

Visual Question Answering for Visually Challenged

43165 - Shubham Mahajan
43212 - Omkar Deshpande
43213 - Devesh Chandak
43253 - Sanya Varghese

Current Group Information

Group Members:

43165 - Shubham Mahajan
43212 - Omkar Deshpande
43213 - Devesh Chandak
43253 - Sanya Varghese

Guide:

Dr. S.C. Dharmadhikari

External Reviewers:

Mr. Jagdish K. Kamble
Ms. Rachana R.Chhajer



Problem Statement

We propose the task of free-form and open-ended Visual Question Answering (VQA) for visually impaired people. Given an image and a natural language question about an image, the task is to provide an accurate natural language answer using visual elements of the image and inference gathered from textual questions. The aim is to mirror real-world scenarios, such as helping the visually impaired or an intelligence analyst.



Abstract

As able-bodied humans, it is easy for us to see an image and answer any question about it using our knowledge. However, there are also scenarios, for instance, a visually-impaired user or an intelligence analyst, where they want to actively elicit visual information given an image. We wish to assist blind people to overcome their daily visual challenges and break down social accessibility barriers. The purpose of the project is to bring sight to blind and low-vision people.



Introduction

Vision-and-language reasoning requires the understanding of visual contents, language semantics, and cross-modal alignments and relationships

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.

Software Tools and Technologies

Software:

- Pytorch
- Tensorflow
- Jupyter Notebook
- React Native/ kivy
- Django Framework
- Google Speech

Hardware:

- Android Phone with Camera
- Cloud Training Hardware (Kaggle/Colab)

Test tools:

- Android Emulator

Human Efforts:

- 4-6 Months (Team of 4)

