# Visuo-Linguistic Question Answering (VLQA) Challenge

**Shailaja Keyur Sampat, Yezhou Yang and Chitta Baral**
Arizona State University
Tempe, AZ, USA
{ssampa17*, yz.yang, chitta}@asu.edu

## Abstract

Understanding images and text together is an important aspect of cognition and building advanced Artificial Intelligence (AI) systems. As a community, we have achieved good benchmarks over language and vision domains separately, however joint reasoning is still a challenge for state-of-the-art computer vision and natural language processing (NLP) systems. We propose a novel task to derive joint inference about a given image-text modality and compile the Visuo-Linguistic Question Answering (VLQA) challenge corpus in a question answering setting. Each dataset item consists of an image and a reading passage, where questions are designed to combine both visual and textual information i.e., ignoring either modality would make the question unanswerable. We first explore the best existing vision-language architectures to solve VLQA subsets and show that they are unable to reason well. We then develop a modular method with slightly better baseline performance, but it is still far behind human performance. We believe that VLQA will be a good benchmark for reasoning over a visuo-linguistic context. The dataset, code and leaderboard is available at https://shailaja183.github.io/vlqa/.

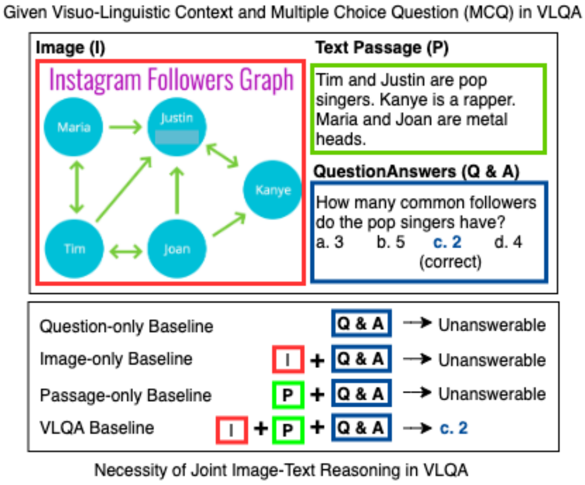

Given Visuo-Linguistic Context and Multiple Choice Question (MCQ) in VLQA
Image (I)
Instagram Followers Graph
Maria
Justin
Kanye
Tim
Joan
Text Passage (P)
Tim and Justin are pop singers. Kanye is a rapper. Maria and Joan are metal heads.
QuestionAnswers (Q & A)
How many common followers do the pop singers have?
a. 3 b. 5 c. 2 d. 4
(correct)
Question-only Baseline
Q & A → Unanswerable
Image-only Baseline
I + Q & A → Unanswerable
Passage-only Baseline
P + Q & A → Unanswerable
VLQA Baseline
I + P + Q & A → c. 2
Necessity of Joint Image-Text Reasoning in VLQA

Figure 1: Example of Visuo-Linguistic Question Answering (VLQA) task for joint reasoning over image-text context.

## 1 Introduction

Question answering (QA) is a crucial way to evaluate the system's ability to understand text and images. In recent years, a large body of natural language QA (NLQA) datasets and visual QA (VQA) datasets have been compiled to evaluate the ability of a system to understand text and images. For most VQA datasets, the text is used merely as a question-answering mechanism rather than an actual modality that provides contextual information. On the other hand, deriving inference from combined visual and textual information is an important skill for humans to perform day-to-day tasks. For example, product assembly using instruction manuals, navigating roads while following street signs, interpreting visual representations (e.g., charts) in various documents such as newspapers and reports, understanding concepts using textbook-style learning, etc. The importance of joint reasoning has also been emphasized in the design of standardized / psychometric tests like PISA (OECD, 2019) and GRE[1], as evident from Figure 2. PISA assessments conducted post 2018 take into account "the evolving nature of reading in digital societies- which requires an ability to compare, contrast and integrate information from multiple sources". The GRE has 'data interpretation' questions that assess a student's ability to "analyze given data as a combination of text and charts."

Both the aforementioned evidence motivate the need to develop Visuo-Linguistic QA (VLQA) system, posing a further challenge to state-of-the-art

*corresponding author

[1] https://www.oecd.org/pisa/, https://www.ets.org/gre/

*Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4606–4616
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

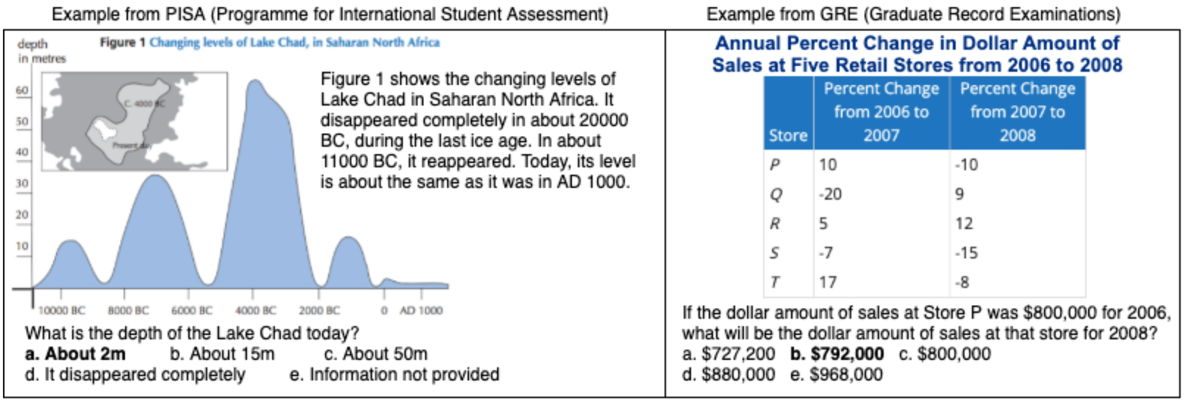


Figure 1 shows the changing levels of Lake Chad in Saharan North Africa. It disappeared completely in about 20000 BC, during the last ice age. In about 11000 BC, it reappeared. Today, its level is about the same as it was in AD 1000.

