

# Bridging by Word: Image-Grounded Vocabulary Construction for Visual Captioning

Zhihao Fan<sup>1</sup>, Zhongyu Wei<sup>1\*</sup>, Siyuan Wang<sup>1</sup>, Xuanjing Huang<sup>2</sup>

<sup>1</sup>School of Data Science, Fudan University, China

<sup>2</sup>School of Computer Science, Fudan University, China

{fanzh18,zywei,wangsy18,xjhuang}@fudan.edu.cn

## Abstract

Existing research for visual captioning usually employs a CNN-RNN architecture that combines a CNN for image encoding with a RNN for caption generation, where the vocabulary is constructed from the entire training dataset as the decoding space. Such approaches typically suffer from the problem of generating N-grams which occur frequently in the training set but are irrelevant to the given image. To tackle this problem, we propose to construct an image-grounded vocabulary that leverages image semantics for more effective caption generation. More concretely, a two-step approach is proposed to construct the vocabulary by incorporating both visual information and relationships among words. Two strategies are then explored to utilize the constructed vocabulary for caption generation. One constrains the generator to select words from the image-grounded vocabulary only and the other integrates the vocabulary information into the RNN cell during the caption generation process. Experimental results on two public datasets show the effectiveness of our framework compared to state-of-the-art models. Our code is available on Github<sup>1</sup>.

## 1 Introduction

Recent years have witnessed growing popularity of research in multimodal learning across vision and language. Image captioning (Xu et al., 2015), one of the most widely studied multimodal tasks, aims at constructing a short text description given an image. Existing research on image captioning usually employs a CNN-RNN architecture with a Convolutional Neural Network (CNN) used for image feature extraction and a Re-



*NIC*: a woman sitting at a table with a cell phone.  
*GT*: a woman sitting next to a display on the ground of jewelry is talking on the phone.

*NIC*: a woman sitting at a table with a cake.  
*GT*: a woman standing over a pan filled with food in a kitchen.

Figure 1: Two images from MS-COCO (Lin et al., 2014) with captions generated by *NIC* (Vinyals et al., 2015) and the corresponding ground truth (*GT*) captions.

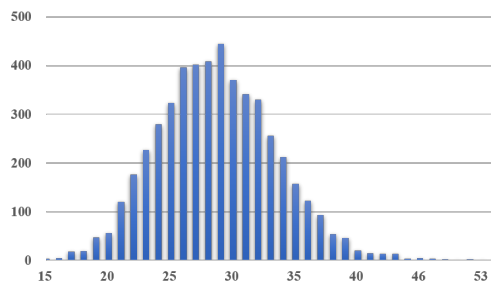


Figure 2: Distribution of images in terms of the number of distinct words used for their descriptions. X-axis: the number of distinct words in all corresponding ground truth captions v.s. Y-axis: the number of instances in MS-COCO.

current Neural Network (RNN) for caption generation (Vinyals et al., 2015). Although impressive results have been achieved, existing models suffer from the problem of generating N-grams which occurred frequently in the training set but are irrelevant to the particular given image (Anderson et al., 2016; Dai et al., 2017).

Examples of caption generation are shown in Figure 1. Two images are presented with model-generated captions (Vinyals et al., 2015) and their

\*Corresponding author

<sup>1</sup><https://github.com/LibertFan/ImageCaption>

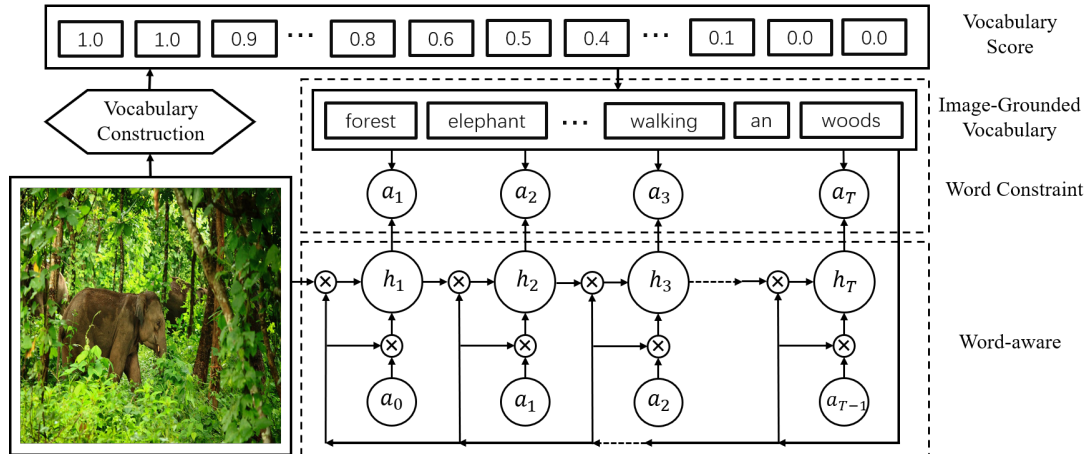


Figure 3: Overall framework of our proposed model.

corresponding human-constructed ones. As we can see, the N-gram “a woman sitting at a table” is generated mistakenly for both images. This is because when generating a text sequence, the RNN-based generator tends to ignore the semantic meaning encoded in the given image and instead generate the text sequences that occurred most often in the training set. Although different image grounding strategies have been proposed to address this problem, they usually consider visual information as external features for caption generation via various attention mechanisms (Xu et al., 2015; You et al., 2016; Lu et al., 2017). We argue that visual information should be embedded into the generation process in a more principled way.

