Received: 5 July 2023 | Revised: 4 September 2023 | Accepted: 13 September 2023

DOI: 10.1002/eng2.12785

REVIEW

Engineering Reports Open Access WILEY

Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods

Mohammad Saif Wajid[1] | **Hugo Terashima-Marin**[1] | **Peyman Najafirad**[2] | **Mohd Anas Wajid**[3]

[1]Department of Computer Sciences, School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Mexico

[2]Department of Information Systems and Cyber Security, University of Texas, San Antonio, Texas, USA

[3]Department of Computer Application in the School of Computing Science & Engineering, Galgotias University, Noida, India

Correspondence
Mohammad Saif Wajid, Department of Computer Sciences, School of Engineering and Sciences, Tecnologico de Monterrey, Monterrey, Mexico.
Emails: mohdsaif06@gmail.com and mohammadsaif.wajid@utsa.edu

Funding information
CONAHCYT and Tecnologico De Monterrey, Mexico

Abstract

Generating an image/video caption has always been a fundamental problem of Artificial Intelligence, which is usually performed using the potential of Deep Learning Methods, Computer Vision, Knowledge Graphs, and Natural Language Processing (NLP). The significant task of image/video captioning is to describe visual content in terms of natural language. Due to a semantic gap, this presents a massive problem in understanding and explaining images or videos syntactically and semantically. The current systems need somewhere to fill the gap between low-level and high-level features while mapping. Therefore, to tackle this problem, there is a need to describe the latest research and methods to overcome difficulties and to propose effective solutions. This work thoroughly analyses and investigates the most related methods (deep learning and knowledge graph-based approaches), benchmark datasets, and evaluation metrics with their benefits and limitations. Here we have also reviewed the state-of-the-art methods related to image/video captioning and their applications in the current scenario. Finally, we provide thorough information on existing research with comparisons of results on benchmark datasets. We have also mentioned the existing challenges and future direction of research.

KEYWORDS

computer vision, dense video captioning, image/video captioning, knowledge graph, natural language processing

1 | INTRODUCTION

Describing visual data automatically of an image/video with artificial intelligence and knowledge graphs catches an eye. It is best served in surveillance, security, violence explanations, and health care systems. Image and video captioning are related but distinct tasks in computer vision and natural language processing.

Image captioning refers to automatically generating a natural language description of an image.[1] It is a challenging task that involves both image understanding and language generation. The goal of image captioning is to generate a sentence that accurately describes the content of an image and is grammatically correct.[2]

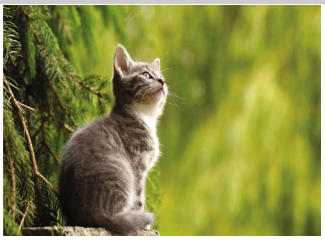
Video captioning also known as video description/video storytelling generates natural language descriptions of the events and actions in a video. This task is like image captioning but more challenging, as it requires understanding visual content and events' temporal dynamics. Recently, deep learning methods such as Convolutional Neural Networks (CNNs),[3] Recurrent Neural Networks (RNNs),[4] and Knowledge Graphs (KGs)[5] have been used to achieve significant improvements in image/video captioning. These methods use neural networks to learn features from images and videos, and then these features are used to generate captions. It involves getting an object, attribute detection, their interaction with other objects, the relationship among things, and finally explaining the same so that anyone can quickly get what is happening in the image/video.

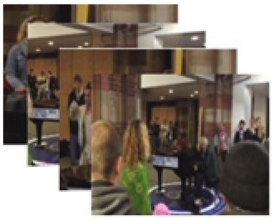
1.1 | Image captioning

Explaining or generating the description of an image is called image captioning or frame captioning.[6] Image captioning refers to finding descriptions of each frame, but the important thing is to find the definition in terms of a sentence that should be correct with proper meaning.[7,8] Mainly image captioning has been tackled via RNNs models,[9-13] and Long Short-Term Memory (LSTMs).[14-18] We have seen that there has been massive progress in vision tasks, whether related to classification[19] work (action classification, image classification, and attribute classification),[20-24] or related to recognition work (object and scene recognition).[25-29]

Generating an automatic description of the image is a new work. As we know, most communication between machines and humans depends on natural language understanding and explaining.[30] That is why various applications of image description take place in a real-world scenario, such as information access and retrieval,[31-33] providing assistance to visually impaired people,[34,35] education, natural language processing, and social media factors,[36] and so forth. Image captioning is gaining popularity and becoming a significant field of artificial intelligence and computer vision study. Some examples of image captioning are given in Table 1.

TABLE 1 Image caption.

Image	Image caption
	"A cat is perched close to a pine tree and is gazing upward".[37]
	"An old man with a black suit is playing Piano".[38]
	"A baby girl wearing a Hijab in preparation to offer Namaz (Salah)".



An old man playing piano in a hall room in front of many people.

FIGURE 1 Video caption.

Somehow, it's a new work to generate the results from the computer in terms of getting an automatic description of the image. As we know most of the communication between machines and humans depends on natural language understanding and explaining.[30] That is why there are various applications of image description taking place in a real-world scenario such as information access and retrieval,[31-33] providing assistance to visually impaired peoples,[34,35] education, social media, natural language processing, and so forth. Image captioning is gaining popularity as a difficult and significant area of artificial intelligence and computer vision study and is becoming more and more crucial.

1.2 | Video captioning

Providing an automatic video content description with human-understandable language is widespread and is termed video captioning.[39] The video captioning task is extremely attractive in artificial intelligence, computer vision, and knowledge graphs. In the past, video captioning was the task of detecting visual content with manually designed characteristics and generating a caption in terms of the sentence.[40,41]

The purpose of creating video captions is to provide a sequence of words to explain the visual content of that video. It is necessary to capture the temporal dynamics to comprehend the video material, in addition to the fact that the video includes significantly more information than a still image.[42] There are many applications where video captioning takes place in a significant way, such as Video Retrieval Systems (VRS),[43] Visual Questioning Answering (VQA),[44,45] assist visually impaired people,[46] Text-to-speech technology,[39] and so forth. An example of video captioning is given in Figure 1.

1.3 | Dense video captioning

Dense captioning of videos is composed of different steps; the first is to identify all events in the video, the second is action recognition, and the third is to do video captioning for all possibilities in a particular video.[38] Localizing noteworthy events from an uncut video and creating textual descriptions (captions) for each identified event is known as dense video captioning. The majority of earlier works on dense video captioning used only visual cues.[47] Dense captioning is popular nowadays, and much research has been done on the job so far.[48-54] An example of a dense video caption[38] is given by Figure 2.

2 | ORGANIZATION

Further the work is organized as follows. Section 3 reviews related work on image/video captioning and dense video captioning. Section 4 describes deep learning methods for image captioning with applications, evaluation metrics, and datasets. Section 5 describes methods of deep learning towards video captioning with applications, datasets, and evaluation metrics. Section 6 describes methods of deep learning for dense video captioning with applications, datasets, and evaluation metrics. Section 7 describes knowledge graph-based methods for image captioning and dense video captioning. In Section 8, we presented an evaluation of existing work with benchmark datasets. We provided some existing challenges in Section 9. In Section 10, we concluded this work with specific points with future directions in image and video captioning.

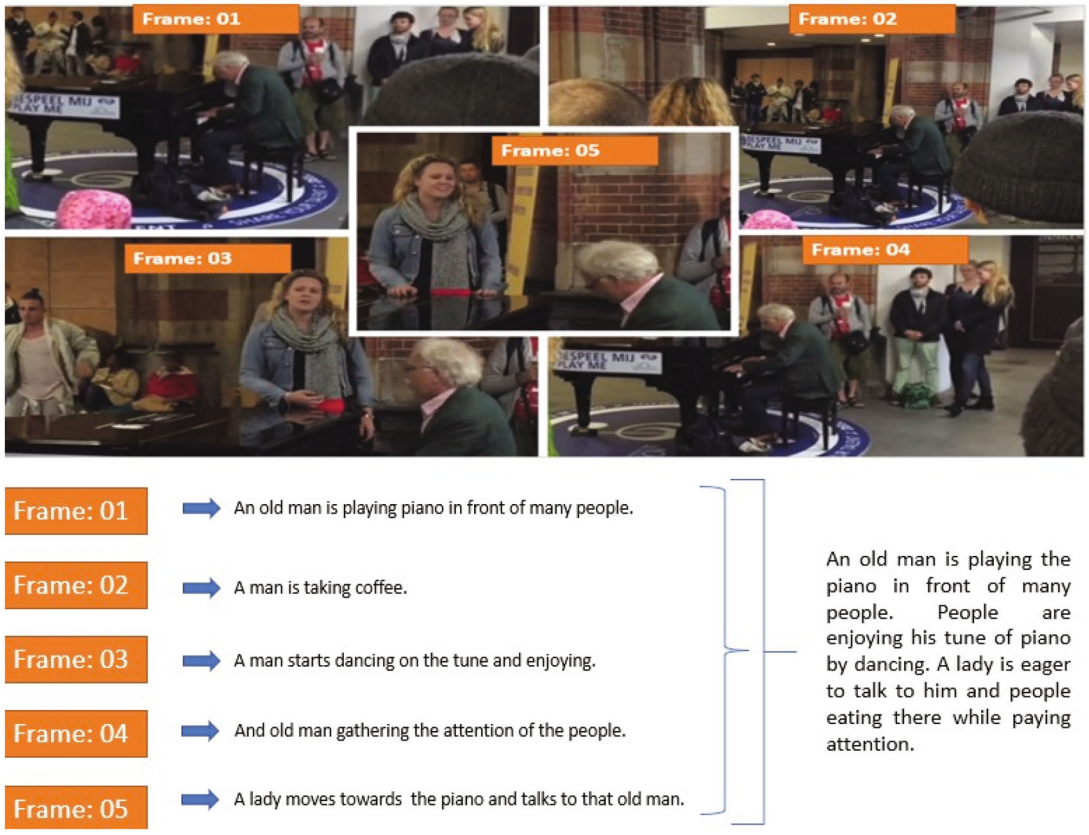


FIGURE 2 Dense video caption.

3 | RELATED WORK

Captioning describes an image/video as a natural language sentence/paragraph. In the field of deep learning and knowledge graph, it has been a vital task. It occurs through recognizing each action and video captioning. We have thoroughly studied existing work regarding image captioning, video captioning, and dense video captioning.

Liunian Li et al.[55] developed a model Grounded Language-Image Pre-training (GLIP) for mastering object-level, linguistically conscious, and semantically rich visual representations. Work suggests integrating text and image deeply, making the detection model linguistically aware and a solid foundational model. Then able to pre-train GLIP on scalable and semantically rich grounding data via reformulation and deep fusion.

Jia et al.[56] developed a straightforward technique "A Large-scale Image and Noisy-text embedding (ALIGN)" for scaling up the learning of visual and vision-language representations from vast amounts of noisy image-text data. They trained dual encoder model using a contrastive loss on benchmark datasets Flickr30K and MSCOCO and got good results.

Maofu Liu et al.[57] tackled the problem of image information and explored visual attention to understand an image using a Fully Convolutional Network (FCN).[58] FCN is used to predict the labels. Work has been carried out on the Chinese Caption Dataset (CCD), and the result is compared with others effectively and feasibly.

Xuelong Li[59] carried out text guided attention and semantic attention to get the most related spatial information and the reduced semantic gap between visual and natural language. At last, the authors gathered all data to produce the required answers in caption form for an optical question answer system.

Songtao Ding[60] proposed attention theory in psychology for image caption and combined low-level features (quality of an image) with high-level features (regions of an image) to focus particular areas of an image. The authors introduced the theory of attention in psychology for appearance captioning and used filter image features.[61] They combined bottom-up attention mechanism with faster R-CNN to get the results with benchmark datasets such as MSCOCO (Microsoft Common Objects in COntext), Flickr30K, PASCAL,[62] and SBU[63] datasets.

Xinlei Chen et al.[64] employed sentence-based explanations and bidirectional mapping between the visuals. Their work led to generating novel captions of an image and was able to reconstruct visual features given in an image description. Authors performed testing of their work in sentence generation, sentence retrieval, and image retrieval. They used different datasets to evaluate the performance of their model, such as the PASCAL sentence,[65] Flickr8K, Flickr30K, and the MSCOCO.[66]

Junhua et al.[9] created a system to produce new captions using unique pairings of elements. The authors proposed a Multimodal Recurrent Neural Network (m-RNN) framework that is tailored for the retrieval and sentence creation tasks. The model was comprised of a CNN and an RNN that interact with one another in a multimodal layer that received three inputs: an image representation, an embedding word layer, and a recurrent layer. A final softmax layer was used to build the probability distribution for the next word.

Mathew et al.[67] proposed a SentiCap system that combined positive and negative attitudes into captions. It was a switching RNN model with word-level regularization that emphasized sentiment.[68] To create styled subtitles, two networks CNN and an RNN, were used. A large dataset of image captions was used to train one network to provide typical factual descriptions, and a smaller dataset containing sentiment polarity was used to prepare the other network. Studies revealed that 74% of the phrases created by SentiCap had the correct sentiment.

Anderson et al.[69] proposed a novel algorithm for training sequence models, such as RNN, on partially specified sequences, which are represented using finite state automata. This method lifted the restriction that previously required captioning image models to be trained on paired image-sentence corpora only. The authors applied their approach to an existing neural captioning model and achieved SOTA (State-of-the-art) outcomes to novel object caption tasks using an MSCOCO dataset. Further, they trained their model to describe new visual concepts from the Open Images dataset while maintaining competitive COCO evaluation scores.

