Visual Question Answering: A Review

Ishmam Tashdeed
Computer Science & Engineering
Islamic University of Technology
Email: ishmamtashdeed@iut-dhaka.edu

Md.Farham Ishmam
Computer Science & Engineering
Islamic University of Technology
Email: farhanishmam@iut-dhaka.edu

Asir Saadat Nipun
Computer Science & Engineering
Islamic University of Technology
Email: asirsaadat@iut-dhaka.edu

***Abstract*—In the past decade, Computer Vision (CV) and Natural Language Processing (NLP) have seen exponential advancements. Visual Question Answering (VQA) is just the culmination of ideas in combining the modalities of vision and language. VQA is a task in which an algorithm is presented with an image and a question about that image and the model has to generate an appropriate answer. To do so, the model needs to have a relevant understanding of the image along with a solid grasp on natural language. In this review, we discuss the techniques and algorithms used in modern VQA models along with the datasets and evaluation metrics of the algorithms that have been shaping up the VQA domain in the last few years alone. Furthermore, we introduce a new diagnostic sub-domain of color-based and shape-based bias in current VQA models. Finally, we conclude by discussing some of our approaches to handling these biases and the future of the field in general.**

I. Introduction

Visual Question Answering (VQA) is a unique problem which requires domain knowledge of both Image Processing and Natural Language Processing. Both of these domains have been predominantly using neural networks to solve research problems of the last decade. The approaches, however, differ as images widely use Convolutional Neural Networks (CNNs), while question-answering uses sequential models such as Recurrent Neural Networks (RNNs), LSTMs and Transformers. VQA tries to combine these fields of technology and produce a model which will be able to answer general questions given a piece of image. It is not difficult to comprehend the interest surrounding this field as it attempts to tackle a fundamental evaluation of intelligence.

VQA is a multimodal task which requires models to have a good understanding of both visual and linguistic modalities. Therefore, it is comparatively more challenging than just single modality tasks. Other domains such as Image Captioning [1, 2, 3], Video Captioning [4, 5], Text-to-Image generative methods [6, 7, 8, 9], Personal Assistant algorithms [10], Spoken Question Answering [11, 12, 13] and many more are examples of multimodal tasks along with VQA. Sometimes, additional textual context is given [14] along with the image and question which bolsters the "perceptivity" of the model. Furthermore, VQA can be used as a visual turing test [15, 16]. These systems can be used to solve various real-world problems such as assistance for visually impaired [17, 18, 19], Video analysis through natural language queries [20, 21, 22], Surveillance footage analysis [23, 24] and likewise others.

In the past decade, the approaches towards VQA developed rapidly from conditional models on Bayesian framework by [26] in 2014 to transformers-based models such as VisualBERT by Li et al [27]. Many other approaches have been used – all showed satisfactory results in their respective subdomains. However, the approaches, especially the papers addressed towards general VQA, fall short in certain types of task. General VQA is a research problem aiming to appropriately answer any type of question given any form of image, and are not limited to certain image, question, and answer constraints common to every other model. The shortcomings of the benchmark datasets in training a general VQA model are identified and addressed by specialized datasets, often called diagnostic dataset, which test the model's performance on a specific category of data (e.g., a certain image or question type). It is to be noted that the shortcomings addressed in this paper are related to the expected performance from a model that can perform general VQA to a certain degree *without* taking the help of an external knowledge base. E.g., in CLEVR [28], the diagnostic dataset presents the problem of state-of-the art models not being able to perform elementary visual reasoning. Knowing these limitations of the current model, pioneers the growth for novel architectures. Later, in this paper, we shall introduce the Color-Shape Bias as a potential area of interest for diagnostic datasets in VQA.

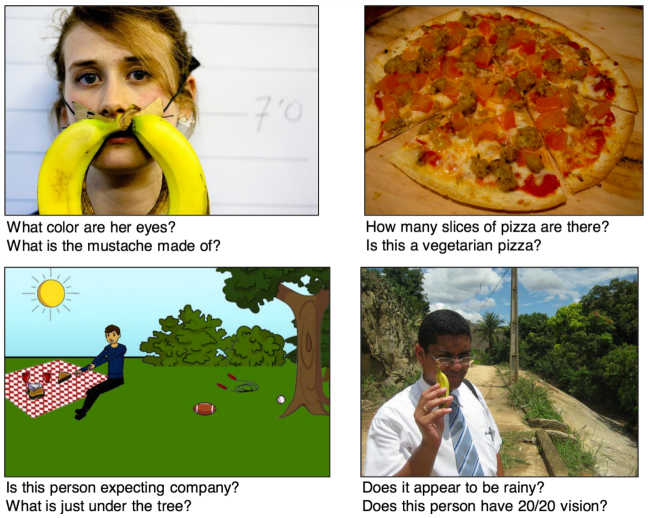


Fig. 1. Examples of questions from the VQA [25] dataset. These are open-ended questions which need a combined understanding of vision and language elements.

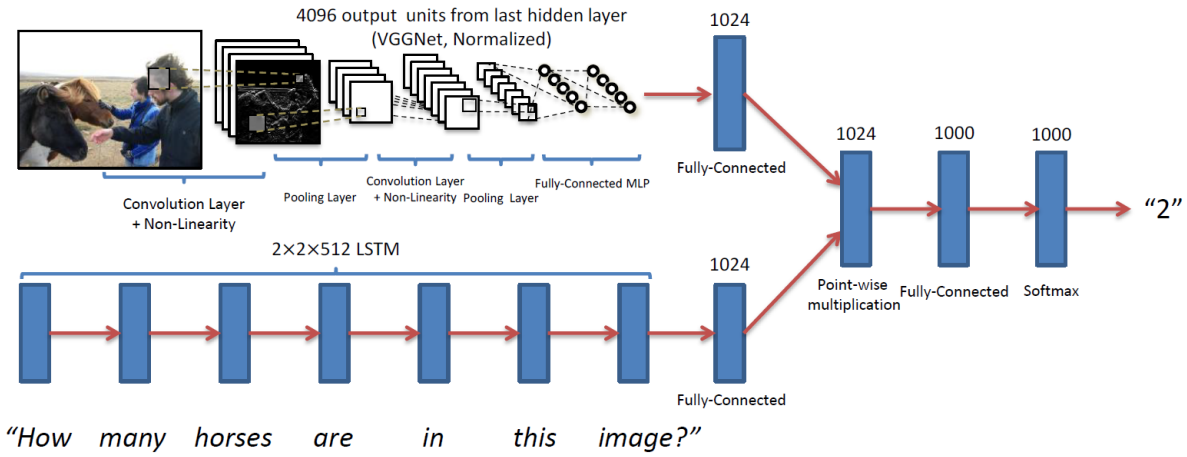


Fig. 2. This model from [25] uses a VGGNet [29] to capture image features and a two layer LSTM [30] to encode the question features. Then it point-wise multiplies them and puts the vector through two fully connected layer which select classify the correct answer from top 1000 answers.

II. Basics of VQA

VQA has been approached by a multitude of ways but the standard approach can be broken down into three separate phases: **Feature extraction, Feature conjugation and Answer generation.**

Generally, we have two separate streams to process the image and the question separately followed by some mechanism to unite these two representations which is then passed to a module that constructs the answer to the given question. Commonly, multiple fully connected layers that classify the correct answer from the top k answers from the dataset [25] as seen on 2. In the following subsections of this review, we will discuss the different methods for feature extraction, feature conjugation and answer generation.

A. Image Feature Extraction

Information of the image is captured in this process and a vector representation of the image is produced. The vector representation should hold an abridged form of information mapped to a space where similar features in different images produce vectors that are close-by. Thus, feature extraction must ensure that the semantic components of the image are conveyed to the model. Also, transforming it into a vector space ensures mathematical manipulation as well as an easier way to group similar images. Before the advent of deep learning models, methods like explicit RGB vector, SVM[31, 32], HAAR [33, 34], HOG [35, 36], SIFT [37], Singular value [38], PCA [39] or simply hand-crafting kernels to extract characteristics [40, 41] which are functional to a certain degree and perform quite well on specific tasks. Furthermore, these methods are easy to compute, but does not tend to generalize well. With the onset of deep learning, Neural networks [42, 43, 44] have become very popular among researchers to retrieve image features. But Convolutional Neural Networks (CNN) are far better at this than regular Feed Forward Neural Networks as shown in [45, 46]. As the years passed, these CNNs boasted deeper layers and larger parameter counts. Models like LeNet [46], AlexNet [45], VGG [29], ResNet [47], InceptionNet [48] are just a few examples — all fueled by the ImageNet [49] competition.

Almost all of the state-of-the art VQA algorithms use these CNN models for image feature extraction where the concept of transfer learning is used. Transfer learning is an early concept in machine learning research [50] which denotes how a learning model trained to do one task can transfer its "knowledge" stored in the gradients to another model being trained on another task [50] and is far quicker than training different algorithms from scratch. The features that the model learns in the lower layers are more primitive and can be used as a building block to capture more complex characteristics which may vary from task to task [50]. Transfer learning is particularly useful for models dealing with the visual modality as training such a model from scratch takes massive amount of data and resources — well documented in [48]. VQA models usually contain VGG [25, 51, 52, 53, 54] and ResNet [55, 56, 57, 58, 59, 60] CNN backbones pre-trained on Imagenet [49] or COCO [61] datasets. The usual approach is to freeze the final few layers of a pre-trained model and fine-tune by training the model on the target dataset [62] and thus, cutting down on training time, taking less resource and fewer data. The gradual shift towards VGG and ResNet provides some insight into the requirements of image feature extraction where simple models like AlexNet and LeNet are not sufficient and InceptionNet does not provide a considerable edge over others. In recent years, ResNet has become a baseline in the image pipeline for VQA models.

