Unpaired Cross-lingual Image Caption Generation with Self-Supervised Rewards

Yuqing Song
Renmin University of China
syuqing@ruc.edu.cn

Shizhe Chen
Renmin University of China
cszhe1@ruc.edu.cn

Yida Zhao
Renmin University of China
zyiday@ruc.edu.cn

Qin Jin*
Renmin University of China
qjin@ruc.edu.cn

ABSTRACT

Generating image descriptions in different languages is essential to satisfy users worldwide. However, it is prohibitively expensive to collect large-scale paired image-caption dataset for every target language which is critical for training descent image captioning models. Previous works tackle the unpaired cross-lingual image captioning problem through a pivot language, which is with the help of paired image-caption data in the pivot language and pivot-to-target machine translation models. However, such language-pivoted approach suffers from inaccuracy brought by the pivot-to-target translation, including disfluency and visual irrelevancy errors. In this paper, we propose to generate cross-lingual image captions with self-supervised rewards in the reinforcement learning framework to alleviate these two types of errors. We employ self-supervision from mono-lingual corpus in the target language to provide fluency reward, and propose a multi-level visual semantic matching model to provide both sentence-level and concept-level visual relevancy rewards. We conduct extensive experiments for unpaired cross-lingual image captioning in both English and Chinese respectively on two widely used image caption corpora. The proposed approach achieves significant performance improvement over state-of-the-art methods.

KEYWORDS

Image Captioning; Cross-lingual; Reinforcement Learning; Self-supervision
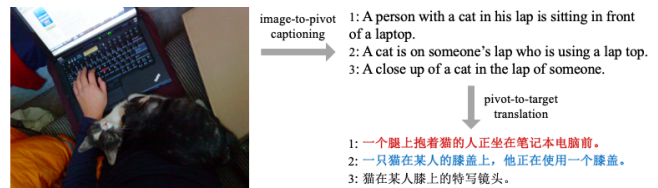
*Corresponding Author

Figure 1: Illustration of cross-lingual Chinese image captioning with English language as pivot. The translated Chinese caption in red suffers from disfluency errors while sentence in blue contains visual irrelevancy errors.

1 INTRODUCTION

Generating natural language sentences to describe the image content, a.k.a image captioning, has received more and more attention in recent years. It could help visually impaired people to better understand the real world, and make it easier to index and retrieve massive images on the web. Thanks to the rapid development of computer vision and natural language generation, remarkable progress has been made in automatic image captioning. However, most of previous works have mainly focused on generating English captions for images. As we know, there are more than 6.6 billion non-native English speakers in the world, and the benefits of image captioning technology should also be brought to these users. Therefore, it is necessary to generate captions in different languages, which is also called the cross-lingual image captioning task.

Since image captioning models are generally data-hungry, the main challenge for cross-lingual image captioning is the lack of large-scale image caption dataset in the target language. It is also prohibitively expensive to collect dataset for each language. Fortunately, great efforts have already been made in collecting large-scale image-caption datasets in English, as well as machine translation datasets from English to other languages. Therefore, for cross-lingual image captioning, a natural way to avoid the demand of paired image-caption data in the target language is to employ another language, such as English, as the pivot to bridge the image and the target language [11], so that the image caption is first generated by an image-to-pivot captioning model, and then translated into the target language by a pivot-to-target machine translation (MT) model. Figure 1 illustrates the idea of utilizing English as pivot for cross-lingual Chinese image captioning.

The major limitation of such language-pivoted approach is that translation errors brought by the pivot-to-target MT model cannot

be corrected which seriously affects the quality of generated captions, especially when the MT model is trained in a different domain from image captions. In order to alleviate the domain mismatch, Gu *et al.* [11] propose to share parameters between the image-to-pivot captioning model and the pivot-to-target MT model, and jointly train these two models, which can enforce the MT model to adapt styles towards image captions. However, this approach is hard to generalize and computationally expensive to employ state-of-the-art MT models for pivot-to-target translation. Lan *et al.* [20] instead directly take advantage of the state-of-the-art translator [1] to generate pseudo image-target caption pairs to train the image-to-target captioning model. They propose to re-weight the translated captions by the language fluency. However, in addition to the disfluent sentences, the imperfect translations may also contain fluent but visually irrelevant sentences as shown in Figure 1, which also greatly affect the accuracy of cross-lingual caption generation.

In this paper, we propose a self-supervised rewarding model (SSR) to deal with both disfluency and visual irrelevancy errors in language-pivoted unpaired image captioning. Our model is based on the reinforcement learning framework, which utilizes two types of rewards learned from self-supervisions to encourage the caption generator to correct above errors. Specifically, to improve the caption fluency, we propose a fluency reward based on a target language model, which is trained with self-supervision loss on mono-lingual sentences in the target language. In order to improve the visual relevancy, we propose a multi-level visual semantic matching model (ML-VSE) to provide relevancy rewards, which employs self-supervised pseudo image-target caption pairs from the pivot-to-target translation model for training. The ML-VSE model contains both sentence-level and concept-level visual semantic matching between images and captions, which provides coarse- and fine-grained rewards respectively. Extensive experiments on two widely used image caption datasets show that our model significantly outperforms prior works on all the caption performance metrics.

The main contributions of this work are summarized as follows:

- We propose to employ the reinforcement learning framework to deal with errors in language-pivoted approaches for unpaired cross-lingual image captioning.
- Introspective self-supervisions with respect to the fluency and visual relevancy of generated captions are designed as the rewards to improve the quality of cross-lingual captions.
- Extensive experiments for both unpaired English image captioning and Chinese image captioning demonstrate that our proposed approach achieves significant improvement over previous methods on both objective caption metrics and human evaluation.

