Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.046155

REVIEW

Trends in Event Understanding and Caption Generation/Reconstruction in Dense Video: A Review

Ekanayake Mudiyanselage Chulabhaya Lankanatha Ekanayake[1,2], **Abubakar Sulaiman Gezawa**[3,*] and **Yunqi Lei**[1]

[1]Department of Computer Science, Xiamen University, Xiamen, 361005, China

[2]Main Library, Wayamba University of Sri Lanka, Kuliyapitiya, 60200, Sri Lanka

[3]School of Information Engineering, Sanming University, Sanming, 365004, China

*Corresponding Author: Abubakar Sulaiman Gezawa. Email: 20220505@fjsmu.edu.cn

Received: 20 September 2023 Accepted: 06 December 2023 Published: 26 March 2024

ABSTRACT

Video description generates natural language sentences that describe the subject, verb, and objects of the targeted Video. The video description has been used to help visually impaired people to understand the content. It is also playing an essential role in devolving human-robot interaction. The dense video description is more difficult when compared with simple Video captioning because of the object's interactions and event overlapping. Deep learning is changing the shape of computer vision (CV) technologies and natural language processing (NLP). There are hundreds of deep learning models, datasets, and evaluations that can improve the gaps in current research. This article filled this gap by evaluating some state-of-the-art approaches, especially focusing on deep learning and machine learning for video caption in a dense environment. In this article, some classic techniques concerning the existing machine learning were reviewed. And provides deep learning models, a detail of benchmark datasets with their respective domains. This paper reviews various evaluation metrics, including Bilingual Evaluation Understudy (BLEU), Metric for Evaluation of Translation with Explicit Ordering (METEOR), Word Mover's Distance (WMD), and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) with their pros and cons. Finally, this article listed some future directions and proposed work for context enhancement using key scene extraction with object detection in a particular frame. Especially, how to improve the context of video description by analyzing key frames detection through morphological image analysis. Additionally, the paper discusses a novel approach involving sentence reconstruction and context improvement through key frame object detection, which incorporates the fusion of large language models for refining results. The ultimate results arise from enhancing the generated text of the proposed model by improving the predicted text and isolating objects using various keyframes. These keyframes identify dense events occurring in the video sequence.

KEYWORDS

Video description; video to text; video caption; sentence reconstruction

1 Introduction

Humans can easily describe a video image or event in normal language, it is difficult for machines or algorithms to do so. It uses several computer vision and Natural Language Processing (NLP) approaches to comprehend multiple entities and identify various event occurrences in a video scene [1]. These issues become even more challenging when the videos are crowded or event-related. The background landscape, light and shadows, objects, human, human-object collaboration, and other events simultaneously occur in a video. The NLP techniques must be used to accurately express all of this information. Recently, computer vision and NLP have been integrated to address the issues in understanding and describing the video sequence [2]. The video description has been used in auto video subtitling, robot environment interaction, and video surveillance. It also helped the visually impaired people to understand the content by generating the description of the environment using speech synthesis and reading out the video description. Presently, these processes are expensive, exhausting, and operate as a manual process [3]. The definition of sign language is another application of video description. It may also provide textual instructions for people or robots by turning actions into basic steps that are shown in a video. For example, there are some events like demonstrating how to download video songs, cooking some foods, walking with dogs, or driving [4]. These developments in video description unfold massive opportunities in various functional domains. It is predicted that humans will cooperate with robots similar to humans in the future. Assume that video description is advanced enough to comprehend various real-world occurrences and transmit them in spoken language. In that scenario, it enables service robots or cutting-edge phone applications to identify human actions and various events to interact with people in a way that is more evocative and understandable. For instance, they might answer a user's inquiry about their requirements or expectations to think about what they ought to offer visitors. In an industrial setting, they may assist workers with any procedure that was lacking from their routine as support work [5]. The automatic generation of stories from video frames is not science fiction. The use of the deep neural network in natural language processing and computer vision has increased the large dataset in the field. The video description generated by NLP that explains the video content has two essential factors: The first step is understanding the content of the video in the second phase is the construction of grammatically correct language sentences. This type of work has originated from the robotics community and can be classified as language-grounded, meaning from vision to robotic perception [6]. The recent field in natural language processing involves establishing words in an image, illustrating the picture in natural language sentences, and enabling the robots to understand the natural language instructions [7]. One more exciting field of NLP is Visual Information Retrieval (VIR), which performs content-related search by using mixed content, including visuals (drawings, images, sketches) and text (tags, sentences, keywords). The MS COCO [8] and Flickr30k datasets [9] helped the researchers with image captioning using effective image questioning.

Additionally, video description methods should describe the pace, direction of different objects co-occurring, and interconnection of events and things. Lastly, events of videos could be of unpredictable intervals and may even result in a possible intersection of events. The visual description can be subdivided into image and video content. The video content is then divided into four subcategories. The video captioning, dense video captioning, video description graph generation, and video questioning and answering as shown in Fig. 1.

Although video captioning and dense Video captioning look similar, they are far more different from each other. The dense video captioning is far more complex than the simple video translation. A simple video caption is just a sentence explaining the events and objects in a video frame, while dense

Video captioning is a series of multiple events and objects co-occurring. In dense Video captioning, one should have produced a unique text description for individual events, as shown in Fig. 2.

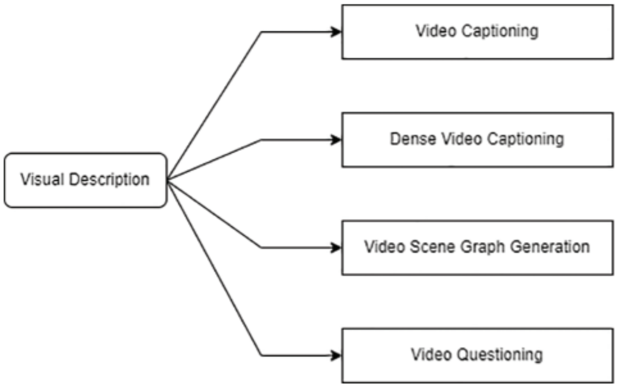

Figure 1: Shows the classification of visual content description

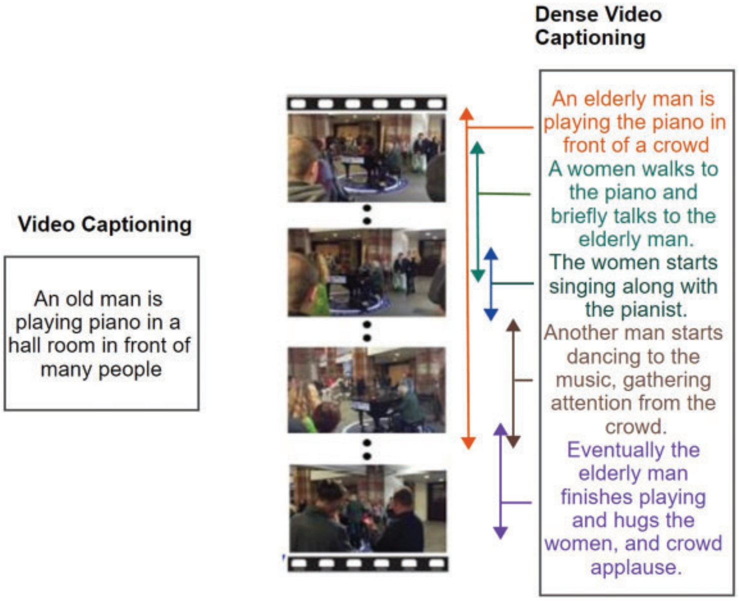
Video Captioning
An old man is playing piano in a hall room in front of many people
Dense Video Captioning
An elderly man is playing the piano in front of a crowd
A women walks to the piano and briefly talks to the elderly man.
The women starts singing along with the pianist.
Another man starts dancing to the music, gathering attention from the crowd.
Eventually the elderly man finishes playing and hugs the women, and crowd applause.

Figure 2: The difference between a simple video translation and dense video captioning [1]

Another problem with the dense video caption is the event overlapping during the frames. Therefore, these properties make a dense video more challenging than a simple Video. Therefore in this review, the main idea aims to discuss several video captioning methods related to video processing and acquire dense events by identifying key scenes of particular video segments.

1.1 Structure of This Survey

The current research lacks a complete and methodical review covering all aspects of video description research, models, datasets, evaluation processes, results, related contests, and video Q&A challenges. This paper deals with this gap and comprehensively reviews further advanced research done in dense video description generation using advanced NLP and CV models. The process of creating descriptive captions that are pertinent to the surrounding context and the actions shown in a particular video sequence is the main idea behind this approach. To ensure that the automatically generated captions accurately reflect the spirit of the developing events in the video, this necessitates a detailed knowledge of the temporal and spatial links between the visual content and the accompanying narrative. The key scene extraction and particular object detection can leverage the meaning of the event by adding extra information.

In Section 1, this review features the critical applications and developments of video description and then categorizes automatic video description techniques into two different clusters, providing the details of the methods for each class in Section 2. Section 3 explains the existing datasets used in the video description. Additionally, this review evaluates the evaluation metrics used for comparative analysis of the video descriptions in Section 4. Section 5 discusses the future directions that can improve the field. In Section 6, conclude our review paper and consider several insights and findings.

