International Journal of Electrical and Electronics Engineers
Volume 18, Issue No. 01, Jan-June 2026
ISSN (0) 2321-2055 ISSN (P) 2321-2045

IMAGE CAPTION GENERATION USING DEEP LEARNING

Sanagapalli . Sai Lakshmi Harika
Depatment Of ECE
Tirumala Engineering College
Jonnalagadda, India
sailakshmisanagapalli@gmail.com

Sadineni . Naga Vamsi
Department Of ECE
Tirumala Engineering College
Jonnalagadda , India
sadineninagavamsi1234@gmail.com

Pudoka . Mahesh
Department Of ECE
Tirumala Engineering College
Jonnalagadda, India
maheshpudoka@gmail.com

Supuri . Harshitha
Department Of ECE
Tirumala Engineering College
Jonnalagadda, India
Harshasupuri123@gmail.com

***Abstract*— In today's highly visual digital landscape, images have emerged as a common mode of communication across social media platforms. However, attaching meaningful and contextually relevant captions to these images remains a significant challenge. This paper presents an automated image caption generator designed to overcome these hurdles by leveraging the power of deep neural networks. The proposed approach employs a hybrid architecture that combines convolutional neural networks (CNNs) and recurrent neural networks with long short-term memory (LSTM) units. CNNs, specifically the VGG16 and ResNet-50 models, are utilized for their proficiency in extracting salient visual features from images. The Transformer architecture, widely used in natural language processing, is compared against these CNNs for its attention-based mechanisms and language modelling capabilities. LSTMs facilitate the generation of coherent and linguistically accurate textual descriptions. The system demonstrates an excellent performance across standard evaluation metrics like BLEU, METEOR and ROUGE on the Flickr8K dataset. Extensive experiments conducted on the Flickr8K dataset validate the efficacy of the system in capturing the essence of diverse images and producing relevant captions. The modular design and straightforward training methodology enable convenient deployment across various platforms and applications. This work introduces an adaptable framework poised to revolutionize visual storytelling by bridging the gap between imagery and text in our increasingly interconnected visual world. The automated image caption generator represents a significant stride towards enhancing communication and understanding through the seamless fusion of visual and linguistic modalities.**

Keywords—Image Captioning, Deep Learning, Natural Language Processing, Computer Vision, Multimodal Learning, Flickr8K, BLEU, METEOR, ROUGE.

I. INTRODUCTION

In today's highly visual digital landscape, images have emerged as a ubiquitous mode of communication across social media platforms and various domains. However, attaching meaningful and contextually relevant textual descriptions to these images remains a significant challenge. Automated image captioning aims to bridge this gap by generating accurate and coherent captions that can effectively convey the essence of visual content. Recent years have witnessed a surge of research efforts dedicated to developing advanced image captioning systems leveraging the power of deep learning and artificial intelligence. These endeavours have explored diverse techniques and architectures, each with its own strengths and limitations, in pursuit of generating human-like captions that can enhance visual understanding and communication.

One prominent line of research has focused on exploiting transformer models, which have demonstrated remarkable success in natural language processing tasks. [1] Many researchers investigated the use of pretrained Vision Transformers (ViT) and text transformers for image captioning in the Thai language, highlighting the potential of attention-based architectures in capturing visual and nuances. Another promising approach has been the integration of capsule networks and transformer neural networks [2]. They introduced a novel method that combines these architectures to capture spatial and geometrical attributes of objects, aiming to produce more detailed and meaningful captions, particularly regarding positional and orientational details. Traditional deep learning techniques, such as convolutional neural networks (CNNs) and recurrent neural networks (RNNs) with long short-term memory (LSTM) units, have also been extensively explored in the image captioning domain. [3] Such models were employed in conjunction with beam search and argmax algorithms, demonstrating their effectiveness in generating captions for datasets like Flickr8k.

Despite the significant progress made in image captioning, several challenges and limitations remain, such as caption diversity, dataset bias, deployment constraints, and the ability to capture intricate visual and contextual nuances effectively. Ongoing research efforts aim to address these limitations, paving the way for more robust, generalizable, and practical image captioning systems that can revolutionize visual communication and enhance our understanding of the visual world around us.