2 RELATED WORKS

2.1 Image Caption Generation

Image caption generation is a challenging task which connects computer vision and natural language processing. With the rapid development in deep learning, great breakthroughs have been made in image captioning [9, 14, 16, 22, 28, 33]. Vinyals *et al.* [28] first propose an end-to-end image captioning model based on the encoder-decoder framework [6]. A convolutional neural network (CNN) [17] is used to encode the image into a fix-dimensional feature vector and a recurrent neural network (RNN) [15] is used as the decoder to generate captions based on the encoder output. The model is jointly optimized by maximizing the log probability of groundtruth descriptions.

Later, many improvements based on such encoder-decoder framework are proposed. Xu *et al.* [31] propose the spatial attention mechanism for image captioning, which divides the image into grids, and teaches the model to attend to the corresponding grid at each decoding step. Anderson *et al.* [1] replace the grids with detected objects in a bottom-up attention to enhance the previous top-down attention method. You *et al.* [32] propose semantic attention which pre-defines a list of visual concepts to be attended to in the decoding step. Gu *et al.* [12] propose to explore both long-term and temporal information in captions with a CNN-based image captioning model. Recently, Biten *et al.* [3] propose to integrate contextual information into the captioning pipeline to deal with the out-of-vocabulary named entity generation.

Besides model structures, the training target also plays an important role in image captioning. The model trained by traditional maximum likelihood target suffers from exposure bias and evaluation mismatch. The exposure bias is caused by the training setting called "Teacher-Forcing" [2], where the model has never been exposed to its own predictions in the training progress, which results in the error accumulation at test time. The evaluation mismatch exists because cross entropy is used as the training loss, but metrics such as BLEU, CIDEr and METEOR are instead used for caption performance evaluation. Therefore, reinforcement learning approaches are proposed to address these two problems. Rennie *et al.* [25] propose a new training method based on reinforcement learning with a baseline reward called "Self-Critical". They provide "reward" for captions sampled from model distribution, and the reward is directly evaluated by CIDEr. In order to enhance the stability of training, they use the reward of captions generated at test time as a baseline reward. Works in [23, 30] propose to train the captioning model by providing rewards of discriminability to improve the diversity of generated captions. Our training strategy is similar to Rennie *et al.* [25] except that we use self-supervision with respect to fluency and relevancy as rewards for model learning.

2.2 Cross-lingual Image Captioning

Cross-lingual image captioning is a more challenging captioning task which has not been well investigated yet, since most previous works have mainly focused on generating English captions. Tsutsui *et al.* [26] propose to generate image captions in Japanese by collecting a large-scale parallel image-caption dataset in Japanese. However, it may not be feasible for many languages due to the expensive cost of dataset collection. Feng *et al.* [10] propose an unsupervised image captioning model with a visual concept detector which is trained on Visual Genome dataset [19]. Although they do not need paired image-caption corpus, a large-scale dataset with images and grounded objects annotations is also difficult to collect in any language. Recently, cross-modal pivoted approaches are popularly used in solving zero-resource learning problems. Chen

et al. [4, 5] propose to utilize images as pivots for zero-resource machine translation. While, Gu *et al.* [11] and Lan *et al.* [20] utilize language as pivot for cross-lingual image captioning. Gu *et al.* [11] propose to train the English image captioning model on images paired with Chinese captions and English-Chinese parallel translation pairs. The model is performed in two steps through language pivoting, which has an inherent deficiency due to translation error accumulation. Lan *et al.* [20] instead directly take advantage of the state-of-the-art translator to generate pseudo image-target caption pairs to train the captioning model. They propose to re-weight translated captions by language fluency to alleviate the disfluency errors brought about by the translator. However, in addition to disfluent sentences, the translation errors also contain fluent but visually irrelevant sentences, which are ignored in their works.

3 UNPAIRED CROSS-LINGUAL IMAGE CAPTIONING WITH SELF-SUPERVISION

In this section, we will describe our self-supervised rewarding (SSR) model for unpaired cross-lingual image captioning. We first present the overview of the model framework in Section 3.1, which is based on reinforcement learning with two types of rewards to address the error accumulation problem in language-pivoted approaches. Then in Section 3.2 and Section 3.3, we describe the proposed self-supervised fluency and relevancy rewards in details.

3.1 Overview

The goal of unpaired cross-lingual image captioning is to generate a natural language sentence to describe the image content in the target language without image-target caption pairs for training. We tackle this problem via a pivot language with the supervision from the help of image-caption pairs in the pivot language and the pivot-to-target translation model. We refer to the pivot-to-target translation model as $f_{P \to T}$, and the image caption dataset in pivot language as $D_P = \{(I^{(i)}, d_P^{(i)})\}_{i=1}^N$, where $I^{(i)}$ refers to an image instance, $d_P^{(i)}$ refers to its corresponding sentence description in the pivot language, and N is the total number of such image-caption pairs. Therefore, although we don't have manually annotated image-caption pairs in the target language, we can generate pseudo pairs $D_T = \{(I^{(i)}, d_T^{(i)})\}_{i=1}^N$ based on $f_{P \to T}$ and D_P where $d_T^{(i)} = f_{P \to T}(d_P^{(i)})$ for training.

