Survey: Visual Question Answering Techniques

Arushi Kumar (2014023), Avneet Kaur (2014027) , Purusharth Dwivedi(2014081)

***Abstract*— With the growth in the field of camera technologies and the advent of social media applications, it has become a common activity to click and upload pictures of anything and everything around the world. As a result, there is a plethora of information on the Internet available in the form of natural language as well as images. Hence, there is a need to develop a system for answering factoid as well as descriptive questions about or based on these images, in order to present knowledge in a readable way without human intervention. This has an application in developing search engines that provide to the point answers to image based, topic specific questions rather than presenting unnecessary information in the form of an entire document pertaining to a particular topic.**

I. INTRODUCTION

Language and vision problems like visual question answering(VQA (http://www.visualqa.org/) have gained popularity in recent years towards solving multi-modal problems. A VQA system takes an image and a natural language question as an input and produces an answer to that question as an output. These questions could target different areas of the image like background context, object details, etc. The questions could be based on object detection (Eg. How many balls are there?), activity recognition (Eg. Is this man laughing?), fine-grained recognition (Eg. What is the color of her eyes?). This kind of an AI system will enable perception in computers in the form of visuals, processing questions as natural language and answering them, thus mimicking the human brain. Also, it has an application in knowledge representation and reasoning. (Figure 2)

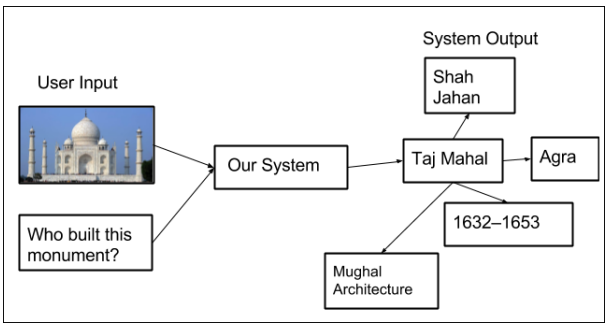


Fig. 1. Representation of Knowledge and Reasoning

One might argue that such factoid questions can be answered using the recent work done in image captioning which is another problem involving an amalgam of computer vision and natural language based techniques.However, descriptive explanations, supportive documents relating to the natural language question cannot be provided unless we a have a complete understanding of the image relating to what it talks about, and what is it describing. There is a lot of extra information that is required which may be encyclopedic knowledge about specific things or it may be common sense knowledge of day to day things. Thus, the VQA problem becomes an AI driven task as it requires knowledge from multiple areas. This leads to an ever increasing interest in VQA for the researchers.

Through this survey we aim to discuss various techniques that have been applied to this problem. In the first part of this survey we explore the related work, followed by the various datasets used in this problem till date. Then we present the various techniques classified as non-deep learning based approaches, deep learning based approaches, methods involving support from external knowledge bases and finally, region specific attention based models applied to this problem.

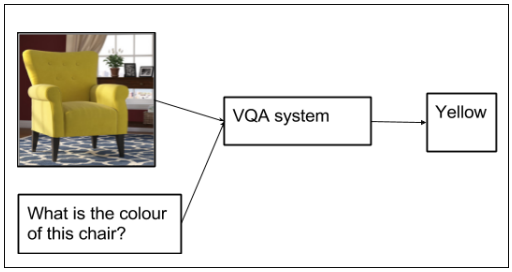


Fig. 2. System

II. RELATED WORK

A. Generating Image attributes

This problem is about finding ways to detect various parts of an image, objects , scenes, backgrounds, actions etc. and provide an input image with natural language attributes. This is the fundamental problem to be solved in case we wish to understand what are the image descriptors in natural language. Efforts have been made in both probabilistic image tagging as well as deep learning based approaches in [1] and [2]. Words and sentences are required to understand the semantics of an image. Recent work has also been done in generating images given natural language attributes of an image in [3].

B. Generating Image Captioning

This problem, again an amalgam of natural language and computer vision problems is a widely studied problem,extending the problem of generating image attributes and detecting objects, where the goal is to generate phrases to describe what is going on in the image. It focuses on

generating short phrases to describe the content of an image rather than long descriptions. Many probabilistic based as well as deep learning based approaches have been applied to the same.

C. Text based Question Answering

An important task in the field of natural language processing is text based question answering, a well-studied problem . Here, essentially we need to be able to answer both factoid as well as descriptive questions. Interesting work has been done in this field by [4], [5], [6]. Some of the related tasks include, sentence completion as explored in [5], measuring sentence as well as document similarity in order to understand the natural language question and provide suitable answers. Also, this field includes open domain [7] as well as common sense question answering. There are some state of the art techniques which use methods like semantic parsing in order to arrive at a logical form in order to capture the intended meaning of a particular answer as in [8] and [9]. Many deep learning based models including recurrent neural networks have also been applied for this task as in [10]. This is a widely studied problem, and forms the basis of Visual Question Answering explored in this paper.

D. Generating Image Descriptions

This is another problem explored in [11], citeKarpathy2017, [12], to generate natural language descriptions of a visual that is provided. This problem extends from the problem of generating short natural language captions for a particular image. This problem is quite similar to visual question answering. If we have certain description about an image, it might help us to answer certain questions related to it in natural language. Describing visual content in natural language has been addressed in various research areas. Some ideas involve the use of a recurrent neural network in order too achieve this task. The model first observes visual content and later it is trained to learn a sequence of words describing the content. The VQA task is an extension of this problem.

E. Visual Turing Test

Recently, a lot of work has been done in this field as in [13]. This problem explores the problem of visual question answering from a different perspective. The problem framed here is to generate a set of binary question and answers about an image in order to depict what is going on in the scene. This is done for a varied set of question answer pairs talking about different aspects of the image.

