**eISSN: 2976-1190**

**Journal of Engineering and Sciences**

Vol. 2 Issue 1 November 30, 2023

*Journal homepage:* ***Journal of Engineering and Sciences (nepjol.info)***

*Publisher: Research Management Cell, Pashchimanchal Campus, Lamachour, Pokhara*

# Image Captioning in Nepali Using CNN and Transformer Decoder

Rabin Budhathoki[1], Suresh Timilsina[1]

[1]*Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal*

*(Manuscript Received:31/08/2023; Revised:21/11/2023; Accepted: 12/11/2023)*

## Abstract

Image captioning has attracted huge attention from deep learning researchers. This approach combines image and text-based deep learning techniques to create the written descriptions of images automatically. There has been limited research on image captioning using the Nepali language, with most studies focusing on English datasets. Therefore, there are no publicly available datasets in the Nepali language. Most previous works are based on the RNN-CNN approach, which produces inferior results compared to image captioning using the Transformer model. Similarly, using the BLEU score as the only evaluation metric cannot justify the quality of the produced captions. To address this gap, in this research work, well known "Flickr8k" English data set is translated into Nepali language and then manually corrected to ensure accurate translations. The conventional Transformer is comprised of encoder and decoder modules. Both modules contain a multi-head attention mechanism. This makes the model complex and computationally expensive. Hence, we propose a noble approach where the encoder module of the Transformer is completely removed and only the decoder part of the Transformer is used, in conjunction with CNN, which acts as a feature extractor. The image features are extracted using the MobileNetV3 Large while the Transformer decoder processes these feature vectors and the input text sequence to generate appropriate captions. The system's effectiveness is measured using metrics, such as the BLEU and Meteor scores, to judge the caliber and precision of the generated captions.

*Keywords*: BLEU; CNN; Flickr8k; Meteor; MobilentV3; RNN; Transformer

## 1. Introduction

A machine learning technique called image captioning automatically converts the provided image into a text description. The machine needs to produce the textual description of a given image [1]. It is one of the very recent and challenging works in Artificial Intelligence. It integrates Computer Vision with Natural Language Processing, two distinct branches of artificial intelligence [2]. The model must be able to identify objects in a picture, understand their relationships, and represent them in a natural language in order to address this problem [1]. There must be a method for extracting and describing an image's features in natural language.

The visually challenged population of Nepal can be greatly aided by the usage of image captions in Nepali. Similarly, an enormous number of images found on the internet can be automatically labeled in our own language, which can be useful in e-commerce sites using this proposed method.

Caption generation is one of the sectors where the usage of deep learning has expanded due to intensive field study and better results.

Previous works on caption generation have mainly been done in English, as insufficient datasets are available in other languages. Most work is based on CNN encoder and RNN decoder, which can only recall previous data and produce subpar results over long sequences. Recurrent Neural Network is prone to vanishing gradient problems and due to its sequential nature, it requires more time and resources to process the given data. Similarly, it cannot capture the contextual information from long sentences and thus suffers long-range dependency problems. Transformer, on the other hand, utilizes a self-attention mechanism to identify the contextual information in a sentence using self-attention mechanism. It can process all the input words simultaneously, requiring less time than the RNN-CNN model. The Transformer can correctly process longer sentences and produce longer captions [3]. Hence, the proposed method uses the Transformer model to generate the captions with better results. The recurrent Neural Network is replaced by the transformer, which increases the exploitation of contextual information from the input sentences and can work at a higher speed due to its parallelization. The input image must be first converted into feature vectors to extract the image's key details using

*Corresponding author. Tel.: +977- 9860565068,
E-mail address: talk2riban@gmail.com

Convolutional Neural Net. The input image must be first converted into feature vectors to extract the image's key details using Convolutional Neural Network (CNN).

