International Journal of Electrical and Computer Engineering (IJECE)
Vol. 15, No. 3, June 2025, pp. 3257~3266
ISSN: 2088-8708, DOI: 10.11591/ijece.v15i3.pp3257-3266

A comprehensive survey on automatic image captioning-deep learning techniques, datasets and evaluation parameters

Harshil Narendrabhai Chauhan[1], Chintan Thacker[2]
[1]Computer Science and Engineering, Parul University, Vadodara, India
[2]Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Parul University, Vadodara, India

Article Info

Article history:

Received Jun 10, 2024
Revised Dec 17, 2024
Accepted Jan 16, 2025

Keywords:

Attention mechanism
Convolutional neural network-recurrent neural network
Description generation
Encoder-decoder
Image caption
Transformer

ABSTRACT

Automatic image captioning is a pivotal intersection of computer vision and natural language processing, aiming to generate descriptive textual content from visual inputs. This comprehensive survey explores the evolution and state-of-the-art advancements in image caption generation, focusing on deep learning techniques, benchmark datasets, and evaluation parameters. We begin by tracing the progression from early approaches to contemporary deep learning methodologies, emphasizing encoder-decoder based models and transformer-based models. We then systematically review the datasets that have been instrumental in training and benchmarking image captioning models, including MSCOCO, Flickr30k, Flickr8k, and PASCAL 1k, discussing image count, types of scenes, and sources. Furthermore, we delve into the evaluation metrics employed to assess model performance, such as bilingual evaluation understudy (BLEU), metric for evaluation of translation with explicit ordering (METEOR), recall-oriented understudy for gisting evaluation (ROUGE), and consensus-based image description evaluation (CIDEr), analyzing their domains, bases, and measurement criteria. Through this survey, we aim to provide a detailed understanding of the current landscape, identify challenges, and propose future research directions in automatic image captioning.

Corresponding Author:

Harshil Narendrabhai Chauhan
Computer Science and Engineering, Parul University
Vadodara, Gujarat, India
Email: harshil.chauhan18838@paruluniversity.ac.in

1. INTRODUCTION

Humans can describe visual things in their native language effortlessly. It is a common habit that whenever we encounter visual content, we extract its features and describe its contents. This task is straightforward for humans, but challenging for machines. Transforming image in to the caption through computer vision and natural language processing, requiring information such as object recognition, relationship identification, location, and activity recognition. Applications for image caption generation include video captioning, medical imaging, aiding the visually impaired, automatic picture retrieval, and more [1], [2].

Image caption generation typically involves a two-step process: first, comprehending the visual content within the image, and then translating this understanding into natural language. Extracting visual information may involve tasks such as object detection, recognition, and identifying relationships [1]. In the early stages, image captioning relied on rule-based or retrieval-based methods [3]. Subsequently, more sophisticated models emerged using deep learning techniques, where a convolutional neural network (CNN)

extracts visual features as an encoder, and a recurrent neural network (RNN), such as long short-term memory (LSTM) [4] or gated recurrent unit (GRU) [3], decodes these features to generate coherent captions.

This paper is structured as follows: The second section offers a comprehensive review of the existing literature related to this area of study. The third section will explore current deep learning techniques employed for generating automatic image captions. The fourth section will outline various benchmark datasets. The fifth section will examine the different evaluation parameters used to examine captioning models. Lastly, the sixth section will present the conclusions derived from this research.

2. RELATED WORK

This provides a detailed examination of different deep learning techniques used for image caption generation. Early work primarily concentrated on retrieval-based and template-based methods. Template-based methods employed fixed templates with blank slots for generating descriptions [5], [6]. However, these methods were limited in their ability to produce variable-length captions due to their fixed structure.

On the other hand, retrieval-based approaches retrieved captions from existing datasets by searching for similar images in the training dataset. While capable of generating syntactically correct captions, they often lacked semantic accuracy. In recent years, the use of deep learning-based approaches in natural language processing and computer vision has grown considerably. These advanced models have overcome the constraints of traditional template- and retrieval-based methods.

Sasibhooshan *et al.* [7] proposed an encoder-decoder model enhanced with visual attention and spatial relation extraction. Their approach included a wavelet transform-based CNN and a virtual attention prediction network (VAPN) serving as an encoder to capture inter-spatial and inter-channel relationships. The decoder utilized LSTM. Despite its strengths, this model faced challenges with incorrect object recognition and failed to detect activities in complex scenes. The authors recommended exploring advanced transformer networks and extending the application to video captioning as future directions.

