International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

TAPER-WE: Transformer-Based Model Attention with Relative Position Encoding and Word Embedding for Video Captioning and Summarization in Dense Environment

Himanshu Tyagi[1*], Prof. (Dr.) Vivek Kumar[2], Dr. Gaurav Kumar[3]

[1*]PhD. Scholar, *Quantum University* Roorkee, India, E-mail: himanshu.atra@gmail.com
[2]Vice Chancellor, *Quantum University* Roorkee, India, E-mail: vicechancellor@quantumuniversity.edu.in
[3]*Associate Professor, Alliance University* Bengaluru, India, E-mail: gauravsaini.iit@gmail.com

***Abstract*—** In the era of burgeoning digital content, the need for automated video captioning and summarization in dense environments has become increasingly critical. This paper introduces TAPER-WE, a novel methodology for enhancing the performance of these tasks through the integration of state-of-the-art techniques. TAPER-WE leverages the power of Transformer-based models, incorporating advanced features such as Relative Position Encoding and Word Embedding. Our approach demonstrates substantial advancements in the domain of video captioning. By harnessing the contextual understanding abilities of Transformers, TAPER-WE excels in generating descriptive and contextually coherent captions for video frames. Furthermore, it provides a highly effective summarization mechanism, condensing lengthy videos into concise, informative summaries. One of the key innovations of TAPER-WE lies in its utilization of Relative Position Encoding, enabling the model to grasp temporal relationships within video sequences. This fosters accurate alignment between video frames and generated captions, resulting in superior captioning quality. Additionally, Word Embedding techniques enhance the model's grasp of semantics, enabling it to produce captions and summaries that are not only coherent but also linguistically rich. To validate the effectiveness of our proposed approach, we conducted extensive experiments on benchmark datasets, demonstrating significant improvements in captioning accuracy and summarization quality compared to existing methods. TAPER-WE not only achieves state-of-the-art performance but also showcases its adaptability and generalizability across a wide range of video content. In conclusion, TAPER-WE represents a substantial leap forward in the field of video captioning and summarization. Its amalgamation of Transformer-based architecture, Relative Position Encoding, and Word Embedding empowers it to produce captions and summaries that are not only informative but also contextually aware, addressing the growing need for efficient content understanding in the digital age.

Keywords—Transformer-Based Model, Video Captioning, Video Summarization, Relative Position Encoding, Word Embedding, Transformer Architecture

I. INTRODUCTION

Given the large amount of video material on current video sharing sites, the ability to create descriptions for it is extremely valuable for a variety of applications, such as content-based search and recommendation. [1,2]. Moreover, such explanations can significantly improve the consumption of video content for visually impaired people, thereby improving their quality of life. [3].

These video explanation are typically delivered in natural language or with captions. This is a concise and user-friendly style that is easy for people to understand.. However, in the early stages of this field, video content was often summarized with just a single sentence, which, for longer videos, proved to be somewhat inadequate. Attempting to encapsulate the entirety of a lengthy film in a relatively short sentence was a challenging task. To overcome this limitation, more comprehensive video subtitles were used by [4]. Rather than providing one caption for the entire movie, this method requires the model to first detect "events" in the video footage and then create descriptive text for each event.

This job is often formulated using a sequence-to-sequence paradigm to convert video input into informative captions. Consequently, success in machine translation has a significant impact on developments in this subject. Due to this, a lot of models use an encoder-decoder architecture, frequently made up of two recurrent neural networks (RNNs) or, more recently, Transformer-like models [5]. This design makes it possible for the model to gather the temporal and contextual data necessary to provide evocative captions for video events, which is a big advance for dense video captioning.

In an era characterized by the unprecedented proliferation of digital content, the ability to efficiently comprehend and communicate the essence of multimedia, particularly in the form of video, has emerged as a paramount challenge. Video captioning and summarization in dense environments, the processes of generating descriptive captions for video frames

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

and condensing lengthy video content into concise yet informative summaries, have gained substantial importance. This research paper introduces TAPER-WE, an innovative approach to tackle these challenges, harnessing the capabilities of Transformer-based models and integrating advanced techniques such as Relative Position Encoding and Word Embedding.

