Intention Oriented Image Captions with Guiding Objects

Yue Zheng, Yali Li and Shengjin Wang

Department of Electronic Engineering, Tsinghua University
zhengy17@mails.tsinghua.edu.cn, {liyali13, wgsgj}@tsinghua.edu.cn

Abstract

Although existing image caption models can produce promising results using recurrent neural networks (RNNs), it is difficult to guarantee that an object we care about is contained in generated descriptions, for example in the case that the object is inconspicuous in the image. Problems become even harder when these objects did not appear in training stage. In this paper, we propose a novel approach for generating image captions with guiding objects (CGO). The CGO constrains the model to involve a human-concerned object when the object is in the image. CGO ensures that the object is in the generated description while maintaining fluency. Instead of generating the sequence from left to right, we start the description with a selected object and generate other parts of the sequence based on this object. To achieve this, we design a novel framework combining two LSTMs in opposite directions. We demonstrate the characteristics of our method on MSCOCO where we generate descriptions for each detected object in the images. With CGO, we can extend the ability of description to the objects being neglected in image caption labels and provide a set of more comprehensive and diverse descriptions for an image. CGO shows advantages when applied to the task of describing novel objects. We show experimental results on both MSCOCO and ImageNet datasets. Evaluations show that our method outperforms the state-of-the-art models in the task with average F1 75.8, leading to better descriptions in terms of both content accuracy and fluency.

1. Introduction

Generating descriptions for images, namely image captioning, is a challenging task in computer vision. It can be used in many practical applications, such as robotic scene understanding and assistant systems for visually impaired users. In the past few years, deep neural networks are extensively used in image captioning [25, 12, 40, 17, 3], achieving fluent and accurate descriptions in commonly used datasets, *e.g.* MSCOCO [21]. However, they are limited in the control of the generation process. For instance, a picture may contain many objects but a description sentence usually contains only one or a small number of objects, as shown in Fig. 1. Although we can accurately classify or detect objects in the image with existing methods [13, 36, 32], we cannot force the language model to describe the object we care about. This can be important in practice because the model may be queried for a specific object. Including a novel object in the description is even harder, in the cases where the object has not been seen in the training data. Several recent works have studied the task of describing novel objects but it is still an open question. In this paper, we propose a novel approach that generates image captions with guiding objects (CGO). CGO can ensure that the user-selected guiding object is contained in a fluent description. The selected object can be any object detected from the image, even if it is unseen in the image caption training data.

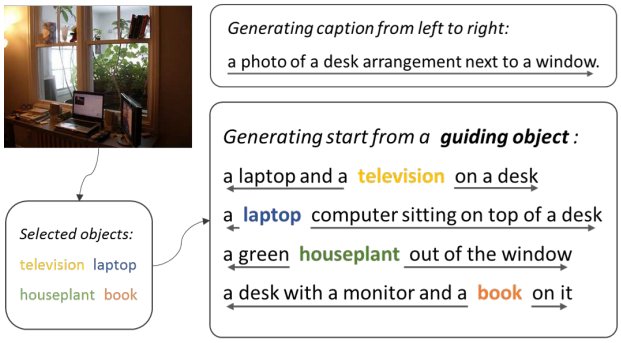

Figure 1. Existing models for image captioning generate the descriptions from left to right. Our CGO approach start generating with a selected object instead. CGO enables us to incorporate the selected objects into descriptions precisely, and generate a set of diverse and comprehensive descriptions for an image.

The encoder-decoder structure is most widely used in recent image caption works and recurrent neural networks (RNNs) are often used as the language model to generate

descriptions. In current approaches, the description is usually generated one by one as a word sequence from left to right. CGO is built on encoder-decoder structure, but instead of generating sequences from left to right, CGO generates sentences based on selected objects. We call them guiding objects as they guide the content of sequences in the generating process. A guiding object is the object that we want to include in the description. It may appear at any position in the sequence. We design a novel framework combining two LSTMs [14] to generate the left part and the right part of the sequence around the object. In this process, it is important that the content of the two sequences are coherent. In CGO, each LSTM encodes information of the other part of the sequence and then generates a sequence conditioned on the encoded sequence and visual features from the image. This helps the two sequences to connect with the guiding object fluently. It also enables us to generate multiple different descriptions for each selected object by providing different information sequences to the LSTMs.

Some earlier works on image caption tasks are template-based methods [20, 11]. These methods detect visual concepts from an image and fill them into templates directly. Although this enables us to control the presence of selected objects in the descriptions, the generated sentences are in limited forms and lack diversity. In CGO approach, the guiding object does not go through the encoding-decoding process and thus it acts like in template-based methods. At the same time, as the sequences on both sides are generated by the LSTMs, the sentences can be more fluent and diverse comparing with template methods. This makes CGO better at dealing with the novel object captioning task.

