Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 9743123, 9 pages
https://doi.org/10.1155/2022/9743123

Research Article

Context-Fused Guidance for Image Captioning Using Sequence-Level Training

Junlong Feng and Jianping Zhao

School of Computer Science and Technology, Changchun University of Science and Technology, Changchun, China

Correspondence should be addressed to Jianping Zhao; zjp@cust.edu.cn

Received 24 November 2021; Accepted 13 December 2021; Published 5 January 2022

Academic Editor: Ahmed Mostafa Khalil

Recent image captioning models based on the encoder-decoder framework have achieved remarkable success in humanlike sentence generation. However, an explicit separation between encoder and decoder brings out a disconnection between the image and sentence. It usually leads to a rough image description: the generated caption only contains main instances but neglects additional objects and scenes unexpectedly, which reduces the caption consistency of the image. To address this issue, we proposed an image captioning system within context-fused guidance in this paper. It incorporates regional and global image representation as the compositional visual features to learn the objects and attributes in images. To integrate image-level semantic information, the visual concept is employed. To avoid misleading decoding, a context fusion gate is introduced to calculate the textual context by selectively aggregating the information of visual concept and word embedding. Subsequently, the context-fused image guidance is formulated based on the compositional visual features and textual context. It provides the decoder with informative semantic knowledge. Finally, a captioner with a two-layer LSTM architecture is constructed to generate captions. Moreover, to overcome the exposure bias, we train the proposed model through sequence decision-making. The experiments conducted on the MS COCO dataset show the outstanding performance of our work. The linguistic analysis demonstrates that our model improves the caption consistency of the image.

1. Introduction

Image captioning, which analyses and converts the image content into a natural language description automatically, is drawing considerable attention in the artificial intelligence field. As a typical multimodal task, the image captioning system combines both computer vision and natural language processing. Therefore, it should not only recognize the salient image objects and other visual properties (attributes, locations, and relations) but also depict the image content with natural and coherent descriptions [1]. Over the past few years, image captioning task has been applied on a wide area of aspects, such as assistance for visually impaired people [2].

For current image captioning system, the encoder-decoder architecture has been a widely adopted pipeline for its conspicuous performance. In general, it employs a convolutional neural network (CNN) to encode the image into a set of feature vectors and a long short-term memory (LSTM) network to generate the captions. Moreover, to steer the model into focusing and capturing informative visual features on a particular image region, the attention mechanisms are introduced as well [3–5].

The encoder-to-decoder framework has achieved remarkable advances in humanlike caption generating, but there are still some issues to be concerned.

First, to capture the visual and textual information simultaneously, some prior networks [3, 4] were designed to learn the sentence structure at a global level. Strictly, the generated caption can only depict the image roughly because during decoding, the network may discard some useful image objects or scenes unexpectedly. This reduces the consistency between image and text description. As a solution, the guidance vector is adopted [6–8]. In [6], the time-independent guidance was implemented as a joint text-image embedding. However, as pointed out in [7], their approach is short of consideration from two aspects: (1) from the view of computer vision, visual evidence is not always essential for the decoder because the description sentence usually contains salient objects that correspond to visual features; (2) the explicit separation between encoder and decoder usually leads to a representational disconnect between the learned feature vectors and generated captions. To handle these issues, they constructed a semantic image guidance, which is conditioned on textual context and image features. It provides the decoder with semantic information from n-gram word and sentence levels. Through this, the generated captions include richer image instances than [6]. Nevertheless, their approach neglects the information about motions and locations of image objects. In addition, although the sentence-level guidance achieved the best performance, it is not a very efficient approach because of the prepositions, articles, and conjunctions in the sentence. Considering the fact that the instances in region image are not always corresponding to the words in the vocabulary, in [8], they concatenated the global image representation with the visual concept [9] as the guidance vector. The visual concept is a set of frequent words that describe the salient image objects, which enhances the correlation between image and text at regional level. However, there is a latent drawback: an inappropriate word in visual concept will mislead the language model to generate unexpected captions.

Second, as indicated in [10], for the models trained with maximum likelihood estimation (MLE), the vanilla encoder-decoder framework may cause the problem of exposure bias. The error accumulation caused by MLE probably results in a word mismatching during caption generating. To address this issue, the reinforcement learning (RL) strategy is introduced in the image captioning task. However, due to the high variance of gradient estimation, it is extremely difficult to train the model with RL strategy directly. To meet this criterion, the self-critical sequence training (SCST) framework [11] is proposed to apply the RL strategy by sequence-level training. During the inference stage, SCST utilizes the generating samples as the baseline to normalize the rewards. Consequently, the network can use nondifferentiable sequence-level metrics (e.g., CIDEr [12]) to evaluate the language quality rather than the cross-entropy loss in word level. Based on this framework, a number of approaches were proposed [13–15]. Particularly, in [14], they proposed the CAVP to accomplish the visual decision-making task. The CAVP captures the visual context that is crucial for compositional reasoning and attends to complex visual compositions over time. Through this, it significantly boosted the caption consistency to image content.

Therefore, to boost the caption consistency of image by utilizing reasonable semantic information and informative visual features, an image captioning system within context-fused guidance (CFG) is proposed in this paper. The main idea is illustrated in Figure 1. The CFG utilizes compositional visual features for multilevel image learning.

