https://doi.org/10.3126/juem.v3i1.84867

AUTOMATIC NEPALI IMAGE CAPTIONING USING CNN-TRANSFORMER MODEL

Swarup Singh Tharu[1*], Savin Basnet[1], Arun Thapa[1], Prashant Poudel[2]

[1] Department of Computer Engineering, United Technical College, Bharatpur -11, Chitwan, Nepal [2] Faculty of Computer Engineering, United Technical College, Bharatpur-11, Chitwan, Nepal

Corresponding Author : swaruptharu123@gmail.com (S. S. Tharu)

Submission Date: 21 July 2025 Accepted Date: 3 August 2025
Revised Date: 31 July 2025 Published Date: 30 Sept. 2025

Cite this: Tharu, S. S., Basnet, S., Thapa, A. and Poudel, P. (2025)., Automatic Nepali Image Captioning using CNN-Transformer Model, JUEM 3(1), 189 – 197, https://doi.org/10.3126/juem.v3i1.84867

ABSTRACT

Image captioning has gained significant attention, with most of the research efforts directed toward the English language. While some work has been explored in regional languages such as Hindi and Bengali, Nepali remains largely underrepresented in this domain. Furthermore, publicly accessible Nepali-language datasets for image captioning are extremely limited. This study leverages an existing pre-trained dataset that includes Nepali image captions and employs deep learning methods to automatically generate descriptions in the Nepali language. The architecture used integrates a Convolutional Neural Network (CNN) for image understanding and a Transformer model for sequence generation. In our approach, EfficientNetB0, a pre-trained CNN model, is utilized to extract high-level features from images. These features are then fed into the Transformer, which generates the corresponding captions in Nepali. The experimental results demonstrate encouraging performance, suggesting the approach is effective and holds potential for further refinement in future research.

Keywords: *Deep Learning, Pré-trained Dataset, Nepali Image Captions, Convolutional Neural Network (CNN),Transformer Model, EfficientNetB0, Feature Extraction, Sequence Generation*

1. Background

Image captioning has evolved significantly, progressing from basic object detection to advanced scene understanding, largely due to the rise of deep learning. Tools such as Convolutional Neural Networks (CNNs) have enabled automatic extraction of features from images, especially when trained on large datasets like ImageNet. This has led to the development of powerful applications, including commercial systems like amazon Rekognition and Google Vision AI, which can recognize complex image content with high precision.

Combining computer vision with natural language processing has opened new possibilities, such as generating automatic image descriptions that benefit both general users and individuals with visual impairments. This project focuses on improving image captioning for the Nepali language, using CNNs for visual analysis and Transformer models for language generation—modern alternatives that outperform earlier RNN-based approaches (Satti et al., 2023).

Despite the progress, image captioning still faces challenges, including time-consuming manual labeling, difficulty processing complex scenes, and a strong bias toward global languages like English. This creates accessibility barriers for users in regions where local languages like Nepali are spoken. To tackle these issues, the project proposes building a web-based image captioning platform that detects objects and generates natural-sounding Nepali sentences using a Transformer-based model (Shrestha et al., 2021). The goal is to promote AI inclusivity, improve digital accessibility, and support native language content development (Adhikari & Ghimire, 2019).

The system is designed for diverse use cases—education, media, accessibility, and public service—and includes APIs for integration into other platforms (Budhathoki & Timilsina, 2023). However, it also faces limitations like reliance on quality training data, computational resource needs, and difficulty handling complex or overlapping image elements. Building robust real-time Nepali datasets remains a key challenge (Subedi et al., 2024).

2. Related work

Numerous studies have been conducted in the field of image captioning, particularly in well-resourced languages like English. Considerable research has also been carried out in Hindi and Bengali, which share linguistic similarities with the Nepali language. The only known research effort in Nepali was done by (Adhikari, 2019), who developed two encoder-decoder models with and without visual attention. Their models used ResNet-50 as the encoder and LSTM/GRU as the decoder. These models were trained on the MS COCO dataset after translation and preprocessing.