In this paper, we first demonstrate the characteristics of our method on MSCOCO to generate descriptions for each detected object in the image. Usually only a small portion of objects in each image is mentioned in image caption labels. With CGO, however, we can extend the ability of description to the objects which are neglected and thus provide a set of more comprehensive and diverse descriptions for an image (as in Fig.1). Then we apply CGO to the novel object captioning task and show its advantages when facing with unseen objects. We test our proposed approach on MSCOCO dataset and exhibit descriptions generated for ImageNet [34] objects. Experiments show that our method outperforms the state-of-the-art models on multiple evaluation metrics, such as METEOR [8], CIDEr [37], SPICE [2] and novel object F1 scores. The generated descriptions are improved in terms of both content accuracy and fluency.

2. Related Work

Image captioning. In earlier image captioning studies, template-based models [20, 11] or retrieval-based models [10] were commonly used. The template-based models detect visual concepts from a given image and fill them into templates to form sentences. Thus the generations usually lack diversity. The retrieval-based models find the most consistent sentences from existing ones and cannot generate new descriptions. In recent works, the structure of encoder-decoder with deep neural networks is widely used [40, 17]. In [43, 12, 23], attention mechanism is used to make the language model pay attention to different areas of the image at each time step. In [31, 33, 22, 45], reinforcement learning algorithms are applied to train the language models, enabling non-differentiable metrics to be used as training objectives.

Diverse descriptions. Controllability in generating process and diversity of descriptions are studied in recent years [6, 15, 35, 39, 41, 46]. GAN-based methods [6, 35] and VAE-based methods [15, 41] are used to improve diversity and accuracy of descriptions. In [26], generated sentences can contain words of different topics. [9] proposed a method to constrain the part-of-speech of words in generated sentences. Different from CGO, these methods do not precisely control the inclusion of objects in the descriptions. [7, 6] studied generating descriptive paragraphs for images. [16] generates descriptions for each semantic informative region in images. These approaches require additional labels in dataset, *e.g*. Visual Genome [19]. CGO approach does not need additional labels. The descriptions are generated based on the whole image with CGO, so the objects may have richer relationships with each other.

Describing novel objects. The novel object captioning task is first proposed by Hendricks *et al*. [5]. The proposed model DCC is required to describe objects unseen during training. In NOC [38], a joint objective is used to train object classifiers and language models together. LSTM-C [44] applied copying mechanisms in NLP to incorporate novel words in generations. NBT and DNOC [24, 42] use language models to generate templates with slots or placeholders and then fill them with objects recognized from images. Unlike [5, 38, 44, 24, 42], novel objects are not predicted by the language model in CGO, making it possible to contain novel words in sentences precisely. CBS [1] constrains the objects contained in generated sentences by adding constraints in beam search process. Different from CBS, novel words does not participate in the calculation of probability when decoding in CGO. Some works [29] in NLP researches also use approaches of generating sentences with constrained words.

3. Approach

Given an image, CGO is able to incorporate the selected guiding object into generated sentences. In this process, two LSTMs are combined to generate the partial sequences on two sides of the guiding object. We use LSTM-L to denote the LSTM generating the left part of the sequence and LSTM-R to denote the other that generates the sequence on

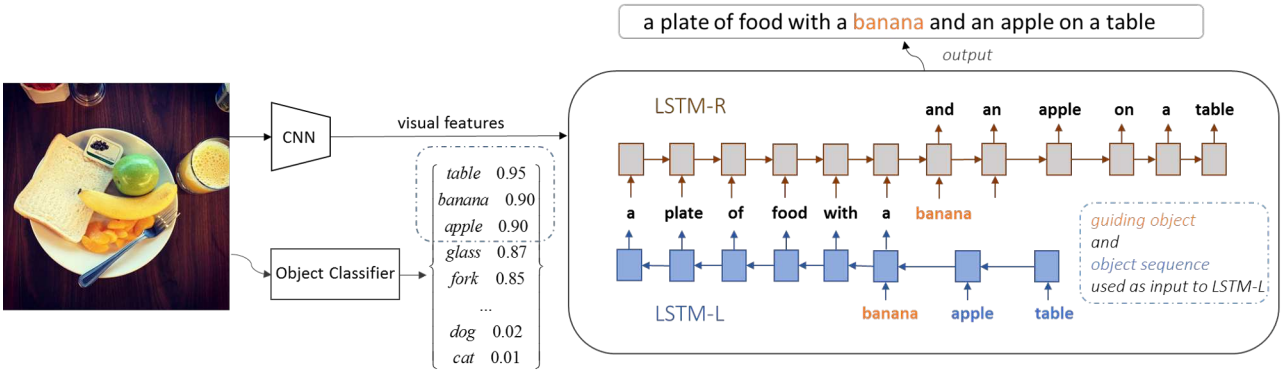

Figure 2. Our CGO approach. We select the guiding object and an object sequence according to the output of an object classifier. The sequence of objects is used as input to LSTM-L, providing with information about an assumed right-side sequence. LSTM-L generates the left-side sequence according to visual features and the input object sequence. The generated left-side sequence is then used as input to the LSTM-R to generate the right-side sequence. Then we connect the two partial sequences with the guiding object to get a complete description.

the right side. CGO can be applied flexibly to other existing RNN language models.

