Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online) 2026, Vol. 5, No. 1

A Comparative Analysis of Image Captioning in Nepali Language Using Deep Learning

Prabhas Parajuli
Advanced College of Engineering and Management
Tribhuvan University, Nepal
prabhas.077bct053@acem.edu.np

Sumita Dangal
Advanced College of Engineering and Management
Tribhuvan University, Nepal
sumita.077bct090@acem.edu.np

Prafulla Shrestha
Advanced College of Engineering and Management
Tribhuvan University, Nepal
prafulla.077bct054@acem.edu.np

Saru Pradhan
Advanced College of Engineering and Management
Tribhuvan University, Nepal
saru.077bct077@acem.edu.np

Bipun Man Pati
Advanced College of Engineering and Management
Tribhuvan University, Nepal
bipunmanpati@acem.edu.np

Ukesh Thapa
Advanced College of Engineering and Management
Tribhuvan University, Nepal
ukesh.thapa@acem.edu.np

Abstract*—Image captioning incorporates the knowledge of both image processing and Natural Language Processing (NLP). The complexity of Nepali grammar, written in the Devanagari script, presents a significant challenge in generating a grammatically correct description from an image. Challenges arise from both the morphological richness and free word order of Nepali grammar, as well as the intricacies of the Devanagari script. This work presents a comparative analysis of deep learning architectures for sequence modeling, with a particular emphasis on Transformer-based and Long Short-Term Memory (LSTM)-based models. In particular, we assess the effectiveness of feature extractors combined with sequential processors, such as Convolutional Neural Network (CNN)+Transformer, VGG16+LSTM, ResNet101+LSTM, EfficientNetB0+LSTM, Vision Transformer (ViT)+Transformer, and Global Context Vision Transformer (GCViT)+Transformer. The models are rigorously assessed on a Nepali-language captioning dataset using a comprehensive set of performance metrics: BLEU-1, BLEU-2, BLEU-3, and BLEU-4 for n-gram-based linguistics, along with METEOR. Results indicate that LSTM-based models consistently outperform their Transformer-based counterparts across all metrics, achieving higher BLEU scores and superior METEOR scores. In particular, the ResNet101-LSTM demonstrates particularly strong performance, suggesting the efficacy of end-to-end attention-based architectures is not always guaranteed in Transformer-based models while dealing with smaller datasets. The findings provide clear empirical benchmarks for model selection in Nepali language image captioning, highlighting the superior capability of LSTM-based approaches in generating linguistically accurate and semantically rich captions, which is vital for enhancing accessibility and digital inclusion for Nepali-speaking communities.

***Index Terms*—Image Processing, Nepali Image Captioning, Transformer Models, Self-Attention**

I. Introduction

In today's digital age, there has been an exponential growth in digital imagery, resulting in a pressing need for the automatic interpretation of visual content. This interpretation extends beyond identifying objects within an image and includes recognizing visual entities and their relationships that contribute to the image's overall meaning. This challenging task is often tackled using the image captioning technique. Image captioning not only identifies the objects within the image but also captures spatial relationships and incorporates them into a grammatically correct textual description. Consequently, it represents a multidisciplinary issue at the intersection of computer vision and NLP.

Automated image captioning has made substantial advancements over the years. Early methods relied on handcrafted features such as Local Binary Patterns (LBP) and Histogram of Oriented Gradient (HOG), combined with predifined language templates. However, these approaches were limited in vocabulary diversity, flexibility, and contextual accuracy [1]. The adoption of deep learning, particularly the encoder-decoder framework, has transformed the field of image captioning by enabling automatic extraction of visual features using CNNs [2] [3]. Subsequently, these features are then decoded into a word sequence using recurrent networks, such as LSTMs [4]. More recently, Transformer-based architectures [5], which leverage self-attention to capture long-range dependencies, have emerged as state-of-the-art models.

