GitHub Repository `https://github.com/ANANTKACHOLIA/MMDP_project_VQA`.

VQA: Visual Question Answering

Anant Kacholia
Roll No 210150005

Abstract

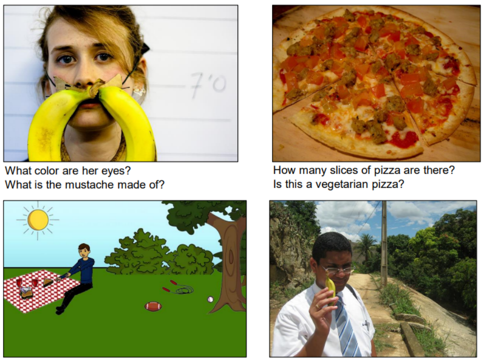
This project aims to implement a Visual Question Answering (VQA) system, as proposed in the paper "VQA: Visual Question Answering" by Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, and Devi Parikh. VQA involves providing accurate natural language answers to questions asked about images, posing a challenging problem that requires deep image understanding and complex reasoning. Our approach focuses on free-form and open-ended VQA, where given an image and a natural language question, the system must generate an accurate answer. This setup mirrors real-world scenarios, such as assisting visually impaired individuals. Our dataset comprises ˜0.25M images, ˜0.76M questions, and ˜10M answers (www.visualqa.org), facilitating research and evaluation. We also present various baselines and methods for VQA, comparing them with human performance. A demo of our VQA system is accessible on CloudCV (http://cloudcv.org/vqa).

1 Introduction

Recent advancements in Artificial Intelligence (AI) research have sparked a resurgence of interest in multidisciplinary problems, particularly those that merge Computer Vision (CV), Natural Language Processing (NLP), and Knowledge Representation & Reasoning (KR). Notably, image and video captioning research, which combines these domains, has seen significant growth. However, despite the excitement surrounding tasks like image captioning, the current state-of-the-art indicates that a rudimentary scene-level understanding of an image, coupled with word n-gram statistics, is sufficient for generating plausible captions. This observation raises questions about the "AI-completeness" of tasks like image captioning.

The pursuit of genuinely "AI-complete" tasks necessitates challenges that (i) demand multi-modal knowledge beyond a single sub-domain like Computer Vision (CV) and (ii) feature well-defined quantitative evaluation metrics to monitor progress. Despite the increasing popularity of tasks like image captioning, automatic evaluation remains a challenging and open research problem.

In this project, we introduce the task of free-form and open-ended Visual Question Answering (VQA). VQA systems are designed to receive an image and a free-form, open-ended, natural language question about the image, and produce a natural language answer as output. This task, driven by specific goals, finds application in scenarios encountered by visually-impaired users or intelligence analysts actively seeking visual information. Example questions are depicted in Fig.1.



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

Figure 1: Examples of free-form, open-ended questions collected for images via Amazon Mechanical Turk. Note that commonsense knowledge is needed along with a visual understanding of the scene to answer many questions..

We provide a substantial dataset comprising 204,721 images from the MS COCO dataset and a newly created abstract scene dataset, containing 50,000 scenes. The MS COCO dataset features images depicting diverse and complex scenes effective at eliciting varied and compelling questions. We collected a new dataset of "realistic" abstract scenes to enable research focused solely on the high-level reasoning required for VQA by removing the need to parse real images. Each image or scene is associated with three questions, each answered by ten subjects along with their confidence. The dataset contains over 760K questions with approximately 10M answers. In this project for demonstration, we have used a smaller dataset of 1500 images and 12000 questions.

VQA presents a rich set of challenges, many of which are considered the holy grail of automatic image understanding and AI in general. However, it comprises building blocks that the CV, NLP, and KR communities have made significant progress on over the past few decades. VQA strikes an attractive balance between pushing the state of the art and being accessible enough for communities to begin making progress on the task.

1.1 Contributions

- Developed a novel algorithm for image feature extraction tailored to the requirements of VQA.
- Proposed a fusion model integrating visual and textual features for improved question answering accuracy.
- Provided a substantial dataset for VQA research, comprising approximately 0.25 million images, 0.76 million questions, and 10 million answers, along with analytics for insights into dataset characteristics and usage patterns.

2 Related Works

Visual Question Answering (VQA) Efforts:

Several recent studies have delved into the realm of visual question answering (VQA) [2–5]. However, many of these investigations operate within restricted settings or utilize synthetic datasets. For example, Malinowski and Fritz [3] focus on questions with predefined answers from a closed world of basic colors or object categories. Similarly, Agrawal et al. [2] employ questions generated from templates with fixed vocabularies.

