International Journal of Applied Mathematics

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

RESIDUAL AND BOTTLENECK CNN ARCHITECTURES WITH LSTM FOR IMPROVED VIDEO CAPTION GENERATION

Amruta Rajendra Chougule[1], Dr. Shankar D. Chavan[2]

[1,2]Department of Electronics and Telecommunication

[1,2] Dr. D.Y.PATIL INSTITUTE OF TECHNOLOGY PIMPRI, Pune, India

[1]Corresponding author (Email):chougaleamruta5543@gmail.com

[1]ORCID ID: 0000-0002-8635-1105

[2]Contributing Author (Email):shankar.chavan@dypvp.edu.in

Abstract

Video captioning has become a pivotal research domain at the interface of computer vision and natural language processing, applications in multimedia retrieval, assistive systems, and human–computer interaction. Despite substantial progress, many existing approaches, including Vid2Seq, Positive-Augmented Contrastive Learning, GL-RG, and TextKG, continue to encounter limitations in jointly modeling fine-grained spatial details and long-term temporal dependencies. These challenges hinder the generation of captions that are both semantically accurate and contextually coherent. It proposes novel video captioning framework that leverages a convolutional neural network (CNN)-based encoder integrated with residual and bottleneck blocks to capture rich temporal–spatial features while mitigating gradient degradation. The encoder's design ensures efficient feature propagation and robust representation of video content. To model sequential dependencies and maintain contextual consistency, the extracted features are processed by a recurrent decoder based on long short-term memory (LSTM) networks. This hybrid architecture effectively balances feature extraction with sequential modeling, thereby addressing critical shortcomings of prior methods. Extensive evaluations were conducted on three benchmark datasets—MSR-VTT, MPII Cooking 2, and M-VAD. The proposed framework achieved a peak BLEU score of 51 . Beyond accuracy improvements, the architecture demonstrated reduced computational complexity, confirming its suitability for large-scale video captioning tasks. In conclusion, the integration of CNN-based residual encoding with LSTM-based recurrent decoding offers a streamlined yet powerful solution for video captioning. The proposed model advances the field by achieving a balance between efficiency and accuracy, thereby contributing a significant step toward the development of high-quality, contextually rich video descriptions in vision–language research.

Keywords: Video captioning, CNN encoder, Residual blocks, LSTM, Temporal-spatial features.

Received: July 18, 2025

1. Introduction

In recent years, the intersection of computer vision and natural language processing has witnessed remarkable advancements, particularly in understanding and generating content-rich multimedia data. Video understanding, in particular, presents a multifaceted challenge due to its sequential and dynamic nature, encompassing both visual and textual modalities. In this context, the fusion of Convolutional Neural Networks (CNNs) for visual feature extraction with Long Short-Term Memory (LSTM) networks for sequential modeling has emerged as a potent methodology for video understanding tasks.

This article introduces a novel model architecture that combines a CNN encoder for extracting rich visual features from videos with an LSTM decoder for generating descriptive text. The CNN encoder is adept at capturing both semantic information and temporal dynamics inherent in videos, while the LSTM decoder processes tokenized textual input to generate coherent and contextually relevant descriptions. The proposed model is trained on two widely used benchmark datasets, MSRVTT and YouCook2, to demonstrate its effectiveness across diverse video content and domains.

Generating captions from visual content, such as images or videos, requires the extraction of rich semantic features using advanced deep learning techniques. These visual cues are then semantically aligned with natural language expressions. To achieve the best performance, it is essential to refine these extracted features in a way that correlates with linguistic representations. This is typically accomplished by processing the input images through multiple layers of a deep neural architecture, where each layer incrementally learns and encodes different visual attributes. The culmination of this process results in a coherent sentence that describes the image's context.

When extended to video captioning, the concept remains similar but becomes more complex due to the temporal dynamics involved [4]. Instead of static images, video-captioning systems are trained on datasets composed of video sequences coupled with their textual annotations [5]. The model learns to capture spatio-temporal representations, encapsulating movement, transitions, and interactions within scenes over time [6]. These representations form the basis for generating descriptive sentences that convey the video's events.

Since effective caption generation involves both understanding the visual domain and encoding linguistic structure, studying image-captioning methods provides foundational insights for developing video-captioning systems. Accordingly, the proposed model in this work draws inspiration from multiple prior methods, focusing on integrating techniques for sequence-specific descriptions. Emphasis is placed on context-aware captioning, including event-level tagging, object-level identification, and sequence-length-aware narration. Additionally, the way different architectures attend to temporal flow and spatial layout across frames greatly influences the quality and specificity of the generated captions.

The significance of this work lies in its potential to advance the state-of-the-art in video understanding and description generation. By leveraging both visual and textual modalities, the proposed model offers a holistic approach to interpreting and summarizing video content.

Received: July 18, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

This is particularly valuable in applications such as video summarization, content retrieval, and assistive technologies for the visually impaired, where accurate and informative descriptions are essential for enhancing user experience and accessibility.