Another recent trend in computer vision is the emergence of visual transformers first introduced in [64]. The image is broken down into 16×16 image patches and is passed through a Transformer [65] architecture which uses multi-headed self-attention[64, 65]. The procedure was later perfected by Swin Transformers [63] which uses a shifting window to apply the

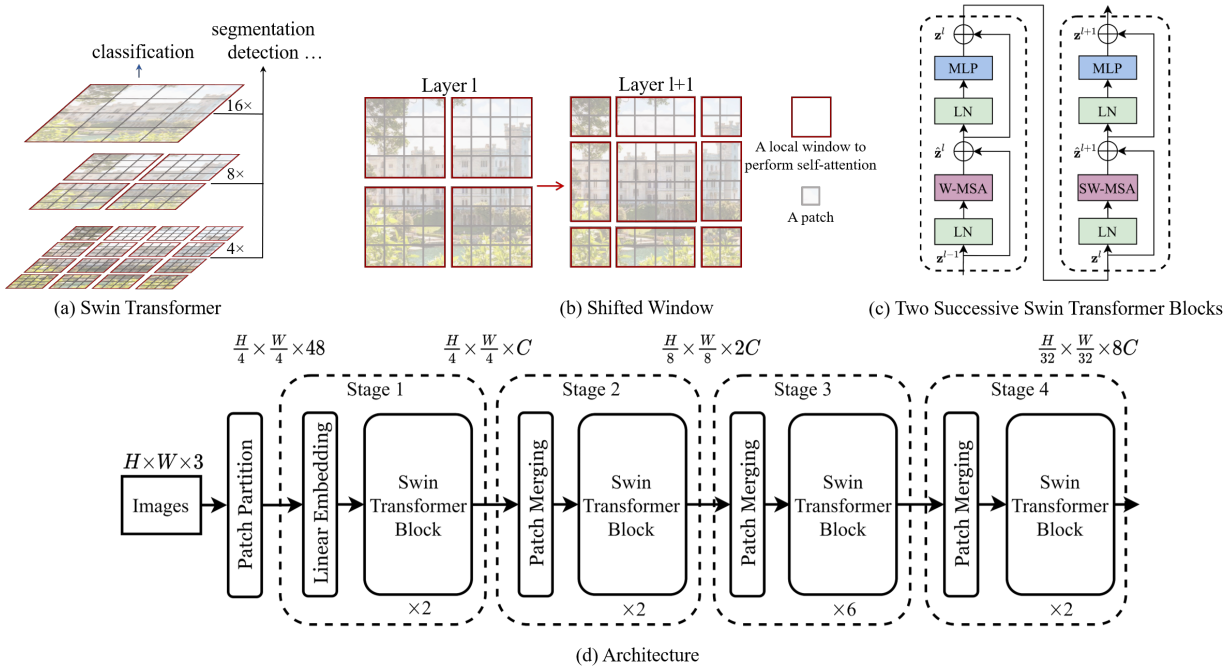


Fig. 3. The Swin Transformer architecture [63] (a) breaking down the image into regions called patches and self-attention is only applied to a small region within the window compared to VIT [64] which applies attention to all the image patches, it learns to merge image patches in deeper layers. (b) The window mechanism. (c) Two Swin Transformer blocks along with (d) that combines these blocks to create the full architecture.

self-attention mechanism mimicking the convolution operation of a CNN and offers relatively better performance in tasks which require more detailed feature extraction e.g., semantic segmentation and object detection. But convolutions are still relevant as seen on [66], the newly proposed ConvNext architecture can perform as well as or even better than transformer based models in tasks such as classification, object detection and semantic segmentation. The architecture borrows certain ideas from modern transformer models and applies them to the classic convolution based architecture. Recent VQA models are using these backbones for visual feature extraction as these methods are proven to be quite effective [67, 68, 69, 70]. Even though these techniques may adhere to better results, they are resource intensive and computationally expensive which might discourage researchers from using them at a greater scale.

B. Question feature extraction

Textual data from the question should be processed in such a manner that the output represents the meaning of the words in a vectorized manner. Researchers have been seeking ways to improve word representation in vector spaces which are also called Word Embeddings. Unlike images, words are not continuous, so representing them in a continuous vector space is a challenge. One of the more earlier approaches is to represent each word as a one-hot vector representation where each n word in the corpus can be represented using an n dimensional vector. The downside of this rather simple representation is that all the words are independent of each other with no cosine similarity i.e. the vectors are orthogonal to each other and equidistant. Additionally, the vector size depends on the vocabulary length, and the method does not account for typos and other misrepresentations. Count-based methods were also popular in the early days of NLP. Such methods rely on a matrix made by the frequency of word occurrences [71] which can be further broken down by computing the singular value decomposition (SVD) [72] and getting the best fit approximation of lower dimensions.

But these representations are still not good enough as they do not embody the semantic information stored in the words. In a proper Word Embedding, words with similar meanings should be nearby. The inability to convey this gave rise to prediction based models which would predict the *following* word given a *current* context. In this way, the model learns the word representations and expresses each word in an n dimensional space through capturing semantic information. Earlier work done using Neural networks [75, 76] paved the way for more complex models such as Recurrent Neural

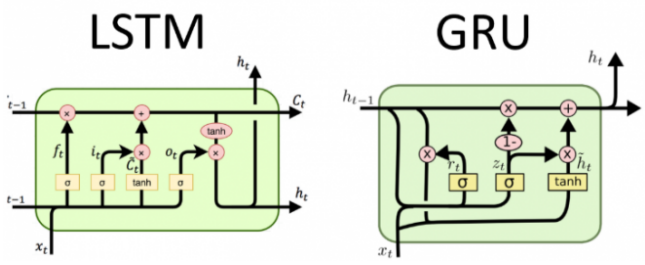


Fig. 4. LSTM [73] and GRU [74] architectures side by side.

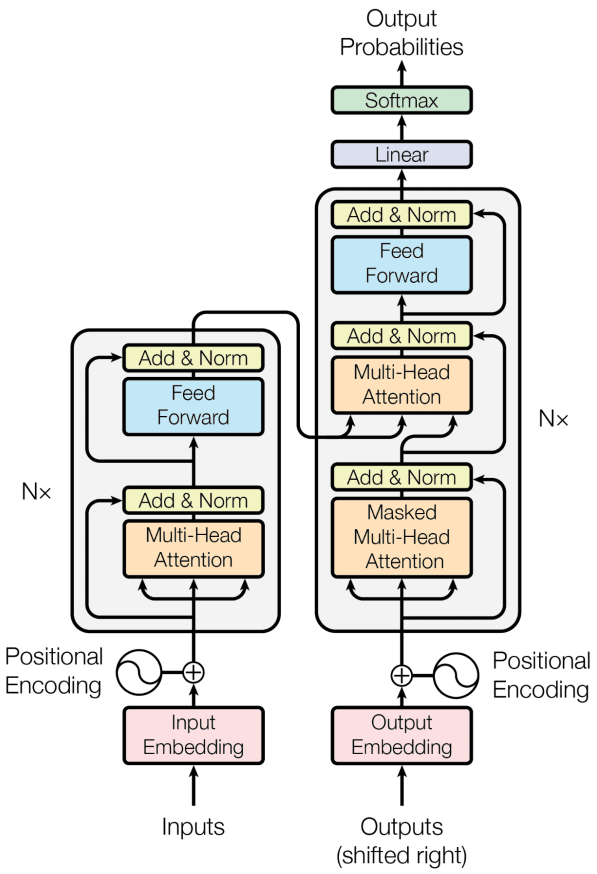


Fig. 5. A transformer [65] encoder and decoder block comprising of multi-headed self-attention. The implementation uses multiple of these encoder-decoder blocks for machine translation.

Networks (RNNs) to take over [77]. Methods like Word2Vec [78], CBOW and Skip-Gram [77] were developed to learn word representations. They are essentially prediction models trained to predict a masked word given the context surrounding it. By trying to predict certain words, the models learn how to represent those words in a vector space which is extracted and used to derive the features.

Further research on the RNN architecture provided us with GRU [74] and LSTMs [30] — currently used in VQA models to extract question features. Some methods [55] use a CNN to extract question features as well. Among all these, LSTMs and GRUs are quite popular as they preserve long range contextual information. The recent trend in the field of NLP marks the rise of the transformer architecture [65] which outperforms regular LSTMs and GRUs. Methods [79, 80, 81] use transformers. [82] discusses what the transformer is learning and focuses on the interpretation.

C. Feature Conjugation

After extracting the image and questions features from their separate streams, the model has to join them together to form a joint feature vector. The joining procedure is sometimes done by stacking or concatenating the vectors [83, 84, 85], element wise multiplication or addition [25, 86, 87, 88, 89]. The methodologies will produce a simple yet effective joint embedding which can be further passed through the latter layers of the model. Another approach to this problem is to design an end-to-end neural network model to enumerate the two embeddings. [90] proposes a Multimodal Compact Bi-Linear Pooling mechanism where-as [91] addresses Multi-modal Tucker Fusion model.

Recently, the usage of attention mechanism seen on [55] utilizes question-image co-attention to join the embeddings. Word to region attention [60] bridges the gap between key-words detected in questions with image regions. Being inspired by [92] "hard attention" was introduced [93] to formulate image and question attention which filters all unnecessary image features and works as an effective feature selection method for visual features. Hence, it works well with cluttered and noisy data.

