Image Captioning in Nepali Language using Full Transformer Network

Bipin Maharjan[#1], Dr. Basanta Joshi[*2]
[#]*Department of Computer Science*
Nepal College of Information Technology, Pokhara University
Kathmandu, Nepal
[1]bipin.maharjan@ncit.edu.np
[*]*Pulchowk Campus, Institute of Engineering, Tribhuvan University*
Kathmandu, Nepal
[2]basanta@ioe.edu.np

***Abstract*—Image captioning is a rapidly growing field in computer vision and natural language processing. Its goal is to generate descriptive text for images, similar to human-level captions, using deep learning techniques. This research utilizes a full transformer model for image feature extraction through a vision transformer and caption generation using a language transformer. This study utilizes Nepali datasets from Flick8k and Flickr30k. Transformers have shown promising performance in vision and language tasks compared to CNN and RNN models. Swin is used as a vision transformer and GPT-2 is used as a language transformer. Experimental results on both datasets demonstrate that while both models successfully generate captions, the larger dataset shows better results in predictions and accuracy. These findings contribute to the field of image captioning by providing insights for developing effective models tailored to the Nepali language.**

***Index Terms*—image captioning, computer vision, natural language processing, vision transformer, language transformer**

I. Introduction

Image captioning is a challenging task that aims to automatically generate descriptive and coherent textual descriptions for images, bridging the gap between visual information and natural language. The image captioning process involves two fields to work together i.e. computer vision and natural language processing [5]. This task requires vision processing to extract the image features. The image features are further processed in multiple steps in the natural language processor for generating captions. Therefore, it is a challenging and interesting problem to solve [6]. The applications of image captioning are far-reaching, encompassing fields such as accessibility, content retrieval, robotics, and aiding visually impaired individuals, engaging descriptions of images in social media and e-commerce platforms. While remarkable progress has been achieved in image captioning for major languages, such as English, many languages have limited research in this domain [1] [3].

As most of the captioning task initially follows the CNN-RNN model and recently with the development and success of the transformer model in NLP, RNN is replaced by the transformer model in the decoder portion due to its capability of parallel processing and to solve the vanishing gradient issue of RNN. The first work carried out in this domain in the Nepali language uses the CNN-RNN model [3] showing promising results but leaving room for improvement. Recently released work has used the CNN-Transformer model to overcome the limits of RNN [1]. As transformers are gaining popularity, study on transformers for image classification also increases, Vision Transformer uses transformer to classify images without using any convolution algorithms [7], and it performs admirably, especially after being pretrained on a massive dataset, after which few papers have used transformer in encoder portion in place of CNN and obtained good results in captioning.

The proposed model's foundation lies to use of the encoder decoder based full transformer network for image captioning. On the encoder end, vision transformer is used for generating image feature vectors, and on the decoder end, the language transformer is used for decoding the captions. The full transformer model has resulted in being better than the CNN-Transformer model [5] [2]. Since the CNN-Transformer model was successful in Nepali image captioning, we could utilize the power of the transformer in the encoder part for vision processing instead of CNN. By using vision and language transformer architectures, we attempt to create an image captioning system that grasps the intricate visual features within images and generates contextually coherent Nepali captions. The rest of this paper is organized as follows: Section 2 reviews related works, Section 3 outlines the methodology used, and Section 4 presents the results and discussion, concluding with a summary and suggestions for future research.

II. Related Works

Many research papers have been published by scholars exploring diverse architectures for image captioning. These approaches involve the utilization of CNNs to extract visual features, coupled with RNNs or Transformer-based models to facilitate language generation. Noteworthy contributions, such as "Show and Tell: A Neural Image Caption Generator" authored by Vinyals et al. [8], along with "Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering" by Anderson et al. [6], have been pivotal in

shaping the bedrock of multimodal image captioning systems. Advancement in image captioning has been propelled by the adoption of deep learning methodologies, notably CNNs and RNNs. Researchers have proposed various improvements after the Show and Tell model, such as attention mechanisms, multimodal fusion, and pre-training on large datasets. Recently, transformer-based models have emerged, demonstrating state-of-the-art performance across various benchmark datasets [5].

Stefanini et al. findings indicate that region-based visual encoding methods outperform grid-based and global approaches, particularly evident in higher CIDEr and SPICE scores. Transformer-based and BERT-like language models consistently outshine LSTM-based and CNN-based counterparts across various metrics, showcasing the effectiveness of self-attention and vision-and-language pre-training. However, a trade-off emerges between caption quality and diversity [5].