In the CNN-RNN architecture, the RNN-based generator constructs image captions word by word. In each step, a word is selected from the vocabulary built on the entire training set. Generally, the size of the full vocabulary is on the order of  $10^4$ . When describing a particular image, the possible words to be used should be drawn from a much smaller word set. As an illustration, we show in Figure 2 the statistics of the number of distinct words in human-generated captions for images from MS-COCO (Lin et al., 2014). We can see that the average size of the pool of words used for the description of a particular image is around 30. Based on this observation, we speculate that if we can efficiently constrain the word selection space during the image caption generation process, we should be able to address the irrelevant N-gram problem.

In this paper, we propose to construct an image-grounded vocabulary as a way to leverage the im-

age semantics for image captioning. For vocabulary construction, we propose a two-step approach which incorporates both visual semantics and the relations among words. For text generation, we explore two strategies to utilize the constructed vocabulary. One uses the vocabulary as a hard constraint and the other encodes the weight of each word obtained from the image-grounded vocabulary into the RNN cell as a soft constraint. Experimental results on two public datasets show the effectiveness of using image-grounded vocabulary for visual captioning compared to several state-of-the-art approaches in terms of automatic evaluation metrics. Further analysis reveals that our model has the advantage of generating more novel captions compared to existing approaches.

## 2 Our Approach

The overall architecture of our model is shown in Figure 3, which consists of two main stages, *image-grounded vocabulary construction* and *text generation with vocabulary constraints*. The *image-grounded vocabulary constructor* builds a vocabulary related to a given image by considering the visual information encoded and the relationships among words. The *text generator with vocabulary constraints* generates captions using the constructed vocabulary in two different ways. First, words generated are strictly limited to those in the image-grounded vocabulary. Second, words in the image-grounded vocabulary are re-weighted within the RNN cell such that they are more likely to be generated. We also study the use of the image-grounded vocabulary under the framework of reinforcement learning treating the image-

grounded vocabulary as the action space for caption generation.

## 2.1 Image-Grounded Vocabulary Construction

The *image-grounded vocabulary constructor* aims to identify words required for the description of a given image  $I_i$ . Intuitively, words used to describe an image can be divided into two groups. One group of words are directly related to the image (e.g., entities or objects depicted in the image) and the other group of words are function words or words that do not correspond directly to elements of the image. We assume that the directly-related words can be determined based on the visual information, while the identification of words in the second group requires the consideration of their relationship with those in the first group. Therefore, we propose a two-step strategy to construct the image-grounded vocabulary.

In the first step, we identify words that are directly related to a given image. Taking each word as a label, the construction of the image-grounded vocabulary can be treated as a multi-label classification problem. We take the visual features of the image as input and obtain a probability distribution  $S_i$  for words, indicating the relevance of words for image  $I_i$ . Following Fang et al. (2015), we only consider a list of words with high frequency in the dataset as seeds, denoted as  $H$ . The relevance distribution of words in  $H$  for an image  $I_i$  is computed as follows:

$$S_i^{(H)} = \sigma(M_1(v_i)) \quad (1)$$

where  $v_i$  is the visual features of image  $I_i$  and  $M_k$  is a multi-layer perceptron (MLP) with  $k$  layers (one layer in this case),  $\sigma(\cdot)$  denotes a sigmoid function,  $S_i^{(H)}$  stands for the relevance of words in  $H$  for image  $I_i$  and  $S_i^{(H)}$  is in the same size as  $H$ .

In the second step, we compute the relevance scores of words in the full vocabulary  $V$  given the image  $I_i$  and the probabilities of directly-related words  $S_i^{(H)}$ . Specifically, a 2-layer MLP with sigmoid function is employed. The probability distribution of words in  $V$  considering both visual information and relations among words is computed in Equation 2:

$$S_i^{(V)} = \sigma(M_2([S_i^{(H)}, v_i])) \quad (2)$$

where  $[\cdot, \cdot]$  is the concatenation operation. During inference, we pick the top  $k$  words in terms of their relevance scores to form the image-grounded vocabulary for image  $I_i$ , denoted as  $W_i$ . Note that  $S_i^{(V)}$  stands for the relevance score of words in  $V$  for image  $I_i$  and  $S_i^{(V)}$  is in the same size as  $V$ .

## 2.2 Text Generation with Vocabulary Constraints

In order to utilize the image-grounded vocabulary  $W_i$  and word relevance distribution  $S_i^{(V)}$  for caption generation, we explore two different strategies. One uses  $W_i$  as a hard constraint and the other integrates the relevance of each word into the RNN cell for caption generation. In what follows, we first introduce the basic RNN-based text generator, and then describe each of the two strategies in turn.