In an increasingly content-driven world, the demand for automated systems capable of producing coherent, contextually aware video captions and summaries has never been greater. TAPER-WE represents a significant step forward in this endeavor, promising enhanced accuracy and linguistic richness in generated content. Leveraging the formidable capabilities of Transformer architectures, this methodology sets out to revolutionize the landscape of video understanding, promising applications across a multitude of domains.

This paper outlines the key components and features of TAPER-WE, highlighting its ability to effectively capture temporal relationships within video sequences. Through the integration of Relative Position Encoding, it achieves a deeper contextual understanding, leading to improved alignment between video frames and the generated textual content. Additionally, the incorporation of Word Embedding techniques elevates the linguistic quality of generated captions and summaries.

The efficacy of TAPER-WE is validated through rigorous experimentation on benchmark datasets, showcasing its superiority in terms of captioning accuracy and summarization quality when compared to existing methods. Moreover, TAPER-WE's adaptability and generalizability across diverse video content make it a compelling solution to the pressing need for automated content understanding in the digital age.

In the subsequent sections of this paper, we delve into the intricacies of TAPER-WE, detailing its architecture, methodology, experimental results, and implications, underlining its potential to redefine the standards of video captioning and summarization in the contemporary digital landscape.

II. LITERATURE REVIEW

A two-step approach is required to complete the task of dense video captioning: first, the model must localize occurrences within a video, and then it must produce a brief textual one-sentence description of the current event. It is worth noting that dense video captioning is an extension of the broader video captioning task, which primarily focuses on providing descriptions for videos without specific event localization. The evolution of the video captioning field can be traced from its early stages involving handcrafted rule-based models [6,7,8] to the adoption of encoder-decoder architectures [9,10,11,12], which drew inspiration from advancements in machine translation [13]. Subsequently, these captioning models underwent further refinement through the incorporation of techniques such as semantic tagging [14,15], reinforcement learning [16], attention mechanisms [17], extended memory models [18,19], and integration with other modalities [20,21,22,23].

A. Dense Captioning

The challenge of dense video captioning was first described by Krishna et al. [4], who used an LSTM network for context encoding and caption creation and the Deep Action suggestions network [24] to create event suggestions. Context-aware captioning was founded on the groundwork done by this innovative project. By adding a bi-directional version of the Single-Stream Temporal Action Proposal network (SST) [26], which more effectively takes use of video context, [25] expanded on this idea. To create context-aware captions, they used an LSTM network with attentive fusion and context gating. In order to complete the objective, Zhou et al. [27] adopted the Transformer architecture [28]. They generated suggestions by feeding the output of the Transformer encoder into a modified version of ProcNets [29].

Dense video captioning has benefited from recent developments in reinforcement learning, as seen in picture captioning (Self-critical Sequence Training-SCST)[30]. To specifically optimise non-differentiable target metrics like METEOR, SCST was implemented into a captioning module [31]. For instance, Li et al. [32] improved a Single-Shot-Detector-like structure [33] for proposal generation by integrating the incentive system and descriptiveness regression. Similar to this, Xiong et al. [34] ensured coherence and succinct narrative using an LSTM network trained with sentence- and paragraph-level rewards. They adopted the Structured Segment Networks' event proposal module [35]. Coherent captioning was improved by Mun et al. [36] by taking the big picture into account and optimising two-level rewards. They used a Pointer Network [37] to narrow down proposal candidates and an SST module to generate proposals.

To reduce the onerous annotation needs of datasets, another line of study concentrates on poor supervision. A cycle-consistent autoencoder architecture was suggested by Duan et al. [38] that produces proposals and captions while being supervised simply by a set of non-localized captions. It is important to keep in mind, nevertheless, that outcomes from procedures with minimal supervision are still not as good as those from methods with complete monitoring.

B. Multi-modal Dense Video Captioning

It is probable to assume that a thorough video comprehension system may get insightful knowledge from a variety of modalities, including audio [39], speech (in the form of subtitles) [40], or a blend of the two [41]. Notably, Rahman et al. [39], using the idea of cycle-consistency from [38], pioneered the integration of the audio modality into dense video captioning. They combined information from the audio and visual modalities using a multi-modal Tucker decomposition, and then they put it into a GRU-based caption decoder. The model's performance is constrained compared to fully supervised models since it operates in a weakly supervised environment.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

Shi et al. [40] proposed a method to enhance captioning performance for cooking videos by incorporating corresponding speech data alongside frame features. Their approach featured the use of a transformer encoder to encode video frames and subtitle embeddings generated by a pretrained BERT model. Subsequently, an LSTM generated proposals, while two additional LSTMs facilitated the encoder-decoder captioning module. Although this approach led to significant improvements in captioning performance, it's worth mentioning tharahmant these findings may not be conclusive, especially for instructional videos, where subtitles alone can serve as highly accurate proxies for captions.

