet	–	LSTM	MSVD, M-VAD	271
	Venugopalan et al. [122]	CNN + LSTM	–	LSTM	MSVD, MPII-MD,M-VAD	96
	Pan et al. [81]	19-layer VGG	C3D	LSTM-type RNN	MSVD, MPII-MD,M-VAD	334
	Peris et al. [91]	GoogleNet		LSTM	MSVD	24
	[148]	VGG-Net	C3D	RNN with the GRU (or the gated RNN)	YouTubeClips, TACoS-Multi-Level	376
2015	Yu et al. [146]	GoogleNet	–	LSTM	Youtube 2Text, DVS	716
	Yao et al. [26]	CNN		LSTM	UCF-101, Flick30k, COCO2014, TACoS Multilevel	3802
	Donahue et al. [121]	Alexnet, 16-layer VGG	–	Stack of two LSTM	MSVD, MPII-MD,M-VAD	829
	Rohrbach et al. [98]	DT (Dense Trajectories), LSDA (Large Scale Object Detector), and PLACES (Places-CNN)		LSTM	MPII-MD, M-VAD	90
	Venugopalan et al. [121]	Caffe (variant of AlexNet)	–	Two-layer LSTM	MSVD, Flick30k, COCO2014	657

two-layer LSTM for description generation which showed top result till 2015. Yao et al. [146] have used GoogleNet for visual feature extraction and used LSTM for description generation. All those works are mostly on MSVD, M-VAD, and MPII-MD datasets. We have seen much improvement in the spatio-temporal feature extraction sector with the use of 3D-CNN or C3D in 2016. We have also seen VGG-19, which has some extra layers compared to VGG-16 showing better results in visual feature extraction in Pan et al. [81]. This research work also has used C3D for spatio-temporal features and LSTM for description generation. Also, Venugopalan et al. [122] have shown a combined structure of CNN and LSTM for visual feature extraction. Ballas et al. [8] have also introduced soft attention mechanisms for description generation, which improved the results more. In the year 2017, C3D has become more popular for spatio-temporal feature extraction and research works like [11, 40, 82, 152, 104], etc. have shown impressive results using C3D. Another variant of CNN called Residual Network (ResNet) has been introduced in 2017 for visual feature extraction purposes with varying amounts of layers like ResNet-50 [11] and ResNet-152 [132]. Bidirectional LSTM [40] has also shown better results for description generation this year.

In 2019, the Inception model, Inception-ResNet-v2 [61] had been introduced for both visual and spatio-temporal feature extraction and had shown better results. In the same year, a groundbreaking model for video captioning has been demonstrated in Yan et al. [143] where GoogleNet and Faster RCNN have been used for visual feature extraction, C3D for spatio-temporal feature extraction, and LSTM for description generation showing better performance for video captioning. The year 2020 has introduced us to much more improved models. We have seen that the enhanced model of ResNet (ResNet-200) has been used [60]. In additon, Mask RCNN improved visual feature extraction accuracy [160]. The most significant improvement that has been made in 2020 is the description generation model or language model. The attention mechanism-based language model, Transformer, became the state-of-the-art description generation model and showed better results than before.

This section depicts different DL approaches toward the video captioning system by dissecting models into two parts, Feature Extraction and Sentence Generation. A good number of state-of-the-art works are discussed here for each part dividing them into CNN-RNN based methods and RNN-RNN based methods.

Benchmark Datasets

Some of the most popular benchmark datasets from old times have been analyzed along with their attributes. At the beginning of video data collection, the datasets were created with videos of specific categories like cooking, makeup, movies, etc. Some of them are M-VAD [116], MPII-MD [99], YouCook-II, MP-II Cooking, and so on. The video captioning or description generation methods lacked the proficiency of operating successfully with videos of all domains due to the lack of diversity in datasets. However, with time going on, people realized that the models' full potential is not explored because of the specific domain-based datasets. Some of the open domain datasets made a breakthrough in this field, and some of them are MSVD [14], MSR-VTT [139], HowTo100M, and many others. In the following, we also have shown statistical analysis among the datasets based on different parameters, for example, average sentence per video clip in Fig. 10, the total number of video clips, total vocabulary, total hours of videos, etc. which also helped us to conclude about datasets shown in Table 2.

MSR-VTT

MSR-Video to Text (MSR-VTT) is one of the largest video captioning datasets based on the number of clips with multiple associated sentences consisting of 7180 videos subdivided into 10,000 images [139]. The clips are divided into 20 different categories which are shown in Fig. 8. The dataset is divided into 6513 training, 497 validation, and 2990 test videos. The training, testing, and validation data ratio is shown in Fig. 9a. Each video comprises 20 reference captions annotated by AMT workers. All videos in MSR-VTT are not uniform in characteristics, as some of the videos consist of multiple scenes and some of them have a single scene.

TACoS-Multi-Level

TACoS-Multi-Level [97] corpus were annotated by AMT staffs on the TACoS corpus [95]. For every video in the TACoS corpus, three types of descriptions were compiled, such as -

- a detailed summary of the video with no more than 15 sentences per video
- a summary of 3–5 sentences per video; and, finally
- a single sentence description of the video

Compilation of three types of descriptions approach made a breakthrough in the limitations of "Automatic Video Description Generation" tending to generate a single sentence for a single frame or single level of details. The detailed description portion helps to generate multiple sentences per frame.

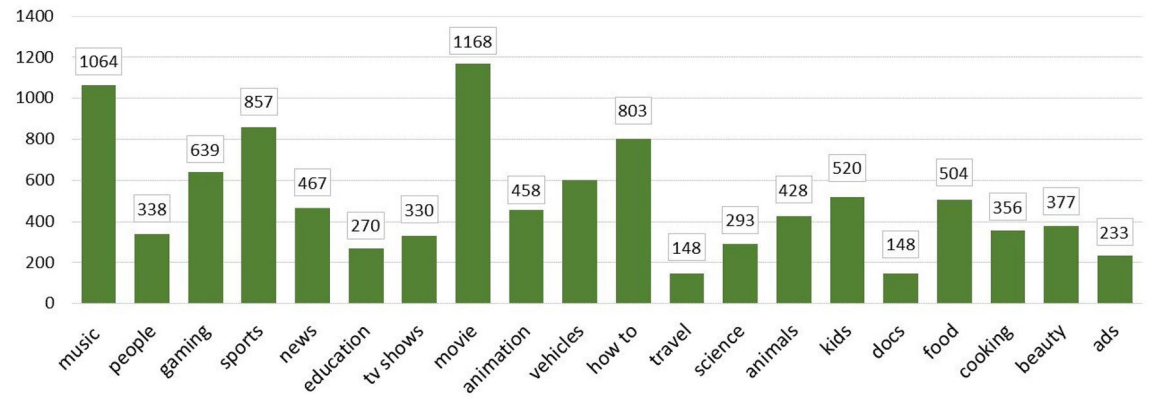
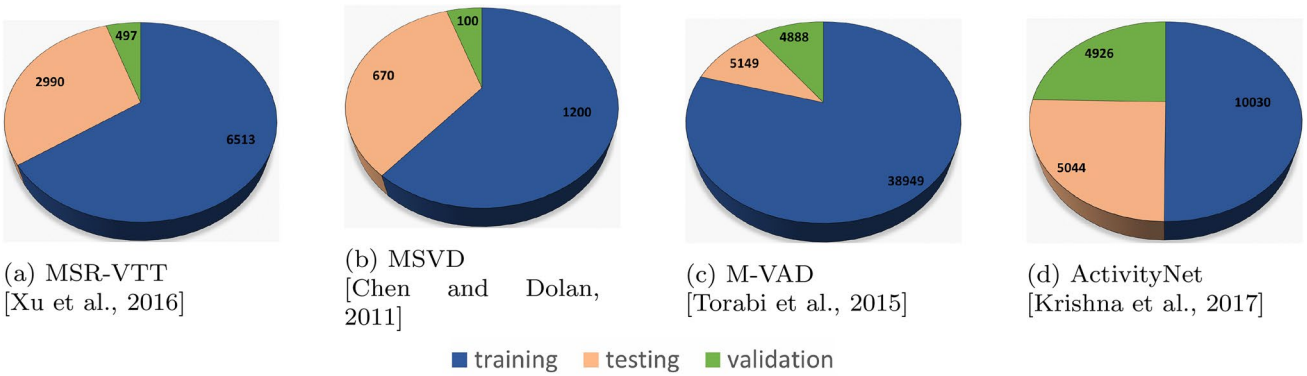