### 2.2.1 RNN-based Generator

RNN-based generator takes the visual features as input, and generates an image caption word by word. In each step, an RNN cell takes the hidden state  $h_{t-1}$  and the output word  $a_{t-1}$  from the previous step as input and computes the hidden state  $h_t$  for the current step. Based on  $h_t$ , a *softmax* layer is used to compute the probability distribution of words in the vocabulary and the top one is selected as the output. The computation process is described in Equation 3:

$$P(w_j|I_i, a_1, \dots, a_{t-1}) = \text{softmax}(M_1(h_t))_j \\ a_t = \underset{w_j}{\text{argmax}} P(w_j|I_i, a_1, \dots, a_{t-1}), w_j \in V \quad (3)$$

In our case, we use an LSTM (Gers et al., 1999) as the RNN cell. Suppose the hidden state, the cell state and the output in the  $(t-1)_{th}$  step are denoted as  $h_{t-1}$ ,  $c_{t-1}$  and  $a_{t-1}$ , respectively, the states and output at the  $t_{th}$  step can be computed as:

$$\begin{aligned} i_t &= \sigma(W_{ia}a_{t-1} + U_{ih}h_{t-1}) \\ f_t &= \sigma(W_{fa}a_{t-1} + U_{fh}h_{t-1}) \\ o_t &= \sigma(W_{oa}a_{t-1} + U_{oh}h_{t-1}) \\ \tilde{c}_t &= \tanh(W_{ca}a_{t-1} + U_{ch}h_{t-1}) \\ c_t &= i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad (4)$$

where  $\odot$  denotes the element-wise multiplication, and  $W_{*a}, U_{*a}, * \in \{i, f, o, c\}$  are the parameters of the LSTM cell.

### 2.2.2 Generator with Hard Constraint

A straightforward way of utilizing the image-grounded vocabulary for text generation is to limit the decoding space to  $W_i$  (refer to *word constraint* in Figure 3). The word selection in each step within the RNN cell can thus be modified as follows:

$$a_t = \underset{w_j}{\operatorname{argmax}} P(w_j | I_i, a_1, \dots, a_{t-1}), w_j \in W_i \quad (5)$$

In practice, a mask operation  $m_i$  is introduced to replace the  $j_{th}$  value in the vector with  $-\infty$  if  $w_j$  is not found in  $W_i$  as shown in Equation 6.

$$m_i(\cdot)_j = -\infty, \forall w_j \notin W_i \quad (6)$$

### 2.2.3 Generator with Soft Constraint

Instead of using the image-grounded vocabulary as the hard constraint, we further explore to integrate the probability distribution  $S_i^{(V)}$  of words in vocabulary  $V$  for the given image  $I_i$  into the decoding RNN cell (refer to *word-aware* in Figure 3). In the  $t_{th}$  step, we simply combine  $a_t$ ,  $h_t$  and  $S_i^{(V)}$  with the element-wise multiplication. The computation steps in the cell are shown below:

$$\begin{aligned} i_t &= \sigma(W_{is}S_i^{(V)} \odot W_{ia}a_{t-1} + U_{is}S_i^{(V)} \odot U_{ih}h_{t-1}) \\ f_t &= \sigma(W_{fs}S_i^{(V)} \odot W_{fa}a_{t-1} + U_{fs}S_i^{(V)} \odot U_{fh}h_{t-1}) \\ o_t &= \sigma(W_{os}S_i^{(V)} \odot W_{oa}a_{t-1} + U_{os}S_i^{(V)} \odot U_{oh}h_{t-1}) \\ \tilde{c}_t &= \sigma(W_{cs}S_i^{(V)} \odot W_{ca}a_{t-1} + U_{cs}S_i^{(V)} \odot U_{ch}h_{t-1}) \\ c_t &= i_t \odot \tilde{c}_t + f_t \odot c_{t-1} \\ h_t &= o_t \odot \tanh(c_t) \end{aligned} \quad (7)$$

where  $W_{*s}, W_{*a}, U_{*s}, U_{*a}, * \in \{i, f, o, c\}$  are the parameters of the cell.

The new RNN cell integrates information about the image-grounded vocabulary so that words in that vocabulary are more likely to be generated.

## 2.3 Reinforcement Learning for Text Generation

Although it is straightforward to impose the hard vocabulary constraint during inference, it is not easy to train the text generator with the hard constraint since words in the ground-truth caption may not appear in the image-grounded vocabulary  $W_i$  constructed for image  $I_i$ . We denote such words as:

$$a_t \in \widetilde{W}_i \setminus W_i \quad (8)$$

where  $\widetilde{W}_i$  is the ground-truth vocabulary for image  $I_i$ . In order to tackle this problem, we employ reinforcement learning to train the generator under the vocabulary constraint so that it is less likely to select words not in  $W_i$ . This strategy not only aligns the behavior of word selection during training and testing, but also makes the generator better accustomed to the distribution of  $W_i$  through the feedback reward.

