Interactive Dual Generative Adversarial Networks for Image Captioning

Junhao Liu,[1,2] Kai Wang,[1] Chunpu Xu,[3] Zhou Zhao,[4] Ruifeng Xu,[5] Ying Shen,[6] Min Yang[1]*
[1]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Huazhong University of Science and Technology
[4]Zhejiang University [5]Harbin Institute of Technology (Shenzhen)
[6]Peking University Shenzhen Graduate School
{jh.liu, kai.wang, min.yang}@siat.ac.cn, cpx@hust.edu.cn, zhaozhou@zju.edu.cn
xuruifeng@hit.edu.cn, shenying@pkusz.edu.cn

Abstract

Image captioning is usually built on either generation-based or retrieval-based approaches. Both ways have certain strengths but suffer from their own limitations. In this paper, we propose an Interactive Dual Generative Adversarial Network (IDGAN) for image captioning, which mutually combines the retrieval-based and generation-based methods to learn a better image captioning ensemble. *IDGAN* consists of two generators and two discriminators, where the generation- and retrieval-based generators mutually benefit from each other's complementary targets that are learned from two dual adversarial discriminators. Specifically, the generation- and retrieval-based generators provide improved synthetic and retrieved candidate captions with informative feedback signals from the two respective discriminators that are trained to distinguish the generated captions from the true captions and assign top rankings to true captions respectively, thus featuring the merits of both retrieval-based and generation-based approaches. Extensive experiments on MSCOCO dataset demonstrate that the proposed *IDGAN* model significantly outperforms the compared methods for image captioning.

1 Introduction

Automatic image captioning aims to produce a textual description (usually a sentence) that verbalizes the visual content of an image. Image captioning methods can be roughly divided into two categories: retrieval-based and generation-based methods. Retrieval-based methods describe images by retrieving pre-existing captions from a repository, while generation-based methods synthesize a textual description (typically a sentence) that verbalizes the query image. Both ways have certain advantages but suffer from their own disadvantages.

When a user issues a query image, the retrieval-based methods search a corresponding caption that best matches the query image in a pre-constructed image-caption repository (Ordonez, Kulkarni, and Berg 2011; Hodosh, Young, and Hockenmaier 2013; Gong et al. 2014). For example, (Hodosh, Young, and Hockenmaier 2013) performed image captioning as a ranking or retrieval task, and introduced a ranking-based method to evaluate systems on a sentence-based image description. (Gong et al. 2014) associated images with descriptive sentences by projecting them into a common latent space. Although retrieval-based methods can produce general and syntactically correct captions, the retrieved captions are not tailored for the query images and limited by the capacity of the pre-constructed repository.

To make a caption tailored appropriately for the query image, a better way is to generate a new one accordingly. A typical generation-based image captioning model is encoder-decoder paradigm (Karpathy and Fei-Fei 2015; Vinyals et al. 2015), which consists of two neural networks: a convolutional neural network (CNN) based encoder encodes a given image into a vector representation, based on which a long short-term memory network (LSTM) decoder decodes the vector representation to generate a variable-length image caption word by word. For example, (Xu et al. 2015) proposed an attentive encoder-decoder neural network to dynamically attend to different locations of the images when decoding different words in the captions. (Mun, Cho, and Han 2017) used associated captions that were retrieved from training data to learn visual attention for image captioning. These methods can synthesize a new sentence as the caption, which brings the results of good flexibility and quality. Nevertheless, a well-known problem for generation-based methods is that they are prone to generate universal or non-fluent captions, which do not appropriately reflect the meaning of the given image.

Previously, the retrieval-based and generation-based systems with their own characteristics have been developed separately. In this paper, we are seeking to absorb their merits. We propose an Interactive Dual Generative Adversarial Network (IDGAN) for image captioning, which bridges the communication between generation- and retrieval-based methods by mutually reviewing each other. The dual adversarial training mechanism is established between two generation- and retrieval-based generators, and two generation- and retrieval-based discriminators, to make the generated captions indistinguishable from the ground-truth captions. Specifically, a language model (LM) generator G_{θ_1} synthesizes tailored captions for the query im-

[*] Min Yang is corresponding author (min.yang@siat.ac.cn).

age, and a generative ranker G_{θ_2} ranks both retrieved and synthetic captions. A discriminator D_{ϕ_1} not only differentiates generated captions from human-written captions but also distinguishes bad generated captions from good ones. Another discriminative ranker D_{ϕ_2} attempts to distinguish the ground-truth captions and the adversarial candidates provided by both generators (G_{θ_1} and G_{θ_2}), which are trained synchronously with the two generators using the adversarial training framework.

This paper has three main contributions listed as follows.

- We introduce an interactive dual generative adversarial framework to mutually enhance both retrieval-based and generation-based image captioning methods, leading to a better ensemble model for image captioning.
- We devise a copy mechanism that naturally incorporates the retrieved guidance captions into the decoding process, enriching the informativeness and diversity of the generated captions.
- Experimental results show that *IDGAN* model significantly outperforms the state-of-the-art image captioning methods on the widely used MSCOCO dataset.

2 Related Work

Existing image captioning methods can be roughly divided into two categories: retrieval-based and generation-based methods. Retrieval-based methods first extracted the visually similar images with their captions from a pre-constructed image-caption repository, forming a candidate captions pool. The final candidate captions for the input image are then chosen from the captions pool by ranking methods (Ordonez, Kulkarni, and Berg 2011; Hodosh, Young, and Hockenmaier 2013; Gong et al. 2014). For example, (Hodosh, Young, and Hockenmaier 2013) treated image captioning as a ranking or retrieval task, and introduced a ranking-based method to evaluate systems on a sentence-based image description. (Gong et al. 2014) associated images with descriptive sentences by projecting them into a common latent space. Although retrieval-based methods can produce general and syntactically correct captions, the retrieved captions are not tailored for the query images and limited by the capacity of the pre-constructed repository.

