Visual Question Answering in Bangla to assist individuals with visual impairments extract information about objects and their spatial relationships in images

ABSTRACT

With millions of speakers worldwide, Bangla is one of the most spoken languages in the world and as a result, a large number of people rely on Bangla as their primary medium of communication. Among them, a lot of individuals have visual impairments including but not limited to central vision loss, peripheral vision loss, and blurry vision. This poses several challenges to these individuals - one of them being extracting information from images. While techniques such as image captioning have been proposed to address this issue, visual question answering or VQA offers a much in-depth and robust way to understand an image. However, there has been a paucity of research into developing a VQA system in Bangla to assist individuals with visual impairments. To reduce this gap, we introduce a VQA system in Bangla designed to assist visually impaired individuals. VQA is a multifaceted problem, and in this paper we focus on finding the spatial relationships between objects. We broke down this problem into four sub-tasks: object detection, object counting, and finally relative positioning for the detected objects. The system takes in questions from the user, understands which subtask to perform and then returns the answer. We leveraged several pre-trained models such as Bangla-BERT, EfficientDet-D7, InceptionResNetV2, and MiDas v2.1. The major aspects of this paper are the introduction of a procedurally generated dataset to train models to identify what action to perform based on the prompt of the user and using image segmentations to identify the relative spatial position between objects in all three spatial dimensions.

CCS CONCEPTS

• Computing methodologies; • Artificial Intelligence; • Natural language processing; • Information extraction;

KEYWORDS

Object Detection, Image Segmentation, Monocular Depth Estimation, Visual Question Answering, Bangla NLP, Spatial Positioning

ACM Reference Format:
. 2022. Visual Question Answering in Bangla to assist individuals with visual impairments extract information about objects and their spatial relationships in images. In Proceedings of ICCTA 2023. ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Many solutions have been proposed to enable individuals with visual impairments to extract information from images. One such solution from the field of deep learning is using image captioning to help such individuals understand what is happening in an image. However, while image captioning can successfully provide a brief overview of an image, it falls short when it comes to answering more specific and in-depth questions about an image. Examples of such questions include, "How many people are there in the image?", "What object is there on top of the table?", "What object is there to the right of the television?", and "What is the person beside the bookshelf doing?". Visual Question Answering provides an answer to this problem by allowing users to query information about an image and helping to provide a more thorough understanding of the image to the users which is very beneficial to users with visual impairments [8].

Despite the promise of Visual Question Answering in addressing this problem, there has been inadequate research done into developing a VQA system in the Bangla language. In Bangladesh alone, a predominantly Bangla speaking country, 650,000 people above the age of 30 suffer from blindness [3]. We believe that by offering an effective VQA system in Bangla, we can vastly improve the ability of the visually impaired to retrieve context-specific information from images.

We break down the task into two parts. In the first part, we perform question featurization. Here, we generate context-sensitive word embeddings from the BERT model [2] which are then fed into an BiLSTM layer [12]. The output from the BiLSTM layer is then fed into two fully connected layers which then classifies which task the system has to perform.

In the second part of the task, we divide the task further into three sub-tasks which are object detection, object counting, and finding relative position of objects in all three dimensions. After performing the required sub-task, the answer is returned to the user. The models we have used while performing these sub-tasks are: EfficientDet-D7 [15], Mask R-CNN Inception ResNet V2 1024x1024 [13], and MiDaS v2.1 small [11].

2 LITERATURE REVIEW

Various efforts have been made in the past to make the architectures in each of the three steps outlined above more expressive while also ensuring their computational requirements are feasible. [4] hypothesized that the outer product of the question and image vectors would produce better results than performing concatenation or element-wise addition or multiplication on them. To tackle the issue of the outer product producing a more computationally demanding higher dimension matrix, they used Multimodal Compact Bilinear Pooling (MCB) to keep the computational requirements feasible. Furthermore, [10] introduced a co-attention model so that

ICCTA 2023, May 10–12, 2023, Vienna, Austria
© 2022 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

the model that understand which words in the question are more relevant to the answer it is looking for. This approach builds on top of previous proposals that focused on visual attention where the models tried to understand which parts of the image were more helpful in answering the question. [5] used question-guided convolutions to capture the visual spatial relationships that are lost in current approaches when trying to merge the image and text vectors.