Recall the goal of reinforcement learning is to maximize the expected reward of the generator with parameter  $\theta$ :

$$L_\theta = \mathbb{E}_{(a_1, \dots, a_T) \sim p_\theta} [r(a_1, \dots, a_T)] \quad (9)$$

The policy gradient of Equation 9 with a baseline is shown in Equation 10.

$$\nabla_\theta L_\theta \approx (r(a_1, \dots, a_T) - b) \nabla_\theta \log p_\theta(a_1, \dots, a_T) \quad (10)$$

Following Rennie et al. (2017), we utilize CIDEr-D (Vedantam et al., 2015) as the reward of the generated sentence  $(a_1, \dots, a_T)$  and set  $b = r(\hat{a}_1, \dots, \hat{a}_T)$  which is the reward obtained by the current model with greedy decoding.

In summary, the training strategy with reinforcement learning under the vocabulary constraint can be described as follows:

---

**Algorithm 1** Caption generation with reinforcement learning

---

- 1: for  $t = 1 : T$  do
  - 2:  $p_t(w_j) = \operatorname{softmax} \left( m_i(M_1(h_t)) \right)_j$
  - 3:  $\hat{a}_t = \operatorname{argmax}_{w_j} p_t(w_j), a_t \sim p_t(w_j)$
  - 4:  $b = r(\hat{a}_1, \dots, \hat{a}_T)$
  - 5:  $L = (r(a_1, \dots, a_T) - b) \log p(a_1, \dots, a_T)$
- 

## 2.4 Training

The overall training procedure of our proposed framework can be described by the following four steps:

---

**Algorithm 2** Training procedure

---

- 1: Train the vocabulary constructor to build the image-grounded vocabulary  $W_i$ .
  - 2: Train the generator  $G_\theta$  with cross-entropy loss under the soft constraints.
  - 3: Train the generator  $G_\theta$  with reinforcement learning according to Equation 3.
  - 4: Train the generator  $G_\theta$  with reinforcement learning under the vocabulary constraints according to Algorithm 1.
-

Model	MS COCO				Flickr30k			
	B-4	R	M	C	B-4	R	M	C
<i>ATT</i> (You et al., 2016)	30.4	-	24.3	-	23.0	-	18.9	-
<i>AdapAtt</i> (Lu et al., 2017)	31.2	53.0	25.0	97.0	23.3	45.5	19.3	48.2
<i>TopDown</i> (Anderson et al., 2018)	32.4	53.8	25.7	101.1	23.7	45.6	19.7	49.8
<i>NIC</i> (Vinyals et al., 2015)	28.6	55.7	25.0	89.2	20.3	48.3	19.1	42.0
<i>NIC+RL</i> (Rennie et al., 2017)	31.5	57.6	25.6	101.4	21.4	49.2	19.6	48.2
<i>NIC+WC</i>	29.3	56.3	25.4	93.1	20.4	48.6	19.8	46.2
<i>NIC+WC+WA</i>	31.5	57.6	26.0	97.6	22.2	50.3	20.4	51.5
<i>NIC+WC+RL</i>	32.2	58.1	26.0	103.7	22.3	50.3	20.4	52.1
<i>NIC+WC+WA+RL</i>	<b>33.0</b>	<b>58.6</b>	<b>26.4</b>	<b>106.6</b>	<b>24.5</b>	<b>51.6</b>	<b>21.5</b>	<b>58.4</b>
<i>NIC+WC(GT)</i>	50.7	67.6	32.7	142.8	37.9	60.2	26.2	86.3

Table 1: Overall performance of different models for image captioning, where B-4, R, M and C are short for BLEU-4, ROUGE, METEOR and CIDEr-D scores, respectively. Numbers in **bold** denote the best performance in each column.

### 3 Experiment

#### 3.1 Dataset

We evaluate our proposed framework on MS-COCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015). In MS-COCO, there are 113,287 images in the training set and 5,000 images in both of the validation and test sets. In Flickr30k, the number of images for the training, validation and test sets is 29,000, 1,000 and 1,000, respectively. Each image contains 5 human annotated captions. We split the dataset following the process described in (Karpathy and Fei-Fei, 2015).

#### 3.2 Implementation Details

For image representation, we rescale the image to  $224 \times 224$  and use ResNet-152 (He et al., 2016) pre-trained on ImageNet (Russakovsky et al., 2015) to extract features of dimension 2,048. The mini-batch size is 64. The dimensions of LSTM hidden unit and the word embedding are 512 and 300, respectively, and the word embedding is initialized with GloVe (Pennington et al., 2014)<sup>2</sup> which is pretrained on Wikipedia 2014 and Gigaword 5. We prune the vocabulary by dropping words appear less than five times. For the generator, We train the model with cross-entropy using Adam (Kingma and Ba, 2014) with an initial learning rate  $1 \times 10^{-3}$  which decreases by a factor of 0.8 every  $2 \times 10^4$  iterations. Then we train the generator with reinforcement learning but without

<sup>2</sup><http://nlp.stanford.edu/data/glove.6B.zip>

hard constraints using Adam with an initial learning rate  $5 \times 10^{-5}$  which decreases by a factor of 0.8 every  $3 \times 10^4$  iterations. Finally, we train the generator with reinforcement learning under the hard constraints using Adam with an initial learning rate  $5 \times 10^{-5}$  which decay at a rate of 0.8 every  $2 \times 10^4$  iterations. For each model, we evaluate on the validation set to select the best parameters with grid search. We set the size of  $W_i$  to 64 for all models with hard constraints.

