

Visual Question Generation in Bengali

Mahmud Hasan, Labiba Islam, Jannatul Ferdous Ruma
Tasmiah Tahsin Mayeesha and Rashedur M. Rahman

North South University

{mahmud.hasan03,labiba.islam,jannatul.ruma,
tasmiah.tahsin,rashedur.rahman}@northsouth.edu

Abstract

The task of Visual Question Generation (VQG) is to generate human-like questions relevant to the given image. As VQG is an emerging research field, existing works tend to focus only on resource-rich language such as English due to the availability of datasets. In this paper, we propose the first Bengali Visual Question Generation task and develop a novel transformer-based encoder-decoder architecture that generates questions in Bengali when given an image. We propose multiple variants of models - (i) image-only: baseline model of generating questions from images without additional information, (ii) image-category and image-answer-category: *guided* VQG where we condition the model to generate questions based on the answer and the category of expected question. These models are trained and evaluated on the translated VQAv2.0 dataset. Our quantitative and qualitative results establish the first state of the art models for VQG task in Bengali and demonstrate that our models are capable of generating grammatically correct and relevant questions. Our quantitative results show that our image-cat model achieves a BLUE-1 score of 33.12 and BLEU-3 score of 7.56 which is the highest of the other two variants. We also perform a human evaluation to assess the quality of the generation tasks. Human evaluation suggests that image-cat model is capable of generating goal-driven and attribute-specific questions and also stays relevant to the corresponding image.

1 Introduction

Visual Question Generation (VQG) is an emerging research field in both Computer Vision and Natural Language Processing. The task of VQG simply uses an image and other side information (e.g. answers or answer categories) as input and generates meaningful questions related to the image. Tasks like cross-modal Visual Question Answering (VQA) (Antol et al., 2015; Cadene et al.,


INPUT		OUTPUT
Image	Given answer category	Generated questions
	count	মাঠে কয়টি গরু আছে ? (How many cows are there in the field?)
	binary	এই ছবিতে কি পাহাড় আছে ? (Are there mountains in this picture?)

Figure 1: Examples of Bengali VQG Predictions with category of answers as additional information.

2019; Peng et al., 2019; Jiang et al., 2020; Guo et al., 2022), Video Captioning (VC) (Chen et al., 2019), Image Captioning (IC) (Vinyals et al., 2015; Karpathy and Fei-Fei, 2017; Xu et al., 2015), and Multimodal Machine Translation (Specia et al., 2016; Elliott et al., 2017; Barrault et al., 2018; Caglayan et al., 2019) are the recent advances in the AI community. While the majority of visuo-lingual tasks tend to focus on VQA, a few recent approaches have been proposed, focusing on the under-researched multi-modal task of VQG. VQG is a more creative and particularly challenging problem than VQA, because the generated questions need to be relevant, semantically coherent and comprehensible to the diverse contents of the given image.

Existing studies on Visual Question Generation (VQG) have been primarily focused on languages that have ample resources, such as English. While some VQA research have been conducted in low-resource languages like Hindi (Gupta et al., 2020), Bengali (Islam et al., 2022), Japanese (Shimizu et al., 2018), and Chinese (Gao et al., 2015), limitations have been identified specifically in the context of Bengali language. While Bengali language has some recent work on reading comprehension based question answering (Mayeesha et al., 2021; Aurpa et al., 2022) and visual question answering (Islam et al., 2022; Rafi et al., 2022), there has been no research conducted for VQG task specifically in Bengali language.

To obtain meaningful questions, some VQG

methods have either augmented the input including additional information such as answer categories, objects in image and expected answers (Pan et al., 2019; Krishna et al., 2019; Vedd et al., 2022). Pan et al. (2019) used ground truth answer with the image as an input, underscoring it to be an effective approach to produce non-generic questions. Krishna et al. (2019) stated that knowing the answers beforehand simply defeats the purpose of generating realistic questions since the main purpose of generating a question is to attain an answer. Instead, they introduced a variational auto-encoder model, which uses the concept of latent space, providing answer categories to generate relevant questions. Vedd et al. (2022), recently, proposed a guiding approach with three variant families that conditions the generative process to focus on specific chosen properties of the input image for generating questions. Inspired by previous work, we also use additional information such as answer and answer categories in our experiments. To summarize, the main contributions of our paper are the following:

- In our study, we introduce the first visual question generation system that leverages the power of Transformer-based encoder-decoder architecture for the low resource Bengali language.
- We conduct experiments of multiple variants considering only the image and also additional information as input such as answers and answer categories.
- We evaluate our novel VQG system with well-established text generation evaluation metrics and report our results as the state of the art in Visual Question Generation in Bengali.
- We perform a human evaluation on our generations to assess the quality and the relevance of the questions.

