Cureus Journal of Computer Science

Open Access Review Article

A Review of Image Captioning Techniques: Types, Deep Learning Advancements, and Limitations

Chaitanya S. Bhosale [1] , Pradip Salve [1] , Vishal Shirsath [1]

1. School of Doctoral Studies, Ajeenkya DY Patil University, Pune, IND

Corresponding authors: Chaitanya S. Bhosale, chaitanya.bhosale@adypu.edu.in, Pradip Salve, pradip.salve@adypu.edu.in, Vishal Shirsath, vishal.shirsath@adypu.edu.in

Review began 02/03/2025
Review ended 07/05/2025
Published 07/14/2025

DOI: https://doi.org/10.7759/s44389-024-01573-w

Abstract

Image captioning involves generating text that describes visual content in images. It integrates computer vision and natural language processing domains. The presented study explores recent advancements and ongoing challenges in the image captioning domain. We have explored various image captioning methodologies, including retrieval-based, template-based, and deep learning-based approaches, and highlighted their respective strengths and limitations. We have also studied the critical role of object and action recognition and mapping relationships in images and how they help create clear and meaningful captions. Datasets that provide the foundation for training and evaluating models are also explored, such as Microsoft Common Objects in Context, Flickr8k, Flickr30k, and Visual Genome. Deep learning models such as recurrent neural networks and transformer-based architectures enhance the performance of captioning tasks. Despite these advancements, there are significant challenges that need to be addressed, such as handling complex scenes in images and generating context-aware captions. This study provides a detailed overview of recent research done in image captioning and offers potential directions for future work, focusing on the accuracy and scalability of the system for real-world applications such as assistive technology for the visually impaired, content monitoring on social media, and e-learning.

Categories: Deep Learning, Natural Language Processing (NLP), Computer Vision
Keywords: image captioning, computer vision, convolutional neural networks (cnns), natural language processing (nlp), long short-term memory (lstm) networks

Introduction And Background

Image captioning showcases how machines can understand visual content and express it in human language, closely imitating human cognitive abilities. The uses of image captioning are extensive, including assistive technology for the visually impaired, automated content creation, improved search engines, and smart control systems in the Internet of Things (IoT) [1-4]. Figure *1* shows the basic diagram of image captioning. An image is passed to an encoder, which extracts visual features from the image. The encoder can be a convolutional neural network (CNN) or a visual transformer. The decoder generates descriptive text from the encoded features using recurrent neural network (RNN), long short-term memory (LSTM), or any other deep learning model designed to handle sequential data. To generate accurate descriptive text, the model needs to be trained on image captioning datasets like Microsoft Common Objects in Context (MS COCO), Flickr8k, and Flickr30k, which contain paired images and captions. Supervised learning is used for training, which optimizes a suitable loss function such as Consensus-based Image Description Evaluation (CIDEr)-based loss.

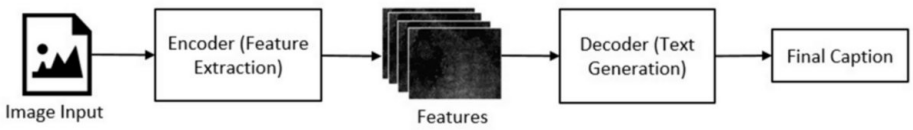

FIGURE 1: Basic image captioning diagram

The main objective of image captioning is to generate captions that are not only grammatically accurate and meaningful, but also contextually fitting and descriptive. This requires an intricate process of identifying objects, actions, and their interrelationships within an image, followed by the generation of a cohesive narrative [5-8]. The challenge lies in the machine's ability to comprehend complex visual scenes and express this understanding in natural language, a task that demands high-level cognitive functions traditionally associated with human intelligence [9-13].

Recent improvements in DL have pushed the field of image captioning forward. DL models such as CNNs have demonstrated exceptional proficiency in object detection and feature extraction, while RNNs,

particularly LSTM networks, have excelled in sequence generation tasks [14-19]. The combination of these techniques has greatly improved the quality and accuracy of image captions. Moreover, innovative architectures like attention mechanisms and transformers have further refined the ability of models to give attention to relevant parts of an image during the captioning process [20-22].

Despite aforementioned advancements, the domain of image captioning continues to face numerous challenges. One of the primary issues is the handling of diverse and complex scenes, where the presence of multiple objects and intricate interactions can complicate the generation of accurate captions [23]. Additionally, ensuring linguistic coherence and avoiding common pitfalls such as repetitive or overly generic descriptions remain significant hurdles [24]. The contextual understanding required for generating truly descriptive and relevant captions is another area that necessitates further research. Here, paper survey is divided into three types of image captioning methods: retrieval-based, template-based, and DL-based methods. Retrieval-based methods focus on finding the most relevant pre-existing captions for a given image from a database. Template-based methods rely on predefined sentence structures to generate captions. In contrast, DL-based approaches utilize neural networks to learn from image features and directly generate captions.

Types of image captioning

Retrieval-Based Image Captioning

Retrieval-based image captioning is a technique that generates captions for images by selecting and adapting pre-existing captions from a database, rather than creating new ones from scratch. This approach leverages the idea that many images share similar features and contexts, allowing the reuse of captions for visually and contextually similar images [23]. Figure *2* shows the process of retrieval-based image captioning, where the caption of most similar image is selected for the input image. Here, the similarity score of image (d) is higher than that of the rest of the images in the dataset, so the caption of image (d) gets assigned to input image.

