MULTI-MODAL LEARNING WITH TEXT MERGING FOR TEXTVQA

Changsheng Xu, Zhenlong Xu, Yifan He, Shuigeng Zhou[*]

School of Computer Science
Fudan University
Shanghai 200438, China

Jihong Guan

Dept. of Computer Sci. & Techl.
Tongji University
Shanghai 201804, China

ABSTRACT

Text visual question answer (TextVQA) is an important task of visual text understanding, which requires to understand the text generated by text recognition module and provide correct answers to specific questions. Recent works of TextVQA have tried to combine text recognition and multi-modal learning. However, due to the lack of effective preprocessing of text recognition output, existing approaches suffer from serious contextual information missing, which leads to unsatisfactory performance. In this work, we propose a **M**ulti-**M**odal **L**earning framework with **T**ext **M**erging (MML&TM in short) for TextVQA, where we develop a text merging (TM) algorithm, which can effectively merge the word-level text obtained from the text recognition module to construct line-level and paragraph-level texts for enhancing semantic context, which is crucial to visual text understanding. The TM module can be easily incorporated into the multi-modal learning framework to generate more comprehensive answers for TextVQA. We evaluate our method on a public dataset ST-VQA. Experimental results show that our TM algorithm can obtain complete semantic information, which subsequently helps MML&TM generate better answers for TextVQA.

Index Terms— Visual text understanding; Text visual question answer; Multi-modal learning; Text merging.

1. INTRODUCTION

Text is an essential means of human communication, and the content of text conveys semantic information, while image is a basic medium for human understanding the world and conveys visual information. Image text embeds text in image, which combines both textual and visual information together and can convey richer message. In many intelligent human-computer interaction scenarios, image (or visual) text understanding is challenging but also particularly important.

Image text understanding plays an indispensable role in various computer vision (CV) applications such as image search [1], instant translation [2], autonomous driving [3], and visual question answering (VQA) [4]. Though all process text, image text understanding is quite different from text understanding in natural language processing (NLP), in at least three aspects: 1) Different representation. Image text is represented as a sparse pixel matrix, while text in NLP is simply a character string. 2) Different semantic organization. The words in image text may be arranged in any form and direction, whereas text in NLP is usually arranged in a left-to-right or right-to-left order. 3) Different complexity of semantic context. The understanding of image text may be greatly affected by image quality, as the factors such as blurring, distortion, and occlusion will result in textual information loss, while text in NLP is relatively regular and easy to handle. Therefore, NLP text understanding techniques cannot be directly applied to image text understanding. Image text understanding usually consists of two stages: *text recognition* and *text understanding*. The former recognizes the textual information from the image through text recognition techniques, and the latter generates decision for a specific task by analyzing the previously acquired textual information through text understanding techniques.

Visual question answering (VQA) is increasingly studied as a visual inference problem with different datasets and methods [5, 6, 7, 8, 9, 10]. However, these datasets and methods mainly focus on the visual information of the scene while ignoring the text in the image that can provide important information for scene understanding and inference. Hence, the image text understanding task is proposed known as *text visual question answering* (TextVQA) [11, 12]. TextVQA requires the model to first understand the input of three types: text description of question, the visual content of image and the text of image, and then choose the correct answer from the text of the image for the question.

Several approaches based on text recognition have been proposed to solve the TextVQA problem [13, 14]. For example, LoRRA [4] extends the VQA model by introducing an attention branch for text recognition and adding text recognition result as a dynamic word list to the answer classifier. The answer classifier randomly selects a word from the dynamic word list as the answer. OCR-VQA [12] chunks the text recognition result based on named entity recognition (NER)

[*] Correspondence author.

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

for TextVQA. However, limited by the NER mechanism, its text chunking result only covers simple information such as name, year, and month. Although the above methods can acquire text from image, contextual information loss is severe. Considering that the text recognition result is word-level, while the ground truth answers are paragraph-level text with contextual semantic information, thus existing methods usually generate incomplete answers. What is worse, they do not adopt multi-modal fusion mechanism, limiting the information interaction and mutual compensation between text and image, which also makes the answer less accurate.

