Visual Question Answering: An Overview, Different Approaches, Use Cases, Future in Industry *

Kshitij Duraphe
Department of Electrical and Computer Engineering
Boston University
Boston, USA
kshitijd@bu.edu

***Abstract*—With advancements in computer vision and natural language processing, it is natural to ask whether the intersection of these two fields helps to produce a true strong AI. Visual Question Answering (VQA) is the natural evolution of both of these fields. Broadly speaking, VQA involves having a computer answering a question asked in Natural Language about an image given as an input. The ultimate goal of VQA is to produce a system that can answer any question asked about an image.**

***Index Terms*—computer vision, natural language processing, knowledge representation, vqa, deep learning**

I. Introduction

Given an image, a reasonable human being can answer basic questions on it. However, getting similar responses from a machine is an open question in the field of artificial intelligence today. For example, a child may be able to answer ’there are two green balloons tied to the tree’ in response to the question ’what is tied to the tree?’. Getting an answer anywhere close to this from a machine requires intensive computation and the answer may not be accurate at all.

In simple terms, a VQA system must analyze an image, extracting information from it. It must then be able to parse a question asked in natural language, converting it to a form understood by the machine. VQA is a natural successor to Textual Question Answering (TQA), which is a well-known and well-studied problem in Natural Language Processing (NLP). Compared to common image processing tasks such as captioning images and text-to-image retrieval, VQA is more complex. The main reason behind this is that the questions are asked ’on-the-fly’ and are not fixed. A VQA system must be able to classify the question, process it, and generate an answer based on the image regardless of the type of question asked. Consider the simple case of an NLP system tasked with answering the question ’how many windows does the White House have?’. The system merely needs to classify the type of question (this is a /emphhow many question, so the answer must be a number), extract the object to count (windows), and extract the context where the task must be performed (for this particular case, the White House). After the system analyzes a question, it generates an internal query for its internal database and relies on a knowledge base to get the answer. This is not trivial, but is a well-studied problem and thus has many databases one can use.

The same question can be asked to a VQA system. The difference, however, is that the VQA system must answer the question based on the image presented to it. It must analyze the image, develop a temporary database to store this data, and answer the question asked appropriately. For modern databases, the response generally consists of a few words or a small phrase. The search and reasoning part of the analysis must be performed over the course of an image. The system must be able to perform a variety of tasks such as detecting objects (i.e. if asked ’how many humans are in the image?’, the system must be able to detect that there are humans present (performing *object detection*)), classifying scenes (i.e. if asked ’is it cloudy?’, the system must be able to tell that there are clouds in the image and the level of sunshine is low compared to normal levels (performing *scene classification*)), and perfrom some level of commonsense reasoning and knowledge reasoning (i.e. if asked ’who’s holding the red ball?’). Many of these tasks have been addressed individually over the years, in different fields. However, VQA aims to bring those fields together and expand upon them.

VQA systems must be able to solve or have a high success rate with a broad spectrum of typical problems faced by the average natural language processing or computer vision system. It is a multidisciplinary field. A system that is capable of answering questions about images has extremely practical applications. Tools that aid the visually impaired as well as tools for online shopping websites that help describe the image are all applications of VQA systems. Virtual assistants can answer questions about products based on the image itself. Medical applications of VQA, such as defining the axis an x-ray image was taken at, are also numerous. Surveillance footage can be analyzed with VQA as well. Museum tours can be conducted with AI assistants instead of real people.

II. Question Answering, Computer Vision, and Knowledge Representation

A. Question Answering

Question Answering is simply the act of a computer answering a question asked to it. The question is most often asked by a human being. The computer must be able to understand the question (asked in natural language) and must be able to generate an appropriate response. It is an important test bed for evaluating how well computer systems understand human language. Since questions can be devised to query any aspect of text comprehension, the ability to answer questions is the strongest possible demonstration of understanding.

