Two-stage Multimodality Fusion for High-performance Text-based Visual Question Answering

Bingjia Li[1], Jie Wang[2], Minyi Zhao[1], and Shuigeng Zhou[1]✉

[1] Shanghai Key Lab of Intelligent Information Processing, and School of Computer Science, Fudan University, Shanghai 200438, China
{bjli20, zhaomy20, sgzhou}@fudan.edu.cn
[2] ByteDance, China
wangjie.bernard@bytedance.com

Abstract. Text-based visual question answering (TextVQA) is to answer a text-related question by reading texts in a given image, which needs to jointly reason over three modalities — question, visual objects and scene texts in images. Most existing works leverage graph or sophisticated attention mechanisms to enhance the interaction between scene texts and visual objects. In this paper, observing that compared with visual objects, the question and scene text modalities are more important in TextVQA while both layouts and visual appearances of scene texts are also useful, we propose a two-stage multimodality fusion based method for high-performance TextVQA, which first semantically combines the question and OCR tokens to understand texts better and then integrates the combined results into visual features as additional information. Furthermore, to alleviate the redundancy and noise in the recognized scene texts, we develop a denoising module with contrastive loss to make our model focus on the relevant texts and thus obtain more robust features. Experiments on the TextVQA and ST-VQA datasets show that our method achieves competitive performance without any large-scale pre-training used in recent works, and outperforms the state-of-the-art methods after being pre-trained.

Keywords: TextVQA · scene text recognition · multimodal information fusion · contrastive learning.

1 Introduction

Nowadays, numerous methods [2, 30, 15, 24] have been proposed to solve the task of visual question answering (VQA) [3], which is to answer questions about images. However, these methods fail in answering text-related questions as they usually focus on objects and scenes while ignoring texts in the images. Actually, text matters in real life as it appears ubiquitously in practical images like advertisements, conveying valuable information that is essential for scene understanding and reasoning. Thus, text-based visual question answering (TextVQA) [28,

Q1: What is the number of the card on right? (a)

Q2: What is written in green font? (b)

Q3: What is written on the top blue sign? (c)

Fig. 1. Some examples of TextVQA. The 1st, 2nd and 3rd rows show the original images, all the scene text areas extracted by an OCR system in each image, and the questions, respectively. All these questions can be correctly answered with only these cropped scene texts in the images. To correctly answer these questions, the model requires to first understand the semantics of the questions and scene texts, and then use some additional scene text information: (a) layout information for Q1, (b) visual appearance for Q2, (c) both layout and visual appearance for Q3.

5, 23] is gaining popularity in recent years as an extension of VQA to answer text-related questions by reading and understanding scene texts in images.

Motivation. In general, the TextVQA task requires the model to read scene texts in images by an OCR system and jointly reason over three modalities - question, visual objects and scene texts. In our points of view, TextVQA research faces two major technical challenges as follows:

1) *How to effectively exploit multimodal information*? Most existing works [19, 9, 14, 11] use an object detector to extract global visual objects and utilize graph or complicated attention mechanisms to enhance the interaction between OCR tokens and visual objects. However, Wang *et al.* [31] pointed out that the accuracy of M4C [12] is almost unaffected after discarding the visual object modality, because the information of visual objects is not utilized well. Then, do global visual objects matter in TextVQA? As we all know, there are three modalities available for TextVQA: question, visual objects and scene texts in each image. And each piece of scene texts contains textual content, layouts and visual appearances (*font, color, background etc.*). Our preliminary study shows that compared to visual objects, scene texts are more important, and for high performance TextVQA, both layouts and visual appearances are indispensable. Concretely, we found that over 70% questions can be answered by using only the scene text areas of the images, and 60% questions can still be correctly answered even after discarding the visual appearances of scene texts while keeping only the layout information. As layouts can provide contextual information, which are not only critical to answer position-related questions, but also helpful for text understanding. Fig. 1 shows some examples of TextVQA. We can see that to answer Q1 (Fig. 1(a)), layouts of scene texts are required, and visual appear-

ances of scene texts are critical to Q2 (Fig. 1(b)). While to correctly answer Q3 (Fig. 1(c)), besides layouts, the blue background of the text *"Park ave"* is also indispensable, without which the answer may become *"STOP"* mistakenly.

2) *How to select proper scene texts to answer the question*? Existing methods try to use as many OCR tokens as possible in the image to provide enough semantic information and make sure that the correct answer texts are included in the input sequence. Although more OCR tokens bring more contextual information, which does help improve the performance, they also introduce noise inevitably. Note that not all scene texts are relevant to the question. For example, in Fig. 1(a), only texts containing numbers (e.g. *"11870"* or *"15"*) need to be considered as the question is about number. Considering that too many unrelated OCR tokens may confuse the model, especially in text-intensive scenarios. An ideal solution is to select the texts most relevant to the question when semantic relationship is not enough to support question answering.

Solution and contributions. In this paper, we pay more attention to the question and scene text modalities. On the one hand, to address the first challenge mentioned above, we propose a two-stage multimodality fusion module to take full advantage of the textual content, layouts and visual appearances of scene texts. In the first fusion stage, our model tries to understand the question and scene texts by combining them and the layouts of scene texts as contextual information with the help of LayoutLM [35]. After the textual and contextual features interaction, visual features are then included in the second fusion stage to handle questions that need the help of visual clues from scene texts. On the other hand, in order to handle the second challenge above, i.e., reducing redundancy and noise in the recognized scene texts, we develop a denoising module that masks irrelevant OCR tokens and uses contrastive loss to integrate the features of positive samples. In such a way, our model is able to focus on relevant texts and obtain more robust features.

