Artificial Intelligence Review (2023) 56:13293–13372
https://doi.org/10.1007/s10462-023-10414-6

Video description: A comprehensive survey of deep learning approaches

Ghazala Rafiq[1] · **Muhammad Rafiq**[2] · **Gyu Sang Choi**[1]

Published online: 11 April 2023

Abstract
Video description refers to understanding visual content and transforming that acquired understanding into automatic textual narration. It bridges the key AI fields of computer vision and natural language processing in conjunction with real-time and practical applications. Deep learning-based approaches employed for video description have demonstrated enhanced results compared to conventional approaches. The current literature lacks a thorough interpretation of the recently developed and employed sequence to sequence techniques for video description. This paper fills that gap by focusing mainly on deep learning-enabled approaches to automatic caption generation. Sequence to sequence models follow an Encoder–Decoder architecture employing a specific composition of CNN, RNN, or the variants LSTM or GRU as an encoder and decoder block. This standard-architecture can be fused with an attention mechanism to focus on a specific distinctiveness, achieving high quality results. Reinforcement learning employed within the Encoder–Decoder structure can progressively deliver state-of-the-art captions by following exploration and exploitation strategies. The transformer mechanism is a modern and efficient transductive architecture for robust output. Free from recurrence, and solely based on self-attention, it allows parallelization along with training on a massive amount of data. It can fully utilize the available GPUs for most NLP tasks. Recently, with the emergence of several versions of transformers, long term dependency handling is not an issue anymore for researchers engaged in video processing for summarization and description, or for autonomous-vehicle, surveillance, and instructional purposes. They can get auspicious directions from this research.

Keywords Deep learning · Encoder–Decoder architecture · Text description · Video captioning techniques · Video description approaches · Video captioning · Vision to text

✉ Muhammad Rafiq
rafiq@kmu.ac.kr

✉ Gyu Sang Choi
castchoi@ynu.ac.kr

Extended author information available on the last page of the article

1 Introduction

The global-village phenomenon is strengthening day by day. Technological advancements, the abundance of devices, automation, and widespread availability of the internet has connected people like never before. People exchange texts, images, and videos for communication, resulting in a massive amount of textual and visual data. These copiously available videos linked with accurate processing can help to address numerous real-world challenges in various disciplines of life. No doubt, human beings are intelligent enough to know that understanding visual aspects and language intricacies is among the inherent capabilities they possess. However, for machines to be rationally sharp enough, proper understanding of images and their consequential interpretations are essential for content description. The primary objective of video description is to provide a concise and accurate textual alternative to visual content. Researchers put considerable effort into understanding visual characteristics and generating an eloquent interpretation, i.e., a video description that is a blend of vision and language encompassing the two prominent domains of CV and NLP (Bhatt et al. 2017). Scientists from both areas have mutually worked on getting appropriate insights from images and videos, and then accurately and precisely interpreting them by considering all the elements appearing in the video frames, like the objects, actions, interactions, backgrounds, overlapping scenes with localization information, and most importantly, their temporal sequence. Table 1 lists abbreviations with full form.

The importance of video description is evident from its practical and real time applications (including efficient searching and indexing of videos on the internet, human-robot relationships in industrial zones, and facilitation of autonomous vehicle driving), and video descriptions can outline procedures in instructional/tutorial videos for industry, education, and the household (e.g., recipes). The visually impaired can gain useful information from a video that incorporates audio descriptions. Long surveillance videos can be transformed into short texts for quick previews. Sign language videos can be converted to natural language descriptions. Automatic, accurate, and precise video/movie subtitling is another important and practical application of the video description task.

1.1 Classical approach

Video description research began with the classical approach (Rohrbach et al. 2013; Kojima et al. 2002; Khan and Gotoh 2012; Barbu et al. 2012; Das et al. 2013; Hakeem et al. 2004) where, after identification of subject, verb, and object in a constrained domain video, the fitting of the SVO in a standard predefined template was performed. These classical methods were effective only for short video clips with a limited number of objects and minimal interactions. For semantic verification, the research in Das et al. (2013) developed a hybrid model addressing the issues in Khan and Gotoh (2012) and Barbu et al. (2012) , combining the best aspects of bottom-up and top-down exploitation of rich semantic spaces of both visual and textual features. They produced high-relevance content beyond simple keyword annotations. The SVO tuple based methods can be split up into two phases for the better performance of video captioning system, i.e., phase-I is the content identification and phase-II is the sentence generation for the identified objects/events/actions in phase-I. Methods for identification (phase-I) include edge detection/color matching (Kojima et al. 2002), Scale Invariant Feature Transform (SIFT) (Lowe 1999) and context based object recognition (Torralba et al. 2003) whereas for sentence generation phase

Table 1 List of Abbreviations

Abbreviation	Full Form	Abbreviation	Full Form
ACT	Adaptive Computation Time	MPII-MD	Max Plank Institute for Informatics - Movie Description
ADs	Audio Descriptions	MSR-VTT	Microsoft research - Video To Text
AMT	Amazon Mechanical Turk	MSVD	Microsoft Video Description
BAST	Bag of aggregated semantic tuples	M-VAD	Montreal Video Annotation Dataset
BERT	Bidirectional Encoder representations from Transformers	NAVC	Non-Auto Regressive Video Captioning
BFVD	Buyer-generated Fashion Video Dataset	NLP	Natural Language Processing
BLEU	Bilingual Evaluation Understudy	NMT	Neural Machine Translation
BP	Brevity Penalty	NN	Neural Network
C3D	3D-CNN	RNN	Recurrent Neural Network
CGAN	Conditional Generative Adversarial Network	Rouge	Recall Oriented Understudy for Gisting Evaluation
CIDEr	Consensus-based Image Description Evaluation	Rouge-L	Recall Oriented Understudy for Gisting Evaluation: Longest Common Subsequence
CNN	Convolution Neural Network	Rouge-N	Recall Oriented Understudy for Gisting Evaluation : n-gram Co-occurrence
CV	Computer Vision	Rouge-S	Recall Oriented Understudy for Gisting Evaluation : Skip-bigram Co-occurrence
DCE	Diverse Captioning Evaluation metric	Rouge-W	Recall Oriented Understudy for Gisting Evaluation : Weighted Longest Common Subsequence
DRL	Deep Reinforcement Learning	SDN	Semantic Detection Network
DVS	Descriptive Video Service	SPICE	Semantic Propositional Image Caption Evaluation
ECN	Efficient Convolution Network	SSIM	Structural Similarity Index Measure
ED	Encoder-Decoder	ST	Standard Transformer
EMD	Earth Mover's Distance	SVO	Subject Verb Object
FFNN	Feed-Forward Neural Network	TACoS	Textually Annotated Cooking Scenes
FFVD	Fan-generated Fashion Video Dataset	TRECVID	Text Retrieval Conference Video Retrieval evaluation
GAN	Generative Adversarial Network	TvT	Two Viewed Transformer
GPU	Graphical Processing Unit	UGVs	User Generated Videos
GRU	Gated Recurrent Unit	UT	Universal Transformer
HRL	Hierarchical Reinforcement Learning	VATEX	Video And Text

Table 1 (continued)

Abbreviation	Full Form	Abbreviation	Full Form
LSH	Locality Sensitive Hashing	VCR	Visual Commonsense Reasoning
LSMDC	Large Scale Movie Description Challenge	VQA	Visual Question Answering
LSTM	Long Short-term Memory	VTW	Video Titles in the Wild
METEOR	Metric for Evaluation of Translation with Explicit Ordering	WMD	Word Mover's Distance
MLE	Maximum Likelihood Estimation	WN	WordNet
MPII	Max Plank Institute of Informatics	XE	Cross Entropy

there exists HALogen (Langkilde-geary and Knight 2002) representation and Head-driven Phrase Structure Grammar (HPSG) (Levine and Meurers 2006).

The methods adopted for the task of image/video captioning can be segregated into two broad categories of retrieval based and template based approaches. In retrieval based methods, the captions are retrieved from a set of existing captions. These methods first find candidate captions, i.e., visually similar frames with their provided captions from the training dataset and then most appropriate and suitable caption is selected from the candidate cations. Although retrieval based captions are grammatically correct, but frame/video specific caption generation is very challenging. The template based approaches have fixed templates with a number of blank slots for generated caption's subject verb and object. These methods are also capable of generating grammatically correct captions but can not generate variable length captions because of the limitation of their dependence on fixed, predefined templates, which are not capable of generating semantically rich natural language sentences, and hence, are not analogous to human annotations.

1.2 Video captioning

The deep learning models employed for video description tasks primarily follow the Encoder–Decoder structure, which is the most productive/beneficial sequence-to-sequence modeling technique. Describing a video can also be defined as a sequence-to-sequence task, since it has a sequence of visual representations as input, and a sequence of generated words as output. The ED architecture gained considerable attention in the earlier research specific to neural machine translations, where the implementation was for text translations from one language domain to another. The task of describing videos can be partitioned into two major sections: the visual model for understanding visual content correctly (without missing any information), and the language model for transforming learned visual information into grammatically correct natural language sentences. Since computers only understand numbers, arrays, and matrices, so learned visual representations are stored as *context vector*. The context vector is a collection of numbers communicating some visual information into the language model. The language model then extracts the connotation of each of the context vectors, and accordingly generates semantically aligned words, one by one. Represented mathematically, we can say that the language model is employed to establish the probability P of generating a word w at time t conditioned on previously generated words $w1, ..., wt-1$, during the preceding time steps (where $1, 2, ...t-1, t$ represents time step), i.e., $P(wt \| w1, ..., wt-1)$ where w_i represents word generated at a certain time.

Figure 2 demonstrates the deep learning-based basic model employing visual and language models for video description. Following the ED architecture for video descriptions, the standard ED structure employs a combination of the convolutional neural network, the recurrent neural network , or the variants LSTM or GRU as encoder and decoder blocks. RNNs for sequential data processing have demonstrated comparable results; but for long sequences, the implementation of the RNN system is not appreciable. The associated vanishing and exploding gradient problems, as well as the recurrent nature involving previous-step computations in the next step, hinders the parallel processing of the sequence, hence degrading overall performance. In order to upgrade the performance of the standard ED architecture, it can be equipped with an attention mechanism, reinforcement learning, or a transformer mechanism. Attention mechanisms focus on specific areas of the frame, and achieve high-quality results. RL employed within the ED architecture can progressively deliver state-of-the-art captions, employing its own agent-environment interactions. The

transformer mechanism is an efficient architecture for robust output. It does not contain any convolution and recurrence, and is developed solely on the basis of self-attention. The transformer allows parallelization along with training on a massive amount of data, with the capability to fully utilize the available GPUs for most machine learning tasks. Drastically reduced training time and efficient model training can take place with high accuracy by using the parallel processing capability of transformers. Recently with the emergence of several versions of transformers, long-term dependency handling is not an issue anymore.

1.3 Dense video captioning/ video description

Comprehending the localized events of a video appropriately and then transforming them accurately into a textual format is called dense video captioning, or simply, video description. This task of describing complex and diverse visual perceptions establishes a connection between the two world-leading realms of computer vision and natural language processing. Capturing the scenes, objects, and activities in a video, as well as the spatial-temporal relationships and the temporal order, is crucial for precise and grammatically correct multi-line text narration.

Nevertheless, the task of automatically describing video is challenging. The model employed for the generation of a caption characterizing a long-duration video or a short clip consisting of a significant number of frames requires not only an understanding of sequential visual data but also the capability to provide a syntactically and semantically accurate translation of that understanding into natural language. Similarly, the proper understanding of a considerable number of objects, events, and actions, and their interactions in the video (as well as their relationships and the order in which they happen) must be captured accurately and explained properly using natural sentences. Whether they belong to an open or a constrained domain, videos mostly contain numerous scenes or events. The dependencies between the events are captured by using contextual information from the previous (past) and coming (future) events, and then all events are jointly described accordingly by using natural language. Analogous to dense image captioning which describes regions in space after localization. Similarly, with the help of transformer, 2D images are transformed into 3D objects with color and texture-aware information by Yuan et al. (2022), for dense captioning. Dense video captioning (Krishna et al. 2017) localizes events in time, and afterwards expresses them. These events can intersect with other events, and hence, are challenging to describe appropriately. Dense video captions capture details of event localization and their co-occurrence (Aafaq et al. 2022).

Terminologies associated with video description have their specific implications. Keeping current research in mind, the task of video captioning can be distributed into two sections: mono-sentence caption generation and multi-sentence (paragraph) caption generation. The mono-sentence is supposed to be a precise, yet fully informative, abstractive representative sentence of the whole video, whereas a multi-sentence (dense) caption is supposed to localize and describe all events in the video temporally, including intersecting and overlapping events. Here, event localization refers to identification of each event in the video with its start and end times; event description means expressing each localized event temporally in a much more detailed way, resulting in the generation of multiple sentences or paragraphs (like a dense summary of the whole video). The generated fine-grained caption is a requirement of such a mechanism that proves to be expressive and subtle. Its purpose is to capture the temporal dynamics of the visuals present in the video, and to then join that with syntactically and semantically correct representations using natural language.

Problem setup: video captioning/description

1. For video captioning (single-sentence): Let us suppose we have a video, V, containing N frames such that $V = \{f_1, f_2, ...f_N\}$ (f representing frame), and our aim is to generate a single-sentence textual caption, T, representing the video content comprising n words such that $T = \{w_1, w_2, ...w_n\}$ (w representing word), and semantically aligned words, one by one, are generated conditioned on previously generated words. At time t, the word wt is generated conditional on probability $P(wt \| w1, ..., wt - 1)$ where w_i represents a word generated at a certain time, i.
2. For video description (dense captioning): Particular to videos containing multiple scenes or events, event localization (Krishna et al. 2017) is the identification of start and end times of a particular event in the video. Comprehending these localized events semantically, and transforming them into precise and grammatically correct multi-sentence natural language explanations, is required. For video V containing N events such that $V = \{E_1, E_2, ...E_N\}$ (E representing event), each event needs to be identified such that $E_1 = \{EST, w_1, w_2, ...w_A, EET\}$ with event start time (EST) and event end time(EET). A certain number of words, A , expresses event E_1, and similarly, localized $E_2 = \{EST, w_1, w_2, ...w_B, EET\}$ has a certain number of words, B, to express event E_2 and so on, until all events in the video are understood. Every event can be expressed with a different number of words (A, B, ...) depending on the duration of the event. The aim is to gather all localized event descriptions and generate a semantically and grammatically correct and coherent paragraph-like description for the video, avoiding redundancy.

This survey aims to present inclusive insights into the deep learning-based techniques implemented in the video description, supported by the most recent research. During the past few years, the field of captioning (image/video) has exhibited remarkable success and has achieved amazing state-of-the-arts. A thorough discussion on these techniques/methodologies adopted from time to time lacks in the current literature. The key motivation behind this research work is to fill this research gap and facilitate the researchers in a clear understanding of the employed approaches. Our *contributions* to this research are as follows.

1. We provide an elaborate view of the latest deep learning-based techniques for video description, with up-to-date supporting articles from the literature.
2. Besides the standard ED architecture, a detailed exploration of deep RL, attention mechanisms, and transformer mechanisms for video descriptions is performed.
3. We categorize and compare the key components of the models, and the substantially crucial information is highlighted for in-depth insights and quick understanding, making it expedient for researchers who are involved in video processing for summarization and description, or for autonomous-vehicle, surveillance, and instructional purposes, to find the state of the art in a single go.
4. Finally, we identify future research directions for further improvement in video description systems.

Outline of the survey This paper is organized as shown in Figure 1. The next section, Section2, offers a brief discussion on the available surveys on the topic. These surveys primarily focus on the simple Encoder–Decoder based models. Section 3 demonstrates detailed deep learning based techniques employed for video description. At first, standard encoder decoder architecture employing CNN-RNN, RNN-RNN and CNN-CNN compositions

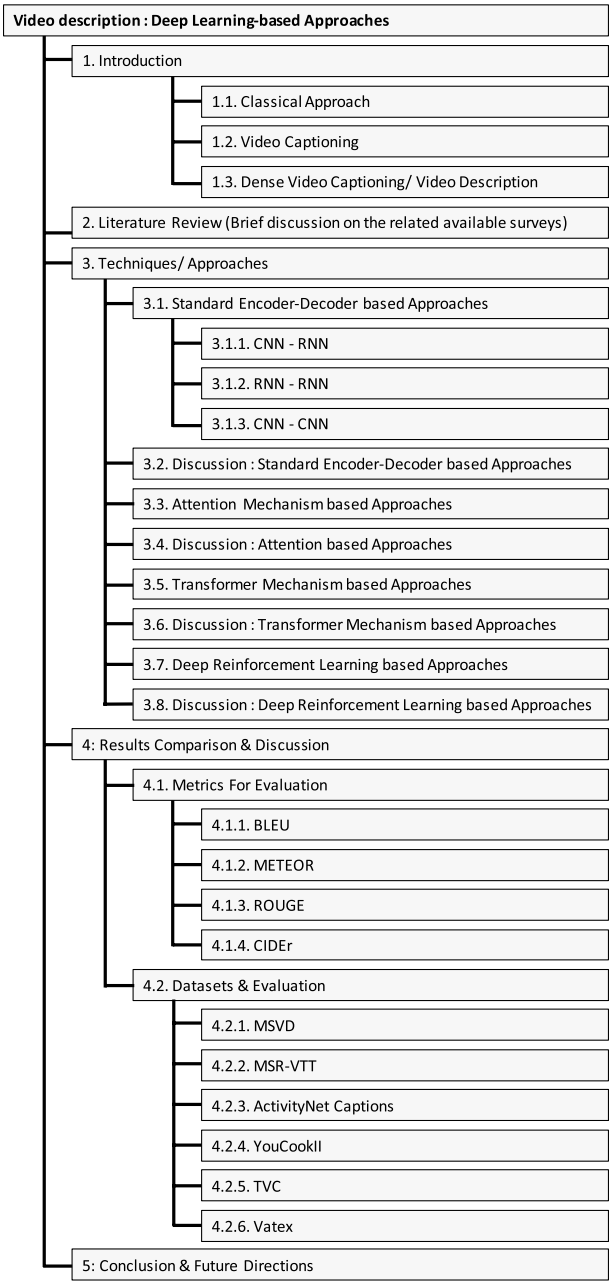

Fig. 1 Hierarchical structure of this paper

followed by a thorough discussion is explore. Secondly, we describe fusion of attention mechanism in encoder decoder system for video captioning models to focus on specific distinctiveness. Thirdly, we present transformer based recent state of the art methods and analyze them for video description generation. Finally, the successful strategies for optimizing

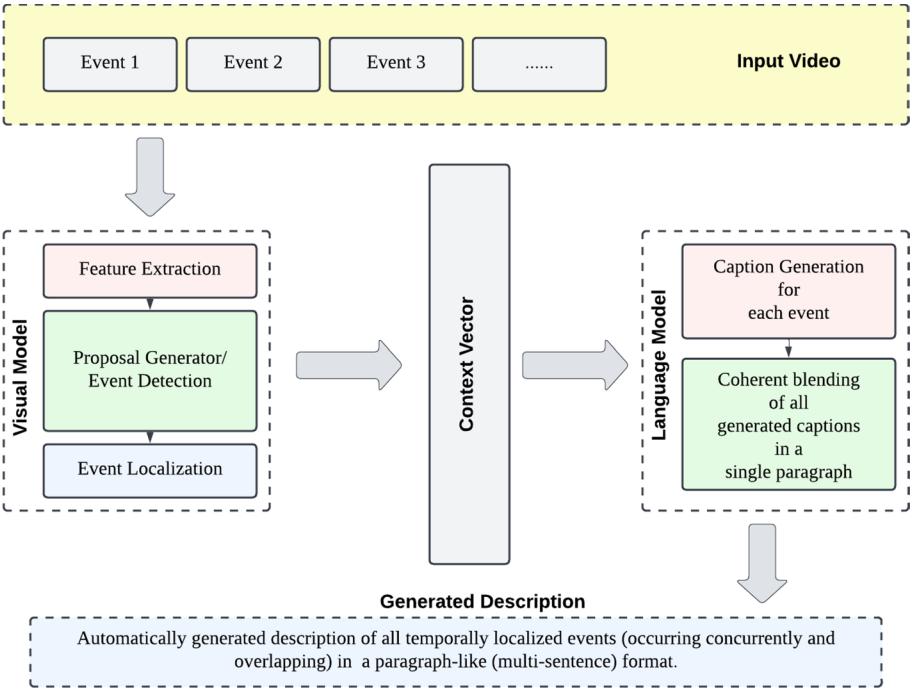

Fig. 2 The basic model for video description (dense video captioning) examines a long video comprising multiple scenes or events, i.e., Event 1, Event 2, up until the last identified event. After localization (identification of start and end times) of each event, the paragraph-like (multi-sentence) description is generated by coherently combining the captions generated for each event, catering to concurrent and overlapping events

the generated descriptions through deep reinforcement learning is discuss in detail. Discussion on limitations and challenges of every technique's is also present with working strategy, computational concept, and literature review in their respective sub-sections. In Section 4, we analyze and compared the benchmark results produced by the state of the art methods, segregated chronologically based on specific dataset. A brief overview of the evaluation metrics and datasets used for video description is also provided in this section. Finally, Section 5 concludes the review with few future directions.

2 Literature review

Computer vision mainly deals with classification, detection, and segmentation tasks (Rafiq et al. 2020; Agyeman et al. 2021). The first part of video captioning, i.e., temporal action recognition, solely belongs under computer vision, whereas the second part (caption generation) bridges computer vision and natural language processing. Captioning is once again split into two types–one for the simple-to-recognize and -describe actions, and the second, for actions too complex to be described with simple and short natural language sentences.

The selection of appropriate components plays a substantial role in the generation of accurate and truthful output. A thorough empirical analysis of each component in the ED

framework was presented in Aafaq et al. (2019b). Significant performance gains were demonstrated by careful selection of an efficient and capable mechanism for four major constituent components: feature extraction, feature transformation, word embedding, and language modeling. The authors emphasized which efficient mechanisms can be adopted for these four factors and how to generate state-of-the-art results. For feature extraction, five different CNN models (3D-CNN, VGG-16, VGG-19, Inception-v3, and Inception-ResNet-v2) were analyzed, and the authors concluded that C3D is a common choice because of its ability to process both individual frames and short video clips. For 2D-CNN models, Inception-ResNet-v2 performed best in feature transformation. For mean pooling, and temporal encoding, temporal encoding was favored since mean pooling will result in considerable loss of information. In contrast, temporal encoding can capture highly reliable temporal dynamics of the whole video without any noteworthy loss of information, creating a positive counterbalance for system performance. In the literature, two methods are commonly referred for word embedding. The first is randomly initializing the embedding vector, and then computing the task-specific embedding, which is not able to capture rich semantics, whereas the second method makes use of pre-trained embedding. The authors examined four pre-trained embeddings–Word2Vec, FastText, and Glove (glove6B, glove 840b)–as well as randomly initialized embedding. FastText, with operative word embedding, performed prominently. Finally, in language modeling, the depth (or number of layers) in the system is crucial for superior performance, along with various hyperparameters, e.g., internal state size, the number of processed frames, fine-tuned word embedding, and dropout regularization.

This research work features deep learning-based frameworks for video description–ED in particular. It is clear from Table 2 that all available surveys on video description were primarily focused on simple ED-based frameworks. Several among them, notably (Li et al. 2019a; Chen et al. 2019b; Aafaq et al. 2019c; Amaresh and Chitrakala 2019; Su 2018) and (Wu 2017), briefly discussed the application of an attention mechanism, and (Li et al. 2019a; Chen et al. 2019b; Aafaq et al. 2019c) just gave an overview of reinforcement learning within the encoder-decoder, but none of them elaborated on these architectures, on

Table 2 A literature review of video captioning/description

References	Deep Learning-based Techniques			
	Std ED	AM	RL	TM
This Research	✓	✓	✓	✓
Aafaq et al. (2019b)	✓	✗	✗	✗
Wang et al. (2020)	✗	✓	✗	✗
Li et al. (2019a)	✓	▼	▼	✗
Chen et al. (2019b)	✓	▼	▼	✗
Aafaq et al. (2019c)	✓	▼	▼	✗
Amaresh and Chitrakala (2019)	▼	▼	✗	✗
Su (2018)	✓	▼	✗	✗
Park et al. (2018)	▼	✗	✗	✗
Wu (2017)	✓	▼	✗	✗

✓: Discussed in detail in the cited work, ✗: Not discussed in the cited work, ▼: Understated in the cited work; AM: Attention Mechanism, RL: Reinforcement Learning, Std ED :Standard Encoder-Decoder, TM: Transformer Mechanism

related articles from the literature in detail, or explored transformer mechanism employment for video captioning. In this survey, all four approaches are described in detail, with state-of-the-art articles proving their worth.

To take full advantage of the advanced state-of-the-art hardware, i.e., GPUs, it is essential to adopt the models/mechanisms that can fully exploit these hardware structures. The sequential nature of RNNs cannot utilize the parallelization found in GPUs, resulting in inferior performance and slow training. As an alternate option to recurrence and convolution, an efficient approach is proposed by the transformer. It is capable of parallel processing, accelerated training, and handling long-term dependencies, and is space-efficient, much faster, solely self-attention-based, and is the model of choice for current advanced hardware.

3 Techniques/approaches

Inspired by technological advancements, researchers have experimented with deep neural networks for the automatic caption generation task. The early frameworks comprised the standard ED structure, but with their methodical rise, new high-tech approaches are fused with the standard structure to produce more expressive and flexible natural language sentences with richer semantics. In this paper, we have classified the adopted techniques into four categories per their technological advancement in time–the standard ED approach, the fusion of attention mechanisms in the standard ED structure, and adoption of the transformer mechanism for robust performance, and the decision-based DRL approaches, which have prominence in accurate natural language caption generation and optimization. The arrangement of these approaches/techniques is based on their technological evolution over time. We discuss these techniques one by one in detail in this section.

3.1 Standard encoder–decoder approaches

The ED approach is a neural network configuration, as shown in Figure 3. The architecture is partitioned into two components, namely, the encoder and the decoder. It has proven to be cutting-edge technology. The modern approach has been employed by the research community around the globe to solve sophisticated tasks, i.e., image captioning, video description, text and video summarization (Rafiq et al. 2020), visual question-answering systems and conversational modeling, and movement classification.

