Visual Question Answering from Multi Models

Kodeti Veerabrahmani
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
A.Durga Devi
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract:

This work proposes an open-ended, free-form task called Visual Question Answering (VQA). The aim for a natural language query regarding the image and a supplied image is to offer an appropriate response in natural language. Answers and questions remain open-ended to reflect real-world situations such as helping the blind. In the various area of an image Visual questions selectively target such as underlying context and background details. Because of this, a system that excels at visual quality assurance (VQA) usually requires an in-depth knowledge of the image and more advanced reasoning than a system that generates generic image descriptions. Furthermore, since many open-ended responses are limited to a few words or a restricted set of responses that may be given in a multiple-choice style, VQA is accessible to computer evaluation. In this project you ask to use ROBERTA model to extract features from questions and answers and then apply BEIT model to extract features from the images. Both features should be fusion in multi-modal to answer for given question and images. To train multi modal you ask to use VQA 2.0 dataset.

Keywords: BEIT, ROBERTA model, VQA.

I. INTRODUCTION

Visual Question Answering (VQA) has emerged as a result of the convergence of computer vision and natural language processing. The goal of Visual Question Answering is to enable machines to understand and react to queries about images or videos in order to reduce the comprehension gap between language and visual perception.

This interdisciplinary approach has far-reaching implications, from enhancing accessibility for visually impaired individuals to enabling advanced human-computer interaction in various applications.

In this paper, we leverage state-of-the-art models to tackle the VQA task, combining the power of language understanding provided by the ROBERTA model with the visual understanding capabilities of the BEIT model. While many questions have a straightforward "yes" or "no" response, figuring out the right response is usually far from simple. Furthermore, since inquiries pertaining to images frequently ask specific questions, one- to three-word answers work well for a lot of questions. In these kinds of situations, counting how many questions a suggested algorithm properly answers makes it simple to assess. The multiple-choice task just asks an algorithm to select an answer from a predetermined list of options, in contrast to the open-answer assignment, requiring for a free-form response.

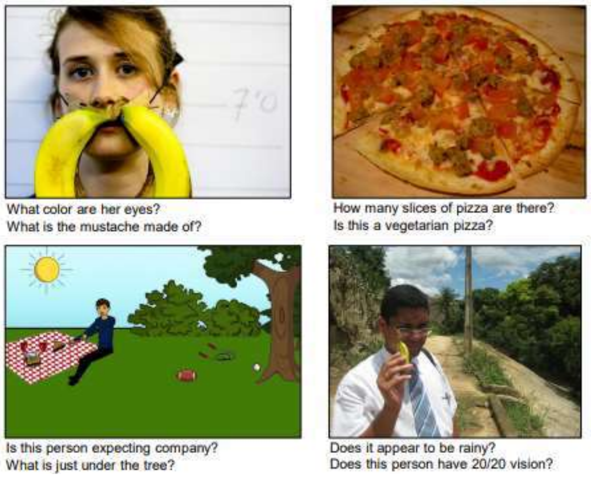

What color are her eyes?
What is the mustache made of?
How many slices of pizza are there?
Is this a vegetarian pizza?
Is this person expecting company?
What is just under the tree?
Does it appear to be rainy?
Does this person have 20/20 vision?

Fig1.1: Examples of free-form, open-ended questions collected for images

In this project, we aim to leverage advanced deep learning models, specifically the ROBERTA model for question and answer understanding and the BEIT (BERT for Image Text) model for visual feature extraction, to create a powerful VQA system. In this project you ask to use ROBERTA model to extract features from questions and answers and then apply BEIT model to extract features from the images. Both features should be fusion in multi-modal to answer for given question and images. To train multi modal you ask to use VQA 2.0 dataset and we used same dataset to train above model.

In order to facilitate research focused solely on the high-level reasoning needed for VQA, we gathered a new dataset of "realistic" abstract settings, eliminating the need to parse actual images. Each image or scene prompted three questions to be answered. Ten individuals responded to each question, including their level of confidence.

