Proceedings of 14th IOE Graduate Conference
Peer Reviewed
Year: 2023 Month: December Volume: 14
ISSN: 2350-8914 (Online), 2350-8906 (Print)

Nepali Image Captioning using End to End Transformer

Chanda Hyoju [a], Basanta Raj Joshi [b]

[a,b] *Pulchowk Campus, IOE, Tribhuwan University*
[a] 078msice.chanda@pcampus.edu.np, [b] basanta@pcampus.edu.np

Abstract
Image captioning, a growing research field in computer vision and machine learning, employs deep learning to generate textual descriptions of images. While significant progress have been made in English image captioning, research in South Asian Language remains limited. This study addresses the gap by utilizing an end-to-end transformer model for image feature extraction and caption generation, leveraging the Flickr8k Nepali dataset. Transformers have demonstrated significant success in natural language processing tasks. This research aims to leverage the power of transformers by employing a vision transformer for extracting visual features from images. To achieve this, images is divided into patches and self attention mechanism is applied within each patch. This enables the model to capture relevant visual information effectively. Additionally, caption is generated by attending to both the visual features and the linguistic context simultaneously. This is achieved through a cross-modal attention mechanism, which allows the model to generate captions that effectively combine information from both modalities. These techniques is utilized to enhance the quality of the generated captions for images.

Keywords
Image Captioning, Deep Learning, Transformer, Attention Mechanism

1. Introduction

Image captioning is an emerging field at the intersection of computer vision and Natural Language Processing, driven by technological advancements and AI. It entails generating descriptive text for images using machine learning, a challenging task that requires the model to understand the image's context and convert it into human-like language. Deep learning and Neural Networks have made significant strides in image captioning, incorporating object analysis, feature extraction, and NLP techniques. Key factors include object significance, features, spatial relationships, and attention mechanisms, enabling machines to mimic human visual understanding. Image captioning finds utility in diverse applications, from guiding autonomous systems to enhancing image search and aiding the visually impaired, reflecting its relevance in the expanding realm of Artificial Intelligence and technology.

Nepali image captioning is relatively understudied compared to English, some research have been carried out in Hindi, and Bengali which have similar grammatical structure to Nepali language. This research aims to fill this gap by leveraging insights from previous studies in related languages. Most image captioning tasks traditionally used a CNN-RNN model. However, due to the success of transformer models in NLP, recent works have replaced RNNs with transformers in the decoder, benefiting from parallel processing capabilities. Additionally, Transformers have demonstrated strong performance in image feature extraction without convolutional algorithms. Some recent works have also integrated transformers in the encoder portion instead of CNNs, achieving promising results in image captioning. This study endeavors to generate Nepali image captions by referencing existing research and employing an end-to-end transformer model in both the encoder and decoder.

2. Related Works

Extensive research has been done in the field of image captioning, particularly in languages that have ample linguistic resources like English. Furthermore, there is a substantial body of research in languages such as Hindi and Bengali, which exhibit linguistic affinity with Nepali.

The first Nepali paper on image captioning was proposed by Adhikari and Ghimire in 2019 [1], they employed both encoder-decoder models, with and without visual attention, these models utilized ResNet-50 and Inception V3 in encoder and LSTM and GRU in decoder, and trained on MS COCO datasets to generate Nepali caption. In 2021, Mishra et. al. [2] created Hindi caption dataset and employed ResNet101 for image encoding, and used transformers as caption decoders, allowing for effective capture of long-range dependencies and semantic relationships between visual features and textual captions. Similar CNN-Transformer encoder-decoder model is used by Palash et. al. [3] using ResNet101-Transformer and Shah et. al. [4] using InceptionV3- Transformer in Bengali language and obtained promising results being inspired from many English captioning research works.

Recently Subedi and Bal [5] in 2022 have used CNN Transformer model to generate Nepali captions using ResNet101 and EfficientNetB0, and measure the accuracy in terms of BLEU score where EfficientNetB0 performed slightly better than ResNet101 with identical parameters. Liu et. al. [6] proposed complete transformer network, eliminating the need for separate CNN and RNN components. Their end-to-end transformer based architecture leverages self-attention mechanisms and cross-modal attention, demonstrating the capability to generate accurate and contextually relevant image captions, attaining comparable results in comparison to the most advanced models.