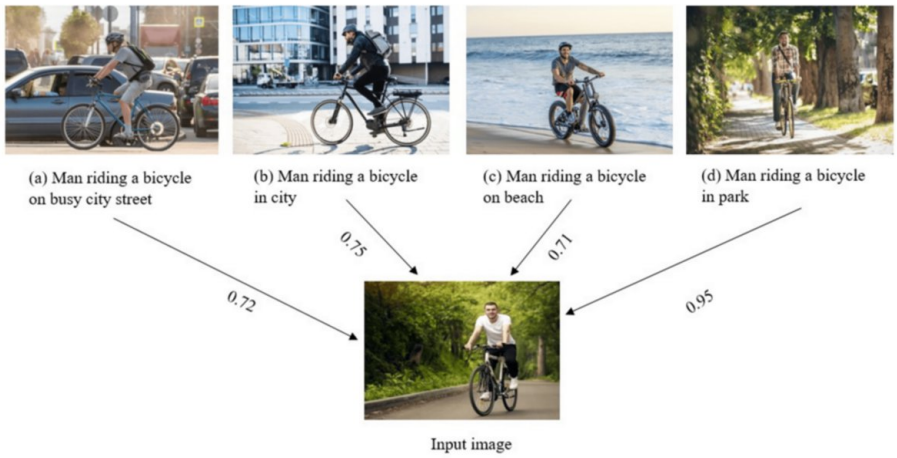

FIGURE 2: Retrieval-based image captioning

The process and components involved in retrieval-based image captioning are listed as follows:

• Image Feature Extraction: Extracting features from the input image is the first step in retrieval-based image captioning. This is typically done using CNNs, which are highly effective in capturing the essential visual characteristics of an image. Popular CNN architectures like Visual Geometry Group (VGG), ResNet, or Inception are often used to transform the input image into a high-dimensional feature vector that encapsulates its visual content. A CNN contains three layers through which the image is processed [22]:

- Convolutional Layers: These layers analyze the input image to detect local patterns such as edges, textures, and fundamental shapes.

- Pooling Layers: In these layers, image data size is reduced, making it easier and faster to process.

- Fully Connected Layers: These layers take the processed image data and turn it into a fixed-size feature vector that summarizes the important information about the image. The final output of the CNN is a feature vector that provides a detailed representation of what is in the image.

• Feature Database: A database of images paired with their corresponding captions is prepared in advance. For each image in this database, features are extracted and stored alongside the captions. This database forms the foundation from which captions are retrieved.

• Similarity Matching: The feature vector of the submitted image is compared to the feature vectors of images in the database to find the most similar ones. Similarity metrics such as Euclidean distance, cosine similarity, or more advanced techniques like k-nearest neighbors (k-NN) are employed to measure the closeness between the feature vectors.

• Caption Retrieval: On the basis of similarity scores, the most similar images from the database are identified. The captions associated with these similar images are retrieved. This step often involves selecting the top-k most similar images and considering their captions as potential candidates.

• Caption Selection and Refinement: The final caption for the input image can be selected directly from the retrieved candidates or through further refinement. Several strategies can be employed here:

- Direct Selection: Selecting the caption of the image with highest similarity.

- Caption Re-ranking: Using additional criteria or models to rank the retrieved captions and select the best one.

- Combination and Refinement: Merging elements from multiple retrieved captions to generate a more accurate or descriptive caption.

Template-Based Image Captioning

With this method, captions are generated by filling predefined sentence structures, or templates, with relevant content derived from an image. This approach combines the predictability and grammatical correctness of fixed sentence patterns with the flexibility to describe various aspects of an image. Components in template-based captioning are listed as follows:

• Image Feature Extraction: As with other image captioning techniques, the process begins with extracting features from the input image. CNNs are typically used to generate a feature vector that helps to detect objects, actions, and attributes.

• Object Detection and Attribute Recognition: Once the image features are extracted, techniques such as region-based CNNs (R-CNN) or YOLO (You Only Look Once) are used to identity objects present in the image. These techniques are also used to identity attributes and actions depicted in the image. Attributes such as colors, shapes, and sizes, as well as actions like running or eating, are also identified during this step.

• Template Selection: Predefined templates form the backbone of this approach. These templates are sentence structures with placeholders for different types of content. For instance, as shown in Figure *3*, a template might look like:

"A [object] is [action] in the [location]."

"The [object] has [attribute]."

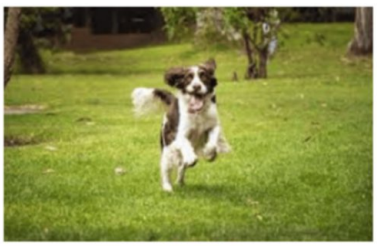

FIGURE 3: Template-based image captioning

Templates are designed to cover a variety of common scenes and descriptions. They ensure grammatical correctness and provide a structured way to generate captions.

• Content Insertion: With the detected objects, attributes, and actions in hand, the next step is to fill the placeholders in the selected templates. This involves:

- Object Insertion: Placing detected objects into the relevant slots in the template.

- Attribute Insertion: Adding recognized attributes to describe the objects.

- Action Insertion: Including any identified actions associated with the objects.

For example, if the system detects a dog running in a park, a suitable template might be:

"A [dog] is [running] in the [park]."

• Template Matching and Ranking: If multiple templates are applicable, the system may use additional criteria to select the most suitable one. This can involve matching the detected objects and actions to templates that best fit the scene or using scoring mechanisms to rank templates based on relevance and completeness.

