Nature-based Bengali Picture Captioning using Global Attention with GRU

Fatema Tuz Zohora *, Sujit Biswas †, Anupam Kumar Bairagi‡, Kashif Sharif§
* Department of Computer and Information Science, Northumbria University, London, United Kingdom
† Department of Computer Science, City University of London, London, United Kingdom
‡ Department of Computer Science and Engineering, Khulna University, Khulna, Bangladesh
‡ School of Computer Science and Technology, Beijing Institute of Technology (BIT), Beijing, China
Email: fatematuzzohora1012@gmail.com, sujit.biswas@city.ac.uk, anupam@ku.ac.bd, kashif.sharif@ieee.org

***Abstract*—Automatic picture captioning is a prominent research area of artificial intelligence technology (AI). Its ability to enhance AI models by translating observed data into human language opens up a wide range of real-time applications. In this study, we explore picture captioning in the Bengali language using the global attention mechanism. Given the limited prior research in this area, we provide a comprehensive assessment of two distinct global attention approaches: the general approach and the concatenation approach. Additionally, we evaluate the performance of two CNN encoders, VGG19 and InceptionV3, within these models. The models are trained on a secondary dataset consisting of 4,849 nature-based images, each annotated with a single caption, enabling the models to gain a broad understanding of related categorical information. To achieve our research objectives, we developed and trained four separate models using this new dataset. Our analysis, both qualitative and quantitative, demonstrates that these algorithms are capable of generating human-like captions for similar images. The results indicate that models using the concatenation approach, particularly with the InceptionV3 encoder, performed best, achieving a BLEU-1 score of 84.85. In contrast, the model using the general approach with VGG19 underperformed in generating satisfactory captions.**

***Index Terms*—Bengali image captioning, Global attention, CNN, GRU, Attention Mechanism, Bengali dataset.**

I. Introduction

Automatic picture captioning is a hybrid technique that focuses on describing real-world scenes and has gained popularity with the advancement of artificial intelligence. It serves as a bridge between the language and vision of an artificial intelligence model that learns to communicate by viewing the world and understanding all elements and activities, as well as being able to express it in human comprehensible language. With the advancement of this technology, artificial intelligence models are constantly improving their ability to communicate with the human world and perform previously impossible tasks. This technology is used in smart blind guidance [1], medical report writing [2], robotics [3], composed image retrieval [4], surveillance [5], and video captioning [6].

Automatic captioning models focus on explaining a scenario in an image or scene as accurately as possible in human language while avoiding grammatical or semantic errors. As a result, understanding the image's context is necessary for choosing the right words to describe the situation adequately. The model's performance depends on developing grammatically correct language descriptions based on the photo's attributes and conditions.A typical captioning model works by first identifying the objects, following their actions, and then integrating the pieces of information gathered to come up with a caption. Figure 1 presents the overall steps for the automatic picture captioning. Recently, advanced deep-learning models like Convolution Neural Network (CNN), Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), Transformer models, Bidirectional Encoder Representations from Transformers (BERT) architectures, and others have been used together to help study images and come up with the best captions for them [7]. Moreover, these models are currently trained in many languages, not just English, potentially eliminating linguistic barriers to global technological advancements in the future. Similarly, several efforts have been made to adopt this technology for speakers of "Bengali," one of the world's most widely spoken languages [8]. Attention approaches have also been quite effective, allowing these models to effectively generate captions considering the picture regions [9]. However, given the diverse expressions of real-time scenes and their infinite attributes, objects, and activities, the models remain far from perfection.

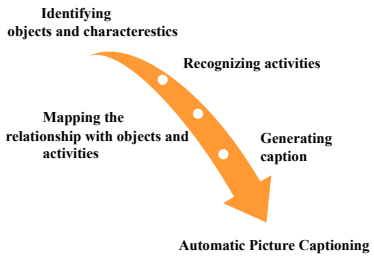


Fig. 1: Essential steps of Automatic Picture Captioning.