Many research have been carried out using CNN-Transformer model and have achieved promising results, but all transformer model is yet to be explored in Nepali image captioning and other language exhibiting linguistic affinity with Nepali language so End to End Transformer model is explored in this research work.

3. Methodology

Encoder Decoder model using transformer is studied in this paper. Here encoder converts the information from image into features and decoder converts the features into meaningful captions. Classic image caption generator model is as below:

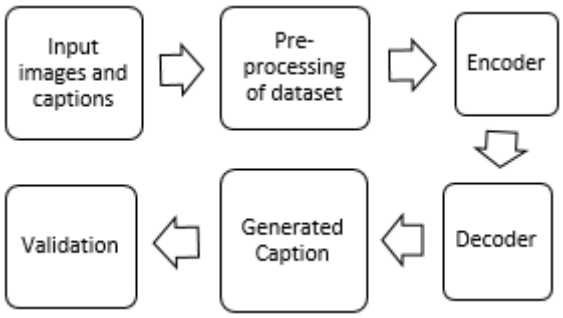


Figure 1: General block diagram

3.1 Dataset

Nepali Flickr8k dataset is used for this study, it consist of 8000 images with 5 Nepali captions for each image. The dataset is divided into train, validation and test set as below:

Table 1: Dataset used

Data	Training	Validation	Testing
Images	6000	1000	1000
Captions	30000	5000	5000

Datasets are preprocessed, captions are broken down into words; punctuation's, special characters, numbers and white spaces are removed. Unique word dictionary is created and tokenized by converting to vector, then vectors are padded to create a consistent length.

3.2 Model Design

End to end transformer model using encoder decoder approach is studied where images are first downsized to set resolution $X = R^{3*H*W}$, then images are split up into N number of patches, $N = \frac{H}{P} X \frac{W}{P}$, where P = patch size. Then, each patch is flattened and reshaped into a 1D patch sequence, $Xp = R^{3*N*P^2}$. Learnable 1D position embedding is added to the patch features after the flattened patch sequence is translated to latent space using a linear embedding layer. This result is the input to the Transformer encoder, Pa = [p1,..., pN]. [6]

The transformer encoder consists of a positional feed forward sublayer and a multi head self attention sublayer, each followed by layer normalization.[7]

Captions are broken down to words, words are tokenized. The word embedding features are combined with positional embedding on the decoder side, with the addition results and output features from encoder serving as the input. The transformer decoder is made up of a positional feedforward sublayer, a multi head cross attention sublayer, and a masked multi head self attention sublayer, each followed by normalization layer.[7] The output feature of the final decoder layer is utilized to predict the next word in a linear layer whose output dimension is equal to the vocabulary size. Two transformer based model has been studied for this research work:

3.3 Model A: ViT + GPT2 model

The core architecture of the image captioning model A revolves around the fusion of a Vision Transformer (ViT) encoder and a GPT-2 decoder. This method is devised to generate textual descriptions for input images seamlessly.

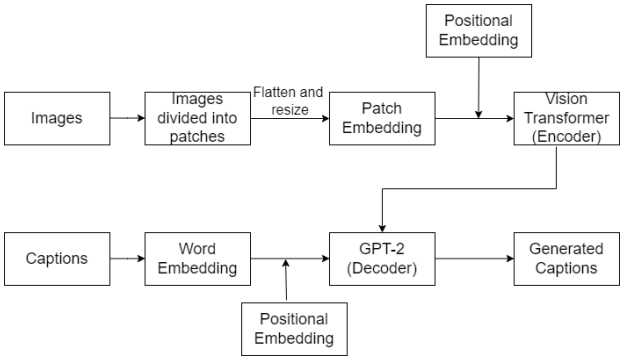


Figure 2: Overview of approached model A

This image captioning model consists of a VIT Encoder based on the ViT architecture for processing images using self-attention and multi-head attention mechanisms. It also includes a GPT Decoder, built on GPT-2, for generating captions in a transformer-based manner, using visual features from the encoder. These components are integrated into an Encoder-Decoder class that combines image encoding and caption generation. The Training Loop optimizes the model over epochs, evaluating it on validation data, using gradient accumulation for efficient training. The model's performance is assessed by loss calculation and optimized with an Adam optimizer.