The vast majority of research on VQA has been conducted using questions and answers in English. The morphological and synctactic complexity of Bangla varies quite significantly from that of English and these differences need to be captured when making VQA effective for the Bangla language. As opposed to English, VQA datasets in Bangla are not readily available - which is why we address the issue of VQA in Bangla in a manner different from the methods listed above.

3 BACKGROUND ARCHITECTURES

1. BERT model
Bidirectional Encoder Representations from Transformers, or BERT, was introduced in 2018 by [2]. BERT utilizes the transformer architecture, more specifically the encoders, by stacking encoders together. BERT is more powerful than previous models as it can understand context in a sentence in a truly bidirectional way. This is achieved by training BERT on two tasks - Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM task, some words in input sentences were replaced with MASK tokens and the model had to predict what the words were based on the context provided by the sentences. In the NSP task, the model was given a pair of sentences A and B, and it had to determine whether sentence B follows sentence A.
BERT offers a much more powerful understanding of language and context which has given state of the art results in multiple Natural Language Processing (NLP) tasks and hence is used in a wide array of NLP applications.
For our research, we used the BERT-Bangla-Base [1] model which is specifically trained on a large corpus of Bangla text on the MLM and NSP tasks described above.

2. EfficientDet-D7
EfficientDet-D7 [15] is a part of the family of EfficientDet architectures. It provides state of the art results in object detection on the Common Objects in Context (COCO) [9] test-dev dataset with an AP score of 55.1. It uses a novel weighted bi-directional feature pyramid network (BiFPN) as well as a compound scale method to uniformly scale resolution, depth, and width for all backbone , feature network, and box/class prediction networks. These optimizations allow EfficientDet-D7 to achieve state of the art results in object detection while also being far smaller than previous state of the arts and drastically reducing the number of computations needed.

3. Mask R-CNN Inception ResNet V2 1024x1024
Mask R-CNN Inception ResNet V2 1024x1024 [13] is based on the family of Inception networks introduced in [14] which incorporates residual networks which were introduced in [6]. Inception networks consist of inception modules with filters of different sizes so that convolution operations are not restricted to one filter size. Residual networks help the flow of information through deep networks by introducing skip connections which passes input from shallower layers to deeper ones.
4. MiDaS v2.1 small MiDaS v2.1 small belongs to the MiDaS group of architectures introduced in [11] for performing monocular depth estimation on images. These architectures were pre-trained on a diverse range of images over extended periods of time. They find the inverse depth of the images, meaning that a higher value is assigned to pixels that are closer to the camera than the ones that are farther away.

4 METHODOLOGY

2.1 Understanding user prompts

2.1.1 Task overview

Given a question from the user, the system has to respond to four categories of questions:

I) Object Detection: Here, the users ask what objects are present in the image. The system detects the objects in the images and returns the classes of the objects detected in the image.

II) Object Counting: Here, the users ask how many instances of a particular object are present in the image. The system returns the count of the specific object in the image.

III) Positioning Questions: Here, the users asks what objects are present in a particular direction of the object. The directions they can specify are: left, right, above, below, front, and behind.

IV) Positioning and Counting Questions: Here, the users asks how many objects are present in a particular direction of an object. It is a positioning question followed by a counting question and the system responds accordingly.

For this project, we have 80 classes of objects all belonging to the Common Objects in Context (COCO) dataset [9] with all of their labels translated to Bangla by native Bangla speakers. Given a question from the user, the system performs a classification task. This classification task consists of 1041 classes - 1 class for object detection task, 80 classes for object counting with each object category having its own class. 480 classes are for detecting objects in a particular direction relative to another object. Since we have 6 directions - left, right, above, below, front, and behind - for 80 objects, we get 480 classes. A further 480 classes are actions for questions which requires the system to detect objects in a particular direction followed by counting them. A full breakdown of the classes is given in Table 1.

