Show, Adapt and Tell: Adversarial Training of Cross-domain Image Captioner

Tseng-Hung Chen[†], Yuan-Hong Liao[†], Ching-Yao Chuang[†], Wan-Ting Hsu[†], Jianlong Fu[‡], Min Sun[†]

[†]Department of Electrical Engineering, National Tsing Hua University, Hsinchu, Taiwan
[‡]Microsoft Research, Beijing, China

{tsenghung@gapp, andrewliao11@gapp, cychuang@gapp, hsuwanting@gapp, sunmin@ee}.nthu.edu.tw
jianf@microsoft.com

Abstract

Impressive image captioning results are achieved in domains with plenty of training image and sentence pairs (e.g., MSCOCO). However, transferring to a target domain with significant domain shifts but no paired training data (referred to as cross-domain image captioning) remains largely unexplored. We propose a novel adversarial training procedure to leverage unpaired data in the target domain. Two critic networks are introduced to guide the captioner, namely domain critic and multi-modal critic. The domain critic assesses whether the generated sentences are indistinguishable from sentences in the target domain. The multi-modal critic assesses whether an image and its generated sentence are a valid pair. During training, the critics and captioner act as adversaries – captioner aims to generate indistinguishable sentences, whereas critics aim at distinguishing them. The assessment improves the captioner through policy gradient updates. During inference, we further propose a novel critic-based planning method to select high-quality sentences without additional supervision (e.g., tags). To evaluate, we use MSCOCO as the source domain and four other datasets (CUB-200-2011, Oxford-102, TGIF, and Flickr30k) as the target domains. Our method consistently performs well on all datasets. In particular, on CUB-200-2011, we achieve 21.8% CIDEr-D improvement after adaptation. Utilizing critics during inference further gives another 4.5% boost.

1. Introduction

Datasets with large corpora of "paired" images and sentences have enabled the latest advance in image captioning. Many novel networks [9, 21, 17, 33] trained with these paired data have achieved impressive results under a domain-specific setting – training and testing on the same domain. However, the domain-specific setting creates a huge cost on collecting "paired" images and sentences in each domain. For real world applications, one will prefer a "cross-domain" captioner which is trained in a "source" domain with paired data and generalized to other "target" domains with very little cost (e.g., no paired data required).

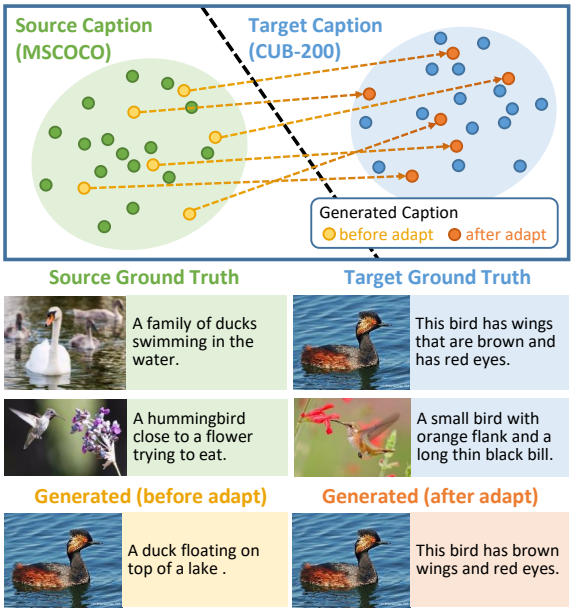


Figure 1: We propose a cross-domain image captioner that can adapt the sentence style from source to target domain without the need of paired image-sentence training data in the target domain. Left panel: Sentences from MSCOCO mainly focus on location, color, size of objects. Right panel: Sentences from CUB-200 describe the parts of birds in detail. Bottom panel shows our generated sentences before and after adaptation.

Training a high-quality cross-domain captioner is challenging due to the large domain shift in both the image and sentence spaces. For instance, MSCOCO [23] mostly consists of images of large scene with more object instances, whereas CUB-200-2011 [34] (shortened as CUB-200 in the following) consists of cropped birds images. Moreover, sentences in MSCOCO typically describe location, color and size of objects, whereas sentences in CUB-200 describe parts of birds in detail (Fig. 1). In this case, how can one ex-

pect a captioner trained on MSCOCO to describe the details of a bird on CUB-200 dataset?

A few works propose to leverage different types of unpaired data in other domains to tackle this challenge. [14, 31] propose to leverage an image dataset with category labels (e.g., ImageNet [8]) and sentences on the web (e.g., Wikipedia). However, they focus on the ability to generate words unseen in paired training data (i.e., word-level modification). Anderson et al. [2] propose to leverage image taggers at test time. However, this requires a robust cross-domain tagger. Moreover, they focus on selecting a few different words but not changing the overall style.

We propose a novel adversarial training procedure to leverage unpaired images and sentences. Two critic networks are introduced to guide the procedure, namely domain critic and multi-modal critic. The domain critic assesses whether the generated captions are indistinguishable from sentences in the target domain. The multi-modal critic assesses whether an image and its generated caption is a valid pair. During training, the critics and captioner act as adversaries – captioner aims to generate indistinguishable captions, whereas critics aim at distinguishing them. Since the sentence is assessed only when it is completed (e.g., cannot be assessed in a word by word fashion), we use Monte Carlo rollout to estimate the assess of each generated word. Then, we apply policy gradient [30] to update the network of the captioner. Last but not least, we propose a novel critic-based planning method to take advantage of the learned critics to compensate the uncertainty of the sentence generation policy with no additional supervision (e.g., tags [2]) in testing.

To evaluate, we use MSCOCO [23] as the source domain and CUB-200 [34, 28], Oxford-102 [26, 28], Flickr30k [37] and TGIF [22] as target domains. Our method consistently performs well on all datasets. In particular, on CUB-200, we achieve 21.8% CIDEr-D improvement after adaptation. Utilizing critic during inference further gives another 4.5% boost. Our codes are available at https://github.com/tsenghungchen/show-adapt-and-tell. Finally, the contributions of the paper are summarized below:

- We propose a novel adversarial training procedure for cross-domain captioner. It utilizes critics to capture the distribution of image and sentence in the target domain.
- We propose to utilize the knowledge of critics during inference to further improve the performance.
- Our method achieves significant improvement on four publicly available datasets compared to a captioner trained only on the source domain.