The ED framework comprises two neural networks (NNs):

$$\varphi(F) = R \tag{1}$$

where

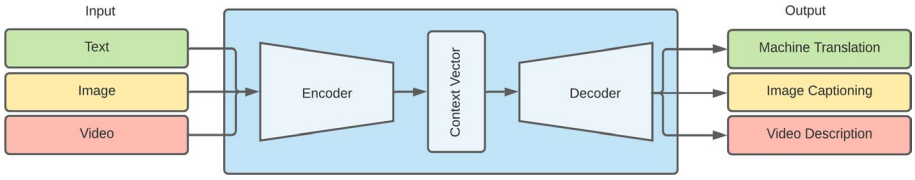

Fig. 3 The standard Encoder–Decoder architecture

$$R = \{r_1, r_2, ..., r_n\} \tag{2}$$

The vector R in (2) is an internal representation that captures the context and meaning of the input, and is known as a *context vector* or *thought vector*. The choice of encoder structure mainly depends on the type of *input*, e.g., for text, the best encoder architecture is the RNN. For an image/frame or video clips as input, the CNN structure proved to be best suited for context-vector/visual-feature extraction (Xu et al. 2015). However, deliberation regarding CNN or RNN selection (Yin et al. 2017), and their behavioral differences for NLP, is under way among researchers. The fusion of these two architectures has accomplished outstanding results, since they both process information through altered techniques, and complement one another.

Context vector R generated from the encoder is input to the second neural network in the system, i.e., the *decoder*. The decoder generates the corresponding output. Selection of the decoder architecture depends on the type of *output*. In the video description task, when meaningful textual information is required as output from the input video, the RNN is the architecture mostly employed for this purpose. RNN variants like long short-term memory and the gated recurrent unit are popular in research involving natural language processing because of their ability to handle long-term dependencies. Decoder RNN θ functionality at any given time t is

$$\begin{bmatrix} O_t \\ h_t \end{bmatrix} = \theta\left(h_{t-1}, O_{t-1}, R\right) \tag{3}$$

where O_t represents output at time t, and h_t is the internal/hidden state of the RNN, whereas $h_{(t-1)}$ and $O_{(t-1)}$ represent the hidden state and the output of the previous time step, $(t-1)$. The RNN repeatedly works until the end-of-sequence $\langle EOS \rangle$ token is generated. LSTM and GRUs with improved performance replace the basic RNN structure.

Specific to the video description, the encoder can be treated as the visual model for the system, whereas the decoder is responsible for language modeling. Two-dimensional (2D) or 3D convolutional neural networks are mostly used as an encoder for computing the context vector of a fixed or variable length. The context vector can be called a *vector representation* or a *visual feature*. After extraction, certain transformations are applied to these visual features, i.e., *Mean/Max pooling* or *temporal encoding*. The resultant transformed visual features are then entered into the language model for description generation. The ED framework is the most popular paradigm for video description tasks in recent years, so the authors in Aafaq et al. (2019b) partitioned the ED structure for video description into four essential components: a CNN model for visual feature extraction, the types of transformations applied to extracted visual features, and the language model and the word embedding within the language model. Since the involvement of each of these components in the performance of the system is of high importance, intelligent selection is essential. By keeping in mind the pros and cons of each selected component, one can straight forwardly determine the overall performance of the description system. Blohm et al. (2018) explored the behavioral variance between CNNs and RNNs using a MovieQA dataset with 11 models trained for different random initializations of both an RNN-LSTM and a CNN, and finally, they observed that RNN-LSTM models outperformed CNN models by a large margin, although both models share the same weaknesses. Considering limitations or weaknesses, they test the transferability of the adversarial examples across models to investigate the CNN models on the adversarial examples optimization to fool the RNN models and vice versa.

Degradation in performance was observed for both CNNs and RNNs and was fixed by including some adversarial examples in the training data.

Three compositions of encoder and decoder for video description available in the literature (CNN-RNN, RNN-RNN, and CNN-CNN, are summarized in Table 3 for convenience along with their visual & language components, contributions, and shortcomings (if any). Figure 4 shows these compositions as percentages. In recent research works, the transformers are also exploited as visual or language component of the ED structure, Seo et al. (2022) employed ViViT (video vision transformer) Arnab et al. (2021) & BERT as encoder and GPT-2 based decoder. Zhao et al. (2022) used an encoder composed of transformer encoder blocks for video features extraction in a global view resulting in reduced loss of intermediate hidden layer information.

3.1.1 CNN–RNN

The conventional ED pipeline typically comprises a CNN as a visual model for extracting visual features from each frame of the video, employing an RNN as a language model for generating the captions, word by word. VSJM-Net (Aafaq et al. 2022) presented a visual and semantic joint embedding network which is employed to detect proposals as well as learn the visual and semantic space. vc-HRNAT (Gao et al. 2022) using hierarchical representations is capable to learn in a self-supervised environment with multi-level semantic representation learning of video concepts. However, the system lacks the ability to visualize concepts of objects and actions that are absent or unclear in videos. VNS-GRU (Chen et al. 2020), a semantic GRU model with variational dropout and layer normalization, is trained using professional learning. For feature generation, the system utilizes ResNetXT-101 pre-trained on ImageNet (Deng et al. 2009) at the frame level, and an efficient convolutional Network (ECN) (Zolfaghari et al. 2018) pre-trained on Kinetics-400 at the video level. The model can learn unique words and delicate grammar based on vocabulary and tagging mechanisms. Similarly, a system comprising 2D and 3D ConvNets with a semantic detection network (SDN) as the encoder and a semantic-assisted LSTM as the decoder was proposed in (Chen et al. 2019a) to overcome the limitations of short and inappropriate descriptions, deprived training approaches, and the non-availability of critical semantic features. Static spatial as well as dynamic spatio-temporal features are involved, along with a scheduled sampling strategy for self-learning of long sentences. A proposal for sentence-length modulated loss reassures optimization as well as thorough and detailed captions.

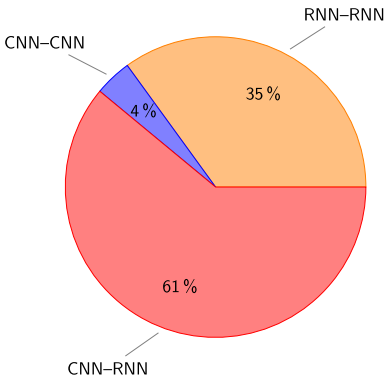

Fig. 4 Composition of the standard ED architecture for video description based on the literature explored for this research work

Table 3 Standard encoder–decoder based models for video captioning with visual and language components, short contributions and shortcomings (if any) are also mentioned with each approach

S/N	References	Year	Approach	Model	
				Visual	Language
A. CNN-RNN					
1	Gao et al. (2022)	2022	vc-HRNAT (Video Captioning - Hierarchical Representation Network with Auxiliary Tasks)	CNN	LSTM
Contributions: An end-to-end framework utilizing hierarchical representation learning and auxiliary tasks in a self-supervised manner. The framework is capable of learning multi-level semantic representation of video concepts.					
Shortcomings: Visulaization of absent or ambiguous concepts of objects and actions in videos.					
2	Seo et al. (2022)	2022	MV-GPT (Multimodal Video Generative Pretraining)	ViViT Arnab et al. (2021) based visual encoder, BERT based text encoder	modified GPT-2 based decoder.
Contributions: Jointly trainable Encoder–Decoder model where no more manually annotated captions are required. Instead utterances at different time steps of the same video can be utilized. The encoder is trained directly from the pixels and words.					
Shortcomings: Considering pre-training, system suffers from performance degradation for inputs in a different domain.					
3	Aafaq et al. (2022)	2022	VSJM-Net (Visual-Semantic Joint Embedding Network)	2D CNN	Vanilla Transformer
Contributions: In the proposed Visual-Semantic Joint Embedding Network, the Visual-Semantic Embedding (ViSE) jointly learns the visual and semantic space while detecting proposals. Video Level Sequence Encoder (VLSE) detects event boundaries across frames in given video. The ViSE embedding are transformed into descriptor vectors with a Hierarchical Descriptor Transformer (HDT). The transformed features are used in proposal generation network along with Linguistic information.					
Shortcomings: NA					
4	Madake (2022)	2022	Dense Video Captioning	EfficientNetB7	bi-LSTM + LSTM
Contributions: Event detection and using information from future and past contexts in the video. EfficientNetB7 neural network is used for visual features extraction. Bi-LSTM and LSTM is employed for caption generation.					
Shortcomings: As length of video increases from a few seconds to several minutes the BLEU and METEOR scores decrease as it is hard for LSTM to learn long term dependencies.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
5	Hammoudeh et al. (2022)	2022	Soccer Captioning	ConvNet (CNN-img, CNN-flow, and CNN-vae)	transformer
Contributions: A dataset of 22k video-caption pairs along with features extracted from 500 hours of SoccerNet videos and model employing semantic related losses while captioning soccer actions.					
Shortcomings: NA					
6	Zhao et al. (2022)	2022	Transformer-LSTM-RL	ResNet-152, ResNeXt-101, ViT	LSTM
Contributions: Used an encoder composed of Transformer Encoder blocks to encode video features in a global view, thereby reducing the loss of intermediate hidden layer information. Further, Introduced the Policy Gradient reinforcement learning method to improve the accuracy of the model.					
Shortcomings: In the video captioning task, the collection and labeling of training data often consumes a lot of manpower and material resources.					
7	Perez-Martin et al. (2021a)	2021	Attentive Visual Semantics Specialized Network (AVSSN)	2D/3D CNN	LSTM
Contributions: Proposed specialized visual and semantic LSTM layers along with Adaptive Attention Gate for integrating different temporal representations into decoder. The model is capable of accurately capturing visual and semantic context representations.					
Shortcomings: Model is unable to capture the multiple events.					
8	Zheng et al. (2020)	2020	Syntax-Aware Action Targeting (SAAT)	Extractor encoder (Cxe)	Extractor decoder (Cxd) LSTM
Contributions: An action guided captioner models relationship among video objects and dynamically fuse information from the predicate and previously generated words.					
Shortcomings: Global temporal information captured by 3D CNN is not enough to learn finer actions.					
9	Chen et al. (2020)	2020	VNS-GRU (Decoder with variational dropout and layer normalization, professional learning strategy)	CNN (ResNetXt-101)	Semantic GRU with variational dropout and layer normalization
Contributions: Variational dropout and layer normalization is combined in the decoder to prevent overfitting and sustain convergence speed. Professional learning training strategy is adopted for efficient model training.					
Shortcomings: Professional learning training strategy used to train model by optimizing losses needs further experimentation.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
10	Hou et al. (2019)	2019	Joint Syntax Representation Learning and Visual Cue Translation (JSRL-VCT)	CNN (C3D, ResNet, Inception)	POS Tag Generator
Contributions: An end-to-end trainable network capable of capturing the syntactic structure of sentences via video POS tagging and perceive intrinsic semantic primitives. Word bias problem caused by imbalanced classes is also addressed.					
Shortcomings: Evaluation on B@4 is not remarkable due to B@4 lexical basis rather than syntactic matching.					
11	Chen et al. (2019a)	2019	Semantic detection network (SDN), Semantic Compositional Network (SCN) & SDN trained with scheduled sampling	2D-CNN for static features, 3D-CNN for spatio-temporal features	Semantic Compositional Network (SCN), a variant of LSTM
Contributions: A semantic-assisted captioning model with scheduled sampling to bridge the gap between training and testing in the Teacher Forcing algorithm. Sentence-length-modulated loss function is also proposed to keep the model in a balance between language redundancy and conciseness. The proposed SDN (Semantic Detection Network) extracts high-quality semantic features for video.					
Shortcomings: NA					
12	Aafaq et al. (2019a)	2019	GRU–Enriched Visual Encoding (EVE) with hierarchical Fourier transform	2D/3D CNN	2-layered GRU
Contributions: Visual encoding technique that effectively encapsulates spatio-temporal dynamics of the videos and embeds relevant high-level semantic attributes in the visual codes for video captioning. Use of hierarchical Fourier Transform to capture the temporal dynamics of videos.					
Shortcomings: NA					
13	Olivastri (2019)	2019	End-to-end network-Inception ResNet V2 (EtENet-IRv2)	CNN (Inception ResNet V2, GoogLeNet)	LSTM with soft attention (SA-LSTM)
Contributions: End-to-End trainable framework designed to learn task-specific features based on two staged training strategy. At first stage, freeze the weights of the pre-trained encoder to train the decoder resulting in low memory requirement and fast process execution. At the second stage, the whole network is trained end-to-end while freezing the batch normalisation layer.					
Shortcomings: Evaluation on BLEU is not remarkable as BLEU score's lack of explicit word matching between translation and reference. Training a deep neural network end-to-end requires significant computational resources.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
14	Zhang et al. (2019a)	2019	Object-aware aggregation with bidirectional temporal graph (OA-BTG)	Convolutional gated recurrent unit (C-GRU)	GRU with attention
Contributions: video captioning approach based on object- aware aggregation with bidirectional temporal graph (OA- BTG), which captures detailed temporal dynamics for the salient objects in video via a bidirectional temporal graph, and learns discriminative spatio-temporal video representations by performing object-aware local feature aggregation on object regions.					
Shortcomings: Modeling salient objects with their trajectories along with interaction and relationship among objects is required for accurate descriptions generation of actions.					
15	Liu et al. (2020)	2018	SibNet (sibling convolutional encoder for video captioning)	CNN (content and semantic branches)	RNN
Contributions: A dual branch achitecture composed of visual branch and semantic branch. Content branch to encode salient visual content information and the semantic branch to encode high-level semantic information with the guidance of ground truth captions brought by visual-semantic joint embedding. The content branch, the semantic branch and the decoder are trained jointly by minimizing the proposed loss function. TCB (Temporal Convolution Block) is proposed providing more efficient video temporal encoding than RNN with less number of parameters.					
Shortcomings: NA					
16	Lee and Kim (2018)	2018	SeFLA (Semantic feature learning and attention-based caption generation)	2D/3D CNN for visual features extraction, LSTM for semantic features extraction	LSTM
Contributions: Semantic Feature Learning and Attention-Based Caption Generation for effective video captioning by utilizing both visual and semantic (dynamic and static) features.					
Shortcomings: Relatively inefficient when predicting consecutive words demonstrating the model's ineffectiveness in generating prepositional and postpositional particles. Low BLEU@2, BLEU@3, and BLEU@4 scores.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
17	Wang et al. (2018a)	2018	Reconstruction Network (RecNet)	CNN (Inception-V4)	LSTM with temporal attention
Contributions: RecNet with the encoder-decoder-reconstructor architecture for video captioning, which exploits the bidirectional cues (video to sentence, i.e., forward and sentence to video, i.e., backward) between natural language description and video content. Video global and local structures are restored byncustomized reconstructor networks. The forward likelihood and backward reconstruction losses are jointly modeled to train the proposed network.					
Shortcomings: RecNet-global under-performed as compared to RecNet-local due to the temporal dynamic modeling and employment of mean pooling for video representation reproduction. This simple temporal attention mechanism cannot capture the internal relationships of key information. Ji et al. (2022)					
18	Pan et al. (2017)	2017	Long Short-Term Memory with Transferred Semantic Attributes(LSTM-TSA)	2D/3D CNN	LSTM with high-level semantic attributes
Contributions: Proposal of LSTM-TSA for addressing the issue of exploiting the mutual relationship between video representations and attributes for boosting video captioning. A transfer unit is designed to dynamically control the impacts of semantic attributes from the two sources (images and videos) on sentence generation.					
Shortcomings: NA					
19	Shen et al. (2017)	2017	Lexical-FCN (lexical fully convolutional neural network)	CNN (Lexical-FCN model)	LSTM
Contributions: Dense video captioning by weakly supervised learning utilizing only video-level sentence annotations. Proposed approach modeled visual cues with Lexical-FCN, discovering region-sequence with submodular maximization, and decodes language outputs with sequence-to-sequence learning.					
Shortcomings: For result comparison with oracle, need to strengthen the evaluator network. Diversity score is slightly worse than the best of the clustered ground-truth sentences.					
20	Zhang et al. (2017)	2017	Task-driven data fusion (TDDF)	2D/3D CNN (VGG-19, GoogLeNet, C3D)	TDDF-based LSTM
Contributions: To reduce ambiguity in video description, the proposed system adaptively choose different fusion patterns according to the task status. The dynamic fusion model can attend to certain visual cues that are most relevant to the current word. Appearance- centric, motion-centric and correlation-centric fusion patterns are designed to support the recognition of visual entities.					
Shortcomings: The system failed to describe animation films due to different description context information during training and testing.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
21	Lowell et al. (2014)	2015	Translating videos into natural language using deep recurrent neural networks	CNN (Caffe)	LSTM with transfer learning
Contributions: End-to-end deep model for video-to-text generation that simultaneously learns a latent meaning state, and a fluent grammatical model of the associated language.					
Shortcomings: Model trained and evaluated on random frames from the video, and not necessarily a key-frame or most-representative frame. Moreover, training on images alone do not directly perform well on video frames, and a better representation is required to learn from videos.					
22	Rivera-soto and Ordóñez (2013)	2013	Sequence-to-sequence models for generating video captions	ResNet-50 Wang et al. (2018c), VGG16, LSTM	LSTM
Contributions: In sequence to sequence model pre-trained convolution network extract visual features from video frames and fed to LSTM encoder for encoding. LSTM decoder is employed to generate natural language description.					
Shortcomings: The responsibility of both encoding the input features and decoding the natural language description in one set of weights complicates the convergence of the network.					
23	Yan et al. (2010)	2010	Crowd Video Captioning (CVC)	2D/3D CNN (Inception, ResNet, C3D)	LSTM/GRU (S2VT)
Contributions: The proposed model aim to generate captions for the crowd video, i.e., describing the off-site audiences or visitors crowd. Created a dataset based on World-Expo'10.					
Shortcomings: Small dataset with simple captions. The number of videos and the complexity of captions needs to be increased in the dataset.					
B. RNN-RNN					
24	Zhang et al. (2021)	2021	RCG (Retrieve-Copy-Generate)	Bi-directional LSTM	att-LSTM + Lang LSTM
Contributions: End-to-end trainable Retrieve-Copy-Generate network, where a pluggable video-to-text retriever is constructed to retrieve sentences as hints from the training corpus effectively, and a copy-mechanism generator is introduced to extract expressions from multi-retrieved sentences dynamically.					
Shortcomings: NA					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
25	Xiao and Shi (2019z)	2019	Diverse Captioning Model (DCM), Conditional GAN (a CNN as a generator/ discriminator)	Bi-directional LSTM	Stacked LSTM
Contributions: An efficient model for generating accurate descriptions with the aim to be consistent with human behaviour. A conditional GAN is proposed to explore the diversity of the descriptions. Diverse Captioning Evaluation (DCE) is also proposed to evaluate not only the differences among captions but also consider the rationality of the generated descriptions.					
Shortcomings: Better accuracy assessment method required to decrease the gap between DCM and Ground Truth under DCE evaluation.					
26	Babariya and Tamaki (2020)	2019	Object Attention, Meaning (OAM)-guided LSTM ED model + metric learning	LSTM	LSTM with attention
Contributions: the proposed approach can describe objects detected by object detection, and generate captions having similar meaning with correct captions.					
Shortcomings: Object detector (mistaken identification of objects) directly affect the encoder.					
27	Wang et al. (2019a)	2019	GFN-POS (Controllable video captioning with POS sequence guidance based on a gated fusion network)	LSTM temporal encoder	2-layer LSTM
Contributions: A gated fusion network incorporating multiple features information together and a POS sequence generator predicting the global syntactic POS information of the generated sentence. Also proposed across gating (CG) strategy to effectively encode and fuse different representations. The global syntactic POS information is adaptively and dynamically incorporated into the decoder to guide the decoder to produce more accurate description in terms of both syntax and semantics.					
Shortcomings: Under performed on BLEU@4 score due to mainly focusing on optimizing the CIDEr metric with reinforcement learning.					
28	Hammad et al. (2019)	2019	Effects and interaction of multi-modal features, seq-to-seq video description	Stacked LSTM	Stacked LSTM with attention
Contributions: Model is based on S2VT(Sequence-Video to Text) focusing on the characterization of the impact of features utilization from pre-trained model to implement video captioning. 2D object recognition features, scene recognition features, 3D action recognition features, audio features, and object recognition Intermediate Features are employed for abstract information about the different objects in the frame and their relations.					
Shortcomings: Contrasting concatenation techniques for size reduction of multi-modal input data along with the model's increased capacity by adding more LSTM nodes and better regularization techniques should be investigated.					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
29	Zhao et al. (2018)	2018	Tube features (Faster-RCNN) object detection + feature extraction + LSTM-based ED with attention	Bi-directional LSTM	LSTM with attention
Contributions: Video caption generator conditioned on tube features. Where tubes are formed by the object trajectories. Each object tube is constructed by the faster-RCNN based detected objects and their corresponding regions in different frames. The edge between each pair of bounding boxes in adjacent frames are labeled with a similarity score. Bidirectional LSTM captures the dynamic information by encoding each tube.					
Shortcomings: Restricted performance due to the visual input to LSTM, that is just the average pooling of frame features.					
30	Donahue et al. (2017)	2017	Long-term Recurrent Convolutional Network (LRCN)	LSTM	LSTM with CRF (max or probabilities)
Contributions: An end-to-end LRCN (Long-term Recurrent Convolutional Networks), a class of recurrent-convolution architectures for visual recognition and description which combines convolutional layers and long-range temporal recursion. The proposed model is specifically for video activity recognition, image caption generation, and video description tasks. The recurrent convolutional models are two times deep in a way that they learn compositional representations in space and time.					
Shortcomings: NA					
31	Wang and Song (2017)	2017	S2VTK (S2VT with knowledge)	LSTM	LSTM
Contributions: video captioning approach aiming at knowledge base information fusion with frame features of the video. LSTM based caption generator is trained by maximizing the probability of correct caption given a video.					
Shortcomings: Only BLEU and METEOR scores are reported. The proposed model is not evaluated for CIDEr and ROUGE scores.					
32	Venugopalan et al. (2015)	2015	S2VT (end-to-end sequence-to-sequence stacked LSTM)	Stacked LSTM	Stacked LSTM
Contributions: A pioneer sequence to sequence model in video captioning. The proposed model learns to map sequence of frames to a sequence of words directly. The optical flow is computed to model the temporal aspects of the events in the video.					
Shortcomings: Model evaluated only for METEOR score (single evaluation metric can not guarantee the algorithm's superiority).					

Table 3 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
33	Cho et al. (2014)	2014	RNN ED for machine translation from English to French	RNN (translation model)	RNN
Contributions: The proposed RNN Encoder–Decoder framework with a hidden unit that adaptively remembers and forgets, is evaluated on the task of NMT (Neural Machine Translation) from english to french.					
Shortcomings: Enhanced performance can be achieved by using neural net language model.					
C. CNN-CNN					
34	Chen et al. (2019b)	2019	TDConvED (CNN for both encoding and decoding with temporal attention)	CNN (VGGNet, ResNet)	CNN (VGGNet, C3D, ResNet)
Contributions: The system contributed by exploiting a fully convolutional sequence learning architecture that relied on CNN-based encoder and decoder for video captioning. Moreover, it explored the temporal deformable convolutions and temporal attention mechanism to extend and utilize temporal dynamics across frames/clips.					
Shortcomings: NA					

Three CNN, RNN compositions are presented, i.e., CNN-RNN, RNN-RNN, CNN-CNN

NA Not available, *S/N* serial number)

In order to enhance the visual encoding mechanism for captioning purposes, GRU-EVE (Aafaq et al. 2019a) was the first to emphasize feature encoding for semantically robust descriptions using a 2D/3D CNN with short Fourier transform as a visual model and a two-layered GRU as a language model for capturing spatio-temporal video dynamics. A 2D-CNN (InceptionResNetv2 Szegedy et al. 2017) pre-trained on an ImageNet dataset, and a 3D-CNN (C3D Tran et al. 2015) pre-trained on a Sports 1M dataset (Karpathy et al. 2014) are used for feature extraction. Then, the extracted features are processed hierarchically with short Fourier transformation, and the visual features are semantically improved. The approach proved that application of short Fourier transformation on a 2D-CNN produces improved results compared to the 3D-CNN. Feature extraction techniques play a significant role in the generation of an accurate caption. Both static and dynamic feature extraction were explored in SEmantic Feature Learning and Attention-Based Caption Generation (SeFLA) (Lee and Kim 2018). The paper suggested a multi-modal feature learning system with an attention mechanism. This research explains the prominence of semantics acquired using LSTM along with broad-spectrum visual features extracted using a ResNet CNN for generating accurate descriptions. Semantics was further categorized as static or dynamic, where static (a noun in the description) refers to the object, person, and background, whereas dynamic (a verb in the description) corresponds to the action taking place within the input video, as shown in Figure 5.

Systems with multiple independently trained models from different domains utilized in a pipeline fashion, focusing only on the input and output and skipping all the intermediate steps to get the required output, are called end-to-end systems. In video description, we have visual and language models for vision and language processing. If we train them independently, and then plug them into a pipeline, they are end-to-end systems. The first end-to-end trainable deep RNN (Zhang et al. 2017) proposed a description model employing Caffe CNN (Jia et al. 2014), a variant of AlexNet, fused with a two-layered LSTM accompanied by transfer learning, forming the ED to describe videos in an efficient fashion.

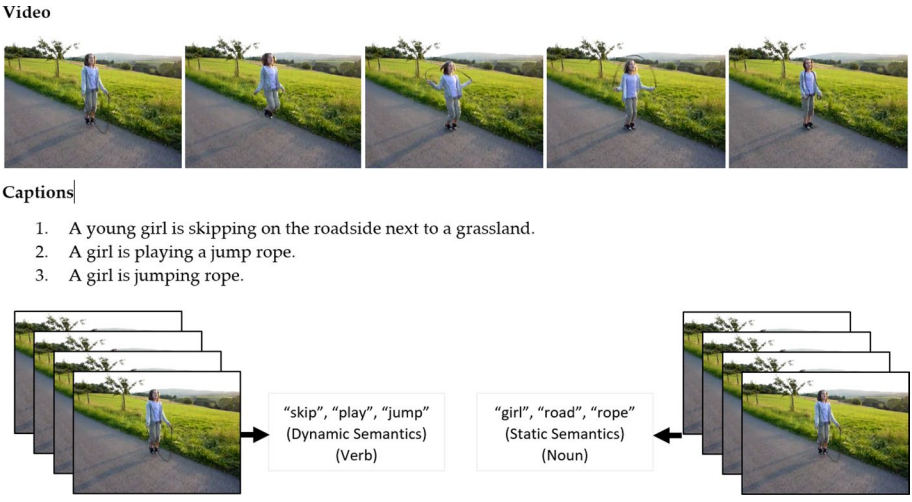

Fig. 5 Semantic feature categorization as either static or dynamic, where *static* refers to the object, the person, and/or the background, and dynamic corresponds to the action taking place within the input video. Sample video frames and reference captions were taken from the Microsoft Video Description (MSVD) dataset

The model is trained on the popular ImageNet dataset, and trained weights are utilized for initialization of the LSTM-based language model, boosting the training speed. Feature extraction, aggregation, and caption generation, are all steps involved in the process that require memory for computation and evaluation. Limitations associated with memory requirements while generating captions is addressed in EtENet-IRv2 (Olivastri 2019), which is also an end-to-end trainable ED architecture proposing a gradient accumulating strategy employing Inception-ResNet-v2 (Szegedy et al. 2017) and GoogLeNet (Szegedy et al. 2015) with two-stage training for encoding. Evaluation of benchmark datasets (Rafiq et al. 2021) showed significant improvement, but with a limitation on the computational resources required for end-to-end training.

