Generating Questions via Unexploited OCR Texts: Prompt-Based Data Augmentation for TextVQA

Mingjie Han[1], Ting Jin[1]* , Wancong Lin[1], Can Li[2], and Liang Qiao[3]
[1]School of Computer Science and Technology, Hainan University, HaiKou, China
[2]School of Electronic Information and Electrical Engineering, Shanghai Jiaotong University, Shanghai, China
[3]School of Computer Science and Technology, Zhejiang University, Hangzhou, China
{mingjiehan,jinting,wanconglin}@hainanu.edu.cn, volcano@alumni.sjtu.edu.cn, qiaoliang@zju.edu.cn

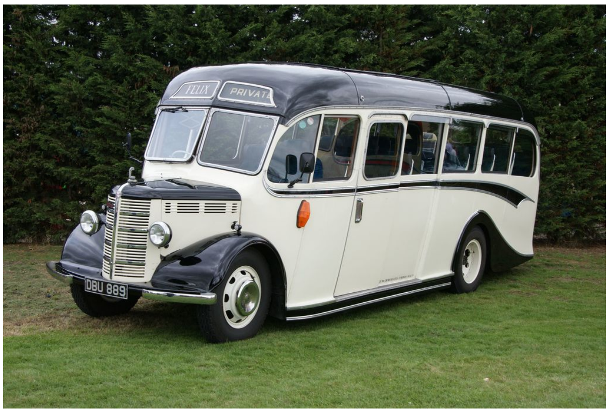
***Abstract*—Text-based Visual Question Answering (TextVQA) tasks rely on Optical Character Recognition (OCR) text to answer. There have been many models successfully exploring multi-modal features fusing and knowledge reasoning. However, current TextVQA datasets are few and the cost of using manual annotation is too high. So generating pseudo-labeled data is a better choice. In this paper, a prompt-based data augmentation method is proposed. The problems of current data augmentation are solved: 1) the distribution of the number of answer words in the pseudo-labeled data is not consistent with the real dataset. 2) the question forms in the pseudo-labeled data are not diverse. Specifically, prompt words are first matched to the constraints in the questions by finding the same words in the vocabulary. So, our generating model can generate different types of questions when the different prompt words are input. Experiments show that our method is significantly better than other state-of-the-art methods on TextVQA.**

***Index Terms*—Data Augmentation, Prompt, TextVQA, OCR texts**

I. Introduction

The Visual Question Answering (VQA) task is an image-text multi-modal question answering task proposed by Agrawal in 2015 [1]. The task requires multimodal models with the ability to read, understand, fuse, and reason. In natural scenes, textual information appears in many places, such as car licenses, store names, and clothing logos. These OCR texts are indispensable supplementary information for visual question answering. This kind of task using OCR text is called Text-based Visual Question Answering [2]. An example is shown in Fig. 1.

Nowadays, many scholars agree that OCR information is very important for the TextVQA task [3], [4]. Based on this view, there are two directions of improvement: the OCR recognition system and the OCR feature extractor. Tap [5] can significantly improve the accuracy rate from 39.40% to 44.50% after using the Microsoft OCR recognition system instead of Rosetta en OCR recognition system. M4C [6] uses different perspectives to mine OCR information, such as word vector features, character statistics features, and visual features, to enrich input features.



Question: What is the license plate number?
Answer: DUB 889

Fig. 1. An example of TextVQA dataset. ”DBU 889”, ”PRIVATE” and ”FELIX” are the OCR tokens in this image. ”DBU 889” is used in the question-answer pair. Others are not.

However, such important OCR information is underutilized in the construction of the TextVQA dataset: At most two OCR tokens are selected for questioning during manual annotation [2]. Statistically, the TextVQA dataset has an average of 12 OCR tokens per image. The number of OCR words in the real world is more. When constructing question-answer pairs, a large amount of OCR tokens that can be used is wasted because of the limitation of manual annotation. Therefore, it is very important to find out how to mine more OCR tokens with question answering through limited datasets. There are usually two approaches: 1) Constructing question-answer pairs using more manual or crowdsourcing platforms; 2) Generating pseudo-label question-answer pairs using question generating models. Compared to the latter method, the former requires more human and time investment, and cannot be applied to different domains and different datasets. The question generating approach can be automatically used in different datasets to make full use of the OCR information on images that are not utilized.

Recently, TAG [7] has proposed a question generating method for TextVQA. They proposed an inverse question answering structure that interchanges the question input and the answer input of the original question answering model. The model is trained to have the ability to map from answer

*Corresponding author

information, image information, and OCR information to question information. Then, based on the intuition that the OCR token with the largest area has the most meaningful semantic information, the largest OCR token is considered as the answer to generate pseudo-label questions. Finally, the Ground Truth label and the pseudo-label information are used together as the input to the question answering model. This inverse question-answer structure is simple and intuitive to build. However, it also has some problems: 1) The answers generated by TAG in the pseudo-label question-answer pairs have only one word, which does not match the TextVQA dataset. In the TextVQA dataset, there are on average 1.6 words per answer. A single word causes a lack of semantic information and can lead to poor quality of generated pseudo-label questions. 2) The forms of generated questions are not diverse enough. When using the strategy of generating questions by answers, the generated questions are forced to be consistent with the real questions. In fact, this forced approach is counter-intuitive, because there can be different questions that all correspond to the same answer. In this case, most pseudo-label questions are influenced by the bias of the dataset, preferring to generate one kind of questions. This bias problem is inevitable in text-based visual question answering datasets because the distribution of question content will hardly be considered when constructing a text-based visual question answering dataset, but will focus more on the distribution of answers.