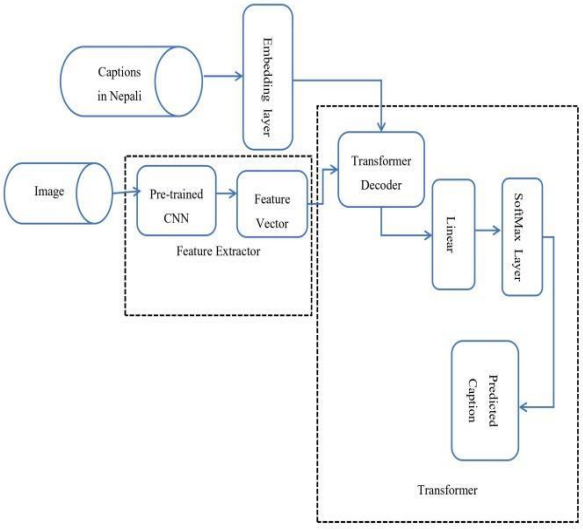


Figure 1: Work Flow of the model

Different pre-trained CNN models like Inceptionv3, ResNet50, VGG-16, VGG-19, Dense Net, etc. Even though InceptionV3 has the highest accuracy, MobilenetV3 Large is used to extract image feature vectors because of its smaller size and higher speed with comparable accuracy. The output from the MobilenetV3 is then given to the Transformer model to predict the caption in the Nepali Language.

The main contributions of this research work are as follows:

- Generation of manually corrected Nepali captions from Flickr8k English dataset.
- Simplification of the Transformer model by utilizing only the decoder part of the original Transformer.

## 2. Related Work

Numerous studies have been done in the area of image captioning. Different deep-learning techniques have been adopted to generate captions in various languages.

In paper [1], Xu et al. used an attention mechanism in the image captioning model. The attention mechanism concentrates on the pertinent area of the image to generate captions. Attention-generated captions are better than the model without attention, which the BLEU and Meteor Score validated. The main contribution of this work is the application of attention mechanisms in the field of image captioning. Xiao et al. [5] have used a number of layers of LSTM stacked on top of each other for image captioning, where the functions of the encoder and decoder are split up using a complex hierarchical structure. The proposed system can effectively utilize deep networks 'capacity for representation to combine high-level linguistic and visual semantics when producing captions. The result obtained was better than that of Xu et al. The use of multilayered LSTM made the model complex and increased the training time. In the paper [6], the Flickr8k data set was utilized by the authors to create captions in Nepali using an encoder-decoder approach. The data set was translated into Nepali using Google Translator. This work contributed by generating captions in the Nepali Language. The main limitation of this research work is the use of translated captions directly from the Google translator, which results in poor captions. Ai Momin Faruk et al. proposed a system capable of generating captions in the Bangla language. They used a bidirectional gated recurrent unit on the decoder side of the system. Argmax and Beam Search methods generated high-quality captions [2]. The author's contribution is using the RNN-CNN model to generate captions in the Bangla Language. The quality of predicted captions is limited by the use of the encoder-decoder model.

The transformer was proposed by Vaswani et al. [7] for machine translation based on attention. This model is based on an attention mechanism that translates English to French and German, surpassing all the traditional models in evaluation matrix scores. This work paved the way for transformer-based models in machine translation and other computer vision-related work, such as image captioning. Inspired by the work in [7], Guang Li et al. have used transformer-based models in image captioning. They have used a number of attention layers and feed-forward layers to eliminate the use of RNN in the decoder of the image captioning model. The use of a transformer enabled the identification of the visual and semantic information in parallel, thereby increasing the performance of the captioning model [3]. This work made a huge contribution to the image captioning field by using Transformer successfully in the image captioning field. Mishra et al. [8] have created a Hindi dataset from the MS COCO dataset and used it to generate captions in the Hindi language using a transformer. Hindi dataset from the existing English dataset is created using Google translator and further correcting the translated captions manually. The author showed that the Transformer model can be used to generate captions in Hindi language. However, they only used the

BLEU score to evaluate the system's performance. In the paper [9], the authors used a pre-trained transformer and CNN to generate captions in Arabic Language using the Flickr8k dataset. The captions generated were better than previous models. The authors contributed by achieving higher scores of evaluation metrics than the conventional model. This work is limited to the accuracy of the pre-trained Transformer model.