Object Detection	1 class
Object Counting	80 classes
Relative Positioning	480 classes
Relative Positioning and Object Counting	480 classes
Total	1041 classes

Table 1: Breakdown of classes corresponding to each question category

2.1.2 Dataset

Due to the scarcity of VQA datasets in Bangla, we used our own procedurally generated dataset. To accomplish this, we defined a set of question formats containing some parameters depending on the question type. An example of such a format is "ছবিতে কয়টি object আছে?" (Translation: "How many [object] are there in the image?"). By replacing the 'object' token with a randomly selected from our 80 object types, we can generate a question for the dataset. The corresponding answers can be generated using a similar format. Given we already know what a sub-task a particular question format is referring to, we can simply generate answers using the corresponding parameters. In the above example, the answer would be generated by simply setting it to 'counting_object' where the object token is the same random object type we had previously selected for the question. A complete breakdown of question-answer formats is given in Table 2 along with English translations.

Question Format	Question Category
1) ছবিতে কি কি জিনিস আছে? (What are the things in the image?) 2) ছবিতে কি দেখা যাচ্ছে? (What can be seen in the image?)	Detection
1) ছবিতে কয়টি obj দেখা যাচ্ছে? (How many obj can be seen in the image?) 2) কয়টি obj আছে ছবিতে? (How many obj can be seen in the image?)	Object Counting
1) obj-এর direction কি আছে? (What is there in the direction of obj) 2) কি আছে obj-এর direction? (What is there is obj's direction?)	Relative Positioning
1) obj-এর direction কয়টি জিনিস আছে? (How many things are there in the direction of obj?)	Relative Positioning and Counting

Table 2: Breakdown of question formats in the dataset

2.1.3 Word Embeddings

We first tokenize the question input from the user. Tokenizing is the process of breaking down words and combinations of words in a sentence into individual tokens. We then use the BERT model to generate word embeddings for the tokens. Word embeddings are vector representations of tokens that can be used as input into various Natural Language Processing (NLP) tasks. Word embeddings produced by BERT offer a way into preserving the semantic meaning of a token in a given context while also offering a token representation that can be fed into deep learning models. Since our data is in Bangla, we used the Bangla-BERT-Base model [1], which is the BERT model trained on a large corpus of Bangla text. One of the benefits word embeddings offer is that similar words have similar vector representations. This ensures that our system can comprehend questions in a much more flexible way and can understand a question properly even if the specific words used by the user to construct the question may have been different from the questions the model was trained on.

2.1.4 BiLSTM layer

LSTMs [7] are a type of Recurrent Neural Network that can detect long range dependencies in sequential data like sentences. It employs a forget gate which can determine which information to pass on to the next RNN unit and which to forget. However, LSTMs are limited in the sense that they can do this in one direction only. BiLSTM [12] is a special type of LSTM which overcomes this by operating in two directions and thereby detecting dependencies both in the previous and subsequent parts of a sentence. The word embeddings generated by the BERT model are passed through a BiLSTM layer with 512 hidden units. It is then fed into two fully connected layer which finally outputs what sub-task the question is asking the system to perform. The summary of this model is provided in Table 3.

Layer (type)	Output Shape	Param
input_2 (InputLayer)	[(None, 20, 768)]	0
bidirectional_1 (Bidirectional)	(None, 20, 1024)	5246976
$\text{flatten}_1(Flatten)$	(None, 20480)	0
dense_2 (Dense)	(None, 1024)	20972544
dense_3 (Dense)	(None, 1041)	1067025

Total parameters: 27,286,545
Trainable parameters: 27,286,545
Non-trainable parameters: 0

Table 3: Model Summary

2.2 Finding relative positions of objects:

In this sub-task, we have to find the direction in which an object is with respect to another object. There are 6 possible answers to this question: left, right, above, below, front, behind. These 6 answers can be grouped into 3 categories: left and right as horizontal relative position, above and below as vertical relative position, and front and behind as the relative depth of the objects in the image. There are two key assumptions made when finding the solution this problem:

I) The question the user is asking is from the perspective of the camera which took the picture

II) Three answers from the three categories mentioned above can be true at the same time.

