Dual Learning for Cross-domain Image Captioning

Conference Paper - November 2017
DOI: 10.1145/3132847.3132920

7 authors, including:

Jianbo Ye
Pennsylvania State University
28 PUBLICATIONS 125 CITATIONS

Yu Qiao
Chinese Academy of Sciences
161 PUBLICATIONS 4,863 CITATIONS

Some of the authors of this publication are also working on these related projects:

- Scene Text Understanding View project
- EmotionW 2017 View project

All content following this page was uploaded by Jianbo Ye on 13 March 2018.

The user has requested enhancement of the downloaded file.
Dual Learning for Cross-domain Image Captioning

Wei Zhao†‡
SIAT, Chinese Academy of Sciences
wei.zhao@siat.ac.cn

Wei Xu
Tencent
weizierxu@tencent.com

Min Yang§
SIAT, Chinese Academy of Sciences
min.yang1129@gmail.com

Jianbo Ye
The Pennsylvania State University
jxy198@ist.psu.edu

Zhou Zhao
Zhejiang University
zhaozhou@zju.edu.cn

Yabing Feng
Tencent
jessefeng@tencent.com

Yu Qiao
SIAT, Chinese Academy of Sciences
yu.qiao@siat.ac.cn

ABSTRACT
Recent AI research has witnessed increasing interests in automatically generating image descriptions in text, which is coined as the image captioning problem. Significant progresses have been made in domains where plenty of labeled training data (i.e. image-text pairs) are readily available or collected. However, obtaining rich annotated data is a time-consuming and expensive process, creating a substantial barrier for applying image captioning methods to a new domain. In this paper, we propose a cross-domain image captioning approach that uses a novel dual learning mechanism to overcome this barrier. First, we model the alignment between the neural representations of images and that of natural languages in the source domain where one can access sufficient labeled data. Second, we adjust the pre-trained model based on examining limited data (or unpaired data) in the target domain. In particular, we introduce a dual learning mechanism with a policy gradient method that generates highly rewarded captions. The mechanism simultaneously optimizes two coupled objectives: generating image descriptions in text and generating plausible images from text descriptions, with the hope that by explicitly exploiting their coupled relation, one can safeguard the performance of image captioning in the target domain. To verify the effectiveness of our model, we use MSCOCO dataset as the source domain and two other datasets (Oxford-102 and Flickr30k) as the target domains. The experimental results show that our model consistently outperform previous methods for cross-domain image captioning.

†Min Yang is the corresponding author.
‡Also with Shenzhen College of Advanced Technology, University of Chinese Academy of Sciences.
§Also with Tencent.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CCS CONCEPTS
• Computing methodologies → Computer vision; Reinforcement learning;

KEYWORDS
Image captioning, Dual learning, Reinforcement learning, Image synthesis

1 INTRODUCTION
Automatic image captioning involves analyzing the visual content of an input image, and generating a textual description (typically a sentence) that verbalizes its most salient aspects [3]. Unlike other computer vision tasks such as image classification and object detection, image captioning requires machine not only understand what facts are presented in an image, but also possess the linguistic capability to describe what it sees. Developing image captioning methodologies connects researches from both the community of computer vision and the community of natural language processing. Its success lies in how to extend the intelligent ability of machines by developing machine learning based visual-language models. Those models should characterize the shared semantics among image and text by jointly learning the multi-modal structure of data. Earlier image captioning methods rely on either using sentence templates [7, 16] or reformulating this problem as a retrieval task such that through ranking candidate sentences from a database, one can select the best matching sentence as the caption [8, 12]. Those approaches usually suffer difficulties in generating sentences at variable lengths to cope with the varied complexity of image contents, or crafting new sentences that are faithfully reflecting what they suppose to describe.

Inspired by the success of the encoder-decoder framework [25] in neural machine translation, most recent image captioning methods employ the encoder-decoder framework to generate image captions [6, 9, 11, 30]. The common idea of these methods is to use a convolutional neural networks (CNN) as an encoder to represent an input image as a compact vector, and then feed this representation to a recurrent neural network (RNN) decoder so as to generate the image descriptions. The encoder-decoder based image captioning methods have become the mainstream mainly because it
enables one to train neural model in an end-to-end manner with the potential that scales to very large training data. Impressive image captioning performances were achieved in domains where plenty of training image and sentence pairs (e.g., MSCOCO) are available. However, such fruitful results are subject to an assumption: the test data should be drawn from the same distribution as the training data. [5] showed traditional image captioning methods are struggling to cope well with generating image captions across different data domains. The model trained on MSCOCO that consists of images of large scenes, is difficult to be generalized to the Oxford-102 that consists of cropped birds images. In a new domain where its own labeled data is in short supply, obtaining more labels is usually expensive and time-consuming. Meanwhile, there may exist plenty of unpaired data where the text descriptions are from the web (e.g., Wikipedia) [1, 5].

Most previous studies based on the encoder-decoder architectures train their neural networks to generate the captions by maximizing the likelihood estimation of the joint image-text distribution. Yet, this strategy has two main shortcomings in generating captions. First, the evaluation metric is different from the training loss. For example, in image captioning systems, the encoder-decoder models are trained using the cross-entropy loss but they are typically evaluated at test time using discrete and non-differentiable metrics such as BLEU [19], ROUGE [13], CIDEr [29]. Second, as discussed in [20], the input of the decoder in each time step is often the previous ground-truth word during training. Nevertheless, when generating captions in the testing phase, the input of the next time step is the previous word generated by the decoder. This exposure bias [20] leads to error accumulation at test time. Once the model generates a “bad” word, the error will propagate and accumulate with the length of the sequence.

