Survey on Multimodal Image Captioning Approaches: Addressing Contextual Understanding, Cross-Dataset Generalization, and Multilingual Captioning

Mr. Nikhil Gopal Khodave[1] and Mr. Prathamesh S. Powar[2]
Student, Computer Science and Engineering [1]
Asst. Professor, Computer Science and Engineering[2]
Ashokrao Mane Group of Institution, Vathar, Kolhapur, India

Abstract: *This paper discusses new advances in multimodal image captioning, and the focus lies on improving access, contextual understanding, and generalization across many datasets. Current state-of-the-art uni-modal approaches are no longer good enough, whereas innovative multimodal techniques combine the visual and textual features to yield more accurate captions and richness of the captions produced. This includes the attention mechanism, scene graphs, and pre-trained transformer models through which more contextual descriptions are made. Further, the paper addresses challenges in cross-dataset generalization and multilingual captioning, pointing to the necessity of systems that adapt to real-world variability and support diversity in linguistic backgrounds. Through synthesizing the research conducted so far, this work outlines future directions for the creation of more inclusive, robust, and effective image captioning technologies, especially for applications in accessibility*

Keywords: Multimodal Image Captioning, Accessibility, Contextual Understanding, Cross-Dataset Generalization

I. INTRODUCTION

Image captioning is the process of creating descriptions for images that bridges the gap between visual and textual data. It has application in multimedia retrieval, autonomous systems, and accessibility to visually impaired persons. Captioning of images enables persons with visual disabilities to understand their surroundings and environment much better by converting visual content into text. According to [1], deep learning approaches for accessibility enhancement are very important for enhancing these interactions.

Traditionally, most captioning algorithms used unimodal approaches that are visual feature based. These approaches have the problem of failing to provide the captions needed for deep understanding as they are mostly context-free and semantically poor. To this end, new multimodal strategies emerged that integrate both visual and textual modalities in order to enhance captioning performance. Authors such as [2] and [3] proposed frameworks that integrate advanced techniques such as attention mechanisms and pretrained transformer models to improve the quality of captions.

Multimodal approaches combine both visual and textual data so that models are in a better position to extract complementary features that improve captioning. For example, [4] proposed BCAN, which integrates diverse visual features for better performance during captioning. [5] Further used multimodal data augmentation with the help of diffusion models to enhance generalization on limited datasets. This would allow models to better understand and generate the proper descriptions of intricate visual content[6]. addressed the problem of accessibility through the creation of dual-attention mechanisms, incorporating both visual and textual attention in captions for greater richness and inclusivity. [7]. emphasized further human-aligned evaluation metrics in enhancing captioning systems for applicability in the real world and, therefore, catering to diverse needs.

DOI: 10.48175/IJARSCT-25104

These improvements show how much of a difference it makes to bring multimodal multi-parters into the image captioning scene so much that it becomes a benchmark for accessible and inclusive technologies. Prospects for future advancement are big in this field for accessibility tools as well as human-computer interaction. However, the research gaps that still remain in the domain are those of contextual understanding, cross-dataset generalization, and multilingual captioning. The current models fail to capture the full context of images, such as object interactions and scene dynamics. Moreover, most of the models rely on specific datasets, limiting generalization across many real-world scenarios. Finally, the lack of multilingual support limits the global accessibility of image captioning systems. The gaps explained here are critical to the further advancement of the accuracy, adaptability, and inclusivity of these models so that more effective and more accessible image captioning technologies will be developed.

II. FOUNDATIONS OF MULTIMODAL IMAGE CAPTIONING

Multimodal image captioning attempts to fill in the gap between visual data and textual descriptions to allow models to produce captions that are both accurate and contextually rich. Core techniques for the transformation of images into meaningful text have been developed, such as visual feature extraction, sequential caption generation, and attention mechanisms, in an effort to overcome these challenges. Collectively, these foundational advancements make it possible to create more sophisticated and accessible captioning systems.

