Improving OCR-based Image Captioning by Incorporating Geometrical Relationship

Jing Wang[1], Jinhui Tang[1]*, Mingkun Yang[2], Xiang Bai[2], and Jiebo Luo[3]
[1]Nanjing University of Science and Technology
[2]Huazhong University of Science and Technology [3]University of Rochester
{jwang,jinhuitang}@njust.edu.cn, {yangmingkun, xbai}@hust.edu.cn, jluo@cs.rochester.edu

Abstract

OCR-based image captioning aims to automatically describe images based on all the visual entities (both visual objects and scene text) in images. Compared with conventional image captioning, the reasoning of scene text is required for OCR-based image captioning since the generated descriptions often contain multiple OCR tokens. Existing methods attempt to achieve this goal via encoding the OCR tokens with rich visual and semantic representations. However, strong correlations between OCR tokens may not be established with such limited representations. In this paper, we propose to enhance the connections between OCR tokens from the viewpoint of exploiting the geometrical relationship. We comprehensively consider the height, width, distance, IoU and orientation relations between the OCR tokens for constructing the geometrical relationship. To integrate the learned relation as well as the visual and semantic representations into a unified framework, a Long Short-Term Memory plus Relation-aware pointer network (LSTM-R) architecture is presented in this paper. Under the guidance of the geometrical relationship between OCR tokens, our LSTM-R capitalizes on a newly-devised relation-aware pointer network to select OCR tokens from the scene text for OCR-based image captioning. Extensive experiments demonstrate the effectiveness of our LSTM-R. More remarkably, LSTM-R achieves state-of-the-art performance on TextCaps, with the CIDEr-D score being increased from 98.0% to 109.3%.

1. Introduction

The task of OCR-based image captioning is to describe images with sentences based on the visual entities (visual objects and scene text) contained in images. Despite the significant improvement made in conventional image captioning [35, 38, 30, 4, 41] and Optical Character Recognition (OCR) [9, 12, 19, 23, 33] in recent years, OCR-based image captioning still faces challenges and leaves much to be desired. To incorporate scene text for generating image descriptions, OCR-based image captioning needs to understand the words, recognize the visual entities of different modalities and reason among them.

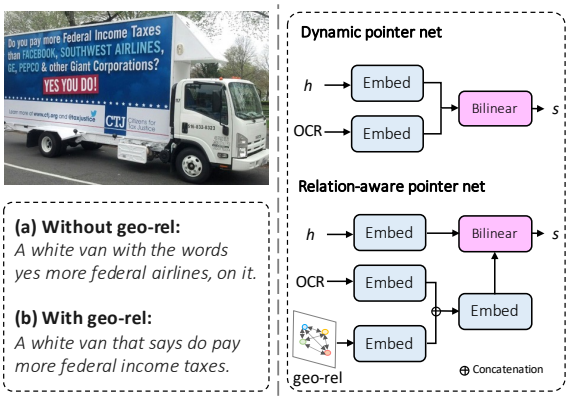


Figure 1. Caption examples from the models with or without geometrical relationship (geo-rel for short) and comparison between the conventional dynamic pointer network (e.g. [31]) and the proposed relation-aware pointer network. $\boldsymbol{h}$ and $\boldsymbol{s}$ denote the output state of the LSTM and the scores of the OCR tokens, respectively.

Among the modalities contained in OCR-based image captioning, the scene text, which is closely related to both visual content and language descriptions, is of vital importance. As has been verified, the model would witness a dramatic decline in the generation performance [31] if the information from scene text was ignored. To achieve better understanding and reasoning of the scene text in images, existing research [31, 37] jointly encodes visual feature, semantic feature and spatial feature of OCR tokens for a richer representation. Although promising results are obtained, the produced descriptions may still encounter problems about the OCR tokens. For example, existing methods (Figure 1(a)) may describe the upper-left image as "a white van with the words yes more federal airlines on it", which reads the OCR tokens in an incorrect order. This in-

*Corresponding author.

dicates that the relations between the OCR tokens learned by existing methods may be imprecise. One of the possible causes is that the visual feature and semantic feature could not always reflect the correlations between the tokens. This is because the spatially-adjacent OCR tokens may be semantically-unrelated or visually-unrelated to each other. In this case, the communication between the OCR tokens is more likely to lie in the size relations and position relations, which are collectively known as the geometrical relationship. Unfortunately, existing methods [31, 37] are limited in the exploration of geometrical relationship. M4C-Captioner [31] directly uses the bounding box coordinates of OCR tokens as spatial feature. MMA-SR [37] tries to find the next token for each OCR token according to the relative angles. These practices can hardly provide strong geometrical relationships for OCR tokens.

In this paper, we propose to improve OCR-based image captioning from the viewpoint of exploiting the geometrical relationship between OCR tokens. To establish the geometrical relationship, we consider the height/width difference, the IoU value, the shortest distance and the relative angle between OCR tokens. These elements constitute the intrinsic relevance between OCR tokens and would be beneficial for describing the images containing scene text. To take full advantage of the learned relation, we equip the captioning model with a newly-devised relation-aware pointer network, which selects OCR tokens under the guidance of the learned geometrical relationship. By this means, the model builds a more thorough understanding of the scene text and the OCR-centered problems appeared in the generated sentences are thus mitigated.