To overcome these drawbacks, we propose a prompt-based method using **U**nexploited **O**CR **T**exts (OUT) to generate higher-quality pseudo-label question-answer pairs. When generating, multi-word OCR texts are more likely to be selected as pseudo-label answers. These OCR tokens are usually spatially adjacent to each other. More semantic information is preserved after the words are concatenated together into a phrase. The rich semantic information of phrases constrains the mapping range of the generating model and makes the generated questions have better quality. On the other hand, a larger number of multi-word samples can also improve the ability of the question answering when facing complex questions. In addition, different prompt words are selectively added to the selected OCR texts as input for the generating model. This approach forces the prompt words to correspond to the question type, and the generating model can obtain different pseudo-label questions when different prompt words are added. Different kinds of pseudo-label questions are more complex and can provide more useful information.

The contributions of this paper are summarized in the following points: 1) We propose a data enhancement method based on prompt words to force the mapping from prompt words to question types. 2) We introduce the importance of multi-word OCR texts, propose an OCR tokens grouping method, and select multi-word OCR texts as input to the generating model. 3) We find that the correctness of the OCR information also has a significant impact on the results when training the generating model, and that better results are obtained using the Microsoft OCR recognition system. 4) Our proposed method gets 47.94% results on TextVQA This shows the effectiveness of our method.