III. Benchmark Datasets

Benchmark datasets in VQA widely vary based on the type of images - real or abstract, type of questions - open-ended or multiple choice, and the use of external knowledge sources. While the external knowledge sources is a different problem domain that won't be addressed in this paper, we will primarily focus on open-ended questions with real images.

A. DAQUAR

The first benchmark dataset, DAQUAR [26], which stands for **Da**taset for **Qu**estion **A**nswering on **R**eal-world images, was introduced in 2014 and contained over 12,000 pairs of question-answers on 1449 RGBD images. The images of this dataset was sourced from NYU-Depth V2 dataset [94], and was annotated based on semantic segmentation of 40 classes. A scene analysis method by Gupta et al [95] was used to map every pixel of the image into 37 object classes and 3 "other" or unknown object classes. The multiple latent worlds model proposed in the paper performed scene analysis using semantic segmentation and hence, required semantic segmentation as the annotation. However, as the latter models hardly opted for semantic segmentation, the annotation became redundant.

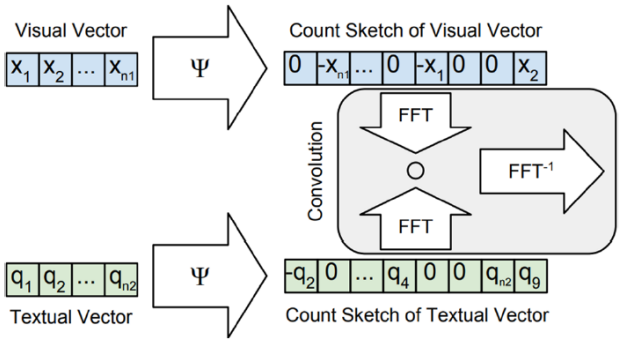


Fig. 6. Multimodal Compact bilinear pooling. [90]

The annotation of question answer pairs were of two types: collected from humans and synthetically generated. Human-collected Q/A pairs were unconstrained and prone to errors as the authors of the paper believed that the architecture of the model should be robust enough to tackle these errors. The Q/A pairs were from a wide range of topics such as basic color-based (based on 16 colors), number of objects, kinds of objects in the image (based on 894 categories) and combinations of these. On average, there were 9 pairs of questions per image.

DAQUAR opened doors for deep neural networks to be trained as it provided a large dataset for training and has been widely established as a benchmark dataset for VQA. However, the key limitation is the question types are constrained to 16 colors and 894 object classes, which might seem a lot at first glance but is not enough for today's standards. Furthermore, the human annotated Q/A pairs showed human biases towards a few of the prominent objects e.g. table and chairs which occurred more than 400 times in the answer.

B. COCO-QA

The success of deep neural networks in computer vision escalated the development of large-scale image datasets - one such prominent dataset being COCO, **C**ommon **O**bjects in **Co**ntext [61]. The corresponding VQA dataset of COCO is the COCO-QA [14] which contains 78,736 train questions and 38,948 test questions over 123,287 images sourced directly from COCO. The questions are synthesized based on image descriptions but are limited to 4 types only - object categorization, number of objects, color of objects and object location while the answers are limited to single word answers only.

The simplicity in Q/A pairs, and the enormous amount of training and testing data popularized the use of COCO-QA in novel architectures. The Q/A pair generation method was used particularly on the MS-COCO image dataset but can be generalized to other datasets as well, which expands the scope of the current dataset even further. COCO-QA questions are also more straightforward from a human's perspective compared to its predecessor DAQUAR, widely due to the human errors present in the latter dataset. However, the dataset is limited to four question types which is far less to train models for general question answering. Similarly, description based answers can never be generated as the answers are also limited to a single world. The simplicity in Q/A pairs is like a double-edged sword - it performs well in directing simpler models towards specific types of questions but will seriously under-perform if the goal is to train and test a general VQA model.

C. FM-IQA

The FM-IQA [96], **F**reestyle **M**ultilingual **I**mage **Q**uestion **A**nswering, uses 158,392 images from the latest version of COCO [61] dataset at that time along with 316,193 human-generated question answer pairs crowd-sourced online from Baidu. Unlike the previous two datasets, the questions are more general and not limited to certain types only. But the questions have to be image-related and not knowledge-based questions (e.g., If the picture of a US president is shown and the question asked is - "Who is this person?", then the model has to check an external knowledge source to get the information, and consequently, such questions are NOT included in the dataset) as seen in 7. Furthermore, the questions can also be common sense-based questions (e.g., a picture of a large orange at the left and a small orange at the right is given and a question is asked - "Which orange is larger?" and the model is expected to answer "The orange at the left."). The answers are also human-generated and are not limited by any means. This freedom of annotation is a trade-off between the diversification of the dataset and the quality assurance of the dataset. Another aspect of the dataset is that the Q/A pairs are available in Chinese and their English translation is used.

The key improvement in FM-IQA over its predecessor is the variety of questions and the high-level reasoning associated with some of the Q/A pairs. The answers are also diversified and varies from single word answers to many word answers, often of the same question (e.g., given a picture of a banana a question is asked - "What is the color of the banana?", and the answer annotations are "The color of the banana is yellow." and "Yellow" - both being correct answers). FM-IQA's Chinese-English Q/A pairs can expand the use of the dataset to other problems such as Visual Machine Translation. However, the Chinese-English Q/A pairs often limit the model's performance in English as the translations are not perfect.

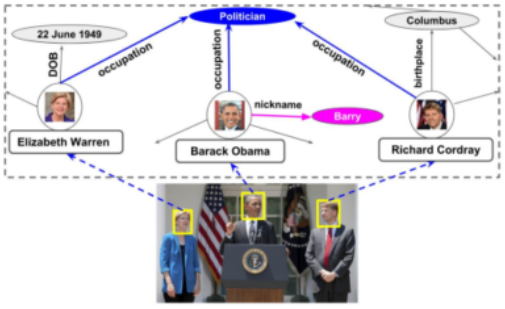
IV. Diagnostic Datasets

The aim of the benchmark datasets is to train a model capable of answering open-ended questions when given any type of image. The diagnostic datasets aim to address shortcomings of the benchmark datasets in certain problem categories; hence the name "diagnostic" dataset.

A. KVQA: Knowledge-Aware Visual Question Answering

Models trained on current datasets are considerably good at answering general questions that require common knowledge or common nouns which can be directly inferred from the question or the image. The ground where it fails to deduce is when any question requires real world knowledge or actual names. For instance, the model may be able to detect how many human beings are present using object detection, but will fail to identify any specific person like Obama or a professional player like Federer as the model has not seen them during the training or learned anything about any specific individual — the training samples were mostly generic. The core idea of KVQA [97] dataset is that it bridges this gap by introducing the idea of models having real world knowledge.

In a nutshell, the KVQA contains the images of people and questions that require external knowledge about them to answer. The datasets have 183K question-answer pairs about more that 18K individuals with 24K images. KVQA is the only dataset that works on 'name entities". The reason why this stands out is because all the other datasets focus



Conventional VQA (Antol et al. 2015; Goyal et al. 2017; Trott, Xiong, and Socher 2018)
Q: How many people are there in the image?
A: 3

Commonsense knowledge-enabled VQA (Wang et al. 2017; 2018; Su et al. 2018; G. Narasimhan and Schwing 2018)
Q: What in the image is used for amplifying sound?
A: Microphone

(World) knowledge-aware VQA (KVQA, this paper):
Q: Who is to the left of Barack Obama?
A: Richard Cordray
Q: Do all the people in the image have a common occupation?
A: Yes
Q: Who among the people in the image is called by the nickname Barry?
A: Person in the center

Fig. 7. Here we can see the types of VQA questions: 1) Simple conventional VQA with the standard VQA questions [25, 87, 98], 2) Commonsense knowledge-enabled VQA [67, 99, 100] and lastly 3) Knowledge aware VQA which requires external specific knowledge [97]

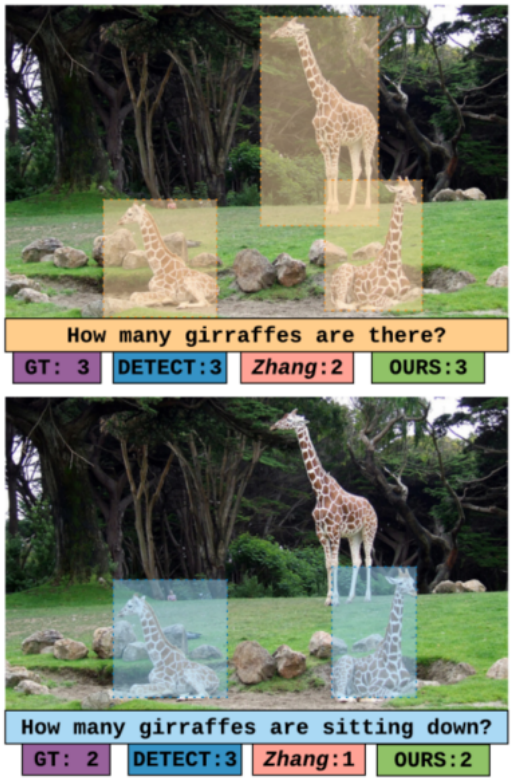


Fig. 8. Here, for the first image, we can see the counting is possible only using object detection, but for complex questions, it needs more than just the object detection. [102]

mostly on the common nouns and are not in need of any external knowledge.During data collection, the list of different individuals were mostly taken from Wikidata [101] which contains images of different politicians, athletes and actors. In summary, KVQA offers the challenging tasks of differentiating a massive range of individuals.