By consolidating the idea of incorporating the geometrical relationship between OCR tokens, we present a Long Short-Term Memory plus Relation-aware pointer network (LSTM-R) architecture for OCR-based image captioning, as depicted in Figure 2. Specifically, the pretrained Faster R-CNN [29] and OCR systems [1, 9] are first adopted to detect object regions and OCR regions from the image. After that, the features of the objects and the OCR tokens are extracted. In the meantime, the geometrical relationships between OCR tokens are exploited and are encoded into fixed-length vectors. Given the features and the learned relation vectors, the image captioner selects a word from the common vocabulary or the OCR tokens according to their scores at each time step. Specifically, the relation-aware pointer network is exploited to determine the scores for the OCR tokens under the guidance of the learned relation vectors. The entire model is optimized with the multi-label loss in the teacher-forcing way. In contrast to [31], we adopt different target labels as the supervision, since the OCR token appearing in captions may have multiple corresponding OCR regions in the image. In addition, considering that the words in captions are from different modalities (common vocabulary or OCR tokens), we design a word encoding method to decide the type of each sentence word and further encode the words according to their types which encourages the maximum use of the rich OCR representations and thus facilitates the model training.

The main contributions can be summarized as follows:

- We propose to explore the geometrical relationship between OCR tokens which enables more thorough understanding of OCR tokens as well as the entire image.
- We devise a relation-aware pointer network to select OCR tokens under the guidance of the geometrical relationship, which also provides an elegant view of how to integrate the learned relationship for OCR-based image captioning.
- We conduct extensive experiments on the TextCaps dataset and achieve the state-of-the-art performance.

2. Related work

Image Captioning. Image captioning witnesses great progress in recent years [35, 38, 42, 30, 4, 40, 39, 28, 20, 36, 14, 41, 18, 10, 25]. *Show and Tell* [35] is one of the pioneering works which first adopts the encoder-decoder framework. The encoder is implemented as CNN while RNN is chosen to be the decoder. Xu *et al*. [38] further intensify such a model by incorporating the attention mechanism, which provides a better attention on different regions of the feature maps in the image. The next breakthrough comes from the use of the self-critical sequence training strategy (SCST) proposed by Rennie *et al*. [30]. SCST adopts a better reward signal normalization method instead of estimating the reward signal, and thus trains the model more stably. Later on, Anderson *et al*. [4] improve the attention mechanism via upgrading it from the region-level to the object-level and present a new bottom-up-top-down combined attention. Most recently, the explorations are mainly along with the directions of learning the relationship between objects [40], improving the attention mechanism [14, 28] and adapting Transformer as the decoder [18, 10, 25]. Despite significant improvements made for image captioning, the models still fail to read from the image text and integrate them into the generated sentences, making OCR-based image captioning still a largely unexplored problem.

Optical Character Recognition (OCR). An OCR system essentially contains two parts, *i.e.*, text detection and text recognition. The earlier methods [21, 15, 11] usually consider text detection and text recognition as two independent tasks. Recent research [12, 19, 23] make an effort to integrate the detection and recognition models into a unified framework, which is gradually becoming mainstream. In [19], Li *et al*. first introduce an end-to-end text spotting framework via combining the text proposal network for text detection with the subsequent attention-based LSTM for

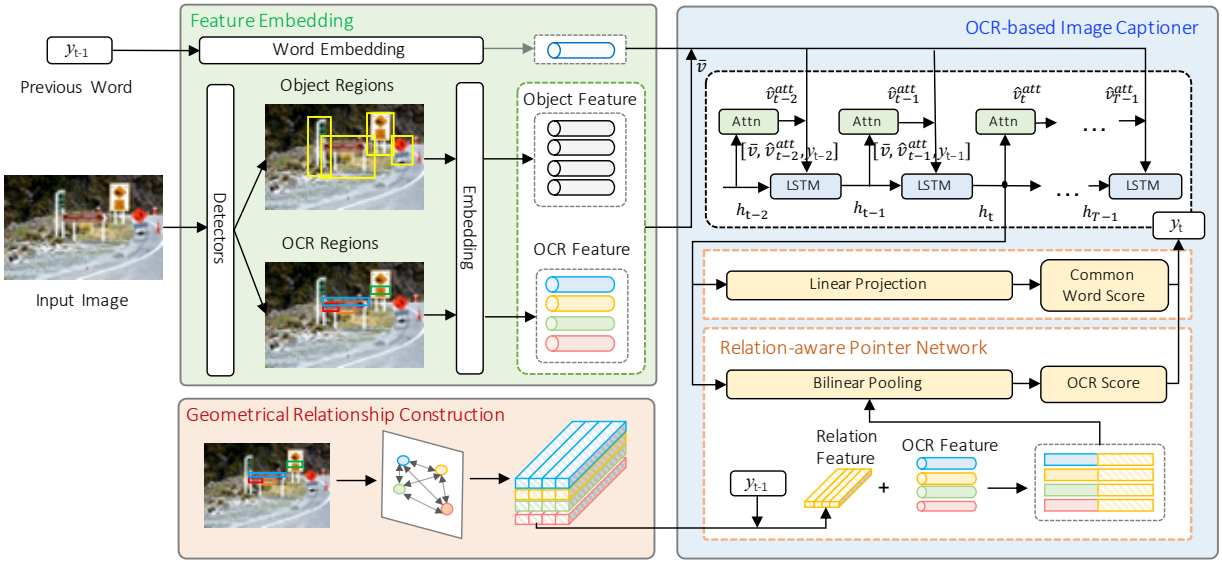


Figure 2. Framework of the proposed LSTM-R for OCR-based image captioning. In the feature embedding module, a set of object regions and a set of OCR regions are first detected, based on which the object features and the OCR features are extracted. In the geometrical relationship construction module, the relationships between OCR tokens are established. During the decoding stage, the LSTM decoder takes the visual features and the previously generated words as input, and produces the decoder state. Based on the decoder state and the learned geometrical relationship, the relation-aware pointer network determines the copying scores for the OCR tokens. A word is finally predicted according to the scores of the common words and the OCR tokens.

text recognition. Liu *et al*. [23] further improve this model by introducing a more practical text detector to cope with more challenging cases. Given the OCR tokens yielded by OCR systems, it is necessary to understand the semantic meaning of the tokens and model the relationships between OCR tokens and visual objects or other tokens during OCR-based image captioning.