Now moving towards the research related to Video Captioning, several prominent researchers gained good results in this field, such as Zhang et al.[70] presented a comprehensive video caption system that included a new structure and effective training strategy to tackle the current problem with video captioning that occurs because current models lack visual presentation due to negligence of interaction among objects. They presented an encoder using an Object Relational Graph (ORG) to capture more specific interaction features and enhance visual representation. Also, they designed a system called Teacher Recommended Learning (TRL) to utilize the successful External Language Model (ELM) fully and incorporated the wealth of linguistic knowledge into the caption model. The ELM generated additional semantically comparable word proposed to address the long-tailed problem, extending the ground truth words utilized in training. Three benchmark datasets, MSVD,[71] MSR-VTT,[72] and VATEX,[73] were used to evaluate the system's performance, which revealed that the proposed ORG-TRL system achieved state-of-the-art performance. Visualizations and extensive ablation studies demonstrated the efficiency of the approach.

Chenggang et al.[74] proposed an encoder-decoder neural network with a brand-new Spatial-Temporal Attention Mechanism (STAT) for video captioning. They suggested that STAT effectively accounted for both the spatial and temporal patterns inside a video clip, causing the word prediction decoder to automatically select the significant portions in the most appropriate temporal segments. Two well-known benchmark datasets, MSVD, and MSR- VTT-10K, were used to assess the spatial-temporal attention mechanism. According to experimental findings, the spatial-temporal attention mechanism performed at the cutting edge on three widely used evaluation metrics BLEU-4, METEOR, and CIDEr.

Yang et al.[75] proposed a novel method for video captioning using adversarial learning and LSTM. The authors worked on the Generative Adversarial Network (GAN)[76] model, which incorporated two different things; a generator (used to generate natural language sentences available in visual content) and another discriminator which controls the accuracy of that particular sentence. They used the LSTM network to implement an already-existing video captioning concept. They suggested a novel realization for the discriminator that used both the sentences and the video features as input and was tailored specifically for the video captioning challenge.

Xiaojuan et al.[77] proposed a semantic descriptor with an objective for scene recognition. The authors presented the statistical information of objects appearing in each scene to compute the distribution of each object across stages, which obtains the co-occurrence pattern of things. To make image descriptors more discriminative, they discarded the patches with non-discriminative objects to enhance the intra class generalized characteristics. They performed their experiment on benchmark scene datasets such as Scene,[78] MIT Indoor,[79] and SUN[80] and achieved good results.

Lianli et al.[81] introduced a unique framework for learning multi-level representations and generating syntax-aware video captions called Hierarchical Representation Network with Auxiliary Tasks (HRNAT). In order to learn how to represent movies hierarchically using the three-level representation of languages as a reference, they used the

cross-modality matching task. In addition to being globally identical to the video material, descriptions created with the help of the syntax-guiding task and the vision-assist task also adhered to the syntax of the ground truth description. Their model's essential elements were universal and easily adaptable to applications requiring both captioning of videos and Video Question Answering (VQA). The effectiveness and superiority of the suggested method over state-of-the-art methods were validated by the authors' evaluation of the framework's performance on several benchmark datasets.

Huaishao et al.[82] proposed a model known as CLIP4Clip. It was utilized to seamlessly convert the knowledge of the CLIP model to video- language retrieval. They used a video encoder and text decoder on datasets like MSR-VTT, MSVC, LSMDC, ActivityNet, and DiDeMo. They adopted the ViT-B/32[83] video encoder, which has 12 layers and a 32-bit patch size. Their work is based explicitly on the pre-trained CLIP (ViT-B/32)[84] and focused chiefly on converting image representation to video representation. The video-text retrieval challenge in their work is successfully completed using the pre-trained CLIP (ViT-B/32).

Tang et al.[85] proposed a CLIP-enhanced video-text matching network as the foundation of the CLIP4Caption system, which enhances video captioning. This approach compels the model to strongly learn text-correlated video features for text creation, making the most of the knowledge from vision and language. Additionally, they employed a Transformer structured decoder network to efficiently learn the long-range visual and language relationship, unlike most previous models that used LSTM or GRU as the sentence decoder and provided a new ensemble method for captioning assignments as well. Experiment results showed how practical their approach is on benchmark dataset like MSR-VTT.

Zhou et al.[86] discovered facial units in video sequences of one or more persons in an unsupervised manner. The methodologies on temporal segmentation,[87] and clustering[88] of sequences containing facial features tried to fill the semantic gap between low-level and high-level features. Portillo et al.[89] explored using CLIP, a language-image model, to produce video representations without requiring the annotations mentioned earlier. This model was specifically designed to discover an area where text and photos could be compared. The approach solely considered visual and text modalities and employed an aggregation function to frame-level characteristics, which is prevalent in other video retrieval works. The authors performed their experiments with two benchmark datasets, MSR-VTT and MSVD, and got good results.

Bang et al.[90] introduced Dual Attribute Prediction, an auxiliary task requiring a video caption model to learn the correspondence between video content and attributes and the co-occurrence relations between attributes. For video captioning, "pre-training and fine-tuning" has become a de facto paradigm, where ImageNet Pre-training (INP) is usually used to help encode the video content, and a task-oriented the network is fine-tuned from scratch to cope with caption generation. They compared INP with CLIP (Contrastive Language-Image Pre-training), their work investigated the potential deficiencies of INP for video captioning and explored the key to generating accurate descriptions. They studied INP versus CLIP and found INP made video caption models tricky to capture attributes' semantics and sensitive to irrelevant background information. By contrast, CLIP's significant boost in caption quality highlighted the importance of attribute-aware representation learning. The authors achieved good results on MSR-VTT datasets. There are many existing works[91-100] regarding video captioning.

Moving on to work on dense video captioning, it generates and describes every event happening in the video in terms of natural language. The extensive image captioning challenge served as a source of inspiration for this.[49] Vladimir Iashin et al.[47] introduced a novel method for dense video captioning that could use a variety of modalities to describe events and demonstrated how audio and speech modalities could enhance a dense video captioning model in particular. An automatic speech recognition system was used to produce a textual description of the address that was temporally synchronized. After that, they treated this description as a distinct input in addition to the audio track and accompanying video frames. To translate multimodal input data into textual descriptions, they structured the captioning task such as an automated translation issue and used the recently described transformer architecture. The authors used the ActivityNet[38] dataset to show off the effectiveness of their approach.

In particular, authors[51] used the concept of the context-aware,[38] generalizing the temporal event proposal module to use both past and future contexts and an attentive fusion to distinguish captions from extremely over lapped events. Meanwhile, Single Shot Detector (SSD)[101] was also used to generate event proposals and reward maximization for better captioning.[52]

To mitigate the intrinsic difficulties of RNNs to model long-term dependencies in a sequence, Zhou et al.[48] tailored the recent idea of Transformer[102] for dense video captioning. In Reference 103, the authors noticed that the

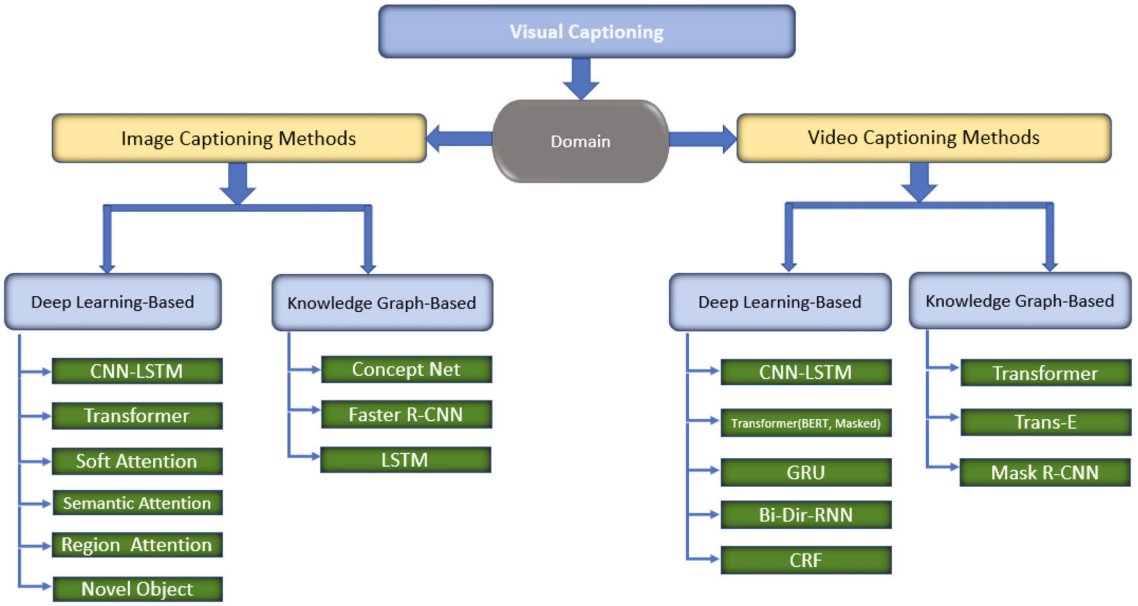


FIGURE 3 Deep learning and knowledge graph based methods for image/video captioning.

captioning may be benefited from interactions between objects in a video and developed recurrent higher order interaction modules to model these interactions. Xiong et al.[104] noticed that many previous models produced redundant captions and proposed to produce captions conditionally on the previous caption in a progressive manner while applying paragraph and sentence-level rewards. Similarly, a "bird view" correction and the reward maximization on two levels for a vast logical narrative telling were also employed.[53] Yu et al.[105] presented Accelerated Masked Transformer (AMT) model[48] for dense video captioning and compared it with the counterpart. AMT is significantly faster while maintaining its performance. The authors worked on two parts. First, they brought a compact, anchor free suggestion and a basic attention strategy to the organization. They used a single shot feature masking technique and a standard attention mechanism. Authors got better results by experimenting with datasets such as ActivityNet Caption and YouCookII.[106]

Though existing captioning techniques made encouraging strides, they could not represent implicit components of the image since the information gained from ground truth captions needed to be more extensive also could not characterize new traits or qualities beyond exercising the collection of knowledge. By adding data from outside sources into the caption generating process, this problem can be resolved. In a previous study, Anne Hendricks et al.[107] used object knowledge from external object recognition datasets or text corpora to enable innovative object captioning.

Li et al.[108] and Gu et al.[109] used Knowledge Graphs to generate scene graphs and answer visual questions, respectively. These two works made their models adaptable to external test situations by embedding the knowledge obtained from external knowledge graphs into a shared area with other data. Zhou et al.[110] suggested about knowledge graphs to improve image captioning, which pre-trains an RNN by extracting phrases that are directly and indirectly connected to entities identified by an object detector using a knowledge graph.

Based on related work, we have categorized visual captioning based on deep learning and knowledge graph based methods for image/video captioning and dense video captioning in Figure 3.

4 | DEEP LEARNING METHODS FOR IMAGE CAPTIONING

Recently lots of work has been done on image captioning using the methods of deep learning. We explored some of the great work by prominent researchers in artificial intelligence and computer vision. Figures 4A and 4B help us grasp the fundamental idea of creating image captions. We used deep learning techniques to feature extraction and describe the working process, respectively. We extracted 2048 feature vectors from an image with an input size of $224 \times 224 \times 3$, which denotes an image with a resolution of 224 pixels in both height and width and three color channels (R, G, and B).

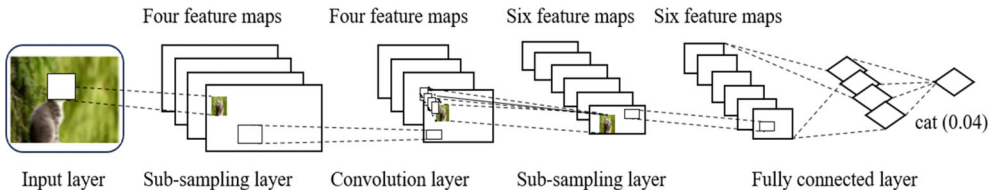


FIGURE 4A Feature processing (CNN layers).

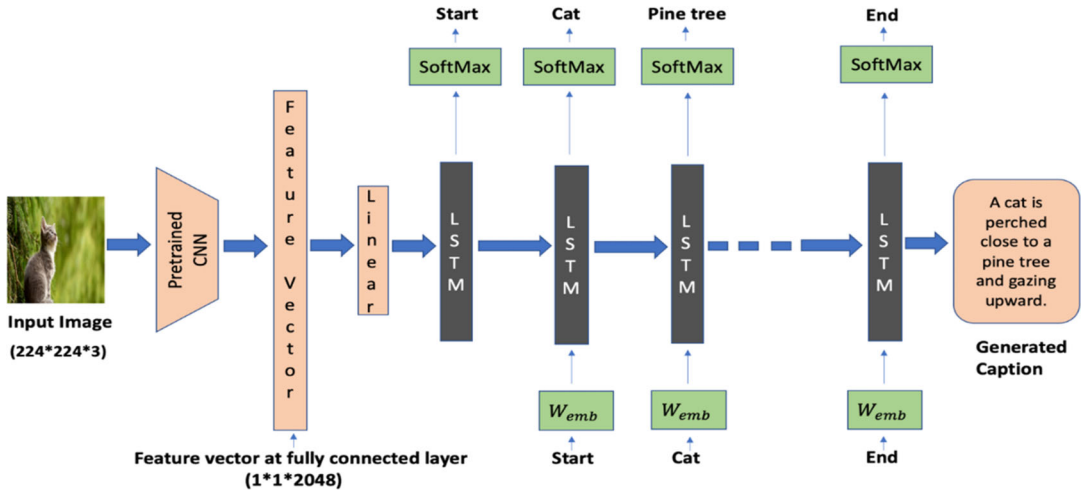


FIGURE 4B General deep learning method for image captioning.

Using a Convolutional Neural Network (CNN) that has already been trained, high-level visual information from images is extracted for use in image captioning. These attributes are then used as inputs to a sequence model, such as a Recurrent Neural Network (RNN), to produce textual captions. By converting the raw picture data into a more condensed and comprehensive representation, feature extraction in image captioning accelerates the procedure. By utilizing the image's visual context, enables the succeeding series model to concentrate on the textual part of caption development.

