Visual Question Answering: Literature review and Exploration

Jiaming Yu
Boston University
jiamingy@bu.edu

***Abstract*— In this paper, we would review and explore VQA: Visual Question Answering. The first part of paper is the literature review of VQA, introducing VQA including its related work, common methods and dataset. The related work includes Visual Dialog, synthetic dataset CLEVR and FVQA. The common methods are divided into four categories: Joint embedding approaches, Attention mechanisms and Bottom-up and Top-down Attention, Compositional Models and Models using external knowledge base according to a 2017 paper. Also this paper introduces dataset like VQA1.0, 2.0, Tally-QA and so on. This paper also reviews and reproduces others' approach on VQA. In this paper, we also explore the user cases of VQA like aided-navigation for blind individuals, image retrieval,automatic querying and so on. Besides, we analyze the current VQA and future work.**

I. INTRODUCTION

In the field of Artificial Intelligence (AI), Computer Vision (CV) and Natural Language Processing (NLP) have been some of the most researched problems. Computer vision is a field of artificial intelligence that trains computers to interpret and understand the visual world. Using digital images from cameras and videos and deep learning models, machines can accurately identify and classify objects and then react to what they "see." In short, Computer Vision teaches machines "how to see". [12] Natural Language Processing concerns with the interactions between computers and human (natural) languages, in particular how to program computers to process and analyze large amounts of natural language data, i.e. Natural Language Processing teaches machines "how to read". They used to develop separately. Recently, a new area named Visual Question Answering combines Computer Vision and Natural Language Processing and push boundaries of both fields.

Visual Question Answering (VQA) is defined as a free-form and open-ended task as is shown in Figure 1, which is to provide an accurate natural language answer given an image and a natural language question about the image. [4] That is to say, a VQA system takes an image and a natural language question about the image as input and then process a natural language answer as the output.Since questions about the image are always trying to seek specific information like counting objects in the image, simple one to three word answers are sufficient for many questions, therefore, the common form of VQA answer is typically a few words or a short pharse. [4] In contrast to the traditional problems like segmentation or object detection, where the question to be answered by an algorithm is determined and only the input image changes. VQA will take the question as unknown. Thus it requires much more understanding of the general image.

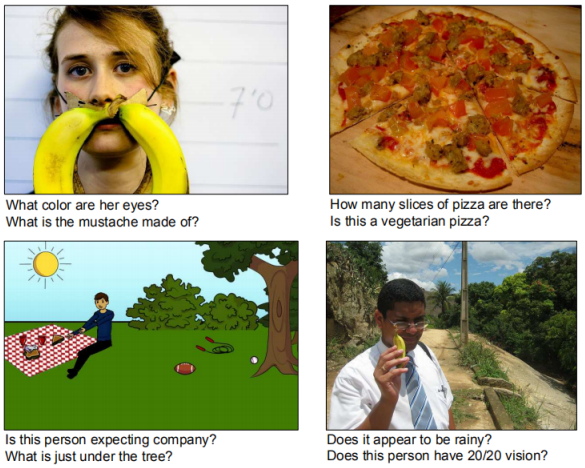


Fig. 1: VQA examples [4]

One of the related work to VQA is text based QA which is a well studied problem in the NLP. One key concern in text is the grounding of problems and thus inspire VQA. VQA is naturally grounded in images, requiring the understanding of both the text and image. [4] And the challenge would be added due to more dimension and noise. Besides, images lack the structure and grammatical rules of language, and there is no direct equivalent to the NLP tools such as syntactic parsers and regular expression matching. Natural language represents higher abstraction than image.

Compared to computer vision, VQA always requires information not present in the image, and thus more complex. The extra information usually range from common sense to encyclopedic knowledge. Luckily, since the answer required by VQA is usually a few words, it may be easier to evaluate.

A. Related Work

The existence of mature techniques in computer vision and NLP and the availability of large-scale datasets have driven the increasing interest in VQA. Therefore, there are many research related to VQA over the last few years. In this section, we would introduce some related work.

