Pattern Recognition Letters
journal homepage: www.elsevier.com

Multimodal grid features and cell pointers for Scene Text Visual Question Answering

Lluís Gómez[a], Ali Furkan Biten[a], Rubèn Tito[a], Andrés Mafla[a], Marçal Rusiñol[a], Ernest Valveny[a], Dimosthenis Karatzas[a]

[a]*Computer Vision Center – Universitat Autònoma de Barcelona, Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola), Barcelona.*

ABSTRACT

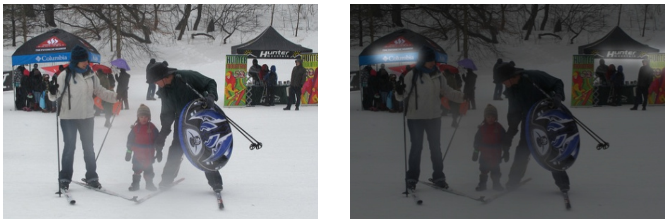
This paper presents a new model for the task of scene text visual question answering. In this task questions about a given image can only be answered by reading and understanding scene text. Current state of the art models for this task make use of a dual attention mechanism in which one attention module attends to visual features while the other attends to textual features. A possible issue with this is that it makes difficult for the model to reason jointly about both modalities. To fix this problem we propose a new model that is based on an single attention mechanism that attends to multi-modal features conditioned to the question. The output weights of this attention module over a grid of multi-modal spatial features are interpreted as the probability that a certain spatial location of the image contains the answer text to the given question. Our experiments demonstrate competitive performance in two standard datasets. In particular we outperform previous state of the art by 4% accuracy on the ST-VQA dataset. Furthermore, we also provide a novel analysis of the ST-VQA dataset based on a human performance study. Supplementary material, code, and data is made available through this link.

1. Introduction

For an intelligent agent to answer a question about an image, it needs to understand its content. Depending on the question, the visual understanding skills required will vary: object/attributes recognition, spatial reasoning, counting, comparing, use of commonsense knowledge, or a combination of any of them. Reading is another skill that can be of great use for Visual Question Answering (VQA) and has not been explored until recently by Biten et al. (2019b) and Singh et al. (2019).

Scene text VQA is the task of answering questions about an image that can only can be answered by reading/understanding scene text that is present in it. An interesting property of this task over standard VQA is that the textual modality is present both in the question and in the image representations. Thus the task calls for a different family of composed models using computer vision (CV) and natural language processing (NLP).

Current state of the art on scene text VQA, Singh et al. (2019), make use of a dual attention mechanism: one attention module that attends the image visual features conditioned to the question, and another that attends to the textual features (OCR text instances) conditioned to the question. A potential issue with this is that it makes difficult for the model to reason jointly about the two modalities, since this can only be done after the late fusion of the two modules. In this paper we propose a solution to this problem, by using a single attention module that attends to multi-modal features as shown in Figures 1 and 2.



Q: What brand name is on the tent with the blue stripe? **A:** COLUMBIA

Fig. 1. Answering scene text visual questions requires reasoning about the visual and textual information. Our model is based on an attention mechanism that jointly attends to visual and textual features of the image.

For that we construct a grid of multi-modal features by concatenation of convolutional features and a spatial aware arrangement of word embeddings, so that the resulting grid combines the features of the two modalities at each spatial location (cell). Then we use an attention module that attends to the multi-modal spatial features conditioned to the question. The output weights of this attention module are interpreted as the probability that a certain spatial location (grid cell) of the image contains the answer to the given question.

It is worth noting that with such an approach we somehow recast the problem of scene text VQA as an answer localization task: given an image and question our model localizes the bounding box of the answer text instance. In this sense the architecture of our model is similar to one-stage object detectors, e.g. Redmon et al. (2016) and Liu et al. (2016), but conditioning their output to a given question in natural language form through an attention layer. This idea also directly links with the pointer networks proposed by Vinyals et al. (2015) and used in Singh et al. (2019), but distinctly to these works, we have a fixed input-output space: the number of grid cells.

Another important difference of our model with current state of the art in both standard VQA and scene text VQA is that we use grid based features for encoding the image, while most current models make use of region based features as in Anderson et al. (2018). Although our motivation here is our belief that visual and textual features must be fused together maintaining their spatial co-relation, this has also other benefits, as the whole model is simplified and the times for training and inference are highly reduced.

The summary of the contributions of this paper is as follows:

- We identify a problem with the dual attention mechanisms used in current state of the art for scene text VQA.
- We propose a new model for fixing this problem.
- We demonstrate that grid visual features from a pre-trained one-stage object detector is a good alternative to bottom-up region based features for this task.
- Our model is faster that previous state of the art in both training and inference.
- We outperform the state of the art by 4% accuracy on the ST-VQA dataset.
- We provide extensive experimental results, a thorough ablation study of our model, and a novel human performance analysis on the ST-VQA dataset.