2 Related Works

The advent of visual understanding has been made possible due to continuous research in question answering and the availability of large-scale Visual Question Answering (VQA) datasets (Antol et al., 2015; Johnson et al., 2017; Mostafazadeh et al., 2017). In the past few years, many methods have been proposed to increase the model’s performance for a VQG task. Earlier studies (Xu et al., 2015; Jain et al., 2017; Mostafazadeh et al., 2016; Serban

et al., 2016; Vijayakumar et al., 2018; Ren et al., 2015) have explored the task of visual question generation through Recurrent Neural Network (RNN), Generative Adversarial Network (GAN), and Variational Auto-Encoder (VAE) which either followed algorithmic rule-based or learning-based approach.

In the visual-language domain, the first VQG paper proposed by Mostafazadeh et al. (2016) introduced question-response generation that takes meaningful conversational dialogues as input to generate relevant questions. Zhang et al. (2017) used an LSTM-based encoder-decoder model that automates the generation of meaningful questions with question types to be highly diverse. Motivated by the discriminator setting in GAN, Fan et al. (2018) formulated a visual natural question generation task that learns two non-generic textual characteristics from the perspective of content and linguistics producing non-deterministic and diverse outputs. Whereas, Jain et al. (2017) followed the VAE paradigm along with LSTM networks instead of GAN to generate large set of diverse questions given an image-only input. During inference, their obtained results nevertheless required the use of ground truth answers. To defeat this non-viable scenario, Krishna et al. (2019) proposed a VAE model that uses the concept of latent variable and requires information from the target, i.e. answer categories, as input with the image during inference. Similarly, Vedd et al. (2022) follows the concept of latent variable, however, their proposed model architecture explores VQG from the perspective of guiding, which involves two variant families, explicit and two types of implicit guiding approach. Our work is closely related to their explicit guiding method excluding the use of latent space. Recently, Scialom et al. (2020) proposed a BERT-gen model which is capable of generating texts either in mono or multi-modal representation from out of the box pre-trained encoders.

3 Methodology

In this section, we introduce our transformer based Bengali Visual Question Generation models which can generate meaningful non-generic questions when shown an image along with additional textual information. Our VQG problem is designed as follows: Given an image $\tilde{i} \in I$, where I denotes a set of images, decode a question q . For each image \tilde{i} , we also have access to textual utterances, such as ground truth answer and answer categories. Note,