Given that there can be multiple answers to a question, our system needs to be able to find the most relevant answer which will help users visualize the relative positions between a pair of objects more easily. We offer a method to do this, later on.

Our goal in this sub-task is to find computationally inexpensive techniques to find the relative positions between objects. There are two reasons for this:

I) Short times to get answers allows users to input queries about the image and hence get a deeper understanding of the spatial relationships of the objects in the image.

II) Users have the ability to pose a greater variety of queries as computationally expensive queries can be answered quickly without disrupting the user experience.

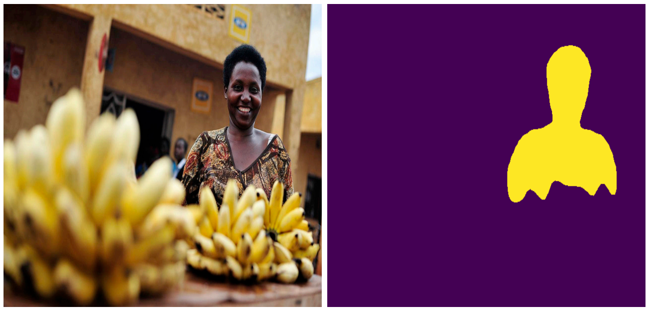

Fig 1. Image Segmentation of a person (right) from original image (left). Images are from COCO dataset [9]

In order to perform this sub-task, we found the semantic segmentations of the image. Through semantic segmentation, we can find each and every pixel that belongs to a given object as shown in Fig. 2. We obtained the image segmentations using the pre-trained model Mask R-CNN Inception ResNet V2 [13]. By using the information given to us by semantic segmentation, we propose the following techniques to find relative position of objects as follows:

2.2.1) Horizontal Relative Position

Every pixel in the semantic segmentation map of the input image can be represented as a tuple (x, y, z) where x and y are the x-coordinate and y-coordinate of the pixel respectively and z is a boolean value which indicates whether the pixel belongs to the object of interest or not. For this task, our input is two objects - A and B - where we are trying to find the relative position of B with respect to A. We first obtain the segmentations of A and B in the image. Then, we find the mean of the unique x-coordinate values of all pixels belonging to B which is denoted as $M^x{}_B$. We find the same quantity for object A and denote it as $M^x{}_A$. If $M^x{}_B$ is greater than $M^x{}_A$, our answer is 'right'. If $M^x{}_B$ is smaller than $M^x{}_A$, our answer is 'left'. If both quantities are the same, we say that are aligned along the horizontal axis.

The intuition behind this operation is that it can simply be seen as an extension to finding the difference between 2 pixels. For example, if A and B were just individual pixels instead of objects, a positive x-coordinate difference between B and A would indicate B is to the left of A while a negative one would indicate B is to the right of A. Thus, by finding the differences between the means of all unique x-coordinates between the objects, we can get an idea of the spatial relationship between the objects in the horizontal axis.

One of the advantages of this method is that the object does not have to be strictly to the left or right of another object for it to give an accurate estimation. This is particularly important when there is a great difference in the sizes of the objects. For example, if a cat is sitting to the left side of the table, the method will be able to identify this even though the cat is not to the left of the table in a strict sense. This method recognizes the flexibility of the Bangla language in conversational settings.

2.2.2) Vertical Relative Position

The method used for this task is the same as previous one except for two differences:

I) We use y-coordinates of the pixels instead of x-coordinates to perform our calculations

II) If the mean of the unique y-coordinates of object B is greater than that of A, then B is below A and vice versa. If they are equal, we say that they are aligned in the vertical direction.