3.4 Model B: ViT with InceptionV3 + Transformer model

This methodology employs a ViT encoder based on the pretrained Inception-V3 model for robust image feature extraction. It utilizes a dynamic attention mechanism to align textual descriptions with relevant image regions, enhancing caption coherence. The transformer-based decoder captures language patterns effectively, improving caption quality. The caption generation process carefully initializes decoder states based on visual features. Dynamic attention ensures that each word prediction considers visual context, and linear layers diversify vocabulary. Strategic state initialization enhances the model's capacity to integrate visual and textual information, resulting in contextually accurate captions.

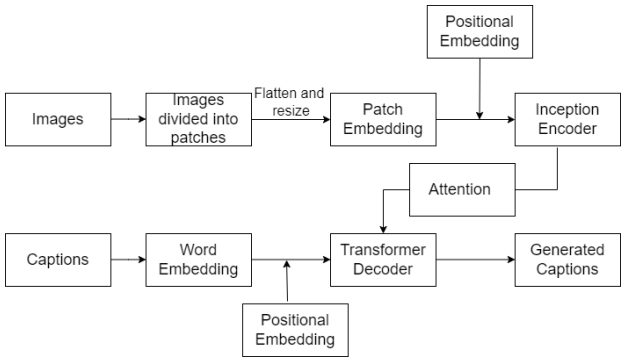


Figure 3: Overview of approached Model B

3.5 Model Parameters

Both the model were trained with different model parameters by changing hyperparameter as learning rate, encoding dimension and decoding dimension. For Model A while learning rate was set to 2e-4, encoding dimension to 4096 and decoding dimension to 1024 and trained for 12 epochs loss analysis seems to be overfitted as training loss decreases with each training but validation loss starts to increase. Similarly learning rate was set to 1e-3, encoding dimension to 4096 and decoding dimension to 1024 for Model B, where validation loss starts to decrease more than training loss, so loss analysis seems to be not correct hence parameters where changed, trained and checked accordingly.

Summary of the model parameters utilized in this research is as below:

Table 2: Model Parameters of Model A and B

Parameters	Model A	Model B
Number of epochs	25	25
Batch size	64	64
Patch size	16	16
Embedding dimension	256	256
Encoding Dimension	768	2048
Decoding Dimension	512	256
Sequence Length	40	40
Number of layers	2	3
Number of head	8	8
Drop out	0.1	0.1
Optimizer	Adam Optimizer	Adam Optimizer
Learning rate	1e-4	3e-4
Loss function	Cross Entropy	Cross Entropy
Validation	BLEU Score	BLEU Score

The model parameters have been meticulously determined by thorough testing using Flickr8k Nepali dataset. These parameter values indicate the good configuration as found by our study. However, it's important to note that further enhancements and refinements may be achieved by incorporating larger datasets and conducting parameter tuning exercises.

4. Results and Discussion

Flick8k Nepali dataset is used, as it contains a diverse set of images that are more free-form and have a wide range of scenes, objects, and activities, making it a good choice for training, validating and testing image captioning models. Nepali caption dataset is selected for this study. The BLEU metric is the most often used metric for text evaluation, which pairs a candidate translation with one or more reference translations, so it is utilized for quantitative examination of the proposed system. [8]

One Sample Image with groundtruth caption is as below:

एउटा खैरो कुकुर चट्टानी किनारमा कालो कुकुरको पछि दौडिरहेको छ

एउटा खैरो कुकुर कालो कुकुरको पछि दौडिरहेको छ

दुईवटा कुकुर समुद्र तटमा खेलिरहेका छन्

दुईवटा कुकुर पानीको छेउमा ढुङ्गा पार गरेर दौडिरहेका छन्

पृष्ठभूमिमा पानी भएको चट्टानी क्षेत्रमा दुईवटा कुकुर एकअर्कातिर दौडिरहेका छन्

Figure 4: Dataset sample with ground truth captions

4.1 Graphical analysis

The capacity of a specific algorithm to model the provided data is determined by its loss function. Loss function measures the dissimilarity between the predicted and real sequences of tokenized captions.

