Modern Question Answering Datasets and Benchmarks: A Survey

Zhen Wang
Delft University of Technology
z.wang-42@student.tudelft.nl

Abstract

Question Answering (QA) is one of the most important natural language processing (NLP) tasks. It aims using NLP technologies to generate a corresponding answer to a given question based on the massive unstructured corpus. With the development of deep learning, more and more challenging QA datasets are being proposed, and lots of new methods for solving them are also emerging. In this paper, we investigate influential QA datasets that have been released in the era of deep learning. Specifically, we begin with introducing two of the most common QA tasks - textual question answer and visual question answering - separately, covering the most representative datasets, and then give some current challenges of QA research.

1 Introduction

Question answering (QA) (Hirschman and Gaizauskas, 2001) aims at providing correct answers to questions base on some given context or knowledge. QA is a traditional research direction that has been proposed half a century ago. People hope to help with everyday life by teaching the program how to answer questions like a real person. Traditional QA systems integrate some information retrieval techniques to find answers. With the development of deep learning, computer programs can now tackle more complex problems. At the same time, in the era of deep learning, more and more datasets are being proposed to measure the capabilities of QA models. These datasets, in turn, facilitate the development of deep learning QA models.

An comprehensive understanding of those datasets and benchmarks is essential before the further research about QA. Therefore, in this paper, we investigate some of the most commonly used datasets nowadays and categorize them according to the capabilities of the QA models involved. Meanwhile, according to the different modes designed, we divide them into three categories: textual QA, image QA, and video QA. In textual QA, all the corpora involved are presented in textual form. A typical sample in textual QA consists a question, an answer, and a paragraph that contains the answer. Image QA and video QA together are generally referred to as Visual Question Answering (VQA) (Antol et al., 2015). In image QA, the question and answer are usually in textual form while the context is an image. And in video QA, the question and answer are the same but the context is a clip of video. We will start by introducing some representative datasets for each of the three types of QA, and conclude by analyzing some of the current challenges and opportunities in QA research from a dataset perspective. We hope this paper will help researchers gain a comprehensive understanding of QA tasks and methods, attracting more attention, leading to greater progress in QA filed.

2 Textual Question Answering

2.1 Reading Comprehension

BoolQ (Clark et al., 2019) is a yes/no reading comprehension QA dataset. Give the passage "The Great Storm of 1987 was a violent extratropical cyclone which caused casualties in England, France and the Channel Islands . . ." and corresponding question "Has the UK been hit by a hurricane?", the correct answer is "Yes". A method is asked to find the correct answer though analyse complex and non-factoid textual information, thus strong inference ability is required.

CNN and Daily Mail (See et al., 2017) using human generated abstractive as questions while one entities is hided, the stories are used as context to do the fill-in-the-blank questions.

CoQA (Reddy et al., 2019) is the first conversational question answering dataset. Given a passage as the conversation context, one person will ask