Fig. 8 Distribution of Videos from MSR-VTT dataset in Twenty Categories

Table 2 Information about Widely used Datasets

Dataset	Context	Total videos	Total clips	Avg. clip length (s)	Total video (h)	Total sentences	Total words	Vocabulary
M-VAD	Movie	92	48,986	6.2	84.6	55,904	519,933	17,609
MSVD	Various/open	1970	1970	10	5.3	70,028	667,339	13,010
MPII-MD	Movie	94	68,337	3.9	73.5	68,375	653,467	24,549
TACoS-multi-level	Cooking	185	14,105	360	27.1	52,593	–	–
MSR-VTT	Open	7180	10,000	20	41.2	200,000	1,856,523	29,316
ActyNet Cap.	Open	20,000	20,000	180	849	100,000	1,348,090	–



(a) MSR-VTT [Xu et al., 2016]

(b) MSVD [Chen and Dolan, 2011]

(c) M-VAD [Torabi et al., 2015]

(d) ActivityNet [Krishna et al., 2017]

Fig. 9 Distribution of Videos from Different Dataset in Training, Test, and Validation set

MSVD

The Microsoft Video Definition (MSVD) dataset [14] consists of 1970 YouTube videos with human-annotated sentences. The data set was also annotated by AMT staff. The data set contains multilingual human-generated captions. Nearly all research groups have split this dataset into training, validation, and testing partitions of 1200, 100, and 670 videos, respectively, and that ratio is shown in Fig. 9b. In all images, the audio is muted. Besides, videos containing subtitles or overlaid text have also been omitted. Usually, the length of each video in this dataset is between 10 and 25 s, much of which indicates only one activity. On average, for each clip, there are 41 single sentence descriptions.

MPII-MD

MPII—Movie Description Corpus includes [99] transcribed audio descriptions taken from 94 Hollywood movies. These movies are categorized into 68,337 clips with an overall

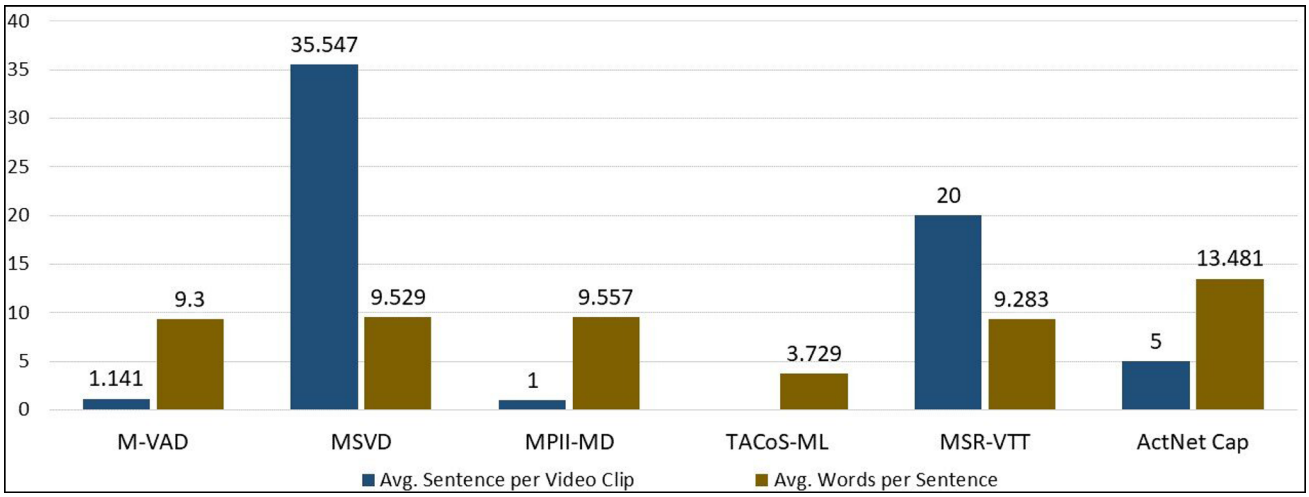


Fig. 10 Average Sentence and Word Comparison of Different Popular Dataset

duration of 3.9 s combined with 68,375 sentences with only one sentence per clip. Every clip is combined with a sentence taken from the movie script and the audio description details. The cumulative time of the dataset videos is almost 73.6 h, and the vocabulary scale is 653,467.

M-VAD

Montreal Video Annotation Dataset (M-VAD) is based on the Descriptive Video Service (DVS) and includes 48,986 video clips from 92 different films [116]. Each clip has an average period of 6.2 s, and the maximum time for the entire data set is 84.6 h. The total number of sentences is 55,904, with few clips linked to more than one sentence. The dataset split consists of 38,949, 4,888, and 5,149 video clips for training, evaluation, and testing purposes, respectively, shown in Fig. 9c.

ActivityNet Captions

This dataset is based on different human activity divided into 200 classes and consists of about 20,000 videos which corresponds to approximately 890 h. ActivityNet Captions dataset [56] has about 100,000 sentences to describe all 20,000 videos including 10030 training videos, 4926 validation videos and 5044 test videos.The dense descriptions of the videos cover 94.6% contents of the corresponding video.

A representation of the average number of sentences and word comparison of the datasets is illustrated in Fig. 10. The distribution (training, testing, and validation) of videos Fig. 9d is depicted following. Also, a brief categorical and statistical information of the dataset is represented in Table 2.

Flickr30k

This data set includes 31,783 photographs of daily activities, events, and scenes (all harvested from Flickr) and 158,915 captions (acquired through crowdsourcing). This includes and expands the corpus of 8,092 images taken by Hodosh et al. [39]. Five annotators unfamiliar with the actual individuals and circumstances represent each image separately, resulting in different captions than the photographs' owners. Besides, multiple annotators use varying degrees of detail, from defining the overall situation to particular actions. This variety of definitions associated with the same picture enables denotational similarities between expressions not trivially related by syntactic rewrite laws.

Throughout this section, we have primarily discussed some of the popular datasets in the field of video captioning. We can see in the Fig. 11 that MSVD potentially outperforms MSR-VTT in almost every evaluation metrics. Again, by far, MSVD performs best with the 'CIDEr' evaluation metric. Besides that, Table 2 shows overall statistics such as the context of the dataset, the total number of video clips, average clip length in seconds, total number of sentences, and vocabulary of all the datasets included in this work.

A combination of different models has been used as a feature extractor and description generator for video captioning at different times. The performance of a video captioning system also depends on datasets. Table 3 shows the year-wise distribution of experimental results of various video captioning models on different datasets measured by the four most popular evaluation metrics. Among all the research works included in this survey, the most used dataset is MSVD. In 2017, Pasunuru et al. [85] had performed probably best with the MSVD dataset in all the evaluation metrics. In the same year Shen et al. [104] brings out quite similar

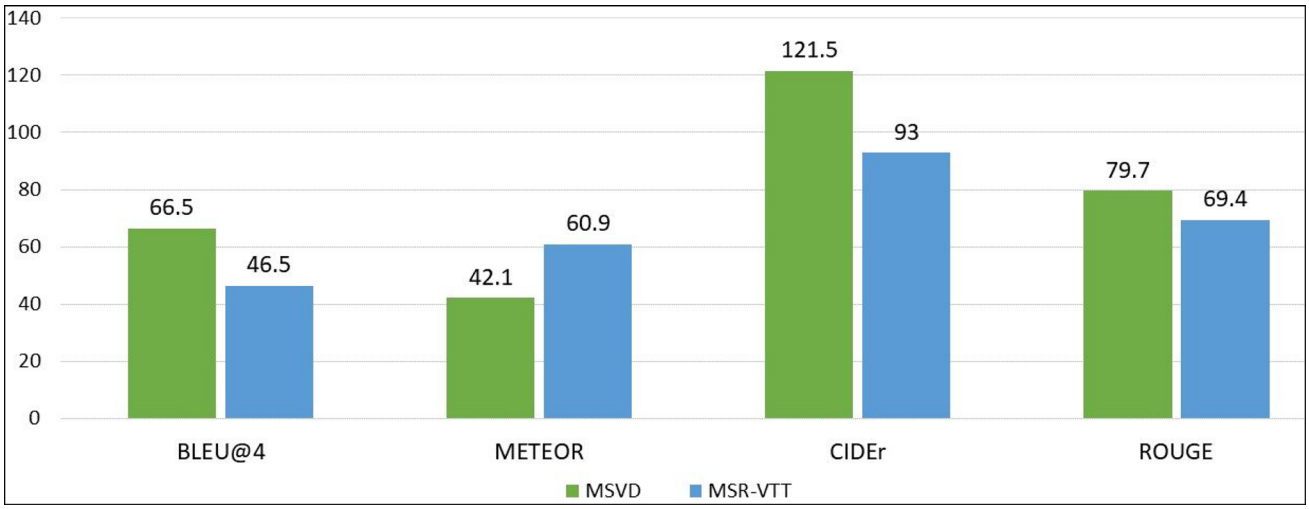


Fig. 11 Performance Comparison Between MSVD and MSR-VTT Dataset Based on Deep Learning Methods

results in all evaluation metrics using the MSR-VTT dataset. With the improvement of LSTM (two layers), [41, 150], and Zhang et al. [155] show some impressive results with the MSVD dataset in 2019. Zhang et al. [155] shows by far the best results in BLEU@4, CIDEr, and METEOR outperforming all the previous video description models. In 2020 (till July), Chen et al. [16] shows possibly the best results of all time with the MSVD dataset. Many other datasets have been used. Rather than MSVD, Xiao et al. [137] shows conceivably the best result using the MSR-VTT dataset.