If the translation model $f_{P \to T}$ is perfect, the pseudo pair $(I^{(i)}, d_T^{(i)})$ can be used as groundtruth to train the target image captioning model in a supervised way. Thus the unpaired cross-lingual image captioning can be converted to a standard image captioning task. In this work, we employ the vanilla image captioning model based on an encoder-decoder framework [28]. The encoder is a deep CNN [17] to encode the image I to a fixed-dimensional feature vector v. The decoder is a RNN [15] to generate descriptions word by word conditioned on v. The whole model is optimized by maximizing the probability of generating each "groundtruth" caption words. The generation loss function can be expressed as:

$$\mathcal{L}_{cap} = -\sum_{i=1}^{N} \sum_{j=1}^{n} \log P(w_{T,j}^{(i)} | w_{T,0:j-1}^{(i)}, v^{(i)}; \theta_{cap}), \quad (1)$$

where $d_T^{(i)} = \{w_{T,1}^{(i)}, \cdots, w_{T,n}^{(i)}\}$, n is the length of $d_T^{(i)}$, $w_{T,0}^{(i)}$ is the sentence beginning signal <BOS>, and θ_{cap} is the parameters of the image caption model.

However, in reality, $f_{P \to T}$ is not perfect and can produce different translation errors such as disfluent translations or visually irrelevant translations as shown in Figure 1. Such translation errors can greatly deteriorate the image captioning performance because the training supervision $\mathcal{L}_{cap}$ for the captioning model relies on the translated sentences. Therefore, extra supervisions are needed to mitigate the negative effects from $f_{P \to T}$, and provide accurate guidance for the caption generator. In this paper, we utilize reinforcement learning framework to improve the caption performance by providing various rewards.

In reinforcement learning framework, the caption generation can be seen as a sequence decision process. The decoder of the captioning model can be seen as an agent, and the generation of each word can be seen as an action taken by the agent in each step. When action decisions are finished, rewards will be fed back to the agent to "tell" how good these actions are. The objective of reinforcement learning is to maximize expected rewards in the end of decision. In order to address the disfluency and visual irrelevancy translation errors, we propose a fluency reward function $r_{flc}(\cdot)$ in Section 3.2 and multi-level visual relevancy reward functions $r_{srlv}(\cdot)$ and $r_{crlv}(\cdot)$ in Section 3.3 to "tell" the captioning model how to improve the generated captions at both coarse and fine-grained levels. Specifically, we adopt the "self-critical" [25] reinforcement learning algorithm to train our model. Firstly, we carry out Monte-Carlo sampling to sample a sentence $s_s^{(i)}$ and evaluate its caption quality with the proposed reward functions. Then we utilize the greedy search algorithm to generate a sentence $s_b^{(i)}$ to provide baseline reward for the stability of reinforcement training.

Therefore, the joint optimization loss function to train the image captioning model consists of three parts:

$$\mathcal{L} = \alpha \mathcal{L}_{cap} + \beta \mathcal{L}_{flc} + \gamma \mathcal{L}_{rlv} \quad (2)$$

where $\mathcal{L}_{flc}$ and $\mathcal{L}_{rlv}$ are the reinforcement learning objectives in fluency and relevancy aspects respectively; α, β and γ are hyper-parameters, which are chosen according to the scale of these loss values and caption performance on the validation set. Figure 2 illustrates the overall framework of our proposed model.

3.2 Self-supervised Fluency Rewards

In order to improve the fluency quality of generated captions, we employ self-supervision from mono-lingual corpus in the target language $S_T = \{s_T^{(i)}\}_{i=1}^N$ to provide the fluency reward. We pre-train a language model on the mono-lingual corpus to evaluate the sentence fluency quality. We utilize the LSTM as our language model which is trained to maximize the probability of generating target sentence $s_T^{(i)}$. Its loss function is expressed as:

$$\mathcal{L}_{lm} = -\sum_{i=1}^{N} \sum_{j=1}^{n} \log P(w_{s,j}^{(i)} | w_{s,0:j-1}^{(i)}; \theta_{lm}), \quad (3)$$

where $s_T^{(i)} = \{w_{s,1}^{(i)}, \cdots, w_{s,n}^{(i)}\}$, n is the length of $s_T^{(i)}$, and θ_{lm} is the parameter of the language model.
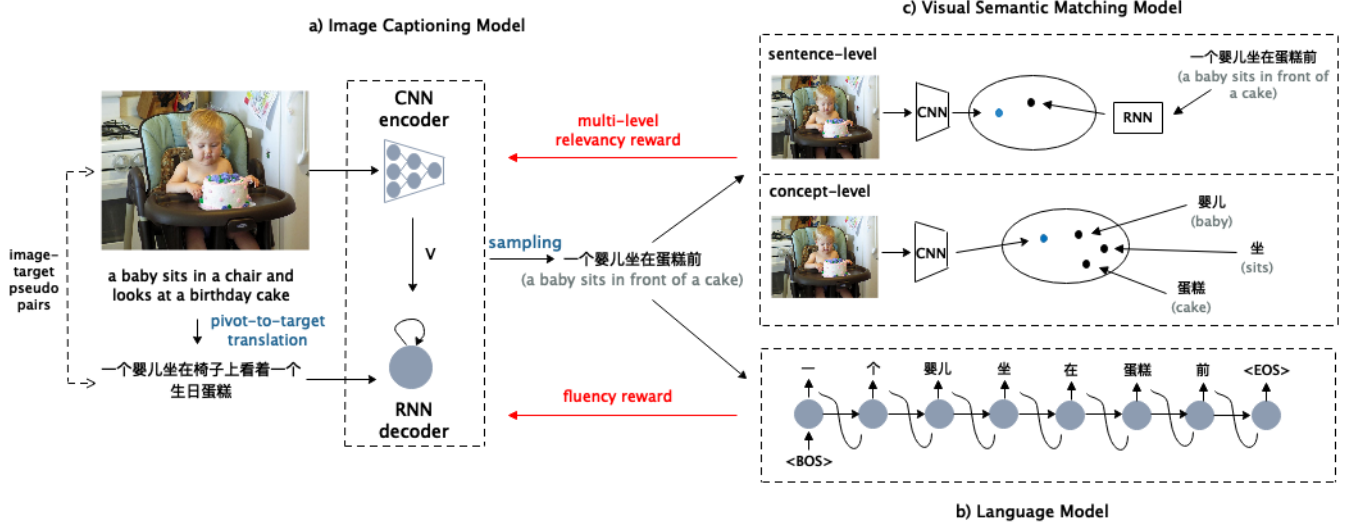
Figure 2: Illustration of the proposed SSR model framework, which consists of three components: a) the image captioning model trained on pseudo image-caption pairs; b) the language model to provide self-supervised fluency reward for the captioning model; c) the visual semantic matching model to provide self-supervised multi-level relevancy rewards. We add the English translation below the sampled Chinese caption in brackets for better understanding.