What is the depth of the Lake Chad today?
**a. About 2m** b. About 15m c. About 50m
d. It disappeared completely e. Information not provided

Annual Percent Change in Dollar Amount of Sales at Five Retail Stores from 2006 to 2008

| Store | Percent Change from 2006 to 2007 | Percent Change from 2007 to 2008 |
|---|---|---|
| *P* | 10 | -10 |
| *Q* | -20 | 9 |
| *R* | 5 | 12 |
| *S* | -7 | -15 |
| *T* | 17 | -8 |

If the dollar amount of sales at Store P was $800,000 for 2006, what will be the dollar amount of sales at that store for 2008?
a. $727,200 **b. $792,000** c. $800,000
d. $880,000 e. $968,000

Figure 2: Examples of joint-reasoning questions in standardized tests[2](boldface represents correct answer)

vision and language research. There are no benchmarking datasets that focus on reasoning over both images and text to our best knowledge. We formalize the task of deriving joint inference, where a system must utilize both visual and textual information to correctly answer the question, as demonstrated in Figure 1. To create a benchmark for this task, we develop and present a new dataset: VLQA (Visuo-Linguistic Question Answering)[3] as our main contribution. VLQA dataset consists of text together with a diverse range of visual elements. Since manuals, documents and books containing texts and visuals are ubiquitous, the VLQA dataset is very much grounded in the real world. The dataset is curated from multiple resources (books, encyclopedias, web crawls, existing datasets, etc.) through combined automated and manual efforts. The dataset consists of 9267 image-passage-QA tuples with detailed annotation, which are meticulously crafted to assure its quality.

We then evaluate the best existing vision-language architectures with respect to our VLQA dataset. This includes LXMERT (Tan and Bansal, 2019), VL-BERT (Lu et al., 2019), ViLBERT (Su et al., 2019) and VisualBERT (Li et al., 2019). Our results demonstrate that despite a significant improvement over vision and language tasks separately, the best existing techniques cannot reason well on the joint tasks. We then propose a modular method HOLE (HOpping and Logical Entailment), which demonstrates slightly better baseline performance and offers more transparency for the interpretation of intermediate outputs. The results indicate that VLQA task is relatively harder compared to existing vision-language tasks due to diversity of figures and additional textual component, demanding the need of better approaches to tackle multi-modal question answering. The VLQA challenge thus has the potential to open new research avenues spanning language and vision.

## 2 Related Work

We identify Image-Text Multi-modality, Multi-hop Reasoning and variants of Visual Question Answering (VQA) closest to VLQA and compare with relevant datasets in these areas (refer Appendix A.1 for comprehensive comparison with more datasets).

### 2.1 Image-Text Multi-modality

Multimodal learning aims to build models that can process and relate information from two or more modalities. Image-Text multi-modality has received growing interest from the Artificial Intelligence (AI) community recently. Diagram QA component of TQA (Kembhavi et al., 2017) and a portion of AI2D (Kembhavi et al., 2016) with additional text are most relevant to ours. They share similarities with VLQA in terms of the presence of additional text, diagram style images and QA style evaluation, but there are important distinctions.

First, TQA uses long lessons ($\sim$50 sentences and 4-5 images) to describe concepts in textbook-style learning, whereas text passages for subsets of AI2D and VLQA are short (1-5 sentences). The goal of TQA aligns with the careful selection of necessary facts from the long-tailed contexts, which is perhaps less important in VLQA as the context is much smaller. At the same time, AI2D aims at AI-based diagram understanding. Contrary to that, we

[2] Often, additional text and question are combined in standardized tests, but we segregate them into Passage and Question for the ease of processing and structured dataset design.

[3] Creation of VLQA is purely research-oriented; By referring standardized tests as an inspiration, comparison with professional organizations like ETS or OECD is not intended.

focus on enhancing the capability of AI models for joint reasoning. Secondly, AI2D and TQA are curated from the school science curriculum whereas, we have a broader horizon of possible reasoning. Lastly, TQA and AI2D do not impose that one must use both modalities while answering, unlike VLQA. For TQA, one can answer 40% of text QA using a single sentence and 50% of diagram QA using the only image. In that case, a significant portion of the dataset becomes analogous to machine comprehension or ordinary VQA, losing out on the actual purpose of multi-modality.

### 2.2 Multi-Hop Reasoning

In the natural language processing (NLP) domain, multi-hop reasoning is proposed to encourage the development of models that can reason about two or more textual contexts. QAngaroo (Welbl et al., 2018) and ComplexWebQuestions (Talmor and Berant, 2018) include multi-hop questions that can be answered by linking entities from a knowledge base (KB). HotpotQA (Yang et al., 2018) is a multi-hop benchmark over pairs of text paragraphs from wikipedia, not being constrained by retrieval from fixed KB schemas. QASC (Khot et al., 2019) dataset made this task further challenging, which first requires to retrieve necessary facts from a large corpus (knowledge ranking) and compose them to answer a multi-hop question.

Solving VLQA examples requires linking information from image and text. Therefore, VLQA can be considered a novel kind of multi-hop task involving images and text, which we believe will drive future vision-language research.

