

An Ensemble of Generation- and Retrieval-Based Image Captioning With Dual Generator Generative Adversarial Network

Min Yang^{ID}, Junhao Liu, Ying Shen, Zhou Zhao^{ID}, *Member, IEEE*, Xiaojun Chen^{ID}, *Member, IEEE*, Qingyao Wu^{ID}, *Member, IEEE*, and Chengming Li^{ID}, *Member, IEEE*

Abstract—Image captioning, which aims to generate a sentence to describe the key content of a query image, is an important but challenging task. Existing image captioning approaches can be categorised into two types: generation-based methods and retrieval-based methods. Retrieval-based methods describe images by retrieving pre-existing captions from a repository. Generation-based methods synthesize a new sentence that verbalizes the query image. Both ways have certain advantages but suffer from their own disadvantages. In the paper, we propose a novel *EnsCaption* model, which aims at enhancing an ensemble of retrieval-based and generation-based image captioning methods through a novel dual generator generative adversarial network. Specifically, *EnsCaption* is composed of a caption generation model that synthesizes tailored captions for the query image, a caption re-ranking model that retrieves the best-matching caption from a candidate caption pool consisting of generated captions and pre-retrieved captions, and a discriminator that learns the multi-level difference between the generated/retrieved captions and the ground-truth captions. During the adversarial training process, the caption generation model and the caption re-ranking model provide improved synthetic and retrieved candidate captions with high ranking scores from the discriminator, while the discriminator based on multi-level ranking is trained to assign low ranking scores to the generated and retrieved image captions. Our model absorbs the merits of both generation-based

and retrieval-based approaches. We conduct comprehensive experiments to evaluate the performance of *EnsCaption* on two benchmark datasets: MSCOCO and Flickr-30K. Experimental results show that *EnsCaption* achieves impressive performance compared to the strong baseline methods.

Index Terms—Image captioning, ensemble generation-retrieval model, adversarial learning.

I. INTRODUCTION

IMAGE captioning has received a significant amount of attention recently and is applicable in numerous scenarios such as image indexing, virtual assistants, and support of the disabled. Different from other tasks in computer vision such as image classification [1], [2] and object detection [3], [4], image captioning requires the machines not only recognize the salient objects in the image, but also have the linguistic capability of verbalizing the most salient aspects of the image. Thus, image captioning connects researches from the communities of both computer vision (CV) and natural language processing (NLP).

Early image captioning systems are usually implemented using retrieval-based methods [5]. When a user issues a query image, the retrieval-based methods retrieve a candidate caption that best describes the query image from a pre-constructed repository. To be more specific, for the query image, the retrieval-based models first extract k nearest images associated with their captions from the pre-constructed image-caption repository, and then use a ranking model to choose the best caption from the extracted candidates as the final caption. The retrieval-based models can produce informative and grammatically correct captions. Nevertheless, these models often struggle to generate innovative and diverse captions that veritably represent the new images. The retrieved captions are limited by the capacity of the existing image-caption repository. Even the best-matched caption from the repository is not guaranteed to be a good caption since the retrieved caption is not tailored for the query image.

To make a caption tailored for the query image, a better way is to generate a new one accordingly. With the availability of large-scale image-caption pairs, generation-based models could achieve impressive results in image captioning. A typical generation-based image captioning method is the sequence-to-sequence (seq2seq) model [6], [7], in which a convolutional neural network (CNN) and a long short-term

Manuscript received February 1, 2020; revised June 18, 2020; accepted September 21, 2020. Date of publication October 15, 2020; date of current version October 22, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61906185 and Grant 61876208, in part by the Natural Science Foundation of Guangdong Province of China under Grant 2019A1515011705, in part by the Key-Area Research and Development Program of Guangdong Province under Grant 2018B010108002, in part by the Shenzhen Basic Research Foundation under Grant JCYJ20180302145607677 and Grant JCYJ20190808182805919, and in part by the Youth Innovation Promotion Association of CAS. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Yun Fu. (*Corresponding authors: Qingyao Wu; Chengming Li.*)

Min Yang, Junhao Liu, and Chengming Li are with the Shenzhen Key Laboratory for High Performance Data Mining, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China (e-mail: min.yang@siat.ac.cn; cm.li@siat.ac.cn).

Ying Shen is with the School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen 510275, China (e-mail: shenyang76@mail.sysu.edu.cn).

Zhou Zhao is with the School of Computing Science, Zhejiang University, Hangzhou 310027, China (e-mail: zhaozhou@zju.edu.cn).

Xiaojun Chen is with the College of Computer Science and Software, Shenzhen University, Shenzhen 518060, China (e-mail: xjchen@szu.edu.cn).

