This ECCV 2018 paper, provided here by the Computer Vision Foundation, is the author-created version.
The content of this paper is identical to the content of the officially published ECCV 2018
LNCS version of the paper as available on SpringerLink: https://link.springer.com/conference/eccv

Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data

Xihui Liu[1][0000−0003−1831−9952], Hongsheng Li†[1][0000−0002−2664−7975], Jing Shao[2][0000−0003−3521−6744], Dapeng Chen[1][0000−0003−2490−1703], and Xiaogang Wang[1]

[1]The Chinese University of Hong Kong [2]SenseTime Research
{xihui-liu@link.,hsli@ee.,dpchen@,xgwang@ee.}cuhk.edu.hk
shaojing@sensetime.com

Abstract. The aim of image captioning is to generate captions by machine to describe image contents. Despite many efforts, generating discriminative captions for images remains non-trivial. Most traditional approaches imitate the language structure patterns, thus tend to fall into a stereotype of replicating frequent phrases or sentences and neglect unique aspects of each image. In this work, we propose an image captioning framework with a self-retrieval module as training guidance, which encourages generating discriminative captions. It brings unique advantages: (1) the self-retrieval guidance can act as a metric and an evaluator of caption discriminativeness to assure the quality of generated captions. (2) The correspondence between generated captions and images are naturally incorporated in the generation process without human annotations, and hence our approach could utilize a large amount of unlabeled images to boost captioning performance with no additional annotations. We demonstrate the effectiveness of the proposed retrieval-guided method on COCO and Flickr30k captioning datasets, and show its superior captioning performance with more discriminative captions.

Keywords: image captioning, language and vision, text-image retrieval

1 Introduction

Image captioning, generating natural language description for a given image, is a crucial task that has drawn remarkable attention in the field of vision and language [2, 5, 14, 21, 22, 26, 35, 41, 43, 47, 49]. However, results by existing image captioning methods tend to be generic and templated. For example, in Fig. 1, although for humans there are non-neglectable differences between the first and second images, the captioning model gives identical ambiguous descriptions "A vase with flowers sitting on a table", while the ground-truth captions contain details and clearly show the differences between those images. Moreover, about fifty percent of the captions generated by conventional captioning methods are exactly the same as ground-truth captions from the training set, indicating that

† Hongsheng Li is the corresponding author.

Fig. 1. Examples of generated captions by conventional captioning models. The generated captions are templated and generic.

the captioning models only learn a stereotype of sentences and phrases in the training set, and have limited ability of generating discriminative captions. The image on the right part of Fig. 1 shows that although the bird is standing on a mirror, the captioning model generates the caption "A bird is sitting on top of a bird feeder", as a result of replicating patterns appeared in the training set.

Existing studies working on the aforementioned problems either used Generative Adversarial Networks (GAN) to generate human-like descriptions [8,36], or focused on enlarging the diversity of generated captions [40,42,44]. Those methods improve the diversity of generated captions but sacrifice overall performance on standard evaluation criteria. Another work [38] generates discriminative captions for an image in context of other semantically similar images by an inference technique on both target images and distractor images, which cannot be applied to generic captioning where distractor images are not provided.

In this study, we wish to show that with the innovative model design, both the discriminativeness and fidelity can be effectively improved for caption generation. It is achieved by involving a self-retrieval module to train a captioning module, motivated from two aspects: (1) the discriminativeness of a caption can be evaluated by how well it can distinguish its corresponding image from other images. This criterion can be introduced as a guidance for training, and thus encourages discriminative captions. (2) Image captioning and text-to-image retrieval can be viewed as dual tasks. Image captioning generates a description of a given image, while text-to-image retrieval retrieves back the image based on the generated caption. Specifically, the model consists of a **Captioning Module** and a **Self-retrieval Module**. The captioning module generates captions based on given images, while the self-retrieval module conducts text-to-image retrieval, trying to retrieve corresponding images based on the generated captions. It acts as an evaluator to measure the quality of captions and encourages the model to generate discriminative captions. Since generating each word of a caption contains non-differentiable operations, we take the negative retrieval loss as self-retrieval reward and adopt REINFORCE algorithm to compute gradients.

Such retrieval-guided captioning framework can not only guarantee the discriminativeness of captions, but also readily obtain benefits from additional unlabeled images, since a caption naturally corresponds to the image it is generated

from, and do not need laborious annotations. In detail, for unlabeled images, only self-retrieval module is used to calculate reward, while for labeled images, both the ground-truth captions and self-retrieval module are used to calculate reward and optimize the captioning model. Mining moderately hard negative samples from unlabeled data further boost both the fidelity and discriminativeness of image captioning.

We test our approach on two image captioning datasets, COCO [6] and Flickr30k [51], in fully-supervised and semi-supervised settings. Our approach achieves state-of-the-art performance and additional unlabeled data could further boost the captioning performance. Analysis of captions generated by our model shows that the generated captions are more discriminative and achieve higher self-retrieval performance than conventional methods.

2 Related Work

Image captioning methods can be divided into three categories [49]. **Template-based methods** [20, 29, 48] generate captions based on language templates. **Search-based methods** [11, 13] search for the most semantically similar captions from a sentence pool. Recent works mainly focus on **language-based methods** with an encoder-decoder framework [7,14–17,28,41,43,46,47], where a convolutional neural network (CNN) encodes images into visual features, and an Long Short Term Memory network (LSTM) decodes features into sentences [41]. It has been shown that attention mechanisms [5, 26, 31, 47] and high-level attributes and concepts [14, 16, 49, 50] can help with image captioning.