II. LITERATURE REVIEW

The literature review examines numerous research papers that apply deep learning techniques like CNNs, RNNs, and transformers to the task of image and video captioning. It encompasses studies that aim to enhance caption quality, address specialized domains like medical imaging and remote sensing, and tackle novel challenges in captioning diverse visual content.

A. *Enhancing Caption Quality with Advanced Architecture*

Papers like [1] and [2] explore improving caption quality using cutting edge architectures like vision transformers and capsule networks. These architectures are designed to capture minute spatial, geometrical, and positional details of objects within photos. By integrating local and global multimodal interaction techniques, These methods seek to produce captions that offer deeper insights into visual contents. This will help with applications that range from assistive technologies to automated image description.

B. *Specialized Image Captioning Domains*

Some papers focus on specialized captioning domains. For instance [1] for annotating images in Thai language that takes into account cultural context and linguistic ease. Meanwhile [5] and [14] center to medical images like X-rays/retinal scans to assist doctors through specialized captioning models. Similarly [6], [9] for remote sensing imagery using joint-training and multiscale multi-interaction networks to enhance captioning accuracy in aerial and satellite imagery applications.

C. *Development in Human-Like and Diverse Caption Generation*

[4] explores using synthetic images from Generative Adversarial Networks for data augmentation to improve image captioning. By incorporating these techniques captioning systems resilience and generalizability can be improved using synthetic data, which makes it possible for them to handle a wider visual scenarios more effectively.A few works [12], [15] introduce novel problem settings like zero-shot captioning of novel objects without extra training data, and captioning images with high noise like rainy scenes by incorporating denoising modules. The mentioned study endeavours enhance the effectiveness and flexibility of captioning solutions in complex visual situations.

D. *Data Augmentation and Novel Problem Settings*

Papers like [1] and [2] explore improving caption quality using cutting edge architectures like vision transformers and capsule networks. These architectures are designed to capture minute spatial, geometrical, and positional details of objects within photos. By integrating local and global multimodal interaction techniques, These methods seek to produce captions that offer deeper insights into visual contents. This will help with applications that range from assistive technologies to automated image description.

E. *Dense Video Captioning*

For video captioning, [11] proposes a gradual approach for dense captioning of untrimmed videos by localizing and describing multiple events while avoiding overlapping regions. These advancements not only contribute to improved video understanding but also enable applications like automated video summarization and content indexing.

F. *Evolution of Image Captioning Research and Evaluation Metrics*

Other papers like [8] analyse the evolution of image captioning research over the past decade using visualization tools. They track the development of captioning models over time, showcasing patterns and breakthroughs in the industry. Commonly used evaluation metrics include BLEU [18], METEOR [19] and ROUGE [20] to assess caption quality and efficacy[3] acting as standards for comparing various methods and directing the course of future study .

G. *Research Gaps*

The papers identify persisting research gaps such as the need for larger, more diverse datasets [16], better capturing of spatial/geometrical understanding, effective cross-domain adaptation, multimodal data fusion, and improving lexical diversity of captions. Challenges also remain in interpreting and ensuring reliability of captions, addressing privacy for applications like dietary monitoring [13], and scaling to long- term real-world deployment scenarios. In order to advance the useful applications and public impact of image and video captioning systems, it will be imperative to address these problems.

Overall, deep learning has enabled significant progress in image and video captioning, but further research is needed to develop robust, scalable systems that can generate high- quality, diverse captions tailored to different domains and deployment scenarios while ensuring reliability, privacy, and trustworthy performance.

III. METHODOLGY

The proposed methodology for image caption generation is based on a deep learning framework that integrates advanced computer vision and natural language processing techniques to produce accurate and context-aware captions. The system begins with image preprocessing, where input images are resized, normalized, and prepared for feature extraction.

A MobileViT model as shown in fig1 is then employed as the visual encoder to extract both local and global features from the image, providing a rich and compact feature vector representation. Unlike traditional CNN models, MobileViT effectively captures spatial relationships and contextual information, improving the understanding of complex scenes.