In summary, the contributions of this paper are:

- Observing that the question and scene text modalities are of first importance in TextVQA, while both layouts and visual appearances of scene texts are useful, we propose a two-stage multimodality fusion based method to take full advantage of these information to boost TextVQA.
- We develop a denoising module with contrastive loss as an auxiliary task to reduce the redundancy and noise of recognized scene texts and thus make the model focus on the relevant texts and get more robust features.
- We validate the effectiveness and superiority of our method on the TextVQA and ST-VQA benchmarks. Experimental results show that our method achieves competitive results without any large-scale pre-training used in recent works, and outperforms the state-of-the-art methods after being pre-trained.

2 Related Work

TextVQA aims to answer text-related questions by first reading scene texts in images and then reasoning over three modalities — question, visual objects

and scene texts. As a pioneer work, Singh *et al.* [28] proposed the first dataset TextVQA with a framework LoRRA by extending the VQA model Pythia [13] with an OCR attention branch. Later, several other datasets were built with texts of different scenarios, *e.g.* ST-VQA [5] in daily natural scenes, OCR-VQA [23] of book covers, STE-VQA [32] with bilingual texts and M4-ViteVQA [39] in video text understanding.

Recent works [10, 12, 9, 19, 14, 11, 40, 36, 29, 21, 31, 37, 38, 4] have tried to improve the performance of TextVQA by various network architectures, more powerful OCR systems or large-scale datasets. Among them, M4C [12] utilizes multimodal transformers to fuse all modalities with a dynamic pointer network supporting multi-step answer decoding, which is the basis of most later works. With M4C, SA-M4C [14] proposes a spatiality-aware self-attention layer and handles different spatial relationships by different attention heads. Similarly, some other works [10, 9, 19, 40, 38] leverage graph or complicated attention mechanisms to emphasize the relationships between objects and OCR tokens, but the performance improvement is mainly gained by stronger OCR systems. TAP [36] is the first work to introduce pre-training to this task and pre-trains the model with three auxiliary tasks. With the help of the Microsoft-OCR system and the proposed large-scale dataset OCR-CC, it significantly boosts the TextVQA performance. LOGOS [21] enhances the model's understanding ability with two grounding tasks to better localize the key information of the image. LaTr [4] bases its architecture on T5 [25] and applies the pre-training strategy on large-scale scanned documents.

Though some latest works emphasize the significance of question and scene texts, yet none of them take full advantage of these two modalities with layouts and visual appearances of scene texts simultaneously. In this paper, we propose a two-stage multimodality fusion based method to comprehensively exploit such information. In addition, we also develop a denoising module with contrastive loss to reduce the redundancy and noise of recognized scene texts, which makes the model focus on the relevant texts. Our experiments verify the effectiveness and advantage of the proposed method.

3 Methodology

3.1 Overview

Fig. 2 shows the architecture of our method, which mainly consists of three components: multimodal feature extraction, two-stage multimodality fusion and denoising. Besides, an optional pre-training component is considered. Given a sample X with an image I and a text-related question Q, we first extract features of the question, scene text and visual object modalities. These features are then progressively fused and reasoned with our two-stage multimodality fusion module, where the first stage focuses on textual and layout information from Q and scene texts in I, and the second stage includes visual appearances of scene texts and utilizes global visual objects as auxiliary information. The denoising module first masks the input OCR tokens, and then utilizes the masked result

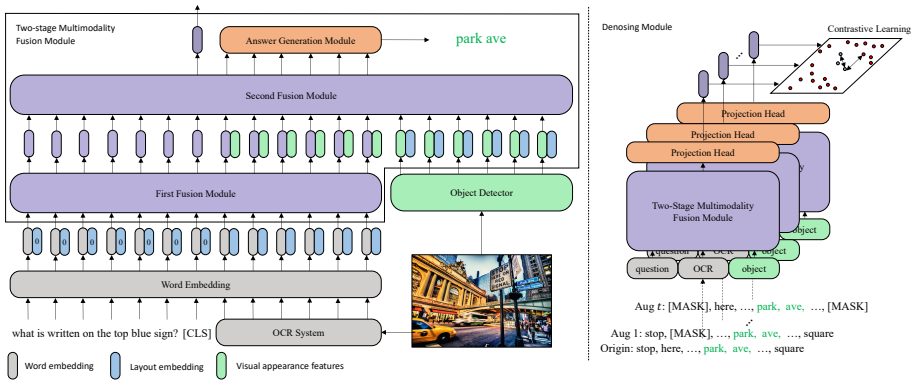


Fig. 2. The architecture of our method. After features of different modalities are extracted, they are fused and reasoned progressively with the two-stage fusion module. The output features of scene texts will be used for further answer decoding with the help of the denoising module, which masks the input and uses contrastive loss as an auxiliary task.

and a contrastive loss as an auxiliary task to make the model focus on the relevant texts. Optionally, our model can be further pre-trained on the question and scene text modalities to boost the performance. Note that in our method, the denoising module is used only in the fine-tuning stage.