VQA and Image Captioning based on Scene Text. The TextVQA task, requiring the model to answer visual questions based on the text appearing in images, has been newly proposed in recent years [32, 7, 24, 6, 13]. The task was first introduced in [32]. The authors extend Pythia [16] with an extra OCR attention branch to predict answers over a predefined vocabulary and OCR tokens presented in the image. To obtain more semantic information, Hu *et al*. [13] improve feature representation for OCR tokens and employ multimodal transformer layers to jointly encode multiple input modalities. Moreover, the answer is predicted by a dynamic pointer network in a multi-step manner. Kant *et al*. [17] further boost the performance by explicitly encoding spatial relations between objects and OCR tokens. The first attempt for OCR-based image captioning is done by Sidorov *et al*. [31]. The proposed model shares the main framework with M4C [13] and utilizes the multi-word answer decoder as the caption decoder. The most related work to our method is [37], which utilizes different attention mechanisms for different visual entities (objects and scene text) and tries to find the next token for each OCR token according to their relative angles.

Unlike [37] which mainly depends on the relative angle relation (usually a scalar), we explore a more exhaustive geometrical relationship which comprehensively considers the height, width, distance, IoU and relative angle relations between OCR tokens. In addition, we devise a relation-aware pointer network to select OCR tokens under the guidance of the learned relations.

3. Approach

In this work, we present a novel Long Short-Term Memory plus Relation-aware pointer network (LSTM-R) architecture for OCR-based image captioning. The overview of the architecture is depicted in Figure 2.

3.1. Overview

Given an image I, the task of OCR-based image captioning is to automatically generate a sentence S based on the visual entities (visual objects and scene text) in images. Technically, in the feature embedding module, a set of object regions $\{x_m^{obj}\}_{m=1:M}$ and a set of OCR tokens $\{x_n^{ocr}\}_{n=1:N}$ are first detected from the image. Appearance feature representations of the two modalities, $\{\boldsymbol{x}_m^{obj\text{-}a}\}_{m=1:M}$ and $\{\boldsymbol{x}_n^{ocr\text{-}a}\}_{n=1:N}$, are then extracted according to their bounding boxes. As OCR tokens contain not only visual information but also semantic information, FastText [8] feature $\{\boldsymbol{x}_n^{ft}\}_{n=1:N}$ and Pyramidal Histogram of Characters (PHOC) [2] feature $\{\boldsymbol{x}_n^{ph}\}_{n=1:N}$ are also extracted as additional supplements for the OCR tokens. Meanwhile, the geometrical relationship construction

module takes OCR tokens as input and establishes the relation vectors between the OCR tokens. Given the visual features and the learned relation vectors, the image captioner generates descriptions word by word in the decoding phrase. The captioner contains two main parts: a LSTM decoder and a relation-aware pointer network. At each time step, the LSTM jointly takes the average feature $\bar{\boldsymbol{v}}$, the embedding of the word $\boldsymbol{y}_{t-1}$ and the attended feature vector $\hat{\boldsymbol{v}}_{t-1}$ generated in the last time step as input, and produces a new hidden state $\boldsymbol{h}_t$. Based on $\boldsymbol{h}_t$, the scores of the common words and the OCR tokens, $\boldsymbol{s}_t^{com}$ and $\boldsymbol{s}_t^{ocr}$, are figured out respectively and a word is finally selected from the common vocabulary or the OCR tokens. While the scores for the words in the common vocabulary are produced by linear projection, a relation-aware pointer network is uniquely devised to compute the scores for the OCR tokens by incorporating the learned geometrical relationships which encourages stronger correlations between the OCR tokens. More details will be elaborated in the following sections.

3.2. Feature Embedding

To detect object regions and OCR regions in images, we first employ pretrained Faster R-CNN [29] and external OCR systems [1, 9] to produce bounding boxes for the two modalities, respectively. After that, the appearance features for the two modalities are both extracted from Faster R-CNN with $\{\boldsymbol{x}_m^{obj\text{-}a}\}_{m=1:M}$ denoting the object features and $\{\boldsymbol{x}_n^{ocr\text{-}a}\}_{n=1:N}$ standing for the OCR features.

Embedding of objects. Before being fed into the captioner, the object features are embeded to the same size as the LSTM hidden states via a linear projection:

$$\boldsymbol{x}_m^{obj} = \sigma(LN(\boldsymbol{W}_1 L_2N(\boldsymbol{x}_m^{obj\text{-}a}))), \tag{1}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{D_h \times D_v}$, L_2N means l_2 normalization, LN represents layer normalization and σ is the activation function (here we use ReLU, the same below).

Embedding of OCR tokens. As OCR tokens contain not only visual clues but also text information, following [13], we also extract FastText [8] feature $\boldsymbol{x}_n^{ft}$, Pyramidal Histogram of Characters (PHOC) [2] feature $\boldsymbol{x}_n^{ph}$ as additional supplements. Coordinates of bounding boxes are also taken into account in the form of $\boldsymbol{x}_n^{bb} = [x_{n,1:4}/w, y_{n,1:4}/h]$, where $x_{n,1:4}$ and $y_{n,1:4}$ are four x-coordinates and four y-coordinates of x_n^{ocr}, w and h are the width and height of the image, respectively. The final OCR embedding is thus obtained as follows:

$$\boldsymbol{x}_n^{ocr} = \sigma(LN(\boldsymbol{W}_2 \boldsymbol{x}_n^{main})) + \sigma(LN(\boldsymbol{W}_3 \boldsymbol{x}_n^{bb})) \; and \tag{2}$$

$$\boldsymbol{x}_n^{main} = [L_2N(\boldsymbol{x}_n^{ocr\text{-}a}); L_2N(\boldsymbol{x}_n^{ft}); L_2N(\boldsymbol{x}_n^{ph})], \tag{3}$$

where $\boldsymbol{W}_2$ and $\boldsymbol{W}_3$ are linear projections, $[\cdot;\cdot]$ denotes the concatenation operation.