Long Short-Term Memory with Transferred Semantic Attributes (LSTM-TSA) (Pan et al. 2017) emphasizes the fusion of jointly exploited semantic attributes for both images and video, along with the significance of its injection in extracted visual features for automatic sentence generation. A transfer unit to model the jointly associated attributes extracted from images and videos was proposed for integrating semantic attributes into sequence learning. The visual model, afterward accompanied by semantic attributes mined from both images and video, is fed into an LSTM for caption generation. Similarly, ResNet50 and VGG-16 CNN architectures coupled with the LSTM structure were exploited in Rivera-soto and Ordóñez (2013) for sequence-to-sequence video description models. Three types of model were proposed: mean pool, a single-layer ED, and a stacked ED. Extensive experimentation, performed on the Microsoft Video Description (MSVD) dataset, proved that a single-layer ED network performs best for machine translation, but complicates the network convergence for video descriptions. Instead, two stacked LSTM networks concentrate efficiently on both visual encoding and natural language decoding.

Both global and local features play roles while captioning a video. Object-aware aggregation with a bidirectional temporal graph-based (OA-BTG) description model (Zhang et al. 2019a) captures in-depth temporal dynamics for significant objects in a video, and learns particular spatio-temporal representations by performing object-aware local feature aggregation on the detected object-aware regions and frames. A bi-directional graph is designed to capture both forward and backward temporal trajectories of a specific object. For learning certain representations, the global frame sequence and object spatio-temporal trajectories are aggregated. The influence of objects at a particular time is differentiated using a hierarchical attention mechanism. Understanding the global contents of a video, as well as the in-depth object information, is essential for the generation of flawless and fine-grained automatic captions. Likewise, RecNet (Wang et al. 2018a) (a novel ED-reconstructor architecture) also exploits the phenomenon of the global and local structure of the video by employing two types of reconstructors and bi-directional flow. The relationship between video frames and generated natural sentences is established and enhanced by incorporating a reconstruction network for video captioning. Global structure is captured by mean pooling, while the attention mechanism is included in the local part of the model to exploit local temporal dynamics for the reconstruction of each frame.

CVC (Yan et al. 2010) proposed a system using the ED approach to describe numerous characteristics of off-site viewers or an audience's crowd (such as the number of people in the crowd), the movement conditions, and the flow direction. The model employs a 2D/3D CNN for crowd feature extraction from video, which then feeds into an LSTM-GRU-based language model for captioning. The authors created their own crowd captioning dataset based on WorldExpo10. Based on the famous S2VT model, the CVC model showed improvement because of the small dataset and simple captions. To deal with the uncertainties faced during inappropriate data-driven static fusion methods employed in the video

description system, TDDF (Zhang et al. 2017) established a task-driven dynamic fusion method. VGG-19 and the GoogLeNet CNN were employed for extraction of appearance features, whereas C3D was utilized for motion feature extraction. The proposed method achieved the best METEOR and CIDEr score when evaluated with the MSVD and Microsoft Research Video to Text (MSR-VTT) datasets, compared to a single-feature system.

One of the significant characteristics required in a generated description is its diversity. Lexical-FCN (Shen et al. 2017) was proposed for generation of multiple diverse and expressive captions based on weak video-level sentence annotations. Although the model is trained with a weakly supervised signal, it produces multiple diverse and meaningful captions with the sequence-to-sequence language model. A convolution-based lexical FCN forms the visual part of the model, whereas the language model follows the state-of-the-art S2VT (Venugopalan et al. 2015) mechanism with a bi-directional LSTM to improve the quality of automatically generated captions. Diversity, coherence, and informativeness of the generated captions ensure the supremacy of the proposed model.

3.1.2 RNN–RNN

In early research, employing an RNN in both encoding and decoding for neural machine translation demonstrated very efficient performance. Researchers explored the horizons for video description by exploiting the RNN for both feature extraction and language modeling. Long-term recurrent convolutional networks (LR-CNs) (Donahue et al. 2017) were proposed with an ED architecture for long sequences with time-varying input and output. Video description is carried out using three variants of the architecture: an LSTM encoder and decoder with a conditional random field (CRF) max, an LSTM decoder with a CRF max, and an LSTM decoder with CRF probabilities. For a broader scope, the research focuses on activity recognition, image captioning, and video description.

A state-of-the-art, sequence-to-sequence video-to-text generator, S2VT (Venugopalan et al. 2015), following the ED architecture uses a stacked two-layer LSTM ED model that takes a sequence of RGB frames as input and produces a sequence of words corresponding to the input sequence. The encoding and decoding of the frame and word representations are learned jointly from a parallel corpus. To model the temporal aspects of activities typically shown in videos, optical flow (Brox et al. 2014) between pairs of consecutive frames is computed. The flow images pass through a CNN and are provided as input to the encoding LSTM. Employing a single LSTM for both encoding and decoding allows parameter-sharing between the two states. Sequential processing at both stages is incorporated because both input and output are of variable, potentially different, lengths. Loss is computed on the decoding side for optimization of the video description system. The model was taken as a basis by many researchers, like in S2VT with knowledge (S2VTK) (Wang and Song 2017), and follows a detect, fetch, and combine approach. It first detects an object in the video, fetches object-related information from a knowledge base DBPedia, and creates a vector using Doc2Vec. Both elements, i.e., visual extracted features and related information regarding the detected object, are then input to the LSTM-based language model for caption generation. Another model based on S2VT (Venugopalan et al. 2015), meaning a guided system (Babariya and Tamaki 2020), was proposed in connection with the object detection module YOLOv3 (Redmon and Farhadi 2018) to generate correct captions having a similar meaning. The proposed model picks the object having the highest abjectness score in the YOLO detector, and after detection, searches for the nearest string describing the detected object. Word2Vec (Demeester et al. 2016) pre-trained on part of the Google

News Dataset, is used for string embedding. Semantic similarity or caption meaning is considered for optimization of the training instead of training using the conventional word-by-word loss. Following the object detection approach, tube features for video description was proposed in Zhao et al. (2018). Trajectories of objects in input videos are captured, employing a Faster-RCNN (Wallach 2017) to extract region proposals, and afterwards, the regions from different frames (but belonging to the same objects) are associated as tubes. A similarity graph is created among the detected bounding boxes, and a similarity score is assigned to a pair of bounding boxes in adjacent frames. A bi-directional LSTM encoder encodes both forward and backward dynamic information of the tubes, and converts each tube into a fixed-sized visual vector, whereas a single LSTM decoder with an attention mechanism to monitor the most correlated tubes, generates the captions.

Dealing with multiple and diverse caption generation, the Diverse Captioning Model (DCM) (Xiao and Shi 2019z) is a conditional Generative Adversarial Network (GAN) with an ED model to describe video content with multiple descriptions. It can describe video content with great accuracy, and can capture both forward and backward temporal relationships to encode the extracted visual features. For a given video, the intermediate latent variables of the conventional encode-decode process are utilized as input to the conditional GAN (CGAN) to generate diverse sentences. Generators comprising different CNNs generate diverse descriptions while the discriminator inspects the worth or quality of the formed captions. Combining the reasonableness and differences between the generated sentences, a diverse captioning evaluation metric (DCE) was also proposed.

Feature extraction from pre-trained models and their sensible arrangement can considerably affect the quality of generated captions. These extracted features or modalities are recognized, and their detailed effects were discussed in Hammad et al. (2019) with an S2VT (Venugopalan et al. 2015) basis. The different video modalities can be recognized as a frame or image, with a scene, the action, and audio (Ramanishka et al. 2016). All these modalities have their own significance while generating the description and inclusion of essential features, and when accompanied by a decoder with an attention mechanism, can help the model to extract the most pertinent information related to the scene and can have a substantial effect on quality improvement of the generated description. A human-like ability to extract the most relevant information from a scene can be incorporated by intelligent selection and accurate concatenation of features.

3.1.3 CNN–CNN

TDConvED (Chen et al. 2019b) was the first and (so far) the only ED approach fully employing CNNs for both visual and language modeling. To address the limitations of vanishing/exploding gradients, as well as the recurrent dependency of the RNN preventing parallelization during sequence training, a system with convolutions in both the encoder and the decoder was proposed. Feed-forward convolution networks are free from recurrent functions, and previous step computations are not considered in the next step, so parallelization of sequence training can be achieved. The proposed model also exploits the temporal attention mechanism for sentence generation. In the encoder, the convolutional block is provided with temporal deformable convolutions by capturing dynamics in temporal extents of actions or scenes. The significant contribution of this research is to use convolutions for sequence-to-sequence learning and for enhancing the quality of video captioning.

3.2 Discussion - ED based approaches

The famous Encoder–Decoder structure for video description configures two neural networks, one for visual information extraction and other for textual narration generation corresponding to the visual perspective. This composition of neural networks involve CNN, RNN, LSTMs, GRUs, and tranformers as its encoding and decoding modules. CNNs are proficient in automatic identification of relavent features without human intervention (Zhang et al. 2019b). As per (Goodfellow et al. 2016) the key charachteristics of CNN are sparse interactions, equivalent representations, and parameter sharing. Scanning regions instead of whole image results in less parameters with simplified and speedy training process and enhance generalization capability to avoid overfitting (Alzubaidi et al. 2021). RNNs are applied mostly in speech and language processing contexts. It uses sequential data to convey the information catering the order of the sequence. It offers recurrent connections to memory blocks in the network and flow of information is controlled through gated units in the network. This algorithm's sensitivity to exploding and vanishing gradients are the main limitations associated with it while dealing with long range dependencies. In comparison with RNNs, CNNs are considered to be more powerful due to its less feature compatibility when compared to CNN (Alzubaidi et al. 2021). It variants LSTMs and GRUs are further enhanced to utilize less training parameters, less memory with more accuracy and faster execution. These deep learning models composition, i.e., CNN-RNN, RNN-RNN and CNN-CNN employed by researcher for video description task demonstrated their findings. A thorough empirical analysis by (Aafaq et al. 2019c) concluded that C3D is commonly employed model for visual features extraction from images and short clips. Inception-ResNet-v2 and temporal encoding achieved compareable results in features transformation. Towards language modelling, the depth or number of layers in the decoder module, internal state size, number of processed frames and word embedding with dropout regularization selection is crucial for high quality description generation. If we train these modules (visual & language) independently, and then plug them into a pipeline, they are end-to-end systems. These systems are pre-trained on large scale datasets and then fine-tuned on video description datasets for the down stream task of video description. The use of deep learning to caption video has been extensively researched, but there are still numerous challenges to be resolved including objects accurate identification and their interactions, generating improved event proposals for dense captioning, utilization of task specific transformers for vision and language accurate comprehension.

3.3 Attention mechanism

An attention mechanism can be characterized as an act of cautiously focusing on the directed, relevant, and important parts in an image, frame, or scene, i.e., considering only the salient contents to be described, while ignoring others. The general structure of the video captioning model supported by the attention mechanism is grounded on various types of cues from the video. These cues are integrated into the basic framework of the ED to get the decoding process to concentrate on specific parts of the video at each time step to generate an appropriate description.

Before establishment of the attention mechanism in a standard ED architecture, the encoder block of the employed model was able to convert image or frame features into a single context vector, which is then fed to the decoder unit for caption generation word

by word. For images loaded with multiple/complicated objects, one intermediate vector is unable to adequately convey the subsequent image features, instigating the loss of important information and substandard caption generation. The fusion of the attention mechanism empowers the encoder to concentrate on the various essential parts of the frame with distinct intensity, generating multiple context vectors, resulting in enhanced quality of the generated natural language sentences.

Let us suppose the video description system takes a video and generates caption Y for that video such that

$$Y = \{W_1, W_2, W_3, ..., W_c\}, W_i \in R^K \tag{4}$$

where K is the size of the vocabulary, and C is the number of words in the caption. Using a 2D/3D CNN to extract the features from each frame/clip of the video, we have an annotation vector as a collection of all the intermediate context vectors or feature vectors, expressed as

$$CV = \{CV_1, CV_2, CV_3, ..., CV_L\}, CV_i \in R^D \tag{5}$$

where L is the number of feature vectors, each of which is a D-dimensional representation corresponding to the relevant part of the frame/clip of the video (Xu et al. 2015; Bahdanau et al. 2015). The attention mechanism permits more direct dependence between the states of the model at different points in time (Raffel and Ellis 2015). A model produces an intermediate context vector, or hidden state CV_t, at time step t. Attention-based models compute a single context vector at time t, SCV_t, as the weighted mean of the state sequence CV_i, expressed in (6) and simplified as (7):

$$SCV_t = \sum_{i=1}^{L} CV_i.\alpha_{ti} \tag{6}$$

or

$$SCV_t = CV_1 * \alpha_{t1} + CV_2 * \alpha_{t2} + CV_L * \alpha_{tL} \tag{7}$$

Each CV_i contains information about the whole input sequence with a strong focus on the parts surrounding the *i'th* word of the input sequence, which is the main essence of the attention mechanism, to find mappings between an input element and its corresponding output. The attention weight computed at each time step t for each feature vector CV_i using Softmax is

$$\alpha_{ti} = \frac{e^{(SCORE_{ti})}}{\sum_{K=1}^{L} e^{SCORE_{ti}}} \tag{8}$$

where

$$SCORE_{ti} = FUNC_{ATT}(CV_i, W_{t-i}) \tag{9}$$

$SCORE_{ti}$ in (9) is the function of attention, which indicates the goodness of input position, and the generated output matches how well the input around the *i'th* position matches output at time t. The score is computed based on hidden state CV_i and decoder-generated output in the previous time step, i.e., W_{t-1}. SCV_t is then concatenated with the word output from the decoder's previous time step, resulting in a concatenated context vector with the weighted feature information conveying where to focus more attention while generating the

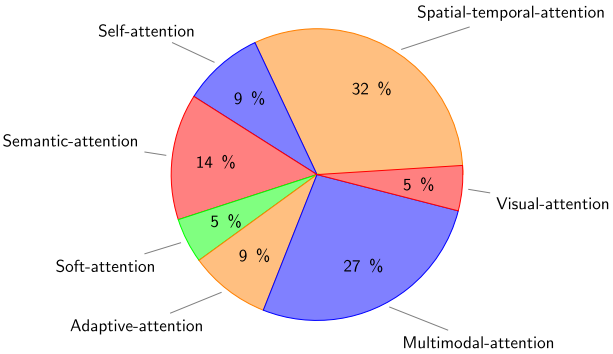

Fig. 6 Distribution of various attentions for video descriptions based on the literature explored for this research work

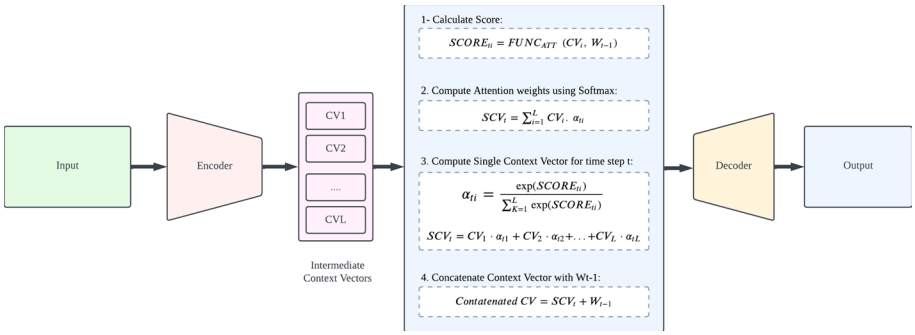

Fig. 7 The attention mechanism in the Encoder–Decoder architecture at time t

word at this particular position. This process continues until reaching the decoder output $\langle END \rangle$ token. The generic review network in Yang et al. (2016) also proposed the same review steps as the concatenation of feature vectors with attention weights, producing a thought vector after each review for input to the decoder attention mechanism. Figure 7 depicts the attention process carried out inside the ED architecture.

Strategies for model optimization during training include the teacher forcing technique (Williams and Zipser 1989), curriculum learning (Bengio et al. 2009), and RL-based optimization techniques. Teacher forcing is a simple way to train RNN-based models while constituting a concatenated context vector. A word is provided from reference annotation instead of the actual generated word at the previous time step to guide word generation. This showed improvements in the model's learning capabilities, and produced better results in the testing phase. Later, (Huszár 2015) proved the biased learning tendency of teacher forcing and curriculum learning, and proposed professor forcing (Goyal et al. 2016) for RNN optimization by adopting an adversarial domain method for alignment of the RNN during training and testing phases (Chen et al. 2019a).

Different types of attention can be applied depending on the nature of the problem and situation. Figure 6 shows adaptation of different types of attention mechanism for video captioning. In Gella et al. (2020), the authors proposed two categories of temporal attention mechanisms (local and global temporal structures) considering the task of video description. The local temporal structure symbolizes fine-grained or detailed in-depth information like picking up the spoon or laying on the bed, whereas the global

temporal structure mentions the sequence of events, objects, shots, and persons in the video. For a video description system to be state of the art, it must selectively concentrate on the most prominent features of the sequence in the video, exploiting both global and local temporal information.

To generate high-quality captions, the model needs to integrate the fine-grained visual clues from the image/frame. Lu et al. (2017) proposed a novel adaptive attention model with a visual sentinel. The spatial and adaptive attention-based model was capable of automatically making decisions on when to count on visual signals and which part of the image to focus on at a particular time, and vice versa. The combination of spatial and adaptive attention with the employed LSTM produced an additional visual sentinel providing a fallback option to the decoder. The sentinel gate helps the decoder get the required information from the image. Supporting that idea, researchers in Gao et al. (2019) and Song et al. (2017) suggested a system of hierarchical LSTM (hLSTMat) based on adaptive and temporal attention to enrich the representation ability of the LSTM. The model's capability to adapt to low-level visual or high-level language information at a certain time step demonstrated robust description for videos.

Two commonly available design strategies in the captioning related literature are top-down and bottom-up. The top-down (or modern) strategy starts from the essence, gist, or central idea of the image/frame and then transforms that gist into appropriate words, whereas the bottom-up (or classical) approach first abstracts the words for the various dynamics of the frame, and then combines those words in a coherent manner. Both approaches suffer from certain limitations; top-down is unable to attend to the fine-grained details, and end-to-end training is not possible for the bottom-up approach. To get the benefits from both strategies You et al. (2016) developed a model to combine both approaches through semantic attention. The proposed model is capable of selectively attending to semantic ideas and regions, guiding when and where to pay more attention. The fusion of attention with the employed RNN structure leads to more efficient and robust performance. Likewise, to tackle the correlation between caption semantics and visual content, Gao et al. (2017) proposed an end-to-end attention LSTM with semantic consistency to automatically generate captions with rich semantics. Attention weights computed from video spatial dynamics are fed into the LSTM decoder, and finally, to bridge the semantic gap between visual content and generated captions, a multi-word embedding methodology is integrated in the system.

The role of spatial and temporal attention exploited for the task of video captioning is very important. Temporal attention refers to the specific case of visual attention that involves focusing attention on a particular instant in time, whereas spatial attention involves some specific location in space. Most of the recent models have adopted spatial-temporal attention to upgrade the accuracy of the model. The studies in Lowell et al. (2014) and Laokulrat et al. (2016) presented early approaches, exploiting temporal attention for sequence-to-sequence learning. To attend to both spatial and temporal information present in video frames, Chen et al. (2018b) presented a visual framework based on saliency spatio-temporal attention (SSTA) to extract the informative visual information better and then transform it into natural sentences using an LSTM decoder. The designed spatial mechanism facilitates capturing the dominant visual notions from salient regions, and the semantic context from the non-salient regions of the video frame. Experimentation on spatial attention demonstrated that employing residual learning for spatial attention feature generation can improve performance. Models with their approach, visual, and language components are summarized in Table 4 for convenience.

Table 4 The attention mechanism employed for video captioning with visual and language components, short contributions and shortcomings (if any) are also mentioned with each approach

S/N	References	Year	Approach	Model	
				Visual	Language
1	Ji et al. (2022)	2022	ADL (Attention-based Dual Learning)	Inception-V4	LSTM + MHDPA (multi-head dot product attention)
Contributions: An attention based dual learning approach (ADL) which can minimize the semantic gap between raw videos and generated captions by minimizing the differences between the reproduced and the raw videos, thereby enhancing the quality of the generated video captions.					
Shortcomings: NA					
2	Peng et al. (2021)	2021	global text combined with local attention enhancement (T-DL)	2D/3D CNN, global Attention	GRU
Contributions: Proposed extraction of 2D/3D video features bidirectional time flow, global image attention, training method of local attention focusing important words of the text. Global dynamic attention is added to text training to generate description text with reference to the context during the training.					
Shortcomings: NA					
3	Ryu et al. (2021)	2021	SGN (Semantic Grouping Network)	2D/3D CNN	LSTM + Semantic Attention
Contributions: The proposed network encodes the video into semantic groups that are in terms of relevant frames and the corresponding word phrases of the partially decoded caption, and adaptively decodes the next word based on the semantic groups. Moreover Contrastive Attention (CA) loss to provide labor-free supervision for the correct visual- textual alignment within each semantic group is proposed.					
Shortcomings: SGN's repeated grouping process reduces the inference speed of about 25%.					
4	Chen et al. (2021)	2021	Scan2Cap	Mask R-CNN for 2D-3D projection, VoteNet for 3D-2D projection	Fusion GRU
Contributions: An end-to-end trainable model capable to detect and describe 3D objects and their relationships in RGB-D scans.					
Shortcomings: Difference in viewpoint, limited field of view and motion blur can cause poor performance.					
5	Chen and Jiang (2021)	2021	EC-SL (Event Captioner-Sentence Localizer)	C3D, ISAB employing multi-head attention	bi-LSTM
Contributions: Integrating the temporal localization and description for events in untrimmed videos under the weakly supervised setting, where temporal boundary annotations are not available. Creation of information communication channels between the tasks for better bridging and unification.					
Shortcomings: Without external training data, the concept learner can not accurately detect concepts that are visually small and still suffers from the long-tailed issue.					

Table 4 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
6	Perez-Martin et al. (2021b)	2021	Visual-Semantic-Syntactic Aligned Network(SemSynAN) - Temporal Attention based on Soft Attention	2D/3D CNN	LSTM
Contributions: Created visual-syntactic embeddings by exploiting the Part-of-Speech (POS) templates of video descriptions. The learning process is based on a match and rank strategy, and ensures that videos and their corresponding captions are mapped close together in the common space. Then map the input video and generate our desired visual-syntactic embedding while generating while producing features for decoder. The proposed video captioning that integrates global semantic and syntactic representations of the input video. It learns how to combine vi- sual, semantic, and syntactic information in pairs.					
Shortcomings: NA					
7	Xu et al. (2020)	2020	Temporal-spatial and channel attention	Inception-V3	LSTM
Contributions: video description model based on temporal-spatial and channel attention is proposed. Model fully utilized the essential characteristics of CNN and added channel features into the attention mechanism of the model. Therefore, the model can use visual features more effectively and ensure the consistency of visual features and sentence descriptions to enhance the effect of the proposed model.					
Shortcomings: The model cannot give the correct word after the article "a," This may be due to the lack of attention mechanism for the ability to model abstract nouns that cannot express specifically. Similarly for articles like "a", because it is not very relevant to the vision, all regions in the video are equally treated, so there would be no salient regions.					
8	Zhang et al. (2020)	2020	Spatial-Temporal attention	Appearance features: InceptionResNetV2 pretrained on ImageNet, Motion features: C3D based ResNetXt-101 pretrained on Kinetics-400, Object features: ResNetXt-101 based Faster-RCNN pretrained on MSCOCO	2 layered LSTM with a temporal- spatial attention module.
Contributions: The proposed model generated video description with the assistance of an original training strategy. A learnable object relational graph to fully explore the spatial and temporal relationships between objects is created. Object representations can be enhanced during the process of relational reasoning. Partial and complete relational graphs are explored in this study. teacher-enforced learning is also introduced to enhance the quality of generated captions.					
Shortcomings: NA					

Table 4 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
9	Yan et al. (2020)	2020	Spatial-Temporal Attention (STAT)	CNN (GoogleNet) + R-CNN (Faster R-CNN)	LSTM
Contributions: A syntax-aware module is proposed that forms a self-attended scene representation to model the relationship among video objects and then decodes syntax components by setting different queries, targeting the action in video clips. An action-guided captioner that learns an attention distribution to dynamically fuse the information from the predicate and previously predicted words, avoiding wrong-action prediction in generated captions.					
Shortcomings: The global temporal information provided by 3D CNNs is not always enough to learn finer actions in video clips.					
10	Gao et al. (2020)	2020	Fused GRU with Semantic-Temporal Attention (STA-FG)	2D/3D CNN	GRU
Contributions: An end-to-end framework incorporating the high-level visual concepts prediction into CNN-RNN approach for videos captioning. Nouns and verbs from the training sentences as used as concepts while training a multi-label CNN. Both low-level visual features and high level semantic representation are fused and a semantic and temporal attention mechanism in a fused GRU network for accurate video captioningis proposed.					
Shortcomings: CIDEr and ROUGE scores are not computed. Only BLEU and METEOR scores are demonstrated during the model evaluation.					
11	Liu et al. (2020)	2020	Soft Attention (SibNet)	GoogleNet + Inception	LSTM
Contributions: A dual branch achitecture composed of visual branch and semantic branch. Content branch to encode salient visual content information and the semantic branch to encode high-level semantic information with the guidance of ground truth captions brought by visual-semantic joint embedding. The content branch, the semantic branch and the decoder are trained jointly by minimizing the proposed loss function. TCB (Temporal Convolution Block) is proposed providing more efficient video temporal encoding than RNN with less number of parameters.					
Shortcomings: NA					
12	Pramanik et al. (2019)	2020	Self-Attention (OmniNet)	ResNet-152	Transformer with a two-step attention mechanism (Spatial & temporal)
Contributions: Proposed an extended transformer towards a unified architecture, which enabled a single model to support tasks with multiple input modalities and asynchronous multi-task learning. Image, text and video peripherals are described as direct conjuction of spatial and temporal phenomenon in this research work. The proposed model can process and store spatio-temporal representation for each of the input domains and then decode predictions across a multitude of tasks.					
Shortcomings: NA					

Table 4 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
13	Yan et al. (2019)	2019	Multi-Granular Attn (GLMSIR)	VGG16 + Faster R-CNN	LSTM
Contributions: Graph-based Learning for Multi-Granularity Interaction Representation (GLMGIR) for fine-grained team sports auto-narrative task. This framework is featured with two key components. firstly, a multi-granular interaction modeling module is proposed to extract among-subjects' interactive actions in a progressive way, for encoding both intra- and inter-team interactions. Secondly, based on the above multi-granular representations, a dense multi-granular attention module is developed to specifically handle the task of spatio-temporal granular feature selection for generating action or event descriptions of multiple spatio-temporal resolutions. The output of both the modules is input into a decoding network, which generates the final descriptions.					
Shortcomings: Explored by the authors that granularity itself is an important threshold, so if the chosen granularity is too big, background noises can jeopardize the learned representations. If the chosen granularity is too small, it cannot provide enough information for generating useful descriptions. A wise selection of granuality threshold is required.					
14	Zhou et al. (2019)	2019	Grounded video description	Feature: ResNeXt-101 + LSTM	LSTM
Contributions: Dataset ActivityNet Entities is created, grounding video description on the noun phrase level to bounding boxes. Wit the bounding boxes supervision, a grounded video description model is proposed.					
Shortcomings: Extra context and region interaction introduced by the self-attention confuses the region attention module and without any grounding supervision makes it fail to properly attend to the right region.					
15	Chen and Jiang (2019)	2019	Motion Guided Spatial Attention (MGSA)	static features: CNN (GoogleNet, Inception-ResNst-V2), Motion information: C3D	LSTM
Contributions: A novel video captioning framework Motion Guided Spatial Attention (MGSA), which utilizes optical flow to guide spatial attention. incorporates optical flow for attention guidance in video captioning. Introduced recurrent relations between consecutive spatial attention maps resulted in boost to captioning performance and designed a recurrent unit called Gated Attention Recurrent Unit (GARU) for this purpose.					
Shortcomings: NA					
16	Gao et al. (2019)	2018	Hierarchical LSTM with Adaptive Attention (hLSTMat)	CNN (ResNet-152 He et al. (2016))	Hierarchical LSTM
Contributions: The proposed hLSTMat framework, with the representation enrichment ability of LSTM, automatically decides when and where to use visual information, and when and how to adopt the language model to generate the next word for visual captioning. Spatial and temporal attention is used to decide where to look at visual information and the adaptive attention decides when to rely on language context information. At each time step, with both LSTM fusion, the low-level visual information and high-level language context information is obtained through hierarchical LSTMs. When connecting LSTMs sequentially, the second LSTM refines the first LSTM.					
Shortcomings: Increased number of parameters and training time for two streams of hierachical LSTM.					

