JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 7, ISSUE 13, 2020

ARTIFICIAL VISION FOR THE VISUALLY IMPAIRED

SIDDHI POTDAR [1*,] SAKSHI GUPTA [2] SIDDHARTH KULKARNI [3] ADITYA KAWALE [4]
PALLAVI DHADE [5]

[1] siddhi.potdar2013@gmail.com, [2] sakshigupta341999@gmail.com, [3] kulksud@gmail.com,
[4]adityamkawale@gmail.com, [5] pallavi.dhade@gmail.com

[1, 2, 3, 4] Student, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

[5] Faculty, Department of Computer Engineering, Pimpri Chinchwad College of Engineering, Pune, India

ABSTRACT: Visual Question Answering (VQA) is a computer vision task where a system is given a textual question about a visual (image) input, and it generates a relevant answer. This approach can be used to devise a system to aid the visually impaired by assisting them in their day-to-day visual tasks. In order to do so the focus areas involved are: 1) Selection of an adequate dataset 2) Promising speech interpretation 3) Effective Object detection 4) Subsume appropriate methodology. We present a survey revolving around the same covering various aspects viz. types, algorithms, techniques and procedure followed in respective areas. Furthermore, we compare them and draw conclusions specific to developing a system for visually impaired.

I. INTRODUCTION

Computer vision is a scientific discipline involving algorithmic basis to analyse, understand and extract useful data from visual data, and NLP or natural language processing involves extraction of useful information from any textual or speech related data. Hence, visual question answering (VQA) can be defined as a multi-discipline problem that in-puts an image and a natural language question to suggest a textual output on the basis of an algorithm. In this sur-vey, we take a look at a few papers aligned with our objectives. The first aspect discussed is the selection of an adequate dataset. We go through well-known datasets to find one more inclined to real-life situations and helps us attain our ultimate goal. The second aspect goes over the area of speech interpretation, which will help us to understand and overcome challenges involved in speech interpretation. The third aspect focuses on object detection methodologies. A state-of-the-art object detection network will improve the efficiency of the algorithm. Finally, the fourth aspect involves tying up all the findings from above three aspects into an efficient methodology to get a result. The algorithm proposed must be able to answer questions asked, prompt in-case of an invalid input and produce appropriate results.

II. LITERATURE SURVEY

A. VQA and Dataset

In order to attain maximized accuracy results it is of due importance that the model is trained on an appropriate dataset. There are various aspects that can be considered in the construction of the dataset. Discussions made in this section are in light of various ways that are used to generate datasets.

1) VQA

Initially, the contribution to VQA dataset included 2 parts, first part consisted of 123,287 training, validation images and 81,434 testing images from [23] COCO-QA, based on MSCOCO dataset. One of the initial datasets that contributed to this research direction was [22] DAQUAR; it was based on real-world images consisting of 5674 testing and 6794 training questions and answers.

2) Complimentary Images

Ideally, datasets are made in no relation amongst involved entities. Mohit B et al. proposed a methodology to enhance dataset by generation of complimentary images.

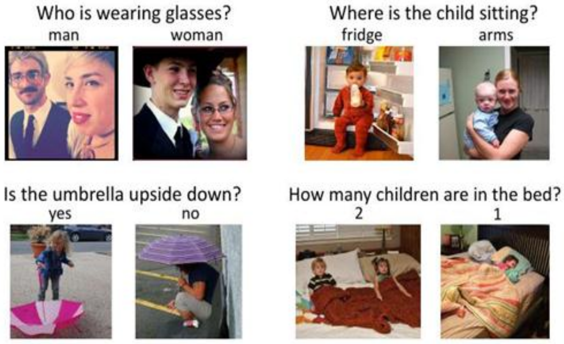
Who is wearing glasses?
man
woman
Where is the child sitting?
fridge
arms
Is the umbrella upside down?
yes
no
How many children are in the bed?
2
1

Figure 1: For every image a given as input(left) a complimentary image(right) is generated. This image is then added to the dataset (as virtual image).

The concept followed is for every image existing in dataset a complimentary virtual image is generated. This increases accuracy of the system in terms of being trained on multiple existing scenarios. Initial attempt to implement this idea [2] consisted of 14,200 initial images in the dataset. The same expanded to 21,004 after few runs on the system. Although the idea was implementable, it gave overhead on the memory with most virtual images hampering the accuracy.

