An Object-Extensible Training Framework for Image Captioning

Yike Wu, Ying Zhang(✉), and Xiaojie Yuan

College of Computer Science, Nankai University, Tianjin 300350, China
wuyike@dbis.nankai.edu.cn, {yingzhang,yuanxj}@nankai.edu.cn

Abstract. Recent years have witnessed great progress in image captioning based on deep learning. However, most previous methods are limited to the original training dataset that contains only a fraction of objects in the real world. They lack the ability to describe other objects that are not in the original training dataset. In this paper, we propose an object-extensible training framework that enables a widely-used captioning paradigm to describe objects beyond the original training dataset (i.e., extended objects) by generating high-quality training data for these objects automatically. Specifically, we design a general replacement mechanism, which replaces the object (An object includes the object region in the image, and the corresponding object word in the caption) in the original training dataset with the extended object to generate new training data. The key challenge in the proposed replacement mechanism is that it should be context-aware to get the meaningful result that complies with common knowledge. We introduce the multi-modal context embedding to ensure that the generated object representation is coherent in the visual context and the generated caption is smooth and fluent in the linguistic context. Extensive experiments show that our method improves significantly over the state-of-the-art methods on the held-out MSCOCO in both automatic and human evaluation.

Keywords: Image captioning · Extended objects · Context-aware replacement

1 Introduction

Image captioning is an important task in the intersection between computer vision and natural language processing. We have witnessed much progress in image captioning based on deep learning. However, most previous methods can only describe objects in the original training dataset, but lack the ability to generate captions for

This research is supported by the NSFC-Xinjiang Joint Fund (No. U1903128), NSFC General Technology Joint Fund for Basic Research (No. U1836109, No. U1936206), Natural Science Foundation of Tianjin, China (No. 18ZXZNGX00110, No. 18ZXZNGX00200), and the Fundamental Research Funds for the Central Universities, Nankai University (No. 63211128).

L. Wang et al. (Eds.): NLPCC 2021, LNAI 13028, pp. 798–809, 2021.
https://doi.org/10.1007/978-3-030-88480-2_64

other objects in the real world. For example, if the original training dataset contains the image-text pairs of "giraffe" but not that of "zebra", a caption model built upon it can describe an image with a giraffe but fails to understand one with a zebra.

The key issue lies in the limitation of the original training dataset that is manually constructed and contains only a small fraction of objects in the real world. Supposing we have a "complete" training dataset covering all objects, we can use it to train a caption model that can describe any object. Therefore, to enable a caption model to describe objects not in the original training dataset, a naive solution is to manually construct additional training data for such objects. However, this process is time-consuming and laborious, which hinders its feasibility in realistic applications. A question naturally arises: can we automatically generate training data for such objects without manual efforts?

We find that the UpDn model [2] provides convenience for us to achieve the automatic generation. It represents the input image by object regions instead of a single feature vector [15] or spatial grids [17], which means it does not require direct access to the original image and only uses the object representation instead. Extensive works (e.g., [4,9,13]) follow this captioning paradigm and all use the object representation, which we define as *UpDn-style caption model*. One merit of such models is that it makes generating training data for an object simple. For example, we want to create a new image-text pair of the object "zebra" that is not in the original training dataset. Suppose that we already have an original image with the caption "a giraffe walking across the grass next to some antelope" as shown in Fig. 1. We could simply replace the giraffe region in the object representation of the original image by the zebra region from another unpaired image[1] to generate the object representation for the new image. And we don't need to generate the new image itself, which is a relatively hard task, as the UpDn-style caption model only needs the object representation as input. To generate the corresponding caption for this new image, we can simultaneously do the replacement of the object word "giraffe" in the original caption and get "a zebra walking across the grass next to some antelope".

In this paper, we propose an object-extensible training framework that enables the UpDn-style caption model to describe objects beyond the original training dataset (i.e., *extended objects*) by generating new training data for these objects automatically. Specifically, we introduce a general replacement mechanism which replaces the object region and object word in the original training dataset simultaneously with the object region and object word of an extended object. The generated data can be used to train any UpDn-style caption model as the input of the UpDn-style caption model is the object representation of an image rather than the image itself. The entire process of data generation and model training is automatic and requires no additional manual efforts.

