Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Multi-Modal Image Captioning to Aid Visually Impaired Persons

S.V. Vasantha[1], S. Ramakrishna[2], B. Jyoshna[3], B. Kiranmai[4], P. Jose[5], S. Hariharan[6]

[1]CSE Department, Vardhaman College of Engineering, Hyderabad, India, s.v.vasantha@gmail.com
[2]CSE Department, GSoT, GITAM(Deemed-to-be-University), Hyderabad, India, rsankara@gitam.edu
[3]CSE(DS) Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, jyoshnabejjam@gmail.com
[4]CSE(DS) Department, Sreyas Institute of Engineering and Technology, Hyderabad, India, kiranmaimtech@gmail.com
[5]CSE Department, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India, drjosep@veltech.edu.in
[6]AI&DS Department, Vardhaman College of Engineering, Hyderabad, India, hariharan.s@vardhaman.org

Image comprehension is vital for the visually impaired, enabling understanding of visual information crucial for daily activities. Access to images aids in education, employment, social interactions, and navigation. Image captioning and audio descriptions are essential for enhancing accessibility and promoting inclusivity for individuals with visual impairments. This study addresses the significant challenge of enhancing image comprehension for the visually impaired by investigating and comparing two distinct models. Model 1 integrates a pretrained ResNet-50 with LSTM and RNN architectures, aiming to enhance language expressiveness using Glove word embeddings and enable robust picture feature extraction. On the other hand, Model 2 employs Vision Transformer (ViT) and GPT-2 architectures. Both models are trained using the Flickr8K dataset, and the generated captions are converted into audio to meet accessibility requirements. Through evaluations utilizing METEOR and ROUGE scores, Model 1 achieves a BLEU score of 58.26, while Model 2 achieves a BLEU score of 69.59. The results highlight the superior performance of the ViT-GPT2 model over the LSTM + RNN-based model, demonstrating its potential for picture captioning tasks and its utility for visually impaired individuals through caption-to-audio conversion.

Keywords: Image Captioning, Deep Learning, LSTM, GPT-2, VIT.

1. Introduction

In today's interconnected world, the omnipresence of visual content is reshaping how information is conveyed and consumed. The online experiences are dominated by pictures and photos, thus understanding images are essential to our digital interactions. A crucial area of research in artificial intelligence is image captioning, which is the process of creating descriptive stories for visual images by combining computer vision and Natural Language Processing (NLP) methods. The objective of this systematic literature review is to thoroughly assess deep learning approaches used for picture captioning, illuminating important methods and standard datasets used in the domain. Using datasets such MS COCO and Flickr8k as standardized benchmarks for performance measurement, the integration of LSTM, CNN, and RNN models has become a popular method for addressing the difficulties of image caption synthesis Al-Shamayleh et. al.(2024). This study attempts to overcome the accessibility gap in the appreciation of visual art for those who are blind or have poor vision (BLV). Two experiments are presented: an evaluation of picture captioning models using artwork datasets, and an exploration of BLV preferences for layered description qualities.

Researchers working on applications for museum engagement and accessible picture captioning are given recommendations that prioritize spatial information access techniques Doore SA et. al.(2024). collaboration with the National Council for the Blind of Ireland (NCBI), outdoor navigation for People with Visual Impairments (PVI) emerges as a significant challenge, underexplored in existing literature. Their questionnaire-based research with 49 PVI participants highlights deficiencies in current navigation applications, particularly in providing essential information and addressing safety concerns. This study underscores the critical need for improved navigation solutions tailored to the specific needs of PVI, guiding future advancements in navigation technology F.E.Z. El-Taher et. al.(2023). Utilizing 60 images from Wikipedia and corresponding descriptions, generated by four state-of-the-art tools, blind evaluations were performed by 76 computer science students. Findings indicate that while Wikipedia descriptions are perceived as most accurate, certain tools show promise for specific image categories, suggesting potential for automated image description addition and improved web accessibility Leotta et. al.(2023).

