# Unsupervised Image Captioning

Yang Feng[♯*] Lin Ma[♮†] Wei Liu[♮] Jiebo Luo[♯]
[♮]Tencent AI Lab [♯]University of Rochester
{yfeng23,jluo}@cs.rochester.edu forest.linma@gmail.com wl2223@columbia.edu

## Abstract

*Deep neural networks have achieved great successes on the image captioning task. However, most of the existing models depend heavily on paired image-sentence datasets, which are very expensive to acquire. In this paper, we make the first attempt to train an image captioning model in an unsupervised manner. Instead of relying on manually labeled image-sentence pairs, our proposed model merely requires an image set, a sentence corpus, and an existing visual concept detector. The sentence corpus is used to teach the captioning model how to generate plausible sentences. Meanwhile, the knowledge in the visual concept detector is distilled into the captioning model to guide the model to recognize the visual concepts in an image. In order to further encourage the generated captions to be semantically consistent with the image, the image and caption are projected into a common latent space so that they can reconstruct each other. Given that the existing sentence corpora are mainly designed for linguistic research and are thus with little reference to image contents, we crawl a large-scale image description corpus of two million natural sentences to facilitate the unsupervised image captioning scenario. Experimental results show that our proposed model is able to produce quite promising results without any caption annotations.*

## 1. Introduction

The research on image captioning has made an impressive progress in the past few years. Most of the proposed methods learn a deep neural network model to generate captions conditioned on an input image [7, 10, 13, 20, 21, 40, 41, 42]. These models are trained in a supervised learning manner based on manually labeled image-sentence pairs, as illustrated in Figure 1 (a). However, the acquisition of these paired image-sentence data is a labor intensive process. The scales of existing image captioning datasets, such as Microsoft COCO [30], are relatively small compared with image recognition datasets, such as ImageNet [34] and Open-Images [25]. The image and sentence varieties within these image captioning datasets are limited to be under 100 object categories. As a result, it is difficult for the captioning models trained on such paired image-sentence data to generalize to images in the wild [37]. Therefore, how to relieve the dependency on the paired captioning datasets and make use of other available data annotations to well generalize image captioning models is becoming increasingly important, and thus warrants deep investigations.

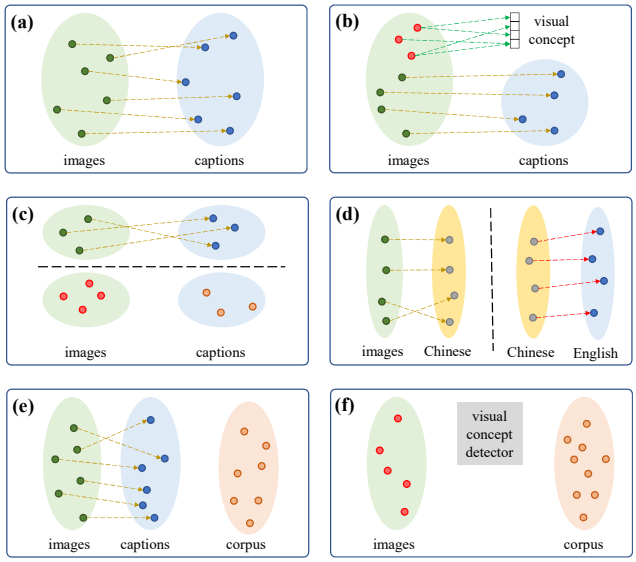


Figure 1. Conceptual differences between the existing captioning methods: (a) supervised captioning [40], (b) novel object captioning [2, 3], (c) cross-domain captioning [8, 43], (d) pivot captioning [15], (e) semi-supervised captioning [9], and (f) our proposed unsupervised captioning.

Recently, there have been several attempts at relaxing the reliance on paired image-sentence data for image captioning training. As shown in Figure 1 (b), Hendricks *et al.* [3] proposed to generate captions for novel objects, which are not present in the paired image-caption training data but exist in image recognition datasets, *e.g.*, ImageNet. As such, novel object information can be introduced into the gener-

*This work was done while Yang Feng was a Research Intern with Tencent AI Lab.
†Corresponding author.

ated captioning sentence without additional paired image-sentence data. A thread of work [8, 43] proposed to transfer and generalize the knowledge learned in existing paired image-sentence datasets to a new domain, where only unpaired data is available, as shown in Figure 1 (c). In this way, no paired image-sentence data is needed for training a new image captioning model in the target domain. Recently, as shown in Figure 1 (d), Gu *et al.* [15] proposed to generate captions in a pivot language (Chinese) and then translate the pivot language captions to the target language (English), which requires no more paired data of images and target language captions. Chen *et al.* [9] proposed a semi-supervised framework for image captioning, which uses an external text corpus, shown in Figure 1 (d), to pre-train their image captioning model. Although these methods have achieved improved results, a certain amount of paired image-sentence data is indispensable for training the image captioning models.

To the best of our knowledge, no work has explored unsupervised image captioning, *i.e.*, training an image captioning model without using any labeled image-sentence pairs. Figure 1 (f) shows this new scenario, where only one image set and one external sentence corpus are used in an unsupervised training setting, which, if successful, can dramatically reduce the labeling work required to create a paired image-sentence dataset. However, it is very challenging to figure out how we can leverage the independent image set and sentence corpus to train a reliable image captioning model.

