Scene Text Visual Question Answering

Ali Furkan Bieten*,1 Rubèn Tito*,1 Andres Mafla*,1 Lluis Gomez1 Marçal Rusiñol1 Ernest Valveny1 C.V. Jawahar2 Dimosthenis Karatzas1

1Computer Vision Center, UAB, Spain 2CVIT, IIIT Hyderabad, India
{abiten, rperez, amafla, lgomez, marcal, dimos}@cvc.uab.es

Abstract

Current visual question answering datasets do not consider the rich semantic information conveyed by text within an image. In this work, we present a new dataset, ST-VQA, that aims to highlight the importance of exploiting high-level semantic information present in images as textual cues in the Visual Question Answering process. We use this dataset to define a series of tasks of increasing difficulty for which reading the scene text in the context provided by the visual information is necessary to reason and generate an appropriate answer. We propose a new evaluation metric for these tasks to account both for reasoning errors as well as shortcomings of the text recognition module. In addition we put forward a series of baseline methods, which provide further insight to the newly released dataset, and set the scene for further research.

1. Introduction

Textual content in man-made environments conveys important high-level semantic information that is explicit and not available in any other form in the scene. Interpreting written information in man-made environments is essential in order to perform most everyday tasks like making a purchase, using public transportation, finding a place in the city, getting an appointment, or checking whether a store is open or not, to mention just a few.

Text is present in about 50% of the images in large-scale datasets such as MS Common Objects in Context [53] and the percentage goes up sharply in urban environments. It is thus fundamental to design models that take advantage of these explicit cues. Ensuring that scene text is properly accounted for is not a marginal research problem, but quite central for holistic scene interpretation models.

The research community on reading systems has made significant advances over the past decade [26, 15]. The current state of the art in scene text understanding allows endowing computer vision systems with basic reading capacity, although the community has not yet exploited this towards solving higher level problems.

At the same time, current Visual Question Answering (VQA) datasets and models present serious limitations as a result of ignoring scene text content, with disappointing results on questions that require scene text understanding. We therefore consider it timely to bring together these two research lines in the VQA domain. To move towards more human like reasoning, we contemplate that grounding question answering both on the visual and the textual in-
formation is necessary. Integrating the textual modality in existing VQA pipelines is not trivial. On one hand, spotting relevant textual information in the scene requires performing complex reasoning about positions, colors, objects and semantics, to localise, recognise and eventually interpret the recognised text in the context of the visual content, or any other contextual information available. On the other hand, current VQA models work mostly on the principle of classical [44] and operand (instrumental) conditioning [51]. Such models, display important dataset biases [23] as well as failures in counting [9, 1], comparing and identifying attributes. These limitations make current models unsuitable to directly integrate scene text information which is often orthogonal and uncorrelated to the visual statistics of the image.

To this end, in this work we propose a new dataset, called Scene Text Visual Question Answering (ST-VQA) where the questions and answers are attained in a way that questions can only be answered based on the text present in the image. We consciously draw the majority (85.5%) of ST-VQA images from datasets that have generic question/answer pairs that can be combined with ST-VQA to establish a more generic, holistic VQA task. Some sample images and questions from the collected dataset are shown in Figure 1.

Additionally, we introduce three tasks of increasing difficulty that simulate different degrees of availability of contextual information. Finally, we define a new evaluation metric to better discern the models’ answering ability, that employs the Levenshtein distance [34] to account both for reasoning errors as well as shortcomings of the text recognition subsystem [15]. The dataset, as well as performance evaluation scripts and an online evaluation service are available through the ST-VQA Web portal.

2. Related Work

The task of text detection and recognition in natural images sets the starting point for a generalized VQA system that can integrate textual cues towards complete scene understanding. The most common approach in the reading systems community consists of two steps, text detection and recognition. Several works have been proposed addressing text detection such as [36, 35, 60, 21] which are mostly based on Fully Convolutional Neural Networks.

Text recognition methods such as the one presented in [22] propose recognizing text at the word level as a classification problem (word spotting) from a 90K English words vocabulary. Approaches that use Connectionist Temporal Classification have also been widely used in scene text recognition, in works such as [47, 7, 57, 12, 38], among others. Later works focus towards end-to-end architectures such as the ones presented by [8, 39, 20], which mostly consist of an initial Convolutional Neural Network (CNN) that acts as an encoder and a Long Short Term Memory (LSTM) combined with attention that acts as the decoder.