III. MOTIVATION

The VQA problem is a much more challenging problem than generating captions or attributes for a particular image and has various applications. It frequently talks about information about things not directly present in the image. Success in this field would also help in solving problems like video question answering, generating descriptions of videos and developing applications for the visually impaired in order to help them see and experience the world around them.

IV. DATASETS

There are many datasets which can be used for research on VQA. At the very least, they contain images, questions, and their answer. Sometimes, additional information could be provided like image captions or image regions supporting the answers. The complexity of datasets and questions, amount of reasoning to get the correct answer vary widely. A summary of datasets is given in Figure 3. Examples of images from these datasets are given in the last page.

A. DAQUAR

The Dataset for Question Answering on Real world images was released in 2015 and was the first dataset designed as benchmark for the VQA test. It takes images from the NYU-Deth v2 dataset which has 1449 RGBD images of indoor scenes along with annotated semantic segmentations. The images are divided into 795 training and 654 test images. There are 894 possible classes and each pixel of in image is labeled with an object class (or no object). The question answer pairs were generated in two ways: 1) Automatic generation using question templates using 9 templates for questions and annotations of the NYU-Depth V2 dataset. 2) Human annotations collected from 5 annotators. The question/answer pairs were focused on colors, numbers, objects and sets of those. 12468 question/answer pairs were collected out of which 6794 are for training and 5674 for testing. The main drawback is that the answers are restricted to a set of 894 object categories and 16 colors. There could be 2 evaluation methods: First is accuracy which is a poor method for multi-word answers. Second is WUPS score. It uses an average matching between the answer and the ground truth answer and generates a score between 0.0 and 1.0. The threshold for WUPS score is 0.9. It means the answer would be correct if the WUPS score is above 0.9.

B. COCO-QA

This dataset is based on MS-COCO (Microsoft Common Objects in Context data). It has 123,287 images out of which 72,783 are for training and 38,948 are for testing. The 4 types of questions that are used are based on object, number, color, and location. Each image has one question and answers are all single word. They were generated automatically by turning the descriptions of the image of the original COCO dataset into question/answer form.

C. FM-IQA

FM-IQA (Freestyle Multilingual Image Question Answering) has 158,392 images which is from the COCO dataset. The questions/answers are provided here by humans through the Amazon Mechanical Turk crowd-sourcing platform. The questions which were given by the annotators could be of any type as long as they were related with the contents of the image leading to a greater diversity of questions. Questions/answers are available in both Chinese and English translation. There are 316,193 questions in this dataset.

D. VQA

The Visual Question Answering dataset is one of the most widely used dataset. It is divided into 2 parts: One contains the natural or real world images, and other one has the abstract clip-art scenes which are made from models of humans and animals so that theres no need to process noisy images and high level reasoning could be performed. Questions and answers are generated from crowd-sourced workers. For each question there are 10 answers which are obtained from these unique workers. Answers are generally one word or a short phrase and 40 % questions have a yes or no answer. For evaluation, two type of formats are available. First, multiple choice questions, which have 18 candidate responses. Second, open ended answer generation for which a machine generated answer is normalized by the visual question answering evaluation system. The score would be evaluated as min(number of humans who provide that exact answer/3, 1). If the answer matches the responses of at least 3 human annotators then only it will be considered as correct. The original VQA dataset has 50,000 abstract images with 150,000 questions and 204,721 MS-COCO images with 614,163 questions.

E. VISUAL7W

This dataset for image captioning, recognition and segmentation is created using images from MS-COCO dataset. The Visual7W dataset is named so because of creation of multiple choice questions of the form (Who,What, Where, When, Why, How and Which). The questions were generated by the workers in Amazon Mechanical Turk (AMT). 3 workers from a separate set rated the questions and those questions with less than 2 positive votes were discarded. Multiple choice answers were created automatically as well as by the AMT workers.

F. VISUAL MADLIBS

It is primarily a dataset consisting of images from MS COCO and fill-in-the blanks and multi-choice questions. The fill-in-the blanks are descriptive and generated automatically using template information. For each question, there can be three answers and which can consist of either a single word or multiple words syntactically correct or not, constituting a phrase. It contains 10,378 images and 360,001 questions.

Name	Number of Images	Number of questions	Average questions per image
DAQUAR	1449	12,468	8.60
COCO-QA	117,684	117,684	1.00
FM-IQA	120,360	316,193	1.99
VQA (COCO)	204,721	614,163	3.00
VQA (Abstract)	50,000	150,000	3.00
VISUAL7W	47,300	327,939	6.93

Fig. 3. Summary of Datasets in VQA (insert ref)

V. METHODS AND TECHNIQUES

A. Non Deep Learning based approaches

1) **ANSWER TYPE PREDICTION***:* One of the ways the user intention can be known is through answer type prediction for a particular question . In this model by [14], a Bayesian framework for answer type prediction is proposed. This approach can be mainly divided into two stages : predicting answer types, and then predicting the probability of an answer. The datasets used are DAQUAR, MS-COCO, VQA, Visual7W. The types of answers may be different in various datasets that have been used. For example, in the DAQUAR dataset, categories (both numerical and categorical) like number, colour, and other, were created by looking at the answers. For the COCO-QA dataset, there are namely four categories defined within the dataset: colour object, location and the number also. For all the questions in the datasets, the answer type is predicted using skip thought vectors and logistic regression. Then, the idea is to predict the probability of an answer , $(A = k)$ and Type $(T = c)$, given the input image x and query q. Hence, we need to find $P(A = k, T = c|x, q)$. Applying the Bayes' Rule and Chain rule for probabilities, we end up with the following equation.

$$P(A = k, T = c|\mathbf{x},\mathbf{q}) =$$

$$\frac{P(x|A = k, T = c, \mathbf{q})P(A = k|T = c, \mathbf{q})P(T = c|\mathbf{q})}{P(x|\mathbf{q})}$$

Here, $P(x|A = k, T = c, \mathbf{q})$ is the probability of an image features given the answer,answer type and the question. $P(A = k|T = c, \mathbf{q})$ is the probability of an answer given an answer type and question,$P(x|\mathbf{q})$ is the probability of an features of an image given a question and $P(T = c|\mathbf{q})$ is the answer type given a question q. In the end, to get an answer, P(A=k—x,q) is calculated by marginalizing over all the types of answers.