Qingyao Wu is with the School of Software Engineering, South China University of Technology, Guangzhou 510640, China, and also with the Key Laboratory of Big Data and Intelligent Robot, Ministry of Education, Guangzhou 510006, China (e-mail: qyw@scut.edu.cn).

Digital Object Identifier 10.1109/TIP.2020.3028651

memory network (LSTM) are used as the encoder and the decoder, respectively. The encoder is to capture the semantics and content of the query image with one or a few distributed and real-valued vectors; the decoder aims at decoding the distributed vectors of the query image to a textual image description (caption). The seq2seq based approaches have become the mainstream in image captioning mainly because they could be trained in an end-to-end manner and scale to the large-scale training data. In addition, these methods can synthesize a new sentence as the caption and bring the results of good flexibility and quality. However, a well-known problem for the generation-based models is that they are prone to produce general, uninformative, and grammatically incorrect captions.

In previous studies, the generation- and retrieval-based approaches with their characteristics have been explored separately. We are seeking to take full advantage of these two kinds of approaches. In this work, we propose an ensemble of retrieval- and generation-based image captioning methods (called *EnsCaption*), which is built on a dual generator generative adversarial network for enhancing both methods mutually. Different from the generative adversarial network (GAN), *EnsCaption* consists of two generators that exploit the complementary properties from the retrieval- and generation-based methods to effectively diversify and improve the produced captions. Specifically, the caption generation model (denoted as G_{θ_1}) synthesizes tailored captions for the query image. We integrate pre-retrieved guidance captions into the word decoding process by a copying mechanism so as to enrich the meaning of the generated captions. The caption re-ranking model (denoted as G_{θ_2}) is built on the top of neural ranking models to select the best caption candidate among the pre-retrieved and generated caption candidates, which follows a two-step procedure. First, we employ a caption encoder (LSTM network) to learn sentence representations for each caption candidate. Second, each caption candidate is re-ranked based on the learned sentence representations via a scoring function (e.g., a feed-forward network). The caption re-ranking model is finally optimized via a triplet ranking-based loss. The discriminator with multi-level ranking (denoted as D_ϕ) attempts to discriminate between the ground-truth captions and the produced captions by the two generators (G_{θ_1} and G_{θ_2}), which is trained synchronously with the two generators using the adversarial training framework. The motivation behind is that through the dual generator generative adversarial network, with G_{θ_1} generating improved candidate captions and G_{θ_2} re-ranking the pre-retrieved and generated captions, the discriminator D_ϕ could be trained to identify captions that feature both generation-based and retrieval-based approaches (e.g., relevant, informative and fluent), which leads an ensemble towards more plausible image captions.

We summarize our main contributions as follows:

- We propose a dual generator generative adversarial network for image captioning, which takes full advantage of both generation- and retrieval-based approaches. Integrating these two kinds of approaches could mutually enhance their performance and lead to a better ensemble image captioning model.
- We propose a multi-level ranking discriminator to learn the multi-level difference between the generated/retrieved captions and the ground-truth captions. During the adversarial training, the discriminator could provide rewards to guide the generation- and retrieval-based models to produce more informative and fluent caption candidates.
- We conduct comprehensive experiments on two benchmark datasets (MSCOCO and Flickr-30K). Experimental results demonstrate that *EnsCaption* outperforms the compared models by a substantial margin, across multiple evaluation metrics.

II. RELATED WORK

A. Retrieval-Based Image Captioning

Image captioning is a multi-modal task that requires the machines to not only understand an image [8]–[10] but also have the capability of describing the image with fluent natural language [11].

Traditional image captioning methods rely on either template-based [12] or retrieval-based techniques [5]. The template-based approaches usually define some templates with plenty of blank slots to produce image captions. For instance, [13] extracted a number of items to fill in the blanks to generate high-quality captions. Template-based methods are able to produce grammatically correct image captions. Nevertheless, the pre-designed templates cannot generate captions with variable lengths. On the other hand, the retrieval-based approaches first extracted a candidate caption set from a pre-constructed image-caption repository with a basic retrieval (pre-retrieval) model. The final best-matching captions for the input image are then chosen from the captions pool by the re-ranking method [5], [14], [15]. For example, [15] treated the image captioning as a ranking or retrieval task, and introduced a ranking-based method to extract image description. Reference [5] associated the query image with a textual description by projecting them into a shared latent space. Although retrieval-based methods can produce syntactically correct captions, the retrieved captions are not tailored for the query images and limited by the size of the pre-constructed image-caption repository.