Even if there are numerous advantages to using open-ended questions, it's still helpful to know what kinds of questions are asked and what kinds of answers different algorithms might be able to provide. In order to achieve this, we examine the different kinds of queries posed and responses given

We illustrate the amazing diversity of the questions posed using a number of images. We also investigate the ways in which questions and responses differ from image captions in terms of information content. We provide multiple methods for baselines that combine text and cutting-edge visual characteristics [23]. We will host an annual challenge and related workshop as part of the VQA project to talk about cutting edge techniques and industry best practices.

II. LITERATURE SURVEY

VQA is a difficult job that is gaining interest from the NLP and computer vision sectors. To determine the proper response given a image and a natural language question, one must use general knowledge and reasoning skills to analyse the image's visual components. We compare contemporary methods to the problem in order to assess the state of the art in the first section of our survey. We categorise techniques based on how they link the textual and visual modes. Specifically, we study the widely used method of mapping queries and images to a common feature space by fusing convolutional and recurrent neural networks. We also discuss modular and memory-augmented architectures for structured knowledge base interfaces. We examine the datasets that are accessible for VQA system evaluation and training in the second section of our investigation. There are questions at different levels of complexity in each of the datatsets, requiring different kinds of thinking and talents. We thoroughly review the question/answer pairs from the Visual Genome project and assess how useful the structured annotations of scene graph-based images are for VQA. [1]

An algorithm must respond to text-based inquiries concerning images in VQA. Many algorithms have been presented and further datasets have been released since the initial VQA dataset was made available in 2014. We do a critical analysis of the state of VQA at the moment, taking into account the phrasing of problems, available datasets, evaluation metrics, and algorithms. Specifically, we address the shortcomings of existing datasets in terms of their suitability for appropriately training and evaluating VQA algorithms. Next, we thoroughly examine all currently available VQA algorithms. Lastly, we discuss about potential future paths for image understanding and VQA research. [2]

In the fields of NLP and computer vision, VQA has attracted a lot of interest, in part because it provides insight into the connections between two crucial information sources. Existing datasets and the models constructed from them have concentrated on answering questions that can be resolved by analysing the query and image separately. Although somewhat small, the set of such questions that can be answered without the need for outside information is intriguing. It does not include questions that, for example, call for the application of common sense or fundamental factual knowledge. Here, we present FVQA (Fact-based VQA), a VQA dataset that necessitates and encourages considerably more in-depth analysis. The majority of the questions in FVQA require outside data to be answered.[3]

An algorithm is required to respond to text-based questions concerning visuals in visual question answering (VQA). Since late 2014, several datasets for VQA have been developed; they are all flawed, both in terms of content and algorithmic evaluation. Consequently, evaluation ratings are inflated and mostly based on answering simpler questions, which makes comparing various approaches challenging. In this study, we use a new dataset, the Task Driven Image Understanding Challenge (TDIUC), comprising over 1.6 million questions categorised into 12 categories, to analyse existing VQA algorithms. In order to offset the prevalence of certain question categories, we suggest novel assessment schemes that facilitate the analysis of algorithms' advantages and disadvantages. We compare the effectiveness of baseline and cutting-edge VQA models, such as recurrent responding units, neural module networks, and multi-modal compact bilinear pooling (MCB). [4]

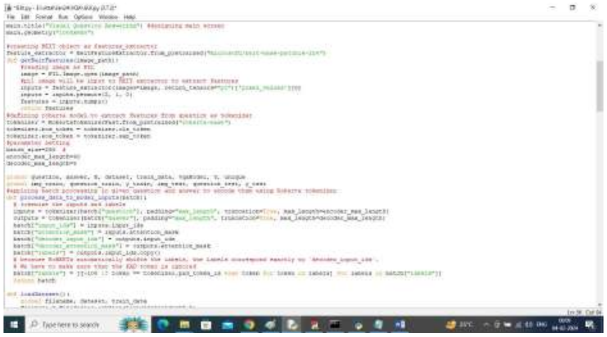
Natural language processing and computer vision are the two main research fields that have given visual question answering (VQA) a significant amount of attention. It has gained widespread acceptance recently as an AI-complete task that can replace the visual testing test. In its most popular form, it is a multi-modal demanding task where the user asks a computer a natural language inquiry regarding an input image and the computer must answer correctly. It draws a lot of deep learning researchers because of its amazing achievements in speech, text, and vision technologies. The present state of VQA research is thoroughly and critically examined in this study, with particular attention paid to datasets, evaluation criteria, and step-by-step solution approaches. Lastly, this study addresses future directions for research on each of the individual VQA components that were previously stated. [5]