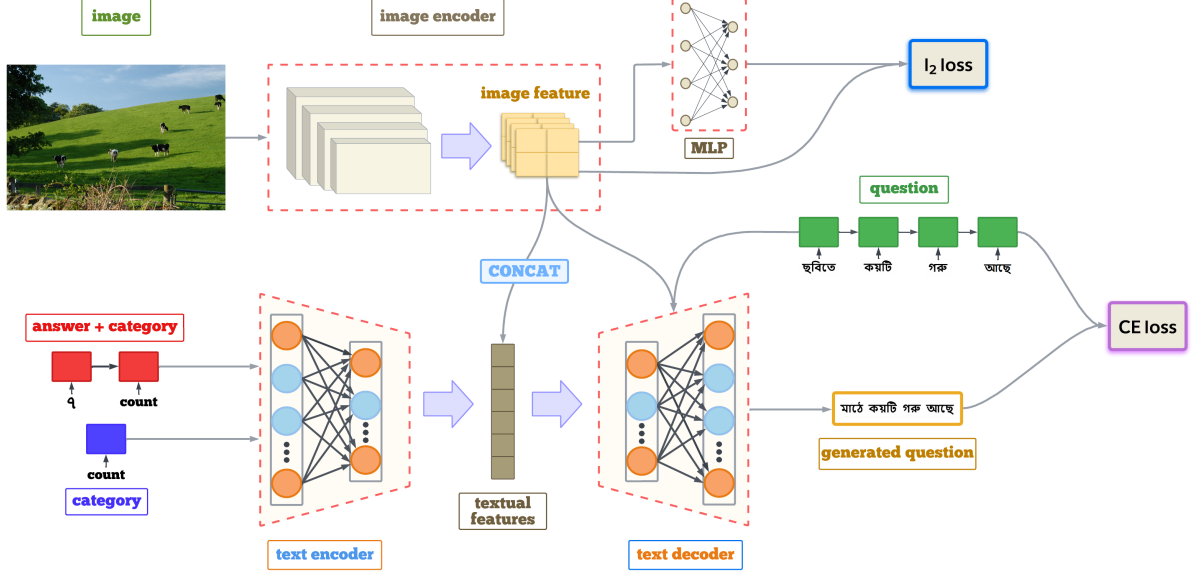


Figure 2: Architecture of the Bengali VQG Model: Given an image, we first extract image features using an image encoder (CNN). Concatenated form answer and category (image-ans-cat) or only category (image-cat) are given as input to the text encoder to obtain textual features which are then concatenated with the obtained image features. Then, this concatenated form of vision and textual modalities combined with target questions are given as inputs to the decoder question generation in Bengali. Finally, we optimize the CE and MSE loss.

we will use terms "answer category" and "category" interchangeably throughout the paper. In our work, we used answer categories from (Krishna et al., 2019) that take 1 out of 16 categorical variables to indicate the type of questions asked. For example, if our model wants to understand answers of "রং (color)" category, then it should generate a question "এটা কি রঙের বাস ? (What color is the bus?)".

Our baseline is an *image - only* model with no additional textual information like answer or category. We present further two variants both of which shares the same architecture but takes different inputs in training. We feed two different textual information to our model during training. The first model is *image - ans - cat* that feeds the concatenated ground truth answer and category to the encoder and is concatenated with the image features. The second model is *image - cat* that takes only the relevant answer category as input to the encoder. In both of the versions, the input image is reconstructed to maximize information between the image and encoded outputs.

Vocabulary: We construct vocabulary considering all the textual utterances: questions, answers and answer categories. Our vocabulary has a total of 7081 entries including the special tokens. We use word level tokenization. We set a default length of 20 token to each of the questions and 5 to each

of the answers. In table 1, we see as maximum length of question in our training dataset is 22 and validation is 21 tokens long, we choose 20 to be the default length. Questions longer than default are truncated and the shorter ones are padded with special <pad> token.

Image Encoder: Given an image \tilde{i} , we can extract image features, $f \in \mathbb{R}^{B \times 300}$ where B is batch size. Our image encoder is a ResNet-18 pretrained CNN model, which is a convolutional neural network with 72-layer architecture consisting 18 deep layers (He et al., 2016). Once obtaining these features, they are passed to a fully connected layer followed by a batch normalization layer. Specifically, given f from image \tilde{i} : $i = BatchNorm(f) \in \mathbb{R}^{B \times 300}$.

Encoder: We build a Transformer encoder (Vaswani et al., 2017) and use Bengali pretrained GloVe (Global Vectors for Word Representation) word vectors (Sarkar, 2019) as the embedding layer of the text encoder. Next, we provide answer or answer categories and image features f as input to the text encoder. Note that, *image-cat* variant only takes answer category c as its input during training and *image-ans-cat* takes concatenated version of answer and category, $[a; c]$ (; operator represents concatenation) as seen in figure 2. For *image-ans-cat* variant, a concatenated ver-

sion of answer and category $[a; c]$ is passed through the embedding layer and projected out as context, $C_{img+ans+cat} = embedding([a; c]) \in \mathbb{R}^{B \times T \times 300}$ where, B is batch size and T is the length of the $[a; c]$. For the *image-cat* variant, we only pass the category, c and similarly generate a context $C_{img+cat} = embedding(c) \in \mathbb{R}^{B \times T \times 300}$, where T is the length of c .

Additionally, we generate padding masks on answer and category $[a; c]_m = generate - mask([a; c]) \in \mathbb{R}^{B \times 1 \times T}$ to avoid <pad> tokens being processed by the encoder as well as the decoder. Same operation is performed on category input c and a masked category is generated c_m . The *image-cat* model takes context, C and masked category c_m as input to the encoder to encode textual feature representation: $S = encoder(C_{img+cat}, c_m) \in \mathbb{R}^{B \times T \times 300}$. We follow the same procedure for the *image-ans-cat* model, where now encoder takes the context, $C_{img+ans+cat}$ and masked concatenated answer and category, $[a; c]_m$.