Al-Malla *et al.* [8] proposed an attention-based encoder-decoder model that leverages pre-trained Xception and YOLOv4 models for feature extraction and object detection, respectively. They evaluated their model on the MSCOCO and Flickr30K datasets, suggesting the adoption of more sophisticated methods and complex language models to enhance accuracy. Wang *et al.* [9] introduced a multilayer dense attention model, where the faster recurrent-CNN is used for feature extraction, and LSTM networks are employed to generate captions. To address the issue of asymmetric information, they implemented a strategy gradient optimization approach within a reinforcement learning framework.

Dhir *et al.* [10] developed Hindi caption generation model using attention mechanism, utilizing ResNet 101 for extracting the features and GRU for sentence generation. Their model was evaluated using the bilingual evaluation understudy (BLEU) metric on the MSCOCO dataset. Mishra *et al.* [11] proposed GPT-2 framework for Hindi image caption generation along with Geometric attention mechanism, integrating region-based convolutional neural networks (R-CNN) and ResNet 101 for object detection. While achieving a good BLEU score, their model struggled with object detection and caption generation in certain cases.

Mishra *et al.* [12] proposed a dynamic convolution-based encoder-decoder framework for Hindi captioning, integrated ResNet 101 with dynamic convolution layer and LSTM for generating descriptions. Despite incorporating various attention mechanisms, the model encountered challenges in activity recognition and object counting. Singh *et al.* [13] introduced an encoder-decoder framework for Hindi caption generation, utilizing visual geometry group16 (VGG16) CNN for feature extraction and LSTM for caption generation. Their model demonstrated promising results, suggesting further enhancements through the incorporation of attention mechanisms. Table 1 (see in appendix) presents a comparative study of various captioning models with respect to evaluation parameters derived from image caption generation on benchmark datasets. Here, the BLEU-1, BLEU-2, BLEU-3, BLEU-4, recall-oriented understudy for gisting evaluation (ROUGE), metric for evaluation of translation with explicit ordering (METEOR), and consensus-based image description evaluation (CIDEr) scores are represented by B1, B2, B3, B4, R, M, and C, respectively.

3. AUTOMATIC IMAGE CAPTIONING TECHNIQUES

This paper discusses advancement in automatic caption generation from image using the deep learning techniques. In recent years two approaches have got researcher's attention and better improvement in image caption generation. These approaches focusing on the incorporating the object detection, attention mechanism, color detection and recognition, object relation mapping [14] and many more to accurately generate the captions.

3.1. Transformer based approach

The approach, first introduced by Vaswani *et al.* [15], consists of an encoder-decoder structure. The encoder is composed of several layers, each containing two key components: a multi-head self-attention mechanism and a position-wise fully connected feed-forward network, as shown in Figure 1. This design enables the encoder to process the input sequence and create continuous representations. Similarly, the decoder also consists of multiple layers, with each layer featuring three main components: a multi-head self-attention mechanism, a multi-head encoder-decoder attention mechanism, and a position-wise fully connected feed-forward network. The decoder's role is to generate the output sequence by utilizing the encoder's representations along with the tokens generated in the previous steps. This approach offers several advantages over traditional encoder-decoder-based approaches, such as increased parallelization during training and inference, the capability to capture long-term dependencies, a simple architecture, scalability, and interpretability.