Maximum Likelihood Estimation (MLE) was adopted for training by many previous works. It maximizes the conditional likelihood of the next word conditioned on previous words. However, it leads to the exposure bias problem [33], and the training objective does not match evaluation metrics. Training image captioning models by reinforcement learning techniques [37] solves those problems [24, 34, 35] and significantly improves captioning performance.

A problem of current image captioning models is that they tend to replicate phrases and sentences seen in the training set, and most generated captions follow certain templated patterns. Many recent works aim at increasing diversity of generated captions [40, 42, 44]. Generative adversarial networks (GAN) can be incorporated into captioning models to generate diverse and human-like captions [8, 36]. Dai *et al*. [9] proposed a contrastive learning technique to generate distinctive captions while maintaining the overall quality of generated captions. Vedantam *et al*. [38] introduced an inference technique to produce discriminative context-aware image captions using generic context-agnostic training data, but with a different problem setting from ours. It requires context information, *i.e*., a distractor class or a distractor image, for inference, which is not easy to obtain in generic image captioning applications.

In this work, we improve discriminativeness of captions by using a self-retrieval module to explicitly encourage generating discriminative captions during training. Based on the intuition that a discriminative caption should be able

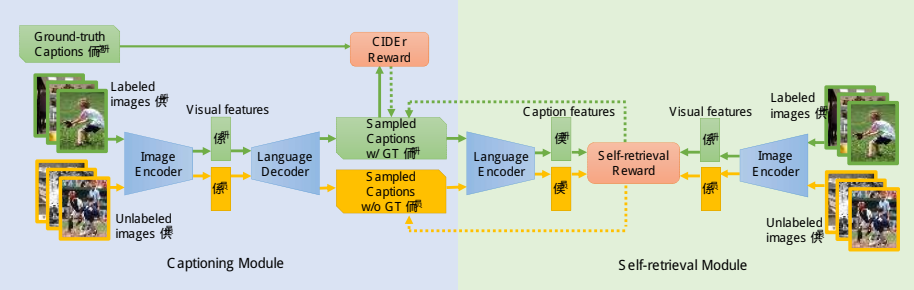

Fig. 2. Overall Framework of our proposed method. The captioning module (left) and the self-retrieval module (right) shares the same image encoder. Dotted lines mean that the reward for each sampled caption is back-propagated by REINFORCE algorithm. The captioning performance is improved by training the captioning module with text-to-image self-retrieval reward. Unlabeled images are naturally handled by our framework.

to successfully retrieve back the image corresponding to itself, the self-retrieval module performs text-to-image retrieval with the generated captions, serving as an evaluator of the captioning module. The retrieval reward for generated captions is back-propagated by REINFORCE algorithm. Our model can also be trained with partially labeled data to boost the performance. A concurrent work [27] by Luo *et al.* also uses a discriminability objective similar to that of ours to generate discriminative captions. However, our work differs from it in utilizing unlabeled image data and mining moderately hard negative samples to further encourage discriminative captions.

3 Methodology

Given an image I, the goal of image captioning is to generate a caption $C = \{w_1, w_2, \ldots, w_T\}$, where w_i denotes the ith word, and we denote the ground-truth captions by $C^* = \{w_1^*, w_2^*, \ldots, w_T^*\}$.

The overall framework, as shown in Fig. 2, comprises of a captioning module and a self-retrieval module. The captioning module generates captions for given images. A Convolutional Neural Network (CNN) encodes images to visual features, and then a Long Short Term Memory network (LSTM) decodes a sequence of words based on the visual features. The self-retrieval module is our key contribution, which is able to boost the performance of the captioning module with only partially labeled images. It first evaluates the similarities between generated captions with their corresponding input images and other distractor images. If the captioning module is able to generate discriminative enough descriptions, the similarity between the corresponding generated-caption-image pairs should be higher than those of non-corresponding pairs. Such constraint is modeled as a text-to-image retrieval loss and is back-propagated to the improve the captioning module by REINFORCE algorithm.

3.1 Image Captioning with Self-retrieval Reward

Captioning module. The captioning module, aiming at generating captions for given images, is composed of a CNN image encoder $E_i(I)$ and an LSTM language decoder $D_c(v)$. The image encoder E_i encodes an image I to obtain its visual features v, and the language decoder D_c decodes the visual features v to generate a caption C that describes the contents of the image,

$$v = E_i(I), \quad C = D_c(v). \tag{1}$$

For conventional training by maximum-likelihood estimation (MLE), given the ground-truth caption words up to time step $t-1$, $\{w_1^*, \ldots, w_{t-1}^*\}$, the model is trained to maximize the likelihood of w_t^*, the ground-truth word of time step t. Specifically, the LSTM outputs probability distribution of the word at time step t, given the visual features and ground-truth words up to time step $t-1$, and is optimized with the cross-entropy loss,

$$L_{CE}(\theta) = -\sum_{t=1}^{T} \log(p_\theta(w_t^* | v, w_1^*, \ldots, w_{t-1}^*)), \tag{2}$$

where θ represents learnable weights of the captioning model.

For inference, since the ground-truth captions are not available, the model outputs the distribution of each word conditioned on previous generated words and visual features, $p_\theta(w_t | v, w_1, \ldots, w_{t-1})$. The word at each time step t is chosen based on the probability distribution of each word by greedy decoding or beam search.

