Journal of Soft Computing Paradigm (ISSN: 2582-2640)
www.irojournals.com/jscp/

Nepali Image Captioning: Generating Coherent Paragraph-Length Descriptions Using Transformer

Nabaraj Subedi[1], Nirajan Paudel[2], Manish Chhetri[3], Sudarshan Acharya[4], Nabin Lamichhane[5]

Electronics and Computer Department, Paschimanchal Campus, TU, Pokhara, Nepal

E-mail: [1]subedinabaraj46@gmail.com, [2]nirajanpaudel33@gmail.com, [3]hsinam12man34@gmail.com, [4]avilashisudarshan@gmail.com, [5]babunabin@gmail.com

Abstract

The advent of deep neural networks has made the image captioning task more feasible. It is a method of generating text by analyzing the different parts of an image. A lot of tasks related to this have been done in the English language, while very little effort is put into this task in other languages, particularly the Nepali language. It is an even harder task to carry out research in the Nepali language because of its difficult grammatical structure and vast language domain. Further, the little work done in the Nepali language is done to generate only a single sentence, but the proposed work emphasizes generating paragraph-long coherent sentences. The Stanford human genome dataset, which was translated into Nepali language using the Google Translate API is used in the proposed work. Along with this, a manually curated dataset consisting of 800 images of the cultural sites of Nepal, along with their Nepali captions, was also used. These two datasets were combined to train the deep learning model. The task involved working with transformer architecture. In this setup, image features were extracted using a pretrained Inception V3 model. These features were then inputted into the encoder segment after position encoding. Simultaneously, embedded tokens from captions were fed into the decoder segment. The resulting captions were assessed using BLEU scores, revealing higher accuracy and BLEU scores for the test images.

DOI: https://doi.org/10.36548/jscp.2024.1.006
Received: 08.03.2024, received in revised form: 04.04.2024, accepted: 18.04.2024, published: 30.04.2024

Keywords: BLEU, Inception V3, Nepali Captions, Transformer

1. Introduction

Understanding and expressing visual input in natural language is pivotal for various applications across domains such as accessibility, tourism, and urban development. The development of systems capable of comprehending visual input and generating natural language descriptions represents a critical challenge at the intersection of computer vision and natural language processing [3]. While considerable progress has been made in this field, much of the existing research has predominantly focused on English-language tasks, leaving languages like Nepali largely unexplored.

Nepali, as a language, presents unique challenges due to its grammatical complexities and limited availability of resources for natural language processing tasks. Hence, the development of systems for describing images in Nepali holds significant promise, particularly in regions like Nepal, where it can enhance accessibility for the visually impaired, improve tourism experiences, and contribute to the advancement of smart city initiatives.

In this context, our research aims to address the gap in Nepali language captioning by developing systems capable of generating coherent paragraph-long descriptions from images. Unlike previous works [4], [5], [6] that typically generate single captions, our approach seeks to provide more comprehensive and context-rich descriptions, mirroring human-like narrative structures.

To facilitate our research, we have curated a novel dataset consisting of images and descriptions of cultural heritage sites in Nepal, along with the Nepali Paragraph dataset derived from the Stanford Paragraph Image Captioning English dataset [1] using google translate API. This dataset not only provides valuable training and evaluation resources but also showcases the linguistic nuances of Nepali captions, which are crucial for accurate image captioning.

In our investigation, we examine the effectiveness of the Transformer architecture [2] for image captioning in Nepali. Transformer models [7] have shown remarkable performance in numerous natural language processing tasks, such as machine translation. This architecture holds promise for accurately generating captions in Nepali.

The key contributions of this research work are:

1. We compiled the Nepali Paragraph dataset for image captioning by manually refining captions from the English Stanford dataset [1] and creating 800 original Nepali cultural image descriptions, while also verifying the accuracy of Google-translated content through human correction.
2. Utilize a Transformer-CNN architecture to generate Nepali Paragraph captions from images. Through our research, we aim to contribute to the advancement of image captioning technology in Nepali, paving the way for improved accessibility, tourism experiences, and urban development initiatives in Nepal and beyond.