In the context of Convolutional Neural Networks (CNNs), a feature vector at the fully connected layer with dimensions of $1 \times 1 \times 2048$ refers to a single-dimensional vector that captures high-level abstract features extracted from the input image. (1×1) indicates the spatial dimensions of the feature map. In this case, it means that the feature map has been reduced to a single spatial location to one row and one column (2048) refers to the number of channels or feature maps in the (1×1) feature map. Each channel captures a different aspect of the image's high-level features. In the mentioned working process of generating a caption of an image, we just took an image of the cat as an input for the model using Deep Residual Networks (ResNet 50).[111] It is a pre-trained model for encoding and extracting features of that image; after that LSTM[112] was used for decoding and generating a caption of the image such as one word at a time.[113] In deep learning methods, features may manage a huge and varied set of images since they are automatically learned from training data. CNNs,[3,114] are primarily used to understand features, and a Softmax is dedicated to classification in the captioning process. Currently, CNNs proceed with RNNs[4] to generate captions of the specific image or image datasets. There has been a lot of research regarding image captioning by deep learning methods, as we can see the Table 2.

4.1 | Datasets

There are some datasets[128] that are widely used to evaluate and compare image captioning, and these can be seen in Table 3.

TABLE 2 Recent works dedicated to image captioning.

Authors	Year	Methods	Dataset
Alexander et al.[115]	2023	GPT-3, Show-Attend-Tell	Open-I, MIMIC-CXR, and MSCOCO
Yiwei et al.[116]	2023	LSTNet	MSCOCO
Yang et al.[117]	2023	CNN-LSTM	MSCOCO
H Wang et al.[118]	2023	RNN	MSCOCO
Peipei et al.[119]	2023	LSTM, CL-HRA	MSCOCO and Flickr30K
Liunian Li et al.[55]	2022	GLIP	MSCOCO
Yang et al.[120]	2022	RNN	MSCOCO and Flickr30K
Poongodi et al.[121]	2022	CNN-LSTM	Flickr 2K
I ̇nce et al.[122]	2022	CNN-LSTM	MSCOCO
Tiago et al.[123]	2022	Encoder Decoder, CNNs, GRUs	Flickr30k and MSCOCO
Kui Qian et al.[124]	2022	CNN, NLP, LSTM	Flickr8k, Flickr30k and MSCOCO
Jie Shao et al.[125]	2022	FVC R-CNN and MT-LSTM	MSCOCO
Chenggang et al.[126]	2022	Task-Adaptive attention module	MSCOCO
Jia et al.[56]	2021	ALIGN	MSCOCO and Flickr30K
Ron et al.[127]	2021	ClipCap	MSCOCO, Conceptual Captions

TABLE 3 Benchmark datasets for image captioning.

Datasets	Description	Images contained
Microsoft common objects in COntext (MSCOCO)[66]	Large-scale object detection, segmentation, and captioning dataset that contains 91 object categories	328,000
Flickr30K[129]	Contains humans involved in everyday activities and events	30,000
Flickr 8K[130]	Contains human and animal images	8000
Open Images V4[131]	Mixed images	9.2M
Pascal1K[132,133]	Mixed images	1000
Visual Genome[134]	Mixed images	108,077

4.2 | Evaluation metrics

When something is developed, it may mean that it will perform poorly. We must evaluate the model, system, architecture, and so forth, with some evaluation metrics. To achieve better results, we must assess image captioning work by standard metrics such as BLEU, ROUGE, METEOR, CIDEr, and SPICE. These are famous and widely used measures to gauge the effectiveness of developed models. The image captioning task focuses on the grammar, the richness of the defined sentence, quality, and the correct semantic way of sentences.[135]

4.2.1 | BLEU

BLUE is a Bilingual Evaluation Understudy.[136] It is a standard approach used to assess the level of generated text/sentence while doing the process of image captioning. The BLEU metric is based on *n*-gram-based precision. It scores each translation segment against a group of examples of reference translations with high quality before estimating the total rate. BLEU uses an *n*-gram matching rule in the picture description as a similarity measurement method. The predicted caption and label have n-grams and can be examined to determine the BLEU assessment metric. The

Euclidean Norm, also known as the L2 norm or 2-norm, is used to determine the accuracy or fine-grained representation when determining the length of a vector. In order to account for potential restrictions and numerical restrictions inherent in calculations of limited precision, this accuracy examines how well the Euclidean Norm captures the vector's extent in Euclidean space. The accuracy or level of information used to represent a vector using the Euclidean Norm, also known as the L2 norm or 2-norm, is referred to as the 2-norm's precision. In Euclidean space, a vector's length is expressed as its 2-norm. A vector's Euclidean Norm (2-Norm) is given as follows for a vector.

$X = [x_1, x_2, x_3, \ldots \ldots \ldots \ldots x_n]$ in n-dimensional space:

$$\left\|\mathrm{x}\right\|2 = \sqrt{\left(\mathrm{x}1^2 + \mathrm{x}2^2 + \ldots + \mathrm{x}n^2\right)}$$

Suppose there are two sentences, one in target sentence and another one in the predicted sentence, then we must find a Precision 1-gram ($p1$), a Precision of 2-gram ($p2$), a Precision of 3-gram ($p3$), a Precision 4-gram ($p4$). Further, combining these Precision Scores using the formula given in Equation (1). It can be computed for different values of N and using different weight values. Typically, we use $N = 4$ and uniform weights $w = N/4$ using Equation (1), where N is the geometric average precision

$$N = \exp\left(\sum_{\mathrm{n}=1}^{\mathrm{N}} \mathrm{w_n}\ \log\ \mathrm{p_n}\right) = \prod_{n=1}^{N} (p_n)\text{^}w_n = (p_1)\text{^}1/4 * (p_2)\text{^}1/4 * (p_3)\text{^}1/4 * (p_4)\text{^}1/4 \tag{1}$$

Now we must calculate a 'Brevity Penalty' by the following Equation (2).

$$Brevity\ Penalty = \begin{cases} 1 \\ e^{1-\frac{r}{c}}, \end{cases} \begin{pmatrix} if\ c > r \\ if\ c \leq r \end{pmatrix} \tag{2}$$

Where c is the predicted length equal to the number of words in the expected sentence, and r is the target length equal to the number of words in the target sentence. The brevity penalty is 1 if $c > r$, and the brevity penalty is $e\,(1 - r/c)$ if $\mathrm{c} \leq r$.

In this example, $c = 8$ and $r = 8$, which means Brevity Penalty = 1.

Finally, to calculate the BLEU Score, we multiply Equation (3) the Brevity Penalty with the Geometric Average of the Precision Scores (GAPS).[137]

$$Blue\ (N) = Brevity\ Penalty * GAPS\ (N) \tag{3}$$

4.2.2 | CIDEr

Each sentence is treated as a document by Consensus-based Image Description Evaluation, which is known as CIDEr.[138] The cosine angle of the word frequency inverse document frequency (TF-IDF) vector is calculated to determine how comparable the description sentence and the label are. The outcome is then produced by averaging the similarity of tuples of various lengths.

The specific formula is given as Equation (4).

$$CIDErn\ (c, S) = \frac{1}{m} \sum_{i=1}^{M} \frac{g^n(C) * g^n(Si)}{\|g^n(C)\ \| * \|g^n(Si)\ \|} \tag{4}$$

where c stands for a potential caption, S for a group of reference captions, n for an n-gram that needs to be assessed, and M for the total number of reference captions in (.) for an n-gram-based TF-IDF vector. This technique enables distinct tuples to have varied weights with various TF-IDFs since tuples that appear more frequently across the whole corpus often carry less information. As a result, CIDEr can accurately and significantly assess the grammar of descriptive sentences.

4.2.3 | METEOR

Another evaluation index used in machine translation is the Metric for Evaluation of Translation with Explicit Ordering (METEOR).[139] The METEOR metric computes the Precision and Recall for a query image caption before averaging the results. Suppose *wt* = word numbers. *wr* = word count in reference. *m* = amount of most used words between *wt* & *wr*. Then, Precision (*P*) = *m*/*wt* and *Recall*(*R*) = *m*/*wr*.

Now the harmonic mean is denoted as F mean by Equation (5).

$$F_{mean} = \frac{PR}{\boldsymbol{\alpha P} + (\mathbf{1} - \boldsymbol{\alpha})\, \boldsymbol{R}} \tag{5}$$

At last, the metric METEOR is computed as follows.

$$METEOR = (1 - pen) * F_{mean}$$

where penalty factor(pen),

$$pen = \gamma\ (cb/m)\ \theta$$

Here, the pen is the penalty factor, and *cb* stands for "chunks," which stands for an ordered block of contiguous space and hyper parameters determined such as α, θ, γ.

4.2.4 | ROUGE

Based on the co-occurrence data of the *N*-tuples in the evaluation abstracts, the ROUGE that is Recall Oriented Understudy for Gisting Evaluation[140] approach analyzes abstracts. It is a technique for measuring the recall rate of N-tuples and is employed to gauge how fluently a machine can translate. Take the frequently employed ROUGE-L as an illustration. *LCS* (*C, R*) can be expressed as the length of the longest common sub-sequence given Candidate *C* and Reference *R* by Equation (6).

$$ROUGE - L = \frac{(1 + \beta 2) * R_{LCS}\, P_{LCS}}{R_{LCS} + \beta 2\, P_{LCS}} \tag{6}$$

4.2.5 | SPICE

Semantic Propositional Image Caption Evaluation (SPICE).[141] With the help of SPICE, objects, properties, and relationships in the description sentence are encoded using a graph-based semantic representation, which is then evaluated on a semantic level. Assume that *S* stands for a collection of reference captions and that *c* is a contender. The candidate's scene graph is denoted as *G*(*c*) =" O(*c*), *E*(*c*), and *K*(*c*)," while *S*′*s* scene graph is designated as *G*(*s*) and *T* (.) denotes the transformation of a scene graph into a set of tuples with the type T (G(c)) O(c) U E(c) U K (c).

The Precision can then be represented as Equation (7).

$$P_{(c,S)} \frac{|\ T(G(c)) \otimes T(G(S))|}{|\ T(G(c))|} \tag{7}$$

and Recall can be expressed as Equation (8)

$$R_{(c,S)} \frac{|\ T(G(c)) \otimes T(G(S))|}{|\ T(G(S))|} \tag{8}$$

As a result, the metric SPICE computation may be written as the sub- sequent in Equation (9).

$$SPICE\ (c, S) = F\ 1(c, S) = \frac{2 * P(c, S) * R(c, S)}{P(c, S) + R(c, S)} \tag{9}$$

Here ⊠ binary matching operator yields a list of tuples that match. While SPICE is better at evaluating semantic information, it cannot assess the natural fluency of sentences since it ignores grammatical criteria. Additionally, as the evaluation primarily examines noun similarity, it is unsuitable for applications like machine translation.

5 | DEEP LEARNING METHODS FOR VIDEO CAPTIONING

The task of creating captions for video is very similar to that of image captioning. The main intention of the video caption process is to describe a video's visual content in natural language form.[39] We can see the use of video captioning in the subtitling video, video surveillance, the interaction of humans with machines, and so forth. The primary deep-learning method of video captioning is given in Figure 5.

Figure 5 shows the procedure for the development of the video captioning process such as

1. Take a video clip/ dataset as input.
2. Generate frames of the input video.
3. Extract Video features and separate the generated elements into visual, motion, auxiliary tag information, and audio segments.[142] It includes 2D, 3D, and semantic features. For 2D feature extraction; we can use ImageNet,[143] VGG,[144] GooglNet,[145] ResNet,[111] Inception Network,[146] EfficientNet.[147] For 3D feature extraction; we can use TSN,[148] C3d,[149] I3D,[150] and S3D,[151] for semantic feature extraction, we can use fast R-CNN,[152] CenterNet[153]; for motion feature we can use C3d,[149] I3D,[150] and S3D[151]; for audio feature extraction we can use Mel-frequency cepstrum (MFC)is a widely used audio feature information,[154] and VGGish-BiGRU network.[155]
4. After getting features we use Encoder-Decoder systems,[156] CNN,[3] YOLO,[157] LSTM,[112] and so forth, and CoreNLP[158] to get the desired caption.

There have been lots of research regarding video captioning by deep learning methods; we can see this in Table 4.

5.1 | Datasets

The key factors influencing the rapid development of this field of study have been the availability of labeled datasets for video descriptions. Except for a few datasets containing many phrases or even paragraphs per video sample, most datasets only assign one caption per video. Here we mentioned benchmark datasets for video captioning process in Table 5.

5.2 | Evaluation metrics

The metrics for the evaluation of machine generated captions of video are the same as image captioning metrics, such as Bilingual Evaluation Understudy (BLEU),[136] Consensus-based Image Description Evaluation (CIDEr),[138] Metric for evaluation of translation with explicit ordering (METEOR),[139] Recall-Oriented Understudy for Gisting Evaluation (ROUGE),[140] Semantic Propositional Image Caption Evaluation (SPICE),[141] and Word Mover's Distance (WMD).[174]

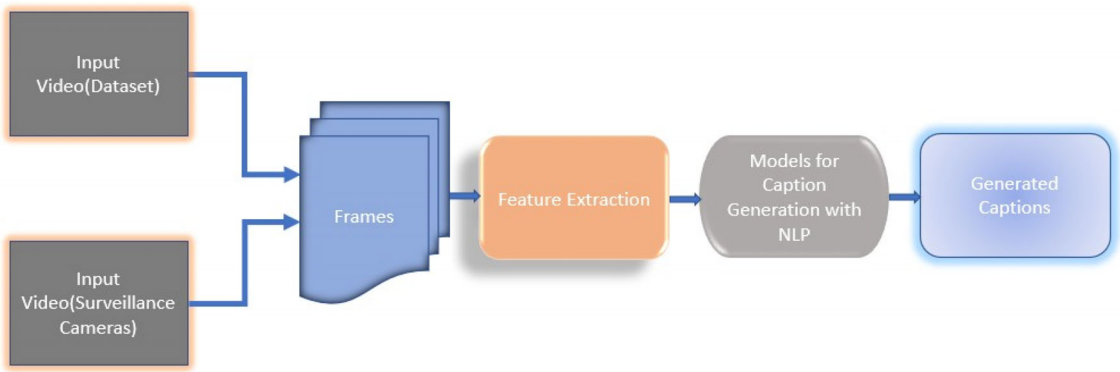


FIGURE 5 General deep learning method for video captioning.