To comprehend events in videos and generate meaningful captions, this work currently presents a comprehensive analysis of the existing literature. It places a particular focus on models within the deep-learning methodology, signifying an advancement beyond conventional deep-learning approaches. The primary objective of this study is to deliver a thorough examination of publicly available datasets used for comprehensive event understanding via video analysis. It highlights the performance of popular techniques on each dataset, enabling readers to identify the most state-of-the-art approaches that yield optimal results. Additionally, the paper presently explores commonly employed evaluation metrics and engages in discussions concerning the inherent limitations of dataset structures. The main contributions of this research are as follows:

- A detailed explanation of video description methods for caption generation and visual content recognition. And discussing vital challenges and new approaches related to the event understanding
- The paper discusses in detail the popular publicly available datasets and their utility for comprehending events-related activities. It also explains video-related datasets.
- In addition to video captioning, our focus in this context centers on in-depth event comprehension within video sequences. This paper introduces innovative methodologies, provides detailed insights into the framework, and explains the process for localizing various events within the same video. Notably, our work introduces novel aspects, including scene content analysis and shot boundary detection through color space analysis.
- Sentence reconstruction and context improvement through key frame object detection currently elucidate a novel approach for refining final results. This approach also involves the fusion of large language models.

This review delves into the intricate realm of dense caption generation and event comprehension within videos, employing a range of advanced deep-learning models and techniques. The primary focus lies in the domains of dense captioning for videos and a comprehensive understanding of events, with a particular emphasis on identifying crucial scenes. The review prioritizes the integration of low-computation encoder models for sentence generation, strategically addressing the intricate complexities associated with video processing.

Moreover, the review explores a proposed methodology for sentence reconstruction, which involves a detailed analysis of object detection within pivotal scenes. This strategic approach aims to enhance the overall quality and coherence of the generated sentences. By concentrating on identifying key scenes and employing sophisticated yet computationally efficient models for sentence generation, this review endeavors to offer insights and methodologies for improving the quality and accuracy of dense captions associated with video content and event comprehension.

2 Video Description Methods

The video description methods are divided into two main groups: the classical methods and the deep learning methods. The visual description worked in classical methods based on traditional computer vision (CV) and natural language processing (NLP) methods. These methods used the (object, action, and scene) entities in the videos and combined them into sentences. However, some researchers categorize them into classical methods. The statistical group enforced the statistical methods for the large data set. At last, the deep learning group consists of machine learning, deep learning, and artificial intelligence methods. These state-of-the-art methods have successfully solved CV and NPL problems [10]. In the next section, this paper will describe these two groups in detail.

2.1 Classical Models

In classical description methods, the SVO (subject, verb, object) methods were used to detect subject, verb, objects, and events. During this era, many efforts were made to translate the video scenes into the caption. The work of Koller et al. [11] developed a model that was successful in explaining the motion of vehicles from a crowded traffic scene. The researchers successfully used natural language words to identify and explain the automobile's motion. In the late 90s, Brand et al. [12] used the Hollywood scene and converted it into a description. This method was called "Inverse Hollywood Problems". In this method, a movie scene is transformed into a series of actions to create a storyboard from video instructions. This system was called a "video gister" that analyzes the video scene into sequences of events to generate a written script that explains the action and events listed in a video scene. This method also developed key frame drawing to detect the event and represent these events in a semantic representation like entering, exiting, adding, removing, and motion. This video gister worked only on human-arm cooperation and explained only five actions, including touch, remove, add, get, and put. The recent work of Zhu et al. [13] introduced several sentence description classification tasks into seven different phases. In that explanation, SVO additionally involved complements (C) and adverbials (A) for sentence constructions. Considering the video description generation, actions and object combinations are described as SV, SVA, SVC, SVO, SVOO, SVOA, and SVOC, respectively. The detailed sentence construction templates are shown in Table 1.

Table 1: The several templates for sentence generation

Templates	
SV	subject + verb
SVA	subject + verb + adverbial
SVC	subject + verb + complement
SVO	subject + verb + object
SVOO	subject + verb + object + object
SVOA	subject + verb + object + adverbial
SVOC	subject + verb + object + complement

2.2 Deep Learning Models

Deep learning technologies have achieved great accomplishments in various fields of computer vision. The Convolution Neural Network is the state-of-art model used for data visualization and object recognition. Long Short-Term Memory (LSTM) [14] models are Deep Recurrent Networks used for sequence modeling, speech recognition, and machine translations. The problem with the traditional model is the data handling and complex nature of datasets. Therefore, it allows deep learning models to solve diverse problem-solving models. The detailed function of deep learning models given in Fig. 3 indicates that different deep learning models can be used in two main stages of video description, event understanding(content extraction) and caption generation.

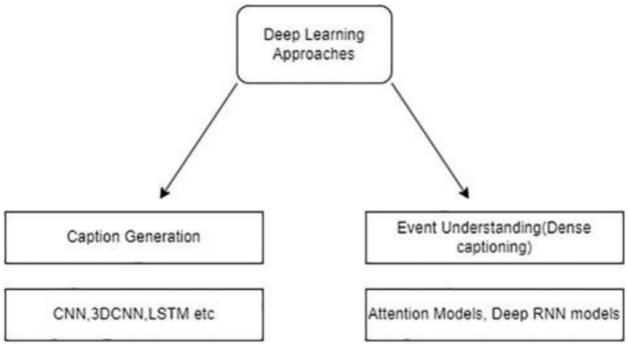
Deep Learning Approaches
Caption Generation
Event Understanding(Dense captioning)
CNN,3DCNN,LSTM etc
Attention Models, Deep RNN models

Figure 3: Deep learning-based models for visual descriptions

The Convolution Neural Network (CNN), Recurrent Nueral Network (RNN), and LSTM are to learn the visual features from the user to be used for the sentence generation phase. The LSTM, deep RNN, and Gated Recurrent Unit are used for text generation. This paper divides this stage into two different architecture models used for the encoding (feature extraction) and decoding stage.

2.2.1 CNN-RNN-Based Description Models

This encoding and decoding process is precise and can be used for a significant scale level. This method allows the variable size of the Video for description and text generation. The work of Olivastri et al. [15] used the encoder-decoder strategy to solve the problem of end-to-end description in a dense environment. They used the CNN model for encoding the network while the RNN model for decoding the model. The CNN model was trained to learn the object and action recognition-related tasks to understand videos' features for visual content learning from Video. At the same time, the decoder is designed to retain the static features to generate the description for the videos. The Microsoft Research Video Description (MSVD) [16] and Microsoft Research Video to Text (MSR-VTT) [17] datasets were used for the experiment. The proposed model was compared with state-of-art models for evaluation. While Li et al. [18] proposed a dense video captioning method by localizing each event's temporal events and sentence generation. The proposed model also consists of two parts: The Temporal Events Proposal (TEP) and Sentence Generation (SG). The model used the 3D Convolution (3DCNN) for descriptive-based temporal attention TEP generation, while LSTM used the model for sentence generation (SG). The model achieved the 12.96% METEOR on the ActivityNet Caption test dataset. To understand human actions and activities in a video, Wang et al. [19] used the incorporation of statistical spatial, short-term motion, and long-term video temporal structure information. They proposed a Hierarchical Attention Network (HAN) to capture complex human action. The HAN can capture the video structure for an extended period and reveal the information for different time slots. The attention process of HAN is illustrated in Fig. 4.

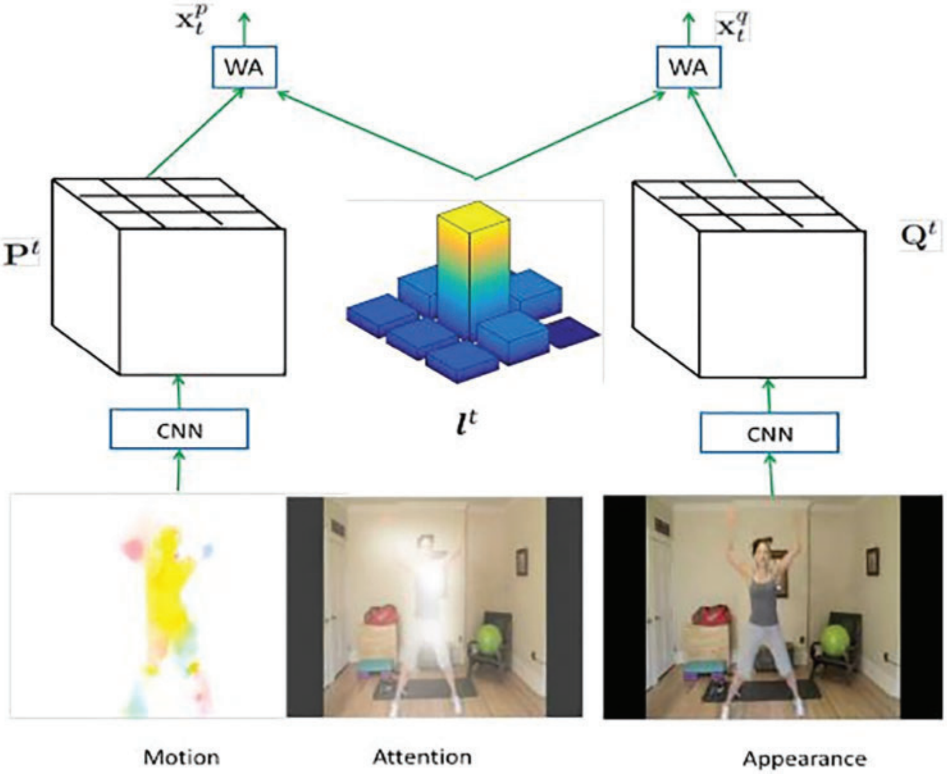

Figure 4: The Attention Process of HAN [19] model

Video captioning is not only about understanding the objects, but it is also very important to understand the relationship among the objects. Most of the existing models fail to understand the interaction among the objects. To address the issue, Zheng et al. [20] proposed a syntax-aware action targeting (SAAT) model to understand the actions of video dynamics. The proposed model used the 2D and 3D CNN models to understand the global dependency between objects. The semantic information also plays a crucial role in understanding the frame information. For this purpose, a Semantic Attended Video Summarization Network (SASUM) [21]. The SASUM model consisted of a frame selector phase and a video descriptor phase. The model used CNN for the extraction of high-level semantic feature information. These extracted features were fed to a network of LSTM models to attain the video description. The video descriptor used the encoder-decoder that translated the visual content into text description.