2. Related Works

Utilizing computer vision algorithms for image captioning has primarily been focused on English language datasets largely due to the inherent complexities of other languages. However, the increasing need for multilingual image captioning has prompted researchers to explore the extension of these techniques to languages beyond English. This expansion not only facilitates image-text retrieval but also enables image captioning and translation in diverse linguistic contexts.

Among the foundational techniques used in image captioning is the Long Short-Term Memory (LSTM) neural network [8], renowned for its ability to maintain long-short term memory, thereby addressing the short-term memory limitations of standard Recurrent Neural Networks (RNNs). This feature is particularly crucial for various tasks such as Natural Language Processing (NLP), object detection, and machine translation. The prevailing sequence translation models typically employ advanced convolutional and recurrent neural networks organized in an encoder-decoder setup, often drawing inspiration from machine translation methodologies.

In the domain of non-English language image captioning, significant strides have been made, particularly in languages such as Hindi and Bengali, which share similarities with Nepali. In Bengali language research, notable studies by S. Paul et al. [9] have explored techniques utilizing convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to generate Bengali captions from images. Subsequent investigations by the same

authors have delved into utilizing transformer models for Bengali image captioning (Shah et al., 2021) [10] focusing on addressing sequential processing challenges.

Similarly, in the Hindi language, S.K. Mishra et al. [11] have introduced a novel image captioning model tailored specifically for Hindi, leveraging transformer networks within an encoder-decoder architecture. This approach, which incorporates a Hindi dataset translated from MSCOCO[12] and refined by human annotators, showcases the potential of transformer-based models in multilingual image captioning tasks.

In contrast, research in Nepali language image captioning remains relatively limited, with only a few studies addressing this area. Notable contributions include the works of Adhikari and Ghimire [4], who introduced image captioning in Nepali, presenting models such as plain encoder-decoder architectures and encoder-decoder with attention mechanisms. Subsequent research by Balkrishna Bal et al. [6] has focused on addressing dataset scarcity and linguistic intricacies in Nepali, proposing a novel approach utilizing a simplified transformer model in conjunction with CNN for feature extraction.

Further related to our work are hierarchical architectures that aim to mirror the hierarchy of language. Notable studies by Li et al. [1], [13], [14], [15] have explored hierarchical auto-encoders and different recurrent units for modeling sentences and words. Most closely aligned with our research is the work of Xu et al. [7] who generate multi-sentence descriptions using hierarchical models. However, our approach, while inspired by these works, simplifies the learning process and emphasizes the interplay between sentences for improved interpretability and generation accuracy.

3. Proposed Work

3.1 Model Architecture

This research utilizes the transformer architecture [7] as shown in Figure 1, wherein the feature vector extracted from InceptionV3[16] serves as input to the encoder component. Subsequently, the resulting embedded tokens are passed to the decoder section, tasked with generating Nepali captions.

