BAN-Cap: A Multi-Purpose English-Bangla Image Descriptions Dataset

Mohammad Faiyaz Khan [1] S.M. Sadiq-Ur-Rahman Shifath [2] Md. Saiful Islam [3]

Shahjalal University of Science and Technology, Sylhet, Bangladesh [1, 2, 3] University of Alberta, Edmonton, Canada [3]
mfaiyazkhan@student.sust.edu [1] sm01@student.sust.edu [2] mdsaifu1@ualberta.ca [3]

Motivation

Multimodal research has been a well sought after topic in the recent past. Unfortunately, for want of a quality dataset, the case is not the same for Bangla, despite being the fifth most spoken language in terms of the number of speakers. We aim to address the issue by proposing a quality human-annotated image-sentence dataset by extending the well-used Flickr8k[4] dataset containing five English and Bangla captions for each image. We provide a baseline evaluation of our dataset on image captioning and investigate text augmentation as a possible direction of improvement. Experimental results show that models trained on our dataset perform better in predicting images in the wild than models trained on other existing datasets. Figure 1 contains a sample of the dataset.

Captions :

1. The children are playing in the water .
   শিশুরা পানিতে খেলছে ।
2. Two boys , one with a yellow and orange ball , play in some water in front of a field .
   দুটি ছেলে, একজন হাতে হলুদ-কমলা রঙের বল নিয়ে, মাঠের সামনে পানিতে খেলছে।
3. Two boys play in a puddle .
   দুটি বালক কর্দমাক্ত ক্ষুদ্র ডোবায় খেলছে ।
4. Two children play with a balloon in mud on a sunny day .
   দুটি ছেলে, একজন হাতে হলুদ-কমলা রঙের বল নিয়ে, মাঠের সামনে পানিতে খেলছে।
5. Two kids are running and playing in some water .
   দুটি বাচ্চা কিছু জলে দৌড়াচ্ছে এবং খেলছে।

Figure 1. A sample of the dataset containing image and English-Bangla caption pairs

The BAN-Cap Dataset

1. Data Collection and Processing

Initial Setup: Our goal was to minimize various human biases in the annotations throughout the data collection process. We adopted the following procedures:

- We divided the annotators into two groups. The first group consisted of twenty native Bangla speakers who studied in the linguistics department at various public universities in Bangladesh. The second group consisted of graduate students with expertise in the Bangla and the English language.
- The first group provided the annotations following a guideline provided by the expert group.
- The gender ratio of males and females in the two groups was 3:2.
- The ages of the members ranged from 18 to 30.
- The annotators represented different demographic regions from all over the country.

Human Annotation and Processing: We developed a website. The annotation page contained an image and an English caption. The annotators were asked to provide a Bangla caption primarily based on their understanding of the image and take help from the provided English caption if necessary. The guideline provided to the annotators by the expert group contained instructions like describing the images following the natural flow and native Bangla sentence structure, avoiding transliterated Bangla words as much as possible, using proper punctuation. An annotator provided only one caption for each image which ensures the variety and vibrancy of the data.

During the data collection process, the expert group repeatedly assessed the quality of the captions by manually checking a subset of the descriptions randomly.

2. Statistical Comparison

Table 1 shows corpus-level statistics and comparison among BAN-Cap and other existing datasets in Bangla. BAN-Cap has higher unique tokens compared to other existing datasets. It has a similar average sentence length compared to the BanglaLekhaImageCaptions [6] while having more than twice as many captions. Also, the recently proposed human-annotated data, Bornon [7], has a significantly lower average sentence length, which is critical for maintaining the details while describing an image. It is also noticeable that there are some structural variations between Bangla and the English captions. BAN-Cap Bangla descriptions have about 87% more unique tokens compared to English. On the other hand, the total number of tokens is about 27% higher in English than in Bangla. Also, an average English description is longer than a Bangla description.