Recently, several models, relying on only monolingual corpora, have been proposed for unsupervised neural machine translation [4, 26]. The key idea of these methods is to map the source and target languages into a common space by a shared encoder with cross-lingual embeddings. Compared with unsupervised machine translation, unsupervised image captioning is even more challenging. The images and sentences reside in two modalities with significantly different characteristics. Convolutional neural network (CNN) [28] usually acts as an image encoder, while recurrent neural network (RNN) [18] is naturally suitable for encoding sentences. Due to their different structures and characteristics, the encoders of image and sentence cannot be shared, as in unsupervised machine translation.

In this paper, we make the first attempt to train image captioning models without any labeled image-sentence pairs. Specifically, three key objectives are proposed. First, we train a language model on the sentence corpus using the adversarial text generation method [12], which generates a sentence conditioned on a given image feature. As illustrated in Figure 1 (f), we do not have the ground-truth caption of a training image in the unsupervised setting. Therefore, we employ adversarial training [14] to generate sentences such that they are indistinguishable from the sentences within the corpus. Second, in order to ensure that the generated captions contain the visual concepts in the image, we distill [17] the knowledge provided by a visual concept detector into the image captioning model. Specifically, a reward will be given when a word, which corresponds to the detected visual concepts in the image, appears in the generated sentence. Third, to encourage the generated captions to be semantically consistent with the image, the image and sentence are projected into a common latent space. Given a projected image feature, we can decode a caption, which can further be used to reconstruct the image feature. Similarly, we can encode a sentence from the corpus to the latent space feature and thereafter reconstruct the sentence. By performing *bi-directional reconstructions*, the generated sentence is forced to closely represent the semantic meaning of the image, in turn improving the image captioning model.

Moreover, we develop an image captioning model initialization pipeline to overcome the difficulties in training from scratch. We first take the concept words in a sentence as input and train a concept-to-sentence model using the sentence corpus only. Next, we use the visual concept detector to recognize the visual concepts present in an image. Integrating these two components together, we are able to generate a pseudo caption for each training image. The *pseudo image-sentence pairs* are used to train a caption generation model in the standard supervised manner, which then serves as an initialization for our image captioning model.

In summary, our contributions are four-fold:

- We make the first attempt to conduct unsupervised image captioning without relying on any labeled image-sentence pairs.
- We propose three objectives to train the image captioning model.
- We propose a novel model initialization pipeline exploiting unlabeled data. By leveraging the visual concept detector, we generate a pseudo caption for each image and initialize the image captioning model using the pseudo image-sentence pairs.
- We crawl a large-scale image description corpus consisting of over two million sentences from the Web for the unsupervised image captioning task. Our experimental results demonstrate the effectiveness of our proposed model in producing quite promising image captions.

## 2. Related Work

### 2.1. Image Captioning

Supervised image captioning has been extensively studied in the past few years. Most of the proposed models use one CNN to encode an image and one RNN to generate

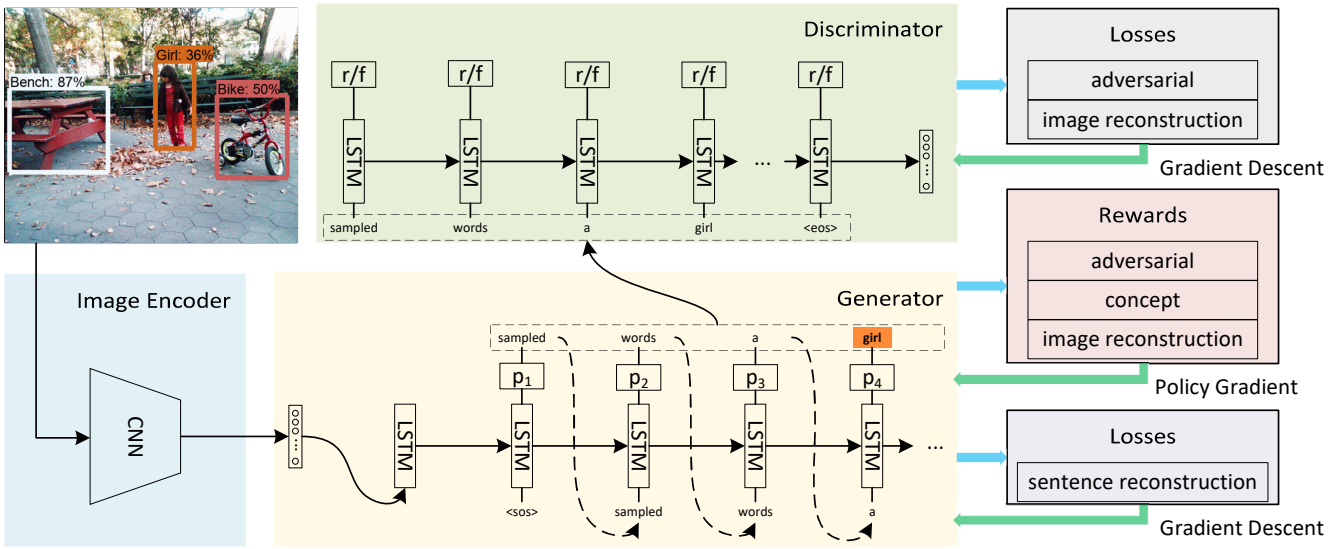


Figure 2. The architecture of our unsupervised image captioning model, consisting of an image encoder, a sentence generator, and a discriminator. A CNN encodes a given image into a feature representation, based on which the generator outputs a sentence to describe the image. The discriminator is used to distinguish whether a caption is generated by the model or from the sentence corpus. Moreover, the generator and discriminator are coupled in a different order to perform image and sentence reconstructions. The adversarial reward, concept reward, and image reconstruction reward are jointly introduced to train the generator via policy gradient. Meanwhile, the generator is also updated by gradient descent to minimize the sentence reconstruction loss. For the discriminator, its parameters are updated by the adversarial loss and image reconstruction loss via gradient descent.