2. Related Work

Scene text visual question answering has been proposed recently with the appearance of two datasets, TextVQA by Singh et al. (2019) and ST-VQA by Biten et al. (2019b).

Along with the ST-VQA dataset, Biten et al. (2019b) presented a baseline analysis including standard VQA models by Kazemi and Elqursh (2017) and Yang et al. (2016), and a variation of those models in which image features where concatenated with a text representation obtained with a scene text retrieval model Gómez et al. (2018) that produces a PHOC representation on its output. Our model takes inspiration from this concatenation of visual and textual features along the spatial dimensions, but we replace the PHOC structural descriptor by semantic word embeddings.

Biten et al. (2019a) organized the ICDAR 2019 Competition on Scene Text Visual Question Answering, in which a total of seven teams evaluated their models on the ST-VQA dataset. The winner entry (VTA) was based on the Bottom-Up and Top-Down VQA model by Anderson et al. (2018) but the textual branch was enhanced with BERT word embeddings, Devlin et al. (2019), of both questions and text instances extracted with an off-the-shelf OCR system.

Mishra et al. (2019) presented a model that represents questions using a BLSTM, images using a pretrained CNN, and OCRed text with their average word2vec representations. They encode each OCRed text block (a group of text tokens) using its coordinate positions, and a semantic tag provided by a named entity recognition model. All these features are concatenated and fed into a MLP network that predicts an answer from a fixed vocabulary (including “yes”, “no”, and 32 predefined book genres) or from one of the OCRed text blocks.

On the Text-VQA side, Singh et al. (2019) proposed the Look, Read, Reason & Answer (LoRRA) method, that extends the well known framework for VQA of Singh et al. (2018) by allowing to copy an OCR token (text instance) from the image as the answer. For this they apply an attention mechanism, conditioned on the question, over all the text instances provided by the OCR model of Borisyuk et al. (2018), and include the OCR token indices as a dynamic vocabulary in the answer classifier's output space. The model uses two attention modules, one attends the visual features and the other attends to textual features, both conditioned on the question. After that the weighted average over the visual and textual features are concatenated and go through a two-layer feed-forward network which predicts the binary probabilities as logits for each answer.

Singh et al. (2019) have also organized the TextVQA Challenge 2019, in which the winner method (DCD_ZJU) extended the LoRRA model by using the BERT embedding instead of GloVE, Pennington et al. (2014), and the Multi-modal Factorized High-order (MFH) pooling proposed by Yu et al. (2018) in both of the attention branches.

The main difference of the model proposed here with the LoRRA and DCD_ZJU models is that we use a single attention branch, that attends jointly to visual and textual features. We also use a different pointer mechanism that directly treats the output weights of the attention module as the probability that a certain cell contains the correct answer to a given question. Notice that this is closer to the original formulation of Pointer Networks Vinyals et al. (2015) since we directly use the predicted weights of the attention module as pointers, without any extra dense layer as in Singh et al. (2019), but slightly different in the sense that our input and output size is fixed by the size of the features' grid. On the other hand in our model we use a one-stage object detector as a visual feature extractor instead of the Faster-RCNN, Ren et al. (2015), used in LoRRA, which implies faster training and inference times.

3. Method

Figure 2 illustrates the proposed model, it consists in four different modules: image encoder (CNN), scene text encoder (OCR + FastText), question encoder (LSTM + FastText), and the answer prediction module. The CNN, OCR, and FastText models are used with pre-trained weights and not updated during training, while the question encoder and answer prediction modules are trained from scratch.

