Pattern Recognition 122 (2022) 108214

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

# A multimodal attention fusion network with a dynamic vocabulary for TextVQA

Jiajia Wu, Jun Du*, Fengren Wang, Chen Yang, Xinzhe Jiang, Jinshui Hu, Bing Yin, Jianshu Zhang, Lirong Dai

*National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP), University of Science and Technology of China, No. 96, JinZhai Road, Hefei, Anhui PR China*

ARTICLE INFO

*Article history:*
Received 2 November 2020
Revised 10 July 2021
Accepted 27 July 2021
Available online 19 August 2021

*Keywords:*
Dynamic vocabulary
Attention map
Multimodal fusion
ST-VQA

ABSTRACT

Visual question answering (VQA) is a well-known problem in computer vision. Recently, Text-based VQA tasks are getting more and more attention because text information is very important for image understanding. The key to this task is to make good use of text information in the image. In this work, we propose an attention-based encoder-decoder network that combines the multimodal information of visual, linguistic, and location features together. By using the attention mechanism to focus on key features to the question, our multimodal feature fusion can provide more accurate information to improve the performance. Furthermore, we present a decoder with attention map loss, which can not only predict complex answers but also deal with a dynamic vocabulary to reduce the decoding space. Compared with softmax-based cross entropy loss which can only handle a fixed-length vocabulary, the attention map loss significantly improves the accuracy and efficiency. Our method achieved the first place of all three tasks in the ICDAR2019 robust reading challenge on scene text visual question answering (ST-VQA).



## 1. Introduction

Visual question answering (VQA) [1] is an emerging research problem at the intersection of computer vision and natural language processing. It has become a hot topic recently since there are many applications in practice such as education of young children and assisting blind people [2]. The performance of VQA has been substantially improved in three important aspects. Firstly, deep learning improves feature representation [3] significantly and better visual and language feature representations [4] are the core parts for boosting VQA performance. Secondly, attention mechanisms [5] make model focus on salient image regions conditioned on question which can improve the accuracy of obtaining information related to the question. Thirdly, multimodal learning [6,7] can capture the high-level interactions between language and visual features.

Many datasets [1,8,9] and methods [10–13] are related to VQA, however most of them only focus on visual part of the scene. If there is text on an image, which is more informative in most cases, one can't answer the question without understanding the text in the image, see Fig. 1. For the above reasons, new VQA datasets [14–16] have been recently proposed with questions that require understanding the text in the image and the new task is called textVQA.

The textVQA task aims to infer answer over multi-modalities and a few methods [15–17] are presented to solve it. For example, the method introduced in [15] uses a text detector to get text in the image, then fuses text embedding and text visual feature through feedforward neural network and predicts answer with a softmax-based classifier. Obviously, there are some limitations in this method. First, it does not take more comprehensive modalities into consideration, lacking important information such as the object in the image and the location of text. Second, the answer to the question is generally related to certain key areas, and the way of feature fusion through DNN can't focus on the salient parts. Finally, the method treats answer prediction as a single-step classification problem, making it difficult to generate complex answers.

In this study, we address the above limitations by proposing a novel encoder-decoder framework to particularly predict complex answers. For getting the most relevant features to the question, we apply the attention mechanism [5] which can select features according to the question. Furthermore, to make better use of multimodal information, we employ a multilayer perceptrons to fuse multimodal features, including text embedding, visual embedding and position embedding to improve the model accuracy. On the other hand, we find that the answer in textVQA is often in the

* Corresponding author.
*E-mail address:* jundu@ustc.edu.cn (J. Du).

https://doi.org/10.1016/j.patcog.2021.108214
0031-3203/


Question: What is the text on the traffic sign?
Answer: school bus stop ahead

Question: Where is the train going to?
Answer: old town

Question: What is the written on the plane?
Answer: british airways

Question: Who holds the copyright?
Answer: charlotte edwards

**Fig. 1.** Some examples of ST-VQA.

text of the image and the task often provides candidate answers such as the task 1 in ST-VQA [18]. Therefore, the attention map loss is introduced to deal with the dynamic vocabulary problem and largely reduces the search space, which significantly improves the performance in terms of accuracy and efficiency, especially when the task provides the candidate answers.

We summarize the main contributions of this study as:

- We present an encoder-decoder framework for textVQA task which can generate complex answers.
- We fully utilize the multimodal features to improve model accuracy with attention mechanism.
- We introduce attention map loss which can address the dynamic vocabulary problem.

The rest of this paper is organized as follows: Section 2 introduces the related works. Section 3 describes the proposed framework of the whole system. Section 4 and Section 5 report the experimental results, and Section 6 presents concluding remarks and future work.

## 2. Related works

In this section, we describe the previous work related to textVQA, including object/text detection, text recognition, VQA based methods and decoding with dynamic vocabulary.

### 2.1. Object/text detection and text recognition

The earliest object detection method based on deep learning adopts R-CNN [19] to generate proposal boxes through selected search and further classify using neural network and SVM. Next, Fast R-CNN [20] and Faster R-CNN [21] extend this method in order to speed up the model. Nowadays, many detection methods [22,23] are based on Faster R-CNN, which are usually called two-stage methods. There are also one-stage methods like [24,25] that are often faster than two-stage methods with a little sacrifice of detection accuracy.

Text detection methods [26,27] usually adopt similar ideas as in general object detection methods. However, there are still some differences between them. For example, the shape of an object box is often rectangular while the shape of a text box can be arbitrary. In order to handle arbitrary shapes, approaches based on semantic segmentation are applied to text detection [28–30] which usually achieve good results.

Early text recognition methods were based on over segmentation, and then merged segments through character models and language models. Since segmentation is inaccurate in complex backgrounds and handwriting scene, the state of the art methods are often segmentation-free. The key of the segmentation-free method is how to align the feature sequence and label sequence. The CTC-based method [31] makes use of CTC loss to align them, ACE [32] uses aggregate cross entropy to align them and methods based on encoder-decoder framework [33] apply the attention mechanism to align feature sequence and label sequence. Recently, there are quite a few research efforts on irregular text recognition [34,35] and structured modeling [36,37].

### 2.2. VQA-based method for TextVQA

A large number of attention-based deep neural networks have been proposed for VQA [10,38]. These methods apply attention mechanism to select salient image regions conditioned on question. As VQA is essentially a vision-and-language task, multimodal learning which learns to fuse vision and language features is important in this task. Recently, BERT [4] as an extractor for text embedding is widely used in natural language tasks and self-attention mechanism [39] is found to be a better way to extract features. Accordingly, many researchers also deal with multimodal learning through self-attention mechanism [39] and BERT [4].

As for solving the textVQA problem, a few VQA-based methods [15–18] have been proposed. OCR-VQA [15] fuses all OCR token features and question embedding features using a feedforward neural network, which is followed by a softmax-based classifier to predict answers directly. LoRRA [16] extends OCR-VQA by selecting proper OCR token features through attention mechanism

conditioned on the question. In addition, a few other approaches such as [18] are based on the bottom-up and top-down attention model architecture [10] which is dedicated to improving the accuracy of attention mechanism. However, there are some drawbacks in existing methods such as a simple approach for multimodal feature learning, lack of ability to predict complex answers and the very large space of answers for the model to search. MM-GNN [40] represents an image as a graph and updates the node features by using graph neural network, which aims to utilize the rich information in the image to help understand the meaning of scene texts. M4C model [17] uses a multimodal transformer architecture to capture the semantic interactions over all the inputs and decodes the answer by choosing words from either the OCR tokens or the fixed vocabulary.