The scope of this research encompasses the development and evaluation of a CNN-LSTM model for video understanding tasks. By integrating CNN-based visual feature extraction with LSTM-based sequential modeling, the model aims to capture both spatial and temporal cues in videos and generate coherent textual descriptions. Furthermore, the application of this model is not limited to specific domains or datasets but can be extended to various video understanding tasks across different domains, making it a versatile tool for multimedia analysis and interpretation.

Summary of Contributions

1. **Novel Architecture**: We propose a novel model architecture that combines CNN-based visual feature extraction with LSTM-based sequential modeling for video understanding and description generation.
2. **Multimodal Fusion**: Our model integrates both visual and textual modalities to provide detailed features of video content, capturing semantic information and temporal dynamics simultaneously.
3. **Benchmark Performance**: Through extensive experimentation on benchmark datasets such as MSRVTT and YouCook2, we demonstrate the effectiveness and generalizability of our proposed approach, achieving state-of-the-art performance in video description tasks.

2. Related Work

Xu et al. [7] proposed a reinforcement learning-based caption generation framework that incorporates a denoising mechanism and grammar refinement unit. Their model maximizes the long-term reward through policy gradient optimization, using the following loss function:

$$\mathcal{L}_{RL} = -\mathbb{E}_{\hat{y}\sim p_\theta}[r(\hat{y})]\ldots(1)$$

Wherep_θ denotes the probability distribution over generated sequences, and $r(\hat{y})$ represents the reward (e.g., CIDEr or BLEU score) for the generated caption $\hat{y}$. This formulation encourages the generation of fluent and semantically accurate captions over time.

D. Yasin et al. [2] developed a method that incorporates object appearance cues using word embeddings and FaceNet features, processed through recurrent layers. The architecture utilizes both LSTM and GRU layers, where the hidden state updates follow:

$$h_t = \text{LSTM}(x_t, h_{t-1}), \quad \text{or} \quad h_t = \text{GRU}(x_t, h_{t-1})\ldots(2)$$

Withx_t representing input vectors composed of image and object appearance features at time t. This allows the model to maintain sequential coherence while integrating visual semantics.

Tang et al. [8] introduced a pyramidal feature extraction mechanism with multi-level attention for fine-grained image recognition. The attention module computes weights α_i for feature maps F_i as:

Received: July 18, 2025

$$\alpha_i = \frac{\exp(e_i)}{\sum_j \exp(e_j)}, \quad \text{where} \quad e_i = \text{score}(h, F_i)\ldots(3)$$

where$h$ is the current decoder state and $\text{score}(\cdot)$ is a similarity function. Their M3Net model [9] enhances this approach by adding additional levels of attention and multi-scale feature aggregation.

Yan et al. [10] addressed the problem of image retrieval based on textual queries using a Channel Attention Filter and Relation-Guided Localization (RGL). The channel attention mechanism dynamically weights feature maps as:

$$F' = \sigma\left(W_2 \cdot \text{ReLU}\left(W_1 \cdot \text{GAP}(F)\right)\right)\ldots(4)$$

whereGAP denotes global average pooling, and W_1, W_2 are trainable weights. The enhanced feature F' guides more precise localization and identification.

In [11], a pre-training strategy was employed to improve visual-language alignment. By optimizing the contrastive loss between paired and unpaired image-text pairs:

$$\mathcal{L}_{contrast} = -\log \frac{\exp\left(\text{sim}(I,T^+)\right)}{\sum_j \exp\left(\text{sim}(I,T_j)\right)}\ldots(5)$$

where$I$ is the image feature vector, T^+ is the correct text, and T_j are negative samples, the model learns stronger cross-modal relationships.

MasoomehNabati et al. [12] proposed a boosted LSTM model designed for iterative training and parallel batch processing. Their architecture improves temporal feature learning by computing sequential representations with enhanced speed and stability.

Chohan et al. [13] focused on encoder-decoder frameworks for image captioning using attention. The decoder at each step attends to relevant image regions via a learned context vector c_t:

$$c_t = \sum_i \alpha_{ti} h_i, \quad \alpha_{ti} = \frac{\exp(e_{ti})}{\sum_j \exp(e_{tj})}\ldots(6)$$

whereh_i are the encoder hidden states, and e_{ti} is a learned attention score.

Mun et al. [14] introduced a temporal-aware video captioning framework that achieved state-of-the-art performance on the ActivityNet Captions dataset by emphasizing sequential dependencies across time frames.

Fujii et al. [15] proposed a model that combines visual embeddings with sentence representations in hidden layers, updating weights to better align the visual-semantic space.Zhang et al. [16] worked on remote sensing image captioning using attribute attention. Their method extracts high-level semantic features and uses an attention-based decoder to link image attributes with descriptive phrases.Kim et al. [17] proposed a relational captioning system based on a triple-stream network, combining visual features, relational information, and part-of-speech tags. The final caption is generated using a fusion mechanism that integrates these three sources of information.