• Final Caption Generation: The final step is generating the complete caption by filling in all placeholders with the appropriate content, resulting in a syntactically and semantically coherent sentence. This caption is then presented as the description for the image.

DNN-Based Image Captioning

This methodology uses advanced DL techniques to generate descriptive captions for images. This approach typically involves two main components: a CNN for extracting features from images and a RNN, often enhanced with LSTM units, for generating the textual descriptions [17,18]. Process of DNN-based image captioning is described as follows:

• Image Feature Extraction: The process begins with extracting high-level features from image using a CNN. CNNs are effective at capturing spatial hierarchies in images. Popular CNN architectures used for this purpose include VGG, ResNet, and Inception [19,21].

• Sequence Generation With RNN: In order to generate image captions, the feature vector extracted from image is fed into a RNN. RNNs are well-suited for sequence generation tasks because they can maintain and update a hidden state that captures the context of previously generated words. In image captioning, LSTM units or gated recurrent units (GRUs) are often preferred because they can handle long-term dependencies more effectively than vanilla RNNs [20].

• Initialization: The feature vector from the CNN initializes the hidden state of the RNN.

• Word Embeddings: RNN takes a word embedding as input at every time step, which represents the previous word in the sequence.

• Hidden State Update: RNN updates its hidden state by considering both the present input and the preceding state.

• Word Prediction: Updated hidden state is used in order to predict the next word in the sequence, typically using a softmax layer to output a probability distribution over the vocabulary.

• Attention Mechanism: The attention mechanism helps enhance the quality of captions by enabling the neural network to focus on various sections of the image while producing each word in the caption, rather than relying on a single fixed feature vector.

• Attention Weights: At each time step, the attention mechanism calculates a set of weights, which determine the significance of various regions of the image.

• Context Vector: A weighted sum of the image features is computed, resulting in a context vector that highlights the relevant parts of image for the current word.

• Combined Input: To generate next word, context vector is merged with the current hidden state and the previous word embedding to generate the next word.

• Training Process: In training process, neural network model is trained using a dataset of images paired with corresponding captions. The training process aims to reduce a loss function, usually the cross-

entropy loss, which calculates the difference between the predicted words as well as actual words in the training captions.

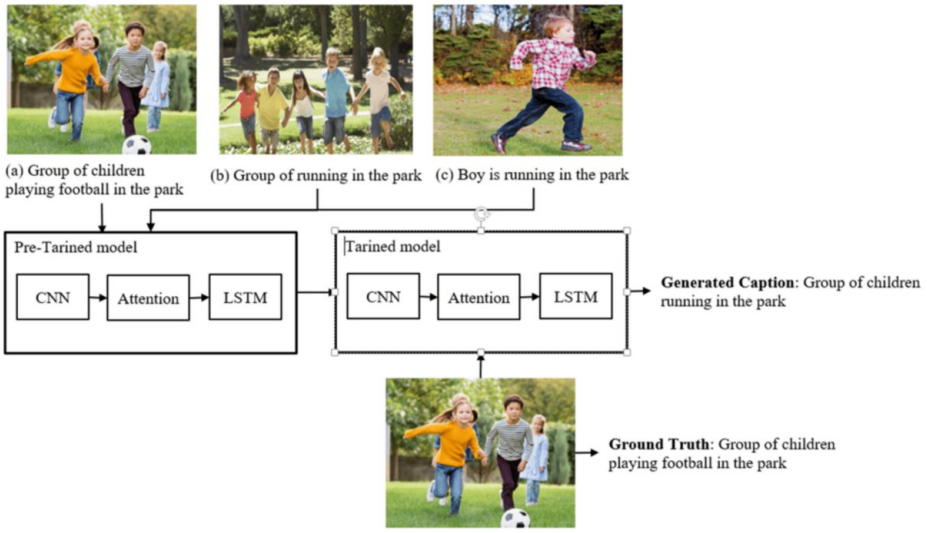

FIGURE 4: DNN-based image captioning

DNN, Deep Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory

As shown in Figure *4*, a model with CNN, attention mechanism, and LSTM is trained on images from the dataset. The trained model is then given the task of captioning similar images it was trained on, and it generates captions that are very similar to the ground truth.

A review was conducted across prominent academic papers published in peer-reviewed journals or conferences such as IEEE Xplore, IEEE Access, Springer, and other Scopus-indexed journals. The studied papers were included based on specific criteria: the paper focuses on methodologies or advancements in image captioning, uses image captioning datasets such as MS COCO, Flickr8k, or Flickr30k, and involves natural language processing and computer vision (CV) domains with various DL models such as CNNs, RNNs, LSTMs, or transformers. From the collected papers, we screened the titles and abstracts of initially selected papers to determine their relevance, followed by full reviews. Papers were selected based on their relevance, methodology, and contribution.

Review

Existing work

DL-based image captioning has improved greatly in recent years, helping to create various applications like improving feature extraction and enhancing models captioning accuracy, which can assist visually impaired individuals. In this paper, we have explored diverse methodologies for image captioning that employ DL architecture, attention mechanisms, and other traditional methodologies.