2. RELATED WORK

The approach in S.R. Chandaran et. al.(2023) combines YOLOv5 for object recognition and a Bi-LSTM layer for feature extraction, resulting in enhanced performance with a notable BLEU score of 0.7 on the Flickr8k benchmark dataset. This refined algorithm significantly improves picture content description accuracy compared to conventional encoder-decoder approaches. To address the lack of Tibetan caption data, the study proposes an approach for generating Tibetan captions using VGG19, InceptionV3, and ResNet101 networks, along with group convolution to improve attention mechanisms J. Xia et al.(2023).

The paper investigates methods for automatically generating text descriptions from images, focusing on complex real-world scenarios and unfamiliar objects, various deep learning techniques, including CNN, RNN, DNN, and LSTM, are explored using datasets like Flickr8K, Flickr30K, and MSCOCO, with an emphasis on MSCOCO's extensive 82783-image

dataset A.P. Singh et. al.(2023).This study investigates combining CNN and LSTM architectures for image caption creation. CNN and LSTM combined with beam search and maximum likelihood estimation produce intelligent captions C. Bhatt et. al.(2023).The model in L. Panigrahi et. al.(2023) focused on relationships between image areas, caption words, and RNN states. Techniques like progressive loading and VGG16 encoding improve performance, as measured by the BLEU metric. This paper uses Xception, VGG-16, and ResNet50 CNN models on the Flickr_8k dataset A. Verma et. al(2023).

Next-LSTM, a novel picture captioning technique, utilizes ResNet for image feature extraction and LSTM network for accurate captioning. Evaluation on the Flickr-8k dataset considers parameters such as Accuracy and BLEU Score Singh et. al.(2023).This study compares LSTMs and Transformers., evaluated on the Flickr8k dataset using the VGG16 model for picture feature extraction, employs the BLEU score metric to assess model performance. Both models demonstrate competence in generating emotive and grammatically correct captions, contributing to the ongoing advancement of image captioning techniques Zouitni et. al.(2023).This paper introduces an innovative image captioning model based on the transformer architecture with a Fourier transform Joshy et. al.(2023).The approach surpasses current state-of-the-art models by capturing key relationships in image elements. Evaluated on the Flickr dataset, the model demonstrates superior performance through metrics like BLEU-n, METEOR, and ROUGE scores.

An improved Transformer-based picture captioning model with a Grid-Augmented Module and a Multi-Featured Attention Module is presented. After a thorough assessment, the model outperforms the Transformer baseline on a number of metrics, including a BLEU-4 score of 0.409 and a CIDEr score of 1.008 with a beam size of 7 in Xian et. al.(2023).This study highlights the importance of image captioning particularly in fields like medical imaging and assisting the blind and visually impaired. Indonesian datasets are examined, using CNN with ResNet as the encoder and Transformer as the decoder. The study also investigates the effects of various pre-trained CNN models, demonstrating that increased model size does not necessarily lead to increased accuracy over training epochs R. Mulyawan et. al.(2022). Uses an attention-based architecture, it integrates convolutional features from a pre-trained Xception CNN and object features from YOLOv4, resulting in a notable 15.04% improvement in the CIDEr score Al-Malla et. al.(2022).

Extracted features from models are inputted into an LSTM for sentence formation, achieving BLEU scores of 0.79 (Xception), 0.75 (VGG-16), and 0.84 (ResNet50). Notably, ResNet50 achieves 84% accuracy in captions N. Goel et. al.(2023).This paper provides a thorough exploration of image captioning models, introducing cutting-edge approaches to generate textual descriptions for visual content TaranehGhandi et. al.(2023). Recent work in related fields has concentrated on programmed captioning, which blends computer vision and speech processing. It is common practice to utilize CNN and RNN/LSTM models with evaluation using BLEU scores and datasets such as MS COCO. Attention mechanisms and encoder-decoder designs are promising possibilities R. Kumar and G. Goel(2023), Yang et. al.(2024).

3. PROPOSED SYSTEM

Image captioning is crucial for individuals with visual impairments, providing them with access to visual content through textual descriptions. By bridging this accessibility gap, image captioning promotes inclusion and equal access to information, empowering visually impaired individuals to engage with a wide range of visual content. This technology enhances overall user experience and fosters inclusivity in various contexts, making it an essential tool for promoting accessibility and ensuring equal participation for all individuals, regardless of their visual abilities. In this study two different multi-models are experimented. First model is based on ResNet-50,GloVe and LSTM and second model is hybrid combination of ViT and GPT-2.