### 2.3. Decoding with dynamic vocabulary

Text tokens in image usually contain answers to the question in textVQA task. Recent works [15,16] have proposed to add text tokens to classifier vocabulary. The main problem of these approaches is that a large pre-defined vocabulary is required, which is quite challenging for model to predict answers, especially when the training data is limited. Therefore it is important to apply a dynamic vocabulary only related to the current image for textVQA task. Prior works such as pointer network [41] address the dynamic vocabulary problem by treating inputs as a vocabulary. But in textVQA, the vocabulary should be determined not only by text tokens in current image but also by fixed high-frequency words in training set. In this study, we introduce attention map loss to handle this problem.

## 3. Proposed method

In this section, we illustrate the main architecture of our method, including three modules. The first one is the encoder module which extracts the tokens in the image and question. The second module selects proper tokens which are related to the question and the last one is the decoder module which predicts the answer word by word. In Section 3.1, we introduce the extraction of the multimodal embedding from an image as the input in detail. In Section 3.2, we elaborate the attention mechanism to locate context features related to the question. In Section 3.3, we present the decoder with a dynamic vocabulary by using attention map loss.

### 3.1. Multimodal embedding

To fully utilize the multimodal information, we design the embedding of three different modalities from vision, sentence and position as the feature extraction as shown in the left part of Fig. 2.

**The embedding of vision.** We first use Mask R-CNN [22] based text detection model to generate the bounding box of text in the image and then produce recognition results through CRNN [31] based text recognition model. Similarly, we employ Faster R-CNN [21] to get the bounding box of object in the image. Next, the detected text and object are resized to $224 \times 224$ as input of pretrained model at ImageNet [42] as illustrated in Fig. 3. We take the global pooling feature of the last layer in the pretrained resnet model as the visual embedding of text and objects, which can be denoted as $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_N\}, \mathbf{v}_i \in \mathbb{R}^{d_v}$.

**The embedding of sentence.** After text/object detection and recognition, we obtain text recognition results and object entities. To generate the embedding of text recognition results and object entities as shown in Fig. 4, we adopt the pretrained BERT [4] model, which is widely used in many natural language and multimodal tasks. First, the input sentence is segmented into tokens through tokenizer. After that, token embedding, segment embedding and position embedding are concatenated as input of pretrained BERT model. Finally, the representation of each token is extracted from the output of the last layer. In practice, we use the BERT representation to produce the embedding of recognition results and object entities $\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_N\}, \mathbf{w}_i \in \mathbb{R}^{d_w}$. The embedding of the question sentence $\mathbf{q} \in \mathbb{R}^{d_w}$ is corresponding to the feature of the special element [CLS].

**The embedding of location.** Considering that image regions lack a natural ordering, we encode the location embedding for each region via a 4-dimensional vector:

$$\mathbf{p}_i = \left( \frac{x_1^i + x_2^i}{W}, \frac{y_1^i + y_2^i}{H}, \frac{x_2^i - x_1^i}{W}, \frac{y_2^i - y_1^i}{H} \right) \tag{1}$$

where $(x_1^i, y_1^i)$ and $(x_2^i, y_2^i)$ denote the coordinate of the bottom-left and top-right corner while W and H are the width and height of the input image. We use $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N\}, \mathbf{p}_i \in \mathbb{R}^{d_p}$ to represent the set of location features.

### 3.2. Multimodal fusion with attention

As shown in Fig. 2, we concatenate the embedding of visual, sentence and location to generate the input of subsequent modules. After obtaining the concatenated features, a two layer MLP is applied to transform the features.

$$\mathbf{F} = \text{RELU}\left(W_2 \,\text{RELU}\left(W_1[\mathbf{V}, \mathbf{W}, \mathbf{P}]\right)\right) \tag{2}$$

where $W_1 \in \mathbb{R}^{d_f \times (d_v + d_w + d_p)}$ and $W_2 \in \mathbb{R}^{d \times d_f}$ are learnable parameters, and the fusion feature $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_N\}, \mathbf{f}_i \in \mathbb{R}^d$. $N$ is the total number of texts and objects in the image and $d$ is the feature dimension. In order to speed up convergence, the weight normalization is used after each linear layer.

Then, an attention module takes the question sentence embedding as query and the fusion features as key-value pairs to obtain the attention weights on the values as the context feature:

$$\alpha_i = \text{softmax}\left(W_5((W_4\mathbf{q}) \odot (W_3\mathbf{f}_i))\right) \tag{3}$$

$$\mathbf{c} = \sum_i^N \alpha_i \mathbf{f}_i \tag{4}$$

where $W_3$ and $W_4 \in \mathbb{R}^{d_h \times d}, W_5 \in \mathbb{R}^{1 \times d_h}$, $d_h$ means the hidden dim of the linear projection, $\odot$ is element wise product. We set the dimension $d$ to be the same as the dimension $d_w$.

In addition, we suppose the attention module output contains most of the information related to the answer. Therefore, we can minimize L2 loss between the context feature vector and the sentence embedding vector of the answer to make the context features close to the answer embedding, which is demonstrated in Fig. 2.

### 3.3. Decoder with dynamic vocabulary

Based on the context features via the attention mechanism, we predict the answer to the question through a LSTM [43]. We decode the answer word by word for a total of $T$ steps until the end symbol is generated, where each decoded word is contained in our vocabulary. The vocabulary consists of frequent answer words in training set, text tokens and object names in current image. Similar to machine translation, we add $<$ begin $>$ and $<$ end $>$ to the vocabulary. $<$ begin $>$ is used as the first input to LSTM decoder and the decoding process stops after $<$ end $>$ is predicted. Since texts and objects in images are different, our vocabulary varies for each training/testing sample.

As shown in Fig. 5, the embedding of each word in our vocabulary is extracted by BERT [4] and denoted as $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$. In the