Although researchers have conducted extensive work in the domain of image captioning, most studies focus on the high-resource languages, particularly English. This concentration has resulted in the development of complex models and extensive benchmarks [6]. However, such an emphasis has led

DOI : https://doi.org/10.3126/jost.v5i1.92659
Article History: Received: 25 Oct. 2025, Revised: 18 Dec. 2025, Accepted: 1 Feb. 2026.

to a significant disparity in research attention towards low-resource languages, which remain relatively underexplored. This oversight is mainly due to the lack of extensive, high-quality annotated datasets and the complexity of morphological structures. [7] The scarcity of parallel image-text datasets in these languages poses substantial obstacles to the training of robust deep learning models. Additionally, the grammatical complexities, such as rich inflectional morphology and free word order, further complicate the generation of captions that are both fluent and syntactically correct. As a result, there exist significant gaps in both research and practical applications for a large number of native languages around the world.

This paper presents a comprehensive comparative analysis of encoder–decoder architectures for automated image captioning in the Nepali language, addressing the aforementioned gap. We systematically compare the performance of various encoder–decoder frameworks by integrating multiple visual encoders with two distinct decoding methods: LSTM and Transformer-based decoders.

Initially, we evaluated several CNN encoders, including VGG16, ResNet101, EfficientNetB0, and DenseNet201, to extract high-level visual features from input images. These encoders were paired with an LSTM decoder to construct CNN–LSTM-based encoder–decoder frameworks. We selected the encoder that demonstrated the best performance within the CNN–LSTM framework and subsequently integrated it with a Transformer-based decoder to create a CNN–Transformer architecture. This integration facilitates a comparative analysis between recurrent and attention-based decoding mechanisms.

Additionally, we implemented fully Transformer-based architectures, utilizing Vision Transformer (ViT) and Global Context Vision Transformer (GCViT) as visual encoders alongside a transformer decoder for caption generation. To evaluate the proposed models, we computed the Bilingual Evaluation Understudy (BLEU) and Metric for Evaluation of Translation with Explicit Ordering (METEOR) scores. The primary aim of this research is to identify the optimal architecture and establish a performance baseline for this field, serving as a foundational reference point for future studies and advancements in Nepali image captioning.

II. Related Work

This section reviews existing studies on image captioning, initially discussing common architectures before focusing on breakthroughs and ongoing challenges for low-resource languages, especially Nepali.

The image captioning techniques have evolved from template-based approaches involving handcrafted features [1] to encoder-decoder networks with a focus on deep architectures. Most of the modern techniques employ a CNN architecture as an image encoder along with a Recurrent Neural Network (RNN) as a language sentence decoder [4]. Recently, Transformer-based models that incorporate the self-attention mechanism have started to replace the RNN, as they demonstrated superior performance through appropriate processing of long-range dependency relationships [5]. Building upon this approach, ViTs [8] have been introduced to interpret images as a sequence of patches, allowing their encoder to operate analogously to a transformer. As a result, there has been a growing shift from CNNs to ViTs as the primary feature extractors in many image captioning models [9]. Present research continues to aim to improve these approaches along with their assessment techniques [10]. A major advancement in this direction is a shift towards Transformer-based architectures such as Swin Transformers, which enables end-to-end vision and language encoding at one stage, negating pre-trained CNN feature extractors [11]. Moreover, recent advancements such as the GCViT, enhance hierarchical transformer architectures by incorporating both local and global self-attention mechanisms. This approach facilitates the efficient modeling of fine-grained visual details while effectively capturing long-range contextual dependencies [12].

Although architectural improvements have significantly improved performance, their effectiveness relies on the availability of large-scale datasets. This poses a substantial challenge when extending image and video captioning to low-resource languages with complex morphology, which often require different methodological considerations [7].

Furthermore, certain languages may require the implementation of language-specific evaluation metrics to effectively assess the performance of various architectures. Typically, studies carried out in the realm of low-resource languages often rely on translated caption corpora from English-language caption corpora. For example, studies [13] [6] carried out image captioning in Hindi and Bengali script, using translated versions of the MSCOCO and Flickr8k corpora, respectively. Additionally, certain languages may require the use of language-specific evaluation metrics to accurately assess model performance. Typically, research in low-resource contexts relies on translated caption corpora that stem from well-established English datasets. For example, studies have investigated image captioning in Hindi and Bengali by utilizing translated versions of the MSCOCO and Flickr8k datasets, respectively.

