Combinatorial Video Captioning With Deep Learning

[1]Ms M Vineela, [2]Angadi Ahalya, [3]Cheelam Aishwarya Laxmi

[1]Assistant professor, Department of CSE, Bhoj Reddy Engineering College for Women, India

[2,3]B.Tech Students, Department of CSE, Bhoj Reddy Engineering College for Women, India

ABSTRACT

Video captioning is the task of generating natural language descriptions for videos by analyzing visual scenes, objects, and actions. Unlike video subtitling, which transcribes spoken dialogue, video captioning provides a comprehensive interpretation of all visual elements. Traditional approaches relied on rule-based and feature-based methods, which struggled with complex videos due to their rigidity and lack of contextual understanding.

Modern techniques leverage deep learning models, such as Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), to extract video features and generate captions. Recent advancements focus on weakly supervised dense video captioning, which generates descriptions without predefined key events. This approach is particularly useful for long, untrimmed videos where multiple overlapping events occur, improving event recognition and caption accuracy. By combining event captioning with caption localization, this method enhances both contextual understanding and flexibility in video captioning tasks.

INTRODUCTION

Video captioning is the process of generating natural language descriptions for videos by understanding scenes, objects, and actions. Unlike video subtitling, which converts spoken audio to text, video captioning describes everything happening in the video visually. This field combines computer vision and natural language processing

modern approach

Modern approaches use deep learning and machine learning, employing models like CNNs and RNNs to extract features from video frames and generate captions

In recent times, many approaches have been used to tackle this problem, like dividing the task into smaller parts, using pre-trained models, or combining different techniques. However, most methods still rely on knowing the key events beforehand, which makes the task easier.

Many methods have been used to make this task easier, like bottom-up and top-down approaches, or using pre-trained models. However, a new approach called weakly supervised desnse video captioning shows promise.This method generates captions without knowing specific key events in advance

Deals with long, untrimmed videos where many things happen at once, and it's tough to provide clear captions for every event. Most existing approaches focus on short videos, but dense video captioning is aimed at longer ones, usually more than 120 seconds. In such videos, many events overlap, making it harder to generate good captions.

The model uses a unique combination of **event captioning** and **caption localization**, working together to improve both event recognition and caption generation

Earlier approach

Earlier methods relied on hand-crafted rules, but they struggled with large, complex videos.

Rule-based approaches used pre-set templates to generate captions, making them rigid and unable to adapt to diverse scenarios. Feature-based methods extracted visual, motion, and audio cues from videos, requiring significant manual effort to design these features. Techniques like Bag-of-Words and Bag-of-Visual-Words treated videos as collections of features but failed to capture the sequence or context of actions. Hidden Markov Models (HMMs) were used to model action sequences but were limited to specific activities and needed labeled datasets.

Template matching combined detected objects and actions with predefined sentence structures but lacked flexibility for novel or complex video content. Overall, these methods struggled with generalization, contextual understanding, and producing dynamic descriptions, paving the way for modern deep learning approaches.

Video captioning involves generating natural language descriptions for videos by analyzing visual scenes, objects, and actions. Traditional approaches used rule-based methods, feature extraction techniques, and statistical models like Hidden Markov Models (HMMs). These methods relied on predefined templates, handcrafted rules, and labeled datasets to generate captions. Additionally, CNN-LSTM architectures have been widely used in modern video captioning models. CNNs extract visual features from video frames, while LSTMs process sequential data to capture temporal dependencies. Various datasets, such as MSVD, MSR-VTT, ActivityNet, and MPII-MD, have been used for training and evaluating video captioning models.

2-LITERATURE SURVEY

wangyu choi, jiasi chen, jongwon yoon conducted a comparative analysis of the most to understand the correlation between video and language within the domain of computer vision. Its objective is to identify multiple temporal bins that encompass significant scenes within a video, followed by providing informative descriptions for each bin in the form of a single sentence. This task holds high potential for application in a variety of video analytics tasks.

Wei chen In this paper, we propose TopicDVC, a novel dense video captioning framework. TopicDVC applies the topic information generated by the topic generator to guide the model in generating more coherent captions. Experiments on the ActivityNet Captions dataset demonstrate that leveraging the topics generated by the diffusion model significantly improves the performance of dense video captioning, producing more accurate and coherent captions. proposed method can generate captions that are richer in contents and can compete with state-of-the-art method without explicitly using video-level features as input.

Andrew shin, Katsunori Ohnishi, Tatsuya Harada are focused on generating single input of aggregated features, which hardly deviates from image captioning process and does not fully take advantage of dynamic contents present in videos. attempt to generate video captions that convey richer contents by temporally segmenting the video with action localization, generating multiple captions from multiple frames, and connecting them with natural language processing techniques, in order to generate a story-like caption.