In contrast, (Mishra et. al, 2021) proposed an advanced transformer-based encoder-decoder framework for Hindi image captioning. Their model employed ResNet-101 to extract image features and used a Transformer as the decoder. They addressed the limitations of traditional RNNs and introduced a stacked attention mechanism to more effectively convert image features into descriptive sentences. Their work reported BLEU-1, BLEU-2, BLEU-3, and BLEU-4 scores of 62.9, 43.3, 29.1, and 19.0, respectively.

Similarly, in the Bengali language domain, (Palash et. al.,, 2021) developed image captioning models using CNNs and Transformers. They used ResNet-101 for feature extraction, pretrained models like InceptionV3 and Xception. They trained and tested their model on the BanglaLekha dataset, whereas initially used the Flickr8k dataset, later expanded using BanglaLekha. The BLEU-1, BLEU-2, BLEU-3, BLEU-4, and METEOR scores reported by Palash et al. were 0.694, 0.580, 0.505, 2.22e-308, and 0.337, which outperform the results from (Mishra et. al.,, 2021).

Despite both approaches utilizing Transformer-based decoders and CNN encoders, the method proposed by (Palash, Nasim, Saha, & Afrin, 2021) incorporates a richer set of image encoders—such as ResNet-101, InceptionV3, and Xception—which may help capture more diverse and detailed visual features. Additionally, their use of language-specific datasets like BanglaLekha and Flickr8k

potentially offers improved linguistic context for caption generation. In contrast, (Mishra et. al., 2021)follow a similar architectural pattern but with fewer encoder variants and more general datasets. Reported results indicate that the former approach achieved relatively higher BLEU-1 to BLEU-3 and METEOR scores, suggesting better performance in generating accurate and semantically meaningful captions, though further comparative analysis would strengthen these findings.

Additionally, focused on remote sensing images (like satellite images) and used a Transformer-based decoder for caption generation (Shen et al., 2020). They added semantic and spatial features to each layer of the decoder to improve caption quality. Their datasets included Sydney Dataset, RSICD, and UCMDataset.

From these studies, it's clear that Transformer models generally perform better than older CNN-RNN combinations. However, Nepali still lacks Transformer-based image captioning systems and publicly available caption datasets (Subedi & Bal, 2022). So, this research takes the first step in that direction by implementing a CNN-Transformer model for Nepali image captioning.

3. Methodology

Dataset

For this research, we utilized a preprocessed dataset developed by Subedi & Bal (2022), specifically curated for image captioning in the Nepali language. As mention in report , a pre-processed version developed based on the publicly available Flickr8k and Flickr30k image captioning datasets, which originally consist of English captions was used. Since no standardized Nepali-language image captioning dataset exists, the dataset had to be extensively modified and localized. The English captions were translated into Nepali using the Google Translate API. Due to API limitations and translation inconsistencies, this process involved translating one caption at a time and writing the results into a new file.

To ensure linguistic and contextual accuracy, all translated captions underwent manual correction and annotation, addressing issues such as mistranslation, grammar errors, and context mismatches. These corrections were manually performed in the lab, with particular attention paid to caption completeness and image-caption alignment. Captions with incomplete translations or missing entries were either corrected or removed, and inconsistencies in image ID naming were handled through frequency-based filtering and manual cross-verification.

Further preprocessing steps included removing punctuation marks, numeric characters (both English and Nepali), and special symbols, along with generating a comprehensive vocabulary. Uniquely, unlike most English-based models that exclude rare words, the Nepali caption dataset retained all unique words—including less frequent ones—to preserve semantic richness. This resulted in vocabularies containing over 14,000 and 35,000 unique words for the Flickr8k and Flickr30k datasets, respectively.

The dataset was then vectorized using the Keras TextVectorization layer and split into training (approximately 75–80%), validation, and test sets. Each caption was mapped to its corresponding image using TensorFlow's Dataset API to ensure compatibility with deep learning pipelines. Although the base image datasets were open-source, this extensive language localization, manual refinement, and formatting effort provided a significant level of originality and made the dataset suitable for image captioning in the Nepali language.(B. Subedi & Bal, 2022).