B. TallyQA: Answering Complex Counting Questions

Most of the VQA datasets fail is to focus on the *complex* counting questions. Let's focus on the term 'complex' here as almost all the datasets contain some simple counting questions. So the question is, what makes the TallyQA [102] unique? It is the largest open-ended counting dataset that requires object detection.

What differs between a simple question from a complex question is shown in the picture 8. Here, for a simple question, we are only counting the number of giraffes to come up with solution of the question. On the other hand, for the complex questions, action or activity is associated with the object as well. The answers for the complex questions require deeper understanding of the image or relationships with the object. In cases of counting, there are ways to optimize for better results by using HowMany-QA [98] and region proposals generated by object detection algorithm but the issue remains the same as the dataset is limited to mostly simple question answers. The model will not be able to answer complex questions if they are not included in the dataset. What truly makes TallyVQA unique is that it checks two attributes for counting — 1) Analyzing both simple and complex questions 2) Requiring object detection. This excels the system to have a better performance in the field of 'counting' which shows better results for the state of the art.

The way it was classified as complex or simple is that it would use SpaCy [103] for the tagging of parts of speech. It will be tagged simple only when it has one noun, no adverbs, and no adjectives or else, it will be ruled as complex. Thus, this paper gives us a better approach by creating a relation between objects and reasoning.

C. Scene Text Visual Question Answering

Although the balanced datasets usually contain questions and answers mostly based on images, the issue is that it contains questions on only objects or any particular action. What is excluded is that images contain text that provide greater information about the particular scene. Out of the necessity to give attention to the texts of the image, ST-VQA [17] was created which provides us with the important semantic information from the images. This dataset increases the difficulty of the model to predict as the model needs to deduce texts from the images.

When working with the images, it is often taken as granted that there might be information that could solve many ques-

Fig. 9. Here, we can see the importance of interpreting the text of images this is the only way to solve QA problems. For instance, the price of the banana can only be know from the texts. [17]

tions. In cases of daily activities, written information is a necessity and can provide a lot of information about the object, scene or action e.g., using public transportation, the name of any store, any particular purchase, etc. If we want to answer questions like "Where is the train going?", the answer can be approached only with the help of the texts from the given image, it does not require any external knowledge. Thus, this idea of getting information from the image has turned into a requirement for basic question answering.

The reason ST-VQA stands out is because it is taken from numerous sources comprising 23,038 images from different datasets. The advantage of different dataset is reducing the dataset biases such as selection, capture and negative biases which brings greater variability as well. Also, another key feature is that, we will be selecting images containing two instances of texts, which means there can be multiple options for any question. Thus bringing us a unique dataset from TextVQA [104]. In conclusion, ST-VQA dataset has 23,038 images with 31,791 questions/answers pairs where 19,027 images ,26,308 questions for training and 2,993 images - 4,163 questions for testing.

D. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning

Current Visual question answering models can be trained in a way where it has strong language biases and can answer questions correctly without any sort of reasoning. This is one of the fundamental flaws of current VQA algorithms. Thus the diagnostic dataset CLEVR [28] is presented. Furthermore, it can be used to analyze visual reasoning of any model and

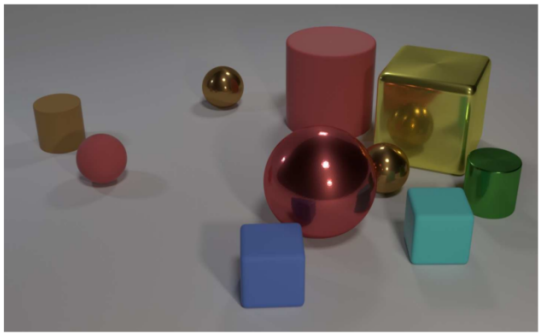


Fig. 10. This is an example of CLEVR [28] questions. It tackles some of the reasoning like comparison, identification and logical actions.

give us a deeper meaning and insight of the model which is the novelty factor.

In cases of training a model, it may be considered as 'cheating' when a model solves the question in a vague manner without any underlying reasoning or understanding. This usually happens because of the biases as many of the questions could just have the same answer. So the question is, how can we understand if the system is capable of reasoning and not just answering using the language biases.

CLEVR contains 100k images that are rendered by Blender [105], about a million automatically generated questions. It stands out for having challenging images and questions that require reasoning such as counting, comparing, logical reasoning and information storing. Also, another fact is ensured as to any external knowledge will not be required. The dataset contains reasoning skills like "factorized representation" for different combinations, requiring short term memory[73] or attending to specific objects [55]. These are the major factors that brings out the novelty of the dataset which other datasets fail to offer. Another fact is that, it may be used for reasoning but it is not a good fit for generalization, thus this should be used with other VQA datasets in order to excel in the field of reasoning.

The dataset is comprised of three object shapes, two materials, eight colors and four positions. The images and scenes are created mostly with objects that are annotated with size, color, shape and position. As mentioned earlier, the images are generated from blender [105]. Lastly, questions are generated from functional programs which are just basic functions that work on the operations of visual reasoning such as value comparison, set counting and query of attributes. CLEVR has countered many different shortcomings like Short-term memory, Spatial relationships, Long reasoning chains and Disentangled representations.

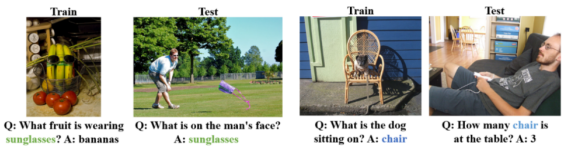


Fig. 11. (a) Is an example of ZSA [106]. Here 'sunglasses' appears in question but not in answer during train, it appears in answer in test. (b) Is an example of ZSQ. Here 'chair' appears in answer but during test, it appears in the question.

TABLE I
TESTING ACCURACIES OF BASELINE METHODS OF DIFFERENT MODEL

	VQA val	Normal	ZSA	ZSQ
LSTM Q+I[107]	54.23% val	47.72%	0.00%	40.03%
SAN [108]	55.86% val	48.70%	0.00%	41.63%
HieCoAtt [55]	57.09% val	41.70%	0.00%	35.46%

E. Zero-Shot Transfer VQA Dataset

Although VQA models and datasets have come a long way, it still falls short when compared to the intellectual behaviors of human beings. The VQA models and the datasets are designed in a way to perform specific tasks and the improvisation or logical reasoning of these models are not similar to human level. This is where the zero-shot transfer VQA dataset[106] comes in , which states that the model should be able to learn from one example (ie. Questions) and transfer that to the other tasks (ie. Predict the answers). Lastly the dataset is used for evaluating different models of VQA to a better understanding if the models address the zero shot transfer problem.

Here, in the example 11 for the first pair, we can see the word 'sunglasses' appears in the question and it appears in the test set as an answer. The word 'sunglasses' is known as 'zero shot word' and this is an example of a 'zero shot answer' or ZSA where the word is in another question of the training set and in the answer of the training set. Likewise, we can see the word 'chair' appearing as an answer in the training set and then in the question of the test set. This is known as 'zero shot question' or ZSQ. With the help of this model, we can also come up with if the model creates any relation between the input and the output and only then can solve the zero-shot learning problems.

The way zero-shot dataset was created is given below-

- Shuffling the training and testing samples and then finding any shared words between questions and answers and creating a word list.
- Randomly sampling 10 percent of words with uniform distribution.
- Dividing the words into two different sub-categories. They are a)ZSA and b)ZSQ words.
- Creating test dataset of ZSA that are from ZSA words.
- Creating test dataset of ZSQ that are from ZSQ words
- Creating both train and test dataset that are not similar with ZSA test and ZSQ test dataset.

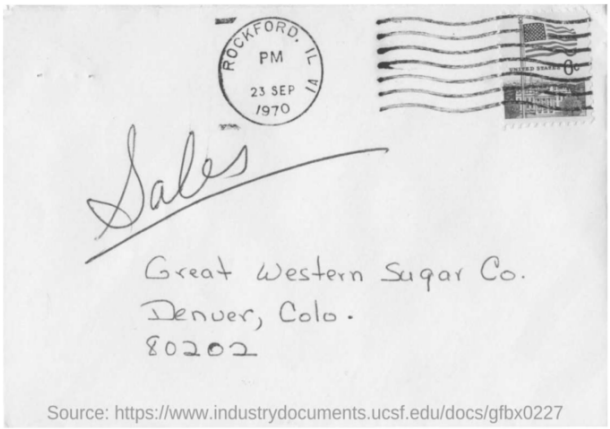


Fig. 12. This is an example of DocVQA[109] question answer.It requires not just the interpretation of the questions but also the format of the documents

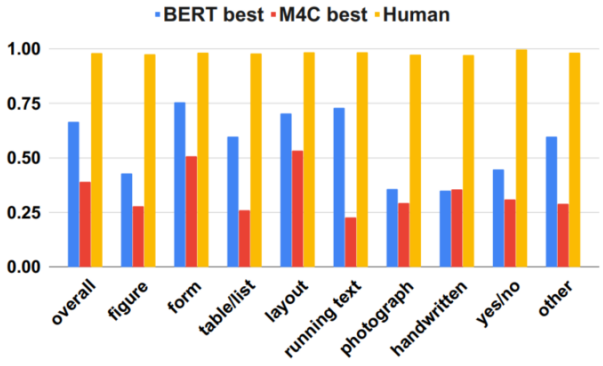


Fig. 13. BERT[110] and M4C[111] compared with human performance.

Lastly, we have used three VQA models for evaluation and I gives us the result. The astounding factor that can be seen is that ZSA has 0% accuracy. This is because the zero-shot words will not be guessed or recognized because the models did not see any of the words or the classes were not created. For ZSA, the bias will be very low because of the negative gradient. Another problem we can define is that answers and questions do not share details to each other, thus a word is used differently in a question compared to being in an answer.