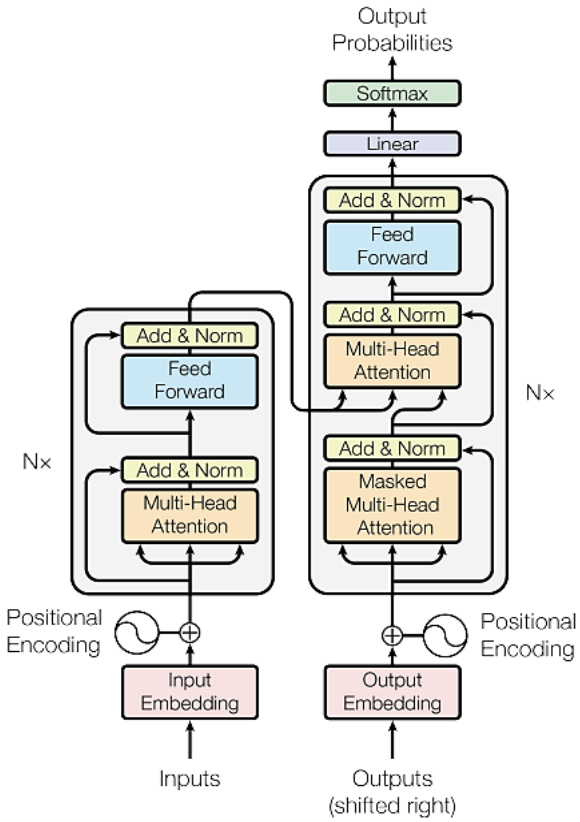


Figure 1. Transformer based model [15]

3.2. Key components

The self-attention mechanism is a fundamental element of transformer-based models. Instead of assigning weights based solely on word positions, it allows the model to assess the relative significance of each word within a sentence.

a. Self-attention mechanism: assists the decoder in grasping the connections among words.
b. Positional encoding: supplies information that indicates the position of each word within the sequence.
c. Feed-forward neural networks: functions independently on every token to enhance the representation of each one.
d. Residual connections and layer normalization: residual connections facilitate the smooth flow of gradients while the training process is stabilized by the layer normalization.

4. ENCODER-DECODER BASED APPROACH

A widely used approach for image captioning, where textual descriptions are generated from images. Two main components of this approach: input image processing through encoder and generating the

corresponding caption through decoder. By incorporating an attention mechanism, the decoder can focus on relevant regions of the image when generating each word in the caption, enhancing the overall performance. For image feature extraction, a pre-trained CNN such as ResNet, Inception, or VGG is utilized by [16]–[18]. The CNN produces a feature map, offering a spatial representation of the image. Typically, the decoder employs architectures like LSTM or GRU to sequentially generate the caption, one word at a time. Figure 2 illustrates this approach with each layer.