3.1. Problem Formulation

In commonly-used encoder-decoder models, convolutional neural networks (CNNs) [36, 13] are usually used as the encoder. Visual features representing information of the image from a CNN are then passed to a language model. RNNs such as the LSTM, are usually used as the language model in the decoding process. Given an image I, we aim to generate a sequence $\mathbf{y} = (y_1, y_2, ..., y_T)$ for description, where T denotes the length of the sequence and y_i is a word in the model vocabulary. The size of the vocabulary is V. Denoting $\boldsymbol{\theta}$ the parameters in the encoder-decoder model, the objective of the learning process is to find the optimal $\boldsymbol{\theta}$ so that

$$\boldsymbol{\theta}^* = \operatorname{argmax} \mathrm{p}(\mathbf{y}^*|\mathrm{I}, \boldsymbol{\theta}) \tag{1}$$

where $\boldsymbol{\theta}^*$ denotes the optimized model parameters, $\mathbf{y}^*$ denotes the ground truth sequence. When the LSTM is used as the language model, at each time step t it predicts the probability of the next word in the sequence according to image features $\mathbf{f_t}$, the input word x_t at this time step and the hidden state $\mathbf{h_{t-1}}$ of the LSTM at time $t-1$. x_t belongs to the model vocabulary.

$$p(y_t|y_1, ..., y_{t-1}) = LSTM(\mathbf{f_t}, x_t, \mathbf{h_{t-1}}) \tag{2}$$

The image features $\mathbf{f_t}$ vary in different model settings. In some models, *e.g.* NIC [40], image features $\mathbf{f}$ is only provided to the language model at time step $t = 0$. Models using attention mechanisms will use attended image features at each time step t as

$$\mathbf{a_t} = ATT(x_t, \mathbf{h_{t-1}}) \tag{3}$$

$$\mathbf{f_t} = \mathbf{f} \odot \mathbf{a_t} \tag{4}$$

where $\mathbf{a_t}$ denotes the attention weight maps at the time step t and $\odot$ denotes element-wise multiplying. The form of the function ATT to calculate attention weights varies with different attention mechanisms.

If we hope the generated sequence contain a specific word, the desired sequence becomes $\mathbf{y} = (y_1, ..., y_{k-1}, y_k, y_{k+1}, ..., y_T)$, where y_k is the specific word. At this time, the model output is conditioned on both image I and the word y_k. Model parameters are trained to be

$$\boldsymbol{\theta}^* = \operatorname{argmax} \mathrm{p}(\mathbf{y^*_{left}}|\mathrm{I}, \mathrm{y_k}, \boldsymbol{\theta}) \mathrm{p}(\mathbf{y^*_{right}}|\mathrm{I}, \mathrm{y_k}, \boldsymbol{\theta}) \tag{5}$$

where $\mathbf{y_{left}} = (y_1, ..., y_{k-1})$ and $\mathbf{y_{right}} = (y_{k+1}, ..., y_T)$. $\mathbf{y^*_{left}}$ and $\mathbf{y^*_{right}}$ are the ground truth partial sequences. The right-side partial sequence $\mathbf{y_{right}}$ could be of arbitrary length. We combine two LSTMs in opposite directions to complete the sequences on both sides of y_k.

3.2. LSTM-L

For the given image I and the word y_k, we first use the LSTM-L to generate the left-side partial sequence. At each time step t, LSTM-L predicts the previous word conditioned on the image features f_t, the input word x_t, and the hidden state h_{t+1}.

$$p(y_t|y_{t+1}, ..., y_k) = LSTM_L(\mathbf{f_t}, x_t, \mathbf{h_{t+1}}) \tag{6}$$

However, there are problems in this process. An image often contains more than one objects. These objects can be arranged in descriptions in different orders. For instance, 'there is an apple and a banana on the table' and 'there is a banana and an apple on the table' are both correct descriptions. These two sentences could appear in the ground truth

labels of an image at the same time. LSTM-L would have no idea about the right-side partial sequence when it is only provided with y_k (Fig. 3(a)). In experiments, we found that the model would tend to output a general and conservative results, such as 'a banana' in such process. It is usually correct in grammar but lacking in variety. In contrast, various objects would occur in the left-side partial sequence in human generated descriptions.

The objects to be described are usually decided before we speak. Similarly in image captioning, we could get sufficient information about objects in the image before generating a description. Therefore, we first assume that a set of objects will appear in the description, and set the order in which these objects are arranged. Then we can get a sequence of object labels that is assumed to occur in the right-side sequence. We denote the object label sequence as $S = \{object_1, ..., object_m\}$, where m is the number of objects in S and can be chosen arbitrarily. Objects in S will not appear in the sequence generated by LSTM-L, but they will affect the contents in the sequence (Fig. 3(b)). Sequence S is used as input to LSTM-L and encoded before y_k. LSTM-L now generates the sequence according to the image I, the assumed sequence S and y_k.

$$p(y_t|y_{t+1}, ..., y_k, S) = LSTM_L(\mathbf{f_t}, x_t, \mathbf{h_{t+1}}) \quad (7)$$

Similar to normal generating processes, when the predicted word is the ending label <END>, the left-side sequence is completed and the sentence reaches the beginning.