3.1 ResNet-50-GloVe+LSTMMulti-Model

The model 1 architecture depicted in Fig.1, involves several key components for efficient image caption generation. Firstly, the ResNet-50 pretrained model is employed to extract features from images, capturing both fine-grained features and high-level abstractions. Simultaneously, GloVe word embeddings are utilized to convert words in captions into vectors, encoding semantic information based on contextual relevance. These image features from ResNet-50 and tokenized captions serve as inputs to an LSTM network, which predicts the next word in the caption sequence. This LSTM network comprehensively analyzes the combined visual and textual features to generate contextually relevant and semantically coherent captions. Subsequently, the generated captions are converted into audio in the chosen language to enhance accessibility.

Finally, the quality of the captions is evaluated using metrics such as BLEU, ROUGE, and METEOR, comparing them against reference or human-generated captions to provide quantitative measures of linguistic similarity and relevance.

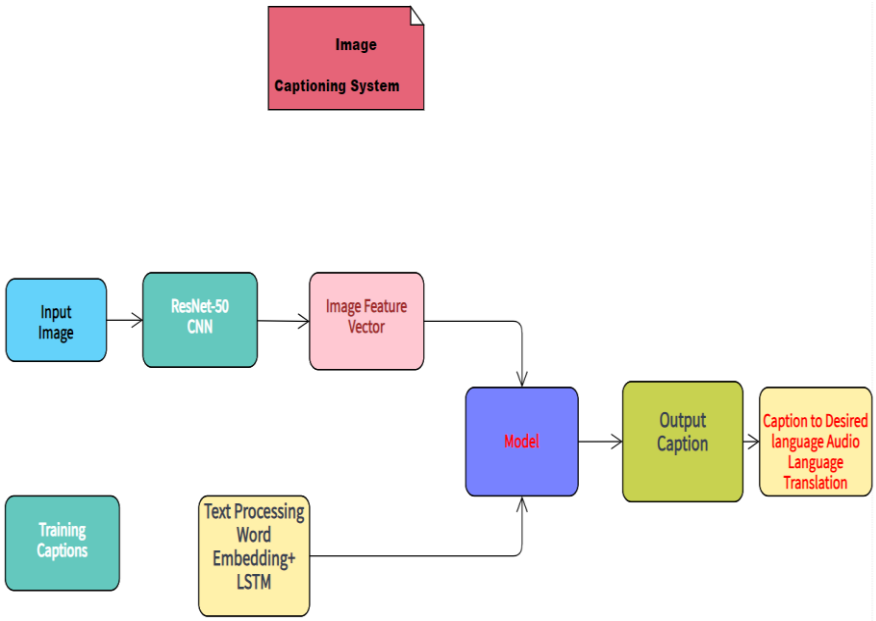

Figure 1. ResNet-50-GloVe+LSTM Multi-Model Architecture

3.2 Vision Transformer (ViT) - GPT-2 Hybrid Multi-Modal

The ViTImageProcessor employs self-attention mechanisms to extract features from images, capturing relevant relationships between different regions. Through fine-tuning with the 'google/vit-base-patch16-224' architecture on the Flickr8k dataset, both the Vision Transformer (ViT) and GPT-2 models adapt their parameters to learn domain-specific knowledge, enhancing performance in generating captions. The fine-tuned ViT model extracts image features, while the fine-tuned GPT-2 model utilizes both visual and textual inputs to generate contextually relevant captions, integrating visual and semantic information effectively.

