2024 6th International Conference on Sustainable Technologies for Industry 5.0 (STI), 14-15 December, Dhaka

Generating Bangla Image Captions with Deep Learning Techniques

Md. Anwar Hossain*, Mirza AFM Rashidul Hasan†, Sajeeb Kumar Ray*, Naima Islam*
*Dept of Information and Communication Engineering, Pabna University of Science and Technology, Bangladesh
†Dept of Information and Communication Engineering, University of Rajshahi, Bangladesh
Email: manwar.ice@pust.ac.bd, mirzahasanice@gmail.com, sajeeb.ray.ice@gmail.com, naimaislam204100@gmail.com

***Abstract*—The importance of Bangla image captioning is rooted in the need to bridge the gap between visual content and textual descriptions in Bengali, thereby facilitating more accessible and inclusive technologies for Bengali-speaking individuals. Image captioning entails the creation of textual descriptions for images through the application of deep learning techniques that combine methodologies from computer vision and natural language processing to accurately recognize and depict visual content. This study presents an innovative methodology that employs EfficientNetB4 and ResNet-50 architectures for the extraction of features. The selection of these models followed a comprehensive evaluation of numerous alternatives, as they exhibited exceptional performance, rendering them well-suited for the objectives of this investigation. A pivotal contribution of this analysis is the introduction of the newly established BanglaView dataset, designed specifically for Bangla captioning tasks, used alongside the widely recognized Flickr30k dataset. The empirical results reveal that EfficientNetB4 surpasses ResNet-50, attaining a BLEU score of 0.54 after merely 10 training epochs, underscoring its efficacy in generating coherent Bangla captions. These outcomes not only illustrate technological progress but also highlight the potential of this methodology to stimulate innovations that promote linguistic diversity and improve user experiences in domains such as accessibility and digital communication.**

Keywords—Bangla Image Captioning, EfficientNetB4, ResNet-50, BanglaView, Flicker30k

I. Introduction

Image captioning requires developing a textual description for an image using techniques from computer vision and natural language processing. This includes scrutinizing the visual components of an image to recognize objects, activities, and settings, and subsequently formulating a consistent and contextually appropriate sentence or group of sentences that depict the image precisely [1]. The objective is to bridge the division between visual data and natural language, empowering machines to comprehend and convey the contents of images in a manner readable by humans. The technique of generating image captions has a wide range of practical uses. It can assist in the automation of driving vehicles, the development of facial recognition systems, as well as the provision of descriptions for visually impaired individuals and the enhancement of photo search capabilities [2]. In image captioning, individuals typically begin by visually examining the image to identify the objects present. The human brain possesses a vast array of local language vocabulary that aids in this process. Following the object identification, the brain accesses the specific vocabulary related to the detected object and formulates a description for the image [3]. Both top-down and bottom-up methods are commonly used for captioning images. In the top-down approach, words are created from the visual information in an image, but in the bottom-up approach, words explaining various parts of a picture are started and then combined into a phrase. These approaches generate coherent phrases by using linguistic models. Convolutional neural networks encode the visual data, and recurrent neural networks decode it to produce captions, in the context of an encoder-decoder employed in recent top-down techniques [5].

Deep learning models have exhibited considerable promise in producing precise and meaningful image descriptions in recent developments within image captioning. This study lies in creating an image captioning system in the Bengali language, utilizing deep learning capabilities. We extracted features from our work using the EfficientNetB4 and ResNet-50 models. In our research, we have made significant advancements in the area of Bangla image captioning through the following initiatives:

- We introduced a sophisticated deep neural network technique employing EfficientNetB4 to create Bangla image descriptions, illustrating the efficacy of the model within this particular linguistic framework.
- We utilized two extensive datasets for this purpose: the Flickr30k dataset, which consists of 31,783 images, and the BanglaView dataset, featuring a substantial 1,58,915 captions. This amalgamation established a strong basis for training and validating our model.
- To meticulously assess the performance of our proposed model, we employed a subset of 960 images for testing, ensuring a comprehensive evaluation of its captioning capabilities.