Visual Question Answering (VQA) aims to come up with an answer to a given natural language question about the image. Since its introduction, VQA has received a lot of attention from the Computer Vision community [4, 11, 46, 16, 23, 2] facilitated by access to large-scale datasets that allow the training of VQA models [4, 16, 33, 58, 52, 40]. Despite VQA’s popularity, none of the existing datasets except TextVQA (reviewed separately next) consider textual content, while in our work, exploiting textual information found in the images is the only way to solve the VQA task.

Related to the task proposed in this paper, are the recent works of Kafle et al. [24] and Kahou et al. [25] on question answering for bar charts and diagrams, the work of Kembhavi et al. [29] on textbook question answering. The Textbook Question Answering (TQA) dataset [29] aims at answering multimodal questions given a context of text, diagrams and images, but textual information is provided in computer readable format. This is not the case for the diagrams and charts of the datasets proposed in [24, 25], meaning that models require some sort of text recognition to solve such QA tasks. However, the text found on these datasets is rendered in standard font types and with good quality, and thus represents a less challenging setup than the scene text used in our work.

TextVQA [50] is a concurrent work to the one presented here. Similarly to ST-VQA, TextVQA proposes an alternative dataset for VQA which requires reading and reasoning about scene text. Additionally, [50] also introduces a novel architecture that combines a standard VQA model [49] and an independently trained OCR module [7] with a “copy” mechanism, inspired by pointer networks [54, 17], which allows to use OCR recognized words as predicted answers if needed. Both TextVQA and ST-VQA datasets are conceptually similar, although there are important differences in the implementation and design choices. We offer here a high-level summary of key differences, while section 3.2 gives a quantitative comparison between the two datasets.

In the case of ST-VQA, a number of different source image datasets were used, including scene text understanding ones, while in the case of TextVQA all images come from a single source, the Open Images dataset. To select the images to annotate for the ST-VQA, we explicitly required a minimum amount of two text instances to be present, while in TextVQA images were sampled on a category basis, emphasizing categories that are expected to contain text. In terms of the questions provided, ST-VQA focuses on questions that can be answered unambiguously directly using part of the image text as answer, while in TextVQA any question requiring reading the image text is allowed.

1 https://rrc.cvc.uab.es/?ch=11
Despite the differences, the two datasets are highly complementary as the image sources used do not intersect with each other, creating an opportunity for transfer learning between the two datasets and maybe combining data for training models with greater generalization capabilities.

## 3. ST-VQA Dataset

### 3.1. Data Collection

In this section we describe the process for collecting images, questions and answers for the ST-VQA dataset, and offer an in-depth analysis of the collected data. Subsequently, we detail the proposed tasks and introduce the evaluation metric.

**Images:** The ST-VQA dataset comprises 23,038 images sourced from a combination of public datasets that include both scene text understanding datasets as well as generic computer vision ones. In total, we used six different datasets, namely: ICDAR 2013[27] and ICDAR2015[26], ImageNet [10], VizWiz[18], IIIT Scene Text Retrieval[42], Visual Genome [33] and COCO-Text [53]. A key benefit of combining images from various datasets is the reduction of dataset bias such as selection, capture and negative set bias which have been shown to exist in popular image datasets[30]. Consequently, the combination of datasets results in a greater variability of questions. To automatically select images to define questions and answers, we use an end-to-end single shot text retrieval architecture [13]. We automatically select all images that contain at least 2 text instances thus ensuring that the proposed questions contain at least 2 possible options as an answer. The final number of images and questions per dataset can be found in Table 1.

<table>
<thead>
<tr>
<th>Original Dataset</th>
<th>Images</th>
<th>Questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coco-text</td>
<td>7,520</td>
<td>10,854</td>
</tr>
<tr>
<td>Visual Genome</td>
<td>8,490</td>
<td>11,195</td>
</tr>
<tr>
<td>VizWiz</td>
<td>835</td>
<td>1,303</td>
</tr>
<tr>
<td>ICDAR</td>
<td>1,088</td>
<td>1,423</td>
</tr>
<tr>
<td>ImageNet</td>
<td>3,680</td>
<td>5,165</td>
</tr>
<tr>
<td>IIIT-STR</td>
<td>1,425</td>
<td>1,890</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>23,038</td>
<td>31,791</td>
</tr>
</tbody>
</table>

Table 1. Number of images and questions gathered per dataset.

