PlacesQA: Towards Automatic Answering of Questions on the Web

Srikanth Muralidharan, Fred Tung, Greg Mori

School of Computing Science,
Simon Fraser University
{smuralid,ftung}@sfu.ca, mori@cs.sfu.ca

Abstract. Web users often post questions: "Does hotel X have a pool?", "Is museum Y wheelchair accessible?". The potential to automate the answering process presents an exciting challenge for AI systems, with many practical applications. However, to the best of our knowledge, there are not yet any public datasets for general question answering on the web. In this paper, we introduce the PlacesQA dataset, which contains 9,750 questions and answers about 750 unique places, including hotels, museums and nightlife venues, derived from questions asked by real users of travel websites. This dataset serves as a testbed for general question answering. For concreteness, we also provide sets of 73,148 and 181,266 images from these 750 places, obtained via web searches. We show that images of these places on the web provide a rich source of information that can be potentially leveraged by an automatic question answering agent.

1 Introduction

Imagine you are planning your next vacation and need to decide where to stay, where to eat, and what local attractions to see. With recommender websites such as TripAdvisor, Expedia, and Yelp becoming ubiquitous, you might look online to inform your travel choices. You may post questions: *Does hotel X have a pool? Are baby strollers allowed in museum Y? Do you have live music?*. Such questions are often answered by human experts: previous clients of the business, or increasingly, hired staff. However, you may have to wait minutes or hours for a reply; for smaller businesses and less travelled locations, you may not receive a satisfactory response at all.

The potential to automate this answering process presents an exciting challenge for AI systems. In this paper, we contribute a new dataset towards automatic general question answering on the web, and show that images on the web provide a rich source of information that can be exploited in this task.

Fig. 1 illustrates how our vision of general question answering on the web differs from traditional visual question answering (VQA) [1–17], and how our new dataset helps bridge the gap towards general question answering. In VQA, the agent is given an image and a natural language question concerning that image, and is tasked with producing the correct answer. Typically, the question

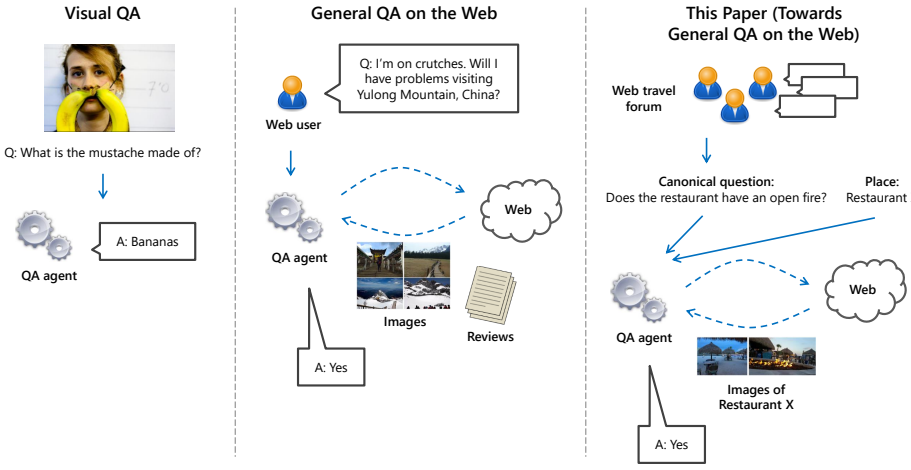


Fig. 1. This paper takes a first step towards general question answering on the web (middle), in which an AI agent is given a user question and is tasked with acquiring relevant images (and other complementary modes of information) from the web to produce an accurate answer. Our PlacesQA dataset consists of "canonical" questions and answers covering 750 unique places, including hotels, museums, and nightlife venues. The visual QA example is from [2].

about the image is created by a human annotator. General question answering on the web starts with real questions that people ask on the web and tries to find automatic ways to answer them. Part of the task is the acquisition of relevant information from the web to correctly answer the question. This paper is a first step in the direction of general question answering. The PlacesQA dataset contains "canonical" questions about 750 unique places that include hotels, museums, and nightlife venues. The canonical questions are derived from real user queries collected from online travel forums.