Questions can be of different types. They can be factoids (e.g. In which direction does the sun rise every day?) and can be non-factoids (e.g. What is the answer to problem 4 on the test?). In either case, an AI system must be able to generate an answer. The answers can be a short span of text, a yes or a no, a paragraph, a list, or a database entry. The context in which these questions are answered can refer to a document, collection of documents, a knowledge base, semi-structured tables, or even images.

The most common form of question answering by an AI that learns is textual question answering - TQA. This is also called 'reading comprehension'. An AI is able to learn from a text by 'reading' it and can answer questions. For example, an AI trained on 'Pride and Prejudice' should reasonably be able to answer the question 'How many sisters does Elizabeth have?'.

To actually generate an answer, a feature-based (for example) model would first generate a list of candidate answers. It can then further narrow this list down by considering only the candidates in parse trees.. It can then define a feature vector based on word frequencies, parse tree matches, length, dependency tables, and so on. The program then applies a multi-class logistic regression model to generate the answer.

B. Computer Vision

Computer vision is a vast field. In general, computer vision refers to the capability of a computer to recognize an image. A computer may detect objects in an image, be able to separate objects from a background, recognize the type of object, and verify that one object in an image matches with another. The general procedure for a computer to learn from an image is to have a set of images and a certain label associated with each image. This label tells the computer whether a certain property is present in the image or not. For example, images may or may not contain a face. The computer then extracts a set of features from the images. The method for extracting those features is different for different algorithms. Supervised learning is generally used to map the features and the associated label.

Image understanding is a core problem in the field of computer vision. Compared to object detection techniques focusing on the 'what is where' problem, we are more interested in mining the semantic hierarchy of object compositions and exploring how these compositions/sub-compositions are organized in an object. Image understanding must solve many more tasks, such as object recognition, object detection, attribute classification, and so on.

It is obvious that to answer any type of question, the generated dataset from an image must necessarily be multidimensional and large enough to answer questions from a broad range of possible ones. Because of this, training VQA systems takes longer when computer to CV systems trained on the same database.

C. Knowledge Representation

Storing member-class relations as 'knowledge' among different neurons mirrors the human brain's method of perceptual-conceptual storage. The ultimate goal of knowledge representation is to give computer systems the ability to use 'common-sense' when asked a question. When faced with a question, the computer uses its knowledge to generate an answer. This knowledge can be objects, procedures, relations, mental states, metaknowledge, and so on. Knowledge representation systems need to adequately represent the required knowledge for a question, be able to manipulate knowledge to produce new knowledge, direct their methods of inference into new productive methods, respond quickly even with limited time and resources, as well as have the ability to gain new knowledge.

The central component of all knowledge representation systems is the knowledge base of the system. Within this base, the system's beliefs, the system's truths, and the available knowledge is stored. Formal knowledge representation systems are used by knowledge recognition systems in order to learn quickly. However, for these systems to work, encoding the available knowledge into a quickly-readable format is required.

III. VQA Datasets

Datasets for VQA have increased in size over the years. However, many of them still follow the same basic fundamentals. Every element of the data set has, at minimum, an image, a question, and a correct answer to that question. Generally, these datasets can be generated automatically, but more recent datasets have been generated through crowdsourcing. This is a time-consuming operation and the recent increases in size represent a major breakthrough in the field of VQA. Datasets are designed for both evaluating and training VQA systems in a supervised setting. It is because of this supervised setting that the size of the datasets must necessarily be large. A 'good' dataset must be large enough to capture the long range of possibilities within questions and image content in real world scenarios. The Microsoft Common Objects in Contexts (COCO) collection of images is often the repository many datasets pull from. COCO contains $3.28 * 10^5$ images with 91 object types 'easily recognizable' by an average 4-year old, with a total of $2.5 * 10^6$ labeled instances.