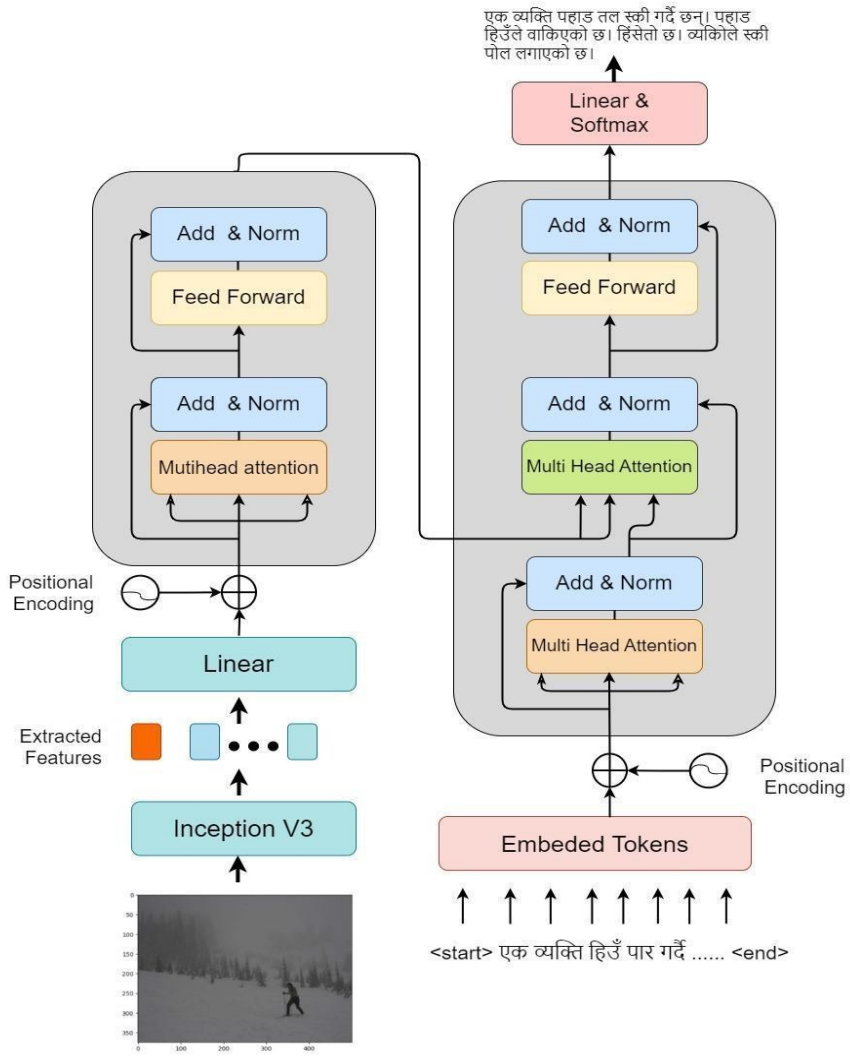

Figure 1. Transformer Architecture for Image Captioning

3.2 Methodology

The overall methodology of this task is shown in Figure 2 below.

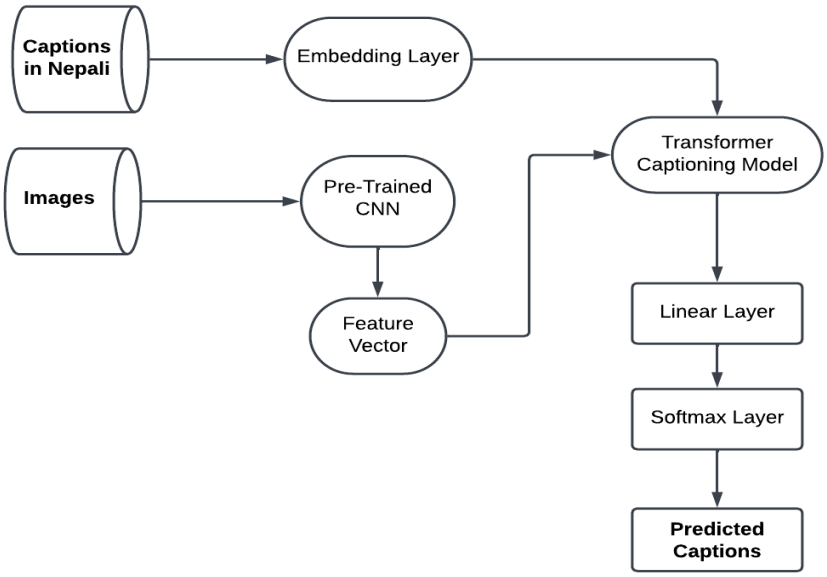

Figure 2. Methodology [5]

3.3 Data Set Collection

The dataset used in this research was created by augmenting the Stanford paragraph captioning dataset [1] with an additional 800 manually curated cultural dataset. The Stanford paragraph captioning dataset is a collection of images along with descriptive paragraphs (as shown in Figure 3, typically used for training models to generate captions for images. By augmenting this dataset with culturally relevant images and their corresponding captions, our research aimed to enhance the diversity and cultural richness of the dataset.

The augmentation process involved carefully selecting and adding 800 images that represent various cultural aspects, such as traditions, customs, landmarks, and artifacts, from the manually curated cultural dataset. These images were then paired with appropriate descriptive paragraphs to create new image-caption pairs.