To ground the task, we propose and evaluate multiple baselines that answer these questions using sets of images acquired using Google image search or Facebook pages. Inferring answers from a set of images is challenging as it requires determining how to fuse positive and negative evidence from across the set. A common fusion strategy is max pooling, in which the fused prediction is based on the most positive evidence in the set. "Max-min" pooling takes into account both the strongest positive evidence as well as the strongest negative evidence (e.g. [18]). Mean pooling averages the predictions across all images in the set. However, it may not be possible to predict the best pooling strategy a priori for a particular question; moreover, the best pooling strategy may be a combination of these common strategies. As a second contribution of this paper, we propose a novel *generalized set fusion* operator that is permutation invariant and learnable end-to-end (Fig. 5). We show that our learned set fusion operator outperforms traditional fusion in answering questions from sets of images.

Fig. 2. A demonstrative example of real world questions and answers, where Google image search results provide evidence for the answers.

We envision effective automatic question answering on the web as an instance of intelligent information retrieval, in which algorithms for understanding visual data can play an important role. Fig. 2 is a motivating example in this direction. The PlacesQA dataset is intended to challenge our algorithms to go beyond the common assumption that we have an image with the answer in front of us, and instead to learn to acquire relevant images (and other complementary modes of information) from the web and use the gathered information to generate accurate answers.

2 Related Work

Visual QA. Given an image and a natural language question about the image, the goal of a visual question answering (VQA) system is to produce an appropriate natural language answer (Fig. 1, left column). Effective VQA requires visual and linguistic inference, and current algorithms draw on advances in both computer vision and natural language processing. For example, many traditional algorithms learn joint embeddings of images and questions using convolutional and recurrent neural networks [3, 8–10]. A common extension is to learn attention models that weigh spatial regions in the image according to their predicted importance [11, 13, 14, 16]. Co-attention models jointly learn image and question attention [7, 15]. Neural module networks assemble network components based on a semantic parsing of the question [1, 4]. VQA algorithms can also benefit from external knowledge bases [12]. Mechanisms for information fusion via efficient decomposition [17] further improve performance.

Representative benchmark datasets for visual question answering include VQA [2], DAQUAR [19], COCO-QA [10], FM-IQA [3], Visual Genome [6], Visual7W [16], and CLEVR [5]. These range in complexity from simple images

to rich structured annotations such as scene graphs [6]. Question-answer pairs are typically sourced from human annotators, such as Amazon Mechanical Turk workers, who are shown an image and asked to write an appropriate question. For example, annotators may be instructed to write questions related to colors, numbers, and objects [19]; challenging, adversarial questions that a "smart robot" would have trouble answering [2]; or precise questions that are answerable if and only if the image is shown [6].

While we show that images provide a rich source of information to answer questions, our focus is on automatically answering general questions on the web. We envision a question answering agent that takes a question, acquires relevant images (and other complementary modes of information) from the web, and uses that information to answer the question. Answers may require reasoning over a single image, multiple images, or generalizing from other places (past experiences). Moreover, in traditional visual QA datasets, the image is often not needed to answer the question due to human bias in question generation [20]. Instead of lab-generated questions, our dataset is built from real questions asked by users of travel websites.

NLP QA. Question answering using information from text is a well studied task in Natural Language Processing (NLP). The TREC-QA dataset [21] is an older but commonly used NLP QA dataset. It contains editor-generated questions and candidate answer sentences selected by matching content words in the question. WikiQA [22] is a collection of 3,047 questions that are answered by sentence selection from a Wikipedia page if the answer exists. In WikiQA, only one third of the questions had correct answers. Many NLP QA datasets focus on answering the question by converting it into a query into a structured database. These datasets do not have to deal with answers occurring in naturally occurring data and the questions are factoid questions. Dong et al. [23] is a recent paper that provides results on a comprehensive list of factoid QA datasets. By far the most popular QA dataset in NLP recently is SQUAD, the Stanford QA Dataset [24]. SQUAD contains 100,000+ question-answer pairs on 500+ Wikipedia articles. However, the questions were not naturally occurring: they were collected from crowdworkers. In SQUAD, the questions were restricted to those that can be answered by span selection and the answer has to be in a given paragraph. Since the crowdworkers could see the paragraph the questions have high lexical similarity with the answer which is precluded in our dataset. The QUASAR-S dataset [25] consists of 37,000 cloze-style (fill-in-the-gap) queries from Stack Overflow and the QUASAR-T dataset [25] consists of 43,000 open-domain trivia questions and their answers obtained from various internet sources. As a result of how these datasets were constructed, NLP QA systems (but not humans) can be fooled by adversarial examples [26]. In general, the level of inference in all of these NLP QA datasets is quite low and methods that can be effectively trained to do pattern matching can obtain a relatively high accuracy.