2.2.3) Finding the relative depth of two objects

For this task, we have to find how far the objects are from the camera taking the picture. Since images have 2 spatial dimensions, the approach used above does not work. As such, we use the MiDaS v2.1 small model to generate a depth map of the image. In the depth map, each pixel can be represented as a 3-tuple (x, y, d) where x is the x-coordinate of the pixel, y is the y-coordinate of the pixel and d is the depth of the pixel. A higher d indicates the pixel is closer to the camera than a lower d. As done previously, we find the segmentations of objects A and B in the image. Then, for both objects A and B, we sum up the depth values of pixels that overlap with the segmentations. More formally, we sum up all values of d of every pixel where $z = 1$ for both object A and object B. We then divide this sum by the total number of pixels in the segmentations of the object for object A and object B. This gives us the average depth of A and B. If the average depth of B is greater than that of A, the answer is 'front' (B is in front of A) and if the average depth of B is lesser than that of A, the answer is 'behind'. If they are the same, we say that the objects are aligned along the depth of the image.

By using the techniques above, a range of queries relating to spatial relationships in images can be answered.

2.3 Finding the best answer to questions about spatial relationships

While answers pertaining to questions about spatial dimensions can be answered in terms of either horizontal direction (left or right), vertical direction (above or below), or in terms of image depth (front or behind), in most circumstances one of these answers is more important to our understanding of the image than the other two. For instance, in Fig 2 below, while the spatial relationship between the person and the surfboard can be answered in terms of any of the three aforementioned categories, the vertical direction is likely to aid people's comprehension the most. That is, the person is 'above' the surfboard.

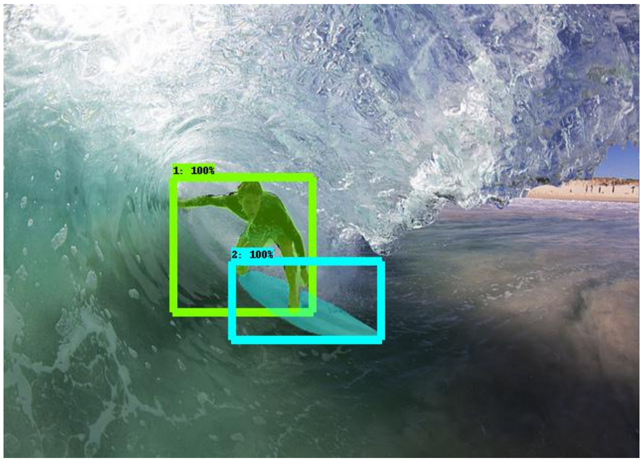

Fig 2. The person is 'above' the surfboard is the answer that is likely to give a person the best understanding of their spatial relationship. Image is from COCO dataset [9]

To determine the best answer, we propose a simple heuristic to determine the best answer. We need three quantities we had calculated previously:

I) The absolute value of the difference between the mean values of the unique x-coordinates of object A and object B.

II) The absolute value of the difference between the mean values of the unique y-coordinates of object A and object B

III) The absolute value of difference in the mean depths of object A and B in the image

We take the maximum of these three quantities and choose the dimension which corresponds to this maximum. That is, if quantity (I) is the maximum, we will answer in terms of horizontal direction (left or right), if quantity (II) is the maximum, we answer in terms of vertical direction (above or below), and if the maximum quantity is (III), we answer in terms of relative depth (front or above).

There is one important detail to keep in mind about quantities (I) and (II). If the width of the image is greater than the height, quantity (II) has to be scaled by a factor of width / height. Conversely, if the height is greater than the width, quantity (I) has to be scaled by a factor of height / image. This is done to ensure that comparison between the dimensions is fair.

Quantity (III) does not require any such scaling since it is not dependent on the height and width of the image.

The idea behind this technique is that the spatial dimension where the difference between the objects is the most pronounced is the one that gives the most relevant answer to the question of spatial dimension between a pair of objects.

Some of the results obtained by using this method are illustrated in Fig 3, Fig 4, Fig 5, and Fig 6 below.