B. Generation-Based Image Captioning

Motivated by the remarkable success of deep neural networks in CV and NLP, the seq2seq paradigm has become the mainstream in image captioning [6], [7], [16], [17], obtaining remarkable improvements over the conventional template-based or retrieval-based methods. The seq2seq models usually use a CNN encoder to learn a distributed image representation, and then employ an LSTM decoder to generate a text description based on the learned image representation. For example, [6] learned the inter-model correspondence between the caption and the image by using the training image-caption pairs.

Attention mechanisms play an essential role in enhancing the performance of the seq2seq models. For example,

an attentive seq2seq model was introduced in [18], which learned to dynamically attend to different locations of the query image at different decoding step. [19] used associated captions that were retrieved from training data to learn visual attention for image captioning. Reference [20] presented a multi-stage decoding method, which consisted of multiple coarse-to-fine decoders for generating high-quality image captions. Reference [21] jointly learned the structure relevance and diversity among groups of images. Reference [22] learned the image representations with multi-layer feature maps, which captured spatial locations and channels by using visual attention. Reference [23] learned an extra guiding vector for the sentence generation (decoder) automatically. The guiding vector modeled the attribute properties of input images. Reference [24] proposed a multi-task framework that enhanced the CNN encoder and LSTM decoder using two auxiliary tasks (object classification and syntax annotation).

There have been increasing interests in combining the seq2seq framework and the reinforcement learning techniques for image captioning, taking advantage of both models [25], [26]. For example, [25] used the policy gradient to optimize the proposed model by maximizing the mixture reward of CIDEr and SPICE. Reference [26] proposed the SCST model, which used the REINFORCE algorithm to train the proposed model. Instead of computing a “baseline” to weaken the variance of the model, SCST utilized its own output to normalize the expected rewards. Reference [27] exploited the copying mechanism to copy words from the retrieved captions and applied adversarial learning to guide the model to generate plausible image captions using a binary classifier as the discriminative model.

Different from the previous studies that only focused on the improvement of either retrieval- or generation-based single model, we fully combine the retrieval- and generation-based image captioning methods by using a dual generator generative adversarial network with two generators and one discriminator. This paper is a significant extension of our previous conference paper [28]. The main differences between this paper and [28] can be summarized as four aspects. In EnsCaption, we propose a multi-level ranking discriminator to learn the multi-level difference between the generated/retrieved captions and the ground-truth captions. During the adversarial training, the multi-level ranking discriminator could provide rewards to guide the generation- and retrieval-based models to produce more informative and fluent caption candidates. Second, EnsCaption designs a retrieval-enhance image encoder with attention mechanism, which learns the distilled essential knowledge from the guidance caption to help locate the salient information of the query image. Third, EnsCaption devises intra-decoder attention to incorporate the information about the previously decoded word sequence into decoding process of the generation-based model. In this way, EnsCaption can alleviate the problem of generating repeated words or phrases. Fourth, comprehensive experiments have been conducted on two benchmark datasets to investigate the effectiveness or limitations of EnsCaption.

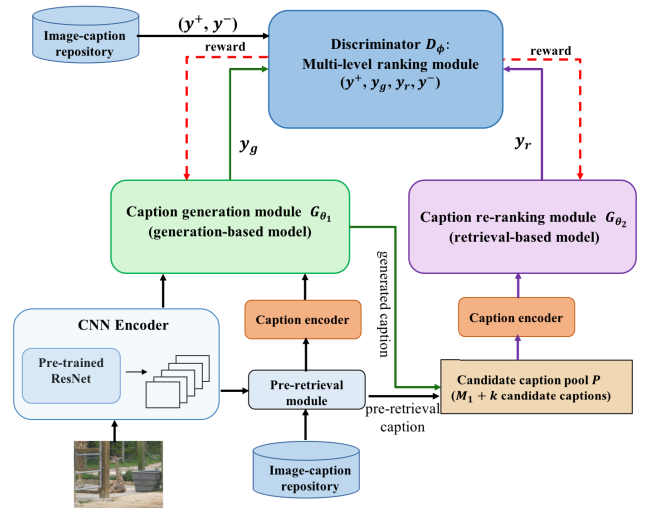


Fig. 1. Illustration of *EnsCaption* architecture (best viewed in color): the generation-based model G_{θ_1} , the retrieval-based model G_{θ_2} , and the discriminator D_ϕ . The caption generation model (G_{θ_1}) synthesizes tailored caption y_g for the query image; the caption re-ranking model (G_{θ_2}) re-ranks the captions in the candidate caption pool (k pre-retrieved captions and M_1 generated captions), and retrieves the best-matching caption y_r for the input image; the discriminator (D_ϕ) adopts a multi-level ranking strategy that learns the multi-level difference between the generated/retrieved captions and the ground-truth captions. D_ϕ calculates the rewards and is trained synchronously with G_{θ_1} and G_{θ_2} via adversarial training.