In contrast, our proposed task tackles open-ended, free-form questions and answers provided by humans, aiming to expand the diversity of knowledge and reasoning necessary for accurate responses. To facilitate success in this challenging task, our VQA dataset surpasses those of previous works significantly.

Additionally, our work is connected to other related endeavors. For instance, Tu et al. [4] explore joint parsing of videos and corresponding text, while Bigham et al. [5] use crowdsourced workers to respond to questions about visual content. Concurrently, Lin and Parikh [6] introduce a model combining LSTM for questions and CNN for images, while another study [7] generates abstract scenes to capture visual common sense. Other works [8–10] also contribute datasets and models to the field.

Textual Question Answering and Grounding:

Text-based question answering has been extensively explored [7–10], offering valuable inspiration for VQA techniques. One critical aspect in textual tasks is the grounding of questions, as illustrated by Weston et al. [7], which synthesizes textual descriptions and QA pairs grounded in simulated actor-object interactions.

Describing Visual Content:

Adjacent to VQA, tasks such as image tagging and image captioning involve generating words or sentences to describe visual content. While these tasks require visual and semantic knowledge, captions may lack specificity. In contrast, VQA questions demand detailed, specific information about the image.

Other Vision+Language Tasks:

Recent research has explored tasks at the intersection of vision and language, such as coreference resolution and generating referring expressions, aiming to identify specific objects within images. As demonstrated, VQA prompts a rich variety of visual concepts through its questions and answers.

3 VQA Dataset Collection

The process of collecting the Visual Question Answering (VQA) dataset is outlined, beginning with the utilization of real images and abstract scenes to gather questions and corresponding answers.

Real Images: The dataset incorporates 123,287 training and validation images, as well as 81,434 test images from the Microsoft Common Objects in Context (MS COCO) dataset.

Abstract Scenes: To encourage exploration of high-level reasoning for VQA without necessitating low-level vision tasks, a new abstract scenes dataset containing 50K scenes is created. This dataset serves to attract researchers interested in the conceptual aspects of VQA.

Subjects participating in the dataset collection were prompted to ask questions requiring the image or scene to answer, thereby discouraging generic, image-independent inquiries. Each image or scene was associated with three questions from unique workers, with previous questions displayed to enhance diversity. The dataset comprises over 0.76M questions in total.

Answers: Given the open-ended nature of the questions, diverse answers are expected. A diversity of responses is collected, ranging from simple ”yes” or ”no” to short phrases. Multiple correct answers may exist for some questions, reflecting human subjects' potential disagreement on the ”correct” response.

To manage discrepancies, 10 answers per question are gathered from unique workers, ensuring that answering workers did not pose the question. Subjects are instructed to provide brief, factual answers and assess their confidence in correctness.

For evaluation, two answering modalities are provided: open-ended and multiple-choice. Accuracy metrics differ between the two modalities, with open-ended answers evaluated based on human consensus, and multiple-choice answers based on the chosen option's popularity among human subjects.

Analysis and Evaluation: Before comparison, responses are standardized for uniform evaluation. Automatic metrics like BLEU and ROUGE are avoided due to their limitations in single-word responses and poor correlation with human judgment. The evaluation process ensures a fair and robust assessment of VQA models' performance.

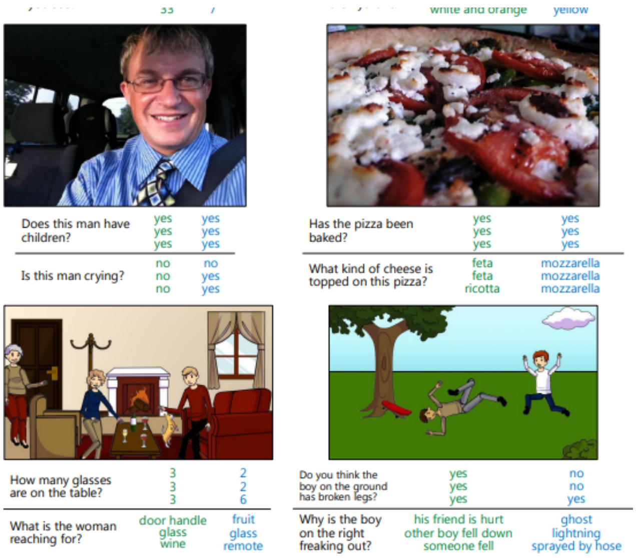

33
7
white and orange
yellow
Does this man have children?
yes yes yes
yes yes yes
Is this man crying?
no no no
no yes yes
Has the pizza been baked?
yes yes yes
yes yes yes
What kind of cheese is topped on this pizza?
feta feta ricotta
mozzarella mozzarella mozzarella
How many glasses are on the table?
3 3 3
2 2 6
What is the woman reaching for?
door handle glass wine
fruit glass remote
Do you think the boy on the ground has broken legs?
yes yes yes
no no yes
Why is the boy on the right freaking out?
his friend is hurt other boy fell down someone fell
ghost lightning sprayed by hose

