Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6855–6865
Marseille, 20-25 June 2022

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

Mohammad Faiyaz Khan[⊤*], **S.M. Sadiq-Ur-Rahman Shifath**[⊤*], **Md Saiful Islam**[*†]
[*]Shahjalal University of Science and Technology, Sylhet, Bangladesh,
{mfaiyazkhan, sm01}@student.sust.edu
[†]University of Alberta, Edmonton, Canada
mdsaifu1@ualberta.ca

Abstract

As computers have become efficient at understanding visual information and transforming it into a written representation, research interest in tasks like automatic image captioning has seen a significant leap over the last few years. While most of the research attention is given to the English language in a monolingual setting, resource-constrained languages like Bangla remain out of focus, predominantly due to a lack of standard datasets. Addressing this issue, we present a new dataset *BAN-Cap* following the widely used Flickr8k dataset, where we collect Bangla captions of the images provided by qualified annotators. Our dataset represents a wider variety of image caption styles annotated by trained people from different backgrounds. We present a quantitative and qualitative analysis of the dataset and the baseline evaluation of the recent models in Bangla image captioning. We investigate the effect of text augmentation and demonstrate that an adaptive attention-based model combined with text augmentation using Contextualized Word Replacement (CWR) outperforms all state-of-the-art models for Bangla image captioning. We also present this dataset's multipurpose nature, especially on machine translation for Bangla-English and English-Bangla. This dataset and all the models will be useful for further research.

Keywords: Image Captioning, Natural Language Processing, Multilingual, Multimodal, Machine Translation

1. Introduction

Image captioning, a variety of multimodal machine learning, is a research area that integrates and models data from different modalities. It generates humanoid descriptions of images by identifying and analysing their contents. It is more involved than other computer vision or natural language processing tasks because it requires object recognition, the inference of their relationships, and the generation of a meaningful and relevant interpretation using a sequence of words. It leverages the heterogeneity and the correlation of data of different modalities to achieve some original goals of artificial intelligence. It has a wide range of applications. For example, an image captioning system can be used in human-computer interaction, develop a hearing-aid system for visually impaired people, perform concept-based image indexing for information retrieval, automate self-driving cars, and many more.

Because of the availability of large-scale image-sentence pair datasets like Flickr8k (Hodosh et al., 2013), Flickr30k (Young et al., 2014), and MS COCO (Lin et al., 2014), research interest in image captioning and similar domains has seen an enormous rise in the last decade. The recent advancements in deep learning have made image captioning a sought-after topic for the research community. Deep learning-based models like Vinyals et al. (2015) introduced significant improvement over the traditional machine learning-based models by following the encoder-decoder architecture, which leverages the widespread sequence generation capability of the Recurrent Neural Network (RNN) conditioned by the image. Later attention-based models like Xu et al. (2015) contained a mechanism that attempted to filter only the necessary features from the image while generating captions. Also, multilingual image description datasets are widely available for other languages. The IAPR TC-12 dataset (Grubinger et al., 2006) has a collection of 20,000 images and a text caption corresponding to each image in up to three different languages (English, German and Spanish). Funaki and Nakayama (2015) curated Japanese translations for the English sentences of the Pascal dataset, which contains 1000 images. Elliott et al. (2016) proposed an image description dataset with English-German sentence pairs. It is an extension of the Flickr30k (Young et al., 2014) dataset that contains 31,014 German translations over a subset of English descriptions and 155,070 German descriptions, which are crowd-sourced independently of the original English descriptions.

Despite being the fifth most spoken language in terms

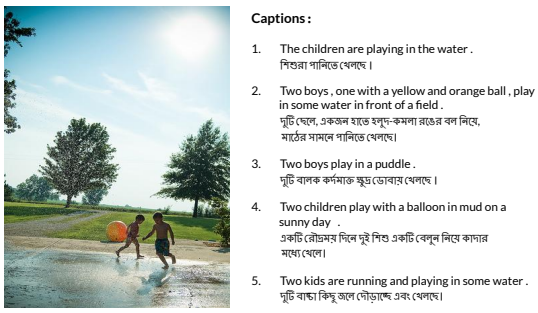


Figure 1: A sample of the dataset containing image and English-Bangla caption pairs

[⊤]These authors contributed equally to this work

of the number of speakers[1], Bangla still lacks the availability of a sizable and high-quality image-sentence dataset. So, research in the related fields in the Bangla language is still in its infancy. Some attempts have been made to create image captioning datasets in Bangla. Rahman et al. (2019) proposed an encoder-decoder model along with a Bangla image captioning dataset (Mansoor et al., 2019). The dataset contains 9,154 images and two captions per image. It lacks adequate captions per image and contains a number of samples with generic captions which do not provide necessary details. Deb et al. (2019) presented a comparative analysis of the CNN-LSTM based methods on a subsampled, machine-translated version of the Flickr8k (Hodosh et al., 2013) dataset. The resulted dataset lacks variety, quality, and usability. Recently, Shah et al. (2021) proposed a transformer-based model along with 4000 images and five human-annotated captions per image. Nevertheless, it has relatively shorter descriptions of an image which result in less detail and lacks usability in domains like multimodal machine translation.