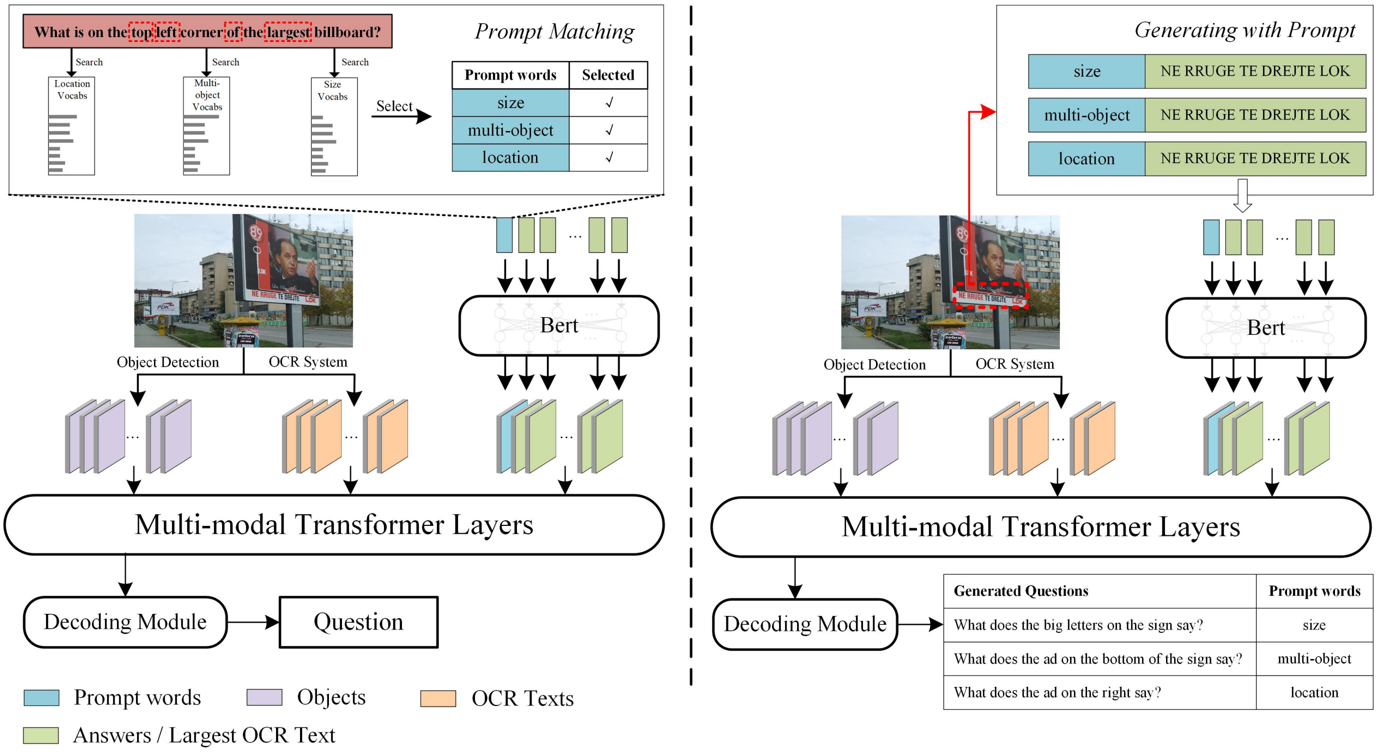
II. Related Work

A. Text-based Visual Question Answering

Singh et. al. [2] proposed a text-based visual question task and released TextVQA, the first dataset for this task, along with the model LoRRA. LoRRA is able to read and understand OCR tokens and answer accurately by the OCR tokens copy module. This is the first model to introduce OCR information into visual question answering. Then, Biten et al. [8] proposed another dataset, STVQA. this dataset is consistent with TextVQA in terms of the length of questions and answers. The difference is that STVQA proposes an evaluation metric Average Normalized Levenshtein Similarity (ANLS). This metric can mildly penalize the errors of OCR recognition. In 2020, Hu et al. [6] proposed the iterative decoding-based text-based visual question model M4C. M4C cleverly utilizes an iterative decoder and a pointer network to select the correct answer among the many OCR tokens and vocabulary words. At the same time, because the text-based visual question task is particularly dependent on OCR tokens, they propose rich OCR features to improve model comprehension. These two modules are the key to the success of M4C. Almost all subsequent models are based on M4C. Han et al. [9] proposed the LaAP-Net model, and they pointed out the problem that masking object features in the M4C model have little effect on the results. To improve the utilization of object features, they fused object features with OCR features for cross-modal attention. The fused features are directly input into Transformer [10], and the original object features are directly discarded. This setup ensures fuller interaction between objects and OCR tokens. The TAP [5] model adds mask language modeling, Relative position prediction, and image text matching pre-training tasks. The additional pre-training tasks help the model to establish the relationship between images and text more than the visual question answering tasks, without adding additional data.

B. Data Enhancement of Visual Question Answering

Kafle et al. [11] propose two methods for data augmentation in visual question answering: 1) Creating new question-answer pairs using semantic segmentation instances in the COCO [12] dataset, which can be classified into ”whether”, ”technology”, ”object recognition”, ”scene recognition”, etc. Based on the type of questions. 2) Generate new question-answer pairs using the LSTM [13] model. Retaining the 3 most common question-answer pairs can avoid generating poor-quality questions. Tang et al. [14] used IFCMS to generate visual samples. The samples were trained by adding the direction of the gradient computation on the input image to make the change in the model prediction error larger. They also used machine translation to translate the questions into other languages and back to English. Li et al. [15] proposed to enhance the visual question and answer model using visual question generating techniques, and they viewed visual question generating and visual question and answer as a dual task. The proposed model



Prompt words	Selected
size	√
multi-object	√
location	√

size	NE RRUGE TE DREJTE LOK
multi-object	NE RRUGE TE DREJTE LOK
location	NE RRUGE TE DREJTE LOK

Generated Questions	Prompt words
What does the big letters on the sign say?	size
What does the ad on the bottom of the sign say?	multi-object
What does the ad on the right say?	location

Fig. 2. The overview of our model in the training (left) and inference (right) process. Different prompt words are chosen depending on how well the key words in questions match the words in the three vocabularies before training. If all three vocabularies are matched at the same time, then the OCR words are concatenated with each of the three prompt words respectively to generate three samples for input into the model for training. By selecting different prompt words and the largest OCR text, different forms of questions can be generated at inference. The green color in the left and right sub-images indicates the answer information and the largest OCR text information, respectively.

can train the visual question generating model along with the visual question answering model. They design a symmetric embedding structure to ensure the duality of the model. Wu et al. [16] on the other hand, view image caption generating and answer prediction as dual tasks. They designed an image caption embedding module that identifies the important parts of the image. This ensures that the generated new image captions are relevant to the question answering. TAG model [7] is the first paper on data augmentation in a text-based visual question answering task. It cleverly interchanges the questions and answers in the input and generates new question-answer pairs using the answer, image, and OCR information. The newly generated question-answer pairs are generated based on the largest OCR words. This method effectively utilizes the OCR words that are not used in the dataset and improves the accuracy of the question answering.

However, the question-answer pairs generated by the above models lack diversity. Only one question-answer pair can be generated for one sample. Moreover, the new question-answer pair has only one-word answers, which does not match the distribution of the original dataset.

III. Proposed Method

In this paper, we propose a prompt-based method, as shown in Figure 2. The M4C-based question answering structure is utilized to generate questions. In the training process of the generating model, the type of each question is matched by different prompt words. The corresponding prompt words are added to the answers of matched samples. In the prediction process of the generating model, spatially related adjacent OCR words are concatenated into phrases. Then these OCR phrases are concatenated with prompt words to construct pseudo-label answers, which are used to generate pseudo-label questions. The generated pseudo-label data is fed into the question answering model along with the original dataset. The basic structure of the generating model is first described below.