The key challenge in the proposed replacement mechanism is to ensure that the replacement result is meaningful and complies with common knowledge. In the example of Fig. 1, if we replace the "giraffe" region-word pair (i.e., object region

[1] This image is not paired with a caption and easy to obtain without manual efforts.

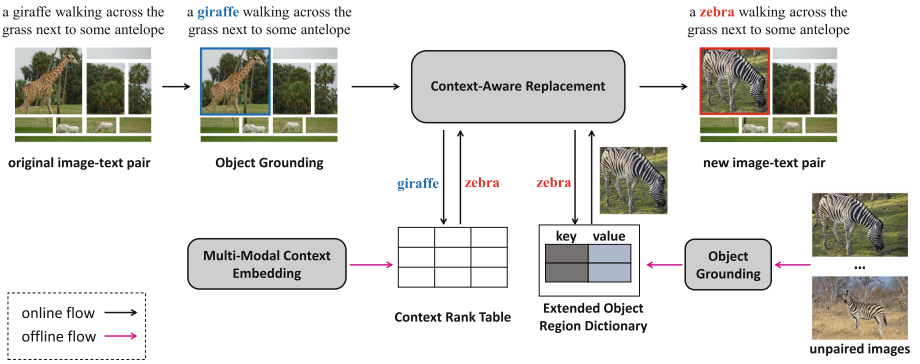

Fig. 1. The framework of the general replacement mechanism.

and object word) with a "bus" region-word pair instead of the "zebra" region-word pair, we will get ridiculous results in two aspects. First, as the bus region is not likely to appear together with grass regions or antelope regions in a real valid image, the resulting object representation is not meaningful. Second, in the caption after replacement, "a bus walking across the grass ...", "bus" does not collocate with "walking" in the natural language. Thus, to ensure that the replacement is meaningful, we need to consider both the visual context of the object region and the linguistic context of the corresponding object word. To this end, we propose the context-aware replacement (CAR), which uses the multi-modal context embedding to find the replacement with the most similar visual context and linguistic context to the given object. In summary, our contributions are three-fold:

- We propose an object-extensible training framework that enables the UpDn-style caption model to describe extended objects by a general replacement mechanism.
- We introduce the multi-modal context embedding to make the replacement process aware of the visual context and linguistic context.
- Extensive experimental results show that the proposed method outperforms the state-of-the-art methods on the held-out MSCOCO dataset.

2 Related Work

In recent years, image captioning methods based on deep learning have made much progress [2,12,13,15,17,19]. However, most of them can only describe the objects in the original training dataset that is manually constructed, and are difficult to be generalized to other objects in the real world.

Some approaches have been proposed to solve this problem. Deep Compositional Captioner (DCC) [3] pretrains a lexical classifier and a language model on unpaired image/text data respectively, and composes them into a caption model. It further trains the caption model on image-text pairs and transfers knowledge between semantically-related words. Venugopalan *et al.* [14] extend

DCC by jointly training the lexical classifier, language model and caption model in an end-to-end manner, which obviates the explicit transfer and achieves better performance. More recently, Yao *et al.* [18] incorporate the copy mechanism into the caption model, which can not only generate a word from the language model but also copy one from objects detected in the image. Li *et al.* [5] further consolidate the method [18] by the pointing mechanism and coverage of objects. In addition, Mogadala *et al.* [8] annotate entity labels for images with the guidance of knowledge base, and build the semantic attention and constrained inference over these entity labels. Another approach [1] proposes the constrained beam search, which forces the visual tags of the image to appear in the generated caption during the inference process. Furthermore, the Decoupled Novel Object Captioner (DNOC) [16] first generates a sentence with placeholders, and then retrieves object words from a key-value object memory to fill them. Neural Baby Talk [7] shares a similar spirit with DNOC, which first generates a sentence with slots tied to object regions in the image, and then fills the slots by the corresponding object words.

Previous works usually design a special model architecture for image captioning to incorporate more objects, which is tightly coupled with the architecture itself and difficultly generalized. In contrast, our solution tackles the problem in a data-driven way, which is fully compatible with any UpDn-style caption model and thus can seamlessly benefit from its potential improvement.