Self-retrieval module. A captioning model trained by MLE training often tends to imitate the word-by-word patterns in the training set. A common problem of conventional captioning models is that many captions are templated and generic descriptions (*e.g.* "A woman is standing on a beach"). Reinforcement learning with evaluation metrics (such as CIDEr) as reward [24, 35] allows the captioning model to explore more possibilities in the sample space and gives a better supervision signal compared to MLE. However, the constraint that different images should not generate the same generic captions is still not taken into account explicitly. Intuitively, a good caption with rich details, such as "A woman in a blue dress is walking on the beach with a black dog", should be able to distinguish the corresponding image in context of other distractor images. To encourage such discriminative captions, we introduce the self-retrieval module to enforce the constraint that the generated captions should match its corresponding images better than other images.

We therefore model the self-retrieval module to conduct text-to-image retrieval with the generated caption as a query. Since retrieving images from the whole dataset for each generated caption is time-consuming and infeasible during each training iteration, we consider text-to-image matching in each mini-batch.

We first encode images and captions into features in the same embedding space a CNN encoder E_i and a Gated Recurrent Unit (GRU) encoder E_c for captions,

$$v = E_i(I), \quad c = E_c(C), \tag{3}$$

where I and C denote images and captions, and v and c denote visual features and caption features, respectively. Then the similarities between the embedded image features and caption features are calculated. The similarities between the features of a caption c_i and the features of the jth image v_j is denoted as $s(c_i, v_j)$. For a mini-batch of images $\{I_1, I_2, \cdots, I_n\}$ and a generated caption C_i of the ith image, we adopt the triplet ranking loss with hardest negatives (*VSE++* [12]) for text-to-image retrieval,

$$L_{ret}(C_i, \{I_1, I_2, \cdots, I_n\}) = \max_{j \neq i}[m - s(c_i, v_i) + s(c_i, v_j)]_+, \tag{4}$$

where $[x]_+ = \max(x, 0)$. For a query caption C_i, we compare the similarity between the positive feature pair $\{c_i, v_i\}$ with the negative pairs $\{c_i, v_j\}$, where $j \neq i$. This loss forces the similarity of the positive pair to be higher than the similarity of the hardest negative pair by a margin m. We also explore other retrieval loss formulations in Sec. 4.4.

The self-retrieval module acts as a discriminativeness evaluator of the captioning module, which encourages a caption generated from a given image by the captioning module to be the best matching to the given image among a batch of distractor images.

Back-propagation by REINFORCE algorithm. For each input image, since self-retrieval is performed based on the complete generated caption, and sampling a word from a probability distribution is non-differentiable, we cannot back-propagate the self-retrieval loss to the captioning module directly. Therefore, REINFORCE algorithm is adopted to back-propagate the self-retrieval loss to the captioning module.

For image captioning with reinforcement learning, the LSTM acts as an "agent", and the previous generated words and image features are "environment". The parameters θ define the policy p_θ and at each step the model chooses an "action", which is the prediction of the next word based on the policy and the environment. Denote $C^s = \{w_1^s, \ldots, w_T^s\}$ as the caption sampled from the predicted word distribution. Each sampled sentence receives a "reward" $r(C^s)$, which indicates its quality. Mathematically, the goal of training is to minimize the negative expected reward of the sampled captions,

$$L_{RL}(\theta) = -\mathbb{E}_{C^s \sim p_\theta}[r(C^s)]. \tag{5}$$

Since calculating the expectation of reward over the policy distribution is intractable, we estimate it by Monte Carlo sampling based on the policy p_θ. To avoid differentiating $r(C^s)$ with respect to θ, we calculate the gradient of the expected reward by REINFORCE algorithm [45],

$$\nabla_\theta L_{RL}(\theta) = -\mathbb{E}_{C^s \sim p_\theta}[r(C^s) \nabla_\theta \log p_\theta(C^s)]. \tag{6}$$

To reduce the variance of the gradient estimation, we subtract the reward with a baseline b, without changing the expected gradient [37]. b is chosen as the reward of greedy decoding captions [35].

$$\nabla_\theta L_{RL}(\theta) = -\mathbb{E}_{C^s \sim p_\theta}[(r(C^s) - b)\nabla_\theta \log p_\theta(C^s)]. \tag{7}$$

For calculation simplicity, the expectation is approximated by a single Monte-Carlo sample from p_θ,

$$\nabla_\theta L_{RL}(\theta) \approx -(r(C^s) - b)\nabla_\theta \log p_\theta(C^s). \tag{8}$$

In our model, for each sampled caption C_i^s, we formulate the reward as a weighted summation of its CIDEr score and the self-retrieval reward, which is the negative caption-to-image retrieval loss.

$$r(C_i^s) = r_{cider}(C_i^s) + \alpha \cdot r_{ret}(C_i^s, \{I_1, \cdots, I_n\}), \tag{9}$$

where $r_{cider}(C_i^s)$ denotes the CIDEr score of C_i^s, $r_{ret} = -L_{ret}$ is the self-retrieval reward, and α is the weight to balance the rewards. The CIDEr reward ensures that the generated captions are similar to the annotations, and the self-retrieval reward encourages the captions to be discriminative. By introducing this reward function, we can optimize the sentence-level reward through sampled captions.

3.2 Improving Captioning with Partially Labeled Images

Training with partially labeled data. The self-retrieval module compares a generated caption with its corresponding image and other distractor images in the mini-batch. As the caption-image correspondence is incorporated naturally in caption generation, *i.e.*, a caption with the image it is generated from automatically form a positive caption-image pair, and with other images form negative pairs, our proposed self-retrieval reward does not require ground-truth captions. So our framework can generalize to semi-supervised setting, where a portion of images do not have ground-truth captions. Thus more training data can be involved in training without extra annotations.