### 2.3 Visual Question Answering (VQA)

Followed by the success of the VQA dataset (Antol et al., 2015), several variants of visual QA have been proposed. The following are most relevant;

**Reasoning-based VQA** Reasoning-based VQA datasets aim at measuring a system's capability to reason about a set of objects, their attributes and relationships. HowManyQA (Trott et al., 2017) and TallyQA (Acharya et al., 2019) have object counting questions over images. SNLI-VE (Xie et al., 2019), VCOPA (Yeo et al., 2018) focus on causal reasoning whereas CLEVR (Johnson et al., 2017), NLVR (Suhr et al., 2017) target spatial reasoning. FigureQA (Kahou et al., 2017), DVQA (Kafle et al., 2018) are testbeds for QA over charts/plots. The objective of VLQA is to equip AI models with diverse reasoning capabilities over the image-text context. A model solving VCR (Zellers et al., 2019) dataset first answers a question in VQA style, then needs to provide a rationale explaining why the answer is true. Therefore, items in VCR could be turned to particular VLQA data items. However, images in VCR are much more specific than ours e.g., they do not have charts, diagrams, or multiple images. Also, the rationale selection is limited to 'Why' questions, not so in VLQA. We identify 10 broad reasoning categories needed to solve VLQA, which is described in Section 3.3.

**Knowledge-based VQA** There are several vision-language tasks that require additional knowledge beyond the provided image and text. F-VQA (Wang et al., 2018), KB-VQA (Wang et al., 2015) and KVQA (Shah et al., 2019) rely on retrieving commonsense or world-knowledge from a Knowledge Base (KB), whereas OK-VQA (Marino et al., 2019) is related to open-ended knowledge extraction from the web. In VLQA, 61% of samples require commonsense or domain knowledge, which is not explicitly stated in image-text context. Knowledge extraction for VLQA is kept open-ended as of now.

## 3 VLQA Dataset

We formally define the VLQA task, explain our approach to curate this dataset and necessary measures for quality assurance below;

### 3.1 Task Overview

A datapoint in VLQA is a 4-tuple <I, P, Q, A>;

**Image(I)** It is provided imagery, which ranges from daily life scenes, a variety of data representations to complex diagrams. A portion of VLQA examples also requires reasoning over multiple images. For the simplicity of processing and retrieval, we compose all images into a single file. Each image is bounded by a red box and provided an explicit detection tag ([0],[1],..) for identification purposes, inspired by VCR (Zellers et al., 2019) annotations. This also provides a convenient way to reference images in passage, question, or answers.

**Passage(P)** It is a textual modality that provides additional contextual information related to the image. The passages in VLQA dataset is composed of 1-5 sentences, which consists of facts, imaginary scenarios or their combination.

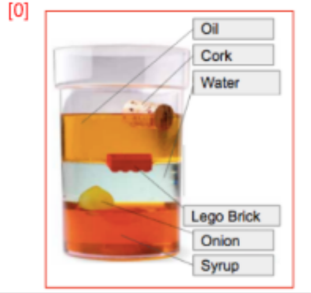
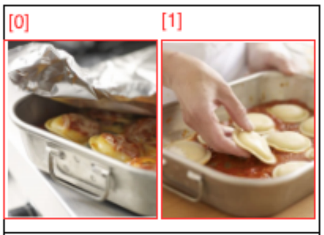
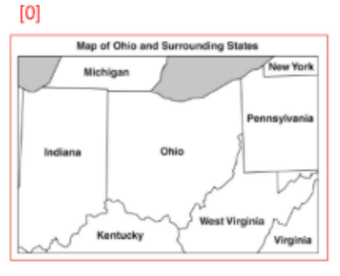
|  | Example 1 | Example 2 | Example 3 |
|---|---|---|---|
| Image(s) (I) | [0]  | [0] [1]  | [0]   Map of Ohio and Surrounding States<br>Michigan<br>New York<br>Pennsylvania<br>Indiana<br>Ohio<br>Kentucky<br>West Virginia<br>Virginia |
| Text Passage (P) | Objects and liquids float on liquids of higher density and sink through liquids of lower density. | I. [0]<br>II. Heat oven to 375F, apply a thin layer of sauce at the bottom of a baking dish.<br>III. Top with spinach, mozzarella and a third of the remaining sauce. Repeat this for another layer.<br>IV. [1] | A child in northern Ohio lost a balloon in a storm. The wind flow during the storm was observed in the North-West direction. |
| Question (Q) | Based on [0], can we say that density of water d(W) is lower than density of Lego brick d(B)? | Choose the correct order of the events I to IV above to prepare Lasagna. | Toward which state would the balloon most likely travel? |
| Answer Choices (A) | a. Yes<br>**b. No** | a. III-I-II-IV b. III-II-I-IV<br>c. II-III-IV-I **d. II-IV-III-I** | a. **Michigan**<br>b. Indiana<br>c. Pennsylvania<br>d. Kentucky |
| Classification | ImageType: Free-form Figure<br>AnswerType: Binary Classification<br>KnowledgeType: No external knowledge required (provided information is sufficient to answer)<br>Difficulty level (human): Easy<br>Source: Encyclopedia | ImageType: Natural Images<br>AnswerType: 4-way Sequencing<br>KnowledgeType: External knowledge (procedural knowledge of cooking) + multi-step Inference<br>Difficulty level (human): Moderate<br>Source: ReceipeQA Dataset | ImageType: Templated Figure (Map)<br>AnswerType: 4-way Text MCQ<br>KnowledgeType: External knowledge (light objects move in the direction of wind) + single-step Inference<br>Difficulty level (human): Easy<br>Source: AI2 Mercury Dataset |

Figure 3: **Examples from VLQA Train Set**. Each example contains image, corresponding text passage and Multiple Choice Question (MCQ) with correct answer choice highlighted by the boldface. Further, each sample is classified based on image type, answer type, knowledge/reasoning type and human annotated difficulty level. (For more examples, refer to A.4 or visit dataset webpage

**Question(Q)** It is a question in natural language that tests the reasoning capability of a model over a given image-passage context. In addition to standard 'Wh' patterns and fact-checking style (True/False), some questions in VLQA are of 'do-as-directed' form, similar to standardized tests.

