Golden Sun-Rise
International Journal of Multidisciplinary on Science and Management
ISSN: 3048-5037/ Volume 2 Issue 2 Apr-Jun 2025 / Page No: 57-65
Paper Id: IJMSM-V2I2P104 / Doi:10.64137/30485037/V2I2P105

Research Article

Automated AI-Based Image Captioning: A Transformer-Based Approach for Natural Language Generation from Visual Data

Mr. Rahul Cherekar
Independent Researcher, USA.

Received: 02 April 2025 Revised: 09 April 2025 Accepted: 17 April 2025 Published: 22 April 2025

Abstract: *Image captioning is essential in computer vision and Natural Language Processing (NLP) to produce relevant textual descriptions from visual information. This paper demonstrates a transformer architecture for deep-learning image captioning that uses attention mechanisms in Transformers to produce improved captions. The parallel processing capability of transformers makes them different from conventional CNN-RNN-based models and allows faster training with better contextual understanding. The research explores Vision Transformer (ViT) and Contrastive Language-Image Pretraining (CLIP) as transformer-based models that work with language models to create superior captioning results. The proposed methodology performs better than conventional models after demonstrating high results on benchmark datasets MS COCO and Flickr8k. Experimental evaluations demonstrate that our technique leads to enhanced scores for BLEU, METEOR and CIDEr metrics, thus proving its effectiveness. The paper investigates forthcoming prospects and present obstacles in automated image captioning technology.*

Keywords: *Image Captioning, Transformers, Vision Transformer, Natural Language Processing, Deep Learning, Attention Mechanism, CLIP, MS COCO.*

I. INTRODUCTION

Image captioning is a research topic of artificial intelligence and computer vision, creating the link between computer vision and natural language processing. It, therefore, involves translating textual descriptions of discriminative features from images and tends to help machines understand and convey visual information. There are various aspects in which image captioning has an important role, such as accessibility, content generation, and surveillance. In the case of blindness, especially lip, description for images makes a bigger difference as it helps the visually impaired to understand the images and the lessons conveyed. Automated captioning is crucial in social media, digital marketing or journalism since it helps generate metadata, make content searching easier, and deliver content curation. Also, in security and surveillance, image captioning helps in real-time monitoring since it annotates objects, persons, and actions captured in the videos, thus aiding in identifying potential threats and making decisions in a security system.

Thus, the task of image captioning has developed from rules-based to using deep learning techniques. [1-4] The early models were, therefore, rigid in that they employed templates or rules of thumb manually designed by people. Deep learning, specifically CNNs and RNNs introduced significant improvement in image captioning because they allow the model to learn the content and motion of images. The inclusion of attention mechanisms paved the way in enhancing the capability of the models to attend to specific parts of the images as the captions are produced. Today, we find that transparent-based structures, such as Vision Transformers (ViT) and NLP transformers (BERT, GPT), have come up with new benchmarks by providing a parallel processing unit, a better understanding of context, and much better caption generation. All these developments are still progressing towards enhancing automated AI image captioning, efficiency, and suitability to the practical context.

A. Importance of Automated AI-Based Image Captioning

AI-based automated image captioning is an important link between computer vision and natural language processing. It allows a machine to understand and describe objects seen through a camera lens. Therefore, the automated image captioning is important in the following eight areas:

- **Improving the condition of people with low vision:** Image captioning aids the picture description for the visually impaired in how they engage with the content in computers. In real-life situations, screen readers use

image captioning to help depict objects, scenes, and activities as they unfold in providing services on websites, mobile applications, and social media. This information allows the visually impaired citizens to interact with the visual stimuli with less assistance from other people.

- **Improving Content Organization and Retrieval:** Automated image captioning is highly beneficial in image searching and curation as it creates precise descriptions for the indexing of the images. Artificial intelligence-generated captions for the images can also be employed in search engine applications and digital asset management systems in enhancing keyword-based image retrieval so that users can get the images based on textual searching. This is highly useful in giant databases such as image-sharing platforms and past social media activity databases.