Captioning models face various challenges, such as the nature of datasets, language differences for the models, the appropriateness of the captions, the production of lengthy descriptions, the insufficient capability in real-time, and the existing issues with the current methods [7]. We frequently

Corresponding author. Email: sujit.biswas@city.ac.uk

train these models using classified image datasets. As a result, some of these classes may include fewer examples than required to develop a sufficient comprehension of that specific domain. Furthermore, diverse geographical regions, people, and cultural knowledge must also be considered when adopting these models.

To address this issue, this study aims to develop a specific niche-based dataset (nature-based); as a result, the model could have strong domain-specific knowledge. Furthermore, while local attention-based models demonstrate impressive efficiency, their focus on "specific picture location" may neglect some significant picture source positions, as only the most relevant sources receive consideration. Global attention ensures that all visual regions are considered while building the caption, so there is no chance of losing any source input during the training phase [10]. Moreover, different ways of implementing a global attention strategy have not previously been used for Bengali captioning. Thus, this study has implemented two global attention approaches—the general and concatenation approaches—to evaluate the effectiveness of these methods for Bengali captioning. The results demonstrate that the concatenation strategy performs better when applying global attention to Bengali captioning.

The following outline summarises the paper's major contributions:

- An advanced ML model for automatic picture captioning in Bengali that uses two different global attention approaches.
- A new niche-based picture dataset, focused explicitly on Bangladeshi outdoor nature, locations, and culture, to train the models.
- The performance analysis of four global-attention models using two pre-trained CNN encoders.

This paper divides the remaining content into five sections. For example, Section II is devoted to a review of prior literature on similar issues. Section III describes the methods used in this research. Then, Section IV addresses the research's data preparation, training specifics, the result findings, and analysis. Finally, Section V at the end of this paper contains a conclusion.

II. Literature Review

A. Picture Captioning in other Languages

The majority of earlier picture captioning studies were conducted using English captions until recently when researchers began to adapt this technique to other languages. Li et al. [11] initially extended this research to the Chinese language in 2016 by creating a dataset translated from English into Chinese. For the experiment, they trained an encoder-decoder framework model already available in English. Several new approaches for Chinese captioning have been adopted after that, including attention models [12] and the use of GAN [13], resulting in better captions in Chinese. Incorporating Chinese captioning into robot-enhanced therapy [14] and remote sensing system [15] were further applications of this technology. Furthermore, Yoshikawa et al. [16] worked with a similar encoder-decoder model in 2017, revealing that these models can also be trained to construct Japanese captions, underlining the superiority of their massive human-generated Japanese captions dataset known as "STAIR captions". Additionally, the research for integrating the Hindi language with this technology has included hybrid models like encoder-decoder models [17], attention models [18], and transformer model [19]. Arabic picture captioning also exists, with several studies done in recent years [20]. Other languages adopting this technology consist of Turkish [21], Indonesian [22], Vietnamese [23], and more languages will likely be added in the future. This paper solely contributes to the experiment by employing Bengali.

B. Bengali Picture Captioning

Rahman et al. [24] published "Chittron" as the first Bengali picture captioning system in 2018, using stacked LSTM layers for encoder-decoder architecture. More advanced techniques are used in [25] such as the newest CNN (InceptionV3) and RNN architecture (GRU). Similarly, in [26], the authors focused on improving grammatically correct Bengali captions using CNN. These papers use BLUE-1, CIDEr, METEOR, ROUGE, and SPICE, and greedy and beam searches produce good captions. However, none of them followed either the local or global attention model.

Meanwhile, in 2020, Ami et al. [27] trained the first local attention-based model using the Bengali-translated Flickr8k dataset. They evaluated their model's captions using the BLEU matrices, the Xception and InceptionV3 encoders, and the GRU decoder. The authors in [8] tested a local-attention-based model using a dataset from Bangladesh. They assessed performance with four encoders. The results were far better than those of previous models. Recently, Das et al. [28] proposed a local-attention model with InceptionV3 and GRU in 2023, utilizing two public datasets as benchmarks.