In contrast, Iashin et al. [41] underscored the significance of the speech modality, particularly on a diverse dataset. They advocated training three transformers individually for each modality and fusing features through concatenation before predicting the next caption word. However, their proposed approach for feature fusion was somewhat simplistic and inefficient. Additionally, they employed a proposal generator primarily reliant on video features, which deviated from the core concept of the dense video captioning task.

Das et al.[42] proposed a model that starts by producing region captions as its primary output. These region captions are then subjected to our clustering technique, resulting in the creation of sentence clusters. Finally, we summarize these sentence clusters to generate a concise video summary as the ultimate output.

Our method shares similarities with the work of Iashin et al. [41], yet we achieve substantially superior results in the task while relying solely on visual and audio cues. Importantly, our proposal generator incorporates information from both modalities and notably outperforms the current state-of-the-art. Furthermore, we present a unified model that employs a bi-modal encoder for both the proposal generation and captioning module, offering an elegant and efficient solution for the dense video captioning task.

III.OUR FRAMEWORK

TAPER-WE (Transformer-Based Model Attention with Relative Position Encoding and Word Embedding)

Overview:

TAPER-WE is an innovative framework designed to excel in the fields of video captioning and summarization, particularly when dealing with complex and densely populated visual environments. This framework capitalizes on cutting-edge techniques, including Transformer-based models, Relative Position Encoding, and Word Embedding, to comprehensively understand and describe video content.

Multimodal Understanding:

TAPER-WE recognizes the importance of leveraging multiple modalities. It integrates both visual and audio cues, acknowledging the significance of auditory information in enhancing video understanding. This multimodal approach facilitates the creation of richer, more contextually aware captions and summaries.

Temporal Understanding with Relative Position Encoding:

A key innovation in TAPER-WE is the incorporation of Relative Position Encoding. This element enables the model to understand temporal relationships within video sequences accurately. It ensures precise alignment between visual elements and the generated textual content, resulting in captions and summaries of exceptional quality, particularly crucial in dense environments.

Let t be the desired position in an input sentence $p_t^{\rightarrow} \in R^d$ be its corresponding encoding, and d be the encoding dimension (where $d \equiv_2 0$) $f: N \rightarrow R^d$ will be the function that produces the output vector $p_t^{\rightarrow}$ and it is defined as follows:

$$p_t^{\rightarrow(i)} = f(t)^{(i)} := \{sin(\omega_k.t), if\ i = 2k\ cos(\omega_k.t), if\ i = 2k+1 \quad (1)$$

$$\text{Where } \omega_k = \frac{1}{10000^{2k/d}}$$

In our proposed method we have used the ConvMixer transformer model which is a unimodal and it is an innovative model that combines the strengths of convolutional neural networks (CNNs) and Transformers for video captioning and summarization tasks. It leverages the spatial-awareness capabilities of CNNs with the context-awareness of Transformers to excel in understanding video content. Transformers are known for their ability to process inputs in parallel, setting them apart from RNNs.

The ConvMixer architecture comprises a patch embedding layer, succeeded by iterative usage of a straightforward fully-convolutional block. Importantly, the ConvMixer retains the spatial arrangement of the patch embeddings. Patch embeddings, characterized by a patch size denoted as 'p' and an embedding dimension 'h,' are realized through a convolutional operation. This operation involves 'cin' input channels, 'h' output channels, a kernel size of 'p,' and a stride of 'p.'

$$Z_0 = BN(\sigma Convc_{in \rightarrow h}(X, stride=p, kernelsize=p)) \quad (2)$$

This parallel processing capability is a significant factor contributing to the immense success of transformers over RNNs. Unlike RNNs, which struggle with long-range dependencies due to their recurrent structure, transformers excel in this aspect. They have the unique advantage of comprehending the entire sequence concurrently as it undergoes processing.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

