Scene Text Visual Question Answering

Ali Furkan Biten*,1 Rubèn Tito*,1 Andres Mafla*,1 Lluis Gomez[1]
Marçal Rusiñol[1] Ernest Valveny[1] C.V. Jawahar[2] Dimosthenis Karatzas[1]

[1]Computer Vision Center, UAB, Spain [2]CVIT, 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 is 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-

Q: What is the price of the bananas per kg?
A: $11.98

Q: What does the red sign say?
A: Stop

Q: Where is this train going?
A: To New York
A: New York

Q: What is the exit number on the street sign?
A: 2
A: Exit 2

Figure 1. Recognising and interpreting textual content is essential for scene understanding. In the Scene Text Visual Question Answering (ST-VQA) dataset leveraging textual information in the image is the only way to solve the QA task.

*Equal contribution.

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 operant (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[1].

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 Kise at al. [32] on QA for machine printed document images, and 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.

Original Dataset	Images	Questions
Coco-text	7,520	10,854
Visual Genome	8,490	11,195
VizWiz	835	1,303
ICDAR	1,088	1,423
ImageNet	3,680	5,165
IIIT-STR	1,425	1,890
Total	23,038	31,791

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 recently presented TextVQA. It can be observed that the length statistics of the two datasets are closely related.

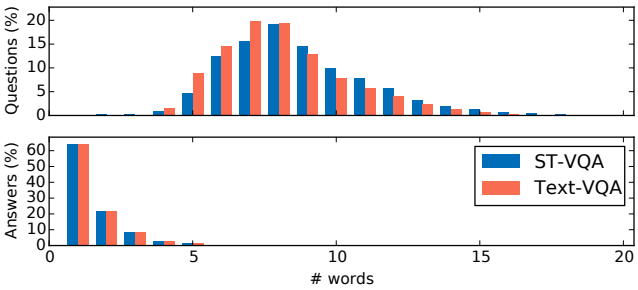


Figure 2. Percentage of questions (top) and answers (bottom) that contain a specific number of words.

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.

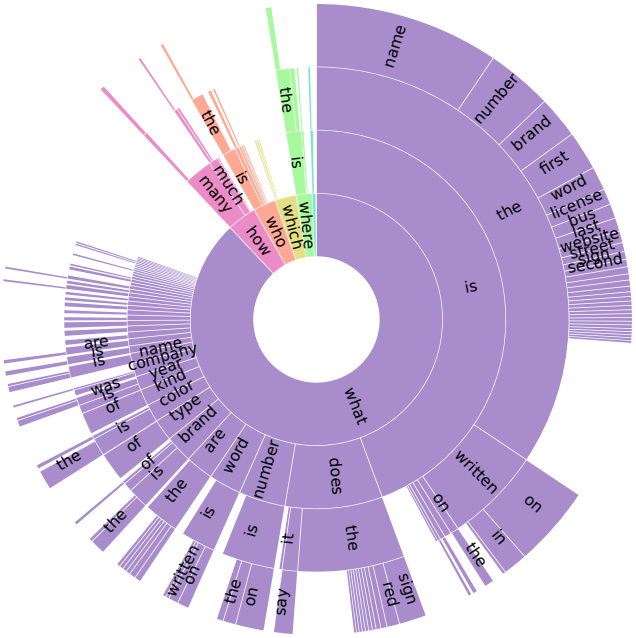


Figure 3. Distribution of questions in the ST-VQA train set by their starting 4-grams (ordered from center to outwards). Words with a small contribution are not shown for better visualization.

