On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering

Xinyu Wang[1*], Yuliang Liu[1,2*], Chunhua Shen[1†], Chun Chet Ng[3], Canjie Luo[2]
Lianwen Jin[2], Chee Seng Chan[3], Anton van den Hengel[1], Liangwei Wang[4]

[1] University of Adelaide [2] South China University of Technology [3] University of Malaya [4] Huawei Noah's Ark Lab

Abstract

Visual Question Answering (VQA) methods have made incredible progress, but suffer from a failure to generalize. This is visible in the fact that they are vulnerable to learning coincidental correlations in the data rather than deeper relations between image content and ideas expressed in language. We present a dataset that takes a step towards addressing this problem in that it contains questions expressed in two languages, and an evaluation process that co-opts a well understood image-based metric to reflect the method's ability to reason. Measuring reasoning directly encourages generalization by penalizing answers that are coincidentally correct. The dataset reflects the scene-text version of the VQA problem, and the reasoning evaluation can be seen as a text-based version of a referring expression challenge. Experiments and analyses are provided that show the value of the dataset. The dataset is available at www.est-vqa.org.

1. Introduction

The fact that Visual Questions Answering [3] methods are able to answer natural language questions that relate to a wide variety of image contents has been an incredible development. The limitations of existing methods, and particularly their tendency to focus on spurious correlations in the data, have been repeatedly identified (see [1, 7, 10], for example). This is visible in the tendency of methods to answer questions on the basis of text alone. The answer to 'How many' questions, for instance, is predominantly 'Two'.

Focusing on coincidental correlations in the data represents a failure to generalize. These correlations are not stable across datasets, meaning that once the test data moves beyond the distribution of the training set, the correlations fail to hold, and methods that exploit them fail to work. The underlying reasoning, in contrast, is stable across datasets.

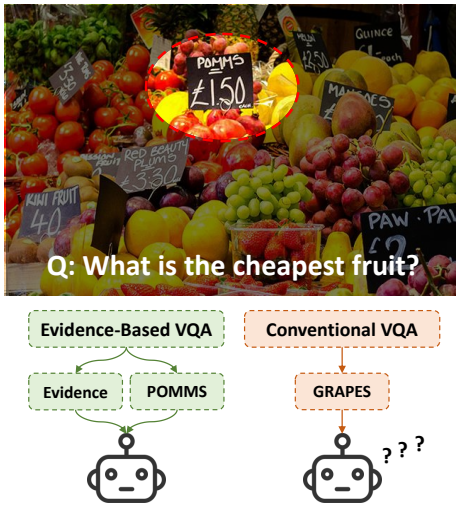


Figure 1. By requiring that vision-and-language methods provide evidence for their decisions encourages the development of approaches that depend on reasoning, and thus that are better able to generalize to new situations. It also helps to build up confidence in the provided answer.

Encouraging VQA methods to reason about the image content is thus critical to achieving methods that generalize.

One of the underlying problems with encouraging VQA methods to generalize has been that it is impossible to tell whether a method arrived at the right answer through for the right reasons. An answer is equally correct whether it results from analysis of the underlying reasoning or through exploiting a coincidental correlation in the data. A series of works have developed more sophisticated measures of performance for vision and language problems [2, 7, 36], and this work falls in this category. What distinguishes this approach is that it uses image-based grounding to encourage generalization, despite the fact that it is not actually required to achieve the desired task.

We propose here an approach to measure VQA performance that encourages generalization by demanding that the algorithm justifies its reasoning (see Fig. 1). Previous methods have applied the same rational, but suffered be-

*XW and YL contributed equally. YL's contribution was made when visiting The University of Adelaide.
†Corresponding author, e-mail: chunhua.shen@adelaide.edu.au

cause the form in which the reason must be provided is constrictive [33, 35]. We show here that it is possible instead to evaluate reasoning by only requiring a method to provide a relatively simple indication of which area of the image it has based its answer on. If the method provides the correct answer and the correct image region then it is likely that it has employed the right reasoning. Using image regions, or more accurately bounding boxes, as an evaluation metric also has the advantage that Intersection-over-Union (IoU) measures are well understood in the field.

The version of the VQA problem that we apply this approach to is Scene Text VQA. Several recent works [4, 29] have revealed that current VQA models perform badly on text VQA datasets, so it represents a compelling challenge falling within the existing framework. The various forms of text VQA problem are also of great practical importance, because text represents a critical cue to understand the content of an image. More than this, text VQA problems are typically less susceptible to solve through exploiting coincidental correlations in the data.

A variety of text-based VQA datasets [4, 15, 24, 29] have been proposed. However, there is still a significant gap between current algorithm performance and that required to support practical applications [4, 24, 29]. Another motivating factor in selecting text-based VQA rather than the generic version of the problem is that the text-based version of the problem is less susceptible to n-way classification over a fixed vocabulary. This is due to the fact that the range of text appearing in images is quite broad. The classification-based approach has repeatedly been shown to be susceptible to overfitting [1, 7]. Text-based VQA requires the development of alternative approaches, some of which will hopefully generalize.