We introduce a sizable dataset of images paired with sentences in English and Bangla to mitigate these problems. It is an extension of the Flicrkr8k (Hodosh et al., 2013) dataset with 8091 images and 40455 English-Bangla caption pairs. The annotations are provided by native Bangla speakers who have expertise in the English and the Bangla language. The dataset is post-processed and evaluated by an expert team for quality maintenance. Figure 1 shows a sample of this dataset with an image and its corresponding English-Bangla caption pairs. Additionally, we demonstrate this datasets' usability and efficacy by training and evaluating multiple deep learning-based methods for image description generation and machine translation. Experimental results show our dataset's variety and diversity and validate its multipurpose nature. To our knowledge, this is the first human-annotated image description dataset containing English-Bangla caption pairs. The dataset and code are available at github[2].

Our contributions are the following:

- We present BAN-Cap, an extension of the Flickr8k (Hodosh et al., 2013) image descriptions dataset, by accumulating Bangla captions with competent annotators under various quality control measures to ensure its quality and usability.
- We perform a statistical analysis of the data and present a qualitative and quantitative comparison between our human-annotated dataset and a machine-translated dataset curated using Google Translate.
- We set up the baseline of this dataset by training and evaluating all the recent notable models of Bangla image captioning. Besides, we present a baseline of neural machine translation to demonstrate this dataset's multipurpose nature.
- We present an adaptive attention based model with contextualized word replacement that outperforms current state-of-the-art models in Bangla image captioning. Additionally, we experiment with other text augmentation techniques as a possible direction of improvement the overall models' performances in Bangla image captioning.
- We compare the model's prediction on unseen data while trained on our dataset with the models trained on other existing Bangla caption datasets. We present qualitative human evaluation scores of the predictions that show the model trained on our dataset generates quality captions and has better generalization capability.

2. The BAN-Cap Dataset

2.1. Data Collection

The Flickr8k dataset contains images collected from a community-based online photo hosting website (Hodosh et al., 2013). We used the Flickr8k data as it contains evenly distributed images from various domains. Each image has five descriptions in English, which are collected through crowd-sourcing platforms. The BAN-Cap dataset contains Bangla captions of the Flickr8k images provided by human annotators. It has 8091 images and 40,455 English-Bangla description pairs.

2.1.1. 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 performed the annotation task. The second group provided an overall guideline for the first group and performed error correction and quality evaluation of the annotations.
- The gender ratio of males and females in the two groups was 3:2.
- The ages of the annotators and the expert group members ranged from 18 to 30.
- We ensured that the annotators represented different demographic regions from all over the country.

[1] https://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] https://github.com/FaiyazKhan11/BAN-Cap

Dataset	#Sentences	#Total Tokens	#Unique Tokens	Sentence Length Mean	Sentence Length Variance
Flickr8k (English)	40455	437421	8440	10.81	14.51
BAN-Cap (Bangla)	40455	344574	15846	8.51	10.99
BanglaLekhaImageCaptions (Mansoor et al., 2019)	18308	155249	5720	8.47	20.13
Bornon (Shah et al., 2021)	20500	110566	6228	5.34	4.38

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

All the above measures were taken to build a group of annotators with as much variety as possible. The members were paid a standard amount depending on their types of work.

2.1.2. Human Annotation

We developed a website for collecting the annotations. The annotators were required to log in using their names and registration numbers. 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.

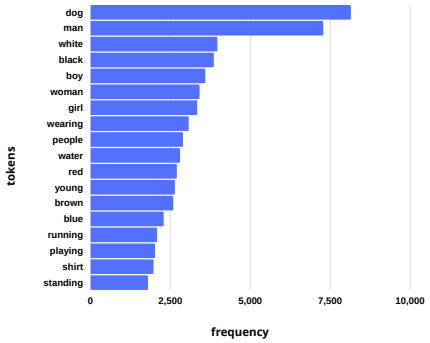
2.1.3. Post Processing

During the data collection process, the expert group repeatedly assessed the quality of the captions by manually checking a subset of the descriptions randomly. After the annotation phase, they performed manual error correction on the data.