Evaluation Metrices and Performance Analysis

Evaluation metric is incomparable in assessing the overall quality of all types of models against ground truth. In video captioning, evaluation metrics play a critical role in measuring how close a model's video captioning is to human annotation. In this section, we have thoroughly described the five most important metrics–Bilingual Evaluation Understudy (BLEU), Consensus-based Image Description Evaluation (CIDEr), Metric for Evaluation of Translation with Explicit ORder-ing (METEOR), Recall-Oriented Understudy for Gisting Evaluation(ROUGE), and Semantic Propositional Image Captioning Evaluation (SPICE). Furthermore, the benchmark results of the studied paper are represented in Table 3.

Bilingual Evaluation Understudy (BLEU)

BLEU is a famous metric, very first introduced in 2002, used to measure the quality of the text generated by the machine [83]. The quality measures the similarity between human outputs and outputs generated by machines. According to BLEU, a high-scoring description should correspond in length to the ground-truth sentence such as the exact match of words as well as their order. BLEU scores take into consideration the similarity between predicted unigrams (single word) or higher n-gram and a set of reference sentences of one or more candidates. BLEU evaluation will show '1' as an output score when an exact match occurs.

Consensus-Based Image Description Evaluation (CIDEr)

CIDEr [119] was first introduced in 2015 and mostly used for evaluating image caption quality. It evaluates the consensus of the corresponding image between a predicted sentence and the reference sentences. Each sentence is treated by CIDEr as a collection of n-grams containing 1 to 4 words. All the words in a sentence are first converted to their stem or root. Encode the consensus between the sentence expected and the reference sentence; it measures the frequency of n-grams co-existing in both sentences. For each n-gram, the weight is calculated using the Term Frequency-Inverse Document Frequency (TF-IDF).

Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

In 2004, ROUGE [65] metric was introduced to determine text summaries. Just like BLEU, ROUGE is also calculated by varying the count of the n-gram. Unlike precision-based BLEU, however, ROUGE is based on the recall values. It calculates the recall score of the generated sentences corresponding to the reference sentences using n-grams. There are different parts of ROUGE for different works. ROUGE-N

Table 3 YEAR-WISE DISTRIBUTION OF RESULTS OF DIFFERENT POPULAR DATASETS USED FOR VIDEO CAPTIONING MODELS

Year	Dataset	Models	METEOR	CIDEr	ROUGE	BLEU@4
2020	MSVD	ResNeXt-101 + ECN (Efficient Convolutional Network) [16]	42.1	–	79.7	66.5
		Inception-Resnet-v2 + C3D [42]	34.7	80.1	71.5	47.9
		InceptionV4 + LSTM-based Guidance Module [153]	33.5	83.1	70.7	52.1
		2D-CNN + 3D-CNN + compositional decoder [89]	39.2	107.7	78.3	62.3
		2D-CNN + 3D-CNN + compositional LSTM [90]	41.9	111.5	79.5	64.4
		GoogLeNet + ResNet-152 + LSTM [134]	26.9	41.7	–	37.5
		Mask R-CNN + GRU [160]	20.0	50.4	45.1	20.2
		Res-Net152 + 3D-CNN + LSTM [37]	34.7	76.0	73.1	50.1
		InceptionV3 + C3D and VGG + LSTM [137]	36.1	85.8	72.8	54.0
		CNN + RNN (with soft attention mechanism) [68]	34.8	88.2	71.7	54.2
		Inception-Resnet-v2(IRV2) + C3D + LSTM [158]	33.5	81.0	69.4	46.5
		Inception-Resnet-v2 + C3D + LSTM [155]	36.4	95.2	73.9	54.3
		ResNet-152 + LSTM [102]	33.2	74.7	69.3	49.9
		VLAD + GRU [126]	34.9	88.2	–	56.1
		ResNet-101 based Faster-RCNN + Two layers SAP-LSTM [44]	35.4	90.8	72.0	53.3
		ResNet-101 + I3D + Transformer [79]	36.9	73.9	93.0	52.2
		GRU + LSTM [35]	29.3	50.9	–	39.5
		2D CNN + R-CNN + 3D CNN + LSTM [113]	36.5	94.4	73.4	54.6
		Sparse attention + Transformer [49]	35.3	89.5	72.3	53.1
		GCN + LSTM [42]	34.7	80.1	71.3	47.9
		VGG-16 + Bi-directional LSTM + Single layer LSTM [93]	–	–	–	18.86
	MSR-VTT	InceptionV3 + C3D and VGG + LSTM [137]	28.7	48.9	62.3	44.7
		I3D + Transformer [68]	28.5	–	–	41.7
		Inception-Resnet-v2 + C3D [42]	27.9	45.3	60.1	40.4
		2D-CNN + 3D-CNN + compositional decoder [89]	31.4	50.6	64.3	45.5
		2D-CNN + 3D-CNN + compositional LSTM [90]	30.1	48.0	63.1	45.6
		ResNet-18 +3D ResNeXt-101 + Bi-LSTM + DSC-N [29]	–	96.4	–	60.0
		CNN + RNN (with soft attention mechanism) [68]	27.5	47.5	60.2	40.9
		Inception-Resnet-v2(IRV2) + C3D + LSTM [158]	33.5	81.0	69.4	46.5
		Inception-Resnet-v2 + C3D + LSTM [155]	–	50.9	62.1	43.6
		ResNet-152 + LSTM [102]	28.2	46.4	60.4	41.1
		ResNet-101 based Faster-RCNN + Two layers SAP-LSTM [44]	28.4	48.5	61.4	41.7
		ResNet-101 + I3D + Transformer [79]	60.9	28.3	47.1	40.5
		2D CNN + R-CNN + 3D CNN + LSTM [113]	28.4	49.6	61.6	42.5
		Sparse attention + Transformer [49]	28.9	51.6	61.5	42.9
		GCN + LSTM [42]	27.9	45.3	60.1	40.4

Table 3 (continued)

Year	Dataset	Models	METEOR	CIDEr	ROUGE	BLEU@4
	ActivityNet Captions	Mask RCNN + GRU [45]	11.72	–	–	2.86
		RestNet-200 + BNInception + Transformer [60]	15.57	22.16	5.44	9.78
		multimodal neuro-symbolic representations using dictionary learning [111]	9.78	29.68	20.42	4.01
		I3D + Transformer [45]	8.44	–	–	1.88
		ResNet-200 + LSTM [109]	5.7	11.68	–	1.1
		LSTM [131]	10.58	39.73	–	–
		3D CNN + LSTM [51]	9.57	28.03	–	4.37
		Self-aware composition Transformer + Transformer [111]	2.95	–	–	5.1
2020	MPII-MD	GRU + LSTM [35]	6.1	10.1	14.6	0.6
	YouCook2	LSTM [131]	13.65	–	–	–
		Self-aware composition Transformer + Transformer [111]	7.34	–	–	0.48
		multimodal neuro-symbolic representations using dictionary learning [111]	10.19	50.22	28.17	4.50
2019	TACOS-MULTILEVEL	JEDDi-Net [138]	23.9	104.0	50.9	18.1
	ActivityNet Captions	C3D + LSTM [154]	10.33	12.93	21.21	2.09
		ResNet-152 + ResNext-101 (R3D) + (LSTM, GAN) [84]	16.48	20.60	–	9.91
		3D CNN + I3D + Bi-directional LSTM + attention-based LSTM [133]	9.96	28.23	21.17	3.68
2019	MSVD	AGHA [150]	35.3	83.3	–	55.1
		STAT [143]	33.5	73.8	–	52.0
		GoogleNet + VGG + Faster RCNN + LSTM [150]	35.3	83.3	–	55.1
		GoogleNet + Faster RCNN + C3D + LSTM [143]	33.3	73.8	–	52.0
		LSTM + LSTM [33]	34.3	75.9	–	52.7
		Inception-Resnet-v2 + LSTM [61]	34.1	87.5	70.8	50.3
		GoogLeNet + Inception-Resnet-v2 + Soft-Attention LSTM [77]	32.2	76.4	69.1	48.3
		ResNet-101(2D Feature) + ResNeXt-101(3D Feature) + GRU with Attention-based Recurrent Decoder [87]	28.1	47.1	60.7	40.4
		CNN + LSTM + Two-layer LSTM [124]	34.9	91.0	72.1	53.9
		CNN + LSTM + Two-layer LSTM [151]	36.2	90.6	–	56.9
		co-attention model based RNN + LSTM [156]	33.4	54.3	69.4	42.4
		CNN + Residual attention-based LSTM [63]	34.3	72.9	–	53.4
		2D CNN + 3D CNN + Neuron-wise Short Fourier Transform + fully-connected layer + multi-layer GRU [1]	35.0	78.1	71.5	47.9
		CNN + temporal deformable convolutional encoder + convolutional decoder + temporal attention mechanism [18]	33.8	76.4	–	53.3
		GoogleNet + Inception-Resnet-v2 + C3D + LSTM [19]	35.0	86.7	–	53.4
	MSR-VTT	STAT [143]	27.1	44.0	–	39.3
		2D CNN + 3D CNN + Neuron-wise Short Fourier Transform + fully-connected layer + multi-layer GRU [1]	28.4	48.1	60.7	38.3
		CNN + temporal deformable convolutional encoder + convolutional decoder + temporal attention mechanism [18]	39.5	42.8	–	38.3
		GoogleNet + Inception-Resnet-V2 + C3D + LSTM [19]	27.6	47.5	–	42.4
		3D CNN + I3D + Bi-directional LSTM + attention-based LSTM [133]	29.4	48.9	62.0	42.2