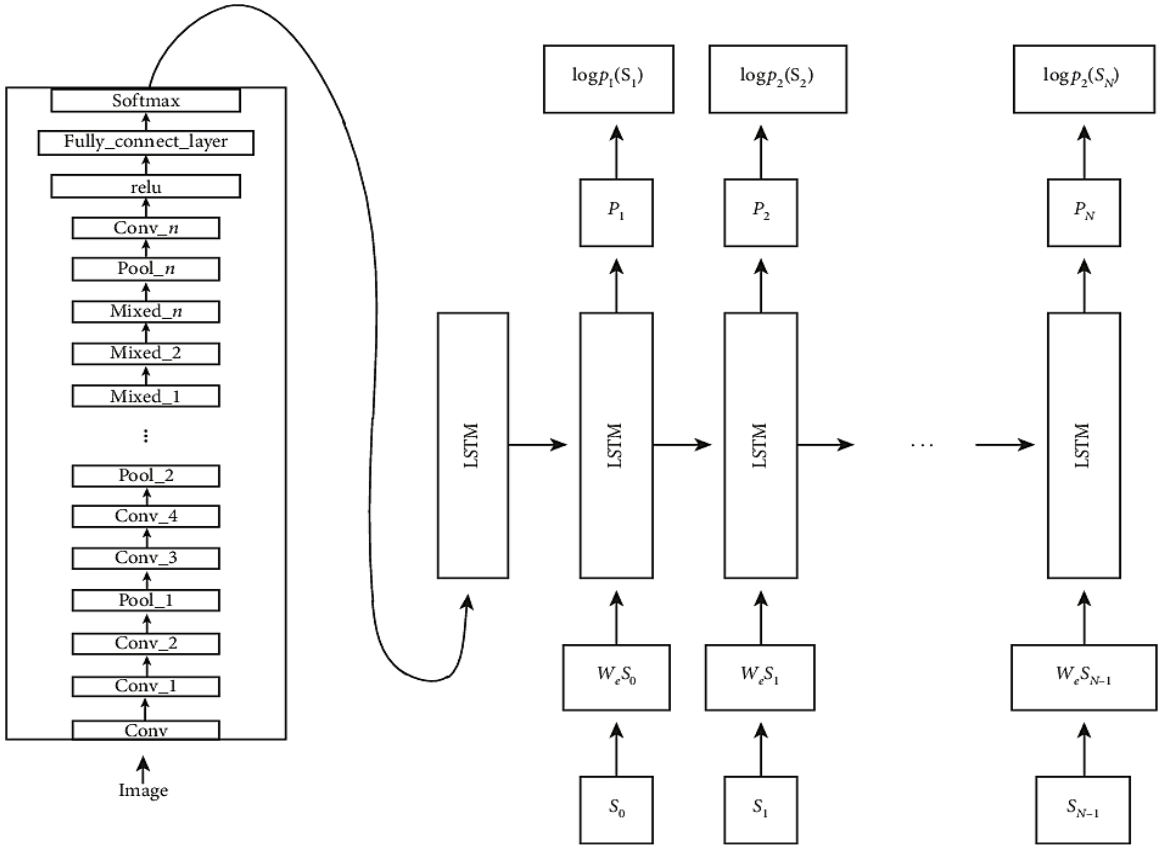


Figure 2. Encoder-decoder based model [4]

4.1. Key components

The encoder plays a crucial to extract the high-level image features by using pre-trained convolution models, such as ResNet, VGG, and Xception. Meanwhile, LSTM or GRU serve as the decoder, generating the caption word by word.

a. Feature extraction: capturing features like details about objects, textures, and colors from the given image done by the pre-trained convolution model.
b. RNN model (like LSTM or GRU): processes the encoded features to produce words in sequence, preserving context from previously generated words through hidden states.
c. Attention mechanism: enables the decoder to concentrate on particular areas of the image during caption generation, thereby enhancing the relevance and quality of the produced captions.

5. BENCHMARK DATASETS

This chapter introduces the available datasets for image caption generation. There are three main elements of artificial intelligence development: data, computational power, and algorithms. It is said that the efficiency of any algorithm or model depends on the dataset. Table 2 provides comparison of different benchmark datasets based on count, reference caption per image, type of scene and its source. Various challenges and dataset complexities which can lead inaccurate caption generation defined in Table 3.

Table 2. Comparison of different benchmark datasets

Sr. No.	Data set	Image count	Captions	Scenes	Source
1.	MSCOCO	328000	5	Mixed	Web
2.	Flickr30K	30000	5	People and animal	Flickr group
3.	Flickr8K	8000	5	People and animal	Flickr group
4.	PASCAL1K	1000	5	Mixed	PASCAL VOC
5.	Visual Genome	108249	5	Mixed	Web

Table 3. Complexity and challenges of dataset on model's performance

Sr. No.	Data set	Complexity	Challenges
1.	MSCOCO	Overlapping items and differing scales heightens the difficulty	Ambiguity, context sensitivity, caption length
2.	Flickr30K	Diverse image compositions and settings	Inconsistency in captions, subjectivity, less rich context
3.	Flickr8K	Limited variety of subjects and interactions	Limited vocabulary, overfitting
4.	PASCAL1K	Focus on specific objects rather than comprehensive scene understanding	Centered around objects, lack of descriptive captions, smaller scale for captioning tasks

5.1. MSCOCO

The MSCOCO dataset, developed by Microsoft [19], stands as a cornerstone benchmark for image caption generation endeavors. It was meticulously curated by the Microsoft team, comprising a diverse array of images spanning various object categories and scenes. Images are drawn from different contexts and situations, covering common objects and interactions. A sample image from MSCOCO dataset shown in Figure 3.