Inspired by the great success of deep learning algorithms in computer vision and natural language processing, generation-based image captioning methods mainly exploit the encoder-decoder architecture to produce sentences with flexible syntactical structures (Karpathy and Fei-Fei 2015; Vinyals et al. 2015; Gu et al. 2017). For example, Karpathy and Fei-Fei (2015) learned about the inter-model correspondences between language and image data by using the training image-caption pairs. Attention mechanism has been proved to be able to significantly improve the performance of the underlying encoder-decoder based methods (Mun, Cho, and Han 2017; Gu et al. 2018; Yang et al. 2018; Zhao et al. 2018). Mun, Cho, and Han (2017) used associated captions that were retrieved from training data to learn visual attention for image captioning. Chen et al. (2017) encoded the images with multi-layer feature maps, capturing the spatial locations and channels via visual attention mechanisms.

There were also several recent generation-based studies exploring GAN and reinforcement learning techniques for image captioning (Liu et al. 2017; Rennie et al. 2017). Rennie et al. (2017) introduced an SCST algorithm by using the REINFORCE algorithm to optimize the model. Rather than estimating a "baseline" to reduce the model variance, SCST used its output to normalize the expected rewards. Recently, some works proposed to employ GAN to generate text descriptions for the input images. Xu et al. (2019) proposed an adversarial learning method for image captioning, which enhances the caption generation with retrieved guidance captions.

3 Our Methodology

3.1 Problem Definition and Model Overview

Given an image x, image captioning aims to generate a text description $y = \{w_1, w_2, ..., w_T\}$ for image x, where T is the length of the text sequence.

As depicted in Figure 1, *IDGAN* consists of the following components: (1) Generative sequence-to-sequence (seq2seq) model G_{θ_1}, which is responsible for synthesizing M_1 image caption candidates $\{\hat{y}_{m=1}^{M_1}\}$ given an image x by the Monte Carlo (MC) roll-out policy. Such process is also noted as $G_{\theta_1}(\hat{y}|x)$; (2) Generative ranking model G_{θ_2}, which computes a relevance score between each image-caption pair and retrieves M_2 caption candidates $\{\langle x, \tilde{y}_{m=1}^{M_2}\rangle\}$. Such process is denoted as $G_{\theta_2}(\tilde{y}|x)$; (3) Discriminative classification model D_{ϕ_1}, which tries to distinguish the human-written captions from adversarial captions generated by G_{θ_1}; (4) Discriminative ranking model D_{ϕ_2}, which inherits from the same ranking model as G_{θ_2}, trying to distinguish the true image-caption pairs from adversarial candidates provided by both generators. By learning over symmetric feedback signals from two dual adversarial discriminators, the generation- and retrieval-based models mutually benefit from each other's complementary targets, leading to better image captioning. For each test image, we use the sentences generated by the generative model G_{θ_1} as the final output caption.

3.2 Generative Seq2Seq module

The sequence to sequence (seq2seq) (Karpathy and Fei-Fei 2015; Vinyals et al. 2015) framework is used as the backbone of our generation-based image captioning model. In encoding, we retrieve the candidate captions to augment the semantic information of the image and thus learn better representation of the image. In the decoding stage, we integrate the guidance captions into the word generation process by designing a copy mechanism so as to enrich the meaning of the generated captions. Next, we will describe the encoder and decoder in detail.

Pre-retrieval Model for Retrieving Candidate Captions The candidate captions are defined as the ground-truth captions of k nearest training images for the query image based on visual similarity. The image features are computed for

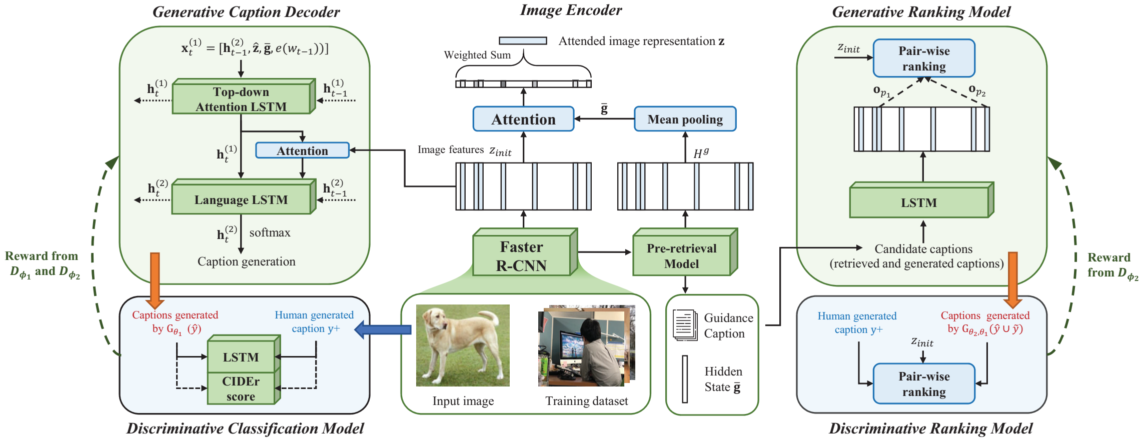
Generative Caption Decoder
$\mathbf{x}_t^{(1)} = [\mathbf{h}_{t-1}^{(2)}, \hat{\mathbf{z}}, \bar{\mathbf{g}}, e(w_{t-1}))]$
$\mathbf{h}_t^{(1)}$
Top-down Attention LSTM
$\mathbf{h}_{t-1}^{(1)}$
$\mathbf{h}_t^{(1)}$
Attention
$\mathbf{h}_t^{(2)}$
Language LSTM
$\mathbf{h}_{t-1}^{(2)}$
$\mathbf{h}_t^{(2)}$
softmax
Caption generation
Reward from D_{ϕ_1} and D_{ϕ_2}
Captions generated by G_{θ_1} $(\hat{y})$
Human generated caption y+
LSTM
CIDEr score
Discriminative Classification Model
Image Encoder
Attended image representation z
Weighted Sum
Attention
$\bar{\mathbf{g}}$
Mean pooling
Image features z_{init}
H^g
Faster R-CNN
Pre-retrieval Model
Input image
Training dataset
Guidance Caption
Hidden State $\bar{\mathbf{g}}$
Generative Ranking Model
z_{init}
Pair-wise ranking
$\mathbf{o}_{p_1}$
$\mathbf{o}_{p_2}$
LSTM
Candidate captions (retrieved and generated captions)
Reward from D_{ϕ_2}
Human generated caption y+
Captions generated by $G_{\theta_2,\theta_1}(\hat{y} \cup \tilde{y})$
z_{init}
Pair-wise ranking
Discriminative Ranking Model