Fig. 2 depicts some of the challenges with existing scene-text based VQA system. For example, Fig. 2(a) is a sample question that can be answered without reference to any textual content; while the question in Fig. 2(b) could have more than one correct answer; the question in Fig. 2(c) requires prior knowledge to answer; and finally in Fig. 2(d), the answer can not be obtained directly from the text in the image, but require other skills.

Empirical results presented in Fig. 3 demonstrate that current VQA approaches rely heavily on a pre-defined answer space constructed by analysis of the answers in the training set, and thus limiting generalization. As shown in Fig. 3(b), their dependence on superficial image features can render conventional VQA methods sensitive to image modifications that do not change the semantics. Fig. 3(c) and 3(d) demonstrate their propensity to generate an answer even when the required information is not present.

Text-VQA [29] employed the generic VQA accuracy as the performance metric, while ST-VQA [4] used a soft score metric inspired by the optical character recognition community. Both of these metrics are results-oriented, which means that a prediction is deemed correct if it is identical to the ground-truth. They do not assess the reasoning process. Such classification-based VQA models are able to achieve impressive performance but they are prone to overfit a fixed answer space and generalize poorly to other datasets.

To address these issues, we propose a new scene-text based VQA dataset called 'Evidence-based Scene Text Visual Question Answering' (EST-VQA). Based on this, three tasks namely *cross language challenge*, *localization challenge* and *traditional challenge* are introduced to motivate the creation of solutions with practical value from various aspects. Also, a series of baseline experiments were conducted to establish a lower bound for these three challenges. The main contributions of this paper are outlined as follows:

- **Dataset:** The EST-VQA dataset provides questions, images and answers, but also a bounding box for each question that indicates the area of the image that informs the answer. We refer to such bounding boxes as *evidence*. The dataset is intended to enable the development of text VQA methods that are closer to the levels of performance required by practical applications, but also to encourage the development of general VQA methods that generalize.
- **Evaluation Metric:** We introduce an Evidence-based Evaluation (EvE) metric, which will require a VQA model to provide evidence to support the predicted answer. For this purpose, a new VQA model is also proposed. Under this new metric, it is anticipated that it will be much more difficult for naive classification models to achieve inflated performance.
- **Bilingual:** To the best of our knowledge, the proposed EST-VQA is the first bilingual scene text VQA dataset that includes both English and Chinese question and answer pairs. The fact that the proposed dataset embodies questions in two languages further rewards methods that generalize well. It is more difficult for a method to exploit superficial correlations in questions expressed in multiple languages. The languages chosen are also particularly grammatically distinct, and reflect culturally distinct populations, which leads to different question statistics, and further encourages generalization.

1.1. Related Work

Visual Question Answering has gained significant attention recently, partly because it seems so unlikely that a method might be capable of answering all possible questions about all possible images[3, 22]. Readers are encouraged to refer to [12, 34] for a complete overview. Due to space constraint, this section only reviews the most relevant works to this paper, *i.e.*, text-based VQA.

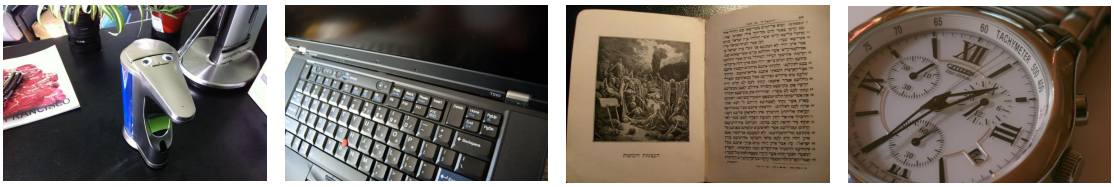


Figure 2. Some example images and QA pairs from the Text-VQA proposed in [29]. Four different types of issues are shown. (a) questions that can be answered without reading image text; (b) questions that have more than one correct answer; (c) questions that require a large amount of external knowledge to answer; (d) questions that require skills that cannot be learned from the training data alone.

Figure 3. A comparison of conventional (LoRRA [29]), and evidence-based VQA methods.

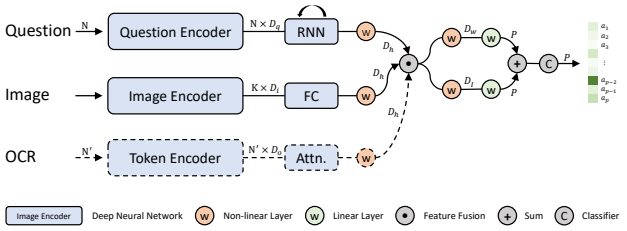


Figure 4. Illustration of the mainstream VQA models. D_q, D_i, D_o and D_h are the dimensions of the word embedding, image feature, OCR token embedding and hidden vector representations respectively. N, $N^{'}$ and P indicate question length, number of OCR tokens and answer space. Blocks with dashed lines are optional modules used for text-based VQA.