The relevant work evaluation makes it evident that the global attention technique has not been considered yet for Bengali. Moreover, no research has been conducted on many global techniques to determine the optimal strategy for this language. This study is the first to develop a dataset with a specific niche to improve the model's capacity to caption. In addition, this paper shows valuable experiments on different global attention mechanisms that might contribute to the advancement of future Bengali captioning techniques.

III. Methodology

A. Overview

We develop the model for enhancing the picture caption using a global attention-based framework, which provides the model with all the necessary connections between words and images to enable it to create appropriate situation-based captions for the scenes. Three technologies comprise the model: a CNN model, an RNN model, and an attention approach. Within the model framework, a fine-tuned CNN encoder collects the attributes of each image region directly from the network's bottom layer. Consequently, the model

extracts features and averages them, giving each region a unique feature. This enables the attention algorithm to define each location according to its relevance to the corresponding term. We deal with sequential caption data using the decoder GRU, an upgraded type of LSTM that has gained popularity due to its extended capabilities. Fig. 2 shows the diagram of the model framework in this paper. The framework illustrates the process of encoding a picture using CNN, which provides visual features such as s1, s2, and ...sp. The decoder GRU then applies an attention technique to these visual features to generate each word for the caption.

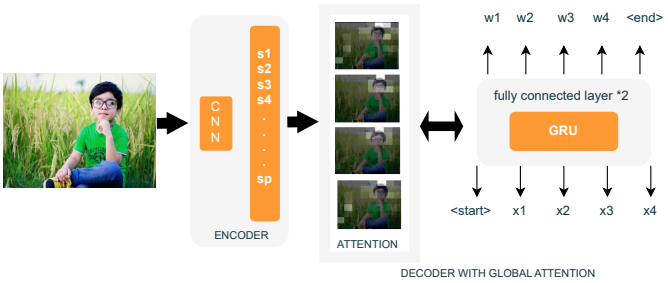


Fig. 2: Model Framework.

The different CNNs are replaced with the decoder model (VGG19 and InceptionV3) to generate captions. Inside the decoder, the GRU is used to analyze the two global attention approaches for calculating the context vector: general and concatenation strategies. As a result, four alternative models were trained for this research, depending on the Encoder and the attention strategy used to test the model.

B. Data Pre-processing

During the pre-processing stage, the <start> and <end> tokens are added at the end of each caption. These tokens are visual cues for the model to determine the appropriate moments to initiate and conclude the caption generating process. Upon identification of a <start> token, the decoder model initiates the construction of a sequential series of words for the caption. Subsequently, upon detection of a <end> token, the model concludes the process. Additionally, a directory is established to house the preprocessed caption dataset and the accompanying photo locations. The captions were later separated into tokens for tokenization, and the paths of the images were recovered to do feature extraction.

C. Tokenization

Each word in the captions had a space between it to tokenize it. Consequently, this approach yields the lexicon of all the unique words in the data, while the vocabulary consists of 2033 words. Subsequently, the word-to-index and index-to-word mappings were established, wherein each word in the captions was associated with a specific index number. Additionally, it was determined that captions must not exceed 27 characters. Consequently, before creating the caption vector, all sequences were extended to match the length of the longest sequence (e.g., 27).

D. Feature Extraction

The model was examined using the pre-trained VGG19 encoder and the InceptionV3 encoder. The photos were pre-processed to conform to the input requirements of the InceptionV3 and VGG19 models. Furthermore, the uppermost layer of these models was removed during feature extraction.

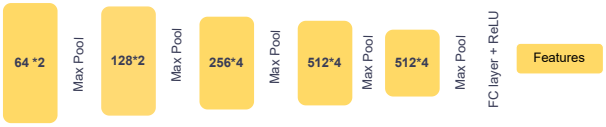


Fig. 3: Fine-tuning of the VGG19 model as an encoder, with the top layer replaced by a fully connected layer.

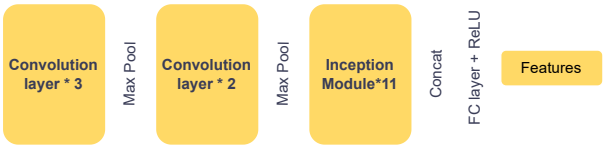


Fig. 4: Fine tuning of the InceptionV3 model as the encoder, with the top layer replaced by a fully connected layer.