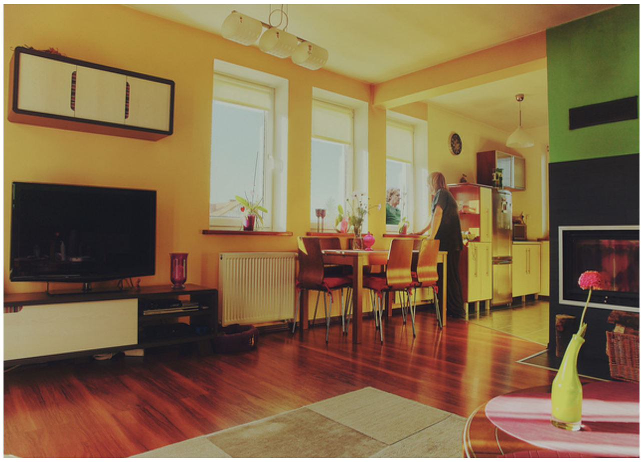

Fig 3. The objects of interest are the television and the flower vase to the left side of the image. The values of quantities (I), (II), and (III) were 95.0, 54.1, and 51.4 respectively. Thus the answer is 'flower vase is to the right of the television'. Image is from COCO dataset [9].

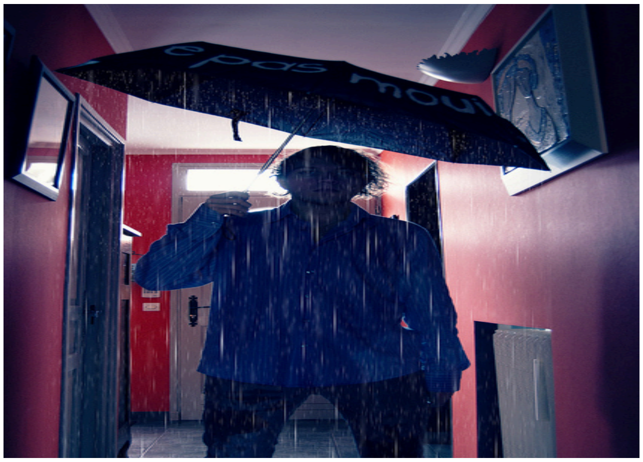

Fig 4. The objects of interest are the person and the umbrella. The values of quantities (I), (II), and (III) were 66.5, 84.1, and 22.2 respectively. Thus the answer is 'umbrella is above the person'. Image is from COCO dataset [9].

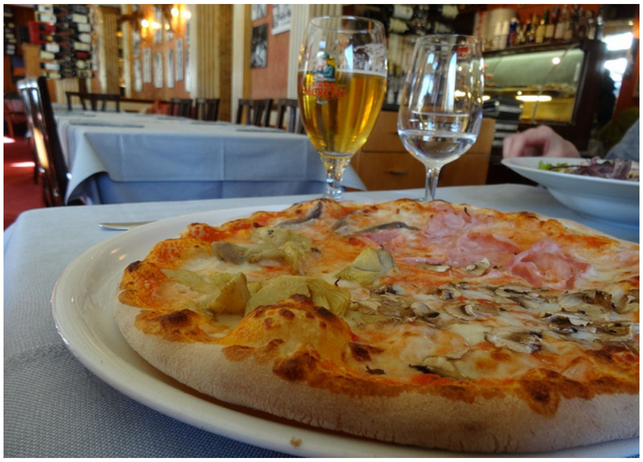

Fig 5. The objects of interest are the pizza and the wine glass. The values of quantities (I), (II), and (III) were 66.0, 323.8, and 642.3 respectively. Thus the answer is 'pizza is behind the wine glass'. Image is from COCO dataset [9].

