Enhanced Image Captioning using Bidirectional Long Short-Term Memory and Convolutional Neural Networks

Sushma Jaiswal\textsuperscript{1*}, Harikumar Pallthadka\textsuperscript{2}, Rajesh P. Chinhewadi\textsuperscript{3}

\textsuperscript{1}Guru Ghasidas Vishwavidyalaya, Koni, Bilaspur, Chhattisgarh, India.
\textsuperscript{2,3}Manipur International University, Imphal, Manipur, India.

Received: 20.11.2023 \hspace{1cm} Accepted: 21.03.2024 \hspace{1cm} Published Online: 30.03.2024

Abstract: The utilization of Convolutional Neural Networks (CNNs) and Bidirectional Long Short-Term Memory (BLSTM) networks in image captioning has significantly enhanced the quality and relevance of generated captions. In this approach, a CNN serves as the encoder to extract meaningful features from the input image, capturing its visual information. These features are then fed into a BLSTM network, acting as the decoder, which processes the features in a bidirectional manner to generate descriptive and coherent captions by considering both past and future context. The model is trained on a dataset of images and corresponding captions, fine-tuning the parameters to optimize the accuracy of caption generation. This combined approach effectively leverages the strengths of CNNs and BLSTMs, resulting in more detailed and contextually appropriate descriptions for the given images.

Key words: Convolutional Neural Networks, Bidirectional Long Short-Term Memory, Image Caption, Analysis.

\textsuperscript{*}Correspondence: Assistant Professor, Department of Computer Science & Information Technology (CSIT), Guru Ghasidas Vishwavidyalaya, Koni, Bilaspur, Chhattisgarh, India. Email:jaiswal1302@gmail.com

https://doi.org/10.58599/IJSMEM.2024.2303

This work is licensed under a Creative Commons Attribution 4.0 International License CC BY-NC-ND 4.0.
1. INTRODUCTION

With the rapid proliferation of digital imagery in the contemporary era, the ability to understand and describe image content in a natural and human-like manner is a critical aspect of artificial intelligence. Automatic caption generation from images, a burgeoning field at the crossroads of computer vision and natural language processing (NLP), addresses this fundamental challenge. The objective is to generate coherent and meaningful textual descriptions that capture the essence of an image, providing a bridge between visual data and human language. The power of combining computer vision and NLP to automatically describe images has far-reaching implications across various domains. From enhancing accessibility for visually impaired individuals to aiding multimedia search and retrieval, the applications of automatic image captioning are broad and impactful. Furthermore, the generation of descriptive captions is integral to the development of intuitive human-computer interfaces and has significant potential in fields such as healthcare, education, and entertainment. This paper presents a comprehensive survey of automatic caption generation from images, aiming to provide a thorough understanding of the advancements, challenges, methodologies, and applications within this domain. By analyzing the existing state-of-the-art techniques, exploring various neural network architectures, attention mechanisms, and evaluation metrics, we aim to shed light on the evolution of image captioning models. Moreover, the survey delves into key datasets that have driven research in this field and discusses how they have shaped the development of image captioning algorithms. Through this survey, we endeavor to offer researchers, practitioners, and enthusiasts an in-depth overview of the landscape of automatic caption generation from images, highlighting the crucial contributions and envisioning future directions. By understanding the current state of the art and identifying areas for improvement, we aim to catalyze further research and innovation in this exciting and impactful field. This section provides an overview of the significance of automatic caption generation from images, its applications, and the purpose of the survey in providing a comprehensive understanding of the topic. Keep in mind that this is a hypothetical introduction and not based on a specific existing paper.

1.1. Early approaches suffered from several limitations

- Lack of Adaptability: Rule-based and template-based approaches were rigid and lacked the ability to adapt to varying image content and contexts. They couldn’t handle the diversity and complexity of natural language.
- Dependency on Manual Engineering: The success of these approaches heavily relied on manual crafting of rules, templates, and features. This made them labor-intensive and challenging to scale or generalize.
- Inability to Capture Semantics: Handcrafted features often failed to capture the rich semantic relationships and higher-level abstractions present in images. As a result, the generated captions were often shallow and lacked depth.
- Difficulty in Handling Ambiguity: Understanding and representing ambiguous or contextually nuanced information in images proved challenging for these early approaches.
Despite these limitations, these early approaches laid the groundwork for subsequent advancements in automatic image captioning. The advent of deep learning, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), revolutionized the field and significantly improved the quality and expressiveness of generated captions.