One is the Visual Dialogue [5] as is shown in Figure 2, which requires an AI agent to hold a meaningful dialog with humans in natural, conversational language about visual content. Given an image and a history of a dialog consisting of a sequence of question-answer pairs, and a natural language follow-up question,the task for the machine is to answer

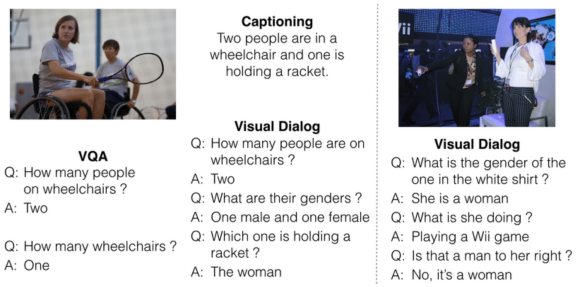
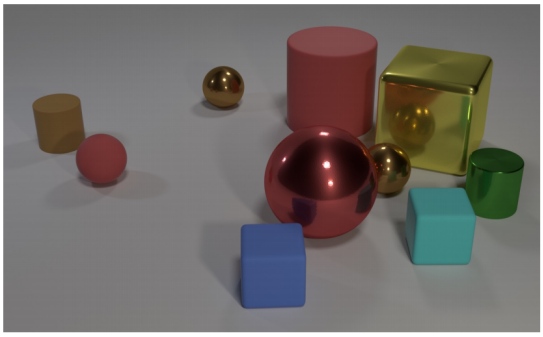


Fig. 2: Visual Dialog [5]



Q: Are there an equal number of large things and metal spheres? **Q:** What size is the cylinder that is left of the brown metal thing that is left of the big sphere? **Q:** There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere? **Q:** How many objects are either small cylinders or metal things?

Fig. 3: CLEVR image [7]

the question in free-form natural language. Therefore, the agent has to ground the question in image, infer context from history, and answer the question accurately. The main difference between Visual Dialogue and normal VQA is the dialogue input.

There is also some research on the VQA dataset. For example, a little different from free-form and open-ended VQA, it introduces one kind of synthetic dataset CLEVR in 2017 using synthetic images and synthetic questions. [7] The difference between synthetic dataset and normal VQA is that it has clean data, predictable number of patterns and less noise. It is designed to analyze VQA model reasoning due to its minimal bias. The image is generated by randomly sampling the scenes which contain objects with random positions, shapes, colors, sizes and materials as is shown in Figure 3.

Another related work to VQA is Fact-based Visual Question Answering (FVQA). [11] As is shown in Figure 4, FVQA always requires much deeper reasoning since if the questions requires no external information would be quite limited. And thus FVQA normally requires common sense knowledge. From large-scale structured knowledge bases, FVQA automatically find the supporting-fact for the visual question. And then FVQA answers the question With the help of common sense knowledge as the supporting fact. Recent development is that questions in FVQA require multi-entity, multi-relation, and multi- hop reasoning over large Knowledge Graphs (KG)

Question: What can the red object on the ground be used for ?
Answer: Firefighting
Support Fact: Fire hydrant can be used for fighting fires.

Fig. 4: FVQA [11]

to arrive at an answer. The most recent FVQA dataset in 2019 contains 24,000 images with 183,100 question-answer pairs employing around 18K proper nouns.

These related work are somehow a little different from the definition of free-form and open-ended VQA. However, they are also significant part of VQA.

B. Paper Organization

The contribution of this paper is to literature review VQA: Visual Question Answering, discussing the related work, common methods and dataset. Also this paper experiments others' work on VQA and explore the user stories and next path.

The paper is organized as follows. The introduction of VQA methods and datasets would be discussed in section 2 and in section 3, we would reproducing others' approach related to VQA. The exploration results shown in section 4 discusses about the user cases, current and future work on VQA.

II. Visual Question Answering Methods and Dataset

A. VQA Methods

In this section, we would introduce some common VQA methods. In [12], those approaches are divided into four categories: Joint embedding approaches, Attention mechanisms, Compositional Models and Models using external knowledge base.