1.2. Text-based VQA

In contrast to generic VQA datasets [12, 34], text-based VQA datasets pay more attention to text related questions where a VQA model is required to read and understand textual content in an image. In [29], the authors proposed a dataset and baseline model, called Text-VQA and LoRRA respectively. LoRRA follows the structure of mainstream VQA models (see Fig. 4) where image features and word embedding are fused to train a classifier. Later, two other similar datasets were introduced, *i.e.*, ST-VQA [4] and OCR-VQA [24]. All these three datasets provide images with text related question and answer pairs. However, there

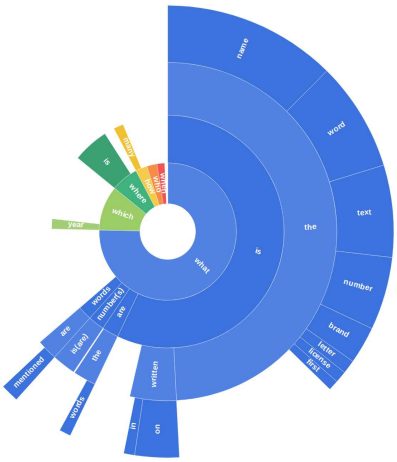
Dataset	Train + Val		Test		Image Source
	# I	# Q	# I	# Q	
[4]	19k	26k	3k	4k	[6, 8, 13, 14, 17, 23, 32]
[24]	180k	900k	20k	100k	[9]
[29]	25k	39k	3k	5k	[16]
ours	21k	23k	4k	5k	[5, 13, 14, 20, 25, 31, 32]

Table 1. A comparison of the amount and source of images between different text-based VQA datasets. #I and #Q indicate the number of images and questions respectively.

are several important differences between them, as well as to our proposed dataset:

Diversity: Table 1 shows the size and image sources of existing datasets and our dataset. Both of the Text-VQA [29] and OCR-VQA [24] images came from a single image database which is Open Images v3 dataset [16] and Book Cover Dataset [9] respectively. While ST-VQA [4] was built upon a combination of public image datasets that include multiple tasks, *e.g.*, text detection [13, 32], image classification [6], generic visual question answering [8], etc. It is noteworthy that although [24] has the highest amount of images and QA pairs, the images are all book covers, thus the diversity of images and questions are very limited. EST-VQA dataset stands out among other text VQA datasets with the consideration that existing datasets pay more attention to the question answering part, and the OCR part is almost ignored in both training and evaluation of the model.

Evaluation Metric: [29] employs a widely used VQA ac-



(a) English Question

(b) Chinese Question

Figure 5. Distribution of first four words in question sets of EST-VQA.

curacy which was first proposed in [7]. Under this metric, each question has 10 answers that are labeled by different human annotators. Supposed that the prediction of a VQA model is ans, then the score for a single sample is calculated as:

$$s_v(ans) = \min\{\frac{\#\text{humans that said } ans}{3},\ 1\} \quad (1)$$

where # indicates the number of human annotated labels that are identical to the predicted answer. This metric is robust against the incorrect answers given by some annotators. However, it is clear that only 4 discrete scores would appear, *i.e.*, $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. In [4], Levenshtein distance [18] was proposed to softly penalize a mistake. Given the predicted answer ans and ground-truth label gt, then the normalized Levenshtein similarity score s_l is given as:

$$s_l(ans, gt) = \begin{cases} 1 - NL(ans, gt), & NL(ans, gt) < \tau \\ 0, & NL(ans, gt) \geq \tau \end{cases} \quad (2)$$

where τ is a penalty threshold, and NL is the normalized Levenshtein distance between ground-truth and prediction.

2. Proposed Dataset: EST-VQA

A fundamental hypothesis in EST-VQA dataset is that a VQA model should answer a question correctly based on the textual content in an image. Therefore, we separate our scene text VQA tasks into two parts, *i.e.*, 1) text spotting and 2) question answering. In this section, we describe the process to build the EST-VQA dataset. Also, we will detail the evidence-based evaluation metric and the new tasks for EST-VQA dataset.

2.1. Data Collection

Images: As EST-VQA dataset is designed for scene text VQA tasks, we collected a total of 20,757 images from publicly available scene text detection and recognition datasets. Specifically, images annotated with English questions and answers are obtained from Total-Text [5], ICDAR 2013 [14], ICDAR 2015 [13], CTW1500 [20], MLT [25], and COCO Text [32]. Whereas, images with Chinese questions and answers are collected from LSVT [31]. All the images originated from these scene text datasets are comprised of daily scenes that include both indoor and outdoor settings.
Questions and Answers: The proposed EST-VQA dataset consists of 15,056 English questions and 13,006 Chinese questions. The question and answer pairs could be formed in cross-language *e.g.*, an English question queries the name of a Chinese restaurant so that the answer could be a Chinese text and vice versa for Chinese question. For the collection of question and answer pairs, annotators were requested to come up with questions that can be answered only by reading texts in the images. In order to avoid the question that does not require reading any text in the image, annotators are enforced to label a corresponding quadrilateral bounding box of the textual answer. The annotated bounding box will then serve as an evidence to support the answer. Moreover, yes/no questions and ambiguous questions that could have multiple correct answers are prohibited. Fig. 5 shows the common types of question, it is clear that most of the English questions start with "what", and follow by 'is' and 'the'. However, the composition of Chinese questions is far more complex than the English questions due to differences in grammar, vocabulary and other characteristics of the Chinese language. Fig. 6 shows the distribution of the length of questions and answers. Different from English words which can be segmented by space directly, Chinese words are composed of multiple Chinese