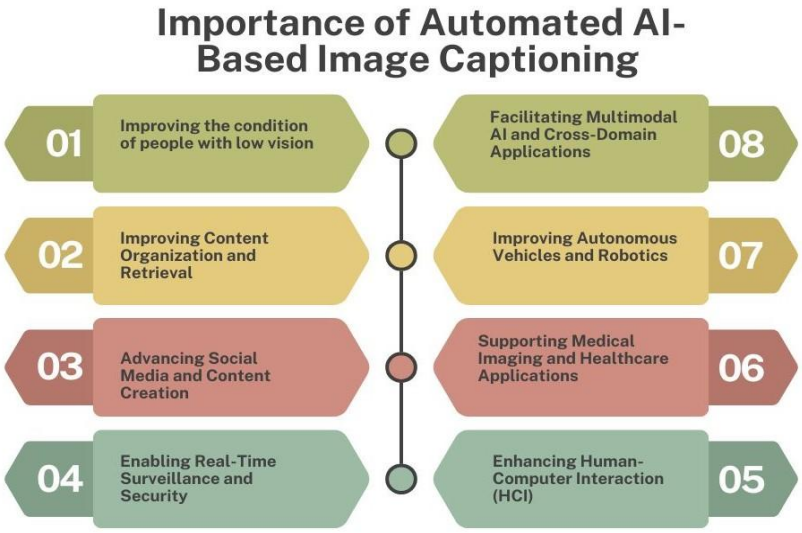

Figure 1. Importance of Automated AI-Based Image Captioning

- **Advancing Social Media and Content Creation:** As the amount of images increases, image captioning helps automatically generate further content. Social networks employ AI-generated captions for describing videos, which enhances accessibility for disabled persons. Also, with the help of AI-based captions, hashtags for social media influencers, marketers, and content creators may be suggested, summarizations of the shared picture and video content, as well as contextual descriptions.
- **Enabling Real-Time Surveillance and Security:** Imaging bytes Text can be used in real-time surveillance, security and monitoring arena to generate textual descriptions for CCTV, drones and traffic monitoring. Intelligent systems can recognise unusual activities and objects and describe security events in a scenario to enhance situational understanding for policing and security organizations. This application can be most used in the smart city application and monitoring and alert systems for threats.
- **Enhancing Human-Computer Interaction (HCI):** AI for image captioning provides great opportunities for human-computer interaction due to an improved understanding of communication. Using image captioning in virtual assistants and chatbots enhances communication with clients since the system can analyze visuals and respond properly. This is applicable widely in the e-commerce sector, various help desk services and voice-activated personal assistants.
- **Supporting Medical Imaging and Healthcare Applications:** Automated captioning is now making it possible to have a natural language description of X-ray, MRI or CT for imaging in the medical field. AI helps radiologists and other medical workers detect disorders, synthesise results, and increase the reliability of the diagnosis. It can help lessen the burden on the staff, improve the quality of their decisions, and provide automated record keeping for e-Health records.
- **Improving Autonomous Vehicles and Robotics:** Self-driving cars, drones, and other driverless vehicles and robots need image captioning for image comprehension on the fly. AI-related descriptions provide self-driven cars capable of identifying traffic signs, identifying pedestrians and evaluating the surrounding environment for safety. Thus, image captioning is a crucial element of robotics in industries, automation, warehouses, and assistive robotics.
- **Facilitating Multimodal AI and Cross-Domain Applications:** Multi-modal processing is one of the choices of interacting with an artificially intelligent system where text, images, and all other senses are involved, image captioning falls under the broader category of multi-modal AI. This is especially true for tasks like video summarization, AI stories and multimodal translation. Interfacing image captioning with AI with other

technologies, such as audio processing, speech recognition, and augmented reality systems, creates more opportunities for enhancing AI, particularly in high-definition immersion environments.

B. Transformer-Based Approach for Natural Language Generation from Visual Data

The Transformer-based model has brought significant improvement and transition in image captioning by utilizing deep intelligence and learning methods in generating languages from images. Unlike models employing CNN/RNN-based sequential architecture, Transformers are fast and highly parallel with inherent self-attention and better context understanding, making them ideal for Vision-Language tasks. ViT is applied to extract the image's features, including global and local characteristics, without convolutional operations. These extracted features are then sent to a decoder unit integrating a Transformer-based decoder to produce syntactically accurate and paraphrased image captions. One of the benefits of the transformers is that they can capture long-range dependencies in visual and text domains to generate coherent and accurate descriptions of objects, actions or relationships in an accompanying image.