At training time, we randomly select an object as y_k from a ground truth caption label and then extract S from the partial sentence on the right side of y_k. The left part of the sentence is provided to LSTM-L as the ground truth sequence. For a given image, we minimize the cross-entropy loss of the model.

$$Loss = -\sum_{t=0}^{k-1} log\mathrm{p}(\mathrm{y}_\mathrm{t}^*|\mathrm{y}_{\mathrm{t}+1}, ..., \mathrm{y}_\mathrm{k}) \quad (8)$$

Note that the loss is only calculated for the generated left-side partial sequence, namely outputs at time steps earlier than $t = k$.

3.3. LSTM-R

After getting the left-side partial sequence from the LSTM-L, LSTM-R takes this sequence as input and complete the other part of the sentence. The model is now trained to be

$$\boldsymbol{\theta}_{\boldsymbol{R}}^* = \mathrm{argmax}\, \mathrm{p}(\mathbf{y}_{\mathbf{right}}^*|\mathrm{I}, \mathrm{y}_{\mathrm{left}}, \mathrm{y}_\mathrm{k}, \boldsymbol{\theta}) \quad (9)$$

In practice, we do not need to process the caption labels in the form as a right-side partial sequence. Instead, we can

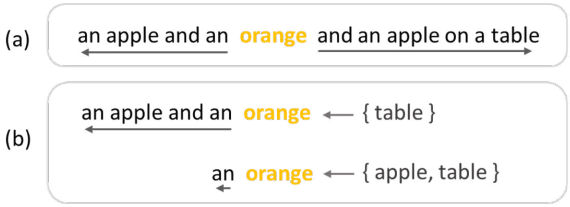

Figure 3. (a) Generated sequences on two sides of the guiding object ('orange') could be incoherent when they are generated independently. (b) Object label sequences are used as input to LSTM-L, providing with information about the right-side sequence. The LSTM-L generates different left-side sequences when the input sequences are different.

simply follow the process of training a normal LSTM that generates the sentences from left to right. The training loss for a given image and a selected y_k is different between these two processes

$$Loss_{normal} = -\sum_{t=0}^{T} log\mathrm{p}(\mathrm{y}_\mathrm{t}^*|\mathrm{y}_0, ..., \mathrm{y}_{\mathrm{t}-1}) \quad (10)$$

$$Loss_{LSTM-R} = -\sum_{t=k+1}^{T} log\mathrm{p}(\mathrm{y}_\mathrm{t}^*|\mathrm{y}_0, ..., \mathrm{y}_{\mathrm{t}-1}) \quad (11)$$

where $Loss_{normal}$ and $Loss_{LSTM-R}$ denote loss functions in the two processes. Note that $Loss_{normal}$ makes stricter restrictions than $Loss_{LSTM-R}$. The process of generating a complete sentence can be seen as a special case where the length of the input sequence is zero. On the other hand, the LSTM trained with complete sequences allows us to use the model more flexibly. When there is no object detected in the image (*e.g.* a picture of the blue sky), or when no object is requested to be contained in the descriptions, we can use LSTM-R as a normal language model and start from the time step $t = 0$. In this case, the process is reduced to normal ones and generates sentences from left to right.

Our approach can be applied to all types of RNN language models for image captioning. In the inference process, various decoding methods can be used, including the greedy sampling and beam search methods.

3.4. Novel Word Embedding

In the encoding process, an input word x_t is represented as a one-hot vector $\mathbf{x_t}$ and is then embedded with the learned parameter W_x. Embedding vector $W_x\mathbf{x_t}$ is used as input to the language model at time step t. A word that is unseen during training will not be generated by the language model when doing inference.

In CGO approach, when a novel object is selected as the guiding object, we can simply use the embedding vector from another seen object that is similar to this object. A

