Guiding Scene Text Question Answering with Region Priors

Anonymous submission
Paper ID: 321

Abstract

TextVQA, attacking increasing interests in recent research, works on the questions related to scene text in an image. The TextVQA task needs models to understand better the scene text (OCR tokens) and its relation with both the visual objects and the question. The question, objects, and OCR tokens contribute to predicting the answers from different perspectives. However, existing works neglect their different roles when performing a joint interaction. In this paper, we analyze and utilize their different roles to explicitly achieve stepwise interactions among them. Specifically, we treat objects as a bridge between the question and the OCR token and propose a Region-aware OCR Query (ROQ) model. The ROQ model performs queries from the question to the objects to the OCR tokens and predicts answers based on the outputs of the queries. The ROQ mainly consists of a Region Query module (RQ) and an OCR Query module (OQ). The RQ module locates the salient region associated with answering the question. The OQ module finds salient OCR tokens relevant to the salient region by grouping similar OCR tokens and querying the OCR tokens according to the salient region. The ROQ model not only significantly outperforms existing state-of-the-art models under the same setting but also can show explainable insights into the stepwise querying process. The code will be publicly available upon publication.

1 Introduction

TextVQA aims to answer questions relevant to scene text and visual components in the scene, drawing increasing attention in recent research studies. Compared to the VQA task, the TextVQA task requires models to understand better scene text and how scene text relates to the visual scene and the question. For example (see Figure 1), to correctly answer the question, the models need to identify the group of OCR tokens in (c) and their relationships to both words "what is written" in the question and the "red part" in the image. Existing works (Hu et al. 2020; Yang et al. 2021; Kant et al. 2020; Zhu et al. 2021) formulate the TextVQA task as a multi-modal understanding problem: with a question (a sequence of words), an image (a set of detected visual objects), and scene text (a set of extracted OCR tokens) as the input, models perform a joint understanding among the three modalities of the question words, objects, and OCR tokens.

It is vital to identify the different roles of the three modalities in the TextVQA task. The reasons are: (1) the three modalities contribute to predicting the answers from different perspectives; (2) The three modalities' different roles provide preliminary information to understand better how the three modalities can be associated. The question can be treated as guidance to predict the answers, and it reveals which objects and OCR tokens the models require to pay more attention to. The objects can be a bridge between the question and the answers, which builds the connections between the question words and the OCR tokens. For example (see Figure 1), the object (see b) mentioned by the "red part" helps locate the answers of the OCR tokens shown in (c). The OCR tokens are closely related to the answers, whose relationship with the question can be modeled directly or established through the objects indirectly. However, existing works neglect the different roles of the three modalities. Some works (Hu et al. 2020; Yang et al. 2021; Biten et al. 2022) independently and respectively project all entities of the three modalities into a common embedding space. Each entity is allowed to indiscriminately build connections with other entities without regard to the entities' modalities in the embedding space. Some other approaches (Kant et al. 2020;

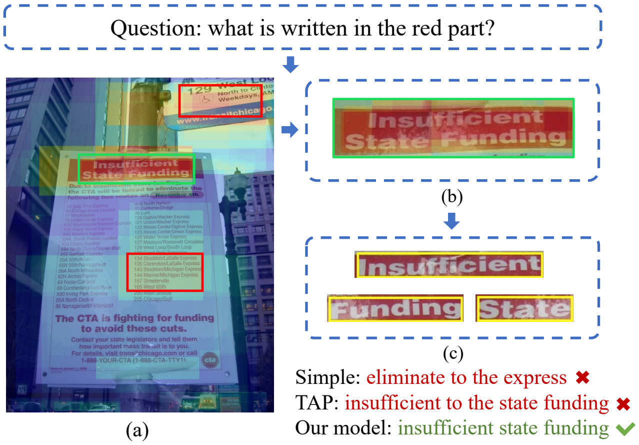

Question: what is written in the red part?
(a)
(b)
(c)
Simple: eliminate to the express ✗
TAP: insufficient to the state funding ✗
Our model: insufficient state funding ✓

Figure 1: Example of TextVQA. Our ROQ model performs region-aware OCR query, which first adopts the question as the query to locate the salient region (region a) and then identify salient OCR tokens "insufficient state funding" relevant to the salient region.

Gao et al. 2021, 2020; Zhu et al. 2021) consider the different roles of the question and the other two modalities. However, they still fail to distinguish the roles between objects and OCR tokens. Some of them (Kant et al. 2020; Gao et al. 2020) make the entities in the two modalities attend to each other equivalently. Alternatively, the others (Zhu et al. 2021; Gao et al. 2021) make the question parallelly attend to the two modalities without distinguishing them.

Besides, insights of models to illuminate the answering process and how they predict the answers based on objects and OCR tokens is also a critical perspective to explore the TextVQA task. However, most prior models (Zhu et al. 2021; Hu et al. 2020; Yang et al. 2021) give the answers without any explanation, which cannot satisfy users in most cases. They usually jointly fuse and understand the three modalities via a holistic multi-modal transformer. It is hard to know how the entities of the three modalities interact by analyzing the multi-head attention distributions in multiple layers. Some methods (Gao et al. 2021, 2020) use complex graph-based encoders to capture the relationships among entities of the three modalities. Based on the graph attention scores, they label one or two objects relevant to answering the question to provide a few insights. However, they fail to show explainable visual evidence of the step-by-step prediction process from the input to the answers.

