Bengali Visual Genome: A Multimodal Dataset for Machine Translation and Image Captioning

Arghyadeep Sen[1], Shantipriya Parida[2], Ketan Kotwal[2], Subhadarshi Panda[3], Ondřej Bojar[4], and Satya Ranjan Dash[1]

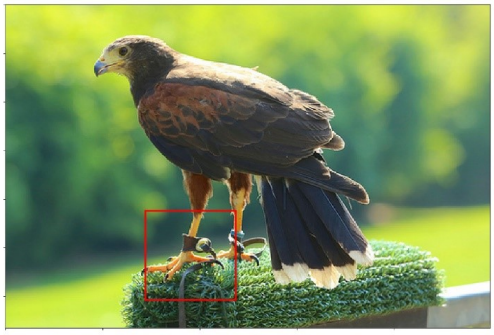
[1] KIIT University, Bhubaneswar, India
{2081012,sdashfca}@kiit.ac.in
[2] Idiap Research Institute, Martigny, Switzerland
{firstname.lastname}@idiap.ch
[3] Graduate Center, City University of New York, USA
spanda@gradcenter.cuny.edu
[4] Charles University, MFF, ÚFAL, Prague, Czech Republic
bojar@ufal.mff.cuni.cz
* Corresponding author: sdashfca@kiit.ac.in (Satya Ranjan Dash) .

Abstract Multimodal machine translation (MMT) refers to the extraction of information from more than one modality aiming at performance improvement by utilizing information collected from the modalities other than pure text. The availability of multimodal datasets, particularly for Indian regional languages, is still limited, and thus, there is a need to build such datasets for regional languages to promote the state of MMT research. In this work, we describe the process of creation of the Bengali Visual Genome (BVG) dataset. The BVG is the first multimodal dataset consisting of text and images suitable for English-to-Bengali multimodal machine translation tasks and multimodal research. We also demonstrate the sample use-cases of machine translation and region-specific image captioning using the new BVG dataset. These results can be considered as the baseline for subsequent research.

Keywords: Machine Translation · Multimodal Dataset · CNN · RNN · Image Captioning.

1 Introduction

In the broad area of machine learning or deep learning, multimodal processing refers to training models based on combined information sources such as image, audio, text, or video. Multimodal data facilitates learning features from various subsets of information sources (based on data modality) to improve the accuracy of the prediction. The multimodal machine translation includes information from more than one modality in the hope that additional modalities will contain useful alternative views of the input data [13]. Though machine translation performance reached near-human level for several language pairs (see e.g. [10]), it remains challenging to translate low resource languages or to effectively utilize other modalities (e.g. image, [9]).



English Text: The sharp bird talon.
Bengali Text: ধারালো পাখি টালন

Figure 1. A sample from the BVG dataset: an image with a specific region marked and its description in English and Bengali for text-only, multimodal and caption generation tasks.

Bengali (also known as Bangla)[5] is an Indo-Aryan language widely spoken in India and Bangladesh and considered as the 6th most spoken language of the world with approximately 230 million speakers. It follows the subject–object–verb (SOV) word order, and it is written in *Brahmic* script. To enrich the Bengali language with natural language processing (NLP) resources, we have developed the **BVG**: the first multimodal dataset for English-Bengali multimodal translation and also suitable for multimodal research.

The objective of the paper is twofold:

1. To describe the process of building the multimodal dataset for Bengali language suitable for English-to-Bengali machine translation and multimodal research.
2. To demonstrate some sample use cases of the newly created multimodal dataset: BVG.

2 Related Work

For the Bengali language, very limited work has been done in multimodal research including image captioning, and none in multimodal machine translation to the best of our knowledge due to lack of multimodal bi-lingual corpus.

An end-to-end image captioning framework for generating Bengali captions using the *BanglaLekhaImageCaptions* dataset [4]. In their work, the image features are extracted using the pretrained ResNet-50 and text (sentences) features

[5] `https://en.wikipedia.org/wiki/Bengali_language`

using a one-dimensional CNN. An automatic image captioning system "Chittron" proposed by [11] which uses VGG16 for generating image features and a staked LSTM for caption generation. An another image captioning dataset in Bengali which consists of 500 images of lifestyle, festivals along with its associated captions is available for research [3].

3 Dataset Preparation

To avoid any bias, we did not use any machine translation system. We worked with human volunteers for the sentence-level translation. The dataset statistics are provided in Table 1.

Dataset	**Number of items**
Training dataset	28930
Development Set (D-Test)	998
Evaluation Set (E-Test)	1595
Challenge Test Set (C-Test)	1400

Table 1. Brief details of the Bengali Visual Genome (BVG) dataset.

3.1 Training Set Preparation

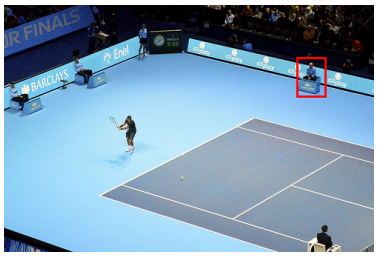
We follow the same selection of short English segments (captions) and the associated images from Visual Genome as HVG 1.1[6] has. For BVG, volunteers manually translated these captions from English to Bengali taking the associated images and their region into account as shown in Figure 1. The translation is performed by human volunteers (native Bengali speakers) without using any machine translation system.

