Visual Question Answering for Visually Impaired

1st Devesh Chandak
IT
Pune Institute of Computer Technology
Pune, India
dschandak@gmail.com

2nd Omkar Deshpande
IT
Pune Institute of Computer Technology
Pune, India
desphandeomkar77@gmail.com

3rd Sanya Varghese
IT
Pune Institute of Computer Technology
Pune, India
varghese.sanya@gmail.com

4th Shubham Mahajan
IT
Pune Institute of Computer Technology
Pune, India
svmahajan8899@gmail.com

***Abstract*—The lack of access to basic visual information like text labels, icons, and colors can cause exasperation and decrease independence for the visually impaired. We thus propose a visual question answering system to mitigate concerns about undesired consequences from today's status quo for blind people, that is, relying on able-bodied humans to answer visual questions. We wish to assist blind people to overcome their daily visual challenges and break down social accessibility barriers.**

In this paper, we discuss in detail the idea of an open ended visual question answering system to assist visually impaired. The task requires an in-depth understanding of visual and language features, finally, we need to assess the relationship between the two modalities to use co-attention. The proposed model is goal-oriented, stressing on the images and questions generated by blind people.

***Index Terms*—Visual Question Answer(VQA), Self-Attention(SA), Guided-Attention(GA), Long short-term memory(LSTM), Modular Co-Attention(MCA), Optical Character Reader(OCR), Bottom-Up and Top-Down(BUTD)**

I. Introduction

Vision-and-language reasoning requires the understanding of visual contents, language semantics, and cross-modal alignments and relationships. Considerable amount of work has been done in both the fields, vision and language. [1] [2]. Despite these distinguished single-modality works, studies for the modality-pair of vision and language, principally, pretraining and fine-tuning are still under developed. The proposed model focuses on learning vision-and-language interactions, especially for representations of a single image and its illustrative sentence.

The aim of the model is to address the following two tasks: (1) predict the answer to a visual question and (2) predict the answerability of visual question. The overall solution can be divided into two parts: (1) data processing, where we reweigh the label according to the confidence and frequency of the ten answers and (2) we structure the questions with extra semantics, adding objects and attributes embedding from object detection to the language part.

In this task we will be focusing on the VizWiz-VQA dataset. It originates from a natural visual question answering setting where blind people each took an image and recorded a spoken question about it, together with ten crowdsourced answers per visual question. Compared to the previous version of this dataset this new version dataset contains more train-validation images with local obfuscation which prevents privacy disclosure.

The evaluation metric followed:

$$Accuracy = min(humans\ that\ answered/3, 1) \quad (1)$$

If minimum 3 humans answered the same answer then the prediction is successful otherwise a partial score is offered. Text, color, counting and object identification parameters were used to evaluate the model.

II. Literature Survey

A. Ensembling Rich Image Features

Recent approaches have introduced the visual attention mechanism into VQA by adaptively learning the attended image features for a given question, and then performing multimodal feature fusion to obtain the accurate prediction [3]. Beyond understanding the visual contents of the image, VQA also requires to fully understand the semantics of the natural language question. Therefore, it is necessary to learn the textual attention for the question and the visual attention for the image simultaneously.

The focus of this approach is Modular Co-Attention(MCA) Layer. The MCA layer is a modular composition of the two basic attention units, i.e., the self-attention (SA) unit and the guided-attention (GA) unit The model uses GloVe + LSTM to extract question features and uses Faster RCNN to extract image features. Self Attention and Guided Attention are applied to build encoder and decoder. It uses MLP to output these features, which are fused together. The loss function is called Multi label BCE loss.

To get precise convolutional features the architecture resizes an image to get FC features from ResNet 152. Bottom up attention to generate image features using ResNet 101. For question representation the paper uses the following steps. In the beginning a question is trimmed to maximum of 14 words. The vectors from pretrained GloVE which generates word embeddings are fed into an LSTM.

Lastly, by ensembling 14 models the paper achieves an accuracy of 56.20.