In this paper, we aim to address the challenges mentioned above: (1) To achieve a better joint understanding of the three modalities by utilizing their different roles in the TextVQA task; (2) To realize the explainable prediction process by explicitly performing stepwise interactions among the three modalities. Specifically, we make full use of objects (components of a visual region) as a bridge between the question and the OCR tokens to implement a Region-aware OCR Query (ROQ) from the question to the salient region to the salient OCR tokens. On the one hand, the salient region and OCR tokens directly contribute to the answer prediction. On the other hand, the stepwise query process to these salient entities can provide insights that can be visualized and explained explicitly. As shown in Figure 1, ROQ first identifies the salient region shown in (a) according to the question and then further grounds salient OCR tokens in (c) based on the salient region. ROQ mainly consists of a Region Query module (RQ) and OCR Query module (OQ). The RQ module locates the salient region according to the question via a stack of self-attention and cross-attention blocks. To facilitate the model to learn to locate the salient region, we propose a region-aware loss which minimizes the difference between the soft target of the region and the model's prediction. The OQ module finds salient OCR tokens relevant to the region by conducting the region-OCR and OCR-OCR queries. The OCR-OCR query models the relationships among similar and close OCR tokens based on their spatial locations. Finally, ROQ generates the answers based on the salient region and salient OCR tokens via an answer generative decoder (Hu et al. 2020).

To demonstrate the effectiveness of our proposed ROQ model, we evaluate it on two common benchmark datasets, *i.e.*, TextVQA (Singh et al. 2019) and ST-VQA (Biten et al. 2019). Experimental results show that our model ROQ consistently and significantly outperforms all the state-of-the-art models under the same setting. The main contributions are:

- We propose a Region Query module (RQ) to find the salient region associated with answering the question, supervised by the proposed region-aware loss. The salient region can serve as an intermediate between the question and OCR tokens.
- We propose an OCR Query module (OQ) to identify salient OCR tokens from the salient region. The OQ module first groups similar OCR tokens by modeling their spatial and semantic relationships and then queries the OCR tokens according to the salient region.
- We propose a Region-aware OCR Query network (ROQ) consisting mainly of RQ and OQ modules. The ROQ model not only significantly outperforms existing state-of-the-art models under the same setting but also can show explainable insights into the answering process.

2 Related Work

Text Based Visual Question Answering. TextVQA aims to answer questions relevant to scene texts present in a given image. Compared with traditional VQA task (Cao et al. 2018; Andreas et al. 2016; Hu et al. 2017; Wang et al. 2020), TextVQA introduces a new modality - Optical Character Recognition (OCR), which poses a challenge for reading, understanding, and reasoning about OCR tokens and their semantic relations with other visual constituents (*e.g.*, objects) in the image and the question.

LoRRA (Singh et al. 2019) is first introduced together with the Text-VQA dataset, extending VQA model Pythia (Jiang et al. 2018) with OCR attention module. M4C (Hu et al. 2020) first adopts a unite multi-modal transformer framework to better jointly understand the three modalities (*e.g.*, question, image and OCR), and then predict the next word from answer vocabulary or select an OCR token iteratively. SA-M4C (Kant et al. 2020) improves M4C by performing spatial reasoning between objects and OCR tokens. Unlike previous models that treats three modalities indiscriminately, recent models (Zhu et al. 2021; Lu et al. 2021) pay increasing attention to OCR understanding. Simple (Zhu et al. 2021) divides OCR representations into visual and linguistic parts for fine-grained feature matching. While LOGOS (Lu et al. 2021) learns to group and localize OCR tokens first before feeding them and other two modalities into the multi-modal transformer. Some studies (Yang, Li, and Yu 2019; Hu et al. 2019; Santoro et al. 2017; Li et al. 2019a) introduce graph neural networks to encode relations between various modalities. MM-GNN (Gao et al. 2020) decomposes the image into three sub-graphs, and then updates representations of the graph nodes via three aggregators. Apart from these models which optimize the objective of TextVQA directly, pretraining-based models like TAP (Yang et al. 2021) and LaTr (Biten et al. 2022) also gain significant progress and show strong potential of pre-training.

Transformer for Vision-and-Language Tasks. Transformer is first introduced in natural language processing(Vaswani et al. 2017) and has achieved superior perfor-