Prior to being delivered to the VGG19 model, the pictures are resized to dimensions (224, 224, 3). By removing the top layer, the output dimension of the VGG19 encoder has been modified to (49, 512). Similarly, the images transform to conform to the required input format of InceptionV3, resulting in a reshaping of their dimensions to (299, 299, 3). The output shape of the InceptionV3, excluding the top layer, was (64, 2048).

E. Encoder

The encoder generates the features without using the prediction layer. To match the dimension between feature vectors and 2-D picture sections, the features are obtained from the bottom convolutional layer [29]. The encoder generates feature vectors of p integers, each representing a different section of the picture. As a result, there should be a "P" number of feature vectors for all "Q" image areas.

$$S = (s_1, s_2, s_3, \ldots, s_p); s_p \in \mathbb{R}^{2D} \quad (1)$$

Here, S represents the overall feature, and $(s_1, s_2, s_3, \ldots, s_p)$ represents the feature vector for individual picture areas. The retrieved picture features are then input into a fully connected layer, followed by activation functions (i.e., RELU), which supply the encoder's hidden states.

F. Decoder

The decoder utilizes global attention to create a context vector that incorporates all relevant information from the picture segment. This context vector assists the decoder in predicting the target word. Therefore, the concealed states of the encoder and decoder are fed into the attention model. The algorithm then calculates an alignment score using these two

inputs, which indicates the relevance and alignment scores of the caption word based on the corresponding picture sections. The two currently available approaches for calculating alignment scores are the general and concatenation techniques. The general approach multiplies the decoder's output by source data, whereas the concatination technique merges outputs with a neural layer. The general procedure formula [30] is as follows:

$$\text{score}(h_t, \bar{h}_s) = h_t W_a \bar{h}_s \tag{2}$$

Here, h_t stands for the decoder's current hidden state, $barh_s$ stands for all of the picture source data, and W_a represents the weight of the dense layer.

The concatenation approach is followed by Eq. 3.

$$\text{score}(h_t, \bar{h}_s) = v_a^{\mathsf{T}} \tanh(W_1 h_t + W_2 \bar{h}_s) \tag{3}$$

Here, v_a^{T} is the vector for scaling.

Subsequently, the likelihood of significant visual areas, referred to as "attention weights," is computed using Equation 4 [30].

$$\alpha_t = \text{softmax}(\text{score}(h_t, \bar{h}_s)) \tag{4}$$

The final stage of the attention approach is the creation of the Context vector [29], [30].

$$c_t = \sum_i \alpha_t \bar{h}_s \tag{5}$$

After that, the decoder concatenates the created context vector with the current output. The output is then passed through a dense layer with a TANH activation function. Finally, the revised decoder output $\tilde{h_t}$ is sent through another dense layer that is the same length as the vocabulary size (W_v), resulting in the final probability sequence for the caption.

$$\tilde{h_t} = \tanh(W_a[\text{concat}(c_t, h_t)]) \tag{6}$$

$$p(y_t|y_1, \ldots, y_{(t-1)}, x) = (W_v \tilde{h_t}) \tag{7}$$

Here, y_t is the target word computed from the probability vector, and x is the given input sentence.

During the training and evaluation phases, the decoder receives the <start> token as the first input, and it uses the predicted target word from the current phase as the subsequent decoder input. After training, the gradients are computed and applied to the Adam optimizer to calculate the loss. The predicted words for the training and evaluation phases are determined by looking at the words with the highest probability. At last, when the model encounters the <end> token, it terminates the process.

IV. Result and Analysis

A. Testbed Details

Cloud platform Google Collaboratory used the TASLA T4 GPU for the investigations. All critical data was stored in Google Drive throughout testing for quick recovery. The model was built using TensorFlow and a 512-unit-per-layer RNN. The model contains 256 embedding dimensions, 10 batches, and 0.001 learning rate. For training, the dataset was randomly partitioned into 3879 photos and descriptions. Additionally, 970 test photos were reserved to evaluate the model. The 30 period training process took about one hour for all four models.