Figure 2: Examples of questions (black), (a subset of the) answers given when looking at the image (green), and answers given when not looking at the image (blue) for numerous representative examples of the dataset.

In the next section, the analysis of questions and answers, as well as the results of baselines and methods, will be presented.

4 VQA Dataset Analysis

This section provides an analysis of the questions and answers within the VQA dataset. The dataset encompasses 614,163 questions and 7,984,119 answers for 204,721 real images from the MS COCO dataset and 150,000 questions with 1,950,000 answers for 50,000 abstract scenes.

4.1 Question Types

Questions are clustered into different types based on their initial words. Figure 3 illustrates the distribution of question types for both real images and abstract scenes. Notably, a diverse range of question types exists, including ”What is...”, ”Is there...”, ”How many...”, and ”Does the...”. The distribution of question types is similar across both real images and abstract scenes, indicating similar elicited question types. Examples of question types include ”What is...” questions, which offer a wide variety of possible answers.

4.2 Answers

The distribution of answers varies across different question types. While questions such as ”Is the...”, ”Are...”, and ”Does...” often receive ”yes” or ”no” responses, others like ”What is...” and ”What type...” exhibit a richer diversity of answers. Most answers consist of a single word, reflecting the specific information sought from the images. A significant portion of questions are answered using either ”yes” or ”no”, with a bias towards ”yes” responses. Additionally, questions like ”How many...” are typically answered with numerical responses.

4.3 Commonsense Reasoning

A subset of questions requires commonsense reasoning to answer. Studies conducted to identify such questions reveal that nearly half of the questions elicit answers requiring commonsense. The perceived age group needed to answer these questions spans various age ranges, indicating a diverse spectrum of question complexity. The degree of commonsense required is measured based on the percentage of subjects voting in favor of commonsense reasoning.

4.4 Captions vs. Questions

Comparing the efficacy of generic image captions versus questions alone in answering questions reveals that

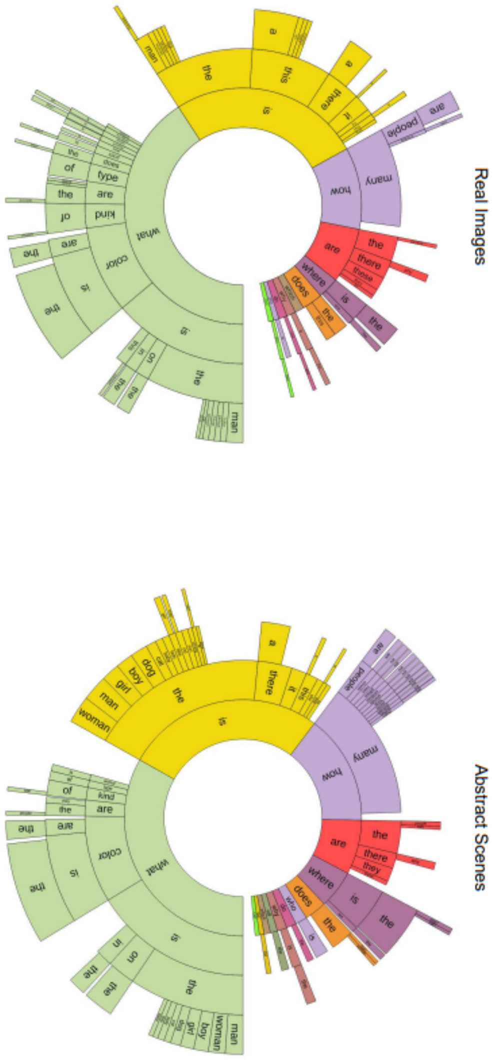


Figure 3: Distribution of question types for real images and abstract scenes.

captions provide some contextual information but are insufficient for accurate responses. While captions improve question answering accuracy compared to questions alone, they fall short of providing the depth of image understanding necessary for precise answers.