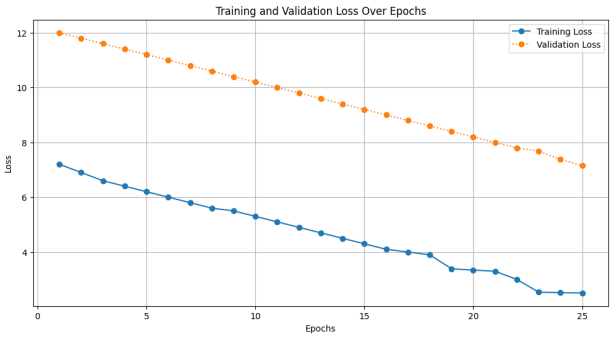


Figure 5: Training and validation loss of model A

We can observe that validation loss is slightly higher than training loss but close to each other which indicates model is not overfitted and good to go with. But conclusion cannot be drawn solely based on this.

The sample output from the attention map generated by attention layer attending over an image to generate caption is as below:

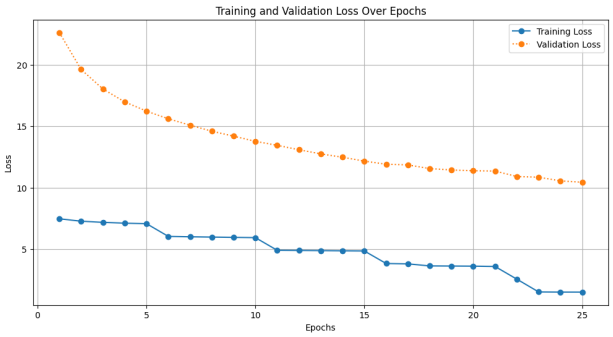


Figure 6: Training and validation loss of model B

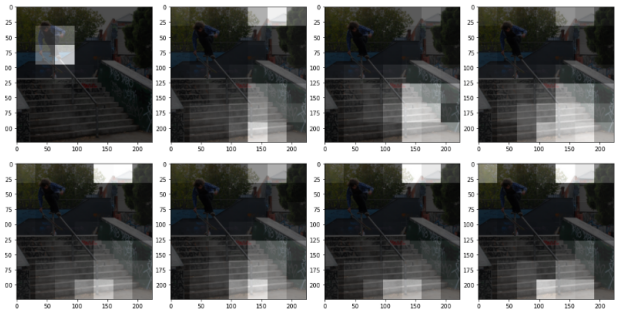
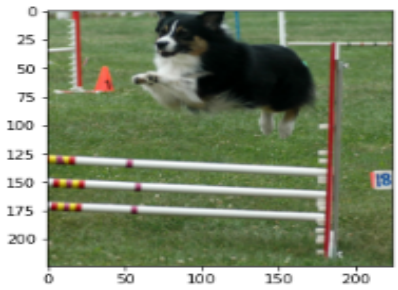


Figure 7: Attention map created by Model A for test image

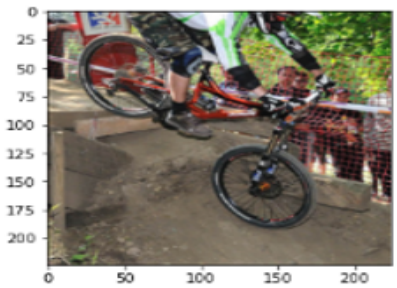
Figure 8: Attention map created by Model B for test image

4.2 Predicted Caption

Captions is predicted for many test images, few test images with their predicted captions are shown below:

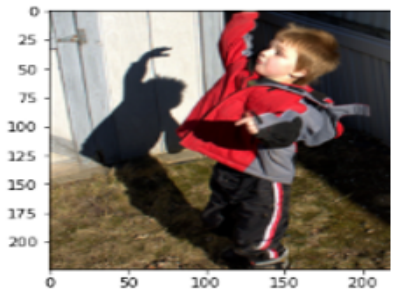


- Model A: सानो कुकुर अवरोध माथि उफ्रन्छ
- Model B: एउटा सानो कुकुर एउटा अवरोध माथि उफ्रन्छ