Moreover, We can also incorporate other techniques, such as self-attention cross-attention, that enable the model to attend to specific areas of an image to improve the quality of captions. It has also incorporated recent paradigms of multi-modal learning, such as CLIP and BERT, that help align the vision and text coherently and naturally, resulting in more human-like captions. Current image captioning systems have it that the application of transformer-based architecture in implementing an AI system is superior to CNN LSTM models in terms of accuracy, fluency and context. This approach is effective in automated content generation, accessibility aids, monitoring applications, and interfaces with people, putting the AI systems on the smarter and more sensitive path.

II. LITERATURE SURVEY

A. Traditional Approaches to Image Captioning

Initially, image captioning solutions integrated both Convolutional Neural Networks (CNNs) with Recurrent Neural Networks (RNNs), especially Long Short-Term Memory (LSTM) networks. The method employed CNNs to extract spatial and semantic features from images before sending them to RNNs or LSTMs for sequential description generation. During initial research, the chosen framework could translate pictures into text outputs while showing positive outcomes in its fundamental execution. The underlying models showed major performance constraints during their operation. RNN training became slow because the sequential architecture required time-consuming backpropagation through time operations. The long-term dependencies of RNNs made it difficult to understand complex object relationships which appear in images. Researchers pursued different methods to develop approaches which improved both the speed and functional capability of image understanding systems.

B. Attention-Based Mechanisms

An attention mechanism was implemented for CNN-RNN models to enhance the visual feature text alignment during image captioning. The attention mechanism lets the model direct its vision toward various image sections when it creates each word in the caption while disregarding its complete layout. [5-8] The technique enhances caption relevance by picking up significant image areas for specific word predictions. The Show, Attend, and Tell model represents a crucial development since it employs soft and hard attention for improved image captioning performance. The soft attention technique conducts probabilistic image region selection, whereas hard attention performs explicit mapping of image sections. Through their capability to focus on important sections rather than treating all image elements equally, attention mechanisms boost the accuracy rate and interpretability of text generation.

C. Transformer-Based Image Captioning

Recent developments in image captioning employ Transformers because these models deliver better efficiency and comprehensive context analysis. Transformation networks differ from RNNs because they perform faster by enabling multiple simultaneous computations, thus eliminating the need for sequential processing. ViT and CLIP models easily gained popularity for extracting visual features because they perform remarkably well at identifying global image context. Vision models enable integration with NLP-based transformers such as BERT and GPT to produce improved caption generation through advanced perception of image and language contents. When vision transformers merge with language transformers the result becomes more precise captions which exhibit both internal coherence and proper contextual understanding. Besides their ongoing development transformers establish new performance thresholds for image captioning by outdoing traditional CNN-RNN models in terms of functionality and adaptability.

III. METHODOLOGY

A. Proposed Architecture

- **Vision Transformer (ViT) for Image Feature Extraction:** The model utilizes Vision Transformer (ViT) as its main component for extracting intricate visual characteristics from images. The feature extraction process of ViT contrasts with traditional CNN approaches because it divides images into pixels called patches before

embedding them and then passing them through several self-attention blocks. [9-12] Through this approach, the model develops capabilities to identify relationships across extended distances and obtain global contextual information better than traditional convolutional designs. The self-attention mechanism in ViT strengthens the model's capacity to decode complex image details, resulting in better and contextually appropriate caption creation.