Querying to Learn Categories. The task we pose in this work is one of general question answering via information gathering. A branch of work in computer vision studies methods that learn how to learn new concepts. Influential work in this domain focused on learning object categories from images based on web supervision [27, 28]. Inspirations from the seminal NELL work [29] for language learning led to continuous image learning systems [30]. Recent work in the domain of video category learning [31] has culminated in approaches for reinforcement learning algorithms for filtering noisy video sets to build accurate classifiers [32]. We believe the automated question answering task presents a fruitful direction for this line of techniques – learning to acquire and filter data relevant to solving a particular task.

3 PlacesQA Dataset

In this section, we describe the PlacesQA dataset. PlacesQA consists of 9,750 question-answer pairs across 750 places from 3 categories. Each category has a set of canonical questions which apply to every place in that category. Table 1 shows more details of our dataset. Along with the dataset, we provide the 73,148 Google and 181,266 Facebook image search results, in order to permit static comparisons of image-based answering systems. However, the dataset permits general-purpose gathering of other images or information sources for the question answering task.

The dataset is constructed in three stages: question collection, where we extract in-the-wild user questions from travel websites; question replication, where we replicate canonical questions across instances of a place category; and answer collection, where we use crowdsourcing to obtain the answers to those questions.

Category	Number of Places	Number of Canonical Questions
Hotels	250	18
Museums	250	15
Nightlife	250	6

Table 1. Statistics of the PlacesQA dataset

3.1 Question Collection

We first collect real-world questions posted by users of travel websites. The questions are about places that can be grouped into three categories: hotels, museums, and nightlife venues.

Real world questions about places on travel websites often contain complex sub-questions that may or may not be be related to each other. They may demand long, complex answers. Moreover, not all of these questions are visually answerable (answerable using evidence from images). We simplify the problem

to demonstrate the idea of automatically answering questions using visual cues from images: we focus solely on questions that (a) contain a single query, (b) can be answered with a yes or no[1], and (c) are visually answerable.

For instance, consider the following set of three questions about a hotel:

1. *I did not see any notes about the $22 per night. I have 3 rooms booked. We don't plan on using the services for the fee. I did not know anything about this until I searched your website. Is there anything that can be done?*
2. *Hi, do the individual rooms in this hotel have a safe you can keep your valuables in?*
3. *Girls travelling on own - is there a bar in the hotel and what time is it open in evening, or any local safe bars?*

The first question is not visually answerable, nor can it be answered with a yes or no. The second question is visually answerable, assuming the QA agent can acquire a set of images of the hotel and its rooms. The third question is interesting as an example that is not visually answerable. Although it has a visually answerable part in it, it has an additional question that is not visually answerable. Thus, of the three questions that are listed here, only the second is visually answerable.

3.2 Question Replication

To scale up our dataset, we extract a set of canonical questions that are transferrable across places within each category (hotels, museums, nightlife venues). For example, canonical questions for the hotel category include

- *Can you get access in a wheel chair?*
- *Do you have microwave in the rooms?*
- *Do you have terrace?*
- *Is there a beach at the hotel?*

These canonical questions are obtained by manually collapsing equivalent real user questions. For example, the following real questions are collapsed into the canonical question "Can you get access in a wheel chair?"

- *Can you get access in a wheel chair?*
- *Do you have disabled access?*
- *Do you disability access?*
- *Can wheelchairs be accommodated here*

Note that if we wished to preserve linguistic variation, we could either sample from the real questions corresponding to a canonical question, or replicate instances of the questions. However, we decide to focus our dataset on the answering of these questions and hence choose a single canonical question for each concept.

[1] We restrict our collection to yes or no questions because evaluating machine-generated natural language answers remains an open challenge.