For a sampled sentence $s^{(i)} = \{w_1^{(i)}, \cdots, w_n^{(i)}\}$ where n is the sentence length, we take the log probability of generating $s^{(i)}$ by the language model as its fluency reward as follows:

$$r_{flc}(s^{(i)}) = \frac{1}{n} \sum_{j=1}^{n} \log P(w_j^{(i)} | w_{0:j-1}^{(i)}; \theta_{lm}). \qquad (4)$$

So the self-critical reinforcement loss function for fluency rewarding is formulated as:

$$\mathcal{L}_{flc} = - \sum_{i=1}^{N} (r_{flc}(s_s^{(i)}) - r_{flc}(s_b^{(i)})) \sum_{j=1}^{n} \log P(w_j^{(i)} | w_{0:j-1}^{(i)}, v^{(i)}; \theta_{cap}). \qquad (5)$$

3.3 Self-supervised Relevancy Rewards

Through the supervision from fluency reward, the caption model is "taught" to generate fluent captions in the target language. However, it cannot guarantee the generated captions are relevant to the given image, especially when the guidance from $\mathcal{L}_{cap}$ is wrong due to the semantically inconsistent translation errors. Therefore, extra relevancy reward is highly required to let the captioning model know what is relevant to the image content and what is not.

We propose to learn a visual semantic matching model to evaluate the relevancy of the generated captions to the image based on the pseudo image-target caption pairs $D_T = \{(I^{(i)}, d_T^{(i)})\}_{i=1}^{N}$. Although such pairs are noisy which might contain disfluent and visually irrelevant translation errors, there are also many content similar images whose descriptions are correctly translated, which enables accurate visual semantic matching. we call this relevancy reward computed by the visual semantic matching model as "self-supervised" reward as no annotated image-target caption pairs are used.

In order to further mitigate noises in the translated sentences, we propose a multi-level visual semantic matching model (ML-VSE) which includes the image-sentence matching at the coarse level and the image-concept matching at the fine-grained level. We use nouns and verbs in the caption sentence as the concepts, which play important roles to deliver semantic information of the sentence. The concepts in the pseudo pairs can be more accurate than sentences since concepts don't suffer from disfluency errors and are easy to translate. We describe the sentence-level and concept-level relevancy rewards computed by the two visual semantic matching models in details below.

Sentence-Level Relevancy Reward. We provide the sentence-level relevancy reward via image-sentence matching. The image is encoded by E_i which consists of a pre-trained CNN and a fully connected embedding layer to generate the image embedding vector v_I. The caption sentence is encoded by E_c which is a bi-directional GRU to generate the caption embedding vector v_c. In order to project v_I and v_c in a common embedding space, we utilize the contrastive ranking loss with hard negative mining [8] for training:

$$\mathcal{L}(I, c) = \max_{c'}[\Delta + s(v_I, v_{c'}) - s(v_I, v_c)]_+ \\ + \max_{I'}[\Delta + s(v_{I'}, v_c) - s(v_I, v_c)]_+, \qquad (6)$$

where Δ severs as a margin hyper-parameter, $[x]_+ \equiv \max(x, 0)$, (I, c) is a pseudo image-caption pair, c' is the negative caption given image I, and I' is the negative image given caption c in the mini-batch. The $s(\cdot)$ means the similarity function between two embedded vectors, which is the cosine similarity in our experiments.

After training, the image-sentence matching model is able to give captions that are relevant to the image higher similarity scores than irrelevant ones. Therefore, our sentence-level visual relevancy reward for the generated caption s of image I is:

$$r_{srlv}(s) = s(E_i(I), E_c(s)). \qquad (7)$$

Concept-Level Relevancy Reward. Similarly to the image-sentence matching model, we utilize E_i to encode the image to vector v_I and encode the concept into the semantic vector v_w with concept embedding matrix E_w. The similar contrastive ranking loss is adopted to the joint image concept embedding space:

$$\mathcal{L}(I, w) = \max_{w'}[\Delta + s(v_I, v_{w'}) - s(v_I, v_w)]_+ \\ + \max_{I'}[\Delta + s(v_{I'}, v_w) - s(v_I, v_w)]_+. \quad (8)$$

The trained image-concept matching model can be used to measure the relevancy of the visual concept and the image. However, the learned similarity score can be greatly influenced by the frequency statistics of concepts. The frequent concepts in pseudo pairs are more likely to obtain high scores than infrequent ones, which biases the captioning model towards frequent concepts. Hence, we normalize the similarity score by the prior probability of the concept in the dataset, so that the concept-level visual relevancy reward is computed as:

$$r_{crlv}(w) = \delta(w)(s(E_i(I), E_w(w)) - \lambda p(w)) \quad (9)$$

where (I, w) is the image-concept pair extracted from the pseudo image-caption pairs, $p(w)$ is the prior probability of w which is its occurrence frequency, $\delta(w)$ denotes whether the word w is a visual concept, and λ is a hyper-parameter.