3.2 Multimodal Features

OCR features. After extracting OCR tokens in image I with an OCR system [7], previous works [12, 19, 36] obtain multiple features of OCR tokens with various pre-trained models and add them together before fusion. Unlike them, here we categorize the features into three types: layout embeddings, visual appearance features and word embeddings. Let $O = \{O_i\}_{i=1}^{N}$ be the OCR tokens after tokenization, where N is the length of the sequence. The layout embedding $x_i^{ocr,l}$ of the i-th OCR token O_i indicates its size and 2-D spatial position, which is defined as follows:

$$x_i^{ocr,l} = E_x(x_i^0) + E_y(y_i^0) + E_x(x_i^1) + E_y(y_i^1) + E_w(w_i) + E_h(h_i) \quad (1)$$

where E_x, E_y, E_w and E_h are learnable embedding layers, $(x_i^0,\ y_i^0)$ denotes the coordinates of the upper left corner, $(x_i^1,\ y_i^1)$ denotes the coordinates of the bottom right corner, and w_i and h_i correspond to the width and height of the detected bounding box. All the coordinates have been scaled to 0-1000. For each OCR bounding box, we use Faster R-CNN [26] to extract the visual appearance feature $x_i^{ocr,v}$. As for texts, the word embedding is represented as $x_i^{ocr,t} = E_t(O_i)$, where E_t is a learnable embedding layer. The final representations of the detected OCR tokens are then defined as $X^{ocr,l} = \{x_i^{ocr,l}\}_{i=1}^{N}$, $X^{ocr,v} = \{x_i^{ocr,v}\}_{i=1}^{N}$ and $X^{ocr,t} = \{x_i^{ocr,t}\}_{i=1}^{N}$.

Question features. Let $Q = \{Q_i\}_{i=1}^{L}$ be the sequence of question tokens after tokenization, where L is the sequence length. For the i-th token Q_i, the word embedding is represented as $x_i^{q,t} = E_t(Q_i)$, which shares the same embedding layer as the OCR word embedding. And the final representation is defined as $X^{q,t} = \{x_i^{q,t}\}_{i=1}^{L}$.

Object features. As we have mentioned above, the visual object modality is not the key factor in the TextVQA task, so we just use it as a supplementary. To unify the inputs, we use the same Faster R-CNN to detect visual objects $V = \{V_i\}_{i=1}^{M}$ in image I, where M is the number of objects, and then extract features of each object V_i as $x_i^{obj,v}$. Similar to the OCR embedding, we obtain the layout embedding $x_i^{obj,l} = E_x^{'}(x_i^{0'}) + E_y^{'}(y_i^{0'}) + E_x^{'}(x_i^{1'}) + E_y^{'}(y_i^{1'}) + E_w^{'}(w_i^{'}) + E_h^{'}(h_i^{'})$. As the features of visual objects are applied only in the second fusion stage, here we just sum them up as $x_i^{obj} = W_1 x_i^{obj,v} + x_i^{obj,l}$, where W_1 is a linear layer to control the dimension. The final object representation is defined as $X^{obj} = \{x_i^{obj}\}_{i=1}^{M}$.

3.3 Two-stage Multimodality Fusion

After the features of different modalities are extracted, a common routine is to add the unimodal features together and fuse them with a multimodal transformer, which is not effective enough in our work. To take full advantage of the text, layout and visual appearance information, we propose two-stage multimodality fusion.

In the **first fusion stage**, the model focuses on understanding the texts, including the question and scene texts. We believe that there is a semantic connection between the scene texts and the question. Most questions are semantically closely related to the texts in the image, so texts can provide valuable clue for answering the question, which constitutes the basis of the TextVQA task. So we put these two modalities at the highest priority and jointly understand them. Besides, previous works [35, 34] on document understanding have shown the value of layout information, which provides contextual information. Similarly in natural scenes, a specific OCR token's 2-D position and positional relationship with its contextual tokens help us to understand the OCR token.

Here, we base the first stage fusion on LayoutLM [35]. LayoutLM is a BERT-like model that incorporates the visually rich layout information and align it with the input texts. With both word embeddings and layout embeddings as input, it is pre-trained with the masked visual language model (MVLM) on the document dataset IIT-CDIP Test Collection 1.0 [18], which can also bring more knowledge to our model. As there is a second fusion stage, we use only the first 6 layers of LayoutLM with the weights from HuggingFace [33] as initialization. To jointly reason over the question and OCR tokens, as shown in Fig. 3, we concatenate the features of the question and OCR tokens. The input of OCR tokens is defined as $X^{ocr,t} + X^{ocr,l}$ and we add a special [CLS] token as the beginning of OCR tokens to represent the whole texts, which will be used in the denoising module. As to the question, because there is no layout embedding, we set all the coordinates as zero. Finally, we input the unified features into the

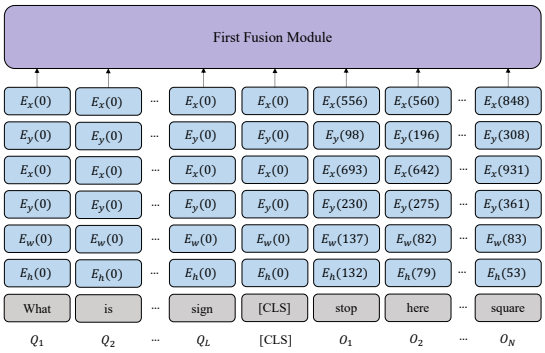


Fig. 3. The input of the first multimodality fusion stage. We concatenate the features of the question and OCR tokens together, including both word embeddings and layout embeddings. All the coordinates for the question are set to zero.

fusion module and obtain

$$[X^{q'}; X^{ocr'}] = \Phi([X^{q,t}; X^{ocr,t} + X^{ocr,l}]) \tag{2}$$

where $X^{q'}$ and $X^{ocr'}$ are the semantically enhanced features for the question and OCR tokens respectively, $\Phi(\cdot)$ is the first fusion module LayoutLM and $[;]$ is a concatenation operation.