We mix labeled data and unlabeled data with a fixed proportion in each mini-batch. Denote the labeled images in a mini-batch as $\{I_1^l, I_2^l, \cdots, I_{n_l}^l\}$, and their generated captions as $\{C_1^l, C_2^l, \cdots, C_{n_l}^l\}$. Denote unlabeled images in the same mini-batch as $\{I_1^u, I_2^u, \cdots, I_{n_u}^u\}$ and the corresponding generated captions as $\{C_1^u, C_2^u, \cdots, C_{n_u}^u\}$. The reward for labeled data is the composed of the CIDEr reward and self-retrieval reward computed in the mini-batch for each generated caption,

$$r(C_i^l) = r_{cider}(C_i^l) + \alpha \cdot r_{ret}(C_i^l, \{I_1^l, \cdots, I_{n_l}^l\} \cup \{I_1^u, \cdots, I_{n_u}^u\}). \tag{10}$$

The retrieval reward r_{ret} compares the similarity between a caption and the corresponding image, with those of all other labeled and unlabeled images in the mini-batch, to reflect how well the generated caption can discriminate its corresponding image from other distractor images.

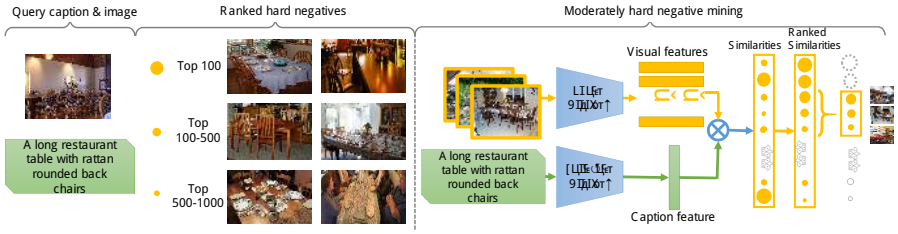

Fig. 3. Moderately hard negative mining. The left part shows a ground-truth caption and its top hard negatives mined from unlabeled images. The right part shows the process of moderately hard negative mining. The circles of different sizes stand for the similarities between each image and the query caption.

As CIDEr reward cannot be computed without ground-truth captions, the reward for unlabeled data is only the retrieval reward computed in the mini-batch,

$$r(C_i^u) = \alpha \cdot r_{ret}(C_i^u, \{I_1^l, \cdots, I_{n_l}^l\} \cup \{I_1^u, \cdots, I_{n_u}^u\}). \tag{11}$$

In this way, the unlabeled data could also be used in training without captioning annotations, to further boost the captioning performance.

Moderately Hard Negative Mining in Unlabeled Images. As described before, the self-retrieval reward is calculated based on the similarity between positive (corresponding) caption-image pairs and negative (non-corresponding) pairs. The training goal is to maximize the similarities of positive pairs and minimize those of negative pairs. To further encourage discriminative captions, we introduce hard negative caption-image pairs in each mini-batch. For example, in Fig. 1, although the first two images are similar, humans are not likely to describe them in the same way. We would like to encourage captions that can distinguish the second image from the first one (*e.g.*, "Creative centerpiece floral arrangement at an outdoor table"), instead of a generic description (*e.g.*, "A vase sitting on a table").

However, an important observation is that choosing the hardest negatives may impede training. This is because images and captions do not always follow strictly one-to-one mapping. In the left part of Fig. 3, we show a ground-truth caption and its hard negatives mined from unlabeled images. The top negative images from the unlabeled dataset often match well with the ground-truth captions from the labeled dataset. For example, when the query caption is "A long restaurant table with rattan rounded back chairs", some of the retrieved top images can also match the caption well. So directly taking the hardest negative pairs is not optimal. Therefore, we propose to use *moderately hard negatives* of the generated captions instead of the hardest negatives.

We show moderately hard negative mining in the right part of Fig. 3. We encode a ground-truth caption C^* from the labeled dataset into features c^* and

all unlabeled images $\{I_1^u, \cdots, I_{n_u}^u\}$ into features $\{v_1^u, \cdots, v_{n_u}^u\}$. The similarities $\{s(c^*, v_1^u), \cdots, s(c^*, v_{n_u}^u)\}$ between the caption and each unlabeled image are derived by the retrieval model. Then we rank the unlabeled images by the similarities between each image and the query caption C^* in a descending order. Then the indexes of moderately hard negatives are randomly sampled from a given range $[h_{min}, h_{max}]$. The sampled hard negatives from unlabeled images and the captions' corresponding images from the labeled dataset together form a mini-batch.

By moderately hard negative mining, we select proper samples for training, encouraging the captioning model to generate captions that could discriminate the corresponding image from other distractor images.

3.3 Training Strategy

We first train the text-to-image self-retrieval module with all training images and corresponding captions in the labeled dataset. The captioning module shares the image encoder with the self-retrieval module. When training the captioning module, the retrieval module and CNN image encoder are fixed.

For captioning module, we first pre-train it with cross-entropy loss, to provide a stable initial point, and reduce the sample space for reinforcement learning. The captioning module is then trained by REINFORCE algorithm with CIDEr reward and self-retrieval reward with either fully labeled data or partially labeled data. The CIDEr reward guarantees the generated captions to be similar to ground-truth captions, while the self-retrieval reward encourages the generated captions to be discriminative. For labeled data, the reward is the weighted sum of CIDEr reward and self-retrieval reward (Eq. (10)), and for unlabeled data, the reward is only the self-retrieval reward (Eq. (11)). The unlabeled data in each mini-batch is chosen by moderately hard negative mining from unlabeled data. Implementation details can be found in Sec. 4.2.