1) Joint embedding approaches: The concept of jointly embedding images and text was first explored for the task of image captioning.It learns the representations in a common feature space of Computer Vision and NLP. Practically, image representations are obtained with convolutional neural networks (CNNs) pre-trained on object recognition while text representations are obtained with word embeddings pre-trained on large text corpora. For example, these features are produced by deep convolutional and recurrent neural networks. They are combined in an output stage, which can take the form of a classifier (e.g. a multilayer perceptron)

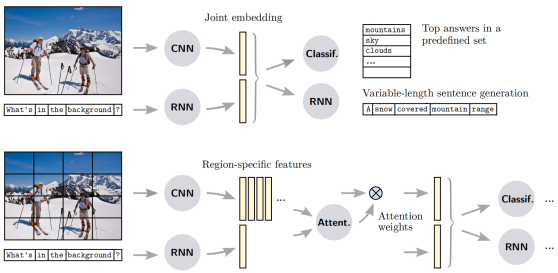


Fig. 5: Joint embedding without Attention (Top) and with Attention (Bottom) [11]

to predict short answers from predefined set or a recurrent network (e.g. an LSTM, a special kind of RNN) to produce variable length phrases.

2) Attention and Bottom-up, Top-down Attention: Attention mechanism [13] has now been quite common in the visual tasks field. The aim of attention mechanisms is to address this issue by using local image features, and allowing the model to assign different importance to features from different regions. Since Xu has been successfully applied Attention mechanism in Image Captioning, Attention has been quite popular to be applied to VQA. Clearly, Attention mechanism can improve the overall accuracy of all VQA datasets by forcing an explicit additional step in the reasoning process that identifies "where to look" before performing further computations. But the study also shows attention may help little on binary (yes or no) questions.

In [2],there is one step further on Visual Feature for this method. Compared to the previous work, where attention is on the uniform grid regions, this method, bottom-up and Top-down Attention focus on the level of objects and other salient image regions, as is shown in Figure 6. In this approach, the bottom-up mechanism (based on Faster R-CNN) proposes image regions, each with an associated feature vector, while the top-down mechanism determines feature weightings. And thus bottom-up and top-down attention has been more and more popular due to better region features.

3) Compositional Models: This approach involves connecting distinct modules designed for specific desired capabilities such as memory or specific types of reasoning. Typically, there are two types: Neural Module Networks (NMN) and the Dynamic Memory Networks (DMN)

Neural Module Networks [3] are specifically designed for VQA, with the intention of exploiting the compositional linguistic structure of the questions. Questions vary greatly in the level of complexity. For example, NMN for answering the question What color is his tie? The attend[tie] module first predicts a heatmap corresponding to the location of the tie. Next, the classify[color] module uses this heatmap to produce a weighted average of image features, which are finally used to predict an output label. Nerual Module Networks seems to be working well on synthetic datasets but so far not so good at handling VQA in real world image. But the potential of this method is quite encouraging. The approach is shown in

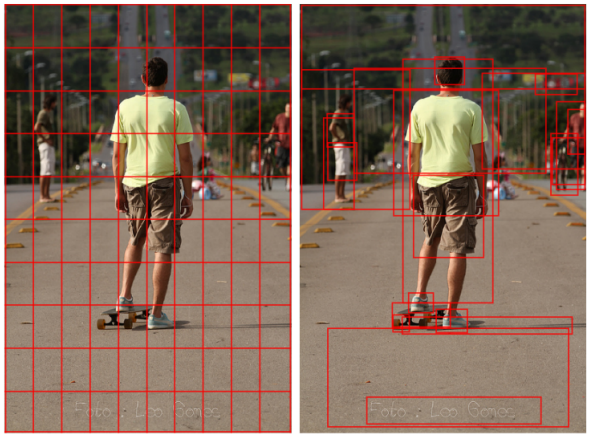


Fig. 6: Typical attention models operated on CNN features corresponding to a uniform grid of equally-sized image regions (left). Bottom-up and Top-down Attention enables attention to be calculated at the level of objects and other salient image regions (right) [2]

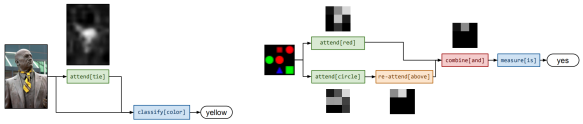


Fig. 7: The approach of Neural Module Networks [3]

Figure 7.