**Answer Choices(A)** VLQA is formed as a classification task over 2-way or 4-way plausible choices, with exactly one of the candidate answers being correct. Answer choices may contain boolean, alphanumeric phrases, image tags or their combination.

**Task** Given the VLQA dataset as a collection of 4-tuple <I, P, Q, A> as shown in Figure 3, the task is to build an AI model that can answer a given question using image-text multi-modal context. The correctness of the prediction is measured against the ground-truth answer. Additionally, we provide rich annotations and classification on several aspects such as image types, question types, required reasoning capability and need for external knowledge. However, this metadata is optional and useful for researchers interested in tackling specific subsets of VLQA.

## 3.2 Constructing VLQA

### 3.2.1 Data Collection

The main goal of our work is to collect a QA dataset that requires to derive joint inference from image-text modality. We classify our data sources as Primary and Secondary;

We obtain raw textual/visual information through primary sources, which can be later used as a modality in VLQA. For example, text crawls from wikipedia containing facts or images crawled by keyword-search can be used as passage and image respectively. Similarly, we collect tabular data from CIA 'world factbook' (Central Intelligence Agency, 2019), WikiTables (Pasupat and Liang, 2015) and convert them into templated figures like bar charts, pie charts, scatter plots, etc. We consider existing structured or semi-structured materials as a secondary data source, which can be quickly manipulated to use for our purpose; educational materials, standardized tests, and existing vision-language datasets are important. We used scrapers to collect textbook exercises, encyclopedias, practice worksheets and question banks. Further, we obtained a subset of interesting samples from existing datasets

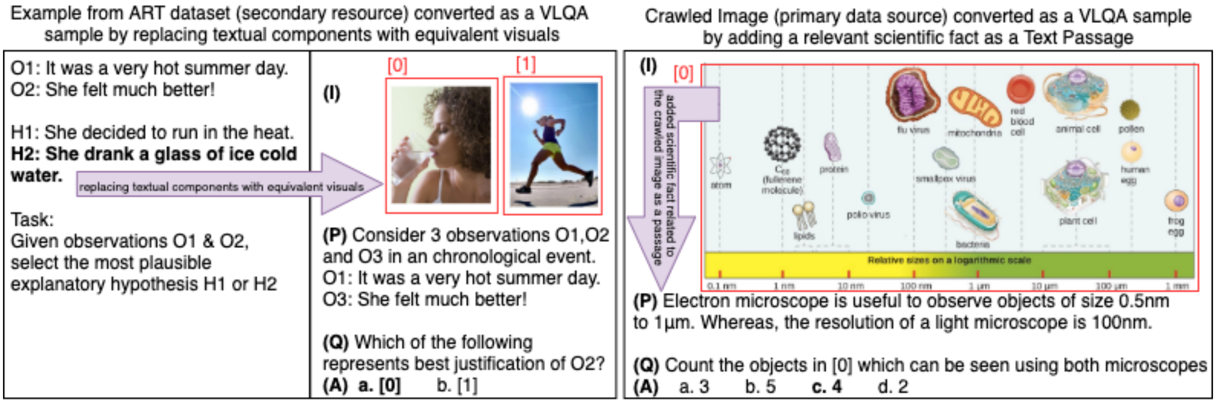

Example from ART dataset (secondary resource) converted as a VLQA sample by replacing textual components with equivalent visuals
O1: It was a very hot summer day.
O2: She felt much better!
H1: She decided to run in the heat.
H2: She drank a glass of ice cold water.
Task:
Given observations O1 & O2, select the most plausible explanatory hypothesis H1 or H2
replacing textual components with equivalent visuals
(I)
[0]
[1]
(P) Consider 3 observations O1,O2 and O3 in an chronological event.
O1: It was a very hot summer day.
O3: She felt much better!
(Q) Which of the following represents best justification of O2?
(A) a. [0] b. [1]
Crawled Image (primary data source) converted as a VLQA sample by adding a relevant scientific fact as a Text Passage
(I)
[0]
added scientific fact related to the crawled image as a passage
atom
C60 (fullerene molecule)
lipids
protein
polio virus
flu virus
smallpox virus
mitochondria
bacteria
red blood cell
animal cell
plant cell
pollen
human egg
frog egg
Relative sizes on a logarithmic scale
0.1 nm
1 nm
10 nm
100 nm
1 µm
10 µm
100 µm
1 mm
(P) Electron microscope is useful to observe objects of size 0.5nm to 1µm. Whereas, the resolution of a light microscope is 100nm.
(Q) Count the objects in [0] which can be seen using both microscopes
(A) a. 3 b. 5 c. 4 d. 2

Figure 4: VLQA data creation process: collect data using primary and secondary sources, then perform post-processing (if any), then finally create question-answers that require joint reasoning.

such as RecipeQA (Yagcioglu et al., 2018), WikiHow (Koupaee and Wang, 2018), PhysicalIQA (Bisk et al., 2019), ART (Bhagavatula et al., 2019) and TQA (Kembhavi et al., 2017).