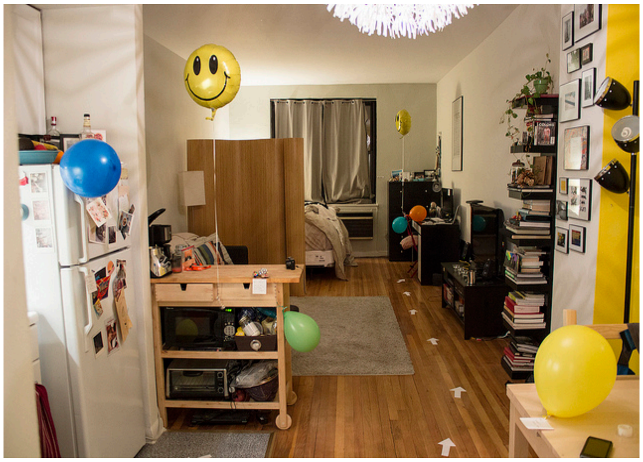

Fig 6. The objects of interest are the refrigerator and the bottle to the left side of the image. The values of quantities (I), (II), and (III) were 35.5, 252.4, and 38.2 respectively. Thus the answer is 'bottle is above the refrigerator'. Image is from COCO dataset [9].

2.4 Object Detection and Object Counting

These tasks are fairly straightforward with the help of modern deep learning architectures. To perform object detection on images using a EfficientDet-D7 architecture, as discussed above, pre-trained on the COCO dataset. We used a threshold score of 0.3, meaning that the obtained confidence score has for an object has to be 0.3 or more for it to be reported as detected. Image is from COCO dataset [9].

After performing object detection, object counting is a trivial problem as we return the number of instances that detected as belonging to a specific class.

5 RESULT ANALYSIS

Understanding User Prompts

In this task, we had to classify questions depending on what sub-task it required us to perform. To make our model flexible and understand questions that may vary from the training data, we obtained word embeddings of our tokenized questions. We then passed them through a BiLSTM layer and finally through two fully connected layers to find the answer.

We used the question format specified in Table 1 to generate 60,896 pairs of questions and answers for the model to train on. The model was trained for 10 epochs with a batch size of 64. A training accuracy of 99.06% was obtained.

Additionally, we also wanted to test how robust and flexible our model was and whether it could adjust to queries of a different format to the ones we used in the training phase but posed the questions as the ones mentioned in the training phase. To obtain this testing partition, we modified the original formats such as using Bangla synonyms for certain terms, changing punctuation, changing the sequence of words, and omitting certain words in the questions to generate 20,000 modified questions. On this data, our model achieved an accuracy of 88.91%. The list of modified question formats can be seen in Table 3. Since some of the changes are quite subtle, their English translations may seem similar as the changes cannot be represented in English. However, all question formats are distinct. The training and testing graphs are given in Fig 7.

Question Format	Question Category
1) ছবিতে কি আছে? (What is in the image?) 2) চিত্রে কি দেখা যাচ্ছে (What can be seen in the picture) 3) কি আছে ছবিতে? (What's there in the picture?)	Detection
1) ছবিতে কয়টি obj আছে? (How many obj are there in the image?) 2) কয়টি obj আছে? (How many obj are there?) 3) obj আছে কয়টি? (How many obj?)	Object Counting
1) কি আছে obj-এর direction (What is there in the direction of obj) 2) আছে কি obj-এর direction? (What is there is obj's direction?)	Relative Positioning
1) obj-এর direction কয়টি জিনিস আছে (How many things are there in the direction of obj) 2) obj-এর direction জিনিস আছে কয়টি? (How many things are there in the direction of obj)	Relative Positioning and Counting

Table 4: Breakdown of modified question formats

Relative Positioning
To test the efficacy of the techniques introduced to find the spatial relationships between objects, we collected 150 object pairs across 33 images from the COCO dataset [9] for each of the three categories - horizontal relative position (left or right), vertical relative position (above or below), and finding relative depths of objects (front or behind) and we obtained accuracies of 99.3%, 99.3%, and 97.3% respectively.

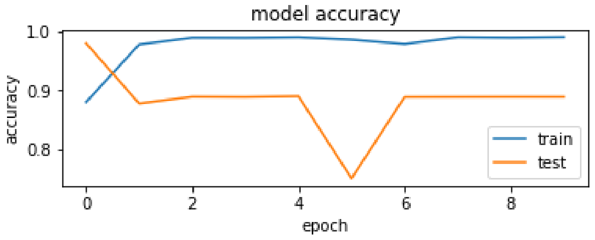

Fig 7. Training and testing graphs of question classification model.