In total, the augmented dataset comprises 20,350 image-caption pairs. This extensive dataset provides a rich resource for training and evaluating image captioning models, especially those intended to be culturally sensitive and inclusive.

3.3.1 Caption Translation to Nepali

Translating English captions into Nepali using Google Translator faces some difficulties:

- Google Translate lacks the capability to consider context, resulting in the loss of meaning when translating captions.
- Frequently, Google Translate produces grammatically incorrect sentences.
- The reliability of Google Translator varies depending on the language combination being translated.

3.3.2 Manual Correction and Annotation

To ensure accurate translations, we corrected errors in the Nepali captions generated by Google Translate. Adhering to Nepali grammar rules, we invested significant time, approximately 3 to 4 months, to rectify the entire dataset. Ambiguous or incorrect captions were manually corrected or removed.

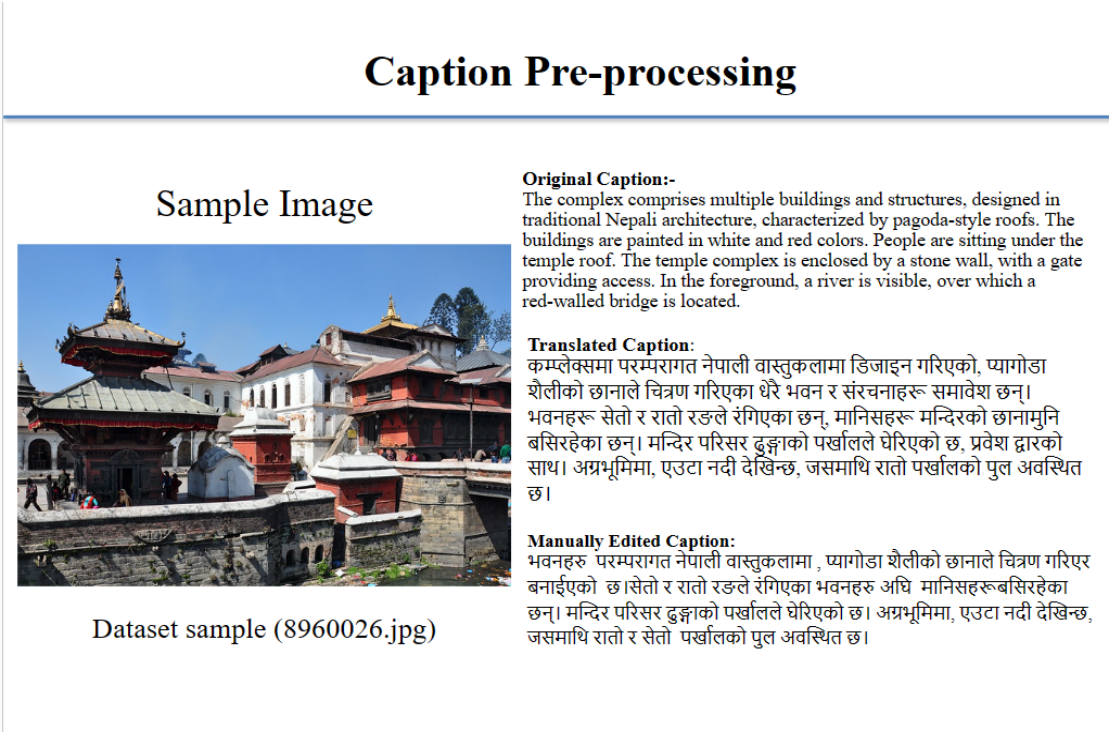

Figure 3. Caption Pre-processing

3.3.3 Data Cleaning

Translated and modified captions undergo further preprocessing to remove punctuation's and numeric values as in Figure 3. Unwanted characters and data are removed in this phase.

3.3.4 Generating Vocabulary and Text Vectorization

Unique words are extracted from the translated captions to create a text vocabulary. These words are then mapped to unique index values using Keras library's built-in Text Vectorizer function for text vectorization. The dataset captions underwent tokenization to separate words by spaces, yielding a vocabulary comprising all unique Nepali words in the dataset, totaling 14,022 words. Consequently, our vocabulary size amounted to 27000 words for the given Nepali dataset.

