Image Captioning with a Joint Attention Mechanism by Visual Concept Samples

JIN YUAN, LEI ZHANG, SONGRUI GUO, YI XIAO, and ZHIYONG LI, Hunan University

The attention mechanism has been established as an effective method for generating caption words in image captioning; it explores one noticed subregion in an image to predict a related caption word. However, even though the attention mechanism could offer accurate subregions to train a model, the learned captioner may predict wrong, especially for visual concept words, which are the most important parts to understand an image. To tackle the preceding problem, in this article we propose Visual Concept Enhanced Captioner, which employs a joint attention mechanism with visual concept samples to strengthen prediction abilities for visual concepts in image captioning. Different from traditional attention approaches that adopt one LSTM to explore one noticed subregion each time, Visual Concept Enhanced Captioner introduces multiple virtual LSTMs in parallel to simultaneously receive multiple subregions from visual concept samples. Then, the model could update parameters by jointly exploring these subregions according to a composite loss function. Technically, this joint learning is helpful in finding the common characters of a visual concept, and thus it enhances the prediction accuracy for visual concepts. Moreover, by integrating diverse visual concept samples from different domains, our model can be extended to bridge visual bias in cross-domain learning for image captioning, which saves the cost for labeling captions. Extensive experiments have been conducted on two image datasets (MSCOCO and Flickr30K), and superior results are reported when comparing to state-of-the-art approaches. It is impressive that our approach could significantly increase BLUE-1 and F1 scores, which demonstrates an accuracy improvement for visual concepts in image captioning.

CCS Concepts: • **Computing methodologies** → **Scene understanding**; *Natural language generation*;

Additional Key Words and Phrases: Image captioning, LSTM, attention, visual concept, cross-domain

ACM Reference format:
Jin Yuan, Lei Zhang, Songrui Guo, Yi Xiao, and Zhiyong Li. 2020. Image Captioning with a Joint Attention Mechanism by Visual Concept Samples. *ACM Trans. Multimedia Comput. Commun. Appl.* 16, 3, Article 83 (July 2020), 22 pages.
https://doi.org/10.1145/3394955

This work was supported by the National Key Research and Development Program of China (2018*YFB*0203904), the Hunan Key R&D Program (2017*GK*2224), and the National Natural Science Foundation of China (61502157, 61502158, and 61502137).

Authors' addresses: J. Yuan, L. Zhang, S. Guo, Y. Xiao, and Z. Li, Hunan University, China, 2 Lushannan Rd, Yuelu Qu, Changsha Shi; emails: {yuanjin, zhangleizl, guosongrui001, yixiao_cess, zhiyong.li}@hnu.edu.cn.

1551-6857/2020/07-ART83 $15.00
https://doi.org/10.1145/3394955

1 INTRODUCTION

Recently, image captioning, which generates a meaningful natural language description of an image, has attracted more and more attention in computer vision. Beyond image classification and object detection, image captioning requires a deeper semantic understanding of an image [46], as well as a flexible and habitual language expression. Technically, this problem connects computer vision and natural language processing.

Generally, there are two main paradigms in existing image captioning approaches: template-based methods and deep network–based ones. The early approaches adopt the template-based methods, which first come up with several semantic fragments of an image and then connect them in a sentence by using a language model. Recently, the state-of-the-art model is the deep network–based methods, which directly generate a sentence based on an input image through an end-to-end architecture like the encoder-decoder framework [14, 64]. The encoder-decoder framework first employs an encoder like the convolution neural network (CNN) to encode an image and then outputs a sentence by using a decoder like the recurrent neural network (RNN). Actually, at each timestep, the decoder generates one caption word, which is only related to a subregion of an image. As a result, the attention mechanism is generated to explore such subregion for better predictions on caption words [1]. The first work was proposed by Xu et al. [50], which employs a visual attention model to attend an important spatial image region for generating the corresponding caption word. Currently, a variety of attention mechanisms are explored to enhance captioning performance from different aspects, including the attention by using different features [6], the attention at different granularity on an image [1], and the attention at different times of decoding [32].

No matter which attention approach is used, one core problem is how to accurately locate a subregion in an image for the desired word, especially when this word represents a visual concept. This is because visual concept words are the most important parts in a sentence to understand an image. Although several attempts have been proposed to enhance the attention accuracy for visual concepts [41], there are still several problems to be solved even with high attention accuracy. First, the attended visual object may be obscured in an image (Figure 1(a)), and thus the decoder cannot well learn this visual concept based on incomplete visual features. Second, given a visual concept, the relevant visual samples may appear rare in the whole dataset (see Figure 1(b)), and thus it is difficult to predict this concept by using limited samples. Third, at each timestep, the attention approaches only offer one subregion to be learned, which limits the capability of the decoder to explore the common characters for a visual concept.

To tackle the preceding problems, in this article we propose Visual Concept Enhanced Captioner (VCEC), which utilizes the joint attention mechanism with multiple visual concept samples (images containing only a visual concept) to strengthen concept recognition in image captioning. VCEC introduces multiple virtual LSTM units in parallel with an actual LSTM unit to construct a decoder. The actual LSTM, aiming to interpret images in sentences, receives image-caption pairs to update parameters. The virtual LSTMs receive multiple visual concept samples to guide the actual LSTM, and it will be removed at the end of training. Based on this design, the joint attention is employed to simultaneously offer multiple subregions, where one is explored by the actual LSTM and the others are received by the virtual LSTMs. Then, the model jointly learns parameters based on these subregions according to a composite loss function. Here, our loss function is composed of two parts: the captioning loss to measure the prediction accuracy between the ground truth and the generated caption words, and the auxiliary loss to measure the prediction accuracy for visual concept samples by virtual LSTMs. Different from traditional attention approaches that explore only one subregion in an image, VCEC adopts the joint attention to offer multiple subregions at each timestep when the desired word is a visual concept (see Figure 1(c)). As a result, richer

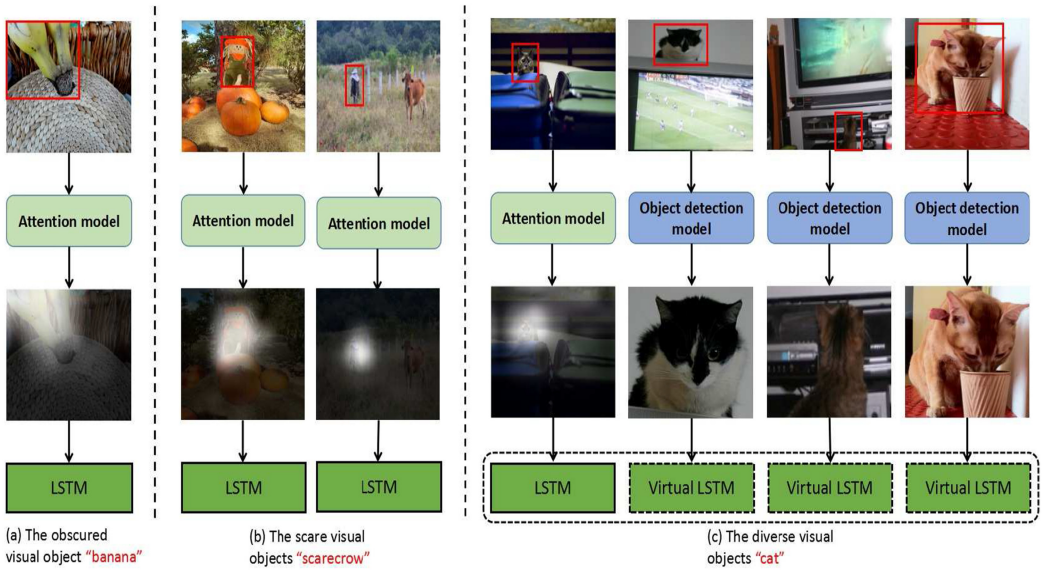


Fig. 1. A comparison between the traditional attention mechanism and the joint attention mechanism, illustrating the process of the traditional attention mechanism (a, b) and that of our joint attention mechanism (c).

information of a visual concept is passed to the decoder to find the common characters for this concept. Moreover, our approach could be extended to tackle the cross-domain problem in image captioning, where a captioning model is jointly learned by exploring two kinds of samples: image-sentence pairs from the source dataset and visual objects from the target dataset. Without extra caption labeling on the target dateset, the learned model integrates rich visual information from both datasets and thus is able to bridge visual bias between them. Experiments on two captioning datasets (MSCOCO and Flickr30K) demonstrate that VCEC achieves promising performance especially for predicting visual concepts.