Figure 1: *IDGAN* architecture consists of two generators G_{θ_1} and G_{θ_2}, and two discriminators D_{ϕ_1} and D_{ϕ_2}.

every image in the training dataset with Eq.(2). The neighbor images are found by exhaustively computing the cosine similarity between the query image and the training images. We maintain a set of top k captions, denoted as $\{c_1, c_2, \ldots, c_k\}$, in terms of the similarity score as candidate captions, which are concatenated into a guidance caption $C = [c_1, c_2, \ldots, c_k]$.

LSTM network is then employed to extract the semantic meanings of the guidance caption. Formally, given the input word embedding $e(w_i^r)$, the hidden state $\mathbf{g}_i$ is computed from the previous hidden state $\mathbf{g}_{i-1}$ as:

$$\mathbf{g}_i = \text{LSTM}(\mathbf{g}_{i-1}, e(w_i^r)) \tag{1}$$

where $e(w_i^r)$ denotes the embedding of the i-th word in retrieved guidance caption. The hidden states of the guidance caption C is represented as $H^g = [\mathbf{g}_1, \ldots, \mathbf{g}_m]$ and m denotes the length of guidance caption C. We use $\bar{\mathbf{g}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{g}_i$ to represent the average vector of the guidance caption.

Image Encoder Following the similar strategy as in (Karpathy and Fei-Fei 2015; Vinyals et al. 2015; Anderson et al. 2018), in this paper, we use the Faster R-CNN (Ren et al. 2015) that is pre-trained on Visual Genome (Krishna et al. 2017) to compact the raw image x into L vectors (with size D), where each vector represents the features learned at different detection of x. Formally, we refer to these annotated vectors as:

$$\mathbf{z}_{init} = \left\{\mathbf{z}_{init,1}, \mathbf{z}_{init,2}, ..., \mathbf{z}_{init,L}\right\} = \text{R-CNN}(x) \tag{2}$$

An attention mechanism is employed to capture the crucial information from the input image. In particular, we take as input the average representation of the guidance caption C as attention source to learn knowledge-aware image representation $\mathbf{z}$ as:

$$\mathbf{z} = \sum_{i=1}^{L} \alpha_i \mathbf{z}_{init,i}, \quad \alpha_i = \frac{\exp\left(\sigma(\bar{\mathbf{g}}, \mathbf{z}_{init,i})\right)}{\sum_{j=1}^{L} \exp\left(\sigma(\bar{\mathbf{g}}, \mathbf{z}_{init,j})\right)} \tag{3}$$

where σ is a feed-forward neural network that converts a vector to a real-valued score, α_i is attention weight for the i-th image feature $\mathbf{z}_{init,i}$.

Caption Decoder The generation of image captions is performed by an LSTM decoder based on the learned image representations and the guidance caption representation.

Two-layer Attention Networks Similar to (Anderson et al. 2018), caption decoder contains a two-layer LSTM network. The first LSTM layer (denoted as LSTM$^{(1)}$) is characterized as a top-down attention model, and the second LSTM layer (denoted as LSTM$^{(2)}$) is a language model. The first LSTM model takes the concatenation of the previous output of the language LSTM (i.e., $\mathbf{h}_{t-1}^{(2)}$), the image vector $\mathbf{z}_{init}$, the attentive image feature $\mathbf{z}$, the average feature vector of the guidance caption $\bar{\mathbf{g}}$, and the word embedding of the previous word (i.e., $e(w_{t-1})$) as input:

$$\mathbf{x}_t^{(1)} = \left[\mathbf{h}_{t-1}^{(2)}, \hat{\mathbf{z}}, \bar{\mathbf{g}}, e(w_{t-1})\right] \tag{4}$$

where $\mathbf{x}_t^{(1)}$ denotes the input of LSTM$^{(1)}$ at time t; $\hat{\mathbf{z}} = [\bar{\mathbf{z}}_{init}, \mathbf{z}]$ and $\bar{\mathbf{z}}_{init} = \frac{1}{L}\sum_{i=1}^{L} \mathbf{z}_{init,i}$. At time step t, the hidden state of LSTM$^{(1)}$ can be calculated as:

$$\mathbf{h}_t^{(1)} = \text{LSTM}^{(1)}\left(\mathbf{h}_{t-1}^{(1)}, \mathbf{x}_t^{(1)}\right) \tag{5}$$

Given the output of the first LSTM at time step t (i.e., $\mathbf{h}_t^{(1)}$), we learn the attentive image representation $\tilde{\mathbf{z}}_t$, which is then fed into LSTM$^{(2)}$. The attentive image vector $\tilde{\mathbf{z}}_t$ ensures that, at each time step, the decoder is capable of getting full information of the initial image representation $\mathbf{z}_{init}$. We compute $\tilde{\mathbf{z}}_t$ when we decode the t-th word as:

$$\tilde{\mathbf{z}}_t = \sum_{i=1}^{L} \beta_{t,i}\mathbf{z}_{init,i}, \quad \beta_{t,i} = \frac{\exp(\sigma(\mathbf{h}_{t-1}^{(1)}, \mathbf{z}_{init,i}))}{\sum_{j=1}^{L} \exp(\sigma(\mathbf{h}_{t-1}^{(1)}, \mathbf{z}_{init,j}))} \tag{6}$$

where σ is a feed-forward neural network, as defined in Eq. (3). The attention weight $\beta_{t,i}$ represents the alignment between the i-th location in the image and the t-th generated word.