Table 3 (continued)

Year	Dataset	Models	METEOR	CIDEr	ROUGE	BLEU@4
2018	MSVD	RecNetlocal [123]	34.1	80.3	69.8	52.3
		ResNet + LSTM + LSTM [135]	34.0	74.9	–	51.7
		Joint LSTMs with adversarial learning [144]	30.4	–	–	42.9
		2D CNN + 3D CNN + LSTM [128]	33.31	–	–	52.82
		Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM) [141]	34.7	79.4	65.9	53.0
		ResNet-152 + C3D + LSTM [69]	33.5	72.1	–	52.0
	MSR-VTT	RecNetlocal [123]	26.6	42.7	59.3	39.1
		Joint LSTMs with adversarial learning [144]	26.1	–	–	36.0
		2D CNN + 3D CNN + LSTM [128]	26.58	–	–	38.13
		Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM) [141]	29.4	46.1	62.3	42.3
		ResNet-152 + C3D + LSTM [69]	26.7	–	–	39.1
	ActivityNet	LSTM [127]	9.60	12.68	19.10	2.30
2018	Charades	ResNet + LSTM + LSTM [135]	17.8	20.8	–	13.5
	M-VAD	Joint LSTMs with adversarial learning [144]	6.3	–	–	–
	MPII-MD	Joint LSTMs with adversarial learning [144]	7.2	–	–	–
		Attentive Multi-Grained Encoder (LSTM) + Dual-Stream Decoder (LSTM) [141]	7.9	–	–	1.9
2017	MSR-VTT	Attension-Fusion [40]	25.7	40.4	–	39.4
		Video prediction+ Entailment generation [85]	28.8	74.1	60.2	40.8
		TDDF (VGG+C3D) [152]	27.8	43.8	59.2	37.3
		Dense video captioning framework [104]	29.4	50.5	62.6	44.2
2017	MSVD	LSTM-TSAIV [82]	33.5	74.0	–	52.8
		Attension-Fusion [40]	34.3	72.4	–	56.8
		DMRM + in5b(g) + DA with SS [145]	33.3	74.8	–	51.1
		Boundary-aware-encoder [11]	32.4	63.5	–	42.5
		Video prediction + Entailment generation [85]	36.0	92.4	72.8	54.5
		TDDF (VGG + C3D) [152]	33.3	73.0	69.7	45.8
2016	MSVD	GoogleNet + Bi-directional GRU-RCN encoder [8]	31.70	68.01	–	49.63
		aLSTM [34]	30.38	61.20	–	44.87
		C3D (MHI + RGB) – Joint [149]	31.1	–	–	-
		Joint-BiLSTM [12]	30.3	–	–	-
		HRNE with attention [80]	33.1	–	–	43.8
		Glove + Deep Ensemble [122]	31.4	–	–	42.1
		LSTM-E (VGG + C3D) [81]	31.0	-	-	45.3
		Objects + BLSTM [91]	32.6	67.2	-	53.6
		h-RNN [148]	32.6	65.8	-	49.9
	M-VAD	C3D (MHI + RGB) – Joint [149]	6.7	-	-	-
		HRNE with attention Pan et al. [80]	6.7	-	-	-
		Glove + Deep Ensemble [122]	6.7	-	-	-
		LSTM-E (VGG+C3D) [81]	6.7	-	-	-
	MPII-MD	C3D (MHI + RGB) – Joint [149]	7.0	-	-	-
		Glove + Deep Ensemble [122]	6.8	-	-	-
		LSTM-E (VGG+C3D) [81]	7.3	-	-	-
	TACOS-MULTILEVEL	h-RNN [148]	28.7	160.2	-	30.5

Table 3 (continued)

Year	Dataset	Models	METEOR	CIDEr	ROUGE	BLEU@4
2015	MSVD	Enc-Dec (Basic) + Local + Global [146]	29.6	51.67	-	41.92
		S2V-RGB(VGG) + FLOW(AlexNet) [121]	29.8	-	-	-
		LSTM-YT [120]	29.07	-	-	33.29
	M-VAD	S2V-RGB(VGG)+FLOW(AlexNet) [121]	6.7	-	-	-
	MPII-MD	S2V-RGB(VGG)+FLOW(AlexNet) [121]	7.1	-	-	-
		Visual-Labels [98]	7.03	9.98	16.02	0.80
	TACOS-MULTILEVEL	LSTM decoder with CRF prob [25]	-	-	-	28.8

measures overlap. It calculates uni-gram, bi-gram, tri-gram, and higher-order overlap. ROUGE-L measures the longest matching sequence of words in between two or more strings using Longest Common Subsequence (LCS). ROUGE-S looks in a sentence for a pair of words in order.

Metric for Evaluation of Translation with Explicit Ordering (METEOR)

METEOR was first introduced in 2005, which claimed to overcome BLEU's deficiencies. Instead of the exact lexical match that BLEU requires, METEOR introduced semantic matching [9]. METEOR matches the word with the same stem and also with synonyms. The calculation of METEOR's score is based on how well the generated and reference sentences match. Each sentence is taken as a set of unigrams, and mapping candidate unigrams and reference sentences perform an alignment. A unigram in the candidate sentence (or reference sentence) should map to either unigram in reference sentence (or candidate sentence) or zero during mapping. The alignment configuration with fewer crossings is preferred in case of multiple options available for alignments between the two sentences.

The performance of a model always varies with the use of different datasets. Table 3 shows video description results of different models from 2015 to 2020 (till September) by four different evaluation metrics. Some of the popular datasets over the last five years are M-VAD, MPII-MD, MSVD, TACoS-Multi-Level, ActivityNet, and so on. Among all these, MSVD is the mostly used dataset and performs potentially best in different video captioning models [32]. The accuracy of the captioning models mostly depends on the number of natural language sentences provided per video clips in a dataset. Video clip to sentence ratio and sentence to word ratio of some popular datasets are shown in Fig. 10. It shows that MSVD has the highest video-clip to sentence ratio and has 41 sentences per video clip. Also, it is an open domain dataset having a large number of activity classes which enables a video captioning model to recognize more activities.

We found that the MSR-VTT dataset has also been largely used throughout the years 2015 to 2020 besides MSVD. Table 2 shows that the MSR-VTT dataset has more video clips, sentences, and vocabulary compared to MSVD. Nevertheless, Fig. 11 clearly shows that MSVD has probably outperformed MSR-VTT in almost every evaluation metric using different models. According to our research, the language model or video caption generator has worked possibly best with a more significant number of natural language sentences per video clip and MSVD has a higher number than MSR-VTT. Another finding caught our eyes that, MSVD has an average clip length of 10 s, having a single event per video clip. However, average clip length in MSR-VTT is 20s leading it to have overlapping of multiple events which makes it challenging for spatio-temporal feature extractor to distinguish among multiple events. Other datasets like TACoS-Multi-Level, ActivityNet, M-VAD, MPII-MD; all these datasets could not bring out the better result as MSVD or MSR-VTT. Our finding is that the domain of the datasets played a vital role here. Except for ActivityNet, all other datasets are of specific domains. For example, M-VAD and MPII-MD are movie dataset, and TACoS-Multi-Level is a cooking dataset. Domain-specific datasets have very few activity classes that limit the overall potential of a model.