The contributions of this work are threefold:

- We propose a joint attention mechanism incorporating multiple visual concept samples to strengthen visual concept recognition in image captioning. To the best of our knowledge, this is the first study that focuses on simultaneously offering multiple visual regions to jointly learn a captioning model.
- We propose a virtual LSTM unit to receive multiple visual regions, as well as a composite loss function to guide the learning process of the model. The architecture is quite simple, and it could be easily embedded into other existing encoder-decoder frameworks.
- Our model is able to bridge visual bias across different domains in image captioning. It is only required to offer visual concept samples by using object detection approaches, and there is no extra labeling cost generated. The experimental results demonstrate that VCEC achieves promising performance and is effective in tackling the cross-domain problem.

The rest of this article is organized as follows. We review related work in Section 2. In Section 3, we elaborate our model. Experimental results are reported in Section 4, followed by the conclusion in Section 5.

2 RELATED WORK

Early image captioning approaches first detect the semantic fragments in an image including objects, attributes, and activities, and then organize them into a sentence [10, 48]. For instance, Kulkarni et al. [21] first detected objects, attributes, and prepositions, jointly reasoned about them through a CRF, and finally filled appropriate slots in a template. Farhadi et al. [10] proposed an intermediate meaning space based on the triplet of object, action, and scene to translate an image into a sentence. Injected by semantic fragments, these methods require to collect human-generated sentences or templates but make the sentence pool hard to be scaled up.

Recently, inspired by the development in machine translation, the encoder-decoder framework has been applied to image caption generations [36, 42, 47]. For example, Vinyals et al. [45] have successfully applied this framework to the image captioning task, using CNN to encode image information and RNN to generate sentences. Song et al. [39] proposed multimodal stochastic RNNs to model the uncertainty propagation in video captioning, whereas Yang et al. [53] combined LSTM and a generative adversarial network to solve the error accumulation problem in video captioning. Based on this framework, more research has focused on an attention mechanism to improve captioning performance [12]. For example, Xu et al. [50] first proposed to incorporate a visual attention model into the framework, where an important spatial image region is noticed at each timestep to generate the corresponding caption word. Chen et al. [6] argued that a visual attention should incorporate spatial and channel-wise attentions, whereas Anderson et al. [1] proved that the attention mechanism on the object level is more effective than that on the grid level. Bin et al. [3] developed a soft attention mechanism working on a global view to generate better global representations for video captioning. Actually, not all of the words have corresponding visual signals, and thus Lu et al. [32] proposed to use a visual sentinel to adaptively attend to an image, knowing when and where to look at each timestep. Song et al. [40] proposed hierarchical LSTMs with adaptive attention to capture information at different scales. Zhou et al. [63] proposed POS-SCAN to serve as a word-region alignment regularization for the captioner's visual attention module. In addition, attention features are crucial for performance improvement. For instance, ALT [58] learns a transformation matrix from the image feature space to the context vector space, then employs a soft threshold regression to predict the spatial attention probabilities to replace the softmax regression due to the fact that it preserves more relevant local regions. GLA [23] integrates object-level features with image-level features to the attention mechanism to select more important regions. Yao et al. [56] further introduced the hierarchy parsing (HIP) architecture to characterize region-level, instance-level, and image-level features, then leveraged a Tree-LSTM to contextually refine these features to acquire better captioning features. Li et al. [22] applied a gate structure on the relationship-enhanced features with semantic relationships that are reasoned by GCN to generate image representations. Liu et al. [28] proposed to use the ground-truth features to ensure the correctness of attention.

In addition to the source image-caption pairs, researchers have found that extra resources like attribute [17, 57, 59], data corpus [16, 44], and evaluation metrics [5, 27, 29] are helpful to improve captioning performance. For example, Wu et al. [49] incorporated high-level semantic concepts into a decoder to achieve a significant improvement. Lu et al. [31] filled in sentence templates with specific attributes retrieved from websites. Chen et al. [4] explored the co-occurrence dependencies among attributes in the captioning task, whereas Yao et al. [55] utilized the visual relationship captured by GCN to better encode images in the captioning task. Yang et al. [52] adopted scene graphs to encode the relationship among attributes for image captioning. Injected by high-level information from attributes, captioning models could achieve a consistent improvement. In addition to attributes, the external data source has also been explored to enhance

captioning performance [44]. For example, Hendricks et al. [16] explicitly transferred the knowledge of semantically related objects to compose the descriptions about novel objects in the proposed Deep Compositional Captioner (DCC). The DCC model is further extended to an end-to-end system by simultaneously optimizing the visual recognition network, LSTM-based language model, and image captioning network with different sources in the work of Venugopalan et al. [44]. Liu et al. [30] aimed to generate discriminative captions by the REINFORCE algorithm, which trains the captioning module with text-to-image self-retrieval reward. Baig et al. [2] leveraged zero-shot learning to inject the novel objects produced by object detection model into the generated captions, whereas Yao et al. [54] introduced a copying mechanism to copy the detection results into image captioning results in the decoding stage, enabling the captioning for novel objects. Li et al. [25] proposed LSTM with a pointing mechanism to capture novel objects in the wild for image captioning. Chen et al. [7] proposed an adversarial training procedure to transfer the target captioning information to the source domain, whereas Yang et al. [51] and Zhao et al. [62] employed multi-task learning to solve the cross-domain problem in image captioning task.

This article implies the joint attention mechanism to better predict visual concepts in the image captioning task. Compared to the previous studies, the joint attention mechanism could simultaneously explore multiple subregions of visual objects, and thus it enhances the prediction accuracy for visual concepts in the captioning task. Moreover, injected by visual concept samples from different domains, our model could bridge visual bias across these domains by simultaneously exploring multiple visual concept samples. Therefore, it greatly saves labeling cost for captions and offers a new way to solve the cross-domain problem in the image captioning task.

3 VISUAL CONCEPT ENHANCED CAPTIONER

In this section, we first briefly introduce the framework of our VCEC and then elaborate the model including notations, model structure, and objective.

3.1 Framework

Figure 2 illustrates the training process of VCEC. Given a training image with the caption words "there is an airplane flying in the sky," the model first employs CNN to encode the image and then LSTM to decode the image to match with the caption words. Different from traditional encoder-decoder models, VCEC introduces the joint attention mechanism with visual concept samples to help learn the model. As Figure 2 shows, several visual concept samples of "airplane" are first selected from the visual concept sample set. Then, these selected samples are encoded by CNN and fed to the top LSTM at certain timesteps to help learn the model for better predictions on visual concepts. Compared to the general attention approaches that only offer one visual region, our model requires multiple visual concept samples at certain timesteps to explore the common characteristics among them, and thus it could learn a better predictor for visual concepts.

3.2 Notations

Given a training image $\mathbf{I}_0$ with the caption words $\mathbf{y} = (y_1, y_2, \ldots, y_T)$ to describe the semantic meanings of the image, our approach first selects a set of visual concept samples $\mathcal{I} = \{\mathbf{I}_1, \mathbf{I}_2, \ldots, \mathbf{I}_M\}$ with respect to each visual word y^{vis} in $\mathbf{y}$ from the visual concept sample set $\mathbf{V}$. Here, a visual word y^{vis} refers to a visual concept such as "dog" or "person," and a visual concept sample of y^{vis} only contains visual objects about y^{vis} (see the images of "airplane" in Figure 2). In experiments, we prebuilt a visual concept sample set $\mathbf{V}$ containing a set of visual concepts. For each visual concept, the relevant visual concept samples are extracted from the training set by using object detection techniques. Then, for each image $\mathbf{I}_m$ ($m = 0$ to M), the encoder employs CNN to extract feature maps $\mathbf{v}^m = (\mathbf{v}_1^m, \mathbf{v}_2^m, \ldots, \mathbf{v}_L^m)$, where each $\mathbf{v}_t^m$ corresponds to a special region in $\mathbf{I}_m$. Finally, the