3) Binary Balancing

There exist certain situations wherein the output/answers are clearly discrete YES/NO. Authors in [2] propose dataset which encapsulates this behavior. Formation of the dataset involved a two-step procedure (1) Language Parsing (2) Visual Verification. The methodology proposes that each question is stored in a type (involving keywords) and answers are generated from thus identified keywords.

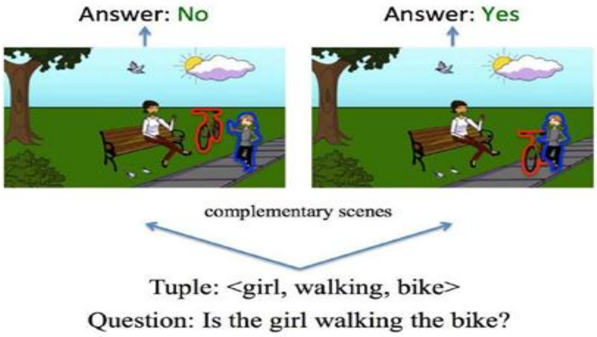
Answer: No
Answer: Yes
complementary scenes
Tuple: <girl, walking, bike>
Question: Is the girl walking the bike?

Figure 2: Binary questions as mentioned, are answered using the method-ology proposed by [2]. Answers are generated with the help of identified keywords stored in tuples.

4) Algorithm Compatible

Oftentimes datasets proposed are not applicable due to the nature and working of varied sub-divisions of an algorithm. Hence [3, 4] propose methodologies that make the model accustomed to the answering mechanism and improve the confidence of answering. Author Peng et al. proposes the idea of a Neural Network capable of learning an algorithm using a co-attention mechanism that looks at the questions, image and facts jointly and gives ranked answers as output. Author Stanislaw et al. [4] proposed a LSTM-RNNs neural network architecture which is used to devise whether the question asked is relevant to the image or not. It proposes a dataset of 121512 visual questions and 9952 non-visual questions.

5) VizWiz

In the previous sections, the ideology behind collection of images to build up the dataset involved co-workers clicking the images and answering to the same. VizWiz dataset involves change in this traditional procedure. In order to build up the VizWiz dataset, an application was deployed that asks the visually impaired or the ultimate end-user to click images. Thus, it was able to captured various aspects that weren't considered by previously proposed datasets for VQA. 72,205 visual questions were collected over a tenure of 4 years. These visual questions served as the starting point for development of the dataset. VizWiz dataset consists of over 31,000 visual questions originating from blind people who each took a picture and recorded a spoken question about it, together with 10 crowd-sourced answers per visual question. In retrospect 2 steps were followed (1) Anonymization and (2) Filtering. In order to maintain confidentiality of the user the question (speech input) was transcribed and spell checked. After this it was given to the crowd-workers in order to generate answers for the same. Filtering of questions was done on following key points (1) Personally-identifying information (PII) (2) Location (3) Complex Scene (4) Low quality images. For every question thus asked there were 10 answers collected from distinct crowd-workers, if in any situation due to bad quality of images, lighting, bad focus the image was inevitable, "unanswerable" option was to be selected. Finally, the analysis of answers was done by: word map, string match and statistical method.