Applications and Challenges on Video Captioning Techniques

Nowadays, the scope of video captioning has increased quite a lot because of the enormous improvement in technology and captioning algorithms. Despite so many enhancements, video captioning still faces many difficulties, and the remaining works also lack the potential to be prepared to solve the real-time usable product. In this section, we will try to comprehensively describe some indispensable applications and probable challenges of the video captioning techniques.

Applications of Video Captioning

Typically, video captioning seems to be only a subject of research on AI. But, it has a vast range of applications in practical life, from mitigating different social problems to ease various technological difficulties.

Aiding Visually Impaired Persons

Video captioning models translate the visual information of a video into natural language outputs. These models can be used for real-time caption generation, where the models would be capable enough to describe the surroundings in natural language. The whole process can be immensely beneficial for visually impaired people as they need a real-time navigation system. The most crucial step for the navigation system is to understand the visual information around the user, where video captioning comes in handy. Generated texts can then be easily converted to audio output for the navigation system.

Helping Speech Impaired People

Interaction with speech-impaired and deaf through the understanding of sign language can be another great example of video captioning. All the previous works on sign language [10] were able to generate corresponding words from hand gestures or hand poses; nothing more. The video captioning model can generate meaningful and complete sentences from sign language, which ultimately results in more eased interaction.

Human Robot Interaction

Using dense video captioning models, instruction sets can be generated in different language text form from directional or instructional videos [64]. With a little advancement, video captioning models can be used to prepare documentation or event details also. These applications will increase the human-robot interactions through which humans will be able to understand the surroundings from the machines' point of view. Again, visual dialogue [22] can be instrumental in the field of human-robot interaction. An intelligent model holds a meaningful conversation with human agents about a video fed into it in a visual dialogue. It is a flourishing research field, and there is scope for a lot of improvement here.

Video Surveillance

Surveillance videos are one of the prime sources of unstructured big data. A lot of works are available on suspicious activity detection [86] from CCTV videos. This tedious and time-consuming job of surveillance can be automated with the help of video captioning. Analysing different types of abnormal activities with object and action detection can be aided through automatic video captioning.

Storage Minimization

For security purposes, all the videos need to be stored in hard drive or cloud storage, which consumes a lot of space in the computer system. Text documents take much smaller space than video documents. So, converting any scenes or activities of the videos into a collection of some caption or text data will save a massive amount of storage space. Aligned with time sequence, generated captions can provide proper information about any activity.

Application of video captioning is quite dynamic and booming day by day. A large part of the technologies that help people with disabilities is occupied by video captioning techniques. Moreover, in the research field of one of the complicated and synthetic topics such as Human-Robot Interaction (HRI), video captioning plays a key role. Additionally, video captioning is continuously contributing immensely in those activities which are comparatively more related to day-to-day life such as video surveillance, storage minimization and much more.

Challenges on Video Captioning

The concept of video captioning has achieved its current state through a long journey, and yet it needs to overcome multiple challenges to get to an ultimate form. This part of the section mentions the challenges video captioning models faced and the step for resolving the problems.

Challenges with Machine Learning methods

- *Short-range of Activity Recognition:* The early works on video caption generation were only limited to text generating regarding human activities [54] based on ML techniques. The main challenges were the lack of robust models to handle the massive computation of frame by frame object / motion calculation and a language model to generate proper sentences describing the whole scene. The problem with objects and activities detection is solved with the combination of approaches such as bag-of-features, dense trajectory, and different ML algorithms like Support Vector Machine (SVM), Conditional Random Field (CRF), etc. [59, 125].
- *Insufficiency of Dataset:* Machine Learning methods need huge amounts of data to be trained with. The popular dataset were UCF101 [107], YouTube2Text, KTH [103] and some collection of 1976 small YouTube videos. These dataset did not have enough videos and suf-

ficient captions for each video. The datasets were also not activity-domain specific which affected the performance of the object and activity recognition modules. Also, the insufficiency of captions lead to less accurate captioning. Different large video dataset such as Microsoft Video Definition (MSVD), MPII-MD, Montreal Video Annotation Dataset (M-VAD), YouCook-II, etc. were introduced to address the dataset challenges. Among them, M-VAD has 48,986 video clips, and others also have a large number of videos. YouCook-II and many other datasets are specified to some specific activities. Some datasets have large caption sets for each video clip, like MSVD has an average of 35 captions per video clip.

- *Lack of Proper Language Generation Techniques:* In the early stages, the language models only generated captions as SVO triplets. Any information regarding place, or background were not included in the generated text data. It was very challenging to describe a whole scenario and complex activity with SVO triplets. Probabilistic methods for calculating the occurrence of verbs [73] and other words are introduced [23] to address the difficulties of language generation models.
- *Less Semantic Information:* Machine learning approaches lacked the efficiency of working with more visual and semantic information. Xu et al. [142] have used deep neural networks with CRF and SVM for object and/or action classification tasks.

Challenges with Deep Learning methods

Using deep learning models like CNNs, RNNs have resolved some of the challenges but can not ensure the optimal output.

- *Lack of Complex sentence Generation:* Different dataset has videos with multiple activities in a single video clip. It is very difficult to describe multiple scenes or actions of a video clip with a single sentence captioning. Generating multiple correlated sentences is a challenging task. Dense Video captioning [56] addressed the problem by generating multiple sentences related to the whole context of the video capturing long dependencies. Fig. 2 shows example of dense video captioning over single captioning.
- *Long Dependencies Incompatibility:* Deep learning models lacked the ability to carry the context of a video to longer steps. While processing a large number of video frames, it becomes very challenging for the models to stick with the main context of the video. Use of LSTM [85, 104, 135] dealt with the challenge allowing to generate better context-related captioning with multiple sentences.
- *Noise from Multiple Objects:* In the current scenario, multiple objects and actions in the video affect the caption generation model leading to a noisy caption. Among different objects, focusing the main object of the video frames has become a great challenge, which is resolved with attention mechanism [87, 111]. Different attention-based models, such as Transformer [118] focuses on the main object and actions associated with that object for better captioning.
- *Small Number of Classes in Dataset:* The performance of the neural networks increases with the number of data available in the dataset. Also the diversity of a model depends on the dataset it is being trained with. Dataset with a larger number of activity class will have a better impact on the performance of the captioning model. All the dataset needs to have a good number of captions against every video clip. A big challenge is that there is no large video dataset with a large number of activity classes containing vast amounts of videos and captions.
- *Hardware Constraints:* Amount of processing depends on the size of the video dataset. More extensive dataset requires more processing power dragging for powerful Graphics Processing Unit (GPU) to process the large amount of data. Number of real life activities and videos available in the dataset are infinite but available technology and processing power are limited. So, need for large dataset indicates the hardware constraint challenge.
- *Lack of Explicit Measuring Metric:* Matrices used for measuring the accuracy of the video captioning models were mostly developed for automatic MT and image captioning tasks. As there is no dedicated evaluation metric explicitly developed for video captioning tasks, evaluation part is somehow depends on human evaluation which does not have any standard metric. So, there remains a limitation of evaluating the models correctly.
- *Unavailability of Real-time Captioning Model:* Even after all these challenges are tackled, real-time video captioning still remains the most formidable challenge to deal with. Because, a video needs to go through few preprocessing steps followed by a visual extraction model and a language generation model. Also, there is no indication of the time span of a single event which increases noise in the caption. Which is why we still do not have such robust model that can process video frames in real-time and generate captions.

The concept of video captioning integrates two of the main branches of modern AI, such as CV and NLP. Recognizing available visual contents, for example, main objects, secondary objects, backgrounds, etc., and their interactions have always been very challenging for the CV domain and video captioning tasks. Because of the narrow field of actions, finding the exact activity based on the object's motion

also remains a significant challenge. Even after extracting all the objects and different semantic information, finding the importance factor stands as challenging for captioning models. Also, establishing the inter-connection of visual and language data is the key challenge of the whole captioning module. Lastly, the lack of a rich dataset takes place as one of the top challenges too.

Conclusion

In this survey, we have represented all the necessary details required in the field of video captioning. We have explained the basic structure of a video captioning system consisting of a visual feature extractor and a language processing portion. This survey has contained almost all the effective methods used for visual extraction, starting from the machine learning methods to the recent works of deep learning. The language processing portion has been handled in the same way as described in all the methods used in different papers. Here, we have accumulated and compared all the state-of-the-art methods and shown statistical analysis regarding the methods. Furthermore, we have gathered different deep learning-based video captioning approaches that are still evolving in recent times. We have tried to show some relative performance comparison among different methods from different periods. We have found out that the performances of the methods increased with the improvement of technologies and research. New approaches also showed unique aspects of captioning problems.