Dataset	Answer Type	Size	Domain	Evaluate Ability
ARC(Clark et al., 2018)	Multi-Choice	7,787	Science	Reasoning
BoolQ (Clark et al., 2019)	Bool	16K	Wikipedia	Reasoning
BioASQ (Tsatsaronis et al., 2015)	Span	282	Biomedical	Articles Indexing
CaseHOLD (Zheng et al., 2021)	Multi-Choice	53,137	Law	Pre-training
bABi (Weston et al., 2015)	Bool/Entity	40K	Open Domain	Reasoning
CBT (Hill et al., 2015)	Entity	20K	Children's Book	Model Memory
CliCR (Šuster and Daelemans, 2018)	Entity	105K	Medical	Domain Knowledge
CNN and Daily Mail (See et al., 2017)	Entity	311K	News	Text Summarization
CODAH (Chen et al., 2019)	Multi-choice	4,149	Open Domain	Commonsense
CommonsenseQA (Talmor et al., 2018)	Multi-choice	12,247	ConceptNet	Commonsense
ComplexWebQuestions (Talmor and Berant, 2018)	Entity	34,689	Freebase	Multi-hop
ConditionalQA (Sun et al., 2021)	Entity/Span	9983	Public Policy	Multi-hop
COPA (Gordon et al., 2012)	Multi-choice	1000	Commonsense	Reasoning
CoQA (Reddy et al., 2019)	Entity	127K	Open Domain	Conversation
DROP (Dua et al., 2019)	Span	96K	Wikipedia	Multi-hop
FinQA (Chen et al., 2021)	Number/Span	8,281	Finance	Multi-hop
HotpotQA (Yang et al., 2018)	Entity	113K	Wikipedia	Multi-hop
JD Production QA (Gao et al., 2019b)	Generation	469,953	E-commerce	Domain Knowledge
LogiQA (Liu et al., 2020)	Multi-choice	8,678	Exam	Reasoning
MCTest (Richardson et al., 2013)	Multi-choice	2,000	Fictional Story	Reading Comprehension
Mathematics Dataset (Saxton et al., 2019)	Numeric	2.1×10^6	Mathematics	Calculate
MS MARCO (Nguyen et al., 2016)	Generation	1,010,916	Web pages	Search
MultiRC (Khashabi et al., 2018)	Multi-choice	6K	Multiple Domain	Multi-hop
NarrativeQA (Kočiskỳ et al., 2018)	Span	46,765	Story	Full Document
Natural Questions (Kwiatkowski et al., 2019)	Span/Passage	323,045	Wikipedia	Search
NewsQA (Trischler et al., 2016)	Span	100,000	CNN news	Reading Comprehension
OpenBookQA (Mihaylov et al., 2018)	Multi-choice	6000	Science Facts	Reasoning
PIQA (Bisk et al., 2020)	Multi-choice	21,000	Physical	Physical
PubMedQA (Jin et al., 2019)	Multi-choice	1K	Medical	Summarization
QASPER (Dasigi et al., 2021)	Extractive	5,049	NLP papers	Reasoning
QuAC (Choi et al., 2018)	Multi-choice Generation	100K	Wikipedia	Dialog
QUASAR (Dhingra et al., 2017)	Span	43,000	StackOverflow/Trivia	search
RACE (Lai et al., 2017)	Multi-choice	100,000	Exam	Reading Comprehension
ReClor (Yu et al., 2020)	Multi-choice	6138	Exam	Logical
SCDE (Kong et al., 2020)	Exam	6K	Exam	Reading Comprehension
SimpleQuestions (Bordes et al., 2015)	Entity	100K	Freebase	Knowledge
SQuAD (Rajpurkar et al., 2016, 2018)	Span	130,319	Wikipedia	Reading Comprehension
TriviaQA (Joshi et al., 2017)	Span	650K	Open Domain	Reading Comprehension
TweetQA (Xiong et al., 2019)	Generation	13,757	Tweet	Reading Comprehension
WikiHop (Welbl et al., 2018)	Multi-choice	51,318	Wikipedia	Multi-hop
WikiQA (Yang et al., 2015)	Sentence	3,047	Wikipedia	Reading Comprehension

Table 1: Statistics of textual QA datasets.

some questions and one person will answer those questions by find the evidences from context. Different from previous QA datasets, in CoQA, each question is related to previous asked questions, thus the answer should not only the context, but also the previous answered QA-pair should be taken into consideration. For example, the first question is " Who had a birthday?", and the second question is "How old would she be?", which one the "she" refer to can only be known by the first question.

MCTest (Richardson et al., 2013) is a multi-choice reading comprehension QA dataset. All the passages in MCTest are children fictional stories written by Amazon Mechanical Turk (AMT). For each story, four multi-choice questions are asked. Specifically, some questions are required must consider multiple sentences to get correct answers.

The candidate passages in MS MARCO (Nguyen et al., 2016) are all from the web pages retrieved by Bing, the questions are the query searched by Bing users, and the answers are generate by human annotators. There are three types of tasks in MS MARCO. First is determine whether a question is answerable, second is generate a proper answer to a give question, and the last is ranking the retrieved passages according to the relevance.

Given previous QA datasets can be answered with some QA irrelevant features such as term frequency, making it impossible to measure the true capabilities of QA models, Deepmind thus propose a new reading comprehension QA dataset called NarrativeQA (Kočiskỳ et al., 2018), where the questions can only be answered after fully review the whole document.

Similar to MS MARCO (Nguyen et al., 2016), Natural Questions (NQ) (Kwiatkowski et al., 2019) is also a reading comprehension dataset collect through the search engine. And the major difference is that NQ not only provide the relevant paragraph as the long answer, but also a human annotated short answer. And the Wikipedia page is used as the context.

NewsQA (Trischler et al., 2016) is aiming at providing a SQuAD-like QA dataset but harder than SQuAD (Rajpurkar et al., 2016). The annotators are asked to generate questions based on CNN news, and the answers are span of text which maybe the person, location, clause, numeric, etc.

