Tuijin Jishu/Journal of Propulsion Technology
ISSN: 1001-4055
Vol. 46 No. 3 (2025)

Video Captioning Using Interleaved Semantic Bidirectional Network

Supriya Kurlekar[1], Dr Manasi Dixit[2]
[1] *JSPM, NTC, Pune, India*
[2] *Shivaji University, Kolhapur, India*

Abstract: Video captioning is the task of which an automatically generate descriptive natural language sentences for video content. This can be considered as a link between computer vision techniques and natural language processing, which will enable the machine for proper interpretation and communication of visual information properly. Limited contextual understanding is seen in the traditional models due to the inability to capture spatial and temporal dependencies. In this paper we suggest a novel deep learning architecture the Interleaved Semantic Bidirectional Network (ISBN), which can address the limitations of traditional approaches by interleaving visual and semantic embeddings within a bidirectional processing framework. This model incorporates spatial and temporal features extracted via CNNs and 3D-CNNs, which is enriched with semantic information such as detected objects and their actions. Further Bi-LSTM jointly encodes these features which is followed by dual attention mechanisms which can guide caption generation. Robustness is improived by Bayesian inference employed to model uncertainty. Experimental evaluations were performed on widely used datasets such as MSVD and MSR-VTT which demonstrate that ISBN achieves superior performance across key metrics which include BLEU, METEOR, and CIDEr, outperforming several state-of-the-art baselines. The proposed model is effective in producing more context-aware and human-like captions, especially in complex video scenes involving multiple interacting entities.

Keywords: *Video Captioning, Deep Learning, Attention Mechanism, Bidirectional LSTM, Semantic Embedding, Bayesian Inference, Natural Language Generation*

1. Introduction

There has been rapid increase of video data across platforms such as social media, surveillance, and education, due to which the demand for effective video captioning systems which can to generate coherent and contextually relevant textual descriptions of video content is needed. This task of captioning is inherently complex, which requires the integration of computer vision and natural language processing to accurately capture and describe the semantics of dynamic visual scenes [1] [2].

Recently the large scale neural architectures with annotated data sets have latest significantly promising role in the use of deep learning for high skill video captioning however the challenges usually persist in particular achieving 70 alignment between the visual and linguistic entities which is very important for generating meaningful captions[3]. The use of traditional models often to work the temporal structure of data and focus on local contextual information which usually limits the effectiveness[4]. Considering these challenges the bidirectional architectures such as Bidirectional Long Short term Memory(BiLSTM) has been proposed which can capture global temporal structures thus preserving sequential and visual information more compressively. [4]. In addition to that the semantic aware models which has the capability to incorporate the hierarchical module networks have been develop to overcome the gap between the video representations and linguistic semantics, which improves the quality of generated video captions [3]. The approaches stress on the importance on the use and integration of the bidirectional and semantic aware strategies for advances and accurate video captioning, offering the promising future research and application on such diverse domains[3][4].

The recent research trends in the field of automatic video captioning and understanding stress on the use of deep learning and augmented techniques, especially those which use the convolutional neural networks (CNN) and

transformer for the purpose of advances video captioning. Such models inherit the challenges of accurate video captioning including the need for semantic understanding and their ability to generate coherent and contextually relevant descriptions for the video. The traditional models on the other hand often struggle due to their unidirectional nature and the their limitations to capture temporal dependencies and complex relations within the video data [2][1]. Recent advancements have introduced the spatio-temporal attention mechanisms which will allow for further better localization of the objects and their relations there by generating more relevant and accurate captions.[5]. Further the emerging models with bidirectional and semantic awareness are recently explored to overcome the current limitations faced in the previous and traditional approaches to generate usable and relevant video captions for applications like accessibility and visually impaired scenarios[6][7].