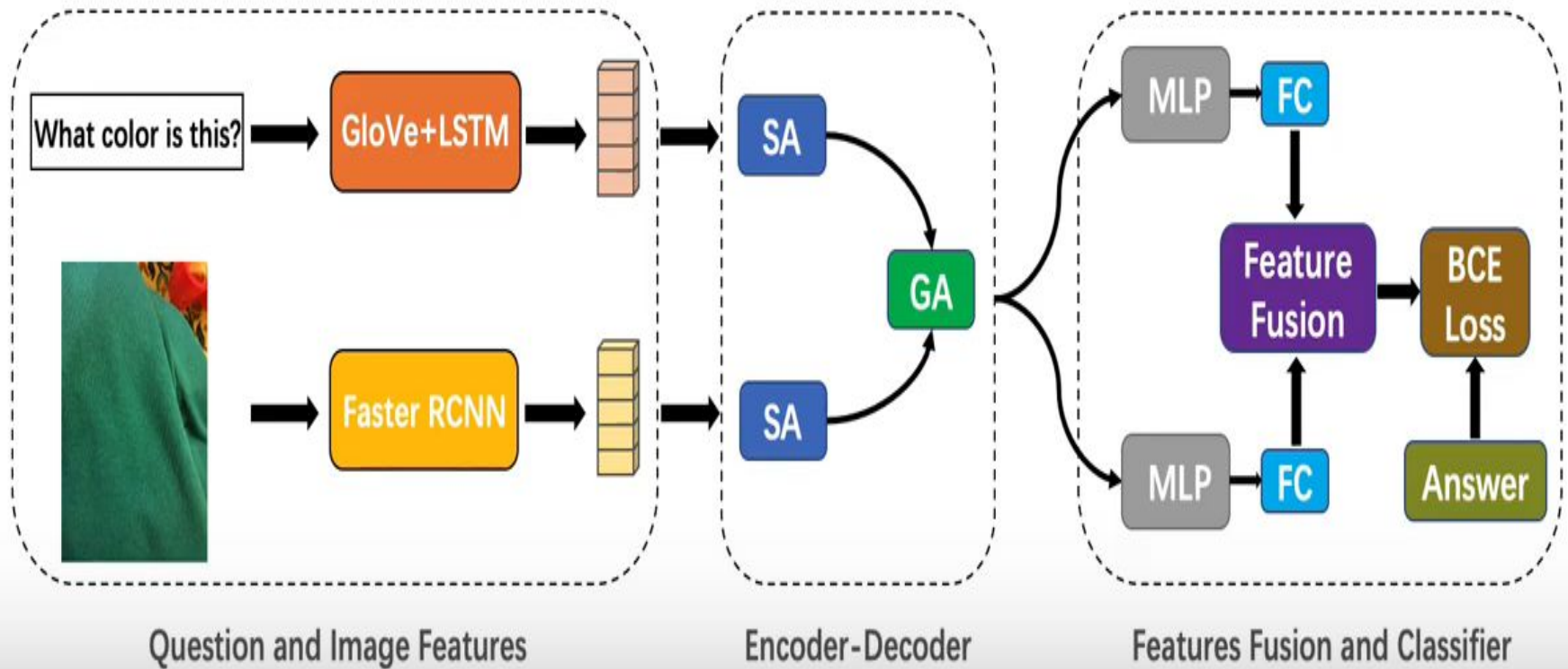


Literature Survey

Ensembling Rich Image Features

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.

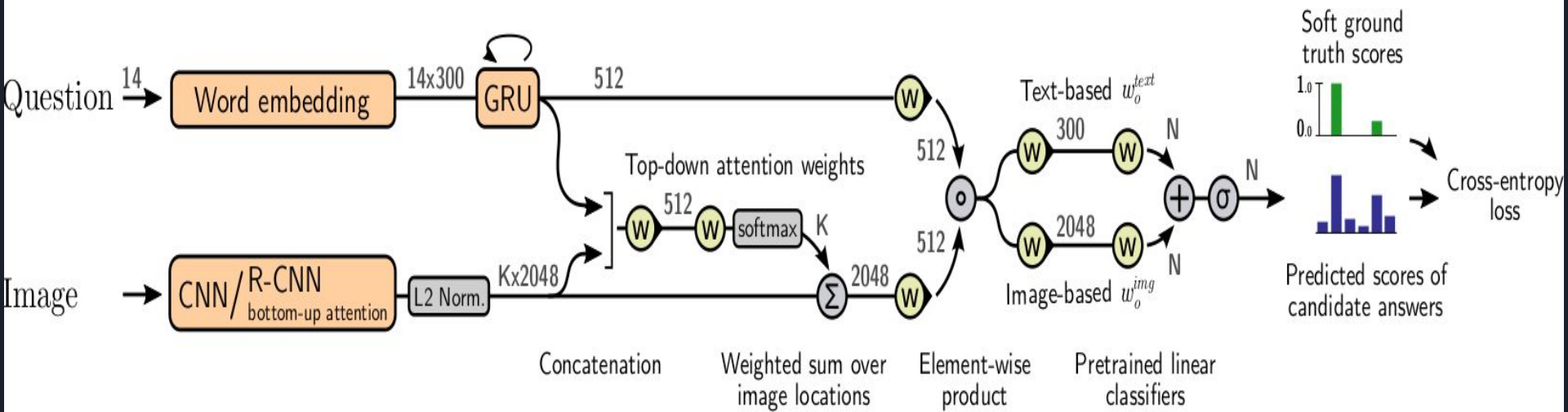




Bottom-Up and Top-Down Attention

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.

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.






Proposed Architecture

We propose a model based on Transformer encoders and novel cross-modality encoder . 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 Bottom - Up and Top - Down

The 3 crucial components of the architecture are:



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.



RoI Feat

Pos Feat

$N_R \times$

Object-Relationship Encoder

$N_X \times$

Cross-Modality Encoder

$N_L \times$

Language Encoder

A woman
riding a bike
with a dog in a
basket.