B. Deep Learning Based Models

Owing to the increasing dataset size of images, text, questions etc, certain developments were made using deep learning approaches. This was done because of advancements in deep learning models for object, scene, background recognition, and natural language understanding. This enabled high level understanding for image understanding and hence VQA. Some of these approaches are described below.

1) **VQA USING CNN***:* In this work by [15] a CNN based approach is explored for the VQA task. 3 CNN models based framework for learning image features , question representations as well as a multi-modal (image + question features) framework for learning the interactions between the two in order to predict the answer is provided. The first model encodes the image content. The second model first finds a word embedding for each word in the question and then considers a convolution unit with a local receptive field and shared weights like in order to obtain important structures between a set of consecutive words. The max

pooling layer is used to select interesting sequence of words, and also filters meaningless ones (eg: ”red chair” would be selected and ”in front of the” would be filtered out.). After several layers of convolution and max pooling are applied, a representation for the question is finally obtained. The third CNN model, in addition to taking the image representation as an individual semantic component, combines it along with the textual representations in order to learn the interactions between the multi-modal data and produces the desired output.

2) **SIMPLE BASELINE IBOWIMG***:* This method based on the work by [16] proposes a simple bag of words (to learn word features) + convolutional neural network (CNN) (to learn image representation) based approach modeled on the much larger version COCO VQA dataset and claims that its performance is comparable to the LSTM models previously proposed under proper setup and training. Moreover, the model trained eventually learns a correlation between discriminative words in the query (input question), the visual concepts in the image and the answer. This approach models the VQA problem as a classification task where the total number of final answer classes that the model needs to learn is actually the number of different types of answers given in the training set. Bag-of words is used as a text feature and combined with the visual features learned from GoogLeNet and finally fed into a soft-max layer to predict the class of the answer. The complete framework is better explained in the form of Figure 4 from [16].

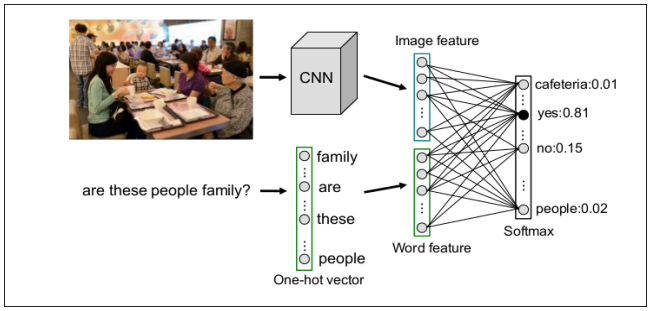


Fig. 4. Framework of the iBOWIMG.[16]

3) **EXPLAINABLE VISUAL QUESTION ANSWERING USING ATTRIBUTES AND CAPTIONS***:* This approach is based on the work of [17] addresses an important concern with the existing algorithms, i.e. , lack of contextual explanation in the answer to a visual based question which might not be readable to the user.This may lead to guessing by the system or incorrect understanding as explained in the figure. The approach hence proposed breaks the VQA pipeline into two steps: explaining (understanding image content) and reasoning (inferences related to the understanding regarding the answer). (see Figure 5) To support this, an image captioning model is used to generate explanations at (1) word level (object level/ attribute level understanding of an image) and (2) sentence level (expressing relationships between objects using captions. This better explained through the following figure 7. The framework consists of three parts :

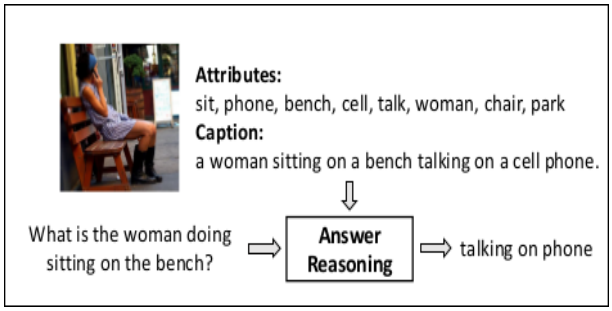


Fig. 5. explanation and reasoning in VQA[17]

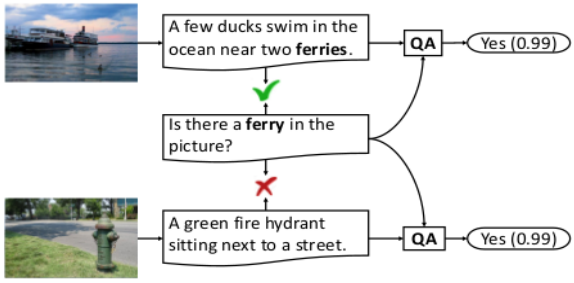


Fig. 6. Two contrasting cases that show how the explana- tions can be used to determine if the system learns from a training set bias or from the image content [17]