a sentence describing the image [40], respectively. These models are trained to maximize the probability of generating the ground-truth caption conditioned on the input image. As paired image-sentence data is expensive to collect, some researchers tried to leverage other data available to improve the performances of image captioning models. Anderson *et al.* [2] trained an image caption model with partial supervision. Incomplete training sequences are represented by finite state automaton, which can be used to sample complete sentences for training. Chen *et al.* [8] developed an adversarial training procedure to leverage unpaired data in the target domain. Although improved results have been obtained, the novel object captioning or domain adaptation methods still need paired image-sentence data for training. Gu *et al.* [15] proposed to first generate captions in a pivot language and then translate the pivot language caption to the target language. Although no image and target language caption pairs are used, their method depends on image-pivot pairs and a pivot-target parallel translation corpus. In contrast to the methods aforementioned, our proposed method does not need any paired image-sentence data.

### 2.2. Unsupervised Machine Translation

Unsupervised image captioning is similar in spirit to unsupervised machine translation, if we regard the image as the source language. In the unsupervised machine translation methods [4, 26, 27], the source language and target language are mapped into a common latent space so that the sentences of the same semantic meanings in different languages can be well aligned and the following translation can thus be performed. However, the unsupervised image captioning task is more challenging because images and sentences reside in two modalities with significantly different characteristics.

## 3. Unsupervised Image Captioning

Unsupervised image captioning relies on a set of images $\mathcal{I} = \{\boldsymbol{I}_1, \ldots, \boldsymbol{I}_{N_i}\}$, a set of sentences $\hat{\mathcal{S}} = \{\hat{\boldsymbol{S}}_1, \ldots, \hat{\boldsymbol{S}}_{N_s}\}$, and an existing visual concept detector, where $N_i$ and $N_s$ are the total numbers of images and sentences, respectively. Please note that the sentences are obtained from an external corpus, which is not related to the images. For simplicity, we will omit the subscripts and use $\boldsymbol{I}$ and $\hat{\boldsymbol{S}}$ to represent an image and a sentence, respectively. In the following, we first describe the architecture of our image captioning model. Afterwards, we will introduce how to perform the training based on the given data.

### 3.1. The Model

As shown in Figure 2, our proposed image captioning model consists of an image encoder, a sentence generator, and a sentence discriminator.

**Encoder.** One image CNN encodes the input image into one feature representation $\boldsymbol{f}_{im}$:

$$\boldsymbol{f}_{im} = \text{CNN}(\boldsymbol{I}). \tag{1}$$

Common image encoders, such as Inception-ResNet-

v2 [36] and ResNet-50 [16], can be used here. In this paper, we simply choose Inception-V4 [36] as the encoder.

**Generator.** Long short-term memory (LSTM), acting as the generator, decodes the obtained image representation into a natural sentence to describe the image content. At each time-step, the LSTM outputs a probability distribution over all the words in the vocabulary conditioned on the image feature and previously generated words. The generated word is sampled from the vocabulary according to the obtained probability distribution:

$$\begin{aligned} \boldsymbol{x}_{-1} &= \text{FC}(\boldsymbol{f}_{im}), \\ \boldsymbol{x}_t &= \boldsymbol{W}_e \boldsymbol{s}_t,\ t \in \{0 \dots n-1\}, \\ [\boldsymbol{p}_{t+1}, \boldsymbol{h}^g_{t+1}] &= \text{LSTM}^g(\boldsymbol{x}_t, \boldsymbol{h}^g_t),\ t \in \{-1 \dots n-1\}, \\ \boldsymbol{s}_t &\sim \boldsymbol{p}_t,\ t \in \{1 \dots n\}, \end{aligned} \tag{2}$$

where FC and $\sim$ denote the fully-connected layer and sampling operation, respectively. $n$ is the length of the generated sentence with $\boldsymbol{W}_e$ denoting the word embedding matrix. $\boldsymbol{x}_t$, $\boldsymbol{s}_t$, $\boldsymbol{h}^g_t$, and $\boldsymbol{p}_t$ are the LSTM input, a one-hot vector representation of the generated word, the LSTM hidden state, and the probability over the dictionary at the $t$-th time step, respectively. $\boldsymbol{s}_0$ and $\boldsymbol{s}_n$ denote the start-of-sentence (SOS) and end-of-sentence (EOS) tokens, respectively. $\boldsymbol{h}^g_{-1}$ is initialized with zero. For unsupervised image captioning, the image is not accompanied by sentences describing its content. Therefore, one key difference between our generator and the sentence generator in [40] is that $\boldsymbol{s}_t$ is sampled from the probability distribution $\boldsymbol{p}_t$, while the LSTM input word is from the ground-truth caption during training in [40].