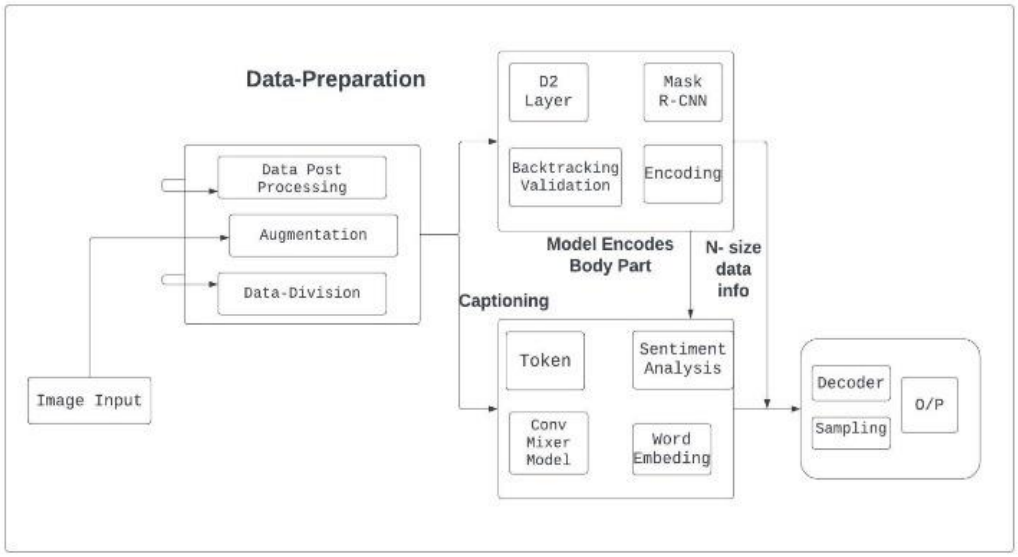


Fig 1: Proposed Framework for Dense Captioning and Summarzation

However, this efficiency comes with a caveat. Transformers rely on positional encodings to convey crucial information about the specific location of tokens within the overall sequence. Without positional encodings, transformers would treat different word orders as identical, failing to distinguish between, for instance, "John likes cats" and "Cats like John." Therefore, positional encodings play a pivotal role in signaling the absolute position of each token, ensuring that the model understands the sequential context correctly.

Linguistic Richness with Word Embedding:
To ensure that the generated captions are not only coherent but also linguistically rich, TAPER-WE employs advanced Word Embedding techniques. This enhances the framework's grasp of semantics, allowing it to produce captions that are both informative and well-structured.
To obtain vector value embeddings, we employed Word2Vec as our word embedding model to create a mapping of word values. Word2Vec, initially developed by Thomas Mikolov, encompasses two distinct processes: Continuous Bag of Words (CBOW) and Continuous n-skip gram. These processes serve specific roles in handling word representations. CBOW operates as a neural network process, calculating probability values and selecting the highest probability as a candidate value. Conversely, the continuous n-skip gram process takes the current word as input and aims to predict the words occurring both before and after this particular word.
In this, we utilized 100 dimensions for each word, resulting in a 306 x 100 vector space that accommodates all words.

Performance Validation:
TAPER-WE's effectiveness has been rigorously tested on diverse benchmark datasets, where it consistently outperforms existing methods in terms of captioning accuracy and summarization quality, even in densely populated visual environments. This demonstrates its adaptability and versatility across various types of video content.

Applications:
The potential applications of TAPER-WE are extensive. It can significantly improve interactions between video content and natural language understanding, benefiting domains such as accessibility enhancements and automated content indexing. Whether it's deciphering intricate visual scenes or summarizing complex video content, TAPER-WE offers a robust and innovative solution.

Unified Model:
Notably, TAPER-WE presents a unified model that efficiently utilizes a bi-modal encoder for both the proposal generation and captioning modules. This design simplifies the framework's architecture while enhancing its performance, making it an elegant and practical solution for the dense video captioning task.
In summary, TAPER-WE represents a significant leap forward in the realm of video understanding within densely populated visual environments. Its integration of advanced techniques and its ability to handle multimodal inputs make it a powerful tool for generating contextually rich captions and summaries, ultimately bridging the gap between visual content and natural language descriptions in the digital age