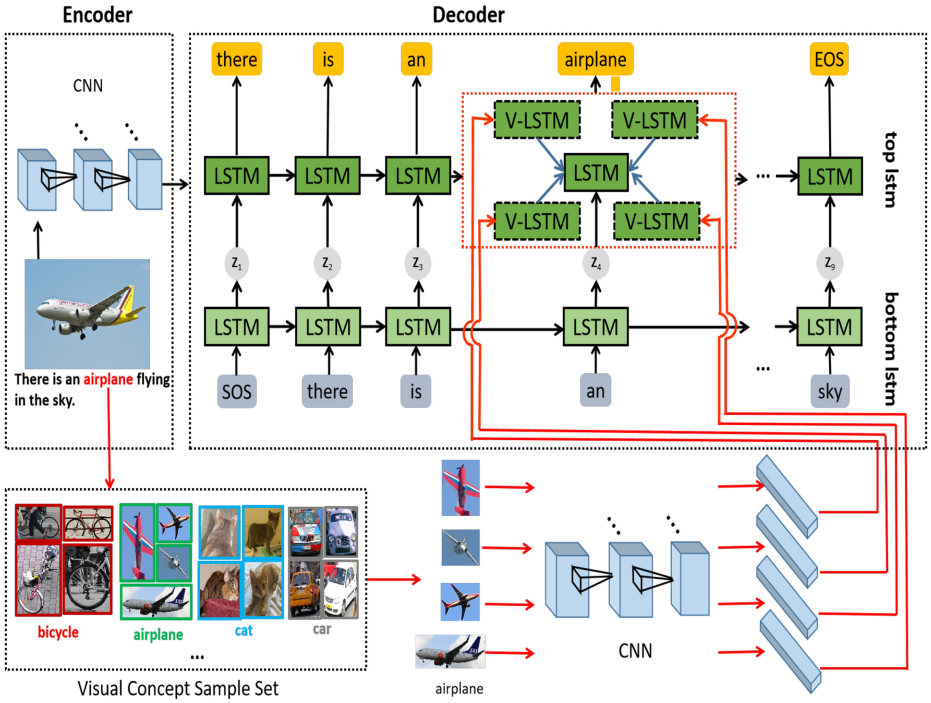


Fig. 2. An example to illustrate the training process of VCEC, where the bottom lstm is used to perform attention and the top lstm is employed to generate caption words. We embed multiple virtual LSTMs (denoted as V-LSTM) into the top lstm, which generates a visual word based on both the attended visual feature $\mathbf{z}_t$ and the visual features extracted from multiple visual concept samples.

text input $\mathbf{y}$ and the visual inputs $\mathbf{v}^m$ are fed to the decoder to generate the caption word by word by using LSTM.

3.3 Model Structure

The traditional decoder aims to predict each word y_t on the basis of an input $\mathbf{s}_t$ at each timestep t. Generally, $\mathbf{s}_t$ is a combination of two parts: a text input $\mathbf{x}_t$ and a visual input $\mathbf{z}_t$. At the training stage, the text input $\mathbf{x}_t$ is the output ground-truth word y_{t-1}, and the visual input $\mathbf{z}_t$ is a weighted sum of region features $\sum_{l=1}^{L} \alpha_{tl}\mathbf{v}_l^0$ from $\mathbf{I}_0$, where α_{tl} is the weight value for the region feature $\mathbf{v}_l^0$ calculated by the attention mechanism [6]. The attention mechanism is able to offer an accurate subregion to the decoder for predicting the corresponding caption word and thus achieves good performance. However, it faces two problems. First, when visual regions of a concept are obscured in images, or scarce in the training set, it is difficult to learn a robust predictor for this concept. Second, since different visual regions of a concept are fed to the decoder at different timesteps, it is difficult to capture the common characters among them.

To tackle the preceding problems, we propose VCEC to employ the joint attention mechanism to strengthen recognition ability for visual concepts in image captioning. In addition to the training sample $\mathbf{I}_0$, VCEC introduces multiple visual concept samples $\{\mathbf{I}_m\}_{m=1}^{M}$ to jointly learn the model. Figure 3 demonstrates the decoder structure of VCEC at a certain timestep, where multiple LSTMs are simultaneously adopted. At each timestep t, $M+1$ visual samples $\{\mathbf{I}_m\}_{m=0}^{M}$ are used to learn the model if y_t is a visual word. As a result, we construct $M+1$ LSTM units $\{LSTM_m\}_{m=0}^{M}$ to

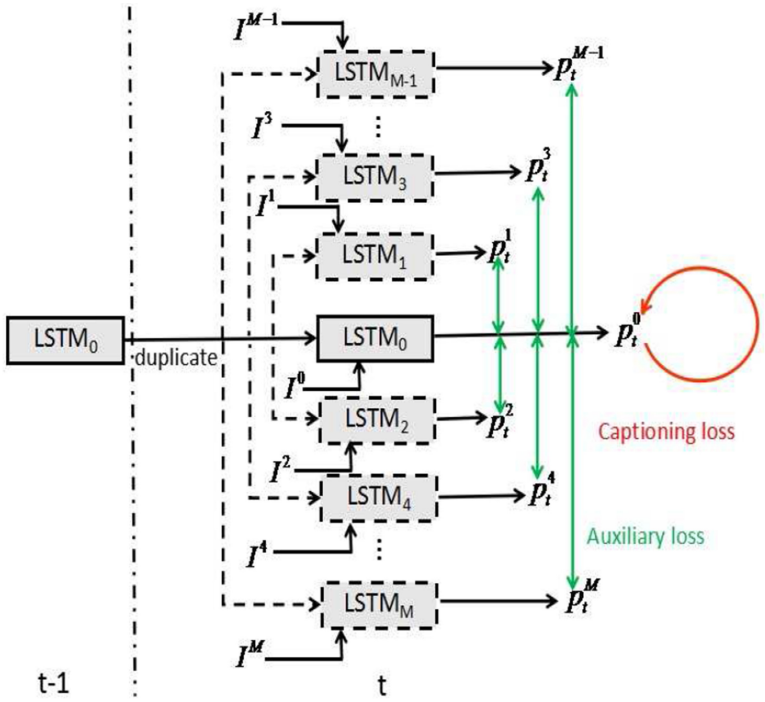


Fig. 3. The structure of the decoder by using the joint attention with multiple LSTMs, where $LSTM_0$ is the actual LSTM, and the others marked by dashed lines are the virtual LSTMs to process visual concept samples.

simultaneously process the $M+1$ visual samples. The first LSTM unit $LSTM_0$ is called the *actual LSTM*, which is used to interpret the semantic meanings of the training sample $\mathbf{I}_0$. The other M LSTM units $\{LSTM_m\}_{m=1}^{M}$ are then called the *virtual LSTM*, which work on the M visual concept samples, respectively, and are used to learn useful visual information from these samples to remedy the inadequacy of the actual LSTM. Concretely, at the beginning of timestep t, the virtual $LSTM_m$ ($m > 1$) is first initialized as a duplicate of $LSTM_0$. Then, $LSTM_m$ updates its state on the basis of an input $\mathbf{s}_t^m$, which is expressed as

$$\mathbf{s}_t^m = \mathbf{M}_t \cdot \mathbf{x}_t + \mathbf{M}_v \cdot \mathbf{z}_t^m, \tag{1}$$

where $\mathbf{M}_t$ and $\mathbf{M}_v$ are the text and visual embedding matrix, respectively. $\mathbf{x}_t$ is the output ground-truth word y_{t-1} in the training stage or the predicted sample token at the testing stage. The visual input $\mathbf{z}_t^0$ is obtained based on $\mathbf{I}_0$ by the attention mechanism, and $\{\mathbf{z}_t^m\}_{m=1}^{M}$ are calculated based on the visual concept samples $\{\mathbf{I}_m\}_{m=1}^{M}$ as Equation (2) shows:

$$\begin{aligned} \mathbf{z}_t^0 &= \sum_{i=1}^{L} \alpha_{ti} \mathbf{v}_{ti}^0, \\ \mathbf{z}_t^m &= U^m \mathbf{v}_t^m, \quad m = 1, 2, \ldots, M, \end{aligned} \tag{2}$$

where U^m is a transition matrix. Injected by different visual inputs, $M+1$ LSTM units separately update the parameters to generate $M+1$ hidden states $\{\mathbf{h}_t^m\}_{m=0}^{M}$ and predict the caption word y_t:

$$\begin{aligned} \mathbf{h}_t^m &= LSTM_m\left(\mathbf{s}_t^m, \mathbf{h}_{t-1}\right), \\ \mathbf{p}_t^m &= softmax\left(\mathbf{W}_g^m \cdot \mathbf{h}_t^m + \mathbf{b}_g^m\right), \end{aligned} \tag{3}$$

where $\mathbf{W}_g^m$ and $\mathbf{b}_g^m$ are parameters, and $\mathbf{p}_t^m$ is a probability distribution over all of the words to predict the caption word y_t by $LSTM_m$. Each probability distribution $\mathbf{p}_t^m$ ($m > 1$) by the virtual LSTM is generated based on a visual concept sample $\mathbf{I}_m$ and could be used to guide the learning process of the actual LSTM for better prediction on the visual concept. Benefiting from this, the actual LSTM could learn useful visual information from the virtual LSTMs to better predict visual concepts. At the end of timestep t, all virtual LSTMs will be cleared out, and it is unnecessary to learn parameters for virtual LSTMs. Therefore, the computational cost remains stable.