4 Experiments

4.1 Datasets and Evaluation Criteria

We perform experiments on COCO and Flickr30k captioning datasets. For fair comparison, we adopt the widely used Karpathy split [17] for COCO dataset, which uses 5,000 images for validation, 5,000 for testing, and the rest 82,783 for training. For data preprocessing, we first convert all characters into lower case and remove the punctuations. Then we replace words that occur less than 6 times with an 'UNK' token. The captions are truncated to be no more than 16 words during training. When training with partially labeled data, we use the officially released COCO unlabeled images as additional data without annotations. The widely used BLEU [30], METEOR [10], ROUGE-L [23], CIDEr-D [39] and SPICE [1] scores are adopted for evaluation.

4.2 Implementation Details

Self-retrieval module. For the self-retrieval module, each word is embedded into a 300-dimensional vector and inputted to the GRU language encoder, which encodes a sentence into 1024-dimensional features. The image encoder is a ResNet-101 model, which encodes an image into 2048-dimensional visual features. Both the encoded image features and sentence features are projected to the joint embedding space of dimension 1024. The similarity between image features and sentence features is the inner product between the normalized feature vectors. We follow the training strategy in [12].

Captioning module. The captioning module shares the same image encoder with the self-retrieval module. The self-retrieval module and image encoder are fixed when training the captioning module. We take the $2048 \times 7 \times 7$ features before the average pooling layer from ResNet-101 as the visual features. For the language decoder, we adopt a topdown attention LSTM and a language LSTM, following the Top-Down attention model in [2]. We do not use Up-Down model in the same paper, because it involves an object detection model and requires external data and annotations from Visual Genome [19] for training.

The captioning module is trained with Adam [18] optimizer. The model is first pre-trained by cross-entropy loss, and then trained by REINFORCE. Restart technique [25] is used improve the model convergence. We use scheduled sampling [3] and increase the probability of feeding back a sample of the word posterior by 0.05 every 5 epochs, until the feedback probability reaches 0.25. We set the weight of self-retrieval reward α to 1. For training with partially labeled data, the proportion of labeled and unlabeled images in a mini-batch is 1:1.

Inference. For inference, we use beam search with beam size 5 to generate captions. Specifically, we select the top 5 sentences with the highest probability at each time step, and consider them as the candidates based on which to generate the next word. We do not use model ensemble in our experiments.

4.3 Results

Quantitative results. We compare our captioning model performance with existing methods on COCO and Flickr30k datasets in Table 1 and Table 2. The models are all pre-trained by cross-entropy loss and then trained with REINFORCE algorithm. The baseline model is the captioning module only trained with only CIDEr reward. The SR-FL model is our proposed framework trained with fully labeled data, with both CIDEr and self-retrieval rewards. The SR-PL model is our framework trained with partially labeled data (all labeled data and additional unlabeled data), with both rewards for labeled images and only self-retrieval reward for unlabeled images. It is shown from the results that the baseline model without self-retrieval module is already a strong baseline. Incorporating the self-retrieval module with fully-labeled data (SR-FL) improves most metrics by large margins. Training with additional unlabeled data (SR-PL) further enhances the performance. The results validate that discriminativeness is

Table 1. Single-model performance by our proposed method and state-of-the-art methods on COCO standard Karpathy test split.

Methods	CIDEr	SPICE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hard-attention [47]	-	-	71.8	50.4	35.7	25.0	23.0	-
Soft-attention [47]	-	-	70.7	49.2	34.4	24.3	23.9	-
VAE [32]	90.0	-	72.0	52.0	37.0	28.0	24.0	-
ATT-FCN [50]	-	-	70.9	53.7	40.2	30.4	24.3	-
Att-CNN+RNN [46]	94.0	-	74.0	56.0	42.0	31.0	26.0	-
SCN-LSTM [14]	101.2	-	72.8	56.6	43.3	33.0	25.7	-
Adaptive [26]	108.5	-	74.2	58.0	43.9	33.2	26.6	-
SCA-CNN [5]	95.2	-	71.9	54.8	41.1	31.1	25.0	53.1
SCST-Att2all [35]	114.0	-	-	-	-	34.2	26.7	55.7
LSTM-A [49]	100.2	18.6	73.4	56.7	43.0	32.6	25.4	54.0
DRL [34]	93.7	-	71.3	53.9	40.3	30.4	25.1	52.5
Skeleton Key [43]	106.9	-	74.2	57.7	44.0	33.6	26.8	55.2
CNNL+RHN [16]	98.9	-	72.3	55.3	41.3	30.6	25.2	-
TD-M-ATT [4]	111.6	-	76.5	60.3	45.6	34.0	26.3	55.5
ATTN+C+D(1) [27]	114.25	**21.05**	-	-	-	**36.14**	27.38	**57.29**
Ours-baseline	112.7	20.0	79.7	62.2	47.1	35.0	26.7	56.4
Ours-SR-FL	114.6	20.5	79.8	62.3	47.1	34.9	27.1	56.6
Ours-SR-PL	**117.1**	21.0	**80.1**	**63.1**	**48.0**	35.8	**27.4**	57.0

Table 2. Single-model performance by our proposed method and state-of-the-art methods on Flickr30k.