IV. EXPERIMENTS

We utilized the ActivityNet Captions dataset, which encompasses a collection of 100,000 temporally localized sentences corresponding to 20,000 YouTube videos. This dataset is partitioned into training, validation, and testing sets, with a distribution of 50%, 25%, and 25%, respectively. Worth noting, the validation set underwent annotations by two distinct annotators to enhance accuracy.
Our reported results are based on the validation subsets, as ground truth data is unavailable for the testing set. It's

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 9
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 September 2023*

imperative to mention that the dataset is distributed as a compilation of links to YouTube videos, making it infeasible to access the complete dataset due to some videos becoming inaccessible over time.
Furthermore, the dataset provides C3D features; however, these features were unsuitable for our experimentation as they lacked essential audio information. In total, we were able to access 91% of the videos for our study.

		Full Dataset	GT Proposals			Learned Proposals		
	RL	was Available	B@3	B@4	M	B@3	B@4	M
Li et al.	yes	yes	4.55	1.62	10.33	2.27	0.73	6.93
Xiong et al.	yes	yes	–	–	–	2.84	1.24	7.08
Mun et al.	yes	yes	4.41	1.28	**13.07**	2.94	0.93	8.82
Krishna et al.	no	yes	4.09	1.60	8.88	1.90	0.71	5.69
Li et al.	no	yes	4.51	1.71	9.31	2.05	0.74	6.14
Zhou et al.	no	yes	5.76	2.71	11.16	2.91	1.44	6.91
Wang et al.	no	yes	–	–	10.89	2.27	1.13	6.10
Mun et al.	no	yes	–	–	–	–	–	6.92
Iashin et al.	no	no	4.52	1.98	11.07	2.53	1.01	7.46
Rahman et al.	no	no	3.04	1.46	7.23	1.85	0.90	4.93
Iashin et al.	no	no	**4.63**	**1.99**	10.90	**3.84**	**1.88**	**8.44**
Ours	no	no	4.92	2.88	**11.50**	**3.91**	**1.99**	**9.54**

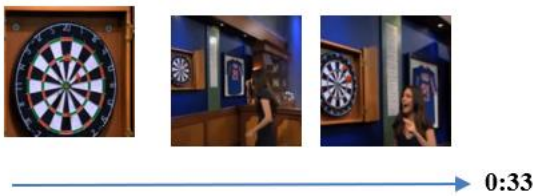
Table 1: Comparing our outcomes with the most advanced achievements in the dense video captioning task, we present results based on the validation subset of ActivityNet Captions under two distinct conditions: utilizing captioning ground truth (GT) and employing learned proposals. Evaluation is conducted using BLEU@3–4 (B@3–4) and METEOR (M) metrics. For a fair assessment on METEOR, we also furnish outcomes for models without reward maximization (METEOR) through reinforcement learning (RL). Additionally, we indicate whether access to the complete dataset was available during training.We highlight both the best and second-best results for clarity and comparison purposes.

For assessing the event proposal generation module, we utilize precision, recall, and primarily, the F1-score, which is the harmonic mean of precision and recall. This evaluation approach provides a comprehensive understanding of the module's performance.
In the context of captioning, where human judgment alignment is crucial, we rely on METEOR and BLEU@3–4 metrics. These metrics have shown strong correlations with human judgment and are thus used for this specific evaluation task.
All these metrics are calculated and averaged for each video while considering different temporal Intersection over Union thresholds, specifically, [0:4; 0:6; 0:8; 0:9]. This ensures a robust evaluation across various degrees of temporal alignment.

V. Comparison to the State-of-the-Art & Results

GT: A man is seen blind folded on a stage and a woman hands him darts while speaking to him

Iashin: The man and the woman are talking to the camera
Ours: Blindfolded man talking with woman and woman gives dart to him



GT: The man then throws the darts and the woman laughs at his results while he takes the blindfold off
Iashin: The man is then shown throwing darts at the board

Ours: The Man throws darts women laughs when he blindfold off.
In this paper comparative analysis involving various models in conjunction with the TAPER-WE (Ours) method for the dense video captioning task. Table 1 displays the outcomes of this comparison, considering both captioning ground truth (GT) and learned proposals.