**Discriminator.** The discriminator is also implemented as an LSTM, which tries to distinguish whether a partial sentence is a real sentence from the corpus or is generated by the model:

$$\left[q_t, \boldsymbol{h}^d_t\right] = \text{LSTM}^d(\boldsymbol{x}_t, \boldsymbol{h}^d_{t-1}),\ t \in \{1 \dots n\}, \tag{3}$$

where $\boldsymbol{h}^d_t$ is the hidden state of the LSTM. $q_t$ indicates the probability that the generated partial sentence $\boldsymbol{S}_t = [\boldsymbol{s}_1 \dots \boldsymbol{s}_t]$ is regarded as real by the discriminator. Similarly, given a real sentence $\hat{\boldsymbol{S}}$ from the corpus, the discriminator outputs $\hat{q}_t, t \in \{1, \dots, l\}$, where $l$ is the length of $\hat{\boldsymbol{S}}$. $\hat{q}_t$ is the probability that the partial sentence with the first $t$ words in $\hat{\boldsymbol{S}}$ is deemed as real by the discriminator.

## 3.2. Training

As we do not have any paired image-sentence data available, we cannot train our model in the supervised learning manner. In this paper, we define three novel objectives to make unsupervised image captioning possible.

### 3.2.1 Adversarial Caption Generation

The sentences generated by the image captioning model need to be plausible to human readers. Such a goal is usually ensured by training a language model on a sentence corpus. However, as discussed before, the supervised learning approaches cannot be used to train the language model in our setting. Inspired by the recent success of the adversarial text generation method [12], we employ the adversarial training [14] to ensure the plausible sentence generation. The generator takes an image feature as input and generates one sentence conditioned on the image feature. The discriminator distinguishes whether a sentence is generated by the model or is a real sentence from the corpus. The generator tries to fool the discriminator by generating sentences as real as possible. To achieve this goal, we give the generator a reward at each time-step and name this reward as *adversarial reward*. The reward value for the $t$-th generated word is the logarithm of the probability estimated by the discriminator:

$$r^{adv}_t = \log(q_t). \tag{4}$$

By maximizing the adversarial reward, the generator gradually learns to generate plausible sentences. For the discriminator, the corresponding adversarial loss is defined as:

$$\mathcal{L}_{adv} = -\left[\frac{1}{l}\sum_{t=1}^{l}\log(\hat{q}_t) + \frac{1}{n}\sum_{t=1}^{n}\log(1-q_t)\right]. \tag{5}$$

### 3.2.2 Visual Concept Distillation

The adversarial reward only encourages the model to generate plausible sentences following grammar rules, which may be irrelevant to the input image. In order to generate relevant image captions, the captioning model must learn to recognize the visual concepts in the image and incorporate such concepts in the generated sentence. Therefore, we propose to distill the knowledge from an existing visual concept detector into the image captioning model. Specifically, when the image captioning model generates a word whose corresponding visual concept is detected in the input image, we give a reward to the generated word. Such a reward is called a *concept reward*, with the reward value indicated by the confidence score of that visual concept. For an image $\boldsymbol{I}$, the visual concept detector outputs a set of concepts and corresponding confidence scores: $\mathcal{C} = \{(c_1, v_1), \dots, (c_i, v_i), \dots, (c_{N_c}, v_{N_c})\}$, where $c_i$ is the $i$-th detected visual concept, $v_i$ is the corresponding confidence score, and $N_c$ is the total number of detected visual concepts. The *concept reward* assigned to the $t$-th generated word $s_t$ is given by:

$$r^c_t = \sum_{i=1}^{N_c} \text{I}(s_t = c_i) * v_i, \tag{6}$$

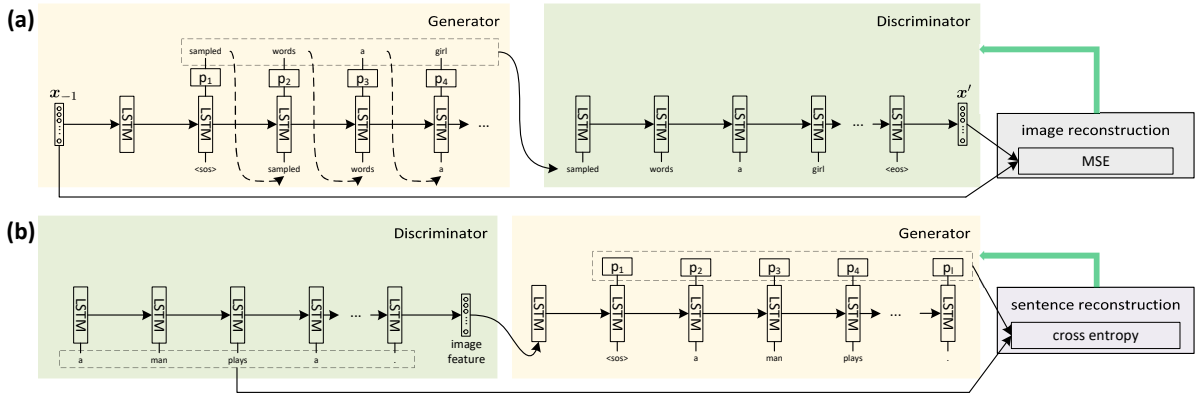


Figure 3. The architectures for image reconstruction (a) and sentence reconstruction (b), respectively, with the generator and discriminator coupled in a different order.

where $\mathrm{I}(\cdot)$ is the indicator function.

#### 3.2.3 Bi-directional Image-Sentence Reconstruction

With the adversarial training and concept reward, the captioning quality would be largely determined by the visual concept detector because it is the only bridge between images and sentences. However, the existing visual concept detectors can only reliably detect a limited number of object concepts. The image captioning model should understand more semantic concepts of the image for a better generalization ability. To achieve this goal, we propose to project the images and sentences into a common latent space such that they can be used to reconstruct each other. Consequently, the generated caption would be semantically consistent with the image.