We employ LSTM$^{(2)}$ to generate an image caption word by word. The concatenation of the output of LSTM$^{(1)}$ and the attentive image representation ($\tilde{\mathbf{z}}_t$) is used as the input of LSTM$^{(2)}$, which is represented as:

$$\mathbf{x}_t^{(2)} = \left[\tilde{\mathbf{z}}_t, \mathbf{h}_t^{(1)}\right] \tag{7}$$

At time step t, the hidden state of LSTM$^{(2)}$ is calculated as:

$$\mathbf{h}_t^{(2)} = \text{LSTM}^{(2)}\left(\mathbf{x}_t^{(2)}, \mathbf{h}_{t-1}^{(2)}\right) \tag{8}$$

Caption Generation We assume a vocabulary $\mathcal{V}^c = \{w_1^c, \dots, w_N^c\}$. The generation model is typically a classifier over the vocabulary $\mathcal{V}^c$. In particular, we feed the representation $\mathbf{h}_t^{(2)}$ into a fully connected layer followed by a softmax layer to generate the image caption. Formally, the generation probability of the t-th word is computed by

$$P_{\theta_1}^c(w_t = w^c) = \text{softmax}(W_c \mathbf{h}_t^{(2)}), \;\; w^c \in \mathcal{V}^c \tag{9}$$

where W^c is the parameter to be learned, θ_1 denotes the parameters of the generation-based model.

Copy Mechanism We also employ copy mechanism to explicitly extract words form the retrieved guidance caption. We assume another set of words $\mathcal{V}^g = \{w_1^g, \dots, w_M^g\}$ for all the unique words in the guidance captions. Since $\mathcal{V}^g$ may contain words not in $\mathcal{V}^c$, copying words in $\mathcal{V}^g$ enables the decoder to output some out-of-vocaulary (OOV) words.

Given the hidden states $H^g = [\mathbf{g}_1, \dots, \mathbf{g}_m]$ for the guidance caption C, we computed the context vector for the guidance caption as a weighted sum of the hidden states H^g:

$$\mathbf{c}_t^g = \sum_{i=1}^{m} \gamma_{t,i}\mathbf{g}_i, \quad \gamma_{t,i} = \frac{\exp(\varrho(\mathbf{g}_i, \mathbf{h}_t^{(2)}))}{\sum_{j=1}^{m}\exp(\varrho(\mathbf{g}_j, \mathbf{h}_t^{(2)}))} \tag{10}$$

where $\mathbf{h}_t^{(2)}$ is defined in Eq. (8), ϱ is a multilayer perceptron.

Formally, the generator selects a word w^g from $\mathcal{V}^g$ at time step t as follows:

$$P_{\theta_1}^g(w_t = w^g) = \text{softmax}(W_g[\mathbf{h}_t^{(2)}; \mathbf{c}_t^g]), \;\; w^g \in \mathcal{V}^g \tag{11}$$

where W_g indicates the learnable parameter.

Finally, at each time step, the caption generator selects a generic word from $\mathcal{V}^c$ or copies a word from $\mathcal{V}^g$ with the following distribution:

$$\lambda_t = \text{sigmoid}(U_o[\mathbf{h}_t^{(2)}; \mathbf{c}_t^g]) \tag{12}$$

$$\hat{w}_t \sim P_{\theta_1}(w_t) = \begin{bmatrix}(1-\lambda_t)P_{\theta_1}^c(w_t = w^c) \\ \lambda_t P_{\theta_1}^g(w_t = w^g)\end{bmatrix} \tag{13}$$

$\lambda_t \in [0, 1]$ is a scalar to balance the choice between generating a content word w_c and copying a word w_g from guidance captions. $P_{\theta_1}(y_t)$ is the final word probability distribution. U_o is learnable parameter.

3.3 Generative Ranking Model

For the retrieval-based approach, we devise a generative ranking model G_{θ_2} to retrieve l competitive candidate captions. We first extract the visually similar images with their captions from the training data set by using the pre-retrieval module defined in Section 3.2. The generated and retrieved captions are denoted as candidate captions P. The captions for the query image are selected from these candidate captions pool P by a ranking model. Specifically, given a query image x, pre-retrieved candidate captions $D = \{d_1, d_2, \dots, d_l\}$, and the captions $\hat{y} = \{\hat{y}_1, \hat{y}_2, \dots, \hat{y}_{M_1}\}$ generated by G_{θ_1}, we compute a relevance score for each (x, p_i) pair, where $p_i \in D \cup \hat{y}$. Each candidate caption p is first encoded into a distributed representation $\mathbf{o}$ by using an LSTM network:

$$\mathbf{h}_i^p = \text{LSTM}(\mathbf{h}_i^p, e(w_i^p)), \quad \mathbf{o} = \mu(\left[\mathbf{h}_1^p, \mathbf{h}_2^p, ..., \mathbf{h}_{L_p}^p\right]) \tag{14}$$

where $\mathbf{h}_i^p$ is the i-th hidden state for p, $e(w_i^p)$ denotes the word embedding of the i-th word in the caption p, L_p is the length of the caption p, μ is the averaging operation.

Instead of an absolute relevance, we optimize IDGAN by using a pair-wise ranking method since relative preference is usually more easily learned. The probability of a caption pair $\langle p_1, p_2\rangle$ with p_1 more relevant than p_2 being correctly ranked can be measured by the distance of their matching degree to the query image x:

$$\begin{aligned} P_{\theta_2}(\langle p_1, p_2\rangle|x) = \rho(g(&f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{p_1})) \\ &- g(f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{p_2}))) \end{aligned} \tag{15}$$

where ρ is the sigmoid function, f_{FC} is a fully connected layer, g is any scoring function (i.e., cosine similarity), θ_2 denotes the parameters of the retrieval-based model. $\mathbf{z}_{init}$, $\mathbf{o}_{p_1}$ and $\mathbf{o}_{p_2}$ are representations of the image x, caption p_1 and caption p_2.