3.3.5 Data Set Creation

The cleaned caption data was divided into three sets: the training set, the validation set, and the test set (Figure 4). Captions were paired with their corresponding images and combined using the TensorFlow 'Dataset' library to create datasets for training and validation. This dataset was accessible on Kaggle.

In more detail, the dataset was divided into three subsets:

1.Training Set: This set contained 15,133 image-caption pairs as shown in Figure 5. The training set was typically used to train the image captioning model, allowing it to learn the relationship between images and their corresponding captions.

2.Validation Set: With 2,603 image-caption pairs, the validation set was used to tune the parameters of the model during training and to assess its performance on unseen data. It helped prevent overfitting and ensured that the model generalized well to new images and captions.

3.Test Set: The test set also contained 2,603 image-caption pairs and was used to evaluate the performance of the trained model after training and validation. It provided an independent assessment of how well the model could generate captions for new, unseen images.

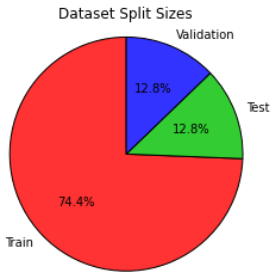

Figure 4. Dataset Spit

	Image_name	Paragraph	train	test	val
0	2356347	यसको अगाडि झ्यालहरूमा बारहरू भएको ठूलो भवन। भव...	False	True	False
1	2317429	एउटा सेतो गोलो प्लेट एउटा टेबलमा छ जसमा प्लास...	True	False	False
2	2414610	नीलो टेनिस पोशाकमा एउटी महिला हरियो टेनिस कोर्...	False	True	False
3	2365091	एउटा ठुलो रातो र सेतो रेलले ग्रामीण क्षेत्र जस...	True	False	False
4	2383120	धेरै सफा र सफा बाथरूम। सबै चीज सफा पोर्सिलेन स...	True	False	False

: 'यसको अगाडि झ्यालहरूमा बारहरू भएको ठूलो भवन। भवन अगाडि मानिसहरु हिडिरहेका छन्। भवनको अगाडि एउटा सडक छ जसमा धेरै कारहरू छन्।'

Figure 5. Dataset Format

3.4 Training and Testing

The Transformer model underwent training with various configurations, and only the results from the best-performing configuration were reported. Hyperparameters were selected empirically during this process.

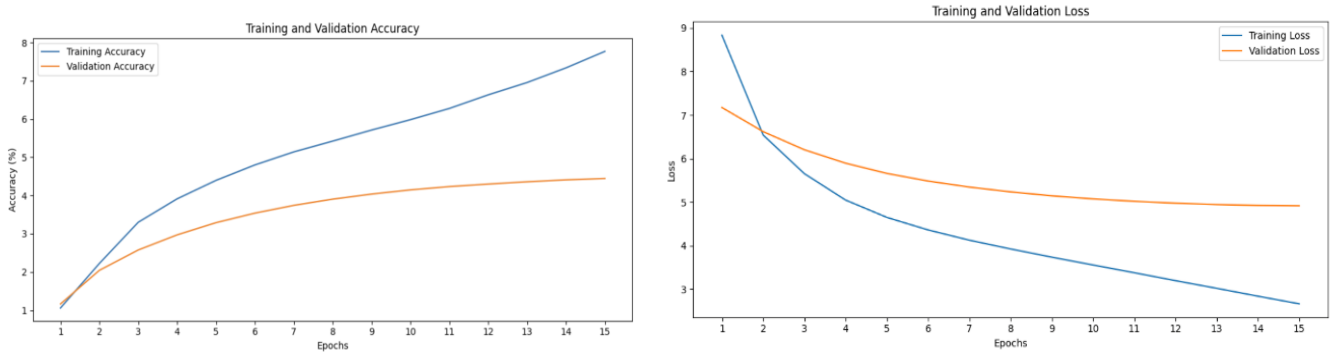

Figure 6. Transformer Accuracy and Loss Curve

The training and validation accuracy (Figure 6) consistently improved while the loss decreased for the first 10 epochs. However, beyond this point, the training curve began to stabilize, and the validation curve showed signs of increasing further. To prevent overfitting, the training process was halted at 15 epochs.