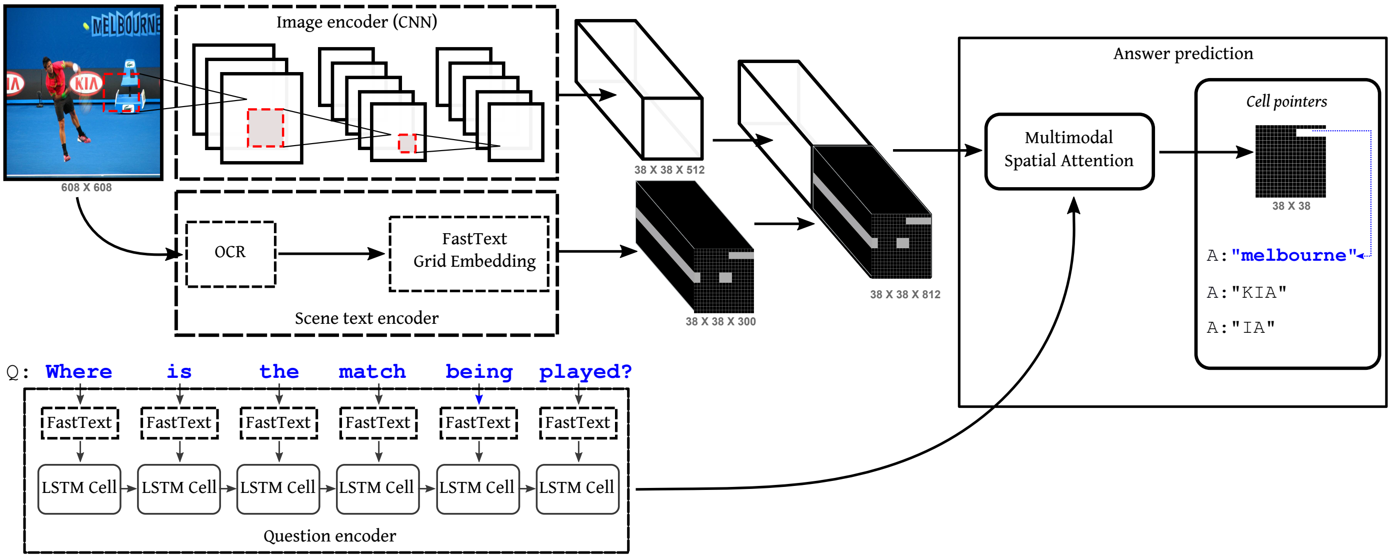


Fig. 2. Our scene text VQA model consists in four different modules: a visual feature extractor (CNN), a scene text feature extractor (OCR + FastText), a question encoder (LSTM + FastText), and the answer prediction model.

3.1. Image encoder

One common component of all visual question answering models is the use of a convolutional neural network as a visual feature extractor. While in the first VQA models it was common to use a single flat vector as a global descriptor for the input image, see Antol et al. (2015) and Kim et al. (2016), with the advent of attention mechanisms grid based features became ubiquitous, see e.g. in Kazemi and Elqursh (2017) and Yang et al. (2016). However, today's standard approach is to use region based convolutional features from a set of objects provided by an object detection network as proposed in Anderson et al. (2018). The rationale is that using objects as the semantic entities for reasoning helps for a better grounding of language.

In this paper we are interested in using grid features, because our whole motivation depends on them. But contrary to previous models using grid features, we propose here to extract them using a one-stage object detector, Redmon and Farhadi (2017), instead of CNN models pretrained for classification. With this we argue that it is possible to maintain a fair trade-off between the use of objects' representations for reasoning and the spatial structure of the grid-based features.

Our visual feature extraction $f_{CNN}(I)$ is based on the architecture of the YOLOv3 model by Redmon and Farhadi (2017) with weights pre-trained on the MS-COCO dataset. The YOLOv3 model has a total of 65 successive 3×3 and 1×1 convolutional layers and residual connections. We extract features from the 61st layer, which produces a feature map with dimensions $38 \times 38 \times 512$ that encode high-level object semantics. This configures the features' grid size in our model to be 38×38. The size of the grid is chosen so that we can quantize the textual information without loosing small words (see next section). A 38×38 grid size means each cell corresponds to a 16×16 patch of the input image (with an 608×608 resolution), which means the smallest possible bounding box of a text instance we expect to find is 16×16.

3.2. Scene text encoder

The first step in our textual feature extractor $f_{ST}(I)$ is to apply an optical character recognition (OCR) model to the input image in order to obtain a set of word bounding boxes and their transcriptions $T = \{(b_1, t_1), (b_2, t_2), \dots, (b_n, t_n)\}$. Text extraction from scene images is still an open research area attracting a lot of interest among the computer vision research community, see e.g. Baek et al. (2019); Liu et al. (2018); Bušta et al. (2018). In this work we have evaluated several publicly available state of the art models as well as the commercial OCR solution of Google[1]

As a standard practice in many applications of natural language processing we embed the words extracted from the OCR module into a semantic space by using a pretrained word embedding model. In our case we make use of the FastText word embedding by Bojanowski et al. (2017), because it allows us to embed out of vocabulary (OOV) words. Notice that OOV words are quite common in scene text VQA because of two reasons: first, some question may refer to named entities or structured textual information that is not present in closed vocabularies, e.g. telephone numbers, e-mail addresses, website URLs, etc.; second, the transcription outputs of the OCR may be partially wrong, either because the scene text is almost illegible, partially occluded or out of the frame.

We use the FastText pretrained model with 1 million $300d$ word vectors, trained with subword information on Wikipedia 2017, UMBC webbase corpus and statmt.org news dataset.