3.2 Test Set Preparation

The development test (D-Test), evaluation test (E-Test), and challenge test (C-Test) sets prepared in the same fashion as the training. The C-Test was created for the WAT2019 multi-modal task[7][7] by searching for (particularly) ambiguous English words based on the embedding similarity and manually selecting those where the image helps to resolve the ambiguity. The surrounding words in the sentence however also often include sufficient cues to identify the correct meaning of the ambiguous word. The sample ambiguous words in used in the challenge test are: *Stand*, *Court*, *Players*, *Cross*, *Second*, Block, *Fast*, *Date*, *Characters*, *Stamp*, *English*, *Fair*, *Fine*, *Press*, *Forms*, *Springs*, *Models*, *Forces*, and *Penalty*

[6] `https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3267`
[7] `https://ufal.mff.cuni.cz/hindi-visual-genome/wat-2019-multimodal-task`

(a) Four men on *court*.



(b) Judge on side of *court* on blue platform.

Figure 2. Sample item from BVG challenge test set. Machine translation system fails to translate correctly into Bengali for the ambiguous word *court* (e.g. tennis or judicial) present in the source English text without using the associated image. The English text in image (b) even more ambiguous containing the words *judge* and *court* refers to the 'tennis court' rather 'judicial court'.

[8]. The NMT system fails to translate the sentence containing one or more listed ambiguous words without referring its associated image as shown in the Fig. 2.

4 Sample Applications of BVG

4.1 Text-Only Translation

For the text-only translation, we have first trained the SentencePiece subword units [6] setting the maximum vocabulary size to 8k. The vocabulary was learned jointly on the source and target sentences. We set the number of encoder and decoder layers to 3 each, and the number of heads was set to 8. The hidden size was set to 128, along with the dropout value of 0.1. We initialized the model parameters using Xavier initialization [2] and used the Adam optimizer [5] with a learning rate of $5e-4$ for optimizing model parameters. Gradient clipping was used to clip gradients greater than 1. The training was stopped when the development loss did not improve for 5 consecutive epochs.

The training, dev, test, and challenge test sizes for the neural machine translation (NMT) experiment are shown in Table 2.

4.2 Bengali Caption Generation

In this section, we demonstrate the use-case of region-specific image caption generation. We provide a baseline method for generating Bengali captions to the area enclosed by the bounding box as provided by the BVG dataset. In [14], O. Vinayls *et al.* have proposed an end-to-end deep neural network for generating the captions for the entire image. Their network consists of a vision CNN (used as

Dataset	#Sentences	#Tokens	
		EN	BN
Train	28930	143156	113993
D-Test	998	4922	3936
E-Test	1595	7853	6408
C-Test	1400	8186	6657

Table 2. Details of the processed BVG for NMT experiments. The number of tokens for English (EN) and Bengali (BN) for each set are reported.

D-Test BLEU	E-Test BLEU	C-Test BLEU
42.8	35.6	17.2

Table 3. Results of text-to-text translation on the BVG dataset.

a feature extractor for images) followed by a language generating RNN (to obtain caption as a sequence). Considering the model from [14] as the reference model, we incorporate the following modification for region-specific caption generation. The overall architecture of the modified network is shown in Figure 3.

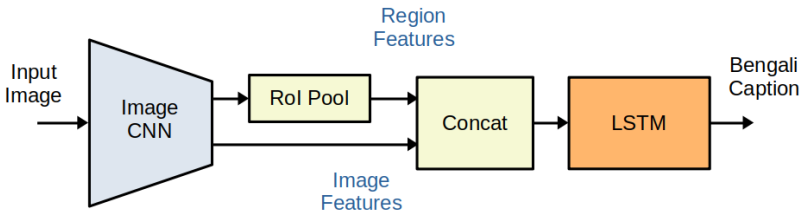


Figure 3. Architecture of the region-specific image caption generator.

The reference model uses features from the last convolutional layer of the vision CNN as the input for subsequent RNN. As our interest lies in obtaining a caption focused on the specific region, we need to consider the features for the region as well, in addition to the whole image features. We compute the scaling factor between the input image size and the size of the final convolutional layer of the vision CNN. Using this factor, we identify the coordinates of the region (bounding box) in this final convolutional layer. We obtain the features for the corresponding region through Region of Interest (RoI) pooling [1]. We generate the final feature vector by concatenation of features from the region and features from the entire image. In the present use case, we consider ResNet-50 as the backbone for the reference model.

In this approach, the encoder module is not trainable, it only extracts the image features however the LSTM decoder is trainable. We used LSTM decoder

using the image features for caption generation using greedy search approach [12]. We have used the cross-entropy loss during training the decoder [15].

D-Test BLEU	E-Test BLEU	C-Test BLEU
2.5	1.3	0.4

Table 4. Results of the region-specific image captioning on the BVG dataset.