Dong et al. [18] explored sentence selection for scene description using multi-scale Word2VisualVec (W2VV). The sentence scoring function is defined as:

$$s = \cos\left(\phi(I), \psi(T)\right)\ldots(7)$$

Received: July 18, 2025

where$\phi(I)$ is the image representation and $\psi(T)$ is the textual embedding, with cosine similarity indicating relevance.

Orozco et al. [19] developed a CNN-LSTM model using the Microsoft Video Description Corpus, designed with accessibility applications in mind. The CNN extracts frame-level features which are passed to an LSTM for sequence modeling.Li et al. [20] performed literature analysis by linking image and text data, encouraging diverse caption generation using reinforcement learning guided by a reward function tailored to linguistic diversity and fluency.Shetty et al. [21] proposed a hybrid captioning framework integrating contextual features with language models, ensuring more semantically meaningful outputs.

Bai et al. [22] improved object relationship modeling through a geometric attention mechanism on MS-COCO. The geometric attention computes spatial relevance using bounding box coordinates and contextual embeddings.Daskalakis et al. [23] introduced a contextual feature-based captioning approach, achieving improved performance on MS-COCO and MSR-VTT by enhancing feature discrimination at the encoder level.Yang et al. [24] developed the Ved2Seq model for activity-based captioning. Their framework aligns scene boundary detection with language model tokens, improving event-specific description accuracy.

Sarto et al. [25] integrated a Vision Transformer encoder with textual processing, using CLIP for pre-alignment. Their similarity loss ensures better alignment between image patches and word tokens.Yan et al. [26] presented the GL-RG model using a granularity-aware learning objective, pairing fine-grained linguistic structures with localized visual representations.Gu et al. [27] proposed a transformer-based TextKG model, where scene-relative captioning is enhanced using knowledge graph embeddings and cross-attention layers.Anderson et al. [28] reviewed encoder-decoder model evolution in video captioning, providing an overview of key improvements and benchmarks.Lee et al. [29] proposed the SeFLA model for semantic event captioning, utilizing both ResNet and C3D pre-trained models. The features are merged before passing into the decoder, achieving competitive results on MSVD and MSR-VTT.

The comprehensive review of existing video captioning techniques provides a strong foundation for the proposed model. The analysis of diverse methods ranging from reinforcement learning-based caption refinement to attention-driven architectures highlights the significance of combining temporal modeling with spatial feature extraction. Many previous works employ encoder-decoder frameworks, attention mechanisms, or knowledge-augmented models to improve caption quality. However, a common limitation observed is the inadequate fusion of spatial and temporal features in a streamlined manner. This motivated the design of a model that leverages the strengths of residual and bottleneck CNN blocks to capture fine-grained spatial patterns from video frames while maintaining computational efficiency. These features are then sequentially processed using an LSTM-based decoder, which excels at modeling temporal dependencies and generating contextually coherent sentences. The integrated architecture ensures that both local spatial details and long-range temporal structures are effectively utilized, addressing gaps identified in earlier models. By

Received: July 18, 2025

grounding the model's design in the strengths and limitations of existing techniques, this study paves the way for more accurate, descriptive, and efficient video caption generation across diverse video datasets.

3. Proposed Work

The proposed work consist of the video frames feature extraction and recurrent neural network (RNN) decoder. The RNN decoder is designed with the use of LSTM. Figure 1 shows the proposed model architecture.

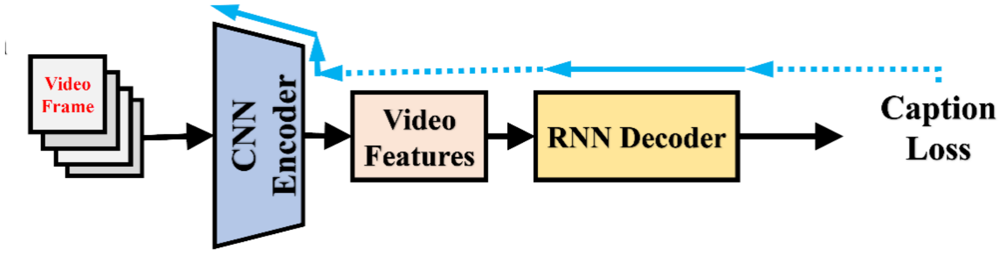


Figure 1: Proposed Model Architecture

Semantic and fine features play crucial roles in various computer vision tasks, including video captioning. Semantic features capture high-level information about the content of an image or video, such as object categories, scene context, and overall context. These features provide a global understanding of the visual content, aiding in tasks like scene recognition, activity detection, and context understanding. Semantic features are essential for generating coherent and contextually relevant captions in video captioning tasks, as they help the model grasp the overall theme and content of the video. On the other hand, fine features capture detailed information at a lower level, including textures, shapes, edges, and colors. Fine features focus on capturing intricate details and nuances within the visual content, enabling the model to discern subtle variations and specific characteristics of objects or scenes. While semantic features provide a holistic understanding, fine features offer granularity and richness to the visual representation, enhancing the model's ability to capture nuances and details crucial for accurate description generation. In video captioning, combining both semantic and fine features allows the model to generate comprehensive and detailed descriptions that effectively convey the content and context of the video. Semantic features provide the overarching structure and context, while fine features contribute nuanced details, resulting in more informative and accurate captions. Therefore, leveraging both semantic and fine features is essential for achieving robust and effective video captioning systems.