#### 3.3 Models for Comparison

We compare our model with the state-of-the-art approaches listed below. In addition, we also performed ablation studies of our proposed model. We denote the hard word constraint mechanism, soft word-aware mechanism and reinforcement learning as *WC*, *WA* and *RL*, respectively.

- *NIC* (Vinyals et al., 2015) is the baseline CNN-RNN model trained with cross-entropy loss. *NIC+RL* is trained with reinforcement learning.
- *ATT* (You et al., 2016) detects a list of visual concepts from a given image, which is used to guide the caption generation process through an attention mechanism.
- *AdapAtt* (Lu et al., 2017) utilizes the context information of RNN cells in the decoder to better predict non-visual words.
- *TopDown* (Anderson et al., 2018) employs the visual attention mechanism with the two-layer LSTM.

- *NIC+WC* uses the hard word constraints (WC). *NIC+WC+RL* is trained with reinforcement learning using the image-grounded vocabulary as the action space.
- *NIC+WC+WA* employs the soft word-aware (WA) mechanism on top of *NIC+WC*. *NIC+WC+WA+RL* is trained with reinforcement learning using the image-grounded vocabulary as the action space.
- *NIC+WC(GT)* utilizes the ground-truth vocabulary  $\tilde{W}_i$  as the word constraints instead of  $W_i$ . This is an oracle.

### 3.4 Overall Performance

We report scores of several widely used metrics for image captioning evaluation, including BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin and Hovy, 2003) and CIDEr-D (Vedantam et al., 2015). The overall performance is shown in Table 1. Several findings stand out:

- Both *NIC+WC* and *NIC+WC+RL* perform better than their counter-part models *NIC* and *NIC+RL* across all metrics. This shows the effectiveness of using the word constraint mechanism for reducing irrelevant words for a given image.
- Both *NIC+WC+WA* and *NIC+WC+WA+RL* outperform *NIC+WC* and *NIC+WC+RL* respectively. This shows that the word-aware mechanism effectively guides the generator to better capturing the semantics of a given image.
- Compared to *NIC+WC* and *NIC+WC+WA*, both *NIC+WC+RL* and *NIC+WA+WC+RL* achieve better performance. This demonstrates that training the generator under the word constraints with reinforcement learning encourages the generator to adhere to the constraints set by the image-grounded vocabulary.
- Our proposed model *NIC+WA+WC+RL* outperforms all the baselines and its variants. However we notice that there is still a large gap between our proposed model and the oracle model *NIC+WC(GT)*. This shows that there is still potential to improve the process for the construction of the image-grounded vocabulary.

We conducted statistical significance tests (Students paired t-test) to verify that the differences seen among the different approaches were

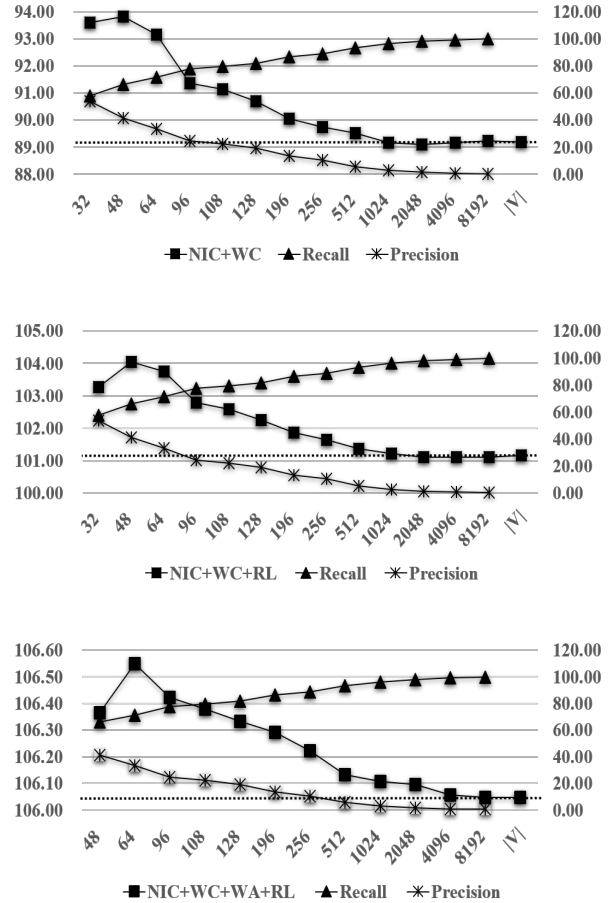


Figure 4: Performance of different models on MSCOCO in terms of CIDEr-D with various sizes of  $W_i$ . X-axis: size of  $W_i$ , Y-axis: score of CIDEr-D (left), precision and recall of  $W_i$  compared to  $\tilde{W}_i$ . The dotted lines are the CIDEr-D scores for models with  $W_i = V$ .

statistically significant. Results showed that *NIC+WA+WC+RL* outperformed *NIC+RL* significantly across all the metrics ( $p < 0.01$ ). Similarly for *NIC+WA+WC* and *NIC* ( $p < 0.01$ ). This confirms the effectiveness of using image-grounded vocabulary to improve visual captioning.