III. PROBLEM DEFINITION AND MODEL OVERVIEW

Given a query image x , the goal of image captioning is to produce a textual description $y = \{w_1, w_2, \dots, w_T\}$ for the query image x , where T denotes the length of the generated caption.

As depicted in Figure 1, the proposed EnsCaption model consists of the following components.

- 1) **Caption generation model** that is responsible for synthesizing M_1 image caption candidates $\{y_g\}_{m=1}^{M_1}$ given an image x by the Monte Carlo (MC) roll-out policy. Such a process is also noted as $G_{\theta_1}(y_g|x)$.
- 2) **Caption re-ranking model** that retrieves the best-matching captions $\{y_r\}_{m=1}^{M_2}$ from a candidate caption pool consisting of M_1 generated captions and k pre-retrieved captions. Such a process is denoted as $G_{\theta_2}(y_r|x)$.
- 3) **Discriminator model with the multi-level ranking** that adopts a multi-level ranking objective to distinguish the gold caption y^+ from the generated caption y_g by generator G_{θ_1} and the retrieved caption y_r by G_{θ_2} .

IV. CAPTION GENERATION MODEL G_{θ_1}

The sequence-to-sequence (seq2seq) [6], [7] framework is used as the backbone of our caption generation model. In encoding, we propose an image encoder that leverages guidance captions to augment the semantic information of the image and thus learn a better representation of the image. In decoding, we employ the top-down attention network to generate the caption word by word. In particular, the top-down

attention network consists of two stacked LSTM layers, where the first LSTM layer is a top-down visual attention network, while the second LSTM layer is a language model. In addition, we devise an intra-decoder attention mechanism to incorporate the information about the previously decoded sequence into the decoder so as to prevent the model from generating repeated words or phrases. Figure 2 illustrates the flow chart of the caption generation model G_{θ_1} . Next, we introduce the encoder and the decoder in detail.

A. Retrieval-Enhanced Image Encoder

1) *Initial Image Representation Learning*: Similar to [6], [7], we utilize the ResNet-101 CNN [29], which is pre-trained on ImageNet [30], to encode the query image x into L feature vectors, denoting the image features obtained at different locations in x . Formally, we represent the learned image feature vectors \mathbf{z}_{init} as:

$$\mathbf{z}_{init} = \{\mathbf{z}_{init,1}, \mathbf{z}_{init,2}, \dots, \mathbf{z}_{init,L}\} = \text{ResNet}(x) \quad (1)$$

2) *Guidance Caption Extraction*: A critic problem with most previous caption generation models is that the generated captions lack informativeness and diversity. Inspired by [19], we retrieve visually similar images from the training set for a given image, which may share salient regions with the query image. The extracted guidance caption would augment the information used in the caption generation model by steering the attention model focus on important and salient regions for image captioning. Different from previous attention mechanisms, the attention of EnsCaption is learned not only with visual features but also with textual features obtained from the guidance captions.

Specifically, the guidance captions are defined as the ground-truth captions of k nearest training images with respect to the given image x in terms of visual similarity¹. In particular, the image feature vectors learned with Eq. (1) is used to retrieve k nearest images from the training set by exhaustively computing the cosine similarity between the query image and the training images. The corresponding k captions for the k nearest images are concatenated to form a guidance caption $C = [g_1, \dots, g_{L_g}]$, where g_i represents the i -th word in the guidance caption and L_g is the length of the guidance caption C . We utilize word2vec embeddings [31] to map each word in C into a distributed vector space, resulting in a word vector $\mathbf{e}_t \in \mathbb{R}^{d_e}$ for each word w_t , where d_e denotes the size of the word embedding. The context representation of the guidance caption can be represented as: $E = [\mathbf{e}_1, \dots, \mathbf{e}_{L_g}]$.

After getting the representation of each word in the guidance caption, the LSTM network is then applied to learn the semantic meanings of the guidance caption. Formally, the hidden state \mathbf{g}_i can be computed from the previous hidden state \mathbf{g}_{i-1} and the word embedding \mathbf{e}_i at index i in the guidance caption:

$$\mathbf{g}_i = \text{LSTM}(\mathbf{g}_{i-1}, \mathbf{e}_i) \quad (2)$$

¹The value of k is chosen optimally for the feature set and typically ranges from 10 to 100.