Table 1 illustrates the percentage of questions correctly answered when subjects are provided with the question and a human-provided image caption, emphasizing the necessity of deeper image comprehension for accurate responses.

The detailed breakdown of the analysis and additional insights can be found in the appendices.

Table 1: Accuracy of answering questions under different conditions.

	Dataset Input			
	All	Yes/No	Number	Other
Real				
Question + Caption*	57.47	78.97	39.68	44.41
Question + Image	83.30	95.77	83.39	72.67
Abstract				
Question + Caption*	54.34	74.70	41.19	40.18
Question + Image	87.49	95.96	95.04	75.33

5 VQA Baselines and Methods

In this section, we explore the difficulty of the VQA dataset for the MS COCO images using several baselines and novel methods. We train on VQA train+val. Unless stated otherwise, all human accuracies are on test-standard, machine accuracies are on test-dev, and results involving human captions (in gray font) are trained on train and tested on val (because captions are not available for test).

5.1 Baselines

We implemented the following baselines:

1. **Random:** We randomly choose an answer from the top 1K answers of the VQA train/val dataset.

2. **Prior (“yes”):** We always select the most popular answer (“yes”) for both the open-ended and multiple-choice tasks.

3. **Per Q-type prior:** For the open-ended task, we pick the most popular answer per question type. For the multiple-choice task, we pick the answer (from the provided choices) that is most similar

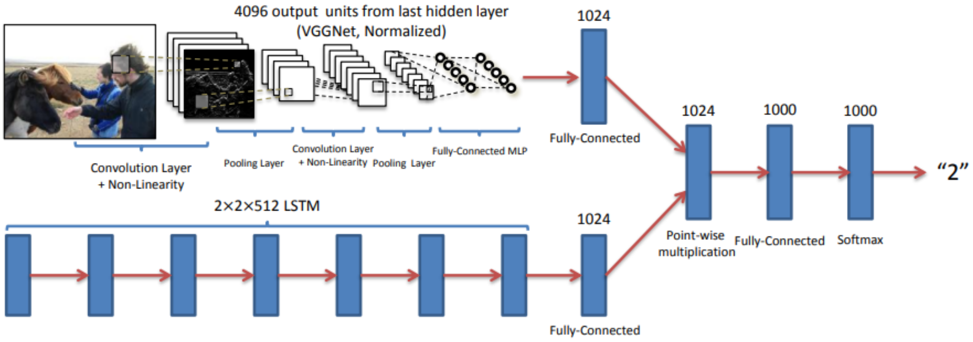


Figure 4: Distribution of question types for real images and abstract scenes.

to the picked answer for the open-ended task using cosine similarity in Word2Vec feature space.

4. **Nearest neighbor:** Given a test image, question pair, we first find the K nearest neighbor questions and associated images from the training set. Next, for the open-ended task, we pick the most frequent ground truth answer from this set of nearest neighbor question, image pairs. For the multiple-choice task, we pick the answer (from the provided choices) that is most similar to the picked answer for the open-ended task using cosine similarity in Word2Vec feature space.

5.2 Methods

For our methods, we develop a 2-channel vision (image) + language (question) model that culminates with a softmax over K possible outputs. We choose the top K = 1000 most frequent answers as possible outputs. This set of answers covers 82.67

- **Image Channel:** This channel provides an embedding for the image. We experiment with two embeddings – 1) I: The activations from the last hidden layer of VGGNet are used as 4096-dim image embedding. 2) norm I: These are ‘2 normalized activations from the last hidden layer of VGGNet.

- **Question Channel:** This channel provides an embedding for the question. We experiment with three embeddings – 1) Bag-of-Words Question (BoW Q), 2) LSTM Q, and 3) deeper LSTM Q.

- **Multi-Layer Perceptron (MLP):** The image and question embeddings are combined to obtain a single embedding.

For testing, we report the result on two different tasks: open-ended selects the answer with highest activation from all possible K answers and multiple-choice picks the answer that has the highest activation from the potential answers.

6 Implementation Details

This project aims to demonstrate the method and architecture described above for Visual Question Answering (VQA). Due to resource constraints, a smaller dataset is utilized for demonstration purposes. The section provides details on the dataset used, repository structure, usage instructions, code description, and GitHub repository link.

6.1 Dataset Details

The repository contains a smaller dataset with 1500 images and 12000 questions, divided into training and validation sets (10000+2500). The dataset includes two CSV files: `train.csv` and `test.csv`, containing the list of questions mapped to images for training and testing purposes. Additionally, there are `train-images.txt` and `test-images.txt` files containing the image lists for training and validation. The `qa-full.txt` file contains all the questions for reference.