Working explanation

The methodology in this research revolves around building a custom deep learning model to generate image captions in the Nepali language. Since no large-scale Nepali image captioning datasets are publicly available, the process begins with creating and preparing a suitable dataset. Once the dataset is ready, the next step is to design and implement a model that can analyze images and produce accurate descriptions in Nepali. The system works by first passing an image through a Convolutional Neural Network (CNN), which extracts key features such as shapes, objects, and patterns. Instead of using a traditional LSTM model, this research uses a Transformer for generating text, which enables better context understanding and faster training. The CNN extracts the image features, and the Transformer uses those features to generate fluent and contextually meaningful captions.

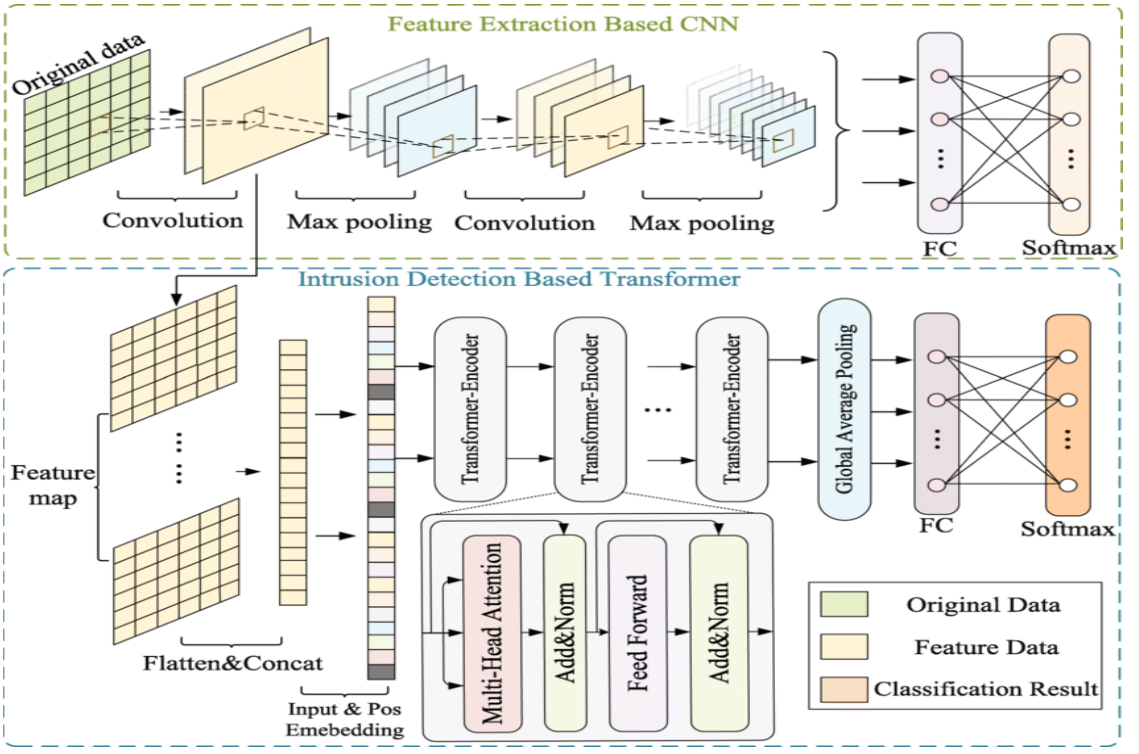


Figure 1: Architecture of system

A Convolutional Neural Network (CNN) is a deep learning model that is specifically designed to analyze visual data like images. It uses a combination of layers—including convolutional layers for feature detection, pooling layers for reducing the spatial size, and fully connected layers for making final predictions. CNNs are excellent at identifying local patterns such as edges, textures, and shapes, which helps in understanding the structure and content of images. Unlike traditional neural networks that treat the image as a flat input, CNNs preserve the spatial relationships in the image. This makes them more efficient and effective for tasks such as image classification, object recognition, and detection. In this research, the CNN is used as the encoder, extracting meaningful features from the input image to pass to the captioning model (Bhensle et al., 2025).