We define a triplet ranking-based loss, which maximizes the relevance between the query image x and the true caption y^+, and minimizes the relevance between the query image x and sampled negative caption y^-:

$$\begin{aligned} L_{rank} = \max(0, \varrho + g(&f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{y^+})) \\ &- g(f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{y^-}))) \end{aligned} \tag{16}$$

where ϱ denotes the desired margin between the similarities, y^- is the negative caption randomly chosen from the entire captions with true and candidate captions excluded.

3.4 Discriminative Classification Model

Ideally, a good image caption should be assigned a high adequacy score and contribute more to updating the generator G_{θ_1}. Therefore, we expect the language model discriminator to not only differentiate generated captions from human-written captions but also distinguish bad generated captions from good ones. We propose a new objective of the language model discriminator to assign a precise score for each generated caption, which is consistent with their adequacy score:

$$D_{\phi_1}(\hat{y}|x) = D_{binary}(x, \hat{y}) + D_{score}(x, \hat{y}) \tag{17}$$

where $D_{binary}(x, \hat{y})$ is a binary classifier implemented by LSTM that aims at distinguishing the input caption as originally generated by humans or synthesized by the generator G_{θ_1}. $D_{score}(x, \hat{y})$ is computed by an evaluation metric (i.e., CIDEr score) that compares the generated caption $\hat{y}$ to the corresponding ground-truth caption y^+.

3.5 Discriminative Ranking Model

The discriminator adopts the same ranking model as G_{θ_2}, which aims to distinguish the ground truth captions from the candidate captions produced by both generators (G_{θ_1} and G_{θ_2}). Concretely, the discriminative model aims to learn the image and caption representations such that the probability that the positive pair $\langle x, y^+\rangle$ is assigned larger similarity score than the negative pair $\langle x, y^{gen}\rangle$, where y^+ and y^{gen} indicate the true caption and the caption generated by the two generators. Here, we have $y^{gen} \in \{\tilde{y} \cup \hat{y}\}$. Similar to Eq.(15), given the query image x, the probability of a caption pair $\langle x, y^{gen}\rangle$ being correctly ranked can be computed using the distance of their relevance scores to the query image x:

$$\begin{aligned} D_{\phi_2}(\langle y^+, y^{gen}\rangle|x) = \rho(g(f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{y^+})) \\ - g(f_{FC}(\mathbf{z}_{init}), f_{FC}(\mathbf{o}_{y^{gen}}))) \end{aligned} \tag{18}$$

where ρ, g, and f_{FC} are defined in Eq. (15). We adopt the same training loss as defined in Eq. (16) to optimize the discriminative ranking model D_{ϕ_2}.

4 Adversarial Dual Objective

In *IDGAN* framework, the two generators G_{θ_1}, G_{θ_2} and the two discriminators D_{ϕ_1}, D_{ϕ_2} form two interactive dual generative adversarial networks, where the two generators attempt to produce fake captions that achieve high scores so as to fool the two discriminators respectively, while the two discriminators on the contrary are expected to score down the generated and retrieved captions. Their minimax game is summarized as the following objective function $\mathcal{L}$:

$$\mathcal{L} = \min_{\theta_1,\theta_2} \max_{\phi_1,\phi_2} (\mathcal{L}_g + \mathcal{L}_r) \tag{19}$$

$$\mathcal{L}_g = \mathbb{E}_{y^+\sim p_{data}}[\log D_{\phi_1}(y^+)] + \mathbb{E}_{\hat{y}\sim G_{\theta_1}}[\log(1 - D_{\phi_1}(\hat{y}))] \tag{20}$$

$$\begin{aligned} \mathcal{L}_r = \mathbb{E}_{\langle y^+,y^-\rangle\sim p_{data}}[\log D_{\phi_2}(\langle y^+, y^-\rangle)] \\ + \mathbb{E}_{y^{gen}\sim G_{\theta_2},G_{\theta_2},G_{\theta_1}}[\log(1 - D_{\phi_2}(\langle y^+, y^{gen}\rangle))] \end{aligned} \tag{21}$$

where $\mathbb{E}$ indicates the mathematical expectation, y^+ is the ground-truth caption for the query image, y^- is the negative caption randomly chosen from the entire captions with true and candidate captions excluded.

4.1 Optimizing Discriminative Models

The objective of discriminative models D_{ϕ_1} and D_{ϕ_2} are to maximize the probability of correctly distinguishing the ground truth captions from the generated captions. For the two generators fixed, we can obtain the optimal parameters for the discriminative models D_{ϕ_1} and D_{ϕ_2} with the following formulation:

$$\phi_1^*, \phi_2^* = \text{argmax}_{\phi_1,\phi_2}(\mathcal{L}_g + \mathcal{L}_r) \tag{22}$$

where $\mathcal{L}_g$ and $\mathcal{L}_r$ are defined in Eq.(20)-(21). This optimization problem is typically solved with gradient descent since D_{ϕ_1} and D_{ϕ_2} are differentiable with respect to ϕ_1 and ϕ_2, respectively.