A. Basic structure our model

The inputs for generating model include answer embedding, object embedding, and OCR embedding. The features between these different modalities are fused by a multimodal Transformer [10] and then passed through a pointer-net based decoder to generate pseudo-label questions.

Answer Embedding. The answer embedding module is the most difference between generating model and question answering models. In question answering models, answer information is considered a label in Teacher Forcing [17]. In the generating model, answer information is treated as an important input to generate questions. A pre-trained Bert [18] model is used to extract language features $X^{ans} = \{x_k^{bert}\}$ (where $k = 1, 2, ..., K$)

Object Embedding. Following the BUTD [19], an object detection model Faster R-CNN [20] is used to extract N object features in images offline. Each object feature consists of a one-dimensional fully connected feature considered as an appearance feature $X^{obj,frcn} = \{x_n^{obj,frcn}\}$ (where $n = 1, 2, ..., N$), and spatial information of the bounding box is considered as a spatial feature $X^{obj,bbox} = \{x^{obj,bbox}\}$. In the generating model, appearance features and spatial features are all projected into the hidden state space. Their dimensions are also aligned with the hidden state dimension of generating model. Then, object features are equal to the sum of appearance features and spatial features as

$$X^{obj} = LN(W_1 X^{obj,frcn}) + LN(W_2 X^{obj,bbox}) \quad (1)$$

where $LN(\cdot)$ is the Layer Normalization [21] function to unify the scale of features and speed up training. W_1 and W_2 are both project features.

OCR Embedding. OCR information is the most important in generating models. The richer information contained in OCR features, the better the model understands OCR words in images. A visual model is used to extract the visual features of OCR regions and a language model is used to extract the language features of OCR words. As with objects embedding, offline OCR appearance features $X^{ocr,frcn} = \{x_m^{ocr,frcn}\}$ (where $m = 1, 2, ..., M$) and spatial features $X^{ocr,bbox} = \{x_m^{ocr,bbox}\}$ are projected and fine-tuned in generating models. Fasttext [22] features $X^{ocr,fast} = \{x_m^{ocr,fast}\}$ and PHOC [23] features $X^{ocr,phoc} = x_m^{ocr,phoc}$ together form the OCR language features. Then OCR features are as

$$\begin{aligned} X^{ocr} = LN(W_3 X^{ocr,frcn} + W_4 X^{ocr,fast} + \\ W_5 X^{ocr,phoc}) + LN(W_6 X^{ocr,bbox}) \end{aligned} \quad (2)$$

where W_3, W_4, W_5 and W_6 are all project features.

Iterative Decoding. The fused features, concatenated by answer features, object features, and OCR features, is input to Transformer [10] and interacts across modalities with self-attention [24]. The output features from Transformer are input into a pointer net [25]. Then the score of each word is generated through iterative decoding. The word with the highest score will be selected for prediction.

B. Data Augmentation.

Prompt Matching. Depending on the constraints, questions are matched with the relevant prompt words. For example, the words ”top left” in the question in Figure 2 contains a spatial relationship constraint. With the constraint, our model focuses more on the bottom right corner of the image and ignores OCR tokens in other locations, such as ”PDK”, and ”zebra”.

Define a vocabulary V that contains the words related to spatial constraints. When the question token is successfully searched in vocabulary V, the prompt word “location” is selected and added to answers. The answer “89” is enhanced to “location 89”. The enhancement answer is input to the generating model for training. In this paper, the prompt word “location” is a special token defined to suggest the spatial constraints in questions.

Similarly, there are other constraints that can be mined in the training dataset, such as the relationship constraints between multiple objects, the object area size constraints, etc. Examples of constraints can be seen in Table I. Specifically, spatial relationship constraints refer to the constraints caused by the spatial words in questions occurring when models are searching for answer space. Relationship constraints between multiple objects refer to the complex constraints involving multiple objects. Object area size constraints refer to the constraints caused by the words about the size. All the above constraints are generalized to any sample and independent of the content information in images. Therefore, when choosing constraints, the word that depends on a specific sample is not considered, such as title, website, and others.

TABLE I
DIFFERENT CONSTRAINTS AND CORRESPONDING PROMPT WORDS WITH EXAMPLE QUESTIONS. WORDS IN RED INDICATE THAT THEY ARE ASSOCIATED WITH CONSTRAINTS.

Constraints	Prompt words	Example questions
Spatial relationship constraints	location	What is the website link found in the bottom right corner?
Relationship constraints between multiple objects	multi-object	What is the name of the company near the right shoulder of this player?
Object area size constraints	size	What is the issue of the magazine after the small gap?

Generating with Prompt. After the generating model is trained on the enhancement answer information, it has learned how to generate different questions with different prompt words when predicting on the same image. In Figure 2, for example, the largest text in the image, ”NE RRUGE TE DREJTE LOK”, is concatenated with three prompt words in sequence to form ”location NE RRUGE TE DREJTE LOK”, ”multi-object NE RRUGE TE DREJTE LOK” and ”size NE RRUGE TE DREJTE LOK”. The prompt words and the OCR words are used together as input to predict the problem under different constraints. As can be seen from the example, after adding the prompt words, the generating model not only keeps its original ability to generate questions, but also learns the relevant constraints, which can generate a more diverse and complex question. These complex questions are very helpful for the question answering model.