Methods	CIDEr	SPICE	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
Hard-attention [47]	-	-	66.9	43.9	29.6	19.9	18.5	-
Soft-attention [47]	-	-	66.7	43.4	28.8	19.1	18.5	-
VAE [32]	-	-	72.0	53.0	38.0	25.0	-	-
ATT-FCN [50]	-	-	64.7	46.0	32.4	23.0	18.9	-
Att-CNN+RNN [46]	-	-	73.0	55.0	40.0	28.0	-	-
SCN-LSTM [14]	-	-	73.5	53.0	37.7	25.7	21.0	-
Adaptive [26]	53.1		67.7	49.4	35.4	25.1	20.4	-
SCA-CNN [5]	-	-	66.2	46.8	32.5	22.3	19.5	-
CNNL+RHN [16]	61.8	15.0	**73.8**	**56.3**	**41.9**	**30.7**	21.6	-
Ours-baseline	57.1	14.2	72.8	53.4	38.0	27.1	20.7	48.5
Ours-SR-FL	61.7	15.3	72.0	53.4	38.5	27.8	21.5	49.4
Ours-SR-PL	**65.0**	**15.8**	72.9	54.5	40.1	29.3	**21.8**	**49.9**

crucial to caption quality, and enforcing this constraint by self-retrieval module leads to better captions.

Qualitative results. Fig. 4 shows some examples of our generated captions and ground-truth captions. Both the baseline model and our model with self-retrieval reward can generate captions relevant to the images. However, it is easy to observe that our model can generate more discriminative captions, while the baseline model generates generic and templated captions. For example, the first and the second images in the first row share slightly different contents. The baseline model fails to describe their differences and generates identical captions "A vase with flowers sitting on a table". But our model captures the distinction, and expresses with sufficient descriptive details "red flowers", "white vase" and "in a garden" that help to distinguish those images. The captions for the last images in both rows show that the baseline model falls into a stereotype and generates templated captions, because of a large number of similar phrases in the training set. However, captions generated by our model alleviates this problem, and generates accurate descriptions for the images.

Fig. 4. Qualitative results by baseline model and our proposed model.

Table 3. Ablation study results on COCO.

Experiment Settings		CIDEr	SPICE	BLEU-3	BLEU-4	METEOR	ROUGE-L
Baseline		112.7	20.0	47.1	35.0	26.7	56.4
Retrieval Loss	VSE++	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	VSE0	116.9	20.9	47.7	35.7	**27.4**	56.8
	softmax	114.5	20.5	46.8	34.6	27.1	56.5
Weight of Self-retrieval Reward α	0	112.7	20.0	47.1	35.0	26.7	56.4
	1	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	4	113.7	20.5	46.5	34.3	27.0	56.5
Ratio between labeled and unlabeled	1:2	115.4	20.5	46.8	34.7	27.2	56.6
	1:1	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**
	2:1	115.0	20.5	46.8	34.7	27.2	56.7
Hard Negative Index Range	no hard mining	114.6	20.7	46.7	34.6	27.3	56.7
	top 100	114.1	20.3	46.6	34.5	27.0	56.4
	top 100-1000	**117.1**	**21.0**	**48.0**	**35.8**	**27.4**	**57.0**

4.4 Ablation Study

Formulation of self-retrieval loss. As described in Sec. 3.1, the self-retrieval module requires a self-retrieval loss to measure the discriminativeness of the generated captions. Besides *VSE++* loss (Eq. (4)), we explore triplet ranking loss without hard negatives, denoted by *VSE0*,

$$L_{ret}(C_i, \{I_1, I_2, \cdots, I_n\}) = \sum_{j \neq i} [m - s(c_i, v_i) + s(c_i, v_j)]_+, \tag{12}$$

and softmax classification loss, denoted by *softmax*,

$$L_{ret}(C_i, \{I_1, I_2, \cdots, I_n\}) = -\log \frac{\exp\left(s(c_i, v_i)/T\right)}{\sum_{j=1}^{n} \exp\left(s(c_i, v_j)/T\right)}, \tag{13}$$

where T is the temperature parameter that normalizes the caption-image similarity to a proper range. We show the results trained by the three loss formulations in Table 3.* All of those loss formulations lead to better performance

* For the reported results in all experiments and ablation study, we tuned hyper-parameters on the validation set and directly used validations best point to report results on the test set.

Table 4. Text-to-image retrieval performance, and uniqueness and novelty of generated captions by different methods on COCO.

Methods	Generated-caption-to-image retrieval			Uniqueness and novelty evaluation	
	recall@1	recall@5	recall@10	unique captions	novel captions
Skeleton Key [43]	-	-	-	66.96%	52.24%
Ours-baseline	27.5	59.3	74.0	61.56%	51.38%
Ours-SR-PL	**33.0**	**66.4**	**80.1**	**72.34%**	**61.52%**

compared to the baseline model, demonstrating the effectiveness of our proposed self-retrieval module. Among them, *vse++* loss performs slightly better, which is consistent with the conclusion in [12] that *vse++* loss leads to better visual-semantic embeddings.

Balance between self-retrieval reward and CIDEr reward. During training by REINFORCE algorithm, the total reward is formulated as the weighted summation of CIDEr reward and self-retrieval reward, as shown in Eq. (10). To determine how much each of them should contribute to training, we investigate how the weight between them should be set. As shown in Table 3, we investigate $\{0, 1, 4\}$ for the weight of self-retrieval reward α, and the results indicate that $\alpha = 1$ leads to the best performance. Too much emphasis on self-retrieval reward will harm the model performance, because it fails to optimize the evaluation metric CIDEr. This shows that both CIDEr and our proposed self-retrieval reward are crucial and their contributions need to be balanced properly.