These contributions highlight the potential of our methodology in enriching the precision and quality of Bangla image captioning, setting the stage for future progress in this field.

II. Related Works

The task of producing a natural language description of video or picture data has long been researched in the field of computer vision. Deep learning is one of the most important methods for image caption, but most of the deep neural networks have many layers and interconnected neurons, which are used to extract features from existing data, such as image data, and to learn the complex relationship between the corresponding features of these data [6, 7]. Huang et al. [8] provide a Multimodal Transformer (MT) model for captioning images that is capable of producing precise captions and

979-8-3503-9431-3/23/$31.00 ©2024 IEEE

carrying out intricate multimodal reasoning. The three-layer LSTM encoder-decoder architecture suggested by Xiao et al. [9] successfully combined images and textual information to generate relevant descriptions. The authors explored the potential of leveraging the output generated by the central layer of LSTMs to enhance the language generation capabilities of the uppermost LSTM. Takashi Miyazaki and N. Shimizu developed a dataset called "YJ captions 26k" for Japanese image captioning. They employed three different learning philosophies in their investigation and found that transfer learning is the best technique for producing Japanese image descriptions [10]. Faiyaz et al. [11] demonstrated a deep learning model for Bengali picture captioning. The ResNet-50 pre-trained model was utilized to train a one-dimensional convolutional neural network (CNN) to retrieve visual characteristics. Chen et al. [12] used a semantic attention model to suggest a new model. CNNs that have already been trained, such as VGG-16 or ResNet50, were compared to it. ResNet-50 produced more encouraging findings than VGG-16, according to their tests. Sharma et al. [13] recommend use it for captioning images in English. For the sequence modeling, they employed a transformer and InceptionResNet0-v2 for the extraction of visual features. Fang et al. [14] suggested a framework using a dataset of images and relevant descriptions. Detectors of frequently occurring terms in the descriptions of the examined image are produced by the system using a poorly supervised learning technique. Next, using this word list, sentences are created and categorized based on how closely they resemble the contents of the images. Kulkarni et al. [15] presented a compact connection to identify things in an image by using recognition models and large semantic data. Compared to previous models, this one is better at extracting pertinent captions for images. Additionally, two advancements in caption retrieval were presented in Ordonez et al. [16] to score the relationship between captions and photos: obtaining all accessible images and retrieving captions based on the geometric distance between the scene and object.

III. Materials and Methods

A. Dataset Loading

In our study both Flickr30k and BanglaView are indispensable assets for the advancement of image captioning models. Flickr30k contributes a well-known dataset for generic image captioning assignments, whereas BanglaView presents a tailored dataset for the Bengali language, facilitating progress in multilingual and culturally relevant image captioning technologies.

Flickr30k [17], the dataset renowned for image captioning, is composed of 31,783 images. Within the realm of image captioning tasks, this dataset is frequently utilized to construct and assess models. And, BanglaView [18] is a recently released dataset designed specifically for tasks involving the Bangla language picture captioning. Captions of BanglaView are directly representing the Flicker30k images, but not automatically translated in Bangla from flicker30k English captions. With 158,915 captions in all, it is a vast collection of captions that makes it an excellent tool for testing and training picture captioning algorithms. With a high vocabulary size of 25,444 words, the dataset allows models to learn and provide a broad range of descriptions and expressions. BanglaView also supports rich, contextual subtitles up to 67 words in length, which aids in grasping the finer points and subtleties of the visual material. With its attention to the linguistic and cultural peculiarities of the Bangla language, this extensive dataset is essential for furthering study and development in the field of picture captioning in Bangla.