Table 4 (continued)

S/N	References	Year	Approach	Model	
				Visual	Language
17	Chen et al. (2018b)	2018	Spatiotemporal Attention	CNN (VGG16)	LSTM
Contributions: Visual saliency information is utilized for arrangement of visual concepts in frames and paying attention to the informative frames in the video. As salient objects point towards important and dominant visual concepts, so a saliency-based spatiotemporal attention mechanism for video captioning is proposed. The model is capable of accurately aligning the visual information with the predicted words, to form a diverse caption.					
Shortcomings: There is a gap between proposed model and other methods while evaluating in terms of METEOR and CIDEr scores. There are two main reasons; one is that the visual feature extractor used are weak, and the second is that fragment-level features are not taken into consideration.					
18	Wang et al. (2018c)	2018	Hierarchically aligned cross-modal attention (HACA)	Image: ResNet, Audio: VGGish	LSTM
Contributions: The proposed model learnt the attentive representations of multiple modalities along with the alignment and fusion of local and global contexts for video understanding and video captioning tasks. Deep audio and visual features are employed for description generation.					
Shortcomings: NA					
19	Yu (2017)	2018	Gaze Encoding Attention Network (GEAN)	Scene: GoogleNet, Motion:C3D, Fovea: GoogleNet	GRU
Contributions: The research work study the effect of supervision by human gaze data on atten- tion mechanisms, particularly for video captioning. A dataset of movie clips with multiple annotations and human gaze tracking labels is created. The proposed GEAN model efficiently incorporates spatial attention by the gaze prediction model with temporal attention in the language decoder.					
Shortcomings: NA					
20	Li et al. (2019b)	2018	Residual attention (Res-ATT)	Static: Google Net, ResNet, Motion: C3D	LSTM
Contributions: (Res-ATT) an attention based model, considering sentence internal information which usually gets lost in the transmission process. Integration of residual mapping into a hierarchical LSTM network to solve the degradation problem is proposed.					
Shortcomings: NA					

NA Not available, *S/N* serial number

Temporal attention commonly captures global features, whereas spatial attention captures local features. Xu et al. (2020) proposed channel attention along with spatial and temporal attention to ensure consistency in the visual features when generating natural language descriptions. Channel features refers to several feature graphs generated by each CNN layer. Spatial (S), temporal (T), and channel (C) attention weights are used to compute the fused features for decoding and caption generation. Eight different combinations of the three attentions were investigated, and S-C-T was the best performing combination, defining the sequence of attention for consideration while capturing features. For end-to-end learning (Chen and Jiang 2019), Motion Guided Spatial Attention (MGSA) is a spatial attention system for exploiting motion between video frames and was developed with a Gated Attention Recurrent Unit (GARU).

Considering attention while incorporating external linguistic knowledge in a captioning system, Zhang et al. (2020) proposed a combination of the object-relational graph (ORG) model with teacher recommended learning (TRL) by Williams and Zipser (1989). The explored external language model (ELM) produces semantically more analogous captions for long sequences. Appearance, motion and object features are extracted by employing 2D and 3D CNNs, reflecting the temporal and spatial dynamics of the given video. The STAT captioning system decoder (Yan et al. 2020) automatically selects important regions for word prediction depending on the local, global, and motion features extracted, exploiting the spatial and temporal structures in the video. The end-to-end semantic-temporal attention (STA-FG) model (Gao et al. 2020) integrated global semantic visual features of a video into the attention network to enhance the quality of generated captions. The hierarchical decoder is comprised of a sematic-based GRU, a semantic-temporal attention block, and a multi-modal decoder for word-by-word semantically rich and accurate caption generation. SibNet (Liu et al. 2020) employs a dual branch structure for video encoding where the first branch deals with visual content encoding, and the second branch captures semantic information in the video, exploiting visual-semantic joint embeddings. The two branches were designed using temporal convolution blocks (TCBs) and fused employing soft attention for caption generation. OmniNet (Pramanik et al. 2019), employing transformer and spatio-temporal cache mechanisms, supports multiple input modalities and can perform parts-of-speech tagging, video activity recognition, captioning, and visual question answering. Due to the efficient capture of global temporal dependencies in sequential data by the employed self-attention mechanism in the transformer architecture, simultaneous shared learning from multiple input domains is possible for accurate and superior performance.

The intrinsic multi-modal nature of video (i.e., static or appearance features, motion features, and audio features) contributes while generating captions. Learning most of these features increases the model's ability to better understand and interpret the visuals, thus improving the overall captioning quality. The video description systems proposed in Wang et al. (2018c), Li et al. (2017), Xu et al. (2017), and Hori et al. (2017) exploit a multi-modal attention mechanism for automatic natural language sentence generation.

More recently, dense video captioning (sports-related) was proposed in Yan et al. (2019) to segment distinct events in time and to then describe them in a series of coherent sentences, particularly focusing on multiple, fine-grain granularities or details of teams. The model auto-narrates the inter-team, intra-team, and individual actions, plus group interactions and all the interactive actions in a progressive manner. Incorporation of a dense multi-granular attention block exploits the spatio-temporal granular feature selection to generate a description. Authors also developed a Sports Video Narrative (SVN) dataset comprising 6k sports videos from YouTube.com and designed an evaluation metric Fine-grained

Captioning Evaluation (FCE) to measure the accuracy of the generated linguistic description, demonstrating fine-grained action details along with the complete spatio-temporal interactional structure for dense caption generation.

3.4 Discussion—attention based approaches

Attention mechanism a general notion of memory, was implemented at first for the performance improvement of Encoder–Decoder based model in the machine translation domain (Bahdanau et al. 2015). Its key concept combines all the encoded input vectors in a weighted manner, with the most salient vectors being given the highest weights. The attention mechanism intended to form a direct connection with each time-step and enable the decoder to utilize the most relevant parts of the input sequence in the most flexible manner. The crucial limitation imposed by ED's fixed length encoding vector for long and complex sequences is its inability to retain long sequences and hinder system performance. Attention mechanism's primary reason for creation was to address the bottleneck of handling long range dependencies. In Implicit attention, the system tend to ignore some of the input parts while concentrating on the other parts. In contrast, explicit attention weigh each part of the input based on previous inputs and concentrate accordingly. Various types of proposed attentions include soft (Vaswani et al. 2017; Liu et al. 2018), hard, self (Pramanik et al. 2019), adaptive, semantic, temporal, spatial (Chen and Jiang 2019), spatio-temporal (Chen et al. 2018b; Yan et al. 2020; Zhang et al. 2020), semantic-temporal (Gao et al. 2020), residual (Li et al. 2019b) , global, and local (Peng et al. 2021) attention. The attention mechanism eliminates the vanishing gradient by providing direct connection between the visual and language modules. The memory in attention mechanism is encapsulated in attention scores computed over time. The attention score acts as a magnifier, directing where to focus in the input for accurate output generation. Several optimization techniques, i.e., teacher forcing, curriculum earning, and reinforcement learning are also combined with the attention mechanism in ED structure to further boost the system performance. With easy to understand nature of attention mechanism, there is a need for more theoretical studies that will contribute to an understanding of the mechanism of attention in complex scenarios.

3.5 Transformer mechanism

Transformer (the first sequence transduction model), which has promptly become the model of choice in language processing, is a novel deep machine learning architecture introduced in 2017. It transforms one sequence into another following the ED architecture employing an attention mechanism, but it differs from the formerly explained ED mechanism in the sense that it does not imply any recurrent networks, i.e., an RNN, a GRU, or an LSTM. Transformers are designed to handle ordered sequences of data. However, unlike RNNs, they do not require ordered processing of the data, resulting in effective and efficient parallelization during training, compared to recurrent architectures. Becoming a fundamental building block of the most natural language-related tasks, it facilitates more parallelization during training, along with training on a huge amount of data. Table 5 lists some transformer-based approaches.

Self-attention, or intra-attention, is an attention mechanism relating different positions of a single sequence in order to compute a representation of the sequence. No doubt, the transformer mechanism theory revolves around self-attention following the

Encoder–Decoder architecture instead of the recurrent network to encode each position. It has multiple layers of self-attention addressing the prohibitive sequential nature, computational complexity, and memory utilization of CNNs and RNNs, and allows for parallelization resulting in accelerated training, handling of global/long-range dependency learning requiring minimal inductive bias (prior knowledge), and facilitating domain-agnostic processing of multiple modalities, i.e., text, images, and videos. Because of the performance-boosting characteristics of transformers, it has become the model of choice in NLP, CV, and the cross-modal tasks relating to the combination of these two world-leading realms. On the other hand, without transformers, self-attention in the recurrent architectures relies on sequential processing of input at the encoding step, resulting in computational inefficiency because the processing cannot be parallelized (Vaswani et al. 2017). Undoubtedly, self-attention-based transformers are a considerable improvement over recurrence-based sequential modeling.

The strong motive behind the development of the transformer was to get rid of the problem faced when learning long-range dependencies in sequences, and to allow for more parallelization by eliminating convolution and recurrence. The transformer does not rely as heavily on the prohibitive sequential nature of input data as the CNNs and RNNs do. Ordered sequential processing in the former deep models is a considerable obstacle in parallelization of the process. If the sequences are too long, it is either difficult to remember the content of distant positions in the sequence, or too difficult to ensure correctness. Although CNNs are much less sequential than RNNs, even then, the number of steps required to collect information from far-off positions, as the sequence grows, increases the computational cost and causes the long-range dependencies issue.

Similar to the ED structure, a transformer also has two key components: an encoder and a decoder. Both encoder and decoder contain a stack of identical units. Each encoder consists of two layers: the self-attention layer and the feed-forward neural network layer. The self-attention layer helps the encoder connect a specific part of the input sequence with other parts. The embedding is performed only in the bottom-most encoder, because only that encoder will get a vector, and all other encoders will get input from the previous encoder; i.e., the output of one encoder will become the input of the next encoder. After embedding the parts of the input sequence, each of them flows through the two layers of the encoder allowing parallel execution. Both attention and feed-forward neural network layers have a residual connection around them and are followed by a normalization layer. The decoder also contains the same layers with an additional ED attention layer to help the decoder focus on the relevant parts of the input sequence while generating captions.

Multi-head attention (refinement of self-attention) improves the performance of the attention layer by efficiently extending the model's ability to focus on different positions of the input sequence, and the attention layer is helped with multiple representation subspaces. In order to determine the position or distance of each word in the input sequence, the transformer adds a vector to each input embedding, i.e., positional encoding. Because of the same specific pattern of this vector, it facilitates efficient learning in the model. This positional encoding is able to scale unseen input lengths.

The output of the top encoder is transformed into attention vectors, K (Keys) and V (Values), and is fed into each decoder's ED attention layer. The attention layer in the decoder can only attend to the earlier positions in the output sequence before the Softmax calculation. The working mechanism of the ED attention layer in the decoder is the same as that of the multi-head attention layer except that it creates its Queries vector from the layer below it, and accepts the keys and values vectors from the top encoder. Logit and Softmax layers at the end of the decoder choose the word with the highest probability.

To solve the issues related to the computational complexity, memory utilization, and long-term dependencies of sequence-to-sequence modeling, several variants of transformers have been proposed in the literature over time. Since video description is a sequence-to-sequence modeling task, these updated versions of transformers can be utilized for reduced complexity and superior performance.

3.5.1 Standard/Vanilla transformer

A simple transduction architecture for sequence modeling is entirely based on an attention mechanism (Vaswani et al. 2017) with the objectives of parallelization (the ability to process all input symbols simultaneously) and a reduction in sequential computation, i.e., a constant number of operations is required to determine the dependency between two input symbols without considering their positional distance in the sequence. The commonly used recurrent layers in the ED architecture are replaced with multi-head self-attention layers where self-attention is about computing the sequence representation by relating different positions or parts of it.

A positional encoding vector, which is used to determine the context based on the position of the words in the sentence, is combined with the input, embedding both in the encoder and decoder stacks, since no recurrence or convolution is involved, so the positional encoding vector will serve the purpose of determining a word's relative or absolute position in the sequence. The dimensions for both input embedding and positional encoding is the same. Sinusoidals (sine and cosine functions of different frequencies) are used to compute these positional encodings. Although both learned positional encoding (Gehring et al. 2017) and sinusoidal positional encoding generate the same results, even then, sinusoidals are preferred because of sequence length extrapolation. Along with multi-head attention layers, each layer in the encoder and decoder section contains a feed-forward neural network (FFN). This FFN has two linear transformations with a rectified linear unit (ReLU) as an activation function.

Three main requirements–total computational complexity per layer, a parallelizable amount of computation, and the path length between long-range dependencies in the network–motivated the use of self-attention for mapping input and output sequences. The number of operations is fixed while computing the representation. Moreover, self-attention layers are considered faster, compared to recurrent layers, and are capable of producing models with increased interpretability. Figure 8 represents the architecture of the standard/vanilla transformer.

3.5.2 Universal transformer

The Universal Transformer (UT) (Uszkoreit and Kaiser 2019) proposed in 2019 by Google introduced recurrence in the transformer to address the issue of the standard/vanilla transformer not being computationally universal. The UT, a generalized form of the standard transformer, is a parallel-in-time recurrent self-attentive sequence model based on the ED architecture, and employs an RNN for representations of every position in both input and output sequences. The recurrence is over the depth, not over the position in the sequence. These representations are revised in parallel following two steps: first is use of a self-attention mechanism for information exchange; second is application of a transition function to the output from self-attention. For the standard transformer and RNNs, the depth (the number of sequential steps in the computation) is fixed because of the fixed number of

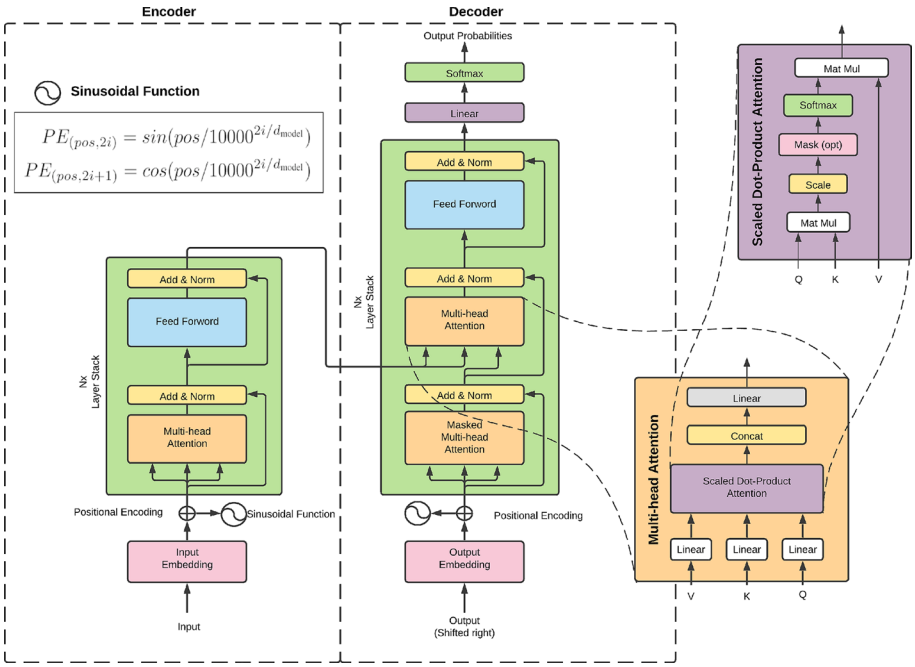

Fig. 8 The standard/vanilla transformer architecture (Vaswani et al. 2017)

layers, whereas there is no limit to the transition function count in a UT, proving its variable depth. This depth is the main difference between the standard transformer and the UT.

In the encoder section of the UT, representations are computed by applying multi-head soft attention at each time step for all positions in parallel tracked by a transition function and the residual connections; dropout and layer normalization are applied. The transition function can use either a separable convolution or a fully connected neural network. A dynamic per-position halting mechanism based on Adaptive Computation Time (ACT) is also incorporated for selection of the number of computational steps required for the refinement of each symbol, resulting in enhanced accuracy for many structured algorithmic as well as linguistic tasks. Bilkhu et al. (2019) employed a UT for single, as well as dense, video captioning tasks, utilizing a 3D CNN for video feature extraction, and reported promising results.

3.5.3 Masked transformer

Dense video captioning is about detecting and describing temporally localized events in a video. An end-to-end masked transformer model was proposed in Zhou et al. (2018) for dense video captioning. The proposed model consists of three parts. The video encoder is composed of multiple self-attention layers (since events are associated with long-range dependencies), so self-attention is required, instead of RNNs, for more effective learning. A proposal decoder following ProcNets (Zhou and Corso 2016) (an automatic procedure segmentation method) decodes the start and end times of events with a confidence score. A captioning decoder takes input from both the video encoder and the proposal decoder,

Table 5 Transformer-based approaches for video description

S/N	References	Year	Approach	Contributions	Shortcomings
1	Wu (2022)	2022	NSVA (Vision Transformer + TimeSformer , Transformer Decoder)	Identity-aware NBA dataset for sports video analysis (NSVA) built on web data. Design of a unified approach to process raw videos into a stack of meaningful features with minimum labelling efforts, showing cross modeling employing transformer architecture.	The bottleneck of the proposed model is player identification.
2	Im and Choi (2022)	2022	UAT: FEGs (feature extraction gates) + UEA (universal encoder attraction)	An end-to-end learnable transformer for video captioning. Proposed a feature extraction gate (FEG) that considers making better features by a fusion of CLS token and patch sequences along with universal encoder layer attention (UEA) constructed to obtain more information from one feature type.	NA
3	Yuan et al. (2022)	2022	Vanilla Transformer+ CMF (Cross-modal fusion) module	Enhanced 3D dense captioning method employing Cross-modal knowledge transfer using Transformer for 3D dense captioning. Proposed captioning through the knowledge distillation enabled by a teacher-student framework.	Due to the limited views of a single image, the performance of image-based dense captioning methods is significantly degraded when directly transferred to 3D scenarios.
4	Vo et al. (2022)	2022	NOC-REK : pre-trained BERT to embed the definition of each word into the embedding.	An end-to-end NOC-REK model which retrieves vocabulary from external knowledge and generates captions using shared-parameter transformers.	All potential objects may not be represented while extracting ROI from a given image. In the case of an heavy vocabulary, training both image features and vocabulary embeddings is required.
5	Wang et al. (2021)	2021	PDVC (parallel Decoding video Captioning)	An end-to-end framework formulating dense video captioning as a parallel set prediction task, significantly simplifying traditional captioning pipeline. An event counter to estimate the number of events in a video is introduced.	employment of transformer captioner for high performance.

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
6	Estevam et al. (2021)	2021	BMT-V+Sm (Bi-Modal Transformer with visual and semantic descriptor)	proposed an unsupervised descriptor that can be easily employed in video understanding tasks and can adequately capture the visual similarity between seen and unseen clips. visual similarity employed to generate event pro- posals replacing the audio signal.	slightly lower performance using semantic descriptor instead of audio modality while comparing with BMT (Bi-Modal Transformer).
7	Liu et al. (2021)	2021	O2NA (Object-Oriented Non-Autoregressive Approach)	Controlled contents video captioning approach focused on practical values than syntatical variations. The proposed approach tackle the controllable video captioning problem by injecting strong control signals conditioned on selected objects, with the benefits of fast and fixed inference time, which are critical for real-time applications.	A more powerful object predictor may be helpful in solving the issue of incorrectly predicted objects.
8	Deng et al. (2021)	2021	SGR (Sketch, Ground, and Refin) Vanilla Transformer	SGR (Sketch, Ground, and Refin) reversed the predominant "detect-then-describe" fashion and proposed to solve dense video captioning from a top-down perspective, i.e., generating a video-level story at first and then ground each sentence in the story to video segment for detailed refinement. By doing this, the event segments are predicted not only based on the visual information, but also the semantic coherence from the text.	NA

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
9	Song et al. (2021)	2021	Vanilla transformer with dynamic video memories enhanced attention	The proposed model avoided the event detection stage and generated paragraphs directly. In vanilla transformer the standard attention module is replaced with dynamic video memories enhanced attention.	NA
10	Zhang et al. (2021)	2021	RSTNet (Relationship-Sensitive Transformer) - Transformer with Adaptive attention	Proposed RSTNet combined with GA and AA. Grid Augmented (GA) unit to integrate the spatial information of raw visual features extracted from image. Adaptive attention (AA) for facilitation of fine-grained captioning by dynamically measuring the involvement of visual and text signals for word prediction.	NA
11	Li et al. (2020)	2020	HERO: Hierarchical encoder for video+language omni-representation pre-training	Cross-modal transformer combining subtitle sentence with its local video frames, followed by temporal transformer to obtain a sequentially contextualized embedding for each video frame is proposed primarily for representation learning.	Unlike Multi- stream, which leverages fine-grained region-level features, HERO's results are reported on global frame- level features. Therefore, it may be difficult for HERO to capture the inconsistency between hypothesis and video content.
12	Ging et al. (2020)	2020	COOT: Cooperative hierarchical transformer for video-text representation learning	An hierarchical transformer architecture with an attention-aware feature aggregation layer and a contextual attention module. Semantic alignment between vision and text features in the joint embedding space is proposed through cross-modal cycle-consistency loss. Both proposed components contribute jointly and individually to an improved retrieval performance.	NA

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
13	Jin et al. (2020)	2020	SBAT: Video captioning with a sparse boundary-aware transformer	boundary- aware pooling operation following the preliminary scores of multihead attention and selection for the features of different scenarios to reduce the redundancy is proposed with the aim to improve the vanilla transformer. Developed a local correlation scheme to compensate the local information loss brought by sparse operation. The developed scheme can be implemented synchronously with the boundary-aware strategy.	NA
14	Lei et al. (2020b)	2020	MMT: Multi-modal transformer	Multi-modal transformer following vanilla transformer taking both video and subtitles as encoder input and generated description employing decoder. At inference, greedy decoding is employed instead of beam search.	Models with both appearance features and motion features performed better than only using one features representation.
15	Lei et al. (2020a)	2020	MART: Memory-augmented recurrent transformer for coherent video paragraph captioning	Memory augmented transformer based paragraph description model conditioned on the given video with pre-defined event segments. The proposed model generated less redundant paragraphs while maintaining relevance to the videos.	Achieved best scores on CIDEr-D and R@4 metrics, not on B@4 and METEOR.
16	Zhu et al. (2018)	2020	Employing X-Lin+Tr Pan et al. (2020b)	Multi-view features and hybrid reward methods are proposed to address the variety of video content and diversity of captions. Weighted Ensemble method has little improvement over Average Ensemble, hence selected the former.	NA

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
17	Fang et al. (2020)	2020	V2C transformer	Creation of dataset annotated with captions and commonsense aspects. Proposed V2C-Transformer architecture that effectively generates relevant commonsense descriptions.	Do not have enough annotations per sample to compute a fair BLEU score for comparisons.
18	Iashin and Rahtu (2020)	2020	Multi-modal dense video captioning	Bi-modal transformer with a bi-modal multi-headed proposal generation module is proposed demonstrating use of audio and visual features while performing dense video captioning.	The dense video captioning system, audio-only model mostly get the signal of "talking", needs more attention.
19	Kitaev et al. (2020)	2020	Reformer	Reformer combines the modeling capacity of a Transformer with an architecture that can be executed efficiently on long sequences and with small memory use even for models with a large number of layers.	The computational cost of a model grows with the number of hashes, so hashing hyperparameter can be adjusted depending on the available compute budget.
20	Dai et al. (2020)	2020	Transformer XL	Extra long transformer is proposed to address the limitation of fixed length context with the notion of recurrence into deep self attention network. The hidden states achieved in previous segments are utilized as memory for current segments resulting in recurrent connections between the segments. The authors also proposed simple yet effective positional encoding formulation.	NA

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
21	Luo et al. (2020)	2020	UniVL: Unified video and language pre-training model for multi-modal understanding and generation	Proposed a multi-modal video-language pre-training model trained on a large-scale instructional video dataset. Model is designed with four modules and five objectives capable of video-language understanding and generative tasks and learning joint representation of video and language and adapt down-stream multi-modal tasks.	Joint loss decreases the generation task a little, although it performs well in the retrieval task. Excessive emphasis on coarse-grained matching can affect the fine-grained description at the generation task.
22	Pan et al. (2020a)	2020	Auto-captions on GIF: A large-scale video-sentence dataset for vision-language pre-training	Large-scale automatically generated pre-training dataset for generic video understanding. Designed Transformer-based Encoder–Decoder structure for vision-language pre-training in video domain.	NA
23	Bilkhu et al. (2019)	2019	Universal transformer	A generalized form of the standard transformer, is a parallel-in-time recurrent self-attentive sequence model based on the ED architecture, and employs an RNN for representations of every position in both input and output sequences. The recurrence is over the depth, not over the position in the sequence. the depth of UT is the main difference between the standard transformer and the UT.	NA
24	Sun et al. (2019b)	2019	VideoBERT: A Joint Model for Video and Language Representation Learning	Method to learn high level video representations that capture semantically meaningful and temporally long-range structure.	Explicitly model visual patterns at multiple temporal scales, instead of the proposed approach, that skips frames but builds a single vocabulary.