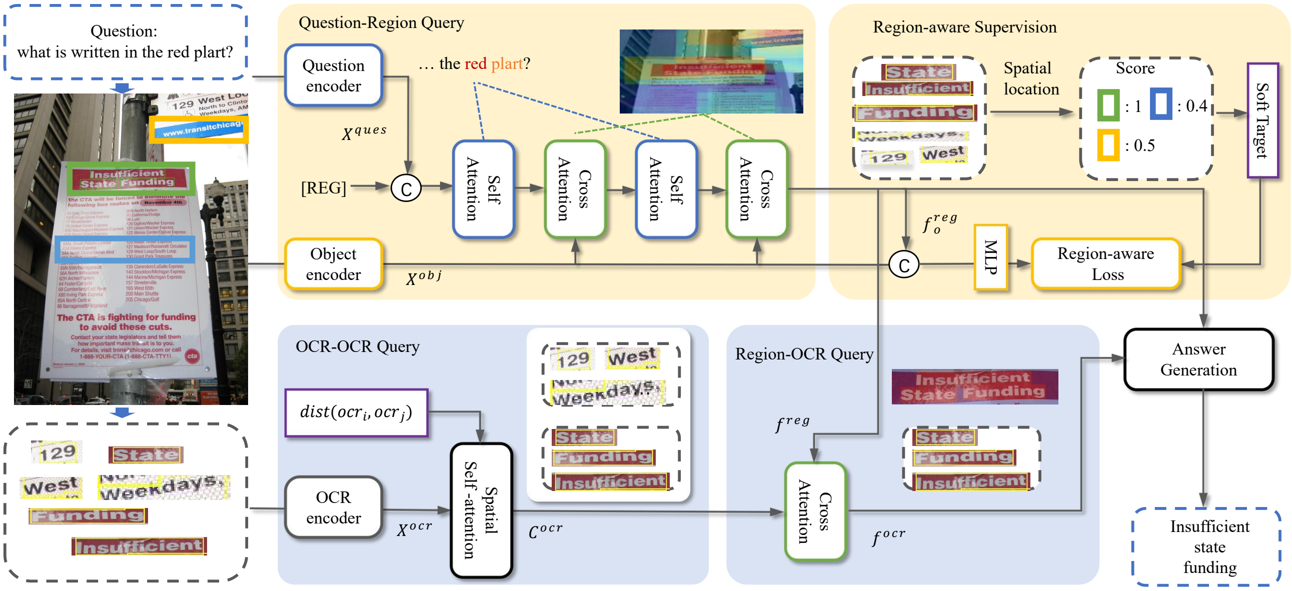


Figure 2: An overview of Region-aware OCR Query network (ROQ). The ROQ mainly consists of two modules: the Region Query module (yellow) and the OCR Query module (blue). Given the question, the image, and OCR tokens, the ROQ first locates the salient region relevant to the question and then identifies salient OCR tokens closely associated with the salient region. Finally, both salient region and OCR tokens contribute to the answer generation.

mance on many tasks like machine translation, question answering, etc.

Recently, progress on transformer is made from solely natural language understanding to vision-and-language multi-modal learning. Inspired by large-scale pre-trained language models such as ELMO (Peters et al. 2018), GPT (Peters et al. 2018) and BERT (Devlin et al. 2018), the cross-modality pre-training for vision and language is proposed to learn task-agnostic multi-modal joint representations based on transformer architectures and multi-modal pre-trained tasks. ViLBERT (Lu et al. 2019) extends standard transformer encoder to co-attention transformer to enable vision-attended language and language-attended visual learning. LXMERT (Tan and Bansal 2019) integrates a cross-modality encoder into two single-modality transformer encoders. While VisualBERT (Li et al. 2019b) and VL-BERT (Su et al. 2020) adopt a single multi-modal transformer to represent visual and linguistic tokens jointly. Apart from vision and language, TAP (Yang et al. 2021) incorporates OCR modality into multi-modal pre-training process, which boosts the performance of textVQA task. In addition to transformer-based pre-training models, some state-of-the-art models (, M4C, SA-M4C, and Simple) utilize the multi-modal transformer as the encoder and iteratively decode the answers based on the multi-modal transformer decoder.

Although details of these models vary, they feed the output of various encoders into a holistic transformer, which fails to consider those different modalities play various roles. In contrast, we propose a region-aware OCR query framework mechanism to stepwisely decode salient regions and salient OCR tokens to identify target OCR tokens.

3 Region-aware OCR Query

The overall architecture of our Region-aware OCR Query (ROQ) is shown in Figure 2. It mainly consists of three modules, *i.e.*, the Region Query module, the OCR Query module and the answer generation module. First, given a question, detected visual objects and OCR tokens in an image as the input, the Region Query module locates the salient region of the image (*i.e.*, the part closely associated with answering the question) to generate a region-aware query vector (in Section 3.2). Next, the OCR Query module finds target OCR tokens and generates an OCR-aware query vector by performing queries among the region-aware query vector and OCR tokens (in Section 3.3). Finally, the answer generation module predicts the answers based on the query vectors generated above (in Section 3.4).

3.1 Feature Encoding

Following existing works (Yang et al. 2021; Lu et al. 2021), we extract the question features $\{x_k^{\text{ques}}\}_{k=1}^{K}$, object features $\{x_m^{\text{obj}}\}_{m=1}^{M}$ and OCR features $\{x_n^{\text{ocr}}\}_{n=1}^{N}$. Specifically, The object feature is composed by the Faster R-CNN feature (Ren et al. 2015) and location feature. And the OCR feature is fused from the Faster R-CNN feature (Ren et al. 2015), the FastText vector (Bojanowski et al. 2017), the BERT linguistic representation (Lu et al. 2021), the PHOC vector (Almazán et al. 2014) and location feature. More details of feature encoding are given in the Appendix.

3.2 Region Query Module

The Region Query module locates the salient region where the question describes and represents the salient region as the region-aware query vector via question-region query (in Section 3.2) and region-aware supervision (in Section 3.2). As shown in Figure 2, given the question and the objects, the question-region query generates the region-aware query vector by querying objects relevant to the question via the cross-attention mechanism (Vaswani et al. 2017). To facilitate the model to learn the query vector, we apply the region-aware supervision to first generate a soft target of the region and then force the model to predict the target based on the query vector by minimizing the cross entropy.

Question-Region Query. As shown in Figure 2, inspired by transformer decoder (Vaswani et al. 2017), question-region query is composed of the pairs of interlaced self-attention and cross-attention blocks. We utilize the self-attention block to model the contexts among the words in the question and enhance the feature representation of words which describe the salient region. With the sequence of word embedding $\{x_k^{\text{ques}}\}_{k=1}^K$ as the input, we obtain the enhanced word representation $\boldsymbol{f}^{\text{ques}} \in R^{d\times(K+1)}$ (where $\boldsymbol{f}^{\text{ques}} = [f_0^{\text{ques}}, f_1^{\text{ques}}, \cdots, f_K^{\text{ques}}]$) as follows,

$$\begin{aligned} \boldsymbol{x}^{\text{ques}} &= [[\text{REG}], x_1^{\text{ques}}, x_2^{\text{ques}}, \cdots, x_K^{\text{ques}}], \\ \boldsymbol{f}^{\text{ques}} &= \text{LN}(\boldsymbol{x}^{\text{ques}} + \text{MSA}(\boldsymbol{x}^{\text{ques}})), \end{aligned} \tag{1}$$

where MSA$(\cdot)$ and LN$(\cdot)$ mean the multi-head self-attention and the layer normalization, respectively. And the [REG] represents a learnable embedding, which is randomly initialized at the begin of the training stage. Inspired by the [CLS] token in machine translation tasks (Devlin et al. 2018), here, we append the [REG] token to abstract the query information relevant to the salient region. The output state of the [REG] token, f_0^{ques}, captures the region-relevant part of the question.