We then refactor textual/ visual information collected from the above sources and mold it as per our task requirements. Figure 4 illustrates this process. Refactoring includes manual or semi-automated post-processing such as replacing given textual/visual attributes with equivalent visual/textual counterparts, adding/removing partial information to/from text or visuals, and creating factual or hypothetical situations around images. Then we standardize all information collected the using above methods as Multiple Choice Questions (MCQ) and get the initial version of the dataset.

Since we impose the condition that a question must be answered through joint reasoning over both the modalities, our annotation process becomes non-trivial and requires careful manual annotation. We opted for a limited number of in-house expert annotators for quality purposes rather than a noisier hard-to-control crowdsourcing alternative.

### 3.2.2 Ensuring dataset integrity

A combined understanding about visual and textual inputs is a key aspect of the VLQA task. As we model it as a classification task, some models might exploit various biases in the dataset to get good performance without proper reasoning. To discourage such models, we employ 3-level verification over the full dataset to ensure the quality.

Firstly, for all collected image-passage pairs, human annotators quickly verify if a portion of image and passage represent identical information. All such image-passage pairs are discarded from the dataset. Secondly, we create 3 baselines- question-only, passage-only and image-only which ignore at least one modality (among image and passage) and try to predict answers. We repeat this experiment 3 times by shuffling answer choices with a fixed seed. We remove samples that are answered correctly by any unimodal baseline in all trials.

Finally, we perform another round of manual quality checks. We instruct workers first to answer a question based only on image(s) and then try to answer a question based only on the text passage. If a question can be answered using a single modality, we suggest annotators to mark the checkbox. Finally, we look over all bad samples and either provide a fix or remove, on a case-by-case basis. (refer Appendix A.2 for detailed explanation on dataset creation process)

### 3.3 VLQA Dataset Analysis

In this section we analyze VLQA on following aspects; Table 1 provides a summary of relevant statistics.

**Multi-modal Contexts** The final version of the VLQA dataset has 9267 unique image-passage-QA items. For each item, the multi-modal context is created by pairing images (roughly 10k collected) with the relevant text passages (roughly 9k retrieved or manually written).

**Text-length Analysis** We provide analysis about lengths of various textual components in our dataset i.e., passages, questions and answers. Length of each textual component is calculated by counting the tokens separated by whitespaces and then averaged out across the dataset. The average passage length of 34.1 tokens indicates that in

VLQA textual contexts are relatively smaller than Reading Comprehension tasks and in most cases, it contains precise context necessary for the joint reasoning. The average question length of 10.0 tokens is larger compared to most other VQA datasets provided in (Hudson and Manning, 2019). Shorter answer lengths (1.7 tokens) suggest that most of the dataset questions have short answers, which provides inherent flexibility if someone wants to leverage generative models to solve this task. The dataset has a vocabulary size of 13259, contributed by all three textual components together.

**Image types** We categorize images in VLQA into 3 major kinds: Natural Images, Template-based Figures and Free-form Figures. Natural images incorporate day-to-day scenes around us, containing abundant objects and actions. Template-based figures are visuals that follow a common structure for information representation. We further categorize template-based figures into 20 sub-types like bar, pie, maps, tables, cycles, processes, etc. The images which neither fit in any templates nor are natural have been put into a free-form category (e.g., science experiments, hypothetical scenarios, etc.). In VLQA, it is also possible that the visual context has multiple related images to reason about.

**Answer types** 4-way or 2-way image MCQ contains 4 and 2 images as plausible answer choices respectively, where the model needs to correctly pick the image best described by the passage and question. 4-way or 2-way text MCQ contains 4 and 2 alphanumeric text as plausible answer choices respectively, where the model needs to reason about given image-text scenario and pick the most likely answer to the question. 4-way Sequencing task assesses a model's capability to order 4 spatial or temporal events represented as a combination of images and text. Binary Classification (Yes/No or True/False) can be considered a fact-checking task where we want to determine the truth value of a question provided image-passage context.

**Knowledge and Reasoning types** 61% of VLQA items are observed to incorporate some commonsense or domain knowledge beyond the provided context. This missing knowledge has to be retrieved through the web. The remaining 39% samples can be answered through a simple join of information from visuo-linguistic context. We observe the following 10 most-frequent reasoning types needed to solve VLQA questions; conditional retrieval, math operations, deduction, temporal, spatial, causal, abductive, logical, and verbal reasoning. We further categorize VLQA samples based on whether it requires a single-step or multi-step inference to answer the question. By multi-step inference, we mean that answering a question involves more than one reasoning types.

| Measure | Stats. |
|---|---|
| **Multimodal Context** | |
| Total #Images | 10209 |
| #Unique Text Passages | 9156 |
| #Questions | 9267 |
| **Text-length Analysis** | |
| Avg. Passage Length | 34.1 |
| Avg. Question Length | 10.0 |
| Avg. Answer Length | 1.7 |
| Vocabulary Size | 13259 |
| **Image types** | |
| Natural Images | 4445 |
| Templated Figures | 3920 |
| Free-form Figures | 1854 |
| **Answer types** | |
| 4-way image MCQ | 1172 |
| 4-way text MCQ | 4647 |
| 4-way Sequencing | 1088 |
| 2-way image MCQ | 1088 |
| Binary Classification (T/F or Yes/No) | 1272 |
| **Knowledge/Reasoning types** | |
| No Ext. Knowledge required | 3145 |
| Ext. Knowledge+Single-step Inference | 2783 |
| Ext. Knowledge+Multi-step Inference | 2939 |
| **Difficulty Level (human annotated)** | |
| Easy | 4188 |
| Moderate | 2943 |
| Hard | 2136 |
| **Dataset Split** | |
| Train (80%) | 7413 |
| Test (10%) | 927 |
| Validation (10%) | 927 |

Table 1: VLQA Statistics and Diversity (MCQ is multiple choice questions, Ext. is External).

