Multimodal Visual Question Answering Using VizWiz Data; A Visual Assistant for the Blind

B. Sreedha and Prashant R. Nair(✉)

Department of Computer Science and Engineering, Amrita School of Engineering, Amrita Vishwa Vidyapeetham, Coimbatore, India
cb.en.p2aid20040@cb.students.amrita.edu, prashant@amrita.edu

Abstract. Visual Question Answering (VQA) is a Multimodal Interaction task between two domains, Computer vision and Natural Language Processing. The task is to develop a system that could answer a question related to an image. The task is easy for a human being, but it is difficult for a system as it involves multiple reasoning over the Image and question to get the most accurate answer. the basic step involved is to extract the image features using a Convolutional Neural Network and the question features extracted via a Recurrent Neural Network. features so extracted undergo multiple reasoning to get the most accurate answer. VQA can act as a visual assistant for the blind to know about the beautiful world around them. This work is basically to develop a VQA assistant for the blind using a real-world dataset known as VizWiz dataset originating from the blind.

Keywords: Visual Question Answering · VizWiz

1 Introduction

Artificial Intelligence (AI) finds its application in every domain across the globe. It is improving the quality of life of people. When it comes to health care, it provides useful insights and innovative solutions to the problems through medical data analysis.

VQA is an AI approach to help blind to know what is happening around them. It can serve as the best visual assistant to the blind. VQA is a combination of computer vision, natural language processing, and knowledge representation and reasoning implemented via deep learning [10, 19]. This can act as an eye for the blind to see the beautiful world around them without the need for human assistance. VQA models should be carefully designed with maximum accuracy as any improper prediction may result in misunderstandings especially while dealing with the blind. real world solutions need a model to be trained on a real-world dataset.

VizWiz is the first real-world dataset originating from blind people where photos are taken by blind photographers and questions were asked based on the photo. In this paper, an innovative deep learning [11] solution is proposed on the VizWiz dataset that will serve as a visual assistant for improving the quality of life of the blind.

S. Mekhilef et al. (Eds.): ICEEE 2022, LNEE 893, pp. 365–372, 2022.
https://doi.org/10.1007/978-981-19-1742-4_30

2 Motivation

Blind people always need external help to know about the things happening around them. however, relying on humans will not be suitable in certain situations. The people may not be available at the time of need, the instantaneous reply cannot be obtained. Moreover, they have to compromise their privacy to get things done. Building an Artificial Intelligent System could greatly help these groups of people, enabling them to get instantaneous answers without compromising their privacy. they no longer need to rely on other people to get their things done.

According to a World Health Organization report, around 40 million people across the globe are blind and 250 million people have visual Impairment. One of the WHO's key areas of work and activities in helping the blind is to support the development and implementation of tools that can assist blind people. VQA could serve the purpose of assisting the blind, hence developing a model, that can accurately answer the question can improve the life of billions of people across the globe.

3 Related Works

Malinowski and Fritz [8] proposed an image question answering model in which the image is analyzed via a Convolutional Neural Network (CNN) [13] and the question together with the visual representation is fed into a Long Short Term Memory (LSTM) [4] network to predict the answer.

One of the earlier attempts in VQA was done by Zhou et al. [14]. They used naive bag-of-words as the text feature and used the deep features from GoogleNet as the visual features. The combined feature is sent to the softmax layer to predict the answer class. Ren et al. [15] introduced a VIS+LSTM model, where the task is considered as a classification task [12].

Antol et al. Introduced the VQA1.0 dataset where questions related to images are asked and is answered by 10 crowdsourced workers. VQA challenge is held every year based on this dataset. At some of the early attempts, textual and images are concatenated together with point-wise multiplication and passed onto a fully-connected classifier. attention mechanism in the model was first introduced by Lu et al. they used a co-attention [17] mechanism that gives equal importance to image and textual features and jointly reasoned the two modalities. Yang et al. introduced the Stacked attention Mechanism where multiple reasoning is done over the visual features based on the question to get the right answer.

The first real-world VQA dataset known as the VizWiz dataset for helping blind people were introduced by Gurari et al., where the dataset is created using the VizWiz application available in android and apple phones [21].