In the proposed video captioning work, combining features from different CNN architectures such as Inception-V3 and MobileNet-V2 enhances the model's ability to capture diverse visual information.

Let $F_{Inception}$ represent the feature vector extracted from the Inception-V3 CNN architecture, and $F_{MobileNet}$ represent the feature vector extracted from the MobileNet-V2 architecture. Both feature vectors are of dimension D.

Received: July 18, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

To combine these features, you can use a weighted sum approach:

$$F_{combined}=\alpha\cdot F_{Inception}+\beta\cdot F_{MobileNet}\ldots(1)$$

Where, Fcombined is the combined feature vector. α and β are the weights assigned to the features from Inception-V3 and MobileNet-V2, respectively. These weights can be learned during the training process or set manually. α+β=1 to ensure that the combined feature vector is a weighted average of the individual feature vectors.

Adjusting the parameters α and β is based on the relative importance or performance of each CNN architecture in specific video captioning task. For example, if Inception-V3 tends to capture more semantic information while MobileNet-V2 captures finer details, in which case assign a higher weight to Inception-V3 (α>β).In summary, combining features from Inception-V3 and MobileNet-V2 for the CNN decoder stage in video captioning involves calculating a weighted sum of the feature vectors extracted from each architecture, with the weights determined based on their relative contributions to the task.

4. Results and Discussion

a. Dataset

The MSR-VTT (Microsoft Research Video to Text) dataset is a large-scale video description dataset that includes 10,000 video clips collected from sources such as YouTube. Each video is 10 to 30 seconds long and annotated with 20 different captions, providing a diverse set of natural language descriptions. The MPII Cooking Activities Dataset contains around 44,000 video clips of 65 cooking activities, annotated with detailed descriptions of actions and objects. This dataset is widely used for recognizing and understanding fine-grained human activities in cooking scenarios. The MVAD (Montreal Video Annotation Dataset) consists of 92 movies with over 48,000 video clips, each annotated with detailed descriptions and aligned with movie subtitles. This dataset is designed for video description, summarization, and understanding tasks, offering a comprehensive resource for developing and evaluating models that aim to understand and generate descriptions for complex video content.

b. Performance Parameters

Performance parameters such as BLEU, METEOR, and CIDEr are crucial for evaluating machine-generated translations or textual descriptions. BLEU measures the likeness between generated and reference texts by comparing n-grams, with a higher score indicating greater similarity. METEOR incorporates precision, recall, and alignment-based measures, producing a single score reflecting overall quality. CIDEr, tailored for image description evaluation, considers both precision and diversity by weighting n-grams based on TF-IDF and including a penalty term for promoting diversity. For BLEU, the score is calculated using modified precision for n-grams, while METEOR combines precision and recall with alignment score computation. CIDEr computation provides the understanding of a weighted sum of TF-IDF scores, scaled by a term frequency penalty. These metrics provide comprehensive

Received: July 18, 2025

Volume 38 No. 2s, 2025

ISSN: 1311-1728 (printed version); ISSN: 1314-8060 (on-line version)

assessments of the quality and diversity of generated texts, essential for refining and optimizing machine-generated outputs

$$.TFIDF_i = \left(\frac{c_i}{|c|}\right) * \log\left(\frac{\sum r_i}{r_i+1}\right) \ldots (1)$$

$$CIDE_r = (\textstyle\sum_{n=1}^{N} w_i * TFIDF_i)/\,N + TF_{penalty} \ldots (2)$$

In the METEOR evaluation metric, a penalty parameter, denoted as α, is assigned for chunk matches, conventionally set to 0.5. Precision (P) and recall (R) are fundamental components of the METEOR computation. This metric calculates both unigram precision and recall, assessing the accuracy and completeness of the captions generated compared to original ground truth. Additionally, METEOR computes an alignment score, which reflects the degree of correspondence between the generated and reference texts. By considering precision, recall, and alignment, METEOR provides a comprehensive evaluation of the quality of machine-generated translations or textual descriptions.

$$P = \frac{Number\ of\ correctly\ matched\ unigrams}{Number\ of\ Unigrams\ in\ the\ candidate\ sentence} \ldots (2)$$

$$R = \frac{(number\ of\ correctly\ matched\ unigrams)}{(number\ if\ unigrams\ in\ the\ reference\ sentence)} \ldots (3)$$

$$Alignement\ Score = \frac{(P*R)}{((1-\alpha)*R + \alpha*P)} \ldots (4)$$

BLEU (Bilingual Evaluation Understudy) is a metric that quantifies the similarity between machine-generated translations and reference texts by comparing n-gram matches, yielding a precision-oriented score.