Compared to traditional attention approaches, VCEC employs the joint attention mechanism to offer multiple important subregions to the decoder at each timestep. The decoder could collect useful visual information from these subregions and explore the common characters among them. As a result, the model could better predict visual concepts in the image captioning task. Moreover, injected by diverse visual concept samples from different domains, our model could learn a robust visual concept predictor to bridge visual bias and thus is able to be extended to tackle the cross-domain problem in image captioning.

3.4 Objective

The goal of the image captioning model is to generate a sentence conditioned on an image $\mathbf{I}_0$, and thus the traditional encoder-decoder model aims to learn the model parameters Θ_0 of $LSTM_0$ to maximize the likelihood of the correct caption words as follows:

$$\Theta_0 = \arg\max_{\Theta_0} \sum_{(\mathbf{I},\mathbf{y})} \log p(\mathbf{y}|\mathbf{I}_0;\Theta_0). \tag{4}$$

Specifically, the captioner takes an action to predict y_t conditioned on the previously generated word y_{t-1} and the image $\mathbf{I}_0$ at timestep t.

Different from traditional approaches, our VCEC employs the joint attention mechanism, which utilizes virtual LSTMs $\{LSTM_m\}_{m=1}^M$ incorporating visual concept samples $\{\mathbf{I}_m\}_{m=1}^M$ to strengthen concept recognition ability for $LSTM_0$. In detail, when y_t is a visual word, the virtual LSTM receives a visual concept sample $\mathbf{I}_m$ and then outputs a probability distribution $\mathbf{p}_t^m$ (see Equation (3)). This $\mathbf{p}_t^m$ indicates the prediction accuracy for y_t by $LSTM_m$ and could be used to guide the status updating for $LSTM_0$. Based on this idea, we define the loss L in our model as two parts: the captioning loss L_C to measure the prediction accuracy between the ground truth and the generated caption words by $LSTM_0$, and the auxiliary loss L_A to measure the prediction accuracy for each visual concept sample by virtual LSTMs. We express the captioning loss L as follows:

$$L = L_C + \lambda L_A, \tag{5}$$

where λ is a weight parameter to balance these two parts. The captioning loss aims to correctly predict y_t based on $\mathbf{I}_0$, which is expressed as

$$L_C = -\sum_t \sum_{k'=1}^K g_{t,k'} \ln p_{t,k'}^0, \tag{6}$$

where K is the number of words in the vocabulary, $g_{t,k'}$ is the ground-truth value over the index k' at timestep t, and $p_{t,k'}^0$ is the output of an vector $\mathbf{p}_t^0$ over the index k'. The auxiliary loss aims to correctly predict each visual concept sample received by the virtual LSTMs. We express it as

$$L_A = -\sum_t B(y_t) \frac{1}{M} \sum_{m=1}^M \sum_{k'=1}^K g_{t,k'} \ln p_{t,k'}^m, \tag{7}$$

where $B(y_t)$ is an indicator for visual concepts. It returns 1 when y_t is a visual word and 0 otherwise. Assume that the actual LSTM aims to predict a visual concept at timestep t and outputs a vector $\mathbf{a}^0 = [a_1^0, a_2^0, \ldots, a_K^0]$ to calculate a probability distribution $\mathbf{p}^0$ by using the softmax function. We calculate the gradient for each output value a_k^0 as follows:

$$\begin{aligned}
\frac{\partial L}{\partial a_k^0} &= \frac{\partial(-\sum_{k'=1}^{K} g_{k'} \ln p_{k'}^0 - \frac{\lambda}{M}\sum_{m=1}^{M}\sum_{k'=1}^{K} g_{k'} \ln p_{k'}^m)}{\partial a_k^0} \\
&= -\sum_{k'=1}^{K} \frac{g_{k'}}{p_{k'}^0}\frac{\partial p_{k'}^0}{\partial a_k^0} - \frac{\lambda}{M}\sum_{m=1}^{M}\sum_{k'=1}^{K}\frac{g_{k'}}{p_{k'}^m}\frac{\partial p_{k'}^m}{\partial a_k^0} \\
&= -\sum_{k'=1}^{K} \frac{g_{k'}}{p_{k'}^0}\left(p_k^0\delta_{kk'} - p_k^0 p_{k'}^0\right) - \frac{\lambda}{M}\sum_{m=1}^{M}\sum_{k'=1}^{K}\frac{g_{k'}}{p_{k'}^m}\left(p_k^m\delta_{kk'} - p_k^m p_{k'}^m\right) \\
&= -g_k + p_k^0 - \frac{\lambda}{M}\sum_{m=1}^{M}\left(g_k - p_k^m\right) \\
&= -(1+\lambda)g_k + p_k^0 + \frac{\lambda}{M}\sum_{m=1}^{M} p_k^m,
\end{aligned} \tag{8}$$

where $\delta_{kk'}$ equals to 1 when $k = k'$ and 0 otherwise. It can be seen that the gradient of a_k^0 is related to the ground truth, the predicted probability by the actual LSTM, and those by the virtual LSTMs. Different from the traditional LSTM model [13], the actual LSTM in our model updates parameters by integrating the information from all virtual LSTMs based on visual concept samples, and this joint training process could explore the common characters among different visual samples, and thus it enhances the detection performance for visual concepts.

By minimizing the loss L, the model is jointly trained based on original training samples and visual concept samples. Consequently, the learned model could recognize visual concepts more accurately. Moreover, when the training samples and visual concept samples come from different domains, our model is able to learn a common description to bridge the visual bias among them.

4 EXPERIMENTS

In this section, we conducted experiments on two captioning datasets. Compared to traditional image captioning approaches, our VCEC introduces visual concept samples to strengthen concept learning for the image captioning task. To validate the effectiveness of our model, we explore to answer the following questions:

(1) Is the auxiliary loss in Equation (5) effective in enhancing captioning performance?
(2) How does the performance change with respect to the number of virtual LSTMs in Figure 3?
(3) Does the joint attention, which simultaneously explores multiple visual concept samples in VCEC, enhance the detection performance for visual concepts in the captioning task?
(4) Given an existing model trained on a source dataset, does our approach by using visual concept samples from a target dataset improve captioning performance on the target dataset?
(5) How does our approach perform compared to other state-of-the-art competitors?

Fig. 4. An example to illustrate visual concept samples on the COCO dataset.

Table 1. Detailed Information of the COCO and Flickr30K Datasets

Dataset Name	Images in the Training Set (#)	Images in the Validation Set (#)	Images in the Testing Set (#)	Words in the Vocabulary (#)	Concepts in the Visual Concept Set (#)
COCO	113,287	5,000	5,000	9,487	80
Flickr30K	29,000	1,014	1,000	8,510	62

4.1 Datasets

4.1.1 MSCOCO Dataset. The Microsoft COCO [26] dataset (COCO for short) is a popular dataset containing 82,783 training images, 40,504 validation images, and 40,755 testing images, respectively. We conducted experiments on 123,287 labeled images (82,783 training images and 40,504 validation images) and used the Karpathy split [18] to divide the images into three parts: 113,287 training images, 5,000 validation images, and 5,000 testing images. We used the training set to train our model, the validation set to tune the parameters, and the testing set to evaluate the performance. For each image, five captions are labeled. We truncated all captions within 16 words and then converted them to lowercase. This generates a vocabulary with 9,487 words, where each word appears at least five times in the captions.

To generate visual concept samples, we first selected the 80 object categories in the object detection task as our visual concepts. We then used the Faster R-CNN [37] pretrained on the MSCOCO dataset to extract visual objects from the training images of COCO. In detail, we set an IOU threshold of 0.7 for region proposal suppression and 0.3 for class suppression, then all candidates, whose detection confidence scores are larger than 0.5, are returned as visual objects. As a result, we constructed a visual concept sample set called *VCOCO* (Figure 4). Table 1 summarizes the details of our COCO dataset.

4.1.2 Flickr30K Dataset. Flickr30K [60] is a captioning dataset containing 31,014 images where each image has five captions. We used the Karpathy split to divide Flickr30K into three parts: 29,000 training images, 1,014 validation images, and 1,000 testing images (see Table 1). Similarly, we converted all caption words to lowercase and truncated the captions within 22 words. This generates a vocabulary with 8,510 words, where each word appears at least three times in the captions.

For the visual concept samples, we also selected the 80 visual concepts on COCO and used the Faster-RCNN on the MSCOCO dataset to extract visual objects from the training images of Flickr30K. As a result, only 62 visual concepts remained, and we denoted it as VFlickr30K.