Therefore, our multi-level self-critical loss to improve the visual relevancy of generated captions is as follows:

$$\mathcal{L}_{rlv} = -\sum_{i=1}^{N}\sum_{j=1}^{n}(r_{srlv}(s_s^{(i)}) - r_{srlv}(s_b^{(i)}) + r_{crlv}(w_j^{(i)})) \\ \cdot \log P(w_j^{(i)}|w_{0:j-1}^{(i)}, v^{(i)}; \theta_{cap}). \quad (10)$$

The overall training process of the proposed self-supervised rewarding model is presented in Algorithm 1.

Algorithm 1 Training algorithm of the proposed self-supervised rewarding model for unpaired cross-lingual image captioning.

Require: pivot image caption dataset D_P; pivot-to-target machine translation model $f_{P\to T}$; target language sentence corpus S_T;
1: Generate pseudo image-target caption pairs D_T based on D_P and $f_{P\to T}$;
2: Pre-train the target language model θ_{lm} based on S_T with Eq (3);
3: Pre-train E_i, E_c, E_w in ML-VSE model based on D_T with Eq (6) and Eq (8) respectively;
4: Initialize θ_{cap} based on D_T with Eq (1);
5: **repeat**
6:     select mini-batch $(I^{(i)}, d_T^{(i)}) \in D_T$;
7:     generate $s_s^{(i)}$ for $I^{(i)}$ via Monte-Carlo sampling;
8:     generate $s_b^{(i)}$ for $I^{(i)}$ via greedy search;
9:     compute fluency self-critic loss for $s_s^{(i)}$ by Eq (5);
10:     compute relevancy self-critic loss for $s_s^{(i)}$ by Eq (10);
11:     update θ_{cap} with Eq (2);
12: **until** θ_{cap} converges

---

4 EXPERIMENTS

We evaluate the unpaired cross-lingual image captioning models in both English and Chinese languages. For unpaired English image captioning, we utilize Chinese as pivot; while for unpaired Chinese image captioning, we utilize English as pivot.

4.1 Evaluation Setting

Datasets. We conduct experiments on the MSCOCO [21] and AIC-ICC [29] image caption datasets in this work. The MSCOCO dataset is annotated in English, which consists of 123,287 images and 5 manually labeled English captions for each image. We follow the public split [21] which utilizes 113,287 images for training, 5,000 images for validation and 5,000 images for testing. The AIC-ICC (Image Chinese Captioning from AI Challenge) dataset contains 238,354 images and 5 manually annotated Chinese captions for each image. There are 208,354 and 30,000 images in the official training and validation set in AI challenge. Since annotations of the testing set are unavailable in the AIC-ICC dataset, we randomly sample 5,000 images from its validation set as our testing set. We use "Jieba" [2] to tokenize Chinese captions. The words with frequency more than 4 are added to our vocabulary. We truncate English captions longer than 20 and Chinese captions longer than 16. The statistics of the two datasets are presented in Table 1.

Table 1: Statistics of the datasets used in our experiments.

Dataset	Lang.	# Images	# Captions	# Vocabulary
AIC-ICC	zh	240K	1200K	7,654
MSCOCO	en	123K	615K	10,368

For unpaired English image captioning, the task is to generate captions in English for images from MSCOCO dataset while no English image-caption pairs are used. In this setting, we use Chinese as the pivot language and utilize the AIC-ICC Chinese image caption dataset. For unpaired Chinese image captioning, the task is to generate captions in Chinese for images from AIC-ICC dataset while no Chinese image-caption pairs are used. In this setting, we use English as the pivot language and utilize the MSCOCO English image caption dataset.

Compared Methods. We compare our proposed model with the following four baseline models:

- Baseline: The vanilla captioning model [28] trained on pseudo pairs D_T with cross-entropy in Eq (1) without any rewards.
- Baseline+: The vanilla captioning model trained on pseudo pairs D_T with CIDEr as reward in the reinforcement learning framework [25].
- 2-Stage pivot Google model [11]: It utilizes a two-stage pipeline for unpaired cross-lingual image captioning. The image-to-pivot captioning model is the vanilla caption model [28] and the pivot-to-target MT model is the online Google translator.
- 2-Stage pivot joint model [11]: It utilizes a two-stage pipeline, including a image-to-pivot captioning model and pivot-to-target MT model. The two models share the same word embedding and are jointly trained to alleviate translation errors on the image caption domain.

---

[2] https://github.com/fxsjy/jieba

Metrics. We utilize the standard caption evaluation metrics to assess the quality of caption sentences, including BLEU [24], METEOR [7] and CIDEr [27]. As an image tells a thousand words, above objective evaluation metrics may not be able to fully measure the caption quality from different aspects. We therefore carry out human evaluation to further assess the caption quality from the fluency and visual relevancy aspects.

4.2 Implementation Details

Image Captioning Model. We extract activations from the last pooling layer of ResNet-101 [13] which is pre-trained on ImageNet as our image features. We encode the image feature to a 512-dimensional vector to initialize the hidden state of LSTM decoder. The LSTM decoder contains 1 layer with 512 hidden units. The dimensionality of the word embedding is set as 512. We use the special token <BOS> and <EOS> to represent the beginning and ending of sentences. At test time, a beam-search decoding with beam size of 10 is used to generate captions. We use the state-of-the-art Baidu translation API [3] as our translation model $f_{P \to T}$.