The Transformer is a state-of-the-art model for handling sequential data, widely used in natural language processing. It uses a self-attention mechanism to understand the relationships between words in a sentence, regardless of their position. Unlike RNNs or LSTMs, Transformers can process all words in a sequence at once, making them much faster and more scalable. In this project, a CNN-Transformer architecture is used, where the CNN handles the visual input and the Transformer generates the output sentence in Nepali. This combined approach benefits from the CNN's strong visual understanding and the Transformer's advanced language modeling. The model can generate accurate, natural-sounding descriptions of images, and its performance makes it ideal for applications like accessibility, content creation, and education in the Nepali language (Rawat et al., 2024).

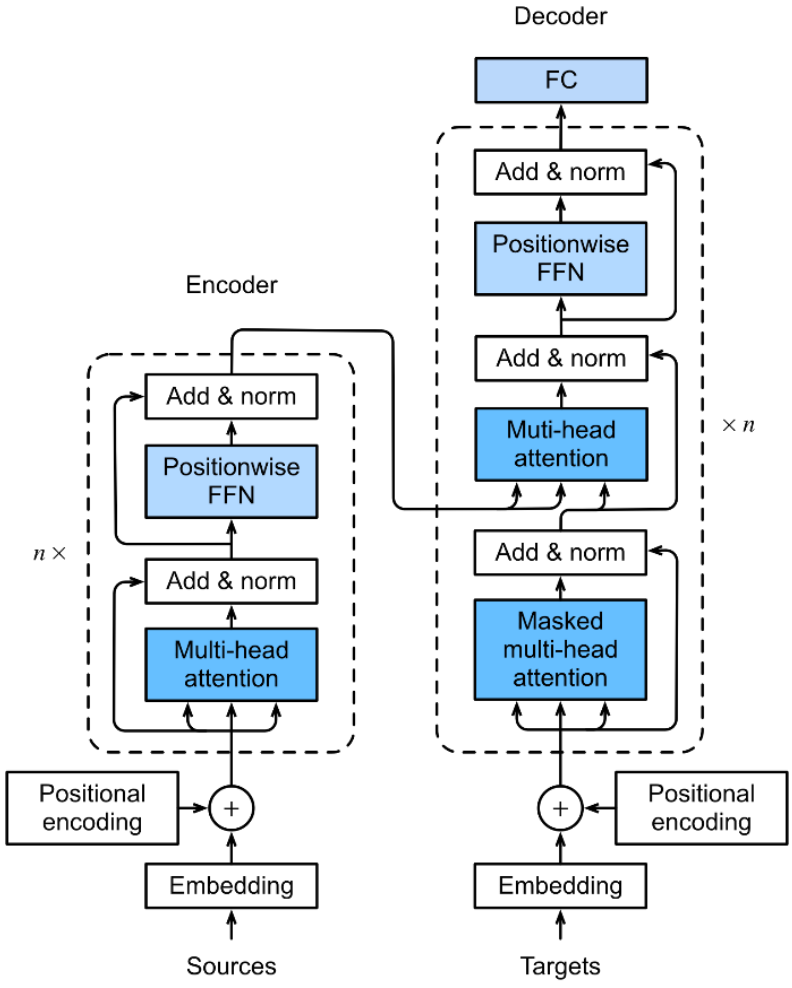


Figure 2: Transformer Architecture

Before training the model, Nepali captions must be tokenized in words or subwords. Tokenization is performed using a custom tokenizer implemented with the Keras TextVectorization layer,

configured to accommodate the unique script and grammatical structure of the Nepali language. The tokenizer is adapted to the entire training corpus and builds a vocabulary limited to the 10,000 most frequent words for computational efficiency. It also standardizes the text by converting all tokens to lowercase and removing punctuation and special characters. Particular attention is given to handling compound words, suffixes, and less predictable linguistic patterns specific to Nepali. Each caption is then padded or truncated to a fixed length of 25 tokens to ensure consistent input dimensions. This step is essential for allowing the Transformer model to learn meaningful patterns in Nepali sentence structure.