4.2 Implementation

Our VCEC was constructed based on the up-down model [1]. Different from the traditional LSTM model, the up-down model employs two different LSTM branches in the captioning task, where one is used to perform attentions and the other is used to predict caption words. This model significantly improves attention accuracies, and thus it achieves excellent performance. Based on this model, we implemented our model as follows:

(1) *Encoder settings*: We employed ResNet101 [15] pretrained on ImageNet [38] as our encoder to extract the spatial image features from the final convolutional layer. Then, the image features were resized to the dimension of $14 \times 14 \times 2048$ by using a bilinear interpolation. In addition, we adopted the bottom-up attention model from Anderson et al. [1] on the COCO dataset. The bottom-up attention model employs Faster R-CNN [37] pretrained on Visual Genome [20] to perform a hard attention and could offer better attention features for the captioning task. We directly downloaded these attention features from the website.[1]

(2) *Decoder settings*: We kept the attention branch and introduced three virtual LSTMs ($M = 3$ in Equation (8)) for the prediction branch. We set 512 dimensions for the embeddings and the size of the LSTM memory. To achieve the best performance, we set $\lambda = 0.15$ in Equation (5) when using ResNet101 features and $\lambda = 0.2$ when using the bottom-up attention features.

(3) *Training and testing settings*: We applied AdaGrad [19] to optimize the network. The initial learning rate was 5×10^{-4} and then gradually reduced by multiplying a factor of 0.8 at every three epochs. The momentum parameter was set to 0.9. In each batch, 32 images are the inputs, then the model updates the parameters. We executed a total of 40 epochs on both datasets. In the testing stage, we used beam search to generate a caption for a given image, where the beam size was 3 in the experiments.

4.3 Evaluation Metrics

We used BLEU-N ($N = 1, 2, 3, 4$) [35], METEOR [9], ROUGE-L [11], and CIDEr [43] scores to evaluate our model. All of these metrics measure the consistency between generated captions and their ground truths, where this consistency is weighted by n-gram saliency and rarity. In addition, we adopted F1 scores to evaluate the performance of predicted visual concept words. Given a testing image $\mathbf{I}$, let $\mathbf{G}$ be a set of visual concepts in the ground truth for $\mathbf{I}$, and let $\mathbf{S}$ be a set of visual

[1] https://github.com/peteanderson80/bottom-up-attention.

Table 2. Performance of VCEC by Using ResNet101 Features with Different λ on the COCO Dataset

λ	0	0.05	0.10	0.15	0.20	0.25	0.30	1
B-1	0.749	0.753	0.756	**0.761**	0.757	0.755	0.736	0.460
F1	0.511	0.520	0.534	**0.541**	0.536	0.528	0.505	0.425

The bold signifies the best result.

Table 3. Performance of VCEC by Using ResNet101 Features with Different λ on the Flickr30K Dataset

λ	0	0.05	0.10	0.15	0.20	0.25	0.30	1
B-1	0.679	0.683	0.685	**0.686**	0.682	0.680	0.663	0.445
F1	0.133	0.154	0.165	**0.196**	0.174	0.135	0.102	0.085

The bold signifies the best result.

concepts in the generated sentence for **I**. The F1 score is calculated as follows:

$$Precision = \frac{|\,\mathbf{S} \bigcap \mathbf{G}\,|}{|\,\mathbf{S}\,|}, \qquad Recall = \frac{|\,\mathbf{S} \bigcap \mathbf{G}\,|}{|\,\mathbf{G}\,|}, \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}. \tag{10}$$

We used the average F1 score over all testing images to evaluate the captioning performance of a model for visual concepts.

4.4 Experimental Results and Analysis

4.4.1 Evaluation on the Auxiliary Loss. In this experiment, we evaluated the utility of the auxiliary loss. For our approach, we set the number of virtual LSTMs to 3 and tried different λ ($\lambda = 0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 1$) in Equation (5) to observe the performance change on both datasets. We mainly use B-1 and F1 to evaluate the performance in the following experiments because our approach only focuses on enhancing the prediction accuracies for visual concept words.

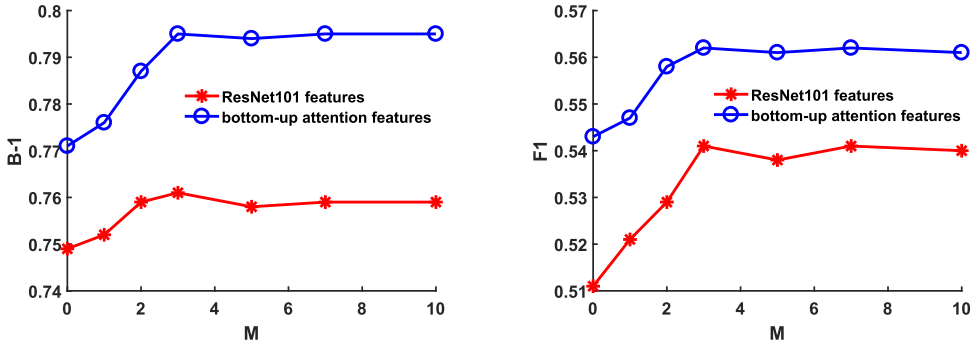
Tables 2 and 3 demonstrate the experimental results by using ResNet101 features on COCO and Flickr30K datasets, respectively. When λ is 0.15, our model performs the best, with a B-1 score of 0.761 on COCO and 0.686 on Flickr30K (an F1 score of 0.541 on COCO and 0.196 on Flickr30K). With λ growing from 0 to 0.15, the performance is improved. This is because the introduction of the auxiliary loss ensures the correct predictions for visual concept samples, and thus it is useful to explore common characters among multiple visual samples for the visual concept. Furthermore, when λ is growing from 0.15 to 1, the B-1 and F1 scores are declining. As Equation (5) shows, too large values of λ overshadow the utility of the captioning loss L_A, resulting in prediction errors for visual concepts. A similar performance trend is illustrated in Table 4 where the bottom-up attention features are employed on the COCO dataset.

4.4.2 Evaluation on the Number of Virtual LSTMs. This experiment evaluates the performance change with respect to the number of virtual LSTMs. We tried different numbers of virtual LSTMs ($M = 1, 2, 3, 5, 7, 10$), and randomly selected a visual concept sample for each virtual LSTM. Therefore, for each training sample, there are M visual concept samples at least to help learn the model.

Table 4. Performance of VCEC by Using Bottom-Up Attention Features with Different λ on the COCO Dataset

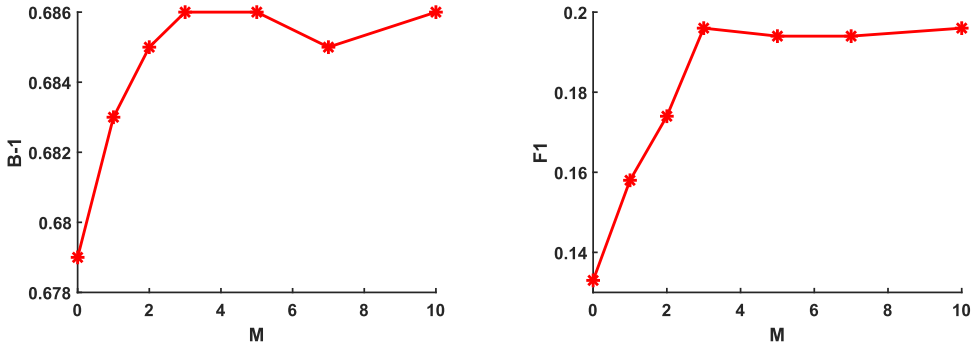
λ	0	0.05	0.10	0.15	0.20	0.25	0.30	1
B-1	0.771	0.783	0.786	0.790	**0.795**	0.770	0.734	0.526
F1	0.543	0.548	0.551	0.557	**0.562**	0.541	0.522	0.474

The bold signifies the best result.



(a) B-1 (b) F1

Fig. 5. The performance change with respect to the number of virtual LSTMs on COCO, where the visual concept samples are randomly selected from VCOCO.



(a) B-1 (b) F1

Fig. 6. The performance change with respect to the number of virtual LSTMs on Flickr30K, where the visual concept samples are randomly selected from VFlickr30K.

Figures 5 and 6 demonstrate the performance trends with respect to the number of virtual LSTMs. Without a virtual LSTM ($M = 0$), our approach by using ResNet101 features achieves a B-1 score of 0.749 on COCO and 0.679 on Flickr30K, whereas the corresponding F1 scores are 0.511 and 0.133, respectively. Comparatively, the introduction of bottom-up attention features brings a significant performance improvement on the COCO dataset. Furthermore, by adding one virtual LSTM, the performance of VCEC is greatly enhanced, especially on F1 scores. This result indicates that a virtual LSTM is effective in enhancing the captioning and detection performance for visual concepts because it utilizes a visual concept sample to better learn the visual concept in the captioning task. As the virtual LSTMs grow, the performance improves and becomes stable after M is

Table 5. Performance Comparison Between the Up-Down Model and Our VCEC by Using Visual Concept Samples from VCOCO Tested on the COCO Dataset

Method	Features	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	F1
Up-Down	ResNet101	0.749	0.586	0.445	0.333	0.258	0.542	**1.034**	0.511
VCEC	ResNet101	**0.761**	**0.589**	**0.456**	**0.335**	**0.260**	**0.546**	1.021	**0.541**
Up-Down	Bottom-up attention	0.771	0.594	0.451	0.361	0.268	0.561	1.130	0.543
VCEC	Bottom-up attention	**0.795**	**0.622**	**0.466**	**0.365**	**0.273**	**0.565**	**1.136**	**0.562**

The bold signifies the best result.