These textual feature representation S from the encoder are then concatenated to the input image features $i \in \mathbb{R}^{B \times 300}$, thus, representing our final encoder outputs as the concatenation (; operator) of textual and vision modality: $X = [S; i] \in \mathbb{R}^{B \times T \times 300}$ where B is the batch size and T is length of S .

Decoder: Our decoder is a Transformer decoder that also uses GloVe embeddings. Following sequence-to-sequence causal decoding practices, our decoder receives encoder outputs from text encoder and ground truth questions during training. We, initially, extract <start> (Start of Sequence) token from encoder outputs which is then taken to the GPU. Each target question is concatenated with a <start> token, forming a tensor.

In our decoder we follow similar steps as we did in our text encoder. We take ground truth questions q and generate target context: $C_q \in \mathbb{R}^{B \times T \times 300}$ and question masks: $q_m \in \mathbb{R}^{B \times 1 \times 300}$. Before, we pass the target context, C_q to the decoder, we concatenate it with the same image features i that were passed as input to the encoder previously. The final target context can be denoted by $Q = [C_q; i] \in \mathbb{R}^{B \times T \times 300}$. Finally, the decoder takes the encoder outputs X from the text encoder, the concatenated target context Q and the source mask ($[a; c]_m$ or c_m) depending on the model variant and target question q_m in the form of a tuple. Our decoder is represented as following:

$\hat{q} = Decoder(X, Q)$ where the decoder outputs a generated question \hat{q} .

4 Experiments

4.1 Datasets

To collect all relevant information for the VQG task in Bengali, we use the VQA v2.0 (Antol et al., 2015) dataset consisting of 443.8K questions from 82.8K images in the training dataset, and 214.4K questions from 40.5K images for validation dataset. From the annotations of previous work (Krishna et al., 2019), 16 categories were derived from the top 500 answers. The top 500 answers cover around 82% of the total VQA v2.0 dataset (Antol et al., 2015). The annotated categories include objects (e.g. “বিড়াল cat”, “ফুল flower”, attributes (e.g. “ঠান্ডা cold”, “পুরাতন old”), color (“লাল red”, “বাদামী brown”), etc.

	Train	Val
Number of Questions	184100	124795
Number of Images	40800	28336
Max Length of Question (by words)	22	21
Min Length of Question (by words)	1	1
Avg Length of Question (by words)	4	4

Table 1: Analysis of the dataset.

Previously in Bengali machine translation research (Hasan et al., 2020), Google translate was found to be competitive with machine translation models trained in Bengali corpora. In another work on Bengali question answering (Mayeesha et al., 2021), synthetic dataset translated by Google translate was again used for creating Bengali question answering models. Due to Bengali being a low resource language, there has been no available VQG dataset. So we translated the VQA v2.0 (Antol et al., 2015) with Google translate following previous works. We maintained the same partitioning as the original dataset. Due to computational constraints we translated a smaller subset of the training and the validation set. We translate the initial 220K questions and answers for training and 150K questions and answers for validation set in Bengali using GoogleTrans library. In table 1, we see out of 220K training and 150K validation questions, 184K training and 124K validation questions were used. It is because these sets of questions

map to top 500 answers in the dataset and we could not use questions and answers that had no mappings to the 16 categories. In figure 3, we can see the samples of our dataset. The 16 categories in our dataset are following in English - “*activity*”, “*animal*”, “*attribute*”, “*binary*”, “*color*”, “*count*”, “*food*”, “*location*”, “*material*”, “*object*”, “*other*”, “*predicate*”, “*shape*”, “*spatial*”, “*stuff*”, “*time*”.




Image	Category	Question	Answer
	count	কয়টি আমেরিকান পতাকা আছে ? (How many American flags are there?)	১ (1)
	object	মাটিতে কি আছে ? (What is on the ground?)	ঘাস (grass)
		কেন্দ্রে কোন ধরনের ভবন দেখা যায় ? (What type of building is seen on the center?)	গির্জা (church)
	binary	ঝরনা কি বড় ? (Is the shower large?)	হ্যাঁ (yes)
		গামছা ঝুলানো আছে ? (Is there towel hanging?)	হ্যাঁ (yes)

Figure 3: Samples from our dataset

4.2 Training and Optimization

Our transformer based encoder-decoder architecture is a variation of explicit guiding variant established by (Vedd et al., 2022) where object labels, image captions and object detected features were used as guiding information. However, we only use answer categories and answers as additional information in our work. Instead of BERT (Devlin et al., 2019) we use Bengali Glove embeddings (Pennington et al., 2014; Sarkar, 2019) for encoding text. We use less number of layers, attention heads and our embedding dimensions and hidden state dimensions are also reduced due to computational constraints. Similar to work done by (Krishna et al., 2019) we use the concept of answer category as our primary textual information and attempt to generate questions that are conditioned towards a specific category.