We introduce a technique that selects image regions relevant to the text-based query in order to learn how to respond to visual questions. Our approach translates visual features and textual inquiries from different regions into a common space, where an inner product is used to compare them for relevancy. Our method shows notable advantages in addressing queries like "what room," where it can identify informative image portions selectively, and "what colour," where a specific place needs to be evaluated. We evaluate our model using the newly available VQA dataset, which consists of open-ended questions and replies with human annotations. [6]

III. PROPOSED METHOD

In this project you ask to use ROBERTA model to extract features from questions and answers and then apply BEIT model to extract features from the images. Both features should be fusion in multi-modal to answer for given question and images. To train multi modal you ask to use VQA 2.0 dataset and we used same dataset to train above model.

In below screen showing code of extracting features from images using BEIT model



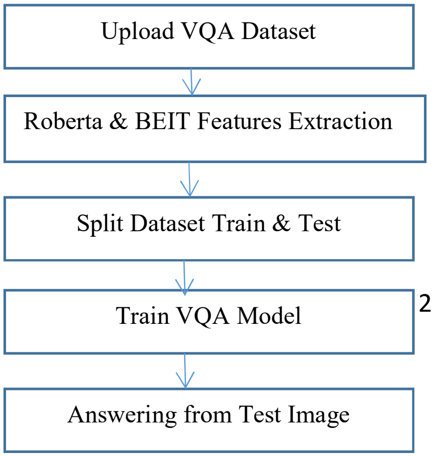
```
main.title("Visual Question Answering") #designing main screen
main.geometry("1300x1200")

#creating BEIT object as features_extractor
feature_extractor = BeitFeatureExtractor.from_pretrained("microsoft/beit-base-patch16-224")
def getBeitFeatures(image_path):
    #reading image as PIL
    image = PIL.Image.open(image_path)
    #pil image will be input to BEIT extractor to extract features
    inputs = feature_extractor(images=image, return_tensors="pt")['pixel_values'][0]
    inputs = inputs.permute(1, 2, 0)
    features = inputs.numpy()
    return features
#defining roberta model to extract features from question as tokenizer
tokenizer = RobertaTokenizerFast.from_pretrained("roberta-base")
tokenizer.bos_token = tokenizer.cls_token
tokenizer.eos_token = tokenizer.sep_token
#parameter setting
batch_size=256 #
encoder_max_length=40
decoder_max_length=8

global question, answer, X, dataset, train_data, vqamodel, Y, unique
global img_train, question_train, y_train, img_test, question_test, y_test
#applying batch processing in given question and answer to encode them using Roberta tokenizer
def process_data_to_model_inputs(batch):
  # tokenize the inputs and labels
  inputs = tokenizer(batch["question"], padding="max_length", truncation=True, max_length=encoder_max_length)
  outputs = tokenizer(batch["answer"], padding="max_length", truncation=True, max_length=decoder_max_length)
  batch["input_ids"] = inputs.input_ids
  batch["attention_mask"] = inputs.attention_mask
  batch["decoder_input_ids"] = outputs.input_ids
  batch["decoder_attention_mask"] = outputs.attention_mask
  batch["labels"] = outputs.input_ids.copy()
  # because RoBERTa automatically shifts the labels, the labels correspond exactly to `decoder_input_ids`.
  # We have to make sure that the PAD token is ignored
  batch["labels"] = [[-100 if token == tokenizer.pad_token_id else token for token in labels] for labels in batch["labels"]]
  return batch

def loadDataset():
    global filename, dataset, train_data
```

Fig.3.1 Dataset

In above screen read red colour comments to know about features extraction from images using BEIT and question features extraction using ROBERTA.