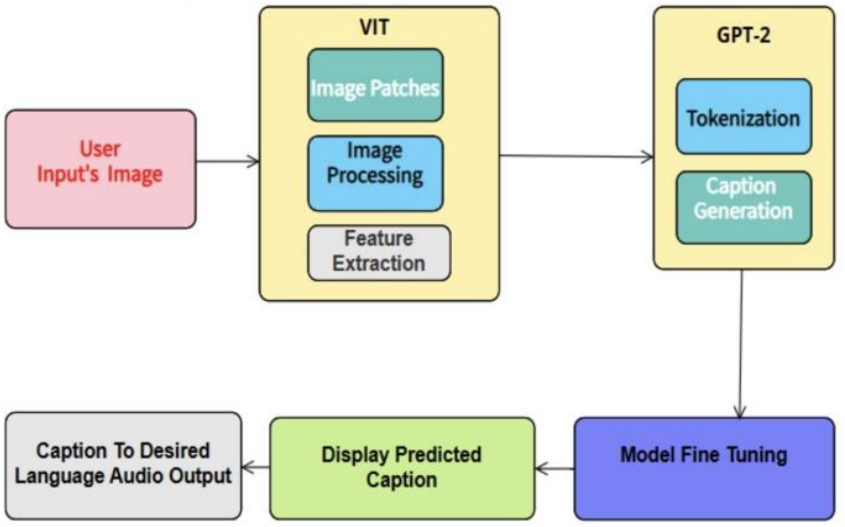

Figure 2. ViT-GPT2 Muti-Model Architecture

3.3 Metrics for Image Captioning

The generated captions were evaluated using various criteria, including BLEU, ROUGE, and METEOR.

BLEU:

let's calculate the BLEU score for unigrams (1-gram) and bigrams (2-grams).

Reference Sentence: A cute cat is sleeping on the sofa.

Hypothesis Sentence: A small cat sleeps on the couch.

Unigram (1-gram) BLEU Score:

Tokenization: Reference - ['A', 'cute', 'cat', 'is', 'sleeping', 'on', 'the', 'sofa'],

Hypothesis - ['A', 'small', 'cat', 'sleeps', 'on', 'the', 'couch'].

Count Matches: $4/7 \approx 0.5714$.

Precision1 = $4/7 \approx 0.5714$.

Bigram (2-gram) BLEU Score:

Tokenization: Reference - ['Acute', 'cutecat', 'catis', 'issleeping', 'sleepingon', 'onthe', 'thesofa'],

Hypothesis - ['Asmall', 'smallcat', 'catsleeps', 'sleepson', 'onthe', 'thecouch'].

Count Matches: 2/6 ≈ 0.3333.

Precision2 = 2/6 ≈ 0.3333.

Geometric mean of Precision1 and Precision2.

BLEU = √(0.5714 × 0.3333) ≈ 0.4388.

Overall BLEU Score:

$$\text{BLEU} = \text{BP} \times \exp\left(\frac{1}{N}\sum_{n-1}^{N}\log(\text{precision}_n)\right)$$

Here, N is the maximum n-gram order considered, and BP is the brevity penalty.

ROUGE:

ROUGE measures the overlap between the system-generated summaries and the reference summaries. There are several variants such as ROUGE-1, ROUGE-2, and ROUGE-L.

ROUGE-N (N-gram overlap):

ROUGE-N = (Overlap of N-grams in reference and hypothesis) / (Total N-grams in reference)

ROUGE-L (Longest Common Subsequence):

ROUGE-L = (Length of longest common subsequence in reference and hypothesis) / (Length of reference)

ROUGE-W (Weighted N-gram overlap):

ROUGE-W = (∑ Weighted overlap of N-grams in reference and hypothesis) / (∑ Weighted total N-grams in reference)

By analyzing the n-gram overlap between the generated and reference texts, ROUGE assesses the recall and precision of shared sequences.

METOR:

METEOR = (β * Precision + (1 - β) * Recall) / ((1 - β) * Precision * Recall)

METEOR is an all-encompassing statistic that provides a thorough assessment of machine-generated text quality by taking into account precision, recall, stemming, and synonymy.

4. RESULTS AND COMPARATIVE ANALYSIS

Two distinct models were developed using different frameworks, each incorporating various components for image captioning. Model 1, implemented using Keras and TensorFlow,

integrated ResNet-50 for image feature extraction, GloVe embeddings for text representation, and an LSTM network for generating captions. On the other hand, Model 2 was built using PyTorch and transformers. Both models were evaluated using standard evaluation metrics such as BLEU, METEOR, and ROUGE, with the results of five test samples presented in Fig. 3, showcasing their proficiency in generating captions. Notably, Model 2 demonstrated superior caption relevance compared to Model 1, as evidenced by better metric scores depicted in Fig. 4. The framework of Model 2 incorporates advanced models such as ViT and GPT, resulting in the enhancement of picture captioning systems, particularly evident in its ability to produce highly relevant captions. Furthermore, our proposed Model 2 was compared with other existing solutions using the same three metrics, the result scores are detailed in Table 1, and it outperformed them across all measures.