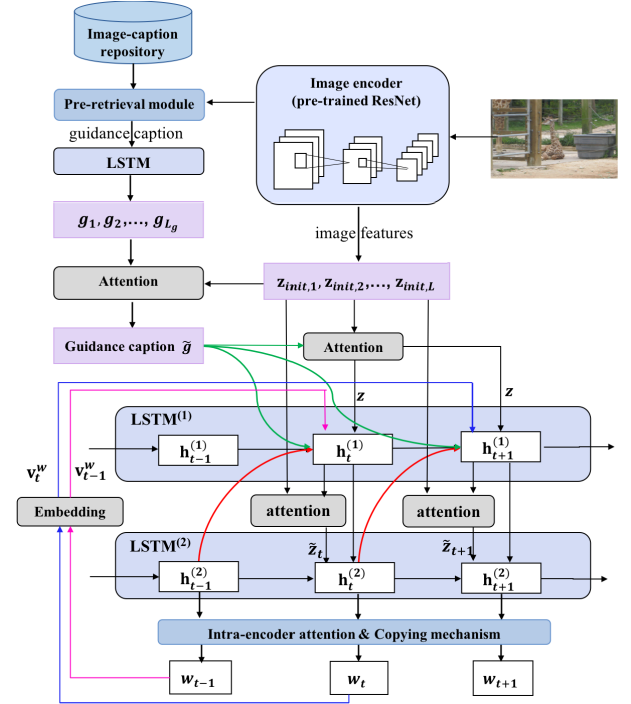


Fig. 2. Flow chart of the caption generation model G_{θ_1} . First, a pre-trained ResNet is applied to learn the initial image features from the raw image. Second, a pre-retrieval module is devised to retrieve guidance captions that augment the caption generation model by steering the attention model focus on important and salient regions for image captioning and help the model to learn a retrieval-enhanced image representation \mathbf{z} . Third, the top-down attention module with two LSTM layers is used as the caption decoder, where the first LSTM layer is a top-down visual attention network and the second LSTM layer is a language model. Fourth, an intra-decoder attention mechanism is proposed to incorporate the information about the previously decoded sequence into the decoder. Fifth, a caption generation model with copy mechanism is employed to generate the caption word by word.

Here, we refer the interested readers to [32] for the implementation details of LSTM in sequence modeling. We denote the hidden states for the guidance caption as: $H^g = [\mathbf{g}_1, \dots, \mathbf{g}_{L_g}]$.

An attention network is designed to learn the distilled representation of the guidance caption representation H^g by using the initial image representation \mathbf{z}_{init} as attention source, which is formally computed as:

$$\tilde{\mathbf{g}} = \sum_{i=1}^m \tau_i \mathbf{g}_i \quad (3)$$

$$\tau_i = \frac{\exp(\rho_1(\mathbf{g}_i, f_{FC}(\mathbf{z}_{init})))}{\sum_{j=1}^m \exp(\rho_1(\mathbf{g}_j, f_{FC}(\mathbf{z}_{init})))} \quad (4)$$

where ρ_1 is a feed-forward neural network. f_{FC} is a fully-connected layer. τ_i represents the attention weight for the i -th word in the guidance caption.

3) *Retrieval-Enhanced Image Representation Learning*: A text-guided attention network is developed to detect crucial information from the input image. In particular, we take as input the contextual representation $\tilde{\mathbf{g}}$ of the guidance caption as attention source to learn guidance-aware image representation

\mathbf{z} as:

$$\mathbf{z} = \sum_{i=1}^L \alpha_i \mathbf{z}_{init,i} \quad (5)$$

$$\alpha_i = \frac{\exp(\rho_2(f_{FC}(\tilde{\mathbf{g}}), \mathbf{z}_{init,i}))}{\sum_{j=1}^L \exp(\rho_2(f_{FC}(\tilde{\mathbf{g}}), \mathbf{z}_{init,j}))} \quad (6)$$

where ρ_2 is a feed-forward neural network. f_{FC} is a fully connected layer. α_i is the attention weight for the i -th image feature $\mathbf{z}_{init,i}$.

B. Retrieval-Enhanced Caption Decoder

The generation of image captions is performed by a top-down attention network based on the learned image representations and the guidance caption representations. An intra-decoder attention mechanism is employed to incorporate the information about the previously decoded sequence into the decoder so as to prevent the model from generating repeated words or phrases.

1) *Top-Down Attention Network*: The top-down attention network is popular in previous image captioning approaches [33], which is composed of two stacked LSTM layers. The first LSTM layer (denoted as LSTM⁽¹⁾) is a top-down visual attention network, while the second LSTM layer (called LSTM⁽²⁾) is a language model.