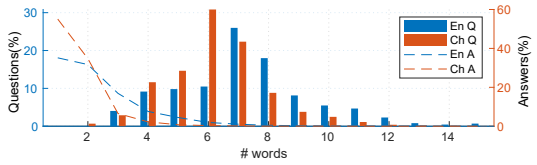


Figure 6. Percentage of question and answer length in EST-VQA dataset. Questions are tokenized by words. En and Ch stand for English and Chinese respectively.

Set	English		Chinese		All	
	# I	# Q	# I	# Q	# I	# Q
Train	11,383	12,638	9,374	10,506	20,757	23,144
Test	2,267	2,514	2,215	2,500	4,482	5,014
Total	13,650	15,152	11,589	13,006	25,239	28,158

Table 2. Volume of the EST-VQA dataset.

characters in a continuous sentence. Therefore, we use [30] to tokenize Chinese questions for counting the percentage of question length. From Fig. 6, it is clear that most of the English and Chinese questions have between 6 to 8 words, and the majority of their answers are of a single word.

In summary, as shown in Table 2, 25,239 images and 28,158 QA pairs are separated into 20,757 images with 23,144 questions for the training set and 4,482 images with 5,014 questions for the testing set.

2.2. Evidence-based Evaluation (EvE) Metric

We observed an intriguing trend among the classification based approaches for scene text VQA task. That is to say, if the ground-truth answer was included in the pre-generated answer dictionary, a generic VQA model may predict a correct answer without reading the textual content. However, such methods rely heavily on the pre-defined answer pool and so, they are unable to handle questions with out-of-vocabulary answers. Therefore, it is unclear whether such models truly have the capability to understand and reason about the questions or they are merely over-fitting to the fixed answer space. Inspired by this observation, we introduce a new evaluation protocol, named Evidence-based Evaluation (EvE) metric, which will require a VQA model to provide evidence to support the predicted answers. Under this metric, it will be much more difficult for naive classification models to achieve inflated performance.

Generally, EvE metric consists of two steps: a) check the answer; b) check the evidence. In the former, we use the normalized Levenshtein similarity score (see Eq. (2)). In the latter, we adopt the widely used IoU metric to determine whether the evidence is sufficient or insufficient. Suppose B_{gt} and B_{det} are the ground-truth and predicted bounding box respectively, then the evidence sufficiency score, E is

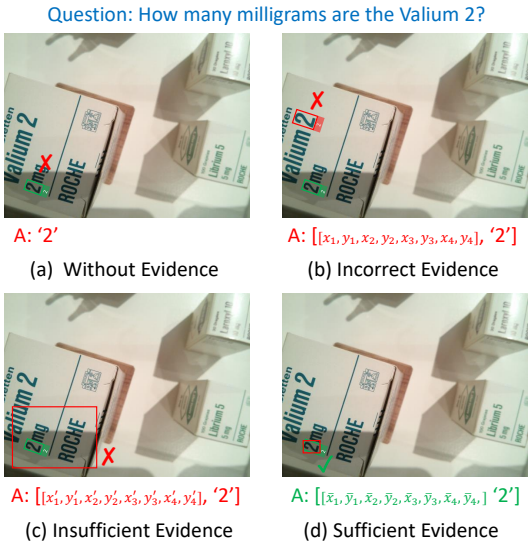


Figure 7. In EvE metric, evidence in the form of bounding box should be provided as well as the predicted answer. Green and red bounding boxes are ground-truth and predicted evidence respectively. **Incorrect**: (a) answer without evidence; (b) answer with inappropriate evidence; (c) answer with insufficient evidence. **Correct**: (d) answer with appropriate evidence. It is worth mentioning that all of the above answers would be marked as correct in the conventional VQA evaluation metric because all of them give the right answer '2'.

defined as:

$$E_\tau^i = f(\frac{B_{gt} \cap B_{det}}{B_{gt} \cup B_{det}}) = \begin{cases} \textit{Incorrect}, & E = 0 \\ \textit{Insufficient}, & 0 < E < \theta \\ \textit{Sufficient}, & E \geq \theta \end{cases} \quad (3)$$

where $\theta = 0.5$ is a predefined threshold. Under the EvE metric, only *correct* answers with *sufficient* evidence contribute to the final performance s_e (see Fig. 7) where it is given by:

$$s_e(ans, gt, E) = \begin{cases} s_l, & \text{if } E \text{ sufficient} \\ 0, & \text{else} \end{cases} \quad (4)$$

where s_l is the normalized Levenshtein similarity score as defined in Eq. (2).