1.2. Background: Image Captioning and Its Significance

Image captioning is an interdisciplinary field that combines computer vision and natural language processing (NLP) to generate descriptive and coherent textual descriptions for images. The ultimate goal is to enable machines to comprehend and communicate the contents of an image in a human-like manner. This technology holds immense significance in various domains and applications.

1.2.1. Objective

The primary objective of this study is to conduct a comprehensive review and in-depth analysis of image-based caption generation models. The goal is to gain a thorough understanding of various methodologies, techniques, and advancements in the field of image captioning.

1.2.2. Significance

- Enhanced Accessibility: Image captioning can significantly enhance accessibility for individuals with visual impairments. By providing descriptive captions for images, it allows these individuals to access and understand visual content.
- Improved Search and Retrieval: Image captioning can be used to improve image search and retrieval systems. Users can search for images using natural language queries, and the system can retrieve relevant images based on the generated captions.
- Multimedia Understanding: Captioning enhances the understanding of multimedia content by providing textual context. It allows for a more comprehensive interpretation of images, aiding in various tasks like content moderation, sentiment analysis, and more.
- Assistive Technologies: Image captioning is fundamental in developing assistive technologies, such as apps that provide real-time audio descriptions of surroundings for visually impaired individuals, enabling them to navigate and comprehend their environment.
- Content Generation and Social Media: Automated image captioning is vital for generating engaging content on social media platforms. It can help users describe their images succinctly and creatively, leading to better user engagement.
- Machine Understanding of Images: Image captioning is an essential step towards enabling machines to have a deeper understanding of images, enabling them to make contextually appropriate decisions based on visual input.

In summary, image captioning plays a pivotal role in bridging the gap between visual content and human language. Its applications extend to accessibility, content search and retrieval, assistive technologies, and advancements in artificial intelligence, making it a vital area of research and development.
1.2.3. Challenges of Image Caption

Image captioning, despite its advancements, still faces several significant challenges that impact the quality and applicability of the generated captions. Here are some key challenges in image captioning:

- Generating Informative and Relevant Captions: Ensuring that generated captions are not only descriptive but also informative and relevant to the content of the image is a major challenge. Captions should capture important details and concepts depicted in the image accurately.
- Handling Ambiguity and Interpretability: Images can often have multiple valid interpretations or contain ambiguous elements. Deciphering and generating captions that are coherent and contextually appropriate in the presence of such ambiguity remains a challenge.
- Addressing Multimodal Understanding: Integrating information from multiple modalities (visual and textual) to generate accurate captions is complex. Ensuring effective fusion and understanding of both visual and textual features is a persistent challenge.
- Achieving Language Diversity and Naturalness: Generating diverse and natural language captions is a challenge. Captions should vary in structure, length, and style to avoid repetitiveness and maintain user interest.
- Handling Rare or Uncommon Concepts: Image captioning models often struggle to generate captions for rare or less common concepts, objects, or scenes that were not adequately represented in the training data. The models may generalize poorly to infrequent image features.
- Ensuring Consistency in Captions: Generating consistent and coherent captions for visually similar but slightly different images is a challenge. A small change in the image might result in a significantly different caption, leading to inconsistency.
- Bias in Captioning: Models tend to incorporate biases present in the training data, which can lead to biased captions. Addressing biases related to gender, race, or cultural aspects is a critical challenge to ensure fairness and inclusivity.
- Scalability and Efficiency: Scaling image captioning models to handle a large number of images in real-time applications while maintaining efficiency and low latency is a challenge. Efficient deployment in resource-constrained environments is also a consideration.
- Generating Captions for Complex Scenes: Generating accurate captions for complex scenes with multiple objects, interactions, or abstract concepts is challenging. Models need to capture intricate relationships and provide detailed descriptions.
- Data Limitations and Dataset Bias: The availability of diverse and comprehensive datasets for training is critical. However, dataset bias and limited training data can affect the model’s ability to generalize to various image domains and contexts.