4.2 Optimizing Generative Models

Generative Seq2seq Model Given a query image x, a caption sequence $\hat{y} = [\hat{w}_0, \hat{w}_1, ..., \hat{w}_T]$ is generated, which can be treated as a decision making process by policy $P_{\theta_1}(w_t|\hat{w}_{1:t-1}, x)$. It is difficult to back-propagate the gradients from the two discriminators to the generator G_{θ_1}, thus we use the policy gradient method to tackle this problem. With the true caption y^+ for image x, the reward of the generated image caption $\hat{y}$ is as follows:

$$J_{\theta_1}(\hat{y}|x) = \mathbb{E}_{\hat{y}\sim G_{\theta_1}}\{[1 - \log D_{\phi_1}(\hat{y})] + [1 - \log D_{\phi_2}(\langle y^+, \hat{y}\rangle)]\} \tag{23}$$

Generative Ranking Model We train G_{θ_2} to generate competitive image caption $\tilde{y}$ that achieves high ranking score from D_{ϕ_2}. More precisely, when given an image x and a scoring function, the probability of G_{θ_2} choosing a caption $\tilde{y}$ from candidate captions pool P is computed by Eq.(15). Formally, we ameliorate G_{θ_2} with the objective function as below:

$$J_{\theta_2}(\tilde{y}|x) = \mathbb{E}_{\tilde{y}\sim G_{\theta_2|\theta_1}}[\log(1 - D_{\phi_2}(\langle y^+, \tilde{y}\rangle|x))] \tag{24}$$

Policy Gradient We apply the policy gradient algorithm (Williams 1992) to update the parameters of the two generators since the sampling process of the generators is non-differential. Formally, with D_{ϕ_1} and D_{ϕ_2} fixed, for each query image x with true caption y^+, the minimization of $\mathcal{L}$ defined in Eq.(19) in response to θ_1 and θ_2 could be computed as follows:

$$\begin{aligned} \min_{\theta_1,\theta_2} \mathcal{L} = \max_{\theta_1,\theta_2}[\mathbb{E}_{\hat{y}_n\sim G_{\theta_1}} J_{\theta_1}(\hat{y}_n|x_n) \\ +\mathbb{E}_{\tilde{y}_n\sim G_{\theta_2|\theta_1}} J_{\theta_2}(\tilde{y}_n|x_n)] \end{aligned} \tag{25}$$

where J_{θ_1} and J_{θ_2} are defined in Eq.(23) and Eq.(24) respectively. T indicates the length of the training sample. Similar to (Rennie et al. 2017), the proximity gradient of the expected rewards of J_{θ_1} can be calculated as follows:

$$\nabla_{\theta_1} J_{\theta_1}(\hat{y}|x) \simeq \sum_{t=1}^{T}(R_1(\hat{y}) - R_1(y'))\nabla_{\theta_1} \log P_{\theta_1}(w_t|\hat{w}_{1:t-1}, x) \tag{26}$$

where y' is a generated caption by the greedy decoding process used as the baseline to reduce the training variance in reinforcement learning. R_1 is the reward function during adversarial training of G_{θ_1}, which is defined as:

$$R_1(\hat{y}) = \gamma_1 D_{\phi_1}(\hat{y}|x) + \gamma_2 D_{\phi_2}(\langle y^+, \hat{y}\rangle|x) + \gamma_3 D_{score}(\hat{y}|x) \tag{27}$$

where $\gamma_1, \gamma_2, \gamma_3$ are parameters that controls the effect of the three kinds of rewards. $D_{score}(\hat{y}|x)$ is computed by an evaluation metric (i.e., CIDEr score) by comparing the generated caption $\hat{y}$ to the corresponding ground-truth caption.

The gradient of the objective function Eq. (25) with respect to parameters θ_2 is:

$$\nabla_{\theta_2} J_{\theta_2}(\tilde{y}|x) \simeq \sum_{\tilde{y}} \nabla_{\theta_2} \log G_{\theta_2|\theta_1}(\tilde{y}|x) R_2(\langle y, \tilde{y}\rangle) \tag{28}$$

where R_2 is the reward function during adversarial training of G_{θ_2}, computed as:

$$R_2(\langle y, \tilde{y}\rangle) \equiv D_{\phi_2}(\langle y, \tilde{y}\rangle|x) \tag{29}$$

Methods	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**METEOR**	**ROUGE**	**CIDEr**
Hard-Attention (Xu et al. 2015)	71.8	50.4	35.7	25.0	23.0	-	-
VAE (Pu et al. 2016)	72.0	52.0	37.0	28.0	24.0	-	90.0
Attributes-CNN (Wu et al. 2016)	74.0	56.0	42.0	31.0	26.0	-	94.0
CNN$_{\mathcal{L}}$+RNN (Gu et al. 2017)	72.3	55.3	41.3	30.6	26.0	-	94.0
PG-SPIDEr-TAG (Liu et al. 2017)	75.4	59.1	44.5	33.2	25.7	55.0	101.3
Adaptive (Lu et al. 2017)	74.2	58.0	43.9	33.2	26.6	54.9	108.5
SCST:Att2all (Rennie et al. 2017)	77.4	60.9	46.0	34.1	26.7	55.7	114.0
TopDown (Anderson et al. 2018)	79.8	63.4	48.4	36.3	27.7	56.9	120.1
StackCap (Gu et al. 2018)	78.4	62.5	47.9	36.1	27.4	56.9	120.4
TextAtt+ResNet (Mun, Cho, and Han 2017)	74.9	58.1	43.7	32.6	25.7	-	102.4
CNN+Att (Aneja and Deshpande 2018)	71.1	53.8	39.4	28.7	24.4	52.2	91.2
GroupCap (Chen et al. 2018)	74.4	58.1	44.3	33.8	26.2	-	-
NBT (Lu et al. 2018)	75.5	-	-	34.7	27.1	-	107.2
DHEDN (Xiao et al. 2019)	80.8	63.7	48.8	36.7	27.2	57.2	117.0
IDGAN (ours)	**81.3**	**65.4**	**50.7**	**38.5**	**28.5**	**58.8**	**123.5**
w/o D_{ϕ_1}	80.7	64.5	49.4	37.5	27.9	58.1	122.1
w/o G_{θ_2}	80.2	64.2	49.0	36.8	27.5	57.5	121.7
w/o D_{ϕ_2}	79.8	63.6	48.7	36.9	27.7	57.6	121.3
w/o copy	80.9	64.9	50.3	38.2	28.1	58.5	122.7

Table 1: The automatic evaluation results of IDGAN and the compared methods on MSCOCO Karpathy test split.