Questions	
1. Hotels	
1. *Does the hotel have a pool?*	10. *Are there safes in the rooms?*
2. *Does the hotel have a bar?*	11. *Do you have air conditioning?*
3. *Is there a nightclub on this property?*	12. *Do the rooms have refrigerators?*
4. *Is there a beach at the hotel?*	13. *Does this hotel have a lift?*
5. *Do you have terrace?*	14. *Do you have parking?*
6. *Do you have tennis court?*	15. *Do you allow well behaved pets?*
7. *Does the hotel have a fitness center?*	16. *Is it suitable for disabled people?*
8. *Does the hotel have spa?*	17. *Is this place kid friendly?*
9. *Do you have microwave in the rooms?*	18. *Do you have golf course?*
2. Museums	
19. *Is there a dinosaur display?*	28. *Is this wheelchair accessible?*
20. *Is there an age limit?*	29. *Is there disabled parking?*
21. *Is there parking*	30. *Are there English translations of the signs for the exhibits?*
22. *Can I bike around the area?*	
23. *Is it air conditioned?*	31. *Is there a picnic area?*
24. *Is there an elevator?*	32. *Is there a place to store my luggage at the museum?*
25. *Are strollers allowed?*	
26. *Are pets allowed?*	33. *Is photography allowed?*
27. *Are there interactive displays for children?*	
3. Nightlife	
34. *Are children welcome ?*	37. *Do you have TV?*
35. *Is there a dress code?*	38. *Do you have live entertainment?*
36. *Is there parking?*	39. *Do they serve food?*

Table 2. Canonical questions used for our PlacesQA task.

We select canonical questions that are broadly applicable across places within a category and for which the answer is not almost always the same (i.e. either yes or no). These selection criteria exclude questions such as "Is there a wax figure of Prince?" (not broadly applicable) or "Does the museum have a gift shop?" (almost always yes). Table 2 lists all the canonical questions from all the categories used for our PlacesQA task.

Next, we replicate the canonical questions across the most popular places for each category; we use only the most popular places to maximize the likelihood that answers are obtainable online (e.g. from website text, images, user reviews). Popularity is estimated by the number of user reviews on a travel recommender website. After replication, we arrive at a set of 9,750 questions for 750 places.

3.3 Answer Collection

Given the set of questions and places from the previous step, we explore the possibility of acquiring accurate annotations using tag information available in travel websites. It offers the potential for a faster and more economical way to collect answers. We found that the hotel category has comprehensive tag

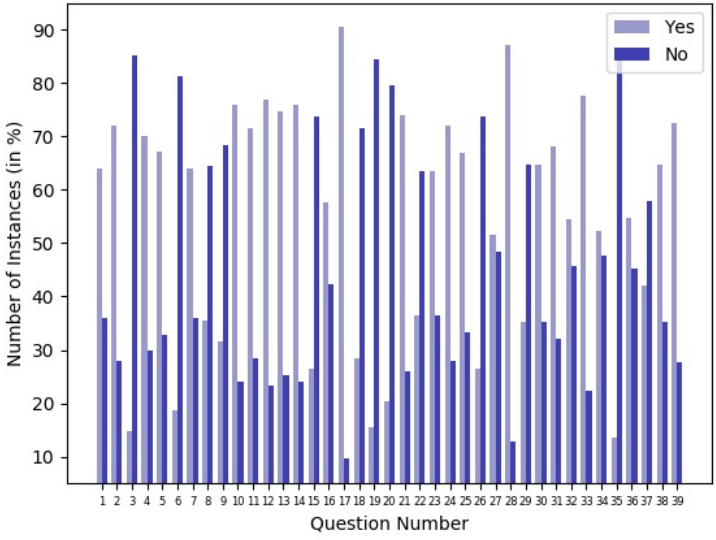


Fig. 3. Yes/No distribution of our canonical questions. Question numbers correspond to the canonical questions listed in Table 2.

information across different travel websites. Therefore, we annotated questions from the hotel category by obtaining tags for all the places from a travel website. We further performed a second round of manual cleaning of the labels for some of the "*no*" answers (relevant tags absent).

We use crowdsourcing for manual annotations for museum and nightlife categories, as we could not find such tag information in their case. A crowd worker is provided a place and the list of questions for the place, and is asked to search online for the answers. The worker may answer each question with *yes* or *no*. Fig. 4 illustrates the annotation workflow.

We also ask crowd workers to indicate the specific sources of evidence, such as a particular image or a user review, that were used to answer questions. We later download this information and add it to the dataset as meta annotation. Thus, our dataset provides verifiability of annotations using information on the web.

Fig. 3 shows the yes/no distribution of all the canonical questions.

4 PlacesQA with Images

In this work, we use freely available web images of places in our dataset as the information source to answer questions. We perform separate experiments using images from two different sources. For the first set of experiments, we download

the top 100, or as many as available, Google image search results by querying the place name, city and country. Google search results for place queries tend to give images of popular and often advertised scenes of the location, and contain rich information.