Most previous works have used variations of the encoder-decoder approach to generate captions. These types of models cannot generate longer meaningful captions. Similarly, the research works are mainly focused on the English Language. Besides the English dataset, image captioning has not fully explored the more powerful Transformer model. Similarly, previous works on other languages have particularly used only one evaluation metric to justify their results. Hence, we have utilized the transformer model's decoder part to generate Nepali captions. Similarly, we have used two different metrics to validate our results.

## 3. Methodology

The model used in this research consists of two main parts: a convolutional neural network and a transformer decoder.

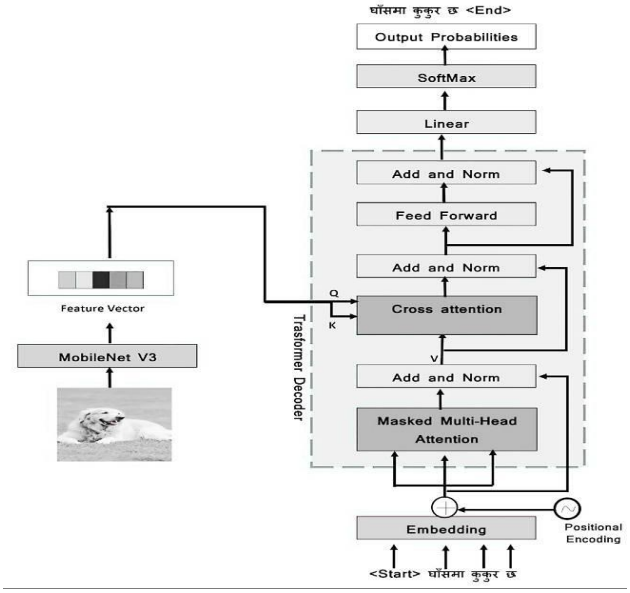


Figure 2: Block Diagram of the Image Captioning

The CNN used in this model is MobileNetV3 Large; its job is to extract the input image's feature maps. The output obtained from the feature extractor is then directly given to the cross-attention module of the Transformer decoder. The Transformer decoder then processes the input feature vector of an image and the sequence of texts to predict the captions of the given image.

### *3.1 MobileNet V3*

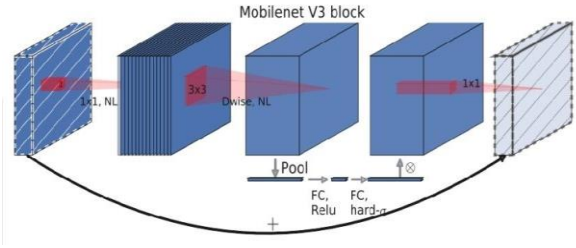


Figure 3: Architecture of MobileNetV3 [12]

MobileNetV3 is highly suited for devices between accuracy and efficiency. To accomplish this constrained computational power, since it is specifically optimized for striking a fair balance b effectiveness, the architecture uses a number of design decisions and methodologies. As shown in Figure 3, it contains depth-wise separable convolutions that divide the standard convolution procedure into two distinct steps: a depth-wise convolution filters each input channel independently, followed by a point-wise convolution that combines the output of the depth wise convolution to produce the final output channels. Prior to using depth-wise convolutions, the number of channels is increased. After applying the convolutions, the channels are projected back to a smaller number. The network can efficiently learn complicated patterns because of this extension and projection. Similarly, compared to conventional RELU activation, MobileNetV3 uses activation functions, including Hard-Swish and Leaky RELU. The quicker training and inference times are a result of these activations.

