CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning

Bipesh Subedi and **Bal Krishna Bal**
Information and Language Processing Research Lab
Kathmandu University, Dhulikhel, Nepal
bipeshrajsubedi@gmail.com
bal@ku.edu.np

Abstract

Many image captioning tasks have been carried out in recent years, the majority of the work being for the English language. A few research works have also been carried out for Hindi and Bengali languages in the domain. Unfortunately, not much research emphasis seems to be given to the Nepali language in this direction. Furthermore, the datasets are also not publicly available in the Nepali language. The aim of this research is to prepare a dataset with Nepali captions and develop a deep learning model based on the Convolutional Neural Network (CNN) and Transformer combined model to automatically generate image captions in the Nepali language. The dataset for this work is prepared by applying different data preprocessing techniques on the Flickr8k dataset. The preprocessed data is then passed to the CNN-Transformer model to generate image captions. ResNet-101 and EfficientNetB0 are the two pre-trained CNN models employed for this work. We have achieved some promising results which can be further improved in the future.

1 Introduction

Image Caption generation is one of the Computer Vision and Natural Language Processing (NLP) tasks which generates the description of an image. It can be used to aid visually impaired people for describing scenarios, for image indexing, creating image-based search engines, annotating news images, and many more. Automatically captioning an image is one of the main goals of scene understanding in Computer Vision. The caption generation model should handle the challenges of identifying the objects in an image as well as capturing and expressing their description in the natural language (Xu et al., 2015). It is a complex process compared to the existing object detection and classification tasks. In this regard, Computer Vision techniques are used to understand the contents and extract features from the image whereas the Natural Language Processing techniques generate words or descriptions from the extracted features in the right order (Srinivasan and Sreekanthan, 2018).

Recent advancements in NLP and Computer Vision have paved the way to perform various tasks using the native language. Image captioning for the Nepali language is one of the least researched topics and very limited literature are available owing to the language complexity and unavailability of datasets. The only work carried out in this domain was proposed by Adhikari and Ghimire (2019) which employs a traditional CNN-RNN encoder-decoder architecture. Hindi, Marathi, and Konkani are some of the other languages that share similar grammatical structures as the Nepali language as all of these languages use the Devanagari script. Additionally, Bengali language also belongs to the Indic language groups along with other languages mentioned above. Hence, this work is an attempt to develop image captioning for the Nepali language with reference to the existing works. Some of the existing works in this field include Hindi image captioning techniques shown by Mishra et al. (2021) and Bengali image captioning proposed by Palash et al. (2021).

In this research, we propose a CNN-Transformer-based Nepali Image Captioning model. The reason behind opting the transformer network is its wide applicability in today's Natural Language Processing domain as well as its performance in image captioning tasks for other languages. It is computationally faster than RNN as it supports parallel processing. Furthermore, transformer networks have not been implemented for Nepali image captioning so far and hence, this work to the best of our knowledge is the first to implement it. The main focus of this study is to create Nepali datasets and develop a CNN-Transformer model to generate Nepali captions. The remainder of this paper is arranged in the following order: Section 2 discusses some of the related works, Section 3 shows

Proceedings of the 19th International Conference on Natural Language Processing (ICON), pages 86 - 91

the methodology used, and Section 4 consists of results and discussions of the work followed by a conclusion and future directions.