By the context fusion gate, CFG adaptively combines the visual concept and word embedding. Using the context-fused image guidance, our model can generate captions with comprehensive descriptions. In short, the main contributions in this paper are as follows:

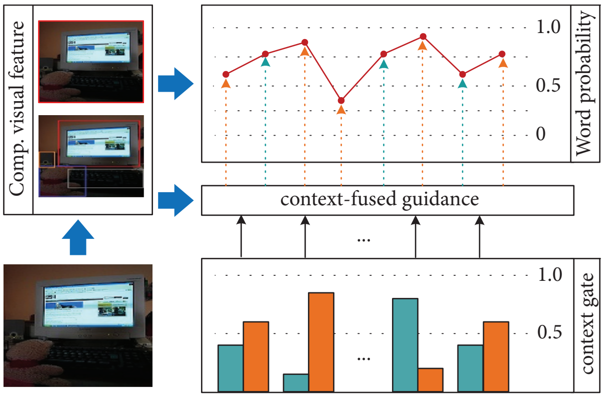


Figure 1: The main idea of our proposed network. The compositional visual feature consists of the image representation at regional and global level. At each decoding step, the context gate calculates the textual context by dynamically aggregating the visual concept and word embedding. The context-fused image guidance is formulated on the compositional visual features and fused textual context.

(1) An image captioning system using sequential decision-making is proposed for a comprehensive caption generation.

(2) A context-fused image guidance is formulated to improve the caption consistency of image. It selectively aggregates the semantic information from the visual concept and word embedding.

(3) Evaluation on the MS COCO dataset shows that our approach outperforms most standard metrics. The linguistic analysis demonstrates that our method enhances the correlation of generated captions and images.

2. Related Works

2.1. Image Captioning. In the past few years, image captioning systems based on encoder-decoder framework have been deeply investigated [3, 16]. In [16], they employed a CNN to encode the image and a recurrent neural network to output a sequence of words. Subsequently, many works were proposed to improve and extend this framework. In [17], they proposed a recurrent fusion network (RFNet) to exploit the complementary information from multiple encoders to understand the image comprehensively. In [18], they extracted the image features at multiple levels to learn accurate subject predictions. As a very recent investigation [19], the editing network generates the image description by refining an existing caption rather than generating a new caption from scratch.

Inspired by the attention mechanism applied in machine translation, several attention-based image captioning systems were proposed. In [3], they integrated the decoder with the proposed hard and soft attention mechanism to capture the highlighting spatial image

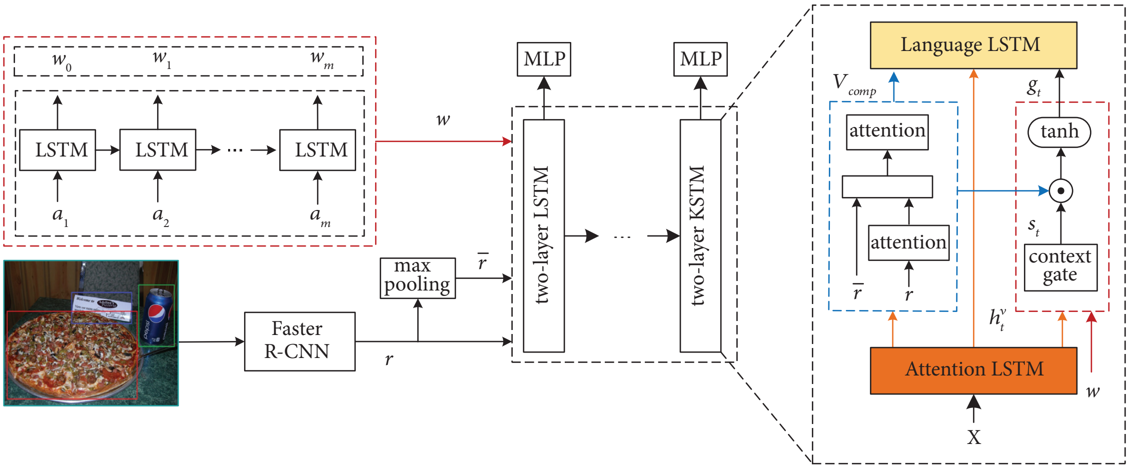


Figure 2: The overview of our proposed network. For the visual concept set $A = \{a_1, a_2, \ldots, a_m\}$, a unidirectional LSTM is adopted to obtain the encoded vector w. The region image feature r is extracted by a Faster R-CNN, and the image representation $\overline{r}$ is obtained by the max pooling applied on r. In decoder, a two-layer LSTM architecture is adopted. s_t indicates the fused textual context. Both V_{comp} and context-fused guidance g_t are passed into the language LSTM along with the hidden state h_t^v from attention LSTM. The input vector X consists of $\overline{r}$, w the word embedding, and the hidden state of language LSTM.

regions. In [4], they constructed a combined bottom-up and top-down attention mechanism. It calculates the attention feature vectors of the objects and other salient regions in image. In [5], the attention-on-attention module employs an attention gate to transform the result from a standard attention mechanism. Moreover, to improve the semantic representation of the generated captions, some approaches also focused on utilizing specific semantic attribute, such as the visual concept [9]. In [8], the guidance vector is equipped with the visual concept to provide the decoder with high-level semantic information. In [20], they proposed a hierarchical attention network to enhance the caption richness by incorporating the visual concept and other visual features.

2.2. Sequential Decision-Making. The models trained on vanilla CNN-LSTM framework often result in the problem of exposure bias [10]. To mitigate this, the reinforcement learning was applied on image captioning by introducing sequential decision-making: agent takes account of the actions, states, and rewards in further sequences. In the case of image captioning, the action corresponds to choosing the next word and image; the state can be the visual context, previous prediction, and other information. The rewards can be any evaluation metric, such as BLEU-N [21] and CIDEr [12]. Several works have applied the sequential decision-making. In [10], the REINFORCE is used to optimize a user-specified evaluation metric during training directly. However, it lacks adequate generalities to other evaluation metrics. In [11], the self-critical sequence training (SCST) framework is proposed. In SCST, the generated captions are evaluated at sentence level. Afterwards, in [13], they incorporated a discriminative loss component into the training objective to produce the caption with high discriminability. To capture crucial compositional information in image, CAVP [14] was proposed to capture complex visual compositions over time. Recently, the B-SCST [15] extended the SCST framework for image captioning models by incorporating Bayesian inference. From the distribution obtained by a Bayesian DNN model, B-SCST generates the baseline reward by averaging predictive quality metrics.