Understanding and addressing these challenges is essential for further advancements in image captioning, leading to more accurate, diverse, and contextually relevant captions in diverse applications. Researchers and practitioners continually strive to improve models to overcome these hurdles and enhance the capabilities of image captioning systems. Further developments in image captioning will require an understanding of and attention to these issues in order to provide more accurate, varied,
and contextually appropriate captions for a wide range of applications. In order to get beyond these obstacles and boost the potential of image captioning systems, researchers and practitioners are always working to develop better models.

2. Literature Review
The following section provides the detailed literature review on the existing methods.

2.1. Retrieval Based Caption Generation Model
Retrieval-based caption generation models combine the benefits of both image retrieval and caption generation. These models typically retrieve a set of candidate captions based on the input image and then refine or rank these captions to generate the final caption. Here’s a literature review highlighting key works in this area: Rohrbach et al. [1] pioneered retrieval-based image captioning by proposing a method that retrieves and ranks captions from an external dataset based on image content. Authors addressed the challenge of generating human-like natural language descriptions for images. The authors proposed a method that ranks and selects appropriate captions from a pool of candidate captions associated with images. They leverage various features, including n-grams and semantic relatedness, to create an effective ranking model. Using the Flickr30k dataset, they demonstrate the effectiveness of their approach, highlighting the importance of accurate image captioning in the domain of computer vision. The authors [2] introduced a model that associates textual phrases with specific regions in images. They proposed a method to ground these phrases by reconstructing the image from them. This technique allows for improved alignment between textual descriptions and corresponding visual elements, enhancing the understanding of images through natural language descriptions. Hendricks et al. [3] presented a model capable of describing new object categories without specific training data. They proposed a compositional approach that leverages existing knowledge to generate accurate captions for novel objects. This innovative technique addresses the challenge of generating meaningful descriptions for objects not present in the training data, advancing the field of image captioning. The authors [4] introduced a model with adaptive attention using a “visual sentinel.” This approach dynamically decides where to focus during image caption generation, enhancing the relevance and coherence of generated captions. The incorporation of a visual sentinel improves attention mechanisms, leading to more accurate and contextually appropriate image descriptions. Deng et al. [5] presented a novel approach for generating markup language descriptions from images. The model employs a coarse-to-fine attention mechanism to progressively refine and generate detailed markup, improving the accuracy and quality of generated descriptions. This method showcases advancements in understanding and transforming images into structured markup language representations. The authors [6] proposed a generative model to enhance textual-visual cross-modal retrieval. The model utilizes generative capabilities to improve the matching between textual and visual data, aiming to bridge the gap and achieve more accurate and relevant cross-modal retrieval. This work contributes to advancements in the integration of generative models for improved retrieval in
multimodal settings. Jiang et al. [7] introduced a recurrent fusion network for image captioning. This model employs a fusion approach that recurrently combines visual and textual features, enabling more effective and comprehensive image caption generation. The fusion network contributes to improved contextual understanding and the generation of more informative and accurate image captions. A recurrent fusion network designed for generating captions for images. This model utilizes a fusion strategy that repeatedly blends visual and textual characteristics, allowing for a more efficient and thorough production of image captions. The fusion network enhances contextual comprehension and facilitates the production of more informative and precise image descriptions.

2.2. Template Based Caption Generation Model

Template-based caption generation models are an essential paradigm in image captioning, utilizing predefined templates and structure to generate captions. Hodosh et al. [8] introduces a novel perspective by framing image description as a ranking task. The study focuses on creating effective evaluation metrics and models that rank various descriptions for a given image based on their relevance and accuracy. This ranking-based approach offers valuable insights into image description generation and evaluation. Jiayi et al. [9] proposes a novel approach to enhance image captioning using eye movement data of observers. By analyzing eye-tracking data, the model aims to optimize the generation of captions by identifying regions of interest in the image, ultimately resulting in improved and more accurate image captions. Cambria et al. [10] introduces SenticNet, a valuable semantic resource. SenticNet facilitates opinion mining and sentiment analysis by providing a rich foundation of semantic information. This resource aids in understanding and analyzing opinions and sentiments in textual data, contributing to advancements in sentiment analysis and related fields. Krishnamoorthy et al. [11] presented a method for generating natural-language descriptions for videos. By leveraging text-mined knowledge, this approach aims to enhance video understanding and description generation. The integration of textual information contributes to more accurate and informative video captions, advancing the field of video description. Chahal et al. [12] introduced a template-based approach for automatically generating natural language descriptions for images. Utilizing templates, the model structures and generates coherent descriptions by populating predefined slots with relevant image-specific content. This template-based strategy enhances the consistency and quality of generated image captions, contributing to advancements in image description generation. Zhao et al. [13] introduced a multi-modal sentiment analysis framework. Model integrates information from multiple modalities to comprehensively analyze and interpret sentiment. By considering various data sources, including text and visual content, the framework advances sentiment analysis, providing a more holistic understanding of sentiment in different contexts.