To address the aforementioned issues, we propose a novel dual learning method for cross-domain image captioning, which aims to leverage the data from both source and target domains effectively. First, we use an encoder-decoder (i.e., CNN-RNN) model to learn the alignment between the representations of images and natural language texts in the source domain, called the pre-training procedure. At the same time, we also train a text-based image generation model that synthesizes a plausible image by reading a meaningful sentence. We expect that the pre-training gives a good initialization of the parameters, and therefore training at the latter stage gives a good generalization performance even if the size of the target domain dataset is limited. Second, we fine-tune both models on the limited paired data (or unpaired data) in the target domain via a dual learning mechanism [32]. Unlike previous work for image captioning, our dual learning procedure simultaneously optimizes the objectives of two coupled tasks: generating the captions for images and generating plausible images from text descriptions, and so the relation of these two tasks can be exploited to improve the performance of image captioning in the target domain. Specifically, we treat image captioning as the primal task, and treat the task of generating image from text descriptions as the dual task. The primal and dual tasks form a closed loop, whose objectives are optimized together to guide each other by employing a reinforcement learning process (i.e., policy gradient) [26, 27]. Such fine-tuning optimization procedure can be done in semi-supervised and unsupervised adaptation settings. In particular, when there is no paired training data in the target domain, we update the two models with policy gradient methods based on their reconstruction rewards; otherwise, we also include the evaluation metrics as feedback signals such as BLEU and CIDEr. By this way, we can adequately leverage both paired and unpaired data in the target domain.

We summarize our main contributions as follows:

- Our method integrates the utility of the encoder-decoder framework in learning the visual concepts of images and compositional semantic meanings of texts. Our method uses reinforcement learning to optimize for highly rewarded captions. Thus, it effectively bypasses exposure bias and non-differentiable task metric issues.
- To the best of our knowledge, we are the first to use dual learning to simultaneously optimize the two tasks: generating captions for images and generating plausible images from text descriptions. Actually, these two tasks are tightly related and their correlation can be explicitly exploited from a modeling perspective, enjoying several performance advantages especially in the cross-domain setting. What’s more, the dual learning mechanism allows to leverage both paired and unpaired target domain data effectively and flexibly. Thus, our model can be trained in both the semi-supervised and the unsupervised adaptation settings by fine-tuning against different rewards.
- To verify the effectiveness of our model on cross-domain image captioning, we use MSCOCO data as the source domain and two other datasets (Oxford-102 and Flickr30k) as the target domains. The experimental results show that our model consistently outperforms previous methods.

The rest of the paper is organized as follows. In Section 2, we discuss the related work. Section 3 presents our dual learning mechanism for image caption via domain adaptation in details. In Section 4, we describe the experimental data and implementation details. We demonstrate and analyze the experimental results in Section 5. Section 6 concludes this paper and indicates the future work.

## 2 RELATED WORK

Generating image captions from images is a challenging problem that has been receiving much attention from the computer vision and natural language processing communities recent years. Bernardi et al. [3] provided a detailed review of most existing approaches, the benchmark datasets, and the evaluation measures for image captioning. Early methods for image captioning either explored template-based approaches [7, 16] or retrieval-based approaches [8, 12]. These models were usually heavily relied on hand-designed features or templates, and it is hard for them to generate novel sentences with new compositions. Recent advances in deep neural networks have substantially improved the performance of image captioning task. A typical image captioning strategy is to combine CNN and RNN [6, 9, 11, 30], where CNN is used to extract the compact representational vector of a whole image, and RNN is used to construct the language model operated on the representation vectors to generate captions. Visual attention has been proven as an effective way to improve
the basic encoder-decoder framework. For example, Xu et al. [33] introduced an attention based model that can automatically learn where to attend when generating image descriptions. The attention is modeled as spatial probabilities that re-weight the feature map of the last convolutional layer in the CNN. Chen et al. [4] dynamically modulated the sentence generation context in multi-layer feature maps, encoding where (i.e., attentive spatial locations at multiple layers) and what (i.e., attentive channels) the visual attention was.

There have been increasing interests in integrating the encoder-decoder framework and reinforcement learning paradigms for image captioning [5, 15, 23]. For example, Liu et al. [15] employed policy gradient (PG) method to directly optimize a linear combination of SPICE and CIDEr metrics, where the SPICE score ensured the captions were semantically faithful to the image, and CIDEr score ensured the captions are syntactically fluent. Rennie et al. [23] proposed a self-critical sequence training (SCST) method by employing the popular REINFORCE algorithm. Instead of estimating a “baseline” to normalize the rewards and reduce variance, SCST utilizes the output of its own test-time inference algorithm to normalize the rewards it experiences.

Recently, domain adaptation techniques have been applied to sequence generation in many applications such as dialogue systems, question answering, and image captioning [5, 17, 36]. Zhang et al. [36] extended the traditional encoder-decoder approach to the personalized response generation by introducing a domain adaptation scheme, which they named as initialization-then-adaptation. Mo et al. [17] proposed a transfer learning framework based on POMDP to learn a personalized dialogue system. The system first learned common dialogue knowledge from the source domain and then adapted this knowledge to the target user. Chen et al. [5] propose an adversarial training procedure for cross-domain image captioning by leverage unpaired data in the target domain. During training, the critics and captioner act as adversaries: captioner aims to generate indistinguishable sentences, whereas critics aim at distinguishing them.

In parallel to our work, dual learning mechanism has been proposed to improve the neural machine translation (NMT) [32], which enables the NMT system to automatically learn from unlabeled data. In the dual-learning mechanism, they use one agent to represent the model for the primal task (English-to-French translation) and the other agent to represent the model for the dual task (French-to-English translation), then let them teach each other through a reinforcement learning process. Based on the reward signals generated during this phase, the two models can be iteratively updated until convergence. In [34], Yi et al. developed a novel dualGAN mechanism, which enabled image translators to be trained from the other direction. The closed loop made by the primal and dual tasks allows images from either domain to be translated and then reconstructed.

In the new language-vision community, generating images from text descriptions has also emerged as a key task. For instance, Reed et al. [22] employed a Generative adversarial networks (GAN) architecture to generate plausible images based on detailed visual descriptions. This is the first end-to-end differentiable architecture from the character level to pixel level. Oord et al. [28] explored conditional image generation with a new image density model based on the PixelCNN architecture. The model can be conditioned on any vector, including descriptive labels or tags, or latent embeddings created by other networks to generate images.