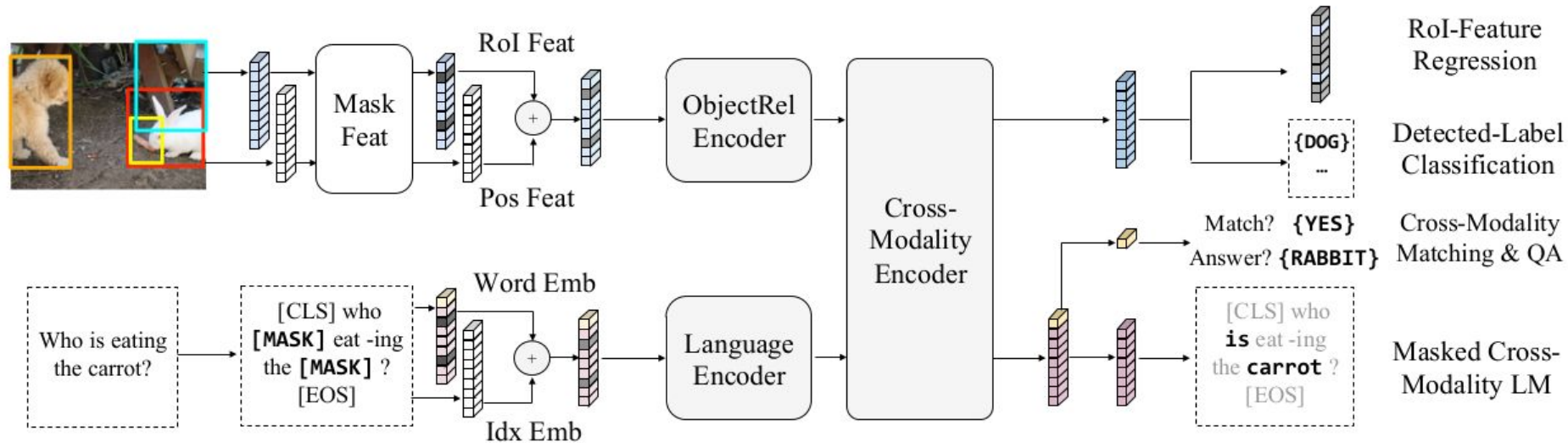
Word Emb

Idx Emb

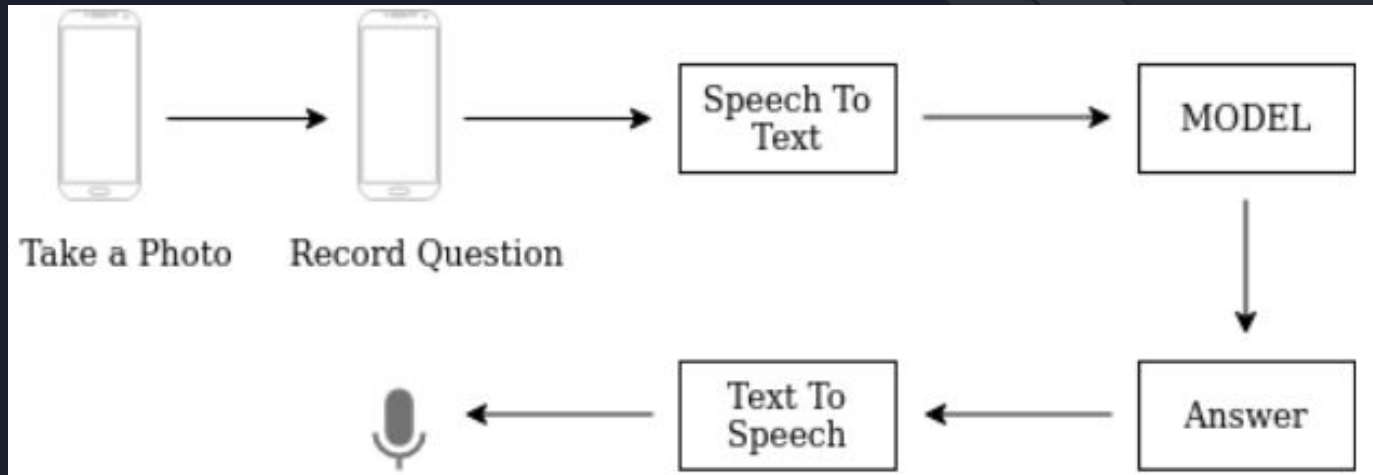
Vision
Output

Cross-
Modality
Output

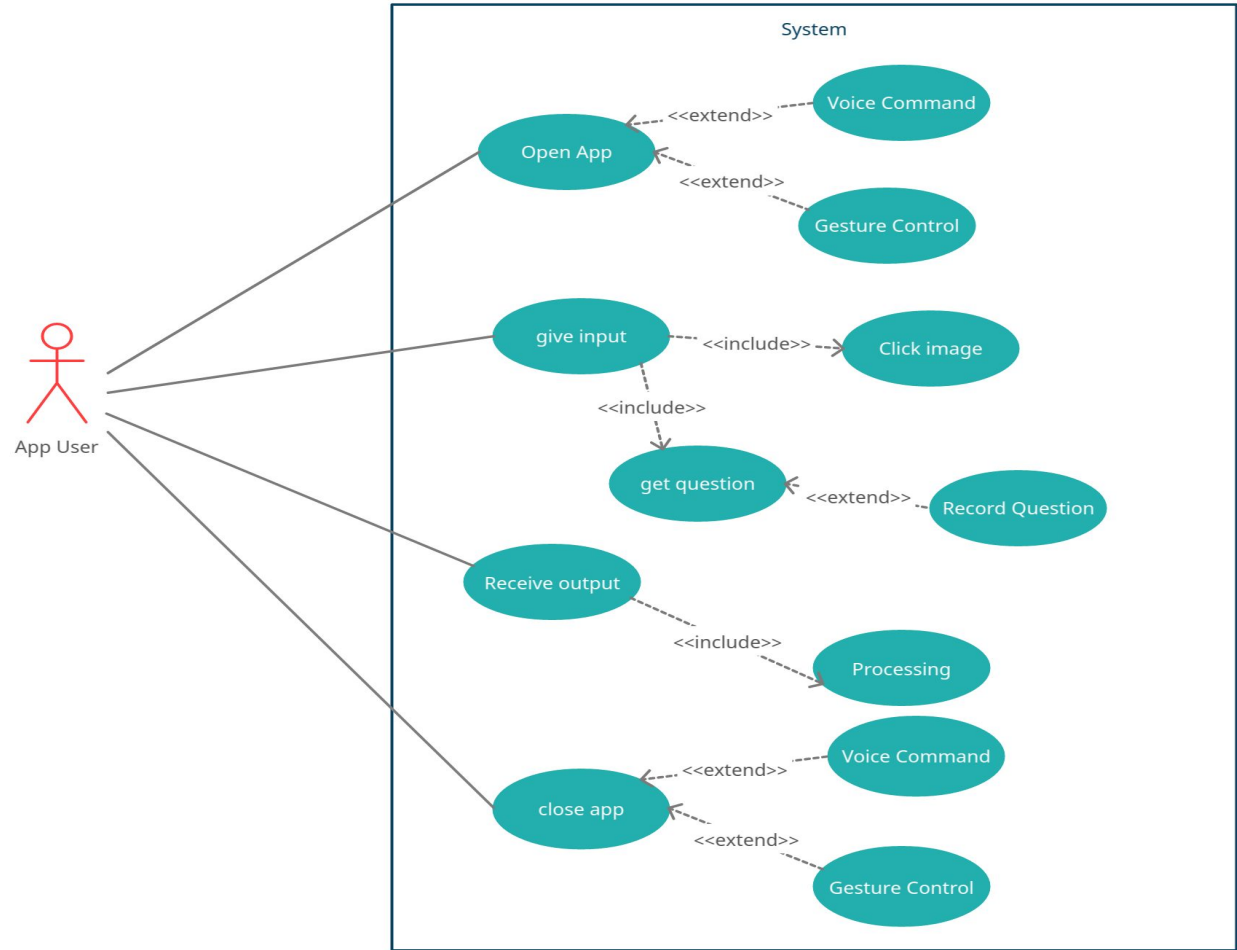
Language
Output



System Design



Use Case Diagram





Proposed Outcome

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.

Gantt Chart

	Week 1	Week 2	Week 3	Week 4	Week 5	Week 6	Week 7	Week 8
<i>Research Algorithms</i>								
<i>Research Frameworks</i>								
<i>Documentation</i>								
<i>Coding</i>								
<i>Bug Fixes</i>								
<i>Result Evaluation</i>								

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." (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."
- [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 23rd 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."

Thank You !