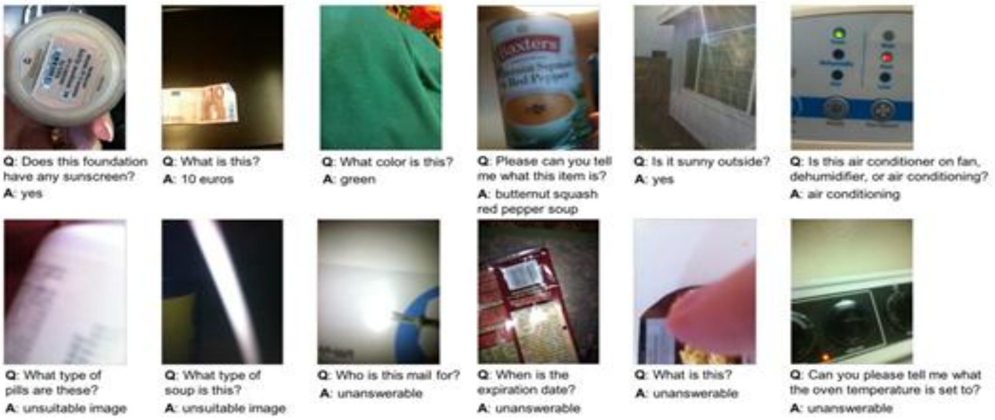
Q: Does this foundation have any sunscreen? A: yes
Q: What is this? A: 10 euros
Q: What color is this? A: green
Q: Please can you tell me what this item is? A: butternut squash red pepper soup
Q: Is it sunny outside? A: yes
Q: Is this air conditioner on fan, dehumidifier, or air conditioning? A: air conditioning
Q: What type of pills are these? A: unsuitable image
Q: What type of soup is this? A: unsuitable image
Q: Who is this mail for? A: unanswerable
Q: When is the expiration date? A: unanswerable
Q: What is this? A: unanswerable
Q: Can you please tell me what the oven temperature is set to? A: unanswerable

Figure 3: The Vizwiz Dataset proposed by [5]. Images when actually clicked by the blind involved number of parameters: light, focus, distance.

B. Speech Interpretation

Speech is the most common means of communication among people. Modelling a relationship between words and their meaning by a computer cannot be done by traditional rule-based approaches, as there exist ambiguities in any language. This certainly gives basis to use of statistical as well as deep-learning based approaches. Steps involved to convert speech into expected output are: (1) Signal processing and Feature extraction (2) Acoustic modelling (3) Language modelling.

1) Signal processing and feature extraction

Speech is an analog quantity, converted to digital to be understood by a computer. Feature extraction takes place which removes noise, surrounding disturbances and under-stands key components like pitch, power etc. According to Usman K et al. several feature extraction techniques include Principal Component Analysis (PCA), Linear Discriminative Analysis (LDA), Linear Predictive Coding (LPC), Mel-Frequency Cepstral Analysis (MFCC) etc. Amongst these MFCC is most commonly used as it directly resonates with the actual human auditory system to give out feature vectors. These vectors when coupled with EBNF grammar files help in forming grammatically correct text.

2) Acoustic Modelling

Acoustic Model helps in integrating relationships between acoustic properties, feature vectors with the phonemes. Which set of feature vectors are suitable for which phoneme is mapped here. Common techniques include GMM (Gaussian Mixture Model), HMM (Hidden Markov Model), DTW (Dynamic Time Warping) etc. GMMs along with EM algorithm which predict parameters for the system are useful for modelling Probability density functions of speech vectors, but fail to capture sequences of information which is important while modelling context related speeches. HMMs are most commonly used for the fact that they can work with sequences and are simple and computationally efficient.

3) Language Modelling

On identification of proper words via AM, it is necessary to model them according to grammatical rules of a language to form sentences. language modelling helps to determine probability of linguistic units: words, sentences. Language Modelling is classified into 2 types: Count based consisting of techniques like Markovian LM and continuous space based which include feed-forward neural probabilistic LM (NPLM), Recurrent neural network (RNN), Long short-term memory (LSTM). Author Alex Sherstinsky et al. elaborates on the same [9]. Count based models rely heavily on statistical methods and exact patterns like word match and sequence matching, which is not efficient considering the con-text of language. To overcome the drawbacks laid by Count based LM, Continuous space LM suggests use of NN with gated memory.

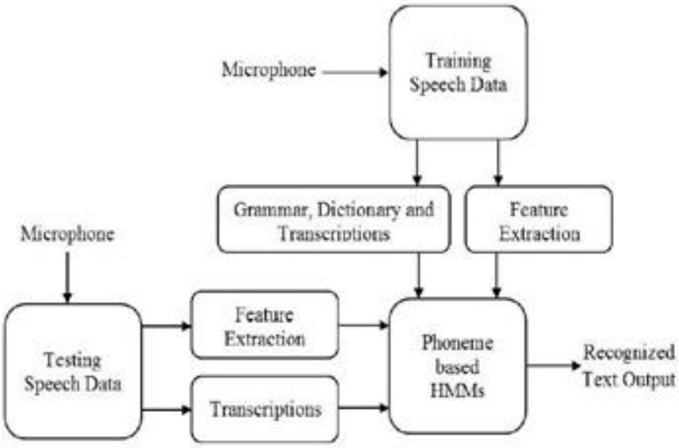

Figure 5: Flow Diagram of Proposed Isolated Word Recognition System

C. Object Detection, Localization and Classification

For any system to detect objects accurately, there is a need for expertise in the areas of Computer Vision. For this, an efficient image enhancement module is needed.