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
25	Lu et al. (2019)	2019	ViLBERT: Vision & language BERT	Extending the popular BERT, the authors developed a model and proxy tasks for learning joint visual-linguistic representations. Two-stream architecture with co-attentional transformer blocks that outperforms sensible ablations and exceeds state-of-the-art when transferred to multiple established vision-and-language tasks.	Considering training, language often only identifies high-level semantics of visual content and is unlikely to be able to reconstruct exact image features. Further, applying a regression loss could make it difficult to balance losses incurred by masked image and text inputs.
26	Child et al. (2019)	2019	Sparse transformers	introduced several sparse factorization of the attention matrix, as well as reconstructed residual blocks, weight initialization for training enhancement of deeper networks, and a reduction in memory usage. Unlike a transformer, where training with many layers is difficult, the sparse transformer facilitates hundreds of layers by using the pre-activation residual block. Instead of positional encoding, learned embedding is useful and efficient.	NA
27	Li and Qiu (2020)	2019	LSTM vs standard transformer	Authors explored attention over space and time with features extracted from Pseudo-3D ResidualNet and compared neural network architectures using temporal attention over 2D-CNN features. Also explored the performance of LSTM vs transformer for video captioning.	Hyperparameter tuning while training model is required. Spatio-temporal attention over P3D features did not improve performance.

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
28	Yang et al. (2019)	2019	NAVC (standard transformer)	Developed a non-autoregressive video captioning model (NAVC) with iterative refinement. Also exploited external auxiliary scoring information to assist the NAVC in precisely focusing on those inappropriate words during iterative refinement. The captioning decoder is capable of predicting target words along with the parallel generation of captions.	Explore an internal auxiliary scoring module to get rid of external constraints.
29	Zhou et al. (2018)	2018	Masked transformer	End to end non-efficient transformer based model employing masking network to restrict attention to the proposal event over the encoding features. The proposed model employs a self attention mechanism.	Small objects, such as utensils and ingredients, are hard to detect using global visual features but are crucial for describing a recipe.
30	Chen et al. (2018)	2018	Two-view transformer	TVT comprises of a backbone of Transformer network for sequential representation and two types of fusion blocks in decoder layers for combining different modalities effectively allowing parallel computing.	Other modalities can be incorporated in the TVT framework for better video captioning.

Table 5 (continued)

S/N	References	Year	Approach	Contributions	Shortcomings
31	Vaswani et al. (2017)	2017	Standard transformer	A simple transduction architecture for sequence modeling entirely based on an attention mechanism with the objectives of parallelization and a reduction in sequential computation. The commonly used recurrent layers in the ED architecture are replaced with multi-head self-attention layers where self attention is about computing the sequence representation by relating different positions or parts of it.	NA

Brief contributions and shortcomings (if any) are also mentioned with each approach

NA Not available, *S/N* serial number

decoding the event proposals into a differentiable mask to restrict attention to the proposed event. Both decoders learn during training to adjust for the best caption generation. Zhou et al. (2018) proposed a differentiable masking scheme by confirming training stability between proposals and captioning decoders. A standard transformer is employed for both encoder and decoder because of its fast self-attention mechanism implemented for accurate and useful performance.

3.5.4 Two-view transformer

Two-view Transformer (TvT) is a video captioning technique derived from the standard transformer and accompanied by two fusion blocks in the decoder layer to combine different modalities effectively. Parallelization, the primary quality of a transformer, leads to efficient and robust training activity, and instead of simple concatenation, two types of fusion blocks are proposed to explore information from frame features, motions, and previously generated words.

TvT (Chen et al. 2018) contains two views of visual representations extracted by the encoder block, i.e., a frame representation obtained using a 2D-CNN (ResNet-152 and NasNet pre-trained on ImageNet) on every frame individually, and motion representation is obtained by employing a 3D-CNN (I3D pre-trained on Kinetics) on connecting frames.

The decoder block contains two types of fusion block: add-fusion and attentive-fusion. The add-fusion block simply combines the frame and motion representation with a fixed weight between 0 and 1. The attentive-fusion block combines the two representations in a learnable way such that these two representations, with previously generated words, can jointly guide the model to accurately generate a description.

3.5.5 Bidirectional transformer

Bidirectional Encoder Representations from Transformers (BERT) (Kenton et al. 1953; Sun et al. 2019b) a conceptually simple yet powerful fine-tuning-based bidirectional language representation model, is the state of the art for several NLP-specific tasks. BERT uses a bidirectional self-attention mechanism to carry out the tasks of masked language modeling and next-sentence prediction. VideoBERT (Sun et al. 2019b) (based on BERT) was proposed basically for text-to-video generation or future prediction, and can be utilized for automatic illustration of instructional videos, such as recipes. VideoBERT is also applied to the task of video captioning following the masked transformer (Zhou et al. 2018) with a transformer ED, but the inputs to the encoder are replaced with features extracted by VideoBERT. VideoBERT reliably outpaces the S3D baseline (Xie et al. 2018), particularly with the CIDEr score. Furthermore, by combining VideoBERT and S3D, the proposed model demonstrated outstanding performance for all metrics. VideoBERT is capable of learning high-level semantic representations, and hence, achieved substantially better results on the YouCookII dataset. Vision & Language BERT (ViLBERT) (Lu et al. 2019) extended BERT to jointly represent text and images, and consists of two parallel streams (visual processing and linguistic processing) interacting through co-attentional transformer layers. The proposed ViLBERT with co-attentional transformer blocks outperformed the ablations and surpassed state-of-the-art models when transferred to multiple established vision-and-language tasks, e.g., visual question answering (VQA) (Antol et al. 2015), visual common sense reasoning (VCR) (Zellers et al. 2019), ground-referring expressions (Kazemzadeh et al. 2014), and caption-based image retrieval (Young et al. 2014).

3.5.6 Sparse transformer

Even with the transformers, the processing of lengthy sequences demands more time and memory, resulting in poor performance and inefficient systems. Sparse transformers (Child et al. 2019) introduced several sparse factorizations of the attention matrix, as well as restructured residual blocks, weight initialization for training enhancement of deeper networks, and a reduction in memory usage. Unlike a transformer, where training with many layers is difficult, the sparse transformer facilitates hundreds of layers by using the pre-activation residual block. Instead of positional encoding, learned embedding is useful and efficient. Gradient checkpoints are incorporated for effective reductions in memory requirements to train deep neural networks. Dropout is applied once, at the end of the residual attention instead of within the residual block. Experimentation with a sparse transformer demonstrated better performance on long-sequence modeling, with less computational complexity.

3.5.7 Reformer (the efficient transformer)

To improve the efficiency of the transformer on long sequences, Reformer (Kitaev et al. 2020) was proposed with reduced complexity and reversible residual layers (Gomez et al. 2017) for storing single-time activations during training. Inside the FFN layer, the activations are split and processed in chunks to save memory inside the FFN. Inclusion of locality sensitive hashing (LSH) in attention, depending on the total number of hashes employed, influences training aspects a lot. It was observed that regular attention is slower for lengthy sequences, but LSH attention speed remains smooth. Experimentation performed on text- and image-generation tasks produced the same results as the standard transformer but with more speed and efficient memory usage.

3.5.8 Transformer-XL

Transformer-XL (Dai et al. 2020) is based on the standard transformer architecture, and deals with better learning of long-range dependencies. Its key technological contributions include the concept of recurrence in a totally self-attentive model and developing an exclusive positional encoding scheme. It introduces a simple but more effective relative positional encoding design that generalizes attention lengths longer than the ones observed during training. For both character-level and word-level modeling, Transformer-XL is the first self-attention model that accomplishes significantly improved results compared to RNNs. Evaluating speed in comparison to the 64-layer standard transformer proposed in Al-Rfou et al. (2019), Transformer-XL achieved speeds up to 1,874 times faster.

3.6 Discussion—transformer based approaches

In the wake of Vaswani et al. (2017) successful implementation of the transformer in natural language processing, the transformer has become increasingly popular in a wide range of fields, including computer vision and speech analysis. Transformers have recently been improved in several variants compared to the vanilla model from the perspective of generalization, parallelization, adaptation, and efficiency. Its first

application in the field of NLP for translation(Vaswani et al. 2017) initiated the tech journey. Recently with the robust representation competences it is proving its worth in the computer vision domain. Particular to transformer for video description (Zhou et al. 2018) introduced the first video paragraph captioning model using masked transformer. Due to the sequential nature of captioning task, unlike RNN which unroll sequence one step at a time, transformers can perform parallel processing of the entire sequence at both ends resulting in efficient and accurate captioning. The transformer enhanced with an external memory block further facilitates history maintenance of the visual & language information and augmentation of the current segment. Dependency among different sequence segments is learned through the self-attention mechanism inside the transformers. Considering long-range dependencies, hard to resolve for RNNs in the case of more extensive sequences is no longer an issue with the use of transformers. The vision transformer (ViT) (Hussain et al. 2022) recognized human activities in surveillance videos and adopted CNN free approach and capture long range dependencies in time to accurately encode relative spatial information. Likewise, video vision transformer (ViViT) (Arnab et al. 2021) factorized the spatial and temporal dimensions of the input video to handle long sequences of tokens encountered in video. Models employing modern transformers demonstrated comparable results handling long-range dependencies on vido description task, still developing efficient transformer models for computer vision's tasks remains an open problem. Transformer models are usually huge and computationally expensive. In spite of their success in various applications, transformer models require a high amount of computing and memory resources, which limits their use on resources-constrained devices such as mobile phones (Han et al. 2022). So to cater resource-limited devices, research in designing efficient transformer models for visio-linguistic tasks need attention.

3.7 Deep reinforcement learning (DRL)

Trial and error, or experience and learn from experience, is the core of reinforcement learning (RL). It is all about taking appropriate actions in a certain environment and accommodating the reward/penalty by following a policy. Deep RL approaches have shown efficient performance in the field of real-world games. Particularly in 2013, Google DeepMind (Mnih et al. 2013, 2015) took the initiative and demonstrated that a single architecture could successfully learn control policies in a range of different environments with minimal prior knowledge. It showed successful integration of RL with deep network architectures. Although many adversities exist for DRL models, compared to conventional learning, even then, DRL has shown extraordinarily proficient performance in the field of captioning. Optimization of evaluation metrics, considered for the reward function, for increasing the generated caption readability, and for training stability and system convergence, is kept under consideration while employing DRL for descriptions. Figure 9 shows the RL agent-environment interaction for video descriptions, and some of the famous DRL approaches and their components (agent, action, environment, reward, and goal) are summarized in Table 6 for a quick view.

An efficacious combination of RL (He et al. 2019) with supervised learning was presented in a multi-task learning framework. The goal of the system is to learn the policy to correctly ground the specific descriptions in the video. Hence, as a reward, the model encourages the agent to better match clips gradually, which is carried out by helping the agent get precise information from the environment, and to maximize the reward by

Table 6 DRL models with their system components

References	Year	Approach	DRL System Components				
			Agent	Action	Environment	Reward	Goal
Li and Qiu (2020)	2019	End-to-end video captioning with multitask RL	Captioning model	Predict next word	Input video with user-annotated captions	CIDEr score	Generate a proper sentence after observing input vide
He et al. (2019)	2019	Read, watch, and move: RL for temporally grounding natural language descriptions in videos	RL (actor-critic-based model)	One of 7 ways to adjust temporal boundaries	Video, the description, temporal grounding boundaries	1, if temporal IoU is within a certain threshold; 0, otherwise	Extraction of well-matched video clips w.r.t. the provided query
Zhang et al. (2019c)	2019	Reconstruction Network (RecNet)	ED model (RecNet)	Predict next word	Video content	ground truth words	CIDEr score Caption generation by metric optimization
Wang et al. (2018b)	2018	Video captioning via hierarchical RL	Manager + Worker + Critic	Selection of words from a dictionary	Textual and video context	Delta CIDEr	Maximize discounted return Rt
Chen et al. (2018a)	2018	Less is more–picking informative frames for video captioning (PickNet)	PickNet Model	Frame picking	ED architecture	Sum of language reward and visual diversity reward	Select informative frames for the task of video captioning
Pasunuru and Bansal (2017)	2017	Reinforced video captioning with entailment rewards	Baseline model	Word generation	Video and caption	CIDEnt (entailment corrected reward)	Minimize the negative expected reward
Ren et al. (2017)	2017	Deep RL-based image captioning with embedding reward	Policy network + value network	Predict next word	Input image and predicted words	Visual-semantic embedding similarities	Generate caption similar to ground truth
Phan et al. (2017)	2017	Consensus-based Sequence Training (CST) for video captioning	LSTM language model	Predict next word	Image/video features and words	CIDEr	Generate captions similar to reference captions

Table 6 (continued)

References	Year	Approach	DRL System Components				
			Agent	Action	Environment	Reward	Goal
Rennie et al. (2017)	2017	Self-critical sequence training (SCST) for image captioning	LSTM	Next word prediction	Words and image features	CIDEr	Generate image captions similar to reference captions
Ranzato et al. (2016)	2016	Sequence-level training with recurrent neural network (MIXER)	Generative model, RNN	Next word prediction	Context vector and words	BLEU, ROUGE-2	Sequence generation

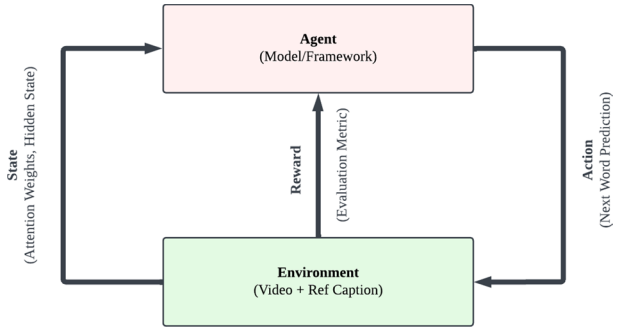

Fig. 9 DRL agent-environment interaction for video description

exploring or exploiting the whole environment, forming a sequential decision case. The actor-critic algorithm (Montague 1999) is employed to generate policy and take appropriate action. The agent is responsible for the iterative adjustment of temporal boundaries until specified conditions are met. After an action is accomplished, the environment (a combination of video, description, and temporally grounded boundaries) is modified accordingly. State vectors combine a description with global, local, and location features, which are then fed into a GRU-FC-based actor-critic module for policy and state-value learning. A penalty mechanism is also defined to keep computational costs within limits. As the name indicates, the agent (model) reads the description, watches the video and localization, and after that iteratively moves the temporal grounding boundaries for best clip matching, according to the description.

The ED architecture intrinsically obstructs the use of end-to-end training because of lengthy sequences in both input and output for the model. Therefore, a multi-task RL model (Li and Qiu 2020) to avoid over-fitting was proposed for end-to-end training. The primary job of the model is to mine or extract as many tasks as possible from human-annotated videos, which can regulate the search space of the ED network. After that, end-to-end training is carried out for video captioning. The auxiliary assignment of the model is to predict the characteristics mined from reference captions and, based on these predictions, maximize the reward defined in the RL system. Specific to RL, the objective of the model is to train an agent to accomplish the tasks in an environment by performing a sequence of actions. For video captioning, the model aims to automatically generate a precise and meaningful sentence after processing the provided video. The agent's action is to predict the next word in the sequence at each time step. The model's reward is defined as the evaluation metric used in the test phase. The CIDEr score functions as a reward signal. Finally, evaluation of multi-task training revealed that domain-specific video representation is more influential than generic image features.

Sequence-to-sequence models optimize word-level cross-entropy loss during training, whereas the video captioning model proposed in Pasunuru and Bansal (2017) optimizes sentence-level, task-based metrics using policy gradients and mixed loss methods for RL. Moreover, an entailment enhanced reward, CIDEnt, was proposed that adjusts phrase-matching-based metrics and, on achieving a low entailment score, penalizes the phrase-matching metric (CIDEr-based) reward. An automatically generated caption gets a high entailment score only when the generated caption has logical matching with the ground truth annotation, instead of word matching.

Cross-entropy loss and reward-based loss are combined as a mixed loss to maintain output fluency and resolve the exposure bias issue. At first, the CIDEr reward

demonstrated significant improvement, and after that, the CIDEnt reward further enhanced system performance.

Most of the captioning systems are trained by maximizing the maximum likelihood estimation (MLE), i.e., the similarity between the generated and reference captions, or by minimizing the cross-entropy (XE) loss. However, the MLE/XE approach suffers from two inadequacies: objective mismatch and exposure bias. Recent research demonstrated that for the captioning task, evaluation metrics could be optimized directly using RL, keeping in mind the associated computational cost and the designed reward for system convergence. The Self-Consensus Baseline (SCB) model proposed in Phan et al. (2017) trains concurrently on multiple descriptions of the same video, and employs human-annotated captions as a baseline for reward calculation instead of creating a new baseline for each generated caption. Following the ED approach, as an encoder, ResNet is used for static image features, C3D is used for short-term motion, and MFCC for acoustic features; GloVe is for word embedding, and LSTM is the decoder employed for language generation. Taking the LSTM language model as an agent in the environment of video features and words, the action is to predict the next word, attaining the dual goal of coming up with an accurate textual alternate of the given video and minimizing the negative expected reward of the model. Compared to the MIXER approach (Ranzato et al. 2016), where RL training gradually mixes into XE training to stabilize the learning, CBT trains both RL and XE simultaneously. A connection between RL and XE training is established in this research, utilizing consensus among multiple reference captions for training improvement, objective mismatch, and exposure bias elimination.

A fully differentiable deep neural network comprising a higher- and a lower-level sequence model, was proposed in Wang et al. (2018b) for video description. Employing a hierarchical-RL (HRL) framework, the agent, environment, action, reward, and goal are defined to efficiently learn semantic dynamics adopting the ED architecture. Each video is sampled at 3fps, extracting ResNet-152 (Zhang et al. 2017) features from the sampled frames. The extracted features are fed into a two-phased encoder, i.e., a low-level bidirectional LSTM (Schuster and Paliwal 1997) and a high-level LSTM (Cascade-correlation and Chunking 1997). In the decoding phase, the HRL agent resembles the decoder. To better capture the temporal dynamics, an attention strategy is employed. The HRL agent is composed of three components: (1) a low-level worker that selects certain actions in each time step to achieve the goal, (2) a high-level manager that set goals, and 3) an internal critic (an RNN structure) to ensure accomplishment of the task and serve the manager accordingly. Both worker and manager are accommodated with the attention mechanism. A strong convergence policy is a challenging area in RL implementation, and the proposed HRL model achieved high convergence by applying cross-entropy loss optimization. The model is able to capture in-depth details of the video content, and can generate more detailed and accurate descriptions.

To avoid redundant visual processing and to lower computational costs, a plug-and-play PickNet model (Chen et al. 2018a) was proposed to perform informative frame selection. The solution comprises two parts; first is PickNet for efficient frame selection, and second is a standard encoder (LSTM) and decoder (GRU) architecture for caption generation. RL-based PickNet selects the informative frames without having full details on the environment, i.e., it makes decisions to pick or drop a frame only on the basis of the current state and the history. The agent selects a subset of frames retaining the maximum visual content (i.e., six to eight frames selected, on average, from a video) while other models commonly need up to 40 frames for analysis. Following flexibility, efficiency, and effectiveness, the selected keyframes are capable of increasing visual diversity and decreasing

textual inconsistency. Visual diversity and language rewards are defined. A negative reward is defined to discourage the selection of too many (or too few) frames. Model training is performed in three phases. First is the supervision phase, where the ED is pre-trained. Second is the reinforcement phase, where PickNet is trained by employing RL. Third is the adaptation phase in which both PickNet and the ED are jointly trained.

3.8 Discussion—deep reinforcement learning (DRL)

In recent years, the Encoder–Decoder structure demonstrated promising results fusing attention and transformer mechanisms. However, due to the long range dependencies handling and semantic gap between the visual and language domain, the generated descriptions contain numerous inaccuracies. These errors can be handled by adopting optimization through deep reinforcement learning. The polishing network (Xu et al. 2021) follows human proofreading mechanism by evaluating and improving the generated captions gradually for revise word errors and grammatical errors. Evaluation mtric selection as reward function also plays role in robust performance. CIDEr score is a choice in most articles. The deep reinforcement learning framework includes an environment, agent, action, reward and goal function. For video captioning, the goal is to generate accurate description aligned with the visual information of the video. The generative language model acts as an agent and takes an action of next word prediction. The provided video and the ground truth descriptions plays the environment role resulting in rewarding the selected evaluation metric on successful word generation or penalize the metric score otherwise. The environment upates the state of attention weights or hidden states based on the employed mechanism. This cycle of agent's action and environment's state and reward update continues to gradually improve the generated description as shown in Figure 9. There has been a growing interest in DRL and hierarchical RL based methods in recent years, which have shown comparable results in the video description.

4 Results comparison & discussion

The benchmark results generated by various models in the recent past are discussed in this section. Dataset-based segregated techniques are further categorized in chronological order according to the approach/mechanism adopted for experimentation.

Video description models are mostly evaluated on MSVD (Chen and Dolan 2011) and MSR-VTT (Xu et al. 2016) datasets because of the wide-ranging and diverse nature of the videos, the availability of multiple ground truth captions for model training and evaluation, and most importantly, task specificity. For models having multiple variants during experimentation, the best performing variant is reported here. Scores shown in bold were the best performing.

4.1 Evaluation metrics

Most of the metrics commonly used for automatically generated captions evaluation, namely, BLEU@(1,2,3,4), METEOR, ROUGE, and WMD, are from the NLP domain, namely, NMT , and document summarization. CIDEr and SPICE evolved as a result of the increased demand for task-specific (captioning) metrics. It is essential for the description to possess the qualities of acceptability, consistency, and expression-fluency, particularly

when considering the evaluations made by humans (Sharif et al. 2018). The evaluation metric is considered best when it exhibits a significant correlation with the human scores (Zhang and Vogel 2010). A short description of the metrics mostly used to evaluate the automatically generated description is given below, For detailed computational concept along with the limitations, please refer to Rafiq et al. (2021).

4.1.1 BLEU

(Bi-Lingual Evaluation Understudy): The evaluation metric proposed by Doddington (2002) measures the numerical proximity between generated captions and their referenced counterparts. It Computes the unigram (overlap of single word) or n-gram(overlap of adjacent n words) between the two texts, i.e., referenced and generated. Multiple reference annotations for a single video can guarantee good BLEU score. The basis for this metric is the precision measure which is the main limitation of this metric. The research work by Lavie et al. (2004) demonstrated that considerably high correlation can be achieved by emphasizing more on recall measure than on the precision score.

4.1.2 METEOR

Metric for Evaluation of Translation with Explicit ORdering In order to ensure the accuracy of this metric (Lavie and Agarwal 2007), an explicit exact word match must be made between the predicted translation and one or more reference annotations. It supports the matching of identical words, synonyms, words with identical stem and also the order of words in referenced and predicted sentences. The computational procedure is based on the harmonic mean of precision and recall of uni-gram matches between the sentences (Kilickaya et al. 2017). Moreover, METEOR score is more closely correlated with human judgment (Elliott and Keller 2014).

4.1.3 ROUGE

Recall-Oriented Understudy for Gisting Evaluation This metric (Lin 2004) belongs to the NLP domain (documents summaries) evaluation metrics family. There are multiple variants in Rouge which are used to determine how closely the generated and reference summaries are comparable. Among these variants, Rouge-N (n-gram Co-occurrence), and Rouge-L (Longest Common Sub-sequence) are related to image and video captioning evaluation. In terms of Rouge-N, it is the n-gram recall between the predicted summary and one or more reference summaries. In contrast, Rouge-L uses a similarity score based on the recall and precision of the longest common sub-sequence between the generated and the reference sentences.

4.1.4 CIDEr

Consensus-Based Image Description Evaluation An image description evaluation metric based on human consensus is proposed by CIDEr (Vedantam et al. 2015). When comparing a generated sentence with the set of reference human annotations provided for an image, the CIDEr understands the underlying concepts of prominence, accuracy, and linguistics. Computational concept involves the cosine similarities between the referenced and generated captions for a provided image. The CIDEr-D variant of CIDEr is famous

for image and video description evaluation. Where verb stem removal in the basic CIDEr metric ensured the usage of correct form of verb and exhibited high spearman's rank correlation with respect to original CIDEr score.

All the evaluation metrics follow the strategy of *the higher, the better*, higher scores are considered better for BLEU, METEOR, ROUGE, and CIDEr. For the models computing BLEU@1, BLEU@2, BLEU@3, and BLEU@4, only BLEU@4 is reported here because of characteristics analogous to human annotations.