Given the embeddings of objects and OCR tokens, the entire region set could be denoted as $\{\boldsymbol{x}_i\}_{i=1:M+N}$, where $\boldsymbol{x}_i$ is the feature embedding of an object or an OCR token.

3.3. Geometrical Relationship Construction

Recently, the spatial relationship of objects is found to be effective for multi-modal tasks such as visual grounding [27] and image captioning [40]. Inspired by this, we propose to upgrade the spatial relationship to the stronger geometrical relationship, which also holds size and position information for reproducing the relations between the OCR tokens appearing in images. Between every two OCR tokens x_i^{ocr} and x_j^{ocr}, we establish a pair of geometrical relationships $\boldsymbol{r}_{ij}^{ocr}$ and $\boldsymbol{r}_{ji}^{ocr}$ considering the following elements: 1) the height and width relation, 2) the distance relation, 3) the IoU relation, and 4) the relative angle relation.

Height and width relation is denoted as a vector with six values: $[w_i, h_i, w_j, h_j, w_d, h_d]$, where w_i, w_j, h_i, h_j are normalized widths and heights of x_i^{ocr} and x_j^{ocr}, $w_d = |w_i - w_j|$ and $h_d = |h_i - h_j|$ are the width difference and height difference between x_i^{ocr} and x_j^{ocr}, respectively.

Distance relation is computed as the shortest distance d_{ij} between the bounding boxes of x_i^{ocr} and x_j^{ocr}.

IoU relation contains three values: $[iou_{ij}, ioa_{ij}, ioa_{ji}]$, where iou_{ij} denotes the IoU between x_i^{ocr} and x_j^{ocr}, ioa_{ij} and ioa_{ji} are obtained via dividing the area of the intersection by the area of x_i^{ocr} and x_j^{ocr}, respectively.

Relative angle relation is defined as a value a_{ij} which reflects the relative angle between two tokens. To compute the angle, we first draw a line to connect the center points of x_i^{ocr} and x_j^{ocr}. Then the angle between the line and the positive direction of x-axis is obtained. According to the angle, we divide the angle relation into eight classes with each class spanning $45°$. Specifically, the angles between $22.5° \sim 67.5°$ are belong to the first class, between $67.5° \sim 112.5°$ are the second class, and so on. Thus the value of a_{ij} is a number between 1 and 8.

Relation embedding. Given the above, the geometrical relationship $\boldsymbol{r}_{ij}$ between x_i^{ocr} and x_j^{ocr} can be denoted as $[w_i, h_i, w_j, h_j, w_d, h_d, d_{ij}, iou_{ij}, ioa_{ij}, ioa_{ji}, a_{ij}]$. Following Section 3.2, $\boldsymbol{r}_{ij}$ is further projected to $\boldsymbol{r}_{ij}^{ocr}$ via:

$$\boldsymbol{r}_{ij}^{ocr} = \sigma(LN(\boldsymbol{W}_4 \boldsymbol{r}_{ij})), \tag{4}$$

where $\boldsymbol{W}_4 \in \mathbb{R}^{D_h \times D_r}$. $\boldsymbol{r}_{ji}^{ocr}$ is obtained in a similar way.

3.4. OCR-based Image Captioner

The OCR-based Image Captioner is composed of a LSTM, an attention module and a pointer network which is strengthened by the learned geometrical relationship. At each time step, we feed the combination of the visual feature and the previously predicted word into the LSTM and obtain the hidden state $\boldsymbol{h}_t$:

$$\boldsymbol{h}_t = \text{LSTM}(\boldsymbol{h}_{t-1}, [\bar{\boldsymbol{v}}; \hat{\boldsymbol{v}}_{t-1}; \boldsymbol{y}_{t-1}]), \tag{5}$$

where $\bar{\boldsymbol{v}} = \frac{1}{M+N}\sum_i^{M+N} \boldsymbol{x}_i$ and $\boldsymbol{x}_i$ is an object feature or an OCR feature. $\hat{\boldsymbol{v}}_{t-1}$ is the attended visual feature obtained from the last time step and it is initialized as $\bar{\boldsymbol{v}}$. $\boldsymbol{y}_{t-1}$ is the embedding of the word which is also generated in the last time step. If y_{t-1} is a common word, $\boldsymbol{y}_{t-1}$ is derived from the common word embedding. If it is copied from OCR tokens, $\boldsymbol{y}_{t-1}$ will be the feature embedding of the copied OCR token. Given the new hidden state $\boldsymbol{h}_t$, the attended vector $\hat{\boldsymbol{v}}_t$ which will be fed back into LSTM at the next time step is then figured out as

$$\hat{\boldsymbol{v}}_t = \sum_{i=1}^{M+N} \alpha_{t,i}\boldsymbol{x}_i \text{ and } \alpha_t = softmax(\boldsymbol{w}^T \tanh(\boldsymbol{W}^h \boldsymbol{h}_t + \boldsymbol{W}^v \mathbf{X})), \tag{6}$$

where $\boldsymbol{w}$, $\boldsymbol{W}^h$ and $\boldsymbol{W}^v$ are learned parameters.

At the end of the time step, a word is selected from the common vocabulary $\{v_k\}_{k=1:K}$ or from the OCR tokens $\{x_n^{ocr}\}_{n=1:N}$ according to their scores based on the hidden state $\boldsymbol{h}_t$. The scores are obtained by a linear projection and a relation-aware pointer network as described below.

Common word scores. The scores of the common words are directly figured out from the hidden state h_t through a linear layer: $\boldsymbol{s}_t^{com} = f(\boldsymbol{h}_t)$.