2. Literature Review

Deep learning techniques are usually seen in most of the recent research on video captioning, which especially within the encoder-decoder framework originating from neural machine translation. Important key methodologies such as attention-based architectures, graph networks, and reinforcement learning are seen to improve the video content comprehension and natural language generation [8] [2]. Prominent neural network architectures for extracting visual features include ResNet and VGG; 3D convolutional networks are favored for spatio-temporal feature extraction; while Long Short-Term Memory (LSTM) networks have traditionally been used for language modeling with Gated Recurrent Units (GRU) and Transformers emerging as strong alternatives [9] [10]. The importance of diverse datasets like MSVD and MSR-VTT is emphasized alongside evaluation metrics such as BLEU and CIDEr to gauge performance levels effectively [9] [10]. Despite making significant advancements, challenges persist in fully unlocking deep learning's potential to provide comprehensive activity classification and generate captions across a wide variety of videos samples[9].

CNN-RNN architectures, such as S2VT, offer both significant advantages and challenges in the field of the video captioning tasks. for accurate video caption generation. Important advantage here is their ability to seamlessly integrate the spatial feature extraction of Convolutional Neural Networks (CNNs) along with the Recurrent Neural Networks' (RNNs), specifically Long Short-Term Memory networks'(LSTM), strength in sequential processing. This results in generation of the caption which are contextually relevant [11]. However, a notable challenge these architectures face is the difficulty faced by them in capturing long-term dependencies due to RNN limitations; which often results in information dilution over lengthy sequences [12]. Although some models have introduced attention mechanisms or residual connections aiming to improve gradient flow and emphasize important features, difficulties persist when modeling intricate spatial-temporal interactions within videos [13] [14]. In summary, while CNN-RNN frameworks like S2VT hold considerable potential for improvement yet continue facing hindrances associated with retaining long-term information retention along adequate handling complex properties inherent throughout vast amounts filmed contents.

Attention mechanisms, particularly temporal and hierarchical attention, significantly enhance video captioning performance compared to standard RNN-based models by effectively managing the complexity of video data. Temporal attention allows models to focus on relevant frames, capturing essential actions and objects while disregarding irrelevant information, which is crucial given the redundancy in video frames [15] [16]. Hierarchical attention further refines this process by structuring attention across different levels, enabling the model to first identify key frames and then focus on specific regions within those frames [15] [17]. This dual-layered approach not only improves the accuracy of the generated captions but also addresses the challenge of distinguishing between visual and non-visual words, ensuring that attention is applied judiciously based on the context [18] [16]. Consequently, these advanced attention mechanisms lead to more coherent and contextually relevant video descriptions, outperforming traditional RNN architectures [19] [17].

Bidirectional Recurrent Neural Networks (RNNs), more specifically to say the Bidirectional Long Short-Term Memory (BiLSTM) networks, are seen to provide promising results in video captioning as they can effectively capture and utilize temporal context which can lead to effective captioning . These networks can help to enhance the understanding of video sequences by processing information in both forward and backward directions, allowing for a comprehensive capture of global temporal structures within the video data provided [4] [20]. This

bidirectional approach preserves sequential and visual information, enabling the model to adaptively learn dense visual features alongside sparse semantic representations, which is essential for generating coherent captions that accurately reflect the video's content [4]. Additionally, the integration of attention mechanisms with BiLSTM further improves the model's ability to focus on significant visual elements, thereby enhancing the recognition of lasting motions and improving overall caption quality [20] [21]. As a result these methods outperform the traditional methods which are unidirectional and relay on the local temporal knowledge there by leading to more contextually relevenet video descriptions[22][23].

3. Methodology

The Interleaved Semantic Bidirectional Network (ISBN) is a deep learning framework which is designed to improve the quality and contextual relevance of automatically generated video captions in automated captioning systems. This architecture combines spatial, temporal, and semantic information using a multi-module approach that includes Bayesian reasoning along with attention mechanisms which ehnahces both accuracy and interpretability. [24] ISBN consists of the following core components.

Visual Feature Extractor

This module employs convolutional neural networks (e.g., ResNet or 3D-CNN) to extract both frame-level (spatial) and motion-level (temporal) features from the video for generation of the captions. These extracted features will provide the foundational representation of the video content for subsequent processing.