In summary, we begin by first passing the image through a Convolutional Neural Network (CNN) to attain a high dimensional encoded representation of image features, i . The image features are passed through an MLP (Multi-layer Perceptron) layer to get a vector representation of reconstructed image features, i_r . Our architecture takes an image and additional information in the form of a concatenated answer and category $[a; c]$ or answer category c as input. We feed these input to our text encoder which then generates the textual S and concatenates the textual S and vision modality rep-

resentations i . Our decoder takes the concatenated form of target context Q , the encoder outputs X , and generates the predicted question, \hat{q} as shown in equation 1.

During training, we optimize the L_q between the predicted \hat{q} and target question q . Additionally, we try to reconstruct the input image from the encoded output, X and minimize the l_2 loss between the reconstructed image features, i_r and the input image features i to maximize mutual information between the input image features and the encoder outputs as mentioned in equation 2.

$$\hat{q} = \text{Decoder}([S; i], [C_q; i]) \quad (1)$$

$$L_q = \text{CrossEntropy}(\hat{q}, q) \quad (2)$$

$$L_i = \|i - i_r\|_2$$

4.3 Inference

During inference, except the image-only variant, both model variants are provided with only answer category (e.g. “*রঙ (color)*”, “*বৈশিষ্ট্য (attribute)*”, “*গননা (count)*”, etc.) alongside an image during inference, because providing answers to the model would violate the realistic scenario (Krishna et al., 2019; Vedd et al., 2022). As a result our model is kept under a realistic inference setting by not providing an answer as input during inference.

4.4 Evaluation Metrics

In our experiments, we followed well established language modelling evaluation metrics BLEU: (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), METEOR (Lavie and Agarwal, 2007), and ROUGE-L (Lin, 2004).

4.5 Implementation details

We use a pretrained ResNet18 as our image encoder to encode image features. Both of our transformer based encoder and decoder uses glove embeddings. We set our transformer encoder and decoder with the following setting: number of layers = 4, number of attention heads = 4, embedding dimension = 300, hidden dimension = 300 and filter size = 300. The model trains a total number of 13000 steps, with a learning rate of 0.003 and batch size of 64. We have implemented our model with pytorch. We expect to release our code and translated dataset publicly at <https://github.com/mahmudhasankhan/vqg-in-bengali>

	466137	85187	82846	349926	82259	567390	
							
Ground Truths	প্লেট কি সাদা ? (Is the plate white?)	এটা কি মার্কিন যুক্তরাষ্ট্রে ? (Is this in the USA?)	পটভূমিতে রক্তগুলা ইটের ভবন আছে ? (How many brick buildings are in the background?)	ফ্রিসবি কি রঙ ? (What color is the Frisbee?)	আবহাওয়া কেমন ? (What is the weather like?)	প্লেটে কি আছে ? (What is on the plate?)	
Image + Ans + Cat	Given Category : Given Answer : Generated Question :	binary হ্যাঁ (yes) এই প্রাণী কি পান ? (What do these animals drink?)	other না (no) এটা কি ইংরেজি ভাষাভাষী দেশ ? (Is it an English speaking country?)	count ২ (2) এটা কি রোদোজ্জ্বল দিন ? (Is it a sunny day?)	color সাদা (white) এই ঘটনা কি ? (What is this event?)	attribute ঠাণ্ডা (cold) এই ছবিতে কি তুষার আছে ? (Is there snow in this picture?)	object হ্যাঁ (yes) পিঞ্জা কি টুকরা করা আছে ? (Is the pizza sliced?)
Image + Cat	Given Category : Generated Question :	binary ট্রেডি বিয়ার কি বিক্রয়ের জন্য ? (Are teddy bears for sale?)	other চিহ্ন গুলো কোন ভাষায় লেখা ? (What language are the signs written in?)	count বাসে কয়টি জানালা আছে ? (How many windows are there in the bus?)	color দলের ইউনিফর্ম কি রং ? (What color is the team's uniform?)	attribute পাহাড়ে কি বরফের তাপমাত্রা ঠাণ্ডা ? (Is the temperature of ice cold in the mountains?)	object প্লেটের আকৃতি কি ? (What is the shape of the plate?)