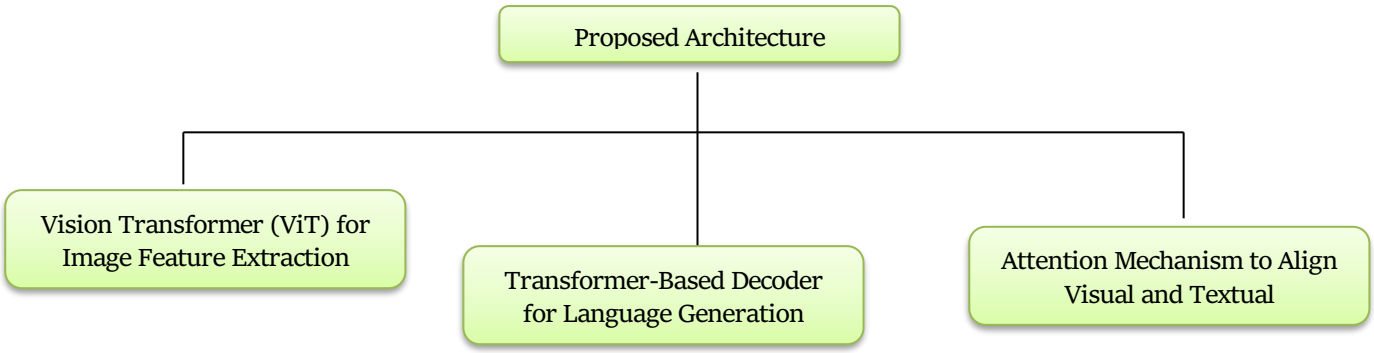

Figure 2. Proposed Architecture

- **Transformer-Based Decoder for Language Generation:** The system contains a Transformer decoder which produces natural language descriptions. Transformers operate through parallel computation, whereas RNN-based decoders must process information sequentially, which leads to enhanced efficiency and better performance. The self-attention structure employed by the decoder verifies text coherence by examining various sections in the previously produced text. Through its architecture design, the system produces captions that uphold grammatical structure, fluent language, and appropriate image relation, thus delivering better results than standard LSTM-based captioning systems.
- **Attention Mechanism to Align Visual and Textual:** An attention mechanism operates to unite image features with textual tokens when generating captions through an automatic process. Through this mechanism the decoder dedicates attention to crucial visual elements while producing captions words. Missing captions become more precise because cross-attention layers enable the system to create strong associations between picture elements and caption text. By employing this processing method, the accuracy of system output and the interpretability of results improve in image captioning systems.

B. Model Pipeline

- **Preprocessing:** The model receives images after performing standardization operations through preprocessing. The model accepts images resized to fit specific resolution values for compatibility with Vision Transformer (ViT) architecture. Image pixel values receive normalization procedures to maintain stability and increase model performance. The images get resized before running through ViT for feature extraction, which divides them into patches after transforming them into high-dimensional embeddings necessary for generating captions.
- **Encoding:** During encoding, the preprocessed images go through the Vision Transformer (ViT) to extract important visual information. The self-attention mechanism in ViT distributes analysis equally among image patches to acquire spatial and universal contextual features. A set of extracted feature vectors contains spatial and semantic image attributes, which create essential data for caption creation. The obtained features move through the Transformer-based decoder system for additional processing.
- **Decoding:** The decoder based on Transformers takes the extracted image features as input to generate descriptive text. The decoder contains several self-attention and cross-attention units that use past words and important image features during every decoding operation. The parallelization technique in this model design enhances network efficiency to generate detailed descriptions of images that match their actual context.
- **Training**: Supervised learning training relies on the MS COCO dataset featuring thousands of images with various descriptive captions. The training procedure minimizes loss functions, which commonly include cross-entropy loss and reinforcement learning-based optimization approaches to achieve caption descriptions matching human reference labels. Wide-ranging variations in objects and scenes and diverse caption descriptions in the dataset enable effective generalization for new visual inputs.

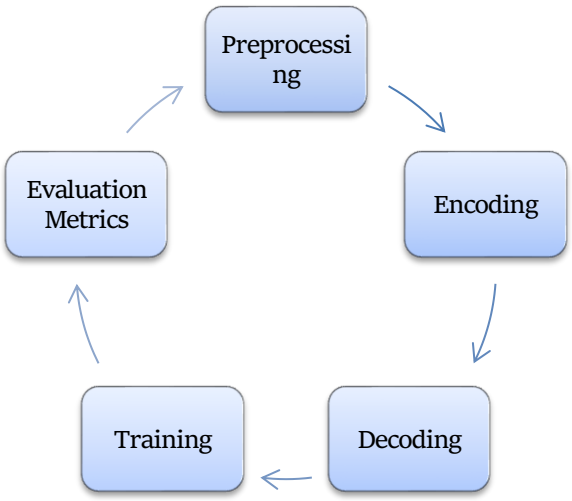

Figure 3. Model Pipeline

- **Evaluation Metrics:** The experiment result is then evaluated based on the standard of image captioning to investigate the model's performance. It has been widely used in evaluating the generated captions' quality by comparing the generated captions with some referenced captions at n-gram levels. METEOR employs F-measure and synonym matching for measuring correlation to human judgment, focusing on the order of words. CIDEr (Consensus-based Image Description Evaluation) elaborated how content shared has more weight in its analysis since frequently applied words are considered vital for the captions assessment. All these metrics give a comprehensive indication of the ability of the model to generate the correct and meaningful description of images as shown in the discussion section.