The most related work to ours is [5], which uses adversarial training procedure to leverage unpaired data in the target domain. Our approach differs from [5] in several aspects. First, we perform cross-domain image captioning with a dual learning mechanism, which can leverage both paired and unpaired target domain data effectively. Second, our dual learning mechanism simultaneously optimizes two tasks: generating captions for images and generating plausible images from text descriptions, the agents of which are not adversaries.

3 MODEL

In this section, we elaborate our cross-domain image captioning model.

We use $D^s$ to denote the data in source domain with paired image $x$ and the ground truth sentence $y$ describing $x$. In the target domain, we are given a set of limited paired data $D^t$, and two separate sets of images and texts $D^t_{xy} = \{ (x)_M, (y)_N \}$, where $M, N$ are the number of unpaired images and texts.

Suppose we have two agents (image captioning agent $A$ and image synthesis agent $B$) that can generate image captions and generate plausible images from text descriptions, respectively. These two agents are pre-trained on the image-text pairs and text-image pairs in source domain ($D^s$). Our goal is to simultaneously fine-tune the image caption agent $A$ and the image synthesis agent $B$ via reinforcement learning on the target domain with limited paired data.

3.1 Model pre-training

3.1.1 Image captioning. CNN-RNN-based image captioning method is originally described in [6, 9], which first uses a CNN encoder to summarize an input raw image as a vector representation, then feeds this representation into an RNN decoder to generate the sentence describing the image. In this paper, we add an attention model to automatically learn where to look depending on the generated words. We use long short-term memory (LSTM) cell as the basic RNN unit.

**CNN encoder.** Similar to [9, 30], we first use a deep CNN to encode the input image $x$ into $L$ vectors, each of which is a $D$-dimensional vector corresponding to the features extracted at different locations of the image. These vectors are referred to as annotation vectors

$$z = \{ z_1, z_2, \ldots, z_L \}$$  \hspace{1cm} (1)

In this paper, we use the Oxford VGGnet-19 [24] pre-trained on ImageNet without fine tuning as our CNN encoder to create the annotations $z$ used by our decoder. Following [33], we extract feature vectors from a lower convolutional rather than use a fully connected layer. In this way, the decoder can selectively attend to certain parts of an image by weighting a subset of the feature vectors. In our experiments, we use the $14 \times 14 \times 512$ feature map of the fifth convolutional layer before max pooling. Finally, our LSTM
decoder operates on the flattened 196 × 512 (i.e., L × D) representations.

LSTM decoder. We use a long short-term memory (LSTM) network that produces a caption by generating one word at every time step conditioned on a context vector, the previous hidden state and the previously generated words.

The generation probability of the $t$-th word is calculated by

$$P(y_t|y_1, \ldots, y_{t-1}, x) = g(s_t, y_{t-1}, c_t)$$

where $g$ is a nonlinear function, $c_t$ is the context vector at time step $t$, and $s_t$ is the hidden state of LSTM decoder at time $t$, computed by

$$s_t = f(y_{t-1}, s_{t-1}, c_t)$$

where $f$ is LSTM cell.

The context vector $c_t$ acts as an extra input into the computation of the hidden states in LSTM decoder to make sure that every time step of the decoder can get full information of the context. The context vector $c_t$ is a dynamic representation of the relevant part of the input image at time $t$. We calculate the context vector $c_t$ when we decode the $t$-th word in the LSTM decoder from the annotation vectors $z = [z_1, z_2, \ldots, z_L]$ corresponding to the features extracted at different image locations by

$$c_t = \sum_{i=1}^{L} \alpha_{t,i} z_i$$

Here, the weight $\alpha_{t,i}$ for the $i$-th annotation of CNN encoder is computed by

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{k=1}^{L} \exp(e_{t,k})}$$

$$e_{t,i} = \sigma(s_{t-1}, z_i)$$

where $\sigma$ is a feed-forward neural network, which maps a vector to a real-valued score. This attention weights $\alpha_{t,i}$ models the alignment between the image content at location $i$ and the output word at position $t$. It allows the decoder to dynamically select and linearly combine different locations of the input image, where weighting factor $\alpha_{t,i}$ determine which part should be selected to generate the new word $y_t$.

3.1.2 Image synthesis. We use the deep convolutional generative adversarial networks (DC-GAN) [22] to implement a basic text-based image synthesis system. Like the standard generative adversarial networks (GANs), DC-GAN also consists of a generator $G$ and a discriminator $D$. Given a text description $y$, the generator tries to generate plausible image $\hat{x}$; while the discriminator evaluates how well an image matches the text description.

Generator $G$. We first sample from the noise prior $z \sim \mathcal{N}(0, 1)$ and we encode the text description $y$ using text encoder $\phi$. The text encoder $\phi$ is implemented by a hybrid character-level convolutional recurrent neural network [21]. The encoded text description is then concatenated to the noise vector $z$. We feed-forward this vector through the generator $G$, and a synthetic image $\hat{x}$ is generated via $\hat{x} \leftarrow G(z, y)$.

Discriminator $D$. The discriminator is implemented as a convolutional neural network. The text-image pairs are viewed as joint observation, and the discriminator is trained to judge the pairs as real or fake. By learning to optimize image/text matching in addition to the image realism, the discriminator can provide an additional signal to the generator. The readers can refer to [22] for the implementation details.

The optimization process can be formalized into a minimax problem as follows:

$$\min_{\theta_B} \max_{\theta_G} \mathbb{E}_{y \sim \mathcal{Y}} \mathbb{E}_{x \sim \mathcal{X}} [\log D_{\theta_B}(y, x)] + \mathbb{E}_{z \sim \mathcal{N}} [\log (1 - D_{\theta_B}(G_{\theta_G}(z), y))]$$

Here, $G_{\theta_G}$ and $D_{\theta_B}$ denote the generator with parameter $\theta_G$ and the discriminator with parameter $\theta_B$. $\mathcal{Y}$ denotes the text descriptions provided in the training set, $\mathcal{N}$ denotes a standard normal
distribution, and $G_{θ_B}(z, y)$ denotes the image generated with $y$ and $z$.

The overall learning procedure alternates between the updating of $G$ and $D$, until they reach an equilibrium.