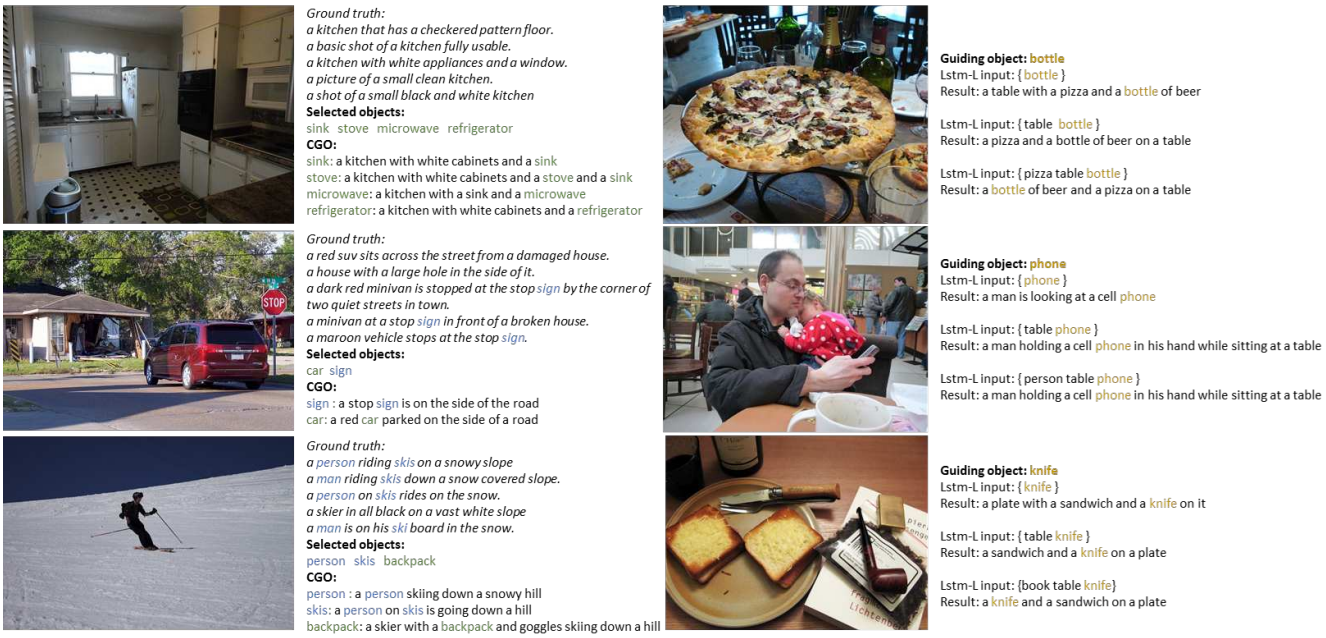

Figure 4. Examples of descriptions for selected objects are shown in the left column. The object at the beginning of each line in CGO results denotes the guiding word used for that description. Objects contained in the ground truth labels are in blue and the others are in green. Examples of diverse descriptions for a fixed guiding object is shown in the right column.

similar object can be chosen according to WordNet [28] or distances between word embedding vectors from word2vec [27] or GloVe [30]. In normal left-to-right generating process, using embedding vectors from a similar object cannot force the language model to generate the novel word. With CGO however, since the novel word is incorporated in the generated sentence directly, without going through the encoding-decoding process, we do not need the language model to predict the novel word. Instead, the novel word is only used in the encoding process, and the embedding result from a similar object is sufficient in this process.

3.5. Model Details

Caption Model. In our experiments, we use the bottom-up and top-down attention model (Up-Down) [3] as our base model. The LSTM-L and the LSTM-R are both Up-Down models. In our experiments, we use the pretrained model features from [4]. It is extracted from a Faster R-CNN model built on ResNet-101 [13] and is pretrained on MSCOCO and Visual Genome.

Object Classifier. The objects in an given image can be recognized with existing object detection models or object classifiers. In our experiments, we follow previous works [5, 1] using a multi-label classifier to determine whether an object appears in the image. We classify the 80 object categories in the MSCOCO object detection dataset. We use the same feature in the classifier as in the language model.

4. Experiments and Results

In this section we show the ability of CGO to incorporate selected objects into descriptions. In subsection 4.1 and 4.2, we show the characteristics of CGO by generating descriptions for each selected object in an image. In subsection 4.2 we show the diversity of the generated descriptions. In subsection 4.3 and 4.4 we apply CGO approach to the novel object captioning task.

Dataset. Models are trained and evaluated on MSCOCO dataset which includes 123287 images. There are 80 object categories labeled for object detection and each image is labeled with 5 human generated descriptions for image captioning. We follow the previous work [12] to preprocess the caption labels that all labels are converted to lower case and tokenized. Words occur less than 5 times are filtered out and the others form a vocabulary of size 9487. We use the Karpathy's splits [17] in subsection 4.1 and 4.2 which is widely used in image caption studies. 113287 images are used in training set, 5000 images in validation set and 5000 in test set. In subsection 4.3 we use splits following [5]. Details are in subsection 4.3. In subsection 4.4 we test the model on the ILSVRC2012 validation set which contains 1000 classes and each image is labeled with its category.

Training details. In our experiments, the object classifiers are optimized with stochastic gradient descent (SGD). The learning rate is set to 1e-4 and decays by 0.1 in every 10 epochs. The classifiers are trained for 20 epochs. The lan-

Model	METEOR	Avg.Num	Avg.R
Base ($b = 1$)	26.6	1.50	0.55
Base ($b = 3$)	27.3	1.68	0.59
Base ($b = 5$)	27.1	1.82	0.62
Base ($b = 10$)	26.7	1.98	0.66
Base (caption GT)	27.3	-	-
CGO ($k = 1$)	24.4	1.62	0.50
CGO ($k = 3$)	24.4	2.43	0.67
CGO ($k = 5$)	24.2	2.77	0.73
CGO ($k = 10$)	24.2	2.92	0.75
Caption GT label	-	2.01	0.61
CGO (caption GT)	28.0	-	-
CGO (det GT)	24.2	3.06	1.00

Table 1. 'Base' denotes the base model used as baseline. b indicates using the top b beam search generations. For CGO we use the top k objects predicted by the object classifier as guiding objects. 'Caption GT label' is the statistical result of the ground truth labels. Base (caption GT) shows descriptions containing at least one object that occur in ground truth caption labels. CGO (caption GT) shows the score of descriptions whose guiding objects occur in ground truth caption labels. CGO (det GT) shows results when we generate descriptions for each object in the image (using object detection ground truth labels). Avg.Num denotes the average number of object categories in descriptions for an image. Avg.R denotes the average recall.

