Pertanika Proceedings 1 (5): 25 - 28 (2025)

PERTANIKA JOURNALS

PERTANIKA PROCEEDINGS

Journal homepage: http://www.pertanika.upm.edu.my/

A Comparative Analysis of Recent Advances in Deep Learning for Image Captioning

Bhargavi Polepalli*, Praveen Kumar Sekharamantry and Konda Srinivasa Rao

Department of CSE, GST, GITAM (Deemed to be University), Gandhi Nagar, Rushikonda, Visakhapatnam-530045, Andhra Pradesh, India

ABSTRACT

Image captioning uses visual perception combined with natural language processing to increase accessibility and support uses including automated monitoring. Transformer models are among the sophisticated methods produced by conventional approaches grounded on recurrent neural networks (RNNs) and convolutional neural networks (CNNs). These models increase the contextual accuracy and grammatical consistency of output captions since they are known to control long-range dependencies. Emphasizing transformer-based developments, this article evaluates and compares modern deep learning methods for image captioning via a comprehensive analysis. To improve caption production, the work investigates models combining Swin Transformers, diffusion models, and depth fusion. Comparative measures, such as BLEU, ROUGE, METEOR, and CIDEr, are used to demonstrate the performance benefits made by these approaches over conventional models, with a focus on efficiency. This paper demonstrates the considerable influence of recent deep learning algorithms on image quality and utility.

Keywords: Deep learning methods, evaluation metrics, image captioning, transformer-based methods

INTRODUCTION

Image Captioning (IC) in AI poses a substantial challenge that integrates visual comprehension and natural language processing. Captioning systems that generate descriptive text for images improve accessibility for visually impaired individuals and facilitate applications including automated surveillance, content indexing, and interactive robotics. Despite great progress, the work facing challenge due to the complexity of effectively processing visual input and producing contextually relevant captions that are syntactically and semantically correct.

ARTICLE INFO

Article history:
Received: 29 August 2025
Published: 31 October 2025

DOI: https://doi.org/10.47836/pp.1.5.005

E-mail addresses:
bpolepal@gitam.in (Bhargavi Polepalli)
psekharm@gitam.edu (Praveen Kumar Sekharmantry)
skonda@gitam.edu (Konda Srinivasa Rao)
* Corresponding author

e-ISSN: 3083-9475

Recently, there has been an increase in the use of deep learning approaches to address these difficulties, with academics suggesting numerous novel methods to improve caption accuracy and contextual sensitivity. Usually depending on CNNs for visual feature extraction and RNNs or transformers to create textual descriptions. This study aims to deliver a thorough assessment and comparison of the latest deep learning techniques for image captioning produced in the last three to five years.

RECENT ADVANCEMENTS

Based on the technologies and techniques in the current use, Figure 1 shows a classification of deep learning-based methodologies applied in IC into several classes. Generally, the techniques fall under Attention-based, Graph-based, Convolutional network-based, and approaches include Unsupervised Learning and Reinforcement Learning. Further divisions of attention-based techniques are those using transformer architectures and those creating multi-style captions; some use graph data for improved contextual understanding. Graph-based techniques are observed to be useful for better representing relationship data between objects in images both with and without attention processes. By using many facets of deep learning technology, these approaches-including Vision Language Pretraining-enchance the relevance, accuracy, and detail of generated captions.

Transformers use self-attention mechanisms to weigh the significance of caption words, making them powerful tools for logical and contextually relevant image descriptions. Key transformer-based techniques in image captioning are compiled in Table 1 together with their creative aspects and methods. Referring to the important contributions from recent studies, it shows how each technique improves the processing and interpretation of visual material for more accurate and context-aware caption production.

In Figure 2 the recent research is compared utilizing metrics such as BLEU, ROUGE, METEOR, and CIDEr demonstrates that Mahmoud et al.'s pre-trained BERT model showcases the advantages of employing advanced NLP techniques in IC, outperforming

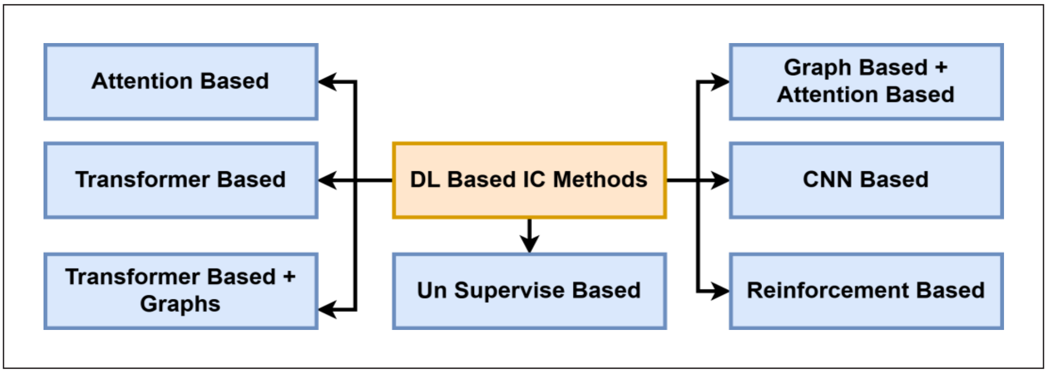

Figure 1. Taxonomy of IC methods in recent literature

most metrics. With the original Xu et al. (2015) scoring the lowest, intermediate performances are shown in the TensorFlow (2022) and Desai and Johnson (2021) models, therefore indicating notable field overtime developments. This visualization emphasizes how well contemporary deep learning approaches increase contextual correctness and caption quality.