B. Dataset Collection

The secondary dataset for this study was derived from the "BNATURE" dataset, created by Al Faraby et al. [25] in 2020. It comprises 8000 photographs measuring 500 x 375, each accompanied by 5 Bengali annotations. All the photographs pertain to the cultural and geographical environment of Bangladesh. For this research, only outdoor nature-based photographs are targeted during the dataset selection process. Consequently, 4849 photographs were collected, and one caption was chosen for each photograph. The dataset was collected manually, considering that the model should be trained with similar categorical photos.

C. Evaluation Metric

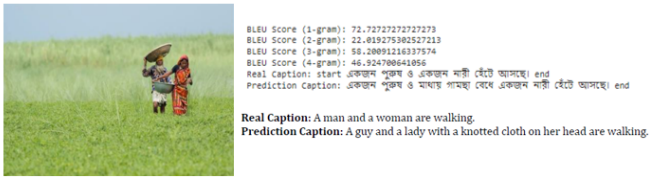
The BLEU matrices are used to assess the quality of predicted captions as they provide a score based on comparing the predicted and reference texts [31]. Machine translation, natural language processing, and captioning model evaluation widely use the BLEU and meteor. The BLEU scores are computed using n-gram matrices, where n varies from 1, 2, 3, and 4. Table I shows the models' performance with the VGG19 and InceptionV3 using average BLEU scores for 1, 2, 3, and 4 grams and meteor. Table II depicts a comparison of our model outcome to new existing models.

TABLE I: The result scores of the trained models.

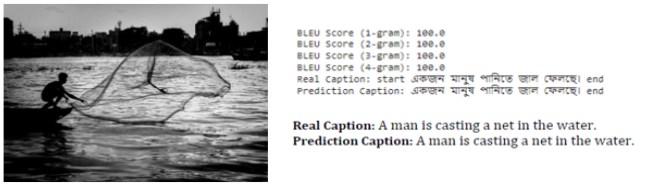
Encoder	Decoder	Strategy	BLEU-1	BLEU-2	BLEU-3	BLEU-4	Meteor
VGG19	Attention + GRU	General	47.91	37.15	29.85	24.05	0.51
VGG19	Attention + GRU	Concat	80.92	77.72	74.97	70.18	0.85
InceptionV3	Attention + GRU	General	78.35	74.47	70.77	66.52	0.79
InceptionV3	Attention + GRU	Concat	84.85	82.13	79.49	75.59	0.87

TABLE II: The existing models with comparison to our model.

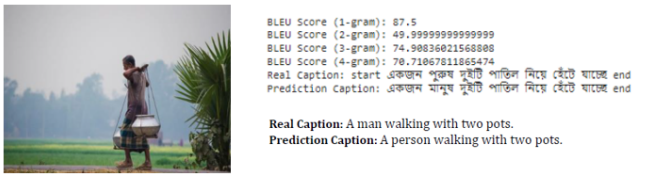
CNN	Model	Dataset	Strategy	BLEU-1	BLEU-2	BLEU-3	BLEU-4
Xception	Atten+GRU [32]	Bornon	Local	60.5	49.2	41.2	35.1
Xception	Atten+GRU [8]	BanglaLekhaImageCaptions	Local	87.18	85.35	83.23	80.60
ResNet-101	Transformer [33]	BanglaLekhaImageCaptions	Self-attention	69.4	58.0	50.5	2.22e-308
InceptionV3	**Atten+GRU (our model)**	**Nature-based dataset**	**Global (concat)**	**84.85**	**82.13**	**79.49**	**75.59**



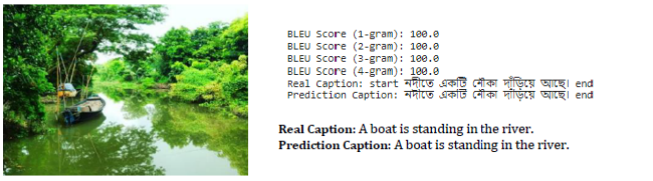
(a) The caption generated by Model 1; the result gives more detail of the picture than the human-generated one.