### 3.2 Model adaptation

In the pre-training procedure, the image captioning agent $A$ and the image synthesis agent $B$ are trained to learn the general alignment between the representations of images and that of natural languages in source domain data $D^s$. Inspired by the success of dual learning in [32], we use dual learning mechanism to simultaneously optimize two tasks: generating captions for images and generating plausible images from text descriptions, and so the relations of these two tasks can be exploited to further improve the performance of image captioning. We describe the model for the game beginning from image captioning agent $A$ in details below:

- We first generate a middle caption $y_{mid}$ for image $x$ in target domain with image captioning agent $A$. We define the agent $A$ as LSTM decoder $P(y_{mid}|x; θ_A)$, where $θ_A$ is its parameters (as described in Section 3.1.1).
- Then, we further generate the image $x$ back from middle caption $y_{mid}$ with the image synthesis agent $B$, denoted as $x'$. We define the agent $B$ as generator $G_{θ_B}(z, y_{mid})$, where $θ_B$ is its parameters and $z$ is the noise prior (as described in Section 3.1.2).
- By evaluating the results from this two-hop generation, we will get a sense of the quality of the image captioning agent $A$ and the image synthesis agent $B$, and be able to improve them accordingly. This process can be iterated for many round until both agents converge via policy gradient algorithm.

Symmetrically, we also update the model for the game beginning from the image synthesis agent $B$:

- We first generate a middle image $x_{mid}$ for text description $y$ in target domain by GAN generator $G_{θ_B}(z, y)$.
- Then, we further generate the text description back (denoted as $y'$) from middle image $x_{mid}$ by $P(y'|x_{mid}; θ_A)$.
- By evaluating the results from this two-hop generation, we update agent $A$ and agent $B$ until both agents converge via policy gradient algorithm.

The overall architecture and dataflow of the dual learning mechanism are illustrated in Figure 1.

#### 3.2.1 Dual learning with policy gradient

In this section, we propose policy gradient algorithm [26, 27] to optimize long-term rewards of the image captioning agent $A$ and the image synthesis agent $B$. We first introduce the state, action, policy and reward of the reinforcement learning architecture below.

**State.** A state $s_p$ for agent $A$ is the vector representation of the input image $x$; the state $s_d$ for agent $B$ is the vector representation of the input text description $y$.

**Policy.** We employ the stochastic policy gradient method to approximate a stochastic policy directly using an independent function approximator with its parameters. In this paper, the policy for the agent $A$ is the LSTM decoder (i.e. $P(y|x; θ_A)$) described in Section 3.1.1; the policy for the agent $B$ is the GAN generator $G_{θ_B}(z, y)$ (described in Section 3.1.2), whose inputs are the representation of the state, whose outputs are action selection probabilities.

**Action.** The action for agent $A$ is the caption $y$ generated from image $x$ by the policy $P(y|x; θ_A)$; the action for agent $B$ is the image $x$ generated from text description $y$ by policy $G_{θ_B}(z, y)$.

**Reward.** The policy gradient algorithm is a type of reinforcement learning methods, which relies upon optimizing parametrized policy with respect to the expected return (long-term cumulative reward) by gradient descent. We assume that our model receives a reward $r$ at each iteration. We discuss the major factors that contribute to the reward for agent $A$ and agent $B$ below, and denote the total reward as $r_A$ and $r_B$, respectively.

For a game beginning from the image captioning agent $A$, the rewards for agent $A$ are:

1. **Reconstruction reward.** For dual learning, given the generated middle caption $y_{mid}$ for image $x$ by agent $A P(y_{mid}|x; θ_A)$, we recover the image from $y_{mid}$ by GAN generator $G_{θ_B}(z, y_{mid})$, denoted as $x'$. Since there is no standard metric to measure the similarity of two images, we define the reconstruction reward for agent $A$ as the negative squared difference between $x$ and $x'$:
   \[
   r^{(rec)}_A = -∥x - x'∥^2
   \]  

2. **Evaluation metrics reward.** When there are paired data $(x, y) ∈ D^p$ in target domain, we also include the evaluation metrics reward for image captioning agent $A$. In this paper, we calculate the evaluation metrics reward as:
   \[
   r^{(eval)}_A = \frac{\text{BLEU}(y_{mid}, y) + \text{CIDEr}(y_{mid}, y)}{2}
   \]  

Finally, we simply adopt a linear combination of the reconstruction reward $r^{(rec)}_A$, and evaluation metrics reward $r^{(eval)}_A$ as the total reward for the image captioning agent $A$.

\[
\begin{align*}
    r_A &= λ^{(rec)}_A r^{(rec)}_A + λ^{(eval)}_A r^{(eval)}_A, \quad \text{if } (x, y) ∈ D^p \\
    &= λ^{(rec)}_A r^{(rec)}_A, \quad \text{if } (x, y) ∈ D^u \\
    \end{align*}
\]

where $λ^{(rec)}_A = λ^{(eval)}_A = 0.5$ and $λ^{(rec)}_A = 1.0$.

For a game beginning from the image synthesis agent $B$, the rewards for agent $B$ are:

1. **Reconstruction reward.** Given the generated middle image $x_{mid}$ for text description $y$ by the generator $G_{θ_B}(z, y)$, we use the log probability of the text description recovered back from $x_{mid}$ by $P(y|x_{mid}; θ_A)$ as the reward of the reconstruction. Mathematically, we define the reconstruction reward for agent $B$ as:
   \[
   r^{(rec)}_B = \log P(y|x_{mid}; θ_A)
   \]

2. **Similarity reward.** When there are paired text-image data $(x, y)$ in target domain, we measure the similarity between $(x, y) ∈ D^p$ and $(x', y') ∈ D^u$ as:
   \[
   r^{(sim)}_B = \frac{1}{(1 + ∥x - x'∥^2)}
   \]

where $x'$ and $y'$ are generated by $G_{θ_B}(z, y)$ and $P(y|x_{mid}; θ_A)$, respectively.
the generated image $x_{mid}$ and the ground truth $x$ as the negative squared difference of the two images:

$$ r_B^{(\text{sim})} = -||x - x_{mid}||^2 $$  \hspace{1cm} (12) 