Table 6. Performance Comparison Between the Up-Down Model and Our VCEC by Using Visual Concept Samples from VFlickr30K Tested on the Flickr30K Dataset

Method	B-1	B-2	B-3	B-4	METEOR	ROUGE-L	CIDEr	F1
Up-Down	0.679	0.502	0.356	0.252	0.205	0.460	**0.539**	0.133
VCEC	**0.686**	**0.505**	**0.360**	**0.264**	**0.212**	**0.462**	0.536	**0.196**

The bold signifies the best result.

larger than 3. This is because the increasing of virtual LSTMs offers richer visual features to help our model better learn visual concepts.

4.4.3 Evaluation on Joint Attention. VCEC employs the joint attention by using multiple virtual LSTMs incorporated with visual concept samples (see Figure 3). This experiment evaluates the effectiveness of the joint attention. We compared the up-down model and our VCEC on both datasets. All training samples are the same for both approaches, except three virtual concept samples, which are extracted from the training set, are used in VCEC.

Tables 5 and 6 demonstrate the performance comparison results on COCO and Flickr30K datasets, respectively. With regard to B-1 and F1 scores, VCEC has significant improvement compared to the up-down model on both datasets. This result demonstrates that the joint attention by VCEC is effective as it simultaneously explores multiple visual concept samples to find the commonality of the concept. On the other criteria, VCEC performs similarly or slightly better. This is because VCEC focuses on improving the prediction accuracy for a single visual concept word, whereas these criteria mainly consider the correlations among multiple caption words.

Figures 7 and 8 show the performance comparison between the up-down model and our VCEC on individual visual concepts by using ResNet101 features. On the COCO dataset, VCEC outperforms the up-down model on 71 concepts, falls behind on 3 concepts, and is equal on 6 concepts. The average F1 score is improved from 0.511 to 0.541. On the Flickr30K dataset, VCEC can improve the average F1 score from 0.133 to 0.196 and only performs worse on 2 concepts. This result indicates that the joint attention by VCEC is more effective in learning visual concepts than the up-down model. The reason is that the joint attention could simultaneously explore multiple visual concept samples to better predict visual concept words.

4.4.4 Evaluation on Cross-Domain Learning. This experiment verifies the ability of VCEC in cross-domain learning. We used VCEC to learn a captioning model, where the training samples come from one dataset (called the *source dataset*), and the visual concept samples are from another (called the *target dataset*). In detail, three visual concept samples are employed in VCEC, and we set $\lambda = 0.15$ for the loss function. Since we have no bottom-up attention features for the Flickr30K dataset, this experiment was conducted based on ResNet101 features.

Tables 7 and 8 demonstrate the results of performance comparisons on Flickr30K and COCO datasets, respectively. From the results, we can draw the following conclusions:

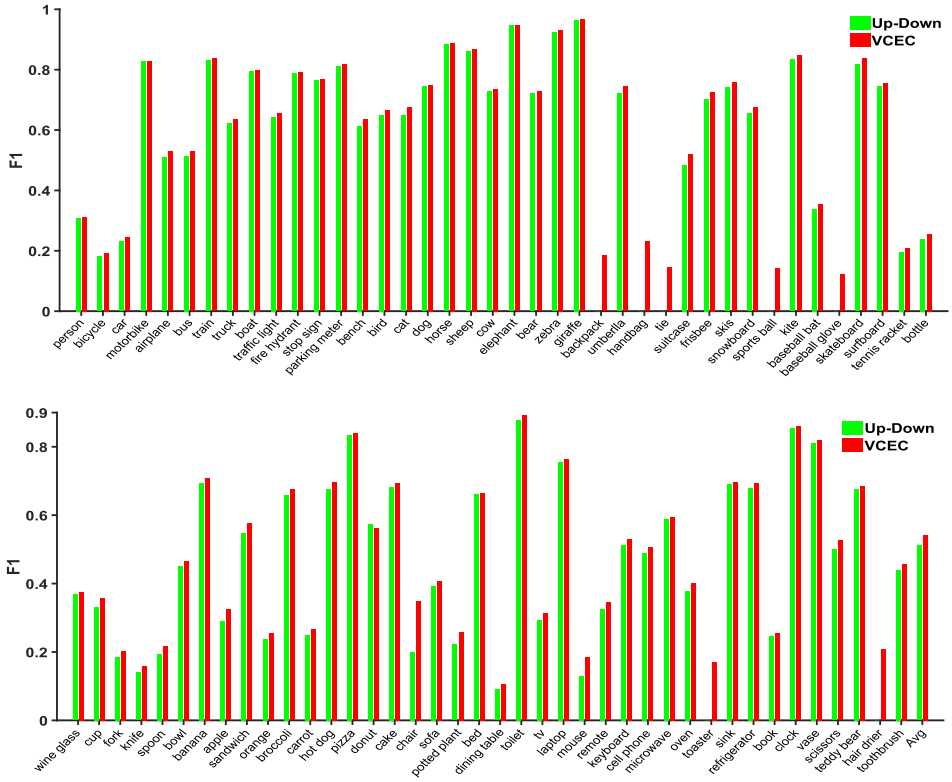


Fig. 7. The performance comparison between the up-down model and VCEC on 80 visual concepts of VCOCO measured by F1.

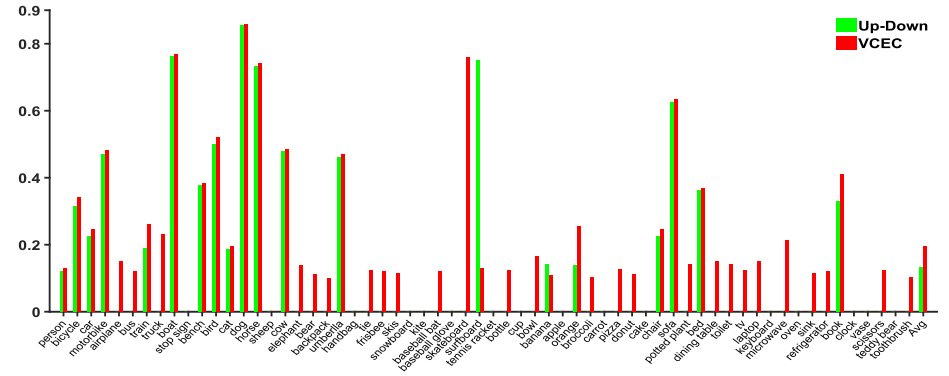


Fig. 8. The performance comparison between the up-down model and VCEC on 62 visual concepts of VFlickr30K measured by F1.

(1) Compared to the up-down model trained on the source dataset, our VCEC achieves better performance, especially on F1 scores, with an improvement from 0.160 to 0.219 on the Flickr30K dataset and from 0.121 to 0.226 on the COCO dataset. This result indicates that the introduction of visual concept samples from the target domain is effective because our VCEC could learn useful visual information from these samples to bridge the visual

Table 7. Performance Comparison Between VCEC and the Up-Down Model in Cross-Domain Learning, Where the Testing Is Performed on the Flickr30K Dataset

Method	Training Set	B-1	B-4	METEOR	F1	M1
Up-Down	COCO	0.594	0.105	0.156	0.160	0.456
VCEC	COCO+VFlickr30K	**0.612**	**0.112**	**0.160**	**0.219**	**0.660**
Up-Down	Flickr30K	0.679	0.125	0.205	0.133	0.676

The bold signifies the best result.

Table 8. Performance Comparison Between VCEC and the Up-Down Model in Cross-Domain Learning, Where the Testing Is Performed on the COCO Dataset

Method	Training Set	B-1	B-4	METEOR	F1	M1
Up-Down	Flickr30K	0.548	0.154	0.153	0.121	0.335
VCEC	Flickr30K+VCOCO	**0.564**	**0.159**	**0.155**	**0.226**	**0.662**
Up-Down	COCO	0.749	0.333	0.258	0.511	0.725

The bold signifies the best result.

bias between the source and target domains. This ability is quite valuable in cross-domain learning, as VCEC only requires the autodetection results by object detection techniques to improve captioning performance in another domain. This greatly saves the labeling cost as compared to traditional transfer learning approaches, which require manually labeling on new domains.