Nepali exhibits a distinct sentence structure in contrast to English, with a Subject-Object-Verb arrangement instead of the typical Subject-Verb-Object structure found in English [9]. Addressing the challenges of NLP in a morphologically rich and intricate language like Nepali involves grappling with its unique sentence structure, setting it apart from more commonly studied languages like English [10]. Hindi, Bengali Marathi, and Konkani exhibit analogous grammatical structures to Nepali, utilizing the Devanagari script. A contribution authored by Adhikari and Ghimire introduces two encoder-decoder models, one integrating visual attention and the other omitting it. These models employ a ResNet-50 encoder paired with LSTM/GRU decoders, with training executed on MSCOCO datasets post-translation and preprocessing [3]. A subsequent study by Subedi et al. unveils a CNN-Transformer-based encoder-decoder model. Their evaluation encompasses ResNet101 and EfficientNetB0 models coupled with a transformer. The reported BLEU1, BLEU2, BLEU3, and BLEU4 scores stand at 0.52, 0.42, 0.37, and 0.34, respectively, employing EfficientNetB0 as the CNN model [1].

A comparative study of models for image captioning in Hindi utilizes Faster R-CNN, LSTM, GRU, and attention mechanisms to enhance localized image recognition and language modeling. By leveraging datasets like the Visual Genome translated into Hindi, these approaches outperform existing methods in generating dense image captions. Key advancements include a novel encoder-decoder attention mechanism with scaled CNN, achieving superior BLEU scores and manual evaluation results. Another approach combines GRU with EfficientNet, utilizing various attention mechanisms, such as Bahdanau, X-linear, and Luong attention, highlighting the versatility of the methods. CNN-Transformer model was able to achieve highest BLEU score of B1 at 62.9 and B4 at 19 among other models [12].

Research in the Bengali language involves approaches includes encoder-decoder framework [13], Bahdanau Attention-Based Bengali Image Caption Generation utilized various CNN models and BiLSTM for caption generation [14]. Furthermore, a study [15] merged the Flickr8k and BanglaLekha datasets to propose a unified method for creating Bengali image descriptions. By employing different architectures and a blend of InceptionResnetV2 and Bidirectional Long Short-Term Memory models showcasing B1 at 54 and B4 at 14.5. Another system named TextMage [16], aimed at understanding visual contexts in Bangladesh, was trained on a dataset consisting of 9,154 photos and two annotations in VGG16 and LSTM resulting in B1 at 66.7 and B4 at 23.8. Image captioning for Urdu has developed ResNet101-V2 and LSTM model achieving a BLEU score of 0.83 [17]. And, another researcher [18] introduces a new dataset and ResNet101 and LSTM resulting with BLEU scores of 72.5, 56.9, 42.8, and 31.6 for generating Urdu image captions.

Transformers use a self-attention mechanism to model long dependencies in input sequences for computer vision tasks. This is unlike LSTMs, which process sequences recursively and are limited to short-term context [19]. Transformers have addressed the vanishing gradient problem in NLG by eliminating recurrence and convolutions, which were problematic for RNNs and LSTMs. The transformer architecture allows for parallel training and captures longer sequence features, leading to more comprehensive and effective language models [20].

A Generative Pre-trained Transformer is a language model showcasing the significant impact of size and scale in Transformer-based models, even without fine-tuning for specific tasks. GPT architecture has undergone transformations in recent years. Transformer-based language models have achieved groundbreaking results in NLG, but they differ in their architectures and applications [20]. Language Transformers, like BERT, GPT and others, leverage this architecture to understand and generate human language.

The Visual Transformer (ViT) brings a breakthrough in computer vision by rethinking how images are represented. The ViT can better handle the challenges of modeling long-range dependencies, capturing global context, and adapting to different input sizes [21]. Researchers about applying transformer to the computer vision field have attracted significant interest. ViT consistently outperforms state-of-the-art results on ImageNet, CIFAR-100, and VTAB's 19 tasks [7]. The CaPtion TransformeR (CPTR) model employs the full transformer architecture for image captioning. Differing from typical "CNN+Transformer" methods, CPTR is an efficient approach circumventing the limitations of traditional "CNN+Transformer" methods [2]. The Swin Transformer builds a hierarchical representation by merging neighboring patches starting from small-sized ones in deeper transformer layers. The Swin Transformer performs self-attention locally within non-overlapping windows, which divides an image and achieves linear computational complexity to image size. Unlike previous Transformer-based architectures that produce feature maps of a single resolution and have quadratic complexity, the Swin Transformer's complexity is linear to the image size making it better than ViT [22].