**Image Reconstruction.** The generator produces a sentence conditioned on an image feature, as shown in Figure 3 (a). The sentence caption should contain the gist of the image. Therefore, we can reconstruct the image from the generated sentence, which can encourage the generated captions to be semantically consistent with the image. However, one hurdle for doing so lies in that it is very difficult to generate images containing complex objects, *e.g.*, people, of high-resolution using current techniques [6, 23]. Therefore, in this paper, we turn to *reconstruct the image features* instead of the full image. As shown in Figure 3 (a), the discriminator can be viewed as a sentence encoder. A fully-connected layer is stacked on the discriminator to project the last hidden state $\boldsymbol{h}_n^d$ to the common latent space for images and sentences:

$$\boldsymbol{x}' = \mathrm{FC}(\boldsymbol{h}_n^d), \tag{7}$$

where $\boldsymbol{x}'$ can be further viewed as the reconstructed image feature from the generated sentence. Therefore, we define an additional image reconstruction loss for training the discriminator:

$$\mathcal{L}_{im} = \|\boldsymbol{x}_{-1} - \boldsymbol{x}'\|_2^2. \tag{8}$$

It can also be observed that the generator together with the discriminator constitutes the image reconstruction process. Therefore, an *image reconstruction reward* for the generator, which is proportional to the negative reconstruction error, can be defined as:

$$r_t^{im} = -\mathcal{L}_{im}. \tag{9}$$

**Sentence Reconstruction.** Similarly, as shown in Figure 3 (b), the discriminator can encode one sentence and project it into the common latent space, which can be viewed as one image representation related to the given sentence. The generator can reconstruct the sentence based on the obtained representation. Such a sentence reconstruction process could also be viewed as a sentence denoising auto-encoder [39]. Besides aligning the images and sentences in the latent space, it also learns how to decode a sentence from an image representation in the common space. In order to make a reliable and robust sentence reconstruction, we add noises to the input sentences by following [26]. The objective of the sentence reconstruction is defined as the cross-entropy loss:

$$\mathcal{L}_{sen} = -\sum_{t=1}^{l} \log\big(p(s_t = \hat{s}_t | \hat{s}_1, \ldots, \hat{s}_{t-1})\big), \tag{10}$$

where $\hat{s}_t$ is the $t$-th word in sentence $\hat{\boldsymbol{S}}$.

#### 3.2.4 Integration

The three objectives are jointly considered to train our image captioning model. For the generator, as the word sampling operation is not differentiable, we train the generator using policy gradient [35], which estimates the gradients with respect to trainable parameters given the joint reward. More specifically, the joint reward consists of *adversarial reward*, *concept reward*, and *image reconstruction reward*. Besides the gradients estimated by policy gradient, the sentence reconstruction loss also provides gradients for

the generator via back-propagation. These two types of gradients are both employed to update the generator. Let $\theta$ denote the trainable parameters in the generator. The gradient with respect to $\theta$ is given by:

$$\nabla_\theta \mathcal{L}(\theta) = -\mathbb{E}\left[\sum_{t=1}^{n}\left(\sum_{s=t}^{n}\gamma^s(\underbrace{r_s^{adv}}_{\text{adversarial}} + \underbrace{\lambda_c r_s^c}_{\text{concept}}\right.\right.$$
$$\left.\left. + \underbrace{\lambda_{im} r_s^{im}}_{\text{image reconstruction}} - b_t\right)\nabla_\theta \log(\boldsymbol{s}_t^\top \boldsymbol{p}_t)\right] + \underbrace{\lambda_{sen}\nabla_\theta \mathcal{L}_{sen}(\theta)}_{\text{sentence reconstruction}}, \tag{11}$$

where $\gamma$ is a decay factor, and $b_t$ is the baseline reward estimated using self-critic [33]. $\lambda_c$, $\lambda_{im}$ and $\lambda_{sen}$ are the hyper-parameters controlling the weights of different terms.

For the discriminator, the adversarial and image reconstruction losses are combined to update the parameters via gradient descent:

$$\mathcal{L}_D = \mathcal{L}_{adv} + \lambda_{im}\mathcal{L}_{im}. \tag{12}$$

During our training process, the generator and discriminator are updated alternatively.

### 3.3. Initialization

It is challenging to adequately train our image captioning model from scratch with the given unlabeled data, even with the proposed three objectives. Therefore, we propose an initialization pipeline to pre-train the generator and discriminator.

Regarding the generator, we would like to generate a pseudo caption for each training image, and then use the pseudo image-caption pairs to initialize an image captioning model. Specifically, we first build a concept dictionary consisting of the object classes in the OpenImages dataset [25]. Second, we train a concept-to-sentence (con2sen) model using the sentence corpus only. Given a sentence, we use a one-layer LSTM to encode the concept words within the sentence into a feature representation, and use another one-layer LSTM to decode the representation into the whole sentence. Third, we detect the concepts of each image by the existing visual concept detector. With the detected concepts and the concept-to-sentence model, we are able to generate a pseudo caption for each image. Fourth, we train the generator with the pseudo image-caption pairs using the standard supervised learning method as in [40]. Such an image captioner is named as feature-to-sentence (feat2sen) and used to initialize the generator.

Regarding the discriminator, parameters are initialized by training an adversarial sentence generation model on the sentence corpus.


Figure 4. Two images and their accompanying descriptions from Shutterstock.