Finally, we adopt a linear combination of the reconstruction reward $r_A^{(\text{rec})}$, language model reward $r_A^{(\text{LM})}$, and evaluation metrics reward $r_A^{(\text{eval})}$ as the total reward for image captioning agent $A$:

$$ r_B = \left\{ \begin{array}{ll}
\lambda_1^{(\text{rec})} B + \lambda_2^{(\text{sim})} B, & \text{if } (x, y) \in D^p \\
\lambda_1^{(\text{rec})} B, & \text{if } (x, y) \in D^u
\end{array} \right. $$  \hspace{1cm} (13) 

### 3.2.2 Updating model parameters. We can explore the state-action space and learn the policies that lead to the optimal expected reward $\mathbb{E}[r_A]$ and $\mathbb{E}[r_B]$. We firstly update the model for the game beginning from the image captioning agent $A$. According to the policy gradient theorem [27, 31], we compute the stochastic gradient of the expected reward $\mathbb{E}[r_A]$ with respect to parameters $\theta_A$ and $\theta_B$:

$$ \nabla_{\theta_A} \mathbb{E}[r_A] = \mathbb{E}[r_A \nabla_{\theta_A} \log P(y_{mid}|x; \theta_A)] $$  \hspace{1cm} (14) 

$$ \nabla_{\theta_B} \mathbb{E}[r_A] = \mathbb{E}(2\lambda_A (x - x') \nabla_{\theta_B} G_{\phi_B}(z, y_{mid})) $$  \hspace{1cm} (15) 

Symmetrically, we update the model for the game beginning from post agent $B$. We compute the stochastic gradient of expected reward $\mathbb{E}[r_B]$ with respect to parameters $\theta_B$ and $\theta_A$:

$$ \nabla_{\theta_B} \mathbb{E}[r_B] = \mathbb{E}[r_B \nabla_{\theta_B} \log P(y|x_{mid}; \theta_A)] $$  \hspace{1cm} (16) 

$$ \nabla_{\theta_A} \mathbb{E}[r_B] = \mathbb{E}(\lambda_B \nabla_{\theta_A} \log P(y|x_{mid}; \theta_A)) $$  \hspace{1cm} (17) 

Once we have a better generator $G_{\phi_B}$, we shall re-train the discriminator model $D_{\phi_B}$ as follows:

$$ \max J(\theta_B^d) = \max \mathbb{E}[\log D_{\phi_B}(x, y)] + \mathbb{E}[\log(1 - D_{\phi_B}(x, y_{mid}))] $$

$$ + \mathbb{E}[\log(1 - D_{\phi_B}(G_{\phi_B}(z, y), y))] $$  \hspace{1cm} (18) 

where the discriminator has three kinds of inputs: real image with matching text (i.e. $D_{\phi_B}(x, y)$), real image with mismatched text (i.e. $D_{\phi_B}(x, y_{mis})$), synthetic image with arbitrary text (i.e. $D_{\phi_B}(G_{\phi_B}(z, y), y)$).

The gradient of the objective function $J(\theta_B^d)$ with respect to $\theta_B^d$ can be derived as:

$$ \nabla_{\theta_B^d} J(\theta_B^d) = \mathbb{E}[\nabla \log D_{\phi_B}(x, y)] + \mathbb{E}[\nabla \log(1 - D_{\phi_B}(x, y_{mis}))] $$

$$ + \mathbb{E}[\nabla \log(1 - D_{\phi_B}(G_{\phi_B}(z, y), y))] $$  \hspace{1cm} (19) 

The reader can refer to [27, 31] for details of policy gradient theorem. After getting the stochastic gradient of $\mathbb{E}[r_A]$ and $\mathbb{E}[r_B]$ with respect to parameters $\theta_A$, $\theta_B^d$ and $\theta_B^d$, we use Adam [10] optimization algorithm to update the parameters. The game can be repeated for many rounds. In each round, we sample an image, generate its caption with agent $A$, then recover the image from the caption with agent $B$. Similarly, we also sample a sentence, generate an image with agent $B$, and recover the caption using agent $A$. We update the agent $A$ and agent $B$ according to the game beginning with the two agents respectively. The details of this process are given in Algorithm 1.
4 EXPERIMENTAL SETUP

4.1 Datasets

In the experiments, we use MSCOCO [14] as the source domain dataset, and use Oxford-102 [18] and Flickr30k [35] as the target domain datasets. The MSCOCO has regular domain shift with Flickr30k, and has large domain shift with Oxford-102. The detailed properties of these datasets are described as follow.

MSCOCO. This original MSCOCO dataset consists of 82,783 training, 40,504 validation, and 40,775 testing images, each of which has 5 ground truth captions. For fair comparison, we adopt the commonly used split proposed in [5, 9], which uses 113,287 images for training.

Flickr30k. This dataset consists of 31,783 images and each of them is annotated with 5 sentences. We follow the public accessible dataset divisions in [9]. In this dataset splits, 29,000/1,000/1,000 images are used for training/validation/testing respectively.

Oxford-102. This dataset consists of 8,189 images of flowers from 102 different categories. Similar to Flickr30k dataset, we choose 1000 images for testing, 1000 images for validation and the rest for training.

We preprocess the captions with the publicly available code. We convert all sentences to lowercase, and discard non-alphanumeric characters. For all the dataset, we prune the vocabulary that includes special Begin-Of-Sentence (BOS) and End-Of-Sentence (EOS) tokens by keeping the words whose frequency is more than 4. All the remaining words are replaced by a special token “UNK”.

4.2 Baseline methods

In the experiments, we evaluate and compare our model with several baseline methods, which we describe below:

Deep Compositional Captioner (DCC) model. This model builds on recent deep captioning models, which consists of a CNN model trained on unpaired images and an LSTM model trained on unpaired texts [1]. The overall DCC model then combines both models with a linear layer trained on paired image-caption data.

Show, attend and tell model (SAdT). This model presents two variants: a “hard” stochastic attention mechanism and a “soft” deterministic attention mechanism to improve the performance of CNN-LSTM based image captioning models [33].