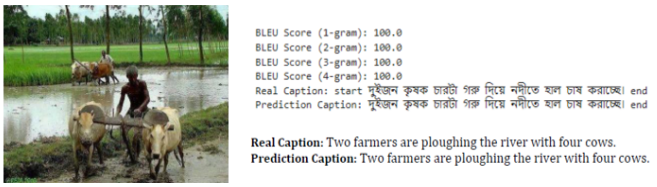
(b) The caption generated by Model 1; The caption is accurate to the human-generated one.



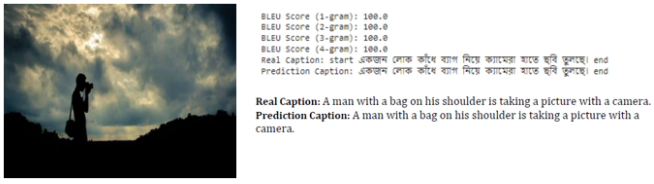
(c) The caption generated by Model 2; The result is almost exact with a small change of a term that has the same meaning.



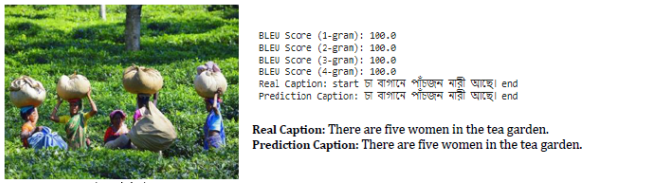
(d) The caption generated by Model 2; The caption is accurate to the human-generated one.



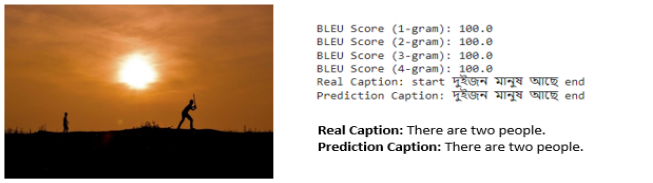
(e) The caption generated by Model 3; The caption is accurate to the human-generated one.



(f) The caption generated by Model 3; The caption is accurate to the human-generated one.



(g) The caption generated by Model 4; The caption is accurate to the human-generated one.



(h) The caption generated by Model 4; The caption is accurate to the human-generated one.

Fig. 5: Predicted captions produced by different models.

D. Result Analysis

The quantitative scores show that the models using the global concatenation approach produce good results for BLUE-1, 2, 3, 4, and meteor. It is also worth noting that the models score more than 80 for Blue 1, which is impressive. This suggests that these models generate more captions similar to the captions provided by humans for the test dataset. Moreover, Model 3 (InceptionV3+general) has almost similar outcomes to Model 2 (VGG19+concat) and Model 4 (InceptionV3+concat). Furthermore, it is demonstrated that Model 1 (VGG19+general) with a global general approach has the lowest proportion of BLUE scores. As a result, this model cannot offer enough captions for the test datasets that are equivalent to human-given annotations.

In terms of qualitative analysis, different samples of captions produced by the four trained models are judged by humans for the test photos. Model 1 (VGG19+general) showed a few generated captions that are identical to the human-derived captions and, in some instances, satisfactory captions that are not similar to the human-given ones. Model 2 (VGG19+concat) and Model 4 (InceptionV3+ concat) showed that these models generated accurate captions for the majority of the pictures. These models could also correctly recognize the unique objects in the image. Moreover, Model 3 (InceptionV3+ general) produced captions that are mostly similar to human-given ones; however, it could also generate captions that are distinct from human-generated descriptions but still match the photograph. Therefore, based on the qualitative and quantitative analysis, Model 4 (InceptionV3+ concat) was the strategy that produced the best results.