**Difficulty Level** Determining difficulty levels is a subjective notion therefore, we asked an odd number of annotators to rate VLQA items as 'easy', 'moderate', or 'hard' based on their personal opinion. Then we take a majority vote of all annotators to assign difficulty level to each question.

**Dataset Splits** VLQA contains 9267 items in <I,P,Q,A> format, with detailed classification based on figure types, answer types, reasoning skills, requirement of external knowledge and difficulty levels as explained above. The data is split in train-test-val (80-10-10%), ensuring the uniform distribution based on the above taxonomies. To preserve the integrity of the test results, we do not release the test set publicly. Note that the use of the metadata for model design is completely optional.

## 4 Benchmarking

**Human Performance** We performed human evaluation on 927 test samples with a balanced variety of questions by image types, answer types, knowledge/reasoning types and hardness. First, we ask 3 in-house experts to take tests in isolation. We also ask them to rate questions based on the difficulty levels (easy/medium/hard) and an option to mark a dataset sample 'ambiguous'. Then we match their predictions against ground-truth answers, which turned out to be 84%.

**Random Baseline** VLQA dataset contains 4-way and 2-way multiple choice questions (MCQs) where each answer choice is likely to be picked with 25% and 50% chance. Based on the answer-type distribution provided in Table 1, the performance of the random baseline is 31.36%.

**Question-only, Passage-only and Image-only Baselines** We use three unimodal baselines only for automated quality assurance of VLQA data (and do no not train) to prevent models from exploiting bias in data. Question-only, Passage-only and Image-only models are implemented using RoBERTa (Liu et al., 2019) finetuned on ARC (Clark et al., 2018), ALBERT (Lan et al., 2019) finetuned on RACE and LXMERT (Tan and Bansal, 2019) finetuned on VQA (Antol et al., 2015) respectively. We report the poor performance of these baselines over resulting VLQA data to indicate the need for joint reasoning over multi-modal context.

**Best Existing Architectures** Recently, several attempts have been made to derive transformer-based pre-trainable generic representations for visuo-linguistic tasks. We pick top-performing single-model architectures VL-BERT (Su et al., 2019), VisualBERT (Li et al., 2019), ViLBERT (Lu et al., 2019) and LXMERT (Tan and Bansal, 2019) that support Visual Question Answering (VQA) downstream task. For the VQA task, the input is an image and a question. To finetune VQA style models with VLQA data, we compose all images into one (in case of multiple images) as a single visual input, and concatenate Passage and Question as a single language input. Hyperparameters and Performance of all 4 architectures is reported in 2 and 3 respectively.

| **Model and Hyperparameters** |
|---|
| **VisualBERT** |
| Ft_VQA: EP=20, BS=256, LR=1e-4, WD=1e-4 |
| Ft_VLQA: BS=16, LR=2e-5, EP=15 |
| **VL-BERT** |
| Ft_VQA: BS=32, LR=2e-5, EP=10 |
| Ft_VLQA: BS=16, LR=1e-5, EP=10 |
| **ViLBERT** |
| Ft_VQA: BS=32, LR=1e-5, EP=20, WR=0.1 |
| Ft_VLQA: BS=32, LR=1e-5, EP=10 |
| **LXMERT** |
| Ft_VQA: BS=32, LR=5e-5, EP=4 |
| Ft_VLQA: BS=16, LR=5e-5, EP=8 |

Table 2: Manual finetuning of best existing architures with VQA followed by VLQA (BS-Batch Size, EP-Epochs, LR-Learning Rate, WD-Weight Decay, WR-Warmup Ratio, Ft.-Manual Finetuning)

## 5 Fusion of HOpping and Logical Entailment (HOLE) to solve VLQA

We propose 'HOLE'- a fusion of modality HOpping (Image-to-passage hop and Passage-to-Image hop) and Logical Entailment as a modular baseline for VLQA, shown in Figure 5. We leverage 'answer types' metadata from the annotations and learn a simple 5-class classifier ('4-way Image', '2-way Image', '4-way Sequencing', 'Binary Classification' or '4-way Text') in order to decide between modality hopping and logical entailment. Note that our model is not end-to-end.

### 5.1 Modality Hopping based Solver

**4-way text MCQ** are solved using modality hopping approach (lower half pipeline in Figure 5). We first compute Image-to-Question Attention (I2Q) and Passage-to-Question Attention (P2Q) scores to determine which modality is important as a starting point for solving a question. I2Q is computed