C. OCR Tokens Grouping.

In the TextVQA dataset, the number of answers with multiple words is 29.23% of the total. Intuitively, the enhanced data should contain both one-word answers and multi-word answers. So all OCR tokens on the original dataset should be grouped in order to select multi-word OCR texts when generating with prompt words. Also, to improve the accuracy of grouping, only OCR tokens that are in the same line are considered. OCR tokens that are not at the same line will be not grouped into the same group. The grouping scheme is presented in Algorithm 1.

Algorithm 1: Pseudo code of our OCR tokens Grouping

```
Data: A set of bounding boxes $b$ with OCR Tokens $t$.
foreach OCR bounding boxes $b_i$ in $b$ do
    Initialize grouping information $g_i = -1$;
    foreach OCR bounding boxes $b_j$ in $b$ do
        if $b_i$ is close to $b_j$ and $b_i$ has similar height with $b_j$ then
            if $t_j$ includes $t_i$ then
                Set grouping information $g_i = 0$
            else
                Set grouping information $g_i = g_j$
            end
        else
            Set grouping information $g_i = max\{g_1, g_2, ..., g_j\} + 1$
        end
    end
end
Result: Grouping information $g$
```

After grouping, if OCR tokens are in the same line, their group information is the same. On the contrary, if they are not in the same line, the group information will be different. If the grouping information of an OCR token is 0, the OCR token might are recognized as duplicate or incorrect. So it will not be considered when selecting the maximum OCR text.

IV. Experiments

A. Datasets and Metrics

TextVQA [2] is the first dataset for text-based visual question answering. All images in the dataset are collected from Open Image v3 dataset. It contains 28408 images, including 21953 images in the training dataset, 3166 images in the validation dataset, and 3289 images in the test dataset. On average, 1-2 questions per image were collected by the annotators. The dataset contains 45336 questions, of which 34602 are in the training dataset, 5000 in the validation dataset, and 5734 in the test dataset.

STVQA [8] is another popular dataset for text-based visual question answering. The images in the dataset are from six different datasets: ICDAR2013 [26], ICDAR2015 [27], ImageNet [28], Vizwiz [29], IIIT Scence Text Retrieval [30], Visual Genome [31], and COCO-Text [32]. 20,238 images and 31,791 questions are included in this dataset, with 19,027 images and 26308 questions in the training dataset, and 2993 images and 4163 questions in the test dataset. We follow previous works and use 17028 images as the training dataset and 1893 images as the validation dataset.

A soft evaluation metric [1] based on human voting is used to calculate the accuracy of a text-based visual question answering task. The rationale behind this metric is that the more occurrences the more correct the answer is.

B. Experiment Settings and Training Details

Our OUT network is followed by M4C [6]. A pretrained model BERT-BASE-UNCASE [18] is used in the answer embedding module. The Faster R-CNN [20] model is pretrained on the Visual Genome dataset same as BUTD [19] model. The backbone of Faster R-CNN is ResNeXT [33]. Microsoft OCR system [5] is used to get the OCR tokens and their bounding boxes. In multi-modal Transformer layers, a triangular matrix is used to ensure the quality of generated questions. The iterative decoding step is 30 to generate longer questions. The numbers K, N, and M are set to 20, 100, and 100.

The dimension of the fused features in multi-modal Transformer layers [10] is 768. The role of the projection matrix W1, W2, W3, W4, W5, and W6 is to have the features of different modes projected in the same feature space. The dimensions of X^{ans}, $X^{obj,frcn}$, $X^{obj,bbox}$, $X^{ocr,frcn}$, $X^{ocr,bbox}$, $X^{ocr,fast}$ and $X^{ocr,phoc}$ are 768, 2048, 4, 2048, 4, 300, 604.

We train our OUT network on two NVIDIA 3090 GPUs. The training iteration is set to 24000 and the number of training samples is 34602. The learning rate is set at 0.001 at the beginning and shrinks by a factor of 10 when the number of iteration steps is 14000 and 19000 respectively. Optimizer is Adam and the batch size is 128.

C. Comparison with State-of-the-art

The results of the comparison of our model and other Text-based VQA models are shown in Table II. Our data augmentation model OUT reach 47.38% accuracy on the test dataset without extra data (line 8). And it reach 47.94% accuracy on the test dataset with STVQA dataset (line 11). M4C [6] (line 1) is the baseline of the Text-based VQA task. M4C† [5] use Microsoft OCR system. The comparison with M4C and M4C†show the importance of OCR recognition accuracy. The OCR recognition accuracy of the Microsoft OCR system is the best now. TAG is the SOTA data augmentation method on Text-based VQA. It boosts M4C by about 1.21% with extra data (line 7). Our method exceeds TAG [7] by about 1.42% without extra data (line 7 and line 8). Moreover, Our method without extra data suppresses TAG with STVQA 1.00% (line 8 and line 10).