3. Proposed Approach

In this section, we introduce the proposed CFG network in detail. As the architecture presented in Figure 2, our model consists of five components: (1) a text encoder, which encodes the visual concept; (2) an image encoder, which encodes the region image features; (3) an attention module, which calculates the attentive compositional visual features; (4) a guidance formulation module, which obtains the fused textual context through the context gate and calculates the context-fused image guidance; and (5) a captioner, which is an extension of the top-down captioner [4] for caption generating.

3.1. Text Encoder. As the visual concept reveals the objects in images explicitly, we introduce it to offset the separation between image and text. In this paper, the visual concept is denoted as $A = \{a_1, a_2, \ldots, a_m\}$, $a_j \in \mathbb{R}^{m \times E}$, where m is the count of the words in visual concept and E is the dimension of word embedding. Specifically, as the word a_j is isolated, therefore a unidirectional LSTM is employed as the text encoder to deal with A as follows:

$$w = \text{LSTM}(E(A)), \qquad (1)$$

where $E(\cdot)$ is the word embedding layer and $w \in \mathbb{R}^{m \times H}$, where H is the size of hidden state. w indicates the encoding semantic vectors of each word in A. It will be used to calculate the fused textual context in the guidance formulation module.

3.2. Image Encoder. For the given image I, to learn the visual information about objects, attributes, and relations, a pretrained Faster R-CNN [22] is adopted to extract the region image representation r as follows:

$$r = W^I [\mathrm{CNN}(I)], \tag{2}$$

where $r = \{r_1, r_2, \ldots, r_k\}$, $r_i \in \mathbb{R}^{2048}$, presents the semantic information of an image region and k indicates the number of selected ROIs according to the ranking scores. To reduce the calculate consumption, a transformation matrix $W^I \in \mathbb{R}^{H\times 2048}$ is applied on r to convert its dimension to $r \in \mathbb{R}^{k\times H}$. Consistent with prior works, the image representation at global level is formulated by a mean-pooling operation as follows:

$$\overline{r} = \frac{1}{k}\sum_{i=1}^{k} r_i, \tag{3}$$

where $\overline{r} \in \mathbb{R}^H$. Both r and $\overline{r}$ are used to compute the attentive compositional visual features.

3.3. Compositional Visual Features. The compositional visual features contain the image information at regional and global levels. As shown in Figure 2 (framed in blue), for the image feature vectors r and $\overline{r}$, an additive attention mechanism is applied to reduce the variance caused by sampling diverse image regions. Without loss of generality, we first introduce the general formulation of the attention computation used in this paper:

$$f_{\mathrm{att}}(\pi, q, h_t) = \mathrm{softmax}\left(w_\pi^T \tanh\left(W_q^\pi q + W_h^\pi h_t\right)\right), \tag{4}$$

where π indicates the attentive weight of the query vector q, and h_t stands for the hidden state output from LSTM unit. w_π^T, W_q^π, and W_h^π are the parameters to be learned. Accordingly, for the region image feature r, the attention computation is presented as follows:

$$\alpha_t = f_{\mathrm{att}}\left(\pi = \alpha, q = r, h_t^v\right). \tag{5}$$

Here, the parameters $w_\alpha^T \in \mathbb{R}^D$, $W_r^\alpha \in \mathbb{R}^{D\times H}$, and $W_h^\alpha \in \mathbb{R}^{D\times H}$ in this case, D indicates the dimension of attention layer, and h_t^v is the hidden state from attention LSTM. Then, the attentive region image feature z_t^r is computed as follows:

$$z_t^r = \sum_{i=1}^{k+1} \alpha_{i,t} \cdot r, \tag{6}$$

where $z_t^r \in \mathbb{R}^H$. Particularly, in contrast to previous works that only integrate the global image representation in the first LSTM layer, similar to equation (5), $z_t^{\overline{r}}$ is computed as the attentive vectors of $\overline{r}$. Then, we combine z_t^r with $z_t^{\overline{r}}$ as the compositional visual features:

$$V_{\mathrm{comp}} = \left[z_t^r; z_t^{\overline{r}}\right], \tag{7}$$

where [;] indicates the vector concatenation. The attentive compositional visual feature z_t^c is obtained as follows:

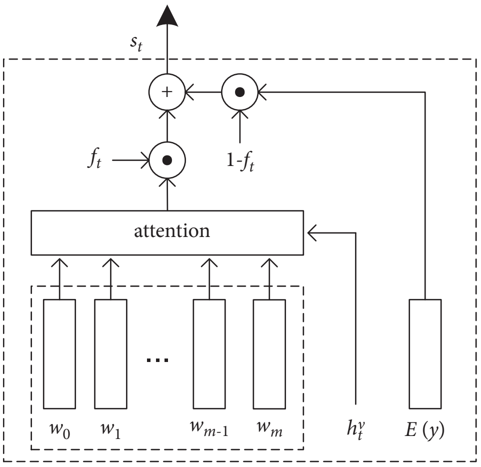


FIGURE 3: An illustration of the context gate. f_t is the scalar factor, s_t is the fused textual context, and *E(y)* indicates the word embedding vectors.

$$\begin{aligned}\beta_t &= f_{\mathrm{att}}\left(\pi = \beta, q = V_{\mathrm{comp}}, h_t^v\right),\\ z_t^c &= \sum_{i=1}^{k+1} \beta_{i,t} \cdot V_{\mathrm{comp}},\end{aligned} \tag{8}$$

where the trained parameters $w_\beta^T \in \mathbb{R}^D$, $W_V^\beta \in \mathbb{R}^{D\times H}$, and $W_h^\beta \in \mathbb{R}^{D\times H}$ here. In comparison to z_t^r, the decoder can capture more comprehensive visual information from z_t^c at each decoding step. Additionally, z_t^c is also utilized to modulate the guidance vectors.