Table 1
Recent advancements in the IC

Author	Model/Approach	Features and Innovations
Liu et al. (2020)	GLIED	Captures spatial and relational image region and attribute word groupings from cross-modal data.
Nguyen et al. (2022)	GRIT	Uses a DETR detector with grid and region-based features for better IC.
Wang et al. (2022)	Pure Transformer-based Model	Utilizes SwinTransformer for image grid feature extraction and employs a refining encoder with multi-head self-attention (MSA) to enhance these features for caption generation.
Luo et al. (2022)	Diffusion Transformer	Employs a diffusion model for generating captions, focusing on improving sentence generation from visual inputs by conditioning the diffusion process on semantic priors.
Ahmed et al. (2023)	Depth Fusion Transformer	Enhances image captioning by integrating depth information with RGB images using a transformer- based encoder-decoder model to address 3D scene descriptions.
Basak et al. (2024)	TICOD	Combines tasks of image captioning and object detection into a single Transformer-based framework, leveraging a shared feature extraction stage to optimize performance.

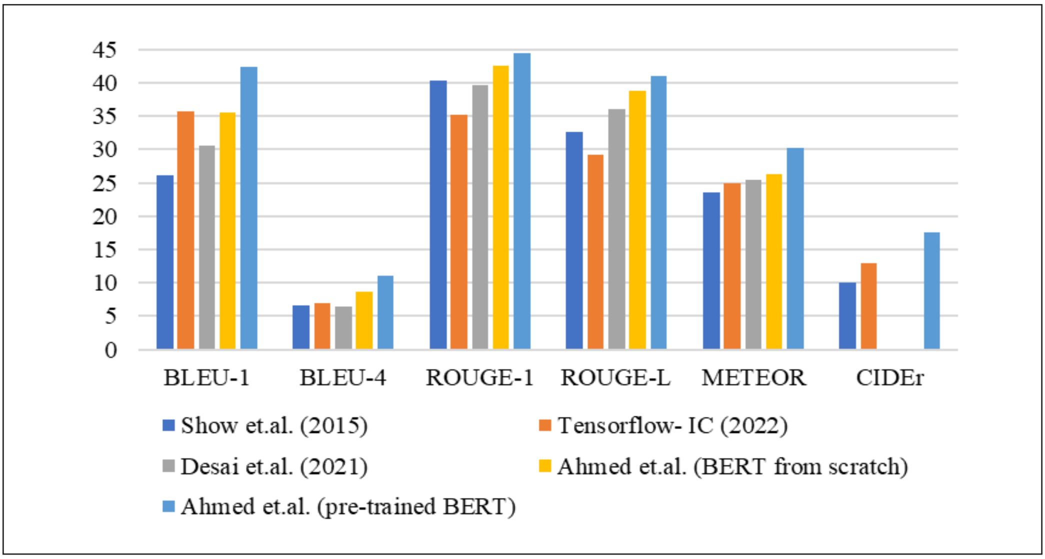

Figure 2. The performance comparison of IC methods

CONCLUSION

The study highlighted the major developments in image captioning enabled by deep learning technologies—especially transformer-based models. Integrating cutting-edge technologies like pre-trained BERT models and SwinTransformers showed successful and established new benchmarks for accuracy across evaluation criteria. Future study will investigate optimizing Enhanced Image Captioning with SST-CNN and fine-tuned BERT.

ACKNOWLEDGEMENT

The authors sincerely acknowledge the support, encouragement, and valuable contributions of all individuals and institutions who assisted us throughout the course of this research.

REFERENCES

Ahmed, A. M., Yousef, M., Hussain, K. F., & Mahdy, Y. B. (2023). Enhancing image captioning with depth information using a Transformer-based framework. *arXiv preprint arXiv:2308.03767*. https://doi.org/10.48550/arXiv.2308.03767

Basak, D., Srijith, P. K., & Desarkar, M. S. (2024). Transformer based multitask learning for image captioning and object detection. In D. Yang, X. Xie, V. S. Tseng, J. Pei, J. W. Huang & J. C. Lin (Eds.) *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 260-272). Springer. https://doi.org/10.1007/978-981-97-2253-2_21

Desai, K., & Johnson, J. (2021). Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11162-11173). IEEE.

Liu, F., Ren, X., Liu, Y., Lei, K., & Sun, X. (2020). Exploring and distilling cross-modal information for image captioning. *arXiv preprint arXiv:2002.12585*. https://doi.org/10.48550/arXiv.2002.12585

Luo, J., Li, Y., Pan, Y., Yao, T., Feng, J., Chao, H., & Mei, T. (2023). Semantic-conditional diffusion networks for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 23359- 23368). IEEE.

Nguyen, V. Q., Suganuma, M., & Okatani, T. (2022). Grit: Faster and better image captioning transformer using dual visual features. In S. Avidan, G. Brostow, M. Cisse, G. M. Farinella & T. Hassner (Eds.) *European Conference on Computer Vision* (pp. 167-184). Springer. https://doi.org/10.1007/978-3-031-20059-5_10

TensorFlow. (2022). *Image captioning with visual attention*. TensorFlow. https://www.tensorflow.org/tutorials/text/image_captioning

Wang, Y., Xu, J., & Sun, Y. (2022). End-to-end transformer based model for image captioning. *Proceedings of the AAAI Conference on Artificial Intelligence,* 36(3), 2585-2594. https://doi.org/10.1609/aaai.v36i3.20160

Xu, K., Ba, J., Kiros, R., Cho, K., Vourville, A., Salakhutdinov, R., Zamel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *arXiv:1502.03044.* https://doi.org/10.48550/arXiv.1502.03044