With all word transcriptions in T embedded in the FastText $300d$ space we construct a $38 \times 38 \times 300$ tensor by assigning each of their bounding boxes to the cells in a 38×38 grid with which they overlap as illustrated in Figure 3, so that the embedding vectors maintain the same relative spatial positions as the words in the original image. In order to overcome small

[1] https://cloud.google.com/vision/

words being overlapped by larger words we do this assignment in order, from larger words to smaller. The cells without any textual information are set to zero value. Finally, we concatenate the outputs of the image encoder and the scene text encoder to obtain the multi-modal grid based features of the image $f_m(I) = [f_{CNN}(I); f_{ST}(I)] \in R^{38\times38\times812}$.

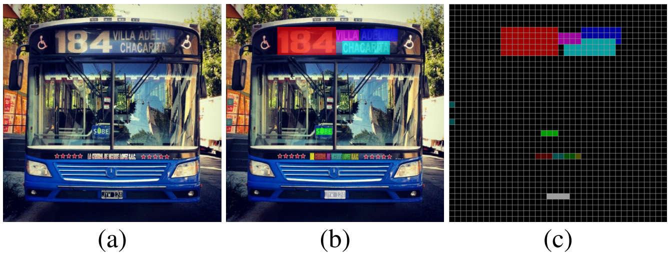

(a) (b) (c)

Fig. 3. Grid cell assignment of the OCR words' bounding boxes. Given an input image (a), the bounding boxes of the words extracted from the OCR model (b) are assigned to their overlapping cells.

3.3. Question encoder

The question encoder is another common module in all VQA models. Recurrent neural networks, either with LSTM or GRU units, are the most common choice of state of the art models, e.g. Yang et al. (2016) Kazemi and Elqursh (2017) Jiang et al. (2018) Anderson et al. (2018) Singh et al. (2019), while the use of CNN has also been explored as an alternative encoding in Yang et al. (2016). In this work we use an LSTM encoder, with the LSTM unit formulation of Gers et al. (2000).

Given a question Q with N words $Q = \{q_1, q_2, \ldots, q_N\}$ we first embed each word with the FastText word embedding function described in section 3.2, and then we feed each word embedding vector into the LSTM. The final hidden layer of the LSTM model is taken as the output of the question encoder:

$$f_q(Q) = LSTM(\tilde{q}_i, h_{i-1}) \forall i \in \{1, 2, \ldots, N\} \tag{1}$$

where $\tilde{q}_i$ is the FastText embedding of word q_i, and h_{i-1} is the output of the LSTM for previous word – we omit the propagation of memory units to simplify the notation. Our LSTM has two dense layers with 256 hidden units and two Dropout layers with a 0.5 drop out rate. The output of the question embedding function $f_q(Q)$ is a vector with 1024 dimensions.

3.4. Answer prediction

The main component of the answer prediction module is an attention mechanism that attends to the spatial multi-modal features $f_m(I)$ conditioned on the question embedding $f_q(Q)$.

Figure 4 illustrates the computation graph of our attention mechanism f_{Att}. First the multimodal grid features $f_m(I)$ are convolved by two 1×1 convolutional layers with 1024 and 512 kernels respectively, resulting in a $38 \times 38 \times 512$ tensor, the question encoded vector $f_q(Q)$ goes through a dense layer with 512 output neurons and is tiled/broadcasted to a shape of $38 \times 38 \times 512$. These two tensors (m_{att} and q_{att}) are added and activated with an hyperbolic tangent (tanh) activation. Finally, the resulting tensor of this operation is convolved with a 1×1 convolutional layer with a sigmoid activation function to produce the output attention map p_{att} with shape $38 \times 38 \times 1$:

$$p_{att} = f_{Att}([f_{CNN}(I); f_{ST}(I)], f_q(Q)) \tag{2}$$

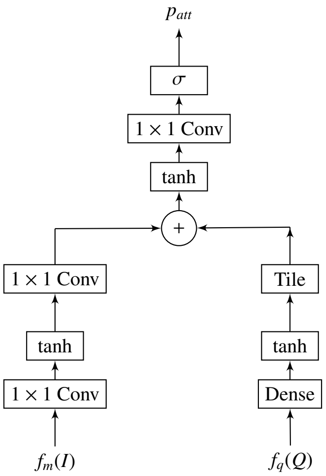


Fig. 4. Computation graph of our attention mechanism f_{Att}.