### 3.5 Further Analysis

Further analysis was conducted to evaluate the sensitivity of our model with respect to parameter and component setting and present case studies to illustrate the merits of our model in comparison to baseline models.

**Influence of the size of  $W_i$**  We explore the influence of the size of the image-grounded vocabulary on the performance of the generator. We test three models, namely, *NIC+WC*, *NIC+WC+RL* and *NIC+WC+WA+RL* using various sizes of  $W_i$ ,

and report the CIDEr-D scores. In addition, we also report the recall and precision of  $W_i$  compared to the ground truth  $\widetilde{W}_i$ . The results are shown in Figure 4. Note that models without word constraints can be interpreted as taking  $|V|$  as  $W_i$ .

We observe a similar trend of CIDEr-D for all the three models. It gradually goes up with the increasing size of  $W_i$ , reaches the peak at 48, 48 and 64, respectively, and then gradually drops with the further increase of the size of  $W_i$ . It is worth noting that the peak numbers are quite close to the average number of words (i.e., around 30) in  $\widetilde{W}_i$  shown in Figure 2. The performance of the generator is poor when the size of  $W_i$  is too small because the possible word choices are too limited. As the size of  $W_i$  gets larger, more irrelevant words are included, which introduces noise to the generator and thus performance drops.

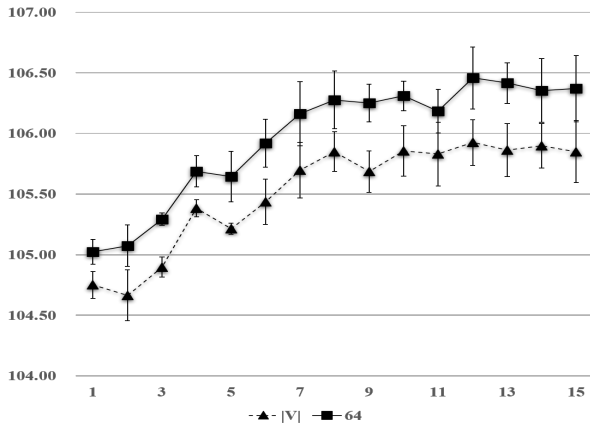


Figure 5: Mean CIDEr-D scores with standard derivation of  $|W_i| = 64$  versus  $|W_i| = |V|$ , for  $NIC+WC+WA+RL$  (3 seeds) on the validation set. X-axis: the number of training iterations ( $2 \times 10^4$ ), Y-axis: CIDEr-D scores.

**Robustness of our model** In Figure 5, we show the mean and standard deviation of CIDEr-D scores of  $NIC+WC+WA+RL$  with  $W_i = 64$  and  $W_i = |V|$  in three different runs for training the generator with reinforcement learning under word constraints. We can see that the model with  $W_i = 64$  consistently outperforms the one with  $W_i = |V|$  across the training iterations.

**Influence of vocabulary constructor** We analyze the influence of the vocabulary constructor on the performance of the generator. Instead of using the vocabulary constructor introduced in section 2.1, we build another baseline model that takes visual features as input and employs a sin-

Model	R@64	P@64	B-4	C
$NIC+WC_b$	69.6	32.7	29.0	92.4
$NIC+WC$	71.2	33.4	29.3	93.1
$NIC+WC_b+RL$	69.6	32.7	31.7	102.3
$NIC+WC+RL$	71.2	33.4	32.2	103.7

Table 2: Performance of different models with various vocabulary constructors on MS COCO.  $R$ ,  $P$ ,  $B-4$  and  $C$  are short for recall, precision, BLEU-4 and CIDEr-D respectively.

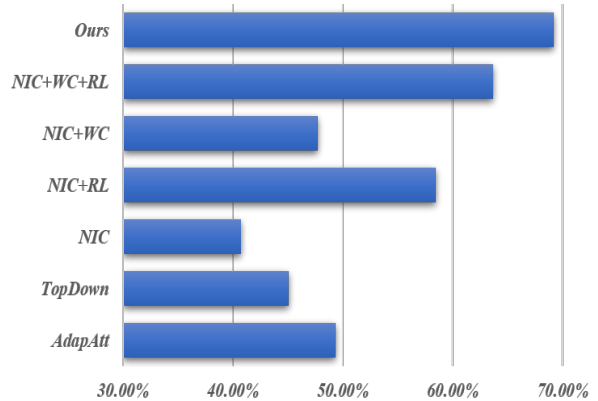


Figure 6: X-axis: novel caption ratio on MS-COCO v.s. Y-axis: different models. Novel captions are those generated during testing that do not appear in the training set.

gle layer MLP with sigmoid for generating the vocabulary. The variants of the models are named as  $NIC+WC_b$  and  $NIC+WC_b+RL$ . Experimental results are shown in Table 2. We report precision and recall of the generated vocabulary to evaluate the constructor directly and report BLEU-4 and CIDEr-D to see their influence on the generator.