2.3. Tasks

Both Text-VQA [29] and OCR-VQA [24] follow the same rules as presented in generic question answering task. Although ST-VQA [4] proposed three tasks, the only difference between each of the tasks is the size of external information (vocabulary), which is insignificant and unreasonable to properly evaluate the models' full capability. For instance, in the strongly contextualized task, all ground-truth answers are provided in a dictionary for every image with a set of distractors, which makes the VQA model prone to overfit the provided vocabulary. Besides, it becomes more

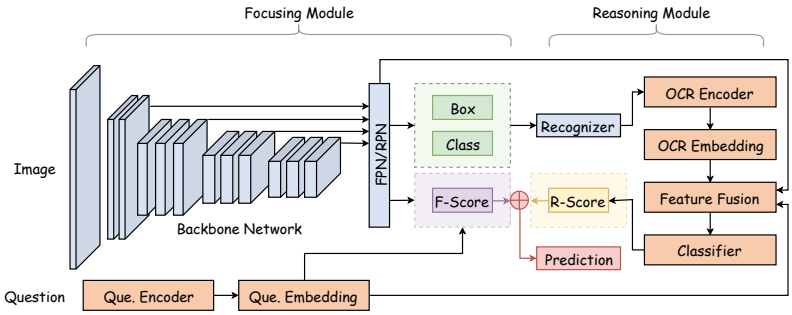


Figure 8. Overview of the QA R-CNN architecture

difficult for these models that are trained on a fixed dictionary to generalize to other datasets.

As a result of this, we propose three related tasks namely as *Cross Language Challenge*, *Localization Challenge*, and *Traditional Challenge* that will be detailed next to improve the task diversity. An online evaluation server will be set up for results submission.

- **Cross Language Challenge (CLC):** As the proposed EST-VQA dataset is a bilingual VQA dataset that contains both English and Chinese QA pairs. This challenge aims to explore a model's ability in extracting common knowledge between different languages. Under this challenge, the candidates are requested to submit results predicted by both the monolingual (*English-only*, *Chinese-only*) and bilingual models with an identical framework (*e.g.* network structure) for evaluations. The proposed EvE metric is used to evaluate the model's performance in this challenge.
- **Localization Challenge (LC):** To gain insights into a VQA model, we encourage candidates to train an evidence based VQA model to simultaneously predict the answer and its corresponding bounding box as evidence, instead of simply employing an off-the-shelf OCR system to obtain the OCR tokens. Hence, the main objective of this challenge is to explore the VQA model's ability in understanding the question and locating the correct image space that contains the answers. That is to say, this challenge requires the VQA model to provide the spatial location where an answer will be most likely to appear in an image based on a question. Compared to the full challenge, LC ignores the text recognition error and the difficulties of combining multiple OCR tokens for long answers. IoU between the predicted and ground-truth bounding box is employed as the performance metric for this challenge.
- **Traditional Challenge (TC):** We maintain the traditional VQA challenge that is consistent with the existing VQA datasets in which this challenge does not consider the evidence for the predicted answers. The normalized Levenshtein similarity score between the prediction and ground-truth is employed as the metric for this challenge.

3. Baselines and Results

3.1. Baseline Methods

This section presents the naive baseline models and two state-of-the-art VQA methods [28, 29] that were employed in the experiments. This helps to show the difficulty of the proposed EST-VQA dataset and the new tasks. The entire EST-VQA dataset is separated into *training* and *testing* sets (see Table 2), and 10% data from the *training* set is used for validation.

Vocabulary Upper Bound: As both [28] and [29] are classification based, two dictionaries are built under the widely used rules. Specifically, a small vocabulary (SV) is built with 927 English and 365 Chinese answers that appeared more than once in the training set and a Larger Vocabulary (LV) is built with 8,102 English and 8,212 Chinese unique answers. We explore the upper bound accuracy of the pre-generated SV and LV. We assume that answers included in the dictionaries can always be predicted correctly with perfect evidence to calculate the upper bound accuracy.

OCR Upper Bound: Since the traditional VQA models cannot obtain OCR tokens and info directly, we employ the state-of-the-art pre-trained text detection and recognition models [21, 27] to extract OCR bounding boxes and characters. To evaluate the effectiveness of the OCR system, we calculate the OCR upper bound accuracy on the test set. All of the answers and evidence are directly obtained from the OCR results (and suppose the correct one can always be selected), it also considers combinations of up to 4 OCR tokens for multi-word answers.

Random OCR Tokens: To assess arbitrary chance, this baseline returns a random OCR token and its bounding box from the OCR results for each question to obtain the random accuracy.