In the **second fusion stage**, the visual appearance features of scene texts are introduced as additional information to help handle questions that require visual clues from scene texts. Here, we combine them with the semantic features obtained in the first fusion stage. In addition, we also use the visual objects as an aid and then fuse the three modalities with a multimodal transformer. The output features are represented as

$$[X^{q''}; X^{ocr''}; X^{obj'}] = \Psi([X^{q'}; X^{ocr'} + W_2 X^{ocr,v}; X^{obj}]) \tag{3}$$

where $X^{q''}$, $X^{ocr''}$ and $X^{obj'}$ are the outputs for the question, scene texts and visual objects, respectively, $\Psi(\cdot)$ is the second fusion module and W_2 is a linear layer that projects $X^{ocr,v}$ to the same dimension as $X^{ocr'}$. Finally, $X^{ocr''}$ will be used in the denoising module and for further answer decoding.

3.4 Denoising Module

More scene texts bring more noise, considering only the semantic relationship between the question and OCR tokens is not enough to search for the relevant texts for answering the question. To relieve this issue, we design a denoising module, which augments the samples by randomly masking the input OCR tokens to reduce modality interaction and understanding difficulty, and utilizes contrastive learning to make the model focus on the relevant texts. In the following, we introduce this module in detail.

Masking strategy. The text of correct answer provides important information, without which the text-related question about the image cannot be answered, so we avoid masking the tokens appearing in the answer. For the remaining OCR tokens, we randomly mask them with a certain probability. It is worth mentioning that masking the OCR tokens by totally deleting them may lead to confusion. For example, for some questions like *"What is the second word on the page?"*. If the first word is masked and we simply mask the final feature of it, the original second word will no longer be the correct answer as it has become the first one. Therefore, we replace the masked OCR tokens with a special [MASK] token while keeping the corresponding layout and visual appearance features. In this way, our model gets the information that there exists a word but does not know what the word is, which effectively reduces the interaction and understanding difficulty. Besides, in order to enlarge the differences between the original sample and the augmented ones, we use **whole word masking**. For example, in our model the input OCR word *"vegeburger"* is tokenized into three subtokens *"ve"*, *"##ge"* and *"##burger"* and the other features will be duplicated for each token. Once the word is chosen, we mask these subtokens of the word separately.

Contrastive learning. With this masking strategy, for a batch of l samples, we augment each sample t times with a probability p. Given a pair of question and answer (a training sample), these augmented samples can be seen as positive samples of the original one while the rest are negative samples. As mentioned above, for the i-th sample X_i in the batch we can get the output features $X_i^{ocr''}$ of scene texts with the two-stage fusion module. We take the feature of the first [CLS] token as the representation of the whole sequence and project it to $z_i \in \mathbb{R}^d$ with a contrastive projection head, which is composed of two linear layers. Similarly, we can get the contrastive output of all the augmented samples using the same module. Then, we utilize a contrastive loss to constrain the distance between positive samples in the latent space as follows:

$$\mathcal{L}_{cont} = \sum_{i \in B} \mathcal{L}_i^{cont} = -\sum_{i \in B} \log \frac{\exp(z_i \cdot z_{j(i)}/\tau)}{\sum\limits_{a \in A(i)} \exp(z_i \cdot z_a/\tau)} \tag{4}$$

where $B = \{1, ..., l * (t+1)\}$ is the index of samples in the batch after augmentation, '$\cdot$' represents the dot production, $j(i)$ denotes the indexes of positive samples of the i-th sample, $A(i) = B \backslash i$, and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter. As only the correct answer will never be masked, it plays a key role in the loss, and the model is then forced to pay more attention on it.

Training. For the main task, we use teacher-forcing technique [17] and multi-label binary cross-entropy loss $\mathcal{L}_{bce}$ to train the model, which is defined as follows:

$$\begin{aligned} \mathcal{L}_{bce} &= -y_{gt}\log(y_{pred}) - (1 - y_{gt})\log(1 - y_{pred}) \\ y_{pred} &= \frac{1}{1 + \exp(-f(X^{ocr''}))} \end{aligned} \tag{5}$$

where y_{gt} is the ground-truth target, y_{pred} is the final prediction of our model and $f(\cdot)$ is the answer generation module following previous works of the TextVQA

task. Finally, the total loss $\mathcal{L}$ is defined as a linear combination of the two losses above as follows:

$$\mathcal{L} = \lambda_1 * \mathcal{L}_{bce} + \lambda_2 * \mathcal{L}_{cont} \tag{6}$$

where λ_1 and λ_2 are hyper-parameters to trade-off the two losses.

3.5 Pre-training

So far, we have introduced the main components of our method. Recent works [36, 4] tend to take advantage of pre-training to improve the performance. As an option, our model also can be pretrained to achieve higher performance. Here, we conduct the masked language modeling (MLM) task on our model. Unlike TAP [36], we directly per-train our model on the question and scene text modalities with both layout and visual appearance features, but do not introduce additional texts to enhance the question modality.