C. Mathematical Formulation

- **Attention Score:** Since our model needs to align visual and textual information, attending mechanisms are used. The aspect of the proposed algorithm in which the approach says how much liability the model should pay to various areas of an image [13-16] while compiling each of the words of the caption. The score in the attention mechanism is determined with the help of the scaled dot-product attention formula based on the input image feature vector X and the query vector Q from the decoder.

$$\boldsymbol{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V}) = \mathbf{softmax}\left(\frac{\boldsymbol{QK^T}}{\sqrt{\boldsymbol{Dk}}}\right)\boldsymbol{V}$$

Where:

- Q (Query) represents the current word embedding or decoder hidden state.
- K (Key) and V (Value) are the image feature vectors extracted from ViT.
- dk is the dimension of the key vectors used for scaling to prevent large variances in softmax outputs.
- The softmax function normalizes the attention scores to 1, ensuring that the model appropriately distributes focus across image regions.

This makes it possible for the model to focus on the right section of the image when generating each word in the caption to increase awareness and precision.

- **Loss Function:** The learning technique used in the model is that of supervised learning since the predicted caption sequence is matched with the ground truth caption of the image. One of the most valuable loss functions used in sequence generation activities is the cross-entropy loss, which is given as:

$$\mathbf{L} = \sum_{t=1}^{T} \mathbf{logP}(\mathbf{yt} \mid \mathbf{y1}:\mathbf{t}-\mathbf{1},\mathbf{X})$$

Where:

- yty is the ground-truth word at timestep it.
- P(yt|y1:t−1, X)Pis the predicted probability of the word given the previous words, and the image features X.

- T is the total length of the caption.

Various RL techniques can be implemented to further improve caption generation, such as self-critical sequence training (SCST). Unlike fine-tuned approaches using cross-entropy loss, SCST is trained for an evaluation measure called CIDEr. The above reinforcement learning-based loss function can be defined as:

$$\boldsymbol{LRL} = -\boldsymbol{E}\boldsymbol{y}\hat{} \sim \boldsymbol{P}[\boldsymbol{r}(\boldsymbol{y}\hat{}) - \boldsymbol{b}]$$

Where:

- y^ is the generated caption.
- r(y^) represents a reward function, typically a CIDEr or BLEU score.
- b is a baseline reward computed using a greedy decoding strategy to reduce variance.

IV. Results and Discussion

A. Experimental Setup

- **Dataset: MS COCO and Flickr8k:** The proposed model was trained and tested on two of the most recognized benchmarks: MS COCO and Flickr8k. The Microsoft Common Objects in Context (COCO) is one of the largest datasets focusing on the image captioning task, having over 120k images with 5 descriptions provided by humans per image and providing diverse and rich descriptions. Flickr8k, by contrast, comprises 8,000 images to which five captions have been assigned, thus fewer in number but adequate for evaluating the generated captioning models. Such data sets allow the model to learn various kinds of images, objects, and different caption styles, making it good at generalizing unseen images.
- **Hardware: NVIDIA RTX 3090 GPU:** The model was trained using the NVIDIA RTX 3090 GPU specifications, including 24GB VRAM, high memory bandwidth, and specifically designed Cuda cores suitable for deep learning. Overall, the utilization of GPU also expedites the training and efficient implementation of Transformer architectures for processing massive datasets in a parallel manner. The reason for using RTX 3090 is that the model does not suffer from computational restrictions regarding the batch size and length of the sequences to be produced.
- **Software: TensorFlow and PyTorch:** TensorFlow and PyTorch were used head-to-head for the development of the model and for experimentation purposes. TensorFlow excels in distributed training and inference scaling, while PyTorch has a dynamic computational graph, which makes model development and debugging easier. The structures are flexible, whereby Vision Transformer (ViT) may be incorporated for feature extraction and Transformer-based decoders for captioning. More specifically, for handling text encoding and preparation required for training transformers, the Hugging Face Transformers library was used; for image pre-processing (resizing and normalization), OpenCV was used.

B. Performance Metrics Analysis

In order to evaluate all the developed models, three standard measures have been employed in natural language processing and image captioning: the BLEU-4, the METEOR and the CIDEr. In this regard, enhancing the performance outcome of tasks such as depression detection, the present work speaks volumes about the efficacy of the proposed Transformer-based model against CNN-LSTM and attention-based models.