Most of the deep neural network lacks adequate visual representation. This problem occurs due to a lack of interaction between objects. Insufficient training and improper word selection raise the long-tailed concern. To solve this problem, Zhang et al. [22] proposed an object-relational graph (ORG) using the GNC and attention-based LSTM as the encoder. For decoding, it used a series of the 2DCNN and 3DCNN networks. The encoder captured the detailed interaction features from the video to attain a more realistic visual representation. The decoder, also called teacher recommendation learning (TRL), establishes a relationship with natural language to make full use of effective and grammatically correct sentences. Sequence-to-sequence model cooperation with attention mechanisms has achieved promising video captioning results because there is rich information between a frame that can be used for an excellent result. For this purpose, Chen and Jiang [23] proposed motion-guided spatial attention (MGSA) that captured the motion from the frames and learned the spatial attention using the custom CNN. Furthermore, the researcher designed a gated recurrent attention unit (GRAU) to cooperate with the CNN network. The model achieved promising results on MSVD and Microsoft Research Video to Text (MSR-VTT) datasets.

2.2.2 RNN-Based Description Models

Although these models are new and not popular enough, these RNN-based Video captioning schemes still provide good results when applied in sequence [24]. Zhao et al. [25] proposed a co-attention-based recurrent neural network (CAM-RNN) for Video text captioning. The model used the most correlated text and visual attributes for caption generation. The CAM was used as an encoder to extract the visual and text features, while RNN acted as a decoder to generate the most related caption. Similarly, Li et al. [26] used the multi-level-attention model-based recurrent neural network (MAM-RNN) for most related text caption generation for dense videos. Most of the video caption methods follow an encoder-decoder approach for video captioning. The problem with this method is that it only concentrates on the source video and cannot obtain the perspective information of a word that appeared multiple times during the video. For this purpose, Pei et al. [27] proposed a memory structure named memory attention recurrent network (MARN) for video text generation. The MARN explored the full spectrum of a word that appeared in a video during model training. This process allowed the model to understand the context of a word from a different perspective. Generating a caption for a dense video is another high-level task as multiple events coincide. Most of the existing models use visual features for text generation. They cannot understand the relationship among the objects that deduces the model efficiency, especially in dense videos. For this purpose, Chang et al. [28] proposed an RNN-based model that used visual and audio features to understand the relationship among the objects in a dense video. On the other hand, Niu et al. [29] proposed a multi-layer memory-sharing network for the text generation of a video which uses a soft attention feature to ignore irrelevant information from the video.

2.2.3 LSTM Based Models

A fundamental model used in computer vision to handle language-related problems is the LSTM model. Its widespread use in language translation emphasizes its importance in this area [30]. In recent years, the LSTM has been a critical model used for video captioning and text generation. The work of Li et al. [31] used the fast-forward LSTM for the YouTube-8 challenge and achieved third place. The model consisted of seven hidden layers that attained an accuracy of 82.75% on the Kaggle Public dataset. The problem with the LSTM model is that although the model is perfect for handling the temporal information for the videos. However, when the scene's length increases, it is complicated for the model to address the temporal dependencies. So, the previously generated sequences do not work for the word prediction. To overcome the problem Zhu et al. [32] used a densely connected LSTM model to handle the prior sequences for word prediction. The dense LSTM allows updating the current information using all the previous information, as shown in Fig. 5.

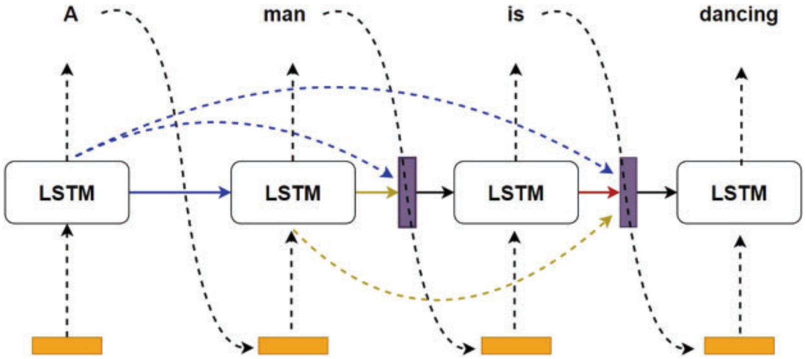

Figure 5: An overview of the Dense LSTM model proposed in [32]

When the LSTM with a large number of layers starts converging the degradation problem arises. This problem reduces the accuracy of the LSTM model during the video caption process. To overcome this problem, Li et al. [14] proposed a Residual Attention-based LSTM (Res-ATT) that uses the sentence internal information lost in the transmission process. The model integrates the residual mapping process into the LSTM network that solves the degradation problem. Ahmed et al. [33] proposed the attention-based LSTM as an encoder and decoder to obtain the textual format. The model consisted of two layers, i.e., bi-LSTM and sequential LSTM for the video caption. The proposed model extracted the temporal features from videos for text generation. The model successfully handled high-level features that helped address flawless sentence generation while adapting the different video modeling approaches. As described above, most models used visual information for caption generation but ignored the semantics information equally beneficial as the visual information. In this regard, Chen et al. [34] proposed an encoder-decoder-based Bi-LSTM conversion gate (BiLSTM-CG) for semantic information extraction and integration. The model used multiple instance learning (MIL) for semantic attribute extraction from a scene frame. The encoder consisted of a 2D and 3D CNN model that encoded the frame and passed these features to the decoder. The Bi-LSTM-CG maps these attributes with textural information for video captioning. Non-visual word prediction is another issue that comprehensively decreases the model's overall accuracy. The case does not end here another problem arises as these non-visual words do not have a visual representation in a video. The words like (a, an the,) are kinds of words that need to be predicated on improving the system's performance. For this purpose, Gao et al. [35] proposed an LSTM to accurately predict the non-visual word from a video. The proposed hierarchical-based LSTM model was designed in such a way that it can understand low-level visual information. This information is combined with high-level context language information to generate a video caption.

2.2.4 Deep Reinforcement Learning Models

Deep Reinforcement Learning (DRL) is a type of learning that works on trial-and-error policy. The agent learns from the environment and acts accordingly. If the decision is good, then it gets rewards. If the decision is wrong, it will be punished [36]. The proposed work of Wang et al. [37] explains hierarchical reinforcement learning (HRL) for video captioning. The proposed model aims to solve the issues faced by reinforcement learning for video captioning. The HRL framework learned the semantic features of a frame. The model also adopts the novel training and deterministic policy. The attention-based model captures the redundant visual feature noise information in video caption, increasing computation cost. Therefore, Chen et al. [38] proposed a reinforcement learning-based model PickNet for the information frame selection. The PickNet works on the standard encoder-decoder framework. It is less likely that DRL models will suffer from training data, overfitting problems, and many others due to their unconventional learning process. Therefore, these models are expected to gain more popularity in video caption generation in the future. The summary of evaluation criteria and models is shown in Table 2.

Table 2: Overview of the caption generation models

Approach	Method	Dataset	Evaluation
Kim et al. [1]	Differentiable neural computer (DNC)	Activitynet	BLEU, METEOR, Consensus-based Image Description Evaluation (CIDEr)
Nabati and Behrad [2]	Multi-sentence	MSVD	n-gram diversity (DIV), BLEU, METEOR, CIDEr, ROUGE
Chen et al. [3]	Multi-modality feature fusion	MSR-VTT, Video And TEX(VATEX)	BLEU, METEOR, CIDEr
Chen et al. [4]	Temporal Deformable Convolutional Encoder-Decoder(TD-ConvED)	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Chen et al. [6]	Dynamic Concept Learner (DCL)	CoLlision Events for Video REpresentation and Reasoning (CLEVRER)	Mean Average Precision (mAP)
Kang and Han [7]	Generate descriptions	Global Action and Interaction(GAI)	BLEU
Young et al. [9]	Denotational similarity metrics	Semantic Textual Similarity(STS)	Pointwise Mutual Information (nPMI)
Aafaq et al. [10]	Short Fourier Transform	MSR-VTT	METEOR, ROUGE
Li et al. [14]	Res-ATT	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Olivastri et al. [15]	Soft-Attention (SA)	MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Li et al. [18]	Dense video captioning	ActivityNet	BLEU, METEOR, CIDEr
Wang et al. [19]	HAN	UCF-101, HMDB-51	
Zheng et al. [20]	SAAT	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Wei et al. [21]	SASUM	SumMe, TVSum	F-Score
Zheng et al. [22]	ORG	MSVD, MSR-VTT, VATEX	BLEU, METEOR, CIDEr, ROUGE

(Continued)

Table 2 (continued)