For the second set of experiments, we use photos from the Facebook page of the place whenever they are available. Images from Facebook focus more on people, and also have larger variety as they include photos of recent events that occurred at the location. We download 400, or as many as available, images considering the increased noise due to the above reasons. Note that Facebook search does not always yield results for a given place. Over the PlacesQA dataset, Facebook search yielded images for 200 hotels, 164 museums and 209 nightlife venues. When no results are returned, we default to Google image results in this experiment.

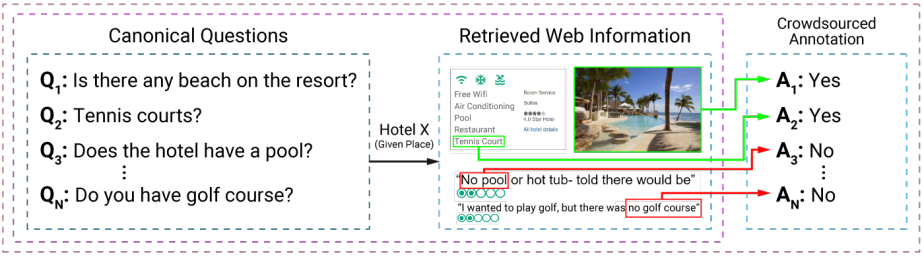


Fig. 4. Overview of our crowdsourced answer gathering pipeline. Given a collection of questions $\{Q_i\}_{i=1}^{N}$ and a place, annotators were asked to answer the questions based on information retrieved from the web (e.g. user reviews, images, official websites) for obtaining the answers $\{A_i\}_{i=1}^{N}$.

5 Set Fusion for PlacesQA

As a first approach, we consider answering the questions in PlacesQA using sets of images acquired via Google image search. Given a place and a canonical question, our baseline QA agent downloads the first 100 images, or as many as available, using the place name and city as the query and passes these images to a trained convolutional neural network (CNN). For each image, the network makes an answer prediction. The predictions across the set of images are pooled using a fusion operator to obtain the agent's final prediction. Common fusion operators include max pooling, "max-min" pooling, and mean pooling. Max pooling outputs the strongest positive evidence in the set. "Max-min" pooling subtracts the strongest negative evidence from the strongest positive evidence. Mean pooling computes an average over all images in the set. However, the best pooling strategy may not be known a priori for a particular question, and moreover, may be a combination of these common strategies.

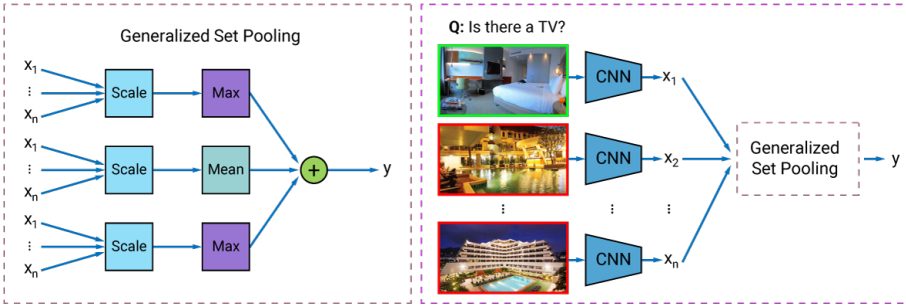


Fig. 5. Left: We propose a new permutation-invariant fusion operator for sets that generalizes common pooling approaches, such as max, mean, and "max-min" pooling, and that can be learned end-to-end. Right: Late fusion model with generalized set pooling.

We therefore propose a novel permutation-invariant fusion operator for sets of inputs that is end-to-end learnable. Our *generalized set pooling* operator is illustrated in Fig. 5, left, and consists of three branches, each of which takes the same variable-size set of input vectors. Each branch consists of a learnable elementwise scaling factor followed by a fixed pooling operator: either max (two branches) or mean. The branch outputs are summed to produce the pooling output. The learnable scaling factors determine how the pooling operator combines the max-pooled and mean-pooled outputs. Formally, given a set of input vectors $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, our set fusion operator computes

$$y = \max_i w_1 \cdot \mathbf{x}_i + \frac{1}{n} \sum_{i=1}^{n} w_2 \cdot \mathbf{x}_i + \max_i w_3 \cdot \mathbf{x}_i \tag{1}$$

where w_1, w_2, w_3 are learned scaling factors of the same dimension as the input vectors. The common non-learned fusion operators discussed above are special cases of generalized set pooling. For instance, if the input vectors have dimension 1:

- Max pooling: scaling factors of max branches sum to 1, scaling factor of mean branch is 0 ($w_1 + w_3 = 1, w_2 = 0$)
- Mean pooling: scaling factors of both max branches are 0, scaling factor of mean branch is 1 ($w_1 = w_3 = 0, w_2 = 1$)
- Max-min pooling: scaling factor of one max branch is 1, scaling factor of the other max branch is -1, scaling factor of mean branch is 0 ($w_1 = 1, w_3 = -1, w_2 = 0$, or $w_1 = -1, w_3 = 1, w_2 = 0$)

We train a separate CNN classifier for each canonical question and perform late fusion on the predictions across the place image set. The model is illustrated in Fig. 5, right. The base network is ResNet-152 [33], pre-trained on ImageNet [34]. We freeze the ResNet-152 weights and replace the final classification layer with two learnable fully-connected layers followed by set fusion – either conventional pooling(max and mean) or generalized set pooling.

6 Experiments

We have 250 places per category in total, out of which we randomly sample 150 places as the training set, 50 places as the validation set, and the rest as the test set. Therefore, our dataset consists of 450 places used as the training set, 150 places used as the validation set and 150 places used as the test set[2].

	Accuracy	Wins vs. Losses	Accuracy	Wins vs. Losses
	Hotels		Museums	
Majority	72.1	n/a	70.2	n/a
Max pooling	72.2	3 vs. 2	69.1	0 vs. 5
Mean pooling	73.5	5 vs. 1	69.4	1 vs. 2
Generalized (ours)	74.9	7 vs. 1	69.5	1 vs. 3
	Nightlife		Overall	
Majority	64.0	n/a	70.1	n/a
Max pooling	63.7	0 vs. 1	69.7	3 vs. 8
Mean pooling	64.0	1 vs. 2	70.4	7 vs. 5
Generalized (ours)	66.0	3 vs. 0	**71.4**	**11 vs. 4**

Table 3. Summary of the results obtained using traditional set fusion methods and our learned generalized set fusion using Google search images. Wins (or losses) indicates the number of questions for which the method performs better (or worse) than answering the majority answer (yes/no) for a particular question.

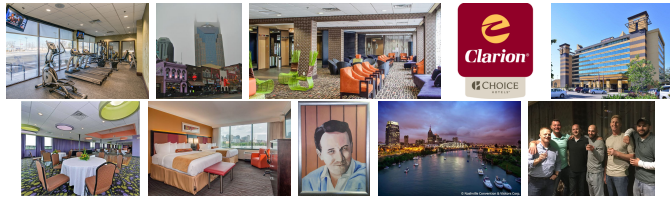
6.1 Classifier settings

Each CNN classifier was trained for 10 epochs with a fixed learning rate of 0.01, and batch size of 1 set. We use binary cross entropy loss to update the parameters of the CNN. During test time, we pick the epoch with maximum accuracy on the validation set and report the test accuracy at that epoch. If there are multiple epochs with the maximum validation accuracy, we pick the epoch with minimum validation loss.

We use accuracy and wins vs. losses to compare the baseline methods with our model. Wins (or losses) indicates the number of questions for which the method performs better (or worse) than answering the majority answer (yes/no) for a particular question. We report the results that are averaged across three independent runs. To determine whether the model has a win or a loss, we compare the average number of correct answers, rounded to the nearest integer, to the number of correct answers obtained by predicting the majority class for that question.

[2] The full PlacesQA dataset, including places, image sets, raw questions, canonical questions, answers, evidence for answers, training-validation-testing splits, etc. will be made available for download.

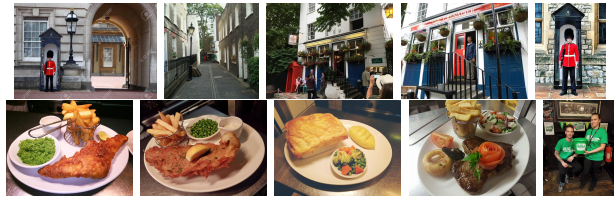
(a) Cordial Green Golf Hotel, Las Palmas, Spain



(b) Clarion Hotel, Nashville, USA

(c) Anne Frank House, Amsterdam, The Netherlands

(d) Munich Residence, Munich, Germany



(e) The Grenadier Bar, London, UK

Fig. 6. Sample place images obtained using Google image search (first row in each example), and Facebook images (second row in each example).