6.2 Repository Structure

The repository consists of the following files:

Jupyter Notebook:

- `VQA_Implementation.ipynb`: Complete implementation of all the Python scripts in a Jupyter notebook format.

Python Scripts:

- main.py: Main script to run the VQA model.
- model.py: Contains the implementation of the VQA model architecture.
- train.py: Script for training the VQA model.
- utils.py: Utility functions used in the project.
- datasets.py: Script for loading and preprocessing the dataset.

Dataset Files:

- train.csv and test.csv: CSV files containing training and testing questions mapped to images.
- train-images.txt and test-images.txt: Text files containing the image lists for training and validation.
- qa-full.txt: Text file containing all the questions in the dataset.

Note: The dataset used for this implementation can be obtained from Kaggle at https://www.kaggle.com/datasets/anantkacholia/mmdp-vqa. Due to the large size of the images (over 400MB), they are not included in this repository.

6.3 Usage Instructions

To use the implementation:

1. Download the notebook and the MMDP-VQA dataset from Kaggle.
2. Update the paths and directory names in the notebook accordingly.
3. Install the required dependencies.
4. Run the notebook (GPU needed for faster processing).

Alternatively, users can upload the notebook to Kaggle, import the dataset, turn on the GPU accelerator, and run the notebook directly. Alternatively:

1. Assemble the Python files into a single directory.
2. Change the directory accordingly to where the Python files and dataset are located.
3. Download the MMDP-VQA dataset from Kaggle or the provided source.
4. Install the required dependencies and packages using the provided requirements.txt file.
5. Ensure that GPU acceleration is available and enabled for faster processing.
6. Execute the assembled Python files by running the main script.

6.4 Code Description

Description of VQA_Implementation.ipynb: The notebook serves as the central hub for implementing the Visual Question Answering (VQA) system, covering all functionalities of the Python scripts in a user-friendly format.

main.py: The main entry point for training the VQA model, handling dataset loading, model initialization, and training.

model.py: Contains the VQA model architecture.

train.py: Manages the training process, including dataset loading, preprocessing, model initialization, and training loop.

utils.py: Provides utility functions for data loading, preprocessing, and dataset splitting.

datasets.py: Implements the dataset class for loading and preprocessing data.

Note: The dataset used for this implementation is available on Kaggle as "MMDP-VQA." There is no alternate source due to the large size of the images (over 400MB), which are not included in the repository.

6.5 GitHub Repository

The implementation is available on GitHub at: https://github.com/ANANTKACHOLIA/MMDP_project_VQA.

6.6 Data Preprocessing

The data preprocessing stage involves setting up the necessary paths for loading the dataset and images. Additionally, it includes importing essential PyTorch and TorchVision modules for building and training the VQA model. The WordNet Wu-Palmer Similarity function, wups, is defined to quantify semantic similarity between words or concepts using NLTK's WordNet interface. Utility functions for loading, preprocessing statements, and generating the answer space are introduced to facilitate data manipulation.

6.7 Model

The model subsection describes the architecture of the VQA model, encapsulated within the `VQAModel` class. Key components such as image embedding, text embedding, LSTM layer, and fully connected layers are outlined to provide an understanding of how the model processes visual and textual information to predict answers. Additionally, file paths for dataset loading, text data preprocessing, and data conversion functions are explained to set the groundwork for subsequent sections.

7 Results

7.1 Research Paper Results

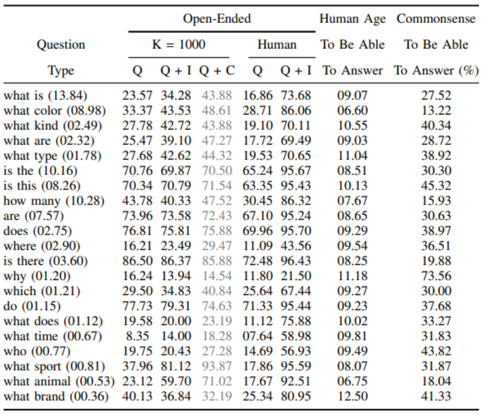
The results demonstrate a nuanced performance across different models and question types in Visual Question Answering (VQA) tasks. While vision-alone models exhibit low accuracy, language-alone methods surprisingly achieve relatively high performance, possibly due to their ability to exploit statistical priors in questions. However, the best-performing model, a deeper LSTM combined with normalized image features, significantly outperforms both vision-alone and language-alone baselines, although falling short of human-level performance. The analysis by question type highlights varying degrees of success, with certain question types benefiting more from scene-level information than others. Overall, while advancements have been made, achieving human-level performance remains a challenging task in VQA.