Dynamic Memory Networks are composed of 4 main modules: the input module which transforms the input data into a set of vectors called "facts"; the question module which computes a vector representation of the question, using a gated recurrent unit (GRU, a variant of LSTM); the episodic memory module retrieves the facts required to answer the question; the answer module uses the final state of the memory and the question to predict the output with a multinomial classification for single words, or a GRU for datasets where a longer sentence is required.

4) Models using external knowledge bases: The task of VQA always requires external common sense knowlege and thus the development of large-scale Knowledge Bases (KB) would be necessary. Linking such knowledge bases to VQA methods is proved to increase the average accuracy.

5) Learning to count objects: In [14], Visual Question Answering (VQA) models have struggled with counting objects in natural images so far. Typically, the fundamental problem is the soft attention.Yan Zhang has proposed a method using VQA v2 dataset with the bottom-up and top-down object proposal features, which provides a better result in counting performance. But it also suggests that the current attention mechanisms and object proposal network are still very inaccurate, and provides further evidence that the balanced pair accuracy is maybe a more reflective measure.

Dataset	# Images	# Questions	Question Type(s)	Venue	Model(s)	Accuracy
DAQUAR [8]	1449	12468	Object Identification	NIPS 2014	AutoSeg [15]	13.75%
VQA [1]	204721	614163	Combining vision, language and common-sense	ICCV 2015	CNN + LSTM	54.06%
Visual Madlibs [16]	10738	360001	Fill in the blanks	ICCV 2015	nCCA (bbox)	47.9%
Visual7W [9]	47300	2201154	7Ws, locating objects	CVPR 2016	LSTM + Attention	55.6%
CLEVR [17]	100000	853554	Synthetic question generation using relations	CVPR 2017	CNN + LSTM + Spatial Relationship	93%
Tally-QA [10]	165000	306907	Counting objects on varying complexities	AAAI 2019	RCN Network	71.8%
KVQA [11]	24602	183007	Questions based on Knowledge Graphs	AAAI 2019	MemNet	59.2%

Fig. 8: Overview of Some VQA datasets [10]

B. Dataset

1) DAQUAR: The first VQA dataset designed as benchmark is the DAQUAR, for dataset for Question Answering on Real-world images. The images of DAQUAR are split to 795 training and 654 test images. Two types of question/answer pairs are collected. First, synthetic questions/answers are generated automatically using 8 predefined templates and the existing annotations of the NYU dataset. Second, human questions/answers are collected from 5 annotators. They were instructed to focus on basic colors, numbers, objects (894 categories), and sets of those. Overall, 12,468 question/answer pairs were collected, of which 6,794 are to be used for training and 5,674 for testing.

2) COCO-QA, FM-IQA and Visual Madlibs: The COCO-QA dataset represents a substantial effort to increase the scale of training data for VQA COCO-QA includes 123,287 images (72,783 for training and 38,948 for testing) and each image has one question/answer pair.

The FM-IQA (Freestyle Multilingual Image Question Answering) dataset sourced from the COCO dataset and has 123,287 images. The difference from COCO-QA is that the questions/answers are provided here by humans through the Amazon Mechanical Turk crowd-sourcing platform.

The Visual Madlibs dataset is designed to evaluate systems on a "fill in the blank" task. The dataset comprises 10,738 images from COCO and 360,001 focused natural language descriptions.

3) Visual7W: The Visual7w dataset is a subset of the Visual Genome that contains additional annotations. Visual Genome is, at this time, the largest dataset for VQA with 1.7 million question/answer pairs. The questions are evaluated in a multiple-choice setting, each question being provided with 4 candidate answers, of which only one is correct. The special part of this dataset is that all the objects mentioned in the questions are visually grounded. That is to say, there are bounding boxes of their depictions in the images and thus objects are located.

4) VQA 1.0 and VQA 2.0: One of the most widely used dataset comes from the VQA team at Virginia Tech, commonly referred to simply as VQA. VQA 1.0 provides a dataset containing 0.25M images, 0.76M questions, and 10M answers. It comprises two parts, one using natural images named VQA-real, and a second one with cartoon images named VQA-abstract. VQA-real comprises 123,287 training and 81,434 test images, respectively, sourced from COCO. And there are at least 3 questions (5.4 questions on average) per image, 10 ground truth answers per question, 3 plausible (but likely incorrect) answers per question and automatic evaluation metric in VQA dataset. It has combined vision, language and common sense.