2 Related Works

Various works have been carried out in the image captioning domain, especially in resourceful languages such as English. A significant number of research works can also be seen in Hindi and Bengali which are closely related to the Nepali language. Adhikari and Ghimire (2019) proposed the only work in Nepali that utilizes the two encoder-decoder models with and without visual attention inspired by (Xu et al., 2015). The models use ResNet-50 as encoder and LSTM/GRU as decoder respectively trained on MS COCO datasets after translating and preprocessing. Mishra et al. (2021) proposed a transformed-based encoder-decoder architecture for image captioning for the Hindi language where ResNet-101 is used as an encoder and Transformer as a decoder. They have outlined the problems with RNN architecture and proposed a transformer model with a stacked attention mechanism to decode the image feature vectors into sentences. The authors have calculated BLEU1, BLEU2, BLEU3, and BLEU4 scores with values of 62.9, 43.3, 29.1, and 19.0 respectively. Similarly, Palash et al. (2021), Shah et al. (2021), and Ami et al. (2020) have proposed image captioning in the Bengali language using CNN and transformer networks. Palash et al. (2021) have used ResNet-101 for extracting image features whereas Shah et al. (2021) and Ami et al. (2020) opted for two pre-trained CNN models: InceptionV3 and Xception. Palash et al. (2021) used BanglaLekha datasets for training and testing the model. Similarly, Ami et al. (2020) have used Flickr8k datasets after preprocessing and the work is extended by Shah et al. (2021) using BanglaLekha datasets. BLEU1, BLEU2, BLEU3, BLEU4, and METEOR scores obtained in Palash et al. (2021) work are 0.694, 0.580, 0.505, 2.22e-308, and 0.337 respectively which is better compared to (Mishra et al., 2021). Moreover, Shen et al. (2020) proposed a new model for remote sensing image captioning tasks for the English language. The transformer-based decoder was used to generate captions from the image features. The semantic and spatial features were extracted from the image using CNN. To capture a deeper relationship between image features and text descriptions, the semantic features were added to the decoder's every single sub-layer. The datasets used for this research were obtained from the Sydney Dataset, Remote Sensing Image Caption Dataset (RSICD), and UCM Dataset.

The above literature suggests that different works have been carried out in the image captioning domain in several languages and most of them have concluded that transformer networks perform better than traditional CNN-RNN architecture. In this regard, the Nepali image captioning system using transformers is yet to be explored and the Nepali caption datasets are also not publicly available. Based on these facts and literature, a CNN-Transformer model is attempted to implement in this research.

3 Methodology

The methodology employed for this research involves experimentation of data on the model architecture. The datasets for Nepali image captions are not available publicly, hence two procedures are followed for developing the Nepali image captioning system. The first step involves dataset preparation, followed by model architecture design and implementation.

3.1 Dataset Preparation

The dataset preparation involves two tasks: Dataset collection and Dataset preprocessing. The dataset thus prepared is divided into three parts - for training, validation, and testing purposes.

3.1.1 Dataset Collection

The datasets used for this research are collected from the Flickr8k public dataset[1]. It consists of more than 8000 images with 5 captions each. These are open-source public datasets. The datasets are originally developed for the English language therefore, they require preprocessing to make them usable for our research purpose.

3.1.2 Dataset Preprocessing

It is an integral part of this research because there are not any publicly available datasets for the Nepali language. Furthermore, the grammatical structure of the Nepali language is a lot more complex compared to the English language. In order to solve these problems and make our research more focused on the Nepali language, the following procedures are performed.

[1] https://forms.illinois.edu/sec/1713398

Caption conversion to Nepali The captions in the Flickr8k dataset are in the English language, therefore, to use these datasets in our context, they should be translated into the Nepali language. Such a translation is done using the Google translate API. Each line of the English caption file is translated and appended to a new text file.

Manual correction and annotation The translated texts using Google translate may contain various errors. The translations may not reflect the context of the image, thereby, generating incorrect captions or incurring a loss in the meaning of the descriptions of the image. We handle such problems through manual human corrections. The incorrect or garbage captions are either corrected or removed depending on their quality. It was not feasible to hire an expert therefore, we performed this task ourselves at the lab. Furthermore, the developed captions are randomly sampled to check for any irregularities.

Data Cleaning The translated and modified captions are further preprocessed to remove punctuations and numeric values. In this phase, all the unwanted characters and data from the captions are removed.

Generating Vocabulary and Text Vectorization A text vocabulary is generated from the translated captions by extracting all the unique words from the image description. In this work, a total of over 14,000 unique words are present in the vocabulary. Moreover, since the machines don't understand the natural language it must be converted to some numerical data to map each vocabulary word to a unique index value. This process is done using a built-in Text Vectorizer function available in the Keras library.