D. Ablation Study

Ablation experiments are based same M4C [6] model. The M4C uses the Microsoft OCR system [5] without extra data here. The ablation results are shown in Table III. Our method is worse than TAG [7] in the same configuration (line 1 and line 2), but is 1.5% higher than TAG after using the three strategies (line 1 and line 10). For prompt strategy, we have tried different prompt words and their combination (line 3-6), and It is found that ”multi-object” works best. The combination of prompt words (line 6) refers to selecting one word of them when generating with prompt words. Although ”location” and

TABLE II
RESULTS OF DIFFERENT STATE-FF-THE-ART MODELS ON TEXTVQA DATASET. THE COLUMN "OCR SYSTEM" REFERS TO THE OCR RECOGNITION SYSTEM USED IN THE QUESTION ANSWERING MODEL.

Model	OCR system	Extra data	Val Acc.	Test Acc.
LoRRA [2]	Rosetta-en	✗	26.56%	27.63%
M4C [6]	Rosetta-en	✗	39.40%	39.01%
LaAP-Net [9]	Rosetta-en	✗	40.68%	40.54%
CRN [34]	Rosetta-en	✗	40.39%	40.96%
SMA [3]	SBD-trans	✗	43.74%	44.29%
SSBaseline [35]	SBD-trans	✗	43.95%	44.72%
M4C† [5]	Microsoft	✗	44.50%	44.75%
M4C†+TAG [7]	Microsoft	✗	45.68%	45.96%
M4C†+OUT	Microsoft	✗	**47.35%**	**47.38%**
M4C† [5]	Microsoft	STVQA	45.22%	-
M4C†+TAG [7]	Microsoft	STVQA	46.63%	46.38%
M4C†+OUT	Microsoft	STVQA	**48.36%**	**47.94%**

”size” have a little effect, the effect is not as good as ”multi-object”. So the performance of their combination is lowered by ”location” and ”size”. This is why we chose only ”multi-object” as a prompt word in the overall strategy (line 10). Another OCR tokens grouping strategy proposed in this paper is proven to be effective (line 2 and line 8). In addition, the Microsoft OCR recognition system has proven to be better than the Rosetta-en system in previous work, but TAG still uses the latter. In this paper, the OCR system of the question generating model was changed to the Microsoft OCR system. It is found that the accuracy rate increased by 1.53% with the Microsoft OCR system (line 7). This is the same conclusion as in the text-based visual question answering model.

TABLE III
ABLATION RESULTS OF OUR METHOD. THE COLUMN "OCR SYSTEM" REFERS TO THE OCR RECOGNITION SYSTEM USED IN THE QUESTION GENERATING MODEL.

Method	Prompt words	Grouping	OCR system	Val Acc.
TAG [7]	✗	✗	Rosetta-en	45.68%
OUT	✗	✗	Rosetta-en	45.05%
OUT	location	✗	Rosetta-en	45.26%
OUT	multi-object	✗	Rosetta-en	46.35%
OUT	size	✗	Rosetta-en	45.36%
OUT	location+multi-object+size	✗	Rosetta-en	46.14%
OUT	✗	✗	Microsoft	46.58%
OUT	✗	✓	Rosetta-en	46.17%
OUT	✗	✓	Microsoft	46.62%
OUT	multi-object	✓	Microsoft	**47.35%**

E. Visualization Analysis

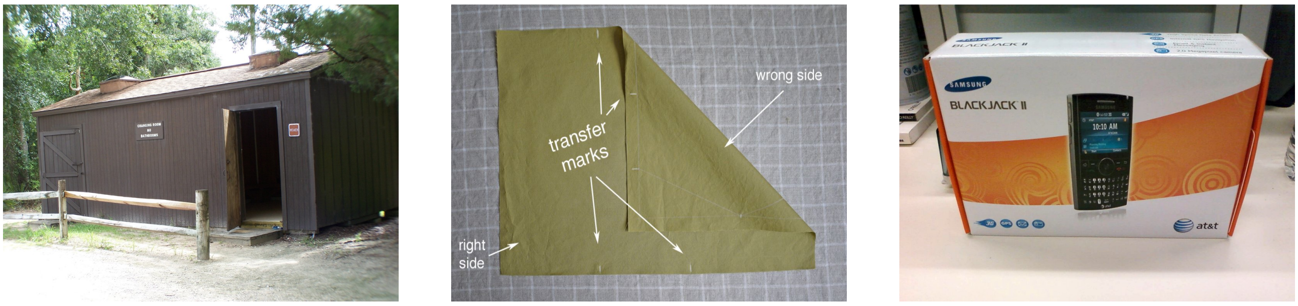
Results of Text-based question answering. Figure 3 shows the answers predicted by the TextVQA model with the OUT model. We use M4C [6] for the TextVQA model and the Microsoft OCR system for the OCR recognition system, with no extra data. For Figure 3 (a) and Figure 3 (b), the prediction is correct. In Figure 3 (a), the model selects ”changing room” among two similar words, which indicates that the model can understand complex linguistic logic after using OUT data augment. Figure 3 (b) shows that the model with OUT understands spatial relations and can answer multiple words. For Figure 3 (c), the prediction is wrong. As can be seen in Figure 3 (c), it is difficult to determine which brand is ”samsung” or ”at&t” just based on location and font, because it requires additional knowledge to answer.