In overall model setup, several key parameters are defined to optimize the image-text processing pipeline. All input images are resized to 299x299 pixels to match the input requirements of the CNN encoder. The extracted image features and tokenized text data are embedded in a 512-dimensional space, allowing the model to relate visual and textual modalities effectively. The Transformer's feed-forward network is also configured with 512 units per layer. To manage computational load and ensure efficient training, a batch size of 64 is used, and the model is trained for 100 epochs. Additionally, TensorFlow's tf.data.AUTOTUNE is utilized to dynamically optimize data loading and pipeline performance, leading to faster and more efficient model training.

.

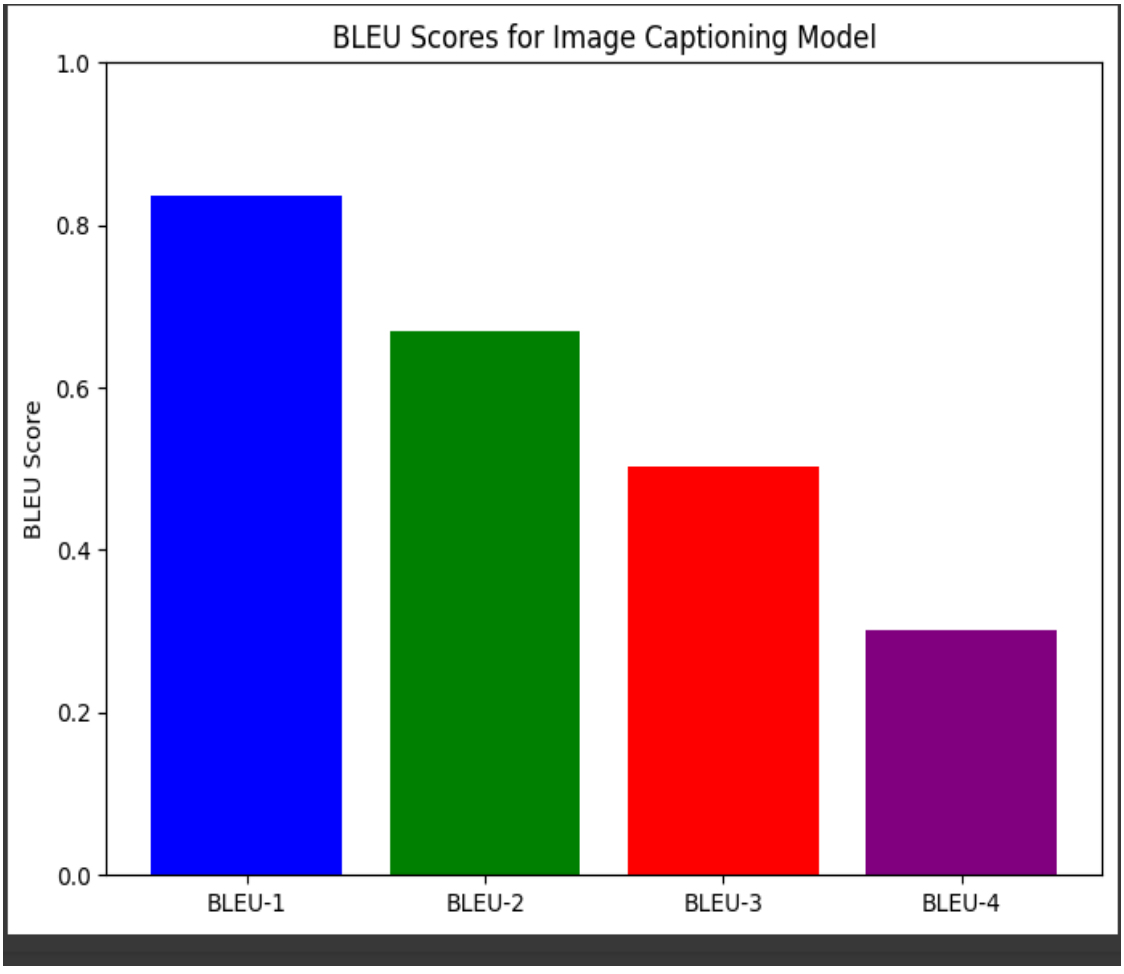
4. Result and Discussion

The outcome of this research includes the development of a Nepali image caption dataset and an evaluation of the model's performance through experimental results. The dataset contains over 40,000 image-caption pairs in Nepali, which are split into training, validation, and testing sets. To assess caption quality, we used the BLEU (Bilingual Evaluation Understudy) metric, which compares machine-generated captions with reference captions. Specifically, BLEU-1 to BLEU-4 scores evaluate how well single words, bigrams, trigrams, and 4-word sequences match between generated and reference text. Scores range from 0 to 1, where values above 0.4 are considered good, and scores between 0.6 and 0.7 are considered excellent (google bleu score Evaluating models, 2022).

In this work, various models were tested: Our mode Caption.ai (optimized CNN-Transformer with parameter tuning) was compare with other model like Model A (baseline CNN-Transformer) (B. Subedi & Bal, 2022). The BLEU scores were calculated on the full test set using the NLTK BLEU library.

Model A achieved BLEU scores of 0.52 (B1), 0.42 (B2), 0.37 (B3), and 0.34 (B4) (B. Subedi & Bal, 2022), whereas the improved Caption.ai model showed further enhancements with scores of 0.83 (B1), 0.67 (B2), 0.50 (B3), and 0.30 (B4). The main reason for the BLEU-1 gain is better preprocessing of Nepali text (especially tokenization and vocabulary selection) combined with parameter tuning and a stronger CNN encoder, which together led to more accurate word predictions.