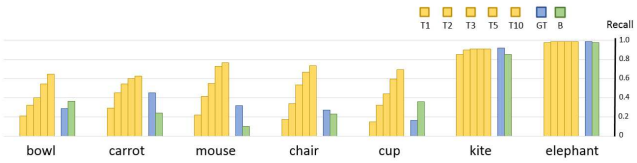

Figure 5. Examples of recalls for a single object category. T1∼T10 denotes CGO results with different numbers of selected objects. GT denotes the statistical results of the ground truth caption labels. B denotes the results of the base model.

guage models are optimized using Adam [18]. The learning rate is set to 1e-4 and decays by 10 in every 20 epochs. The LSTM-L is trained for 80 epochs and LSTM-R is trained for 40 epochs.

4.1. Describe Each Selected Object

To demonstrate the characteristics of our method, we generate one description for each object selected in images to get a set of sentences describing different objects in each image. We choose k objects with the highest probability as the guiding objects according to the outputs of the object classifier. We test the models with $k = 1, 3, 5, 10$ respectively. We count the average number of different object categories involved in the set of descriptions for each image.

	2 objects selected		3 objects selected	
Object label	Uniq.	M	Uniq.	M
Caption label	1.47	26.1	2.08	25.7
Detection label	1.48	24.7	2.62	23.5

Table 2. Results when we select guiding objects from caption labels and object detection labels. M denotes the METEOR score. Uniq. denotes the average number of unique descriptions for each fixed guiding object.

We also count the recall of each object category. That is, whether the object appearing in an image is mentioned in the set of descriptions. It should be noted that whether an object occurs in the image is decided according to the object detection labels, since caption labels often contain only a small portion of the objects appearing in the image.

Both the base model and CGO are trained and tested on MSCOCO Karpathy's splits [17]. Examples are shown in the left column of Fig. 4. In Table 1, we show results generated with the base model using beam search (beam size = b) as baseline, and results generated with CGO. The average recall is the macro average of recalls for all 80 object categories. The average object category number and recall of baseline model are 1.98 and 0.66 with $b = 10$. With CGO, the average number and recall are improved to 2.92 and 0.75 ($k = 10$). We also count the average number and recall for the ground truth caption labels. When $k = 5$ (there are 5 caption labels for each image), the object recall of CGO is 0.73, higher than that of the caption labels (0.61), indicating that CGO can describe the objects which are neglected in caption labels. Note that although the base model can describe more object categories with larger beam size, it cannot control which objects are described in the process.

The METEOR scores of CGO is around 24.2, lower than that of the base model (26.7). The evaluation method only compare the results with ground truth caption labels. Even if objects appearing in the image are described correctly, the scores would be low if the objects do not appear in the ground truth captions. Though the fluency of the generations cannot be evaluated precisely using this score, this provides us with a rough reference. We also evaluate the descriptions whose guiding objects appear in the ground truth labels and the METEOR score is 28.0. This suggests that the generated sentences are fluent when the guiding objects are in-domain for the caption labels.

Figure 5 shows recalls of 7 objects as examples. Comparing with the base model and ground truth caption labels, recall of inconspicuous objects such as "cup" (from 0.15 to 0.69) and "bowl" (from 0.21 to 0.65) can be improved significantly with CGO.

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1
DCC [5]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8
NOC [38]	17.8	68.8	25.6	24.7	69.3	68.1	39.9	89.0	49.1
LSTM-C [44]	29.7	74.4	38.8	27.8	68.2	70.3	44.8	91.4	55.7
CBS+T4 [1]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0
NBT + G [24]	14.0	74.8	42.8	63.7	74.4	19.0	44.5	92.0	53.2
DNOC [42]	33.0	76.9	54.0	46.6	75.8	33.0	59.5	84.6	57.9
CGO (ours)	**45.0**	**79.0**	**69.2**	**64.6**	**87.3**	**89.7**	**75.8**	**95.0**	**75.8**

Table 3. F1 scores of the novel objects on the test split.