Next, we have reviewed all the evaluation metrics used for detecting the accuracy of the generated captions or descriptions. We have presented a summary of results obtained by current video description methods on the benchmark datasets using all metrics and also compared them. We have also tried to find any limitations to the evaluation metrics. We have found that all the metrics being used in an event in recent times are either designed for image captioning purposes or adopted from MT, and no metrics are made or designed dedicatedly for video description generation or captioning. This is why this portion of the system still lacks performance and efficiency. We believe that video captioning research can advance much further if evaluation metrics are improved. One way of improving the performance of these evaluation matrices can be achieved by increasing the number of reference sentences.

However, the most challenging part of video captioning is to generate captions from diversified videos successfully. The problem starts to grow when the videos contain multiple activities, and also, there are multiple objects. It leads to a poor video captioning performance of a model. Again, another problem arises with the long video clips with long activities, as models are designed to capture and encode only the short-term actions from the videos. However, the state of the art methods uses attention mechanism and attention-based models to focus on the spatially and also temporarily significant parts of the video.

Availability of data and material All the data and materials are properly extracted and cited from the existing works.

Compliance with Ethical Standards

Conflict of interest All of the Authors declare that he/she has no conflict of interest.

Code availability There are no execution of codes relating to this work as it is a survey paper.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Aafaq N, Akhtar N, Liu W, Gilani SZ, Mian A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019a;12487–12496.
2. Aafaq N, Akhtar N, Liu W, Mian A. Empirical autopsy of deep video captioning frameworks. 2019b. arXiv preprint arXiv:191109345
3. Aggarwal A, Chauhan A, Kumar D, Mittal M, Roy S, Kim Th. Video caption based searching using end-to-end dense captioning and sentence embeddings. Symmetry. 2020;12(6):992.
4. Agrawal P, Yadav R, Yadav V, De K, Roy PP. Caption-based region extraction in images. In: Proceedings of 3rd International Conference on Computer Vision and Image Processing, Springer, 2020;27–38.
5. Akbari H, Palangi H, Yang J, Rao S, Celikyilmaz A, Fernandez R, Smolensky P, Gao J, Chang SF. Neuro-symbolic representations for video captioning: A case for leveraging inductive biases for vision and language. 2020. arXiv preprint arXiv:201109530
6. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Lawrence Zitnick C, Parikh D. Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision, 2015;2425–2433
7. Ba JL, Kiros JR, Hinton GE. Layer normalization. 2016. arXiv preprint arXiv:160706450
8. Ballas N, Yao L, Pal C, Courville A. Delving deeper into convolutional networks for learning video representations. 2015. arXiv preprint arXiv:151106432
9. Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005;65–72.
10. Bantupalli K, Xie Y. American sign language recognition using deep learning and computer vision. In: 2018 IEEE International Conference on Big Data (Big Data), IEEE, 2018; 4896–4899
11. Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;1657–1666.

12. Bin Y, Yang Y, Shen F, Xu X, Shen HT. Bidirectional long-short term memory for video description. In: Proceedings of the 24th ACM international conference on Multimedia, 2016;436–440 .
13. Blei DM, Jordan MI. Modeling annotated data. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, 2003;127–134.
14. Chen DL, Dolan WB. Collecting highly parallel data for paraphrase evaluation. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, Association for Computational Linguistics, 2011;190–200.
15. Chen H, Lin K, Maye A, Li J, Hu X. A semantics-assisted video captioning model trained with scheduled sampling. 2019a. arXiv preprint arXiv:190900121
16. Chen H, Li J, Hu X. Delving deeper into the decoder for video captioning. 2020a. arXiv preprint arXiv:200105614
17. Chen J, Jin Q. Better captioning with sequence-level exploration. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020;10890–10899.
18. Chen J, Pan Y, Li Y, Yao T, Chao H, Mei T. Temporal deformable convolutional encoder-decoder networks for video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2019b;33:8167–74.
19. Chen S, Jiang YG. Motion guided spatial attention for video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2019;33:8191–8.
20. Chen X, Zhang M, Wang Z, Zuo L, Li B, Yang Y. Leveraging unpaired out-of-domain data for image captioning. Pattern Recognition Letters. 2020b;132:132–40.
21. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y. Learning phrase representations using rnn encoder-decoder for statistical machine translation. 2014. arXiv preprint arXiv:14061078
22. Das A, Kottur S, Gupta K, Singh A, Yadav D, Moura JM, Parikh D, Batra D. Visual dialog. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;326–335.
23. Das P, Xu C, Doell RF, Corso JJ. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2013;2634–2641.
24. Deng J, Krause J, Berg AC, Fei-Fei L. Hedging your bets: Optimizing accuracy-specificity trade-offs in large scale visual recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012;3450–3457.
25. Donahue J, Hendricks LA. Sergio guadar rama. In: Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell, "Long-term recurrent con volutional networks for visual recognition and descrip tion," in Proceedings of the IEEE conference on com puter vision and pattern recognition, 2015;2625–2634.
26. Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T. Long-term recurrent convolutional networks for visual recognition and description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015;2625–2634.
27. Everingham M, Van Gool L, Williams CK, Winn J, Zisserman A. The pascal visual object classes (voc) challenge. International journal of computer vision. 2010;88(2):303–38.
28. Fang Z, Gokhale T, Banerjee P, Baral C, Yang Y. Video2commonsense: Generating commonsense descriptions to enrich video captioning. 2020. arXiv preprint arXiv:200305162
29. Fawzy NK, Marey MA, Aref MM. Video captioning using attention based visual fusion with bi-temporal context and bimodal semantic feature learning. In: International Conference on Advanced Intelligent Systems and Informatics, Springer, 2020; 65–78.
30. Felzenszwalb P, McAllester D, Ramanan D. A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE conference on computer vision and pattern recognition, IEEE, 2008;1–8.
31. Gan Z, Gan C, He X, Pu Y, Tran K, Gao J, Carin L, Deng L. Semantic compositional networks for visual captioning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017;5630–5639.
32. Guadarrama S, Krishnamoorthy N, Malkarnenkar G, Venugopalan S, Mooney R, Darrell T, Saenko K. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In: Proceedings of the IEEE international conference on computer vision, 2013;2712–2719 .
33. Guo Y, Zhang J, Gao L. Exploiting long-term temporal dynamics for video captioning. World Wide Web. 2019;22(2):735–49.
34. Guo Z, Gao L, Song J, Xu X, Shao J, Shen HT. Attention-based lstm with semantic consistency for videos captioning. In: Proceedings of the 24th ACM international conference on Multimedia, 2016;357–361.
35. Hao X, Zhou F, Li X. Scene-edge gru for video caption. In: 2020 IEEE 4th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), IEEE, vol 1, 2020;1290–1295.
36. Hara K, Kataoka H, Satoh Y. Learning spatio-temporal features with 3d residual networks for action recognition. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017;3154–3160.
37. Hemalatha M, Sekhar CC. Domain-specific semantics guided approach to video captioning. In: 2020 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2020;1576–1585.
38. Hochreiter S, Schmidhuber J. Long short-term memory. Neural computation. 1997;9(8):1735–80.
39. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: Data, models and evaluation metrics. Journal of Artificial Intelligence Research. 2013;47:853–99.
40. Hori C, Hori T, Lee TY, Zhang Z, Harsham B, Hershey JR, Marks TK, Sumi K. Attention-based multimodal fusion for video description. In: Proceedings of the IEEE international conference on computer vision, 2017;4193–4202.
41. Hou J, Wu X, Zhao W, Luo J, Jia Y. Joint syntax representation learning and visual cue translation for video captioning. In: Proceedings of the IEEE International Conference on Computer Vision, 2019;8918–8927.
42. Hou J, Jia Y, Qi Y, et al. Video captioning using weak annotation. 2020 arXiv preprint arXiv:200901067
43. Huang G, Pang B, Zhu Z, Rivera C, Soricut R. Multimodal pretraining for dense video captioning. 2020a. arXiv preprint arXiv:201111760
44. Huang Y, Chen J, Ouyang W, Wan W, Xue Y. Image captioning with end-to-end attribute detection and subsequent attributes prediction. IEEE Transactions on Image Processing. 2020b;29:4013–26.
45. Iashin V, Rahtu E. A better use of audio-visual cues: Dense video captioning with bi-modal transformer. 2020. arXiv preprint arXiv:200508271
46. Jain BD, Thakur SM, Suresh K. Visual assistance for blind using image processing. In: 2018 International Conference on Communication and Signal Processing (ICCSP), IEEE, 2018;0499–0503.
47. Jeong D, Kim BG. Dong SY. Deep joint spatiotemporal network (djstn) for efficient facial expression recognition. Sensors. 2020;20(7) 1936.