Results of OCR tokens grouping. Figure 4 shows the grouping of OCR tokens. The red rectangle in the image is the largest outer box of the same group of OCR tokens. The token in the rectangle boxes are selected together when generating new questions, such as ”YOU'VE MADE IT” in Figure 4(a). Figure 4(a) shows the grouping on the printed words. All tokens on the same horizontal line in the image are correctly grouped in the same group, which shows that our grouping algorithm is effective. Figure 4(b) shows a picture of a road sign. From the figure, we can see that most of the OCR tokens are grouped correctly. However, there are some tokens that are too far apart or too close, leading to grouping incorrectly. The curved token in Figure 4(c) poses the greatest challenge to OCR grouping. It is impossible to determine the grouping of curved or skewed token by relying on space or size information alone. This requires adding visual semantic information for grouping.

Results of prediction with prompt. The results of the different question answer pairs generated by TAG and OUT are shown in Table IV. Figure 5 shows the images used in Table IV. Line 1 and line 2 are the original question-answer pairs in the TextVQA dataset. Line 3 is the result of TAG, which selects ”MACAU” as the pseudo-tag answer and generates a relatively simple question. Line 4 to line 6 are the results of OUT generation under different prompt conditions. This shows that even with the same answer input, different types of questions can be generated with different prompt words. And these questions are more complex and informative than the ones generated by TAG.

V. CONCLUSIONS

In this paper, we propose a prompt-based data augmentation method for Text-based VQA. New question-answer pairs are constructed by reusing unexploited OCR texts. Before the answer is entered into the question generating model for training, prompt matching can mine the question for constraints and find prompt words. These prompt words are stitched with the largest OCR text at the time of generating new questions. Using this approach, different prompt words can generate different types of questions, and the quality of the generated questions is also better. In order to approximate the distribution

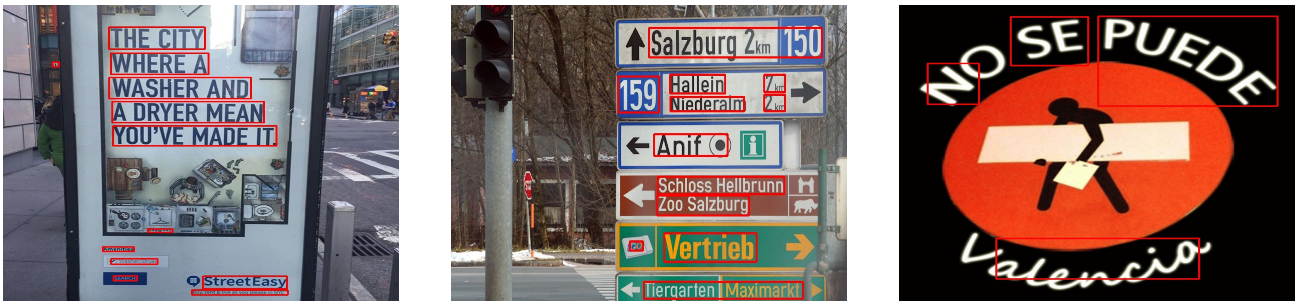


Question: What is this building designed for?
GT Answer: changing room
Prediction: changing room
(a)

Question: What words are written at the bottom left?
GT Answer: right side
Prediction: right side
(b)

Question: What company makes the product?
GT Answer: samsung
Prediction: at&t
(c)

Fig. 3. The visualization results of text-based visual question answering model with our data augmentation method.



(a)

(b)

(c)

Fig. 4. The visualization results of OCR tokens grouping. Tokens in same a red box have same group.

TABLE IV
THE QUESTION-ANSWER PAIRS FROM THE GROUND TRUTH, TAG AND OUR METHOD.

Source	OCR Word	Generated Qestion	Prompt
GT	macau beer	What is the stand for?	-
GT	macau beer	Whats the beer being advertised?	-
TAG [7]	macau beer	What is the name?	-
OUT	macau beer	What is the long's name?	size
OUT	macau beer	what is written on the yellow sign?	location
OUT	macau beer	what is the name of the beer on the left?	multi-object