3.4. Guidance Formulation. In [7], Zhou et al. conditioned the guidance information on the current word $W_e(y_t)$ and used the text-conditional image feature V as the guidance:

$$g_t = \tanh\left(V \odot W_e(y_t)\right), \tag{9}$$

where $W_e(\cdot)$ is a text-conditional embedding matrix. Through this, the model can focus on a part of the semantic image feature when capturing a specific word. In this paper, we extend this formulation with the visual concept vector w. Intuitively, if modulating the semantic image guidance g_t on w only, it may mislead the generating process because of the latent inappropriate word in visual concept set. Hence, it is essential to adaptively incorporate the semantic information from word embedding and visual concept. Inspired by [23], a context fusion gate is introduced. The structure is presented in Figure 3. By this component, our model can learn how much to attend to the context from two different sources. Utilizing the word embedding and visual concept, the context fusion gate is defined as follows:

$$s_t = f_t \odot \left(W_w z_t^w\right) + \left(1 - f_t\right) \odot \tanh\left(W_t\left[E\left(y_t\right)\right]\right), \tag{10}$$

where s_t is the fused textual context. $W_w \in \mathbb{R}^{E\times H}$ and $W_t \in \mathbb{R}^{E\times E}$ are the weight matrix; $\odot$ indicates the elementwise multiplication. The factor $f_t \in (0, 1)$ is calculated by a sigmoid activation function σ as follows:

$$f_t = \sigma\big(W_f [z_t^w; E(y_t)]\big), \tag{11}$$

where W_f is the transformation matrix. z_t^w indicates the attentive semantic vector, which is computed as follows:

$$\begin{aligned} \gamma_t &= f_{att}(\pi = \gamma, q = w, h_t^v),\\ z_t^w &= \sum_{i=1}^{m+1} \gamma_{i,t} \cdot w, \end{aligned} \tag{12}$$

where the parameter $w_\gamma^{\mathrm{T}} \in \mathbb{R}^D$, $W_w^\gamma \in \mathbb{R}^{D\times H}$, and $W_h^\gamma \in \mathbb{R}^{D\times H}$. Through this, z_t^w is equipped with the attentive visual information. Taking V_{comp} and s_t, the context-fused image guidance is formulated as follows:

$$g_t = \tanh\big(V_{\text{comp}} \odot W_s(s_t)\big), \tag{13}$$

where $W_s \in \mathbb{R}^{E\times H}$ is a transformation matrix. In comparison to equation (9), the context-fused image guidance g_t contains richer visual and textual context. It will be passed into the captioner as a time-dependent variable.

3.5. Captioner. The captioner consists of two separated LSTM networks: attention LSTM (AttLSTM) and language LSTM (LangLSTM). The input of AttLSTM is defined as the concatenation of previous word embedding vector $E(y_{t-1})$, the previous hidden state h_{t-1}^l from the LangLSTM, the visual concept vector w, and the image representation $\bar{r}$. That is,

$$\begin{aligned} \mathrm{X}_t &= \big[h_{t-1}^l; \bar{r}; w; E(y_{t-1})\big],\\ h_t^v &= \mathrm{AttLSTM}(\mathrm{X}_t, h_{t-1}^v), \end{aligned} \tag{14}$$

where h_t^v is used to attend over the visual features and semantic vectors, respectively. AttLSTM provides the LangLSTM with the feature vectors at the global level. In LangLSTM, the network focuses on generating the caption with both compositional, image feature V_{comp} and context-fused image guidance g_t:

$$\begin{aligned} X_t^L &= \big[V_{\text{comp}}; h_t^v; g_t\big],\\ h_{t+1}^l &= \mathrm{LangLSTM}\big(X_t^L, h_t^l\big). \end{aligned} \tag{15}$$

Then, we apply a multilayer perceptron (MLP) following by a softmax layer on hidden state h_t^l to obtain the probability distribution of each words as follows:

$$\hat{y}_t \sim p_t = \mathrm{softmax}\big(\mathrm{MLP}\big(h_t^l\big)\big), \tag{16}$$

where each value of p_t indicates the probability of corresponding word in vocabulary. Overall, our proposed network takes full advantage of image and text information to generate captions elaborately.

3.6. Training Strategy. Consistent with prior works [11], the sequence-level training strategy in this paper can be decomposed into two stages: the standard supervised learning with cross-entropy (XE) loss and the reinforcement learning with a self-critical reward. The XE loss is formulated as follows:

$$L(\theta) = -\sum_{t=1}^{N} \log\, p_\theta\big(y_t | y_{1:\,t-1}\big), \tag{17}$$

where N is the length of a generated caption, $y_{1:\,t-1}$ is a target ground-truth sequence, and θ indicates the model parameters. The supervised model is trained by minimizing this value. Then, the one with best performance is chosen as the initial network for next training stage. During reinforcement learning, the negative expected reward is minimized as follows:

$$L(\theta) = -\mathbb{E}_{y^s \sim p_\theta}\big[r\big(y_{1:\,T}\big)\big], \tag{18}$$

where $r(\cdot)$ is the standard metric evaluation (CIDEr [12] in this paper). According to SCST [11], the gradient of $L(\theta)$ can be approximated as follows:

$$\nabla_\theta L(\theta) \approx -\big(r\big(y_{1:\,T}^s\big) - r\big(\hat{y}_{1:\,T}\big)\big)\nabla_\theta \log\, p_\theta\big(y_{1:\,T}^s\big), \tag{19}$$

where $y_{1:\,T}^s$ is the caption sampled from the word distribution and $\hat{y}_{1:\,T}$ is the generated caption by greedy searching. The resulting reward signal $r(y_{1:\,T}^s) - r(\hat{y}_{1:\,T})$ can be treated as a baseline score. The probability of each word in the sampled captions will be increased if $r(y_{1:\,T}^s)$ is higher than $r(\hat{y}_{1:\,T})$, and vice versa.