Approach	Method	Dataset	Evaluation
Chen and Jiang [23]	MGSA	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Zhao et al. [25]	CAM-RNN	MSVD, MSR-VTT, MPII-MD	BLEU, METEOR, CIDEr, ROUGE
Li et al. [26]	MAM-RNN	MSVD, Charades	BLEU, METEOR, CIDEr, ROUGE
Pei et al. [27]	MARN	MSVD,MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Chang et al. [28]	Event-centric multi-modal fusion approach for dense video captioning (EMVC)	ActivityNet, YouCook2	BLEU, METEOR, CIDEr
Niu et al. [29]	Enhanced Gated Recurrent Unit (EN-GRU)	MSVD, MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Li et al. [31]	frame-level features	Youtube- 8M	Global Average Precision (GAP@20)
Zhu and Jiang [32]	Densely Connected Long Short-Term Memory (DenseLSTM)	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Ahmed et al. [33]	Attention-based Bi-LSTM and sequential LSTM (Att-BiL-SL)	MSVD, MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Chen et al. [34]	BiLSTM	MSVD, MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Gao et al. [35]	LSTM adaptive attention	MSVD, MSR-VTT	BLEU, METEOR, CIDEr
Wang et al. [37]	reinforcement learning	MSR-VTT	BLEU, METEOR, CIDEr, ROUGE
Chen et al. [38]	frame picking in video captioning (PickNet)	MSVD, MSR-VTT	BLEU, METEOR, CIDEr, ROUGE

3 Datasets

The dataset plays a critical role in the model training therefore, they are essential for the video analysis. In this section, the paper introduced several datasets used for video analysis. These datasets can be categorized into several classes such as surveillance videos, human activity and actions, social media contributions, movies, and video demonstrations, an overview of each is provided.

3.1 Performance Evaluation of Tracking and Surveillance (PETS2009)

The PETS2009 [39] dataset is crucial for evaluating crowd analysis and surveillance systems, featuring real-world scenarios like urban streets, shopping centers, and train stations in Bristol, UK. It comprises video and sensor data, including infrared, along with manually annotated bounding boxes for individuals in the scenes. PETS2009 contains seven sequences, each around 10 min long, captured from multiple cameras with diverse lighting conditions, crowd densities, and camera motions, challenging the robustness of surveillance systems. This dataset is widely utilized in the research community to assess various algorithms for object detection, tracking, and crowd behavior analysis in computer vision, machine learning, and surveillance research.

3.2 University of California, San Diego (UCSD Dataset)

The UCSD [40] dataset is a widely used dataset for evaluating crowd anomaly detection algorithms. It was collected by the University of California, San Diego, and contains several hours of surveillance video footage of crowded scenes. The dataset includes a variety of crowd scenarios, including normal and anomalous crowd scenes such as fights, accidents, and thefts. The dataset is divided into two parts: Ped1 and Ped2. The Ped1 segment contains 34 video sequences with a total of approximately 2000 frames, while the Ped2 segment contains 16 video sequences with a total of approximately 1000 frames. Each video is labeled with the location and type of anomalous event, as well as the start and end frames of the anomalous event. The dataset also includes ground truth information for the location and number of people in each frame. The UCSD dataset is commonly used in research for evaluating the performance of crowd anomaly detection algorithms and is considered one of the most challenging datasets in this field.

3.3 University of Minnesota (UMN Dataset)

The UMN [41] dataset is a dataset for crowd anomaly detection developed by researchers at the University of Minnesota. The dataset contains a large number of videos of crowded scenes, including both normal and anomalous events [42]. The videos were collected from a variety of sources, including surveillance cameras and YouTube videos. The dataset is designed to be challenging, as it contains a wide variety of crowd scenes, including different types of anomalies, camera angles, and lighting conditions. The UMN dataset is composed of several subsets, each containing a specific type of crowd scene. The subsets include:

UMN-Crowd: This subset contains videos of crowded scenes, including normal and abnormal events such as people congregating, fighting, and falling.

UMN-Mall: This subset contains videos of crowded scenes in a shopping mall, including normal and abnormal events such as people walking, shopping, and loitering.

UMN-Street: This subset contains videos of crowded street scenes, including normal and abnormal events such as people walking, congregating, and loitering. The UMN dataset has been widely used in the research community for evaluating the performance of crowd anomaly detection methods. It is a challenging dataset that can be used to test the robustness and generalization of different algorithms.

3.4 Cooking Datasets

Cooking datasets are collections of videos, images, and textual data that are used to train and evaluate computer vision and natural language processing models for various tasks such as recipe generation, ingredient recognition, and cooking activity recognition. The MAX Plank Institute for Informatics (MP-II) offers a cooking dataset [43] that consists of 65 cooking videos. The dataset contains 888,775 frames for all videos. Textually Annotated Cooking Scenes (TACoS), TacosMultilevel [44,45], and YouCook-II [46] are also popular cooking datasets. These datasets include a variety of cooking activities and recipes, captured from different sources, such as YouTube, and are annotated with detailed information about the ingredients, actions, and other attributes. They are commonly used for training and evaluating models for tasks such as action recognition, recipe generation, and ingredient detection in cooking videos.

3.5 Social Media Datasets

Social media datasets for video captioning refer to collections of videos and their corresponding captions that have been gathered from various social media platforms. These datasets are typically used to train machine learning models for the task of video captioning, which involves automatically generating a written description of the events happening in a video. Social media datasets for video captioning often include a diverse range of videos, from user-generated content to professionally produced videos, and can cover a wide range of topics and genres. These datasets can include videos from platforms such as YouTube, Instagram, TikTok, and Facebook. The captions for the videos in these datasets can be in the form of user-generated subtitles, closed captions, or automatically generated. These datasets are useful for training models for video captioning, video understanding, and natural language processing. The VideoStory dataset [47] is a dataset of 20K social media videos. The dataset aims to target the description generation of long videos. ActivityNet [48] Entities is also a dataset with social media (YouTube) scenes with video descriptions.

3.6 Movie Datasets

The movie dataset contains a collection of video footage from the movies. The goal of this dataset is to provide a resource for training and evaluating video captioning models, which are a type of machine learning model that generates natural language descriptions of the actions and events happening in a video. The dataset typically includes a wide variety of movies with different genres, lengths, and styles, and the videos are accompanied by captions or subtitles that describe the actions and events taking place in the video. The captions are used as the training and testing data for the video captioning models, which are trained to generate captions that accurately describe the content of the video. This dataset can be useful for a wide range of applications, such as video retrieval, video summarization, and video understanding. MPII-MD: MPII-Movie Description Corpus [49] is a 94 Hollywood movie dataset with audio descriptions of these movies. There are 68,337 clips of these moves, along with 68,375 paired sentences. Montreal Video Annotation Dataset (M-VAD) [50] is another movie clips dataset with 48980 videos from 92 different movies and 55,900 sentences.

3.7 Video to Text and Videos in the Wild

MSVD: A Microsoft Video Description (MSVD) [51] is a dataset that contains 1,970 video clips taken from YouTube, along with their associated captions. The dataset was created to support the development of algorithms for video captioning, a task that involves generating a natural language description of the content of a video. The captions in the MSVD dataset were generated by human annotators, and they cover a wide range of topics and styles.

MSR-VTT [16] (Microsoft Research Video to Text) is another dataset that contains 10,000 video clips and 200,000 associated captions. It is also a video captioning dataset, but it contains more diverse and challenging video content than MSVD, such as diverse languages and more complex visual content. The MSR-VTT dataset is designed to help researchers improve the performance of video captioning models by providing more challenging and diverse training data. Both MSVD and MSR-VTT datasets are used for training and evaluating video captioning models and are widely used in the research community. The dataset is created by AMT (Amazon Mechanical Turk) workers. Two other well-known datasets for training models for video caption generation are Charades [43] and Video Titles in the Wild (VTW) [52,53].

3.8 Performance Evaluations

The Mean Absolute Error (MAE) [54] is a metric that is used to measure the accuracy of crowd density estimation methods. It is calculated by taking the average of the absolute differences between the estimated density blob count and the actual density count in the ground truth data across all frames. This is done by first taking the absolute difference between each count in the estimated blob and its corresponding people count in the ground truth data, then averaging the result across all frames, and then calculating the density. The MAE is a widely used metric in evaluating crowd density estimation techniques, as it provides a simple and intuitive measure of the average error in the predicted count. It is an easy-to-understand metric, and it gives an idea of how the estimated count deviates from the actual count on average. It is a measure of the average absolute difference between the estimated density blob count and the actual density count, which is a useful metric for assessing the performance of crowd density estimation methods. The Mean Relative Error (MRE) [54] metric is used to evaluate the performance of different techniques for estimating crowd density. It compares the estimated density maps with the ground truth data by expressing the difference as a percentage of the ground truth density. By doing so, it provides an understanding of how much the predicted density deviates from the true density. This metric is useful in comparing the effectiveness of different crowd density estimation techniques, as it gives a sense of the relative inaccuracy of the predictions concerning the ground truth data. The Root Mean Square Error (RMSE) [54] is a commonly used metric for measuring the accuracy of crowd count estimation methods. It is calculated by taking the square of the difference between the estimated count in a video frame and the ground truth count for that frame, averaging the result across all frames, and then taking the square root of the average. The RMSE provides a measure of the average deviation of the estimated count from the ground truth count and gives an insight into the overall performance of the crowd count estimation technique. It is a commonly used metric for evaluating the effectiveness of different crowd count estimation methods, as it offers a comprehensive measure of the average deviation of the predicted count on a video frame from the actual ground truth data. The performance of video descriptors can be separated into automatic and human evaluations. Automatic evaluations are accomplished using evaluation metrics like BLEU [55], METEOR [56], and WMD [57] are some of the most commonly used evaluation metrics. On the other hand, the human evaluation method is also used to judge the quality of model-generated video captions. But for this purpose, humans should be experts in the relevant field. The summary of all the evaluation metrics is given in Table 3.