COCO can simplify and accelerate VQA dataset building. However, collecting questions that are not ambiguous, cover a large variety of topics, and are convenient for computer systems to process is a difficult task. A good dataset will contain questions that are varied and precise, while also not being biased. For example, if a dataset contains only yes/no questions with a large amount of the answers being 'no', the most frequent class approach will give the answer as 'no' without having to do any VQA processing.

Publicly available datasets vary widely in three main ways. They can have different sizes i.e. the amount of images, questions, and concepts represented. The second way is the amount of reasoning required. For example, can the question be answered by detecting a simple object? If not, does the algorithm need to use multiple concepts and facts to arrive at an answer? The third point to consider is how much information beyond the information present in the actual image is necessary; be it common sense or subject-specific information. Reviews conducted in the field point out that many existing datasets have visual questions. Since state-of-the-art VQA algorithms still struggle with answering these types of questions, the field has a long way to go.

Datasets that are large enough with significant variability must also support a fair evaluation scheme for different VQA models. Minimally biased datasets are also required. VQA algorithms require the ability to recognize objects, attributes, and spatial relationships. The algorithms must also be able to count, logically infer things, and identify relationships between objects. VQA algorithms must also be able to leverage real-world knowledge.

Available VQA datasets can be categorized based on the types of images, the question-answer format, and the use of external knowledge. The types of images available are also variable. These images can be natural, they can be clipart, or can be generated by other algorithms (synthetic images, such as those used to distinguish humans from robots). Question-answer formats can include open-ended question, multiple-choice answers, binary questions, or fill-in-the-blank questions.

The first dataset created for VQA systems was the DAQUAR (DAtaset for QUestion Answering on Real-world images) set. There are 6794 training and 5674 test question-answer pairs, based on images from the NYU-Depth V2 Dataset. However, the major drawback of this dataset is that all of the images are of indoor scenes with varied lighting. The latter makes evaluation difficult, as human beings can answer questions with roughly 50% accuracy. The other drawback of this dataset is that it is too small.

The COCO-QA dataset, an evolution of the COCO dataset, contains 123,287 images coming from the COCO dataset, 78,736 training and 38,948 testing QA pairs. The questions for this dataset were generated using NLP algorithms. It should be noted, however, that all of the questions have a single-word answer. For example, 'what is the color of the bus' has 'yellow' as the answer. The major drawback of this dataset is that the generated questions suffer inherently from NLP generation oddities. In some cases, the questions are simply incorrect. In other cases, the questions aremake no grammatical sense. The dataset also only has four types of questions, and the questions themselves are not equally distributed. Roughly 70% of the questions are based on objects in the image, 17% on color, 7% on counting and 6% on color.

The most common dataset used is the one generated by researched at Virginia Tech. This dataset is simply referred to as the VQA dataset. There are two main parts to this dataset. There are 204721 COCO images which make up the *VQA-real* part of the dataset. There are also $5*10^4$ clip-art images which make up the *VQA-abstract* part of the dataset. VQA-real has 123,287 training and 81,434 test images. Humans generated interesting short questions for these images.Therefore, this dataset allows evaluation in multiple-choice and open-ended scenarios. In the latter case, 17 incorrect answers are also provided for each question. Estimates show that for roughly 18% of the questions, some level of common sense is required, whereas 5.5% of the questions require adult-level knowledge. Therefore, purely visual information might be enough to answer most questions. To evaluate a correct answer in open-ended mode, 3 different instances have to arrive at the same answer. In multiple-choice mode, 10 of the possible 18 options can be marked as *plausible*. However, in this case, there is a drawback. Some questions that simply do not make sense can have plausible answers. For example, an image containing an airplane can have the question 'would you like to fly in that?' which can have 'yes' as one of the plausible answers. To overcome this bias, VQA-abstract was created. It should be noted that VQA-real was recently updated and *VQA2.0* was released in another attempt to overcome bias.