Show, Adapt and Tell (SAdT). This model uses an adversarial training procedure to leverage unpaired data in the target domain [5]. A domain critic network assesses whether the generated sentences are indistinguishable from sentences in the target domain. A multi-modal critic assesses whether an image and its generated sentence are a valid pair.

4.3 Implementation details

Image captioning. We pre-train the image captioner on source domain with ADAM optimizer with learning rate $1 \times 10^{-3}$. We apply batch normalization with 0.95 decay. L2 regularization (with a weight decay value of 0.001) and dropout (with a dropout rate of 0.2) are used to the weights and biases of the decoder LSTM output layer to avoid overfitting. We conduct mini-batch training to train the model with batch size 64.

Image synthesis. We pre-train the image synthesis model on source domain via conditional GAN using ADAM optimizer with learning rate $2 \times 10^{-4}$. We apply batch normalization on all convolutional layers with 0.9 decay and epsilon $1 \times 10^{-5}$ followed by leaky ReLU. We conduct mini-batch training to train the model with batch size 64. To further improve the performance, we use skip-thought based on the sequence to sequence model to generate sentence-level vectors.

4.4 Evaluation Metrics

To quantitatively evaluate our image captioning method, we follow previous work to use BLEU-N (N=1,2,3,4) [19], MENTOR [2], ROUGE [13], CIDEr [29] scores for comparison. For all these metrics, they measure the consistency between n-gram occurrences in generated captions and ground-truth captions, where this consistency is weighted by n-gram saliency and rarity.

5 EXPERIMENTAL RESULTS

In this section, we compare our model with baseline methods quantitatively and qualitatively.

5.1 Quantitative evaluation

Following the experimental setup of [5], we fine-tune our CNN-RNN-based captioner on all paired data in the target domain whose performance serves as a meaningful reference for comparing with other image captioning models (referred to as the Fine-tuning method). Such an approach can not consider the unpaired data in the target domain, and its domain adaptation capability is limited to the size of available paired data in the target domain.

For a fair comparison with other baseline methods, we treat all the training data in target domain as unpaired data. In particular, we train DCC and SAtT models only on MSCOCO, and test them on the target domains, e.g., Flickr30k and Oxford-102. For SAdT and our model, we first pre-train the captioners on MSCOCO training set, and then adapt them with unpaired data in the target domain. Finally, we evaluate SAdT and our model on test sets in the target domains. The experimental results are summarized in Table 1.

<table>
<thead>
<tr>
<th>Method</th>
<th>Flickr30k</th>
<th>Oxford-102</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLEU-4</td>
<td>40.5</td>
<td>37.6</td>
</tr>
<tr>
<td>MENTOR</td>
<td>45.3</td>
<td>42.8</td>
</tr>
<tr>
<td>ROUGE</td>
<td>38.2</td>
<td>35.2</td>
</tr>
<tr>
<td>CIDEr</td>
<td>40.9</td>
<td>38.7</td>
</tr>
</tbody>
</table>

Table 1 shows that our model substantially and consistently outperforms the baseline methods by a noticeable margin on both the target domain datasets. In particular, our model successfully achieves better scores of all evaluation metrics on Oxford-102 than other baselines, even better than the fine-tuning of the CNN-RNN-based captioner. Not surprisingly, SAdT also outperforms DCC and SAtT, since it uses an adversarial training procedure to leverage unpaired training data in the target domain. Our model has more substantial improvement on Oxford-102 than that on Flickr30k. This may be because that Oxford-102 has a larger domain shift from MSCOCO than that from Flickr30k. In a word, the large domain shift between MSCOCO and Oxford-102 exposes a more challenging domain adaptation scenario, differentiating the strengths of different approaches more saliently. Particularly, the results show our model generalizes well to the target domain with a large domain shift without paired training data.
Table 1: Experimental results across two target domain datasets (Oxford-102 and Flickr30). All methods are pre-trained on source domain dataset (MSCOCO).

<table>
<thead>
<tr>
<th>Method</th>
<th>Target</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Meteor</th>
<th>ROUGE</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>Oxford-102</td>
<td>91.2</td>
<td>84.4</td>
<td>77.1</td>
<td>71.6</td>
<td>43.0</td>
<td>82.4</td>
<td>79.7</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>Oxford-102</td>
<td>90.7</td>
<td>83.7</td>
<td>76.2</td>
<td>70.6</td>
<td>42.2</td>
<td>81.7</td>
<td>78.0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Target</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Meteor</th>
<th>ROUGE</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>DCC</td>
<td>Flickr30k</td>
<td>55.0</td>
<td>34.8</td>
<td>20.8</td>
<td>12.6</td>
<td>14.7</td>
<td>38.3</td>
<td>23.2</td>
</tr>
<tr>
<td>SATT</td>
<td>Flickr30k</td>
<td>57.7</td>
<td>32.7</td>
<td>18.9</td>
<td>11.3</td>
<td>14.9</td>
<td>42.5</td>
<td>17.3</td>
</tr>
<tr>
<td>SaDT</td>
<td>Flickr30k</td>
<td>62.1</td>
<td>41.7</td>
<td>27.6</td>
<td>17.9</td>
<td>16.7</td>
<td>42.1</td>
<td>32.6</td>
</tr>
<tr>
<td>Ours</td>
<td>Flickr30k</td>
<td>63.9</td>
<td>44.1</td>
<td>31.8</td>
<td>17.3</td>
<td>16.3</td>
<td>44.2</td>
<td>33.6</td>
</tr>
<tr>
<td>Fine-tuning</td>
<td>Flickr30k</td>
<td>59.7</td>
<td>42.2</td>
<td>27.4</td>
<td>18.3</td>
<td>16.9</td>
<td>44.1</td>
<td>32.8</td>
</tr>
</tbody>
</table>

Table 2: Experimental results of our model by varying the percentage of paired data from 10% to 100%.