Figure 4: Qualitative Examples. Ground truths are target questions for both models.

Model	BLEU			CIDEr	METEOR	ROUGE-L
	1	2	3			
ablations						
image-only	34.84	8.04	3.98	10.62	17.14	36.56
text-only	28.05	7.57	3.65	18.72	19.10	29.68
without-image-recon	11.59	4.85	2.08	26.61	12.34	31.43
variants						
image-cat	33.12	13.52	7.56	22.76	17.18	36.12
image-ans-cat	32.97	11.80	3.82	18.63	18.63	36.90

Table 2: Evaluation results of model variants and ablations.

4.6 Model Ablations

We experiment with a series of ablations performed on our model such as image-only does not include text encoder. Inversely, text-only model does not have image encoder. With respect to without-image-recon, we avoid optimizing the reconstruction l_2 loss between the reconstructed image features and input image features. As for our model variants, image-cat and image-ans-cat, the entire architecture remains intact.

5 Results

5.1 Quantitative Results

We test our model variants except with only categorical information because giving answer to a model beforehand would be unrealistic. We tried to figure out which textual input is more significant and leads to better results. Firstly, our model ablations justify our model architecture as such our intact architecture outperforms all the ablations in BLEU-2 and BLEU-3 (see Table 2). Our baseline image-only model achieves a BLEU-3 score of 3.98 which is higher than image-ans-cat variant.

Moreover, we find that in some metrics image-cat model outperforms the image-ans-cat model and in some metrics stay ahead marginally. As seen in table 2, image-cat model achieves a BLEU-3 score of 7.56 that is almost 4 points ahead of both image-only and image-ans-cat model. Moreover, we notice that image-cat model also performs marginally better in CIDEr metrics. However, both the variants show similar performance on other evaluation metrics except for METEOR and ROUGE-L metric where image-ans-cat variant performs slightly better. In comparison to (Ved et al., 2022) for experiments in explicit image-category setting for English, our BLEU-1 score is 33.12 while for English we see a score of 40.8 with a 7.68 difference, however, BLEU-2 and BLEU-3 scores have higher differences. However, for METEOR in English, the score is 20.8 while our image-cat model scores 17.18 with a 3.62 difference only and for ROUGE the English score is 43.0 while we score 36.12 with a 6.88 point difference. Similar experiments on guided visual generation have not been performed for other languages or Bengali to our knowledge, so we compare only

Model	Experiment	
	1	2
image-cat	47.5%	40%
image-ans-cat	30%	37.5%

Table 3: Human evaluation result of our model variants.

with English. While our scores are lower than English, we train on smaller and translated dataset for computational and data annotation related constraints. Based on the quantitative results we can come to a conclusion that categorical information shows better results overall. In the next section, we see the qualitative results where we shall see that categorical information conditions the image-cat variant to generate category specific questions i.e. goal driven, attribute specific questions rather than generic questions.

5.2 Qualitative Results

In figure 4, we can compare the generated questions from our model variants with the reference ground truth question and answer category more illustratively. Questions generated from the image-cat-ans model although are grammatically and semantically correct but in some cases are not conditioned towards the given category. For example, in image 82846, although the question is grammatically correct, however, the generated question does not follow the given category which is “count”. We see similar behavior for images 349926 and 82259 where questions are grammatically correct and relevant to the image but do not follow the category. In contrast, the image-cat model perfectly conditions its questions towards the given category. The questions are not only grammatically and semantically valid but also follow the given categorical information. The questions from the image-cat model generates goal driven, non-generic and category oriented questions. To understand why this variant of VQG performs well although having less side information during training, is likely due to the fact that in validation step both variants only take category side information. Therefore, the image-cat learns better than image-ans-cat.

Additionally, we notice that both variants are able to decode the semantic information from the input image as well. Both variants can rightly identify the objects and features present in the images.

5.3 Human Evaluation

We conducted a human evaluation to understand the quality of the generated questions similar to work done in (Vedd et al., 2022). In our experiments, we ask three annotators to evaluate our generated questions with two questions. There was no annotator overlap where two annotators annotated the same question. We evaluate category wise question generation by comparing two of our model variants, image-cat and image-ans-cat.