Table 1. Model Parameters for Transformer

Parameter	Value
Vocab Size	27000
Learning Rate	0.01
Batch Size	32
Dropout Rate	0.2
Optimizer	Adam
Epochs	15
Number of Head	8
Row size	8
Column size	8
max position encoding	25001

Transformer utilizes essential parameters such as vocabulary size, learning rate, batch size, dropout rate, and optimizer choice (Adam), along with training epochs as mentioned in Table 1. Additionally, it specifies Transformer-specific parameters like the number of attention heads, row and column size, and maximum position encoding. With a vocabulary size of 27,000, a learning rate of 0.01, and a batch size of 64, the model trains over 15 epochs with a dropout rate of 0.1 to prevent overfitting. It employs 8 attention heads and operates with a row and column size of 8, while the maximum position encoding is set to 25,001 for positional information handling in input sequences.

4. Result and Discussion

BLEU scores [17] (BLEU-1, BLEU-2, BLEU-3, and BLEU-4) are calculated using the NLTK bleu library.

Table 2 presents the obtained results of our work.

Table 2. BLEU Scores

Model	Epochs	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Inception(V3) + Transformer	15	0.23	0.35	0.53	0.59

4.1 Sample Result

The Figure 7 -10 illustrate the sample results observed

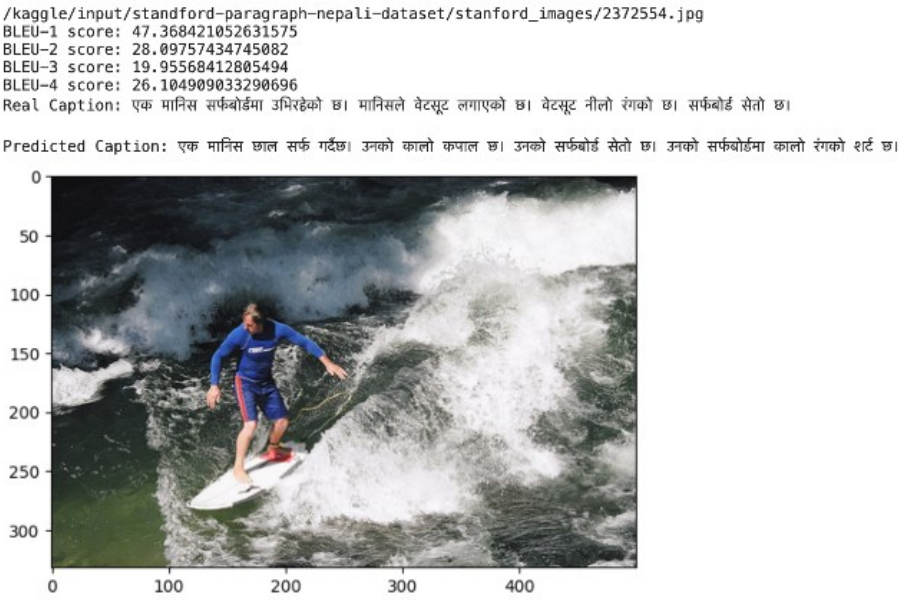

Figure 7. Sample Result 1

Figure 8. Sample Result 2

Figure 9. Sample Result 3

Figure 10. Sample Result 4

The transformer model correctly predicted the object, environment, relations and the actions in the images with coherent sentences.

5. Conclusion

In conclusion, our research work introduces a novel approach for generating detailed Nepali paragraphs to describe images, leveraging both visual and linguistic structures. The transformer's capacity to model long range dependencies on caption to effectively describe the

image in coherent fashion and focus on specific words enhances caption quality and training efficiency, offering promising avenues for future research in Nepali image captioning.

References

[1] Krause, Jonathan, Justin Johnson, Ranjay Krishna, and Li Fei-Fei. "A hierarchical approach for generating descriptive image paragraphs." In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 317-325. 2017.

[2] Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. "Attention is all you need." *Advances in neural information processing systems* 30 (2017).

[3] Ekman, Magnus. Learning deep learning: Theory and practice of neural networks, computer vision, natural language processing, and transformers using TensorFlow. Addison-Wesley Professional, 2021.