## 4. Experiments

In this section, we evaluate the effectiveness of our proposed method. To quantitatively evaluate our unsupervised captioning method, we use the images in the MSCOCO dataset [30] as the image set (excluding the captions). The sentence corpus is collected by crawling the image descriptions from Shutterstock[1]. The object detection model [19] trained on OpenImages [25] is used as the visual concept detector. We first introduce sentence corpus crawling and experimental settings. Next, we present the performance comparisons as well as the ablation studies.

### 4.1. Shutterstock Image Description Corpus

We collect a sentence corpus by crawling the image descriptions from Shutterstock for the unsupervised image captioning research. Shutterstock is an online stock photography website, which provides hundreds of millions of royalty-free stock images. Each image is uploaded with a description written by the image composer. Some images and description samples are shown in Figure 4. We hope the crawled image descriptions to be somewhat related to the training images. Therefore, we directly use the name of the eighty object categories in the MSCOCO dataset as the search keywords. For each keyword, we download the search results of the top one thousand pages. If the number of pages available is less than one thousand, we will download all the results. There are roughly a hundred images in one page, resulting in $100,000$ descriptions for each object category. After removing the sentences with less than eight words, we collect $2,322,628$ distinct image descriptions in total.

### 4.2. Experimental Settings

Following [22], we split the MSCOCO dataset, with 113,287 images for training, 5,000 images for validation, and the remaining 5,000 images for testing. Please note that the training images are used to build the image set, *with the corresponding captions left unused for any training*. All the descriptions in the Shutterstock image description corpus are tokenized by the NLTK toolbox [5]. We build a vocabulary by counting all the tokenized words and removing the

[1] https://www.shutterstock.com

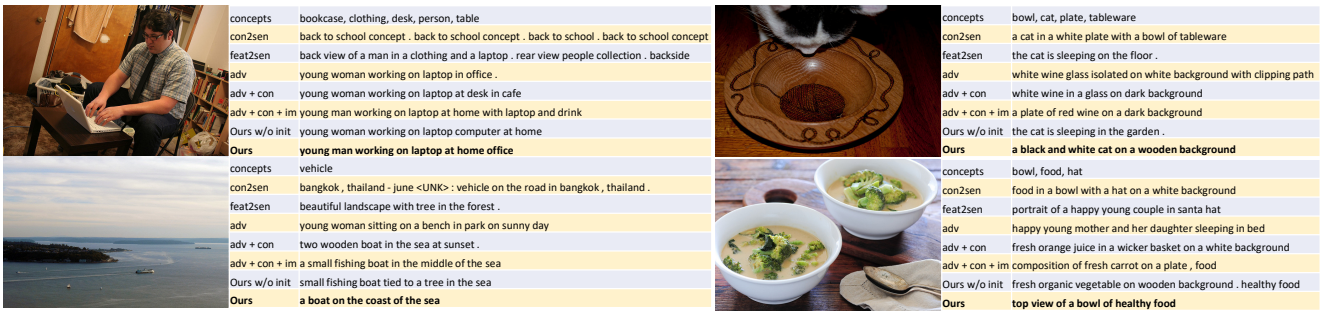


Figure 5. The qualitative results by the unsupervised captioning methods trained with different objectives. Best viewed by zooming in.

Table 1. Performance comparisons of unsupervised captioning methods on the test split [22] of the MSCOCO dataset.

| Method | B1 | B2 | B3 | B4 | M | R | C | S |
|---|---|---|---|---|---|---|---|---|
| Ours w/o init | 38.2 | 20.6 | 9.9 | 4.8 | 11.2 | 27.5 | 22.9 | 6.6 |
| Ours | 41.0 | 22.5 | 11.2 | 5.6 | 12.4 | 28.7 | 28.6 | 8.1 |
| con2sen | 37.2 | 20.0 | 9.6 | 4.7 | 12.3 | 27.3 | 22.5 | 8.2 |
| feat2sen | 38.7 | 21.3 | 10.3 | 5.0 | 12.4 | 28.3 | 23.5 | 8.0 |
| adv | 34.0 | 15.6 | 6.5 | 2.9 | 8.7 | 24.2 | 11.8 | 3.8 |
| adv + con | 37.9 | 19.8 | 9.4 | 4.6 | 11.4 | 26.5 | 24.1 | 7.3 |
| adv + con + im | 37.8 | 19.9 | 9.5 | 4.6 | 11.9 | 26.8 | 25.5 | 7.5 |

words with frequency lower than 40. The object category names of the used object detection model are then merged into the vocabulary. Finally, there are $18,667$ words in our vocabulary, including special SOS, EOS, and an Unkown token. We perform a further filtering process by removing the sentences containing more than 15% Unknown token. After filtering, we retain $2,282,444$ sentences.

The LSTM hidden dimension and the shared latent space dimension are fixed to 512. The weighting hyper-parameters are chosen to make different rewards roughly at the same scale. Specifically, $\lambda_c$, $\lambda_{im}$, $\lambda_{sen}$ are set to be 10, 0.2, and 1, respectively. $\gamma$ is set to be 0.9. We train our model using the Adam optimizer [24] with a learning rate of 0.0001. During the initialization process, we minimize the cross-entropy loss using Adam with the learning rate 0.001. When generating the captions in the test phase, we use beam search with a beam size of 3.

We report the BLEU [31], METEOR [11], ROUGE [29], CIDEr [38], and SPICE [1] scores computed with the coco-caption code [2]. The ground-truth captions of the images in the test split are used for computing the evaluation metrics.