B. Feature extraction models

A standard mission in image captioning involves two primary steps: (1) the extraction of features and (2) the generation of text. Feature extraction acts as the basis of image captioning, as the efficiency of the extracted features significantly influences the quality of the resulting captions. In initial methods of image captioning, features were predominantly crafted manually [4]. For Bangla image captioning we explored several models, after comparing their performance best two models are presented, ResNet-50 [19] and EfficientNetB4 [20] for feature extraction. These models play a vital role in extracting significant features from images, which are subsequently utilized to create precise and contextually suitable captions in Bangla.

ResNet-50: ResNet-50's deep residual network design with 50 layers makes it popular for feature extraction in image captioning applications. By including skip connections, the approach helps to overcome the issue of vanishing gradients and makes it possible to train extremely deep networks. The extraction of more intricate and complicated elements from images is made possible by these skip connections. The pre-trained ResNet50 model offers highly discriminative features that improve the descriptive accuracy and relevancy of output captions when integrated into image captioning systems. To create sophisticated and efficient image captioning models, ResNet50 derived strong and complete picture representations.

EfficientNetB4: EfficientNetB4's balanced and efficient architecture makes it a popular choice for feature extraction in the field of image captioning. EfficientNetB4 is a computationally efficient model that works remarkably well on a variety of picture tasks thanks to its compound scaling strategy, which consistently adjusts depth, width, and resolution. In our study, we use the hierarchical features discovered through extensive image datasets by using a pre-trained EfficientNetB4 model. EfficientNetB4 is a potent tool in the image captioning pipeline because of these extensive properties, which provide a strong basis for producing accurate and contextually appropriate captions.

C. System Architecture

The program generates descriptive captions by utilizing an integrated strategy that combines textual data and visual attributes. Several essential elements make up the system architecture shown in Fig. 1.

Extraction of picture Features: A sequence of dense and normalizing layers are used by the model to handle picture features. These layers improve the visual characteristics and get them ready to be integrated with textual information.

Text Processing: A Gated Recurrent Unit (GRU) layer is used to incorporate and process text data that takes the shape of sequences. In order to capture the temporal relationships in the text, the GRU layer further refines the dense vector representation created by the embedding layer.

Feature Integration: A number of procedures are used to merge the text and picture features that have been processed. The strengths of recurrent neural networks (RNNs) for sequence processing and convolutional neural networks (CNNs) for visual feature extraction are combined in this integration.

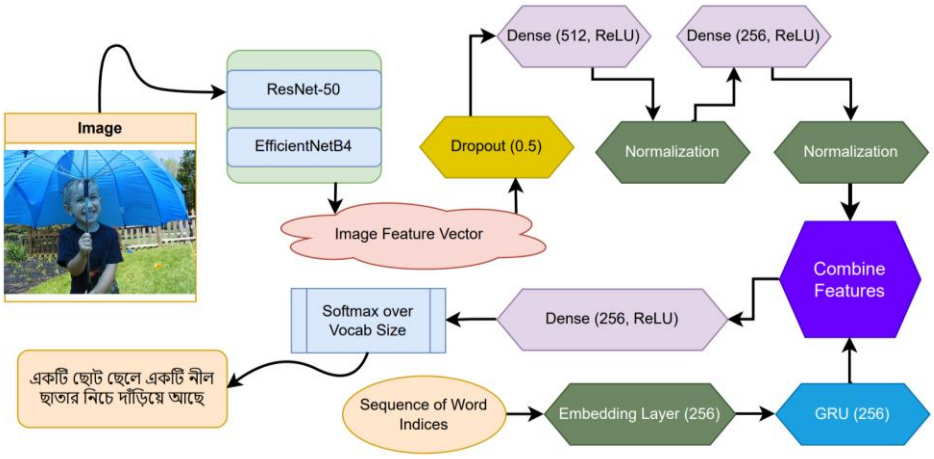


Fig. 1: System Architecture of Bangla caption generation model

Caption Generation: To get the final result, the combined features are run through more thick layers. The result displays the most likely word in the sequence based on a probability distribution over the vocabulary. In order to create every word in the caption, the model does this repeatedly until the sequence is finished.