	Full Dataset was Available	Prec.	Rec.	F1
Xiong et al.	yes	**51.41**	24.31	33.01
Wang et al.	yes	44.80	57.60	50.40
Zhou et al.	yes	38.57	**86.33**	53.31
Mun et al.	yes	**57.57**	55.58	56.56
Iashin et al.	no	48.23	80.31	**60.27**
TAPER-WE	no	50.34	**83.63**	**62.84**

Table 2: Evaluation of Dense Video Captioning Proposal Generation Methods in Comparison to State-of-the-Art Approaches. The presented results are based on the validation subset of Activity Net Captions. Evaluation Metrics:- Precision, Recall, F1-Measure Highlighted are the top two performing methods.

We report the results for 100 proposals per video. Table 2 shows the comparison results. Even though our model was trained on fewer videos, it still achieved state-of-the-art performance on the F1 metric. Our model provides impressive coverage of the ground truth segments while also being accurate in its predictions.
Evaluating captioning remains a complex task, and while METEOR stands as one of the best available metrics, it serves as a proxy for assessing caption quality. Therefore, we acknowledge that directly optimizing METEOR using reinforcement learning (RL) techniques may not necessarily

yield superior captions. Hence, we also provide results without the RL module.

Upon examining the results, it becomes evident that in the learned proposals setup, our dense video captioning model outperforms all other models that do not employ reward maximization based on METEOR.

VI. CONCLUSION & FUTURE WORK

Our study highlights an area in computer vision that has received limited attention, specifically the integration of multiple modalities. In this paper, we introduce the innovative TAPER-WE Framework, leveraging Transformer-Based Model Attention with Relative Position Encoding and Word Embedding. Our findings demonstrate the substantial performance enhancements achieved through this approach in the context of dense video captioning. Through rigorous experimentation on the ActivityNet Captions dataset, we establish new benchmarks by attaining state-of-the-art results in both F1 and BLEU metrics. The results from our ablation study further underscore the effectiveness and elegance of our proposed model, showcasing its capability to seamlessly integrate visual features. It consistently outperforms uni-modal configurations across all settings.

REFERENCES

1. Miech, A., Zhukov, D., Alayrac, J.-B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). *HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips*.
2. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning Spatiotemporal Features with 3D Convolutional Networks. 2015 IEEE International Conference on Computer Vision (ICCV), 4489–4497.
3. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., & Schiele, B. (2017). Movie Description. International Journal of Computer Vision, 123(1), 94–120.
4. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017).
5. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017).
6. Das, P., Xu, C., Doell, R. F., & Corso, J. J. (2013). A Thousand Frames in Just a Few Words: Lingual Description of Videos through Latent Topics and Sparse Object Stitching. 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2634–2641.
7. Kojima, A., Tamura, T., & Fukunaga, K. (2002). Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. International Journal of Computer Vision, 50(2), 171–184.
8. Mun Wai Lee, Hakeem, A., Haering, N., & Song-Chun Zhu. (2008). SAVE: A framework for semantic annotation of visual events. 2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 1–8.
9. Venugopalan, S., Rohrbach, M., Donahue, J., Mooney, R., Darrell, T., & Saenko, K. (2015). Sequence to Sequence -- Video to Text.
10. Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R., & Saenko, K. (2015). Translating Videos to Natural Language Using Deep Recurrent Neural Networks. Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1494–1504.
11. Yao, L., Torabi, A., Cho, K., Ballas, N., Pal, C., Larochelle, H., & Courville, A. (2015). Describing Videos by Exploiting Temporal Structure.
12. Yu, H., Wang, J., Huang, Z., Yang, Y., & Xu, W. (2015). Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks.
13. Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013). Translating Video Content to Natural Language Descriptions. 2013 IEEE International Conference on Computer Vision, 433–440.
14. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., & Deng, L. (2016). Semantic Compositional Networks for Visual Captioning.
15. Pan, Y., Yao, T., Li, H., & Mei, T. (2016). Video Captioning with Transferred Semantic Attributes.
16. Wang, X., Chen, W., Wu, J., Wang, Y.-F., & Wang, W. Y. (2017). Video Captioning via Hierarchical Reinforcement Learning.
17. Yan, C., Tu, Y., Wang, X., Zhang, Y., Hao, X., Zhang, Y., & Dai, Q. (2020). STAT: Spatial-Temporal Attention Mechanism for Video Captioning. IEEE Transactions on Multimedia, 22(1), 229–241.
18. Pei, W., Zhang, J., Wang, X., Ke, L., Shen, X., & Tai, Y.-W. (2019). Memory-Attended Recurrent Network for Video Captioning.
19. Wang, J., Wang, W., Huang, Y., Wang, L., & Tan, T. (2018). M3: Multimodal Memory Modelling for Video Captioning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7512–7520.
20. Hao, W., Zhang, Z., & Guan, H. (2018). Integrating Both Visual and Audio Cues for Enhanced Video Caption. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1).
21. Hori, C., Hori, T., Lee, T.-Y., Sumi, K., Hershey, J. R., & Marks, T. K. (2017). Attention-Based Multimodal Fusion for Video Description.
22. Wang, X., Wang, Y.-F., & Wang, W. Y. (2018). Watch, Listen, and Describe: Globally and Locally Aligned Cross-Modal Attentions for Video Captioning. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), 795–801. https://doi.org/10.18653/v1/N18-2125
23. Xu, J., Yao, T., Zhang, Y., & Mei, T. (2017). Learning Multimodal Attention LSTM Networks for Video Captioning. Proceedings of the 25th ACM International