1) Object Detection

Object classification and localization are functions performed in object detection. Object classification differentiates objects from the background and classifies their class (e.g. human and car). Object localization detects objects by drawing a bounding-box around them.[16] proposed a technique to devise a network in which sharing layers are less semantic in object classification and localization. Based on the feature maps of the last sharing layer, two attention networks are evaluated. Attention maps are generated to fit object classification and localization. The proposed method achieved 36.2 AP. It was 4.6 AP and 1.8 AP higher than Faster R-CNN and R-FCN, respectively. [17]proposed an end to end method to detect objects by devising bounding boxes containing the targeted object and it’s position using CNN. It’s a two step process: (1) Produce bounding boxes from training images to generate the positional coordinates of each object (2) Detect and localize objects simultaneously in the captured image during the testing process. This was performed on two datasets: Washington RGB scene and LIMIARF dataset. Their model is based on the core Tensorbox and Google OverFeat framework. Author of [18] proposed a model which is based on the Markov model for active background subtraction to reduce robustness and improve performance of detection and localization. A Gaussian model is implemented in the proposed system to remove background to track an object. The process is performed using avg. filters to extract target objects from moving frames. This creates an initial background from starting frame to end frame, moving objects are extracted using subtraction process to obtain the avg. background. The equation is given as:

$$\Delta Avg_{(a,b)} = |Frame_{(a,b)} - avg_{(a,b)}|$$

Where (a) is the initial frame and (b) is the final frame. Here, the threshold value is set to 50% or above to remove noisy objects using the equation:

$$BT_{(a,b)} = \left\{ \begin{array}{l} 1, and \Delta Avg_{(a,b)} > \text{Threshold} \\ 0, and \text{ Otherwise} \end{array} \right\}$$

Authors of Author Abinaya et al. [19] proposed an improved model to detect texts from the images captured using the CNN. The dataset is fed into the network which is further provided to the training module of CNN. Features of the images are obtained by each layer of the network which is passed on to other layers for further processing.

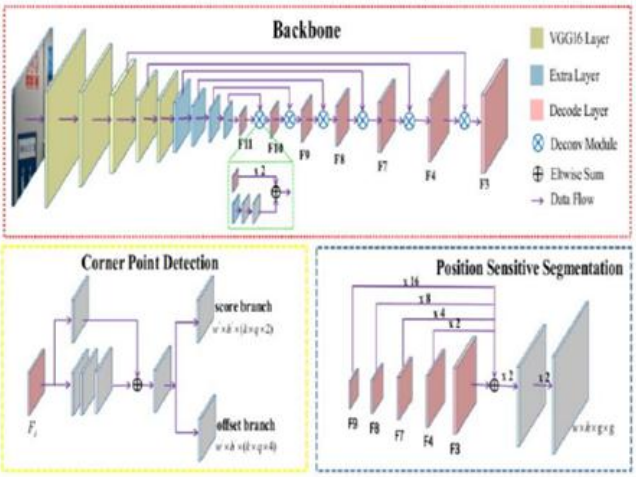

Figure 6: Schematic view of Convolution Neural Network

2) Object Localization and Classification

As a result, the proposed method improves the object detection performance. In article [16] an image is given as an input to the model, the feature maps of the image are encoded in less semantic layers of the backbone network. After this, two attention maps are generated in the attention network for object detection and object classification.