LayoutLM has been pre-trained with both texts and layout information of documents. Considering the domain gap between documents and daily natural scenes, and the fact that natural scenes contain more visual information, we further pre-train our model to align all these features. In the pre-training stage, we randomly mask each text token of both question and scene texts with a probability 15%. The masked tokens are replaced by a special [MASK] with 80% probability, a random token with 10% probability, and remain unchanged with 10% probability. Note that we only mask the input tokens while keeping the corresponding 2-D position embeddings and visual appearance features as we believe that they provide additional contextual information. Tokens with close spatial relationship or sharing similar visual appearance tend to be more related in semantics. Then, the model is required to recover the masked word W_{mask} with two fully-connected layers. In our experiments, we find that the image-text matching (ITM) task used in TAP brings no performance improvement, so we do not use it. Finally, we pre-train our model with cross-entropy loss and use our denoising module in the fine-tuning stage.

4 Experiments

4.1 Datasets and Evaluation Metrics

TextVQA is the first proposed large-scale dataset for the TextVQA task, which contains a total of 45,336 questions about 28,408 images sampled from the Open Image dataset [16]. All the questions are related to the texts in the images, and each of them has 10 answers provided by 10 different annotators.

ST-VQA is similar to TextVQA, it consists of 23,038 images collected from more diverse sources. There are in total 31,791 questions while each question has up to two answers. All the questions can only be answered based on the texts that appear in the images. We report our results on the open dictionary task (task 3), which contains 18,921 training images and 2,971 test images. Following

previous works [12], we split the training images into a training set with 17,028 images and a validation set with 1,893 images.

TextCaps and **OCR-CC** are introduced additionally during the pre-training stage. TextCaps [27] reuses images from the TextVQA dataset and attaches 145,329 captions to them, while OCR-CC was proposed along with TAP [36], which contains 1.367 million text-related image-caption pairs. These captions play the same role as the questions to increase data in the pre-training stage.

Evaluation metrics. For the TextVQA dataset, we report the VQA accuracy [3] measured via the soft voting of the 10 answers. For the ST-VQA dataset, besides VQA accuracy we also report the Average Normalized Levenshtein Similarity (ANLS), which is defined as $1 - d_L(a_{pred}, a_{gt})/\max(|a_{pred}|, |a_{gt}|)$, where a_{pred} and a_{gt} are the predicted and ground-truth answers, while d_L refers to the edit distance. The final result is the average of all scores with those below the threshold 0.5 being truncated to 0.

4.2 Implementation Details

We implement our method based on the code of TAP [36]. The maximum input sequence lengths of question tokens, OCR tokens and object numbers are set to L=20, N=300, and M=100 respectively. For our two-stage fusion module, the base model contains a LayoutLM with 6 layers and a multimodal transformer with 4 layers and 12 attention heads, while '†' refers to a larger model with 8 and 12 layers adopted in LayoutLM and the multimodal transformer respectively. The LayoutLM is initialized with weights from HuggingFace [33] and the multimodal transformer is initialized from scratch. The dimension in the joint embedding space is set to 768. In our denoising module, we use d=128 for the output features of the contrastive projection head. Besides, we set $t = 3$ and $p = 0.15$ to augment each sample. For other hyper-parameters, we use τ=0.07, λ_1=1 and λ_2=1 by experience.

During the training stage, we adopt AdamW [20] as our optimizer and set the batch size l to 32. The learning rate is 1e-4 for the multimodal transformer and 1e-5 for the pre-trained LayoutLM. The warm-up learning ratio and warm-up iteration are set as 0.2 and 1,000. In the pre-training stage, we pre-train the model for 24,000 iterations when only TextVQA or ST-VQA datasets are used, and 240,000 iterations when TextCaps and OCR-CC datasets are included. In the fine-tuning stage, the model is further trained for another 30,000 iterations.

4.3 Experimental Results

Results on TextVQA. Following previous works [36], we evaluate our method under two different settings: the first is the constrained setting that uses only TextVQA for training and Rosetta [7] for OCR detection, the second is the unconstrained setting. All results are presented in Tab. 1.

As shown in Tab. 1, the **top part** of the table reports the results under the constrained setting. Because TAP [36] uses 100 OCR tokens, Latr [4] uses 512 OCR tokens and previous works tend to use 50 OCR tokens, to make a fair

Table 1. Performance comparison on the TextVQA dataset. The top part reports the results in the constrained setting that only uses TextVQA for training and Rosetta for OCR detection, while the bottom part displays the results in the unconstrained setting. We compare our method with several state-of-the-art methods, and our method clearly outperforms them.