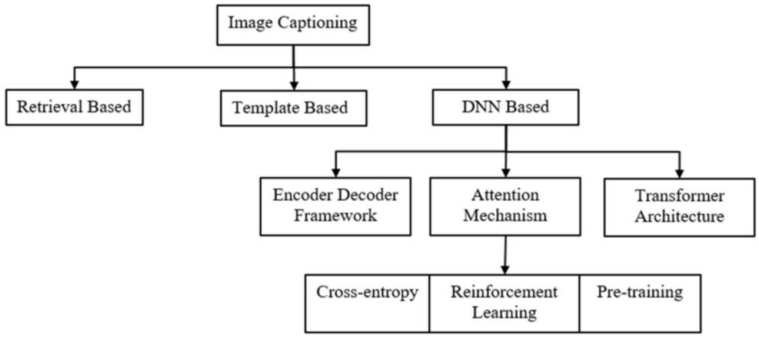

FIGURE 5: Taxonomy of evolution of image captioning

DNN, Deep Neural Network

As shown in Figure *5*, at first, retrieval-based image captioning technique came into existence, which does not generate new captions but instead retrieves the most similar caption from a database based on visual similarity. This technique is simple and fast but is limited to existing data in the database and cannot generalize. Template-based image captioning generates captions from pre-defined sentence templates that are then filled with detected objects, attributes, and actions of these objects. Although the captions are grammatically correct, this technique contains limited expressiveness and adaptability. DNN-based methods are the most advanced, containing an encoder-decoder framework where CNN is used as the encoder and sequential DL models such as RNN and LSTM are used as the decoder. The attention mechanism is integrated to enhance the encoder-decoder framework by allowing the model to focus on specific regions of an image and handle complex scenes. The attention mechanism is paired with cross-entropy, reinforcement learning, and pre-training. Cross-entropy is a standard loss function used in training the model to predict the sequence. Reinforcement learning is used to fine-tune the model using sequence-level rewards. The model is pre-trained using datasets like CLIP and BLIP to enhance performance. There are two aspects of image captioning: multimodal and multilingual. Multimodal image captioning integrates multiple sources of information like images, audio, and video to generate more accurate captions. A multimodal captioning system uses both visual features and audio. For example, a captioning system might use both visual and audio features to describe a scene, such as "A lion roaring at the spectator in the zoo." On the other hand, multilingual image captioning focuses on generating captions in multiple languages, allowing the same visual content to be described in multiple languages like Hindi, English, and Marathi.

Ganesan et al. [20] proposed a DL approach that converts textual descriptions of images embedded in printed texts into voice-based outputs, assisting visually impaired individuals. CNN is used for feature extraction, and LSTM is used for caption generation as well as speech generation. The model is further explored on different architectures like GoogleNet, AlexNet, ResNet, SqueezeNet, and VGG16, with the ResNet architecture achieving the highest accuracy of 83%. Self-Enhanced Attention was introduced by Sun et al. [21] to replace the traditional self-attention mechanism in Transformer decoders. This mechanism adjusts the attention weight matrix to enhance feature representation. Model evaluation was done on the MS COCO dataset, achieving a CIDEr score of 126.8. In a study by Yang et al. [22], the EnsCaption model was introduced to address challenges related to generating informative negative samples. This model integrates generation-based and retrieval-based approaches, employing a dual-generator generative adversarial network. The generation model generates various captions, and the retrieval model selects the best caption candidates. Evaluation was done on the MS COCO and Flickr30k datasets. An attention-based encoder-decoder was proposed by Al-Malla et al. [23], which integrates object detection features from YOLOv4 and convolutional features extracted from Xception. A positional encoding scheme called the "importance factor" was proposed in this study, focusing on large foreground objects over smaller background elements. The model was evaluated on the MS COCO and Flickr30k datasets, showing a CIDEr score improvement of 15.04%. To estimate important image regions, a cyclical training-based localizer was introduced by Suzuki et al. [25], which focuses on critical regions and identifies subject and object words in generated captions. To evaluate the model, the authors created a dataset that defines important regions within images. Comparative analysis shows that the proposed method calculates accurate region estimations better than other saliency-based approaches. Fine-grained image captioning with Attention-Guided Hierarchical Parsing was proposed by Gu and Jin [26]. It uses the segment anything model for segmentation and focuses on important visual elements for feature extraction. The model is integrated with a multi-level encoding-decoding approach and performs hierarchical feature processing through serial decoders. The model was evaluated on the RSTPReid and CUHK-PEDES datasets. To encode image content through an attention mechanism, a cascade semantic fusion architecture was introduced by Wang et al. [27]. CSF integrates three types of visual attention: object-level, image-level, and spatial-level attention. Using a pretrained detector, object-level attention is