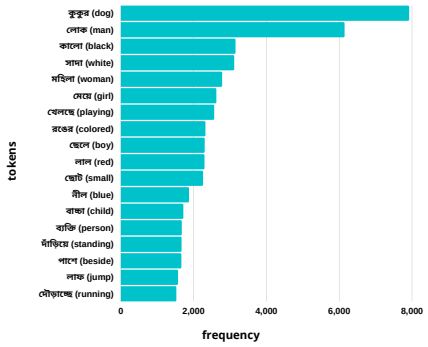
2.2. Statistical Analysis

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 (Mansoor et al., 2019) while having more than twice as many captions. Also, the recently proposed human-annotated data, Bornon (Shah et al., 2021), 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.

Figure 2 illustrates the top twenty most frequent tokens in English and Bangla caption datasets. Both languages have several stop words with a high frequency of occurrence. So we discarded the stop-words in this analysis for the sake of comparison. Both the datasets have a similar token frequency.



(a) Most frequent English tokens



(b) Most frequent Bangla tokens

Figure 2: Most frequent English and Bangla tokens (Descending order)

2.3. Machine Translated Vs Human Annotated Data

Automated machine translation tools such as Google Translate have come a long way. However, when it comes to low resource languages like Bangla, they have a considerable amount of lacking. When compared side by side, we observe the following key factors that contribute to curating a human-annotated dataset rather than using a machine-translated one:

- 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. It is evident in figure 3. We provided English translations for the Bangla captions for the sake of understanding of the non Bangla speaking people.

English caption:
White dog with brown ears standing near water with head turned to one side.

Human Annotated caption:
বাদামি কানওয়ালা সাদা কুকুরটি মাথা একদিকে বাঁকিয়ে পানির কাছে দাড়িয়ে আছে।
The white dog with brown ears is standing by the water with its head bent to one side.

Machine Translated caption:
মাথার সাথে পানির কাছে খয়েরি কানের সাথে সাদা কুকুরটি একদিকে ঘুরে।
The white dog with brown ears turned one side near the water with its head.

English caption:
Two draft horses pull a cart through the snow.

Human Annotated caption:
দুটি বাধাঁ ঘোড়া বরফের মাঝ দিয়ে গাড়ি টানছে।
Two draft horses are pulling a cart through the snow.

Machine Translated caption:
দুটি খসড়া ঘোড়া বরফের মাধ্যমে একটি কার্ট টান।
Two draft horses a cart pull through the ice.

English caption:
There is a girl with pigtails sitting in front of a rainbow painting.

Human Annotated caption:
চুলে বেণী করা ছোট মেয়েটি একটি রংধনু চিত্রের সামনে বসে আছে।
The little girl with the braid in her hair is sitting in front of a rainbow figure.

Machine Translated caption:
পিগটেলসের সাথে একটি মেয়ে আছে যা রামধনু পেইন্টিংয়ের সামনে বসে আছে।
There is a girl with pigtails which sits in front of the rainbow painting.

Figure 3: Examples where the machine translated captions captured the context of the image but failed to maintain syntactical and structural integrity of Bangla sentence (English translations are provided for the understanding of the non Bangla speakers).

English caption:
A dog running in shrubbery along a stream.

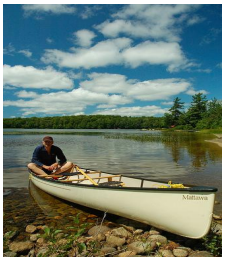
Human Annotated caption:
নদীর পাশ দিয়ে একটি কুকুর দৌড়াচ্ছে।
A dog is running alongside the river.

Machine Translated caption:
একটি কুকুর ঝর্ণা ঝর্ণায় ঝর্ণায় চলছে।
A dog is running along the waterfall in waterfall in waterfall.

English caption:
a biker enjoys a coffee.

Human Annotated caption:
একজন সাইকেল চালক তার কফি উপভোগ করছে।
A cyclist enjoys his coffee.

Machine Translated caption:
একজন বাইকার একটি কফি পান।
A biker gets a coffee.



English caption:
Man sitting in a beached canoe by a lake.

Human Annotated caption:
একজন লোক হ্রদের ধারে ডিঙি নৌকায় বসে আছে।
A man is sitting in a canoe by the lake.

Machine Translated caption:
হ্রদের ধারে একটি সৈকত নোনে বসে আছেন।
A beach is sitting on a none by the lake.

Figure 4: Examples where the machine translated captions very poorly described the image or failed to generate any meaningful sentence at all (English translations are provided for the understanding of the non Bangla speakers).

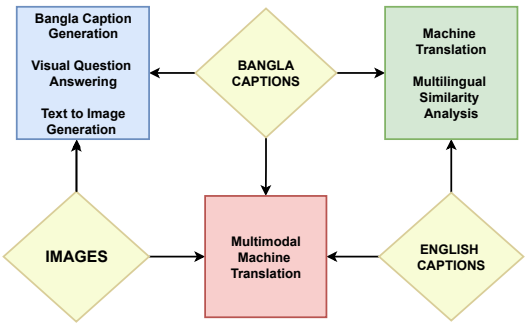


Figure 5: A combination of different components of the dataset can be used for different tasks.