$$\mathrm{BLEU_n} = \frac{(\text{Number of n}-\text{gram matches})}{(\text{Number of n grams in the candidate sentence})} \ldots (5)$$

Received: July 18, 2025

c. Comparative Analysis

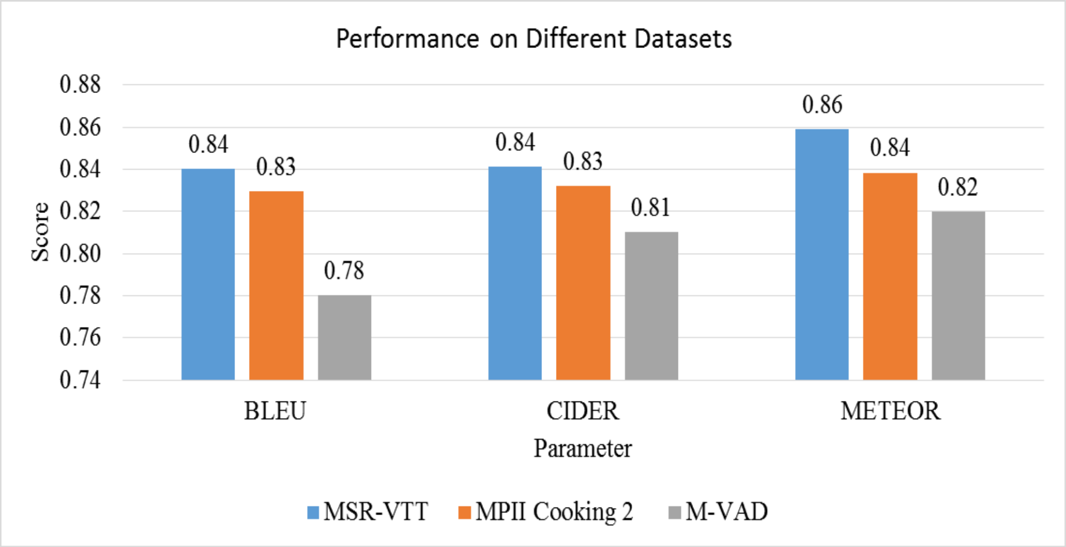
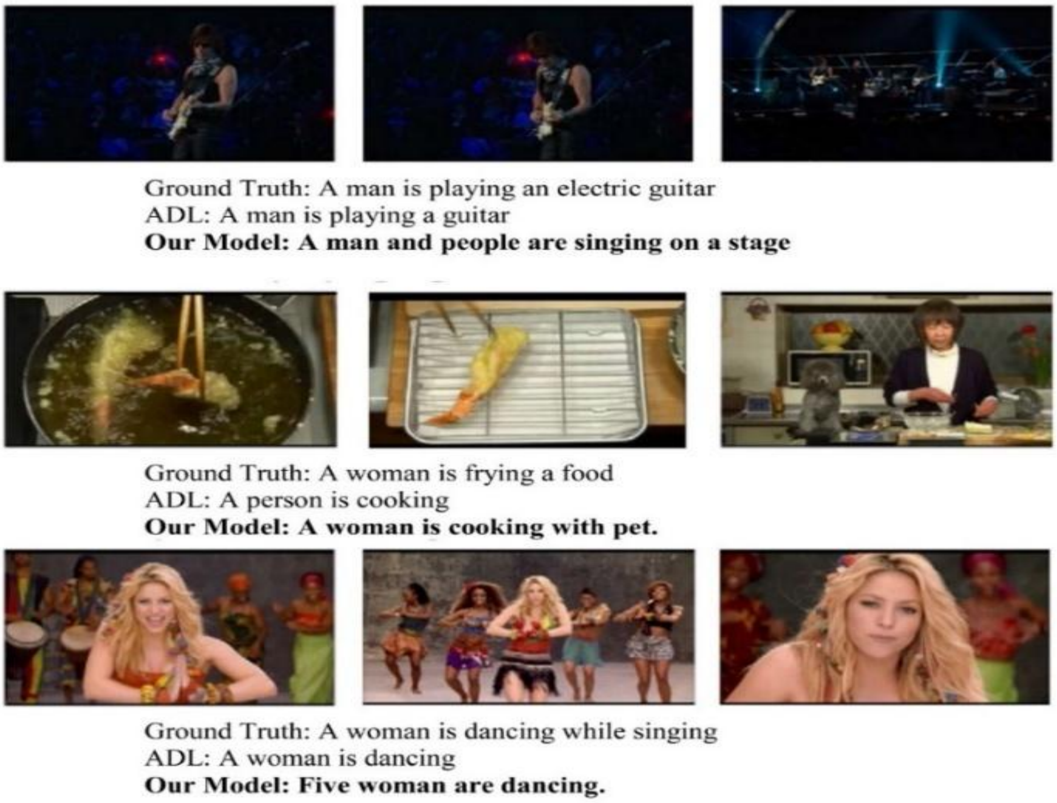


Figure 2: Performance analysis of proposed model on different datasets



Ground Truth: A man is playing an electric guitar
ADL: A man is playing a guitar
Our Model: A man and people are singing on a stage

Ground Truth: A woman is frying a food
ADL: A person is cooking
Our Model: A woman is cooking with pet.