Convolutional Neural Networks (CNNs) have recently been seen as the backbone for extracting features from images, since such networks efficiently encode images into feature vectors that preserve both high-level semantic information and spatial details. [8] showed that pre-trained models of CNN such as VGG16 are helpful to encode images and create captions out of them; similarly, for the precise identification of objects and enhancement of caption accuracy, a backbone architecture uses ResNet architectures to extract local and global features, defined by [4]. The input image data then is processed downstream through the convolutional neural nets. Generally, recurrent neural networks are used, more specifically Long Short-Term Memory networks for producing coherent and contextually correct captions. They feed the visual features extracted from images in sequence; hence the captioned text adheres to the image's content. [9] emphasized that LSTMs are particularly useful in producing text that flows naturally while retaining image-relevant details. Coupled with part-of-speech tagging, [10] further introduced gated recurrent units and improved the system's robustness to contextual accuracy and grammatical soundness in generated descriptions.

The integration of attention mechanisms had improved the captions' quality; models could selectively focus on critical regions within the image. For instance, the bidirectional co-attention network (BCAN) introduced by [4] advances the interaction of visual regions with textual descriptions that achieves state-of-the-art captioning performance as reported. Similarly, [6] proposed dual attention mechanisms that combine visual and textual attention, which enables models to produce semantically richer and more contextually relevant captions. This mechanism allows selective focus on the critical elements of the image and ensures that the generated captions are both accurate and informative.

As depicted in the Fig 1, the multimodal image captioning pipeline is composed of several integral parts which work together in order to render the most meaningful description of theimage. It starts with visual feature extraction using CNNs, sequential caption generation with RNNs, and enhanced by the attention mechanisms, which enhance caption relevance and accuracy. Together, these methods form the bedrock of multimodal image captioning, now able to provide more accurate, efficient, and accessible systems for developing meaningful captions for various applications.

DOI: 10.48175/IJARSCT-25104

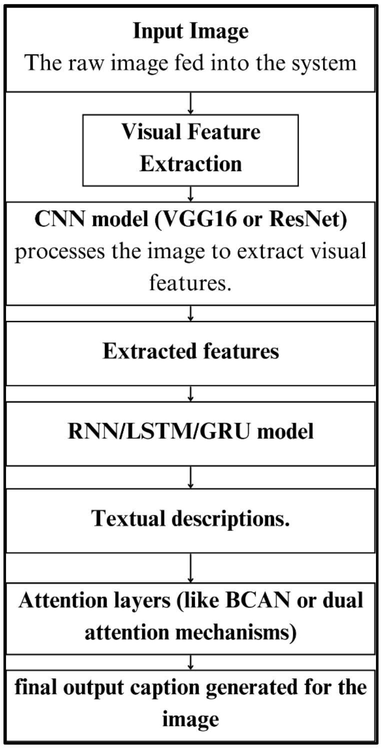
Input Image
The raw image fed into the system
Visual Feature Extraction
CNN model (VGG16 or ResNet) processes the image to extract visual features.
Extracted features
RNN/LSTM/GRU model
Textual descriptions.
Attention layers (like BCAN or dual attention mechanisms)
final output caption generated for the image

Fig 1: Multimodal image captioning pipeline

III. CONTEXTUAL UNDERSTANDING IN CAPTION GENERATION

While modern image captioning models have made major strides in object recognition, they are still far behind in terms of contextual understanding, especially as to describing interactions, functionality, or relationships between objects in a scene. This results in shallow or even misleading captions that seem not to account for the true dynamics of the scene being described.

Most current captioning systems have put emphasis on object detection and attributes identification, ignoring the interaction of those objects. For example, [11] showed that even if a system performs exceptionally well at identifying objects, it lacks understanding in the spatial or semantic interactions between them within the scene [11]. This gap was further stressed by [12], arguing that current models mostly fail to capture significant relational cues, like that "a person holding a knife in a kitchen" evokes action like cutting, which forms the significant part of the contextual scene.