III. Methodology

The methodology begins with a dataset preparation containing pairs of images and captions, followed by splitting the

data. An encoder-decoder image captioning model based on transformers is designed. Vision transformer is used in the encoder end and language transformer is used in the decoder end. Figure 1 shows the basic block diagram for this work.

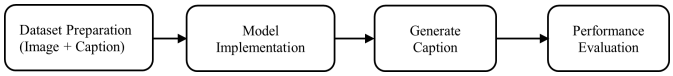


Figure 1: Block diagram of the system

A. Dataset Preparation

1) Dataset Collection: The initial phase entails the compilation of a dataset comprising well-curated images paired with their respective Nepali captions. In pursuit of this objective, the author employed the publicly accessible Flickr8k and Flickr30k dataset, which boasts an extensive collection of over 8000 and 30000 images respectively. Notably, each image is accompanied by five distinct captions. Although the dataset's origin is rooted in the English language, Subedi et al. [1] have transformed it into Nepali. This study utilized the adapted Nepali dataset to fulfill its purpose.

2) Dataset Cleaning: The captions might contain numerous inaccuracies. Punctuation marks and numerical data are excluded from the captions. Extraneous characters and irrelevant information within the captions are meticulously removed during this phase.

3) Special Token Handling: Special tokens, such as the beginning-of-sequence (BOS) token, end-of-sequence (EOS) token, and padding token, are essential for the captioning task. These tokens were properly handled to ensure that the model understands where a caption begins and ends.

4) Language Model Training: We have trained the GPT tokenizer to represent nepali word in vectors because the tokenizer doesn't encompass the Nepali language. This process enables the tokenizer to acquire the ability to learn new words. Furthermore, this fine-tuning adjusts the self-attention weights of the language transformer, enhancing the model's understanding of sentence contexts. This, in turn, contributes to savings in training time and complexity during the image captioning task, as the optimizer can concentrate on refining the cross-attention layers.

5) Dataset Splitting: Once the data was preprocessed, it was divided into three subsets: training, validation, and test data. Both datasets underwent through division of 75% for training, 12.5% for validation, and 12.5% for testing.

B. Model Design and Implementation

The proposed model shown in Fig 2 is encoder-decoder based image captioning model based on transformers. Vision transformer is used in the encoder end for breaking down the images into patches, flattening each patch, and processing it to generate image feature vectors. On the decoder end, the language transformer processes the feature vectors with the reference captions to predict the caption. Swin is used in the vision transformer and GPT-2 is used in the language transformer. Using the combination of encoder and decoder as transformer was motivated by the work done by Liu et. al. [2].