3 Methodology

3.1 Framework Overview

The general replacement mechanism is shown in Fig. 1, which is composed of the online flow and the offline flow. Given an image-text pair $(\boldsymbol{R}, S)$ in the original training dataset $\mathcal{D}_o$, we feed it into the online flow to get a new image-text pair $(\boldsymbol{R}', S')$. We perform this procedure on all image-text pairs in $\mathcal{D}_o$ to obtain an extended training dataset $\mathcal{D}_e$, which contains not only objects in $\mathcal{D}_o$ but also the extended objects. Finally, we use $\mathcal{D}_e$ to train a caption model that can generate captions for all the objects in $W_{obj} \cup W_{ext}$, where W_{obj} and W_{ext} denote the vocabulary of objects in $\mathcal{D}_o$ and that of extended objects respectively.

Online Flow. The input is an image-text pair $(\boldsymbol{R}, S)$ in the original training dataset $\mathcal{D}_o$. The symbol $\boldsymbol{R} = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_o^*, ..., \boldsymbol{r}_M\}$ is the object representation of an image and $S = \{w_1, w_2, ..., w_o^*, ..., w_N\}$ is the corresponding caption, where $\boldsymbol{r}$ and w denote an object region and a word respectively. First, from the input we extract the object word $w_o^* \in W_{obj}$ and identify its corresponding object region $\boldsymbol{r}_o^*$ via the object grounding. Then, we replace the region-word pair $(\boldsymbol{r}_o^*, w_o^*)$ by a new pair $(\boldsymbol{r}_e^*, w_e^*)$ of an extended object through the context-aware replacement. Finally, the online flow outputs a new image-text pair $(\boldsymbol{R}', S')$ for the extended object, where $\boldsymbol{R}' = \{\boldsymbol{r}_1, \boldsymbol{r}_2, ..., \boldsymbol{r}_e^*, ..., \boldsymbol{r}_M\}$ and $S' = \{w_1, w_2, ..., w_e^*, ..., w_N\}$.

Offline Flow. Before the data generation of online flow, we offline construct two data structures leveraged by the context-aware replacement: (1) We build

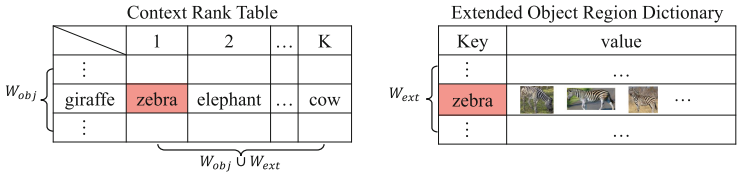

Fig. 2. The details of two data structures constructed by the offline flow.

the *context rank table* with the multi-modal context embedding, which will be used to find the extended object word w_e^* with the most similar visual and linguistic context to w_o^*. (2) We generate the *extended object region dictionary* by the object grounding, which will be queried with w_e^* as the key to find the corresponding object region $\boldsymbol{r}_e^*$.

Caption Model. We employ the UpDn model [2] as a representative of UpDn-style caption models to verify the effectiveness of our method. The model details are elaborated in the previous work [2] and we will not go into them since our focus is the proposed framework in this work.

3.2 Multi-modal Context Embedding

We construct the context rank table based on the similarity of visual and linguistic context between object words, which is measured by the cosine similarity of their multi-modal context embeddings. The structure of context rank table is shown in Fig. 2. Each row corresponds to an object word in W_{obj}, and each column corresponds to a rank value which is assigned to an object word in $W_{obj} \cup W_{ext}$. In the corresponding row of an object word $w_o \in W_{obj}$, we rank each object word in $W_{obj} \cup W_{ext}$ from high to low according to the cosine similarity between its multi-modal context embedding and that of w_o, and only keep the top K rank values to ensure that the object words in the row are similar enough to w_o in both visual and linguistic context.

Now we focus on how to obtain the multi-modal context embedding of an object word. The general idea is to align the visual representation of the object region and the linguistic representation of the corresponding object word in a common latent space. We train a model composed of an object detector, a visual MLP layer f_{vis}, a linguistic MLP layer f_{lin}, and an embedding matrix E initialized with the pretrained GloVe embedding [10]. The input of the model is an image with its object labels $L = \{l\}$, which is composed of the corresponding object words of objects contained in the image. In the training process, we first extract the object representation $\boldsymbol{R}$ of the image by an object detector, and map each object region $\boldsymbol{r} \in \boldsymbol{R}$ into the common latent space by the layer f_{vis}. Then, we also map the corresponding object word (i.e., each label $l \in L$) into the common latent space by applying the layer f_{lin} on its embedding in E. Next, we define the score function which measures how likely the object region $\boldsymbol{r}$ is to contain the label l as follows:

$$\text{sc}(\boldsymbol{r}, l) = \cos\_\text{sim}(f_{\text{vis}}(\boldsymbol{r}), f_{\text{lin}}(E(l))), \tag{1}$$

where cos_sim means cosine similarity. Furthermore, for the entire image $\boldsymbol{R}$, the score function of containing the label l is defined as:

$$\mathrm{sc}(\boldsymbol{R}, l) = \max(\mathrm{sc}(\boldsymbol{r}, l)), \quad \boldsymbol{r} \in \boldsymbol{R}. \tag{2}$$

The greater value of $\mathrm{sc}(\boldsymbol{R}, l)$ means the image $\boldsymbol{R}$ is more likely to contain the label l and vice versa. Finally, we define the training loss on the image $\boldsymbol{R}$:

$$\mathcal{L}(\boldsymbol{R}) = \sum_{l \in L} \sum_{l' \in \{L_U - L\}} \max[0, 0.1 - \mathrm{sc}(\boldsymbol{R}, l) + \mathrm{sc}(\boldsymbol{R}, l')], \tag{3}$$

where L_U denotes the complete set of labels of all images in the training data, and l' is a label that does not appear in the image $\boldsymbol{R}$. Minimizing $\mathcal{L}(\boldsymbol{R})$ is equivalent to increasing $\mathrm{sc}(\boldsymbol{R}, l)$ and decreasing $\mathrm{sc}(\boldsymbol{R}, l')$ simultaneously, which forces the labels in L to approach the image $\boldsymbol{R}$ and keep other labels in $\{L_U - L\}$ away from it in the common latent space. After training, we use $f_{\mathrm{lin}}(E(l))$ as the multi-modal context embedding of l, which is the projection of l in the common latent space.

The cosine similarity of the multi-modal context embeddings can reflect the similarity of object words in both visual and linguistic context. On the one hand, the training loss $\mathcal{L}(\boldsymbol{R})$ makes the labels in similar images (i.e., with a similar visual context) close to each other in the common latent space. On the other hand, we have already incorporated the linguistic context into the training process at the beginning by initializing E with the pretrained GloVe embedding.

3.3 Object Grounding

The object grounding module grounds an object word to its corresponding object region in the image. In the proposed method, we leverage this module to (1) ground the object word w_o^* to its corresponding object region $\boldsymbol{r}_o^*$ in the original image and (2) build the extended object region dictionary. Next, we elaborate how we achieve the above two goals respectively.

Ground w_o^* to $\boldsymbol{r}_o^*$. Given the object word w_o^*, we explore two kinds of strategies to find its corresponding object region $\boldsymbol{r}_o^*$ in the object representation $\boldsymbol{R}$ of the original image. The first kind of strategy requires ground-truth bounding boxes of the image that are manually annotated. We first pick out the ground-truth bounding boxes with the object category corresponding to w_o^*, denoted as $\boldsymbol{B}$, and then identify $\boldsymbol{r}_o^*$ as follows:

$$\boldsymbol{r}_o^* = \{\boldsymbol{r} \in \boldsymbol{R} | \mathrm{IoU}(\boldsymbol{r}, \boldsymbol{b}) > T\}, \quad (\boldsymbol{r}, \boldsymbol{b}) \in \boldsymbol{R} \times \boldsymbol{B}, \tag{4}$$

where $T \in [0.0, 1.0]$ is the threshold value of IoU. The second kind of strategy requires no manual efforts and is more general. It leverages the object categories of object regions in $\boldsymbol{R}$, which are output by the object detector. Specifically, we regard all the object regions in $\boldsymbol{R}$ with the object category corresponding to w_o^* as the object region $\boldsymbol{r}_o^*$.

Build the Extended Object Region Dictionary. The structure of the extended object region dictionary is shown in Fig. 2. Each key corresponds to an extended object word $w_e \in W_{ext}$, and its value consists of a series of object regions corresponding to w_e from different unpaired images. Each object region $\boldsymbol{r}_e$ in the value is obtained by grounding w_e in the object representation $\boldsymbol{R}$ of an unpaired image.