Language Model for Fluency Rewards. For unpaired English image captioning, we use texts in the MSCOCO training set to train the English language model. For unpaired Chinese image captioning, we use texts in the AIC-ICC training set to train the Chinese language model. However, the mono-lingual corpus is not subject to these datasets. We do an ablation study in Section 4.4 to compare the performance of our SSR model with the language model trained on different corpus. The language model is a one-layer LSTM with 512 hidden units. After training, the language model is fixed to evaluate the fluency of target captions.

ML-VSE for Relevancy Rewards. For the image-sentence matching model, we use ResNet-101 [13] pre-trained on ImageNet as the CNN image encoder and one-layer bi-directional GRU with 512 hidden units as the sentence encoder. The dimensionality of image-sentence joint space is set to be 1024. For the image-concept matching model, we extract nouns and verbs as visual concepts from the translated captions via stanford parsing tools [4]. In total, there are 3,231 visual concepts for unpaired English image captioning and 9,107 visual concepts for unpaired Chinese image captioning. The dimensionality of image-concept joint space is set as 512.

Training Details. We pre-train the image captioning model, language model and ML-VSE model using Adam optimizer [18] with a batch size of 128. For the image captioning model, the initial learning rate is 4e-4, while for the language and ML-VSE model, the initial learning rate is 2e-4. In the self-critical reinforcement training, we set the learning rate as 4e-5 and batch size of 256. We set hyper-parameters α, β, γ and λ to 0.05, 0.15, 1.0 and 0.5 respectively. A dropout of 0.3 is applied to all models during training to prevent over-fitting.

4.3 Comparison with the State-of-the-arts

Table 2 presents the unpaired cross-lingual image captioning performance in English and Chinese languages from our proposed approach and the compared baselines. The proposed self-supervised rewarding (SSR) model achieves the best performance among all

---

methods across different languages and evaluation metrics. The "Baseline" method trained with imperfect pseudo pairs is inferior to all other methods. It demonstrates that translation errors in pseudo pairs can significantly deteriorate the captioning performance even if we have utilized the state-of-the-art translation model. In the "Baseline+" method, although self-critical reinforcement learning algorithm is employed to train the model, the improvements over "Baseline" method is marginal since it directly utilizes the noisy translated captions to provide rewards. Our model instead is enhanced with fluency rewards and both coarse- and fine-grained visual relevancy rewards in the reinforcement learning framework. The comparison of our model with "Baseline+" proves that the contribution mainly comes from the proposed self-supervised rewards rather than the "self-critical" reinforcement training.

Our approach also outperforms the 2-stage models in Gu *et al.* [11]. The 2-Stage pivot Google model takes the advantage of the state-of-the-art translation model but ignores the translation errors for unpaired image caption generation. The 2-Stage pivot joint model addresses the translation domain mismatch by joint training but cannot generalize to using the state-of-the-art translation model. To be noted, our model is also more efficient than the 2-stage models in the testing phase since we do not depend on the 2-stage pipeline for caption generation in the target language.

4.4 Ablation Studies

Contributions of different rewards. In Table 3, we ablate the unpaired captioning performance on different self-supervised rewards. The fluency reward alone improves the baseline method on both unpaired English and Chinese image captioning, which demonstrates that the proposed fluency reward can effectively improve the quality of generated caption sentences. However, the fluency reward only promotes the fluency of sentence without considering the visual relevancy. Combining the fluency reward with both sentence- and concept-level visual relevancy rewards achieves additional performance gains on both languages. We notice that the improvements of visual relevancy rewards are larger on unpaired English image captioning than the unpaired Chinese image captioning. Since the diversity of images in the Chinese pivot language AIC-ICC dataset is smaller than that in the MSCOCO dataset, the unpaired English image captioning trained on pseudo image-caption pairs on the AIC-ICC dataset is more likely to suffer from visual irrelevancy problems. Therefore, our proposed visual relevancy rewards can benefit more for unpaired English image captioning.

Multi-level visual-semantic matching performance. We empirically evaluate the performance of ML-VSE model to demonstrate the reliability of the self-supervised relevancy rewards at the sentence- and concept-level. We take the ML-VSE model trained for English captioning as an example. We randomly select 1K images from the AIC-ICC validation set and MSCOCO testing set respectively to evaluate the performance of sentence-level semantic matching model, which is shown in Table 4. We notice that there exists a large performance gap between the MSCOCO testing set and AIC-ICC validation set, which can result from noises in pseudo pairs and image domain mismatch. Therefore, additional fine-grained relevancy reward is requisite. For the image-concept matching model, we visualize the top-10 predicted visual concepts

Table 2: Performance comparison with baseline methods for unpaired English image captioning evaluated on the MSCOCO dataset and unpaired Chinese image captioning evaluated on the AIC-ICC dataset.

Task	Method	Bleu@1	Bleu@2	Bleu@3	Bleu@4	Meteor	CIDEr
Unpaired English Image Captioning	Baseline	42.7	21.4	10.2	5.2	13.5	14.5
	Baseline+	44.0	22.0	10.5	5.3	13.0	14.6
	2-Stage pivot Google model [11]	42.2	21.8	10.7	5.3	**14.5**	17.0
	2-Stage pivot joint model [11]	46.2	24.0	11.2	5.4	13.2	17.7
	Our SSR	**52.0**	**30.0**	**17.9**	**11.1**	14.2	**28.2**
Unpaired Chinese Image Captioning	Baseline	41.1	23.9	13.0	7.1	21.1	11.5
	Baseline+	41.6	24.4	13.3	7.3	21.1	11.6
	Our SSR	**46.0**	**30.9**	**19.3**	**12.3**	**22.8**	**18.3**

Table 3: The contribution of different rewards for unpaired cross-lingual image captioning on MSCOCO and AIC-ICC datasets.