We further adopt the cross-attention block to build the interconnection between the words and the objects to find the salient region. Specifically, the cross-attention block first projects the enhanced word representation $\boldsymbol{f}^{\text{ques}}$ to queries and the object features $\{x_m^{\text{obj}}\}_{m=1}^M$ to key-value pairs, and then performs the mapping between the queries and key-value pairs via the multi-head attention mechanism. The query result $\boldsymbol{f}^{\text{reg}} \in R^{d\times(K+1)}$ corresponding to the enhanced word representation $\boldsymbol{f}^{\text{ques}}$ is computed as follows,

$$\begin{aligned} \boldsymbol{x}^{\text{obj}} &= [x_1^{\text{obj}}, x_2^{\text{obj}}, \cdots, x_M^{\text{obj}}], \\ \boldsymbol{f}^{\text{reg}} &= \text{LN}(\boldsymbol{f}^{\text{ques}} + \text{MCA}(\boldsymbol{f}^{\text{ques}}, \boldsymbol{x}^{\text{obj}})), \end{aligned} \tag{2}$$

where MCA$(\cdot)$ and LN$(\cdot)$ mean the multi-head cross-attention and the layer normalization, respectively. Here, the output corresponding to the initial [REG] token, f_0^{reg}, captures the objects relevant to the salient region. Therefore, we name the f_0^{reg} the region-aware query vector.

Region-aware Supervision. To facilitate the model to learn the region-aware query vector f_0^{reg} and explicitly find the salient region relevant to the question, we propose a region-aware supervision strategy. We leverage the query vector f_0^{reg} to predict the salient region directly and minimize the difference between the prediction and the target region at the training stage.

However, there is no ground-truth annotation for the target region. We thus build a pseudo target region by the following three steps: (1) We set the target as a set of soft labels of objects, and each soft label between 0 and 1 represents the probability that the object associated with the salient region; (2) We first find the target OCR tokens by mapping the ground-truth answers to OCR tokens' texts, and then obtain the minimal bounding box which covers all the target OCR tokens. (3) For each object, we calculate the ratio of the overlapping area of the object and the minimal bounding box to the area of the object. The normalized ratios of the objects are set as the soft labels which are denoted as $\{p_m^{\text{reg}}\}_{m=1}^M$.

We predict the soft labels of the objects based on the query vector f_0^{reg} and minimize the cross entropy between the predicted probability $\{p_m^{\text{pred}}\}_{m=1}^M$ and the target $\{p_m^{\text{reg}}\}_{m=1}^M$, which is formulated as follows,

$$\begin{aligned} p_m^{\text{pred}} &= \text{Sigmoid}(\text{MLP}([f_0^{\text{reg}}; x_m^{\text{obj}}])), \\ L^{reg} &= -\frac{1}{M}\sum_{m=1}^{M} p_m^{\text{reg}} \log(p_m^{\text{pred}}) + (1 - p_m^{\text{reg}}) \log(1 - p_m^{\text{pred}}), \end{aligned} \tag{3}$$

where MLP$(\cdot)$, Sigmoid$(\cdot)$ and $[;]$ and the multilayer perceptron, sigmoid activation function and concatenation operation, respectively.

3.3 OCR Query Module

The OCR Query module identifies target OCR tokens with the guidance of the salient region and generates an OCR-aware query vector by performing queries among OCR tokens via the OCR-OCR query (in Section 3.3) and the query between the region-aware query vector and OCR tokens via the region-OCR query (in Section 3.3). As shown in Figure 2, the OCR-OCR query groups similar OCR tokens by modeling the intra-relationships among OCR tokens based on their semantic features and spatial locations. The region-OCR query generates the ocr-aware query vector by querying the grouped OCR tokens relevant to the region-aware query vector f_0^{reg} (see the definition in Section 3.3).

OCR-OCR Query. As shown in Figure 3, the OCR-OCR query is composed of a stack of the proposed spatial self-attention blocks. Similar to the traditional self-attention mechanism (Vaswani et al. 2017), the spatial self-attention block models the contexts among all the input elements but explicitly considers the spatial relationship among the elements. The relatively spatial locations among OCR tokens indicate their relevance, which is a vital perspective for contextual modeling. We first define the spatial distances among the OCR tokens in an image, and then introduce the spatial self-attention block.

First, for each pair of OCR tokens, we define their spatial distance as the minimum Manhattan distance between any two points on the two OCR tokens' bounding boxes. The reason we use the Manhattan distance is that a group of OCR tokens is usually aligned horizontally or vertically. The spatial distance $d_{i,j}$ between i-th OCR token and j-th

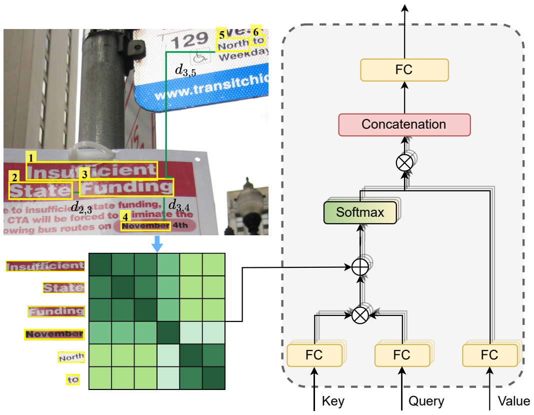


Figure 3: The architecture of the OCR-OCR query block.