**Question and Answers:** The ST-VQA dataset comprises 31,791 questions. To gather the questions and answers of our dataset, we used the crowd-sourcing platform Amazon Mechanical Turk (AMT). During the collection of questions and answers, we encouraged workers to come up with closed-ended questions that can be unambiguously answered with text found in the image, prohibiting them to ask yes/no questions or questions that can be answered only based on the visual information.

The process of collecting question and answer pairs consisted of two steps. First, the workers were given an image along with instructions asking them to come up with a question that can be answered using the text found in the image. The workers were asked to write up to three question and answer pairs. Then, as a verification step, we perform a second AMT task that consisted of providing different workers with the image and asking them to respond to the previously defined question. We filtered the questions for which we did not obtain the same answer in both steps, in order to remove ambiguous questions. The ambiguous questions were checked by the authors and corrected if necessary, before being added to the dataset. In some cases both answers were deemed correct and accepted, therefore ST-VQA questions have up to two different valid answers.

In total, the proposed ST-VQA dataset comprises 23,038 images with 31,791 questions/answers pair separated into 19,027 images - 26,308 questions for training and 2,993 images - 4,163 questions for testing. We present examples of question and answers of our dataset in Figure 1.

### 3.2. Analysis and Comparison with TextVQA

In Figure 2 we provide the length distribution for the gathered questions and answers of the ST-VQA datasets, in comparison to the recent TextVQA dataset. It can be observed that the length statistics of the two datasets are closely related.

To further explore the statistics of our dataset, Figure 3 visualises how the ST-VQA questions are formed. As it can be appreciated, our questions start with “What, Where, Which, How and Who”. A considerable percentage starts with “What” questions, as expected given the nature of the task. A critical point to realize however, is that the questions are not explicitly asking for specific text that appears in the scene; rather they are formulated in a way that requires to have certain prior world knowledge/experience. For example, some of the “what” questions inquire about a brand, website, name, bus number, etc., which require some explicit knowledge about what a brand or website is.

There has been a lot of effort to deal with the language
prior inside the datasets [16, 23, 59]. One of the reasons for having language priors in datasets is the uneven distribution of answers in the dataset. In VQA v1 [4], since the dataset is formed from the images of MSCOCO [37], the answers to the question of “what sport ...” are tennis and baseball over 50%. Another example is the question “is there ...”, having yes as an answer in over 70% of the cases. As can be seen from Figure 4, our dataset apart from the “sign” and “year” questions follows a uniform distribution for the answers, reducing the risk of language priors while having a big vocabulary for the answers.

Figure 3. Distribution of questions in the ST-VQA train set by their starting 4-grams (ordered from center to outwards). Words with a small contribution are not shown for better visualization.

To put ST-VQA in perspective, VQA 2.0 [16], the biggest dataset in the community, contains 1.1 million questions out of which only 8k, corresponding to less than 1% of the total questions, requires reading the text in the image. The TextVQA [50] dataset on the other hand comprises 28, 408 images paired with 45, 336 questions.

As a result of the different collection procedures followed, all ST-VQA questions can be answered unambiguously directly using the text in the image, while in the case of TextVQA reportedly 39% (18k) of the answers do not contain any of the OCR tokens 2. This might be either due to the type of the questions defined, or due to shortcomings of the employed text recognition engine.

The fact that ST-VQA answers are explicitly grounded to the scene text, allows us to collect a single answer per question. To consider an answer as correct, we introduce a soft metric that requires it to have a small edit distance to the correct answer (see section 3.4), factoring this way in the evaluation procedure the performance of the text recognition sub-system. In the case of TextVQA, 10 answers are collected per question and any answer supported by at least three subjects is considered correct. In order to better understand the effects of our approach compared to collecting multiple responses like in TextVQA, we performed an experiment collecting 10 answers for a random subset of 1,000 ST-VQA questions. Our analysis showed that in 84.1% of the cases there is agreement between the majority of subjects and the original answer. The same metric for TextVQA is 80.3%, confirming that defining a single unambiguous answer results in similarly low ambiguity at evaluation time.

3.3. Tasks

We define 3 novel tasks, suitable for the ST-VQA dataset, namely “strongly contextualised”, “weakly contextualised” and “open vocabulary”.