State-of-the-art Approaches: Both state-of-the-art generic [28] and scene text [29] VQA models are employed as baselines to verify the difficulties of the EST-VQA dataset. It is important to note that these methods cannot provide evidence to support their predicted answers. Therefore, we queried the predicted answers from OCR results, *i.e.*, if there are any identical OCR tokens to the predicted answer, then one of the predicted bounding boxes would be ran-

Model	CLC (%)							LC (%)			TC (%)					Δ_r
	Mono.		Bi.					Bi.			Mono.		Bi.			
	En	Ch	En	Ch	S	L	Acc	En	Ch	Acc	En	Ch	En	Ch	Acc	
SV UB	-	-	-	-	-	-	-	-	-	-	31.1	7.8	31.3	8.9	20.1	-
LV UB	-	-	-	-	-	-	-	-	-	-	48.0	16.1	48.3	17.0	32.7	-
OCR UB	33.9	24.5	33.9	24.5	44.1	14.3	29.2	50.0	37.8	43.9	38.5	28.2	38.5	28.2	33.3	-
Random	4.4	1.1	4.7	1.2	5.1	0.8	3.0	15.1	5.1	10.1	5.8	1.5	5.9	1.5	3.7	0.81
P[28]+SV	4.3	0.1	4.5	0.1	4.3	0.2	2.3	17.2	1.8	9.5	8.0	0.7	7.7	0.7	4.2	0.54
P[28]+LV	4.7	0.2	4.4	0.2	4.2	0.3	2.3	17.4	2.4	9.9	9.2	0.8	8.2	0.6	4.4	0.52
L[29]+SV	**8.2**	1.2	8.4	2.0	9.6	0.8	5.2	18.0	5.4	11.7	**12.0**	2.6	**13.2**	3.3	8.2	0.63
L[29]+LV	7.7	0.5	6.8	0.7	6.8	0.7	3.8	**18.5**	3.9	11.2	**12.0**	1.6	11.2	1.7	6.5	0.58
QA R-CNN	7.7	1.4	**8.8**	**3.2**	**10.8**	**1.1**	**6.0**	18.3	**7.3**	**12.8**	9.6	2.2	10.6	4.0	7.3	**0.82**
QA R-CNN w/ tricks	7.4	**1.5**	8.4	2.9	10.3	1.0	5.7	18.3	7.2	**12.8**	11.8	**7.9**	12.7	**9.4**	**11.0**	0.52

Table 3. Quantitative results of the three tasks in EST-VQA dataset. Mono. and Bi. represent monolingual and bilingual model respectively while S and L are short (one word) and long (more than one word) answers. Scores in bold are the best performance across models.

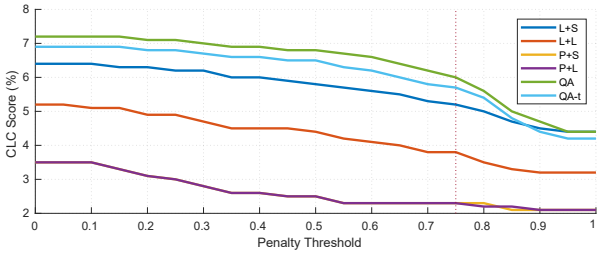


Figure 9. CLC score under different τ

domly selected as evidence, otherwise bounding box of the token which has the smallest normalized Levenshtein distance is selected.

QA R-CNN: It is noteworthy that all of the aforementioned baseline methods cannot simultaneously output the answer and its corresponding bounding box as evidence. Therefore, we propose QA R-CNN. Generally, QA R-CNN consists of two parts: Focusing Module (FM) and Reasoning Module (RM) (see Fig. 8). The core component in FM is a customized Faster R-CNN network trained for text detection task. Compared to the regular Faster R-CNN which only predicts the bounding box and object category, QA R-CNN also outputs a focusing score for each of the bounding boxes. Technically, word embedding of question is first extracted by GloVe [26] for English questions and Word2Vec [19] for Chinese questions. Then, the embedding is fed into LSTM layers to obtain question features. Following this, both question and image features are concatenated to classify the bounding box into answer area and non-answer area. This enables the QA R-CNN to gain the ability to draw its attention to the area that the answer may appear in the image. As such, a straightforward idea is that the model can directly use the underlying text of the bounding box with the highest focusing score as the question's answer. However, the rich semantics of the textual content will not be considered. Therefore, RM is introduced to further improve the pipeline. In RM, we follow the similar architecture in LoRRA where the semantics of detected text are further explored. Specifically, word embedding of the OCR tokens is extracted by FastText [11] models that are pre-trained on English/Chinese Wikipedia, and then the OCR embedding is fused with both image features and question embedding for further classification. Different from other classification-based approaches, we do not use a pre-defined fixed dictionary as the answer space but only use the detected OCR tokens, *i.e.*, only the detected text can be used as the answer. In the end, the weighted score of FM and RM are summed up for the final prediction.

3.2. Results

Quantitative Results: Table 3 summarizes the results of the baselines and our method on the EST-VQA dataset. The penalty threshold τ is practically set to 0.75 during the evaluation to ensure the answer quality. Fig. 9 shows the CLC score under different τ for bilingual models.

We first measure the upper bound performance of the two pre-defined dictionaries SV and LV. Similar to other scene text VQA datasets, SV and LV can achieve high accuracy on English questions, *i.e.*, 31.1 and 48.0 respectively. However, they failed catastrophically on the Chinese questions due to the language features and lower overlapping of answers between the training and testing splits. Hence, it is more difficult for the classification based method to obtain a promising performance on the Chinese split in the EST-VQA dataset. We also provide the upper bound accuracy of the OCR results that are generated by [21, 27], and it achieves better accuracy on Chinese questions compared to the fixed vocabularies. Then a baseline using random OCR token is set as a comparison with other approaches, and this heuristic method only achieves 3.0 and 3.7 overall score for the CLC and TC tasks respectively.