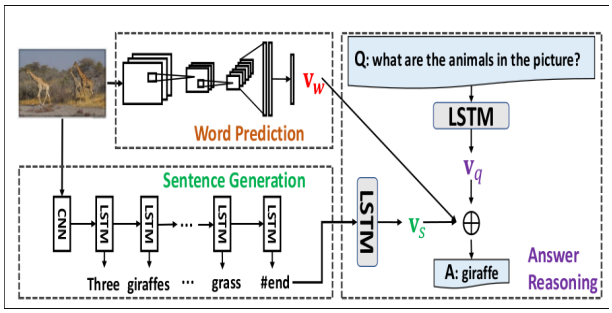


Fig. 7. Word prediction (upper left), sentence generation (lower left), answer reasoning (right). [17]

- Word Prediction : A list of most frequent words from the MS COCO dataset including object names, actions, properties, etc. are extracted and paired with the images allowing one image to have multiple labels. Then the word prediction based on this training set is modeled as a multi-class classification task. ResNet pre-trained on Image net is used for the same.
- Sentence Generation : An image captioning model is trained by maximizing the probability of the correct captions for an image.
- Answer Reasoning.

Advantages over previous approaches The model hence formed, provides an intuitive understanding to the user by providing explanation for whatever answer it predicts. This ensures that the system doesn't guess an answer. Moreover, separating the explaining from the reasoning task lets us see why might the answer predicted may be wrong. e.g.: If the answer does not contain key tags essential to the image, then the problem lies in the explanation step.

4) **ASK YOUR NEURONS(AYN)***:* This method uses a CNN for generating an image embedding vector from an image, to encode its content. A separate LSTM (Long-Short term memory) network is used to encode the questions for which the input at time step t is the word embedding for the t *th* question word qt , as well as the encoded image vector. The final hidden vector generated is the question embedding. Authors use a simple baseline, wherein the final question embedding is the sum of all the word embeddings. Classification is then performed by a fully connected layer followed by a softmax over all the possible answers in the dataset. However, for answer generation, a decoder LSTM is used, which at each time-step takes as input the previously generated word as well as the question and image encoding. The next word is predicted using a softmax over the vocabulary. This model shares some weights between the encoder and decoder LSTMs.

C. Using support from external knowledge bases

1) **MOTIVATION***:* The motivation behind this is that some questions cannot be answered if they are lacking a background contextual information. Moreover, while answering a question, the methods previously proposed may not tell us how they arrived at the answer, they may not be able to explain the reason for arriving at a particular answer. The following methods proposed take this aspect into account and VQA by learning facts about the image from knowledge bases are proposed.

2) **EXPLICIT KNOWLEDGE BASED REASONING FOR VQA***:* This method by [18] proposes an approach to provide reasoning for a question about an image, by using the external information provided by knowledge bases. Also, a dataset called KB-VQA consisting of questions (based on images from VQA dataset) which require support from external knowledge bases for their answers is created for experiments. This method enables indirect question / answering of the image and also provides reasoning for the answer given by the system as explained in the following Figure 7. For this, a two step approach is followed:

- RDF Graph construction: This is done to extract additional information about an image in order to reason about it. For this, first certain visual concepts (objects, scenes, attributes) are extracted from an image using various CNN models, and then these visual concepts are stored as rdf triples. Further, each rdf triple is linked to a DBPedia entity with the same semantic sense.

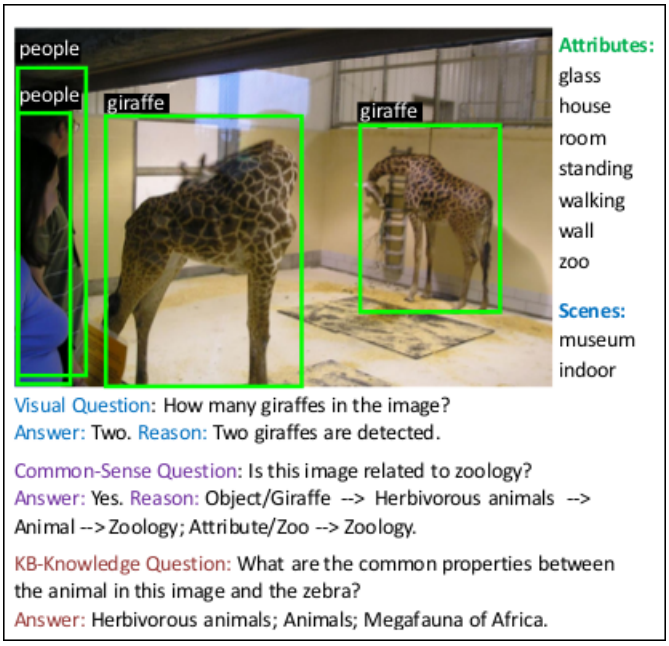


Fig. 8. A real example of the proposed KB-VQA dataset and the results given by the proposed method. [18]

- Answering Questions: In this step, the question(query) is converted to a form which can be input as a Knowledge Base query and then appropriate SPARQL queries are formed. The results hence obtained are used to give a logical reasoning to the question asked.

3) **FREE-FORM VQA USING EXTERNAL KNOWLEDGE BASES***:* Like the previous method, this method by [19] also proposes reasoning based answers to visual questions, taking help from external knowledge bases. However, the novelty in this approach is, representing images in textual form using their attributes and captions instead of using image features directly, and then, combining them using an LSTM framework. An example is shown in Figure 8. First, an attribute based image representation of the image is framed as a multi-class classification problem, based on the already available captions from the MS COCO dataset, using a CNN and following the approach by [20] . Then, using a CNN combined with an recurrent neural network RNN, captions are generated based on the work by [21]. Further, DBPedia is used as an external knowledge source for extracting information about the top 5 predicted attributes for each query. The descriptive information extracted from the comment section of the returned result, is then vectorized using DocToVec, in order to capture the semantics of the returned result. Finally an encoder-decoder LSTM framework is used to generate an answer for a question by maximizing the probability of the correct answer given the question.