The proposed differentiation of tasks can be interpreted by how humans make use of prior knowledge to argue about their current situation. Such prior knowledge in ST-VQA is provided as a dictionary, different for each task. Similar approaches using dynamic per-image dictionaries have been used for DVQA in [24] and for scene text understanding in [26]. Our formulation of the tasks is inspired by the previous notions and the difficulty per task increases gradually. In the strongly contextualised task we capture this prior knowledge by creating a dictionary per image for the specific scenario depicted. In the weakly contextualised task we provide a single dictionary comprising all the words in the answers of the dataset. Finally, for the open dictionary task, we treat the problem as tabula rasa where no a priori and no external information is available to the model.

For the strongly contextualised task (1), following the standard practice used for end-to-end word spotting [27, 26, 55], we create a dictionary per image that contains the words that appear in the answers defined for questions on that image, along with a series of distractors. The distractors are generated in two ways. On one hand, they comprise instances of scene text as returned by a text recogniser applied on the image. On the other hand, they comprise words obtained by exploiting the semantic understanding of the scene, in the form of the output of a dynamic lexicon generation model [43, 14]. The dictionary for the strongly contextualised task is 100 words long and defined per image.

In the weakly contextualised task (2), we provide a unique dictionary of 30,000 words for all the datasets’ images which is formed by collecting all the 22k ground truth words plus 8k distractors generated in the same way as in the previous task. Finally for the open dictionary task (3), we provide no extra information thus we can consider it as

2Presentation of the TextVQA Challenge, CVPR 2019
an open-lexicon task.

By proposing the aforementioned tasks the VQA problem is conceived in a novel manner that has certain advantages. First, it paves the way for research on automatically processing and generating such prior information, and its effect on the model design and performance. Second, it provides an interesting training ground for end-to-end reading systems, where the provided dictionaries can be used to prime text spotting methods.

3.4. Evaluation and Open Challenge

Since the answers of our dataset are contained within the text found in the image, which is dependent on the accuracy of the OCR being employed, the classical evaluation metric of VQA tasks is not optimum for our dataset, e.g. if the model reasons properly about the answer but makes a mistake of a few characters in the recognition stage, like in Figure 6 (first row, third column), the typical accuracy score would be 0. However, the metric we propose named Average Normalized Levenshtein Similarity (ANLS) would give an intermediate score between 0.5 and 1 that will softly penalise the OCR mistakes. Thus, a motivation of defining a metric that captures OCR accuracy as well as model reasoning is evident. To this end, in all 3 tasks we use the normalized Levenshtein similarity [34] as an evaluation metric. More formally, we define ANLS as follows:

\[
\text{ANLS} = \frac{1}{N} \sum_{i=0}^{N} \left( \max_j s(a_{ij}, o_{qi}) \right)
\]

\[
s(a_{ij}, o_{qi}) = \begin{cases} 
(1 - NL(a_{ij}, o_{qi})) & \text{if } NL(a_{ij}, o_{qi}) < \tau \\
0 & \text{if } NL(a_{ij}, o_{qi}) \geq \tau 
\end{cases}
\]

where \(N\) is the total number of questions in the dataset, \(M\) is the total number of GT answers per question, \(a_{ij}\) are the ground truth answers where \(i = \{0, ..., N\}\), and \(j = \{0, ..., M\}\), and \(o_{qi}\) is the network’s answer for the \(i^{th}\) question \(q_i\). \(NL(a_{ij}, o_{qi})\) is the normalized Levenshtein distance between the strings \(a_{ij}\) and \(o_{qi}\) (notice that the normalized Levenshtein distance is a value between 0 and 1). We define a threshold \(\tau = 0.5\) that penalizes metrics larger than this value, thus the final score will be 0 if \(NL\) is larger than \(\tau\). The intuition behind the threshold is that if an output has an edit distance of more than 0.5 to an answer, meaning getting half of the answer wrong, we reason that the output is the wrong text selected from the options as an answer. Otherwise, the metric has a smooth response that can gracefully capture errors in text recognition.

In addition, we provide an online service where the open challenge was hosted [5], that researchers can use to evaluate their methods against a public validation/test dataset.

4. Baselines and Results

The following section describes the baselines employed in this work as well as an analysis of the results obtained in the experiments conducted. The proposed baselines help us to showcase the difficulty of the proposed dataset and its tasks. Aside from baselines designed to exploit all the information available (visual information, scene text, and the question), we have purposely included baselines that ignore one or more of the available pieces of information in order to establish lower bounds of performance. The following baselines are employed to evaluate the datasets:

Random: As a way of assessing aimless chance, we return a random word from the dictionary provided for each task (see section 3.3 for more detail).