Dataset	#Sentences	#Unique Tokens	#Total Tokens	Sent. Len. Mean	Sent. Len. Variance
Flickr8k (English)	40455	437421	8440	10.81	14.51
BAN-Cap (Bangla)	40455	344574	15846	8.51	10.99
BanglaLekhaImageCaptions [6]	18308	155249	5720	8.47	20.13
Bornon [7]	20500	110566	6228	5.34	4.38

Table 1. Statistics of the textual data of BAN-Cap along with existing Bangla image captioning data

Effectiveness

1. Machine Translated Vs Human Annotated Data

Our human annotated dataset has some edges over the machine translated dataset using tools like Google Translate:

- The automatic translators are not optimized yet for Bangla. Only a handful of machine-translated captions maintain coherence with the image's content while retaining the structural integrity of a Bangla sentence.
- Often the machine-translated Bangla captions contain a tremendous amount of misspelled words, erroneous use of punctuation and incomplete sentences, which do not conclude to a meaningful outcome.
- The machine-translated captions often fail to capture any cultural essence. Often the translation occurs in a word by word manner. Also they contain a large amount of transliterated English words which results in bad syntactic and semantic meaning.
- The human-annotated captions provide a wide variety compared to the machine-generated captions. In our case, the machine-translated data contains 14606 unique tokens compared to the 15846 tokens in the human-annotated data. However, when we filter out the tokens with at least a frequency of 3, the number of unique tokens in the machine-translated dataset is 4631 compared to 5636 in the human-annotated Bangla dataset.

2. Use-Cases

The BAN-Cap dataset can be readily used for a number of tasks like Multimodal Machine Translation, Visual Question Answering, Text to Image Generation along with Image Captioning and Machine Translation. Figure 2 contains an illustration of it's usefulness in various domains.

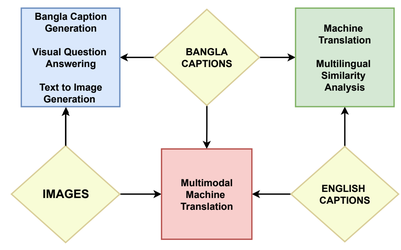

Figure 2. A combination of different components of the dataset can be used for different tasks.

Baseline and Results

1. Baseline Models

We present the baseline of our dataset on the existing Bangla image captioning models like CNN-Merge[3], Visual-Attention [2], Transformer[7]. Additionally, We adopted the highly effective adaptive attention mechanism and trained a similar model existing in the English language [5]. Furthermore, we experimented with different text augmentation techniques like Synonymous Word Replacement (SWR), Back Translation (BT), Contextualized Word Replacement (CWR) and found that CWR, combined with Adaptive Attention model improves the existing results. Also we replicated the model proposed in [8] for machine translation.

2. Results

Quantitative Evaluation: Table 2 contains the evaluation scores of different image captioning and machine translation models on the test set of the main dataset using standard evaluation scores. In Table 2, the CNN-Merge [3] model achieved lowest scores in all evaluation metrics. The Visual-Attention [2] model improves the performance by utilizing the extraction of only important features from an image during a caption prediction. However, despite having a relatively simple architecture, the Transformer [7] model outperforms the Visual-Attention and the CNN-Merge models by utilizing multi-head attention and better context awareness ability of the transformer. The adaptive attention-based model outperforms all the other models in most evaluation metrics by applying visual sentinel to guide the model using the attention mechanism more effectively. Finally, we see a performance boost for every model when applying text augmentation. After experimenting with all the combinations of text augmentation techniques previously described, we find that the Adaptive-Attention model with Contextualized Word Replacement gives us the best evaluation scores. We also present evaluation of the Encoder-Decoder model for machine translation to demonstrate the datasets multipurpose nature.