4.2 Datasets for evaluation

Defining a dataset as a collection of video clips with their respective annotations or descriptions is the act of creating a basis for training, validating, and testing a model. Among the domain-specific datasets are those relating to cooking, movies, social media, wild and human actions. In contrast, a wide variety of videos can be found in open-domain datasets. Following is a brief description of the most widely used benchmark datasets used in recent research for video descriptions.

4.2.1 MSVD—the microsoft video description dataset

Table 7 summarizes the results from popular models using the MSVD dataset. MSVD (Chen and Dolan 2011) is one of the earlier available corpora frequently used by the research community around the globe. It is a collection of 1,970 YouTube video clips provided with human annotations. The collection of these clips was carried out by requesting them from Amazon Mechanical Turk (AMT) workers. They were guided to pick short snippets depicting a single activity and were asked to mute the audio. Each video clip is 10 to 25 seconds long, on average. Afterward, these snippets were labeled with multilingual, mono-sentence captions provided by annotators. Frequently used slices of the dataset for training, validation, and testing comprise 1,200, 100, and 670 video clips. Figure 10 shows histograms for BLEU, ROUGE-L, METEOR, and CIDEr scores employing standard Encoder–Decoder structures and evaluated on MSVD and MSR-VTT datasets. Figure 11 demonstrates performance evaluation of transformer based models on MSVD and MSR-VTT datasets. Figure 12 graphically explains the performance evaluation of DRL based methods employing MSVD and MSR-VTT datasets. Figure 13 depicts the results obtained employing attention based approaches and evaluated on MSVD and MER-VTT datasets.

Considering the standard ED mechanism, benchmarking with MSVD revealed that SeFLA Lee and Kim (2018), a semantic feature learning-based caption generation model, showed a better BLEU score, and VNS-GRU (Chen et al. 2020) achieved best performance results from METEOR, ROUGE, and CIDEr scoring. Advancements in the field of neural machine learning have demonstrated encouraging improvements on the video description task, but models trained using word-level losses cannot correlate well with sentence-level metrics, although all the evaluation metrics are sentence-level. So, metric optimization is critically needed for high-quality caption generation. Deep reinforcement learning is employed for optimization of many techniques. DRL approaches evaluated on the MSVD dataset concluded with the best performance from Pasunuru and Bansal (2017) in all metrics (BLEU, METEOR, ROUGE, and CIDEr).

Models employing a transformer mechanism are progressing at a good pace. Among the transformer-based models, Two-view Transformer (Chen et al. 2018) performed the best in BLEU scoring, whereas Non-Autoregressive Video Captioning with Iterative

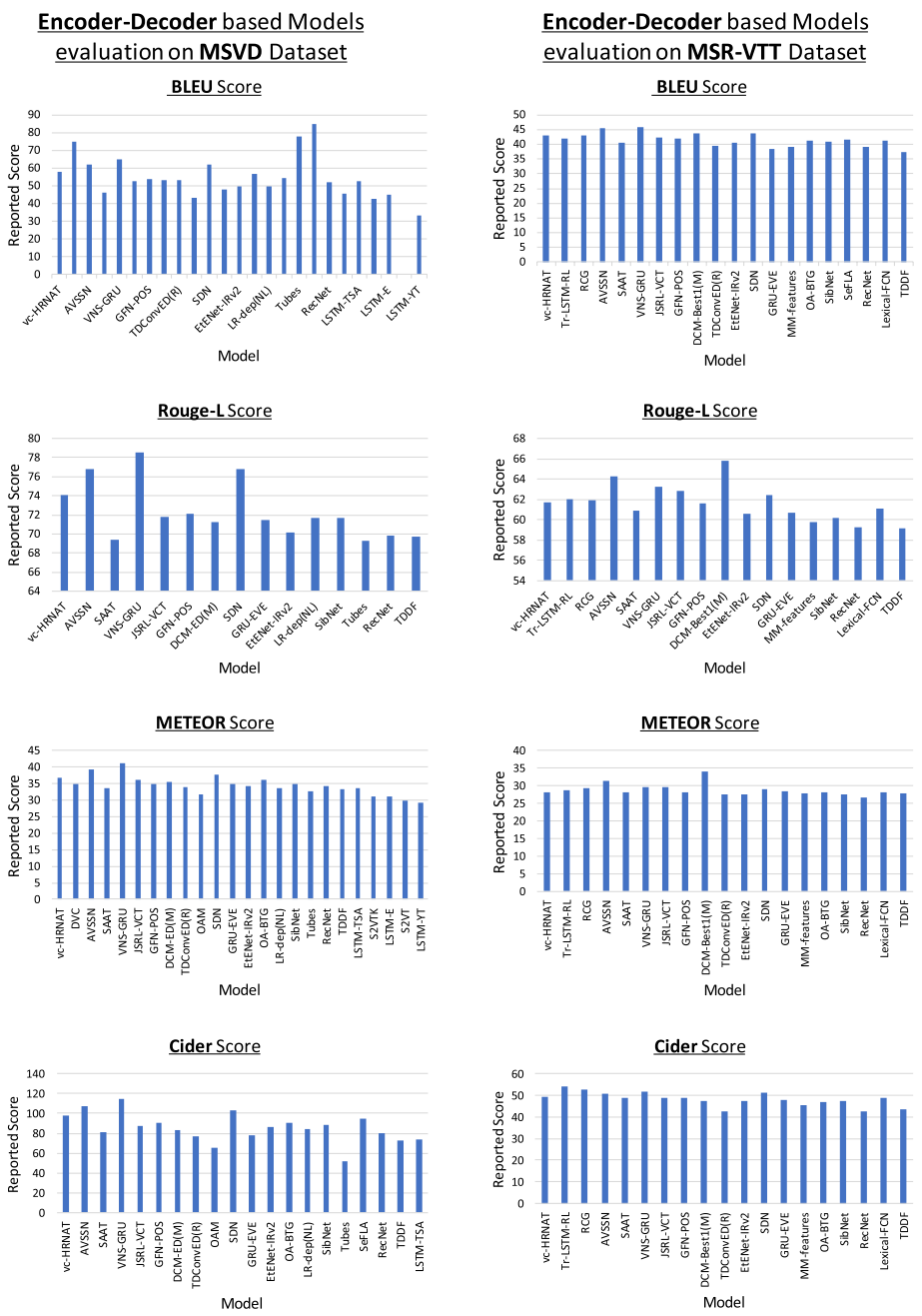

Fig. 10 Performance evaluation of BLEU, METEOR, ROUGE-L and CIDEr on the MSVD & MSR-VTT dataset for standard encoder–decoder approach

Table 7 Video description performance evaluation on the MSVD dataset for all four approaches

References	Model	Year	B	M	R	C
A. Encoder-decoder based approaches						
Gao et al. (2022)	vc-HRNAT	2022	57.7	36.8	74.1	98.1
Madake (2022)	DVC	2022	75	34.7	–	–
Perez-Martin et al. (2021a)	AVSSN	2021	62.3	39.2	76.8	107.7
Zheng et al. (2020)	SAAT	2020	46.5	33.5	69.4	81.0
Chen et al. (2020)	VNS-GRU	2020	64.9	41.1	78.5	**115**
Hou et al. (2019)	JSRL-VCT	2019	52.8	36.1	71.8	87.8
Wang et al. (2019a)	GFN-POS	2019	53.9	34.9	72.1	91.0
Xiao and Shi (2019z)	DCM-ED(M)	2019	53.3	35.6	71.2	83.1
Chen et al. (2019b)	TDConvED(R)	2019	53.3	33.8	–	76.4
Babariya and Tamaki (2020)	OAM	2019	43.5	31.6	–	64.9
Chen et al. (2019a)	SDN	2019	61.8	37.8	76.8	103
Aafaq et al. (2019a)	GRU-EVE	2019	47.9	35.0	71.5	78.1
Olivastri (2019)	EtENet-IRv2	2019	50.0	34.3	70.2	86.6
Zhang et al. (2019a)	OA-BTG	2019	56.9	36.2	–	90.6
Lee et al. (2019)	LR-dep(NL)	2019	49.7	33.7	71.7	84.5
Liu et al. (2020)	SibNet	2018	54.2	34.8	71.7	88.2
Zhao et al. (2018)	Tubes	2018	77.6	32.6	69.3	52.2
Lee and Kim (2018)	SeFLA	2018	**84.8**	–	–	94.3
Wang et al. (2018a)	RecNet	2018	52.3	34.1	69.8	80.3
Zhang et al. (2017)	TDDF	2017	45.8	33.3	69.7	73.0
Pan et al. (2017)	LSTM-TSA	2017	52.8	33.5	–	74.0
Wang and Song (2017)	S2VTK	2017	42.5	31.0	–	–
Pan et al. (2016)	LSTM-E	2016	45.3	31.0	–	–
Venugopalan et al. (2015)	S2VT	2015	–	29.8	–	–
Lowell et al. (2014)	LSTM-YT	2014	33.29	29.07	–	–
B. DRL approaches						
Li and Qiu (2020)	Multi-task RL	2019	50.3	34.1	70.8	87.5
Chen et al. (2018a)	PickNet	2018	49.9	33.1	69.3	76
Pasunuru and Bansal (2017)	CIDEnt	2017	54.4	34.9	72.2	88.6
C. Transformer-based approaches						
Im and Choi (2022)	UAT-FEGs	2022	56.5	36.4	72.8	92.8
Liu et al. (2021)	O2NA	2021	55.4	37.4	74.5	96.4
Jin et al. (2020)	SBAT	2020	53.5	35.3	72.3	89.5
Li and Qiu (2020)	2D-CNN+Tr	2019	40.8	–	–	–
Li and Qiu (2020)	P3D+CNN	2019	35.4	–	–	–
Yang et al. (2019)	NRVC	2019	53.1	35.5	-	89.4
Bilkhu et al. (2019)	I3D+UT	2019	46.0	-	–	–
Chen et al. (2018)	TVT	2018	53.9	35.2	72.0	86.7
D. Attention-based approaches						
Ji et al. (2022)	ADL	2022	54.1	35.7	70.4	81.6
Peng et al. (2021)	T-DL	2021	55.1	36.4	72.2	85.7
Ryu et al. (2021)	SGN	2021	52.8	35.5	72.9	94.3
Perez-Martin et al. (2021b)	SemSynAN	2021	64.4	**41.9**	**79.5**	111.5

Table 7 (continued)

References	Model	Year	B	M	R	C
Zhang et al. (2020)	ORG-TRL	2020	54.3	36.4	73.9	95.2
Yan et al. (2020)	STAT	2020	52.0	33.3	–	73.8
Bin et al. (2019)	BiLTSM	2019	37.3	30.3	–	–
Xiao and Shi (2019b)	Attrib_Sel	2019	56.5	35.4	–	86.1
Sun et al. (2019b)	MSAN	2019	56.4	35.3	–	79.6
Li et al. (2019b)	Res-ATT	2019	53.4	34.3	–	72.9
Gao et al. (2019)	hLSTMat	2019	54.3	33.6	–	73.8
Chen et al. (2018b)	SSTA-R	2018	45.3	30.3	–	59.2
Gao et al. (2017)	aLSTMs	2017	50.8	33.3	-	74.8
Li et al. (2017)	MAM-RNN	2017	41.3	32.9	68.8	53.9
Xu et al. (2017)	MA-LSTM	2017	52.3	33.6	–	70.4
Hori et al. (2017)	A.F	2017	53.9	32.2	–	68.8
Chen and Jiang (2019)	MGSA	2017	53.4	35.0	–	86.7
Laokulrat et al. (2016)	S2S-TA	2016	43.7	32.6	68.1	75.0

For all evaluation metrics, mechanism-wise highest scores are underlined, and the overall highest score (across all four mechanisms) is in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr, NL: Non-Local, A.F: Attentional Fusion

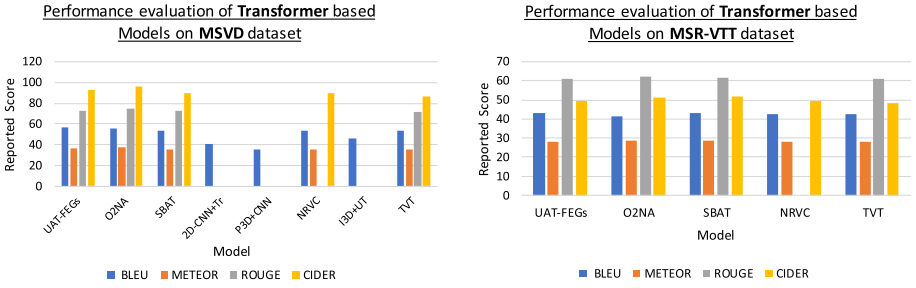

Fig. 11 Performance evaluation of BLEU, METEOR, ROUGE-L and CIDEr on the MSVD & MSR-VTT dataset for transformer mechanism based approach

Fig. 12 Performance evaluation of BLEU, METEOR, ROUGE-L and CIDEr on the MSVD & MSR-VTT dataset for reinforcement learning based approach

Refinement (Yang et al. 2019) performed excellently under METEOR scoring, and the recently proposed SBAT (Jin et al. 2020) outperformed all previous models based on the ROUGE and CIDEr metrics. For attention-based approaches, SemSynAN (Perez-Martin et al. 2021b) outperformed the existing methods based on BLEU@4, METEOR, ROUGE, and CIDEr scores with the MSVD dataset.

For the overall performance evaluation from all four mechanisms on the MSVD dataset, SeFLA (Lee and Kim 2018), a semantic feature learning-based caption generation model, demonstrated an excellent BLEU score; SemSynAN (Perez-Martin et al. 2021b) produced the top METEOR and ROUGE scores, and VNS-GRU (Chen et al. 2020) achieved the best CIDEr score. It is clear from Table 7 that for short video clips comprising a single activity, the standard ED mechanism and the attention-based mechanism achieved top results.

4.2.2 MSR-VTT—microsoft research video to text

Table 8 demonstrates the results reported from using the MSR-VTT dataset (Xu et al. 2016), which is an open-domain, large-scale benchmark with 20 broad categories and diverse video content bridging vision and language. It comprises 10,000 clips that originated from 7180 videos. Being open-domain, it includes videos from categories like music, people, gaming, sports, news, education, vehicles, beauty, and advertisement. The duration of each clip, on average, is 10–30 seconds resulting in a total 41.2 h of video. To provide good semantics from the clips, 1327 AMT workers were engaged to annotate each one with 20 natural sentences. Data were split in Xu et al. (2016), suggesting 6513 videos for training, 497 videos for validation, and 2990 videos for testing purposes.

Considering the standard ED mechanism, benchmarking on the MSR-VTT dataset demonstrated that VNS-GRU (Chen et al. 2020), a variational-normalized semantic GRU-based caption generation model, showed better BLEU and CIDEr scores; DCM (Xiao and Shi 2019z), a diverse captioning model with a conditional GAN, achieved the best performance from the METEOR and ROUGE metrics. Among the DRL-based methods, Consensus-based Sequence Training (CST) (Phan et al. 2017) was trained concurrently on multiple descriptions of the same video. It employed human-annotated captions as a baseline for reward calculation, instead of creating a new baseline for each generated caption resulting in directly optimizing the evaluation metrics. Using DRL performed well based on BLEU, METEOR, ROUGE, and CIDEr metrics with MSR-VTT. Approaches based on a transformer mechanism demonstrated that the recently proposed SBAT (Jin et al. 2020) outperformed all previous models for all four metrics. In the attention-based approaches with the MSR-VTT dataset, the recently proposed SemSynAN (Perez-Martin et al. 2021b) outperformed the existing methods based on the METEOR and ROUGE metrics, whereas MSAN (Sun et al. 2019b), a multi-modal semantic attention network, performed excellently based on BLEU and CIDEr.

Considering the overall performance evaluations from all four mechanisms with the MSR-VTT dataset, MSAN (Sun et al. 2019b) demonstrated an excellent BLEU score, DCM (Xiao and Shi 2019z) (a diverse captioning model with a conditional GAN) achieved the best results for METEOR and ROUGE metrics, and the DRL-based CST model (Phan et al. 2017) achieved the best score from CIDEr.

Table 8 Video description performance on the MSR-VTT dataset for all four approaches

References	Model	Year	B	M	R	C
A. Encoder–decoder based approaches						
Gao et al. (2022)	vc-HRNAT	2022	43	28.2	61.7	49.6
Zhao et al. (2022)	Tr-LSTM-RL	2022	42	28.8	62	**54.2**
Zhang et al. (2021)	RCG	2021	43.1	29.3	61.9	52.9
Perez-Martin et al. (2021a)	AVSSN	2021	45.5	31.4	64.3	50.6
Zheng et al. (2020)	SAAT	2020	40.5	28.2	60.9	49.1
Chen et al. (2020)	VNS-GRU	2020	46.0	29.5	63.3	52.0
Hou et al. (2019)	JSRL-VCT	2019	42.3	29.7	62.8	49.1
Wang et al. (2019a)	GFN-POS	2019	42.0	28.2	61.6	48.7
Xiao and Shi (2019z)	DCM-Best1(M)	2019	43.8	**34.2**	**65.8**	47.6
Chen et al. (2019b)	TDConvED(R)	2019	39.5	27.5	–	42.8
Olivastri (2019)	EtENet-IRv2	2019	40.5	27.7	60.6	47.6
Chen et al. (2019a)	SDN	2019	43.8	28.9	62.4	51.4
Aafaq et al. (2019a)	GRU-EVE	2019	38.3	28.4	60.7	48.1
Hammad et al. (2019)	MM-features	2019	39.2	27.8	59.8	45.7
Zhang et al. (2019a)	OA-BTG	2019	41.4	28.2	–	46.9
Liu et al. (2020)	SibNet	2018	40.9	27.5	60.2	47.5
Lee and Kim (2018)	SeFLA	2018	41.8	–	–	–
Wang et al. (2018a)	RecNet	2018	39.1	26.6	59.3	42.7
Shen et al. (2017)	Lexical-FCN	2017	41.4	28.3	61.1	48.9
Zhang et al. (2017)	TDDF	2017	37.3	27.8	59.2	43.8
B. DRL approaches						
Li and Qiu (2020)	E2E(MT-RL)	2019	40.4	27	61	48.3
Wang et al. (2018b)	HRL	2018	41.3	28.7	61.7	48
Chen et al. (2018a)	PickNet	2018	38.9	27.2	59.5	42.1
Pasunuru and Bansal (2017)	CIDEnt	2017	40.5	28.4	61.4	51.7
Phan et al. (2017)	CST	2017	42.2	28.9	62.3	**54.2**
C. Transformer-based approaches						
Im and Choi (2022)	UAT-FEGs	2022	43	27.8	60.9	49.7
Liu et al. (2021)	O2NA	2021	41.6	28.5	62.4	51.1
Jin et al. (2020)	SBAT	2020	42.9	28.9	61.5	51.6
Yang et al. (2019)	NRVC	2019	42.50	28.0	–	49.40
Chen et al. (2018)	TVT	2018	42.46	28.29	61.07	48.53
D. Attention-based approaches						
Ji et al. (2022)	ADL	2022	40.2	26.6	60.2	44
Peng et al. (2021)	T-DL	2021	42.3	28.9	61.7	49.2
Ryu et al. (2021)	SGN	2021	40.8	28.3	60.8	49.5
Perez-Martin et al. (2021b)	SemSynAN	2021	46.4	30.4	64.7	51.9
Zhang et al. (2020)	ORG-TRL	2020	43.6	28.8	62.1	50.9
Yan et al. (2020)	STAT	2020	39.3	27.1	–	44.0
Bin et al. (2019)	BiLTSM	2019	33.9	26.2	–	–
Xiao and Shi (2019b)	Attrib_Sel	2019	40.1	27.2	–	45.5
Sun et al. (2019b)	MSAN	2019	**46.8**	29.5	–	52.4
Li et al. (2019b)	Res-ATT	2019	37	26.9	–	40.7

Table 8 (continued)

References	Model	Year	B	M	R	C
Gao et al. (2019)	hLSTMat	2019	39.7	27	–	43.4
Wang et al. (2018c)	HACA	2018	43.4	29.5	61.8	49.7
Gao et al. (2017)	aLSTMs	2017	38	26.1	-	43.2
Xu et al. (2017)	MA-LSTM	2017	36.5	26.5	59.8	41
Hori et al. (2017)	A.F	2017	39.7	25.5	–	40.4
Chen and Jiang (2019)	MGSA	2017	45.4	28.6	–	50.1

For all evaluation metrics, mechanism-wise highest scores are underlined, and the overall highest score (across all four mechanisms) is in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr, A.F: Attentional Fusion, MT-RL: Multi-task Reinforcement Learning

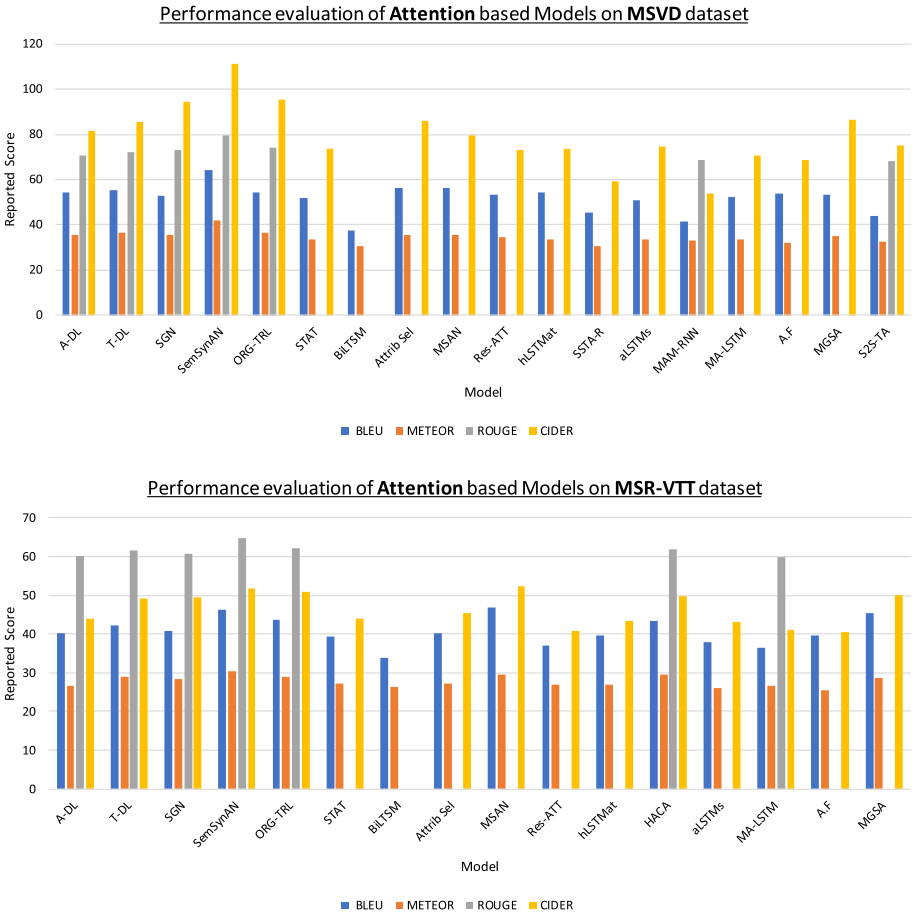

Fig. 13 Performance evaluation of BLEU, METEOR, ROUGE-L and CIDEr on the MSVD & MSR-VTT dataset for attention mechanism based approach

Table 9 Video description performance evaluation on the ActivityNet Captions dataset for three approaches

References	Model	Year	B	M	R	C
A. Standard ED approaches						
Seo et al. (2022)	MV-GPT	2022	6.84	12.31	–	–
Aafaq et al. (2022)	VSJM-Net	2022	3.97	12.89	25.37	26.52
Hosseinzadeh et al. (2021)	VC-FF	2021	2.76	7.02	18.16	26.55
Hou et al. (2019)	JSRL-VCT	2019	1.9	11.30	22.40	**44.20**
Li et al. (2018)	DVC	2018	1.62	10.33	–	25.24
Xu et al. (2019)	JEDDi-Net	2018	1.63	8.58	19.63	19.88
B. Transformer-based approaches						
Wang et al. (2021)	PDVC	2021	10.29	15.8	–	20.45
Estevam et al. (2021)	BMT-V+sm	2021	2.55	8.65	13.62	13.48
Deng et al. (2021)	SGR	2021	1.67	9.07	–	22.12
Song et al. (2021)	TR-Dyn-mem	2021	12.2	16.1	–	27.36
Ging et al. (2020)	COOT	2020	**17.43**	**15.99**	**31.45**	28.19
Iashin and Rahtu (2020)	MDVC	2020	5.83	11.72	–	–
Lei et al. (2020a)	MART	2020	9.78	15.57	–	22.16
Bilkhu et al. (2019)	I3D+UT	2019	**49***	–	–	–
Zhou et al. (2018)	E2E-MskTr	2018	2.23	9.56	–	–
C. Attention-based approaches						
Chen and Jiang (2021)	EC-SL	2021	1.33	7.49	13.02	21.21
Krishna et al. (2017)	DenseCap	2017	3.98	9.5	-	24.6

For all evaluation metrics, mechanism-wise highest scores are underlined, and the overall highest scores are in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr

*: Single metric evaluation

4.2.3 ActivityNet Captions

Results reported from the ActivityNet Captions dataset are presented in Table 9. ActivityNet Captions (Krishna et al. 2017) is a dataset specific to dense captioning events. It covers a wide range of categories, and comprises 20k videos taken from the activity net centered around human activities, with a total duration of 849 hours and 100k descriptions. Overlapping events occurring in a video are provided, and each description uniquely describes a dedicated segment of the video, so it describes events over time. Temporally localized descriptions are used to annotate each video. On average, each video is annotated with 3.65 sentences and 40 words. Event detection is demonstrated in small clips as well as in long video sequences.