VQA-abstract consists of abstract scenes featuring 20 'paper-doll' human models of different races, ages, and genders. There are 8 different possible expressions for each doll. The set also contains 100 objects and 31 animals in different poses. This set was created to overcome another bias inherent in real-world images; real-world images require questions that have a high level of abstraction.

Synthetic images can give great control over what kind of information is presented to the algorithm. Synthetic images can also be attached to XML files containing information about the position of different objects, the scale of the image, and so on. These types of datasets allow the VQA algorithm to focus on high-level semantics rather than waste time visually analyzing the image.

FM-IQA (Freestyle Multilingual-Image Question Answering) consists of COCO images and freestyle, interesting and diversified set of questions which requires a lot of reasoning abilities to answer them. They categorized questions into 8 types: including questions on object actions, object classes and others. Each image has at least two question-answer pairs as annotations.

One of the largest datasets in use today is the Visual Genome dataset. It has $1.7*10^6$ question-answer pairs. Images from the Visual Genome project are used. These images include structured annotations of scene contents in the form of scene graphs. Scene graphs describe the visual elements of each image, their attributes, and the relationships between them. Questions for the images in these dataset come from the seven 'Ws' used in journalism - what, who, where, when, why, how, and which. The 1000 most frequently-given answers correspond to 64% of the correct answers in the dataset. This is in stark contrast to VQA-real, which has a corresponding percentage of 90%. A subset of the Visual Genome dataset, Visual7w, allows for evaluation in multiple-choice settings.

A special type of the Visual7w dataset is the Zero-shot VQA version. In this version, the training/test splits include questions that have no relation to what 's in the images. The question 'how many zebras in the image' can appear for an image containing no zebras. In this case, the algorithm must be able to realize that the question pertains to information beyond the image i.e. that there are no zebras present.

The SHAPES dataset contains shapes of different sizes and colors in different arrangements. The questions in this dataset consist of complex questions about spatial and logical reasoning among multiple shapes. Because of this, learning biases are avoided. A similar dataset is the CLEVR (Compositional Language and Elementary Visual Question Reasoning) dataset. It contains simple 3D objects arranged in much the same fashion. CLEVR is supposed to allow for deeper analysis of visual reasoning abilities of any solution model.

The Knowledge-Based VQA (KB-VQA) dataset is a dataset that is supposed to test an algorithm's ability to answer questions based on higher-level adult reasoning and explicit knowledge. Labels have been attached to each question that estimate the level of knowledge required to answer it. These labels can be 'visual' (the question can be answered directly using visual concepts), 'common-sense' (the question does not require looking up information to answer), or 'KB-Knowledge' (the system must look up the information on a website such as Wikipedia). Fact-Based VQA (FB-VQA) attaches knowledge to the labels themselves in triplets, such as —Fish,CapableOf,Swimming—.

Another VQA dataset based entirely on figures is the FigureQA dataset, containing line plots, dotted line plots, vertical bar graphs, horizontal bar graphs, and pie charts. There are 15 types of questions that ask about the relation between different elements in the graph. For example, the maximum and minimum of a graph, the area under the curve, or the intersection is asked. Data VQA (DVQA) is a dataset developed with FigureQA but tests only bar-graph aspects.

Attempts have been made to create VQA images with video. Datasets have been created using 10^5 videos with $4 * 10^55$ questions attached to them. It has been proposed that the algorithm should be able to answer questions by analyzing the video i.e. answer questions about a movie it has just watched.

IV. VQA PROCEDURES

The general approach to a VQA problem is image and question featurization, joint comprehension, and image generation.