<table>
<thead>
<tr>
<th>Paired data</th>
<th>Target</th>
<th>Bleu-1</th>
<th>Bleu-2</th>
<th>Bleu-3</th>
<th>Bleu-4</th>
<th>Meteor</th>
<th>ROUGE</th>
<th>CIDEr</th>
</tr>
</thead>
<tbody>
<tr>
<td>with 10% paired data</td>
<td>Oxford-102</td>
<td>92.7</td>
<td>86.3</td>
<td>80.1</td>
<td>74.1</td>
<td>44.8</td>
<td>83.9</td>
<td>81.4</td>
</tr>
<tr>
<td>with 20% paired data</td>
<td>Oxford-102</td>
<td>94.4</td>
<td>87.8</td>
<td>81.5</td>
<td>75.9</td>
<td>45.2</td>
<td>84.6</td>
<td>82.5</td>
</tr>
<tr>
<td>with 40% paired data</td>
<td>Oxford-102</td>
<td>95.7</td>
<td>88.6</td>
<td>82.5</td>
<td>77.0</td>
<td>46.2</td>
<td>85.5</td>
<td>83.4</td>
</tr>
<tr>
<td>with 60% paired data</td>
<td>Oxford-102</td>
<td>96.1</td>
<td>89.8</td>
<td>83.1</td>
<td>77.6</td>
<td>46.6</td>
<td>86.3</td>
<td>84.2</td>
</tr>
<tr>
<td>with 80% paired data</td>
<td>Oxford-102</td>
<td>96.5</td>
<td>90.5</td>
<td>83.5</td>
<td>78.2</td>
<td>47.1</td>
<td>86.5</td>
<td>85.2</td>
</tr>
<tr>
<td>with 100% paired data</td>
<td>Oxford-102</td>
<td>96.8</td>
<td>91.0</td>
<td>83.9</td>
<td>78.4</td>
<td>47.3</td>
<td>86.7</td>
<td>86.1</td>
</tr>
<tr>
<td>with 10% paired data</td>
<td>Flickr30k</td>
<td>64.3</td>
<td>45.3</td>
<td>33.2</td>
<td>18.4</td>
<td>17.3</td>
<td>44.9</td>
<td>35.1</td>
</tr>
<tr>
<td>with 20% paired data</td>
<td>Flickr30k</td>
<td>65.0</td>
<td>46.2</td>
<td>33.9</td>
<td>18.6</td>
<td>17.5</td>
<td>45.9</td>
<td>37.9</td>
</tr>
<tr>
<td>with 40% paired data</td>
<td>Flickr30k</td>
<td>66.1</td>
<td>46.9</td>
<td>34.9</td>
<td>19.1</td>
<td>18.1</td>
<td>46.1</td>
<td>38.6</td>
</tr>
<tr>
<td>with 60% paired data</td>
<td>Flickr30k</td>
<td>66.8</td>
<td>47.5</td>
<td>35.4</td>
<td>19.9</td>
<td>18.9</td>
<td>46.5</td>
<td>40.1</td>
</tr>
<tr>
<td>with 80% paired data</td>
<td>Flickr30k</td>
<td>67.4</td>
<td>48.1</td>
<td>36.1</td>
<td>20.5</td>
<td>19.5</td>
<td>47.0</td>
<td>42.5</td>
</tr>
<tr>
<td>with 100% paired data</td>
<td>Flickr30k</td>
<td>67.7</td>
<td>48.5</td>
<td>36.3</td>
<td>20.9</td>
<td>19.9</td>
<td>47.2</td>
<td>43.1</td>
</tr>
</tbody>
</table>

Our model can also be trained in the semi-supervised and unsupervised adaptation settings by choosing different rewards at the stage of fine-tuning. To analyze the performance of our method with respect to the use of different numbers of paired training data in the target domain, we report the results by varying the percentage of paired data from 10% to 100% in the entire training set. Table 2 shows that our approach is robust and achieves excellent performance on different labelling percentages. As one may expect, training with more paired data slightly improves the overall performance. Even when only 10% of the training data in target domain are paired, our method significantly outperforms the fine-tuning results on all evaluation metrics on the two target domain datasets. The main advantage of our model comes from its capability of exploiting both paired and unpaired data in target domain. Another reason for the effectiveness of our approach is to use dual learning to simultaneously optimize image captioning and image synthesis, which are related and can be exploited together to further improve the performance of image captioning.

5.2 Qualitative evaluation

To evaluate the proposed model qualitatively, we show the generated captions and images of our image captioning and image synthesis systems, respectively.

5.2.1 Image captioning. Table 3 shows some examples generated by our model and the baseline methods for images in Oxford-102. It is easy to see that all of these image captioning models can generate reasonably relevant sentences, while the domain adaptation models (SaDT and ours) can generate more accurate and detailed captions. At some cases, the results of our model are more satisfying than SaDT. For example, the sentence "this flower has petals that are orange and has black dots" created by our model is more precise in describing color and content of the flower (the fourth line in Table 3). It partly illustrates that our model has a stronger adaption capability in target domains with significant domain shift but no paired training data. We also provide some qualitative examples for Flickr30k in Table 4. We observe similar results as on Oxford-102. The example in the third row of Table 4 shows that our model describes a specific state of a boy ("a boy in a stripy shirt") accurately while the other methods just mention there is "a boy".

5.2.2 Image synthesis. The generalization ability of the image synthesis is improved by the dual learning mechanism, which leverages the domain-specific knowledge contained in the training signals of image captions. We provided some qualitative examples in Table 5 to evaluate our image synthesis method. We use GAN-CLS [22] trained on MSCOCO dataset (without performing domain adaptation) as our baseline, since it has been the state-of-the-art method in generating plausible images from text descriptions. Due to the space limit, we only illustrate the images generated from two text

Table 3: Some examples of different methods on Oxford-102. The ground-truth captions are in bold text.

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Ours</th>
<th>SAdT</th>
<th>SaDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy is in the garden.</td>
<td>a boy in the garden</td>
<td>a boy is in the garden</td>
<td></td>
</tr>
<tr>
<td>The sun is shining.</td>
<td>The sun is shining</td>
<td>The sun is shining</td>
<td></td>
</tr>
<tr>
<td>The sky is blue.</td>
<td>The sky is blue</td>
<td>The sky is blue</td>
<td></td>
</tr>
<tr>
<td>I am happy today.</td>
<td>I am happy today</td>
<td>I am happy today</td>
<td></td>
</tr>
</tbody>
</table>

Table 4: Some examples of different methods on Flickr30k. The ground-truth captions are in bold text.