### 4.3. Experimental Results and Analysis

The top region of Table 1 illustrates the unsupervised image captioning results on the test split of the MSCOCO dataset. The captioning model obtained with the proposed unsupervised training method achieves promising results, with CIDEr as 28.6%. Moreover, we also report the results of training our model from scratch ("Ours w/o init") to verify the effect of our proposed initialization pipeline. Without initialization, the CIDEr value drops to 22.9%, which shows that the initialization pipeline can benefit the model training and thus boost image captioning performances.

**Ablation Studies.** The results of the ablation studies are illustrated in the bottom region of Table 1. It can be observed that "con2sen" and "feat2sen" generate reasonable results with CIDEr as 22.5% and 23.5%, respectively. As such, "con2sen" can be used to generate pseudo image-caption pairs for training "feat2sen". And "feat2sen" can make a meaningful initialization of the generator of our captioning model.

When only the adversarial objective is introduced to train the captioning model, "adv" alone leads to much worse results. One cause for this is due to the linguistic characteristics of the crawled image descriptions from Shutterstock, which is significantly different from that of the COCO caption. Another cause is that the adversarial objective only enforces genuine sentence generation but does not ensure its semantic correlation with the image content. Because of the linguistic characteristic difference, most metrics also drop even after introducing the concept objective in "adv + con" and further incorporating image reconstruction objective in "adv + con + im". Although the generated sentences of these two baselines may look plausible, the evaluation results with respect to the COCO captions are not satisfactory. However, by considering all the objectives together, our proposed method substantially improves the captioning performances.

**Qualitative Results.** Figure 5 shows some qualitative results of unsupervised image captioning. In the top-left image, the object detector fails to detect the "`laptop`". So "con2sen" model says nothing about the laptop. On the contrary, the other models successfully recognize laptops and incorporate such concepts into the generated caption. In the top-right image, only a small region of the cat is visible. With such a small region, our full captioning model recognizes that it is "`a black and white cat`". The object detector cannot provide any information about color attribute. We are pleased to see that the bi-directional reconstruction objective is able to guide the captioning model to recognize and express such visual attributes in the generated description sentence. In the bottom

[2] `https://github.com/tylin/coco-caption`

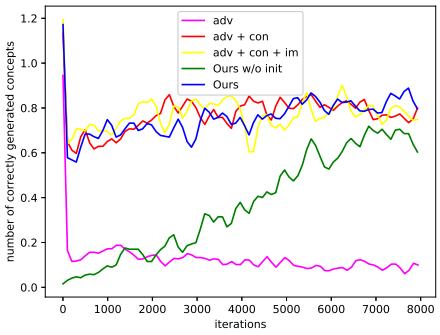


Figure 6. The averaged number of correct concept words in each sentence generated during the training process.

two images, "vehicle" and "hat" are detected by error, which severely affects the results of "con2sen". On the contrary, after training the captioning model with the proposed objectives, the captioning model is able to correct such errors and generate plausible captions.

**Effect of Concept Reward.** Figure 6 shows the averaged number of correct concept words in each sentence generated during the training process. It can be observed that the number of "adv" drops quickly in the beginning. The reason is that the adversarial objective is not related to the visual concepts in the image. "Ours w/o init" continuously increases from zero to about 0.6. The concept reward consistently improves the ability of the captioning model to recognize visual concepts. For "adv + con", "adv + con + im", and "Ours", the number is about 0.8. One reason is that the initialization pipeline gives a good starting point. Another possible reason is that the concept reward prevents the captioning model from drifting towards degradation.

### 4.4. Performance Comparisons under the Unpaired Captioning Setting

The performance of unsupervised captioning model may seem unsatisfactory in terms of the evaluation metrics on the COCO test split. This is mainly due to the different linguistic characteristics between COCO captions and crawled image descriptions. To further demonstrate the effectiveness of the proposed three objectives, we compare with [15] *under the same unpaired captioning setting*, where the COCO captions of the training images are used but in an unpaired manner. Specifically, we replace the crawled sentence corpus with the COCO captions of the training images. All the other settings are kept the same as the unsupervised captioning settings. A new vocabulary with $11,311$ words is created by counting all the words in the training captions and removing the words with frequency less than 4.

The results of unpaired image captioning are shown in Table 2. It can be observed that the captioning model can be consistently improved based on the unpaired data, by including the three proposed objectives step by step. Due to exposure bias [32], some of the captions generated by "feat2sen" are poor sentences. The adversarial objective encourages these generated sentences to appear genuine, resulting in improved performances. With only adversarial training, the model tends to generate sentences unrelated to the image. This problem is mitigated by the concept reward and thus "adv + con" leads to an even better performance. By only including the image reconstruction objective, "adv + con + im" provides a minor improvement. However, if we include the sentence reconstruction objective, our full captioning model achieves another significant improvement, with CIDEr value increasing from 49% to 54.9%. The reason is that the bi-directional image and sentence reconstruction can further leverage the unpaired data to encourage the generated caption to be semantically consistent with the image. The proposed method obtains significantly better results than [15], which may be because that the information in the COCO captions is more adequately exploited in our proposed method.