	Accuracy	Wins vs. Losses	Accuracy	Wins vs. Losses
	Hotels		Museums	
Majority	72.1	n/a	70.2	n/a
Max pooling	72.7	3 vs. 2	69.9	0 vs. 1
Mean pooling	72.5	4 vs. 2	69.8	0 vs. 2
Generalized (ours)	74.3	8 vs. 3	70.0	2 vs. 3
	Nightlife		Overall	
Majority	64.0	n/a	70.1	n/a
Max pooling	63.7	1 vs. 1	70.3	4 vs. 4
Mean pooling	64.0	1 vs. 1	70.1	5 vs. 5
Generalized (ours)	63.7	1 vs. 1	**71.1**	**11 vs. 7**

Table 4. Summary of the results obtained using traditional set fusion methods and our learned generalized set fusion using Facebook images. Wins (or losses) indicates the number of questions for which the method performs better (or worse) than answering the majority answer (yes/no) for a particular question.

6.2 Results

Table 3 and Table 4 show the performance of traditional set fusion (max pooling and mean pooling) and generalized set fusion (our learned pooling method) obtained using Google search images and Facebook images, respectively. From the table we observe that, overall, generalized set fusion outperforms the traditional set fusion baselines both in terms of accuracy, and the wins vs losses metric. Generalized set fusion also performs better than the majority class prediction.

Fig. 6 shows sample place images obtained from Google image search and Facebook pages. The sampled images suggest some of the biases present in the images. The retrieved images include many advertisement-style pictures and pictures taken by visitors, which leads to bias in the nature of images. This bias makes it easier to answer certain questions and more difficult to answer other questions. For example, our model performs very well on questions that are scene related questions, such as *Do you have a pool?*, as the advertisement type pictures from the figure contain images of pool, fitness center, and outdoor views of the places. Similarly, a question like *Do they serve food?* is easier to answer because visitors often take photos of food. On the other hand, having such a bias in the pictures makes it difficult to answer *Do you have air conditioning?* or *Is this wheelchair accessible?* because visitors may not find air conditioning units or accessibility ramps interesting enough to photograph.

Fig. 7 breaks down the performance of our model, using Google image search, on the individual canonical questions.

7 Conclusion

We contributed PlacesQA, a novel dataset for general question answering. PlacesQA pushes the boundaries of requirements for visual question answering systems.

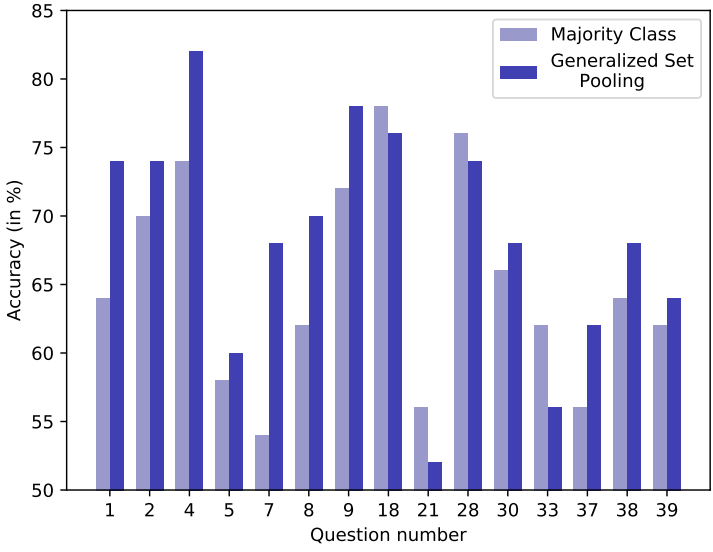


Fig. 7. Performance of generalized set pooling compared to majority class prediction. Question numbers correspond to the canonical questions listed in Table 2. We include only those questions for which the accuracy of generalized set pooling differs from majority class prediction.

Rather than assuming the right image to answer a question is provided as input, PlacesQA challenges a system to find the right information needed to answer the given question. We developed a fusion strategy to demonstrate that image sets collected via search engines can be used in an initial foray into this challenging question answering task. We believe PlacesQA can serve to spur research into algorithms for automated information gathering and harnessing of noisy web image search results to learn to answer complex questions.