Considering the standard ED mechanism, benchmarking on the ActivityNet Captions dataset demonstrated that Joint Syntax Representation Learning and Visual Cue Translation for Video Captioning (JSRL-VCT) (Hou et al. 2019) produced the top METEOR, ROUGE, and CIDEr scores, whereas Video Captioning of Future Frames (VC-FF) (Hosseinzadeh et al. 2021) achieved the best results from the BLEU metric. Among the transformer-based approaches, the recently proposed COOT (Ging et al. 2020), a cooperative hierarchical transformer model, outperformed all previous models for all four metrics. Although the universal transformer approach (Bilkhu et al. 2019) demonstrated the

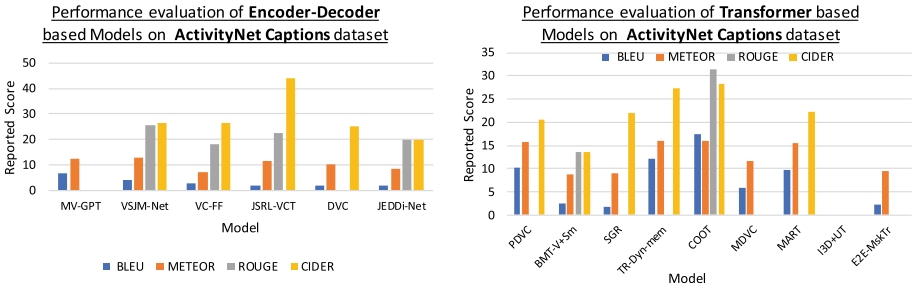

Fig. 14 Performance evaluation of BLEU, METEOR, ROUGE-L and CIDEr on the ActivityNet Captiopns dataset for transformer mechanism & standard Encoder–Decoder based approaches

highest BLEU score, evaluation based only on a single metric cannot guarantee whole-system performance. For attention-based approaches, the pioneer and creator of the ActivityNet Captions dataset (Krishna et al. 2017) reported scores for all four metrics. Related to dense video captioning, in overall performance evaluations of all four mechanisms on ActivityNet Captions, COOT (Ging et al. 2020) outperformed all mechanisms for all four metrics. Figure 14 represents graphical illustration of the results obtained by employing standard Encoder–Decoder and transformer based models on ActivityNet Captipns dataset.

4.2.4 YouCookII

Table 10 presents results reported with the YouCookII (Zhou and Corso 2016) dataset, another dataset mostly utilized to evaluate dense video captioning systems. This dataset comprises 2k YouTube videos that are almost uniformly distributed over 89 recipes from major cuisines all over the world, using a wide variety of cooking styles, components, instructions, and appliances. Each video in the dataset contains 3–16 temporally localized segments annotated in English. There are 7.7 segments per video, on average. About 2,600

Table 10 Video description performance evaluation on the YouCookII dataset for standard encoder–decoder and transformer-based approaches

References	Model	Year	B	M	R	C
A. Standard ED approaches						
Seo et al. (2022)	MV-GPT	2022	**21.88**	**27.09**	**49.38**	2.21
Aafaq et al. (2022)	VSJM-Net	2022	1.09	4.31	10.51	9.07
B. Transformer-based approaches						
Wang et al. (2021)	PDVC	2021	0.89	4.74	–	23.07
Deng et al. (2021)	SGR	2021	–	4.35	–	–
Ging et al. (2020)	COOT	2020	17.97	19.85	37.94	**57.24**
Lei et al. (2020a)	MART	2020	8.0	15.9	–	35.74
Sun et al. (2019b)	VideoBERT	2019	4.33	11.94	28.8	0.55
Zhou et al. (2018)	E2E-MskTr	2018	1.13	5.9	–	–

For all evaluation metrics, mechanism-wise highest scores are underlined, and the overall highest score (across mechanisms) is in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr

Table 11 Video description performance evaluation on the TV show Caption (TVC) dataset for transformer-based approaches

References	Model	Year	B	M	R	C
A. Transformer-based approaches						
Li et al. (2020)	HERO	2020	**12.35**	**17.64**	**34.16**	**49.98**
Lei et al. (2020b)	MMT	2020	10.87	16.91	32.81	45.38

For all evaluation metrics, highest scores are in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr

words are used while describing the recipes. The data split was 67% videos for training, 23% for validation, and 10% for testing purposes.

Only transformer-based model evaluation results are reported using YouCook II dataset, showing that COOT (Ging et al. 2020) again produced excellent scores from all four metrics.

4.2.5 TVC - TV show caption

Results reported from using the TV show Caption dataset are presented in Table 11. The TVC dataset (Lei et al. 2020b) is a multi-modal captioning dataset with 262k captions extended from the TV show Retrieval (TVR) dataset by storing additional descriptions for every single annotated video clip or moment. It involves utilizing both video and subtitles for required information collection and appropriate description generation. The TVC dataset contains 108k video clips paired with 262K descriptions, and on average, there are two to four descriptions per video clip. Human annotators were engaged to write descriptions for video only and video+subtitle if subtitles already existed. The transformer-based MMT model (Lei et al. 2020b) evaluated on TVC for both video and subtitle modalities outperformed the models with only one of the modalities. It establishes the fact that both videos and subtitles are equally valuable for concise and appropriate description generation. Unlike previous datasets employed for video descriptions focusing on captions illustrating visual content, the TVC dataset aims at captions that also describe subtitles.

The creators of the TVC dataset, MMT (Lei et al. 2020b), reported comparable results; however, HERO (Li et al. 2020) demonstrated the highest scores from BLEU, METEOR, ROUGE and CIDEr, demonstrating tough competition from the MMT (Lei et al. 2020b) model.

4.2.6 VATEX - video and TEXt

Table 12 shows the results reported from using the VATEX dataset in both English and Chinese.

VATEX (Wang et al. 2019b) is a large, complex, and diverse multilingual dataset for video descriptions. It contains over 41,269 unique videos covering 600 human activities from kinetic-600 (Kay et al. 2017). There are 10 English and 10 Chinese captions with at least 10 words for English and 15 words for Chinese captions for every clip in the dataset. VATEX comprises 413k English captions and 413k Chinese captions for 41.3k unique videos. Chinese descriptions for each video clip are divided into two parts; half of the descriptions directly describe the video content, while the other half is the paired English translation (done through Google, Microsoft, and a self-developed translation system) of the same clip.

Table 12 Video description performance evaluation on the Video And TEXt (VATEX) dataset for standard encoder–decoder and transformer-based and attention-based approaches

Reerencesf	Model (dataset)	Year	B	M	R	C
A. Standard ED approaches						
Gao et al. (2022)	vc-HRNAT	2022	32.1	21.9	48.4	48.5
Zhang et al. (2021)	RCG	2021	33.9	23.7	50.2	57.5
B. Transformer-based approaches						
Zhu et al. (2018)	X-Lin+Tr(*VATEX-En*)	2020	**40.7**	**25.8**	**53.7**	**81.4**
	X-Lin+Tr(*VATEX-Ch*)		32.6	**32.1**	**56.5**	**59.5**
C. Attention-based approaches						
Zhang et al. (2020)	ORG-TRL(*VATEX-En*)	2020	32.1	22.2	48.9	49.7
Lin et al. (2020)	FAtt(*VATEX-En*)	2020	39.2	25.0	52.7	76.0
	FAtt(*VATEX-Ch*)		**33.1**	30.3	49.7	50.4
Wang et al. (2019b)	ML-Vatex(*VATEX-En*)	2019	28.4	21.7	47.0	45.1
	ML-Vatex(*VATEX-Ch*)		24.9	29.8	51.7	35.0

For all evaluation metrics, mechanism-wise highest scores are underlined, and the overall highest scores are in bold

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr, En: English, Ch: Chinese

For VATEX English and Chinese evaluations of the transformer model, only the X-linear+transformer model is reported, considering it had the highest scores for all metrics. For attention-based systems, Multi-modal Feature Fusion with Feature Attention (FAtt) (Lin et al. 2020) outperformed the baseline with a significant gap, and recorded the highest results for both English and Chinese captioning. However, if we consider the overall performance comparison, the X-linear+transformer model achieved the highest scores for both English and Chinese captioning based on all four metrics.

In Table 13, demonstrating results for miscellaneous datasets, none of the results is highlighted because they were all evaluated on different datasets with different diversities and complexities, so we cannot compare them directly.

From all the above results, we conclude that for simple, single-sentence caption generation, the standard ED & attention mechanisms provide excellent performance, whereas for dense video captioning, the transformer mechanism outperformed the others. For the models to better correlate with sentence-level losses, DRL-based metric optimization is critically needed for high-quality caption generation.

5 Conclusions

Vision and language are the two fundamental systems of human representations, and combining these two into one intelligent and smart system has long been a dream of artificial intelligence.

This survey investigated in detail the four main approaches to video description systems. These deep learning techniques, primarily employing the ED architecture, further accommodate the attention mechanism, the transformer mechanism, and DRL for efficient and accurate output. Owing to the diverse and complex intrinsic structure of video, capturing all the fine-grained detail and complicated spatio-temporal information

Table 13 Video description performance evaluation on misc datasets

References	Model (dataset)	Year	B	M	R	C
A. Standard ED approaches						
Hammoudeh et al. (2022)	Soccer-Cap(Soccer-Dataset)	2022	14.9	27	35	0.99
Zhao et al. (2018)	Tubes(*Charades*)	2018	31.5	19.1	–	18
Pan et al. (2017)	LSTM-TSA(*MPII-MD*)	2017	–	8	–	–
Pan et al. (2017)	LSTM-TSA(*M-VAD*)	2017	–	7.2	–	–
Donahue et al. (2017)	LRCN(*TACoS-ML*)	2015	28.8	–	–	–
Venugopalan et al. (2015)	S2VT(*MPII-MD*)	2015	–	7.1	–	–
Venugopalan et al. (2015)	S2VT(*M-VAD*)	2015	–	6.7	–	–
Yan et al. (2010)	CVC(*WExpo10*)	2010	95.7	67.8	95.8	81.1
B. DRL Approaches						
Ren et al. (2017)	RL-EmbRewd(*MS-COCO*)	2017	30.4	25.1	52.5	93.7
Rennie et al. (2017)	SCST(*MS-COCO*)	2017	31.9	25.5	54.3	106
Phan et al. (2017)	HRL(*Charades*)	2017	18.8	19.5	41.4	23.6
C. Transformer-based approaches						
Wu (2022)	NSVA (NSVA)	2022	24.3	24.3	50.8	113.9
Yuan et al. (2022)	X-Trans2Cap(ScanRefer)	2022	49.07	32.25	65.54	106.11
Yuan et al. (2022)	X-Trans2Cap(Nr3D)	2022	40.51	31.36	68.84	85.4
Vo et al. (2022)	NOC-REK(MSCOCO)	2022	–	32.8	–	138.4
Vo et al. (2022)	NOC-REK(NOCap)	2022	–	–	–	93
Song et al. (2021)	TR-Dyn-mem(CharadesCap)	2021	20.34	20.05	–	27.54
D. Attention-based approaches						
Chen et al. (2021)	Scan2Cap(ScanRefer)	2021	41.49	29.23	63.66	67.95
Gao et al. (2019)	hLSTMat(*LSMDC*)	2019	7.0	5.8	15.0	10.4
Zhou et al. (2019)	GVD(*ANet Ent*)	2019	2.35	11.0	–	45.5
Li et al. (2017)	MAM-RNN(*Charades*)	2017	13.3	19.1	–	18.3
Yu (2017)	GEAN(*VAS*)	2017	–	8.4	22.9	8.4
Yu (2017)	GEAN(*LSMDC*)	2017	–	7.2	15.6	9.3
Laokulrat et al. (2016)	S2S-TA(*M-VAD*)	2016	0.8	7.2	15.9	8.8

Results are reported for references, but are not directly comparable because of the miscellaneous nature of the datasets

B: BLEU, M: METEOR, R: ROUGE, C: CIDEr

present in the video context has not yet been achieved. To accomplish the image captioning task, a lot of research is in progress across the globe on the task of creating video descriptions, and even then, there is a requirement for further achievement and improvement in diverse visual information extraction and accurate description generation.

Deep learning video description mostly revolves around recurrence for sequential data processing, but the main bottleneck from long-term dependencies remains. As an option to recurrence, the transformer mechanism is capable of parallel processing, accelerated training, and handling long-term dependencies; it is space-efficient and much faster than solely self-attention-based methods, and is the model of choice for current advanced hardware. Researchers worldwide have put their efforts into the task of improving generated video descriptions using different state-of-the-art methodologies, but still, even the best performing method cannot match human-generated descriptions.

Despite tremendous improvements, generated descriptions are not yet analogous to human interpretations. So, we can say that the upper bound is still far away, and there is a lot more room for research in this area.

I. There is a need for incorporation of rational expertise in the models to improve the generated captions.
II. The intrinsic multi-modal nature of video contributes to generating captions. Learning multiple features, like visuals, audio, and subtitles (if available in the video), increases the model's ability to better understand and interpret (Ramanishka et al. 2016; Iashin and Rahtu 2020; Wang et al. 2018c; Xu et al. 2017; Hori et al. 2017), thus, improving the overall captioning quality. There is a need to explore this research direction further.
III. The design and development of diversity measuring evaluation metrics to facilitate diverse, efficient, and accurate caption generation is indispensable.
IV. For optimization of video captioning systems, extensive exploration of DRL is required.
V. The unprecedented breakthrough of data-hungry deep learning in various challenging tasks is due to a large number of publicly annotated datasets. The currently available video description datasets lack the visual diversity and language intricacies required to generate human-analogous captions. In particular, for dense video captioning, task-specific dataset creation for improved performance is indispensable. Since the acquisition of high-quality annotations is costly, as an alternative to passive learning (training on a massive labeled dataset), active learning (attempts to maximize a model's performance while annotating the fewest samples possible) can be explored.

We hope this paper will not only facilitate better understanding of video description techniques, but will also accommodate scientists in future research and developments in this specific area.

Funding This work was supported in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education under Grant NRF-2019R1A2C1006159 and Grant NRF-2021R1A6A1A03039493, and in part by the 2022 Yeungnam University Research Grant.

References

Aafaq N, Akhtar N, Liu W, Mian A (2019a) Empirical autopsy of deep video captioning frameworks. arXiv: 1911.09345

Aafaq N, Akhtar N, Liu W, Mian A (2019b) Empirical autopsy of deep video captioning frameworks. arXiv: 1911.09345

Aafaq N, Mian A, Liu W, Gilani SZ, Sha M (2019c) Video description: a survey of methods, datasets, and evaluation metrics 52(6). https://doi.org/10.1145/3355390
Aafaq N, Mian AS, Akhtar N, Liu W, Shah M (2022) Dense video captioning with early linguistic information fusion. IEEE Trans Multimedia. https://doi.org/10.1109/TMM.2022.3146005
Agyeman R, Rafiq M, Shin HK, Rinner B, Choi GS (2021) Optimizing spatiotemporal feature learning in 3D convolutional neural networks with pooling blocks. IEEE Access 9:70797–70805. https://doi.org/10.1109/access.2021.3078295
Al-Rfou R, Choe D, Constant N, Guo M, Jones L (2019) Character-level language modeling with deeper self-attention. Proc AAAI Conf Artif Intell 33 , 3159–3166. https://doi.org/10.1609/aaai.v33i01.33013159arxiv.org/abs/1808.04444
Alzubaidi L, Zhang J, Humaidi AJ, Al-Dujaili A, Duan Y, Al-Shamma O, et al (2021) Review of deep learning: concepts, CNN architectures, challenges, applications, future directions 8(1). https://doi.org/10.1186/s40537-021-00444-8
Amaresh M, Chitrakala S (2019) Video captioning using deep learning: an overview of methods, datasets and metrics. Proceedings of the 2019 IEEE international conference on communication and signal processing, ICCSP 2019 (pp. 656–661). https://doi.org/10.1109/ICCSP.2019.8698097
Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA: visual question answering. Proc IEEE Int Conf Comput Vis. https://doi.org/10.1109/ICCV.2015.279
Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021) ViViT: a video vision transformer. Proceedings of the IEEE international conference on computer vision, 6816–6826. https://doi.org/10.1109/ICCV48922.2021.00676arXiv:2103.15691
Babariya RJ, Tamaki T (2020) Meaning guided video captioning. In: Pattern Recognition: 5th Asian Conference, ACPR 2019, Auckland, New Zealand, November 26–29, 2019, Revised Selected Papers, Part II 5, pp 478–488. Springer International Publishing
Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. 3rd International Conference on Learning Representations, ICLR 2015 -Conference Track Proceedings, 1–15. arXiv:1409.0473
Barbu A, Bridge A, Burchill Z, Coroian D, Dickinson S, Fidler S, Zhang Z (2012) Video in sentences out. Uncertainty Artif Intell–Proc 28th Conf–UAI 2012:102–112 arXiv:1204.2742
Bengio Y, Louradour J, Collobert R, Weston J (2009) Curriculum learning. ACM Int Conf Proc Ser. https://doi.org/10.1145/1553374.1553380
Bhatt S, Patwa F, Sandhu R (2017) Natural language processing (almost) from scratch. Proc IEEE 3rd Int Conf Collaboration Internet Comput CIC 2017 2017:328–338. https://doi.org/10.1109/CIC.2017.00050
Bilkhu M, Wang S, Dobhal T (2019) Attention is all you need for videos: self-attention based video summarization using universal Transformers. arXiv:1906.02792
Bin Y, Yang Y, Shen F, Xie N, Shen HT, Li X (2019) Describing video with attention-based bidirectional LSTM. IEEE Trans Cybern 49(7):2631–2641. https://doi.org/10.1109/TCYB.2018.2831447
Blohm M, Jagfeld G, Sood E, Yu X, Vu NT (2018) Comparing attention-based convolutional and recurrent neural networks: success and limitations in machine reading comprehension. CoNLL 2018–22nd Conference on Computational Natural Language Learning, Proceedings, 108–118. https://doi.org/10.18653/v1/k18-1011arXiv:1808.08744
Brox T, Papenberg N, Weickert J (2014) High accuracy optical flow estimation based on warping–presentation. Lecture Notes Comput Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3024(May):25–36
Cascade-correlation R, Chunking NS (1997) Long Short–Term Memory 9(8):1735–1780
Chen DL, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. Aclhlt 2011–Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies 1 (pp. 190–200)
Chen DZ, Gholami A, Niesner M, Chang AX (2021) Scan2Cap: context-aware dense captioning in RGB-D scans. 3192–3202. https://doi.org/10.1109/cvpr46437.2021.00321arXiv:2012.02206
Chen H, Li J, Hu X (2020) Delving deeper into the decoder for video captioning. arXiv:2001.05614
Chen H, Lin K, Maye A, Li J, Hu X (2019a) A semantics-assisted video captioning model trained with scheduled sampling. https://zhuanzhi.ai/paper/f88d29f09d1a56a1b1cf719dfc55ea61arXiv:1909.00121
Chen J, Pan Y, Li Y, Yao T, Chao H, Mei T (2019b) Temporal deformable convolutional encoder–decoder networks for video captioning. Proc AAAI Conf Artif Intell 33 , 8167–8174. https://doi.org/10.1609/aaai.v33i01.33018167arXiv:1905.01077
Chen M, Li Y, Zhang Z, Huang S (2018) TVT: two-view transformer network for video captioning. Proc Mach Learn Res 95(1997):847–862

Chen S, Jiang Y-G (2019) Motion guided spatial attention for video captioning. Proc AAAI Conf Artif Intel 33:8191–8198. https://doi.org/10.1609/aaai.v33i01.33018191

Chen S, Jiang YG (2021c) Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 1:8421–8431. https://doi.org/10.1109/CVPR46437.2021.00832

Chen S, Yao T, Jiang YG (2019b) Deep learning for video captioning: a review. IJCAI Int Joint Conf Artif Intell 2019:6283–6290. https://doi.org/10.24963/ijcai.2019/877

Chen Y, Wang S, Zhang W, Huang Q (2018) Less is more: picking informative frames for video captioning. Lecture Notes Comput Sci (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 11217:367–384. https://doi.org/10.1007/978-3-030-01261-8_22

Chen Y, Zhang W, Wang S, Li L, Huang Q (2018) Saliency-based spatiotemporal attention for video captioning. 2018 IEEE 4th Int Conf Multimedia Big Data BigMM 2018:1–8

Child R, Gray S, Radford A, Sutskever I (2019) Generating Long Sequences with Sparse Transformers. arXiv:1904.10509

Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, Bengio Y (2014) Learning phrase representations using RNN encoder–decoder for statistical machine translation. EMNLP 2014 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 1724–1734. https://doi.org/10.3115/v1/d14-1179arXiv:1406.1078

Dai Z, Yang Z, Yang Y, Carbonell J, Le QV, Salakhutdinov R (2020) Transformer-XL: Attentive language models beyond a fixed-length context. ACL 2019 -57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 2978–2988. https://doi.org/10.18653/v1/p19-1285arXiv:1901.02860

Das P, Xu C, Doell RF, Corso JJ (2013) A thousand frames in just a few words: lingual description of videos through latent topics and sparse object stitching. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 2634–2641). https://doi.org/10.1109/CVPR.2013.340

Demeester T, Rocktäschel T, Riedel S (2016) Lifted rule injection for relation embeddings. Emnlp 2016—conference on empirical methods in natural language processing, proceedings (pp. 1389–1399). https://doi.org/10.18653/v1/d16-1146

Deng C, Chen S, Chen D, He Y, Wu Q (2021) Sketch, ground, and refine: top-down dense video captioning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn. https://doi.org/10.1109/CVPR46437.2021.00030

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L (2009, June). Imagenet: a large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition, pp 248–255. IEEE

Doddington G (2002) Automatic evaluation of machine translation quality using n-gram co-occurrence statistics, 138. https://doi.org/10.3115/1289189.1289273

Donahue J, Hendricks LA, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2017) Long-term recurrent convolutional networks for visual recognition and description. IEEE Trans Pattern Analys Mach Intell 39(4):677–691. https://doi.org/10.1109/TPAMI.2016.2599174

Elliott D, Keller F (2014) Comparing automatic evaluation measures for image description. 52nd Annu Meet Assoc Comput Linguistics ACL 2014–Proc Conf 2:452–457. https://doi.org/10.3115/v1/p14-2074

Estevam V, Laroca R, Pedrini H, Menotti D (2021) Dense video captioning using unsupervised semantic information. arXiv:2112.08455v1

Fang Z, Gokhale T, Banerjee P, Baral C, Yang Y (2020) Video2Commonsense: generating commonsense descriptions to enrich video captioning. arXiv:2003.05162

Gao L, Guo Z, Zhang H, Xu X, Shen HT (2017) Video captioning with attention-based lstm and semantic consistency. IEEE Trans Multimedia 19(9):2045–2055. https://doi.org/10.1109/TMM.2017.2729019

Gao L, Lei Y, Zeng P, Song J, Wang M, Shen HT (2022) Hierarchical representation network with auxiliary tasks for video captioning and video question answering. IEEE Trans Image Process 31:202–215. https://doi.org/10.1109/TIP.2021.3120867

Gao L, Li X, Song J, Shen HT (2019) Hierarchical LSTMs with adaptive attention for visual captioning. IEEE Trans Pattern Analys Mach Intell 14(8):1–1. https://doi.org/10.1109/tpami.2019.2894139

Gao L, Wang X, Song J, Liu Y (2020) Fused GRU with semantic-temporal attention for video captioning. Neurocomputing 395:222–228. https://doi.org/10.1016/j.neucom.2018.06.096

Gehring J, Dauphin YN (2016) Convolutional Sequence to Sequence Learning. https://proceedings.mlr.press/v70/gehring17a/gehring17a.pdf

Gella S, Lewis M, Rohrbach M (2020) A dataset for telling the stories of social media videos. Proceedings of the 2018 conference on empirical methods in natural language processing, EMNLP 2018:968–974

Ging S, Zolfaghari M, Pirsiavash H, Brox T (2020) COOT: cooperative hierarchical transformer for video-text representation learning. (NeurIPS):1–27. arXiv:2011.00597
Gomez AN, Ren M, Urtasun R, Grosse RB (2017) The reversible resid-ual network: backpropagation without storing activations. Adv Neural Inform Process Syst 2017:2215–2225. arXiv:1707.04585
Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. (http://www.deeplearningbook.org)
Goyal A, Lamb A, Zhang Y, Zhang S, Courville A, Bengio Y (2016) Professor forcing: anew algorithm for training recurrent networks. Adv Neural Inform Process Syst (Nips):4608–4616. arXiv:1610.09038
Hakeem A, Sheikh Y, Shah M (2004) CASE E: a hierarchical event representation for the analysis of videos. Proc Natl Conf Artif Intell:263–268
Hammad M, Hammad M, Elshenawy M (2019) Characterizing the impact of using features extracted from pretrained models on the quality of video captioning sequence-to-sequence models. arXiv:1911.09989
Hammoudeh A, Vanderplaetse B, Dupont S (2022) Deep soccer captioning with transformer: dataset, semantics-related losses, and multi-level evaluation:1–15. arXiv:2202.05728
Han K, Wang Y, Chen H, Chen X, Guo J, Liu Z, Tao D (2022) A survey on vision transformer. IEEE Trans Pattern Analys Mach Intel 8828:1–20. https://doi.org/10.1109/TPAMI.2022.3152247
He D, Zhao X, Huang J, Li F, Liu X, Wen S (2019) Read, watch, and move: reinforcement learning for temporally grounding natural language descriptions in videos. Proceed AAAI Conf Artif Intel 33:8393–8400. https://doi.org/10.1609/aaai.v33i01.33018393. arXiv:1901.06829
He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016:770–778. https://doi.org/10.1109/CVPR.2016.90
Hori C, Hori T, Lee TY, Zhang Z, Harsham B, Hershey JR et al (2017) Attention-based multimodal fusion for video description. Proc IEEE Int Conf Comput Vis 2017:4203–4212. https://doi.org/10.1109/ICCV.2017.450
Hosseinzadeh M, Wang Y, Canada HT (2021) Video captioning of future frames. Winter Conf App Comput Vis:980–989
Hou J, Wu X, Zhao W, Luo J, Jia Y (2019) Joint syntax representation learning and visual cue translation for video captioning. IEEE Int Conf Comput Vis 2019:8917–8926. https://doi.org/10.1109/ICCV.2019.00901
Hussain A, Hussain T, Ullah W, Baik SW (2022) Vision transformer and deep sequence learning for human activity recognition in surveillance videos. Comput Intel Neurosci. https://doi.org/10.1155/2022/3454167
Huszár F (2015) How (not) to train your generative model: scheduled sampling, likelihood, adversary?:1–9. arXiv:1511.05101
Iashin V, Rahtu E (2020) Multi-modal dense video captioning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp 958–959
Im H, Choi Y-S (2022) UAT: universal attention transformer for video captioning. Sensors 22(13):4817. https://doi.org/10.3390/s22134817
Ji W, Wang R, Tian Y, Wang X (2022) An attention based dual learning approach for video captioning. Appl Soft Comput 117:108332. https://doi.org/10.1016/j.asoc.2021.108332
Jia Y, Shelhamer E, Donahue J, Karayev S, Long J, Girshick R, et al. (2014) Caffe: convolutional architecture for fast feature embedding. Mm 2014–proceedings of the 2014 ACM conference on multimedia (pp. 675-678). https://doi.org/10.1145/2647868.2654889
Jin T, Huang S, Chen M, Li Y, Zhang Z (2020) SBAT: Video captioning with sparse boundary-aware transformer. IJCAI Int Joint Conf Artif Intel 2021:630–636. https://doi.org/10.24963/ijcai.2020.88
Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Li FF (2014) Large-scale video classification with convolutional neural net-works. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 1725–1732). https://doi.org/10.1109/CVPR.2014.223
Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, et al. (2017) The kinetics human action video dataset. arXiv:1705.06950
Kazemzadeh S, Ordonez V, Matten M, Berg TL (2014) ReferItGame: referring to objects in photographs of natural scenes:787–798
Kenton M-wC, Kristina L, Devlin J (1953) BERT: pre-training of deep bidirectional transformers for language understanding. (Mlm). arXiv:1810.04805v2
Khan M, Gotoh Y (2012) Describing video contents in natural language. Proceedings of the workshop on innovative hybrid (pp. 27–35)
Kilickaya M, Erdem A, Ikizler-Cinbis N, Erdem E (2017) Re-evaluating automatic metrics for image captioning. 15th conference of the european chapter of the association for computational linguistics, EACL 2017–proceedings of conference (Vol. 1, pp. 199-209). Association for Computational Linguistics (ACL). https://doi.org/10.18653/v1/e17-1019