Table 3: The details of video evaluation metrics

Metrics name	Used for	Evaluation technique
MAE	Accuracy	Absolute error
MRE	Accuracy	Relative error

(Continued)

Table 3 (continued)

Metrics name	Used for	Evaluation technique
RMSE	Accuracy	Mean square error
BLEU [55]	Machine translation	n-gram precision method
METEOR [56]	Machine translation	n-gram synonyms comparing
WMD [57]	Documents similarity	Word2vec
ROUGE [58]	Documents summarization	n-gram recall
Semantic propositional image caption evaluation (SPICE) [59]	Image captioning	n-gram similarity

4 Latest Approaches in the Video Description

Automatic video description (AVD) has gained massive success after using deep learning methods. Still, the performance of these models is deficient compared to human-generated captions. Here is the list of directions that can improve the performance of video descriptors.

4.1 Visual Reasoning and Question Answering

Although the Video Visual Question Answering (VQA) is early, it is a future direction to explore. The Video Q&A is another visual problem. In this methodology, the model gave the answer and the reason for its answer. Consider a scenario where a video displays a parking sign. One potential inquiry directed at the model could be, "Is parking allowed in this area?" The expected response might be "Yes." Subsequently, a follow-up question could be posed: "What is the reason for the permission to park?" To which the model is expected to reply, "Because a parking sign is present in this location." In this approach object detection and model reaction with context is more crucial compared with dense video understanding. In addition, video chat-related models combined with large language models also deal with visual questioning and improve the quality of chat information [60].

Visual reasoning [61] also deals with the logical context of the detection and understanding of is next level of sentence generation. The main visual reasoning datasets and relevant best-approached models are summarized in Table 4.

Table 4: The details of video evaluation metrics

Dataset	Best model	Limitation
Natural language visual reasoning (NLVR) [62]	Multiway transformer (BEiT-3) [63], Visual bidirectional encoder representations from transformers (VisualBERT) [64]	Limited diversity
Winoground [65]	Visual question (VQ2) [66]	Limited context

(Continued)

Table 4 (continued)

Dataset	Best model	Limitation
Gamified association benchmark to challenge vision-and-language (WinoGAViL) [67]	Vision-and-language transformer (ViLT) [68]	Limited image variety
Bongard-openworld [69]	Simple neural attentIve Learner (SNAIL) [70]	Limited size
Visual spartial reasoning (VSR) [71]	Learning cross-modality Encoder representations from transformers (LXMERT) [73]	Annotation difficulty potential ambiguity
Visual analogies of situation recognition (VASR) [72]	Shifted window (Swin) [74]	

4.2 Visual Dialogue

Similar to the audio dialogue found in Siri, Alexa, and Hello Google [75], video dialogue is also a flourishing field. With the interaction with machines and robots, the visual dialogue would be constructive. Real-time video description generation and understanding of the context are more important in this manner.

These AI assistants, in conjunction with compatible devices, had started integrating visual analysis and understanding into their functionalities. While these systems are making strides in understanding real-time video, their capabilities may vary depending on the device, the specific AI assistant, and privacy considerations. They generally rely on machine learning models that analyze visual data and derive information or actions based on that analysis.

4.3 Audio and Video

Most of the explained models focus on the video semantic feature extraction, but the audio can be a perfect candidate to work with video to improve the video descriptor performance. The audio can by providing the video background information [76,77]. Audio can serve as an ideal complement to enhance the performance of video descriptors. When combined with video data, audio can significantly enrich the overall understanding and interpretation of the content. The incorporation of audio cues allows for a more comprehensive analysis of the visual elements in a video. For instance, the sounds within a video, whether ambient noise, dialogue, or music, can provide critical contextual information. Integrating audio data with video descriptors enables a more holistic and nuanced interpretation of the content, leading to improved accuracy in recognizing and describing various elements, scenes, or actions within the video. This fusion of audio and visual data harnesses the synergy between the senses, providing a more robust and comprehensive foundation for video analysis and description algorithms, ultimately advancing the capabilities of these systems. The sound of the sea, traffic sound, rain, and fire sounds. All these sounds help to generate the video description [78]. The audio and video feature content fusion enhances the events-related understanding in the video sequence. The work Ibrahimi et al. [79] introduced separate audio and video attention models instead of audiovisual attention blocks.

5 Future Direction to Improve the Context

There are many ongoing approaches available for understanding video content using methods of deep learning and sequence models. Most of them understand video events, and inappropriate content like specific scene understanding, text generation, content localization, etc. Content generation through video content is widely concerning in many fields especially online content such as video sharing and social media. Human activity detection is one of the major branches of video analysis. Action recognition or identification is a widely used application in video surveillance. In this approach, the model attempts to generate sentences by understanding video contents through isolating key scene filtering. Previously explained some useful techniques to extract spatial information from the given video sequence. Key scene localization is introduced as an extension of morphological image processing by analyzing hue and saturation features in every frame. The methods of pre-data processing for video analysis are explained in this chapter further. The transfer learned CNN and 3D convolution (C3D) feature fusion technique can achieve better performance over the tested dataset. The resultant language generation model is based on the thought vector [80] by generating block-wise C3D convolution over the video image sequence. Finally, this paper concludes new research directions and points out challenges in the area of human activity recognition especially the latest online language models such as ChatGPT [81].

5.1 Proposed Model for Sentence Content Improvement

In this approach, the goal is to develop a language model that can understand the context of a video. The model is based on a classical encoder-decoder architecture, and it uses a combination of the C3D feature vector and the decoder LSTM network for language generation. The decoder LSTM network incorporates key scene attention features, which are generated by detecting the key scenes in the video. In the feature extraction process, relevant key scene frame-blocks are only processed for generating the final feature vector. The C3D Network is fine-tuned by removing the final fully connected network. This allows the network to learn specific features of the video that are relevant to the task of language generation. The key scene detection is performed by estimating the hue and saturation density transition level. To prepare the dataset for this model, the video is first sliced into frames, and these frames and identified keyframes along 16 blocks of bunch fed block-wise into the encoder model and saved. The generated blocks' C3D features are summed along the vertical axes to produce the resultant vector for the language model. Finally, the model is capable of understanding the context of a video and generating language that accurately reflects the content of the video. The proposed model for feature extraction is shown in Fig. 6.

The encoding model starts with the classical 3D convolution operations to perform as a feature extractor. In this process, a fine-tuned C3D model is used with 4096 feature vector dimensions as output. The model input block is kept as 16 frames while extracting features from pre-trained model dimensions. The trained model is based on the Sports-1 M [82] action dataset classification task.

5.2 Key-Scene Extraction

The video sequence (n frames) is analyzed frame by frame to identify variations in hue and saturation levels that indicate a linear transition. Distortions in the linear transition can be used to identify scene separations in the video. The sample video of the MSR-VTT hue and saturation levels graph plot is shown in Fig. 7.