3. METHODOLOGY

VIDEO CLIP CAPTIONING

video captioning aim to describe a video using either a single sentence or more detailed text. Many of these methods follow an encoder-decoder structure, where the encoder typically uses a Convolutional Neural Network (CNN) to extract visual features, which are then processed by a Recurrent Neural Network (RNN) to capture temporal information. The decoder, often an RNN, generates text predictions token by token to form a coherent sentence. However, given that videos often contain rich content with multiple events, recent methods focus on generating comprehensive paragraphs. These paragraphs provide detailed descriptions, with each sentence dedicated to a specific event or occurrence within the video, offering a more complete narrative.

Convolutional Neural Network

A Convolutional Neural Network (CNN) is a type of deep learingalgorithmthat is particularly well-suited for image recognition and processing tasks. It is made up of multiple layers, including convolutional layers, pooling layer and fully connected layers.

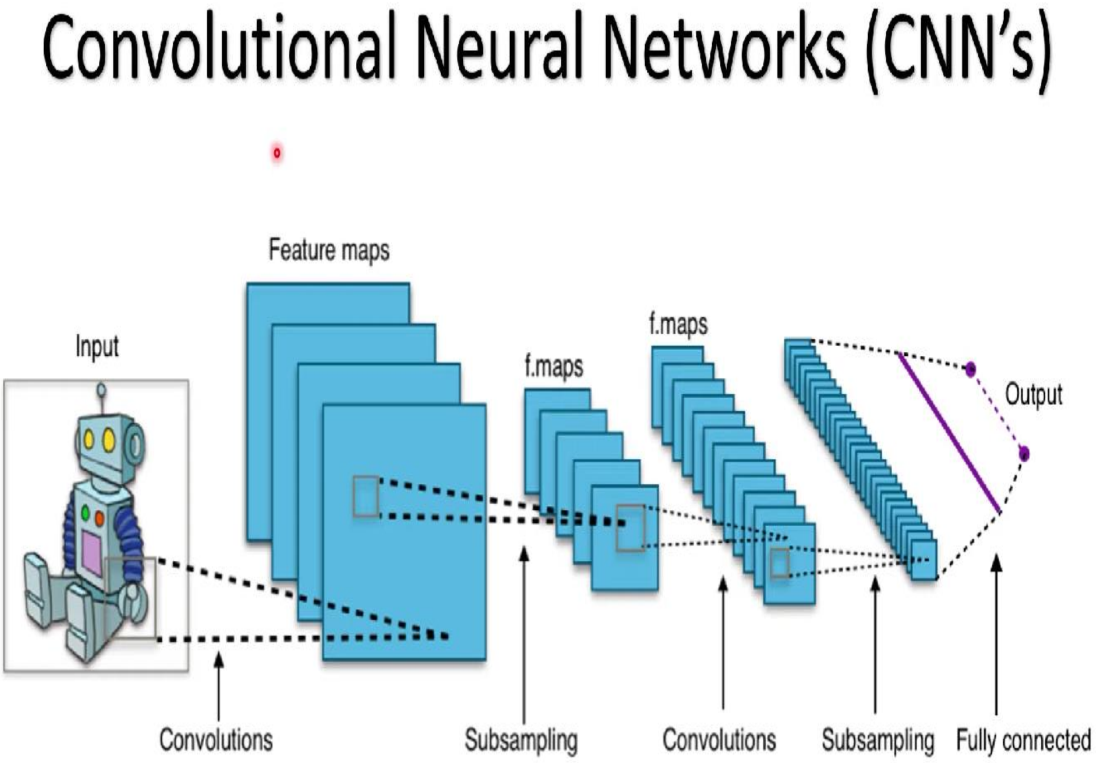


Figure-1 Convolutional Neural Network

Architecture

PWS-DVC (Progressive Weakly Supervised Dense Video Captioning) is a novel approach designed to improve dense video captioning without relying on detailed event information from the video