2. Related Work

Visual description generation. Automatically describing visual contents is a fundamental problem in artificial intelligence that connects computer vision and natural language processing. Thanks to recent advances in deep neural networks and the release of several large-scale datasets such as MSCOCO [23] and Flickr30k [37], many works [9, 21, 17, 33] have shown different levels of success on image captioning. They typically employ a Convolutional Neural Network (CNN) for image encoding, then decoding a caption with a Recurrent Neural Network (RNN). There have been many attempts to improve the basic encoder-decoder framework. The most commonly used approach is spatial attention mechanism. Xu et al. [35] introduce an attention model that can automatically learn where to look depending on the generated words. Besides images, [9, 32, 36, 40] apply LSTMs as video encoder to generate video descriptions. In particular, Zeng et al. [40] propose a framework to jointly localize highlights in videos and generate their titles.

Addressing exposure bias. Recently, the issue of exposure bias [27] has been well-addressed in sequence prediction tasks. It happens when a model is trained to maximize the likelihood given ground truth words but follows its own predictions during test inference. As a result, the training process leads to error accumulation at test time. In order to minimize the discrepancy between training and inference, Bengio et al. [5] propose a curriculum learning strategy to gradually ignore the guidance from supervision during training. Lamb et al. [11] introduce an adversarial training method as regularization between sampling mode and teacher-forced mode. Most recently, there are plenty of works [27, 4, 24, 29] using policy gradient to directly optimize the evaluation metrics. These methods avoid the problem of exposure bias and further improve over cross entropy methods. However, they cannot be applied in cross-domain captioning, since they need ground truth sentences to compute metric such as BLEU.

Reward modeling. In contrast to the above works, we learn the reward function in cross-domain setting and the reward can be computed even during testing to enable our novel critic-based planning method. Several works [13, 38] incorporate auxiliary models as rewards. Hendricks et al. [13] minimize a discriminative loss to ensure generated sentences be class specific. Similar to our method, Yu et al. [38] also introduce a critic to learn a reward function. However, their proposed method is for random sentence generation and not designed for domain adaptation.

Domain adaptation. Conventional DNN-based domain adaptation aim to learn a latent space that minimize the distance metrics (e.g., Maximum Mean Discrepancy (MMD) [25] and Central Moment Discrepancy (CMD) [39]) between data domains. On the other hand, existing adversarial domain adaptation methods use a domain classifier to learn mappings from source to target domains. Ajakan et al. [1] introduce a domain adaptation regularizer to learn the representation for sentiment analysis. Ganin

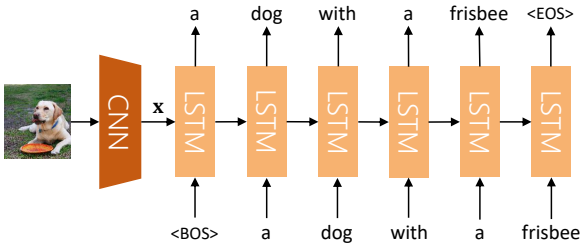


Figure 2: Our captioner is a standard CNN-RNN architecture [33], where predicted word from previous step is serve as input of current step during inference. <BOS> and <EOS> represent the Begin-Of-Sentence and End-Of-Sentence, respectively.

et al. [10] propose a gradient reversal layer for aligning the distribution of features across source and target domain. Hoffman et al. [15] propose an unsupervised domain adversarial method for semantic segmentations in street scenes. Chen et al. [6] further collect a dataset of road scene images across countries for cross-city adaptation. Performance improvement has been shown on sentiment analysis, image classification, person re-identification, and scene segmentation tasks. However, we are not aware of any adversarial domain adaptation approach applied on cross-domain captioning.

3. Cross-domain Image Captioning

We first formally define the task of cross-domain image captioning; then, give an overview of our proposed method.

Cross-domain setting. This is a common setting where data from two domains are available. In the source domain, we are given a set $\mathcal{P} = \{(\mathbf{x}^n, \hat{\mathbf{y}}^n)\}_n$ with paired image $\mathbf{x}^n$[1] and "ground truth" sentence $\hat{\mathbf{y}}^n$ describing $\mathbf{x}^n$. Each sentence $\hat{\mathbf{y}} = [\hat{y}_1, \ldots, \hat{y}_t, \ldots, \hat{y}_T]$ consists of a sequence of word $\hat{y}_t$ with length T. In the target domain, we are given two separate sets of information: a set of example images $\mathcal{X} = \{\mathbf{x}^n\}_n$ and a set of example sentences $\hat{\mathcal{Y}} = \{\hat{\mathbf{y}}^n\}_n$. Note that collecting paired data $\mathcal{P}$ in the source domain is typically more costly than $\mathcal{X}$ and $\hat{\mathcal{Y}}$ in the target domain.

Image captioning. For standard image captioning, the goal is to generate a sentence $\mathbf{y}$ for $\mathbf{x}$, where $\mathbf{y}$ is as similar as the ground truth sentence $\hat{\mathbf{y}}$. For cross-domain image captioning, since the ground truth sentence of each image in $\mathcal{X}$ is not available, the goal becomes the following. For an image $\mathbf{x} \in \mathcal{X}$, we aim at generating a sentence $\mathbf{y}$ such that (1) $\mathbf{y}$ is similar to $\hat{\mathcal{Y}}$ in style, and (2) $(\mathbf{x}, \mathbf{y})$ are a relevant pair similar to pairs in $\mathcal{P}$.

Overview of our method. To achieve the goal of cross-domain image captioning, we propose a novel method consisting of two main components. The first component is a standard CNN-RNN-based captioner (Fig. 2). However, our captioner is treated as an agent taking sequential actions (i.e, generating words). This agent is trained using policy gradient given reward of each generated sentence. Our second component consists of two critics to provide reward. One critic assesses the similarity between $\mathbf{y}$ and $\hat{\mathcal{Y}}$ in style. The other critic assesses the relevancy between $\mathbf{x}$ and $\mathbf{y}$, given paired data $\mathcal{P}$ in the source domain as example pairs. We use both critics to compute a reward for each generated sentence $\mathbf{y}$. Both the captioner and two critics are iteratively trained using a novel adversarial training procedure. Next, we describe the captioner and critics in detail.