Given these limitations in data availability and linguistic diversity, designing appropriate evaluation metrics has been a critical research focus. Significant progress in this area includes the design of domain-specific evaluation benchmarks, like SPICE (Semantic Propositional Image Caption Evaluation) [14] and CIDEr (Consensus-based Image Description Evaluation) [15], along with language-specific evaluation benchmarks, like the THUMB metric, especially in Arabic languages [7]. In addition, widely used NLP metrics—such as BLEU [16], METEOR [17], and ROUGE [18]—have also been adapted for image captioning tasks, despite their original design for machine translation.

In Nepali image captioning, initial research appears limited and has produced inconclusive findings, thus requiring a broader study. Early research by Adhikari et al. [19] comparatively assessed an encoder-decoder model both with and without visual attention, finding that the simpler model (without attention) performed well. Thereby suggesting that standard

attention modules are perhaps not well-suited within Nepali's syntactical framework. Furthermore, Budhathoki et al. [20] employed the Flickr8k dataset and utilised the MobileNet-V3 Large CNN to extract feature maps from the input, which were then fed into the cross-attention module of a Transformer Decoder. However, the model was affected by overfitting due to limitations in the dataset. To address this, they adopted a simpler and less complex Transformer model by eliminating the encoder part of the traditional Transformer architecture. Similarly, Subedi et al. [21] directly compared two CNN-Transformer configurations (ResNet101 and EfficientNetB0) on a translated Flickr 8K image set, deciding that EfficientNetB0 was better. These works are exploratory, often focusing on an architecture or inconclusively discussing model components.

III. Material and Methods

This section details the dataset, preprocessing steps, model architectures, training procedures, and evaluation metrics used in our comparative study.

A. Dataset and Preprocessing

The Flickr8k-Nepali dataset was utilized for this study, which consists of 8,091 images, each annotated with five captions written in the Devanagari script. The entire dataset was divided into training (6,472 images, 80%), validation (810 images, 10%), and test sets (809 images, 10%) according to the standard splits defined in the corpus. For preprocessing of the input images, each input was resized based on the requirements of the encoder. Additionally, normalization was performed using the mean and standard deviation values of the ImageNet dataset. Moreover, caption preprocessing was carried out by removing punctuation and numerical characters. A vocabulary was created from the training captions, resulting in a total of 12,655 unique Nepali tokens. Each caption was tokenized into a sequence of word indices, with a 'startseq' token prepended and an 'endseq' token appended. All sequences were padded to a maximum length of 28 tokens.

B. Model Architectures

All models employed in this research are designed to strictly follow the encoder–decoder architecture. In this framework, a visual encoder was used to obtain a feature representation of the input image, which was subsequently processed by a sequence decoder to generate a caption in the Nepali language. The comparative analysis is focused on two primary variants of sequence decoders, namely LSTM-based and Transformer-based models.

In this study, variants of LSTM-based models were systematically evaluated to identify the most effective visual feature extractor for the task of Nepali image captioning. To ensure a comprehensive comparison across architectural paradigms, a standard LSTM decoder was paired with four distinct CNN backbones, each representing a significant advancement in network design. The classical, deeply stacked VGG16 architecture was employed as a baseline to assess the performance of uniform hierarchical features. Similarly, the widely adopted ResNet101 model, incorporating residual skip connections, was used to represent the modern standard for training very deep networks. In contrast, the compound-scaled EfficientNetB0 architecture was included as a benchmark for parameter efficiency and computational effectiveness. Lastly, the densely connected DenseNet201 architecture was used to examine the hypothesis that extensive feature reuse yields richer representations for sequential language generation. This structured experimental design enabled the impact of visual encoding strategies to be isolated from the sequential decoding process. As a result, it facilitates the identification of the visual feature extractor that provided the most synergistic foundation for generating grammatically correct and semantically accurate Nepali captions when integrated with a recurrent language model. In this context, the final classification layers of each CNN architecture are removed, and feature vectors are extracted from the last pooling layer preceding the classification stage.