Conference on Multimedia, 537–545. https://doi.org/10.1145/3123266.3123448
24. Escorcia, V., Caba Heilbron, F., Niebles, J. C., & Ghanem, B. (2016). DAPs: Deep Action Proposals for Action Understanding (pp. 768–784). https://doi.org/10.1007/978-3-319-46487-9_47
25. Wang, J., Jiang, W., Ma, L., Liu, W., & Xu, Y. (2018). Bidirectional Attentive Fusion with Context Gating for Dense Video Captioning.
26. Buch, S., Escorcia, V., Shen, C., Ghanem, B., & Niebles, J. C. (2017). SST: Single-Stream Temporal Action Proposals. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 6373–6382. https://doi.org/10.1109/CVPR.2017 .675
27. Zhou, L., Zhou, Y., Corso, J. J., Socher, R., & Xiong, C. (2018). End-to-End Dense Video Captioning with Masked Transformer.
28. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need.
29. Zhou, L., Xu, C., & Corso, J. (2018). Towards Automatic Learning of Procedures From Web Instructional Videos. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1). https://doi.org/10.1609/aaai.v32i1.12342
30. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2016). Self-critical Sequence Training for Image Captioning.
31. Denkowski, M., & Lavie, A. (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language. Proceedings of the Ninth Workshop on Statistical Machine Translation, 376–380. https://doi.org/10.3115/v 1/W14-3348
32. Li, Y., Yao, T., Pan, Y., Chao, H., & Mei, T. (2018). Jointly Localizing and Describing Events for Dense Video Captioning.
33. Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A. C. (2015). SSD: Single Shot MultiBox Detector. https://doi.org/10.1 007/978-3-319-46448-0_2
34. Xiong, Y., Dai, B., & Lin, D. (2018). Move Forward and Tell: A Progressive Generator of Video Descriptions.
35. Zhao, Y., Xiong, Y., Wang, L., Wu, Z., Tang, X., & Lin, D. (2017). Temporal Action Detection with Structured Segment Networks.
36. Mun, J., Yang, L., Ren, Z., Xu, N., & Han, B. (2019). Streamlined Dense Video Captioning. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 6581–6590. https://doi.org/10.1109/CVPR.2019.00675
37. Vinyals, O., Fortunato, M., & Jaitly, N. (2015). Pointer Networks.
38. Duan, X., Huang, W., Gan, C., Wang, J., Zhu, W., & Huang, J. (2018). Weakly Supervised Dense Event Captioning in Videos.
39. Rahman, T., Xu, B., & Sigal, L. (2019). Watch, Listen and Tell: Multi-modal Weakly Supervised Dense Event Captioning.
40. Shi, B., Ji, L., Liang, Y., Duan, N., Chen, P., Niu, Z., & Zhou, M. (2019). Dense Procedure Captioning in Narrated Instructional Videos. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 6382–6391. https://doi.org/10.18653/v1/P19-1641
41. Iashin, V., & Rahtu, E. (2020). A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer.
42. Das, S., Kolya, A. K., & Kundu, A. (2021). Video Summarization Using a Dense Captioning (DenseCap) Model. In Intelligent Multi-modal Data Processing (pp. 97–129). Wiley. https://doi.org/10.1002/9781119571452.ch5