D. Image/Regional Attention(focus) based approaches

MOTIVATION The approaches discussed below focus on question answering based on regional attention in images.

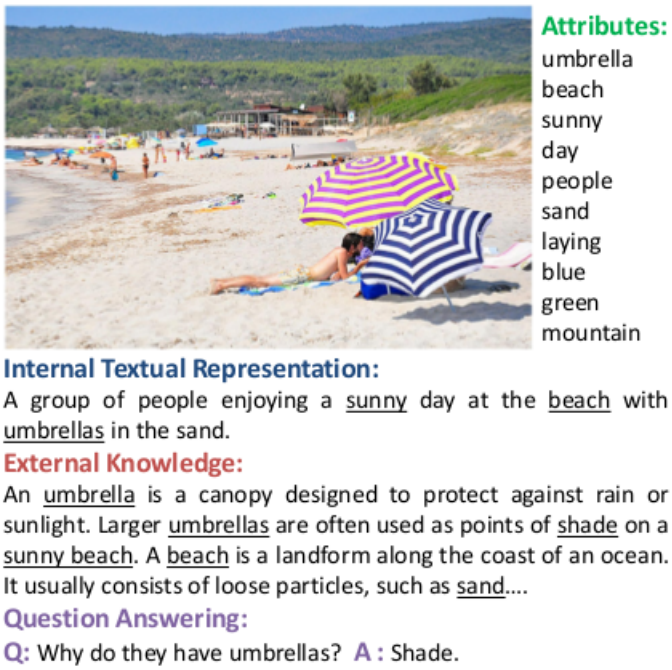


Fig. 9. A real example of the proposed KB-VQA dataset and the results given by the proposed method. [19]

One of the main reasons why this area is being explored is to provide a sense of visual reasoning that is reasoning based on certain focus regions in images. For example, suppose we want to ask, what is on the sofa? , then we want to focus on the sofa and then find out what is on it. Previously, text based reasoning approaches have been followed in order to provide explanations on how to arrive at an answer. This extends from that work for images. The difference between traditional approaches and this approach is summarzsed in the following figure.

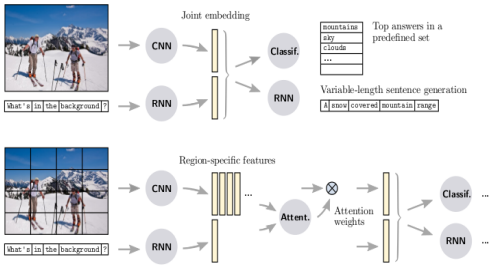


Fig. 10. (Top) A common approach to VQA is to map both the input image and question to a common embedding space. (Bottom) Attention mechanisms build up on this basic approach with a spatial selection of image features. [22]

1) **WHERE TO LOOK***:* This approach in the work by [23], is based on looking at parts of the image relevant to the natural language question and then providing a suitable answer. Figure 9 provides an example for the same. It aims to learn a non linear mapping between an parts of an image and the natural language question. First, an image-region selection mechanism based on CNN learns different parts of an image depicting various objects, relevant to various questions. The layer first projects the image features and then the text features into a shared N dimensional space and then calculates the inner product for each question-answer pair and image regions. Then, a framework is trained for multi-choice VQA, that is, given a question and its various choices, the aim is to maximize the margin between correct and incorrect choices in a structured fashion. Finally, comparison is done between the baselines and this approach using the whole image, without image and with image regions with weighting given to them thus providing an analysis for when selected regions improve the performance of the task.

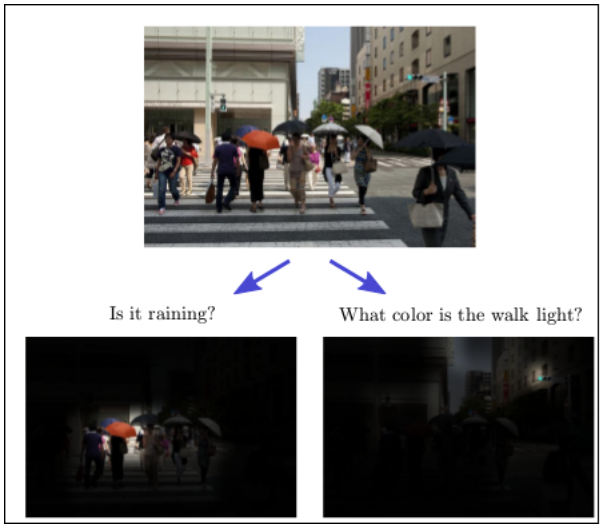


Fig. 11. Our goal is to identify the correct answer for a natural language question, such as What color is the walk light? or Is it raining? [23]

2) **DUAL ATTENTION NETWORK FOR MULTI-MODAL REASONING AND MATCHING***:* This method by [24] makes use of both image as well as textual based attention models in order to find specific regions of attention within the image to answer specific questions about a particular image. The words in the image are encoded as one hot vectors and embedded onto vector space. The image is represented as a set of image regions in vector form. They are then passed through a bidirectional LSTM. This helps to focus on different parts of an image in different time steps.

3) **SHOW, TELL, ATTEND AND ANSWER***:* In this work by [25], a neural as well as probabilistic based approach for generating captions is proposed where the probability of a description given an image is maximized. The VQA answering task is modeled as a classification problem where the features are extracted from to different places: textual features from LSTM and image features from ResNet. These features are then concatenated and then passed through a convolution layer which produce image feature glimpses and a probability distribution over the answers is obtained.