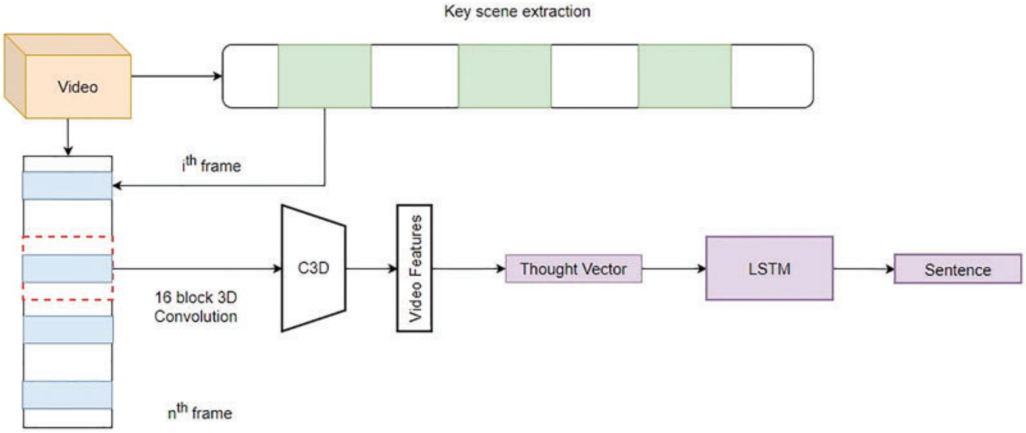

Figure 6: The model flow diagram of the video captioning

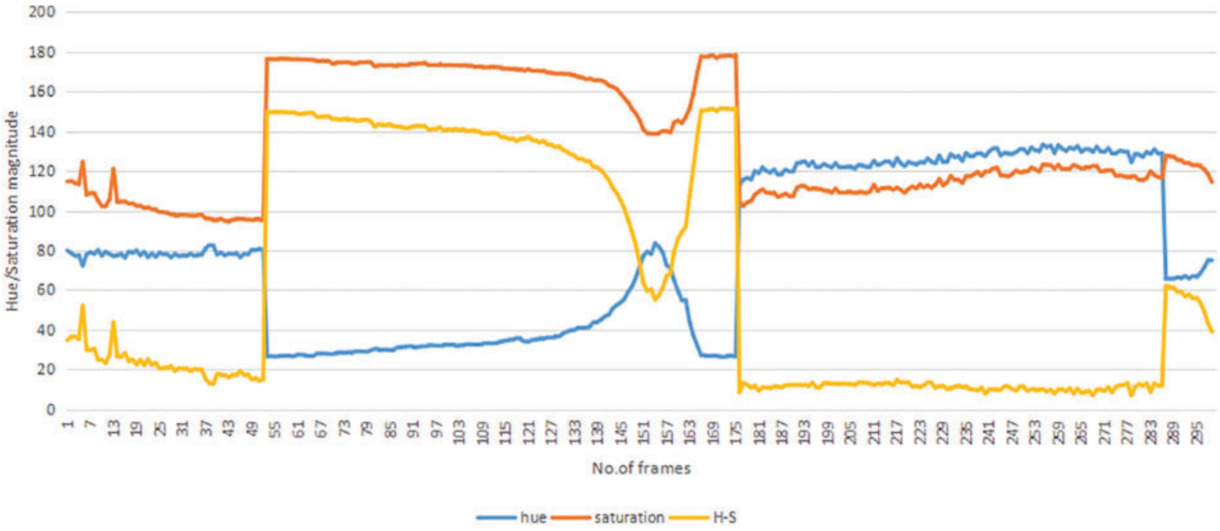

Figure 7: Hue/Saturation magnitude graph for sample video (MSR-VTT)

Hue and saturation both relate to the dominating color in a certain area of the video frame and the degree of that color's strength. Along the video sequence, the hue (blue) and saturation (red) magnitudes vary similarly, and thus clearly demonstrate the dramatic scene shift in a video. The key frame extraction relevant to the hue/saturation is shown in Fig. 8.

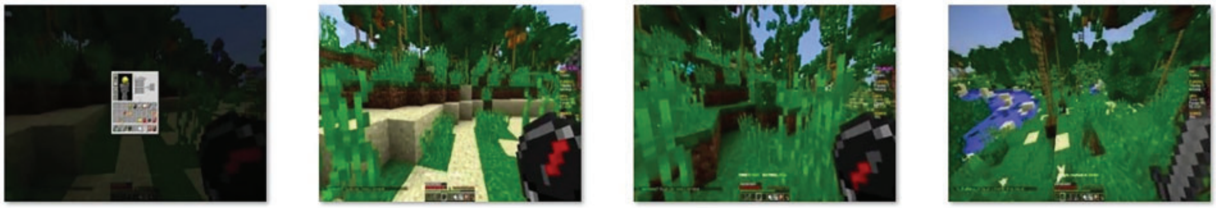

Figure 8: Sample key scene extracted from video MSR-VTT (200)

The preparation of caption ground truth data is an important step in the video captioning task. In this approach, a random sentence is selected from the MSR-VTT video captioning dataset, and four tokens are added to the vocabulary: <pad>, <start>, <end>, and <unk>. These tokens serve specific purposes in the decoding process.

The length of the sentence is kept at 20 for the decoder input. This length is determined based on the maximum expected length of the captions in the dataset. Each word in the sentence starts with a <start> token and ends with an <end> token. This is essential for the decoder section, as the model needs to know where to stop when generating the caption. The word embedding is initiated using word2vec [61] encoding and the embedding vector size is kept at 512 during both training and inference. The use of word2vec encoding ensures that each word in the sentence is represented as a unique vector, allowing the model to distinguish between words with different meanings. The embedding vector size of 512 provides a sufficiently large representation of the words in the sentence, allowing the model to learn complex relationships between words.

In this approach, the Adam optimizer is used, with a cross-entropy loss function and a learning rate of 0.0001. The number of batches is set to 10, which determines the number of times the model will update its parameters during training. The thought vector size is reduced to a dimension of 512 for the LSTM operations. The performance comparison over the MSR-VTT dataset with predicted and ground truth sentences is shown in Fig. 9.

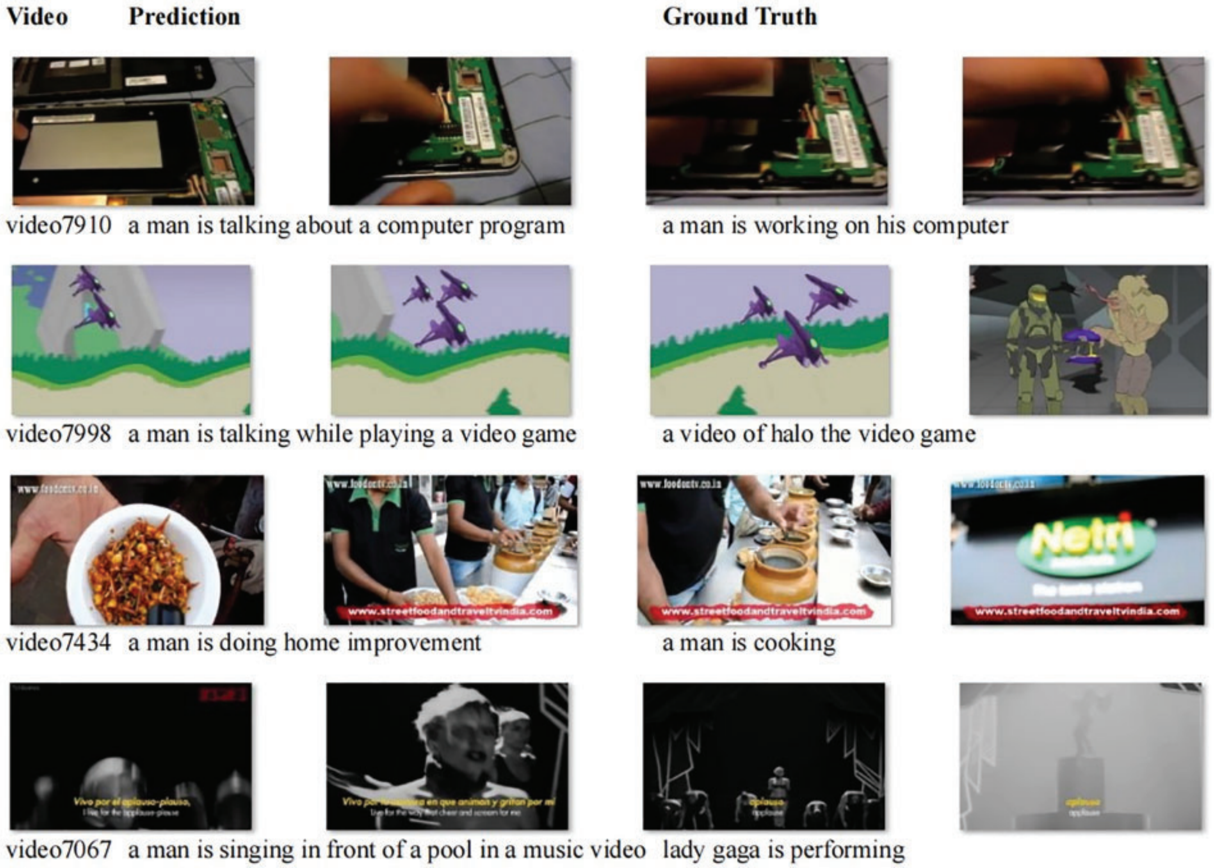
Video
Prediction
Ground Truth
video7910
a man is talking about a computer program
a man is working on his computer
video7998
a man is talking while playing a video game
a video of halo the video game
video7434
a man is doing home improvement
a man is cooking
video7067
a man is singing in front of a pool in a music video
lady gaga is performing

Figure 9: Performance comparison with MSR-VTT dataset

5.3 Sentence Improvement through Object Detection

The MSR-VTT dataset annotation provides extensive information about the corpus, but there are some gaps in the information generated. Our strategy is to enhance the generated sentences by incorporating key scene objects. To achieve this, extract the object detection set into a separate set to improve the content. To extract the key-scene information, the YOLO3 [83] object detection model was employed with the COCO dataset [8]. This combination allows us to identify and extract important objects in each scene effectively. The simplicity of YOLO3 makes it stand out, making it simpler to integrate into the proposed model, especially for those who prefer a traditional approach to object identification. Since it has undergone thorough testing and has been used in a variety of applications, this version is distinguished by its stability and maturity. Additionally, YOLOv3 is appropriate for contexts with limited hardware because it uses fewer computing resources than its later versions. The video processing needs more computational power to train the models and YOLO3 perfectly balances with complexity. Also, it balances efficiency and precision, making it suitable for low-latency and real-time applications such as video surveillance.

The final sentence;

Generated text = Predicted text + Scene objects

The generated sentence and the associated object list text data enable a clear performance improvement. This comparison serves to highlight areas of missing data for improvement in the generating process and offers useful insights into the precision and efficacy of the generated sentences. Sample predicted and reference sentence with object list of MSR-VTT test video (7012) Predicted Text = “a man is standing in the field” Ground Truth = “a man is sitting in the room and talking”, “a man in glasses talks about matters of health”, “ a man in glasses talks about matters of science”, “a man in a brown blazer discussing mental health”, “a man with a blue shirt and glasses is talking”.

Object List = [“person”, “tie”, “chair”], [“person”, “tie”, “sofa”]

The generated text can be used to understand dense events related to the video sequence and sample results present relevant with random ground truth selections (with five sentences). The captured scene segments related to the MSR-VTT (7012) are shown in Fig. 10. After language regeneration using prompt engineering models, Generated sentence = “A person, dressed in a tie, is standing in the field near a chair”. The generated text contains more details about surrounding objects and their relationships.