We have designed following modules To implement this project,

1) Upload VQA Dataset: using this module we will upload dataset to application
2) Roberta & BEIT Features Extraction: this module will extract features from image and text question using BERT and BEIT
3) Split Dataset Train & Test: extracted features will be split in to train and test where application will be using 80% images for training and 20% for testing
4) Train VQA Model: training images and questions will be input to VQA fusion model to train a model and this model will be applied on 20% test images to calculate prediction accuracy
5) Answering from Test Image: using this module will enter some question and upload image to get answer for given image and question.



Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Answering from Test Image

Fig3.2 Flowchart for proposed method

IV. RESULT

To run project double click on 'run.bat' file to get below screen

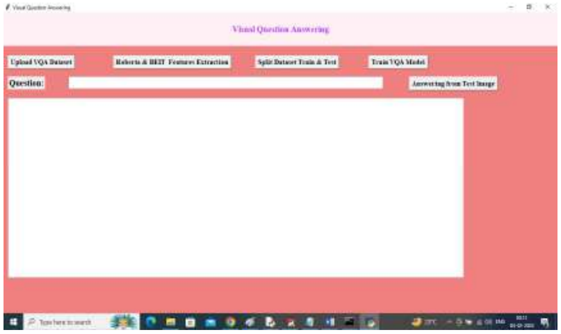

Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

Fig.4.1 Run.bat file

In above screen click on 'Upload VQA Dataset' button to upload dataset and then will get below output

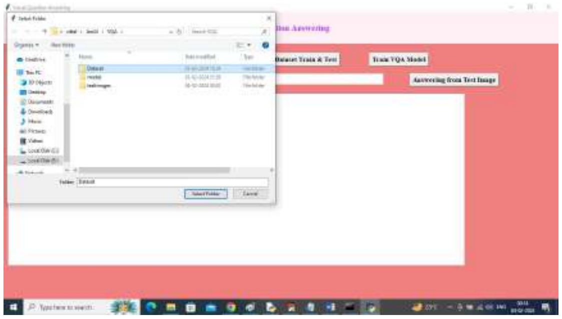
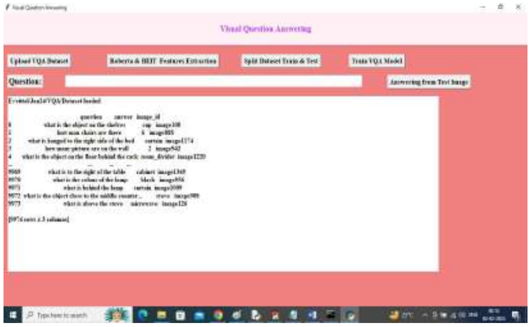


Fig.4.2 Upload VQA Dataset

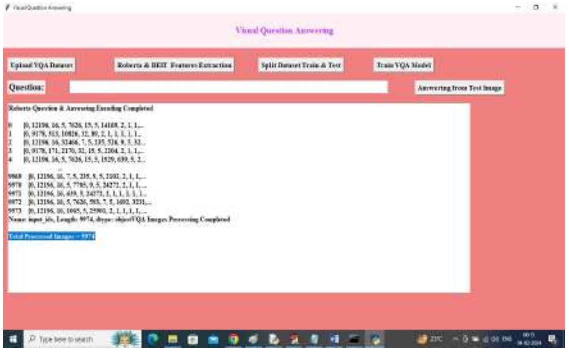
On the above screen, select and upload the entire dataset folder, then click on the 'Select Folder' button to load the dataset, and then you will get the below output,


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

```
E:/vittal/Jan24/VQA/Dataset loaded

                                   question        answer   image_id
0              what is the object on the shelves          cup   image100
1                       how many chairs are there            6   image885
2     what is hanged to the right side of the bed      curtain  image1174
3                 how many picture are on the wall            2   image542
4  what is the object on the floor behind the rack room_divider image1220
...                                            ...          ...        ...
9969              what is to the right of the table      cabinet  image1349
9970                what is the colour of the lamp        black   image956
9971                       what is behind the lamp      curtain  image1009
9972 what is the object close to the middle counter...    stove   image989
9973                       what is above the stove    microwave   image128

[9974 rows x 3 columns]
```