4) **Neural Module Networks***:* This method by [26] combines a variety of image as well as textual features. However instead of utilizing a single large network, this algorithm utilizes a compositional structure to build smaller networks in training time. This algorithm combines the question answering task as a function of many simple tasks which are determined by the question asked. These tasks are performed by a set of modules which are based on neural networks and are combined with other neural network modules in order to predict the answers. A network graph is made with these modules and the structure of this graph is parsed in order to provide a natural language answer to the question about the image.

VI. Image Captioning Techniques For VQA

Motivation Image captioning serves as a subproblem for visual question answering. For a new input image, whose tags may or may not be known, generating captions plays a very important role of describing what the image is talking about. Descriptions must also capture the semantics of an image in terms of how different objects in an image are related to each other. Figure shows an example. Moreover, for understanding natural language, language models are also needed. Recent work in this field focuses on several deep learning based caption generating techniques that have been discussed below.

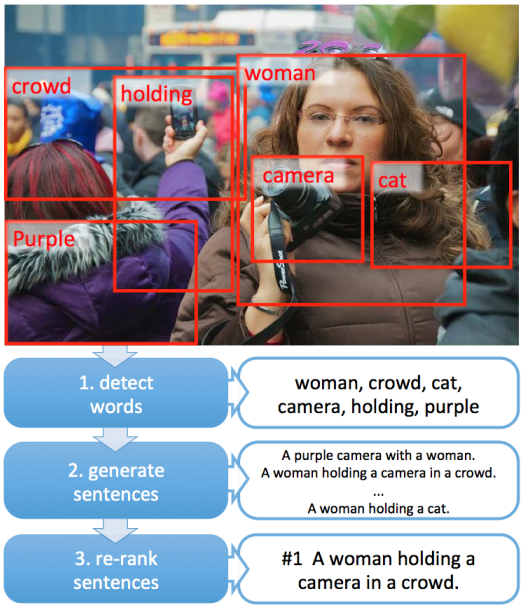


Fig. 12. Image Ref: Google

1) **Image Captioning Based on Attributes and External Knowledge***:* This work by [26] proposed a method to caption images based on attributes and using external knowledge bases. This model is subdivided into two parts: image analysis and caption generation. For the first part, the task of predicting attributes for an image is modeled as a multi-class classification problem and a CNN model is trained for the same. Then, a fixed length vector V is generated for each image whose size equals the size of attribute set for all images. For every dimension of the vector, there is a prediction probability for a particular attribute. Third, for caption generation, an LSTM based sentence generator is applied. In this, the probability of a description given an image is maximized. For the predicted attributes, top five are input to an external knowledge source called DBpedia, to expand the information available to us in the form of captions to the image generated by the model.

2) **Phrase-Based Image Captioning***:* In this work by [27] a bilinear model is proposed that learns a metric to fuse image representations generated by a CNN and natural language descriptions generated using an RNN. The datasets utilized for this task are FLICKER30k and COCO dataset. For an input image, a bilinear model is trained to output a set of sentences that can be used to best describe it. But the difference is, that here we want to generate syntactically correct descriptions which are a subset of noun phrases, verb phrases etc. For this, a trigram based language model is trained, based on the structure of descriptions in the training set. The output is a set of sentences that may or may not be conditioned for the given image and hence, they are re-ranked based on a similarity score to pick the description that is closest to the image.

3) **Language models for image captioning***:* In this method by [22], language models for the task of image captioning is explored. First, a convolution neural network was used to predict a bag of words that are likely to be present in a caption. Then, a maximum entropy language model is used to generate syntactically and grammatically correct sentence descriptions for a particular image. A K nearest neighbor model is employed for finding descriptions of a test image. For each caption, an n-gram overlap F-Score is computed between this particular caption and all others. Intuitively, a single caption that describes many different images similar to a particular test image rather than a caption that describes a single image most similar to the test image is chosen so as to get a broader scope of understanding. The advantage is of a retrieval based approach to image captioning so that false positives are eliminated.

VII. Correlation/Insight

Currently, the best performing methods are all attention based deep-learning methods. This clearly shows the importance of identifying parts of the image that are relevant to the question and then using that region for generating the correct answers. However, there are also models that make use of external knowledge source like DBpedia, Wikidata etc, which in-turn allows them in answering non-trivial questions, answers for which may not be directly available or interpretable from the image content. However, we found that none of the current crop of algorithms make use of both the available source of knowledge, i.e., making use of

	Test-Development					Test-Standard				
	Open Ended				M.C.	Open Ended				M.C.
	Y/N	Number	Other	All	All	Y/N	Number	Other	All	All
iBOWIMG	76.5	35.0	42.6	55.7	-	76.8	35.0	42.6	55.9	-
DPPnet	80.7	37.2	41.7	57.2	-	80.3	36.9	42.2	57.4	-
WTL	-	-	-	-	62.4	-	-	-	-	62.4
AYN	78.4	36.4	46.3	58.4	-	78.2	36.3	46.3	58.4	-
SAN	79.3	36.6	46.1	58.7	-	-	-	-	58.9	-
ATP	80.5	37.5	46.7	59.6	-	80.3	**37.8**	**47.6**	60.1	-
NMN	**81.2**	38.0	44.0	58.6	-	**81.2**	37.7	44.0	58.7	-
CoAtt	79.7	**38.7**	**51.7**	**61.8**	**65.8**	-	-	-	**62.1**	**66.1**
AMA	81.01	38.42	45.23	59.17	-	81.07	37.12	45.83	59.44	-

Fig. 13. Performance of various models on VQA

external knowledge sources for improving performance on more complicated questions as well as using attention cues to identify areas of interest with regards to the question and in turn improve the answering capability and accuracy. This is one area which we want to explore further in our project.