- 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. These are evident in figure 4.
- The machine-translated captions often fail to capture any cultural essence. From all the examples of figure 3 and 4, it is noticeable that the system has mostly translated the source language into the target language word-by-word. Also, they contain a large amount of transliterated English words in the Bangla captions. So the semantic and the syntactic meaning is lost.
- 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.

For the above reasons, systems trained with machine-translated data generally output captions which contains artificiality and lack real world usefulness.

2.4. Usability and Multipurpose Nature

The BAN-Cap dataset can be readily used in various domains along with image captioning. Figure 5 illustrates how different components of this dataset can be used for a variety of tasks. Following are some short descriptions of this dataset's usefulness in other domains:

Multimodal Machine Translation: Multimodal Machine translation (Elliott, 2018; Yao and Wan, 2020; Barrault et al., 2018; Specia et al., 2016; Caglayan et al., 2019) involves gaining information from multiple modalities. Usually, it is assumed that additional modality features will provide an alternative view of the input data. Unlike machine translation, multimodal machine translation is still a field to explore in the Bangla language. The image-text machine translation can perform better compared to the text-to-text machine translation. It can also be used for cross-modal and cross-lingual information retrieval.

Machine Translation: Machine translation (Abujar et al., 2021; Bal et al., 2021; Das and Singh, 2021) is already a vastly researched topic in Bangla natural language processing domain. Although there are some well-curated datasets specifically for the English-Bangla machine translation (Hasan et al., 2020), the BAN-Cap dataset has its niche with five descriptions containing a context, which eventually carries a diversified view of the same content.

Bangla Visual Question Answering: Given an image and one or more textual questions, a Visual Question Answering (VQA) (Sikarwar and Kreiman, 2022; Li et al., 2022; Wang et al., 2021) system produces relevant answers by analyzing the questions and the image. It is now a popular research topic across many languages. It can be used to gain information about visual content as well as in tasks like image retrieval. Most of the datasets used in this domain are derived from an image-text dataset like ours. We expect our dataset to play a vital role to kick-start research in the Bangla visual question answering domain.

Text to Image Generation: Another field of research this dataset can be used is generating an image from text (Cheng et al., 2021; Siddharth and Aarthi, 2021; Zhang et al., 2021). The generated image from a text can serve as a universal language for many applications such as education, language learning, literacy development, summarization of news articles, and data visualization.

3. Baselining

3.1. Existing Models

3.1.1. Image Captioning

To set up the baseline of our dataset, we present the evaluation scores of all the recent state-of-the-art models of Bangla image captioning. We trained the following models on our dataset:

CNN-Merge: Faiyaz Khan et al. (2021) proposed this model following the merge architecture of Tanti et al. (2017).

Visual-Attention: Proposed by Ami et al. (2020), The visual attention model is very similar to the one introduced in Xu et al. (2015).

Transformer: This model, proposed by Shah et al. (2021), is also based on the encoder-decoder architecture. It utilizes the multi-head attention of the transformer during decoding.

3.1.2. Machine Translation

Encoder-Decoder: This model is the replication of the one proposed in Sutskever et al. (2014)

3.2. A Better Approach

In a quest to improve the existing baselines, we adopted the highly effective adaptive attention mechanism and

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE$_L$	SPICE
CNN-Merge (Faiyaz Khan et al., 2021)	0.565	0.355	0.221	0.131	0.178	0.281	0.290	0.042
Visual-Attention (Ami et al., 2020)	0.587	0.368	0.254	0.144	0.195	0.293	0.288	0.033
Transformer (Shah et al., 2021)	0.623	0.396	0.251	0.152	0.198	0.300	0.290	0.038
Adaptive-Attention (Lu et al., 2017)	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**

Table 2: Evaluation of different image captioning models on the BAN-Cap dataset.

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE$_L$	SPICE
CNN-Merge (Faiyaz Khan et al., 2021)	0.458	0.273	0.163	0.094	0.079	0.245	0.204	0.018
Visual-Attention (Ami et al., 2020)	0.505	0.303	0.184	0.107	0.046	0.273	0.195	0.004
Transformer (Shah et al., 2021)	0.458	0.269	0.160	0.091	0.065	0.280	0.179	0.008
Adaptive-Attention (Lu et al., 2017)	0.515	0.310	0.188	0.109	0.055	0.267	0.198	0.005

Table 3: Evaluation of different image captioning models on the machine translated Flickr8k Bangla dataset.

trained a similar model existing in the English language (Lu et al., 2017).