LSTM⁽¹⁾ captures the important information from the image features with the help of guidance caption. In particular, at t -th decoding step, the input of LSTM⁽¹⁾ is the concatenation of the attended image feature \mathbf{z} , the feature vector of the guidance caption $\tilde{\mathbf{g}}$, the word embedding of the previously generated word (i.e., \mathbf{v}_{t-1}^w), and the previous output of LSTM⁽²⁾ (i.e., $\mathbf{h}_{t-1}^{(2)}$), which is defined as:

$$\mathbf{x}_t^{(1)} = [\mathbf{h}_{t-1}^{(2)}, \mathbf{z}, \tilde{\mathbf{g}}, \mathbf{v}_{t-1}^w] \quad (7)$$

where $\mathbf{x}_t^{(1)}$ denotes the input of LSTM⁽¹⁾ at the t -th decoding step. We compute the hidden state of the LSTM⁽¹⁾ as:

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

At the t -th decoding step, given the output of LSTM⁽¹⁾ (i.e., $\mathbf{h}_t^{(1)}$), we calculate the attentive image representation $\tilde{\mathbf{z}}_t$, which is then fed into LSTM⁽²⁾. The attentive image feature vector $\tilde{\mathbf{z}}_t$ ensures that the decoder can obtain full information of the initial image representation \mathbf{z}_{init} at each decoding step. We calculate $\tilde{\mathbf{z}}_t$ when we decode the t -th word by:

$$\tilde{\mathbf{z}}_t = \sum_{i=1}^L \beta_{t,i} \mathbf{z}_{init,i} \quad (9)$$

$$\beta_{t,i} = \frac{\exp(\rho_3(\mathbf{h}_{t-1}^{(1)}, \mathbf{z}_{init,i}))}{\sum_{j=1}^L \exp(\rho_3(\mathbf{h}_{t-1}^{(1)}, \mathbf{z}_{init,j}))} \quad (10)$$

where ρ_3 represents a feed-forward neural network. The attention weights $\beta_{t,i}$ represents the alignment between the i -th image object and the t -th hidden state in LSTM⁽¹⁾.

We employ LSTM⁽²⁾ to generate the target caption word by word, taking the concatenation of the output of LSTM⁽¹⁾ (i.e., $\mathbf{h}_t^{(1)}$) and the attentive image representation (i.e., $\tilde{\mathbf{z}}_t$) as input:

$$\mathbf{x}_t^{(2)} = [\tilde{\mathbf{z}}_t, \mathbf{h}_t^{(1)}] \quad (11)$$

where $\mathbf{x}_t^{(2)}$ represents the input of LSTM⁽²⁾. Then, we compute the hidden state of LSTM⁽²⁾ at t -th decoding step as:

$$\mathbf{h}_t^{(2)} = \text{LSTM}^{(2)}(\mathbf{x}_t^{(2)}, \mathbf{h}_{t-1}^{(2)}) \quad (12)$$

2) *Intra-Decoder Attention*: Being aware of the output words in previous decoding steps can help the proposed model avoid generating repeated information, even though the information was generated many steps away. To avoid generating repeated words or phrases, we devise an attention strategy to incorporate the previously generated caption into the decoding process, inspired by [34]. For each decoding step t , EnsCaption learns a dynamic decoder context vector \mathbf{c}_t^d . We set \mathbf{c}_1^d to a zero vector because the generated sequence is empty at the first decoding step. For $t > 1$, we compute the corresponding decoder context vector as:

$$\mathbf{c}_t^d = \sum_{j=1}^{t-1} \omega_{tj} \mathbf{h}_j^{(2)} \quad (13)$$

$$\omega_{tj} = \frac{\exp(\mathbf{h}_t^{(2)} W_{\text{attn}} \mathbf{h}_j^{(2)})}{\sum_{j'=1}^{t-1} \exp(\mathbf{h}_t^{(2)} W_{\text{attn}} \mathbf{h}_{j'}^{(2)})} \quad (14)$$

where W_{attn} is parameter to be learned, ω_{tj}^d represents the attention weight of the j -th generated word at time step t .

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

$$P_{\theta_1}^{\text{gen}}(w_t = w^v) = \text{softmax}(W_v[\mathbf{h}_t^{(2)}; \mathbf{c}_t^d]), \quad w^v \in \mathcal{V} \quad (15)$$

where W_v denotes a learnable parameter, and θ_1 denotes the set of parameters in the generation-based model.