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE$_L$	SPICE
CNN-Merge [3]	0.565	0.355	0.221	0.131	0.178	0.281	0.290	0.042
Visual-Attention [2]	0.587	0.368	0.254	0.144	0.195	0.293	0.288	0.033
Transformer [7]	0.623	0.396	0.251	0.152	0.198	0.300	0.290	0.038
Adaptive-Attention [5]	0.702	0.466	0.307	0.194	0.297	0.297	0.344	0.055
Adaptive-Attention with CWR	**0.738**	**0.495**	**0.329**	**0.208**	**0.308**	**0.316**	**0.368**	**0.059**
Enc-Dec (MT) (Ban-To-Eng)	0.610	0.375	0.229	0.134	–	0.132	–	–
Enc-Dec (MT) (Eng-To-Ban)	0.656	0.419	0.264	0.158	–	0.306	–	–

Table 2. Evaluation of different image captioning and machine translation models on the BAN-Cap dataset.

Qualitative Evaluation: The metrics mentioned above often fail to summarise how the predictions appear to a human in real-life use-cases. To get a qualitative idea of how a model predicts unseen images outside the datasets it has been trained on, we collected some sample images from an online copyright-free source. We trained the best performing Adaptive-Attention model on our dataset, the Google translated Bangla dataset, the BanglaLekhaImageCaptions dataset, and the Bornon dataset and obtained four different versions of the model. We generated four predictions of each collected image by each of the four versions of the model. Then we asked four experts to assign a score out of five by evaluating the quality of a prediction where a higher score means a better quality caption. The model achieved 3.5/5 on average when trained on our dataset, 2.5/5 on the BanglaLekhaImageCaptions dataset, 2.5/5 on the Bornon, and 1.0/5 on the machine-translated dataset. Figure 3 contains samples of the predictions made by the model when trained on different datasets, along with the corresponding human evaluation score.

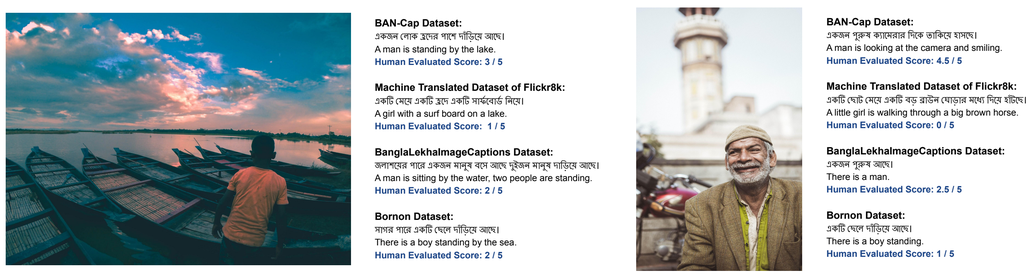

Figure 3. Example of the model's prediction on unseen images while trained on different datasets along with corresponding human evaluation scores. (English translations are provided for the understanding of the non native Bangla speakers)

Conclusion

We present BAN-Cap, a multilingual image descriptions dataset containing English-Bangla caption pairs. We test and evaluate it on various models of image captioning and machine translation to validate its multipurpose nature. Our future works will include investigating the impact of text augmentations on other existing datasets to validate its generalizability and apply this dataset in different research areas.

References

[1] M. Ahmed and D.R. Kim. pcr: An R package for quality assessment, analysis and testing of qPCR data. *PeerJ*, 2018(3), 2018.

[2] Amit Saha Ami, Mayeesha Humaira, Md Abidur Rahman Khan Jim, Shimul Paul, and Faisal Muhammad Shah. Bengali image captioning with visual attention. In *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pages 1–5, 2020.

[3] Mohammad Faiyaz Khan, S. M. Sadiq-Ur-Rahman, and Md. Saiful Islam. Improved bengali image captioning via deep convolutional neural network based encoder-decoder model. In Mohammad Shorif Uddin and Jagdish Chand Bansal, editors, *Proceedings of International Joint Conference on Advances in Computational Intelligence*, pages 217–229, Singapore, 2021. Springer Singapore.

[4] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013.

[5] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[6] Nafees; Kamal Mansoor, Abrar Hasin; Mohammed, Nabeel; Momen, Sifat; Rahman, and Md Matiur. Banglalekhaimagecaptions, mendeley data, 2019.

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

[8] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.