RACE (Lai et al., 2017) is a multi-choice reading comprehension dataset which collected from English exams for Chinese students. Based on a passage, models should find correct answer from four candidate answers to a given question.

Similar to RACE (Lai et al., 2017), ReClor (Yu et al., 2020) is also a Multi-choice reading comprehension dataset. While some researches argue that current human annotated QA datasets contain some bias which may be used by models to cheat to achieve high accuracy, Reclor thus is divided into two datasets based on bias. The EASY set contains bias while the HARD set not. And the experiments prove that models can achieve high performance in EASY set but poor in HARD set.

SCDE (Kong et al., 2020) is used to test models' reading comprehension ability through fill up blanks in passage using given sentences. Given seven sentences where two sentences among them are distractors, models should select the correct five sentences and fill them in the corresponding blanks to complete the whole article.

SimpleQuestions (Bordes et al., 2015) is constructed using Freebase. For each (subject, relationship, object), two of them is used to generate the question and the remained one is used as the answer. The models need to retrieve the correct answers from massive of possible alternatives from freebase, which makes the questions are not easy to be answered.

SQuAD (Rajpurkar et al., 2016, 2018) is the most famous reading comprehension dataset. Given a question and a paragraph as the context, the models need to extract the plausible answer from it. Each answer is a span of text. To make the task harder, in SQuAD 2.0, some unanswerable questions are include into the dataset.

TriviaQA (Joshi et al., 2017) is a reading comprehension dataset focusing on complex and compositional questions where requiring the ability on reasoning over multiple sentences. Specially, there are syntactic and lexical variability among questions, answers and evidences.

The question-answer pairs in TweetQA (Xiong et al., 2019) are all annotated from tweets. Unlike previous reading comprehension datasets using a text span as the answer, the answer in TweetQA can be abstractive, which makes it more difficult to be solve. The experiments show that current SOTA models such as BERT (Devlin et al., 2018) are significantly below human performance.

2.2 Reasoning

AI2 Reasoning Challenge (ARC) (Clark et al., 2018) focuses on hard multi-choice questions to challenge current QA methods. Their questions cannot be answered easily using retrieval based methods or through word correlation. And particularly, they choose some wrong answer question construct a specific Challenge Set contains 2590 questions.

CODAH (Chen et al., 2019) is a common sense reasoning dataset. Given a premise, for example, "A man on his first date wanted to break the ice. He ...", methods are asked to choose a correct reason from four answers. In this sceanario, the answer is "made a corny joke."

Combine totally 20 kinds of tasks, bABi (Weston et al., 2015) aiming at give the answer that can a QA model is able to solve QA problem using some reasoning abilities, such as the chain of facts, induction, deduction operation. When give three fact: (1) Mary went to the bathroom. (2) John moved to the hallway. (3) Mary travelled to the office; and the question "Where is Mary?", the tested QA models should give the correct answer "office".

Build through ConceptNet (Liu and Singh, 2004), CommonsenseQA (Talmor et al., 2018) aiming at generating difficult common sense questions to test models' reasoning ability. The annotators are asked to make the question as difficult as possible by looking for incorrect answers closely related to correct answers to construct the multi-choice.

COPA (Gordon et al., 2012) is a task in SemEval-2012 which is a multi-choice QA dataset aiming at testing the causal reasoning ability of models. Given a premise, the models need to find the best matched cause or result.

LogiQA (Liu et al., 2020) is a logical reasoning multi-choice QA dataset. The questions come from National Civil Servants Examination of China. There are mainly five kinds of reasoning types in LogiQA - categorical reasoning, sufficient conditional reasoning, necessary conditional reasoning, disjunctive reasoning and conjunctive reasoning. The question in LogiQA is extremely hard that current state-of-the-art QA model can only achieve 39% in accuracy.

Different from existing linguistic QA dataset, OpenBookQA (Mihaylov et al., 2018) more focuses on scientific reasoning. The models should find the correct choice by reasoning between questions and the given science facts and common knowledge.

PIQA (Bisk et al., 2020) is a special QA dataset in which the physical reasoning ability is required to solve the questions. The format of the task is select best one from two given answers. For a given scenario in question, the correct physical process needs to be chosen to achieve the expected result in question.

Given the queries from Bing as the questions and users' click, WikiQA (Yang et al., 2015) selects those relevant Wikipedia pages as the context. The annotators are asked to select a correct sentence which can answer a corresponding question. Further, WikiQA also involve some questions that cannot be answered using the context.

2.3 Domain-Specific

CliCR (Šuster and Daelemans, 2018) is a dataset consists of gap-filling questions about medical cases, and the authors find that domain-specific knowledge is a key for success in medical QA.

BioASQ (Tsatsaronis et al., 2015) is a extraction-based QA dataset constructed using biomedical corpus. BioASQ mainly focus on biomedical-style questions such as "What are the physiological manifestations of disorder Y?"

The Children's Book Test (CBT) (Hill et al., 2015) is a dataset to measure the ability that language models understand children's books. This task specifically focuses on syntactic function words prediction.

To determine when domain-specific pre-training is worthwhile, consider current legal related NLP tasks are too easy to challenge transformer-based methods, in CaseHOLD (Zheng et al., 2021), the authors proposed a multi-choice legal QA dataset. They use case as the context, and among some given holding statement, only one is proper. Their experiments shows that compare to normal BERT (Devlin et al., 2018), the BERT pre-trained on legal corpus can achieve better result, indicating that pre-training on domain-specific text is useful when doing difficult domain-specific tasks.

JD Production QA (Gao et al., 2019b) is a large-scale QA dataset focuses on solving the problem that how to generate a plausible answer to a question asked for a particular product in the e-commerce platform.

Mathematics Dataset (Saxton et al., 2019) is the only one QA dataset that only pay attention to mathematical reasoning, which proposed by DeepMind. With this dataset, they evaluate the sequence-to-sequence models' ability in solve mathematical problems.

PubMedQA (Jin et al., 2019) is a biomedical QA dataset which using the abstract from PubMed. For each QA-pair, the question is title or a sentence extract from a article, the article's abstract which exclude the conclusion is used as the context, and the conclusion is used as the long answer. Models should give yes/no/maybe to judge whether the conclusion can answer the question.

In QASPER (Dasigi et al., 2021), 1,585 NLP papers is used as the data source. One part of annotators are asked to give the questions based on papers' title and abstract. And the other part of annotators are asked to answer those questions base on the full text papers while also giving the corresponding support evidences.

QuAC (Choi et al., 2018) is a dialog-specific QA dataset. Based on the context provided by Wikipedia, a dialog between a teacher and a student is constructed. Different from normal QA, the QA in dialog have some differences. The question can usually be open-ended, sometimes there are no answer to a question, and the answers can be meaningful only in current dialog.

The goal of QUASAR (Dhingra et al., 2017) is to test QA models' ability to first retrieve and then generate answers from retrieved documents.

It consists of two sub-datasets. The first one is QUASAR-S which constructed from StackOverflow. User input queries are used to generate fill-in-the-gap questions, and users' posts and comments are used as context. The second one is QUASAR-T which consists of trivia questions and the answers come from internet.

2.4 Multi-hop

ComplexWebQuestions (Talmor and Berant, 2018) uses the simple questions from WebQuestionSP (Yih et al., 2016) build more complex questions which need to be answered by decompose the question into two simple questions. For example, the original question is "What movies have robert pattinson starred in?", and the complex question enhanced by SPARQL is "What movies have robert pattinson starred in and that was produced by Erwin Stoff?".

In ConditionalQA (Sun et al., 2021), a correct answer is applicable only in a certain condition. Thus, the correctly answer questions, the models need to not only give the answer, but also generate the condition. The authors find that for the existing QA models, compare to generate correct answer, how to find the suitable condition is particularly challenging.

HotpotQA (Yang et al., 2018) is the first multi-hop QA dataset and the questions can only be answered by doing inference in multiple documents. Particularly, all the questions in HotpotQA are not depended on any existing knowledge and only the reasoning ability is required to generate the answer. To make the reasoning procedure more reliable, the useful sentence-level facts are also provided to help researches check whether models generate answers using the correct sources.

While previous QA dataset only using single context, the questions in DROP (Dua et al., 2019), however, can only be answered by get multiple fact from different paragraphs and then doing some combine operation to drive the final answer.

Compare to previous multi-hop QA only focuses on textual data, FinQA (Chen et al., 2021), which extracted from financial reports, includes more kinds of information, such as table and number, and therefore some specific numerical operation is needed.

MultiRC (Khashabi et al., 2018) is reading comprehension dataset where the multiple sentences should be taken into consideration to select the correct choice. It covers seven domains: science, news, travel guides, fiction stories, Wikipedia, history, 9/11 reports.

To solve the questions in WikiHop (Welbl et al., 2018), models need to infer multiple documents from Wikipedia to obtain the correct answers. To answer a question, models should first retrieve some support documents, and then find and combine some evident to infer the answer.

3 Image Question Answering

3.1 Recognition

DocVQA (Mathew et al., 2021) collects massive images from documents as the question context and ask questions about the textual information in the images. the models need to correctly recognize the text to generate the answer, which makes it huge different from normal object detection tasks.

HowMany-QA (Trott et al., 2017) mainly focuses on testing the counting ability of models. For each image, some counting questions are asked, such as "How many people are wearing blue shorts?" The images the authors used are from VQA2.0 (Antol et al., 2015) and Visual Genome (Krishna et al., 2017).

Compare to HowManyQA (Trott et al., 2017), the counting questions in TallyQA (Acharya et al., 2019) are more complex and require reasoning ability about the relationship between objects and attribution. Given a simple question "How many girraffes are there?", the complex can be "How many girraffes are sitting down?".

In TDIUC (Kafle and Kanan, 2017), there are totally 12 tasks including Counting, Scene Classification, Sentiment Understanding, etc. The main goal of TDIUC is to provide a solid benchmark to test all the existing VQA algorithms.

While previous VQA datasets merely focus on object detection, TextVQA (Singh et al., 2019) however, aims at the text recognition in images. Give an image, the models need to give a correct text as the answer that contained in the image to correctly answer the corresponding question.

Previous datasets only using the global information between images and QA pairs, which may ignoring some important local region information. Thus in Visual7W (Zhu et al., 2016), the authors first using bounding-box to extract the important objects or regions in a image, and then ask some questions about those special area.

The images in VizWiz (Gurari et al., 2018) are

Dataset	Answer Type	Size	Domain	Evaluate Ability
Image Question Answering				
CLEVR (Johnson et al., 2017)	Open-ended	853K	3D CG	Reasoning
RecipeQA (Yagcioglu et al., 2018)	Multi-choice	36K	Cooking Recipes	Procedural
CRIC (Gao et al., 2019a)	Open-ended	494K	Visual Genome	Scene Reasoning
DocVQA (Mathew et al., 2021)	Open-ended	50,000	Document	Recognition
FVQA (Wang et al., 2017)	Open-ended	5,826	Open Domain	Knowledge
Visual Genome (Krishna et al., 2017)	Open-ended	1,445,322	Open Domain	Recognition
VCR (Zellers et al., 2019)	Multi-choice	290K	Movie	Reasoning
GQA (Hudson and Manning, 2019)	Open-ended/Yes/No	22M	Visual Genome	Reasoning
HowMany-QA (Trott et al., 2017)	Number	106,356	VG/VQA2.0	Counting
TallyQA (Acharya et al., 2019)	Number	287,907	VG/COCO	Counting
TDIUC (Kafle and Kanan, 2017)	Open-ended	1.6M	VG/COCO	Multiple
TextVQA (Singh et al., 2019)	Open-ended	45,336	Open Domain	Text Recognition
VCOPA (Yeo et al., 2018)	Multi-choice	380	Open Domain	Causality
Visual7W (Zhu et al., 2016)	Open-ended/Multi-choice	327,939	VG	Reasoning
VizWiz (Gurari et al., 2018)	Open-ended	31,000	Photo	Recognition
VQA2.0 (Goyal et al., 2017)	Open-ended	1.11M	COCO	Recognition
KVQA (Shah et al., 2019)	Open-ended	183,007	Wikipedia	Knowledge
OK-VQA (Marino et al., 2019)	Open-ended	14,000	Open Domain	Knowledge
R-VQA (Lu et al., 2018)	Open-ended	478,287	VG	Reasoning
KB-VQA (Wang et al., 2015)	Open-ended	2,402	COCO/ImageNet	Knowledge
WebQA (Chang et al., 2021)	Open-ended	25K	Wikipedia	Multi-hop
AQUA (Garcia et al., 2020b)	Open-ended	79,848	Art	Knowledge
Video Question Answering				
CLEVRER (Yi et al., 2019)	Multi-choice	300,000	CG Video	Causal Reasoning
KnowIT VQA (Garcia et al., 2020a)	Open-ended	24,282	TV series	Knowledge
MovieQA (Tapaswi et al., 2016)	Open-ended	14,944	Movie	Reasoning
TVQA (Lei et al., 2018)	Multi-choice	152,545	Movie	Reasoning
PororoQA (Kim et al., 2017)	Multi-choice	8,913	Cartoon	Summarization
Social-IQ (Zadeh et al., 2019)	Multi-choice	7,500	Social	Reasoning

Table 2: Statistics of image and video QA datasets. *RC* means *Reading Comprehension*, *MC* means *Multi-choice*.

all took by blind people with a recorded spoken question. And the answers are annotated by crowdsources. This dataset is mainly used to help the research in using VQA technologies to assist blind people.

VQA2.0 (Goyal et al., 2017) is an improved version of VQA (Antol et al., 2015). The questions in those two dataset are the same, asking questions about the object in the image. But the data distribution in VQA2.0 is more balanced. For each question, there are two similar images with different answers.

3.2 Reasoning

CLEVR (Johnson et al., 2017) aims at testing the reasoning abilities of models. Given the strong bias existing in some reasoning VQA datasets, they use computer generated 3D shapes as the image context and automatically generate some question related to attribute, counting, comparison, logical operations, etc.

Unlike other QA datasets where context is a complete paragraph, context is a procedure in RecipeQA (Yagcioglu et al., 2018). The model needs to understand the sequence of the different steps involved in making a dish in order to answer questions correctly.

While previous VQA datasets only using single object in an image, CRIC (Gao et al., 2019a), however, using the Scene Graph from Visual Genome (Krishna et al., 2017) to generate questions about the relationship between different objects in a same image. For example, the question can be "What utensil can be used for moving the food that is in the bowl?" and models are asked to find the correct object as the answer that is spoon. CRIC challenges the relational reasoning ability of visual models.

Visual Genome (Krishna et al., 2017) is the largest VQA dataset. The questions in Visual Genome always starts with six Ws - what, where, when, who, why, and how, which allow them to test wide range of model abilities such as detection, categorization, commonsense reasoning. Using the annotated scene graphs, human workers are asked to give questions based on region descriptions.

VCR (Zellers et al., 2019) combines object recognition and relation reasoning through asking questions between different objects. For example, given a image contains some people which are already highlighted and annotated by bounding boxes, the question can be "Why the [person1] talk to [person2]?" To correctly answer this question, models should first select the most rational answer from four answers and then should find the most reasonable explanation also from four candidates to explain why the first choice is correct.

GQA (Hudson and Manning, 2019) leverage the scene graph provided by Visual Genome (Krishna et al., 2017) to generate questions about the objects in image. The question involves object recognition, classification of relations between objects and so on, such as "Is the bowl to the right of the green apple?" They use different templates to automatically generate questions and thus the total number of questions achieves 22M.

Inspired by COPA (Gordon et al., 2012), VCOPA (Yeo et al., 2018) researches the causality in VQA scenario. Given an image as the premise, the VQA models need to choose the most possible result that may caused by this premise from two images.

R-VQA (Lu et al., 2018) mainly aims at the relationship reasoning between different entities. They choose three types of relations: (there, is, object), (subject, is, attribute), (subject, relation, object). For each QA pair, they also provide the corresponding relation to help answer the question.

3.3 Commonsense and Knowledge

Different from previous datasets that questions can be answered merely rely on the analysis of image itself, the questions in FVQA (Wang et al., 2017) can only be solved considering some external commonsense and knowledge. This makes the dataset can be easily answered by human but much difficult to models. For instance, given a image contains a red fire hydrant, the question can be "What can the red object on the ground be used for ?" and the correct answer is "Firefighting".

KVQA (Shah et al., 2019) focuses on the world knowledge about the famous people. They collect the images in Wikipedia and asking attribute questions, such as "who is the person in the images?" or "How old is the person?"

For each sample in OK-VQA (Marino et al., 2019), in addition to a question and an answer, they also provide some outside knowledge, in which the correct answer contained. Those knowledge a closely related to the QA pair, and model should analysis both the question and outside knowledge to extract the final answer.

Instead of merely asking questions about visual, questions in KB-VQA (Wang et al., 2015) may also related to commonsense and knowledge. For example, if the visual question is "How many giraffes are there in the image", then the commonsense question can be "Is this image related to zoology?", and the corresponding knowledge question is "What are the common properties between the animal in this image and zebra?".

AQUA (Garcia et al., 2020b) is a VQA dataset specifically about Art. The questions are generate in two different ways. The first is directly generated based on the painting itself, which can be answered only using the painting. And the other way is to generate from the painting's comments which may involve some external knowledge and cannot be answer merely rely on the painting.

3.4 Multi-hop

WebQA (Chang et al., 2021) is currently the only one multi-hop VQA dataset. To ask the questions, the models need to reasoning on some images and text snippets. Some images and text are related to the question while some not. Models should first find the related images and text then combine them together to generate the correct answer.