Ground Truth: A woman is dancing while singing
ADL: A woman is dancing
Our Model: Five woman are dancing.

Figure 3: Comparative Analysis

Received: July 18, 2025

Table 1: Comparative Analysis with Other Existing Methods

Method	Dataset Used	Approach and Steps	Performance
Vid2Seq [24]	MSR-VTTMSVDViTT YouCook2	Implements dense captioning of videos. Identifies event boundaries for generating captions. Utilizes transcripted speech for training. Employs a CLIP-based visual encoder	Attained a peak CIDEr value of 47.2
Positive-Augmented Contrastive Learning [25]	VATEX-EVAL	Encodes video data with CLIP- Augments caption datasets with additional captions for the same video sequences. Utilizes PACS score-based learning. Evaluated through human annotations	Achieved a Machine Correct Caption Score of 0.821
GL-RG [26]	MSR-VTTMSVD	Links descriptive granularity from videos with linguistic expressions. Utilizes incremental learning. Encodes local key frames. Combines temporal features with global frame encoding	Attained a maximum BLEU score of 46.2
TextKG [27]	MSR-VTTYouCook2MSVD	Utilizes a two-stream transformer. Employs self and cross-attention for multimodal data. Encodes appearance information with ResNet-200. Encodes regional information using Faster-R-CNN Models transcription tokens with GloVe. Uses knowledge graph tokens for object relationships	Achieved a BLEU score of 43.7 on MSR-VTT
Proposed	MSR-VTTMPII Cooking 2M-VAD	Extracts temporal global features using FFT. Utilizes an MBConv-based CNN model for spatial local features. Adopts a short video description approach. Generates captions by maximum matching with ground truth	Reached a peak BLEU score of 51

d. Discussion

Various approaches have been proposed for video captioning, each employing unique techniques to extract and utilize video features for generating descriptive captions. This discussion highlights several noteworthy methods and positions the proposed model within this context.

Vid2Seq implements a dense video captioning strategy that focuses on generating captions by identifying event boundaries within videos. The approach uses transcripted speech data for training and leverages a CLIP-based visual encoder to enhance the descriptive

Received: July 18, 2025

quality of the captions. This method has shown impressive results on datasets like MSR-VTT, MSVD, ViTT, and YouCook2, achieving a peak CIDEr value of 47.2.

The Positive-Augmented Contrastive Learning method utilizes a CLIP-based visual-text encoder to represent video content while enriching the training corpus by associating multiple captions with identical video segments. This augmentation strategy enhances semantic diversity, enabling the model to generalize better across varied linguistic expressions. The learning process incorporates a PACS (Positive-Augmented Contrastive Score)-driven objective, supported by human-generated annotations for robust evaluation. On the VATEX-EVAL benchmark, the model achieves a Machine Correct Caption Score (MCCS) of 0.821, indicating its effectiveness in aligning visual scenes with appropriate textual descriptions.

In contrast, the GL-RG (Global-Local Representation Granularity) framework adopts an incremental training strategy that bridges fine-grained visual events with layered language expressions. By progressively mapping temporal cues and spatial details from videos to hierarchical linguistic structures, the model ensures that the generated captions reflect both global scene context and local action details. This alignment between visual granularity and descriptive clarity plays a crucial role in improving caption accuracy and contextual relevance.It encodes local key frames and combines temporal features with global frame encoding. This approach has shown effectiveness on MSR-VTT and MSVD datasets, achieving a maximum BLEU score of 46.2.TextKG utilizes a sophisticated two-stream transformer model that incorporates self and cross-attention mechanisms to handle multimodal data. It encodes appearance information using ResNet-200 and regional information with Faster-R-CNN, and models transcription tokens with GloVe. Additionally, it uses knowledge graph tokens for object relationships. This method has achieved a BLEU score of 43.7 on the MSR-VTT dataset.

In contrast to these methods, the proposed model employs a CNN encoder specifically designed to extract both temporal and spatial features from video frames using a combination of residual and bottleneck blocks. This allows for efficient capture of intricate details and high-level abstractions in the video data. The RNN component of the model utilizes an LSTM-based architecture for generating captions, leveraging its capability to handle sequential data and long-range dependencies effectively. The proposed model adopts a short video description approach and generates captions through maximum matching with the ground truth, evaluated on MSR-VTT, MPII Cooking 2, and M-VAD datasets. This approach has yielded a peak BLEU score of 51, showcasing its effectiveness in accurately describing video content.