48. Jia Z, Li X. icap: Interactive image captioning with predictive text. In: Proceedings of the 2020 International Conference on Multimedia Retrieval, 2020;428–435.
49. Jin T, Huang S, Chen M, Li Y, Zhang Z. Sbat: Video captioning with sparse boundary-aware transformer. 2020. arXiv preprint arXiv:200711888
50. Kim BG, Park DJ. Unsupervised video object segmentation and tracking based on new edge features. Pattern Recognition Letters. 2004;25(15):1731–42.
51. Kim J, Choi I, Lee M. Context aware video caption generation with consecutive differentiable neural computer. Electronics. 2020;9(7):1162.
52. Kim JH, Kim BG, Roy PP, Jeong DM. Efficient facial expression recognition algorithm based on hierarchical deep neural network structure. IEEE Access. 2019;7:41273–85.
53. Koehn P. Statistical machine translation. Cambridge University Press; 2009.
54. Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. International Journal of Computer Vision. 2002;50(2):171–84.
55. Korbar B, Petroni F, Girdhar R, Torresani L. Video understanding as machine translation. 2020. arXiv preprint arXiv:200607203
56. Krishna R, Hata K, Ren F, Fei-Fei L, Niebles JC. Dense-captioning events in videos. In: International Conference on Computer Vision (ICCV). 2017.
57. Krishnamoorthy N, Malkarnenkar G, Mooney R, Saenko K, Guadarrama S. Generating natural-language video descriptions using text-mined knowledge. In: Twenty-Seventh AAAI Conference on Artificial Intelligence 2013.
58. Lan W, Li X, Dong J. Fluency-guided cross-lingual image captioning. In: Proceedings of the 25th ACM international conference on Multimedia, 2017;1549–1557.
59. Laptev I, Marszalek M, Schmid C, Rozenfeld B. Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2008;1–8.
60. Lei J, Wang L, Shen Y, Yu D, Berg TL, Bansal M. Mart: Memory-augmented recurrent transformer for coherent video paragraph captioning. 2020. arXiv preprint arXiv:200505402
61. Li L, Gong B. End-to-end video captioning with multitask reinforcement learning. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019;339–348
62. Li LJ, Su H, Fei-Fei L, Xing EP. Object bank: A high-level image representation for scene classification & semantic feature sparsification. In: Advances in neural information processing systems, 2010;1378–1386
63. Li X, Zhou Z, Chen L, Gao L. Residual attention-based lstm for video captioning. World Wide Web. 2019;22(2):621–36.
64. Li Y, Yao T, Pan Y, Chao H, Mei T. Jointly localizing and describing events for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018;7492–7500.
65. Lin CY. Rouge: A packagefor automatic evaluation of summaries. In: ProceedingsofWorkshop on Text Summarization Branches Out, Post2Conference Workshop of ACL 2004.
66. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, Zitnick CL. Microsoft coco: Common objects in context. In: European conference on computer vision, Springer, 2014;740–755.
67. Liu F, Gao S, Gao X. Segmentation of mr image based on maximum a posterior. In: 2001 Conference Proceedings of the 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, vol 3, 2001;2681–2684.
68. Liu S, Ren Z, Yuan J. Sibnet: Sibling convolutional encoder for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2020.
69. Long X, Gan C, de Melo G. Video captioning with multi-faceted attention. Transactions of the Association for Computational Linguistics. 2018;6:173–84.
70. Luo H, Ji L, Shi B, Huang H, Duan N, Li T, Chen X, Zhou M. Univilm: A unified video and language pre-training model for multimodal understanding and generation. 2020. arXiv preprint arXiv:200206353
71. Malinowski M, Fritz M. A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in neural information processing systems, 2014;1682–1690.
72. Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: Proceedings of the IEEE International Conference on Computer Vision, 2019;2630–2640.
73. Motwani TS, Mooney RJ. Improving video activity recognition using object recognition and text mining. In: ECAI, 2012;vol 1, p 2 .
74. Mukherjee S, Saini R, Kumar P, Roy PP, Dogra DP, Kim BG, et al. Fight detection in hockey videos using deep network. Journal of Multimedia Information System. 2017;4(4):225–32.
75. Mukherjee S, Ghosh S, Ghosh S, Kumar P, Roy PP. Predicting video-frames using encoder-convlstm combination. In: ICASSP 2019–2019 IEEE International Conference on Acoustics. IEEE: Speech and Signal Processing (ICASSP); 2019. p. 2027–31.
76. Mun J, Yang L, Ren Z, Xu N, Han B. Streamlined dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019;6588–6597
77. Olivastri S, Singh G, Cuzzolin F. End-to-end video captioning. In: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019;0–0.
78. Pala P, Santini S. Image retrieval by shape and texture. Pattern Recognition. 1999;32(3):517–27.
79. Pan B, Cai H, Huang DA, Lee KH, Gaidon A, Adeli E, Niebles JC. Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020;10870–10879.
80. Pan P, Xu Z, Yang Y, Wu F, Zhuang Y. Hierarchical recurrent neural encoder for video representation with application to captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016a;1029–1038.
81. Pan Y, Mei T, Yao T, Li H, Rui Y. Jointly modeling embedding and translation to bridge video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016b;4594–4602.
82. Pan Y, Yao T, Li H, Mei T. Video captioning with transferred semantic attributes. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017;6504–6512.
83. Papineni K, Roukos S, Ward T, Zhu WJ. Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics, Association for Computational Linguistics, 2002;311–318.
84. Park JS, Rohrbach M, Darrell T, Rohrbach A. Adversarial inference for multi-sentence video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019;6598–6608.
85. Pasunuru R, Bansal M. Multi-task video captioning with video and entailment generation. 2017. arXiv preprint arXiv:170407489
86. Pawar K, Attar V. Deep learning approaches for video-based anomalous activity detection. World Wide Web. 2019;22(2):571–601.
87. Pei W, Zhang J, Wang X, Ke L, Shen X, Tai YW. Memory-attended recurrent network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019;8347–8356.