4. Experiments

In this section, the dataset and evaluation metrics are introduced first. Then, the implementation details and the comparing models are described. Finally, we discuss the quantitative and qualitative experiments.

4.1. Dataset and Metrics. The MS COCO dataset [24] is one of the most popular benchmark datasets for image captioning task. There are 82,783 images in training set, 40,504 images in validation set, and 40,775 images in test set, respectively. For a fair comparison, the dataset using "Karpathy" split (http://cs.stanford.edu/people/karpathy/deepimagesent/) is adopted in this paper. It contains 113,287 images for training, 5000 images for validation, and 5000 images for test, respectively. The statistics of these two splits are summarized in Table 1. The COCO evaluation toolkit (https://github.com/tylin/coco-caption) is used to report the captioning performance across following metrics: BLEU-N (N = 1, 2, 3, 4) [21], METEOR [25], ROUGHE-L [26], CIDEr [12], and SPICE [27]. In particular, SPICE is defined over the tuples divided into several categories, such as objects, relations, and attributes. It shows a reasonable correlation with human judgments. All of these metrics with a larger score indicate a better effect.

4.2. Implementation Details

4.2.1. Preprocessing. For the region image representation, we use the bottom-up features provided by [4] which extracted top $k = 36$ features in each image as salient regions. The visual concept is detected by a pretrained model [9]. Only object attribute (nouns) is preserved. We convert all the sentences to lowercase, replace the punctuation with space,

Table 1: Statistics of the MS COCO dataset.

Split	Default		Karpathy	
Subset	Image	Caption	Image	Caption
Training	82,783	414,113	113,287	566,738
Validation	40,504	202,654	5000	25,010
Test	40,775	—	5000	25,010

The symbol "—" indicates the data are not public.

and preserve the captions with a length less than 16. The words that occurred less than five times are removed. As a result, there are 10,369 words left in the vocabulary.

4.2.2. Parameter Settings. Only top five attributes in visual concept set are preserved, namely, $m = 5$. The dimension E of word embedding layer is set to 1000. The attention layer size D is set to 1024. For AttLSTM and LangLSTM, the dimension H of hidden state and memory cell is set to 1300. During supervised learning with XE loss, Adam optimizer [28] is adopted with the initial learning rate $5e-4$. We shrink it by 0.8 every 3 epochs. During reinforcement training, the Adam optimizer [28] is initialized with learning rate $5e-5$. We trained the network for 30 epochs with batch size 80 during the first stage. During sequence-level training, we trained the model for 50 epochs with batch size 100. If there is no improvement for 5 epochs during XE training and 8 epochs during sequence-level training, the process is stopped. The whole training takes about 30 hours on a Linux server with an NVIDIA RTX 2080Ti GPU.

4.2.3. Model for Comparison. The following models are chosen for comparison: (1) NIC [16], which is a vanilla CNN-LSTM image captioning model; (2) SCST [11], which uses nondifferentiable metric for optimization; (3) up-down [4], which employs a bottom-up attention mechanism; (4) RFNet [17], which outputs the captions through multiple connections of CNN and LSTM; (5) HAN [20], which uses the hierarchy features to extend the caption richness; and (6) RAtt-Soft [29], which integrates the visual relationship attention and region features to enhance caption generating.

In particular, as the visual features in [7] are extracted by a different CNN, to investigate the performance of different guidance formulation, we also conduct a study on the following ablation models: (1) CFG_V, which only preserves the compositional visual feature and removes the visual concepts, context fusion gate, and context-fused image guidance. (2) CFG_E, which adopts the guidance defined in equation (9) and removes the visual concept, and context fusion gate. It is a 1-gram word-level guidance. (3) CFG_A, in which the factor f_t is removed. The fused textual context s_t is computed by a vector addition directly. Their performance will be discussed in the Ablation Studies section.

4.3. Quantitative Analysis. The evaluation results on the test portion of the Karpathy splits are summarized in Tables 2 and 3. All the scores were inferred by beam searching with size 3. For the cross-entropy loss training (Table 2), our model achieves competitive scores with RAtt-Soft [29]. For the sequence-level optimization (Table 3), our model obtains the scores with advantages across all metrics except for ROUGE-L and SPICE. Optimized by CIDEr, the scores of CFG on all metrics are increased in Table 3. Especially the score on CIDEr is improved from 114.0 to 125.4. The comparison results indicate our model can effectively improve the captioning performance by leveraging the compositional visual feature and context-fused image guidance. Besides, by sequence-level training, our network can significantly promote the results on each evaluation metric and outperform other models. However, it also should be noted that our model fails to achieve an advantage score on SPICE metric on both Table 2 and Table 3. As mentioned, SPICE is defined over the objects, relations and attributes. In [29], RAtt-Soft utilizes the scene graph and visual relation features to precisely map visual relationship information to the semantic description. This indicated a limitation of our proposed network.

4.4. Qualitative Analysis. For an intuitive presentation of the image captioning effect of the model with different guidance formulation, some examples are shown in Figure 4. Compared to CFG_E, the full model CFG can understand the image with detected salient objects (*with a rainbow*, *holding a racket*, *next to glass of beer*, and *with luggage*), but CFG_E neglects these instances and focuses on the main content of the images. In addition, CFG can better recognize the object *remote control*, while CFG_E mistakes it as *computer keyboard*. For the last image, CFG exactly describes the image with clear objects *pizza*, *broccoli*, and *vegetables*, while CFG_E just captures the object *broccoli* and depicts the image at a general level. These examples demonstrate that, in comparison to the guidance modulated on text-conditional embedding, the context-fused guidance is more advantageous to boost the model to depict the image comprehensively. Nevertheless, there are also several shortages in our proposed network, shown as the images presented in red frame. For the first image, our CFG succeeds in depicting the image with main instances, but it misunderstands the "desk" as "table" and generated inappropriate relation information "standing around a table." Similarly, in the last image, our model depicts the image with an incorrect position phrase "in the water." This indicates our network is insufficient to reason accurate relationships, especially among multiply image objects. One possible solution is to introduce the scene graph [30], which contains complex structural representation of image and sentences.