This model was the runner up of the VizWiz-VQA 2020 challenge

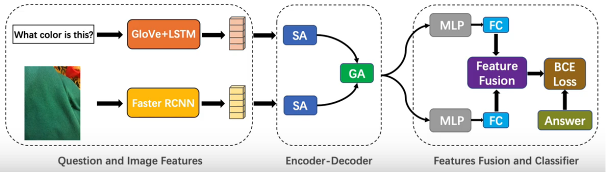

Fig. 1. SUDOKU model

B. Bottom-Up and Top-Down Attention

To achieve the best results the architecture uses the extensively used Top-Down visual attention mechanism. The architecture is well-accepted for image captioning, image description and answering open-ended visual questions(VQA) combined with Bottom-Up attention to capture more salient features [4]. The Bottom-Up mechanism (object detection) puts forward regions of image, each with an associated feature vector, the top-down method determines feature weightings for the feature vector.

The highlights of this architecture are:

1) Image Features: The input image is passed through a Convolutional Neural Network (CNN). Bottom-up attention results in higher performance. The method is based on a ResNet CNN within a Faster R-CNN framework. It is trained to focus on specific elements in the given image

2) Image Attention: The top-down attention, concatenates image features with the question embedding [5]. Establishing a relation between the input text and image later in the architecture.

3) Fusion: The representations of the question and of the image) are passed through non-linear layers and then combined with a simple Hadamard product

The model achieves accuracy of 72 on the VQA dataset. The model was the winner of VQA 2018 challenge

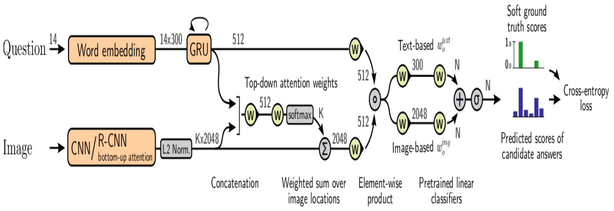

Fig. 2. Bottom-Up Top-Down Attention

III. Proposed Model

We propose a model based on Transformer encoders and novel cross-modality encoder proposed in the paper [14]. The highlight of this architecture is its diverse pre-training tasks and exhaustive dataset. The model firmly adheres to three ideas: Bi-Directional Attention, Transformer, and BUTD [7]. The paper refers to the idea of LXMERT and Oscar inspired addition of semantic words to the language modality. The semantic tags include object and attribute tags. Cross-modality model is built with self-attention and cross-attention layer.

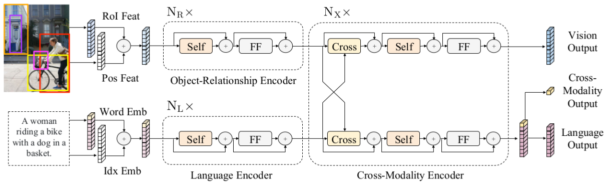

Fig. 3. Proposed Model

1) Encoders: The architecture includes 3 encoders which work mostly on the basis of two kinds of attention layers: Self-Attention layers and Cross-Attention layers. Attention layers aim to extract features from the text-question. It tries to revive context from the question/query. [15].

2) Single Modality Encoders: The language-encoder and image-encoder focus on single modalities. We first separately apply these modalities of the text and the image feature vector.

3) Cross Modality Encoders: Each layer in the Cross-Modality encoder consists of two self-attention sub-layers, one bi-directional cross-attention sub-layer, and two feed-forward sub-layers.