References

1. Andreas, J., Rohrbach, M., Darrell, T., Klein, D.: Neural module networks. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: Visual question answering. In: IEEE International Conference on Computer Vision (ICCV). (2015)
3. Gao, H., Mao, J., Zhou, J., Huang, Z., Wang, L., Xu, W.: Are you talking to a machine? Dataset and methods for multilingual image question answering. In: Advances in Neural Information Processing Systems (NIPS). (2015)
4. Hu, R., Andreas, J., Rohrbach, M., Darrell, T., Saenko, K.: Learning to reason: End-to-end module networks for visual question answering. In: IEEE International Conference on Computer Vision (ICCV). (2017)

5. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.: CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017)
6. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual Genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision (IJCV) **123**(1) (2017) 32–73
7. Lu, J., Wang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Advances in Neural Information Processing Systems (NIPS). (2016)
8. Malinowski, M., Rohrbach, M., Fritz, M.: Ask your neurons: A neural-based approach to answering questions about images. In: IEEE International Conference on Computer Vision (ICCV). (2015)
9. Noh, H., Seo, P.H., Han, B.: Image question answering using convolutional neural network with dynamic parameter prediction. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
10. Ren, M., Kiros, R., Zemel, R.S.: Exploring models and data for image question answering. In: Advances in Neural Information Processing Systems (NIPS). (2015)
11. Shih, K.J., Singh, S., Hoiem, D.: Where to look: Focus regions for visual question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
12. Wu, Q., Wang, P., Shen, C., Dick, A., van den Hengel, A.: Ask me anything: Free-form visual question answering based on knowledge from external sources. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
13. Xu, H., Saenko, K.: Ask, attend, and answer: Exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision (ECCV). (2016)
14. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
15. Yu, Z., Yu, J., Fan, J., Tao, D.: Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: IEEE International Conference on Computer Vision (ICCV). (2017)
16. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7W: Grounded question answering in images. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
17. Ben-younes, H., Cadene, R., Cord, M., Thome, N.: Mutan: Multimodal tucker fusion for visual question answering. In: ICCV. (2017)
18. Durand, T., Thome, N., Cord, M.: Exploiting Negative Evidence for Deep Latent Structured Models. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2018)
19. Malinowski, M., Fritz, M.: A multi-world approach to question answering about real-world scenes based on uncertain input. In: Advances in Neural Information Processing Systems (NIPS). (2014)
20. Jabri, A., Joulin, A., van der Maaten, L.: Revisiting visual question answering baselines. In: European Conference on Computer Vision (ECCV). (2016)
21. Voorhees, E., Tice, D.: Building a question answering test collection. In: Proceedings of SIGIR-2000. (2000) 200–207

22. Yang, Y., Yih, W.t., Meek, C.: Wikiqa: A challenge dataset for open-domain question answering. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. (2015) 2013–2018
23. Dong, L., Mallinson, J., Reddy, S., Lapata, M.: Learning to paraphrase for question answering. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 886–897
24. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016) 2383–2392
25. Dhingra, B., Mazaitis, K., Cohen, W.W.: Quasar: Datasets for question answering by search and reading. arXiv preprint arXiv:1707.03904 (2017)
26. Jia, R., Liang, P.: Adversarial examples for evaluating reading comprehension systems. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. (2017) 2011–2021
27. Li, L.J., Fei-Fei, L.: Optimol: automatic online picture collection via incremental model learning. IJCV **88**(2) (2010) 147–168
28. Schroff, F., Criminisi, A., Zisserman, A.: Harvesting image databases from the web. IEEE transactions on pattern analysis and machine intelligence **33**(4) (2011) 754–766
29. Carlson, A., Betteridge, J., Kisiel, B., Settles, B., Jr, E.R.H., Mitchell, T.M.: Toward an architecture for never-ending language learning. In: AAAI. (2010)
30. Chen, X., Shrivastava, A., Gupta, A.: Neil: Extracting visual knowledge from web data. In: ICCV. (2013)
31. Gan, C., Sun, C., Duan, L., Gong, B.: Webly-supervised video recognition by mutually voting for relevant web im- ages and web video frames. In: European Conference on Computer Vision. (2016)
32. Yeung, S., Ramanathan, V., Russakovsky, O., Shen, L., Mori, G., Fei-Fei, L.: Learning to learn from noisy web videos. In: Computer Vision and Pattern Recognition (CVPR). (2017)
33. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2016)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575 (2014)