It can be observed that our constructor is able to build a better vocabulary compared to the baseline constructor in terms of both precision and recall. This indicates the effectiveness of our two-step approach. Moreover, with our proposed vocabulary constructor, both  $NIC+WC$  and  $NIC+WC+RL$  outperform  $NIC+WC_b$  and  $NIC+WC_b+RL$  respectively, achieving better BLEU-4 and CIDEr-D scores in image captioning.

### Effectiveness of generating novel captions

Novel caption generation is crucial for automatic image captioning because retrieval-based models that simply retrieve existing captions from the training set often produce less human results though they can achieve high scores in terms of au-

omatic evaluation metrics (Devlin et al., 2015b). The worst case of N-gram problem is that the model directly generated the same frequent captions in the training set (Devlin et al., 2015a). Thus the capability of generating novel captions for an image that is not seen in the training set indicates that the generator is able to understand a given image better instead of simply generating frequent N-grams found in the training set.

In this experiment, we consider captions generated by models that are not seen in the training set as novel captions. We show the ratio of novel captions generated by different models in Figure 6. Our proposed model outperforms *NIC* and other two competitive baselines, *TopDown* and *Adap-Att*, by a large margin. Moreover, *NIC+WC* is also able to generate more novel captions compared to *NIC*, indicating that the word constraint mechanism helps reducing generic words.

**Case Study** We show example captions generated by our model in Figure 7. Results from two models are presented, namely *NIC+RL* and *NIC+WA+WC+RL*. In order to show how the image-grounded vocabulary  $W_i$  regulate the generation process, we cross those words in the caption generated by *NIC+RL* but not included in  $W_i$ . The crossed words are entity words such as “grass” and “field” in the first image (up-left), preposition “on” in the second one (up-right) and entity word “bench” in the third one (bottom-left). Examples also indicate the effectiveness of the word-aware mechanism that guides the generator to replace “standing” with “walking” in the first image, “people” with “children” in the third one, “standing” with “is flying over” in the last one (bottom-right).

## 4 Related Work

Research investigation the connection between vision and language has attracted increasing attentions in the past a few years. Popular tasks include image captioning, visual question answering (VQA), and visual question generation. In image captioning, most of the proposed models (Xu et al., 2015; You et al., 2016; Lu et al., 2017; Anderson et al., 2018) employ CNN to extract visual features and RNN to generate captions word by word (2015). Visual question answering (Antol et al., 2015; Goyal et al., 2017) aims to provide an answer to a question related to a given image. Existing architectures designed for VQA (Mali-

nowski et al., 2015) utilize an RNN to encode the question, and a CNN to encode the image. Most efforts are made to align the visual and text information for generating the answer. Visual question generation is a relatively new task that generates natural questions about an image (Mostafazadeh et al., 2016). Approaches have been explored to generate diverse questions (Tang et al., 2017; Zhang et al., 2017; Fan et al., 2018a) and questions with a specific property (Fan et al., 2018b).

Instead of using high-level visual features extracted from the image for text generation, some researchers explore identifying fine-grained information from the image, i.e. objects and attributes, to guide the process of text generation. Traditionally, template-based approaches are used to compose the caption (Farhadi et al., 2010; Kulkarni et al., 2013; Lin et al., 2015). After that, different attention mechanisms are proposed to align visual information and text for better generation (You et al., 2016; Lu et al., 2017; Anderson et al., 2018).

For better aligning visual information and text, some researchers explore identifying semantic concepts related to the image. Jia et al. (2015) employs retrieved sentences as additional semantic information to assist generation. Others (Fang et al., 2015; Wu et al., 2016; You et al., 2016; Gan et al., 2017) utilize high-frequency words as semantic concepts. Fang et al. (2015) develops features based on detected concepts to re-rank the generated captions. You et al. (2016) employs an attention mechanism over concepts to enhance the generator. Gan et al. (2017) applies weight tensors in LSTM units to integrate the semantic concept into the generator. Instead, in our proposed approach, image-grounded vocabulary is built at the word level and imposed as constraints on caption generation.