3.2.1. Adaptive - Attention Model

This model uses "Sentinel Attention" as an addition to spatial attention. The spatial attention is adaptive in the sense that it is dependent on the current hidden state rather than previous hidden states in Xu et al. (2015).

The sentinel attention requires a sentinel gate, which determines what kind of information the model will focus on - visual or textual. A context vector is obtained by combining sentinel and spatial visual information.

3.2.2. Text Augmentation

Text augmentation has been a handy technique in many low resource language-based tasks. To investigate if it has an impact on improving the existing models' performances, we experimented with the following three text augmentation techniques in the context of Bangla image captioning:

1. Synonymous Word Replacement (SWR): Each word of the captions was replaced with a synonymous word using bnltk library[3] and based on the semantic similarity scores with the human-annotated captions, the top 3 captions were selected. (Atliha and Šešok, 2020).

2. Back Translation (BT): Each caption was translated to English and then translated back to Bangla using Google Translate[4]. The back-translated captions were added with the original captions.

3. Contextualized Word Replacement (CWR): Each word of the captions was replaced with a contextually similar word predicted by Bangla-BERT[5] and based on the semantic similarity scores with the human-annotated captions, the top three samples were selected (Atliha and Šešok, 2020).

[3] https://pypi.org/project/bnltk/

[4] https://translate.google.com/

[5] https://huggingface.co/sagorsarker/bangla-bert-base

4. Experimentation Details

4.1. Data Preprocessing

For training and evaluation, we split the dataset using the standard train, test and validation split (Karpathy and Fei-Fei, 2017) for the Flickr8k dataset. We trained on 6000 images, validated on 1000 images, and tested on 1000 images and their corresponding captions. We used only the captions associated with the images in the training, validation, and test sets for machine translation. We applied fundamental preprocessing techniques. We removed punctuations from the captions during tokenization. Each caption begins and ends with a unique starting and ending token. For consistency, captions are either padded or truncated to a fixed size. We set a threshold of five for image captioning and replace all tokens with "unk" that occur less frequently than five times.

4.2. Training Process

We trained the models on the training data and calculated validation loss on the validation data after each training epoch. The model with the lowest validation loss is saved and later used to predict the unseen test data.

5. Result Analysis

We evaluated the models' performances using existing evaluation metrics such as BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002), ROUGE$_L$(Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004), METEOR (Metric for Evaluation of Translation with Explicit Ordering) (Banerjee and Lavie, 2005), CIDEr (Consensus-based Image Description Evaluation) (Vedantam et al., 2015), and SPICE (Semantic Propositional Image Caption Evaluation) (Anderson et al., 2016).

BLEU is calculated by comparing the reference and predicted sentences' n-gram geometric means. However, because the same sentence can be represented in various ways with the same sense, the scores are not

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE$_L$	SPICE
CNN-Merge (Faiyaz Khan et al., 2021)	0.468	0.279	0.167	0.096	0.059	0.256	0.198	0.013
Visual-Attention (Ami et al., 2020)	0.464	0.277	0.166	0.095	0.036	0.224	0.159	0.011
Transformer (Shah et al., 2021)	0.466	0.277	0.166	0.095	0.035	0.255	0.142	0.013
Adaptive-Attention (Lu et al., 2017)	0.480	0.290	0.176	0.102	0.042	0.226	0.171	0.014

Table 4: Evaluation of different image captioning models on the BAN-Cap dataset (two captions per image).

Model Name	BLEU-1	BLEU-2	BLEU-3	BLEU-4	CIDEr	METEOR	ROUGE$_L$	SPICE
CNN-Merge (Faiyaz Khan et al., 2021)	0.483	0.281	0.166	0.094	0.010	0.143	0.192	0.002
Visual-Attention (Ami et al., 2020)	0.484	0.291	0.174	0.100	0.034	0.237	0.167	0.018
Transformer (Shah et al., 2021)	0.487	0.291	0.174	0.100	0.034	0.232	0.179	0.016
Adaptive-Attention (Lu et al., 2017)	0.489	0.293	0.175	0.101	0.038	0.226	0.173	0.018

Table 5: Evaluation of different image captioning models on the BAN-Cap dataset (three captions per image).

always accurate. The predicted captions were evaluated using 1,2,3, and 4-gram BLEU scores. ROUGE$_L$ is calculated by comparing the reference and predicted sentences' longest common subsequences. The sentences' longest common subsequence takes sentence-level structure similarity into account and recognizes the longest co-occurrence in n-grams. CIDEr is calculated using a Term Frequency Inverse Document Frequency (TF-IDF) weighting scheme for the n-gram of each sentence. METEOR is calculated by comparing the actual and predicted sentences word for word and then by calculating the precision and recall harmonic means. SPICE is calculated for sentence pairs based on F-scores on tuples from the scene graphs, semantic representations of the objects, properties, and connections in the captions. CIDER and SPICE are unique metrics for evaluating image captions' syntactic and semantic quality, while BLEU and METEOR are used to assess image captioning and machine translation.

