Visual Question Answering in Bangla

Sheikh Ayatur Rahman
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sheikh.ayatur.rahman@g.bracu.ac.bd

Albert Boateng
Mathematics and Natural Sciences
BRAC University
Dhaka, Bangladesh
albert.boateng@g.bracu.ac.bd

Sabiha Tahseen
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
sabiha.tahseen@g.bracu.ac.bd

Annajiat Alim Rasel
Computer Science and Engineering
BRAC University
Dhaka, Bangladesh
annajiat@gmail.com

Abstract

In recent years, the problem of Visual Question Answering (VQA) has been given a lot of attention. A domain where computer vision and natural language processing intersects, VQA involves asking arbitrary questions about an image such as "What is the person in the red shirt doing?", "What is the cat sitting on?", and "What is the color of the dog?". While there has been extensive study of VQA in English, research on VQA in Bangla has been scarce. VQA has shown promising signs in a wide variety of domains such as assisting visually-impaired individuals and making education more robust. With Bangla having approximately 300 million speakers worldwide, it is therefore imperative for VQA to be implemented in Bangla as it has the potential of benefitting millions of people. Our aim in this paper is to compare various Convolutional Neural Netowrk (CNN) and Recurrent Neural Network (RNN) based models and introduce a labelled dataset for performing Bangla-VQA.

Keywords: BERT, NLP, CV, VQA, Question-Answering

1 Introduction

VQA offers us a new way to look at images beyond just classifying them. By using VQA, we can extract various contextual information about objects of interest from the image such as location, shape, color, quantity, etc. This offers multiple possibilities. For example, visually-impaired users can provide prompts to a VQA model to get specific contextual information about an image rather than solely relying on image captions that is predominantly used in websites today. As VQA models become more robust, the scope of their application grows larger.

As a cross domain problem, VQA relies on techniques used in both computer vision and natural language processing. As such, both CNN and RNN based architectures are used in conjunction with each other. Some of the challenges in VQA comes from the high computational cost required to train a robust VQA model, preparing large labelled datasets containing both images and questions specific to those image, and making VQA models for open ended questions and not only on questions which are in a multiple choice question format. This paper aims to shed light on the application of VQA in Bangla and address the paucity of readily available datasets containing questions and answers in Bangla by introducing a Bangla-VQA dataset.

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. (Fukui et al., 2016) hypothesized that the outer product of the question and image vectors would produce better results than performing concatenation or elementwise addition or multiplication on them. To tackle the issue of the outer product producing a more computationally demanding higher dimension matrix, they used Mutimodal Compact Bilin-

ear Pooling (MCB) to keep the computational requirements feasible. Furthermore, (Lu et al., 2016) introduced a co-attention model so that 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. (Gao et al., 2018) 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 synctatic 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 an issue this paper will utilize different methods to address.

3 Methodology

A. Image Encodings

To obtain the vector representations of the images, we pass them through the pre-trained VGG19 model which has its fully connected layers removed. The VGG19 model consists of a series of convolutional and max pooling layers that help to capture to the features of the image. In the convolutional layers, a filter or a kernel is passed over the image to identify features that are helpful in identifying the important portions of the image. The filters themselves are trainable parameters that can be trained by backpropagation so that they can identify the features that are relevant. The maxpooling layers reduce the dimensions of the vectors that are passed onto subsequent layers which helps to reduce computational requirements. By obtaining the output before it is passed onto the fully connected layers, we obtain a 7x7x512 dimensional matrix which has the identified features of the image encoded within it. This representation is translation invariant, which means that the position of the features do not affect the final result. This allows the model to generalize to images that have the same object in different positions on the image.

B. Word Embeddings

We utilized the BERT architecture to obtain the vector representations of the images. BERT was introduced in (Devlin et al., 2019), and has shown to achieve state of the art performance in many NLP tasks. As opposed to traditional RNN based models, BERT utilizes transformer encoders which can process text in a bidirectional manner which allows it to find the word embedding for each word based on the context surrounding the word.

Since we are dealing with questions and answers in Bangla, we used the Bangla BERT Base, which is the BERT model trained on a corpus of Bangla text in a manner similar to the original BERT.

To obtain the word embeddings, we first pre-process each question by adding the [CLS] and [SEP] tokens which represents the start and end of each question respectively. We then tokenize each question by breaking them down into words and then further splitting the words to obtain all the components of the question. Finally, we pass these tokens through the BERT model. Since the BERT model is 12 layers deep, we use the final hidden layer output as the word embeddings for the questions. Finally, we use zero padding on the embeddings to ensure they have a uniform dimension producing embeddings that have a dimension of 15x768. Padding is done to expediate the training process later on.

C. Merging

On the final step, we have to merge the vector representations of the text and visual data to feed into fully connected layers. We merge the two matrices by concatenating them.

4 Dataset

We created our own procedurally generated dataset which contains 2D images colored images. Images consisted of 3 geometric shapes - rectangles, triangles, and circles in 8 different colors resulting in 24 categories of images. In the training split and testing splits each category consisted of 170 and 45 images respectively of randomized size and location on the image. The training split contained approximately 4000 images while the testing split contained approximately 1000 images.

The questions for the dataset were also procedu-

rally generated with each image having 6 questions associated with it of a specified format. They are as follows (although they are provided in English here, the original dataset consists of question purely in Bangla):

- Does the image contain a {shape}?
- Does the image contain a {color} {shape}?
- Is the shape {color}?
- What is the color of the shape?
- What shape does the image contain?
- What does the image contain?

Randomly chosen shapes and colors were chosen to replace the shape and color in the questions above to produce the final questions.

In the future, we will aim out efforts at developing a dataset containing more complex images and questions.

References

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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Akira Fukui, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell, and Marcus Rohrbach. 2016. 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. Association for Computational Linguistics.

Peng Gao, Hongsheng Li, Shuang Li, Pan Lu, Yikang Li, Steven C. H. Hoi, and Xiaogang Wang. 2018. Question-guided hybrid convolution for visual question answering. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part I*, page 485–501, Berlin, Heidelberg. Springer-Verlag.

Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. 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. Curran Associates Inc.