The architecture is made to efficiently use RNNs to preserve the text's sequential context and CNNs to capture rich picture attributes, allowing for the creation of captions for images that are both coherent and pertinent to their context.

D. Training

The model underwent training utilizing EfficientNetB4 and ResNet50 as the feature extraction modules across 10 epochs, employing an Adam optimizer, a batch size of 32, and employing categorical crossentropy as the loss function. The progression of the training was monitored using metrics including training loss, training accuracy, and validation accuracy, presented in Fig. 2 and Fig. 3.

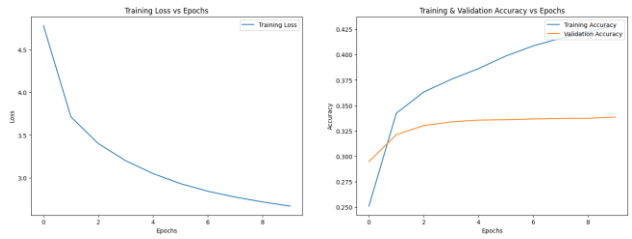


Fig. 2: EfficientNetB4 training and validation graph

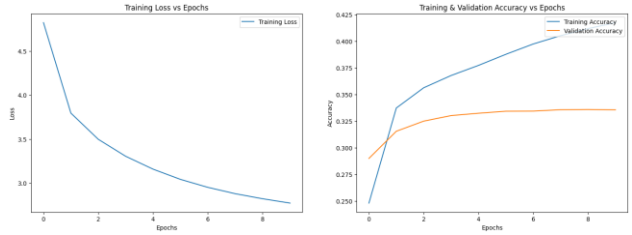


Fig. 3: ResNet50 training and validation graph

In the case of EfficientNetB4, the training loss exhibited a consistent decline from 4.7804 to 2.6693, indicating the model's effective learning from the training dataset. Concurrently, the training accuracy displayed steady enhancement, culminating at 0.4286 towards the conclusion of the training period. The validation accuracy also manifested gradual amelioration, commencing at 0.2947 and stabilizing at approximately 0.3385, as evidenced in the recorded data.

In contrast, for ResNet50, a similar pattern emerged with the training loss diminishing from 4.8227 to 2.7726, signifying proficient learning. The training accuracy escalated from 0.2481 to 0.4174. Likewise, the validation accuracy demonstrated progressive enhancement, initiating at 0.2900 and plateauing at around 0.3356.

These findings imply an increased proficiency of both models in predicting the training data, with EfficientNetB4 achieving marginally superior accuracy compared to ResNet50. Nevertheless, the validation accuracy for both models did not exhibit substantial improvement beyond a specific threshold, indicating a potential issue with overfitting.

E. Performance Evaluation

1) *Bilingual Evaluation Understudy (BLEU):* The metric known as BLEU, as defined by Papineni et al. [21] and symbolized as B, is employed for the automated assessment of machine-translated text. Quantified on a scale from 0 to 1, the BLEU score functions as an indicator of the similarity between a machine-translated text and a set of high-quality reference translations. The range of BLEU scores along with their corresponding descriptions of translation quality can be found in Table I.

TABLE I. THE BLEU SCORE RANGES AND THEIR CORRESPONDING TRANSLATION QUALITY DESCRIPTIONS

Score Range	Quality Description	Score Range	Quality Description
0.00 - 0.10	Very Poor	0.50 - 0.60	Excellent
0.10 - 0.20	Poor	0.60 - 0.70	Outstanding
0.20 - 0.30	Fair	0.70 - 0.80	Near Perfect
0.30 - 0.40	Good	0.80 - 0.90	Exceptional
0.40 - 0.50	Very Good	0.90 - 1.00	Perfect

2) *ROUGE:* ROUGE, standing for Recall-Oriented Understudy for Gisting Evaluation, encompasses a series of metrics utilized to evaluate the effectiveness of summaries and text produced by machines.