Table 2 contains the evaluation scores of different image captioning models on the test set of the main dataset. In Table 2, the CNN-Merge (Faiyaz Khan et al., 2021) model achieved lowest scores in all evaluation metrics. The Visual-Attention (Ami et al., 2020) 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 (Shah et al., 2021) 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.

Table 3 contains the performance of the image captioning models while trained and evaluated on the machine translated Flickr8k Bangla dataset. All the models show a performance drop across all the metrics. A significant drop can be observed in CIDEr and SPICE, specialized evaluation metrics for image captioning.

We hypothesized that training an image captioning system with more captions per image will yield a better and more robust model that will be able to deliver more varied and detailed captions. To test our hypothesis, we trained the image captioning models with datasets containing two and three captions per image and report the result in Table 4 and 5 respectively. The experimental results indeed validate our hypothesis as we see a gradual improvement in the results from Table 4 to Table 5. Finally, in Table 2, we see the highest scores achieved by each model when trained with the dataset containing five captions per image.

Though the metrics mentioned above generally give a numeric estimation of how well a model performs, they 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 (uns, 2022). We trained the best performing Adaptive-Attention model on our dataset, the Google translated Bangla dataset, the BanglaLekhaImageCaptions dataset, and the Bornon dataset. Thus we obtain four different versions of the Adaptive-Attention model. We generated four predictions of each collected image by each of the four versions of the Adaptive-Attention 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 6 contains samples of the predictions made by the model when trained on different datasets, along with the corresponding human evaluation score.

Table 6 provides the experimental results on the Bangla

Model	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR
Bangla-To-English	0.610	0.375	0.229	0.134	0.132
English-To-Bangla	0.656	0.419	0.264	0.158	0.306

Table 6: Evaluation of different models of machine translation on BAN-Cap dataset.

BAN-Cap Dataset:
একজন লোক হ্রদের পাশে দাঁড়িয়ে আছে।
A man is standing by the lake.
Human Evaluated Score: 3 / 5

Machine Translated Dataset of Flickr8k:
একটি মেয়ে একটি হ্রদে একটি সার্ফবোর্ড নিয়ে।
A girl with a surf board on a lake.
Human Evaluated Score: 1 / 5

BanglaLekhaImageCaptions Dataset:
জলাশয়ের পারে একজন মানুষ বসে আছে দুইজন মানুষ দাড়িয়ে আছে।
A man is sitting by the water, two people are standing.
Human Evaluated Score: 2 / 5

Bornon Dataset:
সাগর পারে একটি ছেলে দাঁড়িয়ে আছে।
There is a boy standing by the sea.
Human Evaluated Score: 2 / 5

BAN-Cap Dataset:
একজন পুরুষ ক্যামেরার দিকে তাকিয়ে হাসছে।
A man is looking at the camera and smiling.
Human Evaluated Score: 4.5 / 5

Machine Translated Dataset of Flickr8k:
একটি ছোট মেয়ে একটি বড় ব্রাউন ঘোড়ার মধ্যে দিয়ে হাঁটছে।
A little girl is walking through a big brown horse.
Human Evaluated Score: 0 / 5

BanglaLekhaImageCaptions Dataset:
একজন পুরুষ আছে।
There is a man.
Human Evaluated Score: 2.5 / 5

Bornon Dataset:
একটি ছেলে দাঁড়িয়ে আছে।
There is a boy standing.
Human Evaluated Score: 1 / 5

Figure 6: 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)

to English and English to Bangla machine translation task of the encoder-decoder model. Our primary purpose is not to achieve state-of-the-art results but to demonstrate this dataset's multipurpose nature.

6. Conclusion

We present BAN-Cap, a multilingual image descriptions dataset containing English-Bangla caption pairs. Expert annotations under intense supervision make it a gold standard dataset. To validate this dataset's multipurpose nature, we test and evaluate it on various models of image captioning and machine translation. We also experiment with text augmentation to add variety to the human-annotated captions. 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. We expect the proposed dataset will be helpful in the multimodal and multilingual research domain and hope it will be beneficial to the research community for a variety of other purposes that we cannot predict.

7. Acknowledgements

We want to thank the annotators and evaluators who helped us in data collection and during the human evaluation process. We would also like to thank Natural Language Processing Group, Department of CSE, SUST for their valuable comments on our work.

8. Bibliographical References

Abujar, S., Masum, A. K. M., Bhattacharya, A., Dutta, S., and Hossain, S. A. (2021). English to ben-

gali neural machine translation using global attention mechanism. In *Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2020, Volume 3*, pages 359–369. Springer Singapore.