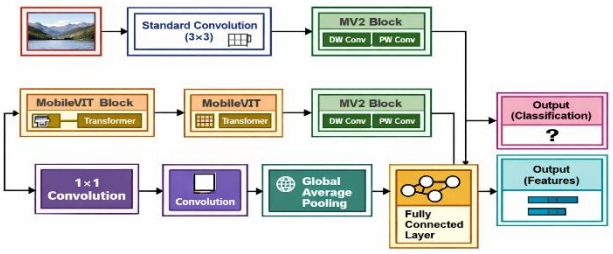


Fig1 : Architecture Of Mobile-Vit

The extracted feature vector is passed to a sequence modeling component, where a Long Short-Term Memory (LSTM) network acts as the decoder. The LSTM processes the visual features and generates captions word by word by learning sequential dependencies in language. During training, the model uses paired image-caption data from the Flickr8K dataset, enabling it to learn the relationship between visual content and corresponding textual descriptions. Techniques such as tokenization, padding, and word embedding are applied to convert textual data into a suitable format for the model.

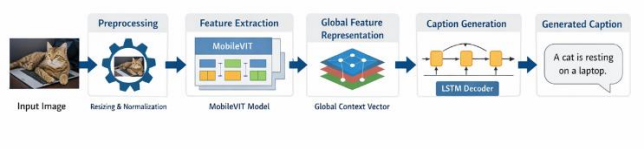
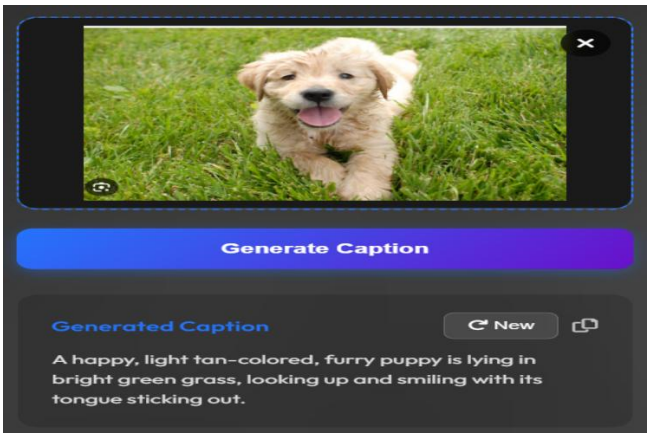


Fig 2 : Working Procedure

The training process as shown in fig2 is carried out using an optimization algorithm to minimize the loss between predicted and actual captions, improving the model's accuracy over time. During inference, the trained model generates captions by predicting the next word in a sequence until a complete sentence is formed. The overall architecture ensures efficient feature extraction, effective language modeling, and improved caption quality. This proposed system enhances performance by generating more meaningful, human-like, and context-aware captions compared to existing methods, while maintaining scalability and real-time applicability.

III . Result Analysis

The results of the proposed image caption generation system were evaluated using the Flickr8K dataset by analyzing both quantitative metrics and qualitative outputs. The generated captions demonstrate that the model is capable of producing meaningful, grammatically correct, and context-aware descriptions of images. Compared to the existing CNN–LSTM approach, the proposed MobileViT–LSTM model shows improved performance in capturing both local features and global context, resulting in more descriptive and relevant captions.



The evaluation metrics such as BLEU, METEOR, and ROUGE indicate a noticeable improvement in caption quality, where higher scores reflect better similarity between generated captions and human-written references. The comparison of generated captions reveals that advanced architectures produce more detailed and precise descriptions. For example, captions generated by improved models include additional attributes such as object color, action, and surrounding context, whereas simpler models tend to produce shorter and more generic sentences.

The training and validation loss graph further supports the effectiveness of the model, showing a consistent decrease in loss values, which indicates proper learning and minimal overfitting. The system is able to generalize well on test images and generate human-like captions with good coherence. However, minor limitations are observed in handling rare or highly complex images, where the model occasionally produces less accurate descriptions due to limited dataset diversity.

Overall, the results confirm that the proposed methodology significantly enhances caption generation performance in terms of accuracy, contextual understanding, and sentence quality. This makes the system suitable for practical applications such as automated image description, social media captioning, accessibility tools for visually impaired users, and content management systems.