F. DocVQA: A Dataset for VQA on Document Images

Here, a new VQA dataset is presented with the images of different documents which is known as DocVQA[109]. The dataset has 50000 questions with 12000 document images. Here, not just the texts are important. The word 'document' plays a crucial role here. Which means, the models need to understand the main structure of the document.

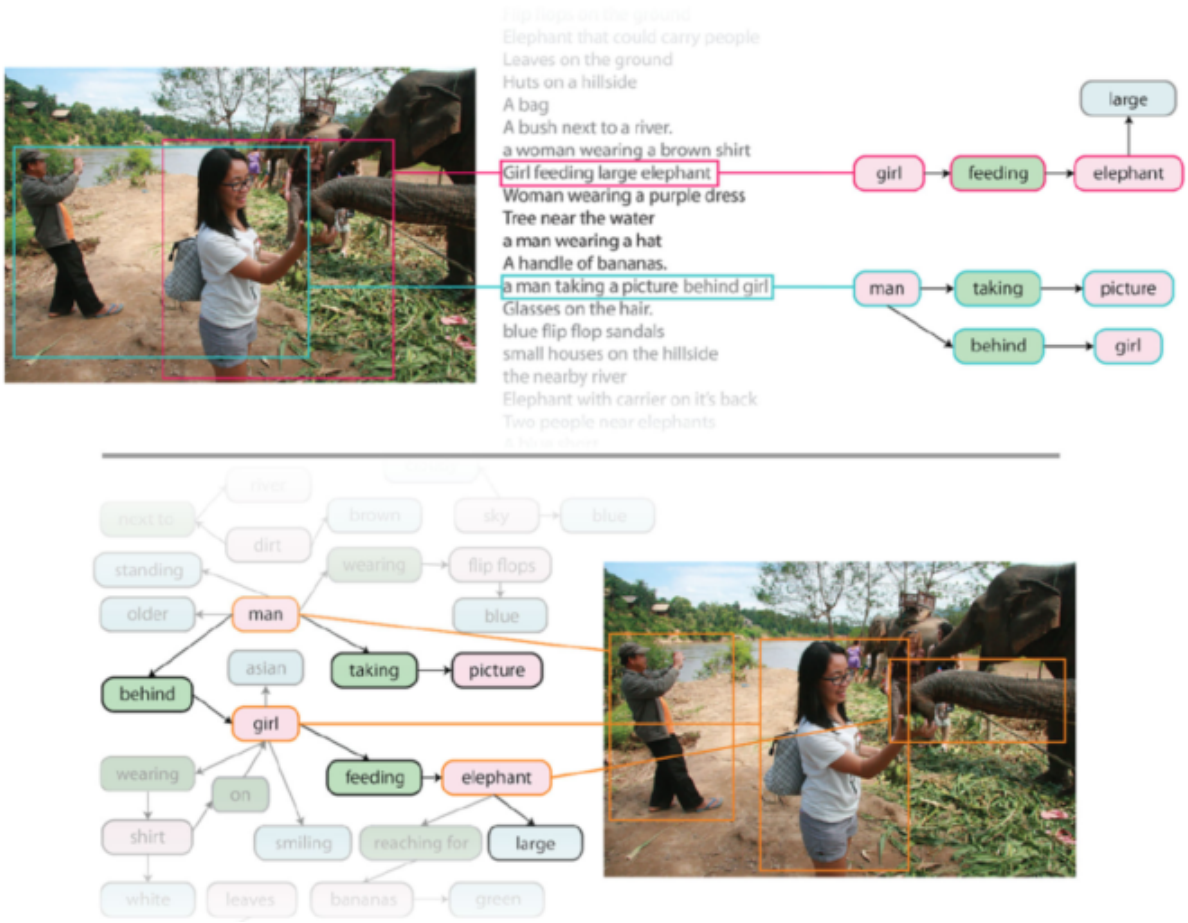


Fig. 14. Here, this is an overview of how the data is annotated with objects, relationships , attributes and descriptions. Here, the object could elephant, relationship could be feeding and attribute could be large

Here, any model with document expertise is expected to respond to requests for information. Here in 12 we can see how the questions can be answered from the format of the document. The system should only get the knowledge from the test, but many other different factors like – layout of the document (tables, forms, boxes, page structure), non-textual elements (diagrams ,marks, separators, tickboxes) and the style (font, colors, highlighting). The reason why this is unique is because in other datasets, they usually focus on specific parts of the documents or specific parts of the books. This dataset offers a generalization of documents. With this, we can also analyze how the models perform on DocVQA. Lastly, the performance of models like LoRRA [104], M4C [111] and BERT[110] were tested and conclusions can be drawn from the analysis.

G. Visual Genome: Connecting Language and Vision Using Crowd-sourced Dense Image Annotations

The problem of the system or any model is that, it performs bad in cases of cognitive tasks for example, image description. The definition of cognitive tasks are, that involves reasoning of the images. Usually, there can be detailed images where the scene might offer a lot of information where the model needs reasoning. Relationships between objects and actions need to be understood. For example, “A guy is riding a horse”, here the guy has a relationship with the horse and it is riding. This is where Visual Genome dataset[112] comes in to create this relationship with the objects and actions.

So what is the definition of ideal model? A model that can describe any object, can see the relationship and speak out the attributes. Also, having a visual understanding of what’s happening in a scene is a key feature for any good model. The problem with the models is that, they would not be able to create a relationship between the objects, rather, their job stays only for object detection.

For example, for 14, it would say that, “Men are standing next to elephants” but it fails to create any relationships. Without the distinction, the models will fail to understand the same scenario in a different image.

To have a true understanding of the images, three key elements must come in the picture. Which are-

- Grounding of visual concepts to language
- Complete set of descriptions and QAs
- Formalized representation

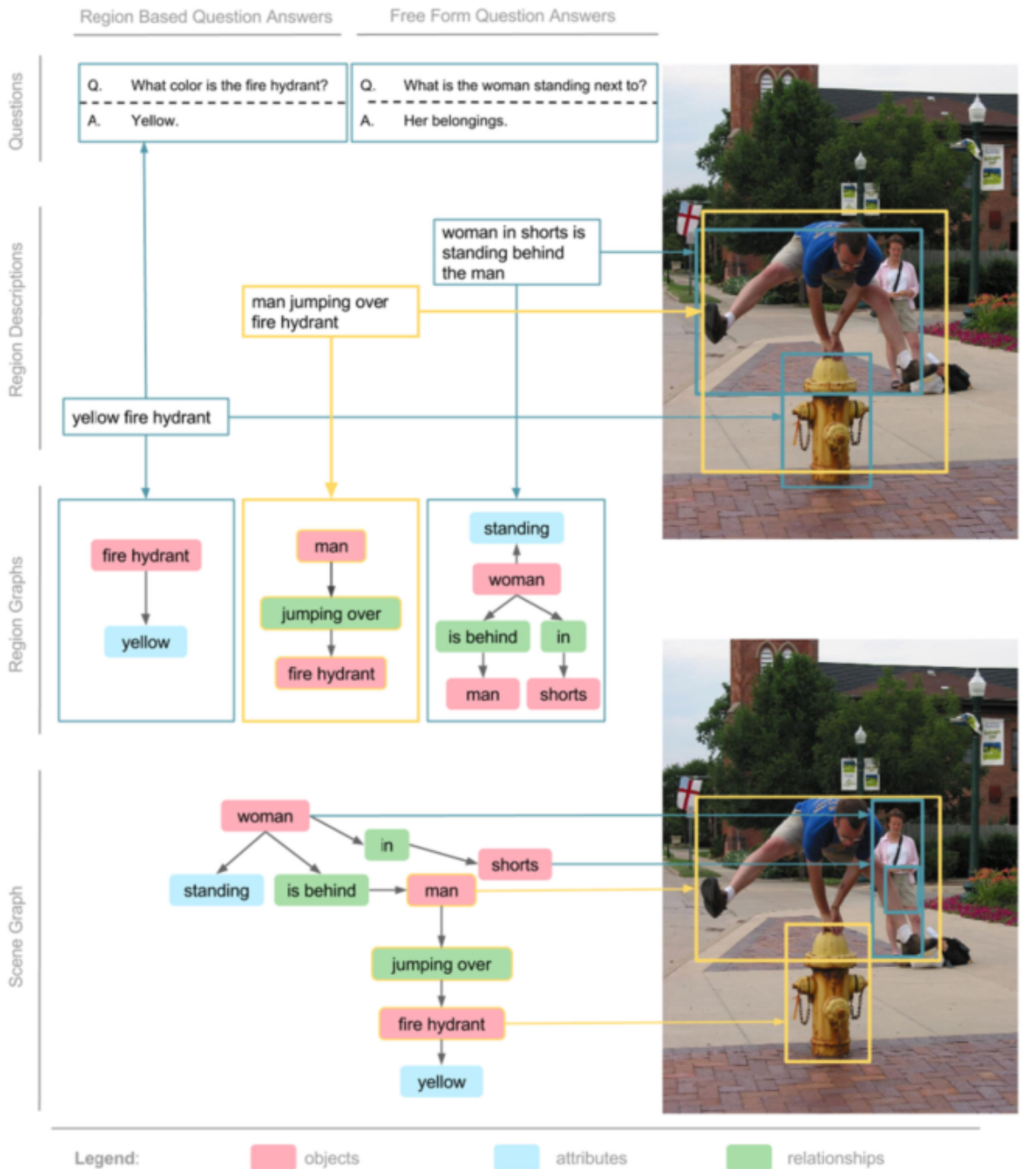


Fig. 15. A representation of Visual Genome dataset.