first extracted, then object-level attention features are merged with spatial features to induce image-level attention. The model was evaluated on the MS COCO dataset. In a study by Bae et al. [28], a bidirectional LSTM (Bi-LSTM) model is used for caption generation, which is incorporated with a part-of-speech guidance module to enhance lexical diversity, thereby generating diverse and grammatically rich captions. The model is evaluated on the Flickr30k and MS COCO datasets. Similarly, a Bi-LSTM was used by Chahkandi et al. [29] to generate descriptions in a multimodal space. The proposed model integrates three main modules: the first is an object detection network, the second is a Bi-LSTM-based decoder for sentence generation, and the third is a semantic space for scoring images and captions. The Flickr8k and Flickr30k datasets are used for model evaluation. A multilayer attention mechanism for video captioning was introduced by Naik and Jaidhar [30], which consists of a visual encoder and a language model based on LSTM networks. Video frames are processed by a stacked Bi-LSTM encoder while a decoder unit generates sentences. The model is evaluated on the MSVD and MSR-VTT datasets. Similarly, an attention mechanism was proposed by Hafeth et al. [31], where a transformer model was used to enhance image caption generation. An external knowledge base, ConceptNet, was used to improve the ability to capture factual knowledge. The model is evaluated on the MS COCO dataset, and experiments prove that incorporating external knowledge enhances the model's reasoning capabilities. The AIC-SSAIDL model was introduced by Arasi et al. [32], which integrates MobileNetv2 for feature extraction and an attention mechanism with LSTM for caption generation. The Sparrow Search Algorithm is used for hyperparameter tuning and optimized using the Fruit Fly Optimization algorithm. Model evaluation is done on the Flickr8k, Flickr30k, and MS COCO datasets. To enhance traditional self-attention, a multi-gate attention layer was introduced by Jiang et al. [33], which uses an additional attention weight gate and a self-gated module to enhance traditional self-attention. Attention weight gate focuses attention on the most relevant objects, and self-gated module eliminates irrelevant information. To refine image features efficiently, a pre-LayerNorm transformer is used. For model evaluation, the MS COCO dataset is used. Wang et al. [34] employed Faster R-CNN for feature extraction, and an LSTM-based decoder with a multilayer dense attention mechanism for caption generation. To optimize model parameters, a gradient optimization technique was used for reinforcement learning. The model is evaluated on the Flickr8k, Flickr30k, MS COCO, and AI Challenge Chinese datasets. Similarly, a hierarchical attention fusion model was introduced by Wu et al. [35], which is based on reinforcement learning. The model employs ResNet multi-level feature maps with hierarchical attention and rewrites the CIDEr score at the word level using a revaluation network. Sentence-level rewards are assigned based on caption quality using a scoring network. The MS COCO dataset is used for model evaluation. Scene graph generation was used by Phueaksri et al. [36] to generate summarized captions from image collections. Scene graphs were employed to extract common objects and relationships across multiple images, which were then passed through a captioning model. To generalize specific words into broader concepts, external knowledge was incorporated. The method was evaluated on the MS COCO dataset. Jing et al. [37] developed a model to generate human-like captions for news images by incorporating real-world knowledge. The proposed method extracts named entities from news articles and extends them using external knowledge. To extract contextual information selectively, a sentence correlation analysis algorithm was proposed, and an entity-linking algorithm was used to establish relationships among named entities. The model was evaluated on a real-world news dataset [37]. A hybrid attention distribution method was proposed by Wang and Feng [38], which addresses the limitations of single attention distribution matrices in self-attention mechanisms. A hybrid attention distribution was introduced to capture intrinsic relationships between query and key vectors. Word embeddings were factorized to reduce the number of model parameters, which enhanced training efficiency. MS COCO dataset is used for model evaluation. Stack-VS, a multi-stage architecture, was proposed by Cheng et al. [39], which integrates top-down and bottom-up attention mechanisms. Stack-VS uses a stack decoder, which consists of multiple LSTM layers to re-optimize attention weights iteratively. Extensive experiments are performed on MS COCO dataset. AICRL, an automatic image captioning model, was introduced by Chu et al. [40], which consists of ResNet50 encoder and an LSTM decoder with a soft attention mechanism. The model focuses on relevant image regions while predicting captions. The model was evaluated on MS COCO dataset. Bi-LSTM is leveraged by Wang et al. [41] to learn learning long-term visual-language interactions. A CNN was used for feature extraction, and a Bi-LSTM network was employed to capture both past and future contextual information. To avoid overfitting, various data augmentation techniques such as multi-crop, multi-scale, and vertical mirroring were used. The model was evaluated on Flickr8k, Flickr30k, and MS COCO datasets. The Topic-Guided Image Captioning Framework was introduced by Yu et al. [42], which generates captions conditioned on predefined topics. A CNN-based multi-label classifier is used to extract topic candidates from a caption corpus. A cross-modal embedding space is trained for images, topics, and captions, ensuring hierarchical organization. The model was evaluated on the MS COCO and Flickr30k datasets. Wu et al. [43] addressed the issue of overly generic captions. To tackle this issue, a new training objective was proposed, incorporating a global discriminative constraint to ensure that the generated captions distinguish images from others in the dataset, and a local discriminative constraint to focus on less frequent but more informative words. The MS COCO dataset was used for model evaluation. Semantic understanding in image captioning was enhanced by Young et al. [44] using Bahdanau Attention with pretrained CNN models. In this model, two pretrained CNN encoders were used, namely Vector Geometry Group and InceptionV3. These encoders provide rich semantic context for the RNN decoder. For model evaluation, the Flickr8k dataset was used, achieving a BLEU score of 62.