(2) Our VCEC performs worse than the up-down model trained on the target dataset. This is because VCEC only employs a part of the visual concept samples and ignores certain useful information on the target dataset, such as language styles and image styles. Moreover, we further explored individual captioning results to analyze this big performance gap. Figure 9 demonstrates several examples. From the figure, we can see that the generated sentences by VCEC actually cover the main semantic meanings in images, but they are quite different from the ground truth in language expressions. This is because the language styles by VCEC are learned from the Flickr30K dataset, while those in the ground truth are inherited from the COCO dataset.

(3) To better demonstrate the effectiveness of our method in cross-domain learning, human evaluation is employed to compare VCEC against the up-down model. We randomly selected 1,000 testing images from each dataset and invited six evaluators to judge whether the generated sentence is correct or not. Given a testing image, a generated sentence is correct only if it satisfies the semantic meaning of the image and resembles the human-generated one. From evaluators' responses, we calculated the M1 score, which equals the percentage of generated correct sentences, and the statistical results are as follows: testing on the Flickr30K dataset, the M1 scores are 0.660 and 0.676 for VCEC (trained on COCO and VFlickr30K) and the up-down model (trained on Flickr30K), respectively, whereas the corresponding values are 0.662 and 0.725 when the testing is performed on COCO. It is demonstrated that VCEC trained on the source dataset by using visual concept samples achieves close performance as compared to the up-down model trained on target dataset, which proves the effectiveness of our approach in handling the cross-domain problem for image captioning.

4.4.5 Comparison with the State-of-the-Art Approaches. Finally, Tables 9 and 10 list the performance comparison between our VCEC and several state-of-the-art approaches on the COCO

Images	Generated sentences	Ground truth
	A small cat is in bathroom sink	① Cat is caught stepping in to the bathroom sink. ② A large cat stands inside of a clean bathroom sink. ③ A cat looks up as it stands in the bathroom sink. ④ A cat climbing into a bathroom sink looking at someone. ⑤ A cute kitty cat in the sink of a bathroom near a brush and other items.
	A man in a black shirt is laying on a bed.	① Man laying on bed with shirt open looking into device for picture. ② Man laying down on a bed with his shirt open. ③ Man lying down on bed with shirt open in bedroom. ④ A may laying in on a bed with his right breasteses hanging out. ⑤ A man with his shirt open lying in bed.
	A woman in a red umbrella is walking on the beach.	① A person standing on top of a beach holding an umbrella. ② The person is walking on the beach with her umbrella up. ③ The sand area of a beach that has water on it and a woman with an umbrella over her head standing on the sand. ④ The woman with an umbrella watches many people in the ocean. ⑤ A woman standing on the beach and holding an umbrella.
	A black and white dog is playing with a ball.	① A dalmatian dog laying on the floor with a ball. ② A spotted dog is laying near someones foot with his ball. ③ A dog biting a ball at a woman's feet. ④ A dog laying on a rug while chewing on a ball next to someone's foot. ⑤ A spotted dog sits protectively with it's toy.

Fig. 9. Several examples on the COCO dataset to illustrate the different language styles between the ground-truth sentences and the generated sentences by VCEC.

and Flickr30K datasets, respectively. In VCEC, we employed three visual concept samples, and set $\lambda = 0.15$ when using ResNet101 features and $\lambda = 0.2$ when using bottom-up attention features. For fair comparison, results are reported for models trained with standard cross-entropy loss and for models optimized for CIDEr on the COCO dataset.

On the Flickr30K dataset, it is demonstrated that VCEC is only inferior to TOMS. This is because TOMS generates multiple sentences in the testing stage, whereas our VCEC only generates one sentence for testing. On the COCO dataset, VCEC outperforms all state-of-the-art approaches over B-n ($n = 1, 2, 3, 4$) and METEOR but has a slight performance drop over CIDEr as compared to GCN-LSTM and SGAE. This is because CIDEr assigns higher weights to infrequent n-gram phrases, whereas our approach mainly focuses on improving the detection performance for common single visual concepts. Therefore, VCEC could achieve a significant performance improvement over B-1, which results from the improvement for detecting visual concepts in the image captioning task.

We further provide several qualitative examples in Figure 10 to illustrate the captioning results by our model. Compared to the up-down model, VCEC could better identify visual objects in images because it employs the joint attention mechanism to offer multiple subregions of visual concepts to better learn the model.

5 CONCLUSION

In this article, we proposed VCEC to strengthen the recognition for visual concepts in image captioning. To the best of our knowledge, this is the first study that employs the joint attention mechanism to offer multiple subregions to better understand visual concepts in image captioning. Compared to the traditional attention mechanism that only offers one subregion each time, the noticed multiple subregions in our model provide richer and more complete information to the decoder,

Table 9. Performance Comparison Between VCEC and the State-of-the-Art Approaches on the COCO Dataset, Where $_{RL}$ Represents the Optimization over the CIDEr Metric, and an Asterisk (*) Indicates the Use of Bottom-Up Attention Features

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr
Soft-Attention [50]	0.707	0.492	0.344	0.243	0.239	0.855
Attribute LSTM [49]	0.740	0.560	0.420	0.310	0.260	0.940
Adaptive Attention [32]	0.742	0.580	0.439	0.332	0.266	1.085
LTG-Review-Net [17]	0.743	0.579	0.442	0.336	0.261	1.039
Up-Down [1]	0.745	—	—	0.334	0.261	1.054
Up-Down* [1]	0.772	—	—	0.362	0.270	1.135
Up-Down$^*_{RL}$ [1]	0.798	—	—	0.363	0.278	1.201
Attribute Driven [4]	0.743	0.579	0.443	0.338	—	1.044
FGSG Attention [61]	0.712	0.514	0.368	0.265	0.247	0.882
obj-R+Rel-A$^*_{RL}$ [24]	0.792	0.632	0.483	0.363	0.276	1.202
GLA-BEAM3 [23]	0.725	0.556	0.417	0.312	0.249	0.964
NBT [33]	0.755	—	—	0.347	0.271	1.072
SR-PL$_{RL}$ [30]	0.801	0.631	0.480	0.358	0.274	1.171
SGAE$^*_{RL}$ [52]	0.808	—	—	0.384	0.286	1.278
GCN-LSTM$^*_{RL}$ [55]	0.809	—	—	0.383	0.286	**1.287**
VCEC	0.761	0.589	0.456	0.335	0.260	1.021
VCEC*	0.795	0.622	0.466	0.365	0.273	1.136
VCEC$^*_{RL}$	**0.823**	**0.638**	**0.486**	**0.386**	**0.287**	1.264

The bold signifies the best result.

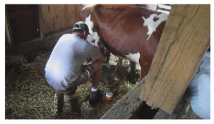
Table 10. Performance Comparison Between VCEC and the State-of-the-Art Approaches on the Flickr30K Dataset

Method	B-1	B-2	B-3	B-4	METEOR	CIDEr
Soft-Attention [50]	0.667	0.434	0.288	0.191	0.185	—
Adaptive Attention [32]	0.677	0.494	0.354	0.251	0.204	0.531
GLA-BEAM3 [23]	0.568	0.372	0.232	0.146	0.166	0.362
TOMS [34]	**0.691**	0.472	0.312	0.208	0.181	—
PMAS [8]	0.615	0.438	0.305	0.213	0.200	0.464
VCEC	0.686	**0.505**	**0.360**	**0.264**	**0.212**	**0.536**

The bold signifies the best result.

and thus the decoder could jointly learn the useful information from them to make more accurate predictions on visual concepts. In addition, our model is able to bridge visual bias across different domains in image captioning. We conducted extensive experiments on two image datasets. The experimental results demonstrate that our model outperforms the state-of-the-art approaches and is more effective in detecting visual concepts.

The proposed model is simple and can be easily embedded into the existing encoder-decoder approaches. In addition, it can be used to deal with the cross-domain problem in image captioning. Despite these advantages, our approach has several limitations. First, the model only focuses on enhancing the recognition ability for visual concepts and ignores other important factors in the image captioning task, such as language styles and motions Second, given a dataset, the current



Up-Down: a brown and white **horse** standing next to a man.

VCEC: a brown and white **cow** standing next to a man.

Up-Down: a woman sitting at a table with a **book**.

VCEC: a woman and man sitting in front of a **laptop**.

Up-Down: a bunch of fruit that are on a **table**.

VCEC: a bunch of fruit and vegetables on a **oven** top.

Up-Down: a man and a **dog** standing next to a **dog**.

VCEC: a man and a group of **sheep** standing in field.

Up-Down: a little girl standing in front of a **bed**.

VCEC: a little girl standing in front of a **chair** with a remote.