[4] A. Adhikari and S. Ghimire, “Nepali Image Captioning,” in 2019 Artificial Intelligence for Transforming Business and Society (AITB), Kathmandu, Nepal: IEEE, Nov. 2019, pp. 1–6. doi: 10.1109/AITB48515.2019.8947436.

[5] R. Budhathoki and S. Timilsina, “Image Captioning in Nepali Using CNN and Transformer Decoder,” J. Eng. Sci., vol. 2, no. 1, pp. 41–48, Dec. 2023, doi: 10.3126/jes2.v2i1.60391.

[6] Subedi, Bipesh, and Bal Krishna Bal. "CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning." In Proceedings of the 19th International Conference on Natural Language Processing (ICON), pp. 86-91. 2022. ”.

[7] Xu, Kelvin, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. "Show, attend and tell: Neural image caption generation with visual attention." In International conference on machine learning, pp. 2048-2057. PMLR, 2015.

[8] Chen, Minghai, Guiguang Ding, Sicheng Zhao, Hui Chen, Qiang Liu, and Jungong Han. "Reference based LSTM for image captioning." In Proceedings of the AAAI conference on artificial intelligence, vol. 31, no. 1. 2017. 3981-3987

[9] 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), DHAKA, Bangladesh: IEEE, Dec. 2020, pp. 1–5. doi: 10.1109/ICCIT51783.2020.9392709.

[10] Muhammad Shah, Faisal, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. "Bornon: Bengali image captioning with transformer-based deep learning approach." SN Computer Science 3 (2022): 1-16.

[11] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, “Image captioning in Hindi language using transformer networks,” Comput. Electr. Eng., vol. 92, p. 107114, Jun. 2021, doi: 10.1016/j.compeleceng.2021.107114.

[12] T.-Y. Lin et al., “Microsoft COCO: Common Objects in Context,” in Computer Vision – ECCV 2014, vol. 8693, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., in Lecture Notes in Computer Science, vol. 8693. , Cham: Springer International Publishing, 2014, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

[13] Y. Bazi, L. Bashmal, M. M. A. Rahhal, R. A. Dayil, and N. A. Ajlan, “Vision Transformers for Remote Sensing Image Classification,” Remote Sens., vol. 13, no. 3, p. 516-534, Feb. 2021, doi: 10.3390/rs13030516.

[14] R. Krishna et al., “Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations,” Int. J. Comput. Vis., vol. 123, no. 1, pp. 32–73, May 2017, doi: 10.1007/s11263-016-0981-7.

[15] X. Shen, B. Liu, Y. Zhou, and J. Zhao, “Remote sensing image caption generation via transformer and reinforcement learning,” Multimed. Tools Appl., vol. 79, no. 35–36, pp. 26661–26682, Sep. 2020, doi: 10.1007/s11042-020-09294-7.

[16] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA: IEEE, Jun. 2016, pp. 770–778. doi: 10.1109/CVPR.2016.90.

[17] 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 on Association for Computational Linguistics - ACL ’02, Philadelphia, Pennsylvania: Association for Computational Linguistics, 2001, p. 311. doi: 10.3115/1073083.1073135.

Author's biography

Nabaraj Subedi - I have currently completed a Bachelor in Electronics, Communication, and Information engineering from IOE,Paschimanchal Campus, with a keen interest in conducting research specifically in the application of deep learning techniques.

Nirajan Paudel - I am an engineering student pursuing electronics, communication and information engineering since 2019 in Paschimanchal Campus, IOE, TU and expect to graduate in 2024, July. I have a deep interest in the field of AI and machine learning.

Manish Chhetri - I am an engineering student pursuing Electronics, communication and Information engineering from 2019 AD and expect to graduate in 2024 AD. Throughout the studies, I have developed passion in the field of AI and Machine Learning.

Sudarshan Acharya, I am an engineering student pursuing Electronics, communication and Information engineering from 2019 AD and expect to graduate in 2024 AD. Throughout the studies, I have developed passion in the field of Web Development and Machine Learning.

Nabin Lamichhhane, Deputy Head of Department of Electronics and Computer Engineering, IOE, TU, Nepal