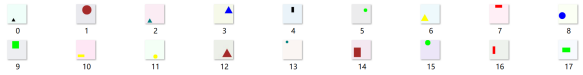


Fig. 9: Easy-VQA images

```
user@DESKTOP-IF638D3 MINGW64 /e/VQA/EasyVQA
$ python gen_data/generate_data.py
Generated 4000 train images and 38222 train questions.
Generated 1000 test images and 9768 test questions.
13 total possible answers.
28044 training questions are yes/no.
7202 testing questions are yes/no.
```

Fig. 10: Generate Easy-VQA dataset

VQA 2.0 in 2017 [6] is by construction more balanced than the original VQA dataset and has approximately twice the number of image-question pairs.

5) CLEVR: As is said in related work section, CLEVR in 2017 [7] is also a dataset designed for VQA. The point is that it is a synthetic dataset. The dataset is generated using three objects in each image, namely cylinder, sphere and cube. These objects are in two different sizes, two different materials and placed in eight different colors. The questions are also synthetically generated based on the objects placed in the image. Thus it reaches a very high accuracy of 93 percent

6) Tally-QA: Very recently, in 2019, the Tally-QA dataset [1] is proposed which is the largest dataset of object counting in the open-ended task. The dataset is quite large in numbers as well as it is 2.5 times the VQA dataset. The dataset contains 287,907 questions, 165,000 images and 19,000 complex questions.

III. Reproducing VQA implementations and Experiment

A. Easy-VQA

In this section, I have reporduced the result of Easy-VQA approaches from others. The easy-VQA dataset is a small dataset which has quite simple images along with simple questions. In the images, there are only several randomly generated colorful shapes,just like a simplified 2D CLEVR dataset as is shown in Figure 9. And the answers are limited to yes/no, color and shapes. We can generate these images and questions/answers easily via python generate-data.py as is shown in Figure 10. Here we generate 80 percent (4k images) of this dataset to be training set and generate 20 per cent (1k images) to the testing set. In fact, we can change the number of training/testing set by changing NUM-TRAIN and NUM-TEST.

The approach for this dataset is to process images, process questions, combine features and assign probabilities to each possible answers. In this approach, we use TensorFlow, which comes packaged with Keras as its high-level API, and use Pillow for image processing. Since it is a simple dataset, CNN can be directly used via TensorFlow to process images. Although RNN like LSTM is more commonly used, since the

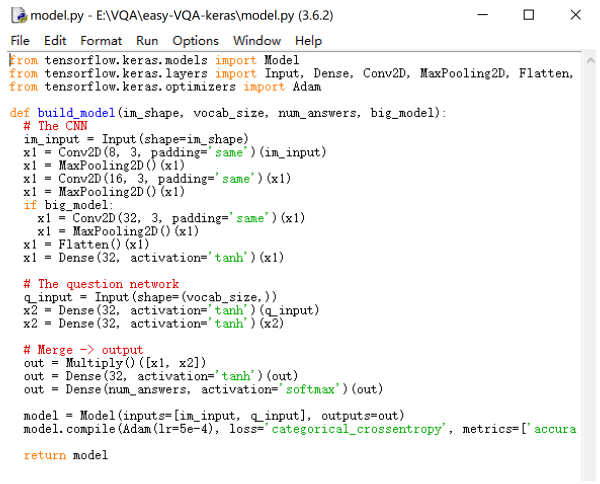


Fig. 11: Easy-VQA model code

questions are also very simple, we would choose a simpler method of turning questions to vector using a Bag of Words (BOW) representation. Keras's Tokenizer class in TensorFlow would help us with that. Then in the model, we would pass the BOW vectors into 2 FC neural network layers, merge image and vectors with element-wise multiplication and use softmax function to turn output value into probabilities. The code of model is shown in Figure 11. As is shown in Figure 12, after 8 epochs training (we can also change the number of epochs in train.py), the accuracy of the model reaches 76 per cent on testing set. Then we generate another different 4k images training set and 1k testing set and train the model again for 10 epochs. The accuracy reaches 84 percent. Then we generate 8k images training set and 2k testing set and train the model for 10 epochs. The accuracy reaches 99 percent.