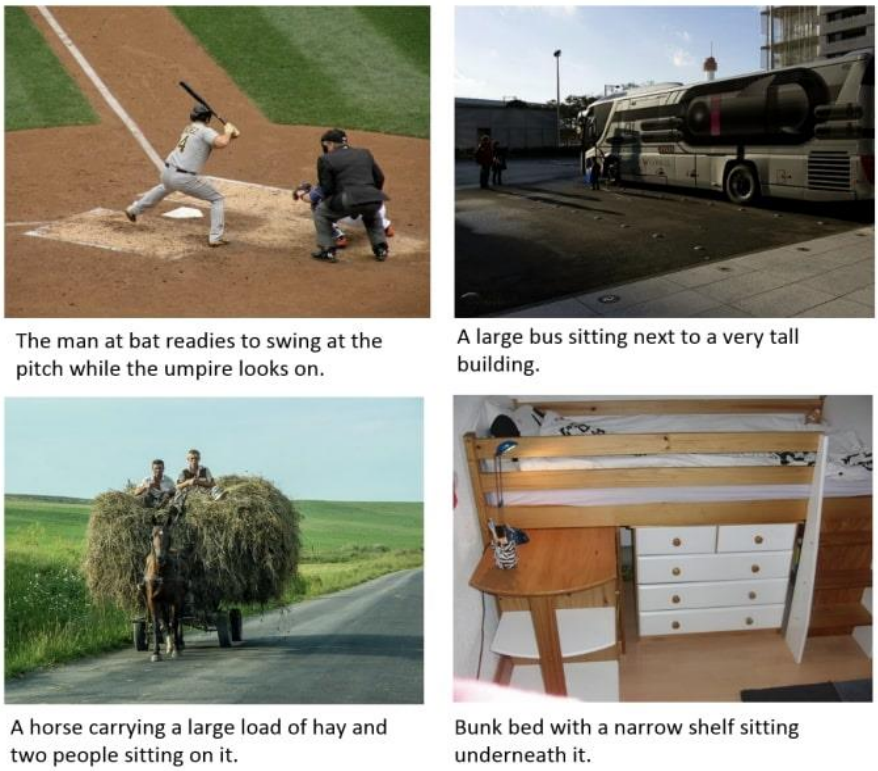
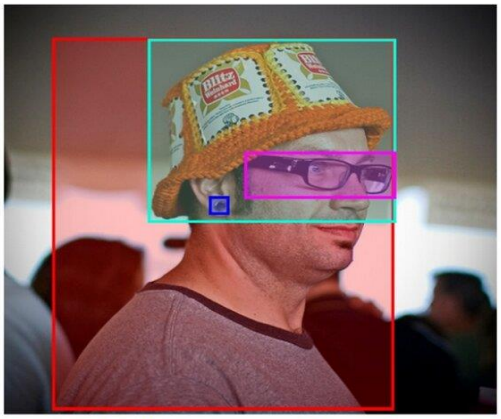


Figure 3. MSCOCO dataset example [19]

5.2. Flickr30K

The Flickr30K dataset, referenced as [20], is a significant resource in the realm of image description datasets. It provides comprehensive ground-truth annotations, creating a detailed mapping of regions in images and corresponding caption phrases [20]. Images come from a variety of sources, showcasing people in diverse scenarios, including everyday activities and events. Example shown in Figure 4.



A man with pierced ears is wearing glasses and an orange hat.
A man with glasses is wearing a beer can crotched hat.
A man with gauges and glasses is wearing a Blitz hat.
A man in an orange hat starring at something.
A man wears an orange hat and glasses.

Figure 4. Flickr30K dataset example [20]

5.3. Flickr8K

The Flickr8K [21] dataset was designed by collecting images from Yahoo's photo album site Flickr. It can contain 8,092 images of actions featuring people in outdoors or social contexts. Each image has five reference captions as shown in Figure 5.

A man is doing tricks on a bicycle on ramps in front of a crowd.
A man on a bike executes a jump as part of a competition while the crowd watch
A man rides a yellow bike over a ramp while others watch.
Bike rider jumping obstacles.
Bmx biker jumps off of ramp.

Figure 5. Flickr8K dataset example [21]

5.4. PASCAL 1K

The PASCAL 1K dataset [22], a subset derived from the well-known pattern analysis, statistical modeling, and computational learning-visual object classes (PASCAL VOC) challenge, a prominent resource in the field. It offers standardized annotations and evaluation systems for images. This dataset comprises 1,000 images, with a subset of 50 images selected randomly. These selected images were then manually annotated with five captions each, leveraging Amazon's Turkish robot service. Sample example given in Figure 6 with reference captions.

Two men playing cards at a table.
The two men are in an intense card game.
Two men playing cards on a kitchen counter are lit by a strong flash.
The men are playing cards
The scene shows two people playing cards.

Figure 6. PASCAL 1K dataset example [22]

6. EVALUATION PARAMETERS

This chapter examines the evaluation parameters used to measure the performance of image caption generation modes. These metrics collectively measure the consistency of n-gram matches between the generated captions and reference captions, offering a thorough evaluation of the model's performance. Table 4 presents the domain, base, and measurement criteria used by these evaluation parameters.