To further extend the comparison beyond convolutional paradigms, Transformer-based vision models were also incorporated. The ViT was utilized to evaluate the effectiveness of global self-attention mechanisms in capturing long-range dependencies within images. Additionally, the GCViT architecture was included to investigate the impact of integrating local inductive biases with global context modeling for enhanced visual representation learning.

Transformer-based models, in contrast, are designed to rely exclusively on self-attention mechanisms. For these models, the output of the visual encoder is formatted as a sequence of feature vectors. Based on the results of the initial CNN-LSTM-based model comparisons (Section IV.A), in which ResNet101 is identified as the most effective visual encoder, this architecture is selected for the primary CNN-Transformer hybrid model. This selection enabled the evaluation of the Transformer decoder when paired with the strongest visual feature extractor identified in the experimental setup. Furthermore, ViT+LSTM and GCViT+LSTM configurations were implemented, where image features were extracted as a sequence of patch embeddings and hierarchical global-context representations and processed by LSTM decoders.

Elaborating upon these frameworks, in the CNN+Transformer configuration, feature maps produced by a CNN backbone were flattened into a spatial feature sequence. In contrast, the Vision Transformer (ViT)+Transformer configuration employed a fully Transformer-based architecture, in which each image was divided into fixed-size patches, linearly embedded, and processed by a standard ViT encoder. Similarly, the GCViT-Transformer configuration was explored, where hierarchical visual features were extracted using GCViT. In all the configurations, captions were generated using a standard Transformer decoder. The decoder attended to the entire encoded visual sequence simultaneously and applied masked self-attention to the partially generated caption, thereby enabling more effective modeling of global context and long-range word dependencies compared to

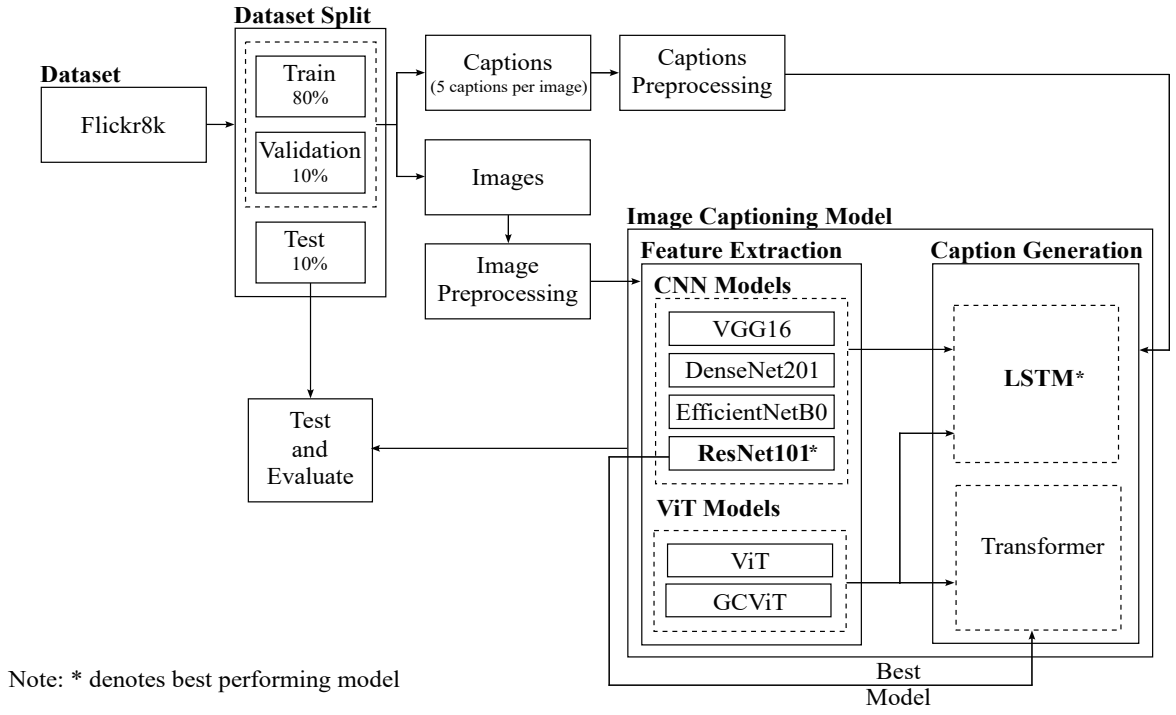

Fig. 1. General Block Diagram

recurrent decoding models.