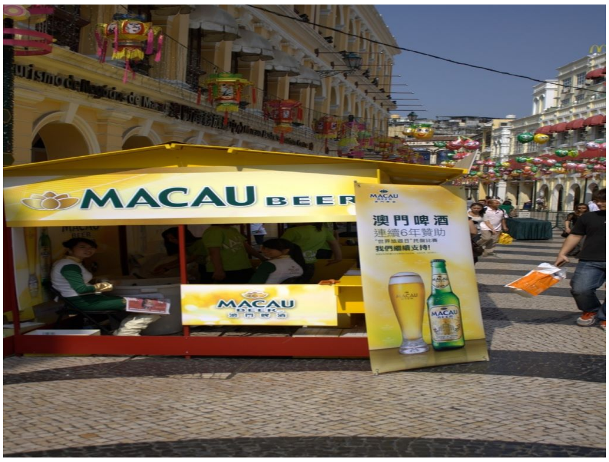


Fig. 5. The image of the sample used in Table IV.

of the number of words in the dataset, we proposed the OCR tokens grouping method. OCR tokens of the same size in the same row are grouped the same. This method also detects the problem of duplicate OCR text recognition. In addition, this paper validates that it is more efficient to use Microsoft OCR data in the generating model. Many experiments have shown that our proposed method is effective and it surpasses the current state-of-the-art. In future work, we will perform data augmentation on other datasets and design an OCR tokens grouping algorithm based on semantic information.

ACKNOWLEDGMENT

This work was supported by National Science Foundation of China (No. 61862021) and Hainan Provincial Natural Science Foundation of China (No. 620RC565).

References

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

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

[3] C. Gao, Q. Zhu, P. Wang, H. Li, Y. Liu, A. Van den Hengel, and Q. Wu, "Structured multimodal attentions for textvqa," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 9603–9614, 2021.

[4] G. Zeng, Y. Zhang, Y. Zhou, and X. Yang, "Beyond ocr+ vqa: involving ocr into the flow for robust and accurate textvqa," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 376–385.

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

[6] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for textvqa," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9992–10 002.

[7] J. Wang, M. Gao, Y. Hu, R. R. Selvaraju, C. Ramaiah, R. Xu, J. F. JaJa, and L. S. Davis, "Tag: Boosting text-vqa via text-aware visual question-answer generation," *arXiv preprint arXiv:2208.01813*, 2022.

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

[9] W. Han, H. Huang, and T. Han, "Finding the evidence: Localization-aware answer prediction for text visual question answering," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 3118–3131.

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

[11] K. Kafle, M. Yousefhussien, and C. Kanan, "Data augmentation for visual question answering," in *Proceedings of the 10th International Conference on Natural Language Generation*, 2017, pp. 198–202.

[12] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 2014, pp. 740–755.

[13] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "Lstm: A search space odyssey," *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[14] R. Tang, C. Ma, W. E. Zhang, Q. Wu, and X. Yang, "Semantic equivalent adversarial data augmentation for visual question answering," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIX 16*. Springer, 2020, pp. 437–453.

[15] Y. Li, N. Duan, B. Zhou, X. Chu, W. Ouyang, X. Wang, and M. Zhou, "Visual question generation as dual task of visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6116–6124.

[16] J. Wu, Z. Hu, and R. J. Mooney, "Generating question relevant captions to aid visual question answering," *arXiv preprint arXiv:1906.00513*, 2019.

[17] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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)*, 2019, pp. 4171–4186.

[19] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "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*, 2018, pp. 6077–6086.

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

[21] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv preprint arXiv:1607.06450*, 2016.

[22] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Transactions of the association for computational linguistics*, vol. 5, pp. 135–146, 2017.

[23] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 12, pp. 2552–2566, 2014.

[24] P. Shaw, J. Uszkoreit, and A. Vaswani, "Self-attention with relative position representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, 2018, pp. 464–468.

[25] O. Vinyals, M. Fortunato, and N. Jaitly, "Pointer networks," *Advances in neural information processing systems*, vol. 28, 2015.

[26] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L. G. i Bigorda, S. R. Mestre, J. Mas, D. F. Mota, J. A. Almazan, and L. P. De Las Heras, "Icdar 2013 robust reading competition," in *2013 12th international conference on document analysis and recognition*. IEEE, 2013, pp. 1484–1493.

[27] D. Karatzas, L. Gomez-Bigorda, A. Nicolaou, S. Ghosh, A. Bagdanov, M. Iwamura, J. Matas, L. Neumann, V. R. Chandrasekhar, S. Lu *et al.*, "Icdar 2015 competition on robust reading," in *2015 13th international conference on document analysis and recognition (ICDAR)*. IEEE, 2015, pp. 1156–1160.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[29] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "Vizwiz grand challenge: Answering visual questions from blind people," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3608–3617.

[30] A. Mishra, K. Alahari, and C. Jawahar, "Image retrieval using textual cues," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 3040–3047.

[31] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma *et al.*, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, pp. 32–73, 2017.

[32] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "Coco-text: Dataset and benchmark for text detection and recognition in natural images," *arXiv preprint arXiv:1601.07140*, 2016.

[33] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

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

[35] Q. Zhu, C. Gao, P. Wang, and Q. Wu, "Simple is not easy: A simple strong baseline for textvqa and textcaps," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3608–3615.