TABLE 4 Recent works dedicated to video captioning.

Authors	Year	Methods	Dataset
Zaoad et al.[159]	2023	LSTM, BiLSTM, and GRUs	MSVD
Tian et al.[160]	2023	MesNet, GRUs	MSVD and MSR-VTT
Yaya et al.[161]	2023	Video-Text Aligned Representations (VTAR)	VATEX and MSR-VTT
Sen et al.[162]	2023	Semantic Discrimination Network (SDN)	MSVD and MSR-VTT
Lialin et al.[163]	2023	OPT, TimeSformer	MSR-VTT
Paul Hongsuck et al.[164]	2022	Encoder-Decoder model, CNN	ActivityNet-Captions, MSR-VTT, YouCook2
Kevin Lin et al.[165]	2022	SwinBERT	MSVD, YouCook2, MSR-VTT, TVC
Wanting Ji et al.[166]	2022	Attention mechanism, DNN, En-coder–Decoder	MSVD and MSR-VTT
Xia Hua et al.[167]	2022	Reinforcement Learning, OS Graph	MSVD and MSR-VTT
Liang Li et al.[168]	2022	Transformer	MSVD and MSR-VTT
Lianli Gao et al.[81]	2021	Hierarchical Representation Network	MSVD and MSR-VTT
Mingkang Tang et al.[85]	2021	CLIP4Caption, Transformer	MSR-VTT

TABLE 5 Benchmark datasets for video captioning.

Datasets	Description	Videos and sentences
MSVD (Microsoft Research Video Description Corpus)[71]	Mixed videos	1970 video clips and 80,839 sentences
MSR-VTT (Microsoft Research Video to Text)[72]	Mixed videos	10,000 video clips and 200,000 sentences
ActivityNet Captions[38]	Mixed videos	20 k videos with 100 k total descriptions
VATEX[73]	Mixed videos	41,250 videos and 825,000 captions
YouCook[169]	Cooking	88 videos and 2688 sentences
YouCook II[106]	Cooking	2000 videos and 15.4 k sentences
TACoS[170]	Cooking	127 videos
LSMDC[171]	Movies	118,081 videos
ViTT[172]	Cooking	8169 videos
BFVD[173]	E-Commerce	43,166 videos

6 | DEEP LEARNING METHODS FOR DENSE VIDEO CAPTIONING

Dense video captioning aims to identify the critical moments in the video input and create thorough captions for each one. Dense video captioning is challenging since it calls for a thorough understanding of the video's contents and contextual reasoning of specific occurrences to maintain accuracy and faith in describing events in videos.[175,176]

Dense video captioning comprises two tasks: event proposing and performing the captioning process on those events.[38] Latest work[47,106,176] follows the two-stage "detect-then-describe" framework, in which the event proposal module first predicts a set of event segments, then the captioning module constructs captions for each candidate event segment. Another line of work[91,177] removes the explicit event proposing process. We can get an idea of the dense video captioning process by the Figure 6.[91]

There is a lot of work regarding dense video captioning[49,123,176,178,179] that we mentioned in Table 6. Here we explained one of those works as mentioned in the above method. The dense caption generation was formulated as a set prediction job, as indicated in the following steps. Authors proposed a straightforward, efficient framework for beginning-to- end dense video captioning with parallel decoding (PDVC) that contains following steps:

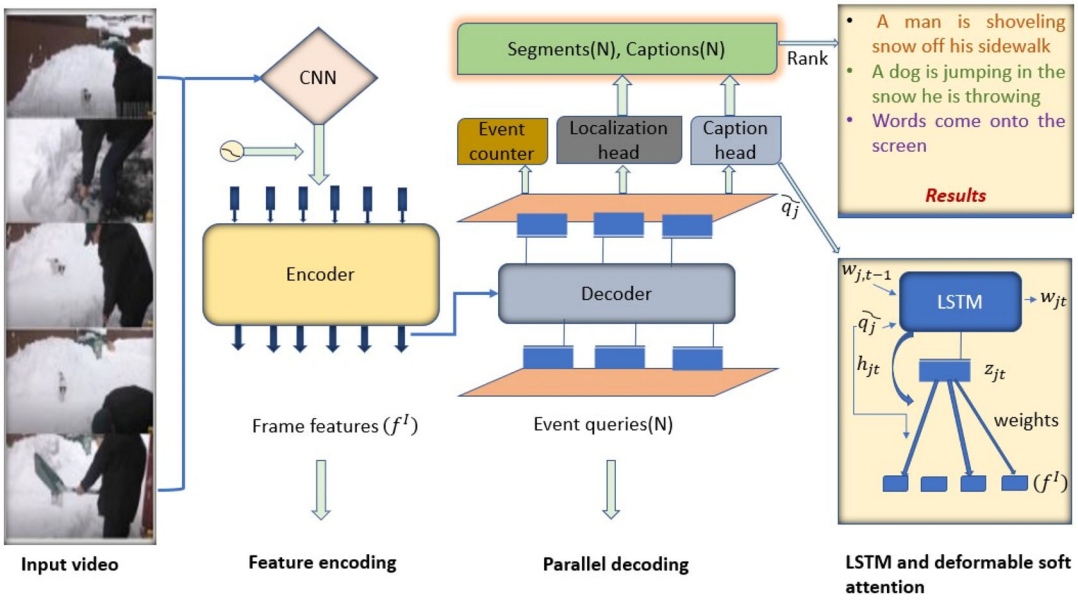


FIGURE 6 General deep learning method for dense video captioning.

TABLE 6 Recent works dedicated to dense video captioning.

Authors	Year	Methods	Dataset
Botian Shi et al.[50]	2023	BERT, LSTM, transformer	YouCook2
Yiwei et al.[180]	2023	Deformable transformer	ActivityNet Caption, YouCook2
Ping et al.[181]	2023	Transformer	ActivityNet Captions
Yang et al.[182]	2023	Vid2Seq	YouCook2, ViTT and ActivityNet
Kevin Lin et al.[165]	2022	Transformers with sparse attention	MSVD, YouCook2, MSR-VTT
Zhi Chang et al.[123]	2022	TSN, VGGish, and TEP	ActivityNet Caption and YouCook2
Nayyer Aafaq et al.[178]	2022	VSJM-Net, HDT	ActivityNet Captions, YouCook2
Wanrong Zhu et al.[176]	2022	Sequence generation	YouCook2 and ViTT
Teng Wang et al.[179]	2022	Encoder and RNN	Kinetics, ActivityNet
Shaoxiang Chen et al.[183]	2021	Sentence localizer, Event captioner	ActivityNet Captions
Gabriel Huang et al.[172]	2020	MASS-style pretraining, transformer	YouCook2, ViTT
Botian Shi et al.[50]	2019	BERT, LSTM, transformer	YouCook2
Jonghwan Mun et al.[53]	2019	Event sequence generation, Sequential network	ActivityNet Captions
Luowei Zhou et al.[48]	2018	Masked transformer	ActivityNet, YouCook2

1. Use a video feature extractor that has been trained.
2. A transformer encoder can be used to retrieve several frame-level features.
3. Then, three prediction heads and a transformer decoder are suggested to anticipate the positions and captions.
4. Quantity of events that an event query can learn.

The work offers two kinds of caption heads based on different LSTM models: vanilla LSTM and deformable soft attention enhanced LSTM. This is required because it does the task without needing to delete redundant data through non-maximum suppression, rank the captioning and localization scores to determine the top identified events during testing.

7 | KNOWLEDGE GRAPH-BASED METHODS FOR IMAGE CAPTIONING AND DENSE VIDEO CAPTIONING

7.1 | Knowledge graph

Knowledge Graphs have been used frequently in research and business, usually in close association with semantic web technologies, linked data, large scale data analytics, and cloud computing. Its popularity is influenced by the introduction of Google's Knowledge Graph in 2012.[5] Significantly, major companies such as Google, Yahoo, Microsoft and Facebook have created their own" Knowledge Graphs" that provide semantic power searches and enable smarter data processing and delivery. It could be envisaged as a network of all things relevant to a specific domain or organization. They are not limited to abstract concepts and relations but can also contain instances of documents and datasets. The knowledge graph can be described as follows:

1. Mostly discusses real world objects and the relationships between them, arranged in a graph.
2. Defines potential entity types and relationships in a schema.
3. Allows for the potential relationship between any two arbitrary entities.
4. Occupies a range of subject areas.[184]

To generate new knowledge, a knowledge graph gathers and incorporates data into an ontology, and with the help of a reasoning engine, it comes with possible solutions. Figure 7 illustrates the combination of these assumptions, which yields an abstract knowledge graph architecture.[5] Most of the prominent open knowledge graphs are Google's Knowledge Graph,[185] Google Knowledge Vault,[186] DBpedia,[187] YAGO (Yet Another Great Ontology),[188] Freebase,[189] and Wikidata,[190] NELL,[191] PROSPERA,[192] cover multiple domains, representing a broad diversity of entities and relationships.

7.2 | Knowledge graph-based methods for image captioning

As per Figure 8, the description reads, "*A woman is standing with luggage*," because the caption would only convey basic elements of the image and not why the woman is standing there. The phrases "lady and luggage," which characterize the image's key components, are actually given more importance in the earlier generated caption than in other words. The caption will read as though she might be looking for a roadside direction sign board to take down an address or waiting for a bus with her bags by combining external knowledge (Knowledge Graph).[193]

To get the caption of an image by Knowledge Graph, we must go with the following steps:

1. Feature extraction and word embedding

   Region proposal network is used to create many rectangular region proposals. Each suggestion is then supplied to three fully linked layers, an ROI pooling layer, and a vector representation of each image region. To provide a general knowledge of a picture with image, we need to generate features $V = v1, v2, \ldots, vL, viRD$ (r to the power d), the mean pooling vector v will be given to initialize the LSTM decoder.
2. Word attention

   Let the description of a picture be a sentence $S = w1, w2, \ldots, wN$, where N is the size of the narrative. These sequential learning methods typically employ RNN or LSTM, with LSTM demonstrating excellent performance, to

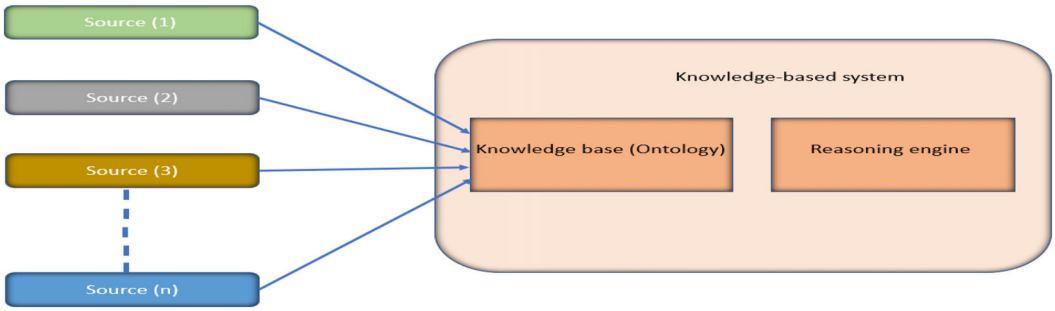


FIGURE 7 Knowledge graph.

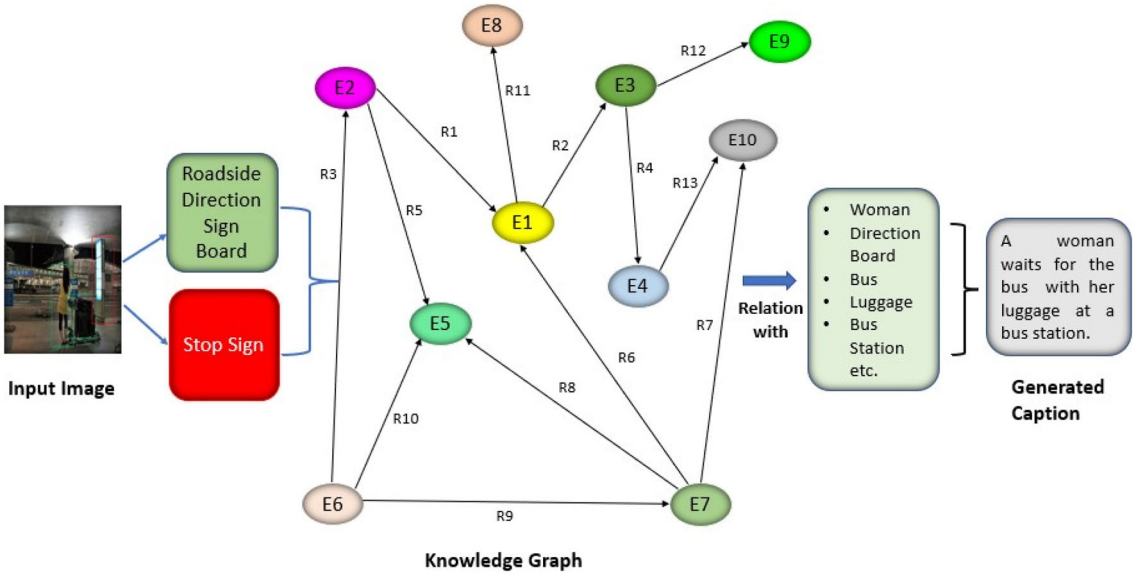


FIGURE 8 Image captioning by knowledge graph.

create every word at every point. Here the word attention is added to develop captions using the caption generator. Ground truth captions are the primary source of word attention during training.