5. Conclusion

In this study, various methods for video captioning were explored, and the distinctive features of the proposed model were highlighted. Existing approaches like Vid2Seq, Positive-Augmented Contrastive Learning, GL-RG, and TextKG leverage different techniques to extract and utilize video features, achieving notable results on various datasets. However, the proposed model stands out by employing a CNN encoder with a combination of residual and

Received: July 18, 2025

bottleneck blocks to effectively capture temporal and spatial features from video frames. The RNN component, based on an LSTM architecture, excels in handling sequential data and long-range dependencies, crucial for accurate caption generation. Evaluated on datasets such as MSR-VTT, MPII Cooking 2, and M-VAD, the proposed model demonstrated superior performance, achieving a peak BLEU score of 51. This highlights its effectiveness in generating high-quality video descriptions. By avoiding reliance on FFT, BiLSTM, and MBConv blocks, the model maintains a streamlined architecture that efficiently balances complexity and performance. The proposed model offers a significant advancement in video captioning by integrating robust feature extraction techniques and sequential data handling, providing a promising solution for generating descriptive and accurate captions for a variety of video content.

References:

[1] S. Elbedwehy, T. Medhat, T. Hamza, and M. F. Alrahmawy, “Enhanced descriptive captioning model for histopathological patches,” *Multimedia Tools and Applications*, pp. 1–20, Jun. 2023, doi: 10.1007/S11042-023-15884-Y/TABLES/6.

[2] G. Rafiq, M. Rafiq, and G. S. Choi, “Video description: A comprehensive survey of deep learning approaches,” *Artificial Intelligence Review 2023 56:11*, vol. 56, no. 11, pp. 13293–13372, Apr. 2023, doi: 10.1007/S10462-023-10414-6.

[3] A. J. Yousif and M. H. Al-Jammas, “Exploring deep learning approaches for video captioning: A comprehensive review,” *e-Prime - Advances in Electrical Engineering, Electronics and Energy*, vol. 6, p. 100372, Dec. 2023, doi: 10.1016/J.PRIME.2023.100372.

[4] S. Han, J. Liu, J. Zhang, P. Gong, X. Zhang, and H. He, “Lightweight dense video captioning with cross-modal attention and knowledge-enhanced unbiased scene graph,” *Complex and Intelligent Systems*, vol. 9, no. 5, pp. 4995–5012, Oct. 2023, doi: 10.1007/S40747-023-00998-5/FIGURES/19.

[5] M. S. Wajid, H. Terashima-Marin, P. Najafirad, and M. A. Wajid, “Deep learning and knowledge graph for image/video captioning: A review of datasets, evaluation metrics, and methods,” *Engineering Reports*, vol. 6, no. 1, p. e12785, Jan. 2024, doi: 10.1002/ENG2.12785.

[6] E. Rashno and F. Zulkernine, “Efficient Video Captioning with Frame Similarity-Based Filtering,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 14147 LNCS, pp. 98–112, 2023, doi: 10.1007/978-3-031-39821-6_7.

[7] W. Xu, J. Yu, Z. Miao, L. Wan, Y. Tian, and Q. Ji, “Deep Reinforcement Polishing Network for Video Captioning,” *IEEE Transactions on Multimedia*, vol. 23, pp. 1772–1784, 2021, doi: 10.1109/TMM.2020.3002669.

[8] H. Tang, C. Yuan, Z. Li, and J. Tang, “Learning attention-guided pyramidal features for few-shot fine-grained recognition,” *Pattern Recognition*, vol. 130, p. 108792, Oct. 2022, doi: 10.1016/J.PATCOG.2022.108792.

Received: July 18, 2025

[9] H. Tang, J. Liu, S. Yan, R. Yan, Z. Li, and J. Tang, “M3Net: Multi-view Encoding, Matching, and Fusion for Few-shot Fine-grained Action Recognition,” *MM 2023 - Proceedings of the 31st ACM International Conference on Multimedia*, pp. 1719–1728, Oct. 2023, doi: 10.1145/3581783.3612221.

[10] S. Yan, H. Tang, L. Zhang, and J. Tang, “Image-Specific Information Suppression and Implicit Local Alignment for Text-Based Person Search,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023, doi: 10.1109/TNNLS.2023.3310118.

[11] S. Yan, N. Dong, L. Zhang, and J. Tang, “CLIP-Driven Fine-Grained Text-Image Person Re-Identification,” *IEEE Transactions on Image Processing*, vol. 32, pp. 6032–6046, 2023, doi: 10.1109/TIP.2023.3327924.

[12] M. Nabati and A. Behrad, “Video captioning using boosted and parallel Long Short-Term Memory networks,” *Computer Vision and Image Understanding*, vol. 190, p. 102840, Jan. 2020, doi: 10.1016/J.CVIU.2019.102840.

[13] M. Chohan, A. Khan, M. S. Mahar, S. Hassan, A. Ghafoor, and M. Khan, “Image Captioning using Deep Learning: A Systematic Literature Review,” *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 5, pp. 278–286, 2020, doi: 10.14569/IJACSA.2020.0110537.

[14] J. Mun, L. Yang, Z. Ren, N. Xu, and B. Han, “Streamlined Dense Video Captioning,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 6581–6590, Apr. 2019, doi: 10.1109/CVPR.2019.00675.