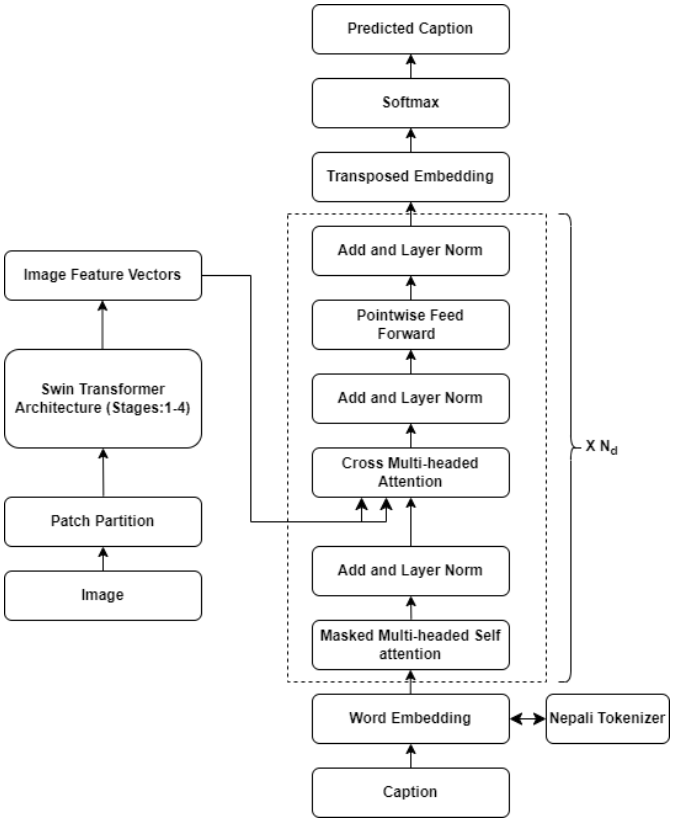


Figure 2: The architecture of Swin + GPT-2 Model

The core architecture of the image captioning model revolves around the fusion of a Swin transformer encoder and a GPT-2 transformer decoder. This allows the model to capture dependencies between image regions and words in the generated captions effectively. This method is devised to generate textual descriptions for input images seamlessly. Initially, the image is resized to a fixed resolution of 224 x 224. The image is then divided into non-overlapping patches, with each patch considered a token. In this implementation, a patch size of 4 x 4 is used, resulting in a feature dimension of each patch being 4 x 4 x 3 = 48. A linear embedding layer is then applied to this raw-valued feature to project it to an arbitrary dimension. It employs window multi-head attention and shifted window multi-head attention in the Swin architecture block and hierarchical structure of merging of feature vectors. The result of the Swin transformer is passed onto the decoder cross-attention layer. The GPT-2 decoder component functions as the textual description generator. As shown in Figure 2, the GPT-2 processes comprising multiple layers of self-attention and feed-forward networks. It takes the visual features obtained from the encoder and generates captions for the images through an autoregressive process, predicting one word at a time based on prior predicted words.

IV. Results and Discussion

The models are implemented using the HuggingFace library and experimented on Flickr8k and Flickr30k dataset. Grid search was used for tuning model hyperparameters. Table 1 shows the model parameters used in this work. The model underwent through multiple experiments by changing number of layers and number of attention heads in encoder and decoder part for varied epochs.

Hyperparameters	Value
Image size	(224, 224)
Max Vocab size	39143
Batch size	24
Embedding size	128
Image patches	16
Learning rate	1e-4
Optimizer	Adam
Loss function	Cross Entropy

Table 1: Model hyperparameters

In multiple experiments, the model demonstrated that the BLEU and METEOR scores improved as the number of epochs increased. Enhancing the number of stages and attention heads in the encoder, as well as increasing the number of layers in the decoder, also yielded positive outcomes. The encoder configuration comprised depths of 2, 2, 18, and 2, each associated with the corresponding number of attention heads in multi-head attention: 4, 8, 16, and 32. The decoder, consisting of 6 layers and utilizing a Nepali tokenizer, achieved the best results. For the Flickr8k dataset, a batch size of 16 is applied whereas, for the Flickr30k dataset, a batch size of 24 is applied. Due to the GPU's memory limitations, executing a larger architecture for the model was not possible, which is a noted limitation of this study. The results of the model are presented in Table 2 below.

Model	Dataset	B1	B2	B3	B4	M
Swin + GPT-2	Flickr30k	0.69	0.54	0.44	0.35	0.22
	Flickr8k	0.59	0.43	0.32	0.23	0.19
Hyoju et al. [4]		0.35	0.32	0.18	0.10	-
Subedi et al. [1]		0.52	0.42	0.37	0.34	-

Table 2: Performance of different models

From the above table 2, we can say that the performance of three models—Swin + GPT-2, Hyoju et al., and Subedi et al. — using the Flickr8k dataset, evaluated with BLEU (B1, B2, B3, B4) and Meteor (M) metrics indicate distinct strengths and weaknesses. The Swin + GPT-2 model achieved the highest BLEU-1 score of 0.59, BLEU-2 score of 0.43, and a Meteor score of 0.19, indicating strong performance in capturing unigram and bigram words while maintaining a balance between precision and recall. Subedi et al., utilizing EfficientNetB0 and a transformer, excelled in generating longer, contextually accurate phrases, achieving the highest BLEU-3 and BLEU-4 scores of 0.37 and 0.34, respectively. In contrast, Hyoju et al., employing ViT GPT-2, demonstrated the lowest performance across all metrics, with a BLEU-4 score of 0.10, reflecting challenges in generating coherent and contextually relevant captions. Notably, training the tokenizer in the Nepali language contributed in improving the evaluation scores whereas similar model experimented by Hyoju et al., which did not use a trained tokenizer, displayed lower performance metrics. With an increase in dataset size from Flickr8k to Flickr30k, there was a notable improvement in the performance of the Swin + GPT-2 model. Specifically, the BLEU-1 score increased from 0.59 to 0.69, and the Meteor score increased from 0.19 to 0.22, suggesting that the model benefits from more data in terms of capturing words and achieving a better balance between precision and recall. These findings indicate that larger datasets, such as Flickr30k, enhance the performance of models in generating more accurate and coherent captions, datasets may restrict the models' ability to generate contextually rich and longer sequences. Some of the predicted captions of this model is presented in Figure 3.