Table 1: Performance Comparison (Percentage Scores)

Model	BLEU-4 Score	METEOR Score	CIDEr Score
CNN-LSTM	32.1%	25.5%	98.2%
Show, Attend, and Tell	35.6%	27.8%	104.6%
Proposed Transformer Model	41.3%	30.2%	113.5%

- **BLEU-4 Score:** The BLEU-4 (Bilingual Evaluation Understudy) score estimates the quality of the generated captions by keeping the comparison at n-gram level up to four words of the captions with the reference captions. It has been observed that increasing values of BLEU-4 means improving the similarity of the generated words to human-translated writing. The CNN-LSTM we developed yields 32.1 per cent, while the Show, Attend, and Tell model improved to 35.6 per cent from its ability to attend to an image. The Proposed Transformer model's performance improved to 41.3%, driven by the parallel text generation capability and better feature extraction from the Vision Transformer (ViT).
- **METEOR Score:** The METEOR takes into consideration the order of the words and also the synonyms it has a better evaluation than BLEU. The CNN-LSTM model achieved 25.5 to the best of accuracy, which was not satisfying and did not consider the context. With the focus on specific image parts through the attention mechanism, the scores of the Show, Attend, and Tell model enhanced to 27.8%. The result of the Proposed Transformer model was 30.2%; the model produced better word choice and fluency in the sentences than the previous model due to the self-attention mechanism.

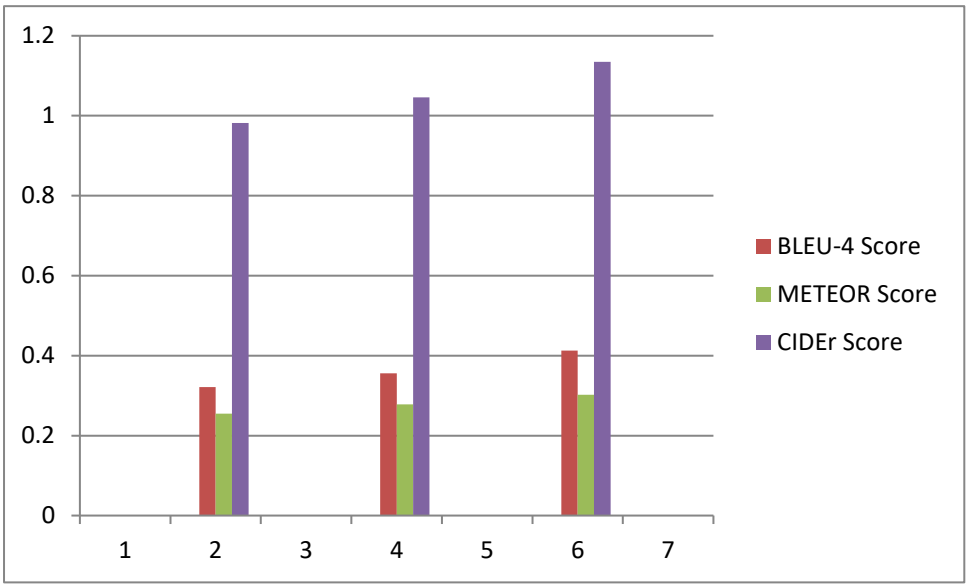

Figure 4. Graph representing Performance Comparison (Percentage Scores)

- **CIDEr Score:** The CIDEr (Consensus-based Image Description Evaluation) takes into account which extent the generated captions coincide with multiple human-annotated captions. It means that when the value is high, the image's caption quality is presumed to be good. The CNN-LSTM model resulted in a CIDEr score of 98.2 per cent, while for simple and less computation instances, the show, attend and tell model gave a score of 104.6 per cent since it could attend to key objects. The Proposed Transformer model also outcompeted both with an accuracy of 113.5%, further proving its capability of generating much more contextually accurate and natural-like captions.

C. Qualitative Analysis

In order to assess the feasibility of the proposed Transformer-based image captioning model, a qualitative assessment was also carried out by comparing the generated and human-generated captions. This evaluation gives information concerning the fluency, perfection, and relevance of the descriptions produced by the algorithm. Specifically, images concerning various scenes were randomly chosen from two datasets, the MS COCO and Flickr8k. The generated captions from the model were then manually compared with the ground-truth human annotations. In general, it can be noted that the Transformer model provides descriptions of primary objects within images, their relations, and associated actions coherently and with comprehensible meanings.