To solve these issues above, in this paper we propose a multi-modal learning framework with text merging (MML&TM in short) for TextVQA. In order to better utilize the contextual information of text in image, we consider the spatial distribution of image text, and develop a text merging algorithm to merge the word-level text obtained from text recognition to construct line-level or paragraph-level text. Furthermore, inspired by M4C [14], we combine our text merging algorithm with multi-modal learning.

Our major contributions are as follows: (1) We propose a multi-modal learning framework with text merging for TextVQA. It extracts and fuses features from text and image, and obtains the answer by encoding and decoding. (2) We design a text merging algorithm to address the contextual information loss issue. The algorithm uses the location information to reconstruct the text recognition result and convert the word-level text into paragraph-level text. The paragraph-level text contains richer contextual information and is more comprehensive as the answer of TextVQA. (3) We conduct extensive experiments on the widely used TextVQA dataset ST-VQA. Experimental results show that the proposed text merging algorithm can obtain adequate contextual information, and MML&TM can achieve outstanding performance.

2. RELATED WORK

Here, we briefly survey related work of VQA and TextVQA.

VQA has received much attention in the field of computer vision since it was proposed. To facilitate the research of VQA, researchers constructed several datasets by automatic generation or manual annotation. For example, Ren et al. [15] built a dataset by automatic generation based on COCO Caption. To solve the VQA problem, Malinowski et al. [16] combined semantic analysis and image segmentation approaches to predict answers. Gao et al. [17] used an autoencoder framework for VQA. It first encodes the images and questions by an LSTM network and then decodes them to generate answers by another LSTM. Ren et al. [15] proposed several neural network-based models, including models based on vanilla LSTM and bi-directional LSTM. Yang et al. [18] applied an attention mechanism over image features to judging their relevance to the question and then figured out the probability distribution of the answer.

Despite its popularity, VQA does not take text in the image into account. Hence, some new datasets were proposed, which require answering the question by using the text in image. This problem is called TextVQA. Researchers have proposed a number of approaches based on VQA models for TextVQA. LoRRA [4] combines the VQA method Pythia with a text recognition method and constructs a text list via an attention mechanism, the answer of TextVQA is selected from the text list. Mishra et al. [12] proposed an approach similar to LoRRA, which pre-chunks text recognition results based on named entity recognition and adds the chunked result to the candidate answers of the VQA model. There are also some approaches that directly introduce text recognition results into existing advanced VQA models [11, 13].

3. METHOD

Here, we present our method MML&TM for TextVQA in detail. We first present the framework, and then introduce the text merging algorithm and our multi-modal learning scheme.

3.1. Framework

Fig. 1 shows the framework of our method. Given a pair of image and question, we first extract the feature of different objects (e.g. vehicles, roads and people) in the input image via the object detection module, combine the OCR system with text merging algorithm to generate feature of the image text, and obtain the feature of question by word segmentation. Then, the features of the three modalities are plugged into the corresponding embedding modules. Each embedding module maps the features of different dimensions to the same dimension, and the embedded features of different modalities are concatenated into a list as subsequent input. The multi-modal information of the above list is processed by a multilayer Transformer so that the information of the same modality and different modalities can all interact and learn from each other to enrich the fusion feature representation. Finally, the fusion representation and the Transformer output of image text are transferred to a bilinear interaction module. The bilinear interaction module calculates the probability of each text in the text list as an answer, then the text with the highest probability is taken as the output answer. In Fig. 1, the answer generated by our method to the question "What is the number on the white truck?" is "020 8887 0101". Our text merging algorithm and multi-modal learning scheme are described in detail in the following subsections.

3.2. Text Merging

We utilize text detection method CRFAT and text recognition method CRNN to construct the OCR system. It can detect and recognize the text of the image and output the text list with corresponding location information. In English text,