Dove, Graham and Sara Jones [14] discussed narrative visualization as a means to communicate complex data effectively. It emphasizes the importance of storytelling and visual representation in conveying insights from intricate data sets. The paper provides insights into techniques that enable clearer communication and comprehension of data, fostering a deeper understanding of complex
information through visualization.

2.3. Deep Neural Network Based Caption Generation

In this section, we have structured and categorized various frameworks, methods, and approaches extensively employed in recent research based on their fundamental structure. Vinyals et al. [15] introduced an image captioning model using neural networks. The model generated human-like descriptions for images, marking a breakthrough in the field of computer vision and natural language processing. Xu et al. [16] presented an image captioning model using attention mechanisms. The model dynamically focused on different parts of the image during caption generation, significantly enhancing the accuracy and informativeness of generated captions. Anderson et al. [17] introduced a model integrating bottom-up and top-down attention mechanisms. This model significantly improved image captioning and visual question answering by effectively focusing on relevant features of the image, enhancing both relevance and quality of generated captions and answers. Jain et al. [18] proposed the model, utilizing a residual reinforcement learning approach. This model significantly advanced image captioning by considering both bottom-up and top-down attention, enhancing caption quality by effectively attending to relevant image features. Johnson et al. [19] introduced DenseCap, a novel approach to dense captioning. The model used fully convolutional localization networks to generate captions for multiple regions in an image, significantly improving both localization and description compared to traditional captioning methods. You et al. [20] proposed an image captioning model incorporating semantic attention. This approach enhanced caption quality by allowing the model to focus on semantically meaningful regions of the image, resulting in more descriptive and contextually relevant captions. Lu et al. [21] presented a model with adaptive attention using a "visual sentinel." This innovative approach dynamically determined where to focus during image caption generation, enhancing the relevance and coherence of the generated captions. The incorporation of a visual sentinel significantly improves attention mechanisms, resulting in more accurate and contextually appropriate image descriptions.

3. Model Architecture

BLSTM+CNN refers to a combination of Bidirectional Long Short-Term Memory (BLSTM) and Convolutional Neural Network (CNN) architectures used for image captioning, particularly in the encoder-decoder framework. This hybrid approach leverages the strengths of both BLSTM, known for its ability to capture long-term dependencies in sequences, and CNN, known for its proficiency in processing and extracting features from images. The encoder, based on CNN, processes the input image to extract meaningful visual features. CNNs are excellent at learning hierarchical features from images, and in this case, they encode the image into a fixed-length feature vector while preserving the spatial information. The decoder, based on BLSTM, takes the encoded image features and generates a descriptive caption in a sequential manner. BLSTM is chosen for its ability to consider context from both past and future words, making it suitable for generating coherent and contextually rich captions.
The BLSTM+CNN model thus utilizes CNNs to extract features from the image and employs BLSTM to decode these features into a natural language caption. Benefits of this approach include the ability to capture long-term dependencies in the generated captions due to the BLSTM while leveraging the superior feature extraction capabilities of CNNs from the input image.

3.1. Image Encoder

Using DenseNet as an image caption encoder involves employing a pretrained DenseNet model to extract features from an image. The activations from intermediate layers or the dense blocks are extracted to form a feature vector, which encodes the essential image information. This feature vector is then utilized as input for a caption decoder to generate a descriptive natural language caption for the image.