In *Experiment 1*, known as the Visual Turing Test, we present annotators with an image, a ground truth question, and a model-generated question. The task of annotators is to discern which question, among the two, they think is produced by the model. *Experiment 2* involves displaying an image to the annotators along with a question generated by the model. Subsequently, the annotators are asked to decide whether the generated question seems relevant to the given image. For each of the experiments we annotate 40 generations for each models, resulting in 80 annotations per experiment. The complete results of our evaluation is listed in table 3.

In *Experiment 1*, the result of our image-cat model outperforms the image-ans-cat variant fooling humans about 47.5% of the time. In a Visual Turing Test, if a model is capable of generating human-like questions, it is expected that its performance would reach approximately 50%. Although close to the desired score of 50%, the image-cat variant represents a promising advancement in surpassing the Visual Turing Test. We evaluate *Experiment 2* on both our model variants where the image-ans-cat model shows a percentage score of 37.5%, outperforming the image-cat model. It is possible that providing the answer with the image and the category helps in generating more relevant questions.

6 Conclusion

We proposed the first VQG work in Bengali and presented a novel transformer based encoder-decoder architecture that generates questions in

Bengali when shown an image and a given answer category. In our work, we presented two variants of our architecture: image-cat and image-ans-cat that differs from what input they receive during training. Both of the variants generate a question based on answer category as guiding information from an image. However, due to having two different input combinations, image-cat performs marginally better in terms of quantitative scores, however, generates goal driven, specific questions conditioned towards the categorical information it receives. In contrast, the image-ans-cat model although generating grammatically valid questions fail to learn about answer categories. Future work could analyze the impact of using more modern CNN architectures and newer pretrained models to generate questions from images.

7 Acknowledgement

This work was funded by the Faculty Research Grant [CTRG-22-SEPS-07], North South University, Bashundhara, Dhaka 1229, Bangladesh