Method	OCR system	Extra data	Val acc.	Test acc.
M4C [12]	Rosetta-en	✗	39.40	39.01
SMA [9]	Rosetta-en	✗	40.05	40.66
CRN [19]	Rosetta-en	✗	40.39	40.96
LaAP-Net [11]	Rosetta-en	✗	40.68	40.54
BOV [37]	Rosetta-en	✗	40.90	41.23
TAP [36]	Rosetta-en	✗	44.06	-
LaTr [4]	Rosetta-en	✗	44.06	-
Ours (100 tokens)	Rosetta-en	✗	44.20	-
Ours	Rosetta-en	✗	**44.75**	-
M4C [12]	Rosetta-en	ST-VQA	40.55	40.46
LaAP-Net [11]	Rosetta-en	ST-VQA	41.02	40.54
SA-M4C [14]	Google-OCR	ST-VQA	45.40	44.6
SMA [9]	SBD-Trans OCR	ST-VQA	-	45.51
BOV [37]	SBD-Trans OCR	ST-VQA	46.24	46.96
TAP [36]	Microsoft-OCR	✗	49.91	49.71
TAP [36]	Microsoft-OCR	ST-VQA	50.57	50.71
LOGOS [21]	Microsoft-OCR	ST-VQA	51.53	51.08
TAP [36]	Microsoft-OCR	ST-VQA, TextCaps, OCR-CC	54.71	53.97
Ours	Microsoft-OCR	✗	53.33	52.35
Ours	Microsoft-OCR	ST-VQA	54.33	54.47
Ours†	Microsoft-OCR	ST-VQA, TextCaps, OCR-CC	**55.96**	**55.33**

comparison, we conduct experiments on different numbers of OCR tokens. As can be seen, when using only 100 OCR tokens, our method has already lifted the accuracy achieved by TAP and LaTr from 44.06% to 44.20%. When using 300 OCR tokens, our method achieves the state-of-the-art accuracy of 44.75%.

The **bottom part** displays results in the unconstrained setting and we also list the OCR system and extra data used by different methods. Following TAP [36], we use Microsoft-OCR to detect scene texts in the images and gradually expand the training data. As we can see, **(1)** when switching to the Microsoft-OCR, without any extra data our method achieves 53.33% and 52.35% on the validation and test set, improving TAP by **+3.42%** and **+2.64%** respectively in the same setting. **(2)** When adding ST-VQA as another training dataset like previous works, our method improves the test accuracy from 51.08% of LOGOS [21] to 54.47% (**+3.39%**), which has already surpassed the final result of TAP after pre-trained on larger datasets (53.97%). Besides, unlike these methods, our accuracy on the test set is better than that on the validation set (54.33%), demonstrating that our method has better generalization ability. **(3)** After introducing image caption datasets when pre-training, our method achieves a final result of 55.33%, increasing the accuracy of TAP by **+1.36%**.

Table 2. Performance comparison on the ST-VQA dataset.

Method	Val acc.	Val ANLS	Test ANLS
M4C [12]	38.05	0.472	0.462
SA-M4C [14]	42.23	0.512	0.504
SMA [9]	-	-	0.466
CRN [19]	-	-	0.483
LaAP-Net [11]	39.74	0.497	0.485
BOV [37]	40.18	0.500	0.472
LOGOS [21]	48.63	0.581	0.579
TAP [36]	50.83	0.598	0.597
Ours	50.49	0.598	0.587
Ours†	**55.51**	**0.646**	**0.634**

Table 3. Ablation study results on the TextVQA dataset. "TMFM" and "DM" refer to the two-stage multi-modality fusion module and denoising module.

Configuration	TMFM	DM	Pre-train	Val acc.
(1) TextVQA	✓	✗	✗	50.71
(2) TextVQA	✓	✓	✗	51.32
(3) w/ ST-VQA	✓	✗	✗	51.49
(4) w/ ST-VQA	✓	✓	✗	52.56
(5) w/ MLM	✓	✗	✓	52.82
(6) w/ MLM, ITM	✓	✗	✓	52.47
(7) w/ MLM	✓	✓	✓	53.33

Results on ST-VQA. Tab. 2 presents the results on the ST-VQA dataset in the unconstrained setting. The base model is pre-trained and fine-tuned on the training set of ST-VQA and † refers to the large model that uses TextVQA, ST-VQA, TextCaps and OCR-CC in pre-training. Unlike the TextVQA dataset, in ST-VQA all the answers are texts in the images. Our method works better on ST-VQA as we focus more on scene texts. As can be seen, without any extra data, our method achieves the accuracy of 50.49% and the ANLS score of 0.598, which is nearly at the same level with the large-scale pre-trained TAP. When pre-trained on larger datasets, the final accuracy and ANLS score of our method reach 55.51% and 0.646, i.e., **+4.68%** and **0.048** higher than that of TAP.

4.4 Ablation Study

Here, we conduct ablation study to demonstrate the effectiveness of each component in our method. All the experiments are done on the TextVQA dataset.

Overall results. As shown in Tab. 3, we first report the overall ablation results of different components in our method. Only by our two-stage multi-modality fusion module but **without any pre-training**, the accuracy reaches 50.71%, which outperforms TAP after being pre-trained (49.91%), while our denoising module increases the performance to 51.32% (+0.61%). Especially, when introducing ST-VQA as an additional training dataset, the improvement gained by our denoising module increases to +1.07% (Row 3 and Row 4), showing that our denoising module works better with more data. One possible reason is that all the answers of the ST-VQA dataset are texts in the images without any external vocabulary, which is beneficial to the denoising module. Finally, our method achieves a competitive result 52.56% without pre-training. On the other hand, Row 5 and Row 6 show the results of our model with pre-training on TextVQA dataset. As can be seen, the MLM task makes a positive impact while introducing the ITM task causes a decrease from 52.82% to 52.47% (-0.35%). This is possibly because that the semantic relationship between scene texts and the image is not so strong (different from the fact that image caption is usually closely related to the image). At last, Row 7 shows the final result of our model

Table 4. Ablation study on the two-stage multimodality fusion module.