OCR token is defined as follows,

$$d_{i,j} = \begin{cases} 0, & \text{if } \mathcal{B}_i \cap \mathcal{B}_j \neq \emptyset \\ \min\{|x_i - x_j| + |y_i - y_j|\}, & \text{otherwise}, \end{cases} \tag{4}$$

where $\mathcal{B}_i$ and $\mathcal{B}_j$ are the bounding boxes of the i-th OCR token and j-th OCR token, and $(x_i, y_i) \in \mathcal{B}_i$, $(x_j, y_j) \in \mathcal{B}_j$. If the $\mathcal{B}_i$ and $\mathcal{B}_j$ interact, their spatial distance would be 0. Moreover, (x_i, y_i) and (x_j, y_j) are the normalized coordinates of any points on the two bounding boxes respectively, and the spatial distance of two disjoint boxes is the minimum L1 distance between any pairs of (x_i, y_i) and (x_j, y_j). Then, we further obtain the spatial similarity $s_{i,j}$ between the i-th OCR token and the j-th OCR token based on their spatial distance $d_{i,j}$, which is formulated as follows,

$$s_{i,j} = 1 - \min(1, d_{i,j}). \tag{5}$$

To focus more on the contextual relationship between OCR tokens with close distances, we set the similarity of OCR tokens whose distance is greater than 1 to 0 and set the others to $1 - d_{i,j}$. The similarity between any two OCR tokens is between 0 and 1.

Next, we utilize the spatial self-attention blocks to realize the spatial similarity weighted contextual modeling among OCR tokens. In the traditional multi-head self-attention layer (Vaswani et al. 2017), the scaled dot-product attention is used to compute the attention scores. We modify the scaled dot-product attention to adapt to the spatial similarity between pairs of queries and keys. With the query q_i, the spatial attention score $\alpha_{i,j}$ of the key k_j is calculated as follows,

$$\alpha_{i,j} = \text{Softmax}\left(\frac{q_i (k_j)^T}{\sqrt{d_h}} + s_{i,j}\right) \tag{6}$$

where the d_h is a scale factor, and $s_{i,j}$ is the spatial similarity between the i-th element and the j-th element.

Finally, with the N OCR tokens $\{x_n^{\text{ocr}}\}_{n=1}^{N}$ and their spatial similarity matrix $\boldsymbol{s} \in R^{N \times N}$ (where $\boldsymbol{s}(i,j) = s_{i,j}$) as the input, we obtain the OCR tokens' contextual features $\boldsymbol{c}^{\text{ocr}} \in R^{d \times N}$ as follows,

$$\begin{aligned} \boldsymbol{x}^{\text{ocr}} &= [x_1^{\text{ocr}}, x_2^{\text{ocr}}, \cdots, x_N^{\text{ocr}}], \\ \boldsymbol{c}^{\text{ocr}} &= \text{LN}(\boldsymbol{c}^{\text{ocr}} + \text{MSSA}(\boldsymbol{x}^{\text{ocr}}, \boldsymbol{s})), \end{aligned} \tag{7}$$

where MSSA$(\cdot)$ and LN$(\cdot)$ mean the spatial self-attention in multi-head version and the layer normalization, respectively.

Region-OCR Query. As shown in Figure 2, following Section 3.2, the region-OCR query is built on the cross-attention blocks to implement the query between the salient region and the OCR tokens. The region-OCR query replaces the question words $\boldsymbol{f}^{\text{ques}}$ and uses the query result $\boldsymbol{f}^{\text{reg}} \in R^{d \times (K+1)}$ from Section 3.2 as the queries. Meanwhile, it replaces the the objects $\{x_m^{\text{obj}}\}_{m=1}^{M}$ and adopts the OCR tokens with contextual features $\{c_n^{\text{ocr}}\}_{n=1}^{N}$ as key-value pairs. The query result $\boldsymbol{f}^{\text{ocr}} \in R^{d \times (K+1)}$ is computed as follows:

$$\boldsymbol{f}^{\text{ocr}} = \text{LN}(\boldsymbol{f}^{\text{reg}} + \text{MCA}(\boldsymbol{f}^{\text{reg}}, \boldsymbol{c}^{\text{ocr}})), \tag{8}$$

where MCA$(\cdot)$ denotes the multi-head cross-attention, and LN$(\cdot)$ means the layer normalization. The input of the first token f_0^{reg} is the region-aware query vector, and the token's corresponding output f_0^{ocr} represents OCR tokens closely associated with the salient region. Therefore, we name the f_0^{ocr} the OCR-aware query vector.

3.4 Answer Generation

Following prior works (Zhu et al. 2021), answer generation module first selects the query results generated in Section 3.2 and 3.3, and then feed these selected query vectors and N OCR tokens with features $\{c_n^{\text{ocr}}\}_{n=1}^{N}$ into the generative decoder (Hu et al. 2020) to generate the answers. We adopt the sum of the multi-label binary cross-entropy loss (Yang et al. 2021) and our region-aware loss to train the model end-to-end. More details are given in the Appendix.

4 Experiments

4.1 Experimental Setup

Datasets and Metrics. The experiments are conducted on two widely-used benchmark datasets including TextVQA (Singh et al. 2019) and ST-VQA (Biten et al. 2019). Following the previous work (Hu et al. 2020; Biten et al. 2019), we adopt the same training/validation/test splits for TextVQA and ST-VQA datasets and use accuracy and ANLS as the evaluation metrics.

Implementation Details. For a fair comparison, we follow prior works (Hu et al. 2020; Lu et al. 2021) to extract features of the question, objects, and OCR tokens. We use the Adam optimizer with the initial learning rate of 1e-4, and we multiply the learning rate by a factor of 0.1 at the 14,000 iterations and the 15,000 iterations. For the pre-training setting, we follow TAP (Yang et al. 2021) to obtain the pre-training feature representations that are used to enhance our original features introduced in Section 3.1.