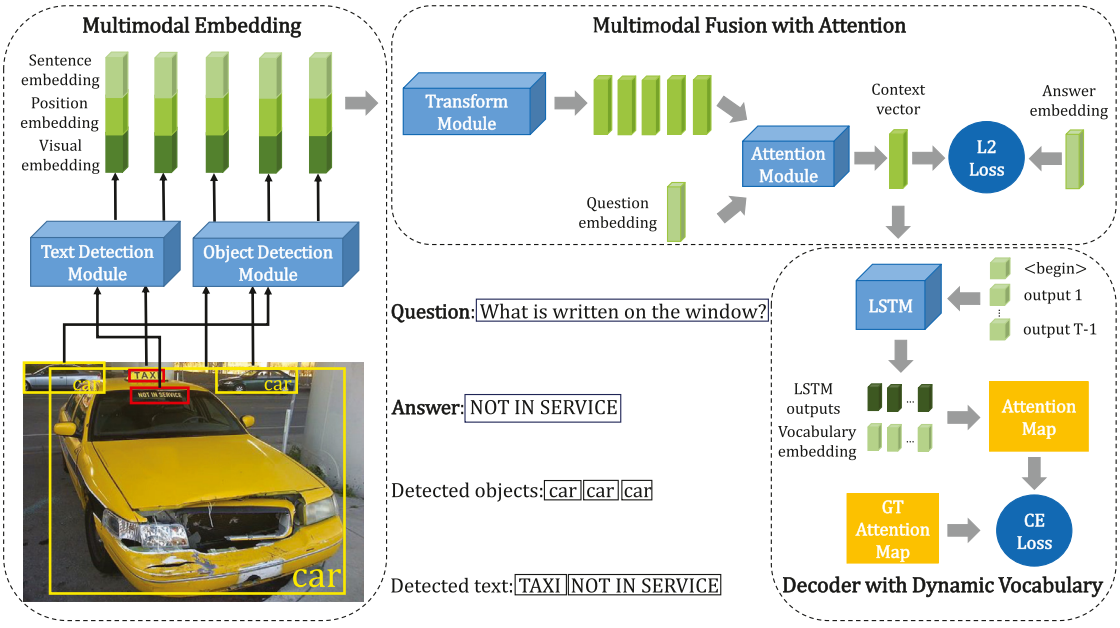



**Fig. 2.** A system overview of the proposed method, including multimodal embedding, multimodal fusion with attention, and decoder with dynamic vocabulary.

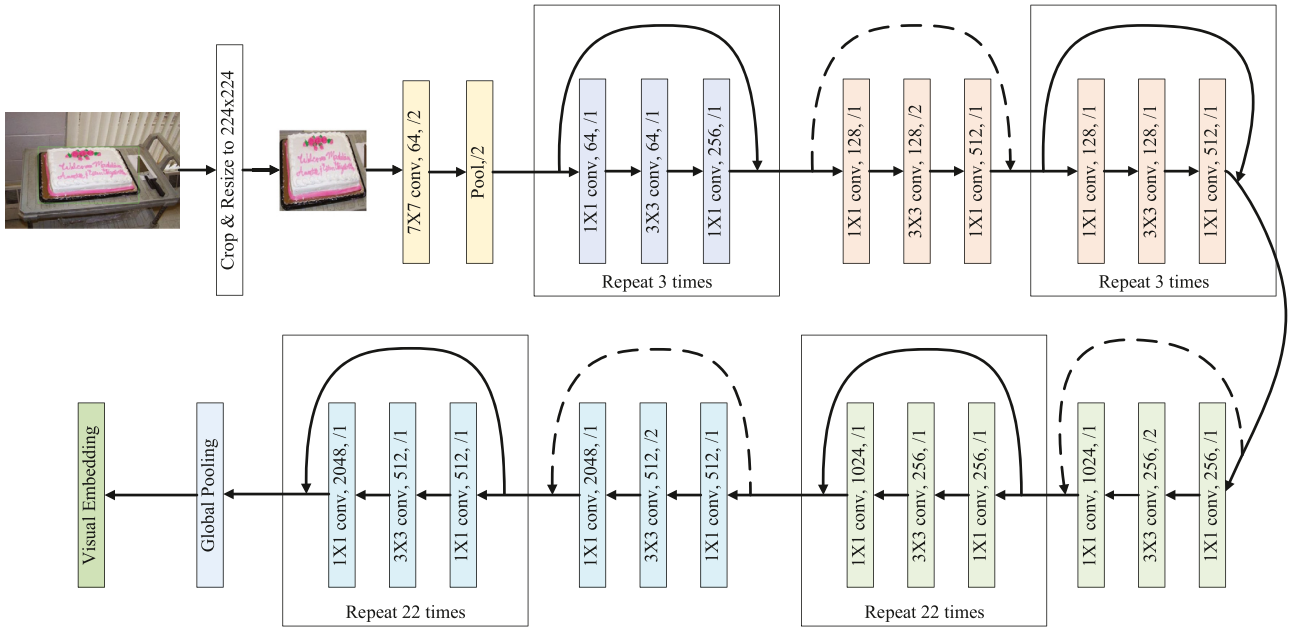



**Fig. 3.** Visual embedding through pretrained model. The backbone of pretrained model is ResNet101, ROI of image is cropped out and resized to $224 \times 224$ as input to ResNet101. '$7 \times 7$ conv, 64, /2' means the size of the convolution kernel is $7 \times 7$, the number of output channels is 64 and stride is 2.

$t$th step, the LSTM decoder receives context features and previous output as input to produce a feature vector in Eq. (5), which is denoted as $\mathbf{h}_t$. Then, we calculate the similarity between $\mathbf{h}_t$ and $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ as:

$$\mathbf{h}_t = \mathrm{LSTM}(\mathbf{y}_t, \mathbf{h}_{t-1}, \mathbf{c}_{t-1}) \tag{5}$$

$$\mathrm{Sim}(\mathbf{h}_t, \mathbf{a}_j) = \frac{\mathbf{h}_t^\top \mathbf{a}_j}{\|\mathbf{h}_t\| \|\mathbf{a}_j\|} \tag{6}$$

$$\beta_{t,j} = \mathrm{softmax}\left(\mathrm{Sim}(\mathbf{h}_t, \mathbf{a}_j)\right) \tag{7}$$

Next, we use softmax function to get normalized attention weights denoted as $\beta_{t,1}, \beta_{t,2}, \ldots, \beta_{t,M}$. After step $T$, we obtain matrix $\mathbf{A}$ with element $\mathbf{A}_{t,j} = \beta_{t,j}$ and we call this matrix attention map. The ground-truth attention map is denoted as $\mathbf{G}$ with element $\mathbf{G}_{t,j} = 1$ if the $t$th word in the answer is the $j$th word in our vocabulary and else $= 0$. We make $\mathbf{A}$ close to $\mathbf{G}$ by minimizing the cross entropy loss between each row of $\mathbf{A}$ and $\mathbf{G}$. Finally, given a ground truth sequence $y^*_{1:T}$ and a VQA model with parameters $\theta$, we minimize the L2 loss and the cross entropy(CE) loss jointly:

$$L(\theta) = \sum_{k=1}^{B} \left( \sum_{i=1}^{T} (\mathrm{CE}\left(\mathbf{A}^k_{i,:}, \mathbf{G}^k_{i,:}\right) + \left\| \mathbf{c}_k - \mathbf{a}^*_k \right\|_2^2 \right) \tag{8}$$

where $B$ is the batch size, $\mathbf{c}_k$ and $\mathbf{a}^*_k$ represent the context feature and the ground-truth embedding in the $k$th sample respectively.