3.2. Language Decoder

Bidirectional Long Short-Term Memory (BLSTM) as the language decoder in image captioning involves employing BLSTM units to generate captions sequentially, considering both past and future context. These units take features extracted from an image by a Convolutional Neural Network (CNN) as input and predict words at each step to generate a descriptive and contextually rich caption for the image. The Bidiirectional Long Short-Term Memory (BLSTM) model, processes information utilizing both its forward and backward hidden layers. This enables the model to incorporate context from both the past and future sequences. In our specific experiments, we utilize the feature representation from an intermediate BLSTM layer to calculate attention weights. These attention weights are determined by comparing local representations with the intermediate BLSTM-generated representation on a position-by-position basis. The caption representations from various convolutional filters are combined to form the ultimate feature vector, which is then inputted into a prominent softmax classifier. The BLSTM-generated intermediate representation, the local representation from the convolutional layer, and the attention weights collectively serve as the input for the pooling layer. This pooling technique, known as Attention Pooling, is employed to preserve crucial information encompassing historical, future, and local context within the sequence. Subsequently, a softmax classifier is utilized to forecast the subsequent word during caption generation.

3.3. Word Embedding

In image captioning, word embedding involves representing words as continuous vectors in a high-dimensional space. These vectors capture semantic relationships between words and are used to input meaningful representations of words into the caption generation model, contributing to the generation of coherent and contextually relevant captions. We employ pre-trained word embeddings to optimize the utilization of syntactic and semantic word associations. Word2vec [22] and GloVe [23] represent popular pre-trained word embedding matrices frequently used for this purpose.
4. Datasets

There are several widely used datasets for image captioning that researchers and practitioners commonly work with. Here are some prominent ones:

- **COCO (Common Objects in Context):** One of the most popular datasets for image captioning, containing a vast collection of images with multiple human-generated captions per image. It’s widely used for training and evaluating image captioning models [24].

- **Flickr30k:** A dataset comprising 31,000 images collected from Flickr, each with five reference captions provided by human annotators. It’s widely used for training and evaluating image captioning models [25].

- **Flickr8k:** A dataset containing 8,000 images sourced from Flickr, each paired with multiple human-written descriptions. It’s commonly used for training and testing image captioning models [26].

- **VisualData (VizWiz):** A dataset of images taken by blind individuals, with accompanying natural language descriptions and questions. It’s used to explore image captioning for visually impaired individuals [27].

- **SBU Captioned Photographic Dataset:** A dataset with over 120,000 images collected from Flickr, each paired with at least one human-generated caption. It’s useful for training and evaluating image captioning models [28].

- **Conceptual Captions:** A dataset with over 3.3 million images and associated captions collected from the web. It’s a large-scale dataset useful for training and testing image captioning models [29].

- **VQA (Visual Question Answering):** While primarily for question answering, VQA datasets also contain images with associated captions that can be utilized for image captioning tasks [30].

- **MSCOCO-Entities:** A variant of COCO dataset with additional entity annotations, providing a richer context for generating detailed and informative captions [31].

- **ImageCLEF:** An annual competition that often includes image captioning tasks and provides specific datasets for evaluation [32].

- **Multi30k:** A dataset for multilingual multimodal machine translation, also used for image captioning. It consists of images paired with English and German captions [33].

These datasets provide a diverse range of images and associated captions, making them valuable resources for training and evaluating image captioning models across different domains and applications. Make sure to review the terms of use and licenses associated with each dataset before usage.

5. Evaluation Metrics

- **BLEU (Bilingual Evaluation Understudy): BLEU-1 to BLEU-4:** BLEU is a standard metric measuring the n-gram overlap between the generated caption and reference captions. BLEU-1 to BLEU-4 assess unigrams to fourgrams overlap, providing a quantitative measure of caption quality [34].

- **METEOR (Metric for Evaluation of Translation with Explicit ORdering):** METEOR is a
comprehensive metric considering unigrams, stems, synonyms, and various linguistic elements to evaluate the generated caption against reference captions. It offers a more holistic evaluation of linguistic quality [35].

- **ROUGE-L (Recall-Oriented Understudy for Gisting Evaluation - Longest Common Subsequence)**: ROUGE-L measures the overlap of longest common subsequences, assessing recall of important phrases in the generated caption compared to reference captions. It is widely used in natural language processing evaluations [36].

- **CIDEr (Consensus-based Image Description Evaluation)**: CIDEr emphasizes consensus by evaluating the similarity of the generated caption to multiple human-written reference captions. It considers consensus phrases and provides insights into agreement with human references [37].