The shortcomings are conspicuously highlighted in dynamic scenes, especially where interactant functions are of essence. For instance, in a kitchen scene, a model may be able to identify that there is a pan, a stove, and utensils, but it will not express functional relationships between them, such as cooking. Such restrictions have been because the models rarely get to perform inferring knowledge such as the common-sense ones or further deeper semantic connections important for achieving more profound sense perceptions of a scene [13]. Similarly, according to [14], models cannot always be able to capture group interactions; for example, people playing ball at a park ignore relational context of the actions, such as catching or throwing a ball.

With these constraints, several approaches have been introduced to improve contextual understanding of caption generation. Interaction modelling and acquisition of external knowledge are two promising approaches towards this direction. Scene graphs have also become a promising direction, bringing objects into a spatial-semantic relationship diagram. [15] proposed scene graph labels that map such relationships resulting in captions that are richer and descriptory. [16]. further extended this methodology by introducing the common-sense knowledge in scene graphs in which the captions became intuitive and contextual. Further relational modelling is provided through scene graph transformers like the one introduced by [17] based on the usage of dual transformer branches for catching global and local interactions, therefore providing much more detailed information regarding the dynamics in the scene.

These advances are a new step toward the deeper contextual understanding of images, and they open up the doors for captions that not only describe what is in an image but are also functionally accurate in a wider range of real-world settings.

IV. CROSS-DATASET GENERALIZATION AND ROBUSTNESS

Although the performance of image captioning models on well-established datasets such as MS-COCO and Flickr8k is remarkable, the application of these models to unseen real-world scenarios generally fails. The primary reason behind this is that the characteristics of narrowly defined datasets cause overfitting and a lack of adaptability in more diverse settings.

This leaves the models with quite a task of trying to identify objects when applying them to independent datasets like Conceptual Captions. According to [18], such overfitting, which results in the failure to generalize effectively over different data distributions, occurs frequently due to training on dataset-specific features such as image quality, scene types, and object representations. This over-reliance on curated datasets limits the model's applicability in real-world environments, since [19] discovered that performance drastically drops in domain-shifted scenarios, where the training data is not aligned with the characteristics of the new data.

Real-world applications, therefore, involve captioning models that can face a wide spectrum of settings which include different forms of scenes or cultural contexts. [20] had pointed out how domain shifts—a change in light, object, or background context—can bring about great hardships for image captioning systems to deal with effectively. The models often fail in providing the captions with high accuracy when applied to diverse populations or settings, noted by [14] since geographic and cultural differences often result in mismatches between the training data and real-world images.

To address these problems, several strategies have been proposed to improve model generalization. [14] proposed a cross-modal retrieval model that generates pseudo image-text pairs for previously unseen domains, thus increasing the adaptability of the model. [21] proposed multi-source domain adaptation, using adversarial optimization to enable models to generalize better across different domains. Moreover, [5] recommended dataset augmentation with adversarial transformations, which would improve the strength of the model to variations in the input data. [19]discussed the usage of multi-domain training where the alignments of feature representations are aligned across the several datasets to achieve generalization and overall better performance of the models.

These developments demand to have some independent approaches in the development of reliable, adaptive, and accurate captioning systems capable of handling the vastness of real-world scenarios.

V. MULTILINGUAL CAPTIONING AND ACCESSIBILITY

Because most image captioning models have concentrated on the English language, these models are less accessible in areas where other languages are more widely spoken. Such language bias confines the usability of these systems in multilingual societies. Moreover, [22] highlighted the fact that such a monolingual bias has further constrained many models' suitability for accommodating and representing a wider range of languages. [23] added to this fact: that a vast majority of annotation datasets available around the world exist only in limited English and create an enormous additional burden for constructing multilingual systems of captioning.

Above all else, a multilingual captioning system bears high importance due to the globalization issue. Not to mention its applicability allows native speakers belonging to different language-based regions, accessibility to visuals is achieved if one understands through one's first language. Hence, [24] suggest that lowresource language models also do have equal prominence in bridge-making, given its data requirements that are as hard to maintain when the former prevails or has the last say. [25] argued further that there is a need for multilingual support to guarantee equitable access to AI-driven tools, especially in education and accessibility, where multiple linguistic backgrounds exist.