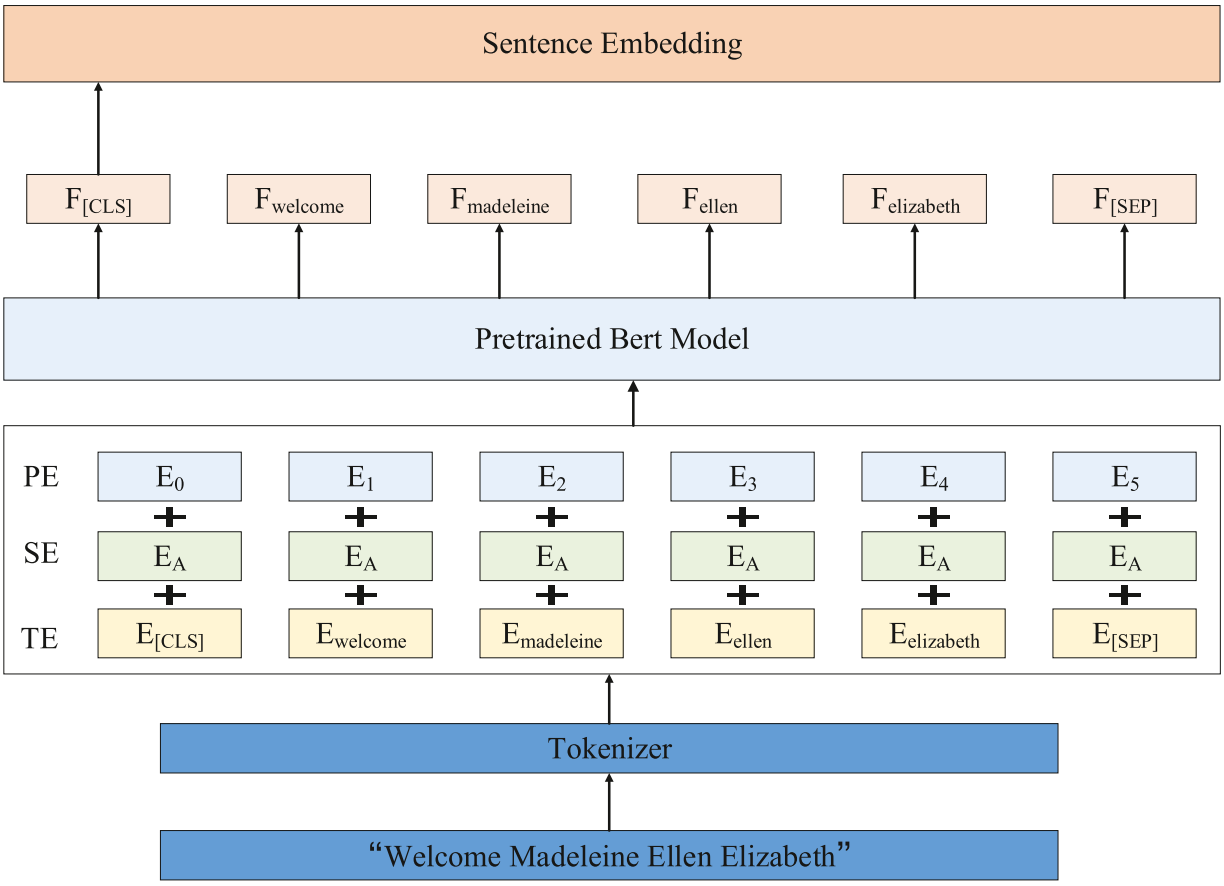



**Fig. 4.** Sentence embedding through BERT. First the sentence is split into words through tokenizer, each word is embedded via three modules which are token embedding (TE), segment embedding (SE) and position embedding (PE). Then the outputs are concatenated as input of BERT. Finally, the feature of the first token [CLS] is chosen as sentence embedding of the whole sentence.

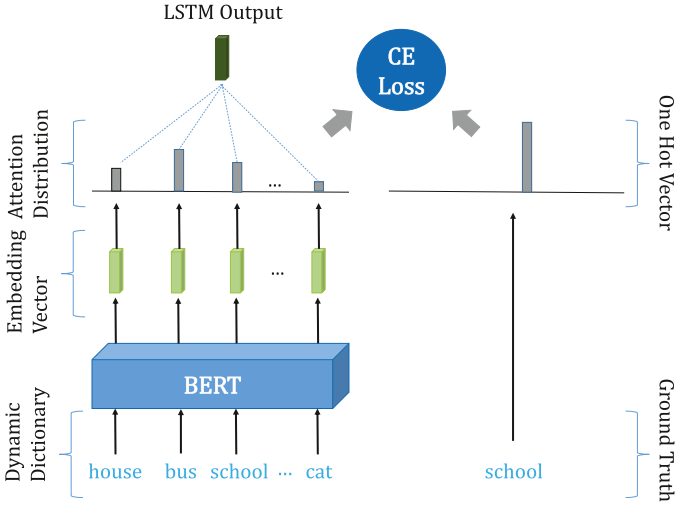



**Fig. 5.** Attention map loss with dynamic vocabulary. First, the embedding vector of each word in dynamic vocabulary is obtained through BERT [4]. Then attention weights are calculated between LSTM output and embedding vectors through (6). After that, we use softmax function to get normalized attention weights which are attention distribution. Ground truth distribution is one hot vector. At last, we make the two distributions(attention coefficients distribution and ground truth distribution) close through cross entropy loss.

If the ground-truth contains multiple words, we will use the sum average of multiple word embeddings as regression targets.

There are three advantages of using attention map loss. The first one is it can deal with variable-length vocabulary unlike softmax-based cross entropy loss used in OCR-VQA [15] which can only handle fixed-length vocabulary. This can greatly reduce the decoding space, therefore improve accuracy and we will show the details in Section 4. The second one is that it can make the training process more stable when the training data is insufficient since the parameters are initialized by BERT [4] and fixed during training. The last one is that it can build dynamic vocabulary based on text tokens in current image and fixed high-frequency words in training set at the same time, unlike pointer network [41] which only treats dynamic inputs as a vocabulary.

## 4. Experiments

### 4.1. Training and testing details

In the training stage, we supervise the model at each decoding step according to the answer label. Similar to machine translation, teacher forcing method in [44] is used to train our model which employs ground-truth inputs to the decoder. In particular, the previous ground-truth word of answer and context features are adopted as inputs to LSTM decoder to predict the next word of answer. For multimodal embedding, we use a Faster R-CNN detector [21] trained on the Visual Genome dataset to detect objects and keep 20 top-scoring objects per image. We extract the global pooling feature vector from pretrained model at ImageNet [42] as the visual embedding of object. Similarly, a text detector based on Mask R-CNN [22] trained on LSVT [45] is employed to detect

text and 2048-dimensional global pooling feature vector from pre-trained model at ImageNet [42] is extracted as the visual embedding of text. We adopt BERT [4] to generate a 768-dimensional feature vector of text and object name for sentence embedding. We use Eq. (1) to get a 4-dimensional feature vector of text and object bounding box for location embedding. The dimensions of $d$, $d_f$ and $d_h$ are set to 768, 1536 and 384 respectively. The LSTM decoder dimension is set to 1536.

In the training stage, the vocabulary consists of task 1 vocabulary, 50 most frequent words in the training set, recognized OCR results and object names for each training sample. The model is trained using the Adamax optimizer for 50 epochs. The learning rate is set to 1e-3 initially and the minimum learning rate is set to 1e-6. We use a cosine annealing decay learning rate schedule. The best model is selected using the validation set accuracy.

In the test stage, after getting the context feature through attention module, the context features and last decoded word are sent to LSTM as inputs in each decoding step. Then we calculate the similarity between the LSTM output and embedding of each word in vocabulary according to the Eq. (6) and the most similar word to LSTM output is the decoded result at this time step. The vocabulary varies for each task accordingly. In task 2/3, we only use recognized OCR results and object names as vocabulary. The maximum decoding step $T$ is set to 10.