Task	Rewards	Bleu@1	Bleu@2	Bleu@3	Bleu@4	Meteor	CIDEr
Unpaired English Image Captioning	No Reward	42.7	21.4	10.2	5.2	13.5	14.5
	r_{flc}	45.9	23.4	11.4	5.8	13.4	16.1
	$r_{flc} + r_{srlv}$	50.6	28.7	17.1	10.6	13.8	26.7
	$r_{flc} + r_{srlv} + r_{crlv}$	**52.0**	**30.0**	**17.9**	**11.1**	**14.2**	**28.2**
Unpaired Chinese Image Captioning	No Reward	41.1	23.9	13.0	7.1	21.1	11.5
	r_{flc}	45.8	30.3	18.6	11.6	22.5	18.0
	$r_{flc} + r_{srlv}$	**46.1**	30.7	19.1	12.1	22.6	**18.5**
	$r_{flc} + r_{srlv} + r_{crlv}$	46.0	**30.9**	**19.3**	**12.3**	**22.8**	18.3

Table 4: Cross-modal retrieval performance using the proposed sentence-level semantic matching model trained on the self-supervised pseudo English pairs on the AIC-ICC training set. R@k represents recall in top k for the cross-modal retrieval.

	Image-to-Sentence		Sentence-to-Image	
	R@1	R@10	R@1	R@10
AIC-ICC val	52.8	85.9	37.7	81.2
MSCOCO test	22.7	58.7	12.8	48.7

Figure 3: Top-10 predicted concepts for examples in MSCOCO test set.

for some images in the MSCOCO testing set in Figure 3. As we can see, the predicted visual concepts are highly relevant to the image content, which cover diverse aspects such as object, action and scene. Both results demonstrate the validity of our proposed relevancy guidance.

Table 5: English Image Captioning performance with language model trained on different mono-lingual corpus.

Corpus	# Sents	B@3	B@4	Meteor	CIDEr
MSCOCO	565K	17.9	11.1	14.2	28.2
AIC-MT	483K	14.4	8.2	13.6	25.6

Language model trained on different mono-lingual corpus. Although we utilize in-domain target corpus to train the language model in Table 2, our SSR model can also benefit from other out-of-domain mono-lingual corpora which are easier to obtain in reality. In Table 5, we present the unpaired English captioning performance with the language model trained on an out-of-domain corpus from AIC-MT [5]. Though using the out-of-domain mono-lingual corpus is not as effective as using in-domain data, it still achieves significant improvements over baseline models in Table 2, which demonstrates the generalization ability of the proposed model to exploit different mono-lingual target corpora.

Comparison with paired target image captioning. Table 6 compares our proposed model with supervised mono-lingual image captioning models with different number of training pairs. We can see that the number of paired image-caption data is critical for the supervised image captioning model. Without sufficient pairs, the captioning performance drops significantly. Our model, however, relies on no supervised image-caption pairs, but achieves performance comparable to the supervised mono-lingual captioning model with 4,000 pairs.

---

[5]https://challenger.ai/competition/ect2018

Baseline+ There is a man in sportswear playing football on the court.
SSR There are a group of children playing football on the grass.

Baseline+ There is a man with glasses in the room standing at the table.
SSR There is a man standing in front of a bookshelf.

Baseline+ In the bright room sat a cat with a cat in his left hand.
SSR There is a cat sitting in the bathtub.

Baseline+ A woman with her hands in her arms stood in the bright room.
SSR A woman in a dress is standing in front of a mirror.

Baseline+ 两个足球运动员在球场上踢足球 (Two football players are playing football on the court)
SSR 一个足球运动员在比赛中摔倒了 (A football player fell down in the match)

Baseline+ 一个穿着蓝色衣服的女人站在一个白色的花瓶里 (A woman in blue is standing in a white vase)
SSR 一个穿着白色衣服的女人站在一个花瓶旁边 (A woman in white is standing next to a vase)

Baseline+ 一个男人和一个女人在一个有两个麦克风的房间里 (A man and a woman in a room with two microphones)
SSR 一个男人和一个女人站在桌子旁交谈 (A man and a woman are talking at the table)

Baseline+ 一个小男孩在树林里玩飞盘 (A little boy is playing frisbee in the woods)
SSR 一个穿着蓝色衣服的小女孩坐在草地上 (A little girl in blue is sitting on the grass)

Figure 4: Examples of the English image captioning from the MSCOCO testing set, and Chinese image captioning from the AIC-ICC testing test. The errors in generated captions are marked in red.

Table 6: Comparison between unpaired English image captioning and supervised English image captioning with different number of training pairs from MSCOCO dataset.

Approach	# Imgs	# Caps	B@4	Meteor	CIDEr
Baseline [28]	82,783	414,113	27.7	23.3	83.9
	40,000	40,000	24.2	21.8	71.0
	10,000	10,000	20.6	18.8	54.6
	4,000	4,000	14.0	14.2	28.5
	3,000	3,000	10.7	12.6	19.1
Our SSR	**0**	**0**	**11.1**	**14.2**	**28.2**

4.5 Human Evaluation and Qualitative Results

Besides the quantitative evaluations in section 4.3, we also conduct human evaluation to verify the effectiveness of the proposed SSR model. We take the unpaired English image captioning as an example. We randomly select 1,000 images from the MSCOCO testing set, and recruit 10 workers who have sufficient English skills to evaluate the quality of generated captions from the "Baseline+" model and our SSR model. Particularly, we measure the caption quality in the fluency and relevancy aspects. The fluency levels consist of 1-very poor, 2-poor, 3-barely fluent, 4-fluent, and 5-human like, and the relevancy levels consist of 1-irrelevant, 2-basically irrelevant, 3-partial relevant, 4-relevant, and 5-completely relevant. Results in Table 7 demonstrate that our approach can generate more fluent and visually relevant captions than the baseline model with the guidance of self-supervised rewards. The example visualization results in Figure 4 for both English and Chinese image captioning also confirm this.