- Model A: एक साइकिल चालक आफ्नो साइकललाई माथि उफ्रन्छ
- Model B: एउटा सानो बालक आफ्नो साइकलमा चढिरहेको छ

- Model A: एउटा कालो टोपी लगाएको जवान महिला
- Model B: एउटा कालो टोपी लगाएको एउटा जवान महिला



- Model A: रातो पोशाक लगाएको एउटा सानो बालक एउटा फुटबाल ' बल समातिरहेको छ
- Model B: रातो पोशाक लगाएको एउटा सानो बालक

4.3 Model Performance evaluation using BLEU

In image captioning, the BLEU statistic is utilized to compare a generated caption to a reference caption. BLEU-1,2 3 and 4 is used validate the generated caption and obtained as below for two model,

Table 3: BLEU Score

Model	B-1	B-2	B-3	B-4
Model A	0.35	0.32	0.18	0.1
Model B	0.51	0.46	0.22	0.15
Subedi et. al. [5]	0.52	0.42	0.37	0.34

By seeing the above BLEU score, Model B seems to work better than Model A. Model B uses a combination of CNN features with Transformer Encoder, whereas Model A only uses Vision transformer in encoder. CNN works better with image dataset whereas transformer works better with text dataset which cause better performance of Model B. Table 4 also includes the BLEU score obtained by Subedi et. al. which is higher than our model as the model uses full CNN Transformer architecture.

In this study, we evaluated the performance of two models employing Vision Transformer (ViT) and GPT-2 and ViT with Inception and GPT-2. While this models may exhibit lower BLEU scores in comparison to the CNN-Transformer model, it is important to note that ViT and GPT-2 possess distinct advantages. ViT excels in capturing global dependencies in images, and GPT-2 showcases strong language modeling capabilities. These characteristics contribute to a holistic understanding of the input data, making our models well-suited for tasks that require a comprehensive fusion of visual and textual information.

5. Conclusion and Future Works

Full transformer model is implemented in this study for Nepali image captioning and generated captions is validated using BLEU score and from the BLEU score obtained above Model B performed better than Model A with the tested model parameters.

In future research and development of image captioning models, several promising directions can be explored. Firstly, there is room for fine-tuning hyper-parameters to further enhance caption quality and ensure training stability. Additionally, expanding the training dataset to encompass more diverse and extensive image data can significantly improve the model's ability to generalize across a wider range of images. Advanced pretraining techniques, such as self-supervised learning and multimodal pretraining that combines text and image data, hold the potential to provide richer contextual information for improved caption generation. Moreover, adopting multimodal approaches by incorporating multiple data modalities, such as text, speech, and visuals, could lead to more versatile and context-aware captioning models, especially relevant in applications where diverse data forms are used together. The development of refined evaluation metrics beyond traditional standards like BLEU and METEOR could better capture the nuances of human language, leading to more accurate assessments of caption quality. Finally, enhancing the interpretability of the attention mechanism within the model can offer valuable insights into the caption generation process, contributing to the creation of more transparent and accountable AI systems. These avenues represent exciting possibilities for advancing the field of image captioning.

Acknowledgments

This work was supported by Tribhuvan University, Institute of Engineering. The authors are grateful and pay deepest gratitude to all the faculty members for providing opportunity and valuable suggestions and feedback during this work.

References

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

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

[3] Md Aminul Haque Palash and MD Abdullah Al. Bangla image caption generation through cnn-transformer based encoder-decoder network. *arXiv preprint arXiv:2110.12442*, 2021.

[4] Faisal Muhammad Shah, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Amit Saha Ami, and Shimul Paul. Bornon: Bengali image captioning with transformer-based deep learning approach. *arXiv preprint arXiv:2109.05218*, 2021.

[5] Bipesh Subedi 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)*, pages 86–91, 2022.

[6] Wei Liu, Sihan Chen, Longteng Guo, Xinxin Zhu, and Jing Liu. CPTR: full transformer network for image captioning. *CoRR*, abs/2101.10804, 2021.

[7] Ashish Vaswani, 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.

[8] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics.