Image Captioning and Visual Question Answering with External Knowledge

APPROVED BY

SUPERVISING COMMITTEE:

Danna Gurari, Supervisor

Kenneth R. Fleischmann

Image Captioning and Visual Question Answering with External Knowledge

by

Chongyan Chen

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN INFORMATION STUDIES

THE UNIVERSITY OF TEXAS AT AUSTIN

May 2020

Dedicated to my parents.

Acknowledgments

I would like to thank my supervisor Dr. Danna Gurari and my second reader Dr. Ken Fleischmann for their guidance and support. I especially wish to express my gratitude to Dr. Danna Gurari, for her invaluable advice, patience, and kindness. I would like to pay my regards to the University of Texas at Austin. I learned a lot at the school of Information in the past two years and hope my future Ph.D. study goes well and can contribute to VQA/image captioning fields. I also would like to recognize the financial support from Microsoft and to thank all the people who collected the datasets as well as all the crowd workers. My thanks also would go to my beloved parents for their continuous support.

Image Captioning and Visual Question Answering with External Knowledge

Chongyan Chen, M.S.Inf.St.
The University of Texas at Austin, 2020

Supervisor: Danna Gurari

The fields of computer vision and natural language processing have made significant advances in visual question answering (VQA) and image captioning. However, a limitation of models in use today is they typically perform poorly when the task requires common sense or external knowledge. Motivated by this observation, this work offers an exploration of the benefits of multi-source external knowledge for these two tasks. Three kinds of external knowledge are evaluated: knowledge base, reverse image search, and image search by text. This work demonstrates the advantage of these external knowledge sources via experiments on two image captioning datasets (COCO-Captions and VizWiz-Captions) and three visual question answering datasets (VQAv2, VizWiz-VQA, and OK-VQA).

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Image captioning [6, 78] and Visual Question Answering (VQA) [4] are two popular research topics in the natural language processing and computer vision communities. Image captioning requires the computer to generate a description for a given image. Visual question answering requires the computer to generate an answer for a given question about an image. Currently, these two tasks benefit visually impaired people by allowing them to get information about both digital environments and their physical surroundings. For example, Facebook [73, 80] provides automated image captioning in social media so that visually impaired people can get information about images. In addition, mobile phone applications such as BeMyEyes, BeSpecular, and TapTapSee empower visually impaired people to learn about pictures they take of their surroundings. More generally, these tasks can be valuable for many real-world problems including for Content-Based Image Retrieval (CBIR) by helping with image indexing, commentary on videos, and analysis in the medical domain.

Some solutions for image captioning and VQA are human-powered. However, this approach is limited because it can be expensive, time-consuming, may not be available 24/7, and have privacy issues, e.g., helping visually im-

paired people to recognize the number on a bank card.

Consequently, lots of AI algorithms are proposed for image captioning and VQA. One classic approach is to build a convolutional neural network (CNN) for image processing and, optionally, a Recurrent Neural Network (RNN) (e.g., Long Short-Term Memory or LSTM) for question (text) processing [50, 54, 70, 67]. Often, such methods learn to perform the tasks by training on a large dataset of examples.

A current challenge is how to ensure automated image captioning and VQA methods learn to move beyond a shallow understanding of the image with, optionally, a question. That is because many images and visual questions may require additional information that is not contained in the image. For example, Figure 1.1 shows some images from a popular computer vision dataset called ImageNet [21] and the label and descriptions provided about the images. Whiles computers may easily know the images show schools and answer related questions about schools, the labels or descriptions are not specific and rich enough to answer questions like "which school is it?" and "what is the history of this school?" As another example, common sense is needed to answer for an image showing "Kit Kat" the question "Is it edible?"——knowledge that Kit Kat is a company producing chocolate and chocolate is edible are possible steps an algorithm could take to arrive at the answer.

Observing that humans who complete image captioning and VQA make use of their common sense, we hypothesize it's also beneficial for machines to get information from external knowledge sources. In this work, we define

Synset: school, schoolhouse has bounding box

School image from ImageNet

Figure 1.1: Examples of images from the popular computer vision dataset ImageNet, and the corresponding label and descriptions for the images.

internal knowledge source as the question and image given by the user, while external knowledge sources could be in any other form. Knowledge Bases (KBs) and search engines are two kinds of information retrieval methods that can provide a larger scale of external information for the VQA and image captioning tasks. For example, results using a Google reverse search is shown in Figure 1.2 to find a description for a given image. In addition, results from querying a knowledge base extracted from Wikipedia called DBpedia is shown in Figure 1.3.

In this paper, we propose using different kinds of external knowledge sources, including knowledge bases and search engines (image search by text and reverse image search) for VQA and image captioning tasks. We evaluate such methods on two image captioning tasks (COCO-Captions and VizWiz-Captions) and three visual question answering tasks (VQAv2, VizWiz-VQA, and OK-VQA). While numerous efforts [84, 69, 72, 68, 51] have explored how knowledge bases can be used for VQA and how reverse image search can be used for image captioning [81, 24, 18], previous research focuses on using a

Image size:
251 × 201

Find other sizes of this image:
All sizes - Small - Medium - Large

Possible related search: ***university of texas, main building tower***

Google reverse search result

Figure 1.2: Use Google image reverse search for an image and the results are "University of texas, main building tower".

"The University of Texas at Austin, informally UT Austin, UT, University of Texas, or Texas in sports contexts, is a public research university and the flagship institution of the University of Texas System. Founded in 1881 as "The University of Texas," its campus is in Austin, Texas—approximately 1 mile (1,600 m) from the Texas State Capitol. The institution has the seventh-largest single-campus enrollment in the nation, with over 50,000 undergraduate and graduate students and over 24,000 faculty and staff."@en

Figure 1.3: 'comment' field returned by query "University of Texas at Austin".

single external knowledge source. In contrast, we tested multiple external knowledge sources to reveal which types of external knowledge are most beneficial for the image captioning and VQA tasks.

Importantly, our analysis addresses a real-world challenge for people

who are visually impaired: since they cannot know whether they took a picture that provides enough information for image captioning and VQA, external knowledge may be regularly needed to perform these tasks. Prior work has shown that, for over 30,000 visual questions asked by blind photographers, roughly 50% of the images suffer from quality issues [17] and 28% of are labeled as "unanswerable" because of missing content of interests or image quality issues [27]. Yet, according to [15], some low-quality images are still answerable when "humans can make inferences". Our analysis highlights a potential benefit of external knowledge to help the computer both when "people are not familiar with the object" and to fill in the missing information gap when the image doesn't give enough information or has low-quality issues.

Chapter 2

Related Work

2.1 Image Captioning

Image captioning has a rich history. For example, in 2006, Bigham et al. proposed WebInSight [13], which mainly makes use of HTML tags like titles of linked pages. More recently, to provide captions for Twitter images, Twitter A11y [18] was proposed to utilize three methods to obtain descriptions. If the images are externally linked preview images, it fetches the descriptions from the external URL. If the image depicts primarily text, it uses OCR text recognition via Google Cloud Vision API to generate a description. In other cases, Amazon Mechanical Turk is used to collect descriptions from humans. Another approach from the artificial intelligence community has entailed introducing (about 20) image captioning datasets publicly to support large-scale training of deep neural network algorithms to automatically perform this task [28, 9]. Automated image captioning methods, according to [30], can be divided into three categories: template-based, retrieval-based, and novel generation. From the latter category, many methods are based on an encoder-decoder framework (e.g., [66, 76]). [24] and [81], in contrast, only use image retrieval to find an exact image match.

2.2 VQA

DAtest for QUestion Answers on Real-world images (DAQUAR) [49] was the first important VQA dataset. It was a small dataset with many low-quality images, and grammar errors in questions and answers sometimes. In 2015, COCO-QA dataset [16] was released. The answers and questions in COCO-QA dataset are generated by a computer adapting the descriptions provided in COCO-Captions [44]. Also in 2015, Antol et al. proposed the VQA v1 dataset [4]. The VQA dataset has two kinds of answer modes: multiple-choice and open-ended.

In 2016, Qi Wu et al. [71] surveyed recently developed methods for VQA and summarized them into four categories: joint embedding approaches, attention mechanisms, compositional models, and models using external knowledge bases. For joint embedding approaches, researches usually use convolutional and recurrent neural networks to extract features separately and feed these features to a classifier [22]. For attention mechanisms, [75] developed a structured spatial attention mechanism. [77] stacked attention networks that reason sequentially to get an answer. [20] found that it seems that humans and deep networks don't pay attention to the same regions when answering visual questions. [2] combined bottom-up (based on Faster R-CNN) and top-down attention mechanisms for both image captioning and visual question answering. [60] proposed Question Type-guided Attention (QTA), which balances bottom-up and top-down visual features based on the question type. For compositional models, [3] decomposed questions, and jointly trained neural modules included

in deep neural networks. Another example for compositional models is [74], consists of four modules: input module, question module, episodic memory module, and answer module. Other researches like [32, 42, 41] explored the VQA explanation and reasoning. In 2017, MUTAN [8], a multi-modal tensor-based Tucker decomposition was proposed. Recently, VisualBERT [40] also reported great results. However, only a few models use external knowledge bases. [72] is one of the most classical experiments that use DBpedia for external knowledge.

Prior work showed that many visual questions need external knowledge to answer them [4]; i.e., 47.43% of studied questions. Prior work also found that crowd workers can provide different answers to the same question for reasons that external knowledge would be useful including [10]: (1) people are not familiar with the image content or (2) the image doesn't give enough information because it is blurred, incomplete, or just missing the information.

Altogether, prior work has examined the benefit of KBs for VQA and shown KBs perform well on image datasets like the COCO-QA dataset [58], where questions were generated by computers based on image captions, or KB-VQA dataset [69], where questions are asked in pre-defined formats. They also proved that customized KBs [61] help VQA on customized datasets. However, to our knowledge, prior work has not explored how general KBs works on a real task in real life that lack pre-formatted questions or additional customized facts.

2.3 VizWiz Challenge

Even though lots of AI algorithms are proposed to fulfill the needs of visually impaired people, most of them are trained on images that are not taken by visually impaired people. Thus, these algorithms perform poorly when it comes to real users' images. That poor performance often is attributed to gap between the visual questions in traditional VQA and the visual questions blind people ask. According to [35], "blind users often know the general object category but are interested in specific characteristics of those objects such as color, kind, flavor, label, brand, and name", while existing traditional VQA dataset consists of images showing a limited number of object categories, e.g., COCO Dataset has 80 object categories and 91 stuff categories. Thus, the VizWiz-VQA challenge [27] and VizWiz-Caption challenge [28] fill an important gap by reflecting visually impaired people's real needs. More generally, recent research have explored different aspects of meeting the real interests of blind people, including what skills are needed, reasons for different answers to their visual questions [10], visual question answerability [27], reasons for poor image quality [17], and privacy issues in their images [26].

2.4 External Knowledge

Knowledge Base A knowledge base is a collection of complex structured and unstructured knowledge. KBs that have been used in VQA are DBpedia [5], ConceptNet [45], and WebChild [63]. Microsoft Concept Graph [34] and Google Knowledge Concept [62] also are relevant knowledge graphs.

Zhu et al. [84] in 2015 introduced a KB construction system that can build a customized KB in several hours to handle an assortment of heterogeneous visual queries using large-scale multiple reference frames. Qi Wu and Peng Wang et al. [69] in 2015 proposed Ahab for explicit reasoning on KB-VQA, a small dataset with 700 images from the MS COCO. Their questions were generated by human beings in well-designed templates and they queried DBpedia for the results. They in 2016 [72] employed a convolutional neural network (CNN) to predict high-level concepts of the image and query the attributes on DBpedia. Then the results of DBpedia were encoded using Doc2Vec and fed into an LSTM to predict the answer. They reported a final model using a Att+Cap+Know-LSTM(Attributes+Caption+KnowledgeBase vectors fed into LSTM). This model outperforms Att+Cap-LSTM by 0.71%. This research suggests the potential of external knowledge. They in 2018 [68] introduced the FVQA dataset, which extends the VQA dataset with additional image-question-answer-supporting fact tuples. They built the KB using the combination of DBpedia, WebChild, and ConceptNet. However, this FVQA dataset just retrieves limited facts from three KBs and the triples cannot represent general knowledge comprehensively. Marino et al. [51] in 2019 offered a knowledge-based VQA dataset named Outside Knowledge VQA(OK-VQA), which only selects images that require external knowledge to answer the question from COCO dataset. They also provided a benchmark for it. [61] proposed the text-KVQA dataset, which contains 257K images, 1.3 million question-answer pairs, and associated three domain-specific knowledge bases:

KB-business, KB-movie, and KB-book. They recognized text in images and conducted knowledge graph reasoning based on the text using gated graph neural networks.

Our method is different from the previous research using KBs for VQA because we examine how KBs work on a real VQA task for people who are blind and compare the performance of numerous knowledge sources on different tasks. Additionally, we evaluate the performance of numerous queries for KBs based on object, text, description, and image search by text compared to just "attributes" [72].

Search Engines: Image search by text and reverse image search Image Search by Text (IST) entails inputting a search query into a search engine in order to receive relevant images and the titles of the images. Reverse Image Search (RIS) entails inputting an image to a content-based image retrieval system to receive information related to this image, which may include where is the image from, image descriptions, and similar images.

Yu Zhong et al.[81] applied reverse image search by extracting key frames from a photostream and matching images against private and public datasets with an IQ Engine to provide additional information for VQA. However, the way they matched photos with datasets remains unknown and the IQ Engine is no longer available (after being acquired by Yahoo, the public service was shut down). [29] used Google reverse image search for object classification. However, Google reverse image search API is no longer publicly provided by Google. In [52], the search engine integrated a multi-modal fusion

technique that considers high-level textual and both high and low-level visual information, supported by tree-based structures. Other options such as VisualSearchApi.com offer different image search characteristics: search by color distributions, pattern, and shapes.

Guinness, Cutrell, and Morris [24] applied reverse image search for web image captioning. They used the Bing Image Insights API to look up the sources of the image. For each web page the API found, the longest caption from the alt text, figure captions, aria-labels, and other metadata was selected. However, their method works only for the images found in multiple places on the internet, and so are not of benefit for personal photos. Moreover, they didn't explore how search by text or KBs can be useful in VQA. . [83] introduced a method that combines common sense from ConceptNet[45] with YOLO9000 [57] for object recognition, CNNs for extracting features, and LSTMs for generating image captions.

Altogether, our work is different from the works mentioned above because none of those works explore the possibility of combining KBs, reverse image search, and search by text. Generally, search engines have been proven useful for web image captioning while KBs has been proven useful in simple image dataset, the performance of search engines on VQA and the performance of KBs on image captioning remains unknown. Our analysis provides insights into the performance of one knowledge base and three visual search engines on two image captioning datasets and three VQA datasets.

Chapter 3

Methods

We now describe the image captioning and VQA tasks and the external knowledge methods we benchmarked.

3.1 Applications

3.1.1 Image Captioning

Image captioning requires a computer to generate a description for a given image. Figure 3.1 shows an image captioning example.

3.1.2 VQA

VQA is the task of providing open-ended questions about images in order to receive answers. This task involves understanding of the language, image, and common sense. Figure 3.2 shows a VQA example.

3.2 Methods

3.2.1 Knowledge Base

- **DBpedia** We select DBpedia because it is the most popular used KBs in VQA[69, 72, 68] and it offers detailed general information about an

Figure 3.1: "VizWiz_val_00001191.jpg" and its five captions collected from five different people: (1) A container of Chobani Greek yogurt that is orange flavored. (2) Package of Greek yogurt, Chobani brand, with white and pinkish coloring. (3) The top of a package of Chobani Greek yogurt. (4) Top of a Chobani Greek yogurt, blood orange flavor. (5) Top of a Chobani yogurt container with dark background.

Figure 3.2: "VizWiz_val_00001932.jpg", question asked and its ten answers collected from crowd workers. The visual question is: "Can you tell if this is vitamin C and what the Milligrams are?" The ten answers are (1) unanswerable (2) yes 500mg (3) 500 mg (4) unsuitable (5) unsuitable (6) yes 500 mg (7) no 500 (8) 500 (9) vitamins (10) vitamin c 500 mg

entity. DBpedia [5] extracts structured content from Wikipedia pages. It is accessed using SPARQL for Resource Description Framework (RDF), which is used for data interchange on the Web. The 2016-04 version of DBpedia contains 6 M entities and 9.3 billion RDF triples (1.3 billion were extracted from the English edition of Wikipedia). Since the 'comment' field gives the most general description of an attribute, we store the 'comment' text returned by the queries. The appendix shows how to query the attribute 'The University of Texas at Austin' in SPARQL language and Figure 1.3 shows the 'comment' field results. To generate queries, we use Microsoft Text Analysis API to extract keywords and entities from the combination of text recognized, object recognized, description, and the result of google image search by text.

3.2.2 Reverse Image Search

We select Google reverse image search and Bing visual search as representative of reverse image search methods.

- **Google reverse image search (GRIS)** We select Google reverse image search because it has been reported useful in many computer vision related tasks, e.g., object recognition [29]. Although Google reverse image search API is no longer provided for public use, we found an alternate Google reverse image search API provided by Zenserp company. By inputting the image or image URL to the API, we can get similar images and their titles.

- **Bing Visual Search** We select Bing Visual Search because it is reported useful for web image captioning [24].

. Bing visual search API is provided by Microsoft. It takes the image or image URL as the input. It returns the 'name' of the result returned as the answers. Other insights provided by Bing include ShoppingSources Insight, RelatedSearches Insight, and Entity Insight.

3.2.3 Image Search by Text

- **Google Image Search by Text** We select Google Image Search by Text (GIST) because 45.33% of images in VizWiz dataset need text skills to answer questions. [61] also explored knowledge bases based on the text recognized in images. Intuitively we believe it helps if more information about the text can be provided. Google image search by text is developed by Google. Since it is not public available, we use the alternative one provided by Zenserp. By inputting the text, we get the top 5 ranked results and the top 5 related search results.

Chapter 4

Image captioning

4.1 Image captioning - Datasets

- **VizWiz-Captions** [28]

  We select VizWiz-Captions dataset because it is collected from real users of a captioning serve and thus reflects visually impaired people's real image captioning needs.

  VizWiz-Captions consists of 39,181 images, and each image is paired with 5 captions. Among them, 23,431 are training images, 7,750 are validation images, and 8,000 test images. We follow the instructions suggested in the VizWiz-Captions Challenge to exclude the pre-canned and spam captions.

- **COCO-Captions** [16]

  We select MS COCO c5 captions dataset because it is popular and well presents the focus of AI community.

  The images are collected from Flickr. The image has at least one object in 80 object categories. MS COCO-Caption has 82,783 images for training, 40,504 images for validation, and 40,775 images for testing. MS COCO Captions c5 contains five captions for each images.

4.2 Image captioning - Evaluation Metrics

Common evaluation metrics for image captioning are BLEU1-4 (BiLingual Evaluation Understudy) [53], ROUGE-L (Longest Common Subsequence (LCS) based statistics) [43], METEOR (Metric for Evaluation of Translation with Explicit Ordering) [7], CIDEr-d (Consensus-based Image Description Evaluation) [65], and SPICE (Semantic Propositional Image Caption Evaluation) [19]. BLEU was proposed in 2002 and its output is the geometric mean of n-gram score. BLEU penalizes shorter answers. ROUGE [43] was proposed in 2004 for evaluating text summaries and its recall encourages detailed descriptions. METEOR was proposed in 2005 based on word-to-word matching and can match synonyms too. CIDEr was proposed in 2014 and rewards methods that match the consensus of different captions collected from multiple people. These are all based on n-gram matching, which is neither sufficient nor necessary for evaluating caption, according to [19]. SPICE is designed to evaluate the a captioning models' ability to recover objects, attributes, and relations. However, SPICE ignores evaluating fluency and grammar.

4.3 Performance on VizWiz-Captions dataset

Table 4.1 provides the results of different methods on the VizWiz-Captions dataset. We see that Microsoft Description outperforms other methods on all evaluation metrics except CIDEr-d. However, its poor performance still suggests that more work is required for image captioning on the real task.

For BLEU-1, text recognition is the second-best method, following by

text+ GIST first and Bing Visual Search. For BLEU-2, GIST top-10 is the second-best method, following by Text Recognition. For BLEU-3 and BLEU-4, Text Recognition and GIST top-10 are the second-best methods. For METEOR, text+GIST top-10 and DBpedia are the second-best methods. For ROUGE-L, DBpedia is the second-best method, following by text+GIST-first, following by Text Recognition, GIST-top10, and Bing Visual Search. For CIDEr-d, GIST-top10 is the best method, following by the Microsoft Description and Text Recognition. SPICE is not available because its algorithm is too complex and fails to handle the input when the average words of methods is too long. The second best method for SPICE is GIST-top10.

Overall the performance of each method is bad because each method only represents part of the images. The answer returned often is either too long or missing too much information. Relatively, Microsoft Description, text recognition, and Google image search by text have better performance.

Table 4.1: Results of evaluation metrics on VizWiz-Captions dataset

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ME-TEOR	ROUGE-L	CIDEr-d	SPICE
Brand Recognition	0	0	0	0	0.001	0.003	0.003	0.001
Object Recognition	0	0	0	0	0.005	0.01	0.006	0.004
Microsoft Description	0.270	0.148	0.081	0.051	0.069	0.258	0.054	0.021
Text Recognition	0.082	0.038	0.022	0.014	0.033	0.047	0.051	-
Bing Visual Search	0.057	0.010	0.003	0	0.019	0.046	0.006	0.005
Google Reverse Image Search	0.031	0.014	0.008	0.005	0.017	0.037	0.029	0.005
GIST-top10	0.008	0.043	0.022	0.013	0.023	0.046	0.057	0.008
Text + GIST-top10	0.049	0.018	0.009	0.006	0.037	0.038	0.001	-
Text + GIST-first	0.078	0.033	0.019	0.012	0.035	0.050	0.032	-
DBpeida	0.035	0.013	0.003	0.001	0.035	0.056	0	-

Table 4.2: Results of evaluation metrics on COCO-Captions dataset

Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	ME-TEOR	ROUGE-L	CIDEr-d	SPICE
Brand Recognition	0.000	0.000	0.000	0.000	0.000	0.001	0.000	0.000
Object Recognition	0.001	0.000	0.000	0.000	0.006	0.004	0.004	0.002
Microsoft Description	0.339	0.142	0.052	0.021	0.073	0.252	0.016	0.012
Text Recognition	0.001	0.000	0.000	0.000	0.003	0.007	0.000	0.000
Bing Visual Search	0.060	0.010	0.000	0.000	0.019	0.048	0.002	0.003
GIST-top10	0.048	0.010	0.000	0.000	0.018	0.021	0.000	0.001
DBpeida	0.058	0.019	0.004	0.000	0.048	0.073	0.000	-

4.4 Performance on COCO-Captions dataset

Table 4.2 presents the results of each methods on COCO-Captions validation dataset. We randomly selected 264 images from that validation set. Overall, Microsoft Description has the best performance for every evaluation metric. DBpedia and Bing Visual Search perform slightly better than other methods. DBpedia is the second-best method For BLEU-2, BLEU-3, METEOR, and ROUGE-L. Bing Visual Search is the second-best method for BLEU-1, CIDEr-d, SPICE. For BLEU-4, none method except Microsoft Description shows the effective results, which reveals that we still need to train a model to generate a complete sentence for caption for future work.

Chapter 5

Visual Question Answering

5.1 VQA - Datasets

We evaluate our methods on three visual question answering datasets: VizWiz-VQA, VQAv2, and OK-VQA.

- **VizWiz-VQA dataset** [27]

  We select the VizWiz dataset because it is collected from visually impaired people in their daily lives. Thus, it can show how external knowledge works on a real VQA task.

  VizWiz dataset is collected using a mobile phone application named VizWiz [12]. It has 20,528 training image/question pairs and 205,280 training answer/answer confidence pairs. We use the validation set for analysis.

- **VQAv2** [23]

  We select this dataset because it is one of the most popular datasets for studying VQA in the artificial intelligence community.

  The VQA dataset v2.0 includes 204,721 balanced COCO images with 1,105,904 questions and 11,059,040 ground-truth answers. The images

are selected from MS COCO [44]. The images in VQA/COCO dataset are images with complex scenes, containing at least one of 80 common object categories. The questions are asked by crowd workers, which the workers think the "smart robot would have trouble answering". Thus the limitation of the VQA dataset is that it is not a task necessarily reflecting daily life needs.

- **OK-VQA dataset** [51]

  We select the OK-VQA dataset because it highlights the importance of external knowledge to VQA.

  This dataset contains images that don't have sufficient information to answer the questions and require external knowledge. There are more than 14,000 questions in OK-VQA dataset.

5.2 VQA - Evaluation Metrics

Both VQAv2 [23] and VizWiz-VQA [26] use

$$Acc(ans) = min\left\{\frac{\#humans\ that\ said\ ans}{3}, 1\right\}$$

as the evaluation metric. This evaluation metric means, if our answer equals to more than three of ten answers from ten different humans, then the accuracy is 100%. If less than three, then divide the number of humans that said the answer by three to get the accuracy.

We adjust this evaluation metric from "equal to" to "be contained in"

as follows:

$$Acc(ans) = min\left\{\frac{\#humans' \ answers \ contained \ in \ ans}{3}, 1\right\}$$

. In other words, we regard a model as successful if it can provide the answer in the information it provides for an answer.

5.3 Performance on VizWiz-VQA validation dataset

This experiment is conducted in the VizWiz-VQA validation dataset with 1300 images. This dataset has four categories: Yes/No, Other, Count, and Unanswerable. We only use the "Other" category in this task because our method is not designed for the more rare other types of visual questions.

From Table 5.1, we see that among four computer vision client libraries provided by Microsoft, using text recognized on the object as the answer to the question can reach the highest accuracy with 25.64%. To our surprise, the object recognition API doesn't perform well. The average words returned by object recognition was only 0.19, which means it often didn't even detect any object for many images. It could be because the objects in VizWiz-VQA are more about everyday use, which are more specific and difficult for a general object recognition model. Poor image quality may also affect its performance.

Table 5.1 also reveals that the accuracy of Google reverse image search outperforms that of Bing Visual Search greatly. One of the limitations of reverse image search is that it seldom returns a highly similar image, which is discussed in a later section and exemplified in Figure 6.1.

For Google Image Search by Text (GIST), the GIST top-10 has a 1.8% increase compared to text recognition. Moreover, the accuracy of the combination of text and GIST top10 gets an 8.41% increase compared to using text recognition. It suggests that GIST top10 can provide 8.41% additional information when the input is text. Table 5.1 also reveals that the performance of GIST first and that of GIST top10 results have a great difference.

For KBs, DBpedia has the longest average words with an accuracy of 17.62%. The combination of text and DBpedia is 35.18%. This shows that solely using DBpedia cannot represent the input text well and that DBpedia can provide 10% more information.

When combining all methods but the external knowledge methods, the accuracy is 36.87%. Adding all external knowledge methods, we observe a 15.57% boost in performance. This shows that the additional external knowledge provides great benefit in VizWiz-VQA. Among all the external knowledge methods, Google search by text performs best, followed by Google reverse image search and DBpedia. Bing Visual Search provides the least information.

Although DBpedia has been reported to be promising in previous VQA research like [72, 69, 68], it doesn't show great value for the VizWiz-VQA dataset. The reason why Google results (Google search by text and Google reverse image search) provides more additional information than DBpedia results may be that DBpedia's information is not as detailed and recent as Google's information.

5.4 Performance on VizWiz-VQA validation dataset with text

Table 5.1 reveals that text in the image contributes greatly to answering questions.

To better highlight the method's benefit for text, we also conduct fine-grained analysis on 874 images in the "other" category with text in the VizWiz-VQA validation set. Images with text are filtered by a "text_detected" flag provided in the VizWiz-Captions dataset. This "text_detected" flag is set to true if at least three of the five crowdsourced workers who analyzed the image indicated text was present. As shown in Table 5.2, the results with respect to "Average words" decrease slightly compared to Table 5.1. It may because Microsoft text API detected text when less than three of the five crowd workers indicated text was present. We also notice a slight decrease in object recognition, description, and Bing Visual Search. Still, when comparing Table 5.1 and Table 5.2, we can see an overall improvement in accuracy related to text recognition. For example, brand recognition slightly improves 0.49%. Text recognition in Table 5.1 increases by 12.38% compared to text recognition in Table 5.2. The combination of all methods leads to the best results with an accuracy of 61.86%, which is 9.42% higher than that in Table 5.1. We believe that text in images and its inferences are two important factors to be considered for future research about the VizWiz-VQA dataset.

5.5 Performance on VQAv2

This experiment is conducted on the VQAv2 validation dataset with 400 randomly selected images. The possible type of answer for each visual question is categorized between: "yes/no", "number", and "other". As done for VizWiz-VQA, we again focus on the "other" category for our analysis.

Among all the external knowledge methods, DBpedia performs best and provides the most additional information. DBpedia 's great performance on VQAv2 confirms the results of previous research [72, 68]. The performance of reverse image search is slightly better than that of "Image Search by Text".

From Table 5.3, we see that the MS Description performs the best from the available computer vision APIs for VQAv2. Text recognition has a much lower accuracy than observed for the VizWiz-VQA dataset. We hypothesize that is because few images in this dataset contain text.

5.6 Performance on OK-VQA

Table 5.4 is the result of 400 randomly selected images in the OK-VQA validation dataset.

Among all the external knowledge, DBpedia has the highest accuracy and can provide the most additional information, which confirms the results of previous research [72][68].

For reverse image search, we are unable to get the results of Google reverse image search because Zenserp API cannot process non-standard image

url (the image urls VQAv2 and OK-VQA provide are non-standard). But comparing the results of "All except Bing Visual Search" to the results of "All methods", we may draw a conclusion that reverse image search may not provide additional information for OK-VQA.

Both OK-VQA and VQAv2 have a low "Image Search by Text" accuracy compared to that observed for the VizWiz-VQA dataset. We hypothesize it is because there is a low text detected rate compared to that of the VizWiz-VQA dataset. Again, we observe MS Description performs best from the computer vision API options, however in this case it performs worse than what was observed for the VQAv2 dataset. More generally, the distribution of the results from the computer vision API options resembles that of the VQAv2 dataset.

Altogether, we conclude that the Google "Image Search by Text" and Bing Visual search don't provide much additional information.

As we expected, the OK-VQA dataset has the lowest overall accuracy compared to VizWiz-VQA and VQAv2 dataset. It makes sense because we expect the questions and images in OK-VQA to be more challenging, requiring more external knowledge and more inference.

5.7 Comparing performance across three VQA datasets

To learn how much improvement can we get with external knowledge, we simply subtract the accuracy of "All except external knowledge" from that

of "All methods" (for VizWiz-VQA, it subtracts the accuracy of "All except external knowledge" from that of "All except reverse image search" because only VizWiz-VQA can get results from reverse image search API).

We observe that VizWiz-VQA dataset has 15.57% of improvement with external knowledge while VQAv2 has a 10.45% increase and OK-VQA has a 6.55% increase. It reveals that the external knowledge we select may not be enough for OK-VQA and that OK-VQA requires deeper and wider external knowledge sources. We leave this for future work to see if adding other knowledge bases (e.g., like ConceptNet) will help.

Surprisingly, from Table 5.4, Table 5.1, and Table5.3, we find the highest overall accuracy for the VizWiz-VQA dataset compared to OK-VQA dataset and VQAv2 dataset, both with or without external knowledge. This may be because a large percentage of the questions in the VizWiz-VQA dataset can be easily answered just by making some inference from the text.

Altogether, we observe that different external knowledge sources are suitable for different datasets. This highlights that there is a gap between real users' tasks and tasks concocted in contrived environments: real tasks require more specific, daily-related information which often may be found on the internet via search engines while traditional VQA tasks requires more general information which may be found on the existing knowledge base.

Table 5.1: Average words and accuracy on VizWiz VQA validation set

Method	Average words	Accuracy
Microsoft Computer Vision APIs		
Brand Recognition	0	1.23%
Object Recognition	0.19	4.13%
Microsoft Description	4.7	10.59%
Text Recognition	12.2	25.64%
Reverse Image Search		
Bing Visual Search	7.09	4.92%
Google Reverse Image Search	4.35	13.46%
Image Search by Text		
Google (the first result)	4.2	19.82%
Google (top 5 results and top 5 related search)	48.01	27.44%
Knowledge Base		
DBpedia	233.12	17.62%
Combination of Methods		
text+Google(top 10)	60.22	34.05%
Object+text+Google(first results)	16.61	34.03%
Object+text+Google(top 10)	60.4	37.56%
Text+DBpedia	245.33	35.18%
All except external knowledge	17.06	36.87%
All except GIST	261.63	47.41%
All except Google reverse image search	305.30	49.15%
All except DBpedia	76.52	49.92%
All except Bing Visual Search	302.56	50.9%
All	309.65	52.44%

Table 5.2: Average words and accuracy on VizWiz VQA validation set with text

Method	Average words	Accuracy
Microsoft Computer Vision APIs		
Brand Recognition	0	1.72%
Object Recognition	0.11	2.75%
Microsoft Description	3.07	6.18%
Text Recognition	12.14	38.02%
Reverse Image Search		
Bing Visual Search	4.97	3.89%
Google Reverse Image Search	2.95	14.68%
Image Search by Text		
GIST-first	4.02	29.10%
GIST-top10	45.79	40.39%
Knowledge Base		
DBpedia	203.87	19.07%
Combination of Methods		
text+GIST-top10	57.93	50.23%
Object+text+GIST-first	16.29	46.87%
Object+text+GIST-top10	58.05	52.10%
Text+DBpedia	216.00	45.19%
All except external knowledge	15.33	44.28%
All except GIST-top10	227.12	54.58%
All except Google reverse image search	269.95	58.77%
All except DBpedia	69.04	59.73%
All except Bing Visual Search	267.94	61.1%
All	272.9	61.86%

Table 5.3: Average words and accuracy on VQA v2 validation set

Method	Average words	Accuracy
Microsoft Computer Vision APIs		
Brand Recognition	0	0.61%
Object Recognition	0	10.3%
Microsoft Description	1	20.51%
Text Recognition	0	4.27%
Reverse Image Search		
Bing Visual Search	1	6.55%
Google Reverse Image Search	-	-
Google Search by Text		
GIST-first	0	2.06%
GIST-top10	5	3.91%
Knowledge Base		
DBpedia	39	20.06%
Combination of Methods		
text+Google(top 10)	5	5.52%
Object+text+GIST-first	1	14.27%
Object+text+GIST-top 10	6	15.12%
Text+DBpedia	39	23.33%
GIST-top10+text+DBpedia	45	24.09%
All except external knowledge	3	28.21%
All except GIST-top10	44	38.21%
All except DBpedia	10	32.12%
All except Bing Visual Search	48	36.30%
All	49	38.7%

Table 5.4: Average words and accuracy on OK-VQA validation set

Method	Average words	Accuracy
Microsoft Computer Vision APIs		
Brand Recognition	0	1.36%
Object Recognition	0	7.73%
Microsoft Description	0	8.64%
Text Recognition	0	2.73%
Reverse Image Search		
Bing Visual Search	0	4.09%
Google Reverse Image Search	-	-
Image Search by Text		
GIST-first	0	1.36%
GIST-top10	1	4.09%
Knowledge Base		
DBpedia	14	13.64%
Combination of Methods		
text+Google(top 10)	1	4.09%
Object+text+GIST-first	0	9.09%
Object+text+GIST-top 10	1	10.45%
Text+DBpedia	14	13.95%
GIST-top10+text+DBpedia	16	14.73%
All except external knowledge	3	15.00%
All except GIST-top10	16	20.78%
All except DBpedia	4	13.49%
All except Bing Visual Search	17	21.55%
All	18	21.55%

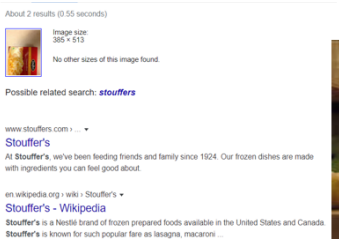
Chapter 6

Qualitative Examples for Each Method

6.1 Example of each methods

Original Image:
VizWiz_val_00001185

Bing Visual Search Insights



Google reverse image search

Google reverse image search

Description: A close up of food on a table
Text Recognition: Stouffer's. . . SIGNATURE CLASSICS fettuccine alfredo FETTUCCINI PASTA IN A CREAMY ALFREDO SAUCE MADE WITH PARMESAN CHEESE AND BLACK PEPPER NO PRESERVATIVES 1/2 0Z (326g)
DB-Pedia:
Stouffer's is a Nestlé brand of frozen prepared foods available in the United States and Canada. Stouffer's is known for such popular fare as lasagna, macaroni and cheese, meatloaf, ravioli, and salisbury steak. It also produces a line of reduced-fat products under the banner Lean Cuisine.

Figure 6.1: Example 1: As shown, reverse image search can return similar images, but not similar enough to be useful for image captioning or visual question answering.

As shown in Figure 6.1, the Microsoft description returned a general description of the image. The Microsoft text recognition works quite well. And the Microsoft brand recognition API successfully recognized it as Stouffer's.

Knowledge Bases: Google/ Microsoft Knowledge Concept returned

very similar results to the results of DBpedia. We used DBpedia because it is the most commonly used KBs in papers about VQA.

Search Engines: Most image search engines can return images somewhat similar to the original one, however, their accuracy is not satisfactory. It may because they just make use of low-level features such as color, shape, light without much consideration of mid-/high-level features about objects and scenes. Often, it appeared they didn't focus on the text on the images either. For example, shown in Figure 6.3, the similar images they found are all about a book with a yellow object at the bottom and a red title. Sometimes, they may find images according to parts of the text: shown in Figure 6.1, Qwant and Google reverse image search found the right brand (Stouffer's), but they mistook "pasta" with "meat sauce". For future work, we recommend searching on online shopping websites because a large percentage of VQA for the blind is about daily using products.

6.2 Reverse Image Search

Figure 6.2 shows how reverse image search works well for image captioning and can answer questions that are crowd workers struggle to answer. The question of this image is "What is the picture?" The ten answers returned by ten different crowd workers are (1) group people (2) kids walking bear, (3) anime characters, (4) group people walking down path, (5) unsuitable, (6) blurry cartoon, (7) men, (8) kids marching on path, (9) pokeymon, and (10) cartoon. The five captions are (1) A blurry picture of a cartoon showing sev-

Figure 6.2: Example 2: Reverse image search can work great on image captioning and VQA. The question to the image is "What is the picture?".

eral adult humans and a little pig(?) in a Red Hat. (2) A cartoon image of a bunch of characters walking along a dirt path in what might be Holland. (3) A cartoon image of various people and a small animal dancing (4) A group of animated people walking down a dirt path. (5) A part of a comic book has an animal and people marching past a windmill in the country. The only external knowledge methods that answer the question correctly are Bing visual search and Google reverse image search. The answer returned by Bing visual search is "One Piece Pictures Anime Pictures" while the answer by Google reverse image search is "One Piece Color Walk Compendium, Eiichiro Oda" (One Piece is a Japanese manga series).

6.3 Image Search by Text

Figure 6.3 shows how image search by text works well for image captioning and can answer questions that are crowd workers struggle to answer. As shown in Figure 6.3, the user asked "Can you please tell me the title of

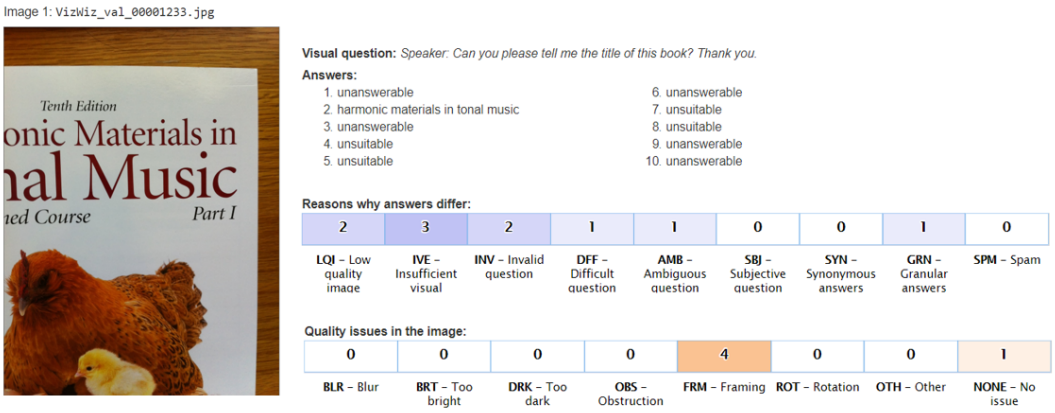
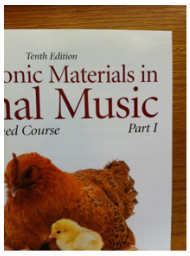


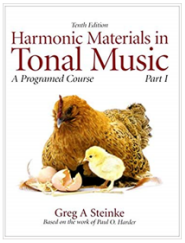
Figure 6.3: Example 3 (a) - The user asked "Can you please tell me the title of this book? Thank you." While marked as unanswerable in the VizWiz-VQA dataset, it is answerable with external knowledge. This image both suffers from framing quality issues and leads to different answers because of insufficient visual evidence in the image.

this book? Thank you." Most human answers to this question are unsuitable or unanswerable. Only one of ten people gave the right answer: the book title is *harmonic material in tonal music*. The answer to "why answers differ" is mainly about "the image provides insufficient visual evidence". The five captions of this image are (1) A book about music sits on top of a brown wooden surface. (2) A copy of a book, part 1 about music, with a picture of chickens on the front. (3) A textbook about music that is resting on a table. (4) Cover of a book or CD about music with part of title and subtitle at top, photograph of brown chicken with yellow chick, light blue background. (5) The front cover of an educational book about music and science.

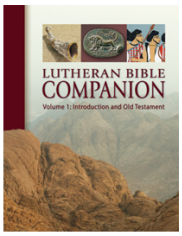
Figure 6.4 shows the text recognized "Tenth Edition onic Materials in al Music red course Part I". Searching this text in Google, we get results



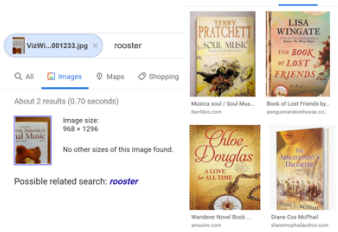
Original Image:
VizWiz_val_00001233



Ideally



Bing Visual Search Insights



Google reverse image search

Description: A bird sitting on top of a book
Text Recognized: Tenth Edition onic Materials in al Music red Course Part I
Object recognized: chicken chicken

Google Search by Text:
['Harmonic Materials in Tonal Music: A Programmed Course ...',
'Harmonic Materials in Tonal Music - Amazon.com',
'Harmonic Materials in Tonal Music: A Programed Course, Part ...',
'Billboard', 'Classical LA.', 'The Illustrated London News',
'Harmony Explained: Progress Towards A Scientific Theory of ...',
'Musical notation - Wikipedia', 'Stone Temple Pilots - Wikipedia',
'harmonic materials in tonal music part 1 pdf',
'harmonic materials in tonal music part 1 answers']

Figure 6.4: Example 3 (b) Using text recognized in an image as input to Google "Search by Text" shows promise.

including “Harmonic Materials in Tonal Music”, which is the right answer.

In this case, Microsoft object recognition API just detected “chicken chicken”. No brand is detected. And the image caption results have nothing to do with the title of the book.

The entities and keys extracted, are “Harmonic Materials”, “Edition onic Materials”, “Music red Course”, “chicken chicken”, “Tonal Music”, “Programmed Course”. None of the entities or keys yield a result from DBpedia. Even when we manually tried to search the real title in DBpedia, it still returned nothing.

Chapter 7

Discussion and Future Work

Altogether, we observe that different external knowledge sources are suitable for different tasks and datasets for VQA and image captioning. Moreover, we find that multi-source external knowledge has the potential to improve the accuracy compared to relying on a single type of external knowledge.

7.1 Image Captioning

The evaluation metrics widely used to evaluate image captioning methods are mostly based on n-gram and mainly evaluate similarity or dissimilarity between generated captions and gold standard captions created by crowd workers. N-gram models have sparcity feature spaces, which is its main shortcoming. The key exception is the SPICE score used for Image captioning datasets like VizWiz-Captions [28]. Still, regardless of this differentiation, it comes with its own limitations in that it is complex, ignores fluency and granularity, time-consuming, and has input words length limitation. For future work, we will explore other evaluation metrics like Word Mover's Distance [36].

Figure 7.1: One example from the COCO image caption dataset. The captions for this image are (1) view over a persons shoulder or from their view as they lay in bed with cats on and around them, (2) a cat laying on top of a blanket next to another cat, (3) a black cat is standing by a book, (4) someone reading a book a bed with 3 cats laying around them, and (5) several cats surrounding someone as they are laying down and reading.

7.2 VQA

In this stage, we only examined if the returned output includes enough information for the answer as well as the average words returned for VizWiz-VQA. We aim to explore additional evaluation metrics in future work.

For future work, we would also like to explore how to train a model to embrace the external information when generating answers. This could

include how to represent the top 10 Google results and DBpedia results. One plausible solution is to follow [72]'s setting to extract the semantic meanings from the descriptions returned by SPARQL query and train a Doc2Vec model. This also could include considering other KBs such as ConceptNet.

For future work, we suggest researchers working on images taken by people who are blind (e.g., the VizWiz dataset) explore the text on the images more and "image search by text", since it had a significant positive benefit for the VizWiz-VQA task. In future work, we can also consider building a personalized datset for the visually impaired people based on their subjects and matches objects in their own personalized dataset using systems like [84].

Chapter 8

Conclusion

8.1 Conclusion

We provide four main findings from our analysis of external knowledge. First, image search by text can provide much more information for the VizWiz-VQA dataset but not for VQAv2 or OK-VQA. Second, we show that knowledge bases like DBpedia can be very helpful for VQA (e.g., VQAv2, OK-VQA) but does not seem very beneficial for real users' needs (i.e., VizWiz dataset). Third, reverse image search doesn't offer much help to all studied VQA datasets. For personal images taken by visually impaired people, the visual search engine cannot return the exact same match image from other online sources, nor the image captioning for the target image. Although the visual search engine can return images which are somewhat similar to the target personal image, the similarity is not high enough for answering the related visual question. Reverse image search seems more suitable for web image captioning. Fourth, we show a possibility of answering questions about low-quality images taken by visually impaired people by using external knowledge otherwise deemed "unanswerable" by crowd workers (but are actually answerable with external knowledge).

Appendices

Appendix A

Appendix

A.1 Other possible VQA dataset related external knowledge

KB-VQA dataset [69] We didn't select the KB-VQA dataset [69] because it only has 700 images manually selected from the validation set of MS COCO and the questions are generated in well-defined format rather than in real task.

FVQA [68] We didn't select FVQA dataset because they already added fact triples to the dataset.

A.2 Other possible Knowledge Bases

ConceptNet [45] is made of common sense relationships, such as "related to", "at location", "is a", "used for", and "part of". We didn't select this KB because is more suitable for VQA reasoning rather than offering detailed information about a certain entity.

WebChild [63] WebChild involves comparative relations such as Smaller, Better, and Slower. We didn't select this KB for the same reason as not selecting ConceptNet.

SPARQL:

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX foaf: <http://xmlns.com/foaf/0.1/>
PREFIX dc: <http://purl.org/dc/elements/1.1/>
PREFIX : <http://dbpedia.org/resource/>
PREFIX dbpedia2: <http://dbpedia.org/property/>
PREFIX dbpedia: <http://dbpedia.org/>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
SELECT ?comment WHERE{
?entry rdfs:label"University of Texas at Austin"@en.
?entry rdfs:comment?comment.
}
```

Figure A.1: query attribute 'University of Texas at Austin' in SPARQL language

Google knowledge Concept [62] and Microsoft knowledge Concept [34] To our knowledge, no existing research are using the Google knowledge concept or Microsoft knowledge concept for VQA or image captioning. Thus we didn't select these two KBs.

Qwant Qwant is a French search engine powered by Bing. The images returned by Qwant are more similar to the original image compared to that returned by Bing Visual Search. To our knowledge, no existing research are using Qwant for VQA or image captioning.

A.3 The website we designed for methods visualization.

Figure A.2 shows the website we designed for visualization. The left side shows the original image, visual question, and answers returned by crowd workers. The middle two rows show relevant images returned by reverse image

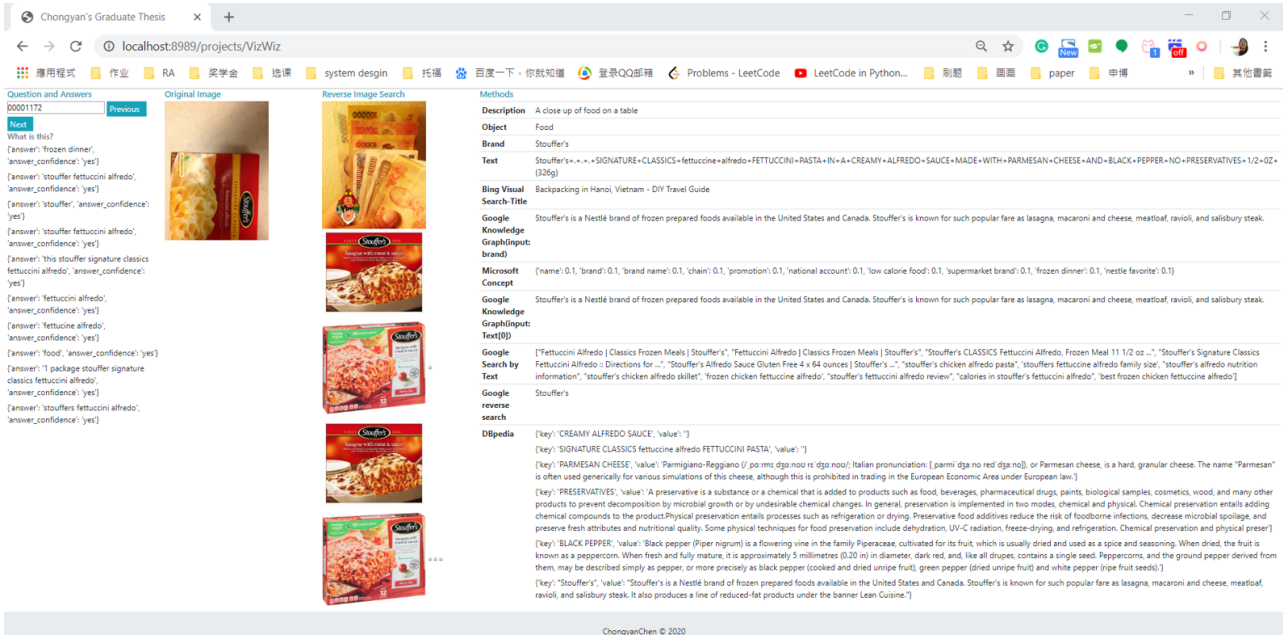


Figure A.2: Website we designed for methods visualization

search methods. The right side shows the results of each method.

Bibliography

[1] Dustin Adams, Lourdes Morales, and Sri Kurniawan. A qualitative study to support a blind photography mobile application. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, page 25. ACM, 2013.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[3] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 39–48, 2016.

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

[5] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. Dbpedia: A nucleus for a web of open data.

In *The semantic web*, pages 722–735. Springer, 2007.

[6] Shuang Bai and Shan An. A survey on automatic image caption generation. *Neurocomputing*, 311:291–304, 2018.

[7] Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

[8] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome. Mutan: Multimodal tucker fusion for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2612–2620, 2017.

[9] Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, 55:409–442, 2016.

[10] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[11] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? *arXiv preprint arXiv:1908.04342*, 2019.

[12] Jeffrey P Bigham, Chandrika Jayant, Andrew Miller, Brandyn White, and Tom Yeh. Vizwiz:: Locateit-enabling blind people to locate objects in their environment. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*, pages 65–72. IEEE, 2010.

[13] Jeffrey P Bigham, Ryan S Kaminsky, Richard E Ladner, Oscar M Danielsson, and Gordon L Hempton. Webinsight: making web images accessible. In *Proceedings of the 8th international ACM SIGACCESS conference on Computers and accessibility*, pages 181–188, 2006.

[14] Ali Furkan Biten, Rubèn Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas. Icdar 2019 competition on scene text visual question answering. *arXiv preprint arXiv:1907.00490*, 2019.

[15] Erin Brady, Meredith Ringel Morris, Yu Zhong, Samuel White, and Jeffrey P Bigham. Visual challenges in the everyday lives of blind people. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2117–2126. ACM, 2013.

[16] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions:

Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[17] Tai-Yin Chiu, Yinan Zhao, and Danna Gurari. Assessing image quality issues for real-world problems. *arXiv preprint arXiv:2003.12511*, 2020.

[18] Amy Pavel Cole Gleason, Emma McCamey, Christina Low, Patrick Carrington, Kris M Kitani, and Jeffrey P Bigham. Twitter a11y: A browser extension to make twitter images accessible.

[19] Yin Cui, Guandao Yang, Andreas Veit, Xun Huang, and Serge J. Belongie. Learning to evaluate image captioning. *CoRR*, abs/1806.06422, 2018.

[20] Abhishek Das, Harsh Agrawal, Larry Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions? *Computer Vision and Image Understanding*, 163:90–100, 2017.

[21] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[22] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. In *Advances in neural information processing systems*, pages 2296–2304, 2015.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[24] Darren Guinness, Edward Cutrell, and Meredith Ringel Morris. Caption crawler: Enabling reusable alternative text descriptions using reverse image search. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–11, 2018.

[25] Danna Gurari, Kun He, Bo Xiong, Jianming Zhang, Mehrnoosh Sameki, Suyog Dutt Jain, Stan Sclaroff, Margrit Betke, and Kristen Grauman. Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s). *International Journal of Computer Vision*, 126(7):714–730, 2018.

[26] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P Bigham. Vizwiz-priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 939–948, 2019.

[27] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 3608–3617, 2018.

[28] Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. *arXiv preprint arXiv:2002.08565*, 2020.

[29] András Horváth. Object recognition based on google's reverse image search and image similarity. In *Seventh International Conference on Graphic and Image Processing (ICGIP 2015)*, volume 9817, page 98170Q. International Society for Optics and Photonics, 2015.

[30] MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36, 2019.

[31] Rabia Jafri, Syed Abid Ali, Hamid R Arabnia, and Shameem Fatima. Computer vision-based object recognition for the visually impaired in an indoors environment: a survey. *The Visual Computer*, 30(11):1197–1222, 2014.

[32] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766, 2017.

[33] Chandrika Jayant, Hanjie Ji, Samuel White, and Jeffrey P Bigham. Supporting blind photography. In *The proceedings of the 13th international ACM SIGACCESS conference on Computers and accessibility*, pages 203–210. ACM, 2011.

[34] Lei Ji, Yujing Wang, Botian Shi, Dawei Zhang, Zhongyuan Wang, and Jun Yan. Microsoft concept graph: Mining semantic concepts for short text understanding. *Data Intelligence*, 1(3):238–270, 2019.

[35] Hernisa Kacorri, Kris M Kitani, Jeffrey P Bigham, and Chieko Asakawa. People with visual impairment training personal object recognizers: Feasibility and challenges. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 5839–5849. ACM, 2017.

[36] Mert Kilickaya, Aykut Erdem, Nazli Ikizler-Cinbis, and Erkut Erdem. Re-evaluating automatic metrics for image captioning. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 199–209, Valencia, Spain, April 2017. Association for Computational Linguistics.

[37] JUHO KIM. Efficient elicitation approaches to estimate collective crowd answers. 2019.

[38] Walter S Lasecki, Rachel Wesley, Jeffrey Nichols, Anand Kulkarni, James F Allen, and Jeffrey P Bigham. Chorus: a crowd-powered conversational assistant. In *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pages 151–162. ACM, 2013.

[39] Kyungjun Lee, Jonggi Hong, Simone Pimento, Ebrima Jarjue, and Hernisa Kacorri. Revisiting blind photography in the context of teachable object recognizers. 2019.

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

[41] Qing Li, Jianlong Fu, Dongfei Yu, Tao Mei, and Jiebo Luo. Tell-and-answer: Towards explainable visual question answering using attributes and captions. *arXiv preprint arXiv:1801.09041*, 2018.

[42] Qing Li, Qingyi Tao, Shafiq Joty, Jianfei Cai, and Jiebo Luo. Vqa-e: Explaining, elaborating, and enhancing your answers for visual questions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 552–567, 2018.

[43] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

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

[46] Chi-kiu Lo, Anand Karthik Tumuluru, and Dekai Wu. Fully automatic semantic MT evaluation. In *Proceedings of the Seventh Workshop on Statistical Machine Translation*, pages 243–252, Montréal, Canada, June 2012. Association for Computational Linguistics.

[47] Lin Ma, Zhengdong Lu, and Hang Li. Learning to answer questions from image using convolutional neural network. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[48] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. *arXiv preprint arXiv:1705.00601*, 2017.

[49] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690, 2014.

[50] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[51] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external

knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3195–3204, 2019.

[52] Anastasia Moumtzidou, Ilias Gialampoukidis, Theodoros Mironidis, Dimitris Liparas, Stefanos Vrochidis, and Ioannis Kompatsiaris. A multimedia interactive search engine based on graph-based and non-linear multimodal fusion. In *2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI)*, pages 1–4. IEEE, 2016.

[53] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[54] Annapurna P Patil, Amrita Behera, P Anusha, Mitali Seth, et al. Speech enabled visual question answering using lstm and cnn with real time image capturing for assisting the visually impaired. In *TENCON 2019-2019 IEEE Region 10 Conference (TENCON)*, pages 2475–2480. IEEE, 2019.

[55] Sameer S. Pradhan, Wayne H. Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. Shallow semantic parsing using support vector machines. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 233–240, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

[56] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. Overcoming language priors in visual question answering with adversarial regularization. In *Advances in Neural Information Processing Systems*, pages 1541–1551, 2018.

[57] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7263–7271, 2017.

[58] Mengye Ren, Ryan Kiros, and Richard Zemel. Image question answering: A visual semantic embedding model and a new dataset. *Proc. Advances in Neural Inf. Process. Syst*, 1(2):5, 2015.

[59] Olga Russakovsky, Li-Jia Li, and Li Fei-Fei. Best of both worlds: human-machine collaboration for object annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2121–2131, 2015.

[60] Yang Shi, Tommaso Furlanello, Sheng Zha, and Animashree Anandkumar. Question type guided attention in visual question answering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 151–166, 2018.

[61] Ajeet Kumar Singh, Anand Mishra, Shashank Shekhar, and Anirban Chakraborty. From strings to things: Knowledge-enabled vqa model that can read and reason. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4602–4612, 2019.

[62] Thomas Steiner, Ruben Verborgh, Raphaël Troncy, Joaquim Gabarro, and Rik Van de Walle. Adding realtime coverage to the google knowledge graph. In *11th International Semantic Web Conference (ISWC 2012)*. Citeseer, 2012.

[63] Niket Tandon, Gerard De Melo, and Gerhard Weikum. Acquiring comparative commonsense knowledge from the web. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[64] Marynel Vázquez and Aaron Steinfeld. An assisted photography framework to help visually impaired users properly aim a camera. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 21(5):25, 2014.

[65] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726, 2014.

[66] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[67] Cheng Wang, Haojin Yang, Christian Bartz, and Christoph Meinel. Image captioning with deep bidirectional lstms. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 988–997, 2016.

[68] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2018.

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

[70] Peng Wang, Qi Wu, Chunhua Shen, and Anton van den Hengel. The vqa-machine: Learning how to use existing vision algorithms to answer new questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1173–1182, 2017.

[71] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017.

[72] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4622–4630, 2016.

[73] Shaomei Wu, Jeffrey Wieland, Omid Farivar, and Julie Schiller. Automatic alt-text: Computer-generated image descriptions for blind users on a social network service. In *Proceedings of the 2017 ACM Conference*

on Computer Supported Cooperative Work and Social Computing, pages 1180–1192. ACM, 2017.

[74] Caiming Xiong, Stephen Merity, and Richard Socher. Dynamic memory networks for visual and textual question answering. *CoRR*, abs/1603.01417, 2016.

[75] Huijuan Xu and Kate Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[76] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

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

[78] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016.

[79] Xiaoyu Zeng et al. *Understanding & predicting the skills needed to answer a visual question*. PhD thesis, 2019.

[80] Yuhang Zhao, Shaomei Wu, Lindsay Reynolds, and Shiri Azenkot. A face recognition application for people with visual impairments: Understanding use beyond the lab. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 215. ACM, 2018.

[81] Yu Zhong, Pierre J Garrigues, and Jeffrey P Bigham. Real time object scanning using a mobile phone and cloud-based visual search engine. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, page 20. ACM, 2013.

[82] Yu Zhong, Walter S Lasecki, Erin Brady, and Jeffrey P Bigham. Regionspeak: Quick comprehensive spatial descriptions of complex images for blind users. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 2353–2362. ACM, 2015.

[83] Yimin Zhou, Yiwei Sun, and Vasant Honavar. Improving image captioning by leveraging knowledge graphs. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 283–293. IEEE, 2019.

[84] Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. Building a large-scale multimodal knowledge base system for answering visual queries. *arXiv preprint arXiv:1507.05670*, 2015.