3.1. Captioner as an Agent

At time t, the captioner takes an action (i.e., a word y_t) according to a stochastic policy $\pi_\theta(y_t|\mathbf{x}, \mathbf{y}_{t-1})$, where $\mathbf{x}$ is the observed image, $\mathbf{y}_{t-1} = [y_1, ..., y_{t-1}]$ [2] is the generated partial sentence, and θ is the parameter of the policy. We utilize an existing CNN-RNN model [33] as the model of the policy. By sequentially generating each word y_t from the policy $\pi_\theta(.)$ until the special End-Of-Sentence (EOS) token, a complete sentence $\mathbf{y}$ is generated. In standard image captioning, the following total expected per-word loss $J(\theta)$ is minimized.

$$J(\theta) = \sum_{n=1}^{N} \sum_{t=1}^{T_n} \text{Loss}(\pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n)) \,, \tag{1}$$

$$\text{Loss}(\pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n)) = -\log \pi_\theta(\hat{y}_t^n|\mathbf{x}^n, \hat{\mathbf{y}}_{t-1}^n) \,,$$

where N is the number of images, T_n is the length of the sentence $\hat{\mathbf{y}}^n$, Loss(.) is cross-entropy loss, and $\hat{\mathbf{y}}_{t-1}^n$ and $\hat{y}_t^n$ are ground truth partial sentence and word, respectively. For cross-domain captioning, we do not have ground truth sentence in target domain. Hence, we introduce critics to assess the quality of the generated complete sentence $\mathbf{y}^n$. In particular, the critics compute a reward $R(\mathbf{y}^n|\mathbf{x}^n, \mathcal{Y}, \mathcal{P})$ (see Sec. 3.2 for details) utilizing example sentences $\mathcal{Y}$ in target domain and example paired data $\mathcal{P}$ in source domain. Given the reward, we modify Eq. 1 to train the agent using policy gradient.

Policy gradient. The main idea of policy gradient is to replace per-word loss Loss(.) in Eq. 1 with another computable term related to the state-action reward $Q(s_t, a_t)$, where the state s_t is characterized by the image $\mathbf{x}$ and partial sentence $\mathbf{y}_{t-1}$ while the action a_t is the current generated word y_t. The state-action reward $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t)$ is defined as the expected future reward:

$$E_{\mathbf{y}_{(t+1):T}} \left[R(\left[\mathbf{y}_{t-1}, y_t, \mathbf{y}_{(t+1):T}\right] |\mathbf{x}, \mathcal{Y}, \mathcal{P}) \right] \,. \tag{2}$$

Note that the expectation is over the future words $\mathbf{y}_{(t+1):T} = [y_{t+1}, \ldots, y_T]$ until the sentence is completed at time T. Hence, $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t)$ takes the randomness of future words $\mathbf{y}_{(t+1):T}$ into consideration. Given $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t)$, we aim at maximizing a new objective as below,

[1] We extract image representation $\mathbf{x}^n$ from CNN.

[2] For the partial sentence starting from index 1, we denoted it as $\mathbf{y}_{t-1}$ for simplicity.

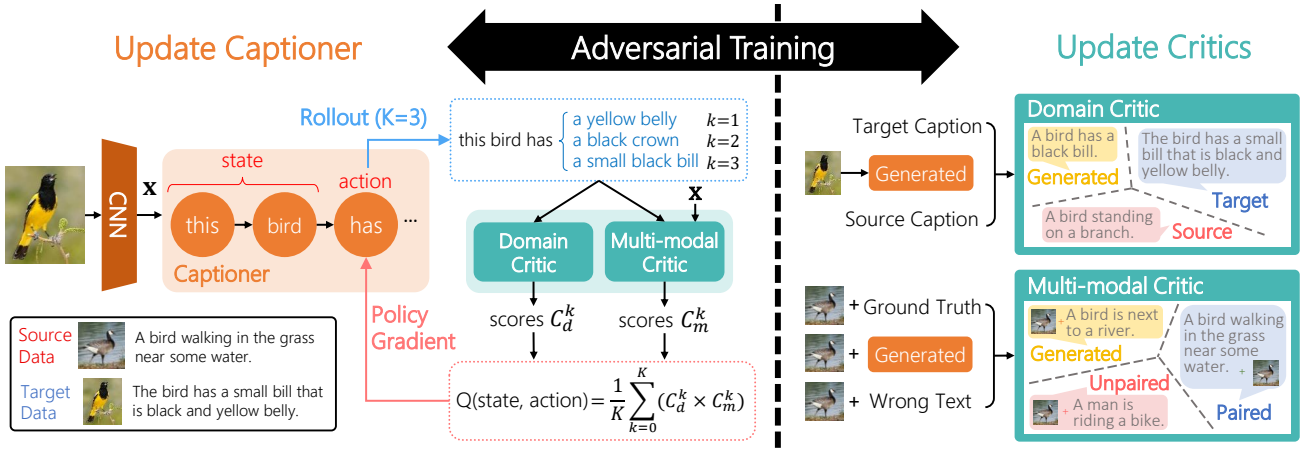


Figure 3: System overview. Left panel: our captioner generates a sentence condition on image representation $\mathbf{x}$. At each step, the expected reward of a newly generated word ("has") is computed from the domain and multi-modal critics using Monte Carlo rollout. We use policy gradient to update the captioner toward generating sentences with higher reward. Right panel: the critics observe sentences generated from the captioner and aim at discriminating them from the true data in target and source domains. During adversarial training, both captioner (Left) and critics (Right) are iteratively updated to achieve competing goals.

$$J(\theta) = \sum_{n=1}^{N} J_n(\theta) \,,$$

$$J_n(\theta) = \sum_{t=1}^{T_n} E_{\mathbf{y}_t^n}\left[\pi_\theta(y_t^n|\mathbf{x}^n, \mathbf{y}_{t-1}^n) Q((\mathbf{x}^n, \mathbf{y}_{t-1}^n), y_t^n)\right] \,,$$

where $\mathbf{y}_t^n = \left[\mathbf{y}_{t-1}^n, y_t^n\right]$ is a random vector instead of ground truth $\hat{\mathbf{y}}_t^n = \left[\hat{\mathbf{y}}_{t-1}^n, \hat{y}_t^n\right]$ as in Eq. 1. However, since the spaces of $\mathbf{y}_t$ [3] is huge, we generate M sentences $\{\mathbf{y}^m\}_m$ to replace expectation with empirical mean as follows,

$$J_n(\theta) \simeq \frac{1}{M}\sum_{m=1}^{M} J_{n,m}(\theta) \,, \qquad (3)$$

$$J_{n,m}(\theta) = \sum_{t=1}^{T_m} \pi_\theta(y_t^m|\mathbf{x}, \mathbf{y}_{t-1}^m) Q((\mathbf{x}, \mathbf{y}_{t-1}^m), y_t^m) \,, \qquad (4)$$