To overcome the challenges of multilingual captioning, several novel approaches have been proposed. [23] developed XLM-R, a transformer-based multilingual model trained on more than 100 languages that has shown significant performance in cross-lingual tasks. [26] suggested a two-phase bilingual distillation method for fine-tuning low-resource languages multilingual models that significantly improved the performance of such models in the low-resource languages. Another effective technique has been shown to be transfer learning, such as in [27], where the authors used embedding alignment methods for transferring knowledge from English to low-resource languages for improving multilingual captioning capabilities.

Such advancements illustrate the massive scope of potential from multilingual captioning systems to further increase access and accessibility. By making use of a multilingual transformer with fine-tuned models over multiple datasets, it is possible to create an image captioning system that can provide more equity and universal accessibility.

VI. SYNTHESIS: ADDRESSING KEY CHALLENGES IN IMAGE CAPTIONING

Advances in image captioning have dramatically improved in recent research, focusing on several critical areas in order to complement previous shortcomings of models. Table 1 and table 2 holding the key findings and contribution of the discussion so far depicts that the central challenge lies in transposing visual and textual data by multimodal techniques. While the earlier attempts of unimodal methods were limited to considering only visual information, the modern attempts by authors like [2] and [3] stressed the use of multimodality and even used the attention mechanism and pre-trained transformer models to get the most accuracy in captioning. BCAN by [4] and diffusion-based multimodal data augmentation by [5]. in 2023 further refine the integration, hence enhancing generalization and contextual richness.

Paper	Focus Areas	Key Findings
[1]	Accessibility enhancement using deep learning	Emphasized the importance of deep learning for improving accessibility for individuals with visual impairments.
[2]	Multimodal captioning frameworks	Integrated advanced techniques such as attention mechanisms and pretrained transformer models for better captioning quality.
[3]	Enhancements in multimodal image captioning	Proposed multimodal frameworks to improve caption accuracy using deep learning models.
[4]	BCAN and attention mechanisms	Developed BCAN for improved interaction between visual features and textual descriptions.
[5]	Multimodal data augmentation using diffusion models	Focused on data augmentation to improve model generalization with limited datasets.
[6]	Dual attention mechanisms	Introduced dual-attention mechanisms for richer and more inclusive captions.
[11]	Contextual understanding in caption generation	Pointed out the failure of models to capture interactions between objects.
[12]	Interaction modeling and context-awareness	Showed that models often overlook interactions between objects and key contextual cues.
[16]	Common-sense knowledge in scene graphs	Integrated common-sense knowledge into scene graphs for better contextual accuracy.
[15]	Scene graph representation for object relationships	Demonstrated how scene graphs enhance caption richness by mapping spatial relationships.
[17]	Scene graph transformers	Proposed scene graph transformers to improve relational modelling and scene dynamics.
[18]	Cross-dataset generalization and domain shift	Addressed issues of over fitting and proposed methods for cross-domain adaptability.
[19]	Domain adaptation techniques	Emphasized multi-domain training and aligning feature representations to improve model performance.
[21]	Multi-source domain adaptation	Proposed adversarial optimization to improve generalization across diverse domains.
[5]	Adversarial transformations for	Suggested augmenting datasets with adversarial transformations

	robust captioning	to handle variations in input data.
[22]	Multilingual captioning challenges and solutions	Highlighted the importance of multilingual models to improve accessibility in diverse linguistic contexts.
[23]	XLM-R multilingual transformer model	Introduced XLM-R for improving multilingual tasks, especially for low-resource languages.
[24]	Multilingual models for low-resource languages	Emphasized the role of multilingual language models like XLM-R in bridging linguistic divides.
[26]	Bilingual distillation for multilingual models	Proposed bilingual distillation to improve performance in low-resource languages.
[27]	Transfer learning for multilingual captioning	Used embedding alignment to transfer knowledge from English to low-resource languages.