[15] T. Fujii, Y. Sei, Y. Tahara, R. Orihara, and A. Ohsuga, “‘Never fry carrots without cutting.’ Cooking Recipe Generation from Videos Using Deep Learning Considering Previous Process,” *Proceedings - 2019 IEEE/ACIS 4th International Conference on Big Data, Cloud Computing, and Data Science, BCD 2019*, pp. 124–129, May 2019, doi: 10.1109/BCD.2019.8885222.

[16] X. Zhang, X. Wang, X. Tang, H. Zhou, and C. Li, “Description Generation for Remote Sensing Images Using Attribute Attention Mechanism,” *Remote Sensing 2019, Vol. 11, Page 612*, vol. 11, no. 6, p. 612, Mar. 2019, doi: 10.3390/RS11060612.

[17] D. J. Kim, J. Choi, T. H. Oh, and I. S. Kweon, “Dense relational captioning: Triple-stream networks for relationship-based captioning,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, pp. 6264–6273, Jun. 2019, doi: 10.1109/CVPR.2019.00643.

[18] J. Dong, X. Li, and C. G. M. Snoek, “Predicting Visual Features from Text for Image and Video Caption Retrieval,” *IEEE Transactions on Multimedia*, vol. 20, no. 12, pp. 3377–3388, 2018, doi: 10.1109/TMM.2018.2832602.

[19] C. I. Orozco, M. E. Buemi, and J. J. Berlles, “Video to Text Study using an Encoder-Decoder Networks Approach,” *Proceedings - International Conference of the Chilean Computer Science Society, SCCC*, vol. 2018-November, Jul. 2018, doi: 10.1109/SCCC.2018.8705254.

[20] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to Text: Survey of Image and Video Captioning," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 4, pp. 297–312, Aug. 2019, doi: 10.1109/TETCI.2019.2892755.

[21] R. Shetty, H. R. Tavakoli, and J. Laaksonen, "Image and Video Captioning with Augmented Neural Architectures," *IEEE Multimedia*, vol. 25, no. 2, pp. 34–46, Apr. 2018, doi: 10.1109/MMUL.2018.112135923.

[22] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018, doi: 10.1016/J.NEUCOM.2018.05.080.

[23] E. Daskalakis, M. Tzelepi, and A. Tefas, "Learning deep spatiotemporal features for video captioning," *Pattern Recognition Letters*, vol. 116, pp. 143–149, Dec. 2018, doi: 10.1016/J.PATREC.2018.09.022.

[24] A. Yang *et al.*, "Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 10714–10726, 2023, doi: 10.1109/CVPR52729.2023.01032.

[25] S. Sarto, M. Barraco, M. Cornia, L. Baraldi, and R. Cucchiara, "Positive-Augmented Contrastive Learning for Image and Video Captioning Evaluation," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2023-June, pp. 6914–6924, 2023, doi: 10.1109/CVPR52729.2023.00668.

[26] L. Yan *et al.*, "GL-RG: Global-Local Representation Granularity for Video Captioning," *IJCAI International Joint Conference on Artificial Intelligence*, pp. 2769–2775, May 2022, doi: 10.24963/ijcai.2022/384.

[27] X. Gu, G. Chen, Y. Wang, L. Zhang, T. Luo, and L. Wen, "Text with Knowledge Graph Augmented Transformer for Video Captioning," pp. 18941–18951, Mar. 2023, doi: 10.1109/cvpr52729.2023.01816.

[28] P. Anderson *et al.*, "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, Jul. 2017, doi: 10.1109/CVPR.2018.00636.

[29] S. Lee and I. Kim, "Multimodal Feature Learning for Video Captioning," *Mathematical Problems in Engineering*, vol. 2018, 2018, doi: 10.1155/2018/3125879.

[30] Mulani, A. O., Birajadar, G., Ivković, N., Salah, B., & Darlis, A. R. (2023). Deep learning based detection of dermatological diseases using convolutional neural networks and decision trees. *Traitement du Signal*, *40*(6), 2819.

[31] Mulani, A.O., Kulkarni, T.M. (2025). Face Mask Detection System Using Deep Learning: A Comprehensive Survey. In: Singh, S., Arya, K.V., Rodriguez, C.R., Mulani, A.O. (eds) Emerging Trends in Artificial Intelligence, Data Science and Signal Processing. AIDSP 2023. Communications in Computer and Information Science, vol 2439. Springer, Cham. https://doi.org/10.1007/978-3-031-88759-8_3.

Received: July 18, 2025

[32] Karve, S., Gangonda, S., Birajadar, G., Godase, V., Ghodake, R., Mulani, A.O. (2025). Optimized Neural Network for Prediction of Neurological Disorders. In: Singh, S., Arya, K.V., Rodriguez, C.R., Mulani, A.O. (eds) Emerging Trends in Artificial Intelligence, Data Science and Signal Processing. AIDSP 2023. Communications in Computer and Information Science, vol 2440. Springer, Cham. https://doi.org/10.1007/978-3-031-88762-8_18.

Received: July 18, 2025