7.2 Project Results

We trained the model for 10 epochs on the MMDP-VQA dataset using the architecture defined above with an initial learning rate of 1×10^{-5}. The results of the project are as follows:

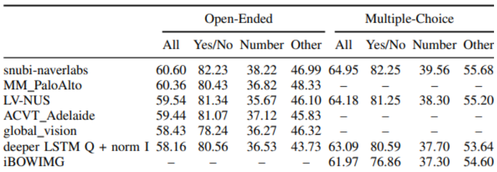
- Epoch 10
  - Training loss: 2.2259
  - Validation loss: 2.2259
  - Validation accuracy: 0.2939

It's worth noting that while accuracy is commonly used as a metric for evaluating model performance, it may not be the most appropriate measure for the VQA problem. Answers to visual questions can be subjective, and there may be multiple correct answers for a single question. Therefore, while accuracy provides a quantitative measure of performance, it may not fully capture the model's ability to understand and respond to the complexities of visual questions.



Question	Open-Ended					Human Age	Commonsense
	K = 1000			Human		To Be Able	To Be Able
Type	Q	Q + I	Q + C	Q	Q + I	To Answer	To Answer (%)
what is (13.84)	23.57	34.28	43.88	16.86	73.68	09.07	27.52
what color (08.98)	33.37	43.53	48.61	28.71	86.06	06.60	13.22
what kind (02.49)	27.78	42.72	43.88	19.10	70.11	10.55	40.34
what are (02.32)	25.47	39.10	47.27	17.72	69.49	09.03	28.72
what type (01.78)	27.68	42.62	44.32	19.53	70.65	11.04	38.92
is the (10.16)	70.76	69.87	70.50	65.24	95.67	08.51	30.30
is this (08.26)	70.34	70.79	71.54	63.35	95.43	10.13	45.32
how many (10.28)	43.78	40.33	47.52	30.45	86.32	07.67	15.93
are (07.57)	73.96	73.58	72.43	67.10	95.24	08.65	30.63
does (02.75)	76.81	75.81	75.88	69.96	95.70	09.29	38.97
where (02.90)	16.21	23.49	29.47	11.09	43.56	09.54	36.51
is there (03.60)	86.50	86.37	85.88	72.48	96.43	08.25	19.88
why (01.20)	16.24	13.94	14.54	11.80	21.50	11.18	73.56
which (01.21)	29.50	34.83	40.84	25.64	67.44	09.27	30.00
do (01.15)	77.73	79.31	74.63	71.33	95.44	09.23	37.68
what does (01.12)	19.58	20.00	23.19	11.12	75.88	10.02	33.27
what time (00.67)	8.35	14.00	18.28	07.64	58.98	09.81	31.83
who (00.77)	19.75	20.43	27.28	14.69	56.93	09.49	43.82
what sport (00.81)	37.96	81.12	93.87	17.86	95.59	08.07	31.87
what animal (00.53)	23.12	59.70	71.02	17.67	92.51	06.75	18.04
what brand (00.36)	40.13	36.84	32.19	25.34	80.95	12.50	41.33

Figure 5: TABLE 2: Open-ended test-dev results for different question types on real images (Q+C is reported on val). Machine performance is reported using the bag-of-words representation for questions. Questions types are determined by the one or two words that start the question. The percentage of questions for each type is shown in parentheses. Last and second last columns respectively show the average human age and average degree of commonsense required to answer the questions (as reported by AMT workers), respectively. See text for details.



	Open-Ended				Multiple-Choice			
	All	Yes/No	Number	Other	All	Yes/No	Number	Other
snubi-naverlabs	60.60	82.23	38.22	46.99	64.95	82.25	39.56	55.68
MM_PaloAlto	60.36	80.43	36.82	48.33	–	–	–	–
LV-NUS	59.54	81.34	35.67	46.10	64.18	81.25	38.30	55.20
ACVT_Adelaide	59.44	81.07	37.12	45.83	–	–	–	–
global_vision	58.43	78.24	36.27	46.32	–	–	–	–
deeper LSTM Q + norm I	58.16	80.56	36.53	43.73	63.09	80.59	37.70	53.64
iBOWIMG	–	–	–	–	61.97	76.86	37.30	54.60

Figure 6: TABLE 3: Test-standard accuracy of our best model (deeper LSTM Q + norm I) compared to test-standard accuracies of other entries for the open-ended and multiple-choice tasks in the respective VQA Real Image Challenge leaderboards (as of October 28, 2016.