There are also two kinds of strategies for grounding w_e to $\boldsymbol{r}_e$. The first needs ground-truth bounding boxes of the image, while the second leverages the object categories with their confidence scores output by the object detector. In the first kind of strategy, we identify $\boldsymbol{r}_e$ as follows:

$$\boldsymbol{r}_e = \underset{\boldsymbol{r} \in \boldsymbol{R}}{\operatorname{argmax}} \operatorname{IoU}(\boldsymbol{r}, \boldsymbol{b}), \quad (\boldsymbol{r}, \boldsymbol{b}) \in \boldsymbol{R} \times \boldsymbol{B}, \tag{5}$$

where $\boldsymbol{B}$ denotes the ground-truth bounding boxes with the object category corresponding to w_e. In the second kind of strategy, we first find out all the object regions with the object category corresponding to w_e in the object representation $\boldsymbol{R}$, and then pick the one with the highest confidence score as $\boldsymbol{r}_e$.

Note that the grounding of the object word w_o^* and that of the extended object words in W_{ext} are slightly different, which makes the replacement more precise and thus improves the quality of the new generated image-text pair. When grounding w_o^* to $\boldsymbol{r}_o^*$, we adopt a relatively loose screening condition to find out all the object regions possibly corresponding to w_o^* in the original image, which means that the notation $\boldsymbol{r}_o^*$ may represent multiple object regions instead of only one. Since the object detector may output multiple object regions that largely overlap for the same object in an image, this loosely grounding can guarantee all of them can be completely removed in the replacement. When building the extended object region dictionary, we ground each extended object word $w_e \in W_{ext}$ to only the most accurate object region in an unpaired image. In this way, when we replace $\boldsymbol{r}_o^*$ with $\boldsymbol{r}_e^*$, we guarantee the object region $\boldsymbol{r}_e^*$ added into the object representation exactly contains the extended object.

3.4 Context-Aware Replacement for Automatic Data Generation

Given an image-text pair from the original training dataset $\mathcal{D}_o$, we generate a new image-text pair of the extended object in the context-aware replacement. We replace the object word w_o^* in the original caption by the extended object word w_e^*, and replace the object region $\boldsymbol{r}_o^*$ corresponding to w_o^* in the object representation of the original image by the extended object region $\boldsymbol{r}_e^*$ corresponding to w_e^*. We describe the context-aware replacement in Algorithm 1.

To ensure the replacement result is meaningful, we need to find the region-word pair $(\boldsymbol{r}_e^*, w_e^*)$ with the most similar context to the region-word pair $(\boldsymbol{r}_o^*, w_o^*)$. First, we extract the corresponding row of w_o^* from the context rank table, and select the most top-ranked extended object word in the row as w_e^*. Then, we take w_e^* as the key to retrieve its corresponding value from the extended object region dictionary, and randomly select an object region as $\boldsymbol{r}_e^*$ from a series of object regions in the value. Note that we do not perform the replacement if there

Algorithm 1. Context-Aware Replacement (CAR)

Input:
An image-text pair $(\boldsymbol{R}, S)$ containing a region-word pair $(\boldsymbol{r}_o^*, w_o^*)$;
Context rank table CRT;
Extended object region dictionary EORD.
Output:
A new image-text pair $(\boldsymbol{R}', S')$.
$\mathrm{CRT}(w_o^*) \leftarrow$ corresponding row of w_o^* in CRT
if $W_{ext} \cap \mathrm{CRT}(w_o^*) \neq \emptyset$ **then**
    $w_e^* \leftarrow$ top-ranked element in $W_{ext} \cap \mathrm{CRT}(w_o^*)$
    $\mathrm{EORD}(w_e^*) \leftarrow$ the value of key w_e^* in EORD
    $\boldsymbol{r}_e^* \leftarrow$ retrieve an object region from $\mathrm{EORD}(w_e^*)$
    $S' \leftarrow$ in S, replace w_o^* by w_e^*
    $\boldsymbol{R}' \leftarrow$ in $\boldsymbol{R}$, replace $\boldsymbol{r}_o^*$ by $\boldsymbol{r}_e^*$
    return $(\boldsymbol{R}', S')$
else
    do not perform the replacement
end if

is no extended object word in the corresponding row of w_o^* in the context rank table, which means we can not find a replacement similar enough to the object in the original image-text pair in both visual and linguistic context.