In Figure 5, we visualized the probabilities of the words the generated sentence and visual concept set, along with the object attention map, respectively. It can be found that the visual concepts are well applied to generate the captions. In the first example, the salient instances (*man*, *horse*, *filed*, and *cows*) are captured and the predicted words are highly corresponding to the detected visual concepts with high probabilities. The image content is well depicted by the generated sentence. This indicates that our model can exploit the high-probability visual concept to generate the relevant words in captions. For the second image, the weights of "bike" (0.34) and "sunset" (0.33) are much lower those of "man" (0.86) and "dock" (0.93), but our model can also

Table 2: Performance comparisons on MS COCO Karpathy test split under cross-entropy training.

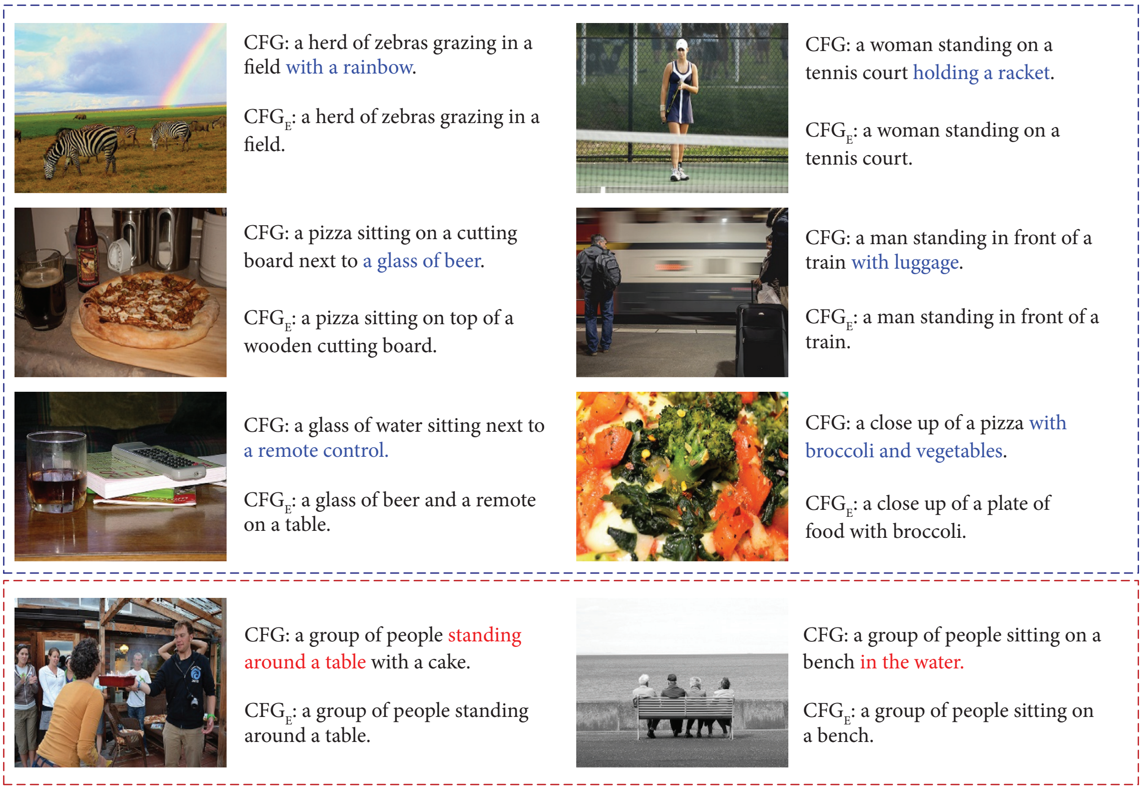
	Cross-entropy loss							
Metric	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE
NIC [16]	—	—	—	29.6	—	52.6	94.0	—
SCST [11]	—	—	—	30.0	25.9	53.4	99.4	—
Up-down [4]	77.2	—	—	36.2	27.0	56.4	113.5	20.3
RFNet [17]	76.4	60.4	46.6	35.8	27.4	56.8	112.5	20.5
HAN [20]	77.2	61.2	47.7	36.2	27.5	56.6	114.8	20.6
RAtt-Soft [29]	**79.2**	**61.8**	47.6	**36.9**	**28.3**	**60.9**	114.3	**20.8**
CFG	77.1	61.5	**47.9**	36.8	27.7	56.7	114.0	**20.8**

The best results (%) are highlighted in boldface. The symbol "—" indicates the results are not reported.

Table 3: Performance comparisons on MS COCO Karpathy test split under CIDEr-D score optimization.

	Sequence-level optimization							
Metric	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr	SPICE
NIC [16]	—	—	—	31.9	—	54.3	106.3	—
SCST [11]	—	—	—	34.2	26.7	55.7	114.0	—
Up-down [4]	79.8	—	—	36.3	27.7	56.9	120.1	21.4
RFNet [17]	79.1	63.1	48.4	36.5	27.7	57.3	121.9	21.2
HAN [20]	**80.9**	64.6	49.8	37.6	27.8	58.1	121.7	21.5
RAtt-soft [29]	80.4	63.4	48.9	37.5	**28.5**	**61.6**	122.1	**22.1**
CFG	80.5	**64.7**	**50.2**	**38.3**	28.2	58.3	**125.4**	21.6

The best results (%) are highlighted in boldface. The symbol "—" indicates the results are not reported.