This research develops an image captioning model for the Nepali language using a full transformer architecture, specifically Swin GPT-2. Unlike previous studies, such as those by Adhikari and Ghimire with CNN-RNN models [3], Subedi et al. with a CNN-Transformer [1], and Hyoju et al. with a ViT GPT-2 model [4], our study shows capabilities of the transformer. We employed Swin in the encoder and GPT-2 in the decoder, trained the tokenizer in Nepali and optimized through hyperparameter tuning involving various depths, layers, and attention heads. Evaluation using BLEU and METEOR scores on the Flickr8k and Flickr30k datasets demonstrated that larger datasets significantly enhance model performance. The Swin transformer effectively captures global image dependencies, while GPT-2 excels in language modeling, indicating that our full transformer model surpasses previous CNN-Transformer approaches in generating high-quality image captions in Nepali.

V. Conclusion and Future Works

In this study, a Nepali image captioning system based on a full transformer architecture is implemented which is trained using Flickr8k and Flickr30k datasets. The output of this experiment demonstrates promising results, with models successfully generating captions for images. Increasing the size of the dataset and training the language model in Nepali resulted in better predictions and accuracy. This study contributes to the image captioning research, offering insights for developing effective models in the Nepali language. Future work could focus on refining the model architecture by exploring a range of CNN, transformer, or hybrid models, such as ViT, BeiT, DeiT, and GcViT for encoders, and ROBERTa, BERT, and Phi for decoders. Expanding evaluation to metrics like CIDEr, SPICE, ROUGE-L, and BERTScore, and incorporating human evaluations and multimodal approaches, will improve semantic understanding and caption diversity. Future studies could scale up with larger datasets, such as MS COCO and ImageNet.



Flickr8k Model Prediction:

टोपी लगाएकी एउटी जवान केटी हाँस्छिन्।

Flickr30k Model Prediction:

गुलाबी पोशाकमा एउटी केटी हाँस्छिन्।



Flickr8k Model Prediction:

केटी स्टेजमा नाच्दै

Flickr30k Model Prediction:

एउटी केटी स्टेजमा नाच्दै।



Flickr8k Model Prediction:

नीलो शर्ट लगाएकी केटी विज्ञान केन्द्रबाट दौडिरहेको छ।

Flickr30k Model Prediction:

केटी इनडोर खेल मैदानमा खेल्न रमाईलो गर्छिन्।



Flickr8k Model Prediction:

दुई घोडाहरूले एउटा सानो बच्चालाई हेरिरहेका छन्।

Flickr30k Model Prediction:

दुई घोडा नजिकै भत्केको संरचना भएको सानो बच्चालाई हेरिरहेका छन्।



Flickr8k Model Prediction:

नीलो शर्ट लगाएको बच्चा जंगलको भुइँमा बसेर दूरबीनबाट हेर्दै।

Flickr30k Model Prediction:

दुर्बीण लगाएकी महिला जङ्गलको भुइँमा बसेर माथि हेर्दै छिन्।

Figure 3: Predicted captions of Swin + GPT-2 model

REFERENCES

[1] B. Subedi and B. K. Bal, "CNN-Transformer based Encoder-Decoder Model for Nepali Image Captioning," aclanthology.org, vol. Proceedings of the 19th International Conference on Natural Language Processing (ICON), p. 86–91, 2022.

[2] W. Liu, S. Chen, L. Guo, X. Zhu, J. Liu, and S. Chen, "CPTR: Full Transformer Network for Image Captioning," https://arxiv.org/abs/2101.10804, 28 Jan 2021.