### *3.2 Transformer*

As shown in Figure 2, the transformer decoder contains N number of decoder layers and two different multi-head attention, linear layer, and feed-forward layers. The decoder's function is to generate the highest probability tokens to generate captions by combining the feature map and the output text generated thus far. The word embedding layer converts the input texts into the index of the input. The position embedding layer then provides information on the position of the input words. The output from the position embedding layer is known as the position vector.

Self-attention computes attention weights of the input image features and generates an output vector with encoded instructions on how each part of the

image should pay attention to every other part of the image in the sequence.
The calculation of self-attention is given by [7]:

$$Scaledattention = softmaxof((QK^{\wedge}T)/\sqrt{dim} \quad (1)$$

Where,
Q = query vector, $k^T$ = transpose of Key vector and dim = dimension of sequence.
The cross-attention layer is similar to the multi-head attention, with the major difference being that it gets the input from the feature extractor as query Q matrix and the key K matrix along with the value matrix V obtained from the masked attention module. The value matrix is obtained from the previously predicted text by the decoder. The cross-attention combines the visual features and the text sequence predicted by the decoder.
The output from the linear layer is the single score of each word. The soft-max then converts each value to the probability score. The word with the highest probability score is chosen at any time which is the predicted word for the given image.

### *3.3 Dataset*

In deep learning, a sufficient amount of data is needed. The proposed model requires images and their corresponding captions. Image Captioning requires two different data, viz., image and captions. The images, along with the captions, are collected from Kaggle [15]. The flickr8k dataset contains 8000 images and 40000 captions in English. Each image contains five different captions. The original flikr8k dataset is officially divided into three sets containing 6000 images as training, 1000 as testing, and 1000 as validation sets.

### *3.4 Image Pre-processing*

Before supplying the images to the model, they must be preprocessed due to their various sizes. Given that we are using MobileNetV3 Large, an image-net pre-trained model, the training and validation images are resized into 224 x 224 pixels and then given to the MobileNet for feature extraction [12].

### *3.5 Captions Pre-processing*

Every image has five different captions in English. All the training and validation group captions must be translated into Nepali. To convert the captions into Nepali Language, Google Translator is used. Google Translator does not provide meaningful translations of many captions. Hence, all the translated captions are manually checked and corrected. The figure below shows Some of the manually corrected translation samples.

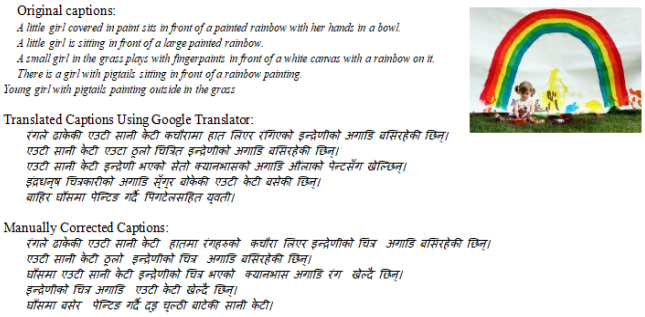


Figure 4: Sample image 1

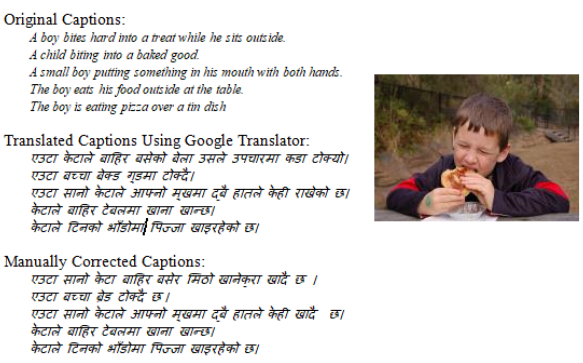


Figure 5 Sample image 2

### *3.6 Data Augmentation*

The model takes one image and corresponding caption at a time. Since there are five captions per image, the image is duplicated five times to fit the number of captions. Hence, the size of the dataset changes to 40000 captions and 40000 images.