Ref. No.	Technical Methods	Datasets Used	Metric	Limitations
[1]	EfficientNet-B3 for feature extraction, RNN for sequential word generation	MS COCO	-	Requires extensive computational resources
[2]	CNN for feature extraction, RNN for descriptive captions, attention processing, beam search	MS COCO, Flickr30k	On MS COCO – Accuracy: 95.26; On Flickr30k – Accuracy: 97.26	Challenges in linguistic coherence and contextually relevant captions
[3]	ResNetRS101 with Bahdanau attention mechanism	MS COCO	BLUE-4: 0.55	Difficulty in capturing complex scene semantics
[4]	CNN for feature extraction, RNN for sequence generation	MS COCO 2014, Flick8k	METEOR: 34; CIDEr: 201	Overemphasis on n-gram and pattern matching in evaluation metrics
[5]	Motion-CNN with object detection	MSR-VTT2016-Image, MS COCO	On MSR-VTT2016 – METEOR: 16.1, CIDEr: 32.7; On MS COCO – METEOR: 26.7, CIDEr: 109.9	Unnecessary motion features decreasing accuracy
[6]	Pre-trained CNN for feature extraction, RNN for language model	Flickr8k	-	Generating concise, fluent, and diverse captions remains challenging
[7]	Areas of attention model, CNN activation grids, object proposals, spatial transformers	MS COCO	METEOR: 23.7; CIDEr: 87.4	Weakly supervised learning needs improvement
[8]	Semantic-driven CNN-LSTM, pre-trained CNN, semantic keywords extraction, facial recognition	Faces dataset	Accuracy: 80.0	Semantic extraction needs further enhancement
[9]	CNN for feature extraction, bidirectional LSTM for sentence generation	Flickr8k, Faces dataset	On Flickr8k – BLUE-4: 0.23, METEOR: 18.8; On Faces – BLUE-4: 0.43, METEOR: 29.4	Needs exploration of more sophisticated language representations
[10]	Convolutional networks for language generation	MS COCO	BLUE-4: 29.4	Comparatively complex and sequential nature of LSTM units

TABLE 1: Comparative analysis of recent image captioning methods

RNN, Recurrent Neural Network; CNN, Convolutional Neural Network; LSTM, Long Short-Term Memory; MS COCO, Microsoft Common Objects in Context

Table *1* presents recent image captioning methodologies, the datasets they use, and their limitations. Most of these methods face key challenges such as high computational costs, generating contextually accurate captions, and weak semantic understanding.

In our research gap analysis, we found that while many DL-based image captioning models have been studied, comparative analysis on diverse datasets is still lacking. This includes domain-specific images, multilingual caption generation, and handling complex scenes. Real-world deployment of image captioning models remains a challenge, especially for edge devices. Ethical issues such as caption bias, fairness, and the ability to elaborate on complex scenes still need to be resolved. Future research should focus on designing lightweight image captioning frameworks optimized for edge devices. Combining retrieval-based image captioning with the flexibility of DL to generate captions can address the challenges of implementing image captioning on edge devices. Extending captioning to work across multiple languages and input modalities like audio and image is an exciting area of research.

Datasets

As presented in Table *2*, there are various benchmark datasets available for CV and natural language processing tasks, specifically for image captioning and object recognition. The MS COCO dataset [45] contains over 3,30,000 images with five human-annotated captions per image. On the other hand, the ImageNet dataset [46] is used for object classification and recognition, containing over 1,281,167 images labeled across 20,000 categories. The Flickr [47] datasets are widely used for image captioning tasks. The Flickr30k dataset contains 30,000 images with five descriptive captions for each image. The Pascal VOC (Visual Object Classes) dataset [48] is designed for object detection, segmentation, and classification,

containing over 11,530 images with 20 object categories.

Dataset	Use Cases	No. of Images	No. of Categories/Classes	No. of Captions per Image
MS COCO	Image Captioning, Object Detection and Segmentation	330,000+	80 object categories	5 captions
ImageNet	Object Classification and Recognition	1,281,167	1000	N/A
Flickr30k	Image Captioning and Retrieval	31,000+	N/A	5 captions
Pascal VOC	Object Detection, Segmentation and Classification	11,530	20	N/A

TABLE 2: Comparison of image captioning and object recognition datasets

MS COCO, Microsoft Common Objects in Context; VOC, Visual Object Classes

Conclusions

The integration of DL architectures such as CNNs for feature extraction and RNNs or LSTM networks for sequence generation has substantially enhanced the quality and precision of generated captions. Despite these advancements, there are several challenges in caption generation, such as handling complex scenes, ensuring linguistic coherence, grammatical correctness, and capturing the full spectrum of image content. Additionally, advanced DL architectures such as transformers have shown significant potential to further improve captioning accuracy and overall performance. This paper explored various DL and traditional image captioning methodologies along with training datasets such as MS COCO and Flickr, highlighting the need for diverse and unbiased data. Using multimodal data, such as incorporating audio and temporal information, can help create richer and more comprehensive captions. Datasets created with real-world scenarios and fewer biases can improve captioning capabilities. One important aspect of image captioning is time. In real-time systems like surveillance monitoring, assistance for the visually impaired, and autonomous self-driving, generating image captions automatically in milliseconds significantly improves operational efficiency. Retrieval-based methods generate captions quickly but lack adaptability. Template-based methods are time-efficient but have limited versatility. While DNN or transformer models generate semantically rich captions, they are computationally heavy. Despite recent advancements, optimizing models for accuracy with response time is a tedious task, especially for edge devices. Future work should focus on designing lightweight models that generate descriptive and accurate captions while meeting time constraints.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Chaitanya S. Bhosale

Acquisition, analysis, or interpretation of data: Chaitanya S. Bhosale, Pradip Salve, Vishal Shirsath

Drafting of the manuscript: Chaitanya S. Bhosale

Critical review of the manuscript for important intellectual content: Chaitanya S. Bhosale, Pradip Salve, Vishal Shirsath

Supervision: Chaitanya S. Bhosale, Pradip Salve, Vishal Shirsath

Disclosures

Conflicts of interest: In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

Acknowledgements

Chaitanya S. Bhosale would like to express his deepest gratitude to his supervisor, Dr. Pradip Salve, for making this work achievable. Dr. Salve's support and guidance were invaluable throughout every phase of the paper's development. Chaitanya is also profoundly grateful to his co-supervisor, Dr. Vishal Shirsath, and to the PhD Coordinator at Ajeenkya DY Patil University.