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

[2] Presentation of the TextVQA Challenge, CVPR 2019

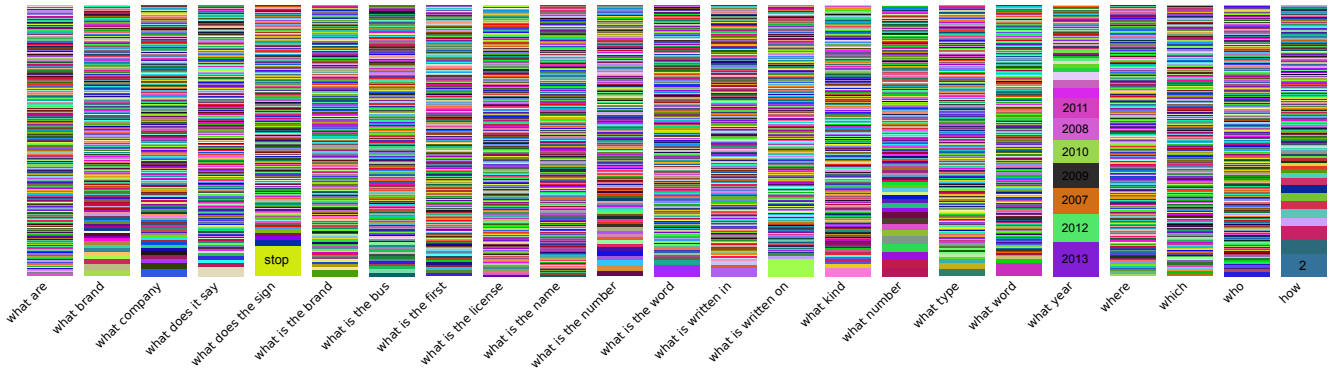


Figure 4. Distribution of answers for different types of questions in the ST-VQA train set. Each color represents a different unique answer.

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:

$$\text{ANLS} = \frac{1}{N}\sum_{i=0}^{N}\left(\max_{j} s(a_{ij}, o_{q_i})\right) \tag{1}$$

$$s(a_{ij}, o_{q_i}) = \begin{cases} (1 - NL(a_{ij}, o_{q_i})) & \text{if } NL(a_{ij}, o_{q_i}) < \tau \\ 0 & \text{if } NL(a_{ij}, o_{q_i}) \geqslant \tau \end{cases}$$

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_{q_i} is the network's answer for the i^{th} question q_i. $NL(a_{ij}, o_{q_i})$ is the normalized Levenshtein distance between the strings a_{ij} and o_{q_i} (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 the NL is larger than τ. 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 VGGN [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. Exam-

Method with	OCR	Q	V	Task 1 ANLS	Task 1 Acc.	Task 2 ANLS	Task 2 Acc.	Task 3 ANLS	Task 3 Acc.	Upper bound ANLS	Upper bound Acc.
Random	✗	✗	✗	0.015	0.96	0.001	0.00	0.00	0.00	-	-
STR [13] (retrieval)	✓	✗	✗	**0.171**	**13.78**	0.073	5.55	-	-	0.782	68.84
STR [13] (bbox)	✓	✗	✗	0.130	7.32	0.118	6.89	0.128	7.21	-	-
Scene Image OCR [20]	✓	✗	✗	0.145	8.89	0.132	8.69	**0.140**	8.60	-	-
SAAA [28] (1k cls)	✗	✓	✓	0.085	6.36	0.085	6.36	0.085	6.36	0.571	31.96
SAAA+STR (1k cls)	✓	✓	✓	0.091	6.66	0.091	6.66	0.091	6.66	0.571	31.96
SAAA [28] (5k cls)	✗	✓	✓	0.087	6.66	0.087	6.66	0.087	6.66	0.740	41.03
SAAA+STR (5k cls)	✓	✓	✓	0.096	7.41	0.096	7.41	0.096	7.41	0.740	41.03
SAAA [28] (19k cls)	✗	✓	✓	0.084	6.13	0.084	6.13	0.084	6.13	0.862	52.31
SAAA+STR (19k cls)	✓	✓	✓	0.087	6.36	0.087	6.36	0.087	6.36	0.862	52.31
QA+STR (19k cls)	✓	✓	✗	0.069	4.65	0.069	4.65	0.069	4.65	0.862	52.31
SAN(LSTM) [56] (5k cls)	✗	✓	✓	0.102	7.78	0.102	7.78	0.102	7.78	0.740	41.03
SAN(LSTM)+STR (5k cls)	✓	✓	✓	0.136	10.34	**0.136**	**10.34**	0.136	**10.34**	0.740	41.03
SAN(CNN)+STR (5k cls)	✓	✓	✓	0.135	10.46	**0.135**	**10.46**	0.135	**10.46**	0.740	41.03

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.