When compared with existing works, (Mishra, Dhir, Saha, puspak, & singh, 2021)reported BLEU scores of 0.66, 0.55, 0.47, and 0.40. While Shah et al.'s model (developed in Hindi language) slightly outperforms our Caption.ai in BLEU-4 (0.40 vs 0.30), Caption.ai achieves significantly higher scores in BLEU-1, BLEU-2, and BLEU-3, indicating that the generated sentences contain more accurate and relevant word sequences, even though some variations in word order remain.




```
Requirement already satisfied: nltk in /usr/local/lib/python3.11/dist-packages (3.9.1)
Requirement already satisfied: click in /usr/local/lib/python3.11/dist-packages (from nltk) (8.2.1)
Requirement already satisfied: joblib in /usr/local/lib/python3.11/dist-packages (from nltk) (1.5.1)
Requirement already satisfied: regex>=2021.8.3 in /usr/local/lib/python3.11/dist-packages (from nltk) (2024.11.6)
Requirement already satisfied: tqdm in /usr/local/lib/python3.11/dist-packages (from nltk) (4.67.1)
BLEU-1 Score: 0.8352
BLEU-2 Score: 0.6700
BLEU-3 Score: 0.5020
BLEU-4 Score: 0.3024
```

Figure 3: BLEU score

In summary, the improved Caption.ai significantly outperforms Model A, demonstrating the effectiveness of parameter tuning and dataset quality in enhancing model performance. This study also highlights that the proposed CNN-Transformer model performs competitively with state-of-the-art approaches in other languages, such as Hindi, and serves as a strong benchmark for future research in Nepali image captioning.

Sample results obtained from our proposed models

Analysis Results
दृश्य विवरण
एउटा केटा फुटबल खेलिरहेको छ ।

Analysis Results
दृश्य विवरण
दुई केटाहरू समुद्र तटमा सँगै पोज दिइरहेका छन् ।

Analysis Results
दृश्य विवरण
एउटा मानिस पहाडको टुप्पोमा उभिएको छ ।

Figure 4 Caption generation in Nepali

5. Challenges and solution

During the development of the Nepali Image Caption Generation system, several challenges were encountered. These included a limited and repetitive dataset, missing annotations, and grammatical issues that reduced caption quality. Additionally, the model initially underperformed due to poor hyperparameter settings and lack of training enhancements like regularization and data augmentation.