Table 4. Evaluation parameter summarization

Sr. No.	Evaluation parameter	Main domain	Base	Measurement criteria
1.	BLEU	Translation	Precision	n-gram overlap
2.	METEOR	Translation	Precision and recall	n-gram's matching and comparison
3.	ROUGE	Translation	Recall	n-gram and sequence overlap
4.	CIDEr	Image captioning	Precision and recall	n-gram's cosine similarity

6.1. BLEU

This bilingual evaluation understudy (BLEU) [23] is a variable length parameter used to measure the performance between computed captions and reference captions. It is a standard metric which most preferable for measurement. It compares the sentences on basis of n-gram. The mathematical formula of the BLUE score:

$$BLEU = \text{BP} * \exp(\sum_{n=1}^{N} w_n \log p_n) \quad (1)$$

BLEU is a widely-used metric that assigns a score between 0 and 1, where 0 signifies a low score and 1 denotes a perfect match. However, BLEU has its limitations. Firstly, if the generated text is longer than the reference text, it may not receive a good score. Secondly, there are instances where the generated text quality is not high, yet the BLEU score remains high.

6.2. METEOR

Metric for evaluation of translation with explicit ordering (METEOR) [24] is automatic metric used for evaluating machine translated caption, developed to address the limitations of BLEU. METEOR measure the alignment between machine translation and human translation. It computes precision, recall, and F-score. METEOR's mathematical formula is as (2):

$$METEOR = \frac{10 * Precesion * Recall}{Recall + 9 * Precesion} \quad (2)$$

6.3. ROUGE

Recall-oriented understudy for gisting evaluation (ROUGE) [25] is a collection of metrics used for evaluating summaries, based on the concept of the Longest Common Subsequence. It measures the similarity between n-grams, word pairs, and word sequences in the generated text compared to those in human-generated summaries. A higher score reflects better performance. This metric is particularly valuable for assessing the fluency and adequacy of machine-generated translations.

$$\text{ROUGE} - \text{N} = \frac{\sum_{\text{S}\in\{\text{Referencessummaries}\}} \sum_{\text{gram}_\text{n}\in\text{S}} \text{Count}_{\text{match}}(\text{gram}_\text{n})}{\sum_{\text{S}\in\{\text{Referencessummaries}\}} \sum_{\text{gram}_\text{n}\in\text{S}} \text{Count}(\text{gram}_\text{n})} \quad (3)$$

6.4. CIDEr

Consensus-based image description evaluation (CIDEr) is designed by [26] to measure the similarities between human-created sentences and machine-generated sentences. CIDEr can measure grammar, saliency, importance, and accuracy (CIDEr). This metric can consider the sentences as a "Document" and represent it as a term frequency–inverse document frequency (TF-IDF) after that it will calculate cosine similarities between human-created sentences and machine-generated sentences. It is also called a vector space model. Cosine Similarity calculated by (4):

$$CIDEr_n(c_i, S_i) = \frac{1}{m}\sum_j \frac{g^n(c_i)^T g^n(S_{ij})}{||g^n(c_i)||.||g^n(S_{ij})||} \quad (4)$$

where vector $\text{g}^\text{n}(\text{c}_\text{i})$ formed by all n-gram of length n. $||\text{g}^\text{n}(\text{c}_\text{i})||$ is a vector magnitude. Same apply for $\text{g}^\text{n}(\text{S}_\text{ij})$. For grammar and semantic properties, higher order n-gram used. From n-gram of varying size, scores combine like:

$$CIDEr_n(c_i, S_i) = \sum_{n=1}^{N} \omega_n CIDEr_n(c_i, S_i) \quad (5)$$

7. CONCLUSION

In this study, we conducted an in-depth investigation into various deep learning-based methodologies opted for image caption generation. We summarized the different approaches used for English and Hindi language with results and findings. Furthermore, our research extended to the examination of various state-of-the-art datasets and the evaluation parameters commonly employed in image captioning tasks. Upon thorough analysis, we arrived at the conclusion that existing models exhibit limitations in accurately identifying both objects and activities depicted within images. This highlights the need for further advancements in image captioning techniques to enhance the models' ability to interpret and describe visual content with greater precision and contextual understanding. For future direction, we suggest optimizing the evaluation parameters and generating the captioning for other languages like Gujarati, Marathi, and Tamil.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Harshil Chauhan	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Chintan Thacker		✓				✓		✓		✓	✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

Data availability is not applicable to this paper as no new data were created or analyzed in this study.