For each image-text pair in $\mathcal{D}_o$, we perform the context-aware replacement to generate a new image-text pair. Finally, we gather all the new image-text pairs, and combine them with $\mathcal{D}_o$ to obtain an extended training dataset $\mathcal{D}_e$. Comparing with training on $\mathcal{D}_o$, the additional computation cost of training on $\mathcal{D}_e$ is empirically sub-linear, since each image-text pair in $\mathcal{D}_o$ yields at most one new image-text pair (sometimes the replacement will not be successfully performed as mentioned above). This indicates that our method can scale up on datasets with different sizes.

4 Experiments

4.1 Experimental Setup

Dataset. For the convenience of comparing with previous works, we evaluate our method on the held-out MSCOCO dataset, a widely-used benchmark [3] for image captioning on objects not in the original training dataset. The dataset consists of four splits: *training*, *validation*, *test* and *rest*. Follow the previous work [3], we employ a subset of MSCOCO [6] training set as the training split, which excludes all the image-text pairs containing at least one of the eight objects: *bottle*, *bus*, *couch*, *microwave*, *pizza*, *racket*, *suitcase*, *zebra*. The eight objects are used as the extended objects in this setting. We use 50% of MSCOCO validation set as the validation split, and set aside the other 50% for the test split. We take the excluded part in MSCOCO training set as the rest split.

Table 1. Performance (%) on held-out MSCOCO test split.

Model	bottle	bus	couch	microwave	pizza	racket	suitcase	zebra	Avg. F1	CIDEr	METEOR	SPICE
DCC [3]	4.6	29.8	45.9	28.1	64.6	52.2	13.2	79.9	39.8	59.1	21.0	13.4
NOC [14]	14.9	69.0	43.8	37.9	66.5	65.9	28.1	88.7	51.8	–	20.7	–
Base+T4 [1]	16.3	67.8	48.2	29.7	77.2	57.1	49.9	85.7	54.0	77.9	23.3	15.9
KGA-CGM [8]	26.4	54.2	42.1	50.9	70.8	75.3	25.6	90.7	54.5	–	22.2	14.6
LSTM-C [18]	29.7	74.4	38.8	27.8	68.2	70.3	44.8	91.4	55.7	–	23.0	–
DNOC [16]	33.0	76.9	54.0	46.6	75.8	33.0	59.5	84.6	57.9	–	21.6	–
NBT [7]	14.0	74.8	42.8	63.7	74.4	19.0	44.5	92.0	53.2	84.0	23.9	16.6
LSTM-P [5]	28.7	75.5	47.1	51.5	81.9	47.1	62.6	93.0	60.9	88.3	23.4	16.6
CAR	29.4	75.7	49.7	56.0	73.5	18.7	50.3	94.4	56.0	**101.9**	**26.1**	**19.3**
CAR + T2	37.4	78.5	52.2	58.7	76.6	39.2	56.1	94.5	**61.7**	100.1	25.8	19.2

In the experiment, we use the training split as the original training dataset $\mathcal{D}_o$ (351134 image-text pairs), and generate 302179 new image-text pairs to obtain the extended training dataset $\mathcal{D}_e$ (653313 image-text pairs). We perform the validation and testing on the corresponding splits respectively. There is no overlap of data between model training and evaluation.

Evaluation. On the one hand, we evaluate the captioning performance on automatic metrics. On the other hand, We also compute F1-score for the eight extended objects respectively. For an image in the test split, we regard it as a true positive example of an extended object only if its generated caption and at least one of its ground-truth captions both mention the object.

Implementation Details. For each image, we take a pretrained Faster R-CNN [11] as the object detector to extract 36 object regions as its object representation. This is aligned with the strong baselines DNOC and NBT which also use the Faster R-CNN feature. Additionally, considering the generality, we perform object grounding based on the output of Faster R-CNN, instead of ground-truth bounding boxes (We discuss the difference in Sect. 4.5). In the context rank table, we set the value of K to 20.