Kitaev N, Kaiser L, Levskaya A (2020) Reformer: the efficient transformer, 1–12. arXiv:2001.04451
Kojima A, Tamura T, Fukunaga K (2002) Natural language description of human activities from video images based on concept hierarchy of actions. Int J Comput Vis 50(2):171–184. https://doi.org/10.1023/A:1020346032608
Krishna R, Hata K, Ren F, Fei-Fei L, Niebles JC (2017) Dense-captioning events in videos. Proc Int Conf Comput Vis 2017:706–715. https://doi.org/10.1109/ICCV.2017.83
Langkilde-geary I, Knight K (2002) HALogen statistical sentence generator. (July):102–103
Laokulrat N, Phan S, Nishida N, Shu R, Ehara Y, Okazaki N, Nakayama H (2016) Generating video description using sequence-to-sequence model with temporal attention. Coling 2015:44–52
Lavie A, Agarwal A (2007) METEOR: an automatic metric for mt evaluation with improved correlation with human judgments. Proceedings of the Second Workshop on Statistical Machine Translation (June):228–231. http://acl.ldc.upenn.edu/W/W05/W05-09.pdf
Lavie A, Sagae K, Jayaraman S (2004) The significance of recall in automatic metrics for MT evaluation. Lecture Notes Comput Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 3265:134–143. https://doi.org/10.1007/978-3-540-30194-3-16
Lee J, Lee Y, Seong S, Kim K, Kim S, Kim J (2019) Capturing long-range dependencies in video captioning. Proc Int Conf Image Process, ICIP, 2019:1880–1884. https://doi.org/10.1109/ICIP.2019.8803143
Lee S, Kim I (2018) Multimodal feature learning for video captioning. Math Prob Eng. https://doi.org/10.1155/2018/3125879
Lei J, Wang L, Shen Y, Yu D, Berg T, Bansal M (2020) MART: memory-augmented recurrent transformer for coherent video paragraph captioning:2603–2614. https://doi.org/10.18653/v1/2020.acl-main.233arXiv:2005.05402
Lei J, Yu L, Berg TL, Bansal M (2020) TVR: a large-scale dataset for video-subtitle moment retrieval. Lecture Notes Comput Sci (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 12366:447–463. https://doi.org/10.1007/978-3-030-58589-1_27
Levine R, Meurers D (2006) Head-driven phrase structure grammar linguistic approach , formal head-driven phrase structure grammar linguistic approach , formal foundations , and computational realization (January)
Li J, Qiu H (2020) Comparing attention-based neural architectures for video captioning, vol 1194. Available on: https://web.stanford.edu/class/archive/cs/cs224n/cs224n
Li L, Chen Y-C, Cheng Y, Gan Z, Yu L, Liu J (2020) HERO: hierarchical encoder for video+language omni-representation pre-training, 2046–2065. https://doi.org/10.18653/v1/2020.emnlp-main.161arXiv:2005.00200
Li S, Tao Z, Li K, Fu Y (2019) Visual to text: survey of image and video captioning. IEEE Trans Emerg Top Comput Intel 3(4):297–312. https://doi.org/10.1109/tetci.2019.2892755
Li X, Zhao B, Lu X (2017) MAM-RNN: Multi-level attention model based RNN for video captioning. IJCAI International Joint Conference on Artificial Intelligence, 2208–2214. https://doi.org/10.24963/ijcai.2017/307
Li X, Zhou Z, Chen L, Gao L (2019) Residual attention-based LSTM for video captioning. World Wide Web 22(2):621–636. https://doi.org/10.1007/s11280-018-0531-z
Li Y, Yao T, Pan Y, Chao H, Mei T (2018) Jointly localizing and describing events for dense video captioning. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 7492–7500). https://doi.org/10.1109/CVPR.2018.00782
Lin C-Y (2004) ROUGE: A Package for Automatic Evaluation of Summaries. In: Text summarization branches out. Association for Computational Linguistics. Barcelona, Spain, pp 74–81. https://aclanthology.org/W04-1013
Lin K, Gan Z, Wang L (2020) Multi-modal feature fusion with feature attention for vatex captioning challenge 2020:2–5. arXiv:2006.03315
Liu F, Ren X, Wu X, Yang B, Ge S, Sun X (2021) O2NA: an object-oriented non-autoregressive approach for controllable video captioning. Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021:281–292. https://doi.org/10.18653/v1/2021.findings-acl.24arXiv:2108.02359
Liu S, Ren Z, Yuan J (2018) SibNet: Sibling convolutional encoder for video captioning. MM 2018 -Proceedings of the 2018 ACM Multimedia Conference, 1425–1434. https://doi.org/10.1145/3240508.3240667
Liu S, Ren Z, Yuan J (2020) SibNet: sibling convolutional encoder for video captioning. IEEE Trans Pattern Analys Mach Intel, 1–1. https://doi.org/10.1109/tpami.2019.2940007
Lowe DG (1999) Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision, Kerkyra, Greece, 1999, pp 1150–1157, vol 2. https://doi.org/10.1109/ICCV.1999.790410

Lowell U, Donahue J, Berkeley UC, Rohrbach M, Berkeley UC, Mooney R (2014) Translating videos to natural language using deep recurrent neural networks. arXiv:1412.4729v3
Lu J, Batra D, Parikh D, Lee S (2019) ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. (NeurIPS), 1–11. arXiv:1908.02265
Lu J, Xiong C, Parikh D, Socher R (2017) Knowing when to look: adaptive attention via a visual sentinel for image captioning. Proc 30th IEEE Conf Comput Vis Pattern Recogn CVPR, 2017:3242–3250. https://doi.org/10.1109/CVPR.2017.345arXiv:1612.01887
Luo H, Ji L, Shi B, Huang H, Duan N, Li T, et al. (2020) UniVL: a unified video and language pre-training model for multimodal understanding and generation. arXiv:2002.06353
Madake J (2022) Dense video captioning using BiLSTM encoder, 1–6
Mnih V, Kavukcuoglu K, Silver D, Graves A, Antonoglou I, Wierstra D, Riedmiller M (2013) Playing atari with deep reinforcement learning, 1–9. arXiv:1312.5602
Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Hassabis D (2015) Human-level control through deep reinforcement learning. Nature 518(7540):529–533. https://doi.org/10.1038/nature14236
Montague P (1999) Reinforcement learning: an introduction, by Sutton RS and Barto AG trends in cognitive sciences 3(9): 360. https://doi.org/10.1016/s1364-6613(99)01331-5
Olivastri S, Singh G, Cuzzolin F (2019) End-to-end video captioning. International conference on computer vision workshop. https://zhuanzhi.ai/paper/004e3568315600ed58e6a699bef3cbba
Pan Y, Li Y, Luo J, Xu J, Yao T, Mei T (2020) Auto-captions on GIF: a large-scale video-sentence dataset for vision-language pre-training. arXiv:2007.02375
Pan Y, Mei T, Yao T, Li H, Rui Y (2016) Jointly modeling embedding and translation to bridge video and language. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016:4594–4602. https://doi.org/10.1109/CVPR.2016.497arXiv:1505.01861
Pan Y, Yao T, Li H, Mei T (2017) Video captioning with transferred semantic attributes. Proc 30th IEEE Conf Comput Vis Pattern Recogn CVPR 2017:984–992. https://doi.org/10.1109/CVPR.2017.111arXiv:1611.07675
Pan Y, Yao T, Li Y, Mei T (2020) X-linear attention networks for image captioning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn, 10968–10977. https://doi.org/10.1109/CVPR42600.2020.01098arXiv:2003.14080
Park J, Song C, Han JH (2018) A study of evaluation metrics and datasets for video captioning. ICIIBMS 2017 -2nd Int Conf Intel Inform Biomed Sci 2018:172–175. https://doi.org/10.1109/ICIIBMS.2017.8279760
Pasunuru R, Bansal M (2017) Reinforced video captioning with entailment rewards. Emnlp 2017—conference on empirical methods in natural language processing, proceedings (pp. 979–985). https://doi.org/10.18653/v1/d17-1103
Peng Y, Wang C, Pei Y, Li Y (2021) Video captioning with global and local text attention. Visual Computer (0123456789). https://doi.org/10.1007/s00371-021-02294-0
Perez-Martin J, Bustos B, Perez J (2021) Attentive visual semantic specialized network for video captioning, 5767–5774. https://doi.org/10.1109/icpr48806.2021.9412898
Perez-Martin J, Bustos B, Pérez J (2021) Improving video captioning with temporal composition of a visual-syntactic embedding. Winter Conference on Applications of Computer Vision, 3039–3049
Phan S, Henter GE, Miyao Y, Satoh S (2017) Consensus-based sequence training for video captioning. arXiv:1712.09532
Pramanik S, Agrawal P, Hussain A (2019) OmniNet: a unified architecture for multi-modal multi-task learning, 1–16. arXiv:1907.07804
Raffel C, Ellis DPW (2015) Feed-forward networks with attention can solve some long-term memory problems, 1–6. arXiv:1512.08756
Rafiq M, Rafiq G, Agyeman R, Jin S-I, Choi G (2020) Scene classification for sports video summarization using transfer learning. Sensors (Switzerland) 20(6). https://doi.org/10.3390/s20061702
Rafiq M, Rafiq G, Choi GS (2021) Video description: datasets evaluation metrics. IEEE Access 9:121665–121685. https://doi.org/10.1109/ACCESS.2021.3108565
Ramanishka V, Das A, Park DH, Venugopalan S, Hendricks LA, Rohrbach M, Saenko K (2016) Multimodal video description. MM 2016 -Proceedings of the 2016 ACM Multimedia Conference, 1092–1096. https://doi.org/10.1145/2964284.2984066
Ranzato M, Chopra S, Auli M, Zaremba W (2016) Sequence level training with recurrent neural networks. 4th international conference on learning representations, ICLR 2016—conference track proceedings (pp. 1–16)
Redmon J, Farhadi A (2018) YOLOv3: an incremental improvement. arXiv:1804.02767

Ren Z, Wang X, Zhang N, Lv X, Li LJ (2017) Deep reinforcement learning-based image captioning with embedding reward. Proc 30th IEEE Conf Comput Vis Pattern Recogn CVPR 2017:1151–1159. https://doi.org/10.1109/CVPR.2017.128

Rennie SJ, Marcheret E, Mroueh Y, Ross J, Goel V (2017) Self-critical sequence training for image captioning. Proc 30th IEEE Conf Comput Vis Pattern Recogn CVPR 2017:1179–1195. https://doi.org/10.1109/CVPR.2017.131

Rivera-soto RA, Ordóñez J (2013) Sequence to sequence models for generating video captions. http://cs231n.stanford.edu/reports/2017/pdfs/31.pdf

Rohrbach M, Qiu W, Titov I, Thater S, Pinkal M, Schiele B (2013) Translating video content to natural language descriptions. Proc IEEE Int Conf Comput Vis. https://doi.org/10.1109/ICCV.2013.61

Ryu H, Kang S, Kang H, Yoo CD (2021) Semantic grouping network for video captioning. arXiv:2102.00831

Schuster M, Paliwal KK (1997) Bidirectional recurrent. Neural Netw 45(11):2673–2681

Seo PH, Nagrani A, Arnab A, Schmid C (2022) End-to-end generative pretraining for multimodal video captioning, 17959–17968. arXiv:2201.08264

Sharif N, White L, Bennamoun M, Shah SAA (2018) Learning-based composite metrics for improved caption evaluation. ACL 2018 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Student Research Workshop, 14–20. https://doi.org/10.18653/v1/p18-3003

Shen Z, Li J, Su Z, Li M, Chen Y, Jiang YG, Xue X (2017) Weakly supervised dense video captioning. Proc 30th IEEE Conf Comput Vis Pattern Recogn, CVPR 2017 2017:5159–5167. https://doi.org/10.1109/CVPR.2017.548c

Song J, Gao L, Guo Z, Liu W, Zhang D, Shen HT (2017) Hierarchical LSTM with adjusted temporal attention for video captioning, 2737–2743

Song Y, Chen S, Jin Q (2021) Towards diverse paragraph captioning for untrimmed videos. Proceedings of the IEEE Comput Soc Conf Comput Vis Pattern Recogn, 11240–11249. https://doi.org/10.1109/CVPR46437.2021.01109arXiv:2105.14477

Su J (2018) Study of Video Captioning Problem. https://www.semanticscholar.org/paper/Study-of-Video-Captioning-Problem-Su/511f0041124d8d14bbcdc7f0e57f3bfe13a58e99

Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019) VideoBERT: a joint model for video and language representation learning. Proc IEEE Int Conf Comput Vis 2019:7463–7472. https://doi.org/10.1109/ICCV.2019.00756

Sun L, Li B, Yuan C, Zha Z, Hu W (2019) Multimodal semantic attention network for video captioning. Proc IEEE Int Conf Multimedia Expo 2019:1300–1305. https://doi.org/10.1109/ICME.2019.00226. arxiv.org/abs/1905.02963

Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-ResNet and the impact of residual connections on learning. 31st AAAI Conf Artif Intel AAAI 2017:4278–4284

Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. (2015) Going deeper with convolutions. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (07-12-June, pp. 1-9). https://doi.org/10.1109/CVPR.2015.7298594

Torralba A, Murphy KP, Freeman WT, Rubin MA (2003) Context-based vision system for place and object recognition. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV'03, vol 2, pp 273. IEEE Computer Society. https://doi.org/10.5555/946247.946665

Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3D convolutional networks. Proc IEEE Int Conf Comput Vis 2015:4489–4497. https://doi.org/10.1109/ICCV.2015.510

Uszkoreit J, Kaiser L (2019) Universal transformers, 1-23. arxiv.org/abs/arXiv:1807.03819v3

Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, et al. (2017) Attention is all you need. Adv Neural Inform Process Syst (Nips), 5998–6008. http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

Vedantam R, Lawrence Zitnick C, Parikh D (2015) Cider: consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4566–4575

Venugopalan S, Rohrbach M, Donahue J, Mooney R, Darrell T, Saenko K (2015) Sequence to sequence -video to text. Proceedings IEEE Int Conf Comput Vis 2015:4534–4542. https://doi.org/10.1109/ICCV.2015.515

Vo DM, Chen H, Sugimoto A, Nakayama H (2022) NOC-REK: Novel object captioning with retrieved vocabulary from external knowledge. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp 17979–17987. https://doi.org/10.1109/CVPR52688.2022.01747

Wallach B (2017) Developing: a world made for money (pp. 241–294). https://doi.org/10.2307/j.ctt1d98bxx.10

Wang D, Song D (2017) Video Captioning with Semantic Information from the Knowledge Base. Proceedings -2017 IEEE International Conference on Big Knowledge, ICBK 2017 , 224–229. https://doi.org/10.1109/ICBK.2017.26

Wang B, Ma L, Zhang W, Liu W (2018a) Reconstruction network for video captioning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp 7622–7631. https://doi.org/10.1109/CVPR.2018.00795

Wang X, Chen W, Wu J, Wang YF, Wang WY (2018b) Video captioning via hierarchical reinforcement learning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn, 4213–4222. https://doi.org/10.1109/CVPR.2018.00443arXiv:1711.11135

Wang X, Wang, Y-f, Wang WY (2018c) Watch , listen , and describe: globally and locally aligned cross-modal attentions for video captioning, 795–801

Wang B, Ma L, Zhang W, Jiang W, Wang J, Liu W (2019a) Controllable video captioning with pos sequence guidance based on gated fusion network. Proc IEEE Int Conf Comput Vis 2019:2641–2650. https://doi.org/10.1109/ICCV.2019.00273. arXiv:1908.10072

Wang X, Wu J, Chen J, Li L, Wang Y-F, Wang WY (2019b) VATEX: a large-scale, high-quality multilingual dataset for video-and-language research. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp 4580–4590. https://doi.org/10.1109/ICCV.2019.00468

Wang H, Zhang Y, Yu X (2020) An overview of image caption generation methods. Computational Intelligence and Neuroscience 2020. https://doi.org/10.1155/2020/3062706

Wang T, Zhang R, Lu Z, Zheng F, Cheng R, Luo P (2021) Endto-End Dense Video Captioning with Parallel Decoding. Proceedings of the IEEE International Conference on Computer Vision, 6827–6837. https://doi.org/10.1109/ICCV48922.2021.00677arXiv:2108.07781

Williams RJ, Zipser D (1989) A learning algorithm for continually running fully recurrent neural networks. Neural Comput 1(2):270–280. https://doi.org/10.1162/neco.1989.1.2.270

Wu D, Zhao H, Bao X, Wildes RP (2022) Sports video analysis on large-scale data (1). arXiv:2208.04897

Wu Z, Yao T, Fu Y, Jiang, Y-G (2017) Deep learning for video classification and captioning. Front Multimedia Res, 3–29. https://doi.org/10.1145/3122865.3122867arXiv:1609.06782

Xiao H, Shi J (2019a) Diverse video captioning through latent variable expansion with conditional GAN. https://zhuanzhi.ai/paper/943af2926865564d7a84286c23fa2c63 arXiv:1910.12019

Xiao H, Shi J (2019b) Huanhou Xiao, Jinglun Shi South China University of Technology, Guangzhou China, 619–623

Xie S, Sun C, Huang J, Tu Z, Murphy K (2018) Rethinking spatiotem-poral feature learning: speed-accuracy trade-offs in video classification. Lecture Notes Comput Sci (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 11219:318–335. https://doi.org/10.1007/978-3-030-01267-0_19

Xu H, Li B, Ramanishka V, Sigal L, Saenko K (2019) Joint event detection and description in continuous video streams. Proc 2019 IEEE Winter Conf App Comput Vis, WACV 2019:396–405. https://doi.org/10.1109/WACV.2019.00048. arXiv:1802.10250

Xu J, Mei T, Yao T, Rui Y (2016) MSR-VTT: a large video description dataset for bridging video and language. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016:5288–5296. https://doi.org/10.1109/CVPR.2016.571

Xu J, Wei H, Li L, Fu Q, Guo J (2020) Video description model based on temporal-spatial and channel multi-attention mechanisms. Appl Sci (Switzerland). https://doi.org/10.3390/app10124312

Xu J, Yao T, Zhang Y, Mei T (2017) Learning multimodal attention LSTM networks for video captioning. MM 2017 -Proceedings of the 2017 ACM Multimedia Conference, 537–545. https://doi.org/10.1145/3123266.3123448

Xu K, Ba JL, Kiros R, Cho K, Courville A, Salakhutdinov R, et al. (2015) Show, attend and tell: neural image caption gener-ation with visual attention. 32nd International Conference on Machine Learning, ICML 2015 3:2048–2057. arXiv:1502.03044

Xu W, Yu J, Miao Z, Wan L, Tian Y, Ji Q (2021) Deep reinforcement polishing network for video captioning. IEEE Trans Multimedia 23:1772–1784. https://doi.org/10.1109/TMM.2020.3002669

Yan C, Tu Y, Wang X, Zhang Y, Hao X, Zhang Y, Dai Q (2020) STAT: spatial-temporal attention mechanism for video captioning. IEEE Trans Multimedia 22(1):229–241. https://doi.org/10.1109/TMM.2019.2924576

Yan L, Zhu M, Yu C (2010) Crowd video captioning. arXiv:1911.05449v1

Yan Y, Zhuang N, Bingbing Ni, Zhang J, Xu M, Zhang Q, et al (2019) Fine-grained video captioning via graph-based multi-granularity interaction learning. IEEE Trans Pattern Analys Mach Intel. https://doi.org/10.1109/TPAMI.2019.2946823

Yang B, Liu F, Zhang C, Zou Y (2019) Non-autoregressive coarse-to-fine video captioning. In: AAAI Conference on Artificial Intelligence. https://doi.org/10.1609/aaai.v35i4.16421
Yang Z, Yuan Y, Wu Y, Salakhutdinov R, Cohen WW (2016) Review networks for caption generation. Adv Neural Inform Process Syst (Nips), 2369–2377. arXiv:1605.07912
Yin W, Kann K, Yu M, Schütze H (2017) Comparative study of CNN and RNN for natural language processing. arXiv:1702.01923
You Q, Jin H, Wang Z, Fang C, Luo J (2016) Image captioning with semantic attention. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2016:4651–4659. https://doi.org/10.1109/CVPR.2016.503. arXiv:1603.03925
Young P, Lai A, Hodosh M, Hockenmaier J (2014) From image descriptions to visual denotations? New similarity metrics for semantic inference over event descriptions 2:67–78
Yu Y, Choi J, Kim Y, Yoo K, Lee SH, Kim G (2017) Supervising neural attention models for video captioning by human gaze data. Proc 30th IEEE Conf Comput Vis Pattern Recogn 2017:6119–6127. https://doi.org/10.1109/CVPR.2017.648. arXiv:1707.06029
Yuan Z, Yan X, Liao Y, Guo Y, Li G, Li Z, Cui S (2022) X-Trans2Cap: cross-modal knowledge transfer using transformer for 3D dense captioning, 3–4. arXiv:2203.00843
Zellers R, Bisk Y, Farhadi A, Choi Y, (2019) From recognition to cognition: visual commonsense reasoning. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2019:6713–6724. https://doi.org/10.1109/CVPR.2019.00688
Zhang J, Peng Y (2019) Object-aware aggregation with bidirectional temporal graph for video captioning. https://zhuanzhi.ai/paper/237b5837832fb600d4269cacdb0286e3 arXiv:1906.04375
Zhang Q, Zhang M, Chen T, Sun Z, Ma Y, Yu B (2019) Recent advances in convolutional neural network acceleration. Neurocomputing 323:37–51. https://doi.org/10.1016/j.neucom.2018.09.038. arXiv:1807.08596
Zhang W, Wang B, Ma L, Liu W (2019) Reconstruct and represent video contents for captioning via reinforcement learning. IEEE Trans Pattern Analys Mach Intel, 1–1. https://doi.org/10.1109/tpami.2019.2920899arxiv.org/abs/1906.01452
Zhang X, Gao K, Zhang Y, Zhang D, Li J, Tian Q (2017) Task-driven dynamic fusion: reducing ambiguity in video description. Proc 30th IEEE Conf Comput Vis Pattern Recogn CVPR 2017:6250–6258. https://doi.org/10.1109/CVPR.2017.662
Zhang X, Sun X, Luo Y, Ji J, Zhou Y, Wu Y, Ji R (2021) RSTnet: captioning with adaptive attention on visual and non-visual words. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 1:15460–15469. https://doi.org/10.1109/CVPR46437.2021.01521
Zhang Y, Vogel S (2010) Significance tests of automatic machine translation evaluation metrics. Machine Transl 24(1):51–65. https://doi.org/10.1007/s10590-010-9073-6
Zhang Z, Qi Z, Yuan C, Shan Y, Li B, Deng Y, Hu W (2021) Open-book video captioning with retrieve-copy-generate network. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn, 9832–9841. https://doi.org/10.1109/CVPR46437.2021.00971arXiv:2103.05284
Zhang Z, Shi Y, Yuan C, Li B, Wang P, Hu W, Zha Z (2020) Object relational graph with teacher-recommended learning for video captioning. arXiv:2002.11566
Zhao B, Li X, Lu X (2018) Video captioning with tube features. IICAI Int Joint Conf Artif Intel 2018:1177–1183. https://doi.org/10.24963/ijcai.2018/164
Zhao H, Chen Z, Guo L, Han Z (2022) Video captioning based on vision transformer and reinforcement learning. Peer J Comput Sci 8(2002):1–16. https://doi.org/10.7717/PEERJ-CS.916
Zheng Q, Wang C, Tao D (2020) Syntax-Aware Action Targeting for Video Captioning. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition 13093–13102. https://doi.org/10.1109/CVPR42600.2020.01311
Zhou L, Corso JJ (2016) Towards automatic learning of procedures from web instructional videos. arXiv:1703.09788v3
Zhou L, Kalantidis Y, Chen X, Corso JJ, Rohrbach M (2019) Grounded video description. Proc IEEE Comput Soc Conf Comput Vis Pattern Recogn 2019:6571–6580. https://doi.org/10.1109/CVPR.2019.00674. arXiv:1812.06587
Zhou L, Zhou Y, Corso JJ, Socher R, Xiong C (2018) End-to-End Dense Video Captioning with Masked Transformer. Proceedings of the IEEE computer society conference on computer vision and pattern recognition (pp. 8739–8748). https://doi.org/10.1109/CVPR.2018.00911
Zhu X, Guo L, Yao P, Lu S, Liu W, Liu J (2019) Vatex video captioning challenge 2020: multi-view features and hybrid reward strategies for video captioning. arXiv:1910.11102
Zolfaghari M, Singh K, Brox T (2018) ECO: efficient convolutional network for online video understanding. Lecture Notes Comput Sci (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) 11206:713–730. https://doi.org/10.1007/978-3-030-01216-8-43

Authors and Affiliations

Ghazala Rafiq[1] · Muhammad Rafiq[2] · Gyu Sang Choi[1]

Ghazala Rafiq
ghazala@ynu.ac.kr

[1] Department of Information & Communication Engineering, Yeungnam University, Gyeongsan-si 38541, South Korea

[2] Department of Game & Mobile Engineering, Keimyung University, 1095 Dalgubeol-daero, Dalseo-gu, Daegu 42601, South Korea