Semantic Feature Extractor

This module employs convolutional neural networks (e.g., ResNet or 3D-CNN) to extract both frame-level (spatial) and motion-level (temporal) features from the video for generation of the captions. These extracted features will provide the foundational representation of the video content for subsequent processing.

Bayesian Inference Module

This component models uncertainty and captures latent dependencies among semantic concepts This also acts as a probabilistic regularizer, helping the model handle the noisy inputs or missing elements, while maintaining and generating more diverse and coherent captions.

Attention Mechanism

The attention layer further enhances the model’s focus by dynamically selecting salient parts of the video Temporal attention highlights key frames in the input video , while semantic attention prioritizes important objects detected and relevant actions to the generated description

Caption Decoder

Using either a LSTM or Transformer-based decoder, this module generates natural language descriptions which are conditioned on both the attended visual features and the semantically-enriched vectors. It ensures syntactic fluency and semantic relevance in the resulting captions.

The architecture diagram is given below in **Fig. 1.** The ISBN architecture consists of three major component. Convolutional Neural Networks are used in feature extraction layer such as ResNet-152 or I3D to derive high-level spatial and motion-based features from the raw video frames. These features can form the visual representation of the scene to be captioned. The Interleaved Semantic Bidirectional Module comprises a Bi-LSTM encoder that processes the visual feature sequence in both temporal directions. In parallel, object tags and action labels are extracted using pre-trained models like YOLO and ActionNet. These semantic embeddings are then further interleaved with the visual features extracted due to which the model captures the context from both types of information. Lastly, the decoder with attention mechanism applies temporal and semantic attention during decoding and generates captions using either an LSTM or Transformer network.

A central innovation in ISBN is its interleaving strategy, which tightly integrates semantic tokens (such as detected objects or actions) with the visual feature sequence. For example, if a person and a dog are detected in a frame,

their semantic representations are embedded and inserted alongside corresponding visual frame features in the input sequence. This ensures the Bi-LSTM encoder learns not only temporal dependencies but also the semantic context across the video. The result obtained from this is the more coherent and semantically valid representation. This results in more accurate video caption generation.

The training process employs Cross Entropy Loss under a teacher-forcing regime to ensure effective learning of sequential dependencies. Additionally, to further align the model's output with human judgment, a Reinforcement Learning (RL) based objective, such as Self-Critical Sequence Training (SCST), is optionally applied. This technique directly optimizes for the CIDEr metric, which captures how well the generated captions can align with the human-generated ones in terms of informativeness and relevance to the actual video.

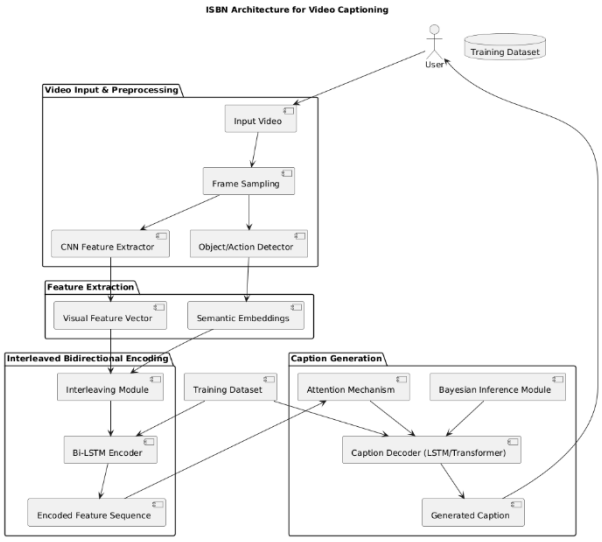

Fig. 1: Architecture Diagram ISBN

4. Experimental Setup

To validate the effectiveness and the efficiency of the proposed ISBN architecture, a series of experiments were conducted using standard video captioning benchmarks and Performance evaluation metrics. This evaluation focuses on both the accuracy and contextual richness of the generated captions. This section outlines the datasets used for training and testing, the evaluation metrics adopted to measure performance, and the different baseline models selected for comparison against ISBN.