4.2 Comparsion with State-of-the-Arts

We compare our model ROQ with state-of-the-art approaches (Biten et al. 2022; Yang et al. 2021; Lu et al. 2021; Gao et al. 2021; Zhu et al. 2021; Liu et al. 2020; Kant et al. 2020; Hu et al. 2020), and the results are shown in Table 1 and Table 2. The ROQ outperforms existing methods on both TextVQA and ST-VQA dataset by a large margin.

Method	OCR System	Pre-Training Data	Extra Finetune	Val Acc.	Test Acc.
M4C (Hu et al. 2020)	Rosetta-en	✗	✗	39.40	39.01
SMA (Gao et al. 2021)	Rosetta-en	✗	✗	40.39	40.86
CRN (Liu et al. 2020)	Rosetta-en	✗	✗	40.39	40.96
LaAP-Net (Han, Huang, and Han 2020)	Rosetta-en	✗	✗	40.68	40.54
Non-Pretraining TAP (Yang et al. 2021)	Rosetta-en	✗	✗	39.55	-
LaTr-Small (Biten et al. 2022)	Rosetta-en	✗	✗	41.84	-
ROQ(Ours)	Rosetta-en	✗	✗	**43.32**	**43.11**
SA-M4C (Kant et al. 2020)	Google-OCR	✗	ST-VQA	45.40	44.60
Simple (Zhu et al. 2021)	SBD-Trans OCR	✗	✗	43.95	44.72
Simple (Zhu et al. 2021)	SBD-Trans OCR	✗	ST-VQA	45.53	45.66
ROQ(Ours)	SBD-Trans OCR	✗	✗	46.47	-
ROQ(Ours)	SBD-Trans OCR	✗	ST-VQA	**48.14**	**47.68**
Non-Pretraining TAP (Yang et al. 2021)	Microsoft-OCR	✗	ST-VQA	45.22	-
TAP (Yang et al. 2021)	Microsoft-OCR	TextVQA	✗	49.91	49.71
TAP (Yang et al. 2021)	Microsoft-OCR	TextVQA,ST-VQA	✗	50.57	50.71
LOGOS (Lu et al. 2021)	Microsoft and Rosetta-en OCR	Visual Genome	✗	50.79	50.65
LOGOS (Lu et al. 2021)	Microsoft and Rosetta-en OCR	Visual Genome	ST-VQA	51.53	51.08
LOGOS (Lu et al. 2021)	Microsoft OCR	Visual Genome	ST-VQA	50.02	-
ROQ(Ours)	Microsoft-OCR	✗	✗	50.30	49.87
ROQ(Ours)	Microsoft-OCR	✗	ST-VQA	51.35	50.47
ROQ(Ours)	Microsoft-OCR	TextVQA,ST-VQA	✗	**52.83**	**52.73**

Table 1: Comparison to state-of-the-art models on TextVQA dataset. Note that Non-Pretraining TAP is denoted by M4C† in the original paper (Yang et al. 2021).

Method	Val Acc.	Val ANLS	Test ANLS
M4C (Hu et al. 2020)	38.05	0.472	0.462
SA-M4C (Kant et al. 2020)	42.23	0.512	0.504
SMA (Gao et al. 2021)	-	-	0.466
CRN (Liu et al. 2020)	-	-	0.483
LaAP-Net (Han, Huang, and Han 2020)	39.74	0.497	0.485
LOGOS (Lu et al. 2021)	44.10	0.535	0.522
TAP (Yang et al. 2021)	45.29	0.551	0.543
ROQ	**47.11**	**0.562**	**0.558**
LOGOS(w/ TextVQA)	48.63	0.581	0.579
ROQ(w/ TextVQA)	**51.14**	**0.612**	**0.593**

Table 2: Comparison to state-of-the-art methods on ST-VQA dataset.“w/ TextVQA” refers that the models are trained on both TextVQA and ST-VQA datasets.

TextVQA dataset. As shown in Table 1, our ROQ consistently outperforms prior works under different settings, including different OCR engines, and pre-training and extra finetuning datasets: (1) For the non-pretraining setting, ROQ outperforms state-of-the-art models under different OCR engines, which improves prior best-performing models using Rosetta-en OCR by 1.48%, SBD-Trans OCR by 2.61% (from 45.53% to 48.14%), and Microsoft OCR by 6.13% (from 45.22% to 51.35%) on validation set, respectively. (2) Although compared with pre-training models like TAP, our non-pretrained ROQ still achieves a comparable accuracy and even a slight performance gain. (3) Moreover, our ROQ can also benefit from pre-training strategy, which boosts the accuracy to 52.83% and 52.73% on the validation and test sets, respectively. The ROQ outperforms the pre-training TAP and LOGOS by 2.04% and 2.71% on the validation set, respectively. (4) Similar to the founding in prior works, ROQ benefits from the joint finetuning with the ST-VQA dataset, which obtains a performance gain of 1.05% on the validation set.