References

- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- Tanjim Taharat Aurpa, Richita Khandakar Rifat, Md Shoaib Ahmed, Md. Musfique Anwar, and A. B. M. Shawkat Ali. 2022. Reading comprehension based question answering system in bangla language with transformer-based learning. *Heliyon*, 8(10):e11052.
- Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, Belgium, Brussels. Association for Computational Linguistics.
- Remi Cadene, Corentin Dancette, Hedi Ben younes, Matthieu Cord, and Devi Parikh. 2019. Rubi: Reducing unimodal biases for visual question answering. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.
- Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4159–4170, Minneapolis, Minnesota. Association for Computational Linguistics.
- Shaoxiang Chen, Ting Yao, and Yu-Gang Jiang. 2019. Deep learning for video captioning: A review. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6283–6290. International Joint Conferences on Artificial Intelligence Organization.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. 2017. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation*, pages 215–233, Copenhagen, Denmark. Association for Computational Linguistics.
- Zhihao Fan, Zhongyu Wei, Siyuan Wang, Yang Liu, and Xuanjing Huang. 2018. A reinforcement learning framework for natural question generation using bi-discriminators. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1763–1774, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question answering. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS’15*, page 2296–2304, Cambridge, MA, USA. MIT Press.
- Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Qi Tian, and Min Zhang. 2022. Loss re-scaling vqa: Revisiting the language prior problem from a class-imbalance view. *Trans. Img. Proc.*, 31:227–238.
- Deepak Gupta, Pabitra Lenka, Asif Ekbal, and Pushpak Bhattacharyya. 2020. A unified framework for multilingual and code-mixed visual question answering. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 900–913, Suzhou, China. Association for Computational Linguistics.
- Tahmid Hasan, Abhik Bhattacharjee, Kazi Samin, Masum Hasan, Madhusudan Basak, M. Sohel Rahman, and Rifat Shahriyar. 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. Association for Computational Linguistics.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. **Deep residual learning for image recognition**. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.
- S M Shahriar Islam, Riyad Ahsan Aunor, Minhajul Islam, Mohammad Yousuf Hossain Anik, A. B. M. Alim Al Islam, and Jannatun Noor. 2022. **Note: Towards devising an efficient vqa in the bengali language**. In *ACM SIGCAS/SIGCHI Conference on Computing and Sustainable Societies (COMPASS), COMPASS '22*, page 632–637, New York, NY, USA. Association for Computing Machinery.
- U. Jain, Z. Zhang, and A. Schwing. 2017. **Creativity: Generating diverse questions using variational autoencoders**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5415–5424, Los Alamitos, CA, USA. IEEE Computer Society.
- H. Jiang, I. Misra, M. Rohrbach, E. Learned-Miller, and X. Chen. 2020. **In defense of grid features for visual question answering**. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10264–10273, Los Alamitos, CA, USA. IEEE Computer Society.
- J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Zitnick, and R. Girshick. 2017. **Clevr: A diagnostic dataset for compositional language and elementary visual reasoning**. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1988–1997, Los Alamitos, CA, USA. IEEE Computer Society.
- Andrej Karpathy and Li Fei-Fei. 2017. **Deep visual-semantic alignments for generating image descriptions**. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):664–676.
- R. Krishna, M. Bernstein, and L. Fei-Fei. 2019. **Information maximizing visual question generation**. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2008–2018, Los Alamitos, CA, USA. IEEE Computer Society.
- Alon Lavie and Abhaya Agarwal. 2007. **METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments**. In *Proceedings of the Second Workshop on Statistical Machine Translation*, pages 228–231, Prague, Czech Republic. Association for Computational Linguistics.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Tasmiah Tahsin Mayeesha, Abdullah Md Sarwar, and Rashedur M. Rahman. 2021. **Deep learning based question answering system in bengali**. *Journal of Information and Telecommunication*, 5(2):145–178.
- Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios Spithourakis, and Lucy Vanderwende. 2017. **Image-grounded conversations: Multimodal context for natural question and response generation**. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 462–472, Taipei, Taiwan. Asian Federation of Natural Language Processing.
- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. **Generating natural questions about an image**. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1802–1813, Berlin, Germany. Association for Computational Linguistics.
- Liangming Pan, Wenqiang Lei, Tat-Seng Chua, and Min-Yen Kan. 2019. **Recent advances in neural question generation**.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 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. Association for Computational Linguistics.
- Gao Peng, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. 2019. **Multi-modality latent interaction network for visual question answering**. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5824–5834.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. **GloVe: Global vectors for word representation**. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
- Mahamudul Hasan Rafi, Shifat Islam, S. M. Hasan Imtiaz Labib, SM Sajid Hasan, Faisal Muhammad Shah, and Sifat Ahmed. 2022. **A deep learning-based bengali visual question answering system**. In *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pages 114–119.
- Mengye Ren, Ryan Kiros, and Richard S. Zemel. 2015. **Exploring models and data for image question answering**. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2, NIPS'15*, page 2953–2961, Cambridge, MA, USA. MIT Press.
- Sagor Sarkar. 2019. <https://github.com/sagorbrur/glove-bengali>.

- Thomas Scialom, Patrick Bordes, Paul-Alexis Dray, Jacopo Staiano, and Patrick Gallinari. 2020. [What bert sees: Cross-modal transfer for visual question generation](#). In *International Conference on Natural Language Generation*.
- Iulian Vlad Serban, Alberto García-Durán, Caglar Gulcehre, Sungjin Ahn, Sarath Chandar, Aaron Courville, and Yoshua Bengio. 2016. [Generating factoid questions with recurrent neural networks: The 30M factoid question-answer corpus](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 588–598, Berlin, Germany. Association for Computational Linguistics.
- Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. 2018. [Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1918–1928, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Lucia Specia, Stella Frank, Khalil Sima’an, and Desmond Elliott. 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, Berlin, Germany. Association for Computational Linguistics.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. [Cider: Consensus-based image description evaluation](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.
- Nihir Vedd, Zixu Wang, Marek Rei, Yishu Miao, and Lucia Specia. 2022. [Guiding visual question generation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1640–1654, Seattle, United States. Association for Computational Linguistics.
- Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2018. [Diverse beam search: Decoding diverse solutions from neural sequence models](#).
- O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. 2015. [Show and tell: A neural image caption generator](#). In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3156–3164, Los Alamitos, CA, USA. IEEE Computer Society.
- Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. [Show, attend and tell: Neural image caption generation with visual attention](#). In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37, ICML’15*, page 2048–2057. JMLR.org.
- Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. [Automatic generation of grounded visual questions](#). In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 4235–4243, United States of America. Association for the Advancement of Artificial Intelligence (AAAI). International Joint Conference on Artificial Intelligence 2017, IJCAI 2017 ; Conference date: 19-08-2017 Through 25-08-2017.