C. Evaluation Metrics

For the quantitative assessment of captions generated by the models, BLEU and METEOR were employed as evaluation metrics. BLEU works by comparing the candidate translation to one or more human reference translations and measuring the overlap of n-grams (sequences of words) between them. It mainly focused on precision and applies a brevity penalty to discourage short output. In contrast, METEOR is based on unigram matching between the candidate translation and the human reference translation. Unlike BLEU, which relies on exact n-gram matches, the unigrams in METEOR can be matched based on their surface forms, stemmed forms, and meanings. It computes the harmonic mean of precision and recall and gives a penalty for the incorrect word order. The score range for the metrics is between 0 and 1, where the higher value indicates better performance of the model.

D. Simulation Setup

This study compares LSTM- and Transformer-based image caption generation using the Flickr8k-Nepali dataset. The LSTM-based approach employs four distinct CNN models and two ViT models as encoders to facilitate the extraction of image features. Among these, the CNN model that yielded the highest performance is integrated with a transformer decoder to assess its efficiency relative to the LSTM. Additionally, fully Transformer-based models are also evaluated in this study, where ViT and GCViT are used as an encoder, with the Transformer architecture employed for the caption generation task. The performance of all models is evaluated using the BLEU, and METEOR.

IV. Results

This section presents the results of our study on image captioning for the Flickr8k-Nepali dataset. We evaluated model performance through various metrics, including computational latency (inference time), BLEU, and METEOR. Table I summarizes the performance of the six LSTM-based models, specifically VGG16-LSTM, ResNet101-LSTM, EfficientNetB0-LSTM, DenseNet201-LSTM, ViT-LSTM, and GCViT-LSTM. Furthermore, Table II illustrates the performance of three Transformer-based models, namely CNN-Transformer, ViT-Transformer, and GCViT-Transformer. We further analyze the trade-off between performance and computational cost by measuring inference time.

A. LSTM-Based Model

From Table I, it was observed that ResNet101, when employed as an encoder, outperformed other CNN models, even though all models employed the same LSTM decoder architecture. It achieved a BLEU-4 score of 0.1479 and a METEOR score of 0.2589. In contrast, using DenseNet201 as the encoder yielded the worst performance among the CNN encoders, with BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR scores of 0.4515, 0.2755, 0.1424, 0.0730, and 0.1603, respectively. Nevertheless, although the DenseNet201-LSTM architecture exhibited the weakest performance, it illustrated a trade-off

TABLE I
PERFORMANCE COMPARISON OF LSTM-BASED MODELS

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Inference Time (Sec)
DenseNet201-LSTM	0.4515	0.2755	0.1424	0.0730	0.1603	0.5032
EfficientNetB0-LSTM	0.4902	0.3023	0.1737	0.0949	0.1967	0.6867
VGG16-LSTM	0.4948	0.3102	0.1879	0.1078	0.2126	0.6031
ResNet101-LSTM	0.5636	0.3763	0.2399	0.1479	0.2589	0.6515
ViT-LSTM	0.4404	0.2688	0.1420	0.0754	0.1527	0.4460
GCViT-LSTM	0.4342	0.2651	0.1404	0.0726	0.1372	0.5302

TABLE II
PERFORMANCE COMPARISON OF TRANSFORMER-BASED MODELS

Models	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	Inference Time (Sec)
ResNet101-Transformer	0.5460	0.3579	0.2205	0.1312	0.2409	0.6515
ViT-Transformer	0.4142	0.2373	0.1253	0.0659	0.1469	0.4460
GCViT-Transformer	0.4314	0.2541	0.1313	0.0660	0.1493	0.7994

between computational efficiency and predictive accuracy, as it exhibited the lowest inference time among all the models analyzed.