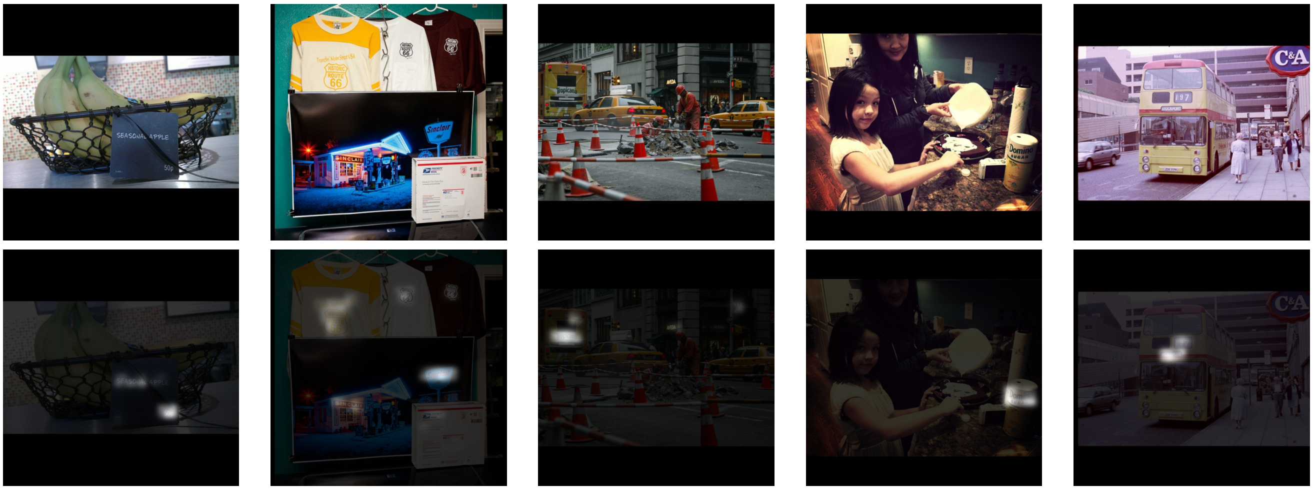
At this point we interpret the values in the output attention map p_{att} as the probability of each image cell to contain the correct answer to the given question Q. Notice that by applying a sigmoid activation function to the last convolution layer we treat the probability for each cell as an individual binary classification problem. This is intentional as in most of the cases the bounding box of the correct answer will cover more than one cell. We train our model using the binary cross entropy loss function:

$$E = -\sum_{i=1}^{38}\sum_{j=1}^{38}\left[g_{i,j}\log p_{i,j} + (1 - g_{i,j})\log(1 - p_{i,j})\right] \tag{3}$$

where $p_{i,j}$ is the probability value of the cell on the ith row and jth column on the output attention map p_{att}, and $g_{i,j}$ is the ground truth value for that cell: 1 if the cell contains the answer, 0 otherwise. At inference time, the predicted answer is the OCR token assigned to the cell with maximum probability.

The attention mechanism described so far can be used within several design variations such as the stacked attention of Yang et al. (2016), or the question-image co-attention of Lu et al. (2016) and Nam et al. (2017). In particular we have adopted the stacked design in our model and empirically found an improvement over using a single attention layer (see the ablation study in section 4.4 for the details). For this we stack two attention layers, and in the first one we combine the weighted average over the multimodal spatial features (using the output probability map as weights) with the question embedding (by addition), and this combination is fed to the second attention layer as the question embedding.

Moreover, we notice that since our model is made fully convolutional (including the image encoder) on all the visual branch, we can perform inference at different input scales using the same learnt weights.



Q: How much is the rate of one seasonal apple fruit?
A: 50p

Q: What is the word on the white sign?
A: sinclair

Q: What is the brand advertised on the back of the bus?
A: tropicana

Q: What is the sugar brand?
A: domino

Q: What is the destination of the bus ?
A: stockport

Fig. 5. Examples of questions from the ST-VQA tests and correctly predicted answers by our model.

4. Experiments

In this section we present a set of experiments performed on the ST-VQA and TextVQA datasets. First, we briefly introduce both datasets and their metrics. Second we present a comparison of different OCR systems on the ST-VQA dataset. Then we compare the performance of the proposed model with the state of the art on both datasets, and present an ablation study of the proposed model. Finally, we present an extension of the ST-VQA dataset and analyze human performance on a subset of its test set.

4.1. Datasets

The ST-VQA dataset comprises $23,038$ images and $31,791$ question/answer pairs. The images were collected from seven different public data sets with the only requirement to contain at least 2 text tokens, so there is always some inherent confusion. The dataset is split into two sets with $19,027$ images / $26,308$ questions for training, and $2,993$ images / $4,163$ questions for testing. The annotation process was carried out by human annotators who received specific instructions to ask questions based on the text present in each image, so that the answer to the questions should always be a token (or a set of tokens) of legible text in the image.

The evaluation metric on the ST-VQA dataset is the average normalized Levenshtein similarity (ANLS) that assigns a soft score s to a given pair of predicted and ground-truth answers (ans_{pred} and ans_{gt}) based on their normalized Levenshtein edit distance (d_{LN}): $s(ans_{pred}, ans_{gt}) = 1 - d_{NL}(ans_{pred}, ans_{gt})$.