OCR scores by relation-aware pointer network. As the OCR tokens are various in different images and have no explicit order, we design a relation-aware pointer network to compute scores for the tokens. First, according to the input word y_{t-1}, the relation features $\boldsymbol{r}_n^{ocr}$ are fetched from the learned geometrical relationship $\boldsymbol{R}^{ocr}$ as:

$$\boldsymbol{r}_n^{ocr} = \begin{cases} \boldsymbol{r}_{pn}^{ocr} & y_{t-1} \in \{x_n^{ocr}\} \ and \ y_{t-1} = x_p^{ocr}, \\ zero & y_{t-1} \in \{v_k\}. \end{cases} \tag{7}$$

If $y_{t-1} \in \{v_k\}_{k=1:K}$, y_{t-1} is a common word and it has no geometrical relation with OCR tokens. In this case, the relation feature is replaced with the zero vector. Otherwise, $\boldsymbol{r}_n^{ocr}$ will be the corresponding geometrical relationship between y_{t-1} (i.e. x_p^{ocr}) and x_n^{ocr}.

Given the above relation feature $\boldsymbol{r}_n^{ocr}$ and the OCR representation $\boldsymbol{x}_n^{ocr}$ from Section 3.2, the augmented feature $\boldsymbol{v}_n^{ocr}$ of x_n^{ocr} is given by:

$$\boldsymbol{v}_n^{ocr} = [\boldsymbol{x}_n^{ocr}; \boldsymbol{r}_n^{ocr}] \tag{8}$$

We then compute the copying scores for the OCR tokens via a bilinear pooling operation based on the decoder state $\boldsymbol{h}_t$ and the augmented OCR features $\mathbf{V}^{ocr}$ as:

$$s_{t,n}^{ocr} = (\boldsymbol{W}^{hs}\boldsymbol{h}_t + \boldsymbol{b}^{hs})^T (\boldsymbol{W}^{ocr}\boldsymbol{v}_n^{ocr} + \boldsymbol{b}^{ocr}), \tag{9}$$

where $\boldsymbol{W}^{hs} \in \mathbb{R}^{D_h \times D_h}$, $\boldsymbol{W}^{ocr} \in \mathbb{R}^{D_h \times 2D_h}$, and $\boldsymbol{b}^{hs}, \boldsymbol{b}^{ocr} \in \mathbb{R}^{D_h}$.

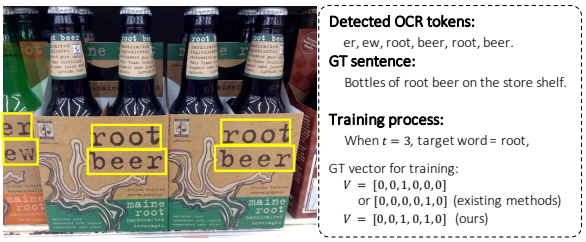


Figure 3. Example of the multi-label training process.

Combining the two scores, the probability of all the candidate words is figured out as $p(y_t) = Sigmoid\ ([\boldsymbol{s}_t^{com}; \boldsymbol{s}_t^{ocr}])$ and the word with the maximum probability will be selected at time step t.

3.5. Training and Inference

Word encoding method. We adopt the widely-used teacher-forcing strategy to train the model, i.e. LSTM takes words from ground truth sentences as input. In the conventional image captioning, only a few words are captured from scene text in images. Thus methods for conventional image captioning often build a vocabulary which contains most of the caption words and learn embedding for the words in the vocabulary. However, the problem is somewhat different in OCR-based image captioning, since words may come from both the common vocabulary and the OCR tokens. Unfortunately, such information of word type is not included in the annotated captions. Existing research [31] randomly assigns the word as a common word or an OCR token if the word can be found in both modalities. This may cause negative effects on the model training. As the frequency of most OCR tokens is quite low in the captions [31], their word embedding in common vocabulary is extremely hard to learn, which is known as the long-tail phenomena.

To mitigate the problem, we propose to determine the type of the words in advance and encode each word according to its type. In this case, a word can only be regarded as a common word or an OCR token. Through studying the annotated captions, we find out that the OCR tokens in the sentences are more likely to begin with capital letters, indicating that it is copied from the scene text. Based on this observation, we mark the words that appear in the detected scene text or begin with capital letters (except the beginning of the sentences) as OCR tokens, and the rest are considered as common words. Such an encoding method encourages the maximum use of the rich OCR representations and thus can facilitate the model training.

Training under multi-label sigmoid loss. Considering that an image may contain multiple OCR tokens that share same semantic meaning, we propose to optimize the model by minimizing the multi-label sigmoid loss. The training process is depicted in Figure 3. Suppose the detected OCR tokens in the image are "er, ew, root, beer, root, beer" and the ground truth sentence is "Bottles of root

Table 1. Performance (%) of our LSTM-R and other baseline methods on the TextCaps validation set.

#	Method	BLEU-4	METEOR	ROUGE-L	SPICE	CIDEr-D
1	Up-Down [4]	20.1	17.8	42.9	11.7	41.9
2	AoA [14]	20.4	18.9	42.9	13.2	42.7
3	M4C-Captioner w/o OCRs [31]	15.9	18.0	39.6	12.1	35.1
4	M4C-Captioner [31]	23.3	22.0	46.2	15.6	89.6
5	MMA-SR [37]	24.6	23.0	47.3	16.2	98.0
6	LSTM-R	**27.9**	**23.7**	**49.1**	**16.6**	**109.3**

Table 2. Performance (%) of the proposed LSTM-R and other baseline methods from the online TextCaps test server.