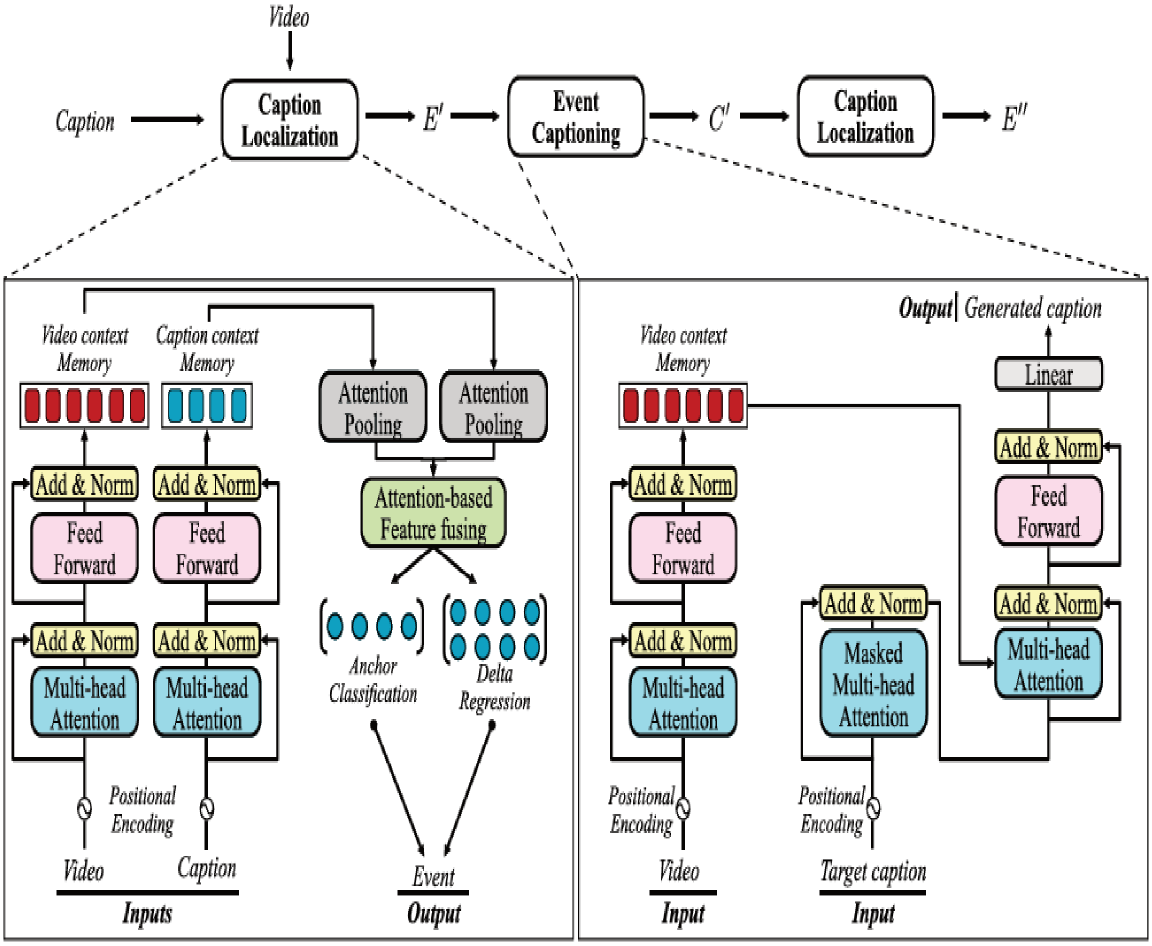


Figure-5 Progressive Weakly Supervised Dense Video Captioning

Caption Localization

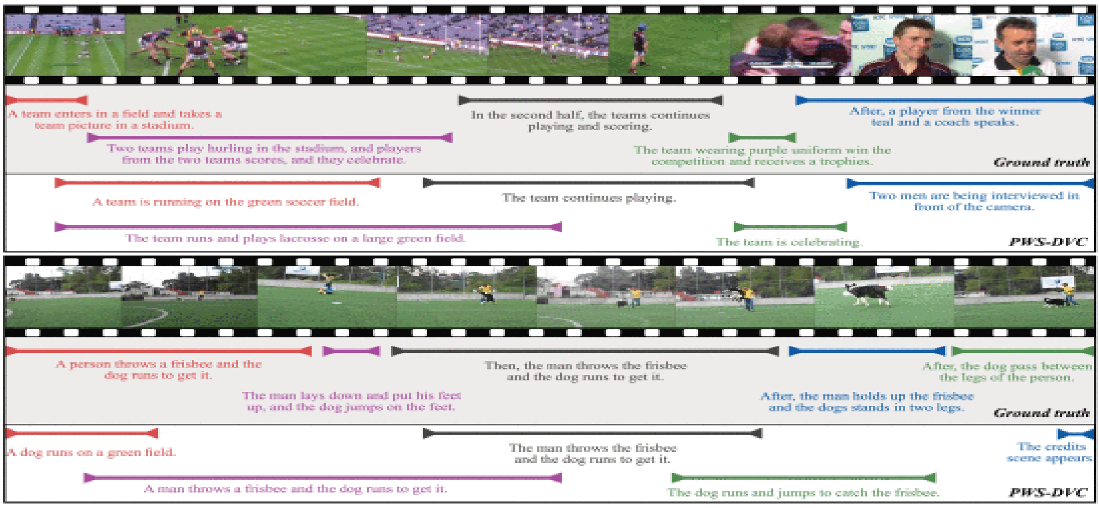
The process starts with Caption Localization on the video input, where the model identifies relevant segments of the video that correspond to the caption.