REFERENCES

[1] S. Bai and S. An, "A survey on automatic image caption generation," *Neurocomputing*, vol. 311, pp. 291–304, Oct. 2018, doi: 10.1016/j.neucom.2018.05.080.
[2] B. Makav and V. Kilic, "A new image captioning approach for visually impaired people," in *2019 11th International Conference on Electrical and Electronics Engineering (ELECO)*, Nov. 2019, pp. 945–949, doi: 10.23919/ELECO47770.2019.8990630.
[3] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys*, vol. 51, no. 6, pp. 1–36, Nov. 2019, doi: 10.1145/3295748.
[4] H. Wang, Y. Zhang, and X. Yu, "An overview of image caption generation methods," *Computational Intelligence and Neuroscience*, vol. 2020, pp. 1–13, Jan. 2020, doi: 10.1155/2020/3062706.
[5] A. K. Poddar and D. R. Rani, "Hybrid architecture using CNN and LSTM for image captioning in Hindi language," *Procedia Computer Science*, vol. 218, pp. 686–696, 2023, doi: 10.1016/j.procs.2023.01.049.
[6] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," *IEEE Transactions on Multimedia*, vol. 21, no. 11, pp. 2942–2956, Nov. 2019, doi: 10.1109/TMM.2019.2915033.
[7] R. Sasibhooshan, S. Kumaraswamy, and S. Sasidharan, "Image caption generation using visual attention prediction and contextual spatial relation extraction," *Journal of Big Data*, vol. 10, no. 1, p. 18, Feb. 2023, doi: 10.1186/s40537-023-00693-9.
[8] M. A. Al-Malla, A. Jafar, and N. Ghneim, "Image captioning model using attention and object features to mimic human image understanding," *Journal of Big Data*, vol. 9, no. 1, p. 20, Dec. 2022, doi: 10.1186/s40537-022-00571-w.
[9] K. Wang, X. Zhang, F. Wang, T.-Y. Wu, and C.-M. Chen, "Multilayer dense attention model for image caption," *IEEE Access*, vol. 7, pp. 66358–66368, 2019, doi: 10.1109/ACCESS.2019.2917771.

[10] R. Dhir, S. K. Mishra, S. Saha, and P. Bhattacharyya, "A deep attention-based framework for image caption generation in Hindi language," *Computación y Sistemas*, vol. 23, no. 3, Oct. 2019, doi: 10.13053/cys-23-3-3269.
[11] S. K. Mishra, S. Sinha, S. Saha, and P. Bhattacharyya, "Dynamic convolution-based encoder-decoder framework for image captioning in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 4, pp. 1–18, Apr. 2023, doi: 10.1145/3573891.
[12] S. K. Mishra, S. Chakraborty, S. Saha, and P. Bhattacharyya, "GAGPT-2: a geometric attention-based GPT-2 framework for image captioning in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 10, pp. 1–16, Oct. 2023, doi: 10.1145/3622936.
[13] A. Singh, T. D. Singh, and S. Bandyopadhyay, "An Encoder-Decoder based framework for Hindi image caption generation," *Multimedia Tools and Applications*, vol. 80, no. 28–29, pp. 35721–35740, Nov. 2021, doi: 10.1007/s11042-021-11106-5.
[14] Z. U. Kamangar, G. M. Shaikh, S. Hassan, N. Mughal, and U. A. Kamangar, "Image caption generation related to object detection and colour recognition using transformer-decoder," in *2023 4th International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Mar. 2023, pp. 1–5, doi: 10.1109/iCoMET57998.2023.10099161.
[15] A. Vaswani *et al.*, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 2017-Decem, no. Nips, pp. 5999–6009, 2017.
[16] S. K. Mishra, Harshit, S. Saha, and P. Bhattacharyya, "An object localization-based dense image captioning framework in Hindi," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 2, pp. 1–15, Mar. 2023, doi: 10.1145/3558391.
[17] M. Kaur and H. Kaur, "An efficient deep learning based hybrid model for image caption generation," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 3, 2023, doi: 10.14569/IJACSA.2023.0140326.
[18] S. Parida, O. Bojar, and S. R. Dash, "Hindi visual genome: a dataset for multi-modal English to Hindi machine translation," *Computacion y Sistemas*, vol. 23, no. 4, pp. 1499–1505, 2019, doi: 10.13053/CyS-23-4-3294.
[19] X. Chen *et al.*, "Microsoft COCO captions: data collection and evaluation server," *ArXiv:1504.00325*, 2015, doi: 10.48550/arXiv.1504.00325.
[20] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k entities: collecting region-to-phrase correspondences for richer image-to-sentence models (Version 4)," *International Journal of Computer Vision*, vol. 123, no. 1, pp. 74–93, 2017, doi: 10.1007/s11263-016-0965-7.
[21] M. Hodosh, P. Young, and J. Hockenmaier, "Framing image description as a ranking task: data, models and evaluation metrics," *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2015-Janua, pp. 4188–4192, 2015.
[22] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier, "Collecting image annotations using Amazon's Mechanical Turk," in *Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Mturk 2010 at the 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2010 - Proceedings*, 2010, pp. 139–147.
[23] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 2001, p. 311, doi: 10.3115/1073083.1073135.
[24] S. Banerjee and A. Lavie, "METEOR: an automatic metric for mt evaluation with improved correlation with human judgments," *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization, Proceedings of the Workshop ACL 2005*, pp. 65–72, 2005.
[25] S. California and A. W. Marina, "ROUGE: a package for automatic evaluation of summaries," *Information Sciences*, 2001.
[26] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: consensus-based image description evaluation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2015, pp. 4566–4575, doi: 10.1109/CVPR.2015.7299087.