References

1. Kavitha R, Shree Sandhya S, Betes P, Rajalakshmi P, Sarubala E: Deep learning-based image captioning for visually impaired people. E3S Web of Conferences. 2023, 399:04005. 10.1051/e3sconf/202339904005
2. Dongare Y, Hardas BM, Srinivasan R, Meshram V, Aush MG, Kulkarni A: Deep neural networks for automated image captioning to improve accessibility for visually impaired users. International Journal of Intelligent Systems and Applications in Engineering. 2023, 12:267-281.
3. Raut R, Patil S, Borkar P, Zore P: Image captioning using ResNet RS and attention mechanism. International Journal of Intelligent Systems and Applications in Engineering. 2023, 11:606-613.
4. Rao BS, Meenakshi K, Kalaiarasi K, Babu PR, Kavitha J, Saravanan V: Image caption generation using recurrent convolutional neural network. International Journal of Intelligent Systems and Applications in Engineering. 2023, 12:76-80.
5. Iwamura K, Louhi Kasahara JY, Moro A, Yamashita A, Asama H: Image captioning using Motion-CNN with object detection. Sensors. 2021, 21:1270. 10.3390/s21041270
6. Kharate N, Patil S, Shelke P, Shinde G, Mahalle P, Sable N, Chavhan PG: Unveiling the resilience of image captioning models and the influence of pre-trained models on deep learning performance. International Journal of Intelligent Systems and Applications in Engineering. 2023, 11:01-07.
7. Pedersoli M, Lucas T, Schmid C, Verbeek J: Areas of attention for image captioning. arXiv. 2017, 10.48550/arXiv.1612.01033
8. Tateno K, Takagi N, Sawai K, Masuta H, Motoyoshi T: Method for generating captions for clothing images to support visually impaired people. 2020 Joint 11th International Conference on Soft Computing and Intelligent Systems and 21st International Symposium on Advanced Intelligent Systems (SCIS-ISIS), Hachijo Island, Japan. 2020, 1-5. 10.1109/SCISISIS50064.2020.9322767
9. Abisha Anto Ignatious L, Jeevitha S, Madhurambigai M, Hemalatha M: A semantic driven CNN - LSTM architecture for personalised image caption generation. 2019 11th International Conference on Advanced Computing (ICoAC), Chennai, India. 2019, 356-362. 10.1109/ICoAC48765.2019.246867
10. Aneja J, Deshpande A, Schwing AG: Convolutional image captioning. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA. 2018, 5561-5570. 10.1109/cvpr.2018.00583
11. Alzahrani N, Al-Baity HH: Object recognition system for the visually impaired: a deep learning approach using Arabic annotation. Electronics. 2023, 12:541. 10.3390/electronics12030541
12. Tiwary T, Mahapatra RP: An accurate generation of image captions for blind people using extended convolutional atom neural network. Multimedia Tools and Applications. 2023, 82:3801-3830. 10.1007/s11042-022-13443-5
13. Ashiq F, Asif M, Ahmad MB, Zafar S, Masood K, Mahmood T: CNN-based object recognition and tracking system to assist visually impaired people. IEEE Access. 2022, 10:14819-14834. 10.1109/access.2022.3148036
14. Alva AS, Nayana R, Raza N, Sampatrao GS, Reddy KBS: Object detection and video analyser for the visually impaired. 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS), Coimbatore, India. 2023, 1405-1412. 10.1109/ICAIS56108.2023.10073662
15. Kumar DR, Thakkar HK, Merugu S, Gunjan VK, Gupta SK: Object detection system for visually impaired persons using smartphone. ICDSMLA 2020. 2022, 783:1631-1642. 10.1007/978-981-16-3690-5_154
16. Wang M, Song L, Yang X, Luo C: A parallel-fusion RNN-LSTM architecture for image caption generation. 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 2016. 2016, 4448-4452. 10.1109/icip.2016.7533201
17. Hilal AM, Alrowais F, Al-Wesabi FN, Marzouk R: Red deer optimization with artificial intelligence enabled image captioning system for visually impaired people. Computer Systems Science and Engineering. 2023, 46:1929-1945. 10.32604/csse.2023.035529
18. Makav B, Kılıç V: A new image captioning approach for visually impaired people. 2019 11th International Conference on Electrical and Electronics Engineering (ELECO), Bursa, Turkey. 2019, 945-949. 10.23919/ELECO47770.2019.8990630
19. Ahsan H, Bhalla N, Bhatt D, Shah K: Multi-modal image captioning for the visually impaired. arXiv. 2021, 10.48550/arXiv.2105.08106
20. Ganesan J, Azar AT, Alsenan S, Kamal NA, Qureshi B, Hassanien AE: Deep learning reader for visually impaired. Electronics. 2022, 11:3335. 10.3390/electronics11203335
21. Sun Q, Zhang J, Fang Z: Self-enhanced attention for image captioning. Neural Processing Letters. 2024, 56:131. 10.1007/s11063-024-11527-x
22. Yang M, Liu J, Shen Y, Zhao Z, Chen X, Wu Q, Li C: An ensemble of generation- and retrieval-based image captioning with dual generator generative adversarial network. IEEE Transactions on Image Processing. 2020, 29:9627-9640. 10.1109/tip.2020.3028651
23. Al-Malla MA, Jafar A, Ghneim N: Image captioning model using attention and object features to mimic human image understanding. Journal of Big Data. 2022, 9:20. 10.1186/s40537-022-00571-w
24. Chahal A, Belani M, Bansal A, Jadhav N, Balakrishnan M: Template based approach for augmenting image descriptions. ICCHP 2018. 2018, 10896:104-112. 10.1007/978-3-319-94277-3_19
25. Suzuki T, Sato D, Sugaya Y, Miyazaki T, Omachi S: Important region estimation using image captioning. IEEE Access. 2022, 10:105546-105555. 10.1109/access.2022.3211260
26. Gu Z, Jin J: Attention-guided hierarchical parsing for fine-grained person-centric image captioning. IEEE Access. 2024, 12:86293-86301. 10.1109/access.2024.3416207
27. Wang S, Lan L, Zhang X, Dong G, Luo Z: Cascade semantic fusion for image captioning. IEEE Access. 2019, 7:66680-66688. 10.1109/ACCESS.2019.2917979
28. Bae JW, Lee SH, Kim WY, Seong JH, Seo DH: Image captioning model using part-of-speech guidance module for description with diverse vocabulary. IEEE Access. 2022, 10:45219-45229. 10.1109/access.2022.3169781