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

Anderson, P., Fernando, B., Johnson, M., and Gould, S. (2016). Spice: Semantic propositional image caption evaluation. In *ECCV*.

Atliha, V. and Šešok, D. (2020). Text augmentation using bert for image captioning. *Applied Sciences*, 10(17).

Bal, S., Mahanta, S., and Mandal, L. (2021). Bilingual machine translation: Bengali to english. In *Proceedings of International Conference on Computational Intelligence, Data Science and Cloud Computing: IEM-ICDC 2020*, volume 62, page 393. Springer Nature.

Banerjee, S. and Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Barrault, L., Bougares, F., Specia, L., Lala, C., Elliott, D., and Frank, S. (2018). Findings of the third shared task on multimodal machine translation. In *THIRD CONFERENCE ON MACHINE TRANSLATION (WMT18)*, volume 2, pages 308–327.

Caglayan, O., Madhyastha, P., Specia, L., and Barrault, L. (2019). Probing the need for visual context in multimodal machine translation. *arXiv preprint arXiv:1903.08678*.

Cheng, J., Wu, F., Tian, Y., Wang, L., and Tao, D. (2021). Rifegan2: Rich feature generation for text-to-image synthesis from constrained prior knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*.

Das, A. and Singh, T. D. (2021). Development of english-to-bengali neural machine translation systems. In *Proceedings of the International Conference on Computing and Communication Systems: I3CS 2020, NEHU, Shillong, India*, volume 170, page 55. Springer.

Deb, T., Ali, M., Bhowmik, S., Firoze, A., Ahmed, S. S., Tahmeed, M. A., Rahman, N. S. M. R., and Rahman, R. (2019). Oboyob: A sequential-semantic bengali image captioning engine. *J. Intell. Fuzzy Syst.*, 37:7427–7439.

Elliott, D., Frank, S., Sima'an, K., and Specia, L. (2016). Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany, August. Association for Computational Linguistics.

Elliott, D. (2018). Adversarial evaluation of multimodal machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2974–2978.

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

Funaki, R. and Nakayama, H. (2015). Image-mediated learning for zero-shot cross-lingual document retrieval. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 585–590, Lisbon, Portugal, September. Association for Computational Linguistics.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The iapr tc-12 benchmark – a new evaluation resource for visual information systems.

Hasan, T., Bhattacharjee, A., Samin, K., Hasan, M., Basak, M., Rahman, M. S., and Shahriyar, R. (2020). Not low-resource anymore: Aligner ensembling, batch filtering, and new datasets for Bengali-English machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2612–2623, Online, November. Association for Computational Linguistics.

He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep residual learning for image recognition.

Hodosh, M., Young, P., and Hockenmaier, J. (2013). Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899.

Karpathy, A. and Fei-Fei, L. (2017). Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.

Li, Y., Fan, J., Pan, Y., Yao, T., Lin, W., and Mei, T. (2022). Uni-eden: Universal encoder-decoder network by multi-granular vision-language pre-training. *arXiv preprint arXiv:2201.04026*.

Lin, T.-Y., Maire, M., Belongie, S. J., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *ECCV*.

Lin, C.-Y. (2004). ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July. Association for Computational Linguistics.

Lu, J., Xiong, C., Parikh, D., and Socher, R. (2017). 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.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July. Association for Computational Linguistics.

Rahman, M., Mohammed, N., Mansoor, N., and Momen, S. (2019). Chittron: An automatic bangla image captioning system. *Procedia Computer Science*, 154:636–642. Proceedings of the 9th International Conference of Information and Communication Technology [ICICT-2019] Nanning, Guangxi, China January 11-13, 2019.

Shah, F. M., Humaira, M., Jim, M. A. R. K., Ami, A. S., and Paul, S. (2021). Bornon: Bengali image captioning with transformer-based deep learning approach. *CoRR*, abs/2109.05218.

Siddharth, M. and Aarthi, R. (2021). Text to image gans with roberta and fine-grained attention networks.

Sikarwar, A. and Kreiman, G. (2022). On the efficacy of co-attention transformer layers in visual question answering. *arXiv preprint arXiv:2201.03965*.

Specia, L., Frank, S., Sima'An, K., and Elliott, D. (2016). A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Z. Ghahramani, et al., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2015). Rethinking the inception architecture for computer vision.

Tanti, M., Gatt, A., and Camilleri, K. (2017). What is the role of recurrent neural networks (RNNs) in an image caption generator? In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 51–60, Santiago de Compostela, Spain, September. Association for Computational Linguistics.

(2022). Unsplash photos. `https://unsplash.com/photos/`. Accessed: 2022-01-17.