A. Image Featurization

Image featurization is the process of describing the image as vector withso that different mathematical operations can be applied to it. One of the simplest possible vectors is an RGB vector. Other transforms that can be applied are the Scale-Invariant Feature Transform, HAAR Transform, Histogram of Oriented Gradients, and others. However, with the advent of deep learning, the neural network itself creates features. Pre-trained models help in featurizing an image easily. The neural network generally chosen for this is a Convolutional Neural Network (CNN). Successful CNN models include AlexNet, ZFNet, VGGNet, ResNet, and GoogleNet. Out of these, ResNet has to lowest error rate of 3.54%. lower than human beings. State-of-the-art VQA systems use CNNs with their last layer removed. Sometimes, the results are normalized and the dimensions are reduced to represent feature data as a numerical vector. VGGNet and ResNet have been the most successful in the field so far. The main drawback of ResNet is that it requires high computational power, which may not be available to the average user. The motivation for a pretrained network is to take advantage of the vast amounts of training data available for image recognition, relative to the amounts of data annotated for VQA. The pretrained network is used as a generic feature extractor, by discarding the final classification layers, and using the features produced within the CNN prior to this classification.

B. Question Featurization

The input question must be processed to obtain a fixed-size representation of its content. Every word in the question is represented as an input vector which then refers to the lookup table. Another implementation involves having a one-hot vector, which then is multiplied by a dense-weight matrix that contains the embedding of all words. All word vectors are then collapsed into a single vector. TO do this, algorithms such as the Bag of Words algorithm are used, which take the average of all the word vectors. Another more sophisticated option is to feed the word vectors into Recurrent Neural Networks (RNN) such as Long Short-Term Memory (LSTM) networks. RNNs process words sequentially and capture relationships between them. The main drawback of one-hot encoding is that similarity in words is ignored.

Practical approaches convert co-occurrence counts of all words to probabilities and assign those probabilities to a co-occurrence matrix. To reduce the computational power required, the co-occurrence matrix is then converted to a low-rank approximation by using singular value decomposition.

To use directly learn word representation, neural networks are used. GoogleNet uses a skip-gram model to represent words, whereas the Continuous Bag-of-Words (CBOW) is used for other algorithms. CBOW networks predict the next word given a bag of context words and a string which to analyze next. Skip-gram models predict context words on both sides of the input string.

Count-based methods rely on the co-occurrence of words globally. Prediction based methods learn word representations using co-occurrence information. Hybrid methods combine these two methods in order to produce a vector called a 'Global Vector' that is then used to generate word embeddings.

Recent advances have used different types of LSTM networks and Gated Recurrent Unit (GRU) networks in order to featurize questions.

C. Joint comprehension of image and text

Historically, images and text have been processed separately by VQA networks. Joint comprehension is the next logical step and has a variety of methods.

The basic methods used are concatenation, element-wise addition and multiplication, and dot products of similar regions. These methods are all 'classical' and are subject to mathematical rigidity. Newer methods include models that are trained end-to-end with neural networks. However, these models are structured differently and there is no general flowchart to depict them.

Basic end-to-end neural network models use the one-layer removed method discussed above with slightly different activation functions. Slightly advanced neural network models use Multimodal Compact Bilinear Pooling. Other models use single CNNs where weights are determined by custom Dynamic Parameter Prediction Networks (DPPNs). Deep residual networks work on the intuition that deeper versions of good shallow network could learn identity transformations in the new layers.

An emerging area of interest is the *joint attention* version of VQA training. VQA networks generally ignore the semantic relationship between image attention and question attention. Semantic cross-model correlation along with attention has produced some interesting results in attempts to solve the VQA problem. Joint attention is a subclass of *attention models*, which are advanced VQA techniques.

Humans may swiftly comprehend visual representations by focusing on certain areas of the image rather than taking in the entire scene at once. Deep neural networks have been successfully employed for machine translation, reading comprehension, textual question answering, object recognition, and picture captioning. Most contemporary VQA models also make use of this technique. Allowing the model to concentrate on specific areas of the image is the main goal of attention mechanisms. Utilizing region-specific picture characteristics and integrating multiplicative interactions in the neural network work are key components of the technique.