### 4.2. Dataset and metric

The experiments are conducted on ST-VQA dataset [18] and TextVQA dataset [16], which are widely used public datasets for the task.The ST-VQA dataset comprises images from seven different public datasets: ICDAR 2013 [46], ICDAR2015 [47], ImageNet [42], VizWiz [2], IIIT Scene Text Retriev [48], Visual Genome [49] and COCO-Text [50]. The final version of the dataset consists of 23038 images and 31791 question/answer pairs. The dataset involves three tasks, namely task 1, task 2 and task 3. Task 1 provides for each image a different vocabulary of 100 words that includes the correct answer. In task 2, a global vocabulary of 30,000 words for all images is given. In task 3, no vocabulary is provided and the open vocabulary task is the most generic and challenging one among all the three tasks. The TextVQA dataset [16] contains about 28k images from the Open Images dataset [51], and is split into the training set (about 22k images and 34k QA pairs), the validation set (about 3k images and 5k QA pairs) and the test set (about 3k images and 5k QA pairs). Each question in the TextVQA dataset has 10 human annotated answers, and the final accuracy is measured via soft voting of the 10 answers, see Eq. (10).

The ST-VQA dataset adopts Average Normalized Levenshtein Similarity (ANLS) [52] as an evaluation metric:

$$\text{ANLS} = \frac{1}{N}\sum_{i=0}^{N}\left(\max_{j\in\{0,\ldots,M\}} s\left(a_{ij}^{\text{gt}}, a_i^{\text{pred}}\right)\right) \tag{9}$$

$$s(a_{ij}^{\text{gt}}, a_i^{\text{pred}}) = \begin{cases} 1 - \text{NL}(a_{ij}^{\text{gt}}, a_i^{\text{pred}}) & \text{if NL}(a_{ij}^{\text{gt}}, a_i^{\text{pred}}) < \tau \\ 0 & \text{if NL}(a_{ij}^{\text{gt}}, a_i^{\text{pred}}) \geq \tau \end{cases}$$

where $\text{NL}(a_{ij}^{\text{gt}}, a_i^{\text{pred}})$ is the Normalized Levenshtein distance between the ground truth string $a_{ij}^{\text{gt}}$ and the predicted string $a_i^{\text{pred}}$, $N$ is the total number of questions, $M$ is the number of ground truth answers per question. The threshold $\tau$ is set to 0.5.

The evaluation metric in TextVQA is the same as the VQA accuracy metric [53]:

$$\text{Accuracy}(ans) = \min\{\frac{\#\text{ humans that said } ans}{3}, 1\} \tag{10}$$

**Table 1**
Comparison of different methods on the ST-VQA competition set with the metric ANLS.

| Method | Task 1 | Task 2 | Task 3 |
|---|---|---|---|
| USTB-TQA [18] | 0.455 | 0.173 | 0.170 |
| USTB-TVQA [18] | 0.124 | 0.093 | 0.095 |
| Focus [18] | 0.295 | 0.080 | 0.088 |
| VQA-DML [18] | 0.141 | - | - |
| TMT [18] | 0.055 | - | - |
| QAQ [18] | - | - | 0.256 |
| Clova AI OCR [18] | - | - | 0.215 |
| SAN+STR [14] | - | - | 0.135 |
| STR [54] | 0.130 | 0.118 | 0.128 |
| Scene Image OCR [55] | 0.145 | 0.132 | 0.140 |
| Our method | **0.506** | **0.279** | **0.282** |

**Table 2**
Comparison of different methods on TextVQA with the metric accuracy.

| Method | Accuracy | |
|---|---|---|
| | Val | Test |
| OCR Max [16] | 0.0976 | 0.116 |
| BAN [38] | 0.123 | - |
| LoRRA [16] | 0.2656 | 0.2763 |
| Our method | **0.2842** | **0.289** |

Comparing the two metrics, ANLS is less strict than Accuracy, since ANLS takes into account the similarity between the predicted string and the ground truth string and Accuracy requires the two string to be exactly the same.

## 5. Results and analysis

In this section we will show the effectiveness of our proposed method on the ST-VQA and TextVQA datasets.

### 5.1. Comparison on the ST-VQA dataset

Table 1 compares the performance of different methods for three tasks on the ST-VQA competition set. The first block consists of 7 submitted systems to ST-VQA competition [18], namely from "USTB-TQA" to "Clova AI OCR" while the second block includes other three state of the art methods. Among them, "Focus" uses an algorithm similar to BUTD [10] with open-ended answer generation, "TMT" adopts a model similar to Dynamic Networks [56] and "Clova-AI OCR" employs a method similar to MAC network [57] with BERT [4] and pointing mechanism. "SAN+STR" [14] combines SAN for VQA and Scene Text Retrieval for answer vocabulary retrieval. "STR" is based on the scene text retrieval method presented in [54], which jointly predicts word bounding boxes and a compact text representation of words given in a PHOC [58] encoding. "Scene Image OCR" [55] employs a state of the art end-to-end scene text spotting model. From the table, obviously our method achieves higher ANLS (0.506 for task 1, 0.279 for task 2, and 0.282 for task 3[1]) than these methods on all three tasks.

### 5.2. Comparison on the TextVQA Dataset

Table 2 compares the performance of different methods on the TextVQA dataset. OCR Max [16] is a heuristic method which predicts the OCR token that is detected maximum times in the image as answer. BAN [38] is a VQA state-of-the-art method that does

[1] Since the test data of task2 and task3 are different, the performance of them have slight change.

**Table 3**
The ablation experiment of the proposed method on the task 1 of ST-VQA with the metric ANLS.

| multimodal embedding | | | question | L2 | attention | ANLS | speed(ms) |
|---|---|---|---|---|---|---|---|
| position | visual | sentence | embedding | loss | map | | |
| ✓ | | | | | ✓ | 0.4087 | 25.36 |
| | ✓ | | | | ✓ | 0.4415 | 27.04 |
| | | ✓ | | | ✓ | 0.4495 | 29.64 |
| ✓ | ✓ | ✓ | | | ✓ | 0.4571 | 31.01 |
| ✓ | ✓ | ✓ | ✓ | | ✓ | 0.4938 | 31.01 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.5049 | 33.14 |
| ✓ | ✓ | ✓ | ✓ | ✓ | × | 0.0693 | 32.86 |