	Out-of-Domain Scores				In-Domain Scores		
Model	SPICE	METEOR	CIDEr	Avg. F1	SPICE	METEOR	CIDEr
DCC [5]	13.4	21.0	59.1	39.8	15.9	23.0	77.2
NOC [38]	-	21.4	-	49.1	-	-	-
LSTM-C [44]	-	23.0	-	55.7	-	-	-
CBS + T4 [1]	15.9	23.3	77.9	54.0	18.0	24.5	86.3
NBT + G [24]	16.6	23.9	84.0	53.2	18.4	25.3	94.0
CGO (p_o = 0.5)	17.7	23.9	89.1	**75.8**	18.0	25.1	94.7
CGO (p_o = 0.7)	17.7	23.9	88.2	**75.8**	18.4	25.3	95.8
CGO (p_o = 0.9)	**18.1**	**24.2**	**90.0**	**75.8**	**19.6**	**26.3**	**103.3**

Table 4. Descriptions generated for in-domain and out-of-domain images are evaluated using image caption metrics. p_o is the threshold for selecting guiding objects which are in-domain. When the probability of occurrence of an object predicted by the object classifier exceeds p_o, it is used as the guiding object. We choose the object with the highest probability when more than one object meet the requirement.

4.2. Diverse Descriptions for Each Object

In this part we show CGO's ability to generate diverse descriptions with a fixed guiding object. We randomly select an object as the guiding object from an image and choose $n = 1$ or 2 other objects to form the LSTM-L input sequence. With $n = 1$, the input sequence can be <Guiding object> or <Guiding object, Object1>, 'Object1' denotes the chosen object for the LSTM-L input sequence. With $n = 2$ we test with 3 different input sequences with length 1, 2 and 3.

Results are shown in Table 2. When we use objects chosen from object detection labels, the average number of unique descriptions is 1.48 with 2 different inputs, and 2.62 with 3 different inputs. This shows that CGO can generate different descriptions even with a fixed guiding object. Examples are shown in the right column in Fig. 4.

4.3. Novel Object Captioning

In this part, we demonstrate the effectiveness of CGO when applied to the novel object captioning task. Following [5], 8 object categories, 'bus, bottle, pizza, microwave, couch, suitcase, racket, zebra' are selected as novel objects. At training time, images are excluded from the MSCOCO training set if their caption labels contain the novel objects. Half of the MSCOCO validation set is used as validation set and the other half as the test set. F1 score is used to evaluate the accuracy of containing the novel objects. For each novel object category, if the generated description and the ground truth label contain the object at the same time, it is regarded as true positive. The average F1 score is the macro average across the 8 categories. Image caption evaluating metrics are used to evaluate the quality of the generated sentences, including SPICE [2], METEOR [8] and CIDEr [37]. Descriptions for out-of-domain images (containing a novel object) and in-domain images are evaluated respectively.

Similar with prior work [5, 1], labels for the object classifier is obtained from caption labels. The full training set is used when training the classifier, including the images which contain novel objects. A novel object is used as the guiding object if it appears in the image. We determine whether a novel object appears in an image according to the results from the object classifier. The probability thresholds of using an object as the guiding object are chosen to maximize the F1 score on the validation set. For novel words, we simply replace their word embedding vectors with other in-domain objects under the same super categories, *e.g.* 'bottle' $\rightarrow$ 'cup'. The results are shown in Table 3 and Table 4. Examples of descriptions are shown in Fig. 6.

Comparing to existing models, CGO significantly improves the F1 score of novel objects, with the average F1 score 75.8. In fact, the F1 score of the output directly de-

bus -> car
Ground truth: bus coming down the street from the intersection
Base model: a yellow van is driving down the street
CGO: a bus is driving down the street with a taxi on the back

bottle -> bowl
Ground truth: a pastry meal sitting on a trey next to a bottle of orange soda
Base model: a hot dog with a bun and a drink
CGO: a hot dog and a bottle of soda on a table

couch -> chair
Ground truth: a dog laying on top of a couch looking sleepy
Base model: a dog laying on a bed with a blanket
CGO: a dog laying on a couch with its head on a pillow

racket -> bat
Ground truth: a young woman holding a tennis racquet on a tennis court
Base model: a woman playing tennis on a court with a ball
CGO: a female tennis player holding a racket on a court

pizza -> cake
Ground truth: a cooked and cut pizza topped with red onions
Base model: a plate of food with a sandwich and a salad
CGO: a pizza with meat and vegetables on a plate

microwave -> toaster
Ground truth: a crock pot on top of a microwave on top of a refrigerator
Base model: a white stove top oven sitting on top of a counter
CGO: a microwave sitting on top of a counter

suitcase -> handbag
Ground truth: a cat is sitting on top of a suitcase
Base model: a cat is sitting on a bag on a floor
CGO: a cat sitting on top of a suitcase on a floor

zebra -> giraffe
Ground truth: a zebra and an ostrich standing in a large grassland
Base model: two geese walking in a field with a field
CGO: a zebra and a bird walking through a field

Figure 6. Examples of descriptions generated for out-of-domain objects (in blue). $O_1 \rightarrow O_2$ indicates that we use O_1 as the guiding object which is encoded using the word embedding vector of an in-domain object O_2. Errors are underlined.

pends on the accuracy of the classification results, just like the template-based models. Note that using different RNN models or different CNN features in the language model does not affect the F1 result. On the other hand, CGO takes advantage of the LSTM language model and generates fluent descriptions. METEOR scores are improved to 24.2 for out-of-domain images and 26.3 for in-domain images. We test the CGO with different probability thresholds p_o for in-do-main objects. An in-domain object is used as the guiding object when probability predicated by the object classifier exceed the threshold. The generating process is reduced to the usual left-to-right generation process when the classifier is not certain about the objects contained in the image.