4.2 Comparison with SOTA Methods

We compare our method context-aware replacement (abbreviated as CAR) with state-of-the-art methods in Table 1. We can see that our method CAR achieves comparable average F1-score (Avg. F1) of extended objects compared to the SOTA methods, which shows that our approach successfully generates captions for extended objects. By using the constrained beam search [1] (CAR + T2), our method achieves the best average F1-score (61.7%) while maintaining decent captioning performance. However, the results on F1-score can only reflect that the generated caption correctly mentions the corresponding object word of the extend object that appears in the image. We should also focus on the overall captioning performance. We observe that CAR significantly outperforms all the SOTA methods on automatic metrics. Particularly, CAR improves over the competitive baseline LSTM-P by 13.6% on CIDEr, 2.7% on METEOR and 2.7% on SPICE. This indicates that the new generated training data is high-quality enough for training a caption model to generate natural and fluent captions.

Table 2. Human evaluation (%) on a sampled subset of held-out MSCOCO test split. The notation "both" means the judgement holds in both criteria.

Judgement	CAR vs. UpDn			CAR vs. NBT		
	object coverage	consistency	both	object coverage	consistency	both
CAR is better	**69.67 ± 0.02**	**43.00 ± 0.03**	**38.33 ± 0.06**	**46.33 ± 0.06**	**37.67 ± 0.11**	**23.00 ± 0.02**
UpDn/NBT is better	9.00 ± 0.09	25.33 ± 0.00	7.00 ± 0.09	24.33 ± 0.16	35.67 ± 0.11	16.67 ± 0.04
two models are equal	21.33 ± 0.10	29.33 ± 0.04	13.67 ± 0.11	31.67 ± 0.02	26.67 ± 0.39	13.00 ± 0.08

4.3 Human Evaluation

To complement the automatic metrics, we re-implement the strong baseline NBT [7], and perform the human evaluation on a sampled subset of the held-out MSCOCO test split to compare our method CAR with it. We also take the UpDn model [2] for comparison. For each image, we generate three captions with the compared models respectively, and randomly shuffle them to avoid potential bias. We ask three human evaluators to compare the generated captions in pair.

Evaluation Criteria. Given two captions generated by different models for the same image, the evaluators make a judgement about which one is better in two aspects respectively. The first is *object coverage*. This criterion reflects how well the caption covers the objects in the image. If the image contains an extended object, we also tell the evaluators to focus more on it. The second is *consistency*. It measures how consistent the caption is with the image content.

Evaluation Results. We report the results of human evaluation in Table 2. Comparing with both UpDn and NBT, our method CAR generates more captions which are better on either object coverage or consistency. Considering the two criteria simultaneously, our approach also outperforms the other methods.

4.4 Qualitative Examples

As shown in Fig. 3, our method CAR describes the extended objects in all the examples while the other methods not, which verifies its effectiveness of incorporating the extended objects into the caption generation. The red bounding box in an image indicates the object region with the largest attention weight when CAR generate the highlighted word. We observe that red bounding boxes fit well with the extended objects in the images, which reflects that our method really learns to ground the extended objects in the images correctly.

4.5 Discussion

Ablation Study. We compare our method CAR with its two variants in Table 3a: 1) UpDn [2]. It represents an UpDn-style caption model which is trained only on the original training dataset. 2) General Replacement. Besides the original training dataset, it also generates new training data for extended objects by the proposed replacement mechanism to assist the model training. However, it

Table 3. Discussion on the ablation study of our approach CAR and the effectiveness of using the ground-truth bounding boxes.

(a) Ablation study to demonstrate contributions from "replacement (R)" and "context-aware (CA)" in CAR.

Model	R	CA	Avg. F1
UpDn			0.0
General Replacement	✓		48.4
CAR	✓	✓	56.0

(b) Performance on held-out MSCOCO test split without/with leveraging the ground-truth bounding boxes.

Model	Avg. F1	CIDEr	METEOR	SPICE
CAR	56.0	101.9	26.1	19.3
CAR + bbox	56.6	102.2	26.3	19.4

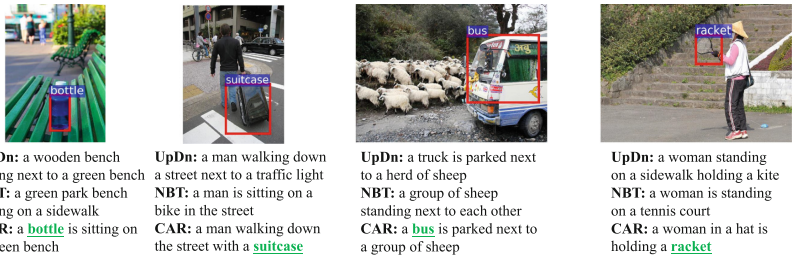

Fig. 3. Qualitative examples of captions generated by different methods.

does not consider the visual context and linguistic context, and just randomly selects an extended object as the replacement.