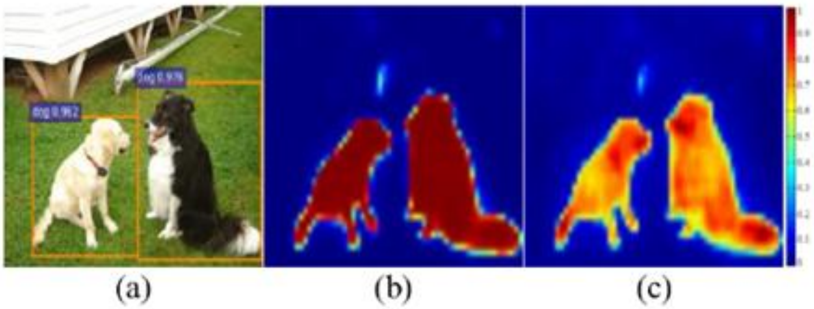

Figure 7: An example depicting generated attention maps in [16]. (b) shows a generated attention map for object localization. (c) shows the generated attention map for object classification, which concentrates on partial areas of the object.

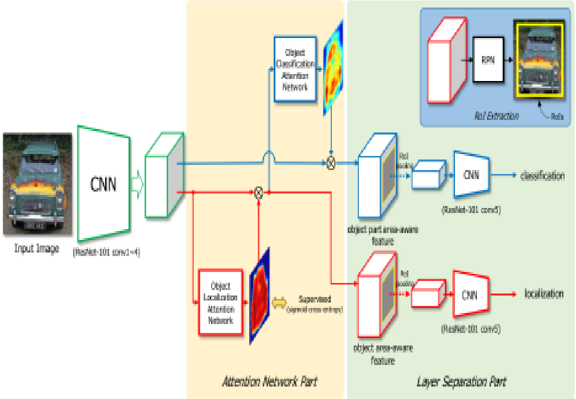
Input Image
CNN
(ResNet-101 conv1~4)
Object Classification Attention Network
Object Localization Attention Network
Supervised (sigmoid cross-entropy)
Attention Network Part
RPN
RoI Extraction
RoIs
RoI pooling
CNN
classification
(ResNet-101 conv5)
object part area-aware feature
localization
object area-aware feature
Layer Separation Part

Figure 8: The figure depicts a system performing object classification and localization through attention networks. Character binary segmentation is used to allow attention maps to localize objects in the image. Object localization attention networks are trained in a supervised manner using a binary segmentation map.

D. Methodology Selection

In this section we go through several recent architectures proposed for performing VQA.

1) Learning by Asking Questions

The authors Misra et al. [10] have introduced an interactive learning framework for the development of VQA systems. LBA differs from typical VQA training as here, instead of training on a fixed large-scale dataset, the system receives only images and decides what questions to ask. The answers to these questions are given by an oracle which may be human supervised. The approach of building an LBA

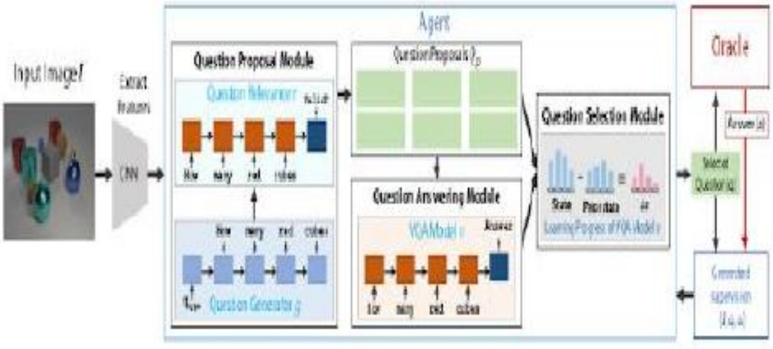

Figure 9: Illustrates the model proposed by the authors in [10]. The model takes an input image, generates questions which are filtered according to relevance and then answered by the oracle.

agent is divided into three modules: (1)a question proposal 4 module, which generates questions using LSTM and checks its relevance (2)a question answering module, which provides answers for a collection of questions Qp and (3)a question selection module, which selects most informative questions Qp based on a selection policy defined by it.

2) End-to-End Module Networks (N2NMN)s

The goal of End-to-End Module Networks (N2NMNs)[11], is to reason instance-specific network layouts without the assistance of a parser to solve visual-answering tasks. This model has two main components: a collection of co-attentive neural modules that lay out parameters for solving sub-tasks and a layout policy that specifies the layout of the assembled task-specific neural network. A neural module is represented as a function:

$$y = f_m(a_1, a_2, a_3, \ldots; x_{vis}, x_{txt}, \theta_m)$$

, where a1, a2, a3,.. are attention maps, xvis; xtxt, are features from visual and textual data inputs and _m is an internal parameter. The layout policy outputs are in the form of a probability distribution p(l|q), where l is the predicted layout for question q to obtain the highest probable network layout for the input question.