Fig.4.3 questions, answers and images dataset loaded

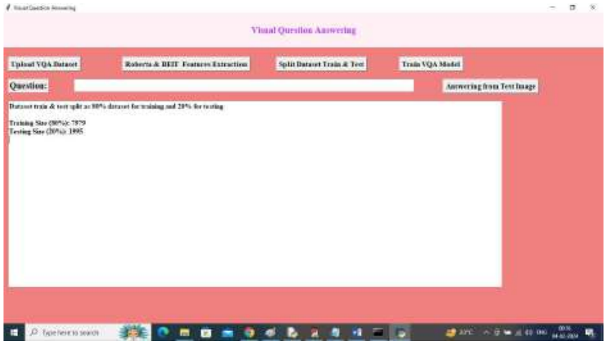
In above screen questions, answers and images dataset loaded and now click on 'Roberta & BEIT Features Extraction' button to extract features from questions and image and then will get below output


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

```
Roberta Question & Answering Encoding Completed

0     [0, 12196, 16, 5, 7626, 15, 5, 14168, 2, 1, 1,...
1     [0, 9178, 313, 10826, 32, 89, 2, 1, 1, 1, 1, 1...
2     [0, 12196, 16, 32466, 7, 5, 235, 526, 9, 5, 32...
3     [0, 9178, 171, 2170, 32, 15, 5, 2204, 2, 1, 1,...
4     [0, 12196, 16, 5, 7626, 15, 5, 1929, 639, 5, 2...
                          ...
9969  [0, 12196, 16, 7, 5, 235, 9, 5, 2103, 2, 1, 1,...
9970  [0, 12196, 16, 5, 7705, 9, 5, 24272, 2, 1, 1,...
9971  [0, 12196, 16, 639, 5, 24273, 2, 1, 1, 1, 1, 1...
9972  [0, 12196, 16, 5, 7626, 593, 7, 5, 1692, 3231,...
9973  [0, 12196, 16, 1065, 5, 25900, 2, 1, 1, 1, 1,...
Name: input_ids, Length: 9974, dtype: objectVQA Images Processing Completed

Total Processed Images = 9974
```

Fig.4.4 all text data converted to numeric BERT vector

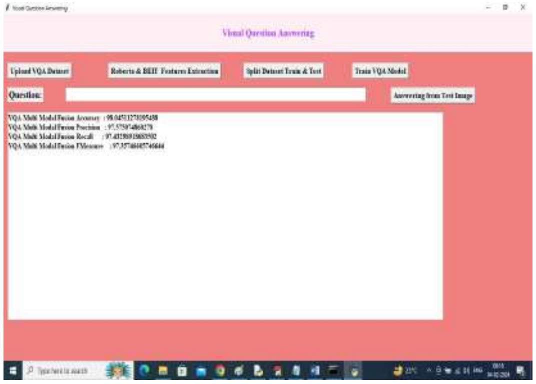
In above screen all text data converted to numeric BERT vector and then can see total images used to extract features and now click on 'Split Dataset Train & Test' button to split dataset and get below output


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

```
Dataset train & test split as 80% dataset for training and 20% for testing

Training Size (80%): 7979
Testing Size (20%): 1995
```

Fig.4.5 Selecting images for training and testing

In above screen can see application using 7979 images for training and 1995 images for testing and now click on 'Run VQA model' button to train model and then will get below output


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

```
VQA Multi Modal Fusion Accuracy  : 98.04511278195488
VQA Multi Modal Fusion Precision  : 97.57587486802
VQA Multi Modal Fusion Recall     : 97.43259915885592
VQA Multi Modal Fusion FMeasure   : 97.35746405746444
```

Fig.4.6 VQA model training completed

In above screen VQA model training completed and it got accuracy as 98% and can see other metrics like precision, recall and FSCORE and now enter some and upload image to get answers