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

Given the hidden states $H^g = [\mathbf{g}_1, \dots, \mathbf{g}_m]$ for the image caption g , at time step t , the guidance vector \mathbf{c}_t^g for the guidance caption can be computed as a weighted sum of the hidden states H^g :

$$\mathbf{c}_t^g = \sum_{i=1}^m \gamma_{t,i} \mathbf{g}_i \quad (16)$$

$$\gamma_{t,i} = \frac{\exp(\rho_4(\mathbf{g}_i, \mathbf{h}_t^{(2)}))}{\sum_{j=1}^m \exp(\rho_4(\mathbf{g}_j, \mathbf{h}_t^{(2)}))} \quad (17)$$

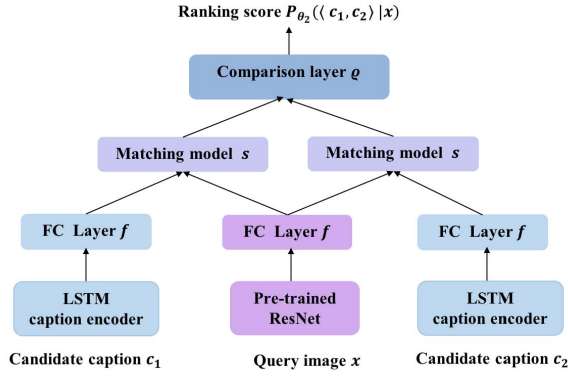


Fig. 3. Flow chart of the caption re-ranking model G_{θ_2} . G_{θ_2} re-ranks the captions in the candidate caption pool P (M_1 generated captions and k pre-retrieved captions), and retrieves the best-matching caption y_r for the input image x . We train the model by using the pair-wise ranking since relative preference is often more easily captured. Here, c_1 and c_2 represent two arbitrary candidate captions from the candidate caption pool P to be ranked.

where $\mathbf{h}_t^{(2)}$ is defined in Eq. (12), ρ_4 is a feed-forward neural network. $\gamma_{t,i}$ represents the alignment between the i -th word in guidance caption and the t -th hidden state in LSTM⁽²⁾.

Formally, the seq2seq model selects a word w^g from \mathcal{V}^g at time step t as follows:

$$P_{\theta_1}^{copy}(w_t = w^g) = \text{softmax}(W_g[\mathbf{h}_t^{(2)}; \mathbf{c}_t^g; \mathbf{c}_t^d]), \quad w^g \in \mathcal{V}^g \quad (18)$$

where W_g indicates the learnable parameter.

b) *Hybrid Method for Caption Generation*: Finally, at each time step, the caption generator selects a generic word from \mathcal{V} or copies a word from the retrieved guidance caption with the following distribution:

$$\lambda_t = \text{sigmoid}(U_o[\mathbf{h}_t^{(2)}; \mathbf{c}_t^g; \mathbf{c}_t^d]) \quad (19)$$

$$\hat{w}_t \sim P_{\theta_1}(w_t) = \begin{bmatrix} (1 - \lambda_t) P_{\theta_1}^{gen}(w_t = w^v) \\ \lambda_t P_{\theta_1}^{copy}(w_t = w^g) \end{bmatrix} \quad (20)$$

where $\lambda_t \in [0, 1]$ controls the choice between generating a content word w^v and coping a word w^g from guidance captions. $P_{\theta_1}^{gen}$ and $P_{\theta_1}^{copy}$ are the distributions over content and caption words, respectively. P_{θ_1} denotes the final word probability distribution. U_o is a learnable parameter.

V. CAPTION RE-RANKING MODEL G_{θ_2}

In the retrieval-based approach, we first extract k visually similar images with their captions from the training dataset by using the guidance caption extraction module defined in Section IV-A.2. The k pre-retrieved candidate captions (by pre-retrieval model) along with M_1 generated captions (by G_{θ_1}) are used as the candidate caption pool P . The caption re-ranking model G_{θ_2} re-ranks the captions in the candidate caption pool P , and retrieves the best-matching caption y_r for the input image x . Figure 3 illustrates the flow chart of the caption re-ranking model G_{θ_2} .

A. Caption Encoder

Given a query image x and a candidate caption pool with $M_1 + k$ captions $\{c_1, c_2, \dots, c_{M_1+k}\}$, we compute a relevance score between each (x, c_i) pair. Concretely, each candidate caption c is first encoded into a distributed representation \mathbf{o} by using an LSTM network:

$$\mathbf{h}_i^c = \text{LSTM}(\mathbf{h}_i^c, e(w_i^c)), \quad \mathbf{o} = \mathbf{h}_{L_c}^c \quad (21)$$

where \mathbf{h}_i^c denotes the i -th hidden state. $e(w_i^c)$ denotes the word embedding of the i -th word in the candidate caption c . L_c is the length of the caption c .

B. Caption Re-Ranking

Instead of learning an absolute relevance (i.e., point-wise ranking), we train caption reranking model by using the pair-wise ranking since relative preference is usually more easily learned. The probability of a caption pair $\langle c_1, c_2 \rangle$ with c_1 more relevant than c_2 being correctly ranked can be evaluated by the normalized distance of their matching relevance to query image x :

$$P_{\theta_2}(\langle c_1, c_2 \rangle | x) = \varrho_1(s(f(\mathbf{z}_{init}), f(\mathbf{o}_{c_1})) - s(f(\mathbf{z}_{init}), f(\mathbf{o}_{c_2}))) \quad (22)$$

where ϱ_1 is an activation function in G_{θ_1} , f is a fully-connected layer, $s(\cdot)$ is any scoring function (matching model), θ_2 denotes the parameters of the retrieval-based model. \mathbf{z}_{init} , \mathbf{o}_{c_1} and \mathbf{o}_{c_2} indicate the distributed representations of the image x , caption c_1 and caption c_2 .