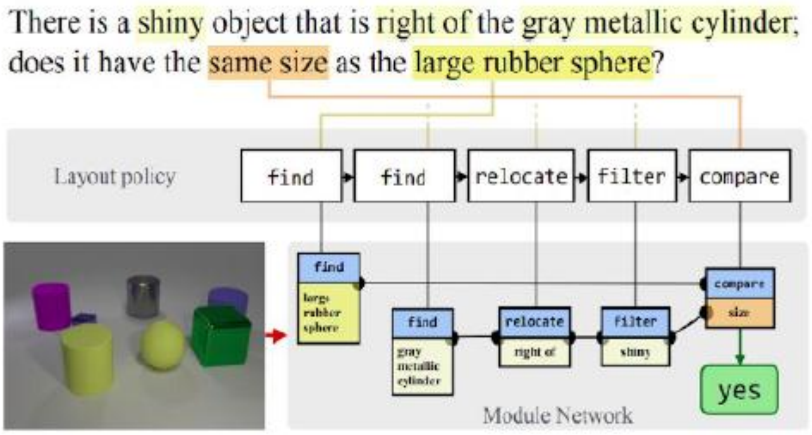

Figure 10: The figure illustrates the layout policy mapped using the N2NMNs to answer the relational questions. [11].

3) A Simple Neural Network Module for Relational Reasoning

Relational reasoning basically means using our intelligence to gather facts, understand relations between them to reach a conclusion. These capabilities of reasoning are baked into an Relational Network (RN) [12] architecture without needing to be learned. This greatly helps in solving relational questions as depicted in Figure 10. The authors of [12] claim RN to be a data efficient method of dealing with relational questions and has a simple architecture.

4) Inferring and Executing Programs for Visual Reasoning

Generally modules that implement VQA attempt to map inputs and outputs using a black-box architecture without modelling the underlying reasoning architecture. This hinders the learning process of the model as it exploits previous generated results. To cumber this, authors of [13] proposed a methodology, consisting of two components, namely: (1)program generator and (2)execution engine. The program generator works on the textual input and generates a program in the form of an LSTM sequence-to-sequence model. The execution engine works on the image input to form an NMN [21]. Finally these are combined to output the VQA result.

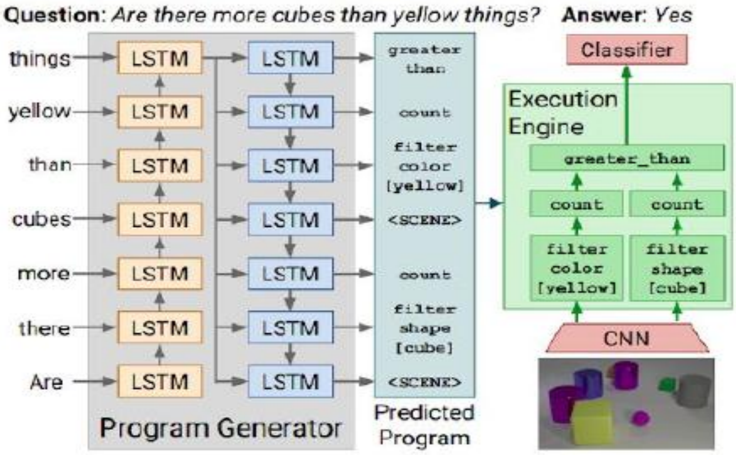

Figure 11: A system overview proposed by authors of [13]

5) FiLM: Visual Reasoning with a General Conditioning Layer

The authors of [14] proposed a methodology that helps solve various multi-input tasks, particularly testing on the image related questioning as it is a challenging multi-input task. Feature-wise Linear Modulation or FiLM combines the processing of image and text with a gain and bias producing a new feature as Fi;c. Each feature here acts as an activation, and hence multiple FiLM layers may be added throughout the architecture. FiLM has achieved state-of-the-art results for various VQA tasks. It is robust and operates in a coherent manner.

6) Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering

Top-down attention has been mainly used to perform image captioning and VQA tasks. The authors in [15] propose a 5 combination of bottom-up and top-down attention mechanisms to enable attention at both object level and other necessary regions in an image. The model takes image features k and question as an input.