4 Video Question Answering

CLEVRER (Yi et al., 2019) focuses on the temporal causal reasoning abilities of models. The authors use CG technology to simulate the movement and collision between different simple geometric objects to design questions and answers. Asking questions such as "Which of the following is responsible for the gray cylinder's colliding with the cube?" and the models need to figure out the correct reason for this.

KnowIT VQA (Garcia et al., 2020a) use the clips from *The Big Bang Theory* to generate questions. There totally four kinds of questions: visual, textual, temporal and knowledge. Models should use both the video and the subtitles, as well as some external knowledge about the TV series, to answer the questions.

There are there types of source information in MovieQA (Tapaswi et al., 2016): plot, video and subtitle. Combine those different information, models should answer the annotated questions such as "Who kills Neo in the Matrix?".

In TVQA (Lei et al., 2018), some questions can be answer by using video or subtitle independently while some questions can only be answered by combining both of them. Furthermore, the models should also learn to precisely locate where the useful clips are.

PororoQA (Kim et al., 2017) is a VQA dataset specially focusing on cartoon for children and thus the word in it are quite simple. There are totally ten types of questions: Action, Person, Abstract, Detail, Method, Reason, Location, Statement, Causality, Yes/No, Time.

Social-IQ (Zadeh et al., 2019) aims to help the AI researches in social situation. Given some video scenes where people socialize, the annotators are asked to ask some question related to those people. For example, "How is the discussion between the woman and the man in the white shirt?"

5 Challenges and Chances

A comparison of those three different forms of QA shows that, given how early they were introduced, textual QA is the most studied among the three QA forms, which has the most number of datasets, and covers the most types of QA abilities. For VQA, because it is just proposed in the new era of deep learning, there is still a lack of research and datasets, especially for video QA. This is also closely related to the difficulty of corpus collection, annotation difficulty and training resource demand.

5.1 Textual QA

Because textual information are most available in real-world or online websites, Textual QA is both the earliest and the most studied direction. There are two main sources of textual QA today. One is a variety of actual exam questions, and one is human design questions from the annotator. Questions from exams are more difficult and of higher quality, but are also limited in quantity to meet the needs of each particular type of QA. On the contrary, Questions designed by human annotators are more flexible and thus can meet different demands, but the annotate process are expensive and usually contains some errors which cannot measure up to the precision of the exam questions.

Although various datasets are now available, there are still many unsolved problems. The first is the evaluation matrix of QA model performance. QA is currently measured primarily by the overlap degree to which a predicted answer meets a target answer, and the higher the overlap, the more accurate it is considered. But language is complex, and there can be more than one correct answer to the same question, and word coincidence doesn't take actual semantics into consideration, which may lead to some misjudgments. The second is the interpretability of QA models. The current QA models just give the answer directly, we cannot know how the answer is extracted or generated, the whole process is unexplainable. The third is the gap between QA datasets and the real-world scenario. Out of domain problem is a critical disadvantage for all deep learning-related research, and QA is no exception. Although some models currently achieve more than 90% accuracy on some particular datasets, there is always a huge performance degradation when they are applied to other fields or practical applications. How to narrow this gap is also a problem worth studying. And the last is the limitation of question types. In the current QA datasets, questions mainly start with *what*, *when*, *where*, etc., while there are few open-end questions like *why* and *how*. The answer to the former is always short and easily solved by the model, while the answer to the latter may be much longer than the question itself. This is a big challenge for capacity-limited deep learning models. We can't use infinite huge models because of the cost of machines, but knowledge itself is infinite.

5.2 Visual QA

Visual QA, including image QA and video QA, is in general an extension of textual QA. The major difference is in VQA, the context is image or video rather than text. In addition to having some of the same issues as textual QA, VQA has some unique problems. Firstly, images and videos are more difficult to access and annotate than text, making the number of VQA datasets much smaller than those of textual QA, and usually with lower quality. Secondly, in QA tasks, there is a higher requirement for understanding images. Traditional computer vision tasks are mainly used for object classification or object recognition. But to answer the question, the model needs to recognize not only the object, but also the relationship between the objects and even the meaning of the whole image. And lastly, compared with textual QA, there are still many problems to be studied in VQA due to its late start. Luckily, many studies in textual QA can be transferred to VQA and provide some references for VQA.

6 Conclusion

In this work, we investigate and present a comprehensive survey on QA. Specifically we focuses on the QA datasets of three different QA tasks: textual QA, image QA and video QA. For each dataset, we introduce its data statistics and usage. Moreover, through self-comparison and cross-comparison of these three tasks, we analyze some of the current challenges and opportunities QA research facing, to provide some reference and insight for the future QA researches.