Table 1 : Key findings and contributions of discussion

Much has been achieved in object recognition, but the contextual understanding, especially about dynamic relationships and interactions between objects, is still missing. According to[11] and [12], key contextual cues are often missed by current models. Recent advances, such as scene graphs [15] and transformers for relational modeling [17], are promising solutions to this problem, which provide more accurate and semantically rich captions by mapping spatial and semantic relationships between objects.

Thirdly, there remains the crossdataset generalizability challenge. Here, following arguments by [18] and more so by [19] who pointed that typical models usually face failure under new domain shift but are successfully operational in limited ways in actual conditions, improving is possible via means such as the multi-source adaptation of the model [21], through adversarial transforms [5].

Multilingual captioning recently has emerged as an important direction, especially regarding enhancing accessibility across linguistic contexts. [22] and [23] underlined the deficiencies of monolingual models; hence, their proposed solutions should be multilingual models, including XLM-R by [23], and transfer learning by [27] to open accessibility to global perspectives.

These advancements underscore the evolution that is ongoing in image captioning, demonstrating that although a lot of ground has been covered, contextual understanding, cross-dataset generalization, and multilingual support are going to be key to further progress.

VII. CONCLUSION

The work has described transformative advancements in multimodal image captioning through an integration of visual and textual data to enrich quality and access the generated captions. The key development areas such as attention mechanism, scene graph, and multilingual captioning further enhance descriptions and make them contextually rich and accurate. But there remain critical challenges like completely capturing contextual understanding of images, enhancing cross-dataset generalization, and broadening the base of multilingual support.

The contributions of this paper give way to future research in several very important areas: the development of models that combine visual data with externally sourced textual information for richer captions, improved contextual understanding for better recognition of object relationships, extended generalization to diverse datasets, and real-time multilingual captioning capabilities. Furthermore, combining these innovations with accessible devices such as smartphones and smart glasses would make it easier to provide on-the-spot auditory descriptions for those who are blind, thus greatly increasing accessibility and interaction with their environment. The developments will ultimately create more adaptive and inclusive technologies in image captioning.

REFERENCES

[1]. S. S. Patil and P. J. Patel, “Image Captioning using Deep Learning Model for Visually Impaired People,” [Online]. Available: https://api.semanticscholar.org/CorpusID:269497836

[2]. T. Wang et al., “Caption anything: Interactive image description with diverse multimodal controls,” arXiv preprint arXiv:2305.02677, 2023.

[3]. R. Ramos, D. Elliott, and B. Martins, “Retrieval-augmented image captioning,” arXiv preprint arXiv:2302.08268, 2023.

[4]. W. Jiang, W. Wang, and H. Hu, “Bi-directional co-attention network for image captioning,” ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 17, no. 4, pp. 1–20, 2021.

[5]. C. Xiao, S. X. Xu, and K. Zhang, “Multimodal data augmentation for image captioning using diffusion models,” in Proceedings of the 1st Workshop on Large Generative Models Meet Multimodal Applications, 2023, pp. 23–33.

[6]. M. Liu, L. Li, H. Hu, W. Guan, and J. Tian, “Image caption generation with dual attention mechanism,” Inf Process Manag, vol. 57, no. 2, p. 102178, 2020.

[7]. Y. Wada, K. Kaneda, D. Saito, and K. Sugiura, “Polos: Multimodal Metric Learning from Human Feedback for Image Captioning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13559–13568.

[8]. M. K. Panigrahi et al., “Effectiveness and safety of Shankhaprakshalana—a yogic technique—in bowel preparation for colonoscopy: A retrospective study,” Indian Journal of Gastroenterology, vol. 43, no. 4, pp. 785–790, 2024.

[9]. V. Tiwari and C. Bhatnagar, “Automatic caption generation via attention based deep neural network model,” in 2021 9th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), 2021, pp. 1–6.