#	Method	BLEU-4	METEOR	ROUGE-L	SPICE	CIDEr-D
1	Up-Down [4]	14.9	15.2	39.9	8.8	33.8
2	AoA [14]	15.9	16.6	40.4	10.5	34.6
3	M4C-Captioner [31]	18.9	19.8	43.2	12.8	81.0
4	MMA-SR [37]	19.8	20.6	44.0	13.2	88.0
5	LSTM-R	**22.9**	**21.3**	**46.1**	**13.8**	**100.8**
6	M4C-Captioner w/GT OCRs (on a subset) [31]	21.3	21.1	45.0	13.5	97.2
7	Human [31]	24.4	26.1	47.0	18.8	125.5

beer on the store shelf". During training, when time step $t = 3$, the model is expected to generate the word "root", which is read from the OCR tokens in the image. In existing methods [37, 31], the target labels of OCR tokens are randomly assigned as $[0, 0, 1, 0, 0, 0]$ or $[0, 0, 0, 0, 1, 0]$, which indicates that only one OCR token in the image has the meaning of "root". In contrast, we set the target labels as $[0, 0, 1, 0, 1, 0]$, which allows an OCR word in the sentence to correspond to different OCR tokens that appear in the image. Comparing with the uncertain supervision of existing methods, our solution enables more stable training process of the model and thus leads to better performance.

Inference. During Inference, the model does not have input ground truth words. The captioner recursively takes the embeddings of the previously generated word as input to promote the generation process.

4. Experiments

4.1. Dataset and Setting

Dataset and metrics. We conduct experiments to verify the effectiveness of the proposed LSTM-R on the TextCaps [31] dataset which is designed for OCR-based image captioning. Each image in the dataset has five human-annotated captions. Follow [31], we use $21,953$ out of $28,408$ images to train the models, leave the rest $3,166$ and $3,289$ for validation and testing, respectively. We evaluate LSTM-R and other compared methods via the officially released coco-caption code[1], which considers the most widely used metrics (BLEU-4 [26], METEOR [5], CIDEr-D [34], ROUGE-L[22] and SPICE [3]) for image captioning.

Implementation details. Instead of directly using the caption tokens provided in [31], we first clean the captions in the training set by removing the special symbols unless the symbols appear in the detected image text (dubbed as Caption Clean below). To build the common vocabulary, all the words are converted into lowercase and only those who appear more than 10 times are preserved. The unique OCR vocabulary for each image is directly built from the detected image text. A special ⟨UNK⟩ token and an ⟨OCRUNK⟩ token are inserted into the common vocabulary and the OCR vocabulary, respectively. Types of the caption words are determined according to the proposed word encoding method. Given the word type, each caption token is matched with a word in the common or OCR vocabulary. If some words cannot find matches, they will be converted into ⟨UNK⟩ or ⟨OCRUNK⟩. The number of the detected object regions is set to 100 and each object region is represented as a 2048-d appearance feature. To detect OCR regions, both the Rosetta OCR system [9] and the Google OCR system [1] are adopted. The numbers of the detected OCR regions are various in different images and we use 80 OCR tokens at most. Apart from the 2048-d appearance feature and the bounding box coordinates, each OCR token also has a 300-d FastText feature and a 604-d PHOC feature. The hidden size of LSTM is set to 1000 and the maximum sentence length is 20. The mini-batch size is set to 50 in our experiments. The learning rate is initialized as 2×10^{-4} and is decreased every 3 epochs with the annealing factor 0.8. We train the model for about 30 epochs and choose the model with the highest CIDEr-D score to conduct online evaluation on the TextCaps test set.

Compared methods. (1) **Up-Down** and (2)**AoA** are two popular methods for image captioning. Top-Down incorporates the bottom-up and top-down attention into a unified framework. AoA devises an Attention-on-Attention

[1]https://github.com/tylin/coco-caption

Table 3. Ablation of each design (i.e. using Rosetta OCR system, adopting Google OCR system, cleaning the captions, adopting the word encoding method, optimizing with multi-label loss and incorporating geometrical relationship) in LSTM-R on the TextCaps validation set.

#	Rosetta OCR	Google OCR	Caption Clean	Word Encoding	Multi-label Loss	Geo Rel	BLEU-4	METEOR	ROUGE-L	CIDEr-D
base	√						24.8	22.1	46.8	91.2
1	√		√				25.4	22.5	47.4	94.4
2	√		√	√			26.0	22.9	48.0	99.0
3		√	√	√			25.8	22.8	47.7	99.7
4	√	√	√	√			27.1	23.4	48.4	105.8
5	√	√	√	√		√	27.1	23.5	48.5	106.8
6	√	√	√	√	√		27.4	23.5	48.9	107.4
7	√	√	√	√	√	√	**27.9**	**23.7**	**49.1**	**109.3**

module which can be plugged into both the encoder and the decoder. (3) **M4C-Captioner** shares the main framework with M4C [13], which is one of the state-of-the-art approaches for TextVQA. M4C-Captioner fuses information of different modalities by Transformer and computes scores for OCR tokens with a dynamic pointer network. **M4C-Captioner w/o OCRs** is trained without OCR tokens. **M4C-Captioner w/GT OCRs** provides results on a subset by using ground-truth OCR tokens. Up-Down, AoA and M4C-Captioner w/o OCRs do not have OCR tokens as input and the results are from [31]. (4) **MMA-SR** uses different attention modules for different modalities for a more thorough understanding of the image content. (5) **Human** means the sentences used for evaluation are provided by human. (6) **LSTM-R** is the proposed method in this paper.