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Ours</th>
<th>SAdT</th>
<th>SaDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>The flower is beautiful.</td>
<td>The flower is beautiful</td>
<td>The flower is beautiful</td>
<td></td>
</tr>
<tr>
<td>The leaves are green.</td>
<td>The leaves are green</td>
<td>The leaves are green</td>
<td></td>
</tr>
<tr>
<td>The bird is singing.</td>
<td>The bird is singing</td>
<td>The bird is singing</td>
<td></td>
</tr>
<tr>
<td>I love this park.</td>
<td>I love this park</td>
<td>I love this park</td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Some examples of different methods on Oxford-102. The ground-truth images are shown on the left, and the generated images are on the right.

<table>
<thead>
<tr>
<th>Ground-truth</th>
<th>Ours</th>
<th>SAdT</th>
<th>SaDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>A boy in a park.</td>
<td>A boy in a park</td>
<td>A boy in a park</td>
<td></td>
</tr>
<tr>
<td>A tree with green leaves.</td>
<td>A tree with green leaves</td>
<td>A tree with green leaves</td>
<td></td>
</tr>
<tr>
<td>A cat on the floor.</td>
<td>A cat on the floor</td>
<td>A cat on the floor</td>
<td></td>
</tr>
</tbody>
</table>
Table 3: Example captions from different models for Oxford-102.

<table>
<thead>
<tr>
<th>Image</th>
<th>Ground truth captions</th>
<th>Generated Captions</th>
</tr>
</thead>
</table>
| ![Image](image) | 1. this flower has yellow petals with bright green sepal and a green pedicel.  
2. this flower has thick and vivid yellow petals with small red accents.  
3. this flower is yellow in color and has petals that are wavy and curled.  
4. a flower with long and wide petal that is bright yellow.  
5. this flower has petals that are dark yellow with green pedicel. | SAiT: a yellow flower in a clear vase on a table.  
SAiT: this flower has petals that are yellow and has red lines.  
Ours: this flower has petals that are yellow and has yellow stamens. |
| ![Image](image) | 1. the flower has rounded petals that are light pink with yellow larger stamen.  
2. this flower has a yellow center and veined pink petals with rounded edges.  
3. the petals of this flower are big and pink and the pistil is bright yellow.  
4. this flower has overlapping pink petals with veins and rounded edges.  
5. this flower is yellow and pink in color with petals that are multi colored. | SAiT: a large red flower in a flower garden.  
SAiT: this flower has petals that are pink and has a yellow center.  
Ours: this flower has petals that are pink and has yellow stamens. |
| ![Image](image) | 1. this flower has petals that are red and has black dots.  
2. this flower has petals that curl inwards and are red with brown spots.  
3. this flower has an interesting shape and a bright red color.  
4. this is a large orange flower with black dots and long stamen.  
5. this interesting orange red flower houses strange draping stamen. | SAiT: a pair of flowers are in a vase filled with water.  
SAiT: this flower has petals that are pink and has red dots.  
Ours: this flower has petals that are orange and has black dots. |

Table 4: Example captions from different models for Flickr30k.

<table>
<thead>
<tr>
<th>Image</th>
<th>Ground truth captions</th>
<th>Generated Captions</th>
</tr>
</thead>
</table>
| ![Image](image) | 1. two girls participate in a softball game.  
2. two girls playing in a game of softball.  
3. a baseball player in a black shirt just tagged a player in a white shirt.  
4. a woman in a baseball uniform tags another with her mitt.  
5. ten players on white softball team gets tagged by a girl of other team. | SAiT: a baseball player holding a bat on a field.  
SAiT: a young baseball player is sliding into a base.  
Ours: two girls are playing a baseball game. |
| ![Image](image) | 1. a boy leaping in the air towards a soccer ball.  
2. a boy in midair trying to kick a soccer ball.  
3. boy leaning backward to kick a soccer ball in midair.  
4. a child falling back to kick a soccer ball.  
5. a boy in a stripy shirt is playing with a soccer ball in a grassy park. | SAiT: a man in a field playing with a frisbee.  
SAiT: a young boy playing with a soccer ball in a field.  
Ours: a boy in a striped shirt is playing with a soccer ball. |
| ![Image](image) | 1. a dog stirs up the dirt on the ground while running to catch a ball.  
2. a dog turns on the grass to a flying ball.  
3. a light brown dog is chasing after a ball.  
4. a dog is playing fetch with a ball.  
5. a yellow dog anticipates a ball. | SAiT: a dog running with a frisbee in its mouth.  
SAiT: a brown dog is running in the grass.  
Ours: a brown dog is chasing a ball through the grass. |

descriptions in Oxford-102 dataset which has a large domain shift with MSCOCO. From Table 5 we can see that GAN-CLS only gets some color information right, but their contents are completely mismatched and the pictures do not look real, either. However, our model generates more plausible images that are close to what the text descriptions describe.

6 CONCLUSIONS

We proposed a novel dual learning method for cross-domain image captioning. First, we modeled the alignment between the neural representations of images and that of natural languages in the source domain where one can access sufficient labeled data. Second, we fine-tuned the pre-trained model on limited data (or unpaired data) in the target domain via a dual learning mechanism, which simultaneously optimized two tasks: generating image captions in text and generating plausible images from text descriptions. To evaluate the effectiveness of our model, we used MSCOCO dataset as the source domain and two other datasets (Oxford-102 and Flickr30k) as the target domain. The experimental results showed that our method consistently outperformed previous methods for cross-domain image captioning.

In the future, we would like to evaluate the quality of the generated captions and the generated images with more careful human evaluation. To evaluate our model on a more realistic target dataset that consists of a large amount of unpaired data from different sources, we plan to collect a new unpaired dataset where the images are from Oxford-102, while the text descriptions are from the web (e.g., Wikipedia).

ACKNOWLEDGEMENT

This work was supported in part by National HighTech Research and Development Program of China (2015AA042303), National Natural Science Foundation of China (U1613211), and External Cooperation Program of BIC Chinese Academy of Sciences (172644KYSB20160033).
### References