CFG: a herd of zebras grazing in a field with a rainbow.

CFG$_E$: a herd of zebras grazing in a field.

CFG: a woman standing on a tennis court holding a racket.

CFG$_E$: a woman standing on a tennis court.

CFG: a pizza sitting on a cutting board next to a glass of beer.

CFG$_E$: a pizza sitting on top of a wooden cutting board.

CFG: a man standing in front of a train with luggage.

CFG$_E$: a man standing in front of a train.

CFG: a glass of water sitting next to a remote control.

CFG$_E$: a glass of beer and a remote on a table.

CFG: a close up of a pizza with broccoli and vegetables.

CFG$_E$: a close up of a plate of food with broccoli.

CFG: a group of people standing around a table with a cake.

CFG$_E$: a group of people standing around a table.

CFG: a group of people sitting on a bench in the water.

CFG$_E$: a group of people sitting on a bench.

Figure 4: Generated captions by the models with different guidance formulation. The positive cases are framed in blue and the failed cases are framed in red.

reason them as the appropriate words in the caption, which enhances the comprehensiveness of text description. This shows the advantage of the context fusion gate. By selectively fusing the information of the visual concept and word embedding, it can address the issue of misleading decoding as much as possible. Moreover, both these samples demonstrate that our model is able to keep a better consistency with the image content.

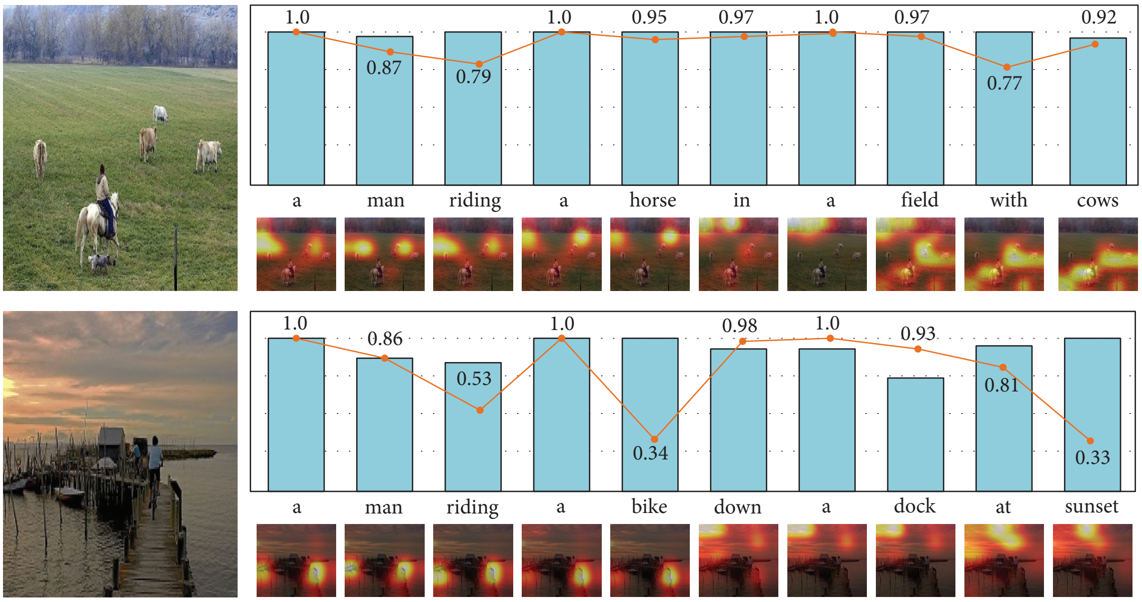


Figure 5: Visualization of the attention map (bottom) and the word probabilities in the generated sentence (histogram) and visual concept set (line). For conciseness, the weights of the visual concept are presented only.

Table 4: Performance comparison of the ablative models.

Model	Cross-entropy training			CIDEr optimization		
Metric	BLEU4	CIDEr	SPICE	BLEU4	CIDEr	SPICE
CFG_V	36.1	112.8	20.3	37.7	123.9	21.0
CFG_E	36.1	112.9	20.5	37.8	124.6	21.1
CFG_A	36.3	113.0	20.6	38.1	124.6	21.4
CFG	**36.8**	**114.0**	**20.8**	**38.3**	**125.4**	**21.6**

4.5. Ablation Studies. The evaluation results of the ablations are given in Table 4. Compared to CFG_V, CFG_E boosts the SPICE from 20.3 to 20.5 on cross-entropy training category, respectively. It suggests the effect of the text-conditional guidance to improve image captioning. In comparison to CFG_E, CFG_A achieved weak advantage results on cross-entropy training. After CIDEr optimization, the scores of BLEU4 and SPICE are boosted from 37.8 to 38.1 and 21.1 to 21.4, respectively. Among these models, CFG still achieved the best performance across all metrics. Particularly, the CDIEr score was significantly improved after sequence-level training. These indicate the following: (1) the introduced visual concept is helpful to boost image captioning. (2) The compositional visual feature and fused textual context are effective to improve the captioning quality. (3) The context fusion gate is beneficial to integrate the context from different sources for a better image captioning performance.

5. Conclusions

In this paper, an image captioning system within fused context guidance is proposed to enhance caption consistency of image. By the compositional visual feature, context fusion gate, and context-fused image guidance, our model further boosts the caption consistency of image. Extensive experiments demonstrate that our proposed model significantly improves the baseline method and outperforms other comparison approaches, which suggests the effect of the explicit consideration of using context-fused guidance.