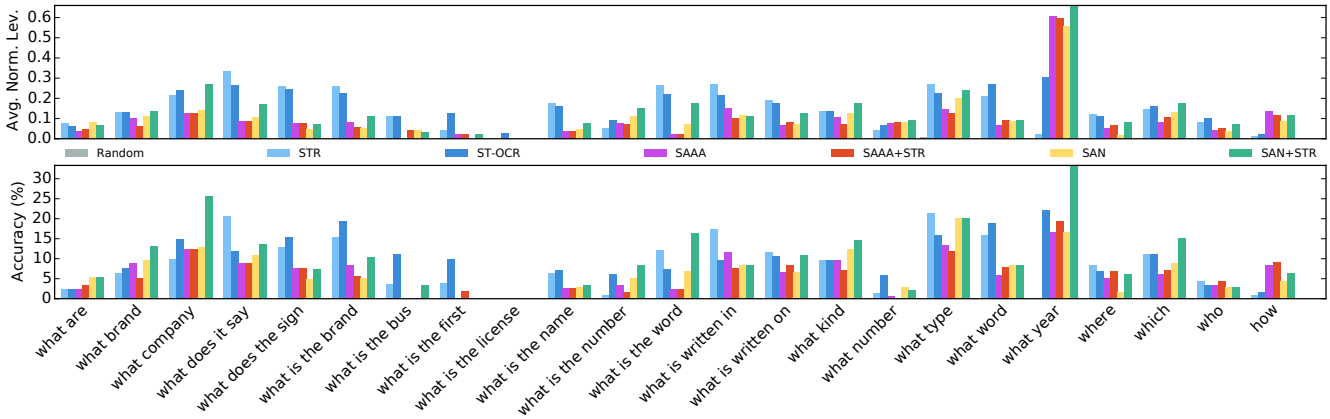


Figure 5. Results of baseline methods in the open vocabulary task of ST-VQA by question type.

ples 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

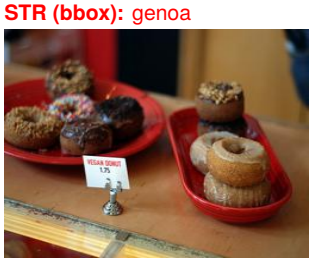
Q: Where is the high court located?
A: delhi
SAN(CNN)+STR: delhi
SAAA+STR: delhi
Scene Image OCR: high
STR (bbox): delhi

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

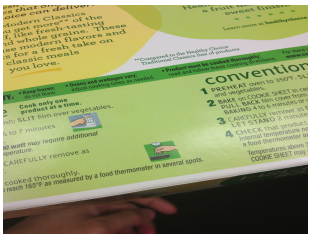
Q: What's the street name?
A: place d'armes
SAN(CNN)+STR: 10th st
SAAA+STR: ramistrasse
Scene Image OCR: d'armes
STR (bbox): dames

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): 0



Q: Which dessert is showcased?
A: donut
A: Vegan Donut
SAN(CNN)+STR: t
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 ($0.5 \leq ANLS < 1$)).

a starting point for datasets that contain text and answers in several languages and out of dictionary words such as license plates, prices, directions, names, etc.

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 as 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

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

[2] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4971–4980, 2018.

[3] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12):2552–2566, 2014.

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

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

[6] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

[7] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 71–79. ACM, 2018.

[8] Michal Busta, Lukas Neumann, and Jiri Matas. Deep textspotter: An end-to-end trainable scene text localization and recognition framework. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2204–2212, 2017.

[9] Prithvijit Chattopadhyay, Ramakrishna Vedantam, Ramprasaath R Selvaraju, Dhruv Batra, and Devi Parikh. Counting everyday objects in everyday scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1135–1144, 2017.

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009.

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

[12] Yunze Gao, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Reading scene text with attention convolutional sequence modeling. *arXiv preprint arXiv:1709.04303*, 2017.

[13] Lluís Gómez, Andrés Mafla, Marçal Rusinol, and Dimosthenis Karatzas. Single shot scene text retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 700–715, 2018.

[14] Lluis Gomez, Yash Patel, Marçal Rusiñol, Dimosthenis Karatzas, and CV Jawahar. Self-supervised learning of visual features through embedding images into text topic spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4230–4239, 2017.