Up-Down: a small **bird** sitting on a tree branch.

VCEC: a bunch of **bananas** that are in a tree.

Up-Down: a man holding a **banana** in his hand.

VCEC: a man holding an **umbrella** in his hand.

Up-Down: a white **airplane** parked in a building.

VCEC: a white bullet **train** parked in a station.

Fig. 10. Several visual examples to illustrate the captioning results by VCEC and the Up-Down model, where red fonts represent the wrong concepts, and green fonts represent the correct.

model only processes the existing visual concepts and cannot be extended to deal with unseen concepts, which limits its applications in the real scenario.

In the future, we will continue our work in the following two aspects. First, we will improve our model by introducing unseen visual concept samples, and thus it can be used to identify new visual concepts in the image captioning task. Second, we plan to improve captioning performance by considering other factors like language styles and motions.

REFERENCES

[1] 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*. 6077–6086.

[2] Mirza Muhammad Ali Baig, Mian Ihtisham Shah, Muhammad Abdullah Wajahat, Nauman Zafar, and Omar Arif. 2018. Image caption generator with novel object injection. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*. 1–8.

[3] Yi Bin, Yang Yang, Fumin Shen, Ning Xie, Heng Tao Shen, and Xuelong Li. 2019. Describing video with attention-based bidirectional LSTM. *IEEE Transactions on Cybernetics* 49, 7 (2019), 2631–2641.

[4] Hui Chen, Guiguang Ding, Zijia Lin, Sicheng Zhao, and Jungong Han. 2018. Show, observe and tell: Attribute-driven attention model for image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 606–612.

[5] Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han. 2018. Temporal-difference learning with sampling baseline for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6706–6713.

[6] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6298–6306.

[7] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan Ting Hsu, Jianlong Fu, and Min Sun. 2017. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *Proceedings of the IEEE International Conference on Computer Vision*. 521–530.

[8] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. 2018. Paying more attention to saliency: Image captioning with saliency and context attention. *ACM Transactions on Multimedia Computing, Communications, and Applications* 14, 2 (2018), Article 48, 21 pages.

[9] Michael J. Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Workshop on Statistical Machine Translation*. 376–380.

[10] Ali Farhadi, Seyyed Mohammad Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David A. Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision*. 15–29.

[11] Carlos Flick. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the Workshop on Text Summarization Branches Out*.

[12] Lianli Gao, Kaixuan Fan, Jingkuan Song, Xianglong Liu, Xing Xu, and Heng Tao Shen. 2019. Deliberate attention networks for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8320–8327.

[13] Alex Graves. 2012. *Supervised Sequence Labelling with Recurrent Neural Networks*. Studies in Computational Intelligence, Vol. 385. Springer, Berlin, Germany.

[14] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6837–6844.

[15] 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*. 770–778.

[16] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond J. Mooney, Kate Saenko, and Trevor Darrell. 2016. Deep compositional captioning: Describing novel object categories without paired training data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–10.

[17] Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. 2018. Learning to guide decoding for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 6959–6966.

[18] Andrej Karpathy and Fei-Fei Li. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

[19] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations*.

[20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, 1 (2017), 32–73.

[21] 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 (2013), 2891–2903.

[22] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4653–4661.

[23] Linghui Li, Sheng Tang, Yongdong Zhang, Lixi Deng, and Qi Tian. 2018. GLA: Global-local attention for image description. *IEEE Transactions on Multimedia* 20, 3 (2018), 726–737.

[24] Xiangyang Li and Shuqiang Jiang. 2019. Know more say less: Image captioning based on scene graphs. *IEEE Transactions on Multimedia* 21, 8 (2019), 2117–2130.

[25] Yehao Li, Ting Yao, Yingwei Pan, Hongyang Chao, and Tao Mei. 2019. Pointing novel objects in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 12497–12506.

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Proceedings of the European Conference on Computer Vision*. 740–755.

[27] Anan Liu, Ning Xu, Hanwang Zhang, Weizhi Nie, Yuting Su, and Yongdong Zhang. 2018. Multi-level policy and reward reinforcement learning for image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 821–827.

[28] Chenxi Liu, Junhua Mao, Fei Sha, and Alan L. Yuille. 2017. Attention correctness in neural image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 4176–4182.

[29] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-aware visual policy network for sequence-level image captioning. In *Proceedings of the ACM International Conference on Multimedia*. 1416–1424.

[30] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European Conference on Computer Vision*. 353–369.

[31] Di Lu, Spencer Whitehead, Lifu Huang, Heng Ji, and Shih-Fu Chang. 2018. Entity-aware image caption generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 4013–4023.

[32] 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*. 3242–3250.

[33] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2018. Neural baby talk. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7219–7228.

[34] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. 2018. Show and tell more: Topic-oriented multi-sentence image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 4258–4264.

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. 311–318.

[36] Marco Pedersoli, Thomas Lucas, Cordelia Schmid, and Jakob Verbeek. 2017. Areas of attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. 1251–1259.

[37] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. 91–99.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[39] Jingkuan Song, Yuyu Guo, Lianli Gao, Xuelong Li, Alan Hanjalic, and Heng Tao Shen. 2019. From deterministic to generative: Multimodal stochastic RNNs for video captioning. *IEEE Transactions on Neural Networks and Learning Systems* 30, 10 (2019), 3047–3058.

[40] Jingkuan Song, Xiangpeng Li, Lianli Gao, and Heng Tao Shen. 2018. Hierarchical LSTMs with adaptive attention for visual captioning. arXiv:1812.11004.

[41] Lingyun Song, Jun Liu, Buyue Qian, and Yihe Chen. 2019. Connecting language to images: A progressive attention-guided network for simultaneous image captioning and language grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8885–8892.

[42] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Annual Conference on Neural Information Processing Systems*. 3104–3112.

[43] 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*. 4566–4575.

[44] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2017. Captioning images with diverse objects. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1170–1178.

[45] 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*. 3156–3164.

[46] Meng Wang, Hao Li, Dacheng Tao, Ke Lu, and Xindong Wu. 2012. Multimodal graph-based reranking for web image search. *IEEE Transactions on Image Processing* 21, 11 (2012), 4649–4661.

[47] Weixuan Wang, Zhihong Chen, and Haifeng Hu. 2019. Hierarchical attention network for image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*. 8957–8964.

[48] Yufei Wang, Zhe Lin, Xiaohui Shen, Scott Cohen, and Garrison W. Cottrell. 2017. Skeleton key: Image captioning by skeleton-attribute decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7378–7387.

[49] Qi Wu, Chunhua Shen, Lingqiao Liu, Anthony R. 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*. 203–212.

[50] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning*. 2048–2057.

[51] Min Yang, Wei Zhao, Wei Xu, Yabing Feng, Zhou Zhao, Xiaojun Chen, and Kai Lei. 2019. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia* 21, 4 (2019), 1047–1061.

[52] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. 2019. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10685–10694.

[53] Yang Yang, Jie Zhou, Jiangbo Ai, Yi Bin, Alan Hanjalic, Heng Tao Shen, and Yanli Ji. 2018. Video captioning by adversarial LSTM. *IEEE Transactions on Image Processing* 27, 11 (2018), 5600–5611.

[54] 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*. 5263–5271.

[55] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2018. Exploring visual relationship for image captioning. In *Proceedings of the European Conference on Computer Vision*. 711–727.

[56] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. 2019. Hierarchy parsing for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*. 2621–2629.

[57] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. 2017. Boosting image captioning with attributes. In *Proceedings of the IEEE International Conference on Computer Vision*. 4904–4912.

[58] Senmao Ye, Junwei Han, and Nian Liu. 2018. Attentive linear transformation for image captioning. *IEEE Transactions Image Processing* 27, 11 (2018), 5514–5524.

[59] 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*. 4651–4659.

[60] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.

[61] Zongjian Zhang, Qiang Wu, Yang Wang, and Fang Chen. 2019. High-quality image captioning with fine-grained and semantic-guided visual attention. *IEEE Transactions on Multimedia* 21, 7 (2019), 1681–1693.

[62] Wei Zhao, Benyou Wang, Jianbo Ye, Min Yang, Zhou Zhao, Ruotian Luo, and Yu Qiao. 2018. A multi-task learning approach for image captioning. In *Proceedings of the International Joint Conference on Artificial Intelligence*. 1205–1211.

[63] Yuenen Zhou, Meng Wang, Daqing Liu, Zhenzhen Hu, and Hanwang Zhang. 2020. More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[64] Zhihao Zhu, Zhan Xue, and Zejian Yuan. 2018. Think and tell: Preview network for image captioning. In *Proceedings of the British Machine Vision Conference*. 82.

Received November 2019; revised April 2020; accepted April 2020