4.2. Main Results

Offline Evaluation on the TextCaps Validation Set. Experimental results of the proposed LSTM-R and other compared methods on the TextCaps validation set are summarized in Table 1. LSTM-R achieves state-of-the-art performance among all the evaluation metrics. Unsurprisingly, the results of the methods (Up-Down, AoA and M4C-Captioner w/o OCRs) without OCR tokens as input are found to be much lower than the approaches (M4C-Captioner, MMA-SR and LSTM-R) with OCR tokens. The CIDEr-D scores of Up-Down, AoA and M4C-Captioner w/o OCRs are just around 40%, which are smaller than half the scores of the methods with OCR tokens. The importance of the incorporation of OCR tokens for OCR-based image captioning is thus verified. By taking the representations of OCR tokens as input, M4C-Captioner boosts the performance by a large margin, with the results being boosted up to 23.3%, 22.0% and 89.6% in BLEU-4, METEOR and CIDEr-D, respectively. MMA-SR further improves the CIDEr-D score from 89.6% to 98.0% via leveraging the angle-based spatial relationship between OCR tokens. This basically confirms that the spatial relationship is a good complement to the visual and semantic feature of OCR tokens. By exploring the geometrical relationship between OCR tokens and incorporating such information with the relation-aware pointer network, the proposed framework is superior to all the other compared methods, including the best existing approach MMA-SR. Compared with MMA-SR, LSTM-R gains absolute rises of 3.3%, 0.7%, 11.3% on BLEU-4, METEOR and CIDEr-D, respectively. The improvement demonstrates the advantage of capturing geometrical relationship between OCR tokens which encourages thorough understanding of scene texts in images.

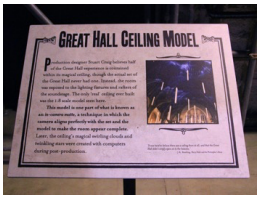
Online Evaluation on the TextCaps test set. We also submit the captions generated by our LSTM-R for TextCaps test set to the online test server [2]. The results in Table 2 present a similar pattern as in Table 1. The methods which are equipped with OCR tokens possess absolute advantage. The proposed LSTM-R outperforms MMA-SR and M4C-Captioner by 3.1% and 4.0% on BLEU-4, 0.7% and 1.5% on METEOR, 12.8% and 19.8% on CIDEr-D, respectively. This confirms the effectiveness of our framework again. Significantly, LSTM-R even exhibits better performance than M4C-Captioner w/GT OCR, whose captions are partially generated from ground truth OCR tokens. Although the results gained on the subset of TextCaps could not precisely reflect the concrete numbers, the gap between M4C-Captioner w/GT OCR and LSTM-R has been narrowed. The seventh line presents the results of human annotated sentences, which can be considered as the upper bound of the task. Although great improvement has been made by LSTM-R, the model still leaves much to be desired when being compared with human annotations.

Ablation Study. To elaborate the influence of each design in our method, we conduct ablation study by training the model without different parts. The results of the downgraded versions on the validation set of TextCaps are summarized in Table 3. Overall, the model is positively affected by each of the components. Cleaning captions (line 1 vs. base) increases CIDEr-D by 3.2%. One possible reason is that the noisy symbols in captions are so abundant that the training of the model is affected. Line 2-4 analyse the effect of leveraging different OCR systems. Specifically, using ei-

[2]https://eval.ai/web/challenges/challenge-page/573/leaderboard

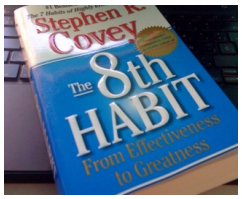
W/O Geo-Rel: A poster of mad singapore that says mad about singapore.
W/ Geo-Rel : Mad about singapore is an exhibition of singapore Instagrammers community.
Human: Mad about Singapore an exhibition of singapore instagram flyer.

W/O Geo-Rel: A blue sign that says alle geht vom volke aus.
W/ Geo-Rel : A sign that says alle macht geht vom volke aus.
Human: The message Alle Macht geht vom Volke aus, is posted on a blue sign.



W/O Geo-Rel: A sign that says great hall model model on it.
W/ Geo-Rel : A poster that says great hall ceiling model on it.
Human: Great Hall Ceiling Model sign that is for people to view and see.

W/O Geo-Rel: A yellow sign that says left left ahead ahead.
W/ Geo-Rel : A yellow sign that says left lane closed ahead.
Human: A sign that says Left Lane Closed Ahead on a highway.



W/O Geo-Rel: A stack of books with the title the habit from greatness.
W/ Geo-Rel : A book titled the 8th habit from greatness.
Human: Stephen Covey's The 8th Habit is sitting on a keyboard.

W/O Geo-Rel: A red sign that says ensure you wear your emergency set.
W/ Geo-Rel : A red sign that says ensure you always wear your escape set.
Human: Red sign hanging which says "Ensure you always wear your emergency escape set".

W/O Geo-Rel: A billboard that says what's a video sharing.
W/ Geo-Rel : A billboard that says what's a hipchat on it.
Human: Large billboard that says "What's a Hipchat" on it.

W/O Geo-Rel: A sign that says ixtlan del rio on it.
W/ Geo-Rel : A sign that says ixtlan del rio guadalajara cd de Mexico.
Human: Highway billboard showing distance to Ixtlan Del Rio, Guadalajara, and Cd. De Mexico.

Figure 4. Exemplar captions generated by LSTM-R w/o geometrical relationship (W/O Geo-Rel), LSTM-R (W/ Geo-Rel) and from human annotations. The words in same colors indicate the exact semantic matches of OCR tokens between the generated captions and the human-provided captions. (Best viewed in color)

ther of the OCR systems behaves similarly while using both of them exhibits much better results, leading to about 6.0% increase in CIDEr-D. We preliminarily thought that the existing OCR systems are good enough for our task. However, after digging into the OCR tokens obtained from the OCR systems, it turns out that both systems miss or mistake some OCR regions and better OCR systems are still needed. As expected, adopting the word encoding method (line 2 vs. line 1) performs well for the model, which certificates the advantage of determining the type of the caption words. Moreover, the benefit of optimizing the model with the proposed multi-label loss is confirmed in line 6 (vs. line 4). By exploring the geometrical relationships and incorporating all the above designs, line 7 further enhances the model and witnesses the most significant performance.