[15] Raul Gomez, Baoguang Shi, Lluis Gomez, Lukas Neumann, Andreas Veit, Jiri Matas, Serge Belongie, and Dimosthenis Karatzas. Icdar2017 robust reading challenge on coco-text. In *2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 1435–1443. IEEE, 2017.

[16] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913, 2017.

[17] Caglar Gulcehre, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. Pointing the unknown words. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016*, pages 140–149. Association for Computational Linguistics (ACL), 2016.

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

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[20] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5020–5029, 2018.

[21] Wenhao He, Xu-Yao Zhang, Fei Yin, and Cheng-Lin Liu. Deep direct regression for multi-oriented scene text detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 745–753, 2017.

[22] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *International Journal of Computer Vision*, 116(1):1–20, 2016.

[23] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, and R. Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910, 2017.

[24] Kushal Kafle, Brian Price, Scott Cohen, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018.

[25] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017.

[26] 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 *2015 13th International Conference on Document Analysis and Recognition (ICDAR)*, pages 1156–1160. IEEE, 2015.

[27] 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 *2013 12th International Conference on Document Analysis and Recognition*, pages 1484–1493. IEEE, 2013.

[28] Vahid Kazemi and Ali Elqursh. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*, 2017.

[29] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4999–5007, 2017.

[30] Aditya Khosla, Tinghui Zhou, Tomasz Malisiewicz, Alexei A Efros, and Antonio Torralba. Undoing the damage of dataset bias. In *European Conference on Computer Vision*, pages 158–171. Springer, 2012.

[31] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[32] Koichi Kise, Shota Fukushima, and Keinosuke Matsumoto. Document image retrieval for QA systems based on the density distributions of successive terms. *IEICE Transactions*, 88-D, 2005.

[33] 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. *International Journal of Computer Vision*, 123(1):32–73, 2017.

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

[35] Minghui Liao, Baoguang Shi, and Xiang Bai. Textboxes++: A single-shot oriented scene text detector. *IEEE transactions on image processing*, 27(8):3676–3690, 2018.

[36] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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

[38] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5676–5685, 2018.

[39] Pengyuan Lyu, Minghui Liao, Cong Yao, Wenhao Wu, and Xiang Bai. Mask textspotter: An end-to-end trainable neural network for spotting text with arbitrary shapes. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 67–83, 2018.

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

[41] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[42] A. Mishra, K. Alahari, and C. V. Jawahar. Image retrieval using textual cues. In *ICCV*, 2013.

[43] Yash Patel, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. Dynamic lexicon generation for natural scene images. In *European Conference on Computer Vision*, pages 395–410. Springer, 2016.

[44] Ivan P Pavlov. Conditioned reflex: An investigation of the physiological activity of the cerebral cortex. 1960.

[45] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[46] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961, 2015.

[47] Baoguang Shi, Xinggang Wang, Pengyuan Lyu, Cong Yao, and Xiang Bai. Robust scene text recognition with automatic rectification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4168–4176, 2016.

[48] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[49] 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, NeurIPS*, volume 2018, 2018.

[50] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *CVPR (To appear)*, 2019.

[51] Burrhus F Skinner. Operant behavior. *American psychologist*, 18(8):503, 1963.

[52] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler.

Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016.

[53] 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.

[54] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, pages 2692–2700, 2015.

[55] Kai Wang and Serge Belongie. Word spotting in the wild. In *European Conference on Computer Vision*, pages 591–604. Springer, 2010.

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

[57] Xu-Cheng Yin, Wei-Yi Pei, Jun Zhang, and Hong-Wei Hao. Multi-orientation scene text detection with adaptive clustering. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1930–1937, 2015.

[58] Licheng Yu, Eunbyung Park, Alexander C Berg, and Tamara L Berg. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the ieee international conference on computer vision*, pages 2461–2469, 2015.

[59] Peng Zhang, Yash Goyal, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Yin and yang: Balancing and answering binary visual questions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5014–5022, 2016.

[60] Xinyu Zhou, Cong Yao, He Wen, Yuzhi Wang, Shuchang Zhou, Weiran He, and Jiajun Liang. East: an efficient and accurate scene text detector. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 5551–5560, 2017.