As the object classifier is independent with the language model, using more advanced models such as object detection models might further improve the F1 scores. In CGO approach, we only guarantee one selected object to be mentioned, but this does not affect its practicability. In many scenarios, novel words do not appear intensively and we are

n02483708
Guiding Object: siamang
Base model: a black bear is standing in the grass
CGO: a black siamang is standing in the grass

n02526121
Guiding Object: eel
Base model: a large fish is sitting in a water
CGO: a eel is sitting on a rock in a river

n04423845
Guiding Object: thimble
Base model: a doll doll sitting on a table next to a toy
CGO: a doll with a thimble sitting on a table

n04147183
Guiding Object: schooner
Base model: a group of boats floating on top of a lake
CGO: a schooner is on the water with a flag in the background

Figure 7. Examples of generated captions for ImageNet objects (in blue). With CGO, an object could be involved in descriptions even if it did not appear in the model vocabulary.

allowed to use more than one description for an image in practice. In addition, CGO can be used in conjunction with other methods such as CBS [1], to contain more objects in the outputs while ensuring that the guiding object is contained in the descriptions.

4.4. Descriptions for ImageNet Objects

Similar to previous works [44, 38], we show qualitative results of our method describing the ImageNet [34] objects. Objects which do not appear in the vocabulary mined from the MSCOCO caption labels are novel to the models trained on the MSCOCO. We use the model trained on Karpathy's training split to generate descriptions. Examples are shown in Fig. 7 and more examples of results can be found in Appendix. Similar to the process in subsection 4.3, word embedding vectors of novel objects are replaced with embedding vectors of the seen objects. *e.g.* 'schooner'→'boat'.

5. Conclusion

We present a novel image captioning approach where the sentence generating process starts from a selected guiding object. Our CGO allows us to include a specific object in generated sentences and describe images in a comprehensive and diverse manner.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant Nos. 61701277, 61771288 and the state key development program in 13th Five-Year under Grant No. 2016YFB0801301.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Guided open vocabulary image captioning with constrained beam search. *arXiv preprint arXiv:1612.00576*, 2016.

[2] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016.

[3] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6077–6086, 2018.

[4] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

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

[6] Dai Bo, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision*, 2017.

[7] Moitreya Chatterjee and Alexander G Schwing. Diverse and coherent paragraph generation from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 729–744, 2018.

[8] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[9] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589*, 2018.

[10] Jacob Devlin, Cheng Hao, Fang Hao, Saurabh Gupta, Deng Li, Xiaodong He, Geoffrey Zweig, and Margaret Mitchell. Language models for image captioning: The quirks and what works. *Computer Science*, 2015.

[11] Desmond Elliott and Arjen de Vries. Describing images using inferred visual dependency representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 42–52, 2015.

[12] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh K Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C Platt, et al. From captions to visual concepts and back. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1473–1482, 2015.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[15] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*, pages 5415–5424, 2017.

[16] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4565–4574, 2016.

[17] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015.

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

[20] Girish Kulkarni, Visruth Premraj, Vicente Ordonez, Sagnik Dhar, Siming Li, Yejin Choi, Alexander C Berg, and Tamara L Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12):2891–2903, 2013.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *Proc. IEEE Int. Conf. Comp. Vis*, volume 3, page 3, 2017.

[23] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. 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 (CVPR)*, volume 6, page 2, 2017.

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

[25] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632*, 2014.

[26] Yuzhao Mao, Chang Zhou, Xiaojie Wang, and Ruifan Li. Show and tell more: Topic-oriented multi-sentence image captioning. In *IJCAI*, pages 4258–4264, 2018.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[28] George Miller. *WordNet: An electronic lexical database*. MIT press, 1998.

[29] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *arXiv preprint arXiv:1607.00970*, 2016.

[30] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[31] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015.

[32] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[33] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *CVPR*, volume 1, page 3, 2017.

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[35] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[36] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[37] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

[38] Subhashini Venugopalan, Lisa Anne Hendricks, Marcus Rohrbach, Raymond J Mooney, Trevor Darrell, and Kate Saenko. Captioning images with diverse objects. In *CVPR*, volume 3, page 8, 2017.

[39] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[41] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017.

[42] Yu Wu, Linchao Zhu, Lu Jiang, and Yi Yang. Decoupled novel object captioner. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 1029–1037. ACM, 2018.

[43] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[44] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Incorporating copying mechanism in image captioning for learning novel objects. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5263–5271. IEEE, 2017.

[45] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales. Actor-critic sequence training for image captioning. *arXiv preprint arXiv:1706.09601*, 2017.

[46] Mingxing Zhang, Yang Yang, Hanwang Zhang, Yanli Ji, Heng Tao Shen, and Tat-Seng Chua. More is better: Precise and detailed image captioning using online positive recall and missing concepts mining. *IEEE Transactions on Image Processing*, 28(1):32–44, 2019.