Table 2. Performance comparisons on the test split [22] of the MSCOCO dataset *under the unpaired setting*.

| Method | B1 | B2 | B3 | B4 | M | R | C | S |
|---|---|---|---|---|---|---|---|---|
| Pivoting [15] | 46.2 | 24.0 | 11.2 | 5.4 | 13.2 | - | 17.7 | - |
| Ours w/o init | 53.8 | 35.5 | 23.1 | 15.6 | 16.6 | 39.9 | 46.7 | 9.6 |
| Ours | 58.9 | 40.3 | 27.0 | 18.6 | 17.9 | 43.1 | 54.9 | 11.1 |
| con2sen | 50.6 | 30.8 | 18.2 | 11.3 | 15.7 | 37.9 | 33.9 | 9.1 |
| feat2sen | 51.3 | 31.3 | 18.7 | 11.8 | 15.3 | 38.1 | 35.4 | 8.8 |
| adv | 55.6 | 35.5 | 23.1 | 15.7 | 17.0 | 40.8 | 45.8 | 10.1 |
| adv + con | 56.2 | 37.2 | 24.2 | 16.2 | 17.3 | 41.5 | 48.8 | 10.5 |
| adv + con + im | 56.4 | 37.5 | 24.5 | 16.5 | 17.4 | 41.6 | 49.0 | 10.5 |

## 5. Conclusions

In this paper, we proposed a novel method to train an image captioning model in an unsupervised manner without using any paired image-sentence data. As far as we know, this is the first attempt to investigate this problem. To achieve this goal, we presented three training objectives, which encourage that 1) the generated captions are indistinguishable from sentences in the corpus, 2) the image captioning model conveys the object information in the image, and 3) the image and sentence features are aligned in the common latent space and perform bi-directional reconstructions from each other. A large-scale image description corpus consisting of over two million sentences was further collected from Shutterstock to facilitate the unsupervised image captioning method. The experimental results demonstrate that the proposed method can produce quite promising results without using any labeled image-sentence pairs. In the future, we will conduct human evaluations for unsupervised image captioning.

## Acknowledgement

This work is partially supported by NSF awards 1704309, 1722847, and 1813709.

## References


[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *ECCV*, 2016.

[2] Peter Anderson, Stephen Gould, and Mark Johnson. Partially-supervised image captioning. In *NeurIPS*, 2018.

[3] Lisa Anne Hendricks, Subhashini Venugopalan, Marcus Rohrbach, Raymond Mooney, Kate Saenko, Trevor Darrell, Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, et al. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016.

[4] Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. Unsupervised neural machine translation. In *ICLR*, 2018.

[5] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[7] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*, 2017.

[8] Tseng-Hung Chen, Yuan-Hong Liao, Ching-Yao Chuang, Wan-Ting Hsu, Jianlong Fu, and Min Sun. Show, adapt and tell: Adversarial training of cross-domain image captioner. In *ICCV*, 2017.

[9] Wenhu Chen, Aurelien Lucchi, and Thomas Hofmann. A semi-supervised framework for image captioning. *arXiv preprint arXiv:1611.05321*, 2016.

[10] Xinpeng Chen, Lin Ma, Wenhao Jiang, Jian Yao, and Wei Liu. Regularizing rnns for caption generation by reconstructing the past with the present. In *CVPR*, 2018.

[11] 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*, 2014.

[12] William Fedus, Ian Goodfellow, and Andrew M Dai. Maskgan: Better text generation via filling in the _. In *ICLR*, 2018.

[13] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.

[14] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[15] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. Unpaired image captioning by language pivoting. In *ECCV*, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[17] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[18] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 1997.

[19] Jonathan Huang, Vivek Rathod, Chen Sun, Menglong Zhu, Anoop Korattikara, Alireza Fathi, Ian Fischer, Zbigniew Wojna, Yang Song, Sergio Guadarrama, et al. Speed/accuracy trade-offs for modern convolutional object detectors. In *CVPR*, 2017.

[20] Wenhao Jiang, Lin Ma, Xinpeng Chen, Hanwang Zhang, and Wei Liu. Learning to guide decoding for image captioning. In *AAAI*, 2018.

[21] Wenhao Jiang, Lin Ma, Yu-Gang Jiang, Wei Liu, and Tong Zhang. Recurrent fusion network for image captioning. In *ECCV*, 2018.

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.

[23] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

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

[25] Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Shahab Kamali, Matteo Malloci, Jordi Pont-Tuset, Andreas Veit, Serge Belongie, Victor Gomes, Abhinav Gupta, Chen Sun, Gal Chechik, David Cai, Zheyun Feng, Dhyanesh Narayanan, and Kevin Murphy. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://storage.googleapis.com/openimages/web/index.html*, 2017.

[26] Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. Unsupervised machine translation using monolingual corpora only. In *ICLR*, 2018.

[27] Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. Phrase-based & neural unsupervised machine translation. In *EMNLP*, 2018.

[28] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 1998.

[29] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[30] 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 *ECCV*, 2014.

[31] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[32] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In *ICLR*, 2016.

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

[34] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. Imagenet large scale visual recognition challenge. *IJCV*, 2015.

[35] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000.

[36] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *AAAI*, 2017.

[37] Kenneth Tran, Xiaodong He, Lei Zhang, Jian Sun, Cornelia Carapcea, Chris Thrasher, Chris Buehler, and Chris Sienkiewicz. Rich image captioning in the wild. In *CVPR Workshops*, 2016.

[38] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015.

[39] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *ICML*, 2008.

[40] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015.

[41] Ting Yao, Yingwei Pan, Yehao Li, Zhaofan Qiu, and Tao Mei. Boosting image captioning with attributes. In *ICCV*, 2017.

[42] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *CVPR*, 2016.

[43] Wei Zhao, Wei Xu, Min Yang, Jianbo Ye, Zhou Zhao, Yabing Feng, and Yu Qiao. Dual learning for cross-domain image captioning. In *CIKM*, 2017.