To address these, the team implemented several solutions: they cleaned and expanded the dataset, used data augmentation techniques, applied attention mechanisms, optimized training settings (like learning rate and batch size), and incorporated regular testing and user feedback to improve accuracy.

For future improvements, the project plans to:

- Train on a larger, high-quality dataset
- Extend to real-time video captioning
- Enable voice-based inputs
- Integrate with wearable devices for accessibility
- Add multilingual and cross-language support for broader usability.

6. Conclusion

The Image Caption Generation project has successfully achieved its goal of connecting people with visual content through meaningful captions in the Nepali language. We have fully developed the system's core framework, including the image analysis module and the Nepali caption generation feature, both of which are now functioning accurately and effectively. These components have laid the foundation for an accessible and user-friendly platform that seamlessly combines traditional Nepali knowledge with modern AI technologies.

The system now accurately recognizes images and generates relevant captions in Nepali, helping users understand visual content in their native language. Image analysis has been refined to capture detailed visual elements, and the caption generation module has been enhanced to produce

grammatically correct and contextually appropriate descriptions.

With all planned features implemented, tested, and validated, the project has been completed successfully. Our objectives have been fully met, and the platform stands as a powerful tool to promote and preserve the Nepali language while making digital visual content more accessible to Nepali-speaking communities.

Reference

Adhikari, A., & Ghimire, S. (2019). Nepali Image Captioning. Artificial Intelligence for Transforming Business and Society (AITB), 1–6. https://doi.org/10.1109/aitb48515.2019.8947436

Bhensle, A. C., Patra, J. P., & Samal, S. (2025). An Efficient Hindi Image Captioning with Transformer Model. In Advances in intelligent systems research/Advances in Intelligent Systems Research (pp. 32–43). https://doi.org/10.2991/978-94-6463-738-0_4

Budhathoki, R., & Timilsina, S. (2023). Image captioniNg in Nepali using CNN and Transformer Decoder. Journal of Engineering and Sciences, 2(1), 41–48. https://doi.org/10.3126/jes2.v2i1.60391

Mishra, S. K., Dhir, R., Saha, S., Bhattacharyya, P., & Singh, A. K. (2021). Image captioning in Hindi language using transformer networks. Computers & Electrical Engineering, 92, 107114. https://doi.org/10.1016/j.compeleceng.2021.107114

Palash, M. a. H., Nasim, M. a. A., Saha, S., Afrin, F., Mallik, R., & Samiappan, S. (2021). Bangla Image Caption Generation through CNN-Transformer based Encoder-Decoder Network. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2110.12442

Rawat, S., Manwal, M., & Purohit, K. C. (2024). Simplifying Image Captioning in Hindi with Deep Learning. 2024 International Conference on Computer, Electronics, Electrical Engineering & Their Applications (IC2E3), 1–7. https://doi.org/10.1109/ic2e362166.2024.10827396

Satti, S. K., Rajareddy, G. N. V., Maddula, P., & Ravipati, N. V. V. (2023). Image Caption Generation using ResNET-50 and LSTM. 2023 IEEE Silchar Subsection Conference (SILCON), 8, 1–6. https://doi.org/10.1109/silcon59133.2023.10404600

Shen, X., Liu, B., Zhou, Y., & Zhao, J. (2020). Remote sensing image caption generation via transformer and reinforcement learning. Multimedia Tools and Applications, 79(35–36), 26661–26682. https://doi.org/10.1007/s11042-020-09294-7

Shrestha, A., Kuikel, S., Neupane, S., & Lamichhane, E. N. (2021). A reflection on machine translation process from Nepali to English.

Subedi, B., & Bal, B. K. (2022, December 1). CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning. ACL Anthology. https://aclanthology.org/2022.icon-main.12/

Subedi, N., Paudel, N., Chhetri, M., Acharya, S., & Lamichhane, N. (2024). Nepali Image captioning: Generating coherent Paragraph-Length descriptions using transformer. Journal of Soft Computing Paradigm, 6(1), 70–84. https://doi.org/10.36548/jscp.2024.1.006