First, by adding the "replacement (R)", general replacement performs much better than UpDn on average F1-score, while UpDn cannot generate captions for any extended object (Avg. F1 is 0.0%). This validates the effectiveness of the proposed replacement mechanism on describing extended objects. Second, by further adding the "context-aware (CA)", CAR increases 7.6% on average F1-score. This indicates that it is necessary to ensure that the replacement result is meaningful and complies with common knowledge, which improves the quality of generated training data and thus is beneficial to model training.

Using Ground-Truth Bounding Boxes. As shown in Table 3, the performance of our method is further boosted by leveraging the ground-truth bounding boxes (CAR + bbox) to perform the object grounding. This is reasonable since better grounding will lead to more precise replacement and thus improve the quality of generated training data.

5 Conclusion

In this paper, we propose an object-extensible training framework based on a general replacement mechanism, which focuses on the training data generation of extended objects and is compatible with any UpDn-style caption model. It paves a new data-driven way to generate captions for extended objects. To ensure that the generated data is meaningful and complies with common knowledge,

we introduce the multi-modal context embedding to make the replacement process aware of both visual context and linguistic context. It guarantees that the generated object representation is coherent in visual context and the generated caption is smooth and fluent in linguistic context. Extensive experiments conducted on held-out MSCOCO shows that our method outperforms the SOTA methods in both automatic and human evaluation.

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Guided open vocabulary image captioning with constrained beam search. In: EMNLP, pp. 936–945 (2017)
2. Anderson, P., et al.: Bottom-up and top-down attention for image captioning and visual question answering. In: CVPR, pp. 6077–6086 (2018)
3. Anne Hendricks, L., Venugopalan, S., Rohrbach, M., Mooney, R., Saenko, K., Darrell, T.: Deep compositional captioning: describing novel object categories without paired training data. In: CVPR, pp. 1–10 (2016)
4. Huang, L., Wang, W., Chen, J., Wei, X.Y.: Attention on attention for image captioning. In: ICCV, pp. 4634–4643 (2019)
5. Li, Y., Yao, T., Pan, Y., Chao, H., Mei, T.: Pointing novel objects in image captioning. In: CVPR, pp. 12497–12506 (2019)
6. Lin, T.-Y., et al.: Microsoft COCO: common objects in context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48
7. Lu, J., Yang, J., Batra, D., Parikh, D.: Neural baby talk. In: CVPR, pp. 7219–7228 (2018)
8. Mogadala, A., Bista, U., Xie, L., Rettinger, A.: Describing natural images containing novel objects with knowledge guided assitance. In: ACM Multimedia (2017)
9. Pan, Y., Yao, T., Li, Y., Mei, T.: X-linear attention networks for image captioning. In: CVPR, pp. 10971–10980 (2020)
10. Pennington, J., Socher, R., Manning, C.D.: Glove: global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
11. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: towards real-time object detection with region proposal networks. In: NeurIPS, pp. 91–99 (2015)
12. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: CVPR, pp. 7008–7024 (2017)
13. Shi, Z., Zhou, X., Qiu, X., Zhu, X.: Improving image captioning with better use of caption. In: ACL, pp. 7454–7464 (2020)
14. Venugopalan, S., Anne Hendricks, L., Rohrbach, M., Mooney, R., Darrell, T., Saenko, K.: Captioning images with diverse objects. In: CVPR, pp. 5753–5761 (2017)
15. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: a neural image caption generator. In: CVPR, pp. 3156–3164 (2015)
16. Wu, Y., Zhu, L., Jiang, L., Yang, Y.: Decoupled novel object captioner. In: ACM Multimedia, pp. 1029–1037 (2018)
17. Xu, K., et al.: Show, attend and tell: Neural image caption generation with visual attention. In: ICML, pp. 2048–2057 (2015)
18. Yao, T., Pan, Y., Li, Y., Mei, T.: Incorporating copying mechanism in image captioning for learning novel objects. In: CVPR, pp. 6580–6588 (2017)
19. Zhao, S., Sharma, P., Levinboim, T., Soricut, R.: Informative image captioning with external sources of information. In: ACL, pp. 6485–6494 (2019)