- Positional encoding

Positional encoding adds a time signal to the input. This is necessary because unlike RNNs, there is no recurrence built into transformers which carries positional information in the network architecture itself

4-RESULT

We select two videos from the validation set of ActivityNet Captions, and show their corresponding events and captions together. Through the examination of these two instances, it is demonstrated that PWS-DVC has the capability to effectively generate events and captions even in the absence of ground-truth events. Furthermore, PWS-DVC produces descriptive sentences by utilizing the event captioning module that has been pretrained on a video clip dataset. Additionally, it accurately identifies the location of each sentence inside the video scene. In the first video, PWS-DVC produces sentences that incorporate certain keywords, such as “green field,” “running,” and “lacrosse,” while also accurately identifying the precise location within the video scene for each sentence. This holds for the second video as well.


A team enters in a field and takes a team picture in a stadium.
In the second half, the teams continues playing and scoring.
After, a player from the winner teal and a coach speaks.
Two teams play hurling in the stadium, and players from the two teams scores, and they celebrate.
The team wearing purple uniform win the competition and receives a trophies.
Ground truth
A team is running on the green soccer field.
The team continues playing.
Two men are being interviewed in front of the camera.
The team runs and plays lacrosse on a large green field.
The team is celebrating.
PWS-DVC
A person throws a frisbee and the dog runs to get it.
Then, the man throws the frisbee and the dog runs to get it.
After, the dog pass between the legs of the person.
The man lays down and put his feet up, and the dog jumps on the feet.
After, the man holds up the frisbee and the dogs stands in two legs.
Ground truth
A dog runs on a green field.
The man throws the frisbee and the dog runs to get it.
The credits scene appears.
A man throws a frisbee and the dog runs to get it.
The dog runs and jumps to catch the frisbee.
PWS-DVC

The collection of these captions results in a story describing the video. During the second process (Text Summarization), this generated story is fed into a process as a document. The purpose of the proposed system is to automatically generate a title and also an abstract for a video clip without manual intervention. In this article, we have provided results based on our experimentation using video clips available from publicly.

A person holding a cell phone in their hand. A woman standing in front of a counter with a bunch of food. A woman and a woman are in a room. A man and a woman are sitting at a table. A person holding a cell phone in their hand. A group of people standing in a parking lot.

5-CONCLUSION

The advancement of video captioning technology is opening up revolutionary opportunities in a variety of industries. Its applications are enhancing security, accessibility, and human-robot interactions, with the promise of making future technology more flexible and intuitive. We should expect increasingly human-like interactions between robots and humans in many spheres of life as video captioning advances

Weakly supervised dense video captioning is promising as it does not rely on event supervision. However, inadequate training in the captioning module negatively affects both localization and captioning. To address this, we generate a robust language model by pretraining event captioning modules on publicly available datasets, such as the MSR-VTT. Our robust language model indirectly enhances the performance of event captioning and caption localization.

Future Scope

video captioning is enhancing dense video captioning for long, untrimmed videos. Current models struggle with multiple overlapping events, and future improvements in event segmentation and attention mechanisms can provide better event recognition and caption generation. Expanding the project to support multiple languages will also make video content more inclusive, allowing global audiences to access captions in their native language.

REFERENCES

1) N.Aafaq,A.Mian,W.Liu,S.Z.Gilani,andM.Shah,‘‘Videodescription: A survey of methods, datasets, and evaluation metrics,’’ ACM Comput. Surv., vol. 52, no. 6, pp. 1–37, Nov. 2020.
2) A.Puscasiu,A.Fanca,D.-I. Gota,andH.Valean,‘‘Automatedimagecaptioning,’’ in Proc. IEEE Int. Conf. Autom., Quality Test., Robot. (AQTR), May 2020, pp. 1–6.
3) Z. Ren, X. Wang, N. Zhang, X. Lv, and L.-J. Li, ‘‘Deep reinforcement learning-based image captioning with embedding reward,’’ in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 1151–1159