Vahid Kazemi and Ali Elqruish proposed a model [2] in which the image features were extracted with pretrained Resnet 152 and the question are embedded with LSTM and it is passed through a stacked attention mechanism to get the output in the VQA1.0 dataset. the reason for the choice of Resnet152 model is that it is pretrained in VQA1.0 dataset.

Anderson et al. [18] proposed a model that uses a pre-trained object detection model (Faster R-CNN) to generate arbitrary regions, where attention is to be estimated. In the bottom-up phase approach, attended images features obtained via Region of Interest (ROI) pooling are combined with the question features obtained from Gated Recurrent Unit (GRU).

In VQA the question may be related to only a certain region of the image, so object detection helps detect the object. but if the images lack quality, it will not be able to detect the Region of Interest (RoI) as the object detection model [7, 20] will be able to recognize the object if the probability of getting detected is greater than the minimum threshold. moreover, it is not able to finetune if the image does not contain tags.

4 Dataset

VQA system could serve as a visual assistant [10] for the blind only if it is trained in the images and questions that originate from the blind. VizWiz is the first real-world VQA dataset introduced by Gurari et al. [1] for helping the blind. This dataset originates from the blind photographers using the VizWiz application that is available on iPhone and Android Phones. The user can take photos and can record the question related to the image. the authors have filtered out those images that contain private information or personally identifying information.

VizWiz Grand Challenge, A VQA Challenge is held every year based on this dataset. for the challenge, the question of the training images is answered by 10 Amazon Mechanical Turk Workers (AMT) workers. for each question, the most frequent answer from the 10 Answers is considered as the ground truth. VizWiz is different from other artificially created VQA datasets as it originates from a real-world setting.

4.1 Challenges in the Dataset

High Uncertainty of Answers: Different from other Artificial VQA Dataset, there is a high disagreement between the annotators, as the images are often blurry or the question may be not be related to the question. since these images are captured by blind photographers, they cannot confirm the quality of the image. Often, they may ask a question, that may not be relevant to the Image.

Conversational Question: In other VQA datasets, the question maybe created with the help of annotation workers. VizWiz is a real dataset from the blind. The questions about the image are recorded with the application. The user may use the salutation words such as hello, thank you, etc. so it is necessary to remove salutations and punctuation from questions.

Relatively Small Size: VizWiz data originate from real-world settings, hence it is difficult to collect data. some data has to be filtered as some images may contain any privacy issue or reveal the identity of the user. the size of VizWiz data is relatively smaller compared to other VQA datasets created from artificial settings.

The Imbalance Between Answerable and Unanswerable Classes: In the VizWiz dataset, some images may lack quality or the questions may not be relevant to the image. So, the number of unanswerable questions is quite high in the VizWiz dataset.

there might also occur high disagreement between the annotators due to this lack of quality issues.

5 Methodology

The goal of the work is to extract the useful information from input images and questions. The extracted features are then passed onto a stacked attention network where the multiple reasoning over the image features based on the question. The output attended image features are combined with the question features and passed onto the classifier to get the answer as shown in Fig. 1.

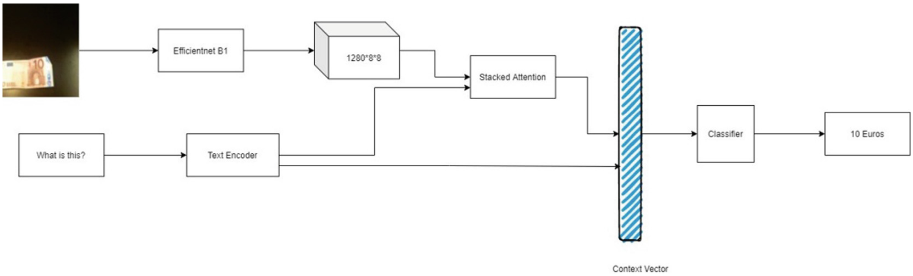

Fig. 1. VQA model

Image Preprocessing: In the preprocessing stage, the images are resized to 240 * 240 dimension, and they are center cropped. input channels of the image are normalized by subtracting the mean and divided by the standard deviation.

Image Feature Extraction: Image features are extracted using a pretrained EfficientNet-B1 [5] model. EfficientNet-B1 model is a convolutional network architecture that uniformly scales in all dimension such as depth, width and compound solution using a compound coefficient.