VIII. Discussion and Future Work

All the VQA methods can be classified essentially into 3 broad groups: Deep-Learning based models without attention, Deep-learning based models with attention and Bayesian methods. Due to the recent successes of Deep-Learning techniques in Computer Vision and NLP tasks, Visual QA shows great promise with Deep-Learning since it involves a multi-modal understanding and fusion of both image and textual features. Therefore, most deep-learning based models only differ in their network structure and how the various embeddings for questions and images are generated to create a final joint embedding.

The answers can be decoded either as a classification problem over all possible answers in the dataset or as a answer generation problem. Most approaches tend to perform a multi-class classification, on all possible answer labels, although some recent papers are also target answer generation.

The deep-learning models have shown great promise and their performance has been steadily improving. It is therefore interesting to note that a non-deep learning approach outperforms the current crop of deep-learning based methods. Answer type prediction model (Reference), which is a non-deep learning outperforms various deep-learning models and even some deep learning models with attention on the VQA dataset. This shows that simply adding convolution or recurrent neural networks is not enough and identifying parts of the image that are relevant to the question in a principled manner is much more important to attain high accuracy.

Ask Me Anything(AMA) is a model, which makes use of external knowledge base like DBpedia for achieving high score. Such a performance might be due to the type of questions available on this dataset, but it can also be because the knowledge base enables the system in answering more general, common-sense based questions. However, the performance of this model is not as good on VQA dataset, which might be because not too many questions in this dataset require external knowledge source.

Thus such a model gives rise to another avenue for future work. If we are able to identify when to use an external source for answering a question, we might be able to see a big improvement in the accuracy of these systems.

Another area of future work is Video Question Answering. Currently, there has been less work done in this field. Since an image is nothing but a series of frames, over a series of frames if we could predict what is happening or changing with respect to the video, we could easily develop an AI based system to answer questions regarding the same. The work done for Visual Question Answering could be easily extended for Video Question Answering.

IX. Conclusion

We present an overview of various techniques and methodologies used for the Visual Question answering task. We compared the various techniques on the basis of approaches used for generating image embeddings for capturing image content and textual embeddings for capturing question semantics. We also compared these algorithms on the basis of their performance on the various VQA datasets and discussed the current trend to move towards a more attention-focused mechanism by various models since the use of attention mechanism has shown significant improvement in the performance of deep-learning approaches. We also discussed how some models employ external data sources/knowledge bases for answer prediction. The external knowledge-bases along with attention could be the next steps in achieving better performance in Visual Question Answering task.

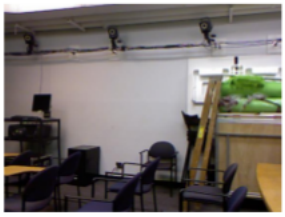
DAQUAR [51]			
	Q: How many white objects in this picture ? A: 9	Q: What color is the chair in front of the wall on the left side of the stacked chairs ? A: blue	Q: What is the largest white object on the left side of the picture ? A: printer

Fig. 14. DAQUAR [28]

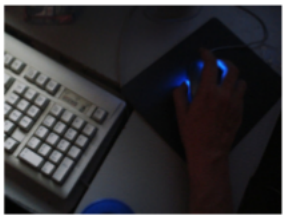
COCO-QA [63]			
	Q: How many giraffes walking near a hut in an enclosure ? A: two	Q: What is the color of the bus ? A: yellow	Q: What next to darkened display with telltale blue ? A: keyboard

Fig. 15. COCO-QA [28]

VQA-real [3]			
	Q: What shape is the bench seat ? A: oval, semi circle, curved, curved, double curve, banana, curved, wavy, twisting, curved	Q: What color is the stripe on the train ? A: white, white, white, white, white, white, white, white, white	Q: Where are the magazines in this picture ? A: On stool, stool, on stool, on bar stool, on table, stool, on stool, on chair, on bar stool, stool

Fig. 16. VQA-real [28]

Visual Genome [41]			
	Q: What color is the clock ? A: Green	Q: What is the woman doing ? A: Sitting	Q: How is the ground ? A: dry

Fig. 17. Visual Genome [28]

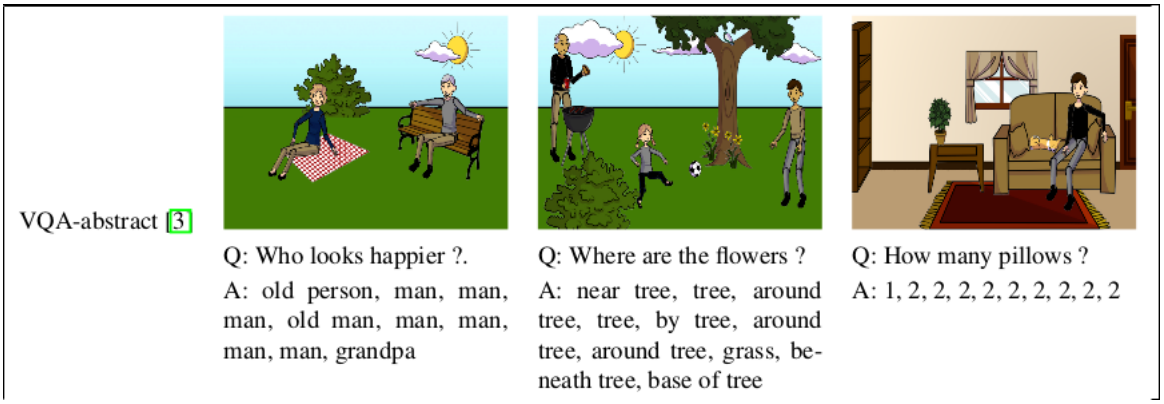


Fig. 18. VQA-abstract [28]

X. EXAMPLES OF VARIOUS DATASETS

REFERENCES

[1] "Midge: Generating Image Descriptions From Computer Vision Detections," *Eacl*, pp. 747–756, 2012.