### *3.7 Training and Testing*

The feature extractor extracts the higher-level feature map of the input, which is then given to the cross-attention layer of the decoder, which provides each of the features a weight in the form of an Attention Score. The higher the attention score, the more focus the decoder gives at that part of the image to generate output words. The ground truth caption and <start> and <end> tokens are then supplied to the masked attention layer and the positional values. The masked attention masks all the future tokens so that the decoder only focuses on the present and past tokens. The soft-max layer then gives a high value to the most probable words from the available vocabulary to generate the first tokens related to the input image. The predicted output is then compared with the first token of the ground truth to compute loss. This loss is then minimized through back-propagation. The decoder stops predicting after it receives the<end> token.

### *3.8 Quality Measure Metrics*

Besides the qualitative analysis of human-generated captions, quantitative analysis of the result can be done using the BLEU score. BLEU stands for Bilingual Evaluation Understudy, a commonly used metric for evaluating Machine Translation. The BLEU score ranges from 1 to 0. The highest agreement between the generated sentence and the reference sentence receives a score of 1, while the lowest receives 0. The BLEU score is determined as [16]:

$$BrevityPenalty(BP) = \min\left(1, e^{1-\frac{g}{p}}\right) \quad (2)$$

$$BLU_N = BP.e^{\frac{1}{N}\sum_{n=1}^{N} logP_N} \quad (3)$$

Where,
P = predicted caption's length and g = ground truth.
METEOR is also used to generate text for speech recognition, image captioning, and text memorization. Metric for Evaluation of Translation with Explicit Ordering, sometimes known as METEOR, is an acronym. It is calculated as [17]:

$$Score = Fmean * (1 - Penalty) \quad (4)$$

Where,
F-mean = harmonic mean of precision and recall and Penalty is calculated as:

$$Penalty = 0.5 * (chunks/(unigrams\_matched))\text{^}3 \quad (5)$$

## 4 Results and Discussions

The image captioning model was successfully built and tested with different parameters. The model's performance is visualized using train and validation loss on different hyper-parameters. The image captioning model is greatly affected by the over-fitting [14]. In order to mitigate over-fitting, we require a large amount of data. However, due to limited resources and time, we experimented with 8000 images. Hence, we used a dropout of 0.5 and a batch normalization technique to reduce overfitting. In addition, early stopping criteria while monitoring validation loss with patience = 3 is used. The model training is stopped whenever the validation loss increases in three consecutive epochs.
After performing a series of experiments on different hyper-parameters settings, the batch size was fixed to 100, and the training data was fixed to 1000 images. Similarly, the vocabulary size is set to 7000. The model with a high BLEU score is chosen.
Above shows the training and validation loss with the learning rate at 5e-4. The model started to overfit after the first epochs and the validation loss did not decrease accordingly. Even though the number of Epochs was set to 50, the training stopped after 6 epochs to prevent over-fitting. We also monitored the BLEU score of all the 1000 test data.