The TextVQA dataset comprises a total of $28,408$ images and $45,336$ questions. It is split into sets of $21,953$ images / $34,602$ questions for training, $3,166$ images / $5,000$ questions for validation, and $3,289$ images / $5,734$ questions for test. All images come from the OpenImages dataset, Kuznetsova et al. (2020), and were sampled on a category basis, emphasizing categories that are expected to contain text. In TextVQA any question requiring reading the image text is allowed, including questions for which the answer does not correspond explicitly to a legible text token (e.g. around 9% are binary (yes/no) questions). Notice that distinct from ST-VQA answering those questions implies the use of a fixed output vocabulary.

The evaluation metric on the TextVQA dataset is the VQAv2 accuracy: $Acc(ans) = min(\frac{h(ans)}{3}, 1)$ where $h(3)$ counts the number of humans that answered ans among the 10 collected human answers for each question. All accuracy values reported in this section are expressed in percentage.

It is worth noting that in parallel to these two datasets Mishra et al. (2019) presented the OCR-VQA dataset, with more than 1 million question-answer pairs about 207K images of book covers. However, we do not consider it in our experiments because the task in OCR-VQA is different in nature to the one our model is designed for, since more than 50% of the questions have answers that are not scene text instances (including for example 40% binary (yes/no) questions and 10% questions about book genres).

4.2. OCR performance analysis

Table 1 shows the answer recall of two different state of the art scene text recognition models and of a commercial OCR system. Answer recall is computed as the percentage of answers in the ST-VQA train set that match with a text token found by the OCR system. The ANLS upper-bound gives us the maximum score we can achieve in this dataset with different OCR systems.

For all experiments reported in this section on the ST-VQA dataset we use the OCR tokens obtained with the Google OCR API. For the experiments on the TextVQA dataset we use the OCR tokens from the Rosetta OCR system, Borisyuk et al. (2018), that are provided with the dataset to showcase comparable results. At training time we discard image/question pairs

Table 1. Answer recall and ANLS upper-bound for different off-the-shelf OCR systems on the ST-VQA training set.

OCR	Answer Recall	ANLS Upper-bound
FOTS – Liu et al. (2018)	37.56	0.47
E2EML – Bušta et al. (2018)	41.37	0.52
Google OCR API	60.19	0.74

for which the answer is not in the OCR tokens' set.

4.3. Performance comparison

Table 2 compares the performance of the proposed model with the state of the art on the ST-VQA dataset. We appreciate that our model clearly outperforms all previously published methods both in ANLS and accuracy, improving more than 10% ANLS compared to the ST-VQA competition models and 5% ANLS over LoRRA. It is important also to recall here that our model is 5× faster than LoRRA at processing an image, as a consequence of using YOLOv3 instead of Faster-RCNN for feature extraction.

Table 2. ST-VQA performance comparison on the test set. Numbers with † are from the official implementation of LoRRA trained on ST-VQA using the same OCR tokens as in our model.

Method	ANLS	Acc.
SAAA Kazemi and Elqursh (2017)	0.087	6.66
SAN Yang et al. (2016)	0.102	7.78
SAN+STR Gómez et al. (2018)	0.136	10.34
QAQ - rep. from Biten et al. (2019a)	0.256	19.19
VTA - rep. from Biten et al. (2019a)	0.282	18.12
LoRRA Singh et al. (2019)	0.331†	21.28
Ours	**0.381**	**26.06**

Figure 5 shows qualitative examples of the produced attention masks and predicted answers for 5 image/question pairs from the ST-VQA test set that are correctly answered by our model. Among them we can see examples in which textual information alone would suffice to provide a correct answer, but also cases where a joint interpretation of visual and textual cues is needed. More qualitative examples are provided as supplementary material of this paper.

Table 3 shows the performance comparison on the validation set of TextVQA. In this case we also compare the accuracy in the specific subset of questions for which the answer is among OCR tokens (indicated as Acc.† in the table), to understand how the presence of answers that do not correspond to scene text instances in the image (e.g. *"yes"*/*"no"* answers) affect the performance of our model. In this subset our model outperforms previous state of the art by a clear margin, while in the whole validation set we observe the opposite. Notice that this is expected because our model has no mechanism for providing valid answers to questions the answers of which are not in the OCR tokens, while the LoRRA model can cope with these questions by using a fixed vocabulary answer output space similar to standard VQA models.

Table 3. TextVQA performance comparison on the validation set. Acc.† refers to the subset of questions with answers among OCR tokens.