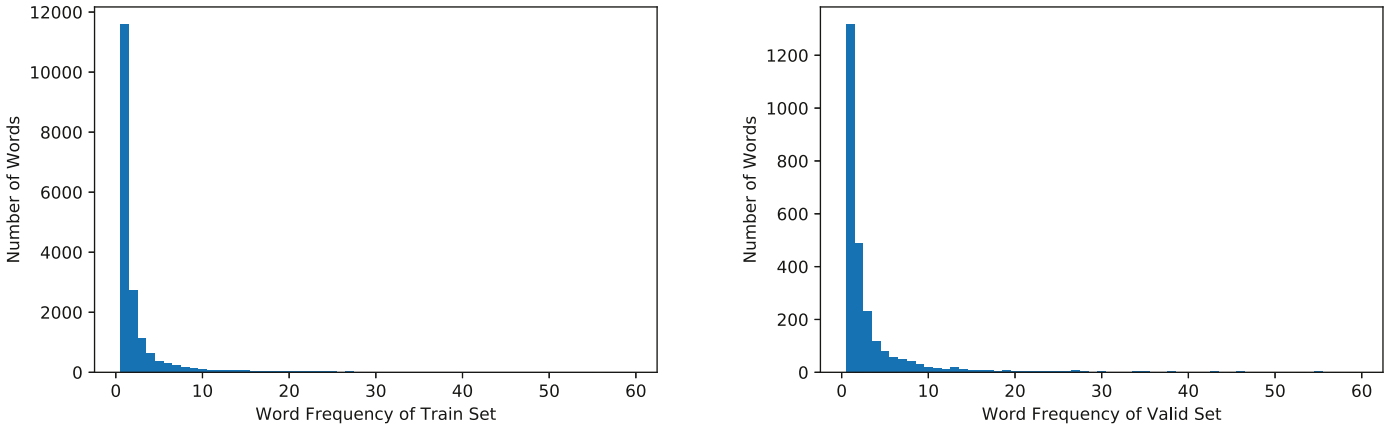



**Fig. 6.** Histogram of training and validation set.

not use image text information. LoRRA [16] adopts the same architecture of the VQA components for getting fused OCR-question features and image-question features which use text information in the image at the same time. From the table, obviously our method achieves higher accuracy than these methods.

### 5.3. Ablation study

As the ST-VQA dataset does not have an official split for training and validation, we randomly select 16,921 images as the training set and use the remaining 2,000 images as the validation set. Table 3 shows the performance of the proposed method and its ablations on validation set of task 1. To evaluate the effect of multimodal embedding, we first test the effect of each modality separately. The ANLS of using position embedding, visual embedding, and sentence embedding is 0.4087, 0.4415 and 0.4495 respectively. By using the embeddings of all three modalities together, the ANLS is increased to 0.4571, demonstrating the complementarity among different embeddings. We then test the effect of question embedding and it is worth noting that the ANLS will achieve 0.4938 if the question embedding is added as input to the decoder. After that, the ANLS will further increase to 0.5049 if the L2 loss is applied to hidden layer. Finally, to demonstrate the effectiveness of attention map loss, we compare it with softmax-based classifier, i.e. do not use the attention map module. We take the vocabulary provided by task 2 for the softmax classifier.

The ANLS of using softmax classifier is only 0.0693 which is significantly lower than that of using attention map loss. To give a better explanation, histograms of training set and validation set are shown in Fig. 6. It is observed that for both the training set and validation set, the sample number of the most commonly used words is less than 10. Therefore, it is difficult to train the model when vocabulary is large and the classification parameters are initialized randomly. However, if the attention map loss is applied, the size of dynamic vocabulary is much smaller, meanwhile classification parameters are pretrained from BERT [4] and fixed during training. These have greatly reduced the difficulty of model learning. We also test the average running time of an image in different configurations on Nvidia Tesla M40 GPU. From the table, we can clearly see that, multimodal embeddings bring a little time cost but with significant performance improvement (ANLS from 0.4087 to 0.5049).

**Table 4**
The ablation experiment of the proposed method on the task 2/3 of ST-VQA with the metric ANLS.

| multimodal embedding | | | question embedding | L2 loss | ANLS |
|---|---|---|---|---|---|
| position | visual | sentence | | | |
| ✓ | | | | ✓ | 0.0791 |
| | ✓ | | | ✓ | 0.1632 |
| | | ✓ | | ✓ | 0.2555 |
| ✓ | ✓ | ✓ | | ✓ | 0.2639 |
| ✓ | ✓ | ✓ | ✓ | ✓ | 0.2821 |

Similarly, we can draw consistent conclusions from the comparison experiment on task 2 and task 3. Please note that we use context vector to predict the answer directly on task 2 and task 3 in ST-VQA competition, leading to the same results for these two tasks. The reason is the dictionary provided by the task 2 is composed of 30,000 words, which is hard to train the model under the design of attention map module as mentioned above, while the task 3 does not give any vocabulary information. From Table 4, we can see that using the multimodal information including both visual and linguistic features can significantly improve the performance compared with the system using the single modality information (e.g., position embedding or visual embedding) as the input. And we also find that the sentence embedding is the most important one in solving the problem of visual question answer-

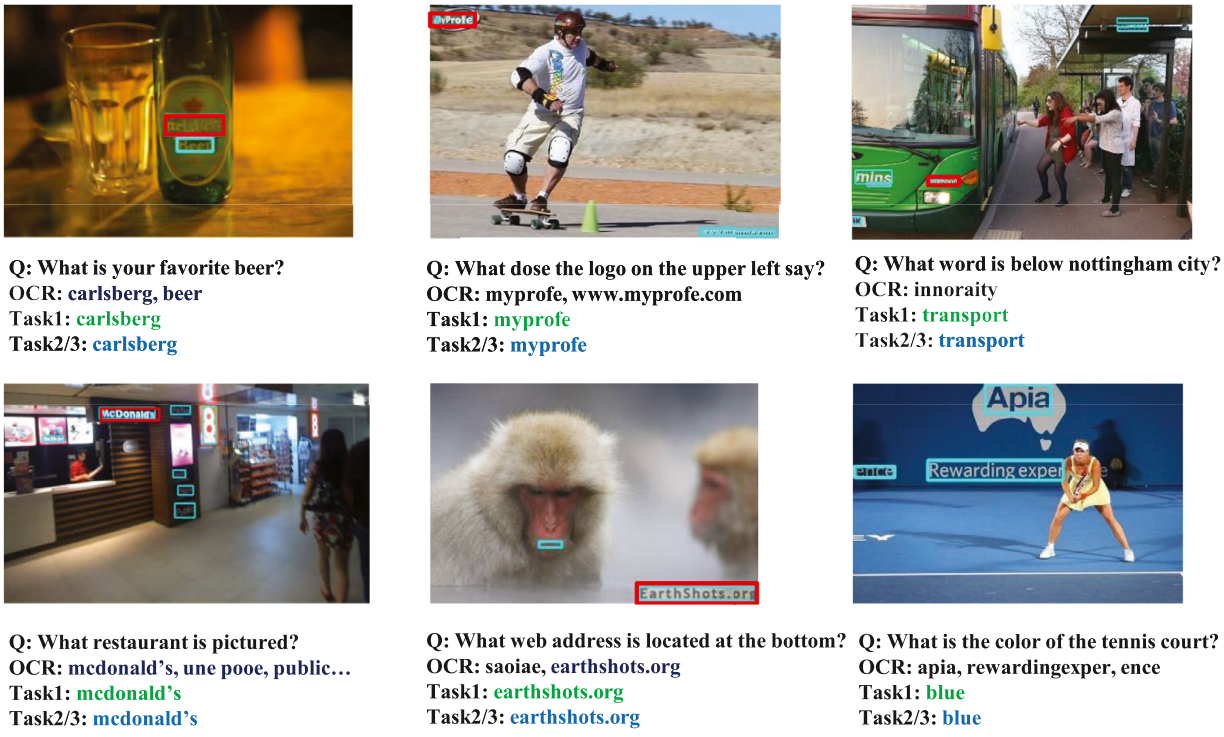



**Fig. 7.** Successful examples analysis. The detected OCR texts are bounded by the green boxes and the red box corresponds to the maximum attention weight.