3. Knowledge graph

   The knowledge contributed by people for caption generation, also known as internal knowledge and is represented by the ground truth annotations for each image in paired image caption datasets. However, it is only possible to include some of the information necessary for captioning tasks in many available datasets, which restricts the advancement of research. Consequently, obtaining information from outside sources to aid in caption production will enhance the generalization capabilities of the captioning model. Knowledge graphs have become increasingly common in artificial intelligence in recent years. Chin et al.[194] used ConceptNet to aid computers in comprehending human intentions. ConceptNet is an open, multilingual knowledge graph containing common sense information intimately tied to daily human life.

   Each knowledge item in the knowledge graph can be seen as a triplet (subject, rel, object), where subject and object stand in for two real- world concepts or entities and rel denotes their relationship. Faster R-CNN is used to identify several things or visual ideas. Then we use these objects or concepts to retrieve semantically related knowledge from the knowledge graph to gain an informative understanding pertinent to the given image.

4. Reinforcement learning-based sequence generation

   This reinforcement learning-based training method's central tenet is that the reward that the inference algorithm receives during testing serves as the baseline for the reinforcement algorithm. This method maintains consistency during training and inference, greatly enhancing the quality of the captions that are created.

7.3 | Knowledge graph-based methods for dense video captioning

As per work,[195] TransE[196] represents the knowledge graph, and Mask R-CNN represents the object detector. Relation represents the projected result of the TransE model, and object category represents the predicted result of the object detector. Figure 9 shows the process of dense video captioning with knowledge graph.

We can understand the work depicted in Figure 9 by following steps.

1. Object detection

   As per the above procedure, Mask R-CNN[197] detects objects in given video frames.

2. Knowledge graph

   For the relationship between the objects TransE[196] the model has been chosen based on the distributed vector representation of entities and relationships as knowledge representation.

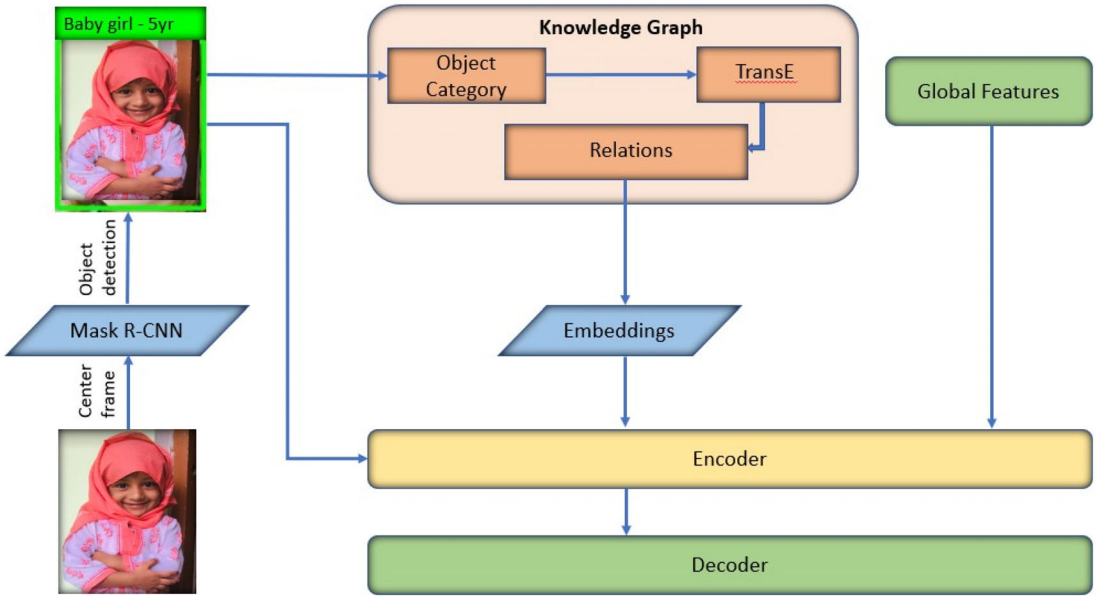


FIGURE 9 Knowledge graph module for dense video captioning.

TABLE 7 Knowledge graph-based methods dedicated to image/video captioning.

Authors	Year	Methods	Task	Dataset
Jiahui et al.[198]	2023	Knowledge Graph	Image captioning	MSCOCO
Ander et al.[199]	2023	Knowledge Graph, Transformer	Image captioning	OK-VQA
Deema et al.[200]	2023	Knowledge Graph, Transformer	Image captioning	MSCOCO
Xin et al.[201]	2023	Knowledge Graph augmented transformer	Video captioning	YouCook2, ActivityNet, MSRVTT, and MSVD
Li et al.[202]	2023	Knowledge Graph, Encoder-Decoder	Video captioning	MSCOCO, NoCaps
Mao Sheng et al.[195]	2022	Knowledge Graph, Object Detection	Video captioning	MSVD, MSR-VTT
Yu Zhang et al.[203]	2021	Knowledge Graph, Transformer	Image captioning	MSCOCO, Flickr30K
Jinglei Lou et al.[204]	2021	Knowledge Graph	Visual relation detection	MSVD
Feicheng et al.[193]	2020	Knowledge Graph, Word and visual attention reinforcement learning	Image captioning	Microsoft COCO dataset, Flickr30k
Jingyi Hou et al.[205]	2020	Knowledge Graph	Image and video captioning	MSVD, MSCOCO

3. Feature representation

   Inception-ResNet-V2 and I3D[150] have been used for feature extraction.

4. Transformer

   Transformers have been taken as the main framework for the above method to enter 2D features, 3D features, and relationship information.

Table 7 shows the recent work regarding image/video captioning by knowledge graph based methods.

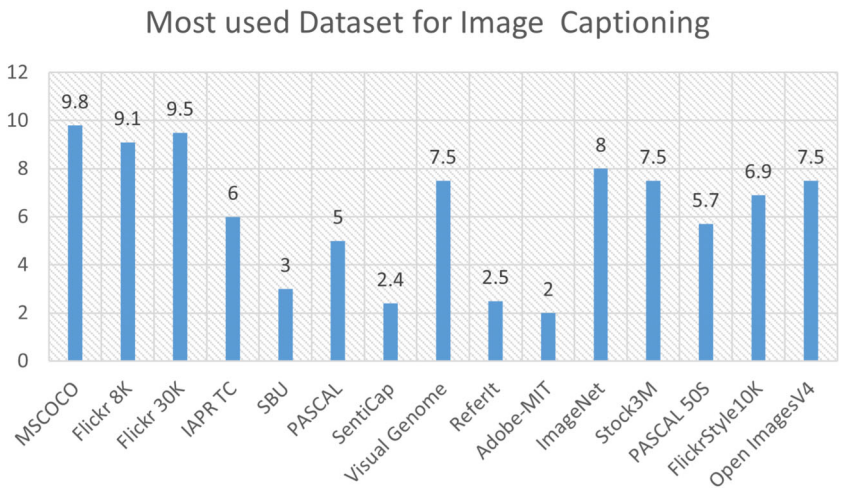


FIGURE 10 Image captioning datasets.

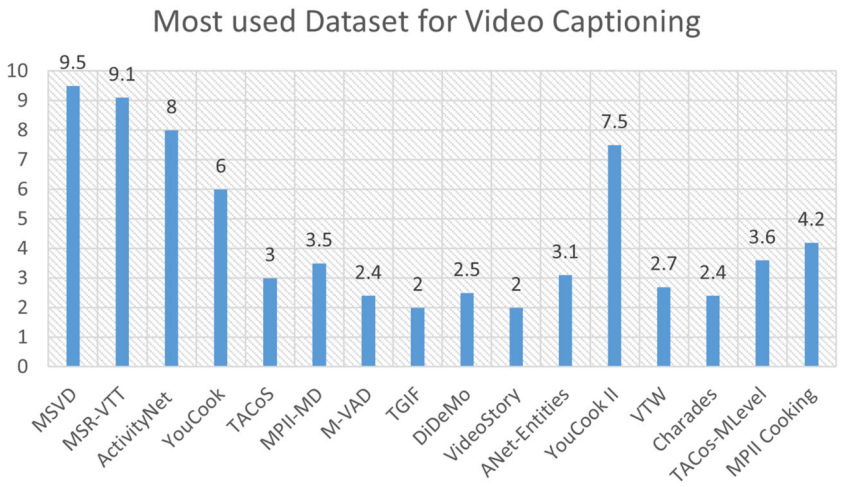


FIGURE 11 Video captioning datasets.

Datasets are the same for image caption whether anyone is taking deep learning approaches or knowledge graph-based methods for image captioning. It's the same in case of video captions, whether we go with deep learning techniques or knowledge graph-based methods for video captioning.

Based on a thorough study of existing work, we have presented in Figure 10 the most used dataset for image captioning, and the most used dataset for video captioning is presented in Figure 11, and evaluation metric in Figure 12.

8 | EVALUATION

8.1 | Comparison of image captioning work with evaluation metrics on benchmark datasets

Finally, Table 8 summarizes the evaluation outcomes of representative image captioning research conducted by academicians on benchmark datasets such as MSCOCO, Flickr30k, and Flickr8k dataset. When analyzing the MSCOCO dataset, Zhilin et al.[206] produced good results in the BLUE metric (0.670), and Marco et al.[207] produced the best results when evaluating CIDEr (0.938). Junqi et al.[208] produced outstanding results when analyzing ROUGE (0.509).

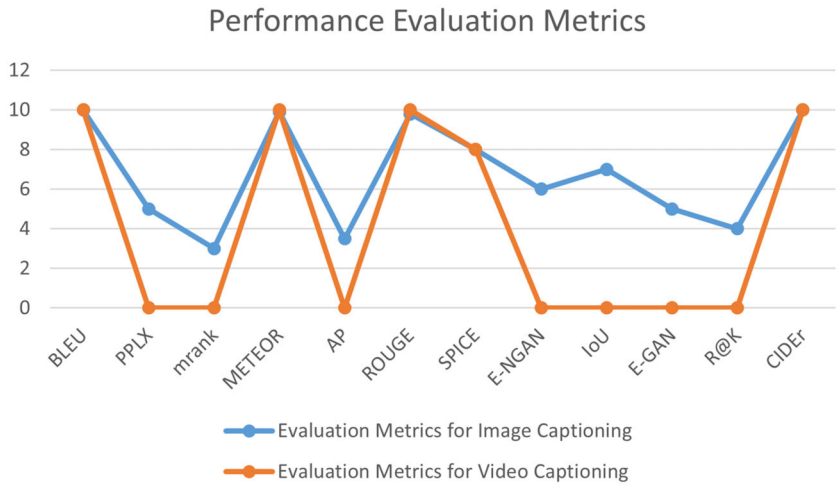


FIGURE 12 Evaluation metrics for image/video captioning.

TABLE 8 Image captioning with evaluation metrics on benchmark datasets.

		Metrics			
Authors	**Dataset**	**BLUE**	**CIDEr**	**METEOR**	**ROUGE**
Kelvin et al.[13]	**Flickr8k**	**0.670**	—	0.203	—
Xu et al.[17]	Flickr8k	0.647	—	0.201	—
Ming Jiang et al.[209]	Flickr8k	0.439	—	**0.418**	—
Jia et al.[17]	**Flickr8k**	0.646	—	0.179	—
Xu et al.[13]	Flickr30k	**0.669**	—	**0.184**	—
Zhilin et al.[206]	**MSCOCO**	**0.740**	—	**0.260**	—
Xu et al.[17]	MSCOCO	0.670	—	0.227	—
Kelvin et al.[13]	MSCOCO	0.718	—	0.230	—
Junqi et al.[208]	MSCOCO	0.282	0.838	0.235	0.509
Zhilin et al.[206]	MSCOCO	0.290	0.886	0.237	—
Marco et al.[207]	MSCOCO	0.307	0.938	0.245	—
Ming Jiang et al.[209]	MSCOCO	0.307	**1.035**	0.259	**0.545**
Qingzhong et al.[210]	MSCOCO	0.272	—	0.258	—

8.2 | Comparison of video captioning works with evaluation metrics on benchmark datasets

On benchmark datasets such as MSVD, MSR-VTT, ActivityNet, YouCook2, M-VAD, MPII-MD, Charades, TACoS MLevel, and LSMDC, researchers have performed representative video captioning work. In Table 9, we explained the evaluation findings of this work. When analyzing performance using METEOR (36.9) and ROUGE (73.9), Lee et al.[211] also produced good results on the MSVD dataset with BLUE (84.8) CIDEr (94.3) metrics.

Pan et al.[100] successfully evaluated BLUE CIDEr (93.0), METEOR (36.8), and ROUGE on the MSR-VTT dataset (73.7). Working with the ActivityNet data collection, Huijuan et al.[212] produced excellent results in CIDEr (19.88) and ROUGE (19.63). Working with the ActivityNet data collection, Wang et al.[213] had excellent results in BLUE (2.30) METEOR (9.60). Sun et al.[214] made positive results using the YouCook2 dataset. On the YouCook2 dataset, Lei et al.[215] obtained the best results in BLUE (8.0), CIDEr (35.74), and METEOR (15.9). The best results were obtained by Yao et al.[216] on the M-VAD dataset 6.1 in BLUE (0.7), CIDEr (6.1), and METEOR by Pan et al.[217] on the same dataset (7.2). With the MPII-MD dataset,

TABLE 9 Video captioning with evaluation metrics on benchmark datasets.