Method	Acc.†	Acc.
SAAA Kazemi and Elqursh (2017)	9.09	13.33
LoRRA Singh et al. (2019)	32.03	**27.48**
Ours	**37.60**	21.88
Ours + SAAA	**37.60**	26.07

In order to provide a fair comparison in the whole validation set of TextVQA we have combined the predictions of our model with the well known standard VQA model SAAA, Kazemi and Elqursh (2017). In this experiment we have trained the SAAA model on TextVQA with a fixed output space of the most common 3,000 answers, and the results of entry Ours+SAAA in Table 3 correspond to an ensemble model in which the the answer is selected with a threshold-based decision. More specifically, the ensemble selects the SAAA answer if its classification confidence is above a given threshold, otherwise it selects the answer of our model. We use a threshold decision over the classification score of the SAAA model and not over ours because we have experimentally found that the confidences of SAAA are better indicators for whether a given question can be answered or not without reading the scene text. The threshold value used was set to 0.5 as in a binary classification problem. We appreciate that this ensemble model achieves competitive performance to the state of the art. While SAAA alone has a marginal performance in TextVQA, the confidences of its predictions are good indicators for whether a given question can be answered without reading the scene text. In such a scenario a model like ours can be leveraged in a mixed dataset where questions may or may not require answers from the OCR tokens' set.

4.4. Ablation study and effect of different pre-trained models

In this section we perform ablation studies and analyze the effect of different pre-trained models in our method's performance. Table 4 shows ablation experiments for different attention mechanisms in our model. **FCN** stands for a Fully Convolutional Network in which three convolutional layers (with respectively 512, 256, and 1 3×3 kernels, ReLU activations and Batch Norm) are applied to the concatenation of features from the YOLOv3 model, the grid of OCR tokens' FastText embedding vectors, and the (tiled) LSTM question embedding. This model has no attention mechanism, but produces at its output a 38×38 grid as in our model and can be trained in the same way. The **FCN + Dual Att.** model uses a dual attention mechanism similar to the LoRRA model: one attention module attends the YOLOv3 features conditioned to the question, and the other attends to the grid of OCR tokens FastText vectors conditioned to the question. The outputs of those two attention modules are then concatenated and fed into a convolutional block (similar as for the FCN model) to produce the 38×38 output. Finally, **FCN**

+ Multi-modal Att. and **FCN + Stack Multi-modal Att** correspond to the proposed model, with one and two multi-modal attention layers respectively as explained in section 3.4. We can point out that the dual attention mechanism is not helping at all under this set-up, while our multi-modal attention layers consistently improve the results of the FCN model.

Table 4. Ablation study using different attention mechanisms in our model.

Method	ANLS
FCN	0.319
FCN + Dual Att.	0.279
FCN + Multi-modal Att.	0.355
FCN + Stack Multi-modal Att.	**0.381**

In Table 5 we study the effect of different pre-trained word embedding models and CNN backbones in our method performance.

Table 5. ST-VQA performance using different pre-trained word embedding models and CNN backbones.

CNN	Q. Emb.	OCR Emb.	ANLS
Inception v2	FastText	FastText	0.319
ResNet-152	FastText	FastText	0.332
YOLO v3	FastText	FastText	**0.381**
YOLO v3	BERT	FastText	0.327
YOLO v3	BERT	BERT	0.310

We observe that the visual features of the YOLOv3 object detection model yield superior performance when compared with pre-trained features of two well known networks for image classification: InceptionV2, Szegedy et al. (2016), and ResNet-152, He et al. (2016). Also in Table 5 we appreciate that the FastText pre-trained word embedding works better than the BERT embedding for both the question and OCR tokens' encoders.

4.5. *ST-VQA extensions and human performance analysis*

With this paper we are releasing an updated version of the ST-VQA dataset that includes the OCR tokens used in all our experiments. This way we make sure any methods using OCR tokens and evaluating in this dataset can be fairly compared under the same conditions. Moreover, in order to understand the nature of the dataset better, we have conducted a study to analyze human performance under different conditions. For this we have asked human participants to answer a subset of $1,000$ questions from the test set given the following information:

- S1: we show the question and the image.
- S2: we show the question and the image but with all text instances blurred (illegible).
- S3: we show the question and a list of words (OCR tokens), no image is shown.

in all three cases participants had the option to mark the questions as *"unanswerable"*.

Table 2 shows the human performance in terms of ANLS and accuracy in the three scenarios described above. We appreciate that S1 is consistent with the human study reported in Singh et al. (2019) in terms of accuracy. Their study shows a human accuracy of 85.0 in TextVQA, but having collected 10 answers per question their accuracy metric is a bit more flexible in accepting diverse correct answers. Moreover, we observe that S2 and S3 demonstrate that the textual cue is much more important than the visual cue in ST-VQA. Another point to stress is that humans are especially good at answering questions without even seeing the image. This is because of the fact that humans use a-priori knowledge of what a number is or what a licence plate is, etc. As an example, an image for which the question is *"What is the price of ..."* can be correctly answered by selecting a unique numerical OCR token since the price has to be a number.

Table 6. Human performance on a subset of $1,000$ questions of the ST-VQA test set under different conditions, depending whether visual (V) or textual (T) information is given.