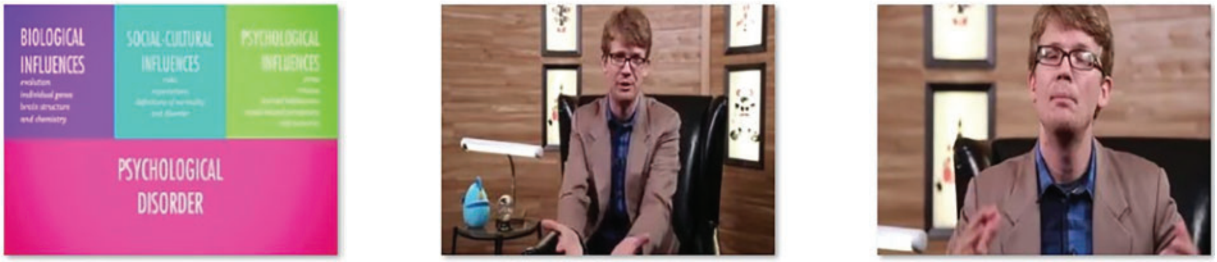

Figure 10: Sample scene list for predicted video MSR-VTT (7012)

6 Conclusion

This paper presents a detailed review of video description research, from classical models to advanced deep learning models. The functionality of various machine learning models employed in video captioning is also highlighted. CNN, RNN, and LSTM are the leading deep learning models improving the video description domain. The paper reviewed the most popular datasets used for training and testing video description models. These datasets have the videos and translated sentences for model training and testing purposes. And also highlighted the most commonly used automatic evaluation metrics to evaluate video descriptions. Lastly, this paper explained some recommendations to guide thinking out of the box. Future trends initially cover how to identify key scenes in a video sequence and how to choose the position of a scene using hue, saturation, and density. The construction of feature vectors is followed by a discussion of feature extraction and language modeling as encoder-decoder approaches. By reference to the key scene index value, needless repetition of video segments can be eliminated while extracting block-level C3D features. Sentence reconstruction methodology object detection is covered in the following section, and sample sentences are generated. The 3D convolution requires greater computational power compared to 2D convolution methods. Consequently, training the simultaneous encoder-decoder model can enhance model parameters. Finally, this work utilizes a low-performance computational setup for training a scene-based model to improve predictions.

Acknowledgement: Not applicable.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: conceptualization and design, supervision, Yunqi Lei; draft manuscript preparation, formal analysis, A.S. Gezawa; validation, visualization, results analysis, writing original draft, E.M.C.L. Ekanayake. All authors have read and agreed to the published version of the manuscript.

Availability of Data and Materials: The data that support the findings of this study are openly available in MSR-VTT at https://www.kaggle.com/datasets/vishnutheepb/msrvtt.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] J. Kim, I. Choi, and M. Lee, "Context aware video caption generation with consecutive differentiable neural computer," *Electron.*, vol. 9, no. 7, pp. 1–15, 2020. doi: 10.3390/electronics9071162.

[2] M. Nabati and A. Behrad, "Multi-sentence video captioning using content-oriented beam searching and multi-stage refining algorithm," *Inf. Process. Manag.*, vol. 57, no. 6, pp. 102302, 2020. doi: 10.3390/electronics9071162.

[3] S. Chen *et al.*, "MM21 Pre-training for video understanding challenge: Video captioning with pretraining techniques," in *Proc. ACM Int. Conf. on Multimedia*, China, 2021, pp. 4853–4857.

[4] J. Chen, Y. Pan, Y. Li, T. Yao, H. Chao and T. Mei, "Temporal deformable convolutional encoder-decoder networks for video captioning," in *Proc. AAAI Conf. on Artif. Intell.*, Hawaii, USA, 2019, pp. 8167–8174.

[5] L. Torrey and J. Shavlik, "Transfer learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*, pp. 242–254, 2010. doi: 10.4018/978-1-60566-766-9.

[6] Z. Chen, J. Mao, J. Wu, K. Y. K. Wong, J. B. Tenenbaum and C. Gan, "Grounding physical concepts of objects and events through dynamic visual reasoning," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–20.

[7] S. H. Kang and J. H. Han, "Video captioning based on both egocentric and exocentric views of robot vision for human-robot interaction," *Int. J. Soc. Robot.*, vol. 15, no. 4, pp. 631–641, 2023. doi: 10.1007/s12369-021-00842-1.

[8] Z. Parekh, J. Baldridge, D. Cer, A. Waters, and Y. Yang, "Crisscrossed captions: Extended intramodal and intermodal semantic similarity judgments for MS-COCO," in *Proc. 16th Conf. of the European. Chap. Assoc. for Comput. Linguist.: Main Vol.*, 2021, pp. 2855–2870.

[9] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," *Trans. Assoc. Comput. Linguist.*, vol. 2, pp. 67–78, 2014. doi: 10.1162/tacl_a_00166.

[10] N. Aafaq, N. Akhtar, W. Liu, S. Z. Gilani, and A. Mian, "Spatio-temporal dynamics and semantic attribute enriched visual encoding for video captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, CA, USA, 2019, pp. 12479–12488.

[11] D. Koller, N. Heinze, and H. H. Nagel, "Algorithmic characterization of vehicle trajectories from image sequences by motion verbs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, HI, USA, 1991, pp. 90–95.

[12] M. Brand, "The "Inverse Hollywood problem": From video to scripts and storyboards via causal analysis," in *Proc. Nat. Conf. Artif. Intell. Ninth Conf. Innovat. App. Arti. Intell.*, RI, USA, 1997, pp. 132–137.

[13] X. Zhu, H. Li, and T. Su, "Autonomous complex knowledge mining and graph representation through natural language processing and transfer learning," *Autom. Constr.*, vol. 155, pp. 105074, 2023. doi: 10.1016/j.autcon.2023.105074.

[14] X. Li, Z. Zhou, L. Chen, and L. Gao, "Residual attention-based LSTM for video captioning," *World Wide Web*, vol. 22, no. 2, pp. 621–636, 2019. doi: 10.1007/s11280-018-0531-z.

[15] S. Olivastri, G. Singh, and F. Cuzzolin, "End-to-end video captioning," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Seoul, Korea, 2019, pp. 1474–1482.

[16] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annual Meet. Assoc. Comput. Linguis.: Human Lang. Tech.*, Oregon, USA, 2011, pp. 190–200.

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

[18] Y. Li, T. Yao, Y. Pan, H. Chao, and T. Mei, "Jointly localizing and describing events for dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, UT, USA, 2018, pp. 7492–7500.

[19] Y. Wang, S. Wang, J. Tang, N. O'Hare, Y. Chang and B. Li, "Hierarchical attention network for action recognition in videos," arXiv preprint arXiv:1607.06416, 2016.

[20] Q. Zheng, C. Wang, and D. Tao, "Syntax-aware action targeting for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, WA, USA, 2020, pp. 13093–13102.

[21] H. Wei, B. Ni, Y. Yan, H. Yu, and X. Yang, "Video summarization via semantic attended networks," in *Proc. AAAI Conf. Artif. Intell.*, Luuisiana, USA, 2018, pp. 216–223.

[22] Z. Zhang *et al.*, "Object relational graph with teacher-recommended learning for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, WA, USA, 2020, pp. 13275–13285.

[23] S. Chen and Y. G. Jiang, "Motion guided spatial attention for video captioning," in *Proc. AAAI Conf. Artif. Intell.*, Hawaii, USA, 2019, pp. 8191–8198.

[24] N. Aafaq, A. Mian, W. Liu, S. Z. Gilani, and M. Shah, "Video description: A survey of methods, datasets, and evaluation metrics," *ACM Comput. Surv. (CSUR)*, vol. 52, no. 6, pp. 1–37, 2019. doi: 10.1145/3355390.

[25] B. Zhao, X. Li, and X. Lu, "CAM-RNN: Co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, 2019. doi: 10.1109/TIP.2019.2916757.

[26] X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based RNN for video captioning," in *Proc. Twenty-Sixth Int. Joint Conf. Artif. Intell.*, Melbourne, Australia, 2017, pp. 2208–2214.

[27] W. Pei, J. Zhang, X. Wang, L. Ke, X. Shen and Y. W. Tai, "Memory-attended recurrent network for video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, CA, USA, 2019, pp. 8339–8348.

[28] Z. Chang, D. Zhao, H. Chen, J. Li, and P. Liu, "Event-centric multi-modal fusion method for dense video captioning," *Neural Netw.*, vol. 146, pp. 120–129, 2022. doi: 10.1016/j.neunet.2021.11.017.

[29] T. Z. Niu *et al.*, “A multi-layer memory sharing network for video captioning,” *Pattern Recognit.*, vol. 136, pp. 109202, 2023. doi: 10.1016/j.patcog.2022.109202.

[30] E. M. Mercha and H. Benbrahim, “Machine learning and deep learning for sentiment analysis across languages: A survey,” *Neurocomputing*, vol. 531, pp. 195–216, 2023. doi: 10.1016/j.neucom.2023.02.015.

[31] F. Li *et al.*, “Temporal modeling approaches for large-scale youtube-8m video understanding,” arXiv preprint arXiv:1707.04555, 2017.

[32] Y. Zhu and S. Jiang, “Attention-based densely connected LSTM for video captioning,” in *Proc. ACM Int. Conf. Multimed.*, Nice, France, 2019, pp. 802–810.