Figure 3. Test Samples results of Model-1 and Model-2

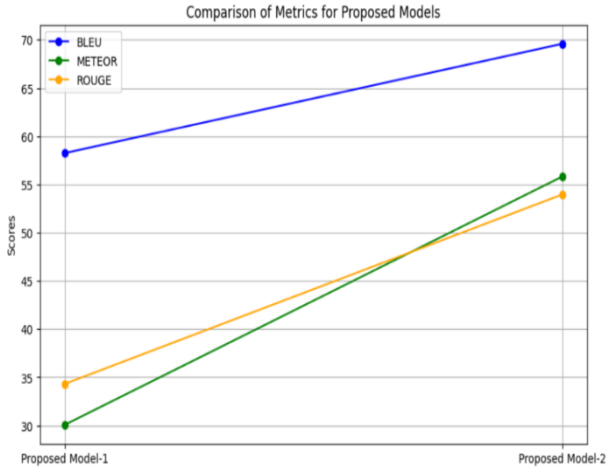

Figure 4. Performance analysis of Model-1 and Model-2

Table 1. Comparison of proposed multi-model with Contemporary Models.

Model/Metric	BLUE	METEOR	ROUGE
PerceptGuideModel[11]	59.8	17.6	42.9
DL Model[12]	66.6	50.0	30.76
Transformer Model[13]	56.00	19.50	44.16
Proposed Model-2(ViT-GPT-2)	69.59	55.82	53.96

5. CONCLUSION

In this work, two new image captioning models that leverage a hybrid technique are introduced, combining the strengths of GPT-2 language models and Vision Transformers (ViT). Our initial model achieved a competitive BLEU score of 58.26 by integrating a pre-trained ResNet-50 for image feature extraction and GloVe word embed-dings for linguistic expressiveness. In contrast, our second model outperformed the first, attaining an impressive BLEU score of 69.59 after refinement on the Flickr8K dataset using ViT and GPT-2. These outcomes underscore the effectiveness of our models in generating insightful and contextually appropriate captions for photos. Our approach, which integrates language and visual settings, holds promise for applications in multimedia content development and assistive technologies for the blind. By introducing novel hybrid architectures, our research contributes to the evolving field of picture captioning models. Future directions in image captioning for the blind involve enhancing context understanding through advanced techniques like attention mechanisms, and integrating multimodal sensory information for a more immersive user experience.

References

1. A. P. Singh, M. Manoria and S. Joshi, A Review on Automatic Image Caption Genera-tion for Various Deep Learning Approaches, 14th International Conference on Computing Communi-

cation and Networking Technologies (ICCCNT), Delhi, India, pp. 1-5, [doi: 10.1109/ICCCNT56998.2023.10308085](2023).