Dataset Creation The cleaned captions data are then split into three sets (Training set, Validation set, and Testing set) with 6000, 1000, and 1000 images respectively. The captions data are then mapped with the respective images and zipped together to create datasets for training and validation using the TensorFlow 'Dataset' library. The dataset created in this work can be found on Github[2].

3.2 Model Design and Implementation

The proposed system for Nepali image captioning comprises a CNN model for image feature extraction and a Transformer network for language modelling. Figure 1 shows the system architecture for this work. The CNN model can be developed from scratch but due to its advancement in recent years more accurate and efficient pre-trained models are available for use. Hence, 'EfficientNet' and 'ResNet101' are used to extract image features. These models are the pre-trained models trained on the Imagenet dataset. The CNN encodes the input image to a vector representation that is used by the decoder to generate captions. Since this work is not a classification task, the final softmax layer is removed from the CNN. Residual networks are deep neural networks that use the concept of skip connections to tackle the vanishing gradient problem (a problem where the gradient values of weights become very low for the machine to train efficiently) (He et al., 2015). ResNet-101 is a variant of a residual network with 101 layers, mainly composed of two blocks: Identity Block and Convolution Block. ResNet is one of the most used pre-trained CNN models in deep learning, especially in the Image Captioning domain. On the other hand, EfficientNet is a CNN model as well as a scaling technique that uses a set of preset scaling coefficients to uniformly scale depth, width, and resolution dimensions, in contrast to standard practice, which scales these variables arbitrarily (Tan and Le, 2019). There are 8 variants of EfficientNet (B0 - B7). EfficeintNetB0 is used for this research because of its simplicity and relatively good performance.

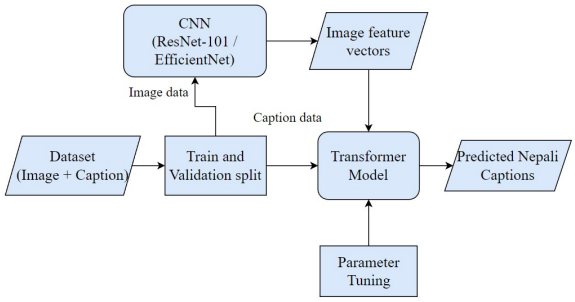

Figure 1: Overall architecture of the system

The transformer model is used to generate captions instead of the conventional Recurrent Neural Network-based architecture because RNN doesn't support parallelization and transformer networks have outperformed RNN in NLP tasks in recent years. Our proposed system follows the transformer architecture proposed by (Vaswani et al.,

[2] https://github.com/bipeshrajsubedi/Flickr8k_Nepali_Dataset

Model	Architecture
Model A	ResNet-101 + Transformer
Model B	EfficientNetB0 + Transformer

Table 1: Different model architectures

Parameters	Value
Image Size	(299,299)
Max. Vocab Size	15000
Sequence Length	20
Embedding Size	512
Batch Size	128
Optimizer	Adam
Loss Function	Categorical-crossentropy

Table 2: Model Parameters

2017). The main components of the transformer networks consist of the encoder, decoder, positional encoding, embeddings, softmax, and multi-headed attention. Attention is utilized in the transformer model to find the relevant collection of values based on a few keys and queries. Attention weights, which are derived using the encoder hidden state (Key) and decoder hidden state (Value), have recently been used to give priority to distinct encoder hidden states (values) in processing the decoder states (query) (Mishra et al., 2021). Single attention-weighted values have been found to be insufficient to capture the many features of the input. The transformer model thus employs multi-headed attention for tackling this challenge. Similarly, positional encoding is used by the transformer networks to keep information about the order of sequence by adding the relative or absolute position of the tokens in sequence (Vaswani et al., 2017). The positional encodings are added to the bottom of the encoder and decoder stacks.

In order to implement our model, the datasets generated are passed to both of the CNN-Transformer architectures mentioned in Table 1. The input images of size (299x299) are passed to the CNN encoder to generate image vectors. The image vectors are then passed to the transformer encoder. The transformer decoder part is fed with the respective captions to train the model. The encoder part comprises a single multi-headed attention head and a normalization layer whereas 2 multi-headed attention heads and 3 normalization layers are used in the decoder. These models are implemented using the TensorFlow Keras library. Table 2 shows the model parameters used in this work.