V. Conclusion

This study developed four global attention-based Bengali photo captioning models to produce captions in the language. Using a previously published dataset that focused on Bangladesh's culture, people, and geographical architecture, we created a new niche-based dataset targeting natural outdoor images. As a result, the model's ability to identify and describe related images of Bangladeshi landscapes, people, and

culture significantly improved. Furthermore, we experimented with two distinct global attention techniques (general and concat) using two CNN encoders, VGG19 and InceptionV3. According to the evaluation results, the model with the InceptionV3 encoder and the concatenation global approach offered cutting-edge performance, scoring 84.85 for BLUE-1 and metoer(0.86). On the other hand, the VGG19 model with the general strategy produced unsatisfactory BLUE scores, 47.91 for BLUE-1, meteor(0.51), with the majority of incorrect or non-similar results. Regarding qualitative analysis, the two alternative models, VGG19 and InceptionV3, created the most qualitative captions for the test dataset using the concatenation global technique. Depending on the applications, we can expand the dataset and annotations in the future and train more niche-based models. We suggest using an InceptionV3-based model with a global concatenation approach for future experiments. Furthermore, future research could explore another global attention method, known as the global dot technique, for Bengali captions.

REFERENCES

[1] J. Y. Jung, T. Steinberger, J. Kim, and M. S. Ackerman, ""so what? what's that to do with me?" expectations of people with visual impairments for image descriptions in their personal photo activities," in *Designing Interactive Systems Conference*, pp. 1893–1906, 2022.

[2] X. Li, M. Li, P. Yan, G. Li, Y. Jiang, H. Luo, and S. Yin, "Deep learning attention mechanism in medical image analysis: Basics and beyonds," *International Journal of Network Dynamics and Intelligence*, pp. 93–116, 2023.

[3] Z. Li, Y. Mu, Z. Sun, S. Song, J. Su, and J. Zhang, "Intention understanding in human–robot interaction based on visual-nlp semantics," *Frontiers in Neurorobotics*, vol. 14, p. 610139, 2021.

[4] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or, "An image is worth one word: Personalizing text-to-image generation using textual inversion," *arXiv preprint arXiv:2208.01618*, 2022.

[5] A. Goyal, M. Mandal, V. Hassija, M. Aloqaily, and V. Chamola, "Captionomaly: A deep learning toolbox for anomaly captioning in social surveillance systems," *IEEE Transactions on Computational Social Systems*, 2023.

[6] A. Yang, A. Nagrani, P. H. Seo, A. Miech, J. Pont-Tuset, I. Laptev, J. Sivic, and C. Schmid, "Vid2seq: Large-scale pretraining of a visual language model for dense video captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10714–10726, 2023.

[7] L. Xu, Q. Tang, J. Lv, B. Zheng, X. Zeng, and W. Li, "Deep image captioning: A review of methods, trends and future challenges," *Neurocomputing*, p. 126287, 2023.

[8] F. T. Zohora and Z. Abedin, "Bangla image captioning with bidirectional gru & attention mechanism," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 306–311, IEEE, 2022.

[9] A. A. Osman, M. A. W. Shalaby, M. M. Soliman, and K. M. Elsayed, "A survey on attention-based models for image captioning," *International Journal of Advanced Computer Science and Applications*, vol. 14, no. 2, 2023.

[10] Z. Zhang, W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun, "Global visual feature and linguistic state guided attention for remote sensing image captioning," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–16, 2021.

[11] X. Li, W. Lan, J. Dong, and H. Liu, "Adding chinese captions to images," in *Proceedings of the 2016 ACM on international conference on multimedia retrieval*, pp. 271–275, 2016.

[12] M. Liu, H. Hu, L. Li, Y. Yu, and W. Guan, "Chinese image caption generation via visual attention and topic modeling," *IEEE Transactions on Cybernetics*, vol. 52, no. 2, pp. 1247–1257, 2022.

[13] J. Gao, Y. Zhou, L. Philip, S. Joty, and J. Gu, "Unison: Unpaired cross-lingual image captioning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, pp. 10654–10662, 2022.