Evaluation metrics

	EXISTING METHOD	PROPOSED METHOD
SCORES	VGG16	MOBILE-VIT
BLEU-1 SCORE	0.529504	0.7746
BLEU-2 SCORE	0.299284	0.5934
BLEU-3 SCORE	0.215704	0.4122
BLEU-4 SCORE	0.107782	0.3112

1) Bilingual Evaluation Understudy (BLEU-1, BLEU-2, BLEU-3, BLEU-4)

These metrics measure the overlap between the predicted and ground truth n-grams, evaluating the precision of the generated captions. The BLEU value ranged from 0 (min) to 1 (max). The higher value was better.

2) Metric for Evaluation of Translation with Explicit Ordering (METEOR)

This metric compares the similarity between the generated and reference captions based on matched unigrams and their synonyms. Unlike BLEU-1, METEOR takes into account the uni-gram precision and the uni-gram recall.

3) Recall-Oriented Understudy for Gisting Evaluation for Longest Common Subsequence (ROUGE-L)

This metric measures the overlap between the longest common subsequence of words between the generated and reference text.

The system achieved strong results across BLEU-1, BLEU-2, BLEU-3, BLEU-4, METEOR, and ROUGE scores,indicating its ability to produce captions containing salient objects, relationships, and attributes present in the images. Some example predictions illustrate descriptive sentences with no obvious repetitions or flaws. However, the model struggles with rare images that lack coverage.

IV . Conclusion

In this work, an efficient image caption generation system using deep learning techniques has been developed to automatically generate meaningful and context-aware descriptions for images. The proposed model integrates MobileViT for effective feature extraction and an LSTM-based decoder for sequential language generation, enabling the system to capture both visual details and linguistic structure. The use of the Flickr8K dataset for training and evaluation ensures that the model learns diverse visual patterns and corresponding textual representations. Experimental results demonstrate that the proposed approach outperforms traditional CNN–LSTM methods by generating more accurate, descriptive, and human-like captions, as reflected in improved evaluation metrics such as BLEU, METEOR, and ROUGE. Although the system performs well in most cases, it still faces limitations when dealing with highly complex or unseen images. Future

improvements can focus on incorporating attention mechanisms, transformer-based models, and larger datasets to further enhance caption quality and robustness. Overall, the proposed system successfully bridges the gap between visual content and natural language, making it suitable for various real-world applications such as accessibility tools, social media, and automated content management.

VI . Future Scope

The proposed image captioning system represents a significant step towards bridging the gap between visual and linguistic modalities. However, there remain numerous avenues for further research and improvement, unlocking the potential for even more advanced and impactful applications.

One crucial direction lies in the expansion and diversification of training datasets. Incorporating larger and more varied collections of images can enhance the model's ability to generalize across a broader range of scenarios, reducing bias and improving caption quality. Additionally, integrating external knowledge sources, such as semantic knowledge bases or commonsense reasoning systems, could provide valuable context, enabling the generation of more nuanced and meaningful captions.

Advancements in multimodal fusion techniques and architectural enhancements, such as attention mechanisms and external memory components, present promising avenues for exploration. Such innovations may lead to more effective ways of combining visual and linguistic information, selectively focusing on the most relevant aspects of input images, and maintaining long-term context, ultimately enhancing caption relevance, coherence, and robustness to rare or unseen scenarios.

Furthermore, efforts towards domain adaptation, transfer learning, and the development of new evaluation metrics and human studies will be instrumental in promoting the scalability, adaptability, and real-world applicability of the image captioning technology. Ensuring accessibility through user-friendly interfaces and catering to individuals with visual impairments will be crucial for widespread adoption and practical impact across various domains, including assistive technologies, augmented reality, and enhanced visual communication.