Methods	Informativeness	Fluency
Adaptive	2.86	2.97
SCST	2.94	2.92
TopDown	3.25	3.18
StackCap	3.21	3.15
TextAtt	2.97	3.14
CNN+Att	2.75	2.83
GroupCap	3.09	3.12
NBT	3.14	3.11
IDGAN (Ours)	3.32	3.35

Table 2: Human evaluation results of the captions generated by our model and several strong baselines.

However, the logarithm may lead to instability of training (Goodfellow et al. 2014). We thus follow (Wang et al. 2017) with the reward advantage function:

$$R_2(\langle y, \tilde{y} \rangle) = 2 * D_{\phi_2}(\langle y, \tilde{y} \rangle | x) - 1 \tag{30}$$

5 Experimental Setup

Dataset We adopt the widely used MSCOCO 2014 (denoted as MSCOCO) image captions dataset (Karpathy and Fei-Fei 2015) as the experimental data. In total, MSCOCO is composed of 82,783 training images, 40,504 validation images, and 40,775 testing images. Each image is corresponding to five reference descriptions. For the off-line testing, we use the Karpathy split setting (Karpathy and Fei-Fei 2015), which has been widely adopted in previous studies. There are 113,287 images for training, 5,000 images for validation, and 5,000 images for testing.

Baseline Methods In this study, we compare IDGAN with several state-of-the-art models, and some representative compared methods are Adaptive model (Lu et al. 2017), SCST (Rennie et al. 2017), Bottom-Up and Top-Down Attention (TopDown) model (Anderson et al. 2018), StackCap model (Gu et al. 2018), Text-Guided Attention (TextAtt) model (Mun, Cho, and Han 2017), Convolutional Image Captioning (CNN+Att) model (Aneja and Deshpande 2018), Group-based Image Captioning (GroupCap) model (Chen et al. 2018), Neural Baby Talk (NBT) model (Lu et al. 2018). In the experiments, the results of baseline methods in Tables 1 are retrieved from previous papers.

Implementation Details Following previous work (Anderson et al. 2018), we use the faster R-CNN to detect objects and extract 100 image region features. In this manner, the decoder is able to attend to specific parts of an image by selecting a subset of the feature vectors. The number of hidden units in LSTM caption encoder is set to 512. The parameters of the LSTM networks are initialized with normal distribution $\mathcal{N}(0, 0.01)$, and the other parameters are initialized by using the uniform distribution [-0.01, 0.01]. We set the number of hidden units in TopDown attention LSTM (LSTM$^{(1)}$) and language model LSTM (LSTM$^{(2)}$) to 1,024. The numbers of hidden units of LSTMs used in Eq. (14) and Eq. (17) are set to 512. During adversarial training, the two generators (G_{θ_1} and G_{θ_2}) produce $M_1 = M_2 = 5$ candidate captions. The value of $\gamma_1, \gamma_2, \gamma_3$ equal to 0.2, 1, 0.8 respectively. We pre-train G_{θ_1} for 30 epochs with maximum likelihood and D_{ϕ_2} for 5 epochs with triplet loss. After that, we optimize the whole model with interactive adversarial training for 30 epochs.

Automatic Evaluation Metrics We adopt the official evaluation metrics of MSCOCO Image Captioning Challenge that are widely used in previous work (Karpathy and Fei-Fei 2015; Vinyals et al. 2015), including BLEU-N (N=1,2,3,4) (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE (Lin 2004), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015). These metrics estimate

	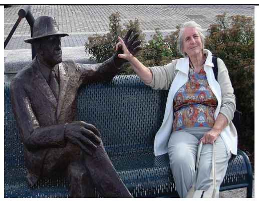	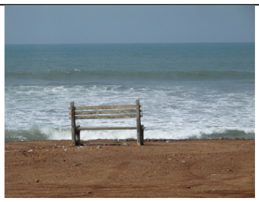	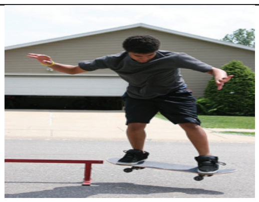	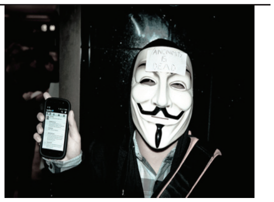
Ground truth	"there is a woman sitting next to a statue on the bench"	"an old bench is right on the oceans edge"	"the boy is doing a trick on his skate board"	"a man with a mask on holding up a phone"
TopDown	"a man and a woman sitting on a bench"	"a bench sitting on top of a beach"	"a man riding a skateboard on a street"	"a man wearing a hat holding a cell phone"
IDGAN	"a woman sitting on a bench with a statue"	"a wooden bench sitting on the beach next to the ocean"	"a man is doing a trick on a skateboard"	"a person in a mask holding a cell phone"

Table 3: Example captions generated by different models.

the consistency between the n-gram existence in the produced text descriptions and the ground truth captions.

6 Experimental Results

6.1 Quantitative Evaluation

We firstly report the model comparison from the quantitative perspective. The experimental results on MSCOCO are summarized in Table 1. The results are calculated using the COCO captioning evaluation tool (Lin et al. 2014). Our model achieves statistically significantly better performance than the state-of-the-art competitors on MSCOCO. Specifically, the proposed IDGAN successfully obtains higher scores over all automatic evaluation measures compared to the TopDown model which adopts the same CNN-LSTM framework as IDGAN.

To investigate the impact of different components of the proposed IDGAN for image captioning, we also conduct the ablation study of IDGAN on MSCOCO by removing the retrieval-based generator G_{θ_2} (w/o G_{θ_2}), the retrieval-based discriminator D_{ϕ_2} (w/o D_{ϕ_2}), the generation-based discriminator D_{ϕ_1} (w/o D_{ϕ_1}), and copy mechanism (w/o copy), respectively. It is no surprise that combining all the factors achieves the best performance for all evaluation metrics. The retrieval-based generator and discriminator (G_{θ_2} and D_{ϕ_2}) contribute a great improvement to our model by providing ranking scores to guide the generation-based generator producing better captions.