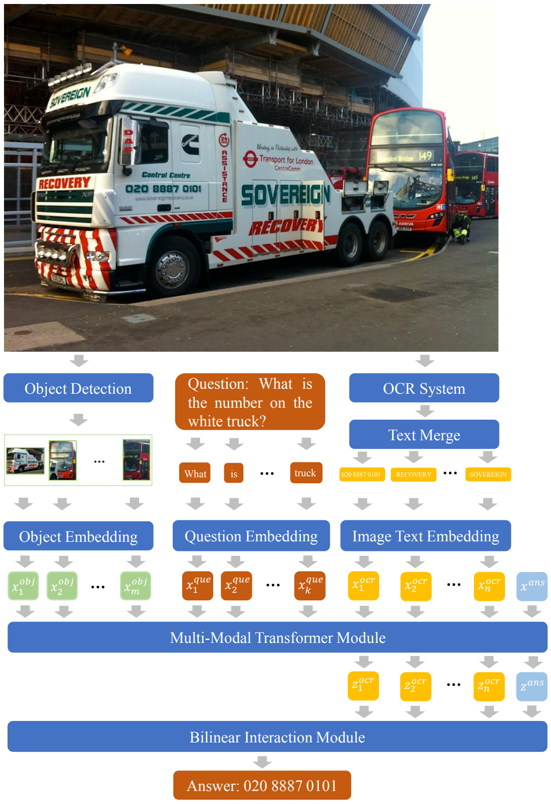

Fig. 1. The framework of MML&TM.

words are usually separated by space, which causes the text list output by OCR system is word-level. The output of OCR system is generally consistent with the order of the bounding boxes generated by the text detection module, which sorts the bounding boxes according to the center or upper-left coordinates of the boxes. However, the bounding boxes belonging to the same paragraph may be different in size and shape, and the coordinates highly correlated with the shape and size are unreliable as the basis for sorting, which will lead to contextual information confusion.

We propose a text merging algorithm to solve this problem, which can process the word-level text list generated by the OCR system to obtain the paragraph-level text list with complete contextual semantic information. The output of the OCR system can be viewed as an array $\mathbf{R}$ of OCR objects consisting of texts and their locations as follows:

$$\mathbf{R} \overset{def}{=} \{(text_i, box_i)\} \tag{1}$$

where $text_i$ represents the i-th text string of $\mathbf{R}$ and box_i is the location of the corresponding bounding box in the image for the text string. The bounding box contains four points: upper-left, upper-right, lower-right, and lower-left, formulated as

$$box_i \overset{def}{=} \{(x_{lu}, y_{lu}), (x_{rd}, y_{rd}), (x_{ru}, y_{ru}), (x_{ld}, y_{ld})\} \tag{2}$$

Our text merging algorithm takes the OCR system output $\mathbf{R}$ as input and finally outputs the merged text list. The algorithm consists of five steps:

1. Create a graph $G = (V, E)$ based on the Euclidean distance between the bounding boxes;
2. Calculate Union over Merge(UoM), Horizontal Intersect over Merge(HIoM) and Vertical Intersect over Merge(VIoM) between pairs of bounding boxes in graph G;
3. Merge the bounding boxes horizontally according to UoM and HIoM, and obtain the new OCR object array $\mathbf{R^h} = \{(text_i^h, box_i^h)\}$;
4. Repeat steps 1, 2 and 3 on $\mathbf{R^h}$;
5. Merge the bounding boxes vertically according to UoM and VIoM, and obtain the final OCR objects array $\mathbf{R^v} = \{(text_i^v, box_i^v)\}$.

Above, UoM measures the aggregation degree of the text bounding boxes in two-dimensional space, HIoM and VIoM measure the aggregation degree of the text bounding boxes in horizontal and vertical directions. For a pair of given bounding box (box_a, box_b), UoM, HIoM and VIoM are calculated as follows:

$$\begin{aligned} \text{UoM} &= \text{Union}(box_a, box_b)/box_m; \\ \text{HIoM} &= \text{Intersect}(\text{Hor}(box_a), \text{Hor}(box_b))/\text{Hor}(box_m)); \\ \text{VIoM} &= \text{Intersect}(\text{Ver}(box_a), \text{Ver}(box_b))/\text{Ver}(box_m)) \end{aligned} \tag{3}$$

where box_m is the merged box, i.e., the minimum bounding rectangle containing box_a and box_b. Hor and Ver indicate the projection operations in the horizontal and vertical directions, respectively. According to the above steps, we output $\{\mathbf{R}, \mathbf{R^h}, \mathbf{R^v}\}$ for the subsequent process.

3.3. Multi-Modal Learning

The input data of TextVQA task have three types: text of question, visual object in image, and image text. We process these data by different feature extraction methods, embed them into the same dimension d, fuse them by the Transformer module, and produce answers through the bilinear interaction module.