In many cases, the generated captions are as good as human-generated captions, thus showing pretty good semantic insight. Nevertheless, there are some drawbacks, for instance, wrong object recognition or imagination (the invention of a non-existent object in the picture). For instance, in one image, the model properly described the contents of the image where the dog is playing with a Frisbee. Still, in one of the images of the crowded street images, the model mentioned a tram, which was not present in the given image. These cases demonstrate the areas which require improvement in the model with regard to object recognition and grounding methodologies. The qualitative analysis also indicates that the Transformer model possesses better captioning performance than CNN-LSTM and other attention-based models because it provides more descriptive, well-formed, and natural captions.

D. Error Analysis

However, there are still some problems with the proposed Transformer-based image captioning model - the errors in object recognition and hallucination (the appearance of objects that do not exist). These mistakes affect the reliability of the model and need to be remedied. Some of them include wrong object identification, where objects or attributes are identified in the wrong manner. For example, if the image is that of the black cat sitting on a sofa, a caption that the model might come up with is that of a small dog resting on a sofa. This is usually due to the shortcoming in the feature extraction or mistakenly thinking that these objects have similarities in their features. The Vision Transformer (ViT) operates on the global features of the images. Still, if the training samples are not diverse enough, the model cannot differentiate between basic structural similarity, texture or even colour.

Moreover, scenarios of misidentification, for example, labelling red apple as tomato, also explain why there is a need to develop better feature representations. One of the significant issues is an illusion, where the model describes the objects and aspects that are not in the picture. For instance, when the input image for the model is a man walking on the beach, the model may produce a caption as "A man walking on the beach with the surfboard", even when there is no such object in the picture. This could be attributed to transformers having strong language priors extracted from large datasets, which the model may reuse incorrectly. If the model tends to hallucinate, it means that based on its training, the two drives frequently occur together then, it generalizes descriptions and does not confine it in the provided visual context. Correcting these errors demands better management of Multi-modal alignment in which the model is crafted to pass good integration between the visual features and the linguistic representations.

V. CONCLUSION

Recent innovations in deep learning discovered that the deep Transformer model has enhanced the image captioning systems more than CNN-LSTM and other attention models. In originality, ViT is adopted for feature extraction, while the choice of the Transformed r-based decoder enhances fluency, accuracy, and context of captions. Transformers do not require sequential processing as do the RNN-based models; therefore, they can easily compute the relations between them and generate coherent descriptions. The attention mechanism is essential for enhancing the vision-language alignment since it enables the model to learn where to pay attention in the images to improve overall caption quality. Analysis of the results obtained on MS COCO and Flickr8k datasets testified that the proposed Transformer-based model performs better than existing models regarding BLEU-4, METEOR, and CIDEr scores. However, there are evident imperfections, including wrong object recognition and hallucination, which call for optimizations. Nevertheless, the described model is a substantial improvement to the initial model towards developing a human-like image captioning system that helps connect computer vision and NLP.

Future Work

Although the proposed model yields promising results with superior performance to state of the art, it is possible to consider the following points to extend the work and increase the practical applicability of the model: One of them is using multi-modal learning where the model also considers other types of data such as audio or text description to come up with even more enlightened captions. For example, including sound information can assist in sorting similar environment environments like a park and a street during the night. Another important research area is few-shot learning, which eliminates reliance on large labeled datasets.

With the future help of meta-learning and self-supervised learning, the models can be trained with much less annotated data across various domains. Furthermore, real-time image captioning in smartphones, drones, and augmented reality systems could open new application areas like assistive vision impaired devices and media content description services. To accommodate the integration into applications for daily use, optimising the model for low power consumption on hardware and real-time inference would prove effective. Finally, enhancing model robustness in terms of hallucination and object misidentification will be appealing, which can be done by contrastive learning, reinforcement learning, and improvements in grounding capabilities. By addressing these areas, future studies in image captioning can advance to make AI-generated vision-language models more precise, flexible, and natural in interpreting and generating descriptions for graphic images.

VI. REFERENCES

1. Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator in Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3156-3164).
2. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... & Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
3. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 6077-6086).
4. Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 7008-7024).
5. Herdade, S., Kappeler, A., Boakye, K., & Soares, J. (2019). Image captioning: Transforming objects into words. Advances in neural information processing systems, 32.
6. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30.
7. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021, July). Learning transferable visual models from natural language supervision. In International conference on machine learning (pp. 8748-8763). PmLR.

8. Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 10578-10587).
9. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., ... & Gao, J. (2020). Oscar: Object-semantics aligned pre-training for vision-language tasks. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16 (pp. 121-137). Springer International Publishing.
10. Zhou, L., Palangi, H., Zhang, L., Hu, H., Corso, J., & Gao, J. (2020, April). Unified vision-language pre-training for image captioning and vqa. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 07, pp. 13041-13049).
11. Huang, L., Wang, W., Chen, J., & Wei, X. Y. (2019). Attention on attention for image captioning. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 4634-4643).
12. Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018, July). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 2556-2565).
13. Wadhwa, V., Gupta, B., & Gupta, S. (2021, December). AI-based automated image caption tool implementation for the visually impaired. In 2021 International Conference on Industrial Electronics Research and Applications (ICIERA) (pp. 1-6). IEEE.
14. Sortino, R., Palazzo, S., Rundo, F., & Spampinato, C. (2023). Transformer-based image generation from scene graphs. Computer Vision and Image Understanding, 233, 103721.
15. Ondeng, O., Ouma, H., & Akuon, P. (2023). A review of transformer-based approaches for image captioning. Applied Sciences, 13(19), 11103.
16. He, S., Liao, W., Tavakoli, H. R., Yang, M., Rosenhahn, B., & Pugeault, N. (2020). Image captioning through image transformer. In Proceedings of the Asian conference on computer vision.
17. Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 3128-3137).
18. Parvin, H., Naghsh-Nilchi, A. R., & Mohammadi, H. M. (2023). Transformer-based local-global guidance for image captioning. Expert Systems with Applications, 223, 119774.
19. Chandy, A. (2019). A review on IoT-based medical imaging technology for healthcare applications. Journal of Innovative Image Processing (JIIP), 1(01), 51-60.
20. Iijima, L., Giakoumoglou, N., & Stathaki, T. (2024). A multimodal approach for cross-domain image retrieval. arXiv preprint arXiv:2403.15152.
21. Cherekar, R. (2023). A Comprehensive Framework for Quality Assurance in Artificial Intelligence: Methodologies, Standards, and Best Practices. International Journal of Emerging Research in Engineering and Technology, 4(2), 43-51. https://doi.org/10.63282/3050-922X.IJERET-V4I2P105
22. Cherekar, R. (2022). Cloud Data Governance: Policies, Compliance, and Ethical Considerations. International Journal of AI, BigData, Computational and Management Studies, 3(2), 24-31. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I2P103
23. Cherekar, R. (2020). DataOps and Agile Data Engineering: Accelerating Data-Driven Decision-Making. International Journal of Emerging Research in Engineering and Technology, 1(1), 31-39. https://doi.org/10.63282/3050-922X.IJERET-V1I1P104
24. Cherekar, R. (2020). The Future of Data Governance: Ethical and Legal Considerations in AI-Driven Analytics. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 3(2), 53-60. https://doi.org/10.63282/3050-9262.IJAIDSML-V3I2P107
25. Cherekar, R. (2022). Cloud Data Governance: Policies, Compliance, and Ethical Considerations. International Journal of AI, BigData, Computational and Management Studies, 3(2), 24-31. https://doi.org/10.63282/3050-9416.IJAIBDCMS-V3I2P103
26. Cherekar, R. (2021). The Future of AI Quality Assurance: Emerging Trends, Challenges, and the Need for Automated Testing Frameworks. International Journal of Emerging Trends in Computer Science and Information Technology, 2(1), 19-27. https://doi.org/10.63282/3050-9246.IJETCSIT-V1I2P104
27. Cherekar, R. (2023). Automated Data Cleaning: AI Methods for Enhancing Data Quality and Consistency. International Journal of Emerging Trends in Computer Science and Information Technology, 5(1), 31-40. https://doi.org/10.63282/3050-9246.IJETCSIT -V5I1P105
28. Rahul Cherekar, "The Integration of Big Data and Business Intelligence: Challenges and Future Directions" International Journal of Multidisciplinary on Science and Management, Vol. 1, No. 2, pp. 38-48, 2024.
29. Cherekar, R. (2020). Integrating AI-Based Image Processing with Cloud-Native Computational Infrastructures for Scalable Analysis. International Journal of Artificial Intelligence, Data Science, and Machine Learning, 6(2), 55-62. https://doi.org/10.63282/3050-9262.IJAIDSML-V6I2P106