Scene Text Retrieval: This baseline leverages a single shot CNN architecture [13] that predicts at the same time bounding boxes and a Pyramidal Histogram Of Characters (PHOC) [3]. The PHOC is a compact representation of a word that considers the spatial location of each character to construct the resulting encoding. This baseline ignores the question and any other visual information of the image.

We have defined two approaches: the first (“STR retrieval”) uses the specific task dictionaries as queries to a given image, and the top-1 retrieved word is returned as the
answer; the second one (“STR bbox”), follows the intuition that humans tend to formulate questions about the largest text instance in the image. We take the text representation from the biggest bounding box found and then find the nearest neighbor word in the corresponding dictionaries.

**Scene Image OCR:** A state of the art text recognition model [20] is used to process the test set images. The detected text is ranked according to the confidence score and the closest match between the most confident text detection and the provided vocabularies for task 1 and task 2 is used as the answer. In task 3 the most confident text detection is adopted as the answer directly.

**Standard VQA models:** We evaluate two standard VQA models. The first one, named “Show, Ask, Attend and Answer” [28] (SAAA), consists of a CNN-LSTM architecture. On one hand, a ResNet-152 [19] is used to extract image features with dimension $14 \times 14 \times 2048$, while the question is tokenized and embedded by using a multi-layer LSTM. On top of the combination of image features and the question embedding, multiple attention maps (glimpses) are obtained. The result of the attention glimpses over the image features and the last state of the LSTM is concatenated and fed into two fully connected layers to obtain the distribution of answer probabilities according to the classes. We optimize the model with the Adam optimizer [31] with a batch size of 128 for 30 epochs. The starting learning rate is 0.001 which decays by half every 50K iterations.

The second model, named “Stacked Attention Networks” [56] (SAN), uses a pre-trained VGNN [48] CNN to obtain image features with shape $14 \times 14 \times 512$. Two question encoding methods are proposed, one that uses an LSTM and another that uses a CNN, both of them yielding similar results according to the evaluated dataset. The encoded question either by a CNN or LSTM is used along with the image features to compute two attention maps, which later are used with the image features to output a classification vector. We optimize the model with a batch size of 100 for 150 epochs. The optimizer used is RMSProp with a starting learning rate of 0.0003 and a decay value of 0.9999.

Overall, three different experiments are proposed according to the output classification vector. The first, is formed by selecting the most common 1k answer strings in the ST-VQA training set as in [4]. For the second one, we selected the 5k most common answers so that we can see the effect of a gradual increase of the output vector in the two VQA models. In the third one, all the answers found in the training set are used (19,296) to replicate the wide range vocabulary of scene-text images and to capture all the answers found in the training set.

**Fusing Modalities - Standard VQA Models + Scene Text Retrieval:** Using the previously described VQA models, the purpose of this baseline is to combine textual features obtained from a scene text retrieval model with existing VQA pipelines. To achieve this, we use the model from [13] and we employ the output tensor before the non-maximal suppression step (NMS) is performed. The most confident PHOC predictions above a threshold are selected relative to a single grid cell. The selected features form a tensor of size $14 \times 14 \times 609$, which is concatenated with the image features before the attention maps are calculated on both previously described VQA baselines. Afterwards the attended features are used to output a probability distribution over the classification vector. The models are optimized using the same strategy described before.

### 4.1. Results

The results of all provided baselines according to the defined tasks are summarized in Table 2. As a way to compare the proposed Average Normalized Levenshtein Similarity (ANLS) metric, we also calculate the accuracy for each baseline. The accuracy is calculated by counting the exact matches between the model predictions and collected answers as is the standard practice in the VQA literature.

The last column in Table 2, upper bound, shows the maximum possible score that can be achieved depending on the method evaluated. The upper bound accuracy for standard VQA models is the percentage of questions where the correct answer is part of the models output vocabulary, while the upper bound ANLS is calculated by taking as answer the closest word (output class) in terms of Levenshtein distance to the correct answer. In the case of the Scene Text Retrieval (STR retrieval) [13] model the upper bound is calculated by assuming that the correct answer is a single word and that this word is retrieved by the model as the top-1 among all the words in the provided vocabularies.