Vedantam, R., Zitnick, C. L., and Parikh, D. (2015). Cider: Consensus-based image description evaluation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Vinyals, O., Toshev, A., Bengio, S., and Erhan, D. (2015). Show and tell: A neural image caption generator. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164.

Wang, W., Bao, H., Dong, L., and Wei, F. (2021). Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *arXiv preprint arXiv:2111.02358*.

Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., and Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In Francis Bach et al., editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2048–2057, Lille, France, 07–09 Jul. PMLR.

Yao, S. and Wan, X. (2020). Multimodal transformer for multimodal machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4346–4350.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Zhang, H., Yang, S., and Zhu, H. (2021). Cjetig: Zero-shot cross-lingual text-to-image generation by corpora-based joint encoding. *Knowledge-Based Systems*, page 108006.

9. Language Resource References

Mansoor, Nafees; Kamal and Abrar Hasin; Mohammed and Nabeel; Momen and Sifat; Rahman and Md Matiur. (2019). *BanglaLekhaImageCaptions, Mendeley Data*.

Appendix

Baseline Model Details

Image Captioning

The general procedure of an image captioning system is generating a sequence of words conditioned by the image and the previously generated words. A convolutional neural network is used to generate image features. The image captioning model tries to find the caption that maximizes the following log probability.

$$\log p(S/I) = \sum_{t=0}^{N} \log p(S_t|I, S_o, S_1, S_2.....S_{t-1}) \quad (1)$$

Here, I is the image, S is the caption, and S_t is the word in the caption at location t. The probability of the word S_t depends on the image I and all the previous words from S_0 to S_{t-1}. Brief explanations of the models we trained on our dataset for baselining are given in the following sections.

CNN-Merge

This model is based on the merged architecture proposed in Tanti et al. (2017) . The image features are encoded using a CNN, and the text features are encoded using another one-dimensional CNN. Here, one dimensional CNN is used instead of LSTM as it performs

Model Name	Batch Size	Learning Rate	Loss Function	Optimizer
CNN-Merge	64	0.01	Cross-Entropy	Adam
Soft-Attention	32	0.0004	Cross-Entropy	Adam
Adaptive-Attention	32	0.0004	Cross-Entropy	Adam
Transformer	128	0.01	Cross-Entropy	Adam
Encoder-Decoder	128	0.01	Negative Log Likelihood	Adam

Table 7: Selection of hyperparameters for different models.

better in capturing details of short sentences in Bangla language (Faiyaz Khan et al., 2021). The image and text feature extraction are independent processes. The two features are merged and passed to a decoder layer for caption generation.

Visual-Attention

Here, ResNet-101 (He et al., 2015) is used to generate feature maps from the image. The relevant location from the feature map is determined, and the location feature vector is passed to the LSTM at each time step. The probability distribution over all the locations is modelled based on the previously generated words. The probability of choosing a location i, denoted by $\alpha_{t,i}$ is proportional to the similarity between vector at that location l_i and the LSTM hidden vector h_t. The context vector z_t is calculated as

$$z_t = \sum_{i=0}^{n} \alpha_{t,i} l_i \tag{2}$$

Equation 1 changes to the following.

$$\log p(S/I) = \sum_{t=0}^{N} \log p(S_t | z_t, S_o, S_1, S_2 ..... S_{t-1}) \tag{3}$$

Instead of the whole image, the only relevant context of the image based on the previously generated words is used.

Transformer

A pre-trained Inception-V3 (Szegedy et al., 2015) extracts the image features. The final classification layer is discarded as only the image vectors are needed. Token and positional embeddings are generated from the captions and passed to a masked multi-head attention layer. This layer's output is passed along with the image features to another multi-head attention layer. The output is routed through a feed-forward layer and then a normalization layer. A softmax layer generates the final output probabilities.

Machine Translation

Encoder-Decoder

In an encoder-decoder architecture for sequence-to-sequence task (Sutskever et al., 2014), an encoder reads an input source sentence and generates a vector representation. We used GRU as an encoder.

$$h_t = f(s_t, h_{t-1}) \tag{4}$$

$$v = g(\{h_1, ..., h_{tx}\}) \tag{5}$$

Here, $s = (s_1, ..., s_{tx})$ is the input sentence, h_t is the hidden state at time t, and v is a vector generated from the hidden states. f and g are nonlinear functions. Another GRU is used as a decoder. The decoder predicts the probability of the next word based on the context vector v and all the previously predicted words.

$$p(y) = \prod_{t=1}^{T} p(y_t | \{y_1, ..., y_{t-1}\}, v) \tag{6}$$

where, $y = \{y_1, ..., y_t\}$ is the sequence of predicted words.

Training Hyperparameters

The selection of hyperparameters for training all the models is summarised in Table 7.