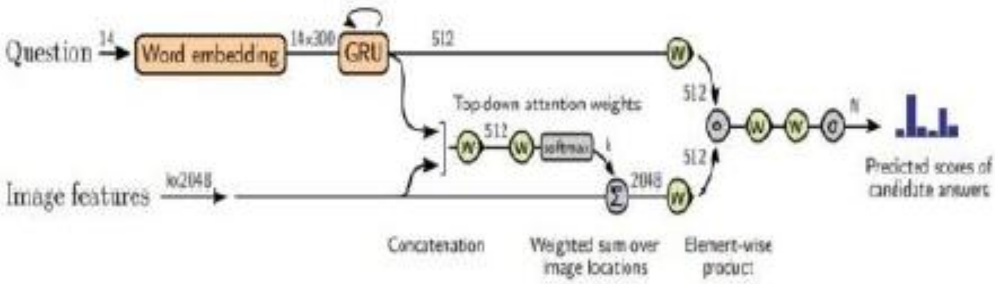

Figure 12: An overview of the model proposed by [15].

A deep neural network implements a joint embedding of the question and image features. Output is generated by a multi-label classifier, over a fixed set answers. Grey numbers indicate the dimensions of the vector representations between layers and yellow elements are learned parameters.

It further converts the question into word embeddings and passes it through a gated recurrent unit (GRU) before concatenating image features with the pre-processed question. On application of top-down attention weights, and further reduction of the matrix into k size, a summation of the matrix values is obtained. An element-wise product of the weighted summation and weighted word embeddings is done to finally obtain a range of predicted answers with their scores. Object detection has been implemented using Faster R-CNN, which is designed to localize objects belonging to certain classes and enclosing them in bounding boxes.

III. RESULTS

The current scenario depicts that in order to capitulate real time working of an efficient system, the VizWiz dataset[5] would help in more accurate, state-of-the-art results as it is made by the user for the user. In terms of speech recognition, MFCC coupled with HMM, in association with LSTM, will yield highly efficient speech recognition. As per the aforementioned object detection, localization and classification algorithms, using the method involving generation of Attention Maps [16] which involves simultaneous object classification and localization bringing forth improved accuracy. The best suited methodology for the implementation VQA, would hence be the bottom-up top-down attention[12], giving state-of-the-art result for VQA tasks. It uses Faster-R-CNN to retrieve the important parts of the image thus improving the focus on the important parts of the image. A 70.34% accuracy is provided by this model, overall on all VQA tasks. A detailed comparison of various datasets and methodologies is given below:

Dataset	Model	Y/N	Num	Unans	Other
Vizwiz	up-down[15]	0.596	0.210	0.805	0.273
VizWiz	vvqa[1]	0.597	0.262	0.805	0.264
MSCOCO	DAN[24]	83.0	39.1	53.9	64.3

Table 1: The following table depicts model accuracies with MSCOCO and VizWiz datasets. All the models use attention mechanisms.

IV. CONCLUSION

In this survey, we looked at various models and techniques which served the purpose to determine a suitable dataset to detect, locate, classify a variety of objects along-side choosing the correct algorithm to incorporate all the studied areas. The comparison of the results based on various models, methods and techniques referred to in this sur-vey is deeply scrutinized in order to get an insight about various approaches involved along with their accuracies and disadvantages which could in turn be used to develop an efficient visual assistance system using VQA.

REFERENCES

[1] M. Bansal, D. Batra, D. Parikh. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. International Journal of Computer Vision. 10.1007/s11263-018-1116-0, 2018. 6

[2] P. Zhang Y. Goyal D. Summers-Stay D. Batra ,D. Parikh. Yin and Yang: Balancing and Answering Bi-nary Visual Questions, 2016 IEEE (CVPR) 2016, pp. 5014-5022. 2

[3] Ray Arijit, Christie Gordon, Bansal Mohit, Ba-tra Dhruv, Parikh Devi. (2016). Question Relevance in VQA: Identifying Non-Visual And False-Premise Questions. 10.18653/v1/D16-1090. 2

[4] A. Agrawal, Jiasen Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Batra, D. Parikh. VQA: Visual Question Answering, arXiv:1505.00468v7 [cs.CL] 2016. 2