In Table 2 we appreciate that standard VQA models that disregard textual information from the image achieve similar scores, ranging between 0.085 to 0.102 ANLS, or 6.36% to 7.78% accuracy. One relevant point is that although in VQA v1 [4] the SAAA [28] model is known to outperform SAN [56], in our dataset the effect found is the opposite, due to the fact that our dataset and task outline is different in its nature compared to VQA v1.

Another important point is that the SAAA model increases both its accuracy and ANLS score when using a larger classification vector size, from 1k to 5k classes; however, going from 5k to 19k classes the results are worse, suggesting that learning such a big vocabulary in a classification manner is not feasible.

It is worth noting that the proposed ANLS metric generally tracks accuracy, which indicates broad compatibility between the metrics. But, in addition, ANLS can deal with border cases (i.e. correct intended responses, but slightly wrong recognized text) where accuracy, being a hard metric based on exact matches, cannot. Such border cases are frequent due to errors at the text recognition stage. Example...
Table 2. Baseline results comparison on the three tasks of ST-VQA dataset. We provide Average Normalized Levenshtein similarity (ANLS) and Accuracy for different methods that leverage OCR, Question (Q) and Visual (V) information.

<table>
<thead>
<tr>
<th>Method with</th>
<th>OCR</th>
<th>Q</th>
<th>V</th>
<th>Task 1</th>
<th>Task 2</th>
<th>Task 3</th>
<th>Upper bound</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ANLS</td>
<td>Acc.</td>
<td>ANLS</td>
<td>Acc.</td>
<td>ANLS</td>
<td>Acc.</td>
<td>ANLS</td>
</tr>
<tr>
<td>Random</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
<td>0.015</td>
<td>0.96</td>
<td>0.001</td>
<td>0.00</td>
</tr>
<tr>
<td>STR [13] (retrieval)</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>0.171</td>
<td>13.78</td>
<td>0.073</td>
<td>5.58</td>
</tr>
<tr>
<td>STR [13] (bbox)</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>0.130</td>
<td>7.32</td>
<td>0.073</td>
<td>5.55</td>
</tr>
<tr>
<td>Scene Image OCR [20]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>0.145</td>
<td>8.89</td>
<td>0.096</td>
<td>7.41</td>
</tr>
<tr>
<td>SAAA [28] (1k cls)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>0.085</td>
<td>6.36</td>
<td>0.085</td>
<td>6.36</td>
</tr>
<tr>
<td>SAAA+STR (1k cls)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.091</td>
<td>6.66</td>
<td>0.091</td>
<td>6.66</td>
</tr>
<tr>
<td>SAAA [28] (5k cls)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>0.096</td>
<td>7.41</td>
<td>0.096</td>
<td>7.41</td>
</tr>
<tr>
<td>SAAA+STR (5k cls)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.084</td>
<td>6.13</td>
<td>0.084</td>
<td>6.13</td>
</tr>
<tr>
<td>SAAA [28] (19k cls)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>0.087</td>
<td>6.36</td>
<td>0.087</td>
<td>6.36</td>
</tr>
<tr>
<td>SAAA+STR (19k cls)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.069</td>
<td>4.65</td>
<td>0.069</td>
<td>4.65</td>
</tr>
<tr>
<td>QA+STR (19k cls)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.102</td>
<td>7.78</td>
<td>0.102</td>
<td>7.78</td>
</tr>
<tr>
<td>SAN(LSTM) [56] (5k cls)</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
<td>0.136</td>
<td>10.34</td>
<td>0.136</td>
<td>10.34</td>
</tr>
<tr>
<td>SAN(LSTM)+STR (5k cls)</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.135</td>
<td>10.46</td>
<td>0.135</td>
<td>10.46</td>
</tr>
</tbody>
</table>

Figure 5. Results of baseline methods in the open vocabulary task of ST-VQA by question type.

Examples of such behaviour can be seen in the qualitative results shown in Figure 6 for some of the answers (indicated in orange color). This also explains why the “Scene Image OCR” model is better ranked in terms of ANLS than of accuracy in Table 2.