29. Chahkandi V, Fadaeieslam MJ, Yaghmaee F: Improvement of image description using bidirectional LSTM. International Journal of Multimedia Information Retrieval. 2018, 7:147-155. 10.1007/s13735-018-0158-y
30. Naik D, Jaidhar CD: A novel Multi-Layer Attention Framework for visual description prediction using bidirectional LSTM. Journal of Big Data. 2022, 9:104. 10.1186/s40537-022-00664-6
31. Hafeth DA, Kollias S, Ghafoor M: Semantic representations with attention networks for boosting image captioning. IEEE Access. 2023, 11:40230-40239. 10.1109/access.2023.3268744
32. Arasi MA, Alshahrani HM, Alruwais N, Motwakel A, Ahmed NA, Mohamed A: Automated image captioning using Sparrow Search Algorithm with improved deep learning model. IEEE Access. 2023, 11:104633-104642. 10.1109/access.2023.3317276
33. Jiang W, Li X, Hu H, Lu Q, Liu B: Multi-gate attention network for image captioning. IEEE Access. 2021, 9:69700-69709. 10.1109/access.2021.3067607
34. Wang EK, Zhang X, Wang F, Wu TY, Chen CM: Multilayer dense attention model for image caption. IEEE Access. 2019, 7:66358-66368. 10.1109/access.2019.2917771
35. Wu C, Yuan S, Cao H, Wei Y, Wang L: Hierarchical attention-based fusion for image caption with multi-grained rewards. IEEE Access. 2020, 8:57943-57951. 10.1109/access.2020.2981513
36. Phueaksri I, Kastner MA, Kawanishi Y, Komamizu T, Ide I: An approach to generate a caption for an image collection using scene graph generation. IEEE Access. 2023, 11:128245-128260. 10.1109/access.2023.3332098
37. Jing Y, Zhiwei X, Guanglai G: Context-driven image caption with global semantic relations of the named entities. IEEE Access. 2020, 8:143584-143594. 10.1109/access.2020.3013321
38. Wang J, Feng J: Hybrid attention distribution and factorized embedding matrix in image captioning. IEEE Access. 2020, 8:154453-154460. 10.1109/access.2020.3018546
39. Cheng L, Wei W, Mao X, Liu Y, Miao C: Stack-VS: stacked visual-semantic attention for image caption generation. IEEE Access. 2020, 8:154953-154965. 10.1109/access.2020.3018752
40. Chu Y, Yue X, Yu L, Sergei M, Wang Z: Automatic image captioning based on ResNet50 and LSTM with soft attention. Wireless Communications and Mobile Computing. 2020, 2020:1-7. 10.1155/2020/8909458
41. Wang C, Yang H, Bartz C, Meinel C: Image captioning with deep bidirectional LSTMs. MM '16: Proceedings of the 24th ACM international conference on Multimedia. 2016, 988-997. 10.1145/2964284.2964299
42. Yu N, Hu X, Song B, Yang J, Zhang J: Topic-oriented image captioning based on order-embedding. IEEE Transactions on Image Processing. 2019, 28:2743-2754. 10.1109/tip.2018.2889922
43. Wu J, Chen T, Wu H, Yang Z, Luo G, Lin L: Fine-grained image captioning with global-local discriminative objective. IEEE Transactions on Multimedia. 2021, 23:2413-2427. 10.1109/tmm.2020.3011317
44. Young P, Lai A, Hodosh M, Hockenmaier J: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics. 2021, 2:67-78. 10.1162/tacl_a_00166
45. Lin TY, Maire M, Belongie S, et al.: Microsoft COCO: Common objects in context. Computer Vision - ECCV 2014. 2014, 8693:740-755. 10.1007/978-3-319-10602-1_48
46. Fang A, Ilharco G, Wortsman M, Wan Y, Shankar V, Dave A, Schmidt L: Data determines distributional robustness in contrastive language image pre-training (CLIP). arXiv. 2022, 10.48550/arXiv.2205.01397
47. Plummer BA, Wang L, Cervantes CM, Caicedo JC, Hockenmaier J, Lazebnik S: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile. 2015, 2641-2649. 10.1109/ICCV.2015.303
48. Everingham M, Eslami SMA, Van Gool L, Williams CKI, Winn J, Zisserman A: The Pascal visual object classes challenge: a retrospective. International Journal of Computer Vision. 2015, 111:98-136. 10.1007/s11263-014-0733-5