#	Method	Acc.(val)
1	Baseline	44.50
2	1 + question-region query	46.74
3	2 + region-aware supervision	47.52
4	3 + OCR-OCR query	49.04
5	4 wo spatial attention + region-OCR query	49.22
6	4 + region-OCR query = Full	50.30
7	6 + pre-training	52.83

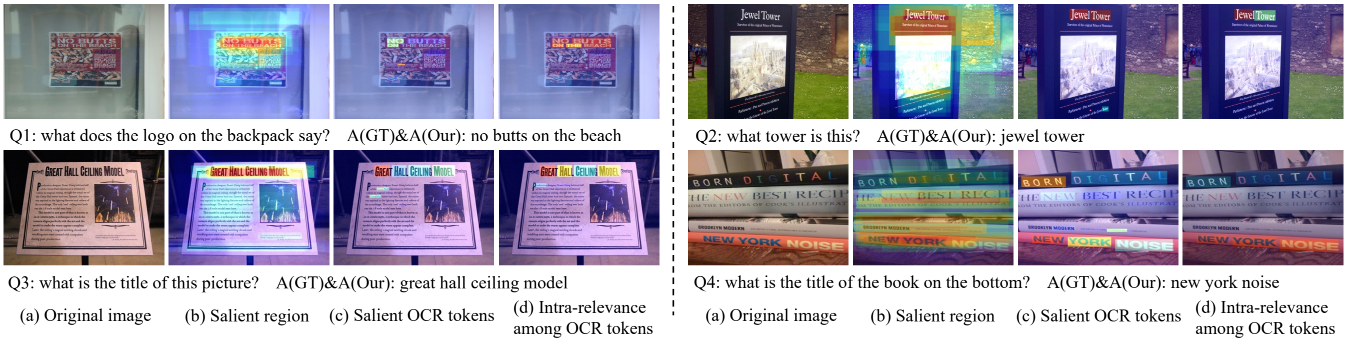
Table 3: Ablation studies of ROQ on TextVQA dataset.

ST-VQA dataset. As shown in Table 2, our ROQ significantly improves the accuracy and ANLS on the ST-VQA dataset compared to existing works. Compared with pre-trained TAP and LOGOS, our ROQ without pre-training strategy improves accuracy by 1.17% on validation set. When using the TextVQA dataset as additional training data, ROQ outperforms LOGOS by 2.51% accuracy, even though LOGOS uses both Microsoft OCR and Rosetta OCR system while we use the Microsoft OCR engine only.

4.3 Ablation Study

To evaluate the effectiveness of the Region Query module and OCR Query module, we have evaluated five additional variants for comparison. The results are shown in Table 3.

Baseline. The baseline is the M4C (Hu et al. 2020) with Microsoft OCR system, which fuses the feature embeddings of



Q1: what does the logo on the backpack say? A(GT)&A(Our): no butts on the beach

Q2: what tower is this? A(GT)&A(Our): jewel tower

Q3: what is the title of this picture? A(GT)&A(Our): great hall ceiling model

Q4: what is the title of the book on the bottom? A(GT)&A(Our): new york noise

(a) Original image (b) Salient region (c) Salient OCR tokens (d) Intra-relevance among OCR tokens

(a) Original image (b) Salient region (c) Salient OCR tokens (d) Intra-relevance among OCR tokens

Figure 4: Qualitative results show explainable visual evidences of the stepwise query processes.

question words, objects, and OCR into a multi-modal transformer for answer generation.

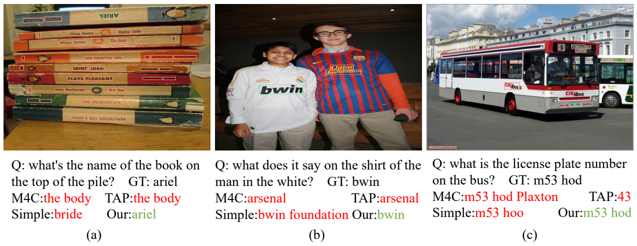
Ablations on Region Query Module. As shown in Table 3, the question-region query (row 2) helps improve the accuracy by 2.24% compared with the baseline (row 1). The sharp performance gain indicates the effectiveness of the question-region query, which grounds the salient regions relevant to the question. We further apply the region-aware supervision (row 3) to the model with question-region query (row 2). The region supervision boosts the 0.78% accuracy, demonstrating the importance of region priors.

Ablations on OCR Query Module. Table 3 shows that the model with the OCR-OCR query (row 4) outperforms the model without the query (row 3) by 1.52%, which clearly validates the effectiveness of building the contexts among OCR tokens via the spatial attention modules. When we further add the region-OCR query (row 6) to row 4, the performance improves by 1.26% compared to row 4. The experiment shows the region-OCR query's effectiveness in facilitating the model to find OCR tokens closely associated with the salient region. Moreover, we evaluate the necessity of spatial attention in the OCR-OCR query by replacing the spatial self-attention with the traditional self-attention (row 5). The spatial attention brings 1.08% improvement, which validates the usefulness in modeling the spatial relationships among OCR tokens.

4.4 Qualitative Evaluation

We visualize the explainable visual evidences of the stepwise query processes (see Figure 4) and predicted results (see Figure 5) to explore in-depth insights into the ROQ.

Figure 4 shows that the ROQ can generate interpretable intermediate processes for stepwisely querying from the question to the salient region to the salient OCR tokens. For the first sample, the ROQ performs queries from the "logo on the backpack" to the salient region to the salient OCR tokens. First, it successfully grounds the salient region ("logo on the backpack") via the question-region query shown in (b). Then, it identifies the salient tokens relevant to the salient region via the region-OCR query shown in (c). The target OCR tokens ("NO", "BUTTS", "ON", "THE" and "BEACH") are closely tied together via the OCR-OCR query shown in (d). For the second and third samples, when answering the tower name or book title, the ROQ highlights the corresponding regions (shown in b) and distinguishes the target OCR tokens ("Jewel Tower" or "Anthony Shaffers") from other tokens. For the fourth sample, the ROQ is able to first identify the book on the bottom and find all target OCR tokens ("new york noise").