8 Limitations

Visual Question Answering (VQA) systems face several limitations across different aspects, ranging from dataset collection challenges to subjective answers and evaluation constraints.

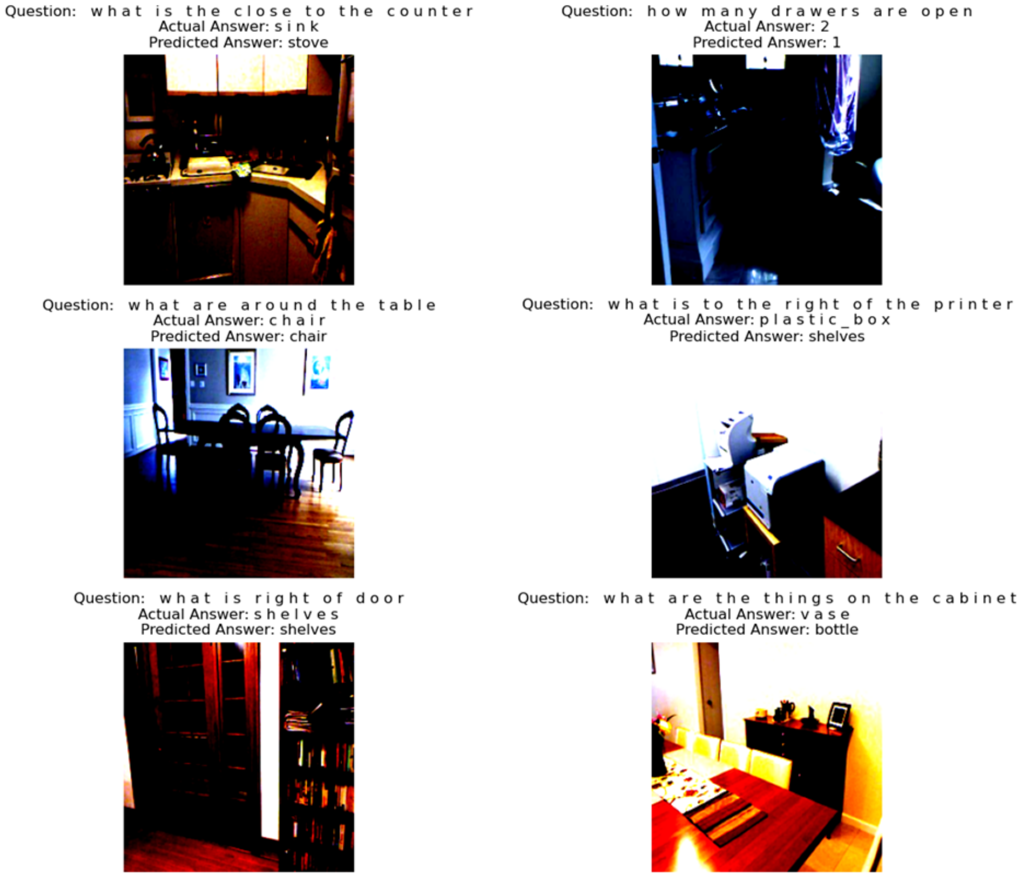

Question: what is the close to the counter
Actual Answer: sink
Predicted Answer: stove
Question: how many drawers are open
Actual Answer: 2
Predicted Answer: 1
Question: what are around the table
Actual Answer: chair
Predicted Answer: chair
Question: what is to the right of the printer
Actual Answer: plastic_box
Predicted Answer: shelves
Question: what is right of door
Actual Answer: shelves
Predicted Answer: shelves
Question: what are the things on the cabinet
Actual Answer: vase
Predicted Answer: bottle

Figure 7: We conducted a detailed analysis of the model's performance on various visual questions. Notably, in several instances, the model provided answers that were contextually correct but differed from the ground truth. For instance, in image row (1), column 1, the question "What is close to the counter?" had the actual answer "sink," while the model predicted "stove," which is also contextually correct. Similarly, in row 1, column 2, the model incorrectly counted the number of open drawers. However, in row 2, column 1, the model accurately identified the chair despite slight variations in perspective. Moreover, in row 3, column 3, both "vase" and "bottle" were present on the cabinet, demonstrating the model's ability to recognize multiple objects. These instances highlight the subjective nature of visual question answering, where answers may vary based on perspective and context. Therefore, while accuracy provides a quantitative measure of performance, it may not fully capture the model's ability to understand and respond to the complexities of visual questions, as evidenced by the discrepancies between predicted and ground truth answers.