where T_m is the length of the generated m^{th} sentence. Note that $\mathbf{y}_t^m = [\mathbf{y}_{t-1}^m, y_t^m]$ is sampled from the current policy π_θ and thus computing $J_{n,m}(\theta)$ becomes tractable. The policy gradient can be computed from Eq. 4 as below,

$$\nabla_\theta J_{n,m}(\theta) = \sum_{t=1}^{T_m} \nabla_\theta \pi_\theta(y_t^m|\mathbf{x}, \mathbf{y}_{t-1}^m) Q((\mathbf{x}, \mathbf{y}_{t-1}^m), y_t^m) =$$

$$\sum_{t=1}^{T_m} \pi_\theta(y_t^m|\mathbf{x}, \mathbf{y}_{t-1}^m) \nabla_\theta \log \pi_\theta(y_t^m|\mathbf{x}, \mathbf{y}_{t-1}^m) Q((\mathbf{x}, \mathbf{y}_{t-1}^m), y_t^m) \,,$$

and the total gradient is

$$\nabla_\theta J(\theta) \simeq \frac{1}{M}\sum_{n=1}^{N}\sum_{m=1}^{M} \nabla_\theta J_{n,m}(\theta) \,. \qquad (5)$$

We apply stochastic optimization with policy gradient to update model parameter θ. Next we describe how to estimate the state-action reward $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t)$.

[3] We remove superscript n for simplification.

Estimating Q. Since the space of $\mathbf{y}_{(t+1):T}$ in Eq. 2 is also huge, we use Monte Carlo rollout to replace expectation with empirical mean as below,

$$Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t) \simeq$$
$$\frac{1}{K}\sum_{k=1}^{K} R\left(\left[\mathbf{y}_{t-1}, y_t, \mathbf{y}_{(t+1):T_k}^k\right]|\mathbf{x}, \mathcal{Y}, \mathcal{P}\right) \,, \qquad (6)$$

where $\{\mathbf{y}_{(t+1):T_k}^k\}_k$ are generated future words, and we sample K complete sentences following policy π_θ. Next, we introduce the critics for computing the reward $R(\cdot)$.

3.2. Critics

For cross-domain image captioning, a good caption needs to satisfy two criteria: (1) the generated sentence resembles the sentence drawn from the target domain. (2) the generated sentence is relevant to the input image. The critics follow these two rules to assign reward to each generated sentence. We introduce the *domain critic* and *multi-modal critic* below.

Domain critic. In order to address the domain shift in sentence space, we train a Domain Critic (DC) to classify sentences as "source" domain, "target" domain, or "generated" ones. The DC model consists of an encoder and a classifier. A sentence $\mathbf{y}$ is first encoded by CNN [18] with highway connection [19] into a sentence representation. Then, we pass the representation through a fully connected layer and a softmax layer to generate probability $C_d(l|\mathbf{y})$, where $l \in \{source, target, generated\}$. Note that the scalar probability $C_d(target|\mathbf{y})$ indicates how likely the sentence $\mathbf{y}$ is from the target domain.

Multi-modal critic. In order to check the relevance between a sentence $\mathbf{y}$ and an image $\mathbf{x}$, we propose a Multi-modal Critic (MC) to classify $(\mathbf{x}, \mathbf{y})$ as "paired", "unpaired", or "generated" data. The model of MC consists

of multi-modal encoders, modality fusion layer, and a classifier as below,

$$\mathbf{c} = \text{LSTM}_\rho(\mathbf{y}) \,, \tag{7}$$

$$f = \tanh(W_x \cdot \mathbf{x} + b_x) \odot \tanh(W_c \cdot \mathbf{c} + b_c) \,, \tag{8}$$

$$C_m = \text{softmax}(W_m \cdot f + b_m) \,, \tag{9}$$

where $\rho, W_x, b_x, W_c, b_c, W_m, b_m$ are parameters to be learned, $\odot$ denotes element-wise multiplication, and C_m is the probabilities over three classes: paired, unpaired, and generated data. In Eq. 7, the sentence $\mathbf{y}$ is encoded by an LSTM-based sentence encoder. Then, in Eq. 8, the encoded image $\mathbf{x}$ and sentence $\mathbf{c}$ representations are fused via element-wise multiplication similar to [3]. Finally, in Eq. 9, the fused representation is forwarded through a fully connected layer and a softmax layer to generate probability $C_m(l|\mathbf{x}, \mathbf{y})$, where $l \in \{paired, unpaired, generated\}$. The scalar probability $C_m(paired|\mathbf{x}, \mathbf{y})$ indicates how a generated caption $\mathbf{y}$ is relevant to an image $\mathbf{x}$. Please see Supplementary for the intuition and empirical studies of the design choices in DC and MC.

Sentence reward. We define the reward $R(\mathbf{y}|.) = C_d(target|.) \cdot C_m(paired|.)$. This ensures a sentence receives a high reward only when (1) DC believes the sentence is from the target domain, and (2) MC believes the sentence is relevant to the image.

Training critics. We introduce the training objective of DC and MC below. For DC, the goal is to classify a sentence into source, target, and generated data. This can be formulated as a supervised classification training objective as follows,

$$\mathcal{L}_d(\phi) = -\sum_{n=1}^{N} \log C_d(l^n|\mathbf{y}^n; \phi)$$

$$l^n = \begin{cases} source & \text{if } \mathbf{y}^n \in \hat{\mathcal{Y}}_{src}, \\ target & \text{if } \mathbf{y}^n \in \hat{\mathcal{Y}}_{tgt}, \\ generated & \text{if } \mathbf{y}^n \in \mathcal{Y}_{\pi_\theta}, \end{cases} \tag{10}$$

$$\mathcal{Y}_{\pi_\theta} = \{\mathbf{y}^n \sim \pi_\theta(.|\mathbf{x}^n, .)\}_n, \mathbf{x}^n \in \mathcal{X}_{tgt} \,,$$

where N is the number of sentences, ϕ is the model parameter of DC, $\hat{\mathcal{Y}}_{src}$ denotes sentences from the source domain, $\hat{\mathcal{Y}}_{tgt}$ denotes sentences from the target domain, and $\mathcal{Y}_{\pi_\theta}$ denotes sentences generated from the captioner with policy π_θ given target domain images $\mathcal{X}_{tgt}$.