The text of question can be considered as a sequence of K words. We fine-tune a pre-trained BERT [19] to embed the question to the representation vector list $\{x_k^{que}\}$ $(k \in 1, ..., K)$, and utilize a pre-trained Faster R-CNN[20] to extract the visual objects in the image. The visual and location features of object are denoted as $\{x_m^{vis}\}$ and $\{x_m^{loc}\}$, respectively. We map visual and location features uniformly into d-dimension by two learnable linear networks. Then, the representation of the object is formulated as:

$$x_m^{obj} = LN(W_1 x_m^{vis}) + LN(W_2 x_m^{loc}) \tag{4}$$

where W_1 and W_2 are parameters of learnable linear network, $LN(.)$ is layer normalization. The image text preprocessed by the text merging algorithm contains N OCR objects $(text_n^{input}, box_n^{input})$. Embedding image text requires not only text character information, but also the appearance (e.g., color, font, and background) and spatial location. Hence, we extract features of image text by various approaches: FastText[21], Faster R-CNN[20] and PHOC[22], then combine the location feature of box_n^{input} to generate the embedding representation of image text:

$$x_n^{ocr} = LN(W_3 x_n^{fastr} + W_4 x_n^{phoc} + W_5 x_n^{fast}) + LN(W_6 x_n^{loc}) \tag{5}$$

After obtaining the embeddings of question, visual object and image text: $\{x_m^{obj}\}$, $\{x_k^{que}\}$ and $\{x_n^{ocr}\}$. We feed them into the multi-modal Transformer module to obtain the representation of image text $\{z_n^{ocr}\}$.

The decoding process is similar to machine translation. We first pass a d-dimensional initial vector x^{ans} through Transformer to obtain a representation z^{ans} containing multi-modal information. Then, we put z^{ans} and z_n^{ocr} into the bilinear interaction module to predict the probability of the n-th image text $text_n^{input}$ as a candidate answer, the formula is as follows:

$$y_n = (W^{ocr} z_n^{ocr} + b^{ocr})^T (W^{ans} z^{ans} + b^{ans}) \tag{6}$$

Finally, we take the image text with the highest probability as the answer.

4. PERFORMANCE EVALUATION

In this section, we conduct experiments on the public TextVQA dataset ST-VQA [11] to demonstrate the superiority of our proposed text merging algorithm and MML&TM.

Biten et al. proposed the ST-VQA dataset along with several baselines; some of them use the common words of the scene as the candidate answer list without an OCR system. We compare the optimal Accuracy and Average Normalized Levenshtein Similarity (ANLS) of different candidate answer lists on ST-VQA. Table 1 shows the experimental results. where the number in parentheses indicates the word list size. From Table 1, it can be seen that the accuracy of OCR is higher than that of the common words (19k), while the ANLS is lower than common words (19k), because each sample of the common words is accurate, while each sample output by the OCR system may be inaccurate due to recognition error. Furthermore, the accuracy and ANLS of our text merging method are the highest because text merging can combine the word-level text into paragraph-level candidate answers.

Table 1. Comparison of candidate answer lists on ST-VQA.

Candidate Answer	Accuracy (%)	ANLS
Common Words (1k)	31.96	0.571
Common Words (5k)	41.03	0.740
Common Words (19k)	52.31	0.862
OCR	68.84	0.782
Text Merging	82.57	0.904

Furthermore, we compare our MML&TM with three existing methods on ST-VQA: SAN+STR [11], VTA [13] and M4C [14], where M4C can be further divided into M4C-cla and M4C-seq according to the decoding method. As shown in Table 2, our MML&TM outperforms the existing methods. Compared with SAN+STR and M4C-cla, MML&TM can produce paragraph-level answers, resulting in higher accuracy and ANLS. It is worth noting that the accuracy of MML&TM is only 2.3% higher than M4C-seq, while the ANLS of MML&TM is 0.05 higher than M4C-seq. This is because M4C does not constrain the spatial aggregation in the sequence decoding stage, while the text merging module of MML&TM calculates the horizontal and vertical aggregation degrees by UoM, and merges the text pairs with larger aggregation degree.

Table 2. Performance comparison on ST-VQA.