[2] X. Zhang, Z. Huang, H. T. Shen, and Z. Li, "Probabilistic image tagging with tags expanded by text-based search," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 6587 LNCS, no. PART 1, 2011, pp. 269–283.

[3] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2Image: Conditional image generation from visual attributes," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, 2016, pp. 776–791.

[4] A. Fader, L. Zettlemoyer, and O. Etzioni, "Paraphrase-Driven Learning for Open Question Answering," *Acl*, pp. 1608–1618, 2013. [Online]. Available: http://turing.cs.washington.edu/papers/acl-2013-fader.pdf

[5] M. Richardson, C. J. C. Burges, and E. Renshaw, "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text," *Empirical Methods in Natural Language Processing (EMNLP)*, no. October, pp. 193–203, 2013.

[6] J. Weston, A. Bordes, S. Chopra, A. M. Rush, B. V. Merri, A. Joulin, and T. Mikolov, "Towards AI Complete Question Answering," *CoRR*, 2016. [Online]. Available: https://arxiv.org/pdf/1502.05698.pdf

[7] "Experiments with open-domain textual question answering," *Proceedings of the 18th ...*, pp. 292–298, 2000. [Online]. Available: http://dl.acm.org/citation.cfm?id=990863

[8] P. Liang, M. I. Jordan, and D. Klein, "Learning Dependency-Based Compositional Semantics," *Computational Linguistics*, vol. 39, no. 2, pp. 389–446, 2013. [Online]. Available: http://www.mitpressjournals.org/doi/10.1162/COLI_a_00127

[9] J. Berant and P. Liang, "Semantic Parsing via Paraphrasing," in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2014, pp. 1415–1425. [Online]. Available: http://aclweb.org/anthology/P14-1133

[10] M. Iyyer, J. Boyd-Graber, L. Claudino, R. Socher, and H. Daumé III, "A Neural Network for Factoid Question Answering over Paragraphs," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 633–644. [Online]. Available: http://aclweb.org/anthology/D14-1070

[11] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677–691, 2017.

[12] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, 2015, pp. 3156–3164.

[13] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual Turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, p. 201422953, 2015. [Online]. Available: http://www.pnas.org/lookup/doi/10.1073/pnas.1422953112

[14] K. Kafle and C. Kanan, "Answer-Type Prediction for Visual Question Answering," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 4976–4984. [Online]. Available: http://ieeexplore.ieee.org/document/7780907/

[15] L. Ma, Z. Lu, and H. Li, "Learning to Answer Questions from Image Using Convolutional Neural Network," in *AAAI*, 2016, p. 7. [Online]. Available: http://arxiv.org/abs/1506.00333

[16] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *CoRR*, vol. abs/1512.02167, 2015. [Online]. Available: http://arxiv.org/abs/1512.02167

[17] J. Y. D. M. T. L. J. Li, Qing; Fu, "Tell-and-Answer: Towards Explainable Visual Question Answering using Attributes and Captions," 2018. [Online]. Available: https://arxiv.org/abs/1801.09041

[18] P. Wang, Q. Wu, C. Shen, A. van den Hengel, and A. R. Dick, "Explicit knowledge-based reasoning for visual question answering," *CoRR*, vol. abs/1511.02570, 2015. [Online]. Available: http://arxiv.org/abs/1511.02570

[19] Q. Wu, P. Wang, C. Shen, A. van den Hengel, and A. R. Dick, "Ask me anything: Free-form visual question answering based on knowledge from external sources," *CoRR*, vol. abs/1511.06973, 2015. [Online]. Available: http://arxiv.org/abs/1511.06973

[20] Y. Wei, W. Xia, J. Huang, B. Ni, J. Dong, Y. Zhao, and S. Yan, "CNN: single-label to multi-label," *CoRR*, vol. abs/1406.5726, 2014. [Online]. Available: http://arxiv.org/abs/1406.5726

[21] Q. Wu, C. Shen, A. van den Hengel, L. Liu, and A. R. Dick, "Image captioning with an intermediate attributes layer," *CoRR*, vol. abs/1506.01144, 2015. [Online]. Available: http://arxiv.org/abs/1506.01144

[22] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell, "Language models for image captioning: The quirks and what works," *CoRR*, vol. abs/1505.01809, 2015. [Online]. Available: http://arxiv.org/abs/1505.01809

[23] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," *CoRR*, vol. abs/1511.07394, 2015. [Online]. Available: http://arxiv.org/abs/1511.07394

[24] H. Nam, J. Ha, and J. Kim, "Dual attention networks for multimodal reasoning and matching," *CoRR*, vol. abs/1611.00471, 2016. [Online]. Available: http://arxiv.org/abs/1611.00471

[25] V. Kazemi and A. Elqursh, "Show, ask, attend, and answer: A strong baseline for visual question answering," *CoRR*, vol. abs/1704.03162, 2017. [Online]. Available: http://arxiv.org/abs/1704.03162

[26] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Deep compositional question answering with neural module networks," *CoRR*, vol. abs/1511.02799, 2015. [Online]. Available: http://arxiv.org/abs/1511.02799

[27] R. Lebret, P. H. O. Pinheiro, and R. Collobert, "Phrase-based image captioning," *CoRR*, vol. abs/1502.03671, 2015. [Online]. Available: http://arxiv.org/abs/1502.03671

[28] Q. Wu, D. Teney, P. Wang, C. Shen, A. R. Dick, and A. van den Hengel, "Visual question answering: A survey of methods and datasets," *CoRR*, vol. abs/1607.05910, 2016. [Online]. Available: http://arxiv.org/abs/1607.05910