B. Transformer-Based Model

This subsection discusses the performance of the Transformer decoder model utilizing ResNet101, ViT, and GCViT as encoders, consistent with Section IV-A. As detailed in Table II, the model utilizing ResNet101 as an encoder outperformed other encoder models with the same Transformer architecture, achieving a METEOR score of 0.2409. It is noteworthy that transitioning from a CNN-based encoder to a Transformer-based model resulted in a significant drop in performance, with the ViT-Transformer and GCViT-Transformer models obtaining BLEU-4 scores of 0.0659 and 0.1493, along with METEOR scores of 0.0660 and 0.1493, respectively. Furthermore, in terms of computational efficiency, the GCViT-Transformer architecture exhibited the highest inference time (0.7994) among all the evaluated models, whereas ViT-Transformer demonstrated the lowest inference time (0.4460).

V. DISCUSSION

In this section, we discuss the results of various encoder-decoder architectures for Nepali image captioning using the Flickr8k-Nepali dataset, focusing on LSTM and Transformer decoders. The results demonstrate that encoder choice significantly impacts the performance. Among the evaluated models, ResNet101 consistently achieves the best performance, obtaining the highest BLEU and METEOR scores due to its deep residual learning framework, which effectively captures essential visual features for accurate caption generation. Moreover, to ensure fair comparison, all decoder architectures and hyperparameters, such as learning rate, batch size, and patch size, were kept constant throughout the experiments.

Similarly, decoder selection also plays a crucial role in overall system performance. The results from Table I and Table II show that LSTM-based decoders consistently outperform Transformer-based decoders across evaluation metrics. Notably, the ResNet101–LSTM configuration achieves superior results compared to the ResNet101–Transformer configuration, demonstrating the effectiveness of LSTMs in capturing sequential dependencies. While Transformer decoders capture long-range dependencies through attention mechanisms, they exhibit relatively lower performance in this context.

Furthermore, Table I shows that DenseNet201 achieves lower performance compared to other CNN encoders, despite its deeper architecture. This indicates that increasing network depth does not necessarily improve performance, as excessive feature reuse may introduce redundancy and reduce the diversity of features essential for caption generation.

Moreover, Table I and II show that Transformer-based encoders (ViT and GCViT) demonstrate comparatively lower performance across both decoding strategies, primarily due to their reliance on large-scale datasets for learning visual representations. Unlike CNN-based architectures, Transformer-based encoders lack inherent advantages for visual understanding, such as capturing local patterns and maintaining robustness to spatial variations. Consequently, they must learn these characteristics directly from the data, which limits their effectiveness on smaller datasets. Although both ViT and GCViT paired with LSTM decoders achieve higher BLEU scores than their Transformer-decoder counterparts, a different trend is observed for METEOR scores. In particular, GCViT with a Transformer decoder outperforms GCViT-LSTM in METEOR, indicating better semantic alignment. Overall, fully Transformer-based encoder–decoder configurations tend to underperform due to increased architectural complexity and optimization challenges. To further assess computational efficiency, we measured inference time, as shown in Table I and II. Among all the models evaluated, ViT-Transformer and DenseNet201-LSTM exhibit the lowest inference time, making them suitable for resource-constrained applications. In contrast, GCViT-transformer shows the highest inference time, highlighting its higher computational cost. Additionally, ResNet-LSTM inference time is also relatively low compared to most other models, while simultaneously achieving strong performance across BLEU and METEOR metrics, indicating a well-balanced trade-off between accuracy and efficiency.

Overall, the findings highlight that model performance in Nepali image captioning depends on the balance between encoder choice, decoder design, and computational efficiency. CNN-based encoders, particularly ResNet101 with LSTM decoders, provide the best trade-off between accuracy and inference time. Although Transformer-based models capture richer semantics, their higher computational cost and reliance on larger datasets limit their effectiveness in this context. These findings emphasize the importance of selecting the architecture according to the availability of data to ensure optimal learning, generalization, and overall performance of caption generation.