- **The SPICE (Semantic Propositional Image Caption Evaluation)**: metric is designed to evaluate the quality of image captions by assessing the semantic similarity between generated captions and human-written reference captions [38]. Unlike other metrics like BLEU or METEOR that rely on n-grams or word-level similarities, SPICE focuses on the semantic content and structure of the captions. SPICE evaluates the quality of a generated caption by analyzing the semantic propositions it contains and comparing them to the propositions in the reference captions. A "semantic proposition" represents a subject-predicate-object triplet, capturing the essential meaning of a sentence. By considering semantic propositions, SPICE offers a more fine-grained evaluation of the generated captions’ content, encouraging models to produce captions that accurately reflect the underlying scene or image.

These evaluation metrics are fundamental in assessing the quality of image captions generated by models. Researchers often use a combination of these metrics to comprehensively evaluate captioning models, considering different aspects of caption quality. Choosing appropriate metrics based on the specific goals of the image captioning task is crucial for accurate evaluation.

6. Implementation Details

In this section, we present the outcomes of our proposed approach using the MSCOCO [24] dataset, while employing standard evaluation metrics like BLEU [34], ROUGE [36], and CIDEr [37] for assessing image captioning performance. The fc7 feature map is used to calculate features. DenseNet’s feature map has a dimension of . In the prediction module, the hidden layer size is 1024. Text characteristics are extracted using BLSTM and CNN. Filter window size and amount of features maps from both BLSTM and CNN structures significantly impact model performance. Select the ideal filter window size of 3. The model’s accuracy can be low with few feature maps. As the size rises, performance may not improve and may potentially worsen due to over-fitting. The model becomes more sophisticated as the number of feature maps increases rapidly. We set the number of feature maps to 200. We embed 512-dimensional words using dropout, learning rate of 0.001, and a linear layer. We train the model with Adam optimizer and mini-batch size 16. After upsampling the word embedding vector
using ReLU activation on the fully connected layer, we use a softmax to calculate the output word probabilities. Our technique was trained for 20 epochs, evaluating metrics on the validation dataset to choose the best model.

6.1. Performance Comparison Based on MSCOCO

We compare our method to various leading image captioning methods. Figure 1 presents the results obtained on the MSCOCO dataset, showcasing various methods that employ distinct CNN encoders for image representation. Specifically, NIC [15] utilize GoogleNet for extracting image features. On the other hand, models employing Soft/Hard attention [39] utilize VGGNet to obtain image-level representation. In contrast, Regarding the BLEU-1 score, which considers bigrams exclusively, NIC [15], COMIC [40], and AICRL [41, 42] achieved scores of 73.1, 56.2, and 41.0, respectively. Overall, the results across the 06 evaluation metrics indicate that our proposed method achieve comparable performances with the state-of-the-art methods. In particular, our method can achieve 73.0, 58.1, 45.1, 35.1, 55.0, and 108.2 in BLEU-1, 2, 3, 4, ROUGH-L, and CIDEr-D respectively, making the superior performance over the methods. The samples of image captions generated by the proposed method is shown in Figure 2.

![Figure 1](image.png)

**Figure 1.** Performance of our method on MSCOCO dataset and popularly used evaluation metrics. Bold indicates the best result and a dash (-) indicates results are unavailable

7. CONCLUSION

The fusion of Bidirectional Long Short-Term Memory (BLSTM) and Convolutional Neural Networks (CNN) in image captioning has demonstrated a remarkable synergy, leading to substantial advancements in generating high-quality captions for images. The CNN, functioning as the encoder, adeptly extracts essential visual features from the input image, while the BLSTM, acting as the decoder, and processes these features bi-directionally, incorporating both past and future context. This synergistic collaboration empowers the model to generate descriptive and contextually coherent captions that are
more meaningful and relevant. The comprehensive experimentation and application of this combined BLSTM and CNN approach on diverse datasets, including the widely used MSCOCO dataset, illustrated its effectiveness. Evaluation metrics such as BLEU, ROUGE, and CIDEr affirmed the improved performance and accuracy achieved in image captioning. By fine-tuning parameters through training on annotated datasets, this approach has proven to be a robust solution, holding great promise in various applications requiring image understanding and contextual captioning. In conclusion, the integration of BLSTM and CNN in image captioning not only enriches the depth of image interpretation but also paves the way for enhanced AI capabilities in describing visual content accurately and informatively. As research in this field continues to evolve, further refinements and innovations in this fusion approach are expected, driving advancements in image captioning towards even more proficient and human-like descriptive capabilities.

References