2. A. Verma, A. Yadav, M. Kumar, & D. Yadav, "Automatic image caption generation us-ing deep learning",, 2022. [doi.org/10.21203/rs.3.rs-1282936/v1](2023).
3. Al-Malla, M.A., Jafar, A. &Ghneim, N. Image captioning model using attention and object features to mimic human image understanding. J Big Data 9, 20 [doi.org/10.1186/s40537-022-00571-w](2022).
4. Al-Shamayleh, A.S., Adwan, O., Alsharaiah, M.A. et al. A comprehensive literature review on image captioning methods and metrics based on deep learning technique. Multimed Tools Appl (2024). https://doi.org/10.1007/s11042-024-18307-8.
5. C. Bhatt, S. Rai, R. Chauhan, D. Dua, M. Kumar and S. Sharma, "Deep Fusion: A CNN-LSTM Image Caption Generator for Enhanced Visual Understanding," 3rd International Con-ference on Innovative Sustainable Computational Technologies (CISCT), Dehradun, India, pp. 1-4, [doi: 10.1109/CISCT57197.2023.10351389](2023).
6. Doore SA, Istrati D, Xu C, Qiu Y, Sarrazin A, Giudice NA. Images, Words, and Imagination: Accessible Descriptions to Support Blind and Low Vision Art Exploration and Engagement. Journal of Imaging. 2024; 10(1):26. https://doi.org/10.3390/jimaging10010026
7. F. E. -Z. El-Taher, L. Miralles-Pechuán, J. Courtney, K. Millar, C. Smith and S. Mckeever, "A Survey on Outdoor Navigation Applications for People With Visual Impairments," in IEEE Access, vol. 11, pp. 14647-14666, 2023, doi: 10.1109/ACCESS.2023.3244073.
8. J. Xia, X. Yang, Q. Ni and D. Gao, Research on Image Tibetan Caption Generation Method Fusion Attention Mechanism, IEEE 4th International Conference on Pat-tern Recognition and Machine Learning (PRML), Urumqi, China, pp. 193-198, [doi: 10.1109/PRML59573.2023.10348351](2023).
9. Joshy, D. M., Das, A., M M, A., Sunil, D. T., & Safar, Enriching Transformer using Fourier Transform for Image Captioning. 3rd International Conference on Intelligent Technologies (CONIT), 1–6. [doi: 10.1109/CONIT59222.2023.10205936](2023).
10. L. Panigrahi, R. R. Panigrahi and S. K. Chandra, Hybrid Image Captioning Model,"OPJU International Technology Conference on Emerging Technologies for Sus-tainable Development (OTCON), Raigarh, Chhattisgarh, India, pp. 1-6, [doi: 10.1109/OTCON56053.2023.10113957](2023).
11. Leotta, M., Mori, F. &Ribaudo, M. Evaluating the effectiveness of automatic image captioning for web accessibility. Univ Access Inf Soc 22, 1293–1313 (2023). https://doi.org/10.1007/s10209-022-00906-7
12. N. Goel, A. Arora, P. Kashyap and S. Varshney, An Analysis of Image Captioning Models using Deep Learning, International Conference on Disruptive Technolo-gies (ICDT), Greater Noida, India, pp. 131-136, [doi: 10.1109/ICDT57929.2023.10151421](2023).
13. R. Kumar and G. Goel, Image Caption using CNN in Computer Vision, Interna-tional Conference on Artificial Intelligence and Smart Communication (AISC), Greater Noida, India, pp. 874-878, [doi: 10.1109/AISC56616.2023.10085162](2023).
14. R. Mulyawan, A. Sunyoto and A. H. Muhammad, Automatic Indonesian Image Cap-tioning using CNN and Transformer-Based Model Approach, 5th International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 2022, pp. 355-360, [doi: 10.1109/ICOIACT55506.2022.9971855](2022).
15. S. R. Chandaran, S. Natesan, G. Muthusamy, P. K. Sivakumar, P. Mohanraj and R. J. Gnanaprakasam, Image Captioning Using Deep Learning Techniques for Partially Impaired People, International Conference on Computer Communication and In-formatics (ICCCI), Coimbatore, India, pp. 1-6, [doi: 10.1109/ICCCI56745.2023.10128287](2023).
16. Singh, P., Kumar, C., & Kumar, A Next-LSTM: a novel LSTM-based image captioning technique. Int J Syst Assur EngManag, 14(5), 1492–1503. [doi: 10.1007/s13198-023-01956-7]

(2023).
17. TaranehGhandi, HamidrezaPourreza, and HamidrezaMahyar , Deep Learning Approaches on Image Captioning: A Review. ACM Comput. Surv. 56, 3, Article 62 , 39 pages.doi.org/10.1145/3617592](2023)
18. Xian, H., Guo, B., & Zhou, Balanced Overall and Local: Improving Image Captioning with Enhanced Transformer Model. 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), 663–667. [doi: 10.1109/AINIT59027.2023.10212804](2023).
19. Yang, X., Yang, Y., Wu, J., Sun, W., Ma, S. and Hou, Z, CA-Captioner: A novel concentrated attention for image captioning. Expert Systems with Applications, 250, p.123847(2024).
20. Zouitni, C., Sabri, M. A., &Aarab. A Comparison Between LSTM and Transformers for Image Captioning. In S. Motahhir& B. Bossoufi (Eds.), Digital Technologies and Applications (pp. 1–13). Springer, Cham. [doi: 10.1007/978-3-031-29860-8_50] (2023).