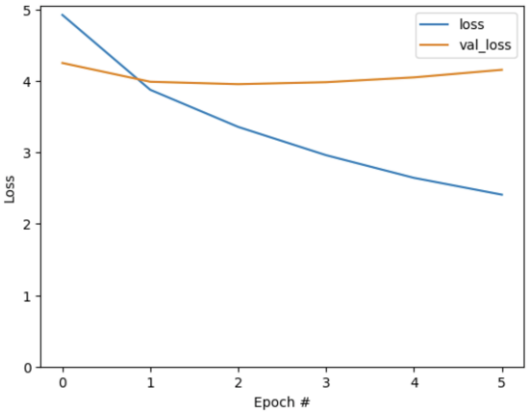


Figure 6: Training and Validation loss curve at learning rate 5e-4

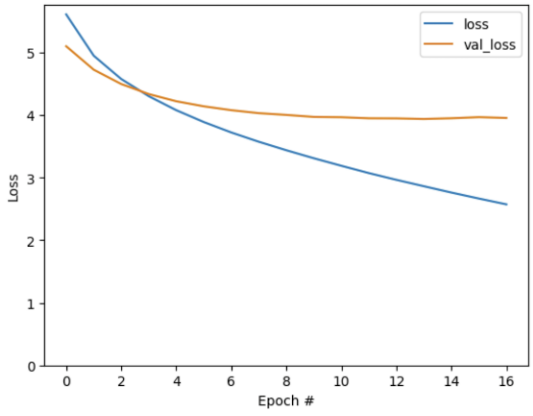


Figure 7: Training and Validation loss curve at learning rate 5e-5

The figure shows the loss curve at learning rate 5e-5. We can see that the model starts to over-fit after 5 epochs, but the validation loss decreases continuously. We continue the training process until the validation starts to increase. After epoch 16, the validation loss increases; hence, the training process is halted. The best loss is then taken as the final loss. We saw the improvement in BLEU as well as the meteor score. Hence, this learning rate is used in the final model.

### *4.1 Sample Outputs*

The sample output is divided into three categories which are good quality (Figures 7 and 8), fairly good (Figures 9 and 10), and bad quality captions (Figure 11). Within each sample, there are two

Table 1: Hyper Parameter Settings

| | |
|---|---|
| Train-Validation (Images) | 5500-1500 |
| Batch size | 100 |
| Learning Rate | 5e-5 |
| Vocabulary Size | 7000 |
| Maximum Sequence Length | 25 |
| Drop Out | 0.5 |
| Loss | Cross Entropy Loss |
| Number of Decoder Layers | 2 |
| Number of Attention Heads | 3 |

images, with the first one showing the original image and the second image representing the region where the attention mechanism is focused in the image to produce the captions.

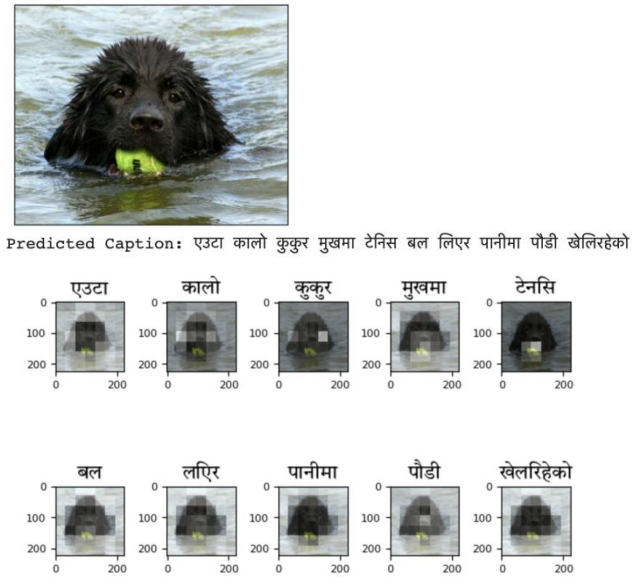


Figure 8: Sample Output 1

The model correctly predicted the dog, its color, background and its detailed action.


Figure 9: Sample Output 2

In the above figure, the model correctly predicted the boy, the color of the clothes and also the background.