The model parameters are chosen based on explicit experimentation on our Nepali dataset. These are the optimal parameter values as per our research which can be further improved with the introduction of larger datasets and via parameter tuning.

4 Results and Discussions

The outcome of this work comprises the Nepali image captions dataset and experimental results of the model performance. A dataset of over 40,000 Nepali image-caption pairs is created which are split into training, validation and testing sets. BLEU metric is used for quantitative analysis of the proposed system which is the most commonly used metric for text evaluation that shows the comparison of candidate translation with one or more reference translations (Google, 2022). Four BLEU scores (BLEU-1, BLEU-2, BLEU-3, and BLEU-4) are typically calculated in the context of image captioning. These scores evaluate merely matching grams of a certain order, such as single words (1-gram) or word pairs (2-gram or bigram), and so forth. BLEU score ranges from 0 to 1 where a score between 0.6 to 0.7 is considered to be the best achievable result but at the same time, a score between 0.3 to 0.4 is considered an understandably good translation and a score greater than 0.4 is considered high-quality translation(Google, 2022). In this work, BLEU score is calculated on the overall test data at once using the NLTK bleu library[3]. Table 3 shows the obtained results from this work as well as results from the existing works in Hindi and Bengali languages proposed by Mishra et al. (2021) and Shah et al. (2021) respectively. The obtained results demonstrate that Model B performs slightly better than Model A keeping the model parameters unchanged. The obtained scores imply that the proposed work has shown promising results and can be further improved in future. On the other hand, the first two BLEU scores are not as good compared to image captioning for Hindi and Bengali but can nevertheless serve as a reference for Nepali image captioning. It can also be seen that the last two BLEU scores are higher than that of Mishra et al. (2021) but lower than Shah et al. (2021). It is found that the predicted sentence generally describes the context of the image where its meaning is preserved but the words do not match with the

[3] https://www.nltk.org/api/nltk.translate.bleu_score.html

Model	B-1	B-2	B-3	B-4
Model A	0.49	0.40	0.37	0.34
Model B	0.52	0.42	0.37	0.34
Shah et al. (2021)	0.66	0.55	0.47	0.40
Mishra et al. (2021)	0.62	0.43	0.29	0.19

Table 3: Performance of different models

reference sentences in a specific order which leads to a lower BLEU score. Nevertheless, this is a pioneer work for Nepali image caption generation using transformers and hence can be used as a reference model. Some of the sample outputs of this work are shown in Figure 2.

5 Limitations

The proposed models have not acquired sufficient accuracy and are not able to generate the desired captions for all input images. Moreover, the datasets are also limited and only two pre-trained CNN models are considered for this work. Similarly, the hyperparameters may not be optimally tuned due to limited experimentations. Such limitations can affect the model's efficiency. We consider addressing them in future.

6 Conclusion and Future works

In this research work, a CNN-Transformer-based Nepali Image Captioning system is implemented. At first, the Flickr8k datasets are translated and preprocessed to create a Nepali captions dataset. The datasets are then fed to two models: Model A and Model B with the same model parameters. The outcome of this experiment shows promising results. Model B performed slightly better than Model A on 40,455 Nepali image-caption pairs. Moreover, these models are able to generate captions from the given input image. On the other hand, this work has some limitations as well which can be addressed in the future. The experimentation on larger datasets and fine-tuning of the hyperparameters can be performed in future which are expected lead to better results. MS COCO and Flickr30k datasets can be used after preprocessing for this purpose. Similarly, other CNN models such as InceptionV3, Xception, EfficientNetB7 etc. can be explored for image captioning tasks. Furthermore, this research work can be extended to video captioning for the Nepali language as well.