To further justify the need for EST-VQA, we trained two state-of-the-art approaches, *i.e.*, Pythia (P) [28] and LoRRA (L) [29]. As shown in Table 3, both methods perform badly on Chinese questions due to a large amount of out-of-vocabulary answers in the test set. Also, as the CLC task requires a model to provide evidence as well as the answer, the accuracy of all of the studied methods dropped significantly when compared to the TC score. This is because the models infer the answers without actually reading the textual content in the images (see Fig. 3(c) and 3(d)), thus they

Figure 10. Visualization of the output answers on the EST-VQA dataset from different models (first four images). Green and Red bounding boxes are ground-truth and predicted evidence by QA R-CNN. (More examples can be found at https://arxiv.org/abs/2002.10215)

can not provide reasonable evidence to support the answer. In contrast, the proposed QA R-CNN shows more robust results on the three tasks (see Table 3).

To further explore the proposed CLC task, we also trained a QA R-CNN with bells and whistles, many heuristic manual rules are adopted to lift the performance. Under this model, it outputs answers predominantly from the vocabulary for a certain type of questions. And if the model failed to detect the corresponding text, question related text would be picked up from the dictionary (e.g. digits for "what number") as the answer. Although this heavy model achieves top performance on the TC task, its CLC score is even lower than the baseline QA R-CNN. Such a scenario suggests that the evaluation protocol used in the current conventional VQA task is not reasonable to some extent, because the VQA models can easily overfit to the answer space by using tricks. Therefore, we introduce a *reasonable score* Δ_r to measure the percentage of answers with sufficient evidence, and it is denoted as $\Delta_r = \frac{\mathrm{CLC}_{all}}{\mathrm{TC}_{all}}$. Lower Δ_r means that the model has outputted many unreasonable but correct answers, which suggests that it might either overfit to the answer pool or use too many manual rules to achieve a higher score under conventional evaluation protocol. As shown in Table 3, the QA R-CNN w/ tricks obtained the lowest reasonable score although it outperforms all other models under the traditional evaluation protocol. Another interesting observation is that all methods achieve extremely low accuracy on the questions that have a longer answer. We believe this is because current models cannot combine multiple texts to generate a long answer. However, how to solve this issue is out of the scope of this paper, and thus we leave it for future work.

Qualitative Results: Fig. 10 illustrates some selected visualization results of the baseline methods. Surprisingly, we found that some models do not learn the concept of question type at all. For example, the 'P+LV' model outputs a word 'caffe' for the question 'What is the room number?' that asks for a number, and 'L+LV' predicts a character '长' (long) for the question '这里是河南中路多少门牌号' (What is the house number of this shop here in Henan Middle Road?) that is also asking a number. Furthermore, incorrect recognition results will cause the models to output incorrect answers. Based on the first sample of Fig. 10, although the bounding box of the answer '708' was predicted correctly, it was however recognized as '8' and was further outputted as the answer. An interesting case is the 'L+LV' model answers the question 'When was this photo uploaded?' with '29/08/2012' when only '2012' appeared in the original image. Such a phenomenon tells us that similar answers in the vocabulary could interfere with the decision of classifier. Another noteworthy example is that 'P+SV' model predicts 'snowbird' for the question 'When was this photo uploaded'. We queried another image with the answer 'snowbird' in the training set (see the last image in Fig. 10) and it shows that the 'P+SV' model outputs the same answer when the image contains similar visual features. Therefore, we believe that this VQA model might rely too heavily on the image feature and learned to map the image feature with the answer space but it does not truly understand the question. Additionally, for the question that requires stronger reasoning ability and image with many texts, such as the third sample in Fig. 10, '伟业水电安装的联系人是谁? (Who is the contact person for Weiye Hydropower Installation?)', none of the models are able to predict the answer correctly.

4. Conclusion

We have introduced a new bilingual scene text+evidence VQA dataset named EST-VQA that is annotated with both English and Chinese QA pairs. Three related challenges are proposed, namely *Cross Language*, *Localization* and *Traditional* that are designed to evaluate the generalization of VQA models. An evidence-based measure of an algorithm's capacity to reason is also proposed that requires the VQA model to provide a bounding box for the predicted answer. This metric aims to uncover whether the VQA model learns deeper relationships between text and image content, rather than overfitting to a pre-defined dictionary. Future work includes extension of the proposed EvE metric to existing VQA datasets in the hope that it might improve generalization and thus the practicality of VQA technologies.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4971–4980, 2018. 1, 2

[2] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3674–3683, 2018. 1

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 2425–2433, 2015. 1, 2

[4] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusiñol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. Scene text visual question answering. *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 2, 3, 4, 5