The preprocessed images are passed to pretrained EfficientNet b1 model. The output of the model is a feature map of 1280 * 8 * 8 where 1028 represent the dimension and 64(8 * 8) represent the number of Image region that corresponds to the 30 * 30-pixel region of the input image.

Question Preprocessing: Questions are made to lower case and punctuation is removed. Questions are tokenized into sequence of words. Word indices are formed from words. If some words are not present in the dictionary, then it is replaced with UNK token.

Question Feature Extraction: Question features is extracted with the text encoder module. Module is composed of 2 units. embedding layer and RNN module.

Embedding Layer: Texts are passed onto the embedding layer. A word embedding is a vector representation of a word that is semantically significant. Instead of representing words in a one-dimensional space per word (one-hot encoding), a dense representation that preserves semantic linkages is used. Word2Vec is a word embedding model introduced by Begino et al. with the goal of learning word representations that can be used to predict the surrounding words in a document. word2vec embedding is used here. input to the Embedding layer module is a list of indices, and the output is the corresponding word embeddings.

Recurrent Neural Network (RNN): Word embedding is passed to an RNN layer to get the semantic representation of the question. RNN will capture the fixed embeddings from text sequences of variable length. Long Short-Term Memory (LSTM) is used as the text encoder. LSTM can get the semantic representation of words in the text. the last hidden state output of LSTM is taken as the question feature vector.

Stacked Attention Network: In most cases, the question may be related to a smaller or specific part of the region. the attention or focus is to be given to the specific part of the image. attention mechanism can help to focus on a specific part of the image. In this work, a stacked attention mechanism model [3] is employed. In Stacked Attention Mechanism (SAN), the semantic representation of a question, to search the image regions that are related to answers. here the image is queried multiple times based on the question, to infer an answer.

$$\begin{aligned} a &= \mathrm{ReLU}\big(W_I x_I \oplus \big(W_Q x_Q + b_Q\big)\big) \\ p_i &= \mathrm{softmax}(W_H a + b_A) \end{aligned} \tag{1}$$

The image feature matrix and question vector are passed onto a single layer of neural network and then to a softmax activation to compute the attention distribution over the images. where W_I, W_Q and W_H are the learning parameters and p_i corresponds to the attention distribution over 64(8 * 8) regions. $x_I \in \mathbb{R}^{dxn}$ corresponds to the image feature vector where d represent the image dimension and n represent the number of image region.$\oplus$ represent the element wise addition of image and question vector. The weighted sum of image vector is obtained based on the attention distribution from the different region of the image. It is then combined with the question vector to form the final query vector. The final vector f contains both visual and question informations. It is then passed onto the classifier.

$$\begin{aligned} \tilde{x}_I &= \sum\nolimits_i^m p_i x_i \\ f &= \tilde{x}_I + x_Q \end{aligned} \tag{2}$$

Classifier: In this work, a two-layer classifier is employed where the first layer is a fully-connected layer with 1024 dimensions with relu [6] nonlinearity and in the second fully connected layer with a size of N, where N corresponds to the most frequent N answer classes.

6 Evaluation Metrics

Each image is associated with a question. the question is answered by ten Amazon Mechanical Turk (AMT) workers. Therefore, corresponding to each image, question pair there are 10 answers. The most frequent answer from the 10 answers is taken as the ground truth answer for each image, question pair for training the model. the best model is saved during training. the model obtained is used to predict the answer for 8000 test images, question pairs. the JSON file containing the predicted answer is submitted to the EvalAI [9] server to get the prediction accuracy of the model in test data.

$$acc(\text{predict}) = \min\left(\frac{\sum_i^{10} \|\text{predict} = \text{answer}_i\|}{3}, 1\right) \tag{3}$$

the evaluation metrics is introduced by Antol et al. [16] for the VQA challenge. the predicted answer is considered correct if at least 3 out of the 10 annotators have given the same predicted answer.

7 Results

We have evaluated the result in test standard set of 2020 VizWiz grand challenge. Results and comparison of the models are shown in Table 1.