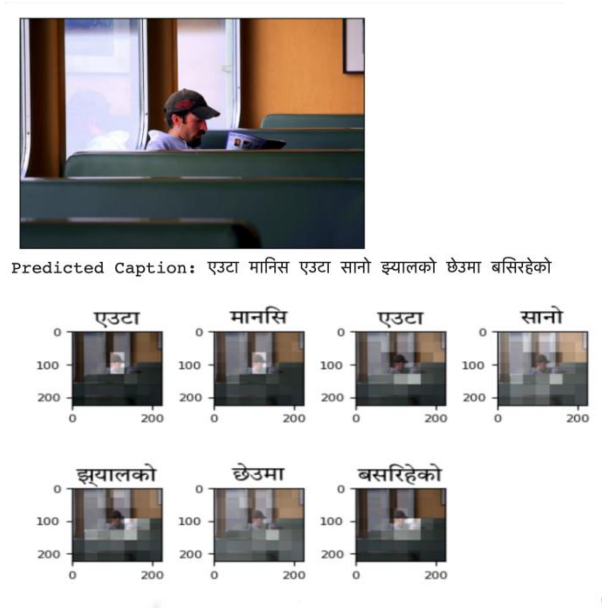


Figure 10: Sample Output

Here, in the above figure even-though the model predicted the object correctly, it failed to describe the action of the person. The caption generated is understandable

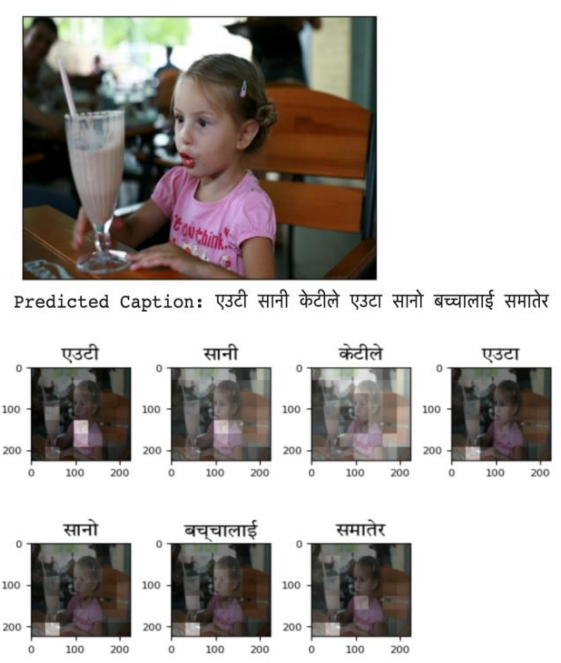


Figure 11: Sample Output 3

The model predicted the little girl in the above figure, but the caption's overall meaning is incorrect.

This sample falls in a very bad prediction. This is because the model is trained with images of small girls, but few data samples in the training dataset describe the action as in the image.

Table 2: Model Comparison

| Models | Dataset | B-1 | B-2 | B-3 | B-4 | METEOR |
|---|---|---|---|---|---|---|
| Our Model | Flickr8k | 0.57 | 0.40 | 0.25 | 0.15 | 0.33 |
| CNN Transformer [8] | MS COCO | 0.629 | 0.433 | 0.291 | 0.19 | --- |
| CNN+ BERT[9] | Flickr8k | 0.391 | 0.246 | 0.151 | 0.093 | 0.317 |

The table shown above compares the proposed model with the two different models. The first model [8] generated captions in Hindi, whereas the second model [9] generated captions in Arabic. Since both are similar to the Nepali language, we can see from Table 2 that our model performs better than the second model, whereas it falls below the first model in terms of BLEU score. The CNN+Transformer model has a slightly better value due to using a larger dataset in training and testing. However, compared with the Arabic image captioning, which uses the same dataset as ours, we achieved better scores in BLEU and METEOR. This is because the Arabic model has used a pre-trained Transformer to generate the captions. The overall quality of captions depends on the input type and pre-trained model.

## Conclusion

This research first developed the manually corrected dataset in the Nepali language and used it to generate the textual description of an image in the Nepali language. We adopted the simpler and less complex Transformer model by eliminating the encoder part of the traditional Transformer model. We achieved comparable results in terms of BLEU score and Meteor score.