Finally, we notice that standard VQA models, disregarding any textual information, perform worse or comparable at best to the “STR (retrieval)” or “Scene Image OCR” models, despite the fact that these heuristic methods do not take into account the question. This observation confirms the necessity of leveraging textual information as a way to improve performance in VQA models. We demonstrate this effect by slightly improving the results of VQA models (SAAA and SAN) by using a combination of visual features and PHOC-based textual features (see SAAA+STR and SAN+STR baselines descriptions for details).

For further analysis of the baseline models’ outputs and comparison between them, we provide in Figure 5 two bar charts with specific results on different question types. In most of them the STR model is better than the “Scene Image OCR” (ST-OCR) in terms of ANLS. The effect of PHOC embedding is especially visible on the SAN model for correctly answering the question type such as “what year”, “what company” and “which”. Also, none of the models is capable of answering the questions regarding license plates, “who” and “what number”. This is an inherent limitation of models treating VQA as a pure classification problem, as they can not deal with out of vocabulary answers. In this regard the importance of using PHOC features lies in their ability to capture the morphology of words rather than their semantics as in other text embeddings [41, 45, 6]; since several text instances and answers in the dataset may not have any representation in a pre-trained semantic model. The use of a morphological embedding like PHOC can provide...
Q: What brand are the machines?
A: bongard
SAN(CNN)+STR: ray
SAAA+STR: ray
Scene Image OCR: zbongard
STR (bbox): 1

Q: Where is the high court located?
A: delhi
SAN(CNN)+STR: delhi
SAAA+STR: esplanade
Scene Image OCR: geno
STR (bbox): genoa

Q: What does the black label say?
A: GemOro
SAN(CNN)+STR: st. george cl.
SAAA+STR: esplanade
Scene Image OCR: gemors
STR (bbox): genoa

Q: What’s the street name?
A: place d’armes
SAN(CNN)+STR: 10th st
SAAA+STR: ramistrasse
Scene Image OCR: d’armes
STR (bbox): dames

Q: What is the route of the bus?
A: purple route
SAN(CNN)+STR: 66
SAAA+STR: 508
Scene Image OCR: 1208
STR (bbox): purple

Q: What is the automobile sponsor of the event?
A: kia
SAN(CNN)+STR: kia
SAAA+STR: kia
Scene Image OCR: kin
STR (bbox): kin

Q: Which dessert is showcased?
A: donut
SAN(CNN)+STR: l
SAAA+STR: donuts
Scene Image OCR: 175
STR (bbox): north

Q: What is preheat oven temperature?
A: 350
SAN(CNN)+STR: 350
SAAA+STR: 0
Scene Image OCR: high
STR (bbox): receivables

Figure 6. Qualitative results for different methods on task 1 (strongly contextualised) of the ST-VQA dataset. For each image we show the question (Q), ground-truth answer (blue), and the answers provided by different methods (green: correct answer, red: incorrect answer, orange: incorrect answer in terms of accuracy but partially correct in terms of ANLS (0.5 ≤ ANLS < 1)).

5. Conclusions and Future Work

This work introduces a new and relevant dimension to the VQA domain. We presented a new dataset for Visual Question Answering, the Scene Text VQA, that aims to highlight the importance of properly exploiting the high-level semantic information present in images in the form of scene text to inform the VQA process. The dataset comprises questions and answers of high variability, and poses extremely difficult challenges for current VQA methods. We thoroughly analysed the ST-VQA dataset through performing a series of experiments with baseline methods, which established the lower performance bounds, and provided important insights. Although we demonstrate that adding textual information to generic VQA models leads to improvements, we also show that ad-hoc baselines (e.g. OCR-based, which do exploit the contextual words) can outperform them, reinforcing the need of different approaches. Existing VQA models usually address the problem as a classification task, but in the case of scene text based answers the number of possible classes is intractable. Dictionaries defined over single words are also limited. Instead, a generative pipeline such as the ones used in image captioning is required to capture multiple-word answers, and out of dictionary strings such as numbers, license plates or codes. The proposed metric, namely Average Normalized Levenshtein Similarity is better suited for generative models compared to evaluating classification performance, while at the same time, it has a smooth response to the text recognition performance.

Acknowledgments

This work has been supported by projects TIN2017-89779-P, Marie-Curie (712949 TECNIOspring PLUS), aBSINTHE (Fundacion BBVA 2017), the CERCA Programme / Generalitat de Catalunya, a European Social Fund grant (CCI: 2014ES05SFOP007), NVIDIA Corporation and PhD scholarships from AGAUR (2019-FIB01233) and the UAB.
References