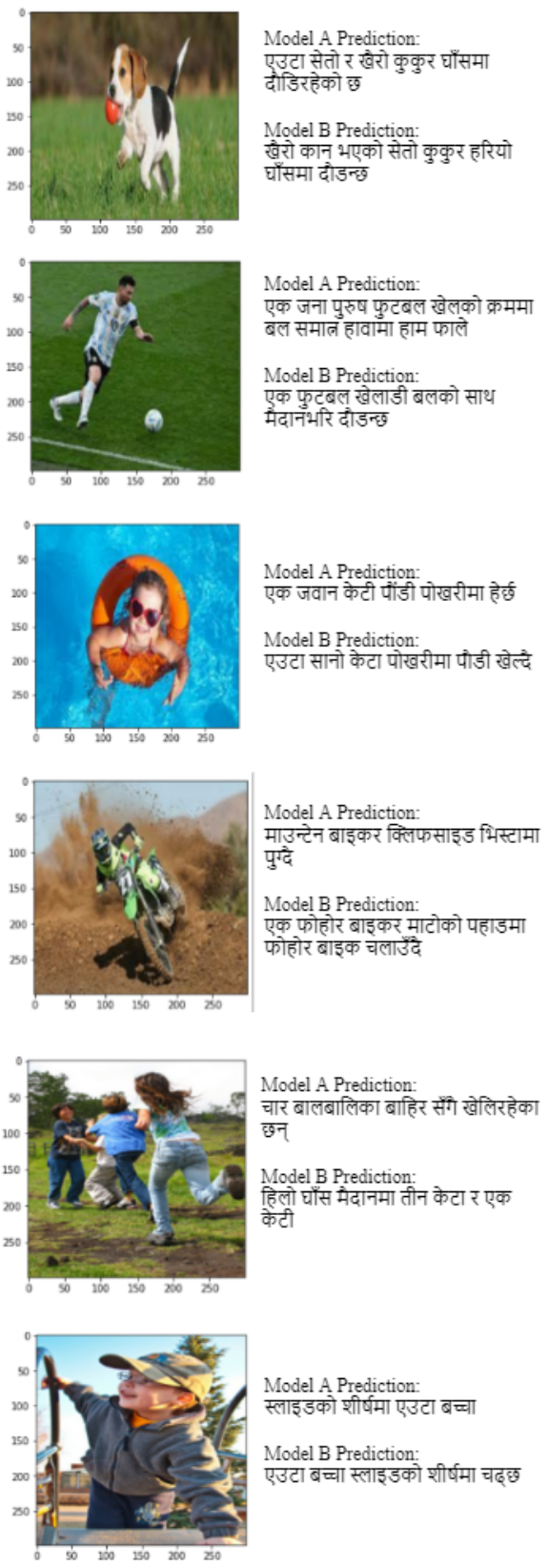

Figure 2: Sample results obtained from our proposed models

References

Aashish Adhikari and Sushil Ghimire. 2019. Nepali image captioning. In *2019 Artificial Intelligence for Transforming Business and Society (AITB)*, volume 1, pages 1–6.

Amit Saha Ami, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Shimul Paul, and Faisal Muhammad Shah. 2020. Bengali image captioning with visual attention. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5.

Google. 2022. Evaluating models.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Deep residual learning for image recognition. *CoRR*, abs/1512.03385.

Santosh Kumar Mishra, Rijul Dhir, Sriparna Saha, Pushpak Bhattacharyya, and Amit Kumar Singh. 2021. Image captioning in hindi language using transformer networks. *Computers Electrical Engineering*, 92:107114.

Md Aminul Haque Palash, M. D. Abdullah Al Nasim, Sourav Saha, Faria Afrin, Raisa Mallik, and Sathishkumar Samiappan. 2021. Bangla image caption generation through cnn-transformer based encoder-decoder network. *CoRR*, abs/2110.12442.

Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. 2021. Bornon: Bengali image captioning with transformer-based deep learning approach. *SN Computer Science*, 3.

Xiangqing Shen, Bing Liu, Zhou Yong, and Jiaqi Zhao. 2020. Remote sensing image caption generation via transformer and reinforcement learning. *Multimedia Tools and Applications*, 79.

Lakshminarasimhan Srinivasan and Dinesh Sreekanthan. 2018. Image captioning-a deep learning approach.

Mingxing Tan and Quoc V. Le. 2019. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *CoRR*, abs/1502.03044.