Visual Genome would be the first to recognize the importance of the relationship between the objects. Thus the dataset comes up with the annotations and labeling which is grounding visual concepts to language. For the dataset, 50 descriptions of different regions of the image were taken having the complete set of descriptions of the scene. Lastly, the dataset is capable of formalized representation of any image which is structured. For example, guy is riding a horse, here riding (guy, horse) is the relationship creating a connection between the horse and the guy.

The dataset consists of 7 different components which are region descriptions, objects, attributes, relationships, region graphs, scene graphs, and question answer pairs.

The images were divided into multiple regions for creating a depth. Bounding boxes are used for each region. Also, each image consists of several objects, each with their bounding boxes. Also, images have on average 26 attributes which can be color, action etc. The relationships connect any objects with each other. After that, a directed graph is created. Lastly, all the scenes are combined to create one scene graph.

TABLE II
DIFFERENT VQA EVALUATION METRICS

Metric	Formula
Accuracy	$\frac{\#correctlyanswered}{\#questionsasked}$
Consensus	$accuracy = min\left(\frac{n}{3}, 1\right)$
Human Evaluation	Judged by humans one by one.
MPT	$\frac{\Sigma A_t}{T}$ or $\frac{T}{\Sigma A_t^{-1}}$
BLEU	$BP.exp\left(\Sigma W_n log P_n\right)$
METEOR	$(1 - Pen) * F_{mean}$

V. EVALUATION METRICS FOR VQA

An evaluation method for a machine learning algorithm is just as important as the model itself. As the models are trying to minimize loss, the evaluation metrics that they are judged upon also need to be well suited for the problem at hand. According to [113], there is no agreed upon evaluation metric for VQA models because they serve a wide variety of datasets, problem statements and approaches. One of the simplest and earliest adapted evaluation metrics is the plain accuracy metric. It is basically the ratio between number of correct answers and total of all the questions asked. This is only applicable to tasks which have only one correct answer via a multiple choice type method. This will not be a good suite for open-ended questions or questions which have more than a single word. For these types of problems we could use Wu Palmer Similarity (WUPS) [114] as seen on [26]. Another popular method of judging VQA methods is Consensus which takes answers from multiple independent sources and the score of the model is to see how many of these does the model agree with. Recently, methods like BLEU and METEOR popular for machine translation are being used for open-ended answers [115]. Other metrics such as human judgment and MPT are used but not as frequently. II gives a clear picture of the VQA evaluation metrics.

VI. COLOR-SHAPE BIAS

The efficient and accurate extraction of the image features from color and shape is critical in VQA. The problem of working with colored images has been addressed since the early days of object recognition. In order to extract stable features from colored images, techniques such as Color Normalization by Graham et al. [116] tried to remove the effect of illumination by preprocessing the image itself. It can be safely inferred that, in object recognition or any problem domain related to feature extraction from colored images, there is a strong correlation between the image colors and feature set extracted from the images. Visual Question Answering is no exception where the extracted features will be used to answer the questions. For the color-shape bias problem we primarily look if the model tends to answer the question by leveraging on the colors or the shapes of the elements in the image. Our work on color-shape bias was motivated by the study by [117] which explores the relationship between the knowledge of colored objects and their representation in the human brain.

A. Incongruently Colored Diagnostic Dataset

As the model tries to generalize the images it has been trained on, it associates colors to shapes and vice-versa. For instance, the model associates the color yellow to an object shaped like a banana, and the color red to an object shaped like an apple. The generalization is logical since apples are primarily red and bananas are primarily yellow. While there can be secondary colors such as green, which both of these fruits have in common, we will overlook this case for the time being. Objects in their primary color are called congruently colored objects. Our key interest lies in the answer generated by the model if a red banana is given or the incongruently colored objects. When asked for the name of the fruit, if the model answers banana, then we can conclude that the model is biased towards the shape of the object. Else, if the model generates the name of an object which is primarily red, for instance - an apple, then we can conclude that the model is biased towards the color of the object.

Bias towards either the shape or the color is not necessarily a bad thing, as the answer can be pretty subjective. Even to human interpreters, when shown a red banana-like object, many would be skeptical to call it a banana. However, the degree of bias is our matter of concern. To what extent is the model willing to favor the color or shape of the image objects. To test this hypothesis, we propose a new dataset on incongruently colored objects and check the accuracy of benchmarks models on this dataset.

The general idea of the experiment is to, first, remove the illumination effect of objects in the images as proposed by [116], and then perform color swapping on images using a python script that will map the primary color to a somewhat

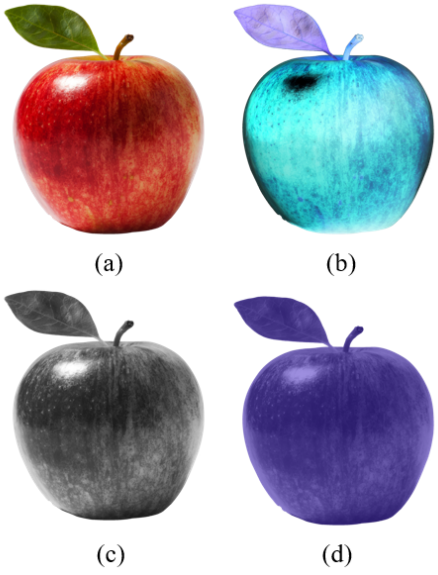


Fig. 16. Image of an apple from our proposed color-shape bias diagnostic dataset where (a) is the original image, (b) is the inverted image, (c) is the grayscale image, and (d) is purple hue added to the grayscale image as proposed in one of our methodologies

"opposite" color. The pipeline of the color swapping algorithm will include object detection, object segmentation and then color swapping. For our experiment we used color inversion, i.e. we took the original color value subtracted from the maximum color value. A bit of random noise was also added to the inversion so that the models capable of performing well on inverted colors can not generalize this.
Another approach is to map the colored images to their corresponding grayscale images with equal luminance. The SHINE toolbox by [118] can be used to desaturate the images. Afterwards, the specific primary hue is applied, for instance, red hue is applied on a grayscale apple. For the incongruently colored objects dataset, we apply a somewhat opposite hue, for instance, blue on a grayscale apple. However, some corner cases on applying opposite hues need to be handled manually, for instance, green can not be applied to grayscale apples as apples can be green too. Both of these approaches should produce similar results, and afterwards, we will end up with two datasets of congruently and incongruently colored objects.

It is to be noted that the questions asked in this dataset are more object-specific. If there is a scene with a banana inside a basket where the color of the banana was swapped to blue then some sample questions will look like - "What is the color of the banana in the object?", "What is inside the basket?", "What is the name of the blue thing inside the basket?", and so on. Another point to keep in mind is that some objects are recognized based on only the color itself e.g. a glass of orange juice won't be called orange juice if we change the color of the juice, and the model can not "taste" the image in the picture, at least not during the time of writing this paper.

B. Grayscale Diagnostic Dataset

Instead of swapping the colors of the objects, we can instead test the performance of the model on grayscale images. The key advantage of using a grayscale dataset is that the questions asked to the model can be more general instead of the object specific questions on incongruently colored objects. However, color specific questions like "What is the color of the banana?" can not be asked, as there is no color information in the picture. However, if the model is trained on grayscale images only, it should be able to infer some of the color features from the grayscale values. The approaches to training models on specific datasets will be discussed in another section.

Evaluating the performance of the model on a grayscale dataset can help us identify how much information is retained from the color of the images. The model should perform noticeably worse on the grayscale dataset compared to the corresponding colored dataset. If the performance drop is insignificant, then we can conclude the models are more biased towards the form or shape of the objects, and if the drop is noticeably large then the bias is towards the color of the objects. Another important aspect of the model can be tested - differentiating lighter and darker regions by asking questions like "What is the dark object in the middle of the basket?" or "What is that light thing at the right?"

C. Edge-Only Diagnostic Dataset

We can take the approach taken in the grayscale dataset a step further by removing the color filling of an object and keeping the edges only. Any edge detection filter can be used to construct this dataset. The dataset also removes the texture of the object, and hence any feature extracted by the model from the texture of the objects can not be used and thus, giving us an even better color vs form comparison. The results are to be analyzed similar to the grayscale diagnostic dataset. However, this model won't be able to differentiate between lighter and darker regions as that information has completely been removed.

D. Incongruently Colored Identification Diagnostic Dataset

Extending the idea of the incongruently colored dataset, we can ask the model what's wrong with the object in the picture and the answer should point towards the unnatural color of the object. E.g. a purple colored banana is shown in the picture and the following question is asked, "What is wrong with the banana?", "What should have been the color of the banana?", or even a general question like "What is unnatural in this picture?" This problem also overlaps with the external knowledge based visual question answering but here, we tend to not ask questions outside the training dataset. For instance - if the model has been trained on bananas which are primarily yellow in color, we expect the model to associate the shape of the banana with the color yellow based on the training data. When the model sees a purple banana then it should, judging by its low confidence on the object being a banana, understand that there is something wrong with the banana and then should also be able to generate an answer that describes the event. It can be "The color of the banana is wrong." or even better if "The color of the banana should be yellow but it is purple." We do not expect the model to perform any kind of logical reasoning apart from the incongruent colors and also do not expect it to take help from external knowledge bases.

VII. Discussion

VQA is one of the fastest growing fields of multimodal research and the sheer volume of work speaks for that. To master VQA, one must first master visual and natural language processing pipelines on their own. III shows the current state of the art in different models on the VQA dataset [25].