[10]. T. do Carmo Nogueira, C. D. N. Vinhal, G. da Cruz Júnior, and M. R. D. Ullmann, “Reference-based model using multimodal gated recurrent units for image captioning,” Multimed Tools Appl, vol. 79, pp. 30615–30635, 2020.

[11]. N. Xu, A.-A. Liu, Y. Wong, W. Nie, Y. Su, and M. Kankanhalli, “Scene graph inference via multi-scale context modeling,” IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 3, pp. 1031–1041, 2020.

[12]. X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, “A comprehensive survey of scene graphs: Generation and application,” IEEE Trans Pattern Anal Mach Intell, vol. 45, no. 1, pp. 1–26, 2021.

[13]. M. J. Khan, J. G Breslin, and E. Curry, “NeuSyRE: Neuro-symbolic visual understanding and reasoning framework based on scene graph enrichment,” Semant Web, vol. 15, no. 4, pp. 1389–1413, 2024.

[14]. L. Zhang, H. Yin, B. Hui, S. Liu, and W. Zhang, “Knowledge-based scene graph generation with visual contextual dependency,” Mathematics, vol. 10, no. 14, p. 2525, 2022.

[15]. K. Nguyen, S. Tripathi, B. Du, T. Guha, and T. Q. Nguyen, “In defense of scene graphs for image captioning,” in Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 1407–1416.

[16]. M. J. Khan, J. G. Breslin, and E. Curry, “Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning,” in European Semantic Web Conference, 2022, pp. 93–112.

[17]. S. Sun, D. Huang, X. Tao, C. Pan, G. Liu, and C. Chen, “Boosting Scene Graph Generation with Contextual Information,” ACM Transactions on Multimedia Computing, Communications and Applications, vol. 20, no. 2, pp. 1–24, 2023.

[18]. Z. Zhao, R. Hu, H. Ma, and X. Zhang, “Scene graph generation based on global embedding and contextual fusion,” in Third International Conference on Computer Science and Communication Technology (ICCSCT 2022), 2022, pp. 917–927.

[19]. Y. Ren et al., “Crossing the gap: Domain generalization for image captioning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 2871–2880.

[20]. Z. Yuan, X. Li, and Q. Wang, “Exploring multi-level attention and semantic relationship for remote sensing image captioning,” IEEE Access, vol. 8, pp. 2608–2620, 2019.

DOI: 10.48175/IJARSCT-25104

[21]. J. Si, H. Shi, T. Han, J. Chen, and C. Zheng, "Learn generalized features via multi-source domain adaptation: Intelligent diagnosis under variable/constant machine conditions," IEEE Sens J, vol. 22, no. 1, pp. 510–519, 2021.

[22]. M. Bhargava, K. Vijayan, O. Anand, and G. Raina, "Exploration of transfer learning capability of multilingual models for text classification," in Proceedings of the 2023 5th International Conference on Pattern Recognition and Intelligent Systems, 2023, pp. 45–50.

[23]. A Conneau, "Unsupervised cross-lingual representation learning at scale," arXiv preprint arXiv:1911.02116, 2019.

[24]. S. Doddapaneni, G. Ramesh, M. M. Khapra, A. Kunchukuttan, and P. Kumar, "A primer on pretrained multilingual language models," arXiv preprint arXiv:2107.00676, 2021.

[25]. J. Hu, M. Johnson, O. Firat, A. Siddhant, and G. Neubig, "Explicit alignment objectives for multilingual bidirectional encoders," arXiv preprint arXiv:2010.07972, 2020.

[26]. A Ansell, E. M. Ponti, A. Korhonen, and I. Vulić, "Distilling Efficient Language-Specific Models for Cross-Lingual Transfer," arXiv preprint arXiv:2306.01709, 2023.

[27]. K. Ding et al., "A Simple and Effective Method to Improve Zero-Shot Cross-Lingual Transfer Learning," arXiv preprint arXiv:2210.09934, 2022