[33] S. Ahmed *et al.*, “Att-BiL-SL: Attention-based Bi-LSTM and sequential LSTM for describing video in the textual formation,” *Appl. Sci.*, vol. 12, no. 1, pp. 1–19, 2022. doi: 10.3390/app12010317.

[34] S. Chen, X. Zhong, L. Li, W. Liu, C. Gu and L. Zhong, “Adaptively converting auxiliary attributes and textual embedding for video captioning based on BiLSTM,” *Neural Process. Lett.*, vol. 52, no. 3, pp. 2353–2369, 2020. doi: 10.1007/s11063-020-10352-2.

[35] L. Gao, X. Li, J. Song, and H. T. Shen, “Hierarchical LSTMs with adaptive attention for visual captioning,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 5, pp. 1112–1131, 2020. doi: 10.1109/TPAMI.2019.2894139.

[36] P. B. Peter Henderson, R. Islam, D. M. Joelle Pineau, and D. Precup, “Deep reinforcement learning that matters,” in *Proc. AAAI Conf. Artif. Intell. (AAAI-18)*, Luuisiana, USA, 2018, pp. 3207–3214.

[37] X. Wang, W. Chen, J. Wu, Y. F. Wang, and W. Y. Wang, “Video captioning via hierarchical reinforcement learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Salt Lake City, Utah, USA, 2018, pp. 4213–4222.

[38] Y. Chen, S. Wang, W. Zhang, and Q. Huang, “Less is more: Picking informative frames for video captioning,” in *Proc. European. Conf. Comput. Vis. (ECCV)*, Munich, Germany, 2018, pp. 367–384.

[39] PETS 2009 Benchmark Data, 2009. Accessed: Dec. 06, 2023. [Online]. Available: http://cs.binghamton.edu~mrldata/pets2009.html

[40] UCSD Anomaly Detection Dataset, 2022. Accessed: Nov. 07, 2022. [Online]. Available: http://www.svcl.ucsd.edu/projects/anomaly/dataset.html

[41] Monitoring Human Activity-Action Recognition, 2022. Accessed: Dec. 06, 2023. [Online]. Available: http://mha.cs.umn.edu/proj_recognition.shtml

[42] A. Jan and G. M. Khan, “Real world anomalous scene detection and classification using multilayer deep neural networks,” *Int. J. Interact. Multimed. Artif. Intell.*, vol. 8, no. 2, pp. 158–167, 2023. doi: 10.9781/ijimai.2021.10.010.

[43] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, “A database for fine grained activity detection of cooking activities,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, 2012, pp. 1194–1201.

[44] P. Das, C. Xu, R. F. Doell, and J. J. Corso, “A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, 2013, pp. 2634–2641.

[45] A. Rohrbach *et al.*, “Coherent multi-sentence video description with variable level of detail,” in *Proc. German Conf. Pattern Recognit. (CGPR)*, Munster, Germany, 2014, pp. 184–195.

[46] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, “VideoBERT: A joint model for video and language representation learning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Long Beach, CA, USA, 2019, pp. 7463–7472.

[47] S. Gella, M. Lewis, and M. Rohrbach, “A dataset for telling the stories of social media videos,” in *Proc. Conf. on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018, pp. 968–974.

[48] L. Ruan, J. Chen, Y. Song, S. Chen, and Q. Jin, “Team RUC_AIM3 technical report at activityNet 2021: Entities object localization,” arXiv preprint arXiv:2106.06138, 2021.

[49] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, “A dataset for movie description,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, 2015, pp. 3202–3212.

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

[51] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional long-short term memory for video description," in *Proc. ACM Int. Conf. Multimed.*, Amsterdam, Netherland, 2016, pp. 436–440.

[52] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Proc. European Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherland, 2016, pp. 510–526.

[53] K. H. Zeng, T. H. Chen, J. C. Niebles, and M. Sun, "Title generation for user generated videos," in *Proc. European Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherland, 2016, pp. 609–625.

[54] T. O. Hodson, "Root-mean-square error (RMSE) or mean absolute error (MAE): When to use them or not," *Geosci. Model Dev.*, vol. 15, no. 14, pp. 5481–5487, 2022. doi: 10.5194/gmd-15-5481-2022.

[55] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proc. Annual Meeting Assoc. Comput. Linguist.*, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.

[56] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. Annual Meeting Assoc. Comput. Linguist.*, Ann Arbor, Michigan, USA, 2007, pp. 228–231.

[57] M. J. Kusner, Y. Sun, N. I. Kolkin, and K. Q. Weinberger, "From word embeddings to document distances," in *Proc. Int. Conf. Mach. Learn. (ICML)*, Lille, France, 2015, pp. 957–966.

[58] C. Y. Lin "Rouge: A package for automatic evaluation of summaries," in *Text Summarization Branches Out*, CA, USA: University of Southern California, 2004, pp. 74–81.

[59] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: Semantic propositional image caption evaluation," in *Proc. European Conf. Comput. Vis. (ECCV)*, Amsterdam, Netherland, 2016, pp. 382–398.

[60] K. Li *et al.*, "VideoChat: Chat-centric video understanding," arXiv preprint arXiv:2305.06355, 2023.

[61] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Long Beach, CA, USA, 2019, pp. 6693–6702.

[62] A. Suhr, M. Lewis, J. Yeh, and Y. Artzi, "A corpus of natural language for visual reasoning," in *Proc. Annual Meeting Assoc. Comput. Linguist.*, Vancouver, CA, 2017, pp. 217–223.

[63] W. Wang *et al.*, "Image as a foreign language: BEIT pretraining for vision and vision-language tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, CA, 2023, pp. 19175–19186.

[64] L. H. Li, M. Yatskar, D. Yin, C. J. Hsieh, and K. W. Chang, "VisualBERT: A simple and performant baseline for vision and language," arXiv preprint arXiv:1908.03557, 2019.

[65] T. Thrush *et al.*, "Winoground: Probing vision and language models for visio-linguistic compositionality," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orleans, LA, USA, 2022, pp. 5228–5238.

[66] M. Yarom *et al.*, "What you see is what you read? improving text-image alignment evaluation," arXiv preprint arXiv:2305.10400, 2023.

[67] Y. Bitton *et al.,* "WinoGAViL: Gamified association benchmark to challenge vision-and-language models," in *Advances in Neural Information Processing Systems (NeurIPS)*, New Orleans, Louisiana, USA: Neural Information Processing Systems Foundation, Inc., 2022, pp. 1–16.

[68] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[69] R. Wu *et al.*, "Bongard-OpenWorld: Few-shot reasoning for free-form visual concepts in the real world," arXiv preprint arXiv:2310.10207, 2023.

[70] N. Mishra, M. Rohaninejad, X. Chen, and P. Abbeel, "A simple neural attentive meta-learner," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, Vancouver, CA, 2018, pp. 1–17.

[71] F. Liu, G. Emerson, and N. Collier, "Visual spatial reasoning," *Trans. Assoc. Comput. Linguist.*, vol. 11, pp. 635–651, 2023. doi: 10.1162/tacl_a_00566.

[72] Y. Bitton, R. Yosef, E. Strugo, D. Shahaf, R. Schwartz and G. Stanovsky, "VASR: Visual analogies of situation recognition," in *Proc. AAAI Conf. Artif. Intell.*, Washington DC, USA, vol. 37, 2023, pp. 241–249.

[73] H. Tan and M. Bansal, "LXMert: Learning cross-modality encoder representations from transformers," in *Proc. Conf. on Empirical Methods in Natural Language Processing and 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, 2019, pp. 5100–5111.

[74] Z. Liu *et al.*, "Video swin transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, New Orlance, LA, USA, 2022, pp. 3192–3201.

[75] M. B. Hoy, "Alexa, Siri, Cortana, and more: An introduction to voice assistants," *Med. Ref. Serv. Q.*, vol. 37, no. 1, pp. 81–88, 2018. doi: 10.1080/02763869.2018.1404391.

[76] T. Han, M. Bain, A. Nagrani, and A. Zisserman, "AutoAD II: The sequel–who, when, and what in movie audio description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, CA, 2023, pp. 13645–13655.

[77] H. Zhang, X. Li, and L. Bing, "Video-LLaMA: An instruction-tuned audio-visual language model for video understanding," arXiv preprint arXiv:2306.02858, 2023.

[78] N. Shvetsova, A. Kukleva, X. Hong, C. Rupprecht, B. Schiele and H. Kuehne, "HowToCaption: Prompting LLMs to transform video annotations at scale," arXiv preprint arXiv:2310.04900, 2023.

[79] S. Ibrahimi, X. Sun, P. Wang, A. Garg, A. Sanan and M. Omar, "Audio-enhanced text-to-video retrieval using text-conditioned feature alignment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Vancouver, CA, 2023, pp. 12054–12064.

[80] Z. Yang, Y. Yuan, Y. Wu, R. Salakhutdinov, and W. W. Cohen "Review networks for caption generation," in *Advances in Neural Information Processing Systems 29 (NIPS 2016)*, 2016, pp. 2369–2377.

[81] C. Wu, S. Yin, W. Qi, X. Wang, Z. Tang and N. Duan, "Visual ChatGPT: Talking, drawing and editing with visual foundation models," arXiv preprint arXiv:2303.04671, 2023.

[82] A. Karpathy and T. Leung, "DeepVideo," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Columbus, OH, USA, 2014, pp. 10–20.

[83] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," arXiv preprint arXiv:1804.02767, 2018.