High ROUGE scores suggest that the Bangla captions created closely resemble the reference captions in terms of individual words and phrase structures. Such scores indicate that the model excels in generating captions that are not only accurate but also logically coherent. Nonetheless, it is imperative to take into account additional metrics like BLEU, METEOR, or assessments by human evaluators to carry out a comprehensive and thorough evaluation of the model's efficiency and efficacy.

reflecting the increasing difficulty in correctly generating longer sequences. Specifically, EfficientNetB4 scored 0.3442 in BLEU-2, 0.2044 in BLEU-3, and 0.1163 in BLEU-4, while ResNet-50 scored 0.3101 in BLEU-2, 0.1763 in BLEU-3, and 0.0944 in BLEU-4. These results indicate that while both models struggle more with longer sequences, EfficientNetB4 generally performs better in generating accurate captions.

TABLE II. PERFORMANCE EVALUATION OF CAPTION GENERATION MODEL USING DIFFERENT FEATURE EXTRACTION MODELS

Feature extraction model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L
EfficientNetB4	0.5396	0.3442	0.2044	0.1163	0.9896	0.9621	0.9893
ResNet-50	0.5011	0.3101	0.1763	0.0944	0.9861	0.9517	0.9861

As depicted in Table II, the model's performance was assessed utilizing various feature extraction models, such as EfficientNetB4 and ResNet50. These models underwent evaluation based on different metrics, including BLEU scores (1 to 4) and ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L), offering a comprehensive evaluation of their caption generation capabilities. The outcomes demonstrate the performance of each feature extraction model in terms of generating captions that closely align with the reference translations, indicating their efficacy in this task.

IV. RESULTS AND DISCUSSION

In this study, we evaluated the performance of our image captioning model using two different feature extraction models: EfficientNetB4 and ResNet-50. We used only 10 epochs to train the model and then tested it using 960 images. The results are presented in the Table II, showcasing various performance metrics including BLEU scores (1 to 4) and ROUGE scores (ROUGE-1, ROUGE-2, and ROUGE-L).

The BLEU scores provide an assessment of the generated captions' accuracy by comparing them to reference captions across different n-grams. The EfficientNetB4 model achieved a BLEU-1 score of 0.5396, which is higher than the ResNet-50's score of 0.5011. As the n-gram size increases, both models show a decline in scores,

The ROUGE (Recall-Oriented Understudy for Gisting Evaluation) scores measure the overlap between the generated captions and reference captions, focusing on n-grams and the longest common subsequence. EfficientNetB4 achieved very high ROUGE scores, with ROUGE-1 at 0.9896, ROUGE-2 at 0.9621, and ROUGE-L at 0.9893, indicating a high overlap with the reference captions. ResNet-50, while slightly lower, still performed well with ROUGE-1 at 0.9861, ROUGE-2 at 0.9517, and ROUGE-L at 0.98618. These high ROUGE scores for both models suggest that the generated captions maintain a strong resemblance to the reference captions in terms of both n-gram and sequence overlap, with EfficientNetB4 demonstrating a marginally better performance.

In Table III the reference captions describe a child, specifically a girl, climbing up the stairs to a playhouse. The EfficientNetB4 model accurately captures the action and setting but misidentifies the child as gender-neutral. In contrast, ResNet50 incorrectly describes the child tearing off a structure, diverging significantly from the actual scene. This discrepancy highlights the challenge in accurately identifying and describing specific actions and participants in an image.