For MC, the goal is to classify a image-sentence pair into paired, unpaired, and generated data. This can also be formulated as a supervised classification training objective as follows,

$$\mathcal{L}_m(\eta) = -\sum_{n=1}^{N} \log C_m(l^n|\mathbf{x}^n, \mathbf{y}^n; \eta) \,,$$

$$l^n = \begin{cases} paired & \text{if } (\mathbf{x}^n, \mathbf{y}^n) \in \mathcal{P}_{src} \,, \\ unpaired & \text{if } (\mathbf{x}^n, \mathbf{y}^n) \in \acute{\mathcal{P}}_{src} \,, \\ generated & \text{if } (\mathbf{x}^n, \mathbf{y}^n) \in \mathcal{P}_{gen} \,, \end{cases} \tag{11}$$

$$\acute{\mathcal{P}}_{src} = \{(\mathbf{x}^i \in \mathcal{X}_{src}, \hat{\mathbf{y}}^j \in \hat{\mathcal{Y}}_{src}); i \neq j\} \,,$$

$$\mathcal{P}_{gen} = \{(\mathbf{x} \in \mathcal{X}_{src}, \mathbf{y} \in \mathcal{Y}_{\pi_\theta})\} \,,$$

Algorithm 1: Adversarial Training Procedure

Require: captioner π_θ, domain critic C_d, multi-modal critic C_m, an empty set for generated sentences $\mathcal{Y}_{\pi_\theta}$, and an empty set for paired image-generated-sentence $\mathcal{P}_{gen}$;

Input: sentences $\hat{\mathcal{Y}}_{src}$, image-sentence pairs $\mathcal{P}_{src}$, unpaired data $\acute{\mathcal{P}}_{src}$ in source domain; sentences $\hat{\mathcal{Y}}_{tgt}$, images $\mathcal{X}_{tgt}$ in target domain;

```
Pre-train π_θ on P_src using Eq. 1;
while θ has not converged do
    for i = 0, ..., N_c do
        Y_{π_θ} ← {y}, where y ~ π_θ(·|x, ·) and x ~ X_tgt;
        Compute g_d = ∇_φ L_d(φ) using Eq. 10;
        Adam update of φ for C_d using g_d;
        Y_{π_θ} ← {y}, where y ~ π_θ(·|x, ·) and x ~ X_src;
        P_gen ← {(x, y)};
        Compute g_m = ∇_η L_m(η) using Eq. 11;
        Adam update of η for C_m using g_m;
    for i = 0, ..., N_g do
        Y_{π_θ} ← {y}, where y ~ π_θ(·|x, ·) and x ~ X_tgt;
        P_gen ← {(x, y)};
        for t = 1, ..., T do
            Compute Q((x, y_{t-1}), y_t) with Monte Carlo
              rollouts, using Eq. 6;
        Compute g_θ = ∇_θ J(θ) using Eq. 5;
        Adam update of θ using g_θ;
```

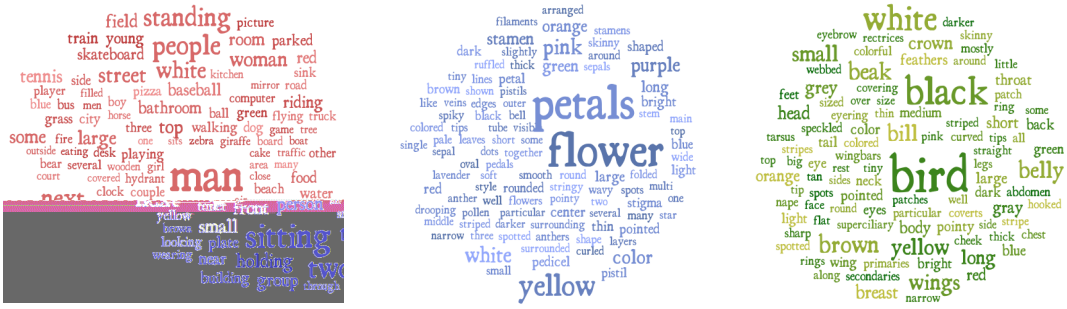
where η is the model parameter of MC, $\mathcal{P}_{src}$ is the paired data from the source domain, $\acute{\mathcal{P}}_{src}$ is the unpaired data intentionally collected randomly by shuffling images and sentences in the source domain, and $\mathcal{P}_{gen}$ is the source-image-generated-sentence pairs.

3.3. Adversarial Training

Our cross-domain image captioning system is summarized in Fig. 3. Both captioner π_θ and critics C_d and C_m learn together by pursuing competing goals as described below. Given $\mathbf{x}$, the captioner π_θ generates a sentence $\mathbf{y}$. It would prefer the sentence to have large reward $R(\mathbf{y}|.)$, which implies large values of $C_d(target|\mathbf{y})$ and $C_m(paired|\mathbf{x}.\mathbf{y})$. In contrast, the critics would prefer large values of $C_d(generated|\mathbf{y})$ and $C_m(generated|\mathbf{x}, \mathbf{y})$, which implies small values of $C_d(target|\mathbf{y})$ and $C_m(paired|\mathbf{x}.\mathbf{y})$. We propose a novel adversarial training procedure to iteratively updating the captioner and critics in Algorithm 1. In short, we first pretrain the captioner using cross-entropy loss on source domain data. Then, we iteratively update the captioner and critics with a ratio of $N_g : N_c$, where the critics are updated more often than captioner (i.e., $N_g < N_c$).

3.4. Critic-based Planning

The quality of a generated word y_t is typically measure by the policy network $\pi(y_t|\cdot)$. For cross-domain caption-



(a) MSCOCO (b) Oxford-102 (c) CUB-200

Figure 4: Word clouds for testing set of (a) MSCOCO, (b) Oxford-102, (c) CUB-200, where font size indicates the frequency of words.

ing, the learned critics can also be used to measure the quality of y_t by computing $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t)$ using Eq. 6. Here, Q is an expected value that models the randomness of future words, so we call our method "critic-based planning". Critic-based planning takes advantage of both the learned policy network as well as the critics. By default, we select $y_t^* = \arg\max_y \pi_\theta(y|\cdot)$ as the generated word. However, when the difference between the maximum probability and the second largest probability of $\pi_\theta(\cdot)$ is below a threshold Γ (where the selection of y_t^* is ambiguous), we take the top J words $\{y_t^j\}_{j=1}^J$ according to $\pi(y|\cdot)$ and evaluate $Q((\mathbf{x}, \mathbf{y}_{t-1}), y_t^j)$ for all j. Then, we select the word with the highest Q value as the generated word. Note that the sentences generated via critic-based planning can be exactly the same as greedy search. Our critic-based planning method obtain further performance improvement typically on dataset with large domain shift (e.g., CUB-200 and Oxford-102).