Method	Accuracy(%)	ANLS
SAN+STR [11]	10.46	0.135
VTA [13]	18.13	0.282
M4C-cla [14]	33.52	0.397
M4C-seq [14]	38.05	0.462
MML&TM (ours)	**40.32**	**0.518**

5. CONCLUSION

In this paper, We present a text merging algorithm that combines word-level OCR output into paragraph-level text based on spatial aggregation degree. Then, we integrate the text merging algorithm into a multi-modal learning framework, and propose a Multi-Modal Learning framework with Text Merging (MML&TM) model for TextVQA. Experimental results suggest that the proposed text merging algorithm can obtain texts with more complete contextual information, which is more likely to be the answer for TextVQA.

Acknowledgement. Jihong Guan was partially supported by National Natural Science Foundation (NSFC) under grant No. U1936205.

6. REFERENCES

[1] Sam S Tsai, Huizhong Chen, David Chen, Georg Schroth, Radek Grzeszczuk, and Bernd Girod, "Mobile Visual Search on Printed Documents Using Text and Low Bit-rate Features," in *ICIP*, 2011, pp. 2601–2604.

[2] Eric Cheung and Kermin Hal Purdy, "System and Method for Text Translations and Annotation in An Instant Messaging Session," 2008, US Patent 7,451,188.

[3] Wen Wu, Xilin Chen, and Jie Yang, "Detection of Text on Road Signs from Video," *IEEE TITS*, vol. 6, no. 4, pp. 378–390, 2005.

[4] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach, "Towards Vqa Models that can Read," in *CVPR*, 2019, pp. 8317–8326.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual Question Answering," in *ICCV*, 2015, pp. 2425–2433.

[6] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh, "Making the V in Vqa Matter: Elevating the Role of Image Understanding in Visual Question Answering," in *CVPR*, 2017, pp. 6904–6913.

[7] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick, "Clevr: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning," in *CVPR*, 2017, pp. 2901–2910.

[8] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and Top-down Attention for Image Captioning and Visual Question Answering," in *CVPR*, 2018, pp. 6077–6086.

[9] Hedi Ben-Younes, Rémi Cadene, Matthieu Cord, and Nicolas Thome, "Mutan: Multimodal Tucker Fusion for Visual Question Answering," in *ICCV*, 2017, pp. 2612–2620.

[10] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining Task-agnostic Visiolinguistic Representations for Vision-and-language Tasks," *NIPS*, 2019.

[11] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas, "Scene Text Visual Question Answering," in *ICCV*, 2019, pp. 4291–4301.

[12] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty, "Ocr-vqa: Visual Question Answering by Reading Text in Images," in *ICDAR*, 2019, pp. 947–952.

[13] Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Minesh Mathew, CV Jawahar, Ernest Valveny, and Dimosthenis Karatzas, "Icdar 2019 Competition on Scene Text Visual Question Answering," in *ICDAR*, 2019, pp. 1563–1570.

[14] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach, "Iterative Answer Prediction with Pointer-Augmented Multimodal Transformers for Textvqa," in *CVPR*, 2020, pp. 9992–10002.

[15] Mengye Ren, Ryan Kiros, and Richard S Zemel, "Exploring Models and Data for Image Question Answering," in *NIPS*, 2015, pp. 2953–2961.

[16] Mateusz Malinowski and Mario Fritz, "A Multi-World Approach to Question Answering About Real-World Scenes Based on Uncertain Input," in *NIPS*, 2014, pp. 1682–1690.

[17] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu, "Are You Talking to A Machine? Dataset and Methods for Multilingual Image Question Answering," in *NIPS*, 2015, pp. 2296–2304.

[18] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alex Smola, "Stacked Attention Networks for Image Question Answering," in *CVPR*, 2016, pp. 21–29.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *NAACL*, 2019, p. 4171–4186.

[20] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE TPAMI*, vol. 39, no. 6, pp. 1137–1149, 2016.

[21] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, "Enriching Word Vectors with Subword Information," *TACL*, vol. 5, pp. 135–146, 2017.

[22] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny, "Word Spotting and Recognition with Embedded Attributes," *IEEE TPAMI*, vol. 36, no. 12, pp. 2552–2566, 2014.