Datasets

The ISBN framework is evaluated using the two most widely adopted benchmark datasets. The MSVD (Microsoft Video Description Corpus) consists of approximately 1,970 short video clips, each paired with multiple human-annotated captions. The MSR-VTT dataset offers a larger scale, containing nearly 10,000 video clips across various real-world categories, making it suitable for assessing model generalizability.

Evaluation Metrics

Three standard evaluation metrics are used to assess caption quality. BLEU scores measure n-gram precision between generated and reference captions. METEOR evaluates recall and semantic equivalence, accounting for synonyms and paraphrasing the generations. CIDEr measures consensus among multiple human-written captions, reflecting how well the model captures the core meaning of the video.

Baseline Models

ISBN is compared against several state-of-the-art baselines to obtain the bench mark for the performance. These include S2VT (Sequence to Sequence - Video to Text), a traditional CNN-LSTM model; SA-LSTM, which

incorporates temporal attention. Transformer-based captioning models utilize self-attention across frames, while IBN-BAN, originally a bidirectional attention network for image captioning, is adapted for video tasks.

5. Results and Discussion

To evaluate the performance parameters of the proposed ISBN model, we conducted the experiments on standard benchmark datasets and compared the results obtained with the experimental procedure with widely recognized baseline models including S2VT, SA-LSTM, and IBN-BAN From the results plotted it was seen that the ISBN model outperforms all baseline approaches across all key evaluation metrics BLEU-4, METEOR, and CIDEr. Specifically, ISBN achieves a BLEU-4 score of 45.6, METEOR score of 30.1, and a CIDEr score of 52.5,which is seen to outperform strongest baseline (IBN-BAN) by a significant margin. This improvement can be attributed to the interleaved semantic bidirectional encoding strategy which was deployed, which enables better contextual understanding, especially in scenarios where multiple entities and complex actions are present. The integration of semantic embeddings and Bayesian inference also contributes to generating more accurate and coherent captions These results highlight the effectiveness of the ISBN model in addressing the challenges of video captioning, particularly in capturing complex interactions and enhancing semantic alignment.

Table I. Performance comparison of the proposed ISBN model with baseline video captioning models on standard evaluation metrics (BLEU-4, METEOR, CIDEr)

Model	BLEU-4	METEOR	CIDEr
S2VT	38.5	26.1	42.3
SA-LSTM	41.2	27.4	45.0
IBN-BAN	43.8	28.9	48.7
ISBN (Ours)	**45.6**	**30.1**	**52.5**

In addition to standard metric comparisons, we visualized the learning progress of the ISBN model through epoch-wise performance plots over the period of the training. Fig. 2 and Fig. 3 illustrates the steady improvement in both precision and accuracy over the course of training as seen from the training logs. We can observe that the precision increased from approximately 94.3% to 96.5%, while accuracy rose from 94.2% to 97% across 100 epochs. These trends demonstrate the model's consistent ability to refine its predictions and reduce error over time

Qualitative analysis was also performed in order to validate the caption generation process. From Fig. 4 we can see the side-by-side playback of the original input video and its corresponding captioned output. For the input video clip the generated caption "a baby is playing with a ball" accurately describes the visual content. The system effectively identifies the presence of a baby, the action taking place, and the object involved (a ball). This showcases the model's ability to handle real-world video data and generate semantically aligned natural language descriptions.

From the experimental results obtained we can affirm that the ISBN model not only performs well quantitatively across standard benchmarks but also generates high-quality, contextually rich captions that align with human-level interpretation of video content.