As evident from III, most of the models do not have a high margin of accuracy which speaks about the current state of the art. Most modes have around 60% accuracy and much work is yet to be done here. VQA datasets were always plagued with bias towards a single modality and future work may be done towards eliminating that and making a model that generalizes well. Another outlet of research might be investigating the color and shape bias of current VQA models.

VIII. Conclusion

In this paper, we looked through the general methodologies of VQA and focused more on the benchmark and diagnostic

TABLE III
ACCURACIES FOR DIFFERENT VQA MODELS

Model	test-dev	test-std
VQA[25]	53.74	54.06
Comp QA [119]	54.8	55.1
DPPNet [120]	-	57.22
ReBaseline [84]	-	65.2
MCB [90]	66.7	66.5
SAN [108]	58.7	58.9
Region-VQA [121]	62.44	62.43
AskNeuron [122]	58.4	58.4
TDAN [123])	-	63.94*
MCAN [124]	70.63*	70.9*
Cycle [125]	69.87*	-
AnswerAll [126]	65.96*	-
Dropout [127]	66.92*	-

* indicates that VQA v2.0 [128] was used.

datasets. The comprehensive study on the basics of VQA, methodologies and datasets should help the reader to get started with VQA. We reviewed some of the latest diagnostic datasets and also explored new areas of improvement in these fields by proposing a diagnostic dataset on color-shape bias. We then expanded upon this idea by categorizing many such approaches to color-shape bias datasets and proposed techniques to construct these datasets. Widely motivated by the visual perception of shape and color in the human brain, we believe that the models should also have optimal methodologies to focus on either color or shape of an image in order to visually answer questions related to those aspects.

REFERENCES

[1] A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

[2] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

[3] J. Johnson, A. Karpathy, and L. Fei-Fei, "Densecap: Fully convolutional localization networks for dense captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4565–4574, 2016.

[4] J. Xu, T. Mei, T. Yao, and Y. Rui, "Msr-vtt: A large video description dataset for bridging video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

[5] Y. Pan, T. Mei, T. Yao, H. Li, and Y. Rui, "Jointly modeling embedding and translation to bridge video and language," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4594–4602, 2016.

[6] X. Yan, J. Yang, K. Sohn, and H. Lee, "Attribute2image: Conditional image generation from visual attributes," in *European conference on computer vision*, pp. 776–791, Springer, 2016.

[7] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee, "Generative adversarial text to image synthesis," in *International conference on machine learning*, pp. 1060–1069, PMLR, 2016.

[8] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, and X. He, "Attngan: Fine-grained text to image generation with attentional generative adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1316–1324, 2018.

[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International Conference on Machine Learning*, pp. 8821–8831, PMLR, 2021.

[10] A. Baevski, W.-N. Hsu, A. Conneau, and M. Auli, "Unsupervised speech recognition," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[11] C. You, N. Chen, and Y. Zou, "Mrd-net: Multi-modal residual knowledge distillation for spoken question answering," in *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21* (Z.-H. Zhou, ed.), pp. 3985–3991, International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track.

[12] C. You, N. Chen, and Y. Zou, "Self-supervised contrastive cross-modality representation learning for spoken question answering," *arXiv preprint arXiv:2109.03381*, 2021.

[13] H. M. Fayek and J. Johnson, "Temporal reasoning via audio question answering," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2283–2294, 2020.

[14] M. Ren, R. Kiros, and R. Zemel, "Image question answering: A visual semantic embedding model and a new dataset," *Proc. Advances in Neural Inf. Process. Syst*, vol. 1, no. 2, p. 5, 2015.

[15] M. Malinowski and M. Fritz, "Towards a visual turing challenge," *arXiv preprint arXiv:1410.8027*, 2014.

[16] D. Geman, S. Geman, N. Hallonquist, and L. Younes, "Visual turing test for computer vision systems," *Proceedings of the National Academy of Sciences*, vol. 112, no. 12, pp. 3618–3623, 2015.

[17] A. F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, E. Valveny, C. Jawahar, and D. Karatzas, "Scene text visual question answering," in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4291–4301, 2019.

[18] C. Chen, S. Anjum, and D. Gurari, "Grounding answers for visual questions asked by visually impaired people," *arXiv preprint arXiv:2202.01993*, 2022.

[19] K. Baker, A. Parekh, A. Fabre, A. Addlesee, R. Kruiper, and O. Lemon, "The spoon is in the sink: Assisting

visually impaired people in the kitchen," in *Proceedings of the Reasoning and Interaction Conference (ReInAct 2021)*, pp. 32–39, 2021.

[20] S. Wang, W. Zhao, Z. Kou, J. Shi, and C. Xu, "How to make a blt sandwich? learning vqa towards understanding web instructional videos," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1130–1139, 2021.

[21] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1686–1697, 2021.

[22] N. Mostafazadeh, I. Misra, J. Devlin, M. Mitchell, X. He, and L. Vanderwende, "Generating natural questions about an image," *arXiv preprint arXiv:1603.06059*, 2016.

[23] I. Chowdhury, K. N. Thanh, S. Sridharan, *et al.*, "Video question answering for surveillance," *TechRxiv*, 2020.

[24] S. Barra, C. Bisogni, M. De Marsico, and S. Ricciardi, "Visual question answering: which investigated applications?," *Pattern Recognition Letters*, vol. 151, pp. 325–331, 2021.

[25] 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*, pp. 2425–2433, 2015.

[26] M. Malinowski and M. Fritz, "A multi-world approach to question answering about real-world scenes based on uncertain input," *Advances in neural information processing systems*, vol. 27, 2014.

[27] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "Visualbert: A simple and performant baseline for vision and language," *arXiv preprint arXiv:1908.03557*, 2019.

[28] 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*, pp. 2901–2910, 2017.

[29] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[30] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[31] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang, "Large-scale image classification: Fast feature extraction and svm training," in *CVPR 2011*, pp. 1689–1696, 2011.

[32] I. Parra Alonso, D. Fernandez Llorca, M. A. Sotelo, L. M. Bergasa, P. Revenga de Toro, J. Nuevo, M. Ocana, and M. A. Garcia Garrido, "Combination of feature extraction methods for svm pedestrian detection," *IEEE Transactions on Intelligent Transportation Systems*, vol. 8, no. 2, pp. 292–307, 2007.

[33] R. Lienhart and J. Maydt, "An extended set of haar-like features for rapid object detection," in *Proceedings. International Conference on Image Processing*, vol. 1, pp. I–I, 2002.

[34] J. Maydt and R. Lienhart, "A fast method for training support vector machines with a very large set of linear features.," in *ICME (1)*, pp. 309–312, 2002.

[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893 vol. 1, 2005.

[36] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision – ECCV 2006* (A. Leonardis, H. Bischof, and A. Pinz, eds.), (Berlin, Heidelberg), pp. 428–441, Springer Berlin Heidelberg, 2006.

[37] D. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the Seventh IEEE International Conference on Computer Vision*, vol. 2, pp. 1150–1157 vol.2, 1999.

[38] Z.-Q. Hong, "Algebraic feature extraction of image for recognition," *Pattern Recognition*, vol. 24, no. 3, pp. 211–219, 1991.

[39] A. Hyvarinen, E. Oja, P. Hoyer, and J. Hurri, "Image feature extraction by sparse coding and independent component analysis," in *Proceedings. Fourteenth International Conference on Pattern Recognition (Cat. No.98EX170)*, vol. 2, pp. 1268–1273 vol.2, 1998.

[40] K. Fukushima and S. Miyake, "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition," in *Competition and cooperation in neural nets*, pp. 267–285, Springer, 1982.

[41] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3642–3649, 2012.

[42] D. A. Pomerleau, "Alvinn: An autonomous land vehicle in a neural network," *Advances in neural information processing systems*, vol. 1, 1988.

[43] A. Sarlashkar, M. Bodruzzaman, and M. Malkani, "Feature extraction using wavelet transform for neural network based image classification," in *Proceedings of Thirtieth Southeastern Symposium on System Theory*, pp. 412–416, 1998.

[44] B. Lerner, H. Guterman, M. Aladjem, and I. Dinstein, "A comparative study of neural network based feature extraction paradigms," *Pattern Recognition Letters*, vol. 20, no. 1, pp. 7–14, 1999.

[45] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, 2012.

[46] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recog-

nition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[47] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[48] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

[49] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

[50] S. Bozinovski, "Reminder of the first paper on transfer learning in neural networks, 1976," *Informatica*, vol. 44, no. 3, 2020.

[51] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4995–5004, 2016.

[52] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4613–4621, 2016.

[53] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 261–277, Springer International Publishing, 2016.

[54] L. Ma, Z. Lu, and H. Li, "Learning to answer questions from image using convolutional neural network," in *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[55] J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," *Advances in neural information processing systems*, vol. 29, 2016.

[56] N. Ruwa, Q. Mao, L. Wang, and M. Dong, "Affective visual question answering network," in *2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR)*, pp. 170–173, 2018.

[57] M. Lao, Y. Guo, H. Wang, and X. Zhang, "Cross-modal multistep fusion network with co-attention for visual question answering," *IEEE Access*, vol. 6, pp. 31516–31524, 2018.

[58] Y. Yuan, S. Wang, M. Jiang, and T. Y. Chen, "Perception matters: detecting perception failures of vqa models using metamorphic testing," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16908–16917, 2021.

[59] A. S. Toor, H. Wechsler, and M. Nappi, "Question action relevance and editing for visual question answering," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2921–2935, 2019.

[60] L. Peng, Y. Yang, Y. Bin, N. Xie, F. Shen, Y. Ji, and X. Xu, "Word-to-region attention network for visual question answering," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 3843–3858, 2019.