VI. Conclusion and Future Work

In this study, we conducted a comprehensive comparative analysis of encoder-decoder architectures for automated image captioning in the Nepali language. Our experiments on the Flickr8k-Nepali dataset demonstrated that LSTM-based decoders consistently outperform Transformer-based decoders across standard metrics, i.e., BLEU-1 through BLEU-4 and METEOR. Specifically, the CNN-LSTM model with a ResNet101 encoder emerged as the most reliable and efficient architecture, achieving the highest overall scores. While the Transformer-based model, which used a ResNet101 encoder, demonstrated competitive performance (METEOR:0.2409), it still could not outperform the best-performing model in LSTM-based models. Furthermore, none of the Transformer-based models outperformed their LSTM counterparts. Additionally, models using Transformer-based encoders, namely ViT and GCViT, exhibited low performance across both decoder types. This work provides a necessary performance benchmark and a methodological foundation for subsequent research in low-resource language image captioning. Furthermore, the analysis of inference time indicates that the Transformer-based model does not always guarantee computational efficiency, as evidenced by the GCViT-Transformer model, which exhibited the highest inference time.

Future research should enhance the Nepali dataset by including a larger variety of images and captions. To further enhance the model's performance, advanced data augmentation techniques and transfer learning strategies can be employed. Image captioning can also be integrated with speech recognition technology to assist visually impaired individuals, allowing them to interact with the system using voice commands, thereby improving accessibility. Furthermore, we can expand our research to generate paragraph-level captions, providing more detailed and context-rich descriptions of the images. Future work could also explore experimenting with other state-of-the-art models, such as BERT or T5, to evaluate their effectiveness in generating high-quality captions.

References

[1] M. Z. Hossain, F. Sohel, M. F. Shiratuddin, and H. Laga, "A comprehensive survey of deep learning for image captioning," *ACM Computing Surveys (CsUR)*, vol. 51, no. 6, pp. 1–36, 2019.

[2] J. Gu, G. Wang, J. Cai, and T. Chen, "An empirical study of language cnn for image captioning," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1222–1231.

[3] J. Aneja, A. Deshpande, and A. G. Schwing, "Convolutional image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5561–5570.

[4] C. Wang, H. Yang, C. Bartz, and C. Meinel, "Image captioning with deep bidirectional lstms," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 988–997.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] A. S. Ami, M. Humaira, M. A. R. K. Jim, S. Paul, and F. M. Shah, "Bengali image captioning with visual attention," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2020, pp. 1–5.

[7] M. T. Lasheen and N. H. Barakat, "Arabic image captioning: the effect of text pre-processing on the attention weights and the bleu-n scores," *Int J Adv Comput Sci Appl*, vol. 13, no. 7, p. 11, 2022.

[8] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[9] O. Ondeng, H. Ouma, and P. Akuon, "A review of transformer-based approaches for image captioning," *Applied Sciences*, vol. 13, no. 19, p. 11103, 2023.

[10] A. Jamil, K. Mahmood, M. G. Villar, T. Prola, I. D. L. T. Diez, M. A. Samad, I. Ashraf *et al.*, "Deep learning approaches for image captioning: Opportunities, challenges and future potential," *IEEE Access*, 2024.

[11] Y. Wang, J. Xu, and Y. Sun, "End-to-end transformer based model for image captioning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 2585–2594.

[12] A. Hatamizadeh, H. Yin, G. Heinrich, J. Kautz, and P. Molchanov, "Global context vision transformers," in *International conference on machine learning*. PMLR, 2023, pp. 12 633–12 646.

[13] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in hindi language using transformer networks," *Computers & Electrical Engineering*, vol. 92, p. 107114, 2021.

[14] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: Semantic propositional image caption evaluation," in *European conference on computer vision*. Springer, 2016, pp. 382–398.

[15] R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.

[16] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[17] S. Banerjee and A. Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.

[18] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.

[19] A. Adhikari and S. Ghimire, "Nepali image captioning," in *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, vol. 1. IEEE, 2019, pp. 1–6.

[20] R. Budhathoki and S. Timilsina, "Image captioning in nepali using cnn and transformer decoder," *Journal of Engineering and Sciences*, vol. 2, no. 1, pp. 41–48, 2023.

[21] B. Subedi and B. K. Bal, "Cnn-transformer based encoder-decoder model for nepali image captioning," in *Proceedings of the 19th International Conference on Natural Language Processing (ICON)*, 2022, pp. 86–91.