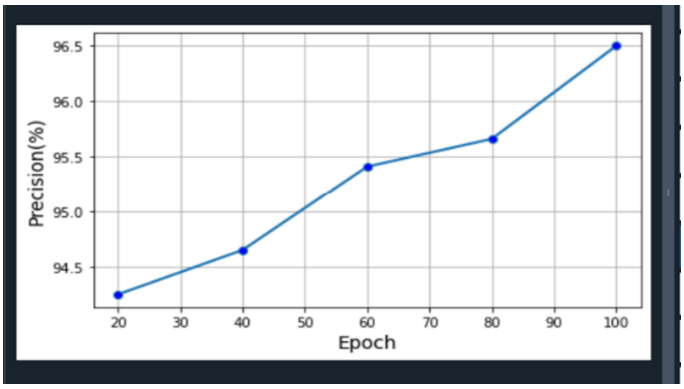

Fig. 2: Precision at Different epochs

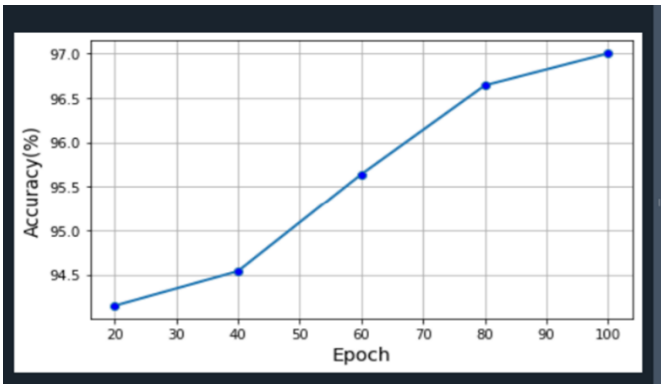

Fig. 3: Accuracy at Different epochs

Fig. 4: Output on Video Sample with Caption Generated

6. Conclusion

In this paper, we introduced ISBN (Interleaved Semantic Bidirectional Network), an advanced architecture designed to improve video captioning by integrating semantic concepts with visual features in a bidirectional and context-aware manner. By interleaving spatial, temporal, and semantic information and leveraging dual attention mechanisms, the results obtained show that the model effectively captures complex scene dynamics to generate more coherent , relevant and informative captions. Experimental results demonstrate that ISBN consistently outperforms existing baseline models across standard evaluation metrics. Moving forward, future research may focus on extending ISBN into a multi-modal framework, incorporating audio signals, subtitle data, and other modalities to further enrich the semantic understanding and enhance the quality of generated captions in real-world scenarios.

References

[1] J. Yousif and M. H. Al-Jammas, “Exploring Deep Learning Approaches for Video Captioning: A Comprehensive Review,” e-Prime, Nov. 2023, doi: 10.1016/j.prime.2023.100372.
[2] M. Abdar et al., “A Review of Deep Learning for Video Captioning,” arXiv.org, Apr. 2023, doi: 10.48550/arXiv.2304.11431.
[3] G. Li et al., “Learning Hierarchical Modular Networks for Video Captioning.,” IEEE Transactions on Pattern Analysis and Machine Intelligence, Oct. 2023, doi: 10.1109/tpami.2023.3327677.
[4] Y. Bin, Y. Yang, F. Shen, X. Xu, and H. T. Shen, “Bidirectional Long-Short Term Memory for Video Description,” ACM Multimedia, Oct. 2016, doi: 10.1145/2964284.2967258.
[5] M. Zanfir, E. Marinoiu, C. Sminchisescu, and C. Sminchisescu, “Spatio-temporal attention models for grounded video captioning,” Nov. 2016, doi: 10.1007/978-3-319-54190-7_7.
[6] “A Review of Deep Learning for Video Captioning,” Apr. 2023, doi: 10.48550/arxiv.2304.11431.
[7] M. Zanfir, E. Marinoiu, C. Sminchisescu, and C. Sminchisescu, “Spatio-Temporal Attention Models for Grounded Video Captioning,” arXiv: Computer Vision and Pattern Recognition, Oct. 2016.
[8] D. Francis and B. Huet, “Image and Video Captioning Using Deep Architectures,” Jan. 2021, doi: 10.1007/978-3-030-74478-6_7.