All of the results are summarized in Table 4.

Task	Train Accuracy	Test Accuracy
Question Classification	99.06%	88.91%
Horizontal Relative Positioning	-	99.3%
Vertical Relative Positioning	-	99.3%
Finding Relative Depth	-	97.3%

Table 5: Breakdown of accuracy of each of the tasks

6 DISCUSSION

From the training and testing metrics of the question classification task, we can see that the model generalizes reasonably well to questions of a different format to the ones used to train it. This performance was achievable due to the large amount of training data we generated procedurally as well the word embeddings generated by the Bangla-BERT model. By providing more data, and a more diverse range of question formats, the performance of the model should improve even further.

For the relative positioning task, we used simple heuristics to convert the numeric outputs provided by the models used to categories that can be easily understood by humans. We see from the accuracies that these methods performs well purely from a spatial perspective. There are some limitations to this method. Firstly, this method will not provide consistent results if certain transformations are applied to the image. For example, if the image is rotated, these methods will be unable to detect it. Secondly, these methods only capture spatial relationships between objects and not their semantic relationships. For example, relationships like 'the person is riding the horse' cannot be detected by this system. Future research can be focused on answering these issues.

Finally, we also introduced a method for determining the best answer among all three spatial dimensions. By ensuring the scale of pixels in all dimensions are the same, we can select the dimension that 'jumps out' the most to the user by choosing the most dimension where the difference between the objects is the most pronounced.

7 CONCLUSION

In this paper, we introduced a visual question answering system which allows individuals with visual impairments to better understand the spatial relationships between the objects. For this, we leveraged various state of the art models. We also constructed a procedurally generated dataset to train a model identify what action to perform based on user prompts as well as use image segmentations to find the spatial relationships between objects.
While relative spatial positioning is a crucial aspect of VQA, there is a myriad of other aspects of VQA such as scene classification

and action classification. Future research can be focused on exploring these applications in Bangla to develop a robust Bangla VQA system.

REFERENCES

[1] Abhik Bhattacharjee, Tahmid Hasan, Wasi Ahmad, Kazi Samin Mubasshir, Md Saiful Islam, Anindya Iqbal, M. Sohel Rahman, and Rifat Shahriyar. Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla. Findings of the Association for Computational Linguistics: NAACL 2022, 2022.

[2] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.

[3] Brendan Dineen, Rupert Bourne, S Ali, D Huq, and G Johnson. Prevalence and causes of blindness and visual impairment in bangladeshi adults: Results of the national blindness and low vision survey of bangladesh. The British journal of ophthalmology, 87:820--8, 07 2003.

[4] Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. Multimodal compact bilinear pooling for visual question answering and visual grounding. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 457--468, Austin, Texas, November 2016. Association for Computational Linguistics.

[5] Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven C. H. Hoi, and Xiaogang Wang. Question-guided hybrid convolution for visual question answering. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, Computer Vision -- ECCV 2018, pages 485--501, Cham, 2018. Springer International Publishing.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770--778, 2016.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9:1735--80, 12 1997.

[8] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. Computer Vision and Image Understanding, 163:3--20, 2017. Language in Vision.

[9] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In European conference on computer vision, pages 740--755. Springer, 2014.

[10] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In Proceedings of the 30th International Conference on Neural Information Processing Systems, NIPS'16, page 289–297, Red Hook, NY, USA, 2016. Curran Associates Inc.

[11] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(3), 2022.

[12] Mike Schuster and Kuldip Paliwal. Bidirectional recurrent neural networks. Signal Processing, IEEE Transactions on, 45:2673 -- 2681, 12 1997.

[13] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In Thirty-first AAAI conference on artificial intelligence, 2017.

[14] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 1--9, 2015.

[15] Mingxing Tan, Ruoming Pang, and Quoc V Le. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10781--10790, 2020.