TABLE III. IMAGE CAPTION PREDICTIONS FOR A CHILD CLIMBING A PLAYHOUSE

Flickr30k Image

BanglaView Captions

C1: গোলাপী পোশাক পরা একটি শিশু প্রবেশ পথে সিঁড়ি বেয়ে উপরে উঠছে (A child in a pink dress is climbing up the stairs at the entrance)

C2: একটি গোলাপী পোশাক পরা ছোট্ট মেয়ে একটি কাঠের ঘরে যাচ্ছে (A little girl in a pink dress is going into a wooden house)

C3: একটি ছোট মেয়ে তার খেলার ঘরের সিঁড়ি বেয়ে উঠছে (A little girl is climbing the stairs to her playhouse)

C4: একটি ছোট মেয়ে একটি কাঠের খেলাঘরে আরোহণ করছে (A little girl is climbing into a wooden playhouse)

C5: একটি মেয়ে একটি কাঠের ভবনে যাচ্ছে (A girl is going into a wooden building)

Model Generated

EfficientNetB4: একটি ছোট শিশু একটি কাঠের খেলাঘরে আরোহণ করছে (A small child is climbing into a wooden playhouse)

ResNet50: একটি ছোট শিশু একটি কাঠের কাঠামো থেকে একটি কাঠামো ছিঁড়ে যাচ্ছে (A little child is tearing off a structure from a wooden structure)

In Table IV the reference captions for this image describe a performer on stage with specific attire and accessories. EfficientNetB4 offers a generic description of a man singing into a microphone, lacking the detailed context provided by the references. ResNet50, however, adds incorrect details, such as a mime outfit, which are not present in the image. This table highlights the models' struggle to accurately capture and relay detailed visual attributes and context.

In Table V the reference captions in this table describe hikers resting in front of mountains. EfficientNetB4's prediction inaccurately depicts people sitting at the top of a mountain, while ResNet50 describes a trekking scene, both diverging from the resting context of the references. These inaccuracies demonstrate the models' difficulty in distinguishing between similar but contextually different activities such as resting and trekking, and the importance of capturing the correct context in image captioning tasks.

Although our study revealed notable progress, it is important to recognize that there were a number of limitations. Despite the large number of captions available in the BanglaView dataset, computational limitations prevented us from taking full advantage of this resource. We were unable to train the models thoroughly because of this restriction. As a consequence, even though our models performed well, more training using more epochs is necessary to fully fulfill their potential and produce more thorough and precise outcomes. Improving the reliability and efficacy of Bangla picture captioning will require addressing these resource constraints in subsequent study.

In forthcoming research, our goal is to tackle the computational intricacies linked with sophisticated neural network architectures such as ResNet-50 and EfficientNetB4.

TABLE IV. IMAGE CAPTION PREDICTIONS FOR A PERFORMER SINGING ON STAGE

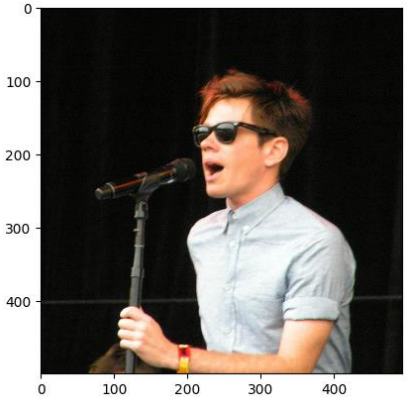
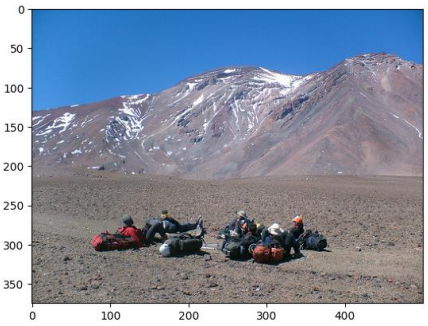
Flicker30k Image	*BanglaView Captions*
	C1: একটি বোতামযুক্ত শার্ট এবং ভিনটেজ-স্টাইলের সানগ্লাস পরে মঞ্চে একজন তরুণ অভিনয়শিল্পী তার গান গাইছেন (A young performer wearing a buttoned shirt and vintage-style sunglasses is singing his song on stage) C2: সানগ্লাস এবং হালকা নীল শর্ট-হাতা শার্ট পরা একজন শ্যামাঙ্গিনী ব্যক্তি একটি মাইক্রোফোনে গান গাচ্ছেন (A brunette person wearing sunglasses and a light blue short-sleeved shirt is singing into a microphone) C3: হালকা নীল শার্ট পরা একজন যুবক মাইক্রোফোনে কথা বলছে বা গান করছে (A young man in a light blue shirt is speaking or singing into a microphone) C4: সানগ্লাস পরা একজন সাদা লোক মাইক্রোফোন নিয়ে গান গাইছে (A white man wearing sunglasses is singing with a microphone) C5: একজন তীক্ষ্ণ পোশাক পরা লোক মাইক্রোফোনে গান গাইছে (A sharply dressed man is singing into a microphone)
	Model Generated
	EfficientNetB4: একজন লোক মাইক্রোফোনে গান গাইছে (A man is singing into a microphone) ResNet50: একটি কালো এবং সাদা মাইম এবং একটি সাদা শার্ট পরা একজন ব্যক্তি একটি মাইক্রোফোনে গান গাইছেন (A person wearing a black and white mime outfit and a white shirt is singing into a microphone)