The work most relevant to ours is from Yao et al. (2017) and Wu et al. (2018). Yao et al. (2017) incorporates a copy mechanism to encourage the generator to generate visually related words. Wu et al. (2018) dynamically construct a vocabulary with a lightweight network and then picks one from this smaller vocabulary with a more complex network to improve computational efficiency. Our model is novel in three ways. First, we observe that the large mismatch between the dataset vocabulary and the vocabulary required for describing an image is one of the main reasons for the generation of irrelevant N-grams. Second, we propose a







		forest: 0.969 elephant: 0.705 walking: 0.653 an: 0.567 two: 0.539 wood: 0.334		sign: 0.966 street: 0.900 signs: 0.891 front: 0.291
<i>NIC+RL</i>	two elephants are standing in the <del>grass</del> in a <del>field</del> .		a <del>group</del> of street signs <del>on</del> a stop sign.	
<i>NIC+WA+WC+RL</i>	an elephant is walking in the woods in a forest.		a stop sign in front of a street.	
		sitting: 0.589 child: 0.281 people: 0.271 kids: 0.231 children: 0.165 field: 0.160		bird: 0.464 flying: 0.415 field: 0.719 grass: 0.565 over: 0.221
<i>NIC+RL</i>	a group of people sitting on a <del>bench</del> .		a bird standing in the grass in a field.	
<i>NIC+WA+WC+RL</i>	a group of children sitting on a field.		a bird is flying over a field in the grass.	

Figure 7: Examples of generated captions and some corresponding words in image-grounded vocabulary. Words that are crossed out are not in the image-grounded vocabulary.

novel two-step approach for image-grounded vocabulary construction. Third, we explore two different strategies for caption generation using the constructed vocabulary.

## 5 Conclusion and Future Work

In this paper, we have proposed a novel framework which constructs an image-grounded vocabulary to leverage the image semantics for image captioning in order to tackle the problem of generating irrelevant N-grams. A novel two-step approach has been proposed to construct the vocabulary considering both visual information and relations among words. Two strategies have then been explored to utilize the constructed vocabulary via hard constraints and soft constraints. Reinforcement learning has been adopted for the training of the generator to encourage it to only choose words from the image-grounded vocabulary. Experiments on two public datasets, namely, MS COCO and Flickr30k, show that image-grounded vocabulary is able to enhance the quality of image captions compared to existing state-of-the-art approaches. In future, we plan to study more effective ways to construct the image-grounded vocabulary. Furthermore, it is also interesting to design a mutual reinforcement mechanisms between the vocabulary constructor and the text generator

to improve both components simultaneously.

## Acknowledgments

This work is partially supported by National Natural Science Foundation of China (No. 61751201, No. 61702106) and Shanghai Science and Technology Commission (No. 17JC1420200, No. 17YF1427600 and No. 16JC1420401).

We thank Yulan He and Michael Johnston for revising the paper carefully to make this camera-ready version. We also thank three anonymous reviewers for their insightful suggestions.

## References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. SPICE: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer.
- Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. 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*, pages 6077–6086.
- Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick,

- and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.
- Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72.
- Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979.
- Jacob Devlin, Hao Cheng, Hao Fang, Saurabh Gupta, Li Deng, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. 2015a. Language models for image captioning: The quirks and what works. *arXiv preprint arXiv:1505.01809*.
- Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. 2015b. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*.
- Zhihao Fan, Zhongyu Wei, Piji Li, Yanyan Lan, and Xuanjing Huang. 2018a. A question type driven framework to diversify visual question generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4048–4054. International Joint Conferences on Artificial Intelligence Organization.
- Zhihao Fan, Zhongyu Wei, Siyuan Wang, Yang Liu, and Xuanjing Huang. 2018b. A reinforcement learning framework for natural question generation using bi-discriminators. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1763–1774.
- Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. 2015. From captions to visual concepts and back. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1473–1482.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer.
- Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. 2017. Semantic compositional networks for visual captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5630–5639.
- Felix A Gers, Jürgen Schmidhuber, and Fred Cummins. 1999. Learning to forget: Continual prediction with lstm.
- Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.
- Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding the long-short term memory model for image caption generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2407–2415.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903.
- Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using N-gram co-occurrence statistics. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 71–78. Association for Computational Linguistics.
- Dahua Lin, Chen Kong, Sanja Fidler, and Raquel Urtasun. 2015. Generating multi-sentence linguistic descriptions of indoor scenes. *arXiv preprint arXiv:1503.00064*.
- Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.
- Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 375–383.
- Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–9.

- Nasrin Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. 2016. Generating natural questions about an image. *arXiv preprint arXiv:1603.06059*.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics.
- Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.
- Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2641–2649.
- Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024.
- Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. 2015. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252.
- Duyu Tang, Nan Duan, Tao Qin, Zhao Yan, and Ming Zhou. 2017. Question answering and question generation as dual tasks. *arXiv preprint arXiv:1706.02027*.
- Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.
- Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony Dick, and Anton Van Den Hengel. 2016. What value do explicit high level concepts have in vision to language problems? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 203–212.
- Yu Wu, Wei Wu, Dejian Yang, Can Xu, and Zhoujun Li. 2018. Neural response generation with dynamic vocabularies. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057.
- Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2017. Incorporating copying mechanism in image captioning for learning novel objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6580–6588.
- Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4651–4659.
- Shijie Zhang, Lizhen Qu, Shaodi You, Zhenglu Yang, and Jiawan Zhang. 2017. Automatic generation of grounded visual questions. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 4235–4243.