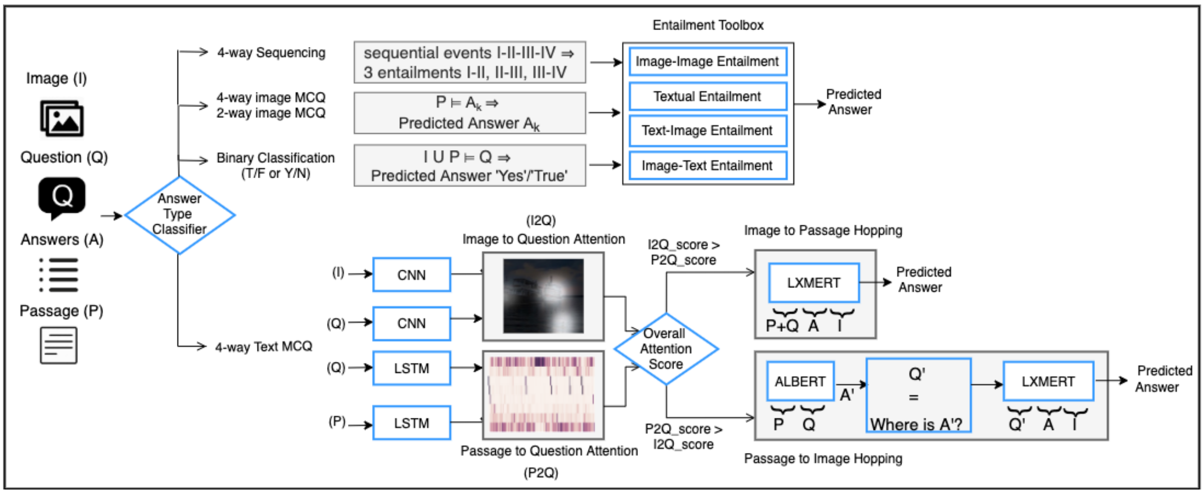


Figure 5: Proposed HOLE method to solve VLQA: Based on the answer type classification, a dataset item is solved as a sequence of Logical Entailment operations or performs Hopping between modalities to find the correct answer.

using Stacked Attention Network (SAN) (Yang et al., 2016), which takes Convolution Neural Network (CNN) encoding of I and Q. Whereas, P2Q is computed using a variant of Bi-Directional Attention Flow (BIDAF) (Seo et al., 2016) trained using Embeddings from Language Models (ELMo) (Peters et al., 2018) over Long-Short Term Memory (LSTM) encoding of Q and P.

A higher I2Q score suggests that Q has more overlap with I than P. Therefore, image modality should be used first and then incorporate passage to compute the answer. This is termed as an 'Image-to-Passage Hop'. This is identical to a Visual Question Answering (VQA) scenario that takes an image and a question as input. Since we have P as an additional text component, we combine passage (P+Q). This is implemented through pre-trained architecture LXMERT (Tan and Bansal, 2019) which is state-of-the-art on VQA that picks the most likely answer choice as a correct answer.

Similarly, a higher P2Q score suggests that Q has more overlap with P than I. Therefore, passage modality should be used first and then incorporate image to compute the answer. This is termed as a 'Passage-to-Image Hop'. This can be achieved by a machine comprehension model followed by a VQA model. We use ALBERT (Lan et al., 2019) as a machine comprehension model which takes in P and Q to generate an open-ended response in the style of SQuAD (Rajpurkar et al., 2016), which we refer to as A'. Now we want to determine where is A' located in the image I. Therefore, we formulate a new question Q' as "Where is A'?", where A' is substituted by the answer from ALBERT. We then use LXMERT (Tan and Bansal, 2019) that takes image I, new question Q' and original answer choices A to pick the most likely one.

### 5.2 Logical Entailment based Reasoner

[4] For all other answer types, we leverage Logical Entailment (upper half pipeline in 5) of image and text to answer questions. We create an 'Entailment Toolbox' which consists of image-image, image-text (Xie et al., 2019), text-image and textual entailment (Khot et al., 2018) sub-modules and use them as required. For image-image and image-text entailment, we augment Visual COPA (Yeo et al., 2018) dataset and train custom network for both. (refer Supplementary Material B for more details)

**4-way or 2-way image MCQ** contains images as an answer choice, which is similar to an Image Selection task (Hu et al., 2019). The goal here is to identify an image that best matches the description of P or mathematically, determine $P \vdash A_k$ (i.e., text-image entailment) with maximum score. $A_k$ represents answer choices where k=4 and k=2 for 4-way and 2-way image problems respectively.

**Binary Classification** can be considered as a fact-checking task where we want to determine the truth value of a question provided image-passage, or mathematically, $P \cup I \vdash Q$. We use textual entailment to determine $P \vdash Q$ and image-text entailment to determine $I \vdash Q$. If both entailment modules' confidence score is above 0.65 then it is determined as True, otherwise False.

[4] $\vdash$ is the symbolic representation of entailment

**4-way Sequencing** task assesses a model's capability to order 4 spatial or temporal events. If we consider I-II-III-IV as a sequence of events, it is equivalent to 3 entailment tasks: I-II, II-III, and III-IV, where each I to IV can be an image or a text. Among the answer choices, the sequence with maximum overall confidence is selected as an answer.

## 6 Results & Discussion

Multi-modality brings both pros and cons while developing new Artificial Intelligence (AI) benchmarks. The presence of multiple modalities provide natural flexibility for varied inference tasks, simultaneously making the reasoning process more complex as information is now spanned across them and requires cross-inferencing. In this work, we focused on joint reasoning over image-text multimodal context and developed a Visuo-Linguistic Question Answering (VLQA) Dataset. Our proposed VLQA dataset has important distinctions from existing VQA datasets. Firstly, it incorporates a text passage that contains additional contextual information. Secondly, it offers various figure types including natural images, templated images and free-form images (unstructured), which is not so common for other VQA datasets. Thirdly, it tests diverse reasoning capabilities, including cross-inferencing between visual and textual modalities.

We then use several baselines and benchmark their performance over the resulting VLQA dataset. As VLQA has multiple choice questions with exactly one correct answer, we use standard accuracy as an evaluation metric. From the results in 3, we can observe that pre-trained vision-language models fail to solve a significant portion of the VLQA items. Our proposed modular method HOLE slightly outperforms them and is more interpretable for analysis. We also report the performance of Question-only, Image-only and Passage-only baselines which we used for quality check. The poor performance of these baselines indicate that the VLQA dataset requires models to jointly understand both image and text modalities and is relatively harder than other vision-language tasks.