Table 7: Human evaluation results on the MSCOCO 1K test.

Measure	Baseline+ model	Our SSR model
Fluency	4.1	4.8
Relevancy	3.3	3.8

5 CONCLUSIONS

In this paper, we propose a novel language-pivoted approach for unpaired cross-lingual image captioning. Previous language-pivoted methods mainly suffer from translation errors brought about by the pivot-to-target translation model, such as disfluency and visually irrelevancy errors. We propose to alleviate negative effects from such errors by providing fluency and visual relevancy rewards as guidance in the reinforcement learning framework. We employ self-supervisions from mono-lingual sentence corpus and machine translated image-caption pairs to obtain the reward functions. Extensive experiments with both objective and human evaluations on both unpaired English and Chinese image captioning tasks demonstrate the effectiveness of the proposed approach.

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*.

[2] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled Sampling for Sequence Prediction with Recurrent Neural Networks. In *NIPS*.

[3] Ali Furkan Biten, Lluis Gomez, Marçal Rusiñol, and Dimosthenis Karatzas. 2019. Good News, Everyone! Context driven entity-aware captioning for news images. In *CVPR*.

[4] Shizhe Chen, Qin Jin, and Jianlong Fu. 2019. From Words to Sentences: A Progressive Learning Approach for Zero-resource Machine Translation with Visual Pivots. In *IJCAI*.

[5] Shizhe Chen, Qin Jin, and Alexander Hauptmann. 2019. Unsupervised Bilingual Lexicon Induction from Mono-lingual Multimodal Data. In *AAAI*.

[6] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

[7] Michael Denkowski and Alon Lavie. 2014. Meteor Universal: Language Specific Translation Evaluation for Any Target Language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 376–380. https://doi.org/10.3115/v1/W14-3348

[8] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *spotlight presentation at British Machine Vision Conference (BMVC)*. https://arxiv.org/abs/1707.05612?context=cs.CV

[9] Hao Fang, Saurabh Gupta, Forrest Iandola, Rupesh Srivastava, Li Deng, Piotr Dollár, Jianfeng Gao, Xiaodong He, Margaret Mitchell, John C. Platt, C. Lawrence Zitnick, and Geoffrey Zweig. 2015. From Captions to Visual Concepts and Back. In *CVPR*.

[10] Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2018. Unsupervised Image Captioning. https://arxiv.org/abs/1811.10787

[11] Jiuxiang Gu, Shafiq Joty, Jianfei Cai, and Gang Wang. 2018. Unpaired Image Captioning by Language Pivoting. In *Computer Vision – ECCV 2018*. Springer International Publishing, Cham, 519–535.

[12] Jiuxiang Gu, Gang Wang, Jianfei Cai, and Tsuhan Chen. 2017. An Empirical Study of Language CNN for Image Captioning. In *ICCV*. 10.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *CVPR*.

[14] Julian Hitschler, Shigehiko Schamoni, and Stefan Riezler. 2016. Multimodal Pivots for Image Caption Translation. In *ACL*.

[15] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-term Memory. *Neural computation* 9 (12 1997), 1735–80. https://doi.org/10.1162/neco.1997.9.8.1735

[16] Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars. 2015. Guiding Long-Short Term Memory for Image Caption Generation. In *ICCV*.

[17] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. (2014). https://arxiv.org/abs/1408.5882

[18] Diederik Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. https://arxiv.org/abs/1602.07332

[20] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-Guided Cross-Lingual Image Captioning. In *ACM Multimedia*. ACM, 9. https://doi.org/10.1145/3123266.3123366

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

[22] Daqing Liu, Zheng-Jun Zha, Hanwang Zhang, Yongdong Zhang, and Feng Wu. 2018. Context-Aware Visual Policy Network for Sequence-Level Image Captioning. In *ACM Multimedia*. ACM, 9. https://doi.org/10.1145/3240508.3240632

[23] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[24] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.

[25] Steven J. Rennie, Etienne Marcheret, and Youssef Mroueh et al. 2017. Self-Critical Sequence Training for Image Captioning. In *CVPR*.

[26] Satoshi Tsutsui and David Crandall. 2017. Using Artificial Tokens to Control Languages for Multilingual Image Caption Generation. (2017).

[27] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*.

[28] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and Tell: A Neural Image Caption Generator. In *CVPR*. IEEE. https://doi.org/10.1109/CVPR.2015.7298935

[29] Jiahong Wu, He Zheng, Bo Zhao, Yixin Li, Baoming Yan, Rui Liang, Wenjia Wang, Shipei Zhou, Guosen Lin, Yanwei Fu, Yizhou Wang, and Yonggang Wang. 2017. AI Challenger : A Large-scale Dataset for Going Deeper in Image Understanding. *CoRR* abs/1711.06475 (2017).

[30] Jing Shao Dapeng Chen Xiaogang Wang Xihui Liu, Hongsheng Li. 2018. Show, Tell and Discriminate: Image Captioning by Self-retrieval with Partially Labeled Data. In *ECCV*.

[31] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. abs/1502.03044 (2015). http://dblp.uni-trier.de/db/journals/corr/corr1502.html#XuBKCCSZB15

[32] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image Captioning with Semantic Attention. In *CVPR*.

[33] Wendong Zhang, Bingbing Ni, Yichao Yan, Jingwei Xu, and Xiaokang Yang. 2017. Depth Structure Preserving Scene Image Generation. In *ACM Multimedia*. ACM.