4. Experiments

We perform extensive evaluations on a number of popular datasets. For all experiments, we use MSCOCO [23] as the source dataset and CUB-200 [34], Oxford-102 [26], TGIF [22], and Flickr30k [37] as target datasets. We show that our method generalizes to datasets with large domain shift (CUB-200 and Oxford-102) and datasets with regular domain shift (Flickr30k and TGIF). We also show that critic-based planning can further improve performance during inference on datasets with large domain shift. Finally, we conduct an ablation study on Flickr30k to show the contribution of different components.

4.1. Implementation details

Data preprocessing. For source domain dataset, we select the MSCOCO training split from [17] which contains $113,287$ images, along with 5 captions each. We prune the vocabulary by dropping words with frequency less than 5, resulting in $10,066$ words including special Begin-Of-Sentence (BOS) and End-Of-Sentence (EOS) tokens. We use the same vocabulary in all experiments. For target domain datasets, we remove the training sentences containing out-of-vocabulary words (see Supplementary for detailed statistics).

Pre-training details. The architecture of our captioner is a CNN-LSTM with hidden dimension 512. The image features are extracted using the pre-trained Resnet-101 [12] and the sentences are represented as one-hot encoding. We first pre-train the captioner on source domain dataset via cross entropy objective using ADAM optimizer [20] with learning rate 5×10^{-4}. We apply learning rate decay with a factor of 0.8 every three epoches. To further improve the performance, we use schedule sampling [5] to mitigate the exposure bias. The best model is selected according to the validation performance and serve as the initial model for adversarial training.

Adversarial training details. We train the captioner and critics using ADAM optimizer [20] with learning rate of 5×10^{-5}. We apply dropout in training phase to prevent over-fitting, which also served as input noise similar to [16]. In Monte Carlo rollout, the model samples words until the EOS token under the current policy for $K = 3$ times. These K sentences are them fed to the critics for estimating the state-action value $Q(\cdot)$. Both critics are trained from scratch using the standard classification objective.

4.2. Experimental Results

We first pre-train the captioner on MSCOCO training set. Next, we update the captioner by adversarial training procedure with unpaired data from the training set in target domains. Finally, we evaluate our method on four target domain datasets, representing different levels of domain shift.

Baseline. We re-implement Deep Compositional Captioner (referred to as DCC) [14] as our baseline method. DCC consists of a lexical classifier and a language model. The former is a CNN model trained to predict semantic attributes and the latter is an LSTM model trained on unpaired text. In the end, the overall DCC model combines both models with a linear layer trained on paired image-caption data. For fair comparison, we apply the following settings on DCC, where the lexical classifier is a ResNet-101 model and the language model is trained on target domain sentences. Note that the ResNet-101 is fine-tuned with 471 visual concepts (pre-defined in [14]) extracted from captions. Finally, we use source domain image-caption pairs to fine-tune DCC. We also fine-tune a pre-trained source domain model di-

Table 1: Results of adaptation across four target domain datasets. Source (MSCOCO) Pre-trained and DCC are two baseline methods. Fine-tuning with paired data in target domain serves as the upper bound performance of our CNN-RNN captioner.

Method	Target (test)	Bleu-1	Bleu-2	Bleu-3	Bleu-4	Meteor	ROUGE	CIDEr	SPICE
Source Pre-trained	CUB-200	50.8	28.3	13.9	6.1	12.9	33	3	4.6
DCC	CUB-200	68.6	47.3	31.4	21.4	23.8	46.4	11.9	11.1
Ours	CUB-200	**91.4**	**73.1**	**51.9**	**32.8**	**27.6**	**58.6**	**24.8**	**13.2**
Fine-tuning	CUB-200	91.3	80.2	69.2	59	36.1	69.7	61.1	17.9
Source Pre-trained	Oxford-102	48.3	21.6	6.2	1.3	10.5	25.8	3.1	4.4
DCC	Oxford-102	51	33.8	24.1	16.7	21.5	38.3	6	9.8
Ours	Oxford-102	**85.6**	**76.9**	**67.4**	**60.5**	**36.4**	**72.1**	**29.3**	**17.9**
Fine-tuning	Oxford-102	87.5	80.1	72.8	66.3	40	75.6	36.3	18.5
Source Pre-trained	TGIF	41.6	23.3	12.6	7	12.7	32.7	14.7	8.5
DCC	TGIF	34.6	17.5	9.3	4.1	11.8	29.5	7.1	7.3
Ours	TGIF	**47.5**	**29.2**	**17.9**	**10.3**	**14.5**	**37**	**22.2**	**10.6**
Fine-tuning	TGIF	51.1	32.2	20.2	11.8	16.2	39.2	29.8	12.1
Source Pre-trained	Flickr30k	57.3	36.2	21.9	13.3	15.1	38.8	25.3	8.6
DCC	Flickr30k	54.3	34.6	21.8	13.8	16.1	38.8	27.7	9.7
Ours	Flickr30k	**62.1**	**41.7**	**27.6**	**17.9**	**16.7**	**42.1**	**32.6**	**9.9**
Fine-tuning	Flickr30k	59.8	41	27.5	18.3	18	42.9	35.9	11.5

rectly on paired training data in the target domain (referred to as Fine-tuning). Ideally, this serves as the upper bound [4] of our experiments.

We further categorize three kinds of domain shift between MSCOCO and other target datasets, namely general v.s. fine-grained descriptions, difference in verb usage and subtle difference in sentence style.

General v.s. fine-grained descriptions. The large domain shift between MSCOCO and CUB-200/Oxford-102 suggests that it is the most challenging domain adaptation scenario. In CUB-200/Oxford-102, descriptions give detailed expressions of attributes such as beak of a bird or stamen of a flower. In contrast, in MSCOCO, descriptions usually are about the main scene and character. We illustrate the differences at word-level distribution among MSCOCO, CUB-200, and Oxford-102 using Venn-style word clouds [7] (see Fig. 4 [5]).