Since it is a synthetic simple dataset, it is easy for the accuracy to reach very high like 99 percent. The demo of other's work which reaches 99 percent accuracy is created via create-react-app, TensorFlow.js and TensorFlow.js converter and is shown in Figure 13. To derive this demo, we could derive model.h5 after the above training and use TensorFlow.js converter to convert it to package.json. Then we just load it to create-react-app in Github.

In short, the approach of Easy-VQA is the joint embedding of CNN plus Bags of words representations turning questions to vectors using Keras in TensorFlow.

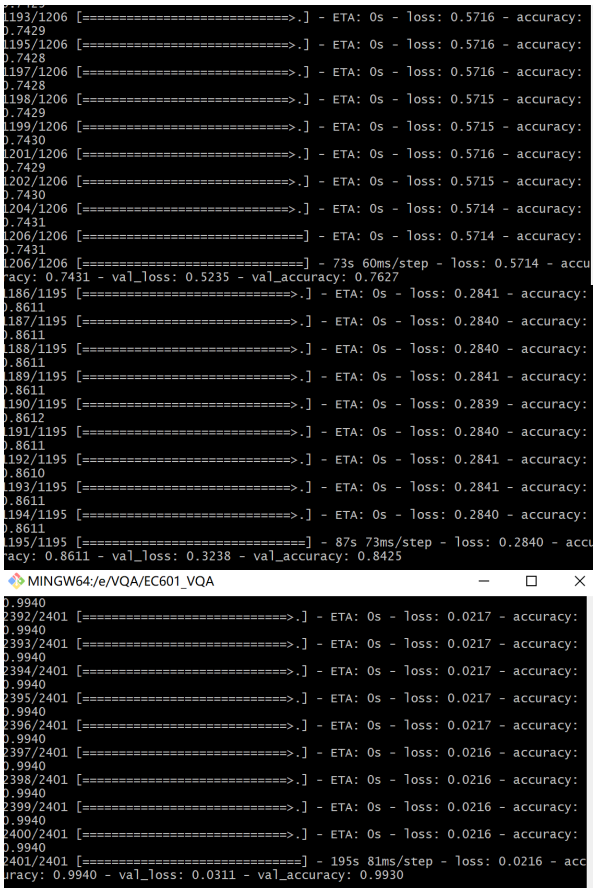


Fig. 12: Training process

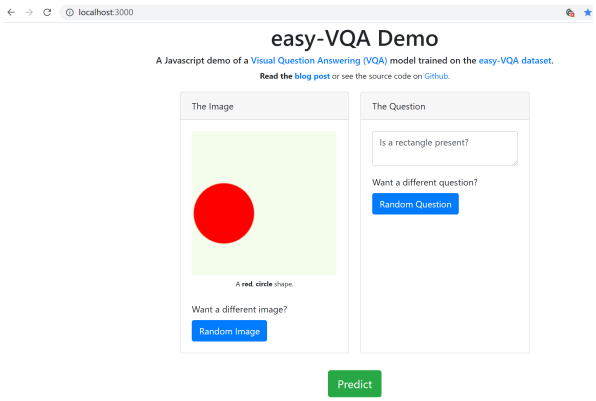


Fig. 13: Easy-VQA demo

IV. VQA Exploration

A. VQA user cases

There are many real life potential applications for VQA. VQA can be directly applicable to a variety of applications of high societal impact that involve humans eliciting situationally-relevant information from visual data; where humans and machines must collaborate to extract information from pictures.

1) aided-navigation for blind individuals: Probably the most direct application is to help blind and visually-impaired users. For example, VQA can aid visually-impaired users in understanding their surroundings such as "What temperature is this oven set to?". This project has the potential to fundamentally improve the way visually-impaired users live their daily lives, and revolutionize how society at large interacts with visual data.

2) image retrieval: Another obvious application is to integrate VQA into image retrieval systems. This could have a huge impact on social media or e-commerce.A VQA system could provide information about an image on the Web or any social media. VQA can also be used with educational or recreational purposes.

3) automatic querying: Besides, VQA can be used in automatic querying of surveillance video. Analysts can rely on VQA to help making decisions based on large quantities of surveillance data like "What kind of car did the man in the red shirt drive away in?"