However, the visual relation bias is not well handled. In the future, we will extend our network with scene graph, because it provides a unified representation that connects the objects, attributes, and their relationship in an image or a sentence. It is more advantageous for the model to employ the scene graph to depict an image with an accurate text description about object relationships.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

References

[1] S. Chen, Q. Jin, P. Wang, and Q. Wu, “Say as you wish: fine-grained control of image caption generation with abstract scene graphs,” in *Proceedings of the 2020 IEEE/CVF Conference Computer Vision and Pattern Recognition*, pp. 9962–9971, Seattle, WA, USA, June 2020.

[2] H. Wang, Y. Zhang, and X. Yu, “An overview of image caption generation methods,” *Computational Intelligence and Neuroscience*, vol. 2020, Article ID 3062706, 13 pages, 2020.

[3] K. Xu, J. Ba, R. Kiros et al., “Show, attend and tell: neural image caption generation with visual attention,” in *Proceedings of the 32nd International Conference Machine Learning*, pp. 2048–2057, Lille, France, July 2015.

[4] P. Anderson, X. He, C. Buehler et al., “Bottom-up and top-down attention for image captioning and visual question

answering," in *Proceedings of the 2018 IEEE Conference Computer Vision and Pattern Recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.

[5] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proceedings of the 2019 IEEE International Conference Computer Vision*, pp. 4633–4642, Seoul, South Korea, October 2019.

[6] X. Jia, E. Gavves, B. Fernando, and T. Tuytelaars, "Guiding long-short term memory for image caption generation," in *Proceedings of the 2015 IEEE International Conference Computer Vision*, pp. 2407–2415, Santiago, Chile, USA, December 2015.

[7] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: image captioning with text-conditional attention," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, pp. 305–313, Mountain View, CA, USA, October 2017.

[8] W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu, "Learning to guide decoding for image captioning," in *Proceedings of the 32nd AAAI Conference Artificial Intelligence*, pp. 6959–6966, New Orleans, LA, USA, February 2018.

[9] H. Fang, S. Gupta, F. N. Iandola, R. K. Srivastava, and L. Deng, "From captions to visual concepts and back," in *Proceedings of the 2015 IEEE Conference Computer Vision and Pattern Recognition*, pp. 1473–1482, Boston, MA, USA, June 2015.

[10] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba, "Sequence level training with recurrent neural networks," in *Proceedings of the 4th International Conf. Learning Representations*, San Juan, Puerto Rico, 2016.

[11] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, "Self-critical sequence training for image captioning," in *Proceedings of the 2017 IEEE Conference Computer Vision and Pattern Recognition*, pp. 1179–1195, Honolulu, HI, USA, July 2017.

[12] R. Vedantam, Lawrence, C. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proceedings of the 2015 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 4566–4575, Boston, MA, USA, June 2015.

[13] R. Luo, G. Shakhnarovich, S. Cohen, and B. Price, "Discriminability objective for training descriptive captions," in *Proceedings of the 2018 IEEE/CVF Conf. Computer Vision and Pattern Recognition*, pp. 6964–6974, Salt Lake City, UT, USA, June 2018.

[14] Z.-J. Zha, D. Liu, H. Zhang, Y. Zhang, and F. Wu, "Context-aware visual policy network for fine-grained image captioning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020, In press.

[15] S. Bujimalla, M. Subedar, and O. Tickoo, "B-SCST: bayesian self-critical sequence training for image captioning," 2020, https://arxiv.org/abs/2004.02435.

[16] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, "Show and tell: a neural image caption generator," in *Proceedings of the 2015 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 3156–3164, Boston, MA, USA, June 2015.

[17] W. Jiang, L. Ma, Y.-G. Jiang, W. Liu, and T. Zhang, "Recurrent fusion network for image captioning," in *Proceedings of the 15th European Conf. Computer Vision*, pp. 510–526, Munich, Germany, September 2018.

[18] K. Zheng, C. Zhu, S. Lu, and Y. Liu, "Multiple-level feature-based network for image captioning," in *Proceedings of the 19th Pacific-Rim Conference on Multimedia*, pp. 95–103, Hefei, China, September 2018.

[19] F. Sammani and L. Melas-Kyriazi, "Show, edit and tell: a framework for editing image captions," in *Proceedings of the 2020 IEEE/CVF Conference Computer Vision and Pattern Recognition*, pp. 4807–4815, Seattle, WA, USA, June 2020.

[20] W. Wang, Z. Chen, and H. Hu, "Hierarchical attention network for image captioning," in *Proceedings of the 33rd AAAI Conference Artificial Intelligence*, pp. 8957–8964, Honolulu, HI, USA, January 2019.

[21] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Association for Computational Linguistics*, pp. 311–318, Philadelphia, PA, USA, July 2002.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[23] Z. Tu, Y. Liu, Z. Lu, X. Liu, and H. Li, "Context gates for neural machine translation," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 87–99, 2017.

[24] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft COCO: common objects in context," in *Proceedings of the 13th European Conference Computer Vision*, pp. 740–755, Zurich, Switzerland, September 2014.

[25] S. Banerjee and A. Lavie, "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*, pp. 65–72, Ann Arbor, MI, USA, June 2005.

[26] C. Y. Lin, "Rouge: a package for automatic evaluation of summaries," in *Proceedings of the ACL Workshop on Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004.

[27] P. Anderson, B. Fernando, M. Johnson, and S. Gould, "Spice: semantic propositional image caption evaluation," in *Proceedings of the 14th European Conference Computer Vision*, pp. 382–398, Amsterdam, Netherlands, October 2016.

[28] D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.

[29] Z. Zhang, Q. Wu, and Y. Wang, "Exploring region relationships implicitly: image captioning with visual relationship attention," *Image and Vision Computing*, vol. 109, pp. 1041–1046, 2021.

[30] X. Yang, K. Tang, H. Zhang, and J. Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the 2019 IEEE Conf. Computer Vision and Pattern Recognition*, pp. 10685–10694, Long Beach, CA, USA, June 2019.