Q: what's the name of the book on the top of the pile? GT: ariel
M4C:the body TAP:the body
Simple:bride Our:ariel

(a)

Q: what does it say on the shirt of the man in the white? GT: bwin
M4C:arsenal TAP:arsenal
Simple:bwin foundation Our:bwin

(b)

Q: what is the license plate number on the bus? GT: m53 hod
M4C:m53 hod Plaxton TAP:43
Simple:m53 hoo Our:m53 hod

(c)

Figure 5: Qualitative results showing the effects of the RQ and OQ modules of our model ROQ.

To demonstrate the effects of our ROQ, we present three examples in Figure 5. In Figure 5(a), our ROQ can locate the book on the top and predict the answer "ariel". In Figure 5(b), the ROQ avoids predicting the redundant token "foundation" like the Simple, which benefits from our Region Query module. The Region Query module facilitates the model to focus on the salient region ("the man in the white") and identifies OCR tokens closely associated with the salient region. Figure 5(c) shows that our ROQ not only identifies the location of the license plate number but also excludes wrong tokens. The TAP does not find the location of license plate number, while the M4C model includes the redundant token "plaxton". More qualitative results are included in the Appendix.

5 Conclusion

We propose the ROQ network for TextVQA. The Region Query module of ROQ is to find the salient region relevant to answering the question. The salient region serves as an intermediate between the question and OCR tokens. The OCR Query module of ROQ is to locate salient OCR tokens from the salient region. It first groups homogeneous OCR tokens by representing their spatial and semantic relations and then queries the OCR tokens according to the salient region. The ROQ significantly outperforms state-of-the-art models.

References

Almazán, J.; Gordo, A.; Fornés, A.; and Valveny, E. 2014. Word spotting and recognition with embedded attributes. *IEEE transactions on pattern analysis and machine intelligence*, 36(12): 2552–2566.

Andreas, J.; Rohrbach, M.; Darrell, T.; and Klein, D. 2016. Learning to compose neural networks for question answering. *arXiv preprint arXiv:1601.01705*.

Biten, A. F.; Litman, R.; Xie, Y.; Appalaraju, S.; and Manmatha, R. 2022. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16548–16558.

Biten, A. F.; Tito, R.; Mafla, A.; Gomez, L.; Rusinol, M.; Valveny, E.; Jawahar, C.; and Karatzas, D. 2019. Scene text visual question answering. In *Proceedings of the IEEE/CVF international conference on computer vision*, 4291–4301.

Bojanowski, P.; Grave, E.; Joulin, A.; and Mikolov, T. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5: 135–146.

Cao, Q.; Liang, X.; Li, B.; Li, G.; and Lin, L. 2018. Visual Question Reasoning on General Dependency Tree. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Gao, C.; Zhu, Q.; Wang, P.; Li, H.; Liu, Y.; Van den Hengel, A.; and Wu, Q. 2021. Structured multimodal attentions for textvqa. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Gao, D.; Li, K.; Wang, R.; Shan, S.; and Chen, X. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12746–12756.

Han, W.; Huang, H.; and Han, T. 2020. Finding the evidence: Localization-aware answer prediction for text visual question answering. *arXiv preprint arXiv:2010.02582*.

Hu, R.; Andreas, J.; Rohrbach, M.; Darrell, T.; and Saenko, K. 2017. Learning to reason: End-to-end module networks for visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 804–813.

Hu, R.; Rohrbach, A.; Darrell, T.; and Saenko, K. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10294–10303.

Hu, R.; Singh, A.; Darrell, T.; and Rohrbach, M. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9992–10002.

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

Kant, Y.; Batra, D.; Anderson, P.; Schwing, A.; Parikh, D.; Lu, J.; and Agrawal, H. 2020. Spatially aware multimodal transformers for textvqa. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, 715–732. Springer.

Li, L.; Gan, Z.; Cheng, Y.; and Liu, J. 2019a. Relation-Aware Graph Attention Network for Visual Question Answering. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10312–10321.

Li, L. H.; Yatskar, M.; Yin, D.; Hsieh, C.-J.; and Chang, K.-W. 2019b. VisualBERT: A Simple and Performant Baseline for Vision and Language. *ArXiv*, abs/1908.03557.

Liu, F.; Xu, G.; Wu, Q.; Du, Q.; Jia, W.; and Tan, M. 2020. Cascade reasoning network for text-based visual question answering. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4060–4069.

Lu, J.; Batra, D.; Parikh, D.; and Lee, S. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NeurIPS*.

Lu, X.; Fan, Z.; Wang, Y.; Oh, J.; and Rosé, C. P. 2021. Localize, group, and select: Boosting text-vqa by scene text modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2631–2639.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep Contextualized Word Representations. In *NAACL*.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Santoro, A.; Raposo, D.; Barrett, D. G.; Malinowski, M.; Pascanu, R.; Battaglia, P.; and Lillicrap, T. 2017. A simple neural network module for relational reasoning. *Advances in neural information processing systems*, 30.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8317–8326.

Su, W.; Zhu, X.; Cao, Y.; Li, B.; Lu, L.; Wei, F.; and Dai, J. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. *ArXiv*, abs/1908.08530.

Tan, H. H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *EMNLP*.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Liu, Y.; Shen, C.; Ng, C. C.; Luo, C.; Jin, L.; Chan, C. S.; van den Hengel, A.; and Wang, L. 2020. On the General Value of Evidence, and Bilingual Scene-Text Visual Question Answering. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10123–10132.

Yang, S.; Li, G.; and Yu, Y. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4644–4653.

Yang, Z.; Lu, Y.; Wang, J.; Yin, X.; Florencio, D.; Wang, L.; Zhang, C.; Zhang, L.; and Luo, J. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8751–8761.

Zhu, Q.; Gao, C.; Wang, P.; and Wu, Q. 2021. Simple is not Easy: A Simple Strong Baseline for TextVQA and TextCaps. In *AAAI*.