V. Evaluation

A VQA system's output step can be viewed as either a generating or a classification task. The ability to create complicated sentences is a benefit of the production of a free-form response. But in reality, developing such a model is challenging. Short answers make up the majority of the data sets currently available, so learning a classifier is a useful alternative. The most frequent answers from the training set are used to predetermine a large collection of candidate replies for this purpose. This inevitably leaves out a few uncommon terms, but a set like this is usually enough to accurately answer more than 90% of test questions. Since this number is significantly higher than the precision of the systems in use, there is no limit to this.

The goal of computer vision research is to create models that can comprehend images similarly to humans. It has been suggested that computer vision systems use a visual Turing test. The majority of recent papers argue that VQA may be used as a substitute for the Visual Turing Test, or that it is a 'AI-complete' problem. Existence of a clearly defined quantitative evaluation metric to monitor progress is one of the crucial requirements for a work to be considered 'AI-complete'.

Open-ended and multiple-choice questions are both included in the VQA datasets. For each question in a multiple-choice situation, there is only one correct response. As a result, evaluating a suggested answer is simple because it is simple to calculate the mean accuracy across test questions. Due to synonyms and paraphrase, there is a chance that there will be more than one correct response to a question in an open-ended environment.

Some methods of VQA evaluation metrics include accuracy, WUPS, consensus, human judgment, MPT, BLEU, and METEOR.

VI. Research review, Industry Review, Importance, Future Scope, Applications

Due to its prospective uses and AI-completeness, has attracted the attention of numerous researchers. Knowledge of research in the fundamental problems of computer vision and natural language processing is a prerequisite for accomplishing this challenging challenge. A tremendous amount of research has been done in this area, leading to both the creation of new datasets and meteoric increases in the functionality of VQA algorithms. However, much work needs to be done in VQA research before it can do as well as humans do when answering questions based on images.

Visual Question Answering is a fledgling industry and is just beginning to be used in different industries. Some applications include virtual assistants which can interpret images from the camera. The medical industry can also use VQA to get information quicker. Assistance to the visually impaired in the form of Google Glass is another application. VQA has already been implemented in some form by social media websites such as Instagram, that analyze images in order to serve better advertisements. Virtual assistants can help answer questions on online commerce websites.

As a personal opinion, VQA is still very much in the development stage and a significant amount of research is needed before AI can answer anything a wide variety of questions about different images. At the moment, industries implement a reduced form of VQA, such as Google's virtual assistant, which can answer questions based on photographs by looking up things on Google. Other assistants such as Apple's Siri also work the same way.

VII. Open Source Implementations

Many open-source implementations exist. Some of the more important ones are: https://github.com/Cadene/vqa.pytorch,
https://github.com/akirafukui/vqa-mcb,
https://github.com/nerdimite/neuro-symbolic-ai-soc,
https://github.com/SHI-Labs/Interpretable-Visual-Reasoning,
https://github.com/cdancette/detect-shortcuts

and many more on https://paperswithcode.com/task/visual-question-answering link. However, I personally feel that many of these papers implement VQA in a very narrow sense, however sophisticated that sense may be. I have personally tried to run https://iamaaditya.github.io/2016/04/visual_question_answering_demo_not and gotten some success; however the execution time was variable and is probably indicative of a deeper problem in my Arch Linux installation.

References

References

[1] Damien Teney, Qi Wu, and Anton van den Hengel, Visual Question Answering: a tutorial, EEE: Signal Processing Magazine, 2017; 34(6):63-75, https://hdl.handle.net/2440/116146

[2] Manmadhan, S., Kovoor, B.C. Visual question answering: a state-of-the-art review. Artif Intell Rev 53, 5705–5745 (2020). https://doi.org/10.1007/s10462-020-09832-7

[3] Muralikrishnna G. Sethuraman, Ali Payani, Faramarz Fekri, J. Clayton Kerce, Visual Question Answering based on Formal Logic, arXiv:2111.04785

[4] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, Devi Parikh, Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering, arXiv:1612.00837