8.1 Dataset Collection Challenges

One significant limitation arises from the challenges in collecting comprehensive and diverse datasets for training and evaluation. Dataset biases, such as imbalanced question types or overrepresented image categories, can affect the model's generalization ability and performance on real-world scenarios.

8.2 Subjective Answers

VQA inherently involves subjective answers, where multiple correct responses may exist for a given question based on context, perspective, or interpretation. This subjectivity introduces ambiguity and variability in evaluating model performance, as answers may differ even among human annotators.

8.3 Evaluation Constraints

Evaluating VQA models is challenging due to the subjective nature of answers. Traditional metrics like accuracy may not adequately capture the model's performance, especially when answers are subjective or context-dependent. Additionally, evaluation datasets may not fully represent the diversity and complexity of real-world scenarios, leading to potential performance discrepancies.

8.4 General Constraints

Other constraints include computational resources required for training and inference, as VQA models often involve complex neural architectures and large-scale datasets. Limited computational resources can restrict the scalability and applicability of VQA systems, particularly in resource-constrained environments or real-time applications.

9 Data and Resource Constraints

During the development of our project, we encountered several data and resource constraints. Limited access to annotated datasets with sufficient diversity and scale posed challenges in training robust models. Additionally, constraints in computational resources, including GPU availability and memory limitations, impacted model training times and experimentation possibilities. These constraints necessitated careful optimization and trade-offs to ensure efficient utilization of available resources while maintaining reasonable performance levels.

10 Conclusion

In this paper, we endeavored to reproduce and extend upon existing approaches in Visual Question Answering (VQA), aiming to address the complexities and challenges inherent in this task. Through comprehensive experimentation and analysis, we strived to contribute insights into the effectiveness and limitations of current VQA models.

Our efforts involved reproducing state-of-the-art architectures and methodologies, training models on the MMDP-VQA dataset for 10 epochs with an initial learning rate of 1×10^{-5}. Despite achieving notable results, it became evident that accuracy alone is not a sufficient measure for evaluating VQA systems, given the subjective nature of answers and the potential for varied interpretations.

Throughout our exploration, we encountered various limitations, including dataset collection challenges, subjective answers, and evaluation constraints. These constraints underscored the need for robust evaluation metrics and diverse, representative datasets to ensure the reliability and generalizability of VQA models.

Despite these challenges, our work provides valuable insights and contributions to the field of VQA. By highlighting the complexities and limitations inherent in VQA systems, we aim to inspire further research and innovation towards developing more robust, interpretable, and context-aware models for visual question answering.

In conclusion, while our study contributes to the ongoing discourse on VQA methodologies and challenges, it also emphasizes the importance of continual refinement and exploration in this dynamic and evolving field.

References

[1] Aishwarya Agrawal, Jiasen Lu, Stanislaw Antol, Margaret Mitchell, C. Lawrence Zitnick, Dhruv Batra, Devi Parik. ”VQA: Visual Question Answering.” *www.visualqa.org*.

[2] H. Agrawal et al., ”Cloudcv: Large-scale distributed computer vision as a cloud service,” in *Mobile Cloud Visual Media Computing*, Springer International Publishing, 2015.

[3] M. Malinowski and M. Fritz, ”A Multi-World Approach to Question Answering about Real-

World Scenes based on Uncertain Input," in *NIPS*, 2014.

[4] K. Tu et al., "Joint Video and Text Parsing for Understanding Events and Answering Queries," *IEEE MultiMedia*, 2014.

[5] J. P. Bigham et al., "VizWiz: Nearly Real-Time Answers to Visual Questions," in *User Interface Software and Technology*, 2010.

[6] X. Lin and D. Parikh, "Don't just listen, use your imagination: Leveraging visual common sense for non-visual tasks," in *CVPR*, 2015.

[7] J. Weston et al., "Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks," CoRR, abs/1502.05698, 2015.

[8] A. Fader et al., "Paraphrase-Driven Learning for Open Question Answering," in *ACL*, 2013.

[9] M. Richardson et al., "MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Text," in *EMNLP*, 2013.

[10] M. Mitchell et al., "Midge: Generating Image Descriptions From Computer Vision Detections," in *ACL*, 2012.