[3] A. Adhikari and S. Ghimire, "Nepali Image Captioning," 2019 Artificial Intelligence for Transforming Business and Society (AITB), pp. 1-6, 2019.

[4] C. Hyoju and B. Joshi, "Nepali Image Captioning using End to End Transformer." Accessed: Apr. 28, 2024. [Online]. Available: http://conference.ioe.edu.np/publications/ioegc14/IOEGC-14-047-C3-5-771.pdf

[5] M. Stefanini, M. Cornia, L. Baraldi, S. Cascianelli, G. Fiameni and R. Cucchiara, "From Show to Tell: A Survey on Deep Learning-Based Image Captioning," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 45, no. 1, pp. 539-559, 2023.

[6] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2018.

[7] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit and N. Houlsby, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," https://arxiv.org/abs/2010.11929, 03 June 2020.

[8] O. Vinyals, A. Toshev, S. Bengio and D. Erhan, "Show and Tell: A Neural Image Caption Generator," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[9] S. Timilsina, M. Gautam and B. Bhattarai, "NepBERTa: Nepali Language Model Trained in a Large Corpus," Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), p. 273–284, November 2022.

[10] S. Thapa, K. Rauniyar, S. Shiwakoti, S. Poudel, U. Naseem and M. Nasim, "NEHATE: Large-Scale Annotated Data Shedding Light on Hate Speech in Nepali Local Election Discourse," Frontiers in Artificial Intelligence and Applications, vol. 372, pp. 2346 - 2353, 2023.

[11] S. Paudel, B. K. Bal and D. Shrestha, "Large Vocabulary Continuous Speech Recognition for Nepali Language using CNN and Transformer," Information and Language Processing Research Lab, 2023.

[12] V. Jayaswal, R. Rani and J. Kaur, "A Comprehensive Survey on Image Captioning for Indian Languages: Techniques, Datasets, and Challenges," Research Square, 13 November 2023.

[13] F. M. Shah, M. Humaira, M. A. R. K. Jim, A. S. Ami and S. Paul, "Bornon: Bengali Image Captioning with Transformer-based Deep learning approach".

[14] R. Rahman, H. Murad, N. N. Rahman, A. K. Saha, S. M. R. A. Masud and A. S. Z. Momtaz, "CapNet: An Encoder-Decoder Based Neural Network Model for Automatic Bangla Image Caption Generation," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 13, no. 8, pp. 1-8, 2022.

[15] M. Humaira, S. Paul, M. A. R. K. Jim, A. S. Ami and F. M. Shah, "A Hybridized Deep Learning Method for Bengali Image Captioning," (IJACSA) International Journal of Advanced Computer Science and Applications, vol. 12, no. 2, pp. 698-707, 2021.

[16] A. H. Kamal, M. A. Jishan and N. Mansoor, "TextMage: The Automated Bangla Caption Generator Based On Deep Learning," 2020 International Conference on Decision Aid Sciences and Application (DASA), pp. 822-826, 2020.

[17] I. Ilahi, H. M. A. Zia, M. A. Ahsan, R. Tabassam and A. Ahmed, "Efficient Urdu Caption Generation using Attention based LSTM," arXiv, 02 Aug 2020.

[18] M. K. Afzal, M. Shardlow, S. Tuarob, F. Zaman, R. Sarwar, M. Ali, N. R. Aljohani, M. D. Lytras, R. Nawaz and S.-U. Hassan, "Generative image captioning in Urdu using deep learning," Journal of Ambient Intelligence and Humanized Computing, vol. 14, no. 6, pp. 7719-7731, 10 April 2023.

[19] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan and M. Shah, "Transformers in Vision: A Survey," ACM Journals, vol. 54, no. 10, pp. 1-41, 13 September 2022.

[20] M. O. Topal, A. Bas and I. v. Heerden, "Exploring Transformers in Natural Language Generation: GPT, BERT, and XLNet," International Conference on Interdisciplinary Applications of Artificial Intelligence, 16 Feb 2021.

[21] B. Wu, C. Xu, X. Dai, A. Wan, P. Zhang, Z. Yan, M. Tomizuka, J. Gonzalez, K. Keutzer and P. Vajda, "Visual Transformers: Token-based Image Representation and Processing for Computer Vision," https://arxiv.org/abs/2006.03677, 20 Nov 2020.

[22] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin and B. Guo, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," arXiv, 17 Aug 2021.