References

[1] Jaknamon, T., Marukatat, S.: ThaiTC: Thai Transformer-based Image Captioning. In: 2022 17th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP), pp. 1–4. IEEE (2022)

[2] J Yu, Z., Fu, K., Jin, H., Bai, J., Zhang, H., Li, Y.: Local and Global Multimodal Interaction for Image Caption. In: 2023 4th International Conference on Electronic Communication and Artificial Intelligence (ICECAI), pp. 164–169. IEEE (2023) Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)

[3] Shambharkar, P.G., Kumari, P., Yadav, P., Kumar, R.: Generating caption for image using beam search and analyzation with unsupervised image captioning algorithm. In: 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 857–864. IEEE (2021)

[4] Hossain, M.Z., Sohel, F., Shiratuddin, M.F., Laga, H., Bennamoun, M.: Text to image synthesis for improved image captioning. IEEE Access 9, 64918–64928 (2021)

[5] Park, H., Kim, K., Park, S., Choi, J.: Medical image captioning model to convey more details: Methodological comparison of feature difference generation. IEEE Access 9, 150560–150568 (2021)

[6] Ye, X., Wang, S., Gu, Y., Wang, J., Wang, R., Hou, B., Jiao, L.: A joint-training two-stage method for remote sensing image captioning. IEEE Trans. Geosci. Remote Sens. 60, 1–16 (2022)

[7] Kastner, M.A., Umemura, K., Ide, I., Kawanishi, Y., Hirayama, T., Doman, K., Satoh, S.I.: Imageability-and length-controllable image captioning. IEEE Access 9, 162951–162961 (2021)

[8] Liu, W., Wu, H., Hu, K., Luo, Q., Cheng, X.: A Scientometric Visualization Analysis of Image Captioning Research From 2010 to 2020. IEEE Access 9, 156799–156817 (2021)

[9] Wang, Y., Zhang, W., Zhang, Z., Gao, X., Sun, X.: Multiscale multiinteraction network for remote sensing image captioning. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 15, 2154–2165 (2022)

[10] Bae, J.W., Lee, S.H., Kim, W.Y., Seong, J.H., Seo, D.H.: Image captioning model using part-of-speech guidance module for description with diverse vocabulary. IEEE Access 10, 45219–45229 (2022)

[11] Choi, W., Chen, J., Yoon, J.: Step by Step: A Gradual Approach for Dense Video Captioning. IEEE Access (2023)

[12] Wu, Y., Jiang, L., Yang, Y.: Switchable novel object captioner. IEEE Trans. Pattern Anal. Mach. Intell. 45(1), 1162–1173 (2022)

[13] Qiu, J., Lo, F.P.W., Gu, X., Jobarteh, M.L., Jia, W., Baranowski, T., Lo, B.: Egocentric image captioning for privacy-preserved passive dietary intake monitoring. IEEE Trans. Cybern. (2023)

[14] Huang, J.H., Wu, T.W., Yang, C.H.H., Worring, M.: Deep Context-Encoding Network for Retinal Image Captioning. arXiv preprint arXiv (2021)

[15] Krisna, A., Parihar, A.S., Das, A., Aryan, A.: End-to-End Model for Heavy Rain Image Captioning. In: 2022 4th International Conference on Advances in Computing, Communication Control and Networking (ICAC3N), pp. 1646–1651. IEEE (2022)

[16] Flickr 8k Dataset. Kaggle, https://www.kaggle.com/datasets/adityajn105/flickr8k (2020)

[17] Ariyadi, M.R.N., Pribadi, M.R., Widiyanto, E.P.: Unmanned Aerial Vehicle for Remote Sensing Detection of Oil Palm Trees Using You Only Look Once and Convolutional Neural Network. In: 2023 10th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), pp. 1–6. IEEE (2023)

[18] Papineni, Kishore & Roukos, Salim & Ward, Todd & Zhu, Wei Jing. (2002). BLEU: a Method for Automatic Evaluation of Machine Translation. 10.3115/1073083.1073135.

[19] Lavie, Alon & Agarwal, Abhaya. METEOR, "An automatic metric for MT evaluation with high levels of correlation with human judgments'. 228-231(2007)

[20] Lin, Chin-Yew. Rouge: A Package for Automatic Evaluation of summaries. Proceedings of the ACL Workshop: Text Summarization Braches Out 2004. 10. (2004)

[21] Yousuf, U., Singh, R.P., Mehra, M.: Caption Generation of Images Using CNN and LSTM. International Journal of Innovative Research in Engineering & Management 9(1), 1-5 (2022)