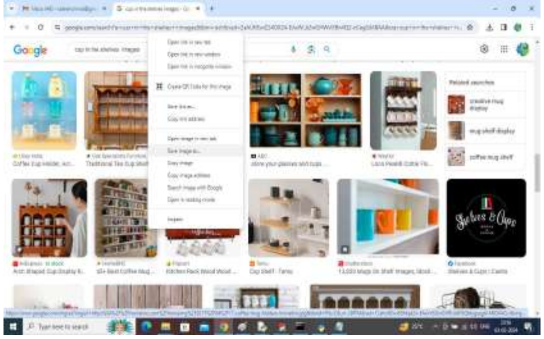
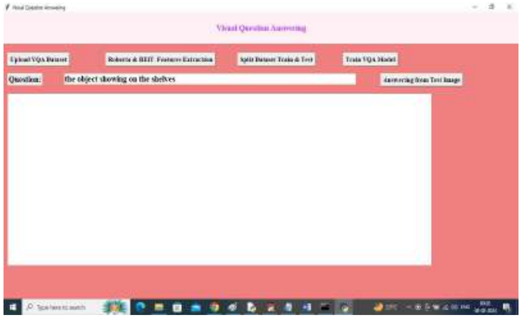


Fig.4.7 downloading some sample images from Google

In above screen from Google downloading some sample images of cup and then will upload to application


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Answering from Test Image

Question: the object showing on the shelves

Fig.4.8 entered some text question

In above screen entered some text question and now upload image

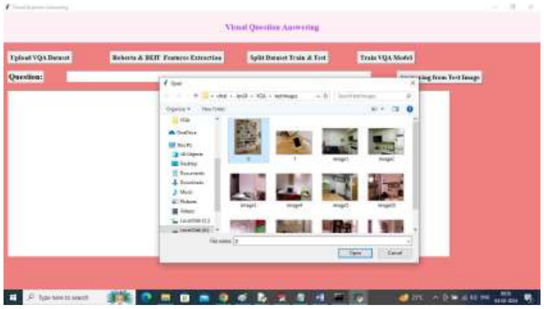

Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image
Open
Cancel

Fig.4.9 uploading image

In above screen uploading image and then click 'Open' button to get below answer

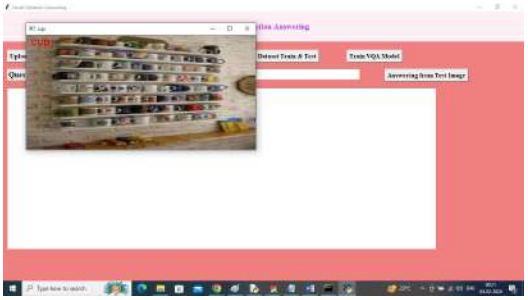
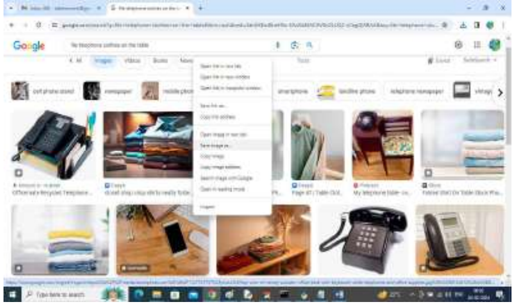
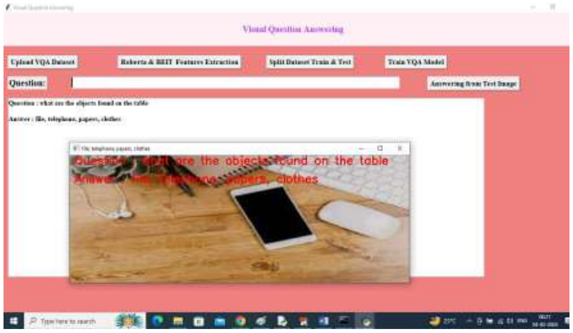


Fig.4.10 got answer as 'cup'

In above screen got answer as 'cup' which is showing in red colour text and below is another example. Now from Google downloading another sample showing in below screen