		Metrics			
Authors	**Dataset**	**BLUE**	**CIDEr**	**METEOR**	**ROUGE**
Aafaq et al.[221]	**MSVD**	47.9	78.1	35.0	71.5
Pan et al.[222]	MSVD	52.2	93.0	**36.9**	**73.9**
Lee et al.[211]	MSVD	**84.8**	**94.3**	—	—
Jin et al.[223]	MSVD	53.5	89.5	35.3	72.3
Zhang et al.[224]	MSVD	56.9	90.6	36.2	—
Pan et al.[222]	**MSR-VTT**	52.2	**93.0**	**36.8**	**73.7**
Zhang et al.[224]	MSR-VTT	**56.8**	90.6	36.2	—
Wang et al.[213]	MSR-VTT	39.1	42.7	26.6	59.3
Chen et al.[225]	MSR-VTT	46	52.0	29.5	63.3
Xiao et al.[226]	MSR-VTT	43.8	47.6	34.2	65.8
Huijuan et al.[212]	**ActivityNet**	1.63	**19.88**	8.58	**19.63**
Wang et al.[51]	ActivityNet	**2.30**	12.68	**9.60**	—
Sun et al.[214]	**YouCook2**	4.33	0.55	11.94	**28.8**
Lei et al.[215]	YouCook2	**8.0**	**35.74**	**15.9**	—
Yao et al.[216]	**M-VAD**	**0.7**	**6.1**	5.7	—
Pan et al.[217]	M-VAD	—	—	**7.2**	—
Lorenzo et al.[218]	**MPII-MD**	**0.8**	**10.8**	7.0	**16.7**
Yingwei et al.[217]	MPII-MD	—	—	**8.0**	—
Wang et al.[219]	**Charades**	18.8	**23.2**	**19.5**	**41.4**
Zhao et al.[220]	Charades	**31.5**	18	19.1	—
Yu et al.[227]	TACoS MLevel	30.5	160.2	28.7	—
Youngjae et al.[228]	LSMDC	—	9.3	7.2	15.6

Lorenzo et al.[218] obtained the results BLUE (0.8), CIDEr (10.8), and ROUGE (16.7). The same dataset was used by Yingwei et al.,[217] who produced positive METEOR results (8.0). Working using the Charades dataset, Wang et al.[219] obtained findings in the CIDEr (23.2), METEOR (19.5), and ROUGE (41.4). Zhao et al.[220] used the same dataset, calculated the BLUE (31.5) measure, and got good results. Additionally, we present new findings from prominent academics using various datasets.

9 | EXISTING CHALLENGES

Though deep learning and knowledge graph based methods have been used to generate captions of images and videos, there are still several unresolved issues, which are needed to be addressed. These are as follows:

1. Existing captioning systems frequently generate captions sequentially, where the next generated word depends on both the previously generated word and the picture characteristic. They often lack composition and naturalness. This frequently results in language structures that are syntactically accurate but semantically meaningless, as well as an absence of diversity in the output captions.
2. Creating rich, inventive, and human like captions bridging the semantic gap between linguistic and visual representations.
3. Current evaluation metrics still need to be improved because they ignore the image. When scoring various descriptive captions, their scoring frequently remains insufficient and misleading. Human assessment continues to be the gold standard for rating captioning systems.

4. There is a requirement to improve the higher-quality video representation approach for video captioning.
5. Need for logic and common sense in scene comprehension.

10 | CONCLUSION AND FUTURE WORK

This article reviews and evaluates most studies on image/video captioning and dense video captioning. The work classifies the captioning methods into two groups: the deep learning approaches and knowledge graph based approaches. Each category is based on each research method's fundamental characteristics and differences. Many researchers have employed various scene interpretation techniques, including the encoder-decoder and attention mechanisms. We mentioned several evaluation measures that are most frequently utilized concerning evaluation measures. We briefly overviewed the most popular datasets and evaluation metrics for dense captioning and simple captioning processes. The most appropriate datasets for image captioning are MSCOCO, Flickr8K, and Flickr30K, and for video captioning, they are MSVD and MSR-VTT. The most recent techniques are then evaluated using benchmark datasets. While carrying out this work, we mentioned numerous methods for extensive image/video captioning. The best models for extracting image/video content are CNN, RNN, and LSTM, which are widely used for language production.

By this thorough review, we also conclude that Knowledge Graph based methods are best for captioning because they can detect objects and predict relations between objects with their attributes. Incorporating knowledge graphs into the image/video captioning systems can improve the semantic understanding, consistency, and coherence of the generated captions, making them more valuable and understandable to humans. Knowledge graphs can be used to refine the performance of image/video captioning systems in several ways:

1. Providing additional contextual information that can help the captioning system to better understand the content of image/video. For example, a knowledge graph containing information about ordinary objects, scenes, and actions could guide the captioning system's attention to relevant parts of the image/video. Additionally, the knowledge graph can provide information about relationships between objects and scenes, which can help the captioning system generate more accurate and detailed captions.
2. Knowledge graphs can be used in training the model. By using knowledge graph entities to anchor the captions, also the model learns to generate semantically consistent captions with the information in the knowledge graph.
3. Knowledge graphs can be used in the post-processing stage of captioning to improve the coherence and consistency of the generated captions.

For prospective future research, this paper explores dense video captioning utilizing knowledge graph based approaches. This comprehensive analysis will assist academics better comprehend the methodology, measurements, and datasets for image/video descriptions and pave the way for future research.

AUTHOR CONTRIBUTIONS

Mohammad Saif Wajid: Carried out the experiments, reviews and wrote the manuscript and have Conceptualization (equal); formal analysis (equal); Hugo Terashima-Marin: Conceptualization (equal); formal analysis (equal); funding acquisition (equal); supervision (equal); writing – review and editing (equal). **Peyman Najafirad:** Supervision (equal); writing – original draft (equal); writing – review and editing (equal). **Mohd Anas Wajid:** Formal analysis (equal); investigation (equal); methodology (equal); resources (equal); software (equal); visualization (equal); writing – original draft (equal); writing – review and editing (equal).

FUNDING INFORMATION

This research is funded by **CONACYT** (Consejo Nacional de Ciencia y Tecnología) & **Tecnológico de Monterrey, Mexico**.

CONFLICT OF INTEREST STATEMENT

Authors declare no conflict of interest relevant to this article.

PEER REVIEW

The peer review history for this article is available at https://www.webofscience.com/api/gateway/wos/peer-review/10.1002/eng2.12785.

DATA AVAILABILITY STATEMENT

Data and information will be available on request.

ORCID

Mohammad Saif Wajid https://orcid.org/0000-0001-7572-8205

REFERENCES

1. Zhang W, Tang S, Su J, Xiao J, Zhuang Y. Tell and guess: cooperative learning for natural image caption generation with hierarchical refined attention. *Multimed Tools Appl*. 2021;80:16267-16282.
2. Cheng C, Li C, Han Y, Zhu Y. A semi-supervised deep learning image caption model based on pseudo label and n-gram. *Int J Approx Reason*. 2021;131:93-107.
3. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proc IEEE*. 1998;86:2278-2324. doi:10.1109/5.726791
4. Dupond S. A thorough review on the current advance of neural net- work structures. *Ann Rev Control*. 2019;14:200-230.
5. Ehrlinger L, Wöß W. Towards a definition of knowledge graphs. Paper presented at: SEMANTiCS (Posters, Demos, SuCCESS). 2016.
6. Aafaq N, Mian A, Liu W, Gilani SZ, Shah M. Video description: a survey of methods, datasets, and evaluation metrics. *ACM Comput Surv (CSUR)*. 2019;52:1-37.
7. Hossain MZ, Sohel F, Shiratuddin MF, Laga H. A comprehensive survey of deep learning for image captioning. *ACM Comput Surv (CsUR)*. 2019;51:1-36.
8. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: a neural image caption generator. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3156-3164.
9. Mao J, Xu W, Yang Y, Wang J, Huang Z, Yuille A. Deep captioning with multimodal recurrent neural networks (m-rnn), arXiv Preprint arXiv:1412.6632. 2014.
10. Cornia M, Baraldi L, Cucchiara R. Show, control and tell: a frame- work for generating controllable and grounded captions. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:8307-8316.
11. Karpathy A, Fei-Fei L. Deep visual-semantic alignments for generating image descriptions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:3128-3137.
12. Vinyals O, Toshev A, Bengio S, Erhan D. Show and tell: lessons learned from the 2015 mscoco image captioning challenge. *IEEE Trans Pattern Anal Mach Intell*. 2016;39:652-663.
13. Xu K, Ba J, Kiros R, et al. Show, attend and tell: neural image caption generation with visual attention. Paper presented at: International Conference on Machine Learning, PMLR. 2015:2048-2057.
14. Tan YH, Chan CS. Phrase-based image caption generator with hierarchical lstm network. *Neurocomputing*. 2019;333:86-100.
15. Soh M. *Learning Cnn-Lstm Architectures for Image Caption Generation*. Dept. Comput. Sci. Stanford Univ. Tech. Rep 1; 2016.
16. Peng Y, Liu X, Wang W, Zhao X, Wei M. Image caption model of double lstm with scene factors. *Image Vis Comput*. 2019;86:38-44.
17. Jia X, Gavves E, Fernando B, Tuytelaars T. Guiding the long-short term memory model for image caption generation. Proceedings of the IEEE International Conference on Computer Vision. 2015:2407-2415.
18. Khademi M, Schulte O. Image caption generation with hierarchical contextual visual spatial attention. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2018:1943-1951.
19. Wajid MA, Zafar A, Wajid MS, Terashima-Marın H. Neutrosophic-cnn-based image and text fusion for multimodal classification. *J Intell Fuzzy Syst*. 2023;45:1-17.
20. Lampert CH, Nickisch H, Harmeling S. Learning to detect unseen object classes by between-class attribute transfer. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. 2009:951-958.
21. Gan C, Yang T, Gong B. Learning attributes equals multi-source domain generalization. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:87-97.
22. Maji S, Bourdev L, Malik J. Action recognition from a distributed representation of pose and appearance. Paper presented at: Cvpr 2011, IEEE. 2011:3177-3184.
23. Chao Y-W, Wang Z, Mihalcea R, Deng J. Mining semantic affordances of visual object categories. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4259-4267.
24. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. *Commun ACM*. 2017;60:84-90.
25. Felzenszwalb PF, Girshick RB, McAllester D, Ramanan D. Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell*. 2010;32:1627-1645.
26. Girshick R, Donahue J, Darrell T, Malik J. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014:580-587.

27. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. *Adv Neural Inf Process Syst*. 2014;27:487-495.
28. Gong Y, Wang L, Guo R, Lazebnik S. Multi-scale orderless pooling of deep convolutional activation features. European Conference on Computer Vision, Springer. 2014:392-407.
29. Shao Z, Han J, Marnerides D, Debattista K. Region-object relation- aware dense captioning via transformer. *IEEE Trans Neural Netw Learn Syst*. 2022:3.
30. Bai S, An S. A survey on automatic image caption generation. *Neurocomputing*. 2018;311:291-304.
31. Wajid MA, Zafar A. Multimodal information access and retrieval notable work and milestones. Paper presented at: 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT), IEEE. 2019:1-6.
32. Smeaton AF, Quigley I. Experiments on using semantic distances between words in image caption retrieval. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 1996:174-180.
33. Li K, Zhang Y, Li K, Li Y, Fu Y. Visual semantic reasoning for image-text matching. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4654-4662.
34. Gurari D, Li Q, Stangl AJ, et al. Vizwiz grand challenge: answering visual questions from blind people. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:3608-3617.
35. Islam MM, Sadi MS, Zamli KZ, Ahmed MM. Developing walking assistants for visually impaired people: a review. *IEEE Sensors J*. 2019;19:2814-2828.
36. Yasir M, Zafar A, Wajid MA. Nep-2020's implementation & execution: a study conducted using neutrosophic Pestel analysis. *Int J Neutrosophic Sci*. 2023;20:86-107.
37. pixabay/cat (2017). https://pixabay.com/photos/cat-young-animal-kitten-gray-cat-2083492/.
38. Krishna R, Hata K, Ren F, Fei-Fei L, Niebles JC. Dense-captioning events in videos. Paper presented at: International Conference on Computer Vision (ICCV). 2017.
39. Chen S, Yao T, Jiang Y-G. Deep learning for video captioning: a review. Paper presented at: IJCAI. 2019.
40. Kojima A, Tamura T, Fukunaga K. Natural language description of human activities from video images based on concept hierarchy of actions. *Int J Comput Vis*. 2002;50:171-184.
41. Guadarrama S, Krishnamoorthy N, Malkarnenkar G, et al. Youtube2text: recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. Paper presented at: 2013 IEEE International Conference on Computer Vision. 2013:2712-2719.
42. Pei W, Zhang J, Wang X, Ke L, Shen X, Tai Y-W. Memory-attended recurrent network for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:8347-8356.
43. Yu Y, Ko H, Choi J, Kim G. End-to-end concept word detection for video captioning, retrieval, and question answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:3165-3173.
44. Ma L, Lu Z, Li H. Learning to answer questions from image using convolutional neural network. Paper presented at: Thirtieth AAAI Conference on Artificial Intelligence. 2016.
45. Zeng K-H, Chen T-H, Chuang C-Y, Liao Y-H, Niebles JC, Sun M. Leveraging video descriptions to learn video question answering. Paper presented at: Thirty-First AAAI Conference on Artificial Intelligence. 2017.
46. Voykinska V, Azenkot S, Wu S, Leshed G. How blind people interact with visual content on social networking services. Proceedings of the 19th Acm Conference on Computer-Supported Cooperative Work & Social Computing. 2016:1584-1595.
47. Iashin V, Rahtu E. Multi-modal dense video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020:958-959.
48. Zhou L, Zhou Y, Corso JJ, Socher R, Xiong C. End-to-end dense video captioning with masked transformer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:8739-8748.
49. Johnson J, Karpathy A, Fei-Fei L. Densecap: fully convolutional localization networks for dense captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4565-4574.
50. Shi B, Ji L, Liang Y, et al. Dense procedure captioning in narrated instructional videos. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:6382-6391.
51. Wang J, Jiang W, Ma L, Liu W, Xu Y. Bidirectional attentive fusion with context gating for dense video captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7190-7198.
52. Li Y, Yao T, Pan Y, Chao H, Mei T. Jointly localizing and describing events for dense video captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7492-7500.
53. Mun J, Yang L, Ren Z, Xu N, Han B. Streamlined dense video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:6588-6597.
54. Rahman T, Xu B, Sigal L. Watch, listen and tell: multi-modal weakly supervised dense event captioning. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:8908-8917.
55. Li LH, Zhang P, Zhang H, et al. Grounded language-image pre-training. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:10965-10975.
56. Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision. Paper presented at: International Conference on Machine Learning, PMLR. 2021:4904-4916.
57. Liu M, Li L, Hu H, Guan W, Tian J. Image caption generation with dual attention mechanism. *Inf Process Manag*. 2020;57:102178.
58. Fang H, Gupta S, Iandola F, et al. From captions to visual concepts and back. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1473-1482.