On the top two rows of Fig. 5 show that our model can describe birds and flowers in detailed and also the appearance of fine-grained object attributes. In the top two blocks of Table 1, our method outperforms DCC and Source Pre-trained models by a considerable margin for all evaluation metrics.

Difference in verb usage. Next, we move towards the verb usage difference between the source and target domains. According to [22], there are more motion verbs (30% in TGIF vs. 19% in MSCOCO) such as dance and shake, and more facial expressions in TGIF, while verbs in MSCOCO are mostly static ones such as stand and sit. Examples in Fig. 5 show that our model can accurately describe human activities or object interactions. On the third panel of Table 1, our method also significantly improves over Source Pre-trained and DCC models.

Subtle difference in sentence style. In order to test the generalizability of our method, we conduct an experiment using similar dataset (i.e. Flickr30k) as target domain. In the bottom block of Table 1, our method also offers a noticeable improvement. In addition, we reverse the route of adaptation (i.e. from Flickr30k to MSCOCO). Our method (CIDEr 38.2%, SPICE 8.9%) also improves over source pre-trained model (CIDEr 27.3%, SPICE 7.6%). To sum up, our method shows great potentials for unsupervised domain adaptation across datasets regardless of regular or large domain shift.

Critic-based planning. Instead of directly generating the word y_t from policy network $\pi(y_t|.)$, we take the advantage of its adversary, critics, during the inference. The results is shown in Table 2. The threshold Γ is set to 0.15 in CUB-200 and to 0.1 in Oxford-102. In every time-step, we choose top $J = 2$ words according to $\pi_\theta(.)$. Out of 16.2% and 9.4% of words are determined by the critics in CUB-200, and Oxford-102, respectively. Compared to greedy search, critic-based planning can achieve better performance in many evaluation metrics, especially in datasets with large domain shift from the source domain dataset (e.g., CUB-200 and Oxford-102). Compared to beam search with beam size 2, critic-based planning also typically gets a higher performance. Beam search method generates the words only depending on captioner itself, while critic-based planning method acquires a different point of view from the critics. For the case of regular domain shift (e.g., TGIF and Oxford-102), critic-based planning achieves comparable performance with beam search and greedy search. Some impressive examples are shown in Fig. 6.

4.3. Ablation Study

We have proposed an adversarial training procedure with two critic models: Multi-modal Critic (MC) and Domain

[4] We find that the model directly trained on all paired data in target domain performs worse than fine-tuning. Please see Supplementary for details.

[5] Visualization generated using http://worditout.com/.

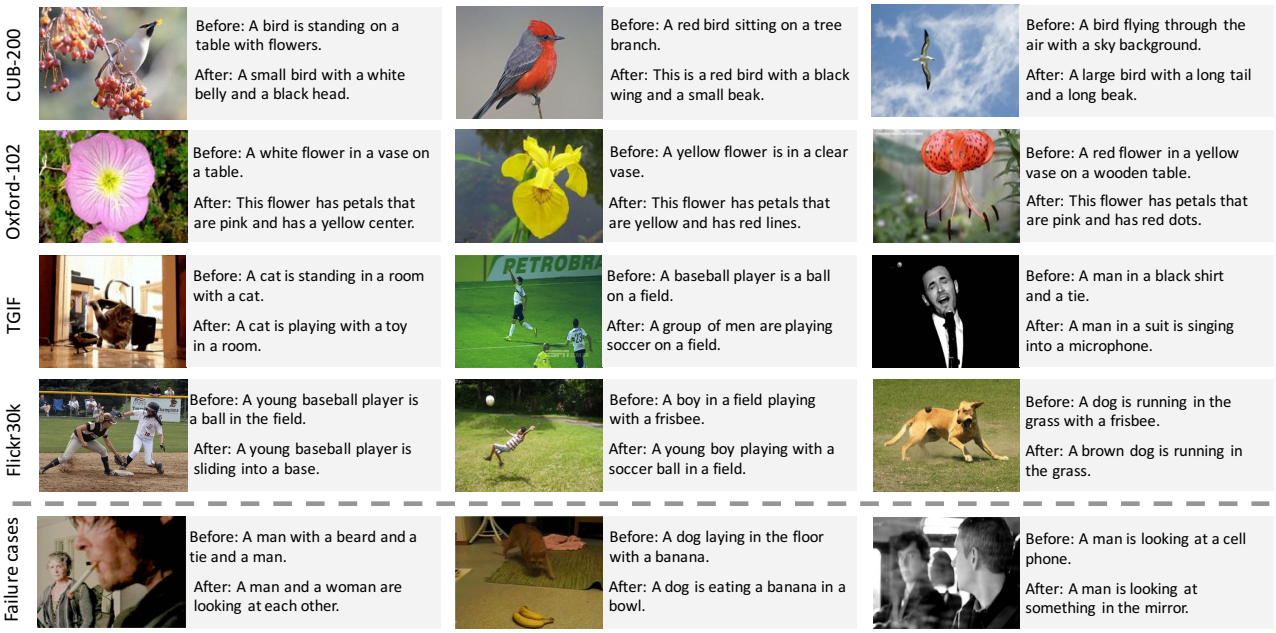


Figure 5: Examples of captions before and after domain adaptation for all four target domain datasets. The last row demonstrates the failure cases, where the generated captions do not accurately describe the images.

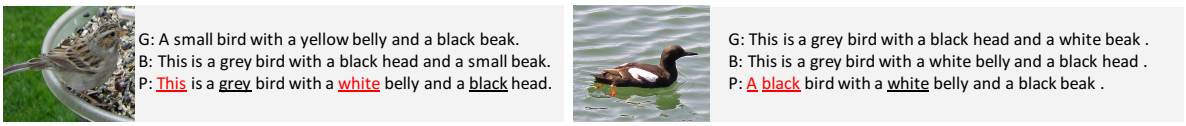


Figure 6: Results of critic-based planning. G stands for greedy search, B for beam search, and P for critic-based planning. The underlined words denote that the difference between the maximum probability and the second largest probability of π is lower than Γ (selected by critic). When critic-based planning does not choose the word with maximum probability of π, the word is colored in red.

Table 2: Results of proposed critic-based planning compared with greedy search and beam search.