Qualitative Results. Figure 4 showcases some exemplar captions generated by LSTM-R w/o geometrical relationship, LSTM-R and human. Generally, our LSTM-R presents smoother sentences with more accurate OCR tokens. Take the first image as an example. Compared with the method without geometrical relationship which omits the "about" in "mad singapore" at the first time and includes repeat tokens in the sentence, LSTM-R describes the OCR tokens in the correct order and covers most of the tokens that appear in the human-annotated caption. In the last image, LSTM-R w/o geometrical relationship only talks about the "ixtlan del rio" whereas LSTM-R details most of the OCR tokens contained in the image. This mainly benefits from the incorporation of the geometrical relationship, which emphasizes the connections between OCR tokens and enhances the capability of the generation model.

5. Conclusions

In this paper, we present a novel Long Short-Term Memory plus Relation-aware pointer network (LSTM-R) architecture which explores the geometrical relationship between OCR tokens for OCR-based image captioning. To incorporate the learned relation into the framework, we devise a relation-based pointer network which copies words from the OCR tokens under the guidance of the geometrical relationship. To facilitate the training of the model, we also design a word encoding method and propose to optimize the model with multi-label loss. Experimental results on TextCaps certify the effectiveness of all the components. More remarkably, our LSTM-R exhibits superior performance over the benchmarks and achieves state-of-the-art performances on the TextCaps dataset.

Acknowledgments

This work was supported by the National Key Research and Development Program of China under Grant 2018AAA0102002, and the National Natural Science Foundation of China under Grant 61925204.

References

[1] Googleocr. https://cloud.google.com/vision/docs/ocr#optical_character_recognition_ocr.

[2] Jon Almazán, Albert Gordo, Alicia Fornés, and Ernest Valveny. Word spotting and recognition with embedded attributes. *IEEE TPAMI*, 36(12):2552–2566, 2014.

[3] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[4] 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 vqa. In *CVPR*, 2018.

[5] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL workshop*, 2005.

[6] 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. *arXiv preprint arXiv:1907.00490*, 2019.

[7] 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.

[8] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *TACL*, 5:135–146, 2017.

[9] Fedor Borisyuk, Albert Gordo, and Viswanath Sivakumar. Rosetta: Large scale system for text detection and recognition in images. In *SIGKDD*, 2018.

[10] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *CVPR*, 2020.

[11] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. Synthetic data for text localisation in natural images. In *CVPR*, 2016.

[12] Tong He, Zhi Tian, Weilin Huang, Chunhua Shen, Yu Qiao, and Changming Sun. An end-to-end textspotter with explicit alignment and attention. In *CVPR*, 2018.

[13] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. In *CVPR*, 2020.

[14] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *ICCV*, 2019.

[15] Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Reading text in the wild with convolutional neural networks. *IJCV*, 116(1):1–20, 2016.

[16] Yu Jiang, Vivek Natarajan, Xinlei Chen, Marcus Rohrbach, Dhruv Batra, and Devi Parikh. Pythia v0. 1: the winning entry to the vqa challenge 2018. *arXiv preprint arXiv:1807.09956*, 2018.

[17] Yash Kant, Dhruv Batra, Peter Anderson, Alexander Schwing, Devi Parikh, Jiasen Lu, and Harsh Agrawal. Spatially aware multimodal transformers for textvqa. In *ECCV*, 2020.

[18] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. Entangled transformer for image captioning. In *ICCV*, 2019.

[19] Hui Li, Peng Wang, and Chunhua Shen. Towards end-to-end text spotting with convolutional recurrent neural networks. In *ICCV*, 2017.

[20] Zechao Li, Jinhui Tang, and Tao Mei. Deep collaborative embedding for social image understanding. *IEEE TPAMI*, 41(9):2070–2083, 2018.

[21] Minghui Liao, Baoguang Shi, Xiang Bai, Xinggang Wang, and Wenyu Liu. Textboxes: A fast text detector with a single deep neural network. In *AAAI*, 2017.

[22] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *ACL workshop*, 2004.

[23] Xuebo Liu, Ding Liang, Shi Yan, Dagui Chen, Yu Qiao, and Junjie Yan. Fots: Fast oriented text spotting with a unified network. In *CVPR*, 2018.

[24] Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. Ocr-vqa: Visual question answering by reading text in images. In *ICDAR*, 2019.

[25] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. X-linear attention networks for image captioning. In *CVPR*, 2020.

[26] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[27] Bryan A. Plummer, Arun Mallya, Christopher M. Cervantes, Julia Hockenmaier, and Svetlana Lazebnik. Phrase localization and visual relationship detection with comprehensive image-language cues. In *ICCV*, 2017.

[28] Yu Qin, Jiajun Du, Yonghua Zhang, and Hongtao Lu. Look back and predict forward in image captioning. In *CVPR*, 2019.

[29] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[30] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017.

[31] Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. Textcaps: a dataset for image captioning with reading comprehension. In *ECCV*, 2020.

[32] 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.

[33] Ray Smith. An overview of the tesseract ocr engine. In *ICDAR*, 2007.

[34] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[35] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[36] Jing Wang, Yingwei Pan, Ting Yao, Jinhui Tang, and Tao Mei. Convolutional auto-encoding of sentence topics for image paragraph generation. In *IJCAI*, 2019.

[37] Jing Wang, Jinhui Tang, and Jiebo Luo. Multimodal attention with image text spatial relationship for ocr-based image captioning. In *ACM MM*, 2020.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2015.

[39] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *CVPR*, 2019.

[40] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *ECCV*, 2018.

[41] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Hierarchy parsing for image captioning. In *ICCV*, 2019.

[42] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.