59. Li X, Yuan A, Lu X. Vision-to-language tasks based on attributes and attention mechanism. *IEEE Trans Cybern*. 2019;51:913-926.
60. Ding S, Qu S, Xi Y, Sangaiah AK, Wan S. Image caption generation with high-level image features. *Pattern Recogn Lett*. 2019;123:89-95.
61. Ishtiaque S, Wajid MS. A review on medical image compression techniques. *Int J Digit Appl Contemp Res*. 2017:17.
62. Bentley P. Pattern analysis, statistical modelling and computational learning 2004-2008. *PASCAL Netw Excellence*. 2008:2.
63. Ordonez V, Kulkarni G, Berg T. Im2text: describing images using 1 million captioned photographs. *Adv Neural Inf Process Syst*. 2011;24:1-9.
64. Chen X, Lawrence Zitnick C. Mind's eye: a recurrent visual representation for image caption generation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:2422-2431.
65. Rashtchian C, Young P, Hodosh M, Hockenmaier J. Collecting image annotations using amazon's mechanical turk. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010:139-147.
66. Lin T-Y, Maire M, Belongie S, et al. Microsoft coco: common objects in context. Paper presented at: European Conference on Computer Vision, Springer. 2014:740-755.
67. Mathews A, Xie L, He X. Senticap: generating image descriptions with sentiments. Proceedings of the AAAI Conference on Artificial Intelligence, volume 30. 2016.
68. Mishra A, Wajid MS, Dugal U. A comprehensive analysis of approaches for sentiment analysis using twitter data on COVID-19 vaccines. *J Inform Electr Electron Eng*. 2021;2:1-10.
69. Anderson P, Gould S, Johnson M. Partially supervised image captioning. *Adv Neural Inf Process Syst*. 2018;31:1-12.
70. Zhang Z, Shi Y, Yuan C, et al. Object-relational graph with teacher-recommended learning for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:13278-13288.
71. Chen D, Dolan WB. Collecting highly parallel data for paraphrase evaluation. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011:190-200.
72. Xu J, Mei T, Yao T, Rui Y. Msr-vtt: a large video description dataset for bridging video and language. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:5288-5296.
73. Wang X, Wu J, Chen J, Li L, Wang Y-F, Wang WY. Vatex: a large-scale, high-quality multilingual dataset for video-and-language research. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:4581-4591.
74. Yan C, Tu Y, Wang X, et al. Stat: spatial-temporal attention mechanism for video captioning. *IEEE Trans Multimed*. 2019;22:229-241.
75. Yang Y, Zhou J, Ai J, et al. Video captioning by adversarial lstm. *IEEE Trans Image Process*. 2018;27:5600-5611.
76. Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial networks. *Commun ACM*. 2020;63:139-144.
77. Cheng X, Lu J, Feng J, Yuan B, Zhou J. Scene recognition with objectness. *Pattern Recognit*. 2018;74:474-487.
78. Lazebnik S, Schmid C, Ponce J. Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. Paper presented at: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06), volume 2, IEEE. 2006:2169-2178.
79. Quattoni A, Torralba A. Recognizing indoor scenes. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. 2009:413-420.
80. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. Sun database: large-scale scene recognition from abbey to zoo. Paper presented at: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. 2010:3485-3492.
81. Gao L, Lei Y, Zeng P, Song J, Wang M, Shen HT. Hierarchical representation network with auxiliary tasks for video captioning and video question answering. *IEEE Trans Image Process*. 2021;31:202-215.
82. Luo H, Ji L, Zhong M, et al. Clip4clip: An empirical study of clip for end-to-end video clip retrieval, arXiv preprint arXiv:2104.08860. 2021.
83. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16×16 words: transformers for image recognition at scale, arXiv preprint arXiv:2010.11929. 2020.
84. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from natural language supervision. PMLR. 2021:8748-8763.
85. Tang M, Wang Z, Liu Z, Rao F, Li D, Li X. Clip4caption: clip for video caption. Proceedings of the 29th ACM International Conference on Multimedia. 2021:4858-4862.
86. Zhou F, De la Torre F, Cohn JF. Unsupervised discovery of facial events. Paper presented at: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE. 2010:2574-2581.
87. Wajid ZA. Neutrosophic image segmentation: An approach for the treatment of uncertainty in multimodal information systems. 2022 https://www.americaspg.com/articleinfo/21/show/1271
88. Mohd RV, Wajid S, Maurya S. Sentence similarity-based text summarization using clusters 4. 2013 https://www.ijser.org/researchpaper/Sentence-Similarity-based-text-summarization-using-clusters.pdf
89. Portillo-Quintero JA, Ortiz-Bayliss JC, Terashima-Marin H. A straightforward framework for video retrieval using clip. Paper presented at: Mexican Conference on Pattern Recognition, Springer. 2021:3-12.
90. Yang B, Zou Y. Clip meets video captioners: attribute-aware representation learning promotes accurate captioning, arXiv preprint arXiv:2111.15162. 2021.
91. Wang T, Zhang R, Lu Z, Zheng F, Cheng R, Luo P. End-to-end dense video captioning with parallel decoding. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2021:6847-6857.
92. Ryu H, Kang S, Kang H, Yoo CD. Semantic grouping network for video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35:2514-2522.
93. Wajid MS, Terashima-Marin H, Wajid MA, et al. Violence detection approach based on cloud data and neutrosophic cognitive maps. *J Cloud Comput*. 2022;11:1-18.

94. Wajid MS, Wajid MA. The importance of indeterminate and unknown factors in nourishing crime: a case study of South Africa using neutrosophy. *Neutrosophic Sets Syst*. 2021;41(2021):15.
95. Wajid MA, Zafar A. Multimodal fusion: a review, taxonomy, open challenges, research roadmap and future directions. *Neutrosophic Sets Syst*. 2021;45:8.
96. Kumar S, Singh AK, Singh P, Khan AM, Agrawal V, Wajid MS. Sentiment analysis based on ai over big data. Proceedings of the International Conference on Data Engineering and Communication Technology, Springer. 2017:641-649.
97. Ji W, Wang R. A multi-instance multi-label dual learning approach for video captioning. *ACM Trans Multimed Comput Commun Appl*. 2021;17:1-18.
98. Yang B, Zou Y, Liu F, Zhang C. Non-autoregressive coarse-to-fine video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2021;35:3119-3127.
99. Perez-Martin J, Bustos B, Pérez J. Attentive visual semantic specialized network for video captioning. Paper presented at: 2020 25th International Conference on Pattern Recognition (ICPR). 2021:5767-5774. doi:10.1109/ICPR48806.2021.9412898
100. Pan B, Cai H, Huang D-A, et al. Spatio-temporal graph for video captioning with knowledge distillation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020:10870-10879.
101. Liu W, Anguelov D, Erhan D, et al. Ssd: single shot multibox detector. European Conference on Computer Vision, Springer. 2016:21-37.
102. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. *Adv Neural Inf Process Syst*. 2017;30:1-11.
103. Ma C-Y, Kadav A, Melvin I, Kira Z, AlRegib G, Graf HP. Attend and interact: higher-order object interactions for video understanding. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:6790-6800.
104. Xiong Y, Dai B, Lin D. Move forward and tell: a progressive generator of video descriptions. Proceedings of the European Conference on Computer Vision (ECCV). 2018:468-483.
105. Yu Z, Han N. Accelerated masked transformer for dense video captioning. *Neurocomputing*. 2021;445:72-80.
106. Zhou L, Xu C, Corso JJ. Towards automatic learning of procedures from web instructional videos. Paper presented at: Thirty-Second AAAI Conference on Artificial Intelligence. 2018.
107. Hendricks LA, Venugopalan S, Rohrbach M, Mooney R, Saenko K, Darrell T. Deep compositional captioning: describing novel object categories without paired training data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:1-10.
108. Li G, Su H, Zhu W. Incorporating external knowledge to answer open-domain visual questions with dynamic memory networks, arXiv preprint arXiv:1712.00733. 2017.
109. Gu J, Zhao H, Lin Z, Li S, Cai J, Ling M. Scene graph generation with external knowledge and image reconstruction. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:1969-1978.
110. Zhou Y, Sun Y, Honavar V. Improving image captioning by leveraging knowledge graphs. Paper presented at: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. 2019:283-293.
111. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:770-778.
112. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*. 1997;9:1735-1780.
113. Geetha G, Kirthigadevi T, Ponsam GG, Karthik T, Safa M. Image captioning using deep convolutional neural networks (cnns). *Journal of Physics: Conference Series*. Vol 1712. IOP Publishing; 2020:012015.
114. Gupta P, Siddiqui MK, Huang X, et al. Covid-widenet—a capsule network for covid-19 detection. *Appl Soft Comput*. 2022;122:108780.
115. Selivanov A, Rogov OY, Chesakov D, Shelmanov A, Fedulova I, Dylov DV. Medical image captioning via generative pretrained transformers. *Sci Rep*. 2023;13:4171.
116. Ma Y, Ji J, Sun X, Zhou Y, Ji R. Towards local visual modeling for image captioning. *Pattern Recognit*. 2023;138:109420.
117. Yang X, Zhang H, Gao C, Cai J. Learning to collocate visual- linguistic neural modules for image captioning. *Int J Comput Vis*. 2023;131:82-100.
118. Wang H, Wang H, Xu K. Evolutionary recurrent neural network for image captioning. *Neurocomputing*. 2020;401:249-256.
119. Zhu P, Wang X, Zhu L, et al. Prompt-based learning for unpaired image captioning. *IEEE Trans Multimed*. 2023:1-13.
120. Yang L, Wang H, Tang P, Li Q. Captionnet: a tailor-made recurrent neural network for generating image descriptions. *IEEE Trans Multimed*. 2021;23:835-845. doi:10.1109/TMM.2020.2990074
121. Poongodi M, Hamdi M, Wang H. Image and audio caps: automated captioning of background sounds and images using deep learning. *Multimed Syst*. 2022;29:1-9.
122. Ince M. Automatic and intelligent content visualization system based on deep learning and genetic algorithm. *Neural Comput Appl*. 2022;34:2473-2493.
123. do Carmo Nogueira T, Vinhal CDN, da Cruz Júnior G, Ullmann MRD, Marques TC. A reference-based model using deep learning for image captioning. *Multimed Syst*. 2022;29:1-17.
124. Qian K, Tian L. A topic-based multi-channel attention model under hybrid mode for image caption. *Neural Comput Appl*. 2022;34:2207-2216.
125. Shao J, Yang R. Controllable image caption with an encoder-decoder optimization structure. *Appl Intell*. 2022;52:1-12.
126. Yan C, Hao Y, Li L, et al. Task-adaptive attention for image captioning. *IEEE Trans Circuits Syst Video Technol*. 2022;32:43-51. doi:10.1109/TCSVT.2021.3067449
127. Mokady R, Hertz A, Bermano AH. Clipcap: clip prefix for image captioning, arXiv preprint arXiv:2111.09734. 2021.