The main problem in image captioning in the Nepali language is the lack of a proper dataset. The Nepali language is also grammatically and morphologically complex. This complex nature of language greatly affects the value of evaluation metrics such as BLEU and METEOR. Similarly, there are no pre-processing tools for the Nepali language and the entire pre-processing task had to be carried out manually.

## Acknowledgment

The Department of Electronics and Computer Engineering, IOE, Pashchimanchal Campus, Tribhuvan University, Nepal, supports this research program.

## References

[1] Xu, K. *et al*. Show, attend and tell: Neural image caption generation with visual attention. *International conference on machine learning*, PMLR, (2015) 2048–2057.

[2] Faraby, A. *et al.* Image to Bengali caption generation using deep CNN and bidirectional gated recurrent unit. *23rd international conference on computer and information technology (ICCIT)*, IEEE, (2020) 1–6.

[3] Li, G. *et al.* Entangled transformer for image captioning. *International Conference on Computer Vision (ICCV), (*2019) 8927–8936. DOI:10.1109/ICCV.2019.00902.

[4] Saha, S.A. Comprehensive guide to convolutional neural networks-the eli5 way. *Towards data science*, 15 (2018) 15.

[5] Xiao, X. *et al.* Deep hierarchical encoder–decoder network for image captioning. *IEEE Transactions on Multimedia,* 21(11) (2019) 29422956. .DOI:10.1109/TMM.2019.2915033.

[6] Adhikari A. and Ghimire S. Nepali image captioning. *Artificial Intelligence for Transforming Business and Society (AITB),* 1 (2019) 16. DOI:10.1109/AITB48515.2019.8947436.

[7] Vaswani, A. *et al*. Attention is all you need, *Advances in Neural Information Processing Systems, I. Guyon, U. V. Luxburg, S. Bengio, et al., Eds.*, 30 (2017). [Online Avaible: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.]

[8] Mishra, S.K. *et al*. Image captioning in Hindi language using transformer networks. *Computers Electrical Engineering*, 92 (2021) 107-114. DOI: https://doi. org/ 10 . 1016 / j. compel ceng. 2021. 107114. [Online]. Available: https : / / www . sciencedirect.com / science / article / pii / S0045790621001191.

[9] Emami, J., Nugues, P., Elnagar, A. and Afyouni, I. Arabic image captioning using pre-training of deep bidirectional transformers. *Proceedings of the 15th International Conference on Natural Language Generation, Waterville, Maine, USA and virtual meeting: Association for Computational Linguistics,* (2022) 40–51. [Online]. Available: https:// aclanthology. org/ 2022 . inlg-main.4.

[10] Liu, W. *et al.* CPTR: full transformer network for image captioning. *CoRR*, vol. abs/2101.10804, (2021). arXiv: 2101.10804. [Online]. Available: https://arxiv.org/abs/2101.10804.

[11] O'Shea, K. and Nash, R. An introduction to convolutional neural networks, (2015). arXiv: 1511.08458 [cs.NE].

[12] Howard, A. *et al*. Searching for mobilenetv3, (2019). arXiv:1905.02244 [cs.CV].

[13] Szegedy C. *et al.* Rethinking the inception architecture for computer vision. *Proceedings of the IEEE conference on computer vision and pattern recognition*, (2016) 2818–2826.

[14] Luo, R. C. and Chang, H. Coping with overfitting problems of image caption models for service robotics applications, 2019 IEEE *International Conference on Industrial Cyber Physical Systems (ICPS)*, (2019) 815–820. DOI: 10. 1109/ICPHYS.2019.8780257.

[15] Flickr 8k Dataset, Flickr8k Dataset for image captioning. https://www.kaggle.com/datasets/adityajn105/flickr8k

[16] Papineni, K. *et al*. Bleu: A method for automatic evaluation of machine translation, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Pennsylvania,USA: Association for Computational Linguistic*s, (2002) 311–318.

[17] Banerjee, S. and Lavie, A. METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation*, (2005).