Proportion of labeled and unlabeled data. When training with partially labeled data, we use a fixed proportion between labeled and unlabeled images. We experiment on the proportion of forming a mini-batch with labeled and unlabeled data. We try three proportions, 1:2, 1:1 and 2:1, with the same self-retrieval reward weight $\alpha = 1$. The results in Table 3 show that the proportion of 1:1 leads to the best performance.

Moderately Hard Negative Mining. In Sec. 3.2, we introduce how to mine semantically similar images from unlabeled data to provide moderately hard negatives for training. We analyze the contribution of moderately hard negative mining in Table 3. Firstly, the performance gain is relatively low without hard negative mining, demonstrating the effectiveness of this operation. Secondly, after ranking unlabeled images based on the similarity between the given ground-truth caption and unlabeled images in the descending order, the index range $[h_{min}, h_{max}]$ of selecting hard negatives also impacts results. There are cases that an unlabeled image is very similar to an image in the training set, and a caption may naturally correspond to several images. Therefore, selecting the hardest negatives is very likely to confuse the model. In our experiments, we found that setting the hard negative index range $[h_{min}, h_{max}]$ to [100, 1000] for the ranked unlabeled images is optimal.

4.5 Discriminativeness of Generated Captions

Retrieval performance by generated captions. Since the self-retrieval module encourages discriminative captions, we conduct an experiment to retrieve im-

ages with the generated captions as queries, to validate that captions generated by our model are indeed more discriminative than those generated by the model without self-retrieval module. Different from the self-retrieval study in [9], which uses the conditional probabilities of generated captions given images to obtain a ranked list of images, we perform self-retrieval by our self-retrieval module. More precisely, we rank the images based on the similarities between the images and a generated query sentence calculated by our retrieval module. We compute the recall of the corresponding image that appears in the top-k ranked images. The retrieval performance is an indicator of the discriminativeness of generated captions. In Table 4, we report retrieval results on COCO Karpathy test split. It can be clearly seen that the our model improves the retrieval performance by a large margin.

Uniqueness and novelty evaluation. A common problem for captioning models is that they have limited ability of generating captions that have not been seen in the training set, and generates identical sentences for similar images [11]. This demonstrates that the language decoder is simply repeating the sequence patterns it observed in the training set. Although our approach is not directly designed to improve diversity or encourage novel captions, we argue that by encouraging discriminative captions, we can improve the model's ability to generate unique and novel captions. Following the measurements in [43], we evaluate the percentage of unique captions (captions that are unique in all generated captions) and novel captions (captions that have not been seen in training) on COCO Karpathy test split. It is shown in Table 4 that our framework significantly improves uniqueness and novelty of the generated captions.

5 Conclusions

In this work, we address the problem that captions generated by conventional approaches tend to be templated and generic. We present a framework that explicitly improves discriminativeness of captions via training with self-retrieval reward. The framework is composed of a captioning module and a novel self-retrieval module, which boosts discriminativeness of generated captions. The self-retrieval reward is back-propagated to the captioning module by REINFORCE algorithm. Results show that we obtain more discriminative captions by this framework, and achieve state-of-the-art performance on two widely used image captioning datasets.

Acknowledgement

This work is supported by SenseTime Group Limited, the General Research Fund sponsored by the Research Grants Council of Hong Kong (Nos. CUHK14213616, CUHK14206114, CUHK14205615, CUHK14203015, CUHK14239816, CUHK419412, CUHK14207814, CUHK14208417, CUHK14202217), the Hong Kong Innovation and Technology Support Program (No.ITS/121/15FX).

References

1. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV (2016)
2. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and vqa. arXiv preprint arXiv:1707.07998 (2017)
3. Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N.: Scheduled sampling for sequence prediction with recurrent neural networks. In: Advances in Neural Information Processing Systems. pp. 1171–1179 (2015)
4. Chen, H., Ding, G., Zhao, S., Han, J.: Temporal-difference learning with sampling baseline for image captioning (2017)
5. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.S.: 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. pp. 5659–5667 (2017)
6. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015)
7. Chen, X., Lawrence Zitnick, C.: Mind's eye: A recurrent visual representation for image caption generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2422–2431 (2015)
8. Dai, B., Fidler, S., Urtasun, R., Lin, D.: Towards diverse and natural image descriptions via a conditional gan. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2970–2979 (2017)
9. Dai, B., Lin, D.: Contrastive learning for image captioning. In: Advances in Neural Information Processing Systems. pp. 898–907 (2017)
10. Denkowski, M., Lavie, A.: Meteor universal: Language specific translation evaluation for any target language. In: Proceedings of the ninth workshop on statistical machine translation. pp. 376–380 (2014)
11. Devlin, J., Cheng, H., Fang, H., Gupta, S., Deng, L., He, X., Zweig, G., Mitchell, M.: Language models for image captioning: The quirks and what works. arXiv preprint arXiv:1505.01809 (2015)
12. Faghri, F., Fleet, D.J., Kiros, J.R., Fidler, S.: Vse++: Improved visual-semantic embeddings. arXiv preprint arXiv:1707.05612 (2017)
13. Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J., Forsyth, D.: Every picture tells a story: Generating sentences from images. In: European conference on computer vision. pp. 15–29. Springer (2010)
14. Gan, Z., Gan, C., He, X., Pu, Y., Tran, K., Gao, J., Carin, L., Deng, L.: Semantic compositional networks for visual captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. vol. 2 (2017)
15. Gu, J., Cai, J., Wang, G., Chen, T.: Stack-captioning: Coarse-to-fine learning for image captioning. arXiv preprint arXiv:1709.03376 (2017)
16. Gu, J., Wang, G., Cai, J., Chen, T.: An empirical study of language cnn for image captioning. In: Proceedings of the International Conference on Computer Vision (ICCV) (2017)
17. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015)