APPENDIX

Table 1. Comparative analysis of evaluation parameters obtained by researchers

Ref#	Encoder	Decoder	Caption language	Dataset	B1	B2	B3	B4	M	R	C
7	Wavelet-CNN + VAPN	LSTM	English	MSCOCO	78.5	62	49.1	38	28.9	58.3	124.2
				Flickr30K	70.1	49.4	35.8	27.2	21.7	-	67.3
				Flickr8K	70.5	50.2	37.3	28.6	24.5	-	-
8	Xception + YOLOv4	Bahdanau attention + GRU	English	MSCOCO	49.2	29.6	17.4	10.1	16.3	35.8	39
				Flickr30K	39.8	22.1	11.6	6.1	12.9	29.8	15
9	Faster R-CNN	LSTM	English	MSCOCO	-	-	51	34	24.8	56.5	118.3
				Flickr30K	-	-	52	39.1	26	51	209.1
				Flickr8K	-	-	47.7	33.4	23.1	46.9	167.5
				Chinese_AI	-	-	66	58.1	41.6	69.5	118.9
10	VGG19	LSTM	English	Flickr30K	53.9	24	10.2	4.6	21	-	-
				Flickr8K	54	22	7.9	3	20	-	-
10	ResNet-101	GRU	Hindi	MSCOCO	57	39.1	26.4	17.3	-	-	-
11	ResNet + Transformer model	GPT-2	Hindi	MSCOCO	69.7	54.2	39.8	27.7	-	-	-
12	Dynamic CNN + X-linear attention	LSTM	Hindi	MSCOCO	68	49	34.2	21.2	-	-	-
13	VGG19	LSTM	Hindi	Visual Genome	66	-	-	-	-	-	-
14	Transformer-Encoder	Transformer-Decoder	English	Flickr8K	85.5	78.4	71	48.5	-	-	-

BIOGRAPHIES OF AUTHORS

Harshil Narendrabhai Chauhan has received M.E. degree in computer engineering from HJD-ITER, Kera, Kutch, Gujarat, India and pursuing Ph.D. in computer science and engineering from Parul University, Vadodara, Gujarat, India. He has 7+ years of experience in academia. Currently he serves as an assistant professor, in the Department of Computer Engineering at Parul University, Vadodara, Gujarat, India. His research interests are in the areas of machine learning, deep learning, and computer vision. He can be contacted at email: harshil.chauhan18838@paruluniverisity.ac.in.

Chintan Thacker received the Ph. D. degree in the domain of artificial intelligence and computer vision from Gujrat Technical University in the year 2022. He had served as head of The Department of Computer Science and Engineering Department at HJD Institute of Technical Education and Research, Kera, India. He has 1+ years of experience in industry and 12+ years of experience in academia. Currently, he serves as an associate professor in Computer Science and Engineering Department in Parul Institute of Technology, Parul University, Vadodara, Gujarat. In addition, he has also guided several doctorate students and has been active in conducting several workshops in the domain of computer vision. He can be contacted at email: chintan.thacker19435@paruluniversity.ac.in.