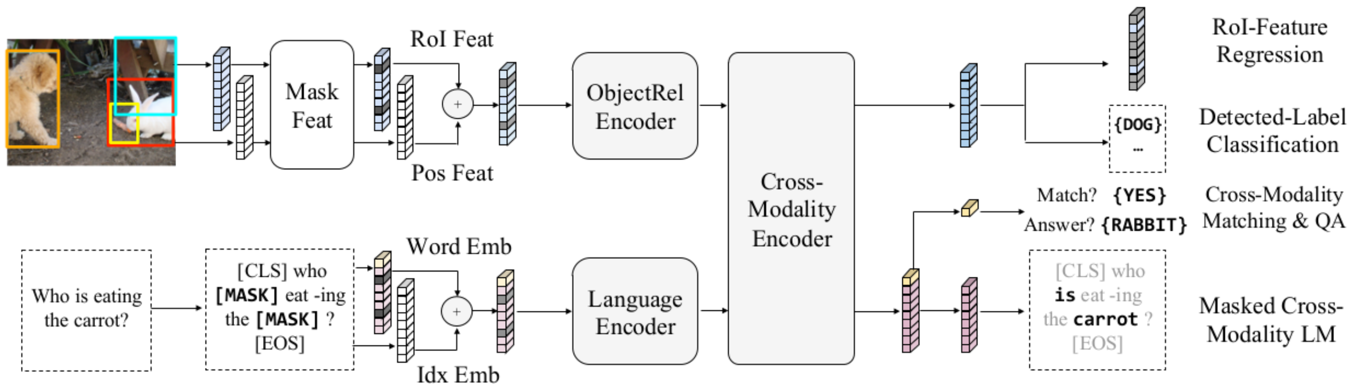

Fig. 4. Overview of the Proposed Model

IV. System Design

The user should be able to click an image with a button, and record their question with a button. To make the application user-friendly, the buttons would be replaced by taps on the screen. The image and the spoken question once fed in to the model would return an answer to the user. The answer would be supported by a speech assistant.The system has been designed to serve as an assisting technology for the blind and visually impaired

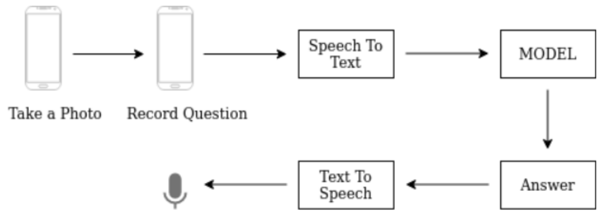

Fig. 5. System Design

V. Future Scope

To improve the architecture an Optical Character Reader(OCR) can be included. It would further increase the viability of the product. With a separate OCR, the app would be able to answer more critical questions and give better results with text based questions.

VI. Conclusion

The objective of the paper was to discuss the system design and methodology to be adopted to design an assistive technology for the blind and visually impaired. We compared different Deep Learning architectures and proposed a real-time application which is user friendly and protects privacy.

REFERENCES

[1] Antol, Stanislaw, et al. ”Vqa: Visual question answering.” Proceedings of the IEEE international conference on computer vision. 2015.

[2] Kazemi, Vahid, and Ali Elqursh. ”Show, ask, attend, and answer: A strong baseline for visual question answering.” arXiv preprint arXiv:1704.03162 (2017).

[3] Anderson, Peter, et al. ”Bottom-up and top-down attention for image captioning and visual question answering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[4] Gurari, Danna, et al. ”Vizwiz grand challenge: Answering visual questions from blind people.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[5] Jiang, Yu, et al. ”Pythia v0. 1: the winning entry to the vqa challenge 2018.” arXiv preprint arXiv:1807.09956 (2018).

[6] Aafaq, Nayyer, et al. ”Video description: A survey of methods, datasets, and evaluation metrics.” ACM Computing Surveys (CSUR) 52.6 (2019): 1-37.

[7] Srivastava, Yash, et al. ”Visual Question Answering using Deep Learning: A Survey and Performance Analysis.” arXiv preprint arXiv:1909.01860 (2019).

[8] Teney, Damien, et al. ”Tips and tricks for visual question answering: Learnings from the 2017 challenge.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[9] Bigham, Jeffrey P., et al. ”VizWiz: nearly real-time answers to visual questions.” Proceedings of the 23nd annual ACM symposium on User interface software and technology. 2010.

[10] Li, Xiujun, et al. ”Oscar: Object-semantics aligned pre-training for vision-language tasks.” European Conference on Computer Vision. Springer, Cham, 2020.

[11] Kafle, Kushal, and Christopher Kanan. ”Answer-type prediction for visual question answering.” Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016.

[12] Shih, Kevin J., Saurabh Singh, and Derek Hoiem. ”Where to look: Focus regions for visual question answering.” Proceedings of the IEEE conference on computer vision and pattern recognition. 2016.

[13] Gurari, Danna, et al. ”Captioning Images Taken by People Who Are Blind.” arXiv preprint arXiv:2002.08565 (2020)

[14] Tan, Hao, and Mohit Bansal. ”Lxmert: Learning cross-modality encoder representations from transformers.” arXiv preprint arXiv:1908.07490 (2019).

[15] Vaswani, Ashish, et al. ”Attention is all you need.” Advances in neural information processing systems. 2017.