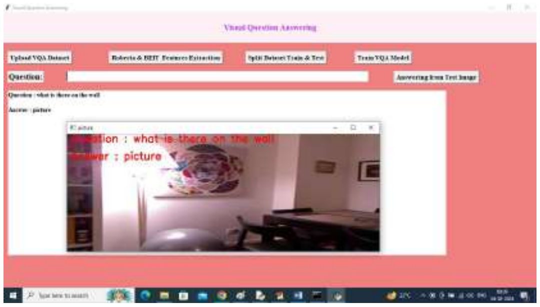

Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

Question : what are the objects found on the table

Answer : file, telephone, papers, clothes

Fig.4.11 question and answer from uploaded other images

In above screen can see question and answer from uploaded other images


Visual Question Answering
Upload VQA Dataset
Roberta & BEIT Features Extraction
Split Dataset Train & Test
Train VQA Model
Question:
Answering from Test Image

Question : what is there on the wall

Answer : picture

Similarly you can upload and test other images

V. CONCLUSION

We introduce the task in conclusion, which is Visual Question Answering (VQA). We use the ROBERTA and BEIT models to develop a powerful VQA system capable of understanding and answering questions about visual content. The aim is to precisely provide a natural language response to an open-ended query regarding an image in natural language. A total of 1995 images were used for testing and 7979 images for training. We show the large range of questions and responses in our dataset, along with the broad range of AI capabilities in natural language processing, computer vision, and common sense reasoning needed to provide accurate answers to these queries. Our human volunteers were asked open-ended questions that were not task-specific. Gathering task-specific questions would be helpful for certain application areas. For example, questions may be collected from visually impaired participants, or they may be domain-specific. It's interesting to observe that generic captions are rarely enough to address these queries. Practical VQA applications may be made possible through training on task-specific datasets.

REFERENCES

1. Wu, Qi, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. "Visual question answering: A survey of methods and datasets." *Computer Vision and Image Understanding* 163 (2017): 21-40.
2. Kafle, Kushal, and Christopher Kanan. "Visual question answering: Datasets, algorithms, and future challenges." *Computer Vision and Image Understanding* 163 (2017): 3-20.
3. Wang, Peng, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. "Fvqa: Fact-based visual question answering." *IEEE transactions on pattern analysis and machine intelligence* 40, no. 10 (2017): 2413-2427.
4. Kafle, Kushal, and Christopher Kanan. "An analysis of visual question answering algorithms." In *Proceedings of the IEEE international conference on computer vision*, pp. 1965-1973. 2017.
5. Manmadhan, Sruthy, and Binsu C. Kovoor. "Visual question answering: a state-of-the-art review." *Artificial Intelligence Review* 53, no. 8 (2020): 5705-5745.
6. Shih, Kevin J., Saurabh Singh, and Derek Hoiem. "Where to look: Focus regions for visual question answering." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4613-4621. 2016.
7. Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. "Vqa: Visual question answering." In *Proceedings of the IEEE international conference on computer vision*, pp. 2425-2433. 2015.
8. S. Antol, C. L. Zitnick, and D. Parikh. Zero-Shot Learning via Visual Abstraction. In ECCV, 2014. 2, 3
9. J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. VizWiz: Nearly Real-time Answers to Visual Questions. In User Interface Software and Technology, 2010. 1, 2, 8
10. K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor. Freebase: A Collaboratively Created Graph Database for Structuring Human Knowledge. In International Conference on Management of Data, 2008. 2
11. A. Carlson, J. Betteridge, B. Kisiel, B. Settles, E. R. H. Jr., and T. M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In AAAI, 2010. 2
12. X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollar, and ´ C. L. Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. arXiv preprint arXiv:1504.00325, 2015. 3
13. X. Chen, A. Shrivastava, and A. Gupta. NEIL: Extracting Visual Knowledge from Web Data. In ICCV, 2013. 2
14. X. Chen and C. L. Zitnick. Mind's Eye: A Recurrent Visual Representation for Image Caption Generation. In CVPR, 2015. 1, 2
15. J. Deng, A. C. Berg, and L. Fei-Fei. Hierarchical Semantic Indexing for Large Scale Image Retrieval. In CVPR, 2011