4) Other applications: Considering when you are looking for something, you can interact with a domestic robot like "Is my laptop in my bedroom upstairs?". And then VQA can easily provide you with the answer and help you organize life better.

B. Current VQA and paths for moving forward

VQA is believed to have the distinctive advantage of pushing the frontiers on "AI-complete" problems, while being amenable to automatic evaluation. VQA is particularly attractive because it constitutes an AI complete task in its ultimate form, i.e. considering open-world free-form questions and answers.

Current VQA research is split into two camps: the first focuses on VQA datasets that require natural image understanding and the second focuses on synthetic datasets that test reasoning. A good VQA algorithm should be capable of both, but only few VQA algorithms perform very well in both situation. [9]

The Visual Question Answering has recently witnessed a great interest and development by the group of researchers and scientists from all around the world. The recent trends are observed in the area of developing more and more real life looking datasets by incorporating the real world type questions and answers. The recent trends are also seen in the area of development of sophisticated deep learning models by better utilizing the visual cues as well as textual cues by different means. The performance of even best model is still lagging and around 60-70 per cent only. [10]

Therefore, we had to admit that there is still a long way to go and thus reduced and limited forms of VQA, e.g. multiple-choice format, short answer lengths, limited types of questions, etc., are reasonable intermediate objectives that seem attainable. That is to say, so far the recent research is still far away from general AI. it is still an open problem to develop better deep learning models as well as more challenging datasets for VQA.

VQA is a recent field that requires the understanding of both text and vision. Since deep learning techniques are significantly improving NLP and CV results, we can reasonably expect that VQA is going to be more and more accurate in the next years. And it is very likely that new datasets and metrics will allow to deepen and refine the notion of quality. Different strategies like object level details, segmentation masks, deeper models, sentiment of the question, etc. can be considered to develop the next generation VQA models.

Future work on VQA involves the creation of larger and far more varied datasets. Bias in these datasets will be difficult to overcome, but evaluating different kinds of questions individually in a nuanced manner, rather than using naive accuracy alone, will help significantly. Further work will be needed to develop VQA algorithms that can reason about image content, but these algorithms may lead to significant new areas of research. [8]

V. CONCLUSIONS

The related work of VQA includes Visual Dialog, synthetic dataset CLEVR and FVQA. The common methods are divided into four categories: Joint embedding approaches, Attention mechanisms, Compositional Models and Models using external knowledge base according to a 2017 paper. We also discuss some interesting methods like Bottom-up and Top-down Attention. Also this paper introduces dataset like CLEVR, VQA 1.0, 2.0, Tally-QA and so on. We also reproduce the approach of Easy-VQA with a joint embedding of CNN plus Bags of words representations using Keras in TensorFlow in this paper. In this paper, we also explore the user cases of VQA like aided-navigation for blind individuals, image retrieval,automatic querying and so on. Besides, we analyze the current VQA and future paths. Current VQA research is split into two camps: the first focuses on VQA datasets that require natural image understanding and the second focuses on synthetic datasets that test reasoning. So far few methods work well on both and thus the recent research is still far away from general AI. Future work would involve the creation of larger datasets and the development of reasoning algorithms.

REFERENCES

[1] M. Acharya, K. Kafle, and C. Kanan. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084, 2019.

[2] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018.

[3] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein. Deep compositional question answering with neural module networks. corr abs/1511.02799 (2015). *arXiv preprint arXiv:1511.02799*, 2015.

[4] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[5] A. Das, S. Kottur, K. Gupta, A. Singh, D. Yadav, J. M. Moura, D. Parikh, and D. Batra. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335, 2017.

[6] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[7] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[8] K. Kafle and C. Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Computer Vision and Image Understanding*, 163:3–20, 2017.

[9] R. Shrestha, K. Kafle, and C. Kanan. Answer them all! toward universal visual question answering models. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 10472–10481, 2019.

[10] Y. Srivastava, V. Murali, S. R. Dubey, and S. Mukherjee. Visual question answering using deep learning: A survey and performance analysis. *arXiv preprint arXiv:1909.01860*, 2019.

[11] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.

[12] Q. Wu, D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[13] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[14] Y. Zhang, J. Hare, and A. Prügel-Bennett. Learning to count objects in natural images for visual question answering. *arXiv preprint arXiv:1802.05766*, 2018.