TABLE V. IMAGE CAPTION PREDICTIONS FOR HIKERS RESTING

Flicker30k Image	*BanglaView Captions*
	C1: একদল হাইকার কিছু পাহাড়ের সামনে মাটিতে বিশ্রাম নিচ্ছে (A group of hikers is resting on the ground in front of some mountains) C2: বেশ কিছু হাইকার একটি পাহাড়ের সামনে তাদের গিয়ার নিয়ে বিশ্রাম নিচ্ছেন (Several hikers are resting with their gear in front of a mountain) C3: পর্বতারোহীদের দল একটি পাহাড়ের সামনে বিশ্রাম নিচ্ছে (A group of mountaineers is resting in front of a mountain) C4: মানুষ পটভূমিতে পাহাড়ের সাথে ক্যাম্প করে (People are camping with mountains in the background) C5: একদল ব্যাকপ্যাকার শুকনো মাটিতে শুয়ে আছে (A group of backpackers is lying on the dry ground)
	Model Generated
	EfficientNetB4: একটি পাহাড়ের কাছে একটি পাহাড়ের চূড়ায় তিনজন লোক বসে আছে (Three people are sitting at the top of a mountain near another mountain) ResNet50: পর্বতারোহীদের একটি দল একটি পাহাড়ের উপর দিয়ে ট্রেক করছে (A group of mountaineers is trekking over a mountain)

Through the optimization of these architectures or the exploration of more efficient alternatives, our aim is to diminish the time required for execution and the resources needed. This enhancement will facilitate more comprehensive evaluations spanning over 10 epochs, potentially yielding more substantial and thorough outcomes. These advancements have the potential to enhance the precision and caliber of Bangla image captioning, thereby paving the way for more pragmatic and efficient applications in this domain.

V. Conclusion

In conclusion, our research emphasizes the efficacy of employing advanced neural network architectures like ResNet-50 and EfficientNetB4 for extracting features in the context of Bangla image captioning. Although these models' intricacy demands significant computational resources and longer processing times, the outcomes exhibit their considerable capabilities. Notably, the EfficientNetB4 model distinguished itself with outstanding performance, delivering noteworthy results in only 10 epochs. This effectiveness highlights the model's appropriateness for our objective, striking a balance between precision and computational feasibility. The human-like captions produced through our methodology further affirm the model's effectiveness, illustrating its capacity to generate contextually accurate and relevant descriptions of Bangla images. These results suggest a promising avenue for future exploration and utilization in Bangla image captioning, emphasizing the potential of EfficientNetB4 to elevate the quality and appropriateness of the generated captions.

Acknowledgment

We acknowledge the support in conducting this research from the Department of Information and Communication Engineering (ICE) at Pabna University of Science and Technology (PUST).

References

[1] 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.

[2] 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, Lecture Notes in Networks and Systems, vol. 437, Singapore: Springer, 2021.