128. Amirian S, Rasheed K, Taha TR, Arabnia HR. A Short Review on Image Caption Generation with Deep Learning. Proceedings of the International Conference on Image Processing, Computer Vision, and Pattern Recognition (IPCV). The Steering Committee of The World Congress in Computer Science, Computer. 2019:10-18.
129. Young P, Lai A, Hodosh M, Hockenmaier J. From image descriptions to visual denotations: new similarity metrics for semantic inference over event descriptions. *Trans Assoc Comput Linguist*. 2014;2:67-78.
130. Hodosh M, Young P, Hockenmaier J. Framing image description as a ranking task: data, models and evaluation metrics. *J Artif Intell Res*. 2013;47:853-899.
131. Kuznetsova A, Rom H, Alldrin N, et al. The open images dataset v4. *Int J Comput Vis*. 2020;128:1956-1981.
132. vision.cs. Collecting image annotations using amazon's mechanical turk. 2010 https://vision.cs.uiuc.edu/pascal-sentences/
133. Callison-Burch C, Dredze M. Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk. 2010.
134. Krishna R, Zhu Y, Groth O, et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int J Comput Vis*. 2017;123:32-73.
135. Luo G, Cheng L, Jing C, Zhao C, Song G. A thorough review of models, evaluation metrics, and datasets on image captioning. *IET Image Process*. 2022;16:311-332.
136. Papineni K, Roukos S, Ward T, Zhu W-J. Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. 2002:311-318.
137. Doshi K. Foundations of Nlp Explained—Bleu Score and Wer Metrics, toward Data Science 9. 2021.
138. Vedantam R, Lawrence Zitnick C, Parikh D. Cider: consensus-based image description evaluation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:4566-4575.
139. Banerjee S, Lavie A. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005:65-72.
140. Lin C-Y, Hovy E. Automatic evaluation of summaries using n-gram co-occurrence statistics. Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. 2003:150-157.
141. Anderson P, Fernando B, Johnson M, Gould S. Spice: semantic propositional image caption evaluation. European Conference on Computer Vision, Springer. 2016:382-398.
142. bozliu.medium.com. Video captioning-automatic description generation from digital video. 2021 https://bozliu.medium.com/video-captioning-c514af809ec
143. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. Paper presented at: 2009 IEEE Conference on Computer Vision and Pattern Recognition, IEEE. 2009:248-255.
144. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034. 2013.
145. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015:1-9.
146. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:2818-2826.
147. Tan M, Le Q. Efficientnet: rethinking model scaling for convolutional neural networks. Paper presented at: International Conference on Machine Learning, PMLR. 2019:6105-6114.
148. Wang L, Xiong Y, Wang Z, et al. Temporal segment networks: towards good practices for deep action recognition. European Conference on Computer Vision, Springer. 2016:20-36.
149. Tran D, Bourdev L, Fergus R, Torresani L, Paluri M. Learning spatiotemporal features with 3d convolutional networks. Proceedings of the IEEE International Conference on Computer Vision. 2015:4489-4497.
150. Carreira J, Zisserman A. Quovadis, action recognition? A new model and the kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6299-6308.
151. Xie S, Sun C, Huang J, Tu Z, Murphy K. Rethinking spatiotemporal feature learning: speed-accuracy trade-offs in video classification. Proceedings of the European Conference on Computer Vision (ECCV). 2018:305-321.
152. Ren S, He K, Girshick R, Sun J. Faster rcnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst*. 2015;28:1-14.
153. Zhou X, Wang KP. Objects as points [j/ol], arXiv preprint arXiv:1904.07850. 2019.
154. Han W, Chan C-F, Choy C-S, Pun K-P. An efficient mfcc extraction method in speech recognition. Paper presented at: 2006 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE. 2006:4.
155. Shi L, Du K, Zhang C, Ma H, Yan W. Lung sound recognition algorithm based on vggish-bigru. *IEEE Access*. 2019;7:139438-139449.
156. Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K. Sequence to sequence-video to text. Proceedings of the IEEE International Conference on Computer Vision. 2015:4534-4542.
157. Redmon J, Farhadi A. Yolo9000: better, faster, stronger. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:7263-7271.
158. Manning CD, Surdeanu M, Bauer J, Finkel JR, Bethard S, Mc- Closky D. The Stanford corenlp natural language processing toolkit. Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations. 2014:55-60.
159. Zaoad MS, Mannan MR, Mandol AB, Rahman M, Islam MA, Rahman MM. An attention-based hybrid deep learning approach for bengali video captioning. *J King Saud Univ Comput Inf Sci*. 2023;35:257-269.

160. Niu T-Z, Dong S-S, Chen Z-D, et al. A multi-layer memory sharing network for video captioning. *Pattern Recognit*. 2023;136:109202.
161. Shi Y, Xu H, Yuan C, Li B, Hu W, Zha Z-J. Learning video-text aligned representations for video captioning. *ACM Trans Multimed Comput Commun Appl*. 2023;19:1-21.
162. Du S, Zhu H, Xiong G, et al. Semantic similarity information discrimination for video captioning. *Expert Syst Appl*. 2023;213:118985.
163. Lialin V, Rawls S, Chan D, Ghosh S, Rumshisky A, Hamza W. Scalable and accurate self-supervised multimodal representation learning without aligned video and text data. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. 2023:390-400.
164. Seo PH, Nagrani A, Arnab A, Schmid C. End-to-end generative pretraining for multimodal video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:17959-17968.
165. Lin K, Li L, Lin C-C, et al. Swinbert: end-to-end transformers with sparse attention for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022:17949-17958.
166. Ji W, Wang R, Tian Y, Wang X. An attention based dual learning approach for video captioning. *Appl Soft Comput*. 2022;117:108332.
167. Hua X, Wang X, Rui T, Shao F, Wang D. Adversarial reinforcement learning with object-scene relational graph for video captioning. *IEEE Trans Image Process*. 2022;31:2004-2016. doi:10.1109/TIP.2022.3148868
168. Li L, Gao X, Deng J, Tu Y, Zha Z-J, Huang Q. Long short-term relation transformer with global gating for video captioning. *IEEE Trans Image Process*. 2022;31:2726-2738. doi:10.1109/TIP.2022.3158546
169. 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. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2013:2634-2641.
170. Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M. Grounding action descriptions in videos. *Trans Assoc Comput Linguist*. 2013;1:25-36.
171. openaccess.thecvf.com. A dataset for movie description. 2015 https://openaccess.thecvf.com/contentcvpr2015/papers/RohrbachADatasetfor2015CVPRpaper.pdf.
172. Huang G, Pang B, Zhu Z, Rivera C, Soricut R. Multimodal pre-training for dense video captioning, arXiv preprint arXiv:2011.11760. 2020.
173. Zhang S, Tan Z, Yu J, et al. Poet: product-oriented video captioner for e-commerce. Proceedings of the 28th ACM International Conference on Multimedia. 2020:1292-1301.
174. Kusner M, Sun Y, Kolkin N, Weinberger K. From word embed- dings to document distances. Paper presented at: International Conference on Machine Learning, PMLR. 2015:957-966.
175. Suin M, Rajagopalan A. An efficient framework for dense video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34:12039-12046.
176. Zhu W, Pang B, Thapliyal A, Wang WY, Soricut R. End-to- end dense video captioning as sequence generation, arXiv preprint arXiv:2204.08121. 2022.
177. Deng C, Chen S, Chen D, He Y, Wu Q. Sketch, ground, and refine: top-down dense video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:234-243.
178. Aafaq N, Mian AS, Akhtar N, Liu W, Shah M. Dense video captioning with early linguistic information fusion. *IEEE Trans Multimed*. 2022;25:2309-2322.
179. Wang T, Liu Z, Zheng F, Lu Z, Cheng R, Luo P. Semantic- aware pretraining for dense video captioning, arXiv preprint arXiv:2204.07449. 2022.
180. Wei Y, Yuan S, Chen M, et al. Mpp- net: multi-perspective perception network for dense video captioning, Available at SSRN 4346395. 2023.
181. Li P, Zhang P, Wang T, Xiao H. Time–frequency recurrent trans- former with diversity constraint for dense video captioning. *Inf Process Manag*. 2023;60:103204.
182. Yang A, Nagrani A, Seo PH, et al. Vid2seq: large-scale pretraining of a visual language model for dense video captioning, arXiv preprint arXiv:2302.14115. 2023.
183. Chen S, Jiang Y-G. Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:8425-8435.
184. P. Cimiano, H. Paulheim, Knowledge Graph refinement: a survey of approaches and evaluation methods, *Sem Ther*. 8 (2017) 489–508. 10.3233/SW-160218
185. blog.google. Introducing the knowledge graph: things, not strings. 2012 https://blog.google/products/search/introducing-knowledge-graph-things-not/
186. Dong X, Gabrilovich E, Heitz G, et al. Knowledge vault: a web-scale approach to probabilistic knowledge fusion. Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2014.
187. Lehmann J, Isele R, Jakob M, et al. Dbpedia-a large-scale, multilingual knowledge base extracted from Wikipedia. *Semant Web*. 2015;6:167-195.
188. Suchanek FM, Kasneci G, Weikum G. Yago: a core of semantic knowledge. Proceedings of the 16th International Conference on World Wide Web, WWW'07, Association for Computing Machinery, New York, NY, USA. 2007:697-706. doi:10.1145/1242572.1242667
189. Bollacker K, Cook R, Tufts P. Freebase: a shared database of structured general human knowledge. Proceedings of the 22nd National Conference on Artificial Intelligence-Volume 2, AAAI'07, AAAI Press. 2007:1962-1963.
190. Vrandeić D, Krötzsch M. Wikidata: a free collaborative knowledge-base. *Commun ACM*. 2014;57:78-85.
191. Carlson A, Betteridge J, Kisiel B, Settles B, Hruschka ER, Mitchell TM. Toward an architecture for never-ending language learning. Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI'10, AAAI Press. 2010:1306-1313.

192. Nakashole N, Theobald M, Weikum G. Scalable knowledge harvesting with high precision and high recall. Proceedings of the Fourth ACM International Conference on Web Search and Data mining, WSDM '11, Association for Computing Machinery, New York, NY, USA. 2011:227-236. doi:10.1145/1935826.1935869
193. Huang F, Li Z, Wei H, Zhang C, Ma H. Boost image captioning with knowledge reasoning. *Mach Learn*. 2020;109:2313-2332.
194. Speer R, Chin J, Havasi C. Conceptnet 5.5: An open multilingual graph of general knowledge. Paper presented at: Thirty-First AAAI Conference on Artificial Intelligence. 2017.
195. Zhong M, Zhang H, Xiong H, Chen Y, Wang M, Zhou X. Kgvideo: a video captioning method based on object detection and knowledge graph, Available at SSRN 4017055. 2022.
196. Bordes A, Usunier N, Garcia-Duran A, Weston J, Yakhnenko O. Translating embeddings for modeling multi-relational data. *Adv Neural Inf Process Syst*. 2013;26:1-9.
197. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. Proceedings of the IEEE International Conference on Computer Vision. 2017:2961-2969.
198. Wei J, Li Z, Zhu J, Ma H. Enhance understanding and reasoning ability for image captioning. *Appl Intell*. 2023;53:2706-2722.
199. Salaberria Saizar A, Azkune Galparsoro G, López de La- calle Lecuona O, Soroa Echave A, Agirre Bengoa E. Image captioning for effective use of language models in knowledge-based visual question answering. 2023.
200. Hafeth DA, Kollias S, Ghafoor M. Semantic representations with attention networks for boosting image captioning. *IEEE Access*. 2023;11:40230-40239.
201. Gu X, Chen G, Wang Y, Zhang L, Luo T, Wen L. Text with knowledge graph augmented transformer for video captioning, arXiv preprint arXiv:2303.12423. 2023.
202. Li W, Zhu L, Wen L, Yang Y. Decap: decoding clip latents for zero- shot captioning via text-only training, arXiv preprint arXiv:2303.03032. 2023.
203. Zhang Y, Shi X, Mi S, Yang X. Image captioning with transformer and knowledge graph. *Pattern Recogn Lett*. 2021;143:43-49.
204. Lou J, Li A, Cao Y, et al. Visual information-oriented knowledge graph. Paper presented at: 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), IEEE. 2021:1-5.
205. Hou J, Wu X, Zhang X, Qi Y, Jia Y, Luo J. Joint commonsense and relation reasoning for image and video captioning. Proceedings of the AAAI Conference on Artificial Intelligence. 2020;34:10973-10980.
206. Yang Z, Yuan Y, Wu Y, Cohen WW, Salakhutdinov RR. Review networks for caption generation. *Adv Neural Inf Process Syst*. 2016;29.
207. Pedersoli M, Lucas T, Schmid C, Verbeek J. Areas of attention for image captioning. Proceedings of the IEEE International Conference on Computer Vision. 2017:1242-1250.
208. Jin J, Fu K, Cui R, Sha F, Zhang C. Aligning where to see and what to tell: image caption with region-based attention and scene factorization, arXiv preprint arXiv:1506.06272. 2015.
209. Jiang M, Huang Q, Zhang L, et al. Tiger: text-to-image grounding for image caption evaluation, arXiv preprint arXiv:1909.02050. 2019.
210. Wang Q, Wan J, Chan AB. On diversity in image captioning: metrics and methods. *IEEE Trans Pattern Anal Mach Intell*. 2020;44:1035-1049.
211. Lee S, Kim I. Multimodal feature learning for video captioning. *Math Probl Eng*. 2018;2018:937-941.
212. Xu H, Li B, Ramanishka V, Sigal L, Saenko K. Joint event detection and description in continuous video streams. Paper presented at: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV), IEEE. 2019:396-405.
213. Wang B, Ma L, Zhang W, Liu W. Reconstruction network for video captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:7622-7631.
214. Sun C, Myers A, Vondrick C, Murphy K, Schmid C. Videobert: a joint model for video and language representation learning. Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019:7464-7473.
215. Lei J, Wang L, Shen Y, Yu D, Berg TL, Bansal M. Mart: memory- augmented recurrent transformer for coherent video paragraph captioning, arXiv preprint arXiv:2005.05402. 2020.
216. Yao L, Torabi A, Cho K, et al. Describing videos by exploiting temporal structure. Proceedings of the IEEE International Conference on Computer Vision. 2015:4507-4515.
217. Pan Y, Yao T, Li H, Mei T. Video captioning with transferred semantic attributes. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:6504-6512.
218. Baraldi L, Grana C, Cucchiara R. Hierarchical boundary-aware neural encoder for video captioning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:1657-1666.
219. Wang X, Chen W, Wu J, Wang Y-F, Wang WY. Video captioning via hierarchical reinforcement learning. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018:4213-4222.
220. Zhao B, Li X, Lu X, et al. Video captioning with tube features. IJCAI. 2018:1177-1183.
221. Aafaq N, Akhtar N, Liu W, Gilani SZ, Mian A. Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:12487-12496.
222. Boxiao P, Haoye C, Huang D-A, et al. Spatio-temporal graph for video captioning with knowledge distillation. 2021.
223. Jin T, Huang S, Chen M, Li Y, Zhang Z. Sbat: video captioning with sparse boundary-aware transformer, arXiv preprint arXiv:2007.11888. 2020.
224. Zhang J, Peng Y. Object-aware aggregation with bidirectional temporal graph for video captioning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019:8327-8336.
225. Chen H, Li J, Hu X. Delving deeper into the decoder for video captioning, arXiv preprint arXiv:2001.05614. 2020.

226. Xiao H, Shi J. Diverse video captioning through latent variable expansion. *Pattern Recogn Lett*. 2022;160:19-25.
227. Yu H, Wang J, Huang Z, Yang Y, Xu W. Video paragraph captioning using hierarchical recurrent neural networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016:4584-4593.
228. Yu Y, Choi J, Kim Y, Yoo K, Lee S-H, Kim G. Supervising neural attention models for video captioning by human gaze data. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017:490-498.

How to cite this article: Wajid MS, Terashima-Marin H, Najafirad P, Wajid MA. Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods. *Engineering Reports*. 2024;6(1):e12785. doi: 10.1002/eng2.12785