[5] Kristen Grauman, Jiebo Luo, Jeffrey P. Bigham, Danna Gurari, Qing Li, Chi Lin. VizWiz Grand Challenge: Answering Visual Questions from Blind Peo-ple, IEEE (CVPR), 2018, pp. 3608-3617. 2, 6

[6] Krupakar Hans, Rajvel Keerthika, Bharathi B., Susee-lan Angel, Krishnamurthy Vallidevi. (2016). A Survey of Voice Translation Methodologies - Acoustic Dialect Decoder. 1-9. 10.1109/ICICES.2016.7518940. 3

[7] Khan Usman, Sarim Muhammad, Ahmad, Maaz, Shafiq Farhan. (2019). Feature Extraction and Mod-eling Techniques in Speech Recognition: A Review. 63-67. 10.1109/ICISE.2019.00020.

[8] Sawant Sai, Deshpande Mangesh.(2018). Isolated Spoken Marathi Words Recognition Using HMM. 1-4. 10.1109/ICCUBEA.2018.8697457.

[9] Sherstinsky Alex. (2020). Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenom-ena. 404. 132306. 10.1016/j.physd.2019.132306. 3

[10] Misra Ishan, Girshick Ross, Fergus Rob, Hebert Martial, Mulam Harikrishna, van der Maaten Lau-rens. (2018). Learning by Asking Questions. 11-20. 10.1109/CVPR.2018.00009. 4

[11] Ronghang Hu, Jacob Andreas, Marcus Rohrbach, TrevorDarrell, Kate Saenko. (2017). Learning to Rea-son: End-to-End Module Networks for Visual Ques-tion Answering. 804-813. 5

[12] Adam Santoro, David Raposo, David Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, Timothy Lillicrap. (2017). A simple neural network module for relational reasoning. 10.1109/ICCV.2017.93. 5, 6

[13] Johnson Justin, Hariharan Bharath, van der Maaten Laurens, Hoffman Judy, Fei-Fei Li, Zitnick C., Girshick Ross. (2017). Inferring and Execut-ing Programs for Visual Reasoning. 3008-3017. 10.1109/ICCV.2017.325. 5

[14] Perez Ethan, Strub Florian, Vries Harm, Dumoulin Vincent, Courville Aaron.(2017). FiLM: Visual Reasoning with a General Conditioning Layer. 5

[15] Anderson Peter, He Xiaodong, Buehler Chris, Teney Damien, Johnson Mark, Gould Stephen, Zhang Lei.(2018). Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. 6077-6086. 10.1109/CVPR.2018.00636. 5, 6

[16] Kim Jung, Ro Yong. Attentive Layer Sep-aration for Object Classification and Object Localization in Object Detection. 3995-3999. 10.1109/ICIP.2019.8803439. (2019) 3, 4, 6

[17] Ouadiay Fatima, Bouftaih Hamza, Bouyakhf El Hous-sine, Himmi M.(2018). Simultaneous Object Detection and Localization using Convolutional Neural Net-works. 10.1109/ISACV.2018.8354045. 3

[18] Angelo Kandavalli.(2018). A novel approach on ob-ject detection and tracking using adaptive back-ground subtraction method. 1055-1059. 10.1109/IC-CMC.2018.8487514. 3

[19] Ulagamuthalvi J. B. J. Felicita, D. Abinaya. An Ef-ficient Object Detection Model Using Convolution Neural Networks, ICOEI, 2019, pp. 142-147. 4

[20] M. Sharma, S. Shukla, Relative object localization us-ing logistic regression, ICACCA, 2017, pp. 1-5.

[21] Andreas Jacob, Rohrbach Marcus, Darrell Trevor, Klein Dan.(2016).Neural Module Networks. 39-48. 10.1109/CVPR.2016.12. 5

[22] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about realworld scenes based on uncertain input. In NIPS, 2014. 1

[23] Mengye Ren, Ryan Kiros, Richard Zemel. 2015. Ex-ploring models and data for image question answering. In NIPS. 1

[24] Dual Attention Networks for Multimodal Reason-ing and Matching. Hyeonseob Nam, Jung-Woo Ha, Jeonghee Kim. arXiv:1611.00471v2 2017 6