Fusion module	$X^{ocr,l}$	$X^{ocr,v}$	X^{obj}	Val acc.
One-stage fusion	✓	✓	✓	49.41
Two-stage fusion	✗	✗	✗	45.12
	✓	✗	✗	49.24
	✓	✓	✗	50.50
	✓	✓	✓	50.71

Table 5. Ablation study on augmentation times t and masking probability p of the denoising module.

	p=0.15	p=0.3	p=0.5
t=1	50.53	51.04	50.41
t=3	51.32	51.16	50.36

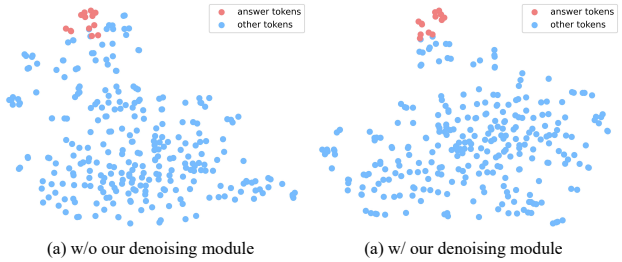
trained on the TextVQA dataset. Our denoising module still works well after the model being pre-trained, increasing the accuracy from 52.82% to 53.33%.

Effect of two-stage multimodality fusion module. We conduct further ablation study on our two-stage multimodality fusion module to verify its effectiveness and show the influence of each feature. First, we test the one-stage fusion paradigm. In this setting, features of different modalities are extracted respectively and unimodal features are added together before fusion. Here, we use BERT [8] to extract the feature of the question following previous works while keeping LayoutLM to extract the semantic features of OCR tokens. The semantic and visual features of OCR tokens are then added together before interacting with the other modalities. As shown in Tab. 4, the final result is 49.41% while our two-stage multimodality fusion module improves the accuracy to 50.71 (+1.30%), which verifies our expectation that jointly considering the question and scene texts are beneficial to the understanding of both modalities.

Then, we display the importance of various features. As we have mentioned in the beginning that texts are the basis, by removing the object features and layout/visual appearance features of scene texts in our two-stage fusion module, the model can still obtain an accuracy of 45.12%. Adding layout features in the first stage to help text understanding brings an increase of +4.12%, and introducing visual appearance features of scene texts based on the former setting achieves additional improvement of +1.26%, which shows the importance of both layout and visual appearance of scene texts. Finally, we add object features as auxiliary information. The gap of the accuracy between our model with or without object features is only about 0.2%, which is consistent with our claim that global visual objects are not the first important role in the TextVQA task.

Effect of denoising module. As shown in Tab. 5, we conduct ablation study on the hyper-parameters in our denoising module by setting different values for augmentation times and masking probability. Augmenting only one time makes no obvious improvement, which may because of the strong randomness. Increasing augmentation times is helpful while increasing masking probability leads to accuracy decrease. At last, we choose to augment each sample 3 times with the masking probability being 15%.

In order to show the superiority of our masking strategy, we further test some other methods. Masking without excluding the correct answers obtains an accuracy of 50.74% (-0.58%) while deleting the whole features achieves 50.13% (-0.19%), which shows that our masking strategy is the best choice.



(a) w/o our denoising module (a) w/ our denoising module

Fig. 4. tSNE results of the output features of OCR tokens in a text-rich sample. (a) Without our denoising module, there are more noise points around the answer tokens. (b) With our denoising module, fewer unrelated tokens are around the correct answers.

As using only data augmentation can also boost performance, we conduct experiment by removing our contrastive loss, and the accuracy decreases to 51.08%, which shows the effectiveness of contrastive learning. Besides, we compare the visualization results of models trained with and without our denoising module. Fig. 4 shows the visualization results of the output features of OCR tokens by tSNE [22]. As seen, when trained without our denoising module, there are more noise points around the correct answer. While with our denoising module, as tokens in the correct answer are not masked and fewer tokens are considered each time, there are fewer noise points around the answer tokens.

Moreover, our proposed denoising module can also work well with previous works. Here we conduct experiments with the common baseline model M4C [12] on the TextVQA dataset. As the input features of M4C are different from ours, here we randomly mask the semantic features of FastText [6] and PHOC [1]. Our experimental results show that the denoising module lifts the accuracy of M4C from 45.55% to 46.24%, which verifies the effectiveness and flexibility of it.

5 Conclusion

In this paper, we observe that compared to visual objects, question and scene text modalities are more important in the TextVQA task while both layout and visual appearance are useful. Based on this observation, we propose a two-stage multimodality fusion based method to boost TextVQA. Besides, in order to alleviate the redundancy and noise of recognized scene texts, we develop a denoising module that utilize contrastive loss to make the model focus on the relevant texts. Extensive experiments on two benchmarks are conducted, which verify the effectiveness and superiority of the proposed method.

Acknowledgments This work was supported in part by a ByteDance Research Collaboration Project.