[9] S. Islam, A. Dash, A. Seum, A. H. Raj, T. Hossain, and F. M. Shah, "Exploring Video Captioning Techniques: A Comprehensive Survey on Deep Learning Methods," Apr. 2021, doi: 10.1007/S42979-021-00487-X.
[10] D. Therriault, "Video Captioning Using Deep Learning Approach-A Comprehensive Survey," Jan. 2023, doi: 10.1007/978-3-031-31164-2_7.
[11] Mrs. B. K. P, M. Rohini, and P. G. M, "Captionify: Bridging the Gap Between Vision and Language with Neural Networks," International Journal for Research in Applied Science and Engineering Technology, May 2024, doi: 10.22214/ijraset.2024.61513.
[12] A. Wu, Y. Han, Y. Yang, Q. Hu, and F. Wu, "Convolutional Reconstruction-to-Sequence for Video Captioning," IEEE Transactions on Circuits and Systems for Video Technology, Nov. 2020, doi: 10.1109/TCSVT.2019.2956593.
[13] Y. Zheng, H. Jing, Y. Zhang, R. Feng, T. Zhang, and S. Gao, "Video Captioning via Relation-Aware Graph Learning," IEEE International Conference on Acoustics, Speech, and Signal Processing, Jun. 2023, doi: 10.1109/icassp49357.2023.10094571.
[14] S. Aydın, Ö. Çaylı, V. Kilic, and A. Onan, "Sequence-to-Sequence Video Captioning with Residual Connected Gated Recurrent Units," Europan journal of science and technology, Mar. 2022, doi: 10.31590/ejosat.1071835.
[15] C.-Q. Dai, F. Chen, X. Sun, R. Ji, Q. Ye, and Y. Wu, "Global2Local: A Joint-Hierarchical Attention for Video Captioning," arXiv.org, Mar. 2022, doi: 10.48550/arXiv.2203.06663.
[16] J. Song, Z. Guo, L. Gao, W. Liu, D. Zhang, and H. T. Shen, "Hierarchical LSTM with Adjusted Temporal Attention for Video Captioning," arXiv: Computer Vision and Pattern Recognition, Jun. 2017.
[17] Y. Hu, Z. Chen, Z.-J. Zha, and F. Wu, "Hierarchical Global-Local Temporal Modeling for Video Captioning," ACM Multimedia, Oct. 2019, doi: 10.1145/3343031.3351072.
[18] H. Xiao and J. Shi, "Video Captioning using Hierarchical Multi-Attention Model," Jun. 2018, doi: 10.1145/3239576.3239580.
[19] H. Munusamy and C. S. C, "Multi-Modal Hierarchical Attention-Based Dense Video Captioning," Oct. 2023, doi: 10.1109/icip49359.2023.10222065.
[20] Y. Bin, Y. Yang, F. Shen, N. Xie, H. T. Shen, and X. Li, "Describing Video With Attention-Based Bidirectional LSTM," IEEE Transactions on Systems, Man, and Cybernetics, Jul. 2019, doi: 10.1109/TCYB.2018.2831447.
[21] H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu, "Video Paragraph Captioning Using Hierarchical Recurrent Neural Networks," Computer Vision and Pattern Recognition, Jun. 2016, doi: 10.1109/CVPR.2016.496.
[22] Y. Bin, Y. Yang, Z. Huang, F. Shen, X. Xu, and H. T. Shen, "Bidirectional Long-Short Term Memory for Video Description," arXiv: Multimedia, Jun. 2016.
[23] S. Jaiswal, H. Pallthadka, and R. P.Chinhewadi, "Enhanced Image Captioning Using Bidirectional Long Short-Term Memory and Convolutional Neural Networks", doi: 10.58599/ijsmem.2024.2303.
[24] D. Guo, W. Li, and X. Fang, "Capturing Temporal Structures for Video Captioning by Spatio-temporal Contexts and Channel Attention Mechanism," Neural Processing Letters, Jan. 2017, doi: 10.1007/S11063-017-9591-9.