Table 1. A comparison of models

Model	Overall	Other	Unanswerable	Yes or No	Number
Show ask attend and answer model	48.43%	35.3%	81.2%	59.6%	18.7%
Our model	49.69%	36.94%	81.57%	59.5%	23.04%

EfficientNet-B1 is 7.6× smaller and. 5.7× faster than ResNet-152 showed an improvement in the overall accuracy in VizWiz dataset by 1.26%. EfficientNet b1 has 7.8 million parameters while the resent152 model has 60 million Parameters. It is difficult to deploy huge models in edge devices. Smaller models are always preferred over the big models in the deployment stage.

8 Future Scope and Conclusion

The main focus of the research was to improve the result of the existing model by bringing about changes in the existing model. Here, we concentrated on improving the image feature extraction part through a robust, fast and small model. This could bring about an improvement in the overall accuracy of the model by 1.26%. We hope the model performance can be further improved by bringing about changes in the text extraction part.

References

1. Li, G.D., et al.: VizWiz grand challenge: answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3608–3617 (2018). https://doi.org/10.1109/cvpr.2018.00380
2. Kazemi, V., Elqursh, A.: Show, ask, attend, and answer: a strong baseline for visual question answering. arXiv preprint: arXiv:1704.03162 (2017)
3. Yang, Z., He, X., Gao, J., Li, D., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
5. Tan, M., Le, Q.: Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, pp. 6105–6114. PMLR (2019)
6. Agarap, A.F. Deep learning using Rectified Linear Units (RELU). arXivpreprint: arXiv:1803.08375 (2018)
7. Le, T., Huy, N.T., Le Minh, N.: Integrating transformer into global and residual image feature extractor in visual question answering for blind people. In: 2020 12th International Conference on Knowledge and Systems Engineering (KSE), pp. 31–36. IEEE (2020)
8. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1–9 (2015)
9. Sarath, S., Amudha, J.: Visual question answering models evaluation. In: 2020 International Conference for Emerging Technology (INCET), pp. 1–5. IEEE (2020)
10. Kandoth, A., Arya, N.R., Mohan, P.R., Priya, T.V., Geetha, M.: Dhrishti: a visual aiding system for outdoor environment. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 305–310 (2020). https://doi.org/10.1109/ICCES48766.2020.9137967
11. Babu, R., Naren, V.S., Soman, K.P: Indian car number plate recognition using deep learning. In: 2019 2nd International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT), vol. 1, pp.1269–1272. IEEE (2019)
12. Saiharsha, B., Diwakar, B., Karthika, R., Ganesan, M.: Evaluating performance of deep learning architectures for image classification. In: 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 917–922. IEEE (2020)
13. Aloysius, N., Geetha, M.: A review on deep convolutional neural networks. In: 2017 International Conference on Communication and Signal Processing (ICCSP), pp. 0588–0592. IEEE (2017)
14. Zhou, B., Tian, Y., Sukhbaatar, S., Szlam, A., Fergus, R.: Simple baseline for visual question answering arXivpreprint: arXiv:1512.02167 (2015)
15. Ren, M., Kiros, R., Zemel, R.: Exploring models and data for image question answering. Adv. Neural. Inf. Process. Syst. **28**, 2953–2961 (2015)
16. Malsa, N., et al.: CERTbchain: a step by step approach towards building a blockchain based distributed application for certificate verification system. In: 2021 IEEE 6th International Conference on Computing, Communication and Automation (ICCCA), pp. 800–806 (2021). https://doi.org/10.1109/ICCCA52192.2021.9666311
17. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. Adv. Neural Inf. Process. Syst. **29**, 289–297 (2016)
18. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6077–6086 (2018)

19. Nirmala, R., Thangavel, S.K.: Develop, implement and evaluate a multimodal system for government organization. In: 2017 4th International Conference on Advanced Computing and Communication Systems (ICACCS), pp.1–8. IEEE (2017)
20. Dushi, D.: Using Deep Learning to Answer Visual Questions from Blind People (2019)
21. Vishnu, R., Krishna Prakash, N.: Mobile application-based virtual assistant using deep learning. In: Sivakumar Reddy, V., Kamakshi Prasad, V. (eds.) Soft Computing and Signal Processing. AISC, vol. 1340, pp. 609–617. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-1249-7_57