	V	T	ANLS	Acc.
S1 human performance	✓	✓	0.85	78.16
S2 human performance	✓	✗	0.21	18.81
S3 human performance	✗	✓	0.52	37.54

The complete results of this human study are provided as supplementary material to this paper. Furthermore, we will include in the new version of the dataset the indices of the $1,000$ test questions used in this study, and the indexes of text questions for which their answer is among the provided OCR tokens, so that interested researchers can analyze the performance of their methods on those test subsets of special interest.

5. Conclusion

We have presented a new model for scene text visual question answering that is based in an attention mechanism that attends to multi-modal grid features, allowing it to reason jointly about the textual and visual modalities in the scene.

The provided experiments and ablation study demonstrate that attending on multi-modal features is better than attending saparately to each modality. Our grid design choice also proves to work very well for this task, as well as the choice of a one-stage object detection backbone instead of a classification one. Moreover, we have shown that the proposed model is flexible enough to be combined with a standard VQA model obtaining state of the art results on mixed datasets with questions that can not be answered directly using OCR tokens.

Acknowledgement

This work has been partially supported by the Spanish research project TIN2014-52072-P and the CERCA Programme / Generalitat de Catalunya. We gratefully acknowledge the support of the NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L., 2018. Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6077–6086.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D., 2015. Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, pp. 2425–2433.

Baek, J., Kim, G., Lee, J., Park, S., Han, D., Yun, S., Oh, S.J., Lee, H., 2019. What is wrong with scene text recognition model comparisons? dataset and model analysis, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4715–4723.

Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Mathew, M., Jawahar, C., Valveny, E., Karatzas, D., 2019a. Icdar 2019 competition on scene text visual question answering, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE. pp. 1563–1570.

Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D., 2019b. Scene text visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, pp. 4291–4301.

Bojanowski, P., Grave, E., Joulin, A., Mikolov, T., 2017. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics 5, 135–146.

Borisyuk, F., Gordo, A., Sivakumar, V., 2018. 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, pp. 71–79.

Bušta, M., Patel, Y., Matas, J., 2018. E2e-mlt-an unconstrained end-to-end method for multi-language scene text, in: Asian Conference on Computer Vision, Springer. pp. 127–143.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186.

Gers, F.A., Schmidhuber, J., Cummins, F., 2000. Learning to forget: Continual prediction with lstm. Neural Computation 12, 2451–2471.

Gómez, L., Mafla, A., Rusinol, M., Karatzas, D., 2018. Single shot scene text retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 700–715.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 770–778.

Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D., 2018. Pythia v0. 1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956 .

Kazemi, V., Elqursh, A., 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. arXiv preprint arXiv:1704.03162 .

Kim, J.H., Lee, S.W., Kwak, D., Heo, M.O., Kim, J., Ha, J.W., Zhang, B.T., 2016. Multimodal residual learning for visual qa, in: Advances in neural information processing systems, pp. 361–369.

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., et al., 2020. The open images dataset v4. International Journal of Computer Vision , 1–26.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.Y., Berg, A.C., 2016. Ssd: Single shot multibox detector, in: European conference on computer vision, Springer. pp. 21–37.

Liu, X., Liang, D., Yan, S., Chen, D., Qiao, Y., Yan, J., 2018. Fots: Fast oriented text spotting with a unified network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5676–5685.

Lu, J., Yang, J., Batra, D., Parikh, D., 2016. Hierarchical question-image co-attention for visual question answering, in: Advances in neural information processing systems, pp. 289–297.

Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A., 2019. Ocr-vqa: Visual question answering by reading text in images, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE. pp. 947–952.

Nam, H., Ha, J.W., Kim, J., 2017. Dual attention networks for multimodal reasoning and matching, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 299–307.

Pennington, J., Socher, R., Manning, C.D., 2014. Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pp. 1532–1543.

Redmon, J., Divvala, S., Girshick, R., Farhadi, A., 2016. You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 779–788.

Redmon, J., Farhadi, A., 2017. Yolo9000: better, faster, stronger, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 7263–7271.

Ren, S., He, K., Girshick, R., Sun, J., 2015. Faster r-cnn: Towards real-time object detection with region proposal networks, in: Advances in neural information processing systems, pp. 91–99.

Singh, A., Natarajan, V., Jiang, Y., Chen, X., Shah, M., Rohrbach, M., Batra, D., Parikh, D., 2018. Pythia-a platform for vision & language research, in: SysML Workshop, NeurIPS.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M., 2019. Towards vqa models that can read, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8317–8326.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z., 2016. Rethinking the inception architecture for computer vision, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2818–2826.

Vinyals, O., Fortunato, M., Jaitly, N., 2015. Pointer networks, in: Advances in neural information processing systems, pp. 2692–2700.

Yang, Z., He, X., Gao, J., Deng, L., Smola, A., 2016. Stacked attention networks for image question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 21–29.

Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D., 2018. Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE transactions on neural networks and learning systems 29, 5947–5959.