References

Manoj Acharya, Kushal Kafle, and Christopher Kanan. 2019. Tallyqa: Answering complex counting questions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8076–8084.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. 2020. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7432–7439.

Antoine Bordes, Nicolas Usunier, Sumit Chopra, and Jason Weston. 2015. Large-scale simple question answering with memory networks. *arXiv preprint arXiv:1506.02075*.

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2021. Webqa: Multihop and multimodal qa. *arXiv preprint arXiv:2109.00590*.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. 2021. Finqa: A dataset of numerical reasoning over financial data. *arXiv preprint arXiv:2109.00122*.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. *arXiv preprint arXiv:1808.07036*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

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*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Difei Gao, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2019a. From two graphs to n questions: A vqa dataset for compositional reasoning on vision and commonsense. *arXiv preprint arXiv:1908.02962*.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019b. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437.

Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. 2020a. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10826–10834.

Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. 2020b. A dataset and baselines for visual question answering on art. In *European Conference on Computer Vision*, pages 92–108. Springer.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. Semeval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In ** SEM 2012: The First Joint Conference on Lexical and Computational*

Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012), pages 394–398.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. 2015. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*.

Lynette Hirschman and Robert Gaizauskas. 2001. Natural language question answering: the view from here. *natural language engineering*, 7(4):275–300.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1965–1973.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: Video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*.

Tomáš Kočiskỳ, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Moritz Hermann, Gábor Melis, and Edward Grefenstette. 2018. The narrativeqa reading comprehension challenge. *Transactions of the Association for Computational Linguistics*, 6:317–328.

Xiang Kong, Varun Gangal, and Eduard Hovy. 2020. Scde: sentence cloze dataset with high quality distractors from examinations. *arXiv preprint arXiv:2004.12934*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. Race: Large-scale reading comprehension dataset from examinations. *arXiv preprint arXiv:1704.04683*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*.

Hugo Liu and Push Singh. 2004. Conceptnet—a practical commonsense reasoning tool-kit. *BT technology journal*, 22(4):211–226.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *arXiv preprint arXiv:2007.08124*.

Pan Lu, Lei Ji, Wei Zhang, Nan Duan, Ming Zhou, and Jianyong Wang. 2018. R-vqa: learning visual relation facts with semantic attention for visual question answering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1880–1889.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2200–2209.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

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

Siva Reddy, Danqi Chen, and Christopher D Manning. 2019. Coqa: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266.

Matthew Richardson, Christopher JC Burges, and Erin Renshaw. 2013. Mctest: A challenge dataset for the open-domain machine comprehension of text. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 193–203.

David Saxton, Edward Grefenstette, Felix Hill, and Pushmeet Kohli. 2019. Analysing mathematical reasoning abilities of neural models. *arXiv preprint arXiv:1904.01557*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

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.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326.

Haitian Sun, William W Cohen, and Ruslan Salakhutdinov. 2021. Conditionalqa: A complex reading comprehension dataset with conditional answers. *arXiv preprint arXiv:2110.06884*.

Simon Šuster and Walter Daelemans. 2018. Clicr: A dataset of clinical case reports for machine reading comprehension. *arXiv preprint arXiv:1803.09720*.

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

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *arXiv preprint arXiv:1811.00937*.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Adam Trischler, Tong Wang, Xingdi Yuan, Justin Harris, Alessandro Sordoni, Philip Bachman, and Kaheer Suleman. 2016. Newsqa: A machine comprehension dataset. *arXiv preprint arXiv:1611.09830*.

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

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):1–28.

Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. 2017. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427.

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.

Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart van Merriënboer, Armand Joulin, and Tomas Mikolov. 2015. Towards ai-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Wenhan Xiong, Jiawei Wu, Hong Wang, Vivek Kulkarni, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Tweetqa: A social media focused question answering dataset. *arXiv preprint arXiv:1907.06292*.

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*.

Yi Yang, Wen-tau Yih, and Christopher Meek. 2015. Wikiqa: A challenge dataset for open-domain question answering. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2013–2018.

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*.

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 *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2019. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*.

Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. 2016. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–206.

Weihao Yu, Zihang Jiang, Yanfei Dong, and Jiashi Feng. 2020. Reclor: A reading comprehension dataset requiring logical reasoning. *arXiv preprint arXiv:2002.04326*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset. *arXiv preprint arXiv:2104.08671*.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.