Method	Bleu-4	Meteor	ROUGE	CIDEr-D
MSCOCO → CUB-200				
Greedy Search	32.8	27.6	58.6	24.8
Beam Search	33.1	27.5	58.3	26.2
Planning	**35.2**	27.4	58.5	**29.3**
MSCOCO → Oxford-102				
Greedy Search	60.5	36.4	72.1	**29.3**
Beam Search	60.3	36.3	72	28.3
Planning	**62.4**	**36.6**	**72.6**	24.9
MSCOCO → TGIF				
Greedy Search	10.3	14.5	37	22.2
Beam Search	10.5	14.2	36.7	22.6
Planning	10.3	14.4	37	21.9
MSCOCO → Flickr30k				
Greedy Search	17.5	16.4	41.9	32.2
Beam Search	18.2	16.4	42.1	33.3
Planning	17.3	16.5	41.7	32.3

Critic (DC). In order to analyze the effectiveness of these two critics, we do ablation comparison with either one and both. Table 3 shows that using MC only is insufficient since MC is not aware of the sentence style in target domain. On the other hand, using DC only contributes significantly. Finally, combining both MC and DC achieves the best performance for all evaluation metrics. We argue that both MC and DC are vital for cross-domain image captioning.

Table 3: Ablation study for two critic models on Flickr30k. MC: Multi-modal Critic, DC: Domain Critic.

Method	Bleu-4	Meteor	ROUGE	CIDEr-D
Source Pre-trained	13.3	15.1	38.8	25.3
+MC	13.7	15.2	38.8	25.9
+DC	17.6	16.3	41.4	32.1
+MC+DC	**17.9**	**16.7**	**42.1**	**32.6**

5. Conclusion

We propose a novel adversarial training procedure (captioner v.s. critics) for cross-domain image captioning. A novel critic-based planning method is naturally introduced to further improve the caption generation process in testing. Our method consistently outperforms baseline methods on four challenging target domain datasets (two with large domain shift and two with regular domain shift). In the future, we would like to improve the flexibility of our method by combining multiple critics in a plug-and-play fashion.

Acknowledgement We thank Microsoft Research Asia, MediaTek and MoST 106-3114-E-007-004 for their support. We also thank Kuo-Hao Zeng and Shao-Pin Chang for useful feedbacks during internal review.

References

[1] H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, and M. Marchand. Domain-adversarial neural networks. In *NIPS workshop on Transfer and Multi-Task Learning: Theory meets Practice*, 2014. 2

[2] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Guided open vocabulary image captioning with constrained beam search. *CoRR*, abs/1612.00576, 2016. 2

[3] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, and D. Parikh. Vqa: Visual question answering. In *ICCV*, 2015. 5

[4] D. Bahdanau, P. Brakel, K. Xu, A. Goyal, R. Lowe, J. Pineau, A. Courville, and Y. Bengio. An actor-critic algorithm for sequence prediction. In *ICLR*, 2017. 2

[5] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *NIPS*, 2015. 2, 6

[6] Y.-H. Chen, W.-Y. Chen, Y.-T. Chen, B.-C. Tsai, Y.-C. F. Wang, and M. Sun. No more discrimination: cross city adaptation of road scene segmenters. In *ICCV*, 2017. 3

[7] G. Coppersmith and E. Kelly. Dynamic wordclouds and vennclouds for exploratory data analysis. In *Workshop on Interactive Language Learning, Visualization, and Interfaces*. 7

[8] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 2

[9] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *CVPR*, 2015. 1, 2

[10] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky. Domain-adversarial training of neural networks. *JMLR*, 17(59):1–35, 2016. 2

[11] A. Goyal, A. Lamb, Y. Zhang, S. Zhang, A. C. Courville, and Y. Bengio. Professor forcing: A new algorithm for training recurrent networks. In *NIPS*, 2016. 2

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[13] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *ECCV*, 2016. 2

[14] L. A. Hendricks, S. Venugopalan, M. Rohrbach, R. Mooney, S. Kate, and T. Darrell. Deep compositional captioning: Describing novel object categories without paired training data. In *CVPR*, 2016. 2, 6

[15] J. Hoffman, D. Wang, F. Yu, and T. Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. 3

[16] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017. 6

[17] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 2, 6

[18] Y. Kim. Convolutional neural networks for sentence classification. *EMNLP*, 2014. 4

[19] Y. Kim, Y. Jernite, D. Sontag, and A. M. Rush. Character-aware neural language models. *AAAI*, 2016. 4

[20] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *ICLR*, 2015. 6

[21] R. Kiros, R. Salakhutdinov, and R. S. Zemel. Unifying visual-semantic embeddings with multimodal neural language models. *TACL*, 2015. 1, 2

[22] Y. Li, Y. Song, L. Cao, J. Tetreault, L. Goldberg, A. Jaimes, and J. Luo. A new dataset and benchmark on animated gif description. In *CVPR*. 2016. 2, 6, 7

[23] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*. 2014. 1, 2, 6

[24] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Optimization of image description metrics using policy gradient methods. *CoRR*, abs/1612.00370, 2016. 2

[25] M. Long, Y. Cao, J. Wang, and M. I. Jordan. Learning transferable features with deep adaptation networks. In *ICML*, 2015. 2

[26] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *ICVGIP*, 2008. 2, 6

[27] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *ICLR*, 2016. 2

[28] S. Reed, Z. Akata, H. Lee, and B. Schiele. Learning deep representations of fine-grained visual descriptions. In *CVPR*, 2016. 2

[29] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. *CoRR*, abs/1612.00563, 2016. 2

[30] R. S. Sutton, D. A. McAllester, S. P. Singh, Y. Mansour, et al. Policy gradient methods for reinforcement learning with function approximation. 2

[31] S. Venugopalan, L. A. Hendricks, M. Rohrbach, R. J. Mooney, T. Darrell, and K. Saenko. Captioning images with diverse objects. *CoRR*, abs/1606.07770, 2016. 2

[32] S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. In *ICCV*, 2015. 2

[33] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *CVPR*, 2015. 1, 2, 3

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical report, 2011. 1, 2, 6

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. *ICML*, 2015. 2

[36] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville. Describing videos by exploiting temporal structure. In *ICCV*, 2015. 2

[37] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *TACL*, 2014. 2, 6

[38] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: sequence generative adversarial nets with policy gradient. *AAAI*, 2017. 2

[39] W. Zellinger, T. Grubinger, E. Lughofer, T. Natschläger, and S. Saminger-Platz. Central moment discrepancy (CMD) for domain-invariant representation learning. In *ICLR*, 2017. 2

[40] K.-H. Zeng, T.-H. Chen, J. C. Niebles, and M. Sun. Title generation for user generated videos. In *ECCV*, 2016. 2