18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
19. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017)
20. Kulkarni, G., Premraj, V., Ordonez, V., Dhar, S., Li, S., Choi, Y., Berg, A.C., Berg, T.L.: Babytalk: Understanding and generating simple image descriptions. IEEE Transactions on Pattern Analysis and Machine Intelligence **35**(12), 2891–2903 (2013)
21. Li, Y., Ouyang, W., Zhou, B., Cui, Y., Shi, J., Wang, X.: Factorizable net: An efficient subgraph-based framework for scene graph generation (2018)
22. Li, Y., Ouyang, W., Zhou, B., Wang, K., Wang, X.: Scene graph generation from objects, phrases and region captions. In: ICCV (2017)
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. Text Summarization Branches Out (2004)
24. Liu, S., Zhu, Z., Ye, N., Guadarrama, S., Murphy, K.: Improved image captioning via policy gradient optimization of spider. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 873–881 (2017)
25. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:1608.03983 (2016)
26. Lu, J., Xiong, C., Parikh, D., Socher, R.: 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). vol. 6 (2017)
27. Luo, R., Price, B., Cohen, S., Shakhnarovich, G.: Discriminability objective for training descriptive captions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6964–6974 (2018)
28. Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., Yuille, A.: Deep captioning with multimodal recurrent neural networks (m-rnn). arXiv preprint arXiv:1412.6632 (2014)
29. Mitchell, M., Han, X., Dodge, J., Mensch, A., Goyal, A., Berg, A., Yamaguchi, K., Berg, T., Stratos, K., Daumé III, H.: Midge: Generating image descriptions from computer vision detections. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 747–756. Association for Computational Linguistics (2012)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th annual meeting on association for computational linguistics. pp. 311–318. Association for Computational Linguistics (2002)
31. Pedersoli, M., Lucas, T., Schmid, C., Verbeek, J.: Areas of attention for image captioning. In: ICCV-International Conference on Computer Vision (2017)
32. Pu, Y., Gan, Z., Henao, R., Yuan, X., Li, C., Stevens, A., Carin, L.: Variational autoencoder for deep learning of images, labels and captions. In: Advances in neural information processing systems. pp. 2352–2360 (2016)
33. Ranzato, M., Chopra, S., Auli, M., Zaremba, W.: Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732 (2015)
34. Ren, Z., Wang, X., Zhang, N., Lv, X., Li, L.J.: Deep reinforcement learning-based image captioning with embedding reward. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on. pp. 1151–1159. IEEE (2017)

35. Rennie, S.J., Marcheret, E., Mroueh, Y., Ross, J., Goel, V.: Self-critical sequence training for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7008–7024 (2017)
36. Shetty, R., Rohrbach, M., Hendricks, L.A., Fritz, M., Schiele, B.: Speaking the same language: Matching machine to human captions by adversarial training. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2017)
37. Sutton, R.S., Barto, A.G.: Reinforcement learning: An introduction, vol. 1. MIT press Cambridge (1998)
38. Vedantam, R., Bengio, S., Murphy, K., Parikh, D., Chechik, G.: Context-aware captions from context-agnostic supervision. In: Computer Vision and Pattern Recognition (CVPR). vol. 3 (2017)
39. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4566–4575 (2015)
40. Vijayakumar, A.K., Cogswell, M., Selvaraju, R.R., Sun, Q., Lee, S., Crandall, D., Batra, D.: Diverse beam search: Decoding diverse solutions from neural sequence models. arXiv preprint arXiv:1610.02424 (2016)
41. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. In: Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on. pp. 3156–3164. IEEE (2015)
42. Wang, L., Schwing, A., Lazebnik, S.: Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In: Advances in Neural Information Processing Systems. pp. 5758–5768 (2017)
43. Wang, Y., Lin, Z., Shen, X., Cohen, S., Cottrell, G.W.: Skeleton key: Image captioning by skeleton-attribute decomposition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 7272–7281 (2017)
44. Wang, Z., Wu, F., Lu, W., Xiao, J., Li, X., Zhang, Z., Zhuang, Y.: Diverse image captioning via grouptalk. In: IJCAI. pp. 2957–2964 (2016)
45. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning **8**(3-4), 229–256 (1992)
46. Wu, Q., Shen, C., Liu, L., Dick, A., van den Hengel, A.: 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. pp. 203–212 (2016)
47. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: Neural image caption generation with visual attention. In: International Conference on Machine Learning. pp. 2048–2057 (2015)
48. Yang, Y., Teo, C.L., Daumé III, H., Aloimonos, Y.: Corpus-guided sentence generation of natural images. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing. pp. 444–454. Association for Computational Linguistics (2011)
49. Yao, T., Pan, Y., Li, Y., Qiu, Z., Mei, T.: Boosting image captioning with attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4894–4902 (2017)
50. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4651–4659 (2016)
51. Young, P., Lai, A., Hodosh, M., Hockenmaier, J.: From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics **2**, 67–78 (2014)