**Table 5**
The ablation experiment of the proposed method on TextVQA validation set with the metric accuracy.

| multimodal embedding | | | question | L2 | attention | Accuracy |
|---|---|---|---|---|---|---|
| position | visual | sentence | embedding | loss | map | |
| ✓ | | | | | ✓ | 0.1837 |
| | ✓ | | | | ✓ | 0.2045 |
| | | ✓ | | | ✓ | 0.2275 |
| ✓ | ✓ | ✓ | | | ✓ | 0.2471 |
| ✓ | ✓ | ✓ | ✓ | | ✓ | 0.2814 |
| ✓ | ✓ | ✓ | ✓ | ✓ | ✓ | 0.2842 |
| ✓ | ✓ | ✓ | ✓ | ✓ | × | 0.0088 |

**Table 6**
The percentage of different error types on the eyeball set.

| Error Type | Percentage |
|---|---|
| w/o visual grounding | 30.65% |
| w/o numerical reasoning | 14.52% |
| w/o commonsense reasoning | 12.90% |
| long answer | 16.13% |
| OCR failure | 8.06% |
| others | 17.74% |

ing, which can achieve 0.2555 ANLS which is much better than the other two.

We also run ablation experiments on TextVQA validation set and the conclusion is basically the same as ST-VQA. From Table 5, we see that the accuracy is increased to 0.2471 by using the embeddings of all three modalities together. Question embedding and L2 loss also help improve the accuracy to 0.2842. Finally, attention map brings a huge performance improvement(0.0088 to 0.2842) than softmax classifier.

### 5.4. Qualitative analysis

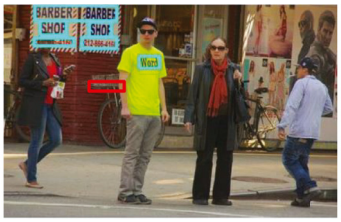
In Fig. 7, we show some successful cases for our model on ST-VQA dataset. The proposed system obtains the correct answer along with reasonable attention results when utilizing the multimodal contexts. The predicted OCR result is bounded in a green box.

In Table 6, we sample 200 image-question pairs randomly from the wrong predictions of the task l validation set as the eyeball set, and then analyze the percentage of different failure reasons. And in Fig. 8, we show some typical failure cases of our model to better illustrate the reasons. We can summarize these reasons or error types into five broad categories:

1. w/o visual grounding: our model lacks the capability of visual grounding, which aims to ground a natural language query (phrase or sentence) about an image onto a correct region of the image. Most errors (30.65%) are caused by the model not establishing a good relationship between the keywords in the sentence (entity, attribute, location, etc.) and the image regions.
2. w/o reasoning: numerical reasoning and commonsense reasoning are vital to complete some hard questions. To alleviate this problem, we need to equip the model with a numerically-aware module or commonsense-aware module.
3. w/o generating long answers: our decoder often generates only a fraction of the answer, which means that the decoder does not pay attention to the different areas of the image dynamically in each step or is not sensitive to the semantic relevance of the text of the long answer.
4. OCR failures: OCR failures can result in various mistakes, a portion of examples can be correctly answered with correct OCR recognition. As shown in Fig. 8, our model attends to the right position, but the OCR system result is wrong.
5. Others: this category includes question ambiguity, difficult to answer even by human and so on, which depends on the annotation of the dataset.

All in all, to build a more effective textVQA system, we need to improve in these aspects. Besides, how to effectively and practica-

Q: Who is the author of the book?
OCR: the tiger, judith kerr, came, to tea
GT: judith kerr
Task1: brooke tea
Task2/3: the tiger
Error Reason: w/o visual grounding



Q: how many pedestrians are in the picture?
OCR: barberebarber, shop, 212-866-4160
GT: Four
Task1: 212-866-4160
Task2/3: 212-866-4160
Error Reason: w/o numerical reasoning

Q: What is the smallest number on the clock?
OCR: 765, 1611613
GT: 1
Task1: 2
Task2/3: 765
Error Reason: w/o commonsense reasoning

Q: What is the name of the bar on the right?
OCR: loa, cure, gourmande
GT: la cure gourmande
Task1: la cure
Task2/3: loa cure
Error Reason: long answer

Question: What is the number on the police hat?
OCR: 41940, pong, ce
GT: 11940
Task1: 11940
Task2/3: 41940
Error Reason: OCR faliure

Q: Which is the second part of the brand name?
OCR: 2413, heatyk, kev2, roundys…
GT: mist
Task1: ocean
Task2/3: roundys
Error Reason: others(question ambiguity)

**Fig. 8.** Failure examples analysis. The detected OCR texts are bounded by the green boxes and the red box corresponds to the maximum attention weight.

bility construct dynamic vocabulary is also an important issue to improve system performance.

## 6. Conclusion and future work

In this study, we propose a novel framework for textVQA. This framework improves model accuracy through fusion of multimodal information, attention mechanism and attention map loss. It makes use of attention mechanism to get the most relevant features related to question and a LSTM decoder to predict complex answer word by word. In addition, we introduce attention map loss which can deal with dynamic vocabulary to greatly reduce the decoding space and significantly improve the model performance compared with softmax-based cross entropy loss. As for future work, we will probe the solutions to handle the dynamic vocabulary under the setting of task 2 and task 3. Besides, we will explore the more effective fusion mechanism to make better use of the complementarities among different modalities.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgment

We thank the reviewers for their constructive comments. This work was supported by the National Key Research and Development Program of China under Grant No. 2020AAA0107900.

## References

[1] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, VQA: visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2425–2433.

[2] D. Gurari, Q. Li, A.J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, 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.

[3] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[4] J.D.M.-W.C. Kenton, L.K. Toutanova, Bert: pre-training of deep bidirectional transformers for language understanding, in: Proceedings of NAACL-HLT, 2019, pp. 4171–4186.

[5] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, Y. Bengio, Show, attend and tell: neural image caption generation with visual attention, in: International Conference on Machine learning, 2015, pp. 2048–2057.

[6] H. Ben-Younes, R. Cadene, M. Cord, N. Thome, Mutan: multimodal tucker fusion for visual question answering, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2612–2620.

[7] J. Lu, D. Batra, D. Parikh, S. Lee, ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, in: Advances in Neural Information Processing Systems, 2019, pp. 13–23.

[8] D.A. Hudson, C.D. Manning, GQA: a new dataset for real-world visual reasoning and compositional question answering, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2019, pp. 6693–6702.

[9] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. Lawrence Zitnick, R. Girshick, CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2901–2910.

[10] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, 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.

[11] Z. Fang, J. Liu, Y. Li, Y. Qiao, H. Lu, Improving visual question answering using dropout and enhanced question encoder, Pattern Recognit. 90 (2019) 404–414.

[12] M. Farazi, S. Khan, N. Barnes, Accuracy vs. complexity: a trade-off in visual question answering models, Pattern Recognit. (2021) 108106.

[13] Z. Bai, Y. Li, M. Woźniak, M. Zhou, D. Li, DecomVQANet: decomposing visual question answering deep network via tensor decomposition and regression, Pattern Recognit. 110 (2021) 107538.