[14] B. Zhang, L. Zhou, S. Song, L. Chen, Z. Jiang, and J. Zhang, "Image captioning in chinese and its application for children with autism spectrum disorder," in *Proceedings of the 2020 12th International Conference on Machine Learning and Computing*, pp. 426–432, 2020.

[15] T. Wei, W. Yuan, J. Luo, W. Zhang, and L. Lu, "Vlca: vision-language aligning model with cross-modal attention for bilingual remote sensing image captioning," *Journal of Systems Engineering and Electronics*, vol. 34, no. 1, pp. 9–18, 2023.

[16] Y. Yoshikawa, Y. Shigeto, and A. Takeuchi, "Stair captions: Constructing a large-scale japanese image caption dataset," *arXiv preprint arXiv:1705.00823*, 2017.

[17] A. Rathi, "Deep learning apporach for image captioning in hindi language," in *2020 International Conference on Computer, Electrical Communication Engineering (ICCECE)*, pp. 1–8, 2020.

[18] R. Dhir, S. K. Mishra, S. Saha, and P. Bhattacharyya, "A deep attention based framework for image caption generation in hindi language," *Computación y Sistemas*, vol. 23, no. 3, pp. 693–701, 2019.

[19] S. K. Mishra, R. Dhir, S. Saha, P. Bhattacharyya, and A. K. Singh, "Image captioning in hindi language using transformer networks," *Computers & Electrical Engineering*, vol. 92, p. 107114, 2021.

[20] J. Emami, P. Nugues, A. Elnagar, and I. Afyouni, "Arabic image captioning using pre-training of deep bidirectional transformers," in *Proceedings of the 15th International Conference on Natural Language Generation*, pp. 40–51, 2022.

[21] B. D. Yılmaz, A. E. Demir, E. B. Sönmez, and T. Yıldız, "Image captioning in turkish language," in *2019 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 1–5, IEEE, 2019.

[22] M. R. S. Mahadi, A. Arifianto, and K. N. Ramadhani, "Adaptive attention generation for indonesian image captioning," in *2020 8th International Conference on Information and Communication Technology (ICoICT)*, pp. 1–6, 2020.

[23] B. G. Do, D. C. Bui, N. D. Vo, and K. Nguyen, "A multi-scale approach for vietnamese image captioning in healthcare domain," in *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, pp. 142–147, IEEE, 2022.

[24] M. Rahman, N. Mohammed, N. Mansoor, and S. Momen, "Chittron: An automatic bangla image captioning system," *Procedia Computer Science*, vol. 154, pp. 636–642, 2019.

[25] H. Al Faraby, M. M. Azad, M. R. Fedous, M. K. Morol, *et al.*, "Image to bengali caption generation using deep cnn and bidirectional gated recurrent unit," in *2020 23rd international conference on computer and information technology (ICCIT)*, pp. 1–6, IEEE, 2020.

[26] M. Faiyaz Khan, S. Sadiq-Ur-Rahman, and M. Saiful Islam, "Improved bengali image captioning via deep convolutional neural network based encoder-decoder model," in *Proceedings of International Joint Conference on Advances in Computational Intelligence: IJCACI 2020*, pp. 217–229, Springer, 2021.

[27] A. S. Ami, M. Humaira, M. A. R. K. Jim, S. Paul, and F. M. Shah, "Bengali image captioning with visual attention," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2020.

[28] B. Das, R. Pal, M. Majumder, S. Phadikar, and A. A. Sekh, "A visual attention-based model for bengali image captioning," *SN Computer Science*, vol. 4, no. 2, p. 208, 2023.

[29] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv preprint arXiv:1409.0473*, 2014.

[30] M.-T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," *arXiv preprint arXiv:1508.04025*, 2015.

[31] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

[32] 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," *arXiv preprint arXiv:2109.05218*, 2021.

[33] M. A. H. Palash, M. A. A. Nasim, S. Saha, F. Afrin, R. Mallik, and S. Samiappan, "Bangla image caption generation through cnn-transformer based encoder-decoder network," in *Proceedings of International Conference on Fourth Industrial Revolution and Beyond 2021*, pp. 631–644, Springer, 2022.