[5] Chee Kheng Ch'ng, Chee Seng Chan, and Chenglin Liu. Total-text: Towards orientation robustness in scene text detection. *Int. J. Doc. Anal. Recognit.*, 23:31–52, 2020. 3, 4

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 248–255. Ieee, 2009. 3

[7] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 6904–6913, 2017. 1, 2, 4

[8] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 3608–3617, 2018. 3

[9] Brian Kenji Iwana, Syed Tahseen Raza Rizvi, Sheraz Ahmed, Andreas Dengel, and Seiichi Uchida. Judging a book by its cover. *arXiv preprint arXiv:1610.09204*, 2016. 3

[10] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, 2017. 1

[11] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. *European Chapter of the Association for Computational Linguistics*, 2016. 7

[12] Kushal Kafle and Christopher Kanan. Visual question answering: Datasets, algorithms, and future challenges. *Comput. Vis. Image Underst.*, 163:3–20, 2017. 2, 3

[13] Dimosthenis Karatzas, Lluis Gomez-Bigorda, Anguelos Nicolaou, Suman Ghosh, Andrew Bagdanov, Masakazu Iwamura, Jiri Matas, Lukas Neumann, Vijay Ramaseshan Chandrasekhar, Shijian Lu, et al. Icdar 2015 competition on robust reading. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1156–1160. IEEE, 2015. 3, 4

[14] Dimosthenis Karatzas, Faisal Shafait, Seiichi Uchida, Masakazu Iwamura, Lluis Gomez i Bigorda, Sergi Robles Mestre, Joan Mas, David Fernandez Mota, Jon Almazan Almazan, and Lluis Pere De Las Heras. Icdar 2013 robust reading competition. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, pages 1484–1493. IEEE, 2013. 3, 4

[15] Aniruddha Kembhavi, Minjoon Seo, Dustin Schwenk, Jonghyun Choi, Ali Farhadi, and Hannaneh Hajishirzi. Are you smarter than a sixth grader? textbook question answering for multimodal machine comprehension. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4999–5007, 2017. 2

[16] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github.com/openimages*, 2017. 3

[17] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 3

[18] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710, 1966. 4

[19] Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. Analogical reasoning on chinese morphological and semantic relations. *Proc. Annu. Meet. Assoc. Comput. Linguist.*, 2018. 7

[20] Yuliang Liu, Lianwen Jin, Shuaitao Zhang, Canjie Luo, and Sheng Zhang. Curved scene text detection via transverse and longitudinal sequence connection. *Pattern Recogn.*, 90:337–345, 2019. 3, 4

[21] Yuliang Liu, Sheng Zhang, Lianwen Jin, Lele Xie, Yaqiang Wu, and Zhepeng Wang. Omnidirectional scene text detection with sequential-free box discretization. *Proc. Int. Joint Conf. Artificial Intell.*, 2019. 6, 7

[22] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1682–1690, 2014. 2

[23] Anand Mishra, Karteek Alahari, and CV Jawahar. Image retrieval using textual cues. In *Proc. IEEE Int. Conf. Comp. Vis.*, pages 3040–3047, 2013. 3

[24] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *Proc. Int. Conf. Doc. Anal. and Recognit.*, 2019. 2, 3, 5

[25] Nibal Nayef, Yash Patel, Michal Busta, Pinaki Nath Chowdhury, Dimosthenis Karatzas, Wafa Khlif, Jiri Matas, Umapada Pal, Jean-Christophe Burie, Cheng-lin Liu, et al. Icdar2019 robust reading challenge on multi-lingual scene text detection and recognition–rrc-mlt-2019. *arXiv preprint arXiv:1907.00945*, 2019. 3, 4

[26] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proc. Conf. Empir. Methods in Natural Language Process.*, pages 1532–1543, 2014. 7

[27] Baoguang Shi, Xiang Bai, and Cong Yao. An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 39(11):2298–2304, 2016. 6, 7

[28] Amanpreet Singh, Vivek Natarajan, Yu Jiang, Xinlei Chen, Meet Shah, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia-a platform for vision & language research. In *SysML Workshop, Advances in Neural Inf. Process. Syst.*, 2018. 6, 7

[29] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 8317–8326, 2019. 2, 3, 5, 6, 7

[30] Junyi Sun. 'Jieba'. *https://github. com/fxsjy/jieba*, 2012. 5

[31] Yipeng Sun, Jiaming Liu, Wei Liu, Junyu Han, Errui Ding, and Jingtuo Liu. Chinese street view text: Large-scale chinese text reading with partially supervised learning. *Proc. IEEE Int. Conf. Comp. Vis.*, 2019. 3, 4

[32] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 3, 4

[33] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton van den Hengel. FVQA: Fact-based visual question answering. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(10):2413–2427, Oct. 2018. 2

[34] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *Comput. Vis. Image Underst.*, 163:21–40, 2017. 2, 3

[35] 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 *Proc. IEEE Conf. Comp. Vis. Patt. Recogn.*, pages 4622–4630, 2016. 2

[36] Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Proc. Advances in Neural Inf. Process. Syst.*, pages 1031–1042, 2018. 1