88. Pendurkar S, Kolpekwar S, Dhoot S, Haribhakta YV, Banerjee B. Attention based multi-modal fusion architecture for open-ended video question answering systems. Procedia Computer Science. 2020;171:446–55.
89. Perez-Martin J, Bustos B, Pérez J. Attentive visual semantic specialized network for video captioning. In: International Conference on Computer Vision 2020.
90. Perez-Martin J, Bustos B, Perez J. Improving video captioning with temporal composition of a visual-syntactic embedding. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021;3039–3049.
91. Peris Á, Bolaños M, Radeva P, Casacuberta F. Video description using bidirectional recurrent neural networks. In: International Conference on Artificial Neural Networks, Springer, 2016;3–11.
92. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision, 2015;2641–2649.
93. Rahman M, Abedin T, Prottoy KS, Moshruba A, Siddiqui FH, et al. Semantically sensible video captioning (ssvc). 2020. arXiv preprint arXiv:200907335
94. Rashtchian C, Young P, Hodosh M, Hockenmaier J. Collecting image annotations using amazon's mechanical turk. In: Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, 2010;139–147
95. Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M. Grounding action descriptions in videos. Transactions of the Association for Computational Linguistics. 2013;1:25–36.
96. Rimle P, Dogan P, Gross M. Enriching video captions with contextual text. 2020. arXiv preprint arXiv:200714682
97. Rohrbach A, Rohrbach M, Qiu W, Friedrich A, Pinkal M, Schiele B. Coherent multi-sentence video description with variable level of detail. In: German conference on pattern recognition, Springer, 2014;184–195.
98. Rohrbach A, Rohrbach M, Schiele B. The long-short story of movie description. In: German conference on pattern recognition, Springer, 2015a;209–221.
99. Rohrbach A, Rohrbach M, Tandon N, Schiele B. A dataset for movie description. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015b;3202–3212.
100. Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B. Translating video content to natural language descriptions. In: Proceedings of the IEEE International Conference on Computer Vision, 2013;433–440.
101. Roy PP, Bhunia AK, Bhattacharyya A, Pal U. Word searching in scene image and video frame in multi-script scenario using dynamic shape coding. Multimedia Tools and Applications. 2019;78(6):7767–801.
102. Sah S, Nguyen T, Ptucha R. Understanding temporal structure for video captioning. Pattern Analysis and Applications. 2020;23(1):147–59.
103. Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local svm approach. In: Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004., IEEE, 2004;3,32–36
104. Shen Z, Li J, Su Z, Li M, Chen Y, Jiang YG, Xue X. Weakly supervised dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;1916–1924.
105. Song J, Guo Y, Gao L, Li X, Hanjalic A, Shen HT. From deterministic to generative: Multimodal stochastic rnns for video captioning. IEEE transactions on neural networks and learning systems. 2018;30(10):3047–58.
106. Song K, Tan X, Qin T, Lu J, Liu TY. Mass: Masked sequence to sequence pre-training for language generation. 2019. arXiv preprint arXiv:190502450
107. Soomro K, Zamir AR, Shah M. Ucf101: A dataset of 101 human actions classes from videos in the wild. 2012. arXiv preprint arXiv:12120402
108. Srivastava N, Mansimov E, Salakhudinov R. Unsupervised learning of video representations using lstms. In: International conference on machine learning, 2015;843–852.
109. Suin M, Rajagopalan A. An efficient framework for dense video captioning. In: AAAI, 2020;12039–12046.
110. Sun C, Nevatia R. Semantic aware video transcription using random forest classifiers. In: European Conference on Computer Vision, Springer, 2014;772–786.
111. Sur C. Sact: Self-aware multi-space feature composition transformer for multinomial attention for video captioning. 2020. arXiv preprint arXiv:200614262
112. Szegedy C, Ioffe S, Vanhoucke V, Alemi A. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Proceedings of the AAAI Conference on Artificial Intelligence 2017.
113. Tan G, Liu D, Wang M, Zha ZJ. Learning to discretely compose reasoning module networks for video captioning. 2020. arXiv preprint arXiv:200709049
114. Tang J. Intelligent Mobile Projects with TensorFlow: Build 10+ Artificial Intelligence Apps Using TensorFlow Mobile and Lite for IOS, Android, and Raspberry Pi. Packt Publishing Ltd 2018.
115. Thomason J, Venugopalan S, Guadarrama S, Saenko K, Mooney R. Integrating language and vision to generate natural language descriptions of videos in the wild. In: Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014;1218–1227.
116. Torabi A, Pal C, Larochelle H, Courville A. Using descriptive video services to create a large data source for video annotation research. 2015. arXiv preprint arXiv:150301070
117. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. In: Proceedings of the IEEE international conference on computer vision, 2015;4489–4497.
118. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Advances in neural information processing systems, 2017;5998–6008.
119. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015;4566–4575.
120. Venugopalan S, Xu H, Donahue J, Rohrbach M, Mooney R, Saenko K. Translating videos to natural language using deep recurrent neural networks. 2014. arXiv preprint arXiv:14124729
121. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K. Sequence to sequence-video to text. In: Proceedings of the IEEE international conference on computer vision, 2015;4534–4542.
122. Venugopalan S, Hendricks LA, Mooney R, Saenko K. Improving lstm-based video description with linguistic knowledge mined from text. 2016. arXiv preprint arXiv:160401729
123. Wang B, Ma L, Zhang W, Liu W. Reconstruction network for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018a;7622–7631.
124. Wang B, Ma L, Zhang W, Jiang W, Wang J, Liu W. Controllable video captioning with pos sequence guidance based on gated fusion network. In: Proceedings of the IEEE International Conference on Computer Vision, 2019a;2641–2650.
125. Wang H, Kläser A, Schmid C, Liu CL. Action recognition by dense trajectories. In: CVPR 2011, IEEE, 2011;3169–3176.

126. Wang H, Gao C, Han Y. Sequence in sequence for video captioning. Pattern Recognition Letters. 2020a;130:327–34.
127. Wang J, Jiang W, Ma L, Liu W, Xu Y. Bidirectional attentive fusion with context gating for dense video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018b;7190–7198.
128. Wang J, Wang W, Huang Y, Wang L, Tan T. M3: Multimodal memory modelling for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018c;7512–7520.
129. Wang J, Wang W, Wang L, Wang Z, Feng DD, Tan T. Learning visual relationship and context-aware attention for image captioning. Pattern Recognition. 2020b;98:107075.
130. Wang S, Lan L, Zhang X, Dong G, Luo Z. Object-aware semantics of attention for image captioning. Multimedia Tools and Applications. 2020c;79(3):2013–30.
131. Wang T, Zheng H, Yu M, Tian Q, Hu H. Event-centric hierarchical representation for dense video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2020d.
132. Wang X, Chen W, Wu J, Wang YF, Yang Wang W. Video captioning via hierarchical reinforcement learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018d;4213–4222.
133. Wang X, Wu J, Zhang D, Su Y, Wang WY. Learning to compose topic-aware mixture of experts for zero-shot video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2019b;33:8965–72.
134. Wu A, Han Y. Hierarchical memory decoding for video captioning. 2020. arXiv preprint arXiv:200211886
135. Wu X, Li G, Cao Q, Ji Q, Lin L. Interpretable video captioning via trajectory structured localization. In: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2018;6829–6837.
136. Xia Q, Huang H, Duan N, Zhang D, Ji L, Sui Z, Cui E, Bharti T, Zhou M. Xgpt: Cross-modal generative pre-training for image captioning. 2020. arXiv preprint arXiv:200301473
137. Xiao H, Shi J. Video captioning with text-based dynamic attention and step-by-step learning. Pattern Recognition Letters. 2020.
138. Xu H, Li B, Ramanishka V, Sigal L, Saenko K. Joint event detection and description in continuous video streams. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE, 2019;396–405.
139. Xu J, Mei T, Yao T, Rui Y. Msr-vtt: A large video description dataset for bridging video and language. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;5288–5296.
140. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y. Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, 2015a;2048–2057.
141. Xu N, Liu AA, Wong Y, Zhang Y, Nie W, Su Y, Kankanhalli M. Dual-stream recurrent neural network for video captioning. IEEE Transactions on Circuits and Systems for Video Technology. 2018;29(8):2482–93.
142. Xu R, Xiong C, Chen W, Corso JJ. Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: Twenty-Ninth AAAI Conference on Artificial Intelligence 2015b.
143. Yan C, Tu Y, Wang X, Zhang Y, Hao X, Zhang Y, Dai Q. Stat: spatial-temporal attention mechanism for video captioning. IEEE transactions on multimedia 2019.
144. Yang Y, Zhou J, Ai J, Bin Y, Hanjalic A, Shen HT, Ji Y. Video captioning by adversarial lstm. IEEE Transactions on Image Processing. 2018;27(11):5600–11.
145. Yang Z, Han Y, Wang Z. Catching the temporal regions-of-interest for video captioning. In: Proceedings of the 25th ACM international conference on Multimedia, 2017;146–153.
146. Yao L, Torabi A, Cho K, Ballas N, Pal C, Larochelle H, Courville A. Describing videos by exploiting temporal structure. In: Proceedings of the IEEE international conference on computer vision, 2015;4507–4515.
147. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics. 2014;2:67–78.
148. Yu H, Wang J, Huang Z, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;4584–4593
149. Zhang C, Tian Y. Automatic video description generation via lstm with joint two-stream encoding. In: 2016 23rd International Conference on Pattern Recognition (ICPR), IEEE, 2016;2924–2929.
150. Zhang J, Peng Y. Hierarchical vision-language alignment for video captioning. In: International Conference on Multimedia Modeling, Springer, 2019a;42–54.
151. Zhang J, Peng Y. Object-aware aggregation with bidirectional temporal graph for video captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019b;8327–8336.
152. Zhang X, Gao K, Zhang Y, Zhang D, Li J, Tian Q. Task-driven dynamic fusion: Reducing ambiguity in video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017;3713–3721.
153. Zhang X, Liu C, Chang F. Guidance module network for video captioning. 2020a. arXiv preprint arXiv:201210930
154. Zhang Z, Xu D, Ouyang W, Tan C (2019) Show, tell and summarize: Dense video captioning using visual cue aided sentence summarization. IEEE Transactions on Circuits and Systems for Video Technology
155. Zhang Z, Shi Y, Yuan C, Li B, Wang P, Hu W, Zha ZJ. Object relational graph with teacher-recommended learning for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020b;13278–13288.
156. Zhao B, Li X, Lu X. Cam-rnn: Co-attention model based rnn for video captioning. IEEE Transactions on Image Processing. 2019;28(11):5552–65.
157. Zhao W, Wu X, Zhang X. Memcap: Memorizing style knowledge for image captioning. In: AAAI, 2020;12984–12992.
158. Zheng Q, Wang C, Tao D. Syntax-aware action targeting for video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020;13096–13105.
159. Zhu F, Hwang JN, Ma Z, Chen G, Guo J. Understanding objects in video: Object-oriented video captioning via structured trajectory and adversarial learning. IEEE Access 2020a.
160. Zhu F, Hwang JN, Ma Z, Jun G. Object-oriented video captioning with temporal graph and prior knowledge building. 2020b. arXiv preprint arXiv:200303715
161. Zhu L, Yang Y. Actbert: Learning global-local video-text representations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020;8746–8755.