6.2 Human Evaluation

Similar to previous works (Xu et al. 2015; Rennie et al. 2017), we also use human annotation to evaluate the image captioning systems quantitatively. We randomly sample 100 images from the MSCOCO test set and invite three human annotators to score the generated captions based on their *Informativeness* (whether the caption is appropriate and natural to an image) and *Fluency* (whether the generated caption is fluent with a proper grammatical structure). The annotators are asked to assign each caption a score of 1 (bad), 2 (poor), 3 (not bad), 4 (satisfactory), 5 (good) for *Informativeness* and *Fluency*, respectively. The human evaluation results are summarized in Table 2. According to Table 2, *IDGAN* substantially outperforms the compared approaches by a noticeable margin on the MSCOCO dataset.

6.3 Case Study

To measure the performance of IDGAN from the qualitative perspective, we report several produced image captions in Table 3. We can easily observe from Table 3 that *IDGAN* is able to generate reasonable and relevant text descriptions of the given images. For example, the sentence "a woman sitting on a bench with a statue" generated by *IDGAN* precisely describes the content of the image. In contrast, the Top-Down method often fails in such cases.

6.4 Error Analysis

To examine the limitations of the proposed model, we additionally carry out an analysis of the errors made by *IDGAN*. Specifically, we randomly choose 100 images from MSCOCO test set whose captions generated by our model have low human evaluation scores. We reveal several reasons for the low evaluation scores, which can be divided into two primary categories. **First**, *IDGAN* fails to generate semantically diverse image captions across visually similar images. For example, *IDGAN* tends to generate the same caption for two different images that are semantically related or have similar objects, ignoring some details of the two images. One possible solution is to devise a new metric to measure the semantic diversity of image captions, and then the diversity score can be used as a reward in reinforcement learning so as to encourage the model to consider both diversity and accuracy. **Second**, *IDGAN* fails to detect some objects in the images that have no high-quality retrieved captions. It suggests that certain object detection strategy needs to be devised in the future so as to generate better captions for specific images.

7 Conclusion

In this paper, we proposed *IDGAN* to enhance a generation-retrieval ensemble model with dual adversarial learning, allowing for both generation-based and retrieval-based image captioning methods to be mutually enhanced. We integrated retrieved guidance captions into word decoding process by a copy mechanism, which enriched the meaning of the generated captions. Extensive experiments revealed that the proposed *IDGAN* model significantly outperformed the compared methods by a remarkable margin.

Acknowledgments

This work was partially supported by the National Natural Science Foundation of China (NSFC) (No. 61906185), Guangdong Basic and Applied Basic Research Foundation (No. 2019A1515011705 and No. 2018A030313943).

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and vqa. In *CVPR*.

Aneja, J., and Deshpande, A. 2018. Convolutional image captioning. In *CVPR*, 5561–5570.

Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL*, volume 29, 65–72.

Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; and Chua, T.-S. 2017. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*.

Chen, F.; Ji, R.; Sun, X.; Wu, Y.; and Su, J. 2018. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *CVPR*, 1345–1353.

Gong, Y.; Wang, L.; Hodosh, M.; Hockenmaier, J.; and Lazebnik, S. 2014. Improving image-sentence embeddings using large weakly annotated photo collections. In *ECCV*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *NIPS*, 2672–2680.

Gu, J.; Wang, G.; Cai, J.; and Chen, T. 2017. An empirical study of language cnn for image captioning. In *ICCV*.

Gu, J.; Cai, J.; Wang, G.; and Chen, T. 2018. Stack-captioning: Coarse-to-fine learning for image captioning. In *AAAI*.

Hodosh, M.; Young, P.; and Hockenmaier, J. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *JAIR* 47:853–899.

Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.-J.; Shamma, D. A.; et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123(1):32–73.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollar, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop on Text Summarization Branches Out*, volume 8.

Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Optimization of image description metrics using policy gradient methods. In *ICCV*.

Lu, J.; Xiong, C.; Parikh, D.; and Socher, R. 2017. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2018. Neural baby talk. In *CVPR*, 7219–7228.

Mun, J.; Cho, M.; and Han, B. 2017. Text-guided attention model for image captioning. In *AAAI*, 4233–4239.

Ordonez, V.; Kulkarni, G.; and Berg, T. L. 2011. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 1143–1151.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Pu, Y.; Gan, Z.; Henao, R.; Yuan, X.; Li, C.; Stevens, A.; and Carin, L. 2016. Variational autoencoder for deep learning of images, labels and captions. In *NIPS*, 2352–2360.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.

Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*, 1179–1195.

Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*.

Wang, J.; Yu, L.; Zhang, W.; Gong, Y.; Xu, Y.; Wang, B.; Zhang, P.; and Zhang, D. 2017. Irgan: A minimax game for unifying generative and discriminative information retrieval models. In *SIGIR*, 515–524. ACM.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning* 8:229–256.

Wu, Q.; Shen, C.; Liu, L.; Dick, A.; and van den Hengel, A. 2016. What value do explicit high level concepts have in vision to language problems? In *CVPR*.

Xiao, X.; Wang, L.; Ding, K.; Xiang, S.; and Pan, C. 2019. Deep hierarchical encoder-decoder network for image captioning. *IEEE Transactions on Multimedia*.

Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhudinov, R.; Zemel, R.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, 2048–2057.

Xu, C.; Zhao, W.; Yang, M.; Ao, X.; Cheng, W.; and Tian, J. 2019. A unified generation-retrieval framework for image captioning. In *CIKM*, 2313–2316. ACM.

Yang, M.; Zhao, W.; Xu, W.; Feng, Y.; Zhao, Z.; Chen, X.; and Lei, K. 2018. Multitask learning for cross-domain image captioning. *IEEE Transactions on Multimedia* 21(4):1047–1061.

Zhao, W.; Wang, B.; Ye, J.; Yang, M.; Zhao, Z.; Luo, R.; and Qiao, Y. 2018. A multi-task learning approach for image captioning. In *IJCAI*, 1205–1211.