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

[15] A. Mishra, S. Shekhar, A.K. Singh, A. Chakraborty, OCR-VQA: visual question answering by reading text in images, in: Proceedings of the International Conference on Document Analysis and Recognition, vol. 1, 2019, p. 5.

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

[17] R. Hu, A. Singh, T. Darrell, 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. 9992–10002.
[18] A.F. Biten, R. Tito, A. Mafla, L. Gomez, M. Rusinol, M. Mathew, C.V. Jawahar, E. Valveny, D. Karatzas, ICDAR 2019 competition on scene text visual question answering, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE, 2019, pp. 1563–1570.
[19] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 580–587.
[20] R. Girshick, Fast R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1440–1448.
[21] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, in: Advances in Neural Information Processing Systems, 2015, pp. 91–99.
[22] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2961–2969.
[23] J. Dai, Y. Li, K. He, J. Sun, R-FCN: object detection via region-based fully convolutional networks, in: Advances in Neural Information Processing Systems, 2016, pp. 379–387.
[24] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: unified, real-time object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 779–788.
[25] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, SSD: single shot multibox detector, in: European Conference on Computer Vision, Springer, 2016, pp. 21–37.
[26] M. Liao, B. Shi, X. Bai, X. Wang, W. Liu, TextBoxes: a fast text detector with a single deep neural network, in: Thirty-First AAAI Conference on Artificial Intelligence, 2017.
[27] X. Zhou, C. Yao, H. Wen, Y. Wang, S. Zhou, W. He, J. Liang, East: an efficient and accurate scene text detector, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5551–5560.
[28] W. Wang, E. Xie, X. Li, W. Hou, T. Lu, G. Yu, S. Shao, Shape robust text detection with progressive scale expansion network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9336–9345.
[29] Y. Zhu, J. Du, TextMountain: accurate scene text detection via instance segmentation, Pattern Recognit. (2020) 107336.
[30] W. He, X.-Y. Zhang, F. Yin, Z. Luo, J.-M. Ogier, C.-L. Liu, Realtime multi-scale scene text detection with scale-based region proposal network, Pattern Recognit. 98 (2020) 107026.
[31] B. Shi, X. Bai, C. Yao, An end-to-end trainable neural network for image-based sequence recognition and its application to scene text recognition, IEEE Trans. Pattern Anal. Mach. Intell. 39 (11) (2016) 2298–2304.
[32] Z. Xie, Y. Huang, Y. Zhu, L. Jin, Y. Liu, L. Xie, Aggregation cross-entropy for sequence recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6538–6547.
[33] B. Shi, X. Wang, P. Lyu, C. Yao, X. Bai, Robust scene text recognition with automatic rectification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4168–4176.
[34] F. Zhan, S. Lu, ESIR: end-to-end scene text recognition via iterative image rectification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2059–2068.
[35] C. Luo, L. Jin, Z. Sun, MORAN: a multi-object rectified attention network for scene text recognition, Pattern Recognit. 90 (2019) 109–118.
[36] J. Zhang, J. Du, L. Dai, Radical analysis network for learning hierarchies of chinese characters, Pattern Recognit. (2020) 107305.
[37] C. Shi, C. Wang, B. Xiao, S. Gao, J. Hu, End-to-end scene text recognition using tree-structured models, Pattern Recognit. 47 (9) (2014) 2853–2866.
[38] J.-H. Kim, J. Jun, B.-T. Zhang, Bilinear attention networks, in: Advances in Neural Information Processing Systems, 2018, pp. 1564–1574.
[39] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[40] D. Gao, K. Li, R. Wang, S. Shan, X. Chen, 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, 2020, pp. 12746–12756.
[41] O. Vinyals, M. Fortunato, N. Jaitly, Pointer networks, in: Advances in Neural Information Processing Systems, 2015, pp. 2692–2700.
[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, 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.
[43] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Comput. 9 (8) (1997) 1735–1780.
[44] D. Bahdanau, K.H. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: 3rd International Conference on Learning Representations, ICLR 2015, 2015.
[45] Y. Sun, Z. Ni, C.-K. Chng, Y. Liu, C. Luo, C.C. Ng, J. Han, E. Ding, J. Liu, D. Karatzas, et al., ICDAR 2019 competition on large-scale street view text with partial labeling-RRC-LSVT, in: 2019 International Conference on Document Analysis and Recognition (ICDAR), IEEE Computer Society, 2019, pp. 1557–1562.
[46] D. Karatzas, F. Shafait, S. Uchida, M. Iwamura, L.G. i Bigorda, S.R. Mestre, J. Mas, D.F. Mota, J.A. Almazan, 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.
[47] 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.
[48] A. Mishra, K. Alahari, C.V. Jawahar, Image retrieval using textual cues, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 3040–3047.
[49] 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, Int. J. Comput. Vis. 123 (1) (2017) 32–73.
[50] A. Veit, T. Matera, L. Neumann, J. Matas, S. Belongie, Coco-text: dataset and benchmark for text detection and recognition in natural images, (2016) arXiv preprint arXiv:1601.07140
[51] I. Krasin, T. Duerig, N. Alldrin, V. Ferrari, S. Abu-El-Haija, A. Kuznetsova, H. Rom, J. Uijlings, S. Popov, A. Veit, 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) (2017) 18.
[52] V.I. Levenshtein, et al., Binary codes capable of correcting deletions, insertions, and reversals, in: Soviet Physics Doklady, vol. 10, Soviet Union, 1966, pp. 707–710.
[53] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the V in VQA matter: elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6904–6913.
[54] L. Gómez, A. Mafla, M. Rusinol, D. Karatzas, Single shot scene text retrieval, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 700–715.
[55] T. He, Z. Tian, W. Huang, C. Shen, Y. Qiao, C. Sun, An end-to-end textspotter with explicit alignment and attention, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5020–5029.
[56] C. Xiong, S. Merity, R. Socher, Dynamic memory networks for visual and textual question answering, in: International Conference on Machine Learning, 2016, pp. 2397–2406.
[57] D.A. Hudson, C.D. Manning, Compositional attention networks for machine reasoning, in: International Conference on Learning Representations, 2018.
[58] J. Almazán, A. Gordo, A. Fornés, E. Valveny, Word spotting and recognition with embedded attributes, IEEE Trans. Pattern Anal. Mach. Intell. 36 (12) (2014) 2552–2566.

**Jiajia Wu** received his B.Eng. degree from the Department of Mathematics and Applied Mathematics, Lanzhou University in 2010. He is currently a Ph.D. candidate of USTC and he also works for iFLYTEK Research. His current research areas include deep learning, handwriting mathematical expression recognition, Chinese document analysis and text detection.

**Jun Du** received his B.Eng. and Ph.D. degrees from the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC) in 2004 and 2009, respectively. From 2004 to 2009, he was with iFlytek Speech Lab of USTC. During the above years, he worked as an Intern for two 9-month periods at Microsoft Research Asia (MSRA), Beijing. In 2007, he worked as a Research Assistant for 6 months in the Department of Computer Science, University of Hong Kong. From July 2009 to June 2010, he worked at iFLYTEK Research on speech recognition. From July 2010 to January 2013, he joined MSRA as an Associate Researcher, working on handwriting recognition, OCR, and speech recognition. Since February 2013, he has been with the National Engineering Laboratory for Speech and Language Information Processing (NEL-SLIP) of USTC.