[61] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*, pp. 740–755, Springer, 2014.

[62] K. Kafle and C. Kanan, "Answer-type prediction for visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4976–4984, 2016.

[63] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10012–10022, 2021.

[64] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[65] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[66] Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A convnet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.

[67] W. Su, X. Zhu, Y. Cao, B. Li, L. Lu, F. Wei, and J. Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," *arXiv preprint arXiv:1908.08530*, 2019.

[68] Y. Hirota, N. Garcia, M. Otani, C. Chu, Y. Nakashima, I. Taniguchi, and T. Onoye, "A picture may be worth a hundred words for visual question answering," *arXiv preprint arXiv:2106.13445*, 2021.

[69] H. Xue, Y. Huang, B. Liu, H. Peng, J. Fu, H. Li, and J. Luo, "Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[70] G. Luo, Y. Zhou, X. Sun, Y. Wang, L. Cao, Y. Wu, F. Huang, and R. Ji, "Towards lightweight transformer via group-wise transformation for vision-and-language tasks," *IEEE Transactions on Image Processing*, 2022.

[71] G. A. Miller and W. G. Charles, "Contextual correlates of semantic similarity," *Language and cognitive processes*, vol. 6, no. 1, pp. 1–28, 1991.

[72] C. Eckart and G. Young, "The approximation of one matrix by another of lower rank," *Psychometrika*, vol. 1, no. 3, pp. 211–218, 1936.

[73] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[74] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, "Empirical evaluation of gated recurrent neural networks on

sequence modeling," *arXiv preprint arXiv:1412.3555*, 2014.

[75] W. Xu and A. Rudnicky, "Can artificial neural networks learn language models?," *Sixth International Conference on Spoken Language Processing*, pp. 202–205, 01 2000.

[76] Y. Bengio, R. Ducharme, and P. Vincent, "A neural probabilistic language model," *Advances in Neural Information Processing Systems*, vol. 13, 2000.

[77] T. Mikolov, W.-t. Yih, and G. Zweig, "Linguistic regularities in continuous space word representations," in *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: Human language technologies*, pp. 746–751, 2013.

[78] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[79] Z. Yang, N. Garcia, C. Chu, M. Otani, Y. Nakashima, and H. Takemura, "A comparative study of language transformers for video question answering," *Neurocomputing*, vol. 445, pp. 121–133, 2021.

[80] A. F. Biten, R. Litman, Y. Xie, S. Appalaraju, and R. Manmatha, "Latr: Layout-aware transformer for scene-text vqa," *arXiv preprint arXiv:2112.12494*, 2021.

[81] Z. Yang, Y. Lu, J. Wang, X. Yin, D. Florencio, L. Wang, C. Zhang, L. Zhang, and J. Luo, "Tap: Text-aware pre-training for text-vqa and text-caption," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8751–8761, 2021.

[82] C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, and C. Wolf, "How transferable are reasoning patterns in vqa?," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4207–4216, 2021.

[83] B. Zhou, Y. Tian, S. Sukhbaatar, A. Szlam, and R. Fergus, "Simple baseline for visual question answering," *arXiv preprint arXiv:1512.02167*, 2015.

[84] A. Jabri, A. Joulin, and L. van der Maaten, "Revisiting visual question answering baselines," in *Computer Vision – ECCV 2016* (B. Leibe, J. Matas, N. Sebe, and M. Welling, eds.), (Cham), pp. 727–739, Springer International Publishing, 2016.

[85] J.-H. Huang, C. D. Dao, M. Alfadly, and B. Ghanem, "A novel framework for robustness analysis of visual qa models," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33-01, pp. 8449–8456, 2019.

[86] P. Zhang, Y. Goyal, D. Summers-Stay, D. Batra, and D. Parikh, "Yin and yang: Balancing and answering binary visual questions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5014–5022, 2016.

[87] 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*, pp. 6904–6913, 2017.

[88] X. Lin and D. Parikh, "Leveraging visual question answering for image-caption ranking," in *European conference on computer vision*, pp. 261–277, Springer, 2016.

[89] D. Teney and A. van den Hengel, "Visual question answering as a meta learning task," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 219–235, 2018.

[90] A. Fukui, D. H. Park, D. Yang, A. Rohrbach, T. Darrell, and M. Rohrbach, "Multimodal compact bilinear pooling for visual question answering and visual grounding," *CoRR*, vol. abs/1606.01847, 2016.

[91] H. Ben-Younes, R. Cadene, M. Cord, and N. Thome, "Mutan: Multimodal tucker fusion for visual question answering," in *Proceedings of the IEEE international conference on computer vision*, pp. 2612–2620, 2017.

[92] 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*, pp. 6077–6086, 2018.

[93] M. Malinowski, C. Doersch, A. Santoro, and P. Battaglia, "Learning visual question answering by bootstrapping hard attention," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 3–20, 2018.

[94] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*, pp. 746–760, Springer, 2012.

[95] S. Gupta, P. Arbelaez, and J. Malik, "Perceptual organization and recognition of indoor scenes from rgb-d images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 564–571, 2013.

[96] H. Gao, J. Mao, J. Zhou, Z. Huang, L. Wang, and W. Xu, "Are you talking to a machine? dataset and methods for multilingual image question," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[97] S. Shah, A. Mishra, N. Yadati, and P. P. Talukdar, "Kvqa: Knowledge-aware visual question answering," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33-01, pp. 8876–8884, 2019.

[98] A. Trott, C. Xiong, and R. Socher, "Interpretable counting for visual question answering," *CoRR*, vol. abs/1712.08697, 2017.

[99] P. Wang, Q. Wu, C. Shen, and A. Hengel, "The vqa-machine: Learning how to use existing vision algorithms to answer new questions," 12 2016.

[100] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual

visual question answering," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 451–468, 2018.

[101] D. Vrandečić and M. Krötzsch, "Wikidata: a free collaborative knowledgebase," *Communications of the ACM*, vol. 57, no. 10, pp. 78–85, 2014.

[102] M. Acharya, K. Kafle, and C. Kanan, "Tallyqa: Answering complex counting questions," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33-01, pp. 8076–8084, 2019.

[103] M. Honnibal and I. Montani, "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing." To appear, 2017.

[104] A. Singh, V. Natarajan, M. Shah, Y. Jiang, X. Chen, D. Batra, D. Parikh, and M. Rohrbach, "Towards vqa models that can read," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8317–8326, 2019.

[105] B. O. Community, *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018.

[106] Y. Li, Y. Yang, J. Wang, and W. Xu, "Zero-shot transfer vqa dataset," *arXiv preprint arXiv:1811.00692*, 2018.

[107] J. Lu, X. Lin, D. Batra, and D. Parikh, "Deeper lstm and normalized cnn visual question answering model," *GitHub repository*, vol. 6, 2015.

[108] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola, "Stacked attention networks for image question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 21–29, 2016.

[109] M. Mathew, D. Karatzas, and C. Jawahar, "Docvqa: A dataset for vqa on document images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2200–2209, 2021.

[110] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[111] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9992–10002, 2020.

[112] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.

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

[114] Z. Wu and M. Palmer, "Verb semantics and lexical selection," *arXiv preprint cmp-lg/9406033*, 1994.

[115] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617, 2018.

[116] G. D. FINLAYSON and G. Y. TIAN, "Color normalization for color object recognition," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 13, no. 08, pp. 1271–1285, 1999.

[117] L. Teichmann, G. L. Quek, A. K. Robinson, T. Grootswagers, T. A. Carlson, and A. N. Rich, "The influence of object-color knowledge on emerging object representations in the brain," *Journal of Neuroscience*, vol. 40, no. 35, pp. 6779–6789, 2020.

[118] V. Willenbockel, J. Sadr, D. Fiset, G. O. Horne, F. Gosselin, and J. W. Tanaka, "Controlling low-level image properties: the shine toolbox," *Behavior research methods*, vol. 42, no. 3, pp. 671–684, 2010.

[119] J. Andreas, M. Rohrbach, T. Darrell, and D. Klein, "Learning to compose neural networks for question answering," *arXiv preprint arXiv:1601.01705*, 2016.

[120] H. Noh, P. H. Seo, and B. Han, "Image question answering using convolutional neural network with dynamic parameter prediction," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 30–38, 2016.

[121] K. J. Shih, S. Singh, and D. Hoiem, "Where to look: Focus regions for visual question answering," 2015.

[122] M. Malinowski, M. Rohrbach, and M. Fritz, "Ask your neurons: A deep learning approach to visual question answering," *CoRR*, vol. abs/1605.02697, 2016.

[123] M. Li, L. Gu, Y. Ji, and C. Liu, "Text-guided dual-branch attention network for visual question answering," in *Advances in Multimedia Information Processing – PCM 2018* (R. Hong, W.-H. Cheng, T. Yamasaki, M. Wang, and C.-W. Ngo, eds.), (Cham), pp. 750–760, Springer International Publishing, 2018.

[124] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," *CoRR*, vol. abs/1906.10770, 2019.

[125] M. Shah, X. Chen, M. Rohrbach, and D. Parikh, "Cycle-consistency for robust visual question answering," *CoRR*, vol. abs/1902.05660, 2019.

[126] R. Shrestha, K. Kafle, and C. Kanan, "Answer them all! toward universal visual question answering models," *CoRR*, vol. abs/1903.00366, 2019.

[127] Y. Feng, X. Zhu, Y. Li, and Y. Ruan, "Learning capsule networks with images and text," *NIPS*, 2019.

[128] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.