Table 5 shows a sample output of text-only translation and Bengali captions generated.

5 Conclusion and Future Work

We have presented the first multimodal English-Bengali dataset suitable for multimodal research applications such as- *(a)* multimodal translation, *(b)* Bengali caption generation including e-commerce product catalog labelling, and *(c)* product development for visually impaired persons.

To exploit the BVG by the research community, we plan to include this dataset in the multimodal shared tasks for Bengali image captioning as well as the tasks related to English-Bengali multimodal machine translation. We also plan to extend the BVG multimodal dataset for visual question answering.

Our "Bengali Visual Genome" is available for research and non-commercial use under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License[8] at `http://hdl.handle.net/11234/1-3722`.

Acknowledgments

The author Ondřej Bojar would like to acknowledge the support of the grant 19-26934X (NEUREM3) of the Czech Science Foundation.

References

1. Girshick, R.: Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV). pp. 1440–1448 (2015). https://doi.org/10.1109/ICCV.2015.169
2. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Teh, Y.W., Titterington, M. (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (13–15 May 2010), `http://proceedings.mlr.press/v9/glorot10a.html`

[8] `https://creativecommons.org/licenses/by-nc-sa/4.0/`

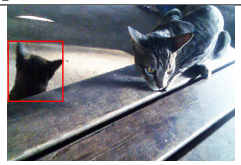
	Text-Only MT	**Region-specific image caption**
Image	—	
Source Text	fine thin red hair	—
System Output	পাতলা লাল চুল	একটি বিড়ালের মাথা
Gloss	Thin red hair	A cat's head
Reference Solution	সূক্ষ্ম পাতলা লাল চুল	একটি বিড়ালের মাথা পিছনে।
Gloss	fine thin red hair	Behind the head of a cat.
Image	—	
Source Text	a bunch of books on book stand	—
System Output	স্ট্যান্ডে একটি বইয়ের তাক	একটি কালো এবং সাদা ছবি
Gloss	A bookshelf on the stand	A black and white picture
Reference Solution	বইয়ের স্ট্যান্ডে একগুচ্ছ বই	কালো এবং সাদা বর্গ চিহ্ন।
Gloss	A bunch of books on the book stand	Black and white square marks.

Table 5. The sample outputs of text-only translation and region-specific image captioning for the BVG dataset.

3. Kamruzzaman, T.: Dataset for image captioning system (in bangla) (2021). https://doi.org/10.21227/4tsj-yn92, https://dx.doi.org/10.21227/4tsj-yn92
4. Khan, M.F., Sadiq-Ur-Rahman, S., Islam, M.S.: Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. In: Proceedings of International Joint Conference on Advances in Computational Intelligence. pp. 217–229. Springer (2021)
5. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization (2014), http://arxiv.org/abs/1412.6980, cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
6. Kudo, T., Richardson, J.: SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 66–71. Association for Computational Linguistics, Brussels, Belgium (Nov 2018). https://doi.org/10.18653/v1/D18-2012, https://www.aclweb.org/anthology/D18-2012

7. Nakazawa, T., Doi, N., Higashiyama, S., Ding, C., Dabre, R., Mino, H., Goto, I., Pa, W.P., Kunchukuttan, A., Oda, Y., Parida, S., Bojar, O., Kurohashi, S.: Overview of the 6th workshop on Asian translation. In: Proceedings of the 6th Workshop on Asian Translation. pp. 1–35. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5201, https://www.aclweb.org/anthology/D19-5201
8. Parida, S., Bojar, O., Dash, S.R.: Hindi visual genome: A dataset for multi-modal english to hindi machine translation. Computación y Sistemas **23**(4) (2019)
9. Parida, S., Motlicek, P., Dash, A.R., Dash, S.R., Mallick, D.K., Biswal, S.P., Pattnaik, P., Nayak, B.N., Bojar, O.: Odianlp's participation in WAT2020. In: Proceedings of the 7th Workshop on Asian Translation. pp. 103–108 (2020)
10. Popel, M., Tomkova, M., Tomek, J., Kaiser, Ł., Uszkoreit, J., Bojar, O., Žabokrtský, Z.: Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. Nature communications **11**(1), 1–15 (2020)
11. Rahman, M., Mohammed, N., Mansoor, N., Momen, S.: Chittron: An automatic bangla image captioning system. Procedia Computer Science **154**, 636–642 (2019)
12. Soh, M.: Learning cnn-lstm architectures for image caption generation
13. Sulubacak, U., Caglayan, O., Grönroos, S.A., Rouhe, A., Elliott, D., Specia, L., Tiedemann, J.: Multimodal machine translation through visuals and speech. Machine Translation **34**(2), 97–147 (2020)
14. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3156–3164 (2015). https://doi.org/10.1109/CVPR.2015.7298935
15. Yu, J., Li, J., Yu, Z., Huang, Q.: Multimodal transformer with multi-view visual representation for image captioning. IEEE transactions on circuits and systems for video technology **30**(12), 4467–4480 (2019)