For human evaluation of the VLQA test-set, the reported accuracy is 84.0%. For 148 wrongly predicted answers, we group them according to 4 reasons for failures, which are listed in 4. The results demonstrate a room for significant improvement in existing vision-language models that are far behind the human performance. This stimulates the need for more complex reasoning capabilities of AI models. We suspect that VLQA questions that purely rely on facts might be exploited by the latest language models, despite strong measures taken through manual and automated quality control during the creation of the dataset. We would like to explore this further in the future.

| Method | Test(%) | Val(%) |
|---|---|---|
| Human | 84.00 | – |
| Random | 31.36 | 31.36 |
| Question-only: RoBERTa$_{ARC}$ | 28.56 | 29.42 |
| Passage-only: ALBERT$_{RACE}$ | 30.16 | 30.25 |
| Image-only: LXMERT$_{VQA}$ | 29.48 | 30.56 |
| Vision-Language | | |
| VL-BERT | 35.92 | 34.60 |
| VisualBERT | 33.17 | 34.17 |
| ViLBERT | 34.70 | 35.25 |
| LXMERT | **36.41** | 37.82 |
| HOLE (Proposed Model) | **39.63** | 40.08 |

Table 3: Performance benchmarks over test-set of VLQA task and corresponding validation results

| Underlying reason for incorrect answer provided by test-taker | #incorrect/148 (%incorrect) |
|---|---|
| Lacked necessary knowledge | 27 (18.2%) |
| Misunderstood the provided info | 47 (31.7%) |
| Mistake in deduction/calculation | 63 (42.5%) |
| Felt that data item is ambiguous | 11 (7.4%) |

Table 4: Classification of incorrectly predicted answers in Human-evaluation of VLQA test-data

## 7 Conclusion

In this work, we introduced the Visuo-Linguistic Question Answering (VLQA) challenge that we believe has the potential to open new research avenues in areas of joint vision & language. Our experiments show that a system equipped with state-of-the-art vision-language pre-training does not perform well on the task that requires joint image-text inference. There is a room for significant improvement in capability of these models to tackle multi-modal contexts. Our future work would include further expansion of this dataset and building generic AI models that can learn novel visual concepts from a small set of examples.

## Acknowledgments

We are thankful to the anonymous reviewers for the feedback. This work is partially supported by the National Science Foundation grant IIS-1816039.

## References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *AAAI*, volume 33.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *IEEE ICCV*, pages 2425–2433.

Chandra Bhagavatula, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Scott Wen tau Yih, and Yejin Choi. 2019. Abductive commonsense reasoning.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. Piqa: Reasoning about physical commonsense in natural language. *arXiv preprint arXiv:1911.11641*.

DC Central Intelligence Agency, Washington. 2019. The world factbook.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Hexiang Hu, Ishan Misra, and Laurens van der Maaten. 2019. Evaluating text-to-image matching using binary image selection (bison). In *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE CVPR*.

Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. In *IEEE CVPR*, pages 5648–5656.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: A figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *ECCV*.

Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. 2017. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *IEEE CVPR*, pages 4999–5007.

Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. 2019. Qasc: A dataset for question answering via sentence composition. *arXiv preprint arXiv:1910.11473*.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Mahnaz Koupaee and William Yang Wang. 2018. Wikihow: A large scale text summarization dataset. *arXiv preprint arXiv:1810.09305*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *IEEE CVPR*.

OECD. 2019. Pisa: Programme for international student assessment. *Recuperado el*.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. *arXiv preprint arXiv:1508.00305*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *arXiv preprint arXiv:1611.01603*.

Sanket Shah, Anand Mishra, Naganand Yadati, and Partha Pratim Talukdar. 2019. Kvqa: Knowledge-aware visual question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8876–8884.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2019. Vl-bert: Pre-training of generic visual-linguistic representations.

Alane Suhr, Mike Lewis, James Yeh, and Yoav Artzi. 2017. A corpus of natural language for visual reasoning. In *55th ACL (Vol 2: Short Papers)*.

Alon Talmor and Jonathan Berant. 2018. The web as a knowledge-base for answering complex questions. *arXiv preprint arXiv:1803.06643*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *EMNLP*.

Alexander Trott, Caiming Xiong, and Richard Socher. 2017. Interpretable counting for visual question answering. *arXiv preprint arXiv:1712.08697*.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2018. Fvqa: Fact-based visual question answering. *IEEE PAMI*.

Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*.

Johannes Welbl, Pontus Stenetorp, and Sebastian Riedel. 2018. Constructing datasets for multi-hop reading comprehension across documents. *Transactions of the Association for Computational Linguistics*, 6:287–302.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Semih Yagcioglu, Aykut Erdem, Erkut Erdem, and Nazli Ikizler-Cinbis. 2018. Recipeqa: A challenge dataset for multimodal comprehension of cooking recipes. *arXiv preprint arXiv:1809.00812*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. *arXiv preprint arXiv:1809.09600*.

Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola. 2016. Stacked attention networks for image question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 21–29.

Jinyoung Yeo, Gyeongbok Lee, Gengyu Wang, Seungtaek Choi, Hyunsouk Cho, Reinald Kim Amplayo, and Seung-won Hwang. 2018. Visual choice of plausible alternatives: An evaluation of image-based commonsense causal reasoning. In *11th LREC*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *IEEE CVPR*.