References

1. Almazán, J., Gordo, A., Fornés, A., Valveny, E.: Word spotting and recognition with embedded attributes. IEEE transactions on pattern analysis and machine intelligence **36**(12), 2552–2566 (2014)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: 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 (2018)
3. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: Vqa: Visual question answering. In: Proceedings of the IEEE international conference on computer vision. pp. 2425–2433 (2015)
4. Biten, A.F., Litman, R., Xie, Y., Appalaraju, S., Manmatha, R.: Latr: Layout-aware transformer for scene-text vqa. arXiv preprint arXiv:2112.12494 (2021)
5. Biten, A.F., Tito, R., Mafla, A., Gomez, L., Rusinol, M., Valveny, E., Jawahar, C., Karatzas, D.: Scene text visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4291–4301 (2019)
6. Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)
7. Borisyuk, F., Gordo, A., Sivakumar, V.: 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 (2018)
8. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
9. Gao, C., Zhu, Q., Wang, P., Li, H., Liu, Y., Van den Hengel, A., Wu, Q.: Structured multimodal attentions for textvqa. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
10. Gao, D., Li, K., Wang, R., Shan, S., Chen, X.: 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. pp. 12746–12756 (2020)
11. Han, W., Huang, H., Han, T.: Finding the evidence: Localization-aware answer prediction for text visual question answering. arXiv preprint arXiv:2010.02582 (2020)
12. Hu, R., Singh, A., Darrell, T., Rohrbach, M.: Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9992–10002 (2020)
13. Jiang, Y., Natarajan, V., Chen, X., Rohrbach, M., Batra, D., Parikh, D.: Pythia v0. 1: the winning entry to the vqa challenge 2018. arXiv preprint arXiv:1807.09956 (2018)
14. Kant, Y., Batra, D., Anderson, P., Schwing, A., Parikh, D., Lu, J., Agrawal, H.: Spatially aware multimodal transformers for textvqa. In: European Conference on Computer Vision. pp. 715–732. Springer (2020)
15. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021)
16. Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., et al.: Openimages: A public dataset for large-scale multi-label and multi-class image classification. Dataset available from https://github. com/openimages **2**(3), 18 (2017)

17. Lamb, A.M., ALIAS PARTH GOYAL, A.G., Zhang, Y., Zhang, S., Courville, A.C., Bengio, Y.: Professor forcing: A new algorithm for training recurrent networks. Advances in neural information processing systems **29** (2016)
18. Lewis, D., Agam, G., Argamon, S., Frieder, O., Grossman, D., Heard, J.: Building a test collection for complex document information processing. In: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 665–666 (2006)
19. Liu, F., Xu, G., Wu, Q., Du, Q., Jia, W., Tan, M.: Cascade reasoning network for text-based visual question answering. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 4060–4069 (2020)
20. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017)
21. Lu, X., Fan, Z., Wang, Y., Oh, J., Rosé, C.P.: Localize, group, and select: Boosting text-vqa by scene text modeling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2631–2639 (2021)
22. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
23. Mishra, A., Shekhar, S., Singh, A.K., Chakraborty, A.: Ocr-vqa: Visual question answering by reading text in images. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 947–952. IEEE (2019)
24. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X.S., Wen, J.R.: Counterfactual vqa: A cause-effect look at language bias. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12700–12710 (2021)
25. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. arXiv preprint arXiv:1910.10683 (2019)
26. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
27. Sidorov, O., Hu, R., Rohrbach, M., Singh, A.: Textcaps: a dataset for image captioning with reading comprehension. In: European Conference on Computer Vision. pp. 742–758. Springer (2020)
28. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8317–8326 (2019)
29. Singh, A., Pang, G., Toh, M., Huang, J., Galuba, W., Hassner, T.: Textocr: Towards large-scale end-to-end reasoning for arbitrary-shaped scene text. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8802–8812 (2021)
30. Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., Dai, J.: Vl-bert: Pre-training of generic visual-linguistic representations. In: International Conference on Learning Representations (2019)
31. Wang, Q., Xiao, L., Lu, Y., Jin, Y., He, H.: Towards reasoning ability in scene text visual question answering. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2281–2289 (2021)
32. Wang, X., Liu, Y., Shen, C., Ng, C.C., Luo, C., Jin, L., Chan, C.S., Hengel, A.v.d., Wang, L.: On the general value of evidence, and bilingual scene-text visual question answering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10126–10135 (2020)

33. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Huggingface's transformers: State-of-the-art natural language processing. arXiv preprint arXiv:1910.03771 (2019)
34. Xu, Y., Xu, Y., Lv, T., Cui, L., Wei, F., Wang, G., Lu, Y., Florencio, D., Zhang, C., Che, W., et al.: Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. arXiv preprint arXiv:2012.14740 (2020)
35. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., Zhou, M.: Layoutlm: Pre-training of text and layout for document image understanding. In: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp. 1192–1200 (2020)
36. Yang, Z., Lu, Y., Wang, J., Yin, X., Florencio, D., Wang, L., Zhang, C., Zhang, L., Luo, J.: Tap: Text-aware pre-training for text-vqa and text-caption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8751–8761 (2021)
37. Zeng, G., Zhang, Y., Zhou, Y., Yang, X.: Beyond ocr+ vqa: Involving ocr into the flow for robust and accurate textvqa. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 376–385 (2021)
38. Zhang, X., Yang, Q.: Position-augmented transformers with entity-aligned mesh for textvqa. In: Proceedings of the 29th ACM International Conference on Multimedia. pp. 2519–2528 (2021)
39. Zhao, M., Li, B., Wang, J., Li, W., Zhou, W., Zhang, L., Xuyang, S., Yu, Z., Yu, X., Li, G., et al.: Towards video text visual question answering: Benchmark and baseline. In: Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2022)
40. Zhu, Q., Gao, C., Wang, P., Wu, Q.: Simple is not easy: A simple strong baseline for textvqa and textcaps. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 3608–3615 (2021)