[3] M. A. Hossain, M. Hasan, E. Hossen, A. Md, M. Faruk, A. F. M. A. Abadin, and M. Ali, "Automatic Bangla Image Captioning Based on Transformer Model in Deep Learning," International Journal of Advanced Computer Science and Applications, vol. 14, 2023.

[4] K. Zhao and W. Xiong, "Exploring region features in remote sensing image captioning," International Journal of Applied Earth Observation and Geoinformation, vol. 127, pp. 103672, 2024.

[5] A. M. Rinaldi, C. Russo, and C. Tommasino, "Automatic image captioning combining natural language processing and deep neural networks," Results in Engineering, vol. 18, pp. 101107, 2023.

[6] C. Cheng, C. Li, Y. Han, and Y. Zhu, "A semi-supervised deep learning image caption model based on Pseudo Label and N-gram," International Journal of Approximate Reasoning, vol. 131, pp. 93-107, 2021.

[7] M. A. Hossain, S. K. Ray, N. Islam, A. Alamin, and M. A. F. M. R. Hasan, "Enhanced human activity recognition through deep multi-layer perceptron on the UCI-HAR dataset," International Journal of Advances in Applied Sciences, vol. 13, pp. 429-438, 2024.

[8] J. Yu, J. Li, Z. Yu, and Q. Huang, "Multimodal Transformer With Multi-View Visual Representation for Image Captioning," IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 12, pp. 4467-4480, Dec. 2020.

[9] X. Xiao, L. Wang, K. Ding, S. Xiang, and C. Pan, "Deep hierarchical encoder–decoder network for image captioning," IEEE Transactions on Multimedia, pp. 2942-2956, 2019.

[10] T. Miyazaki and N. Shimizu, "Cross-lingual image caption generation," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pp. 1780–1790, 2016.

[11] M. F. Khan, S. M. S. 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, M. S. Uddin and J. C. Bansal, Eds., Singapore: Springer, 2021.

[12] Q. Chen, W. Li, Y. Lei, X. Liu, C. Luo, and Y. He, "Cross-Lingual Sentiment Relation Capturing for Cross-Lingual Sentiment Analysis," in Advances in Information Retrieval. ECIR 2017, J. Jose et al., Eds., Lecture Notes in Computer Science, vol. 10193, Springer, 2017.

[13] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning," in Annual Meeting of the Association for Computational Linguistics, 2018.

[14] H. Fang et al., "From captions to visual concepts and back," in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 1473-1482, 2015.

[15] G. Kulkarni et al., "BabyTalk: Understanding and Generating Simple Image Descriptions," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 12, pp. 2891-2903, Dec. 2013.

[16] V. Ordonez, X. Han, P. Kuznetsova, et al., "Large Scale Retrieval and Generation of Image Descriptions," Int J Comput Vis, vol. 119, pp. 46–59, 2016.

[17] B. A. Plummer, L. Wang, C. M. Cervantes, J. C. Caicedo, J. Hockenmaier, and S. Lazebnik, "Flickr30k Entities: Collecting Region-to-Phrase Correspondences for Richer Image-to-Sentence Models," in 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 2641-2649, 2015.

[18] M. A. Hossain and M. A. F. M. R. Hasan, "BanglaView: A Bangla Image Captioning Dataset," Mendeley Data, 2023.

[19] S. K. Satti, G. N. V. Rajareddy, P. Maddula, and N. V. V. Ravipati, "Image Caption Generation using ResNET-50 and LSTM," in 2023 IEEE Silchar Subsection Conference (SILCON), Silchar, India, pp. 1-6, 2023.

[20] S. V. Patnaik, R. Mukka, R. Devpreyo, and A. Wadhawan, "Image Caption Generator using EfficientNet," in 2022 10th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, pp. 1-5, 2022.

[21] K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a Method for Automatic Evaluation of Machine Translation," in Annual Meeting of the Association for Computational Linguistics, 2022.