C. Objective Function

The ranking score of the retrieval caption y_r (by G_{θ_2}) is restricted to be lower than the ground-truth caption y^+ but higher than a caption \tilde{y} from the candidate caption pool P . A triplet ranking-based loss can be formulated as:

$$L_{\text{rank}} = \max(0, \varepsilon - s(f(\mathbf{z}_{init}), f(\mathbf{o}_{y^+})) + s(f(\mathbf{z}_{init}), f(\mathbf{o}_{y_r}))) + \max(0, \varepsilon - s(f(\mathbf{z}_{init}), f(\mathbf{o}_{y_r})) + s(f(\mathbf{z}_{init}), f(\mathbf{o}_{\tilde{y}}))) \quad (23)$$

where ε denotes the desired margin between the similarities.

VI. DISCRIMINATOR D_ϕ : MULTI-LEVEL RANKING MODULE

The previous work [27] usually employs a binary classifier to implement the discriminator in GAN, which is insufficient to recognize the progressive relationship between the diverse candidate captions. In our EnsCaption model, we propose a multi-level ranking module which learns the multi-level difference between the generated/retrieved captions and the ground-truth captions. Concretely, we learn two sets of ranking relations, which are introduced as follows:

- The generated captions by G_{θ_1} should be scored above the randomly chosen captions, but below the ground-truth captions, called the *generation path*.

- The retrieved captions by G_{θ_2} should be scored above the captions from the candidate caption pool P , but below the ground-truth captions, called the *re-ranking path*.

Formally, we assume that y^+ denotes the ground-truth caption, while y_g , y_r , \tilde{y} , and y^- refer to the generated caption, the retrieved caption, the caption from P , and the random caption. The *generation path* guides the ranking model to capture the fine-grained relationships between the three kinds of captions. Specifically, the ranking scores of generated captions are restricted to be lower than the gold captions but higher than the random captions. Given the query image x , the ranking function D_ϕ^1 for *generation path* can be defined as:

$$D_\phi^1 = \varrho_2[s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y^+})) - s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y_g}))] + \varrho_2[s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y_r})) - s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y^-}))] \quad (24)$$

where ϱ_2 is an activation function in D_ϕ . s and f are defined in Eq. (22). The ranking model is encouraged to enlarge the matching scores between strong caption distractors, which would meet the requirement in the real-life testing scenario.

The *re-ranking path* plays a similar role to the *generation path*, where the ranking score of the re-ranked caption y_r is restricted to be lower than the ground-truth caption y^+ but higher than the caption \tilde{y} from P . The ranking function D_ϕ^2 for the *re-ranking path* can be defined as:

$$D_\phi^2 = \varrho_2[s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y^+})) - s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y_r}))] + \varrho_2[s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{y_r})) - s(f(\mathbf{z}_{\text{init}}), f(\mathbf{o}_{\tilde{y}}))] \quad (25)$$

Subsequently, we integrate the generation path and the re-ranking path by adding their objective functions. The multi-level ranking scores can be computed as:

$$D_\phi = D_\phi^1 + D_\phi^2 \quad (26)$$

Intuitively, a good ranking model could not only distinguish good captions from random ones (usually totally irrelevant) as adopted in the conventional discriminator with binary classification objective [27], but also capture the fine-grained differences of the matching scores among competitive candidate captions.

VII. ADVERSARIAL TRAINING FOR EnsCaption

In *EnsCaption* framework, both G_{θ_1} and G_{θ_2} aim at producing candidate captions that achieve high ranking scores so as to fool the discriminator D_ϕ , while the discriminator D_ϕ attempts to score down the produced captions by G_{θ_1} and G_{θ_2} , inspired by [35]. We summarize this minimax game with the following objective function \mathcal{L} :

$$\mathcal{L} = \min_{\theta_1, \theta_2} \max_{\phi} (\mathcal{L}_1 + \mathcal{L}_2) \quad (27)$$

$$\mathcal{L}_1 = \mathbb{E}_{y_g \sim G_{\theta_1}} [\log(1 - D_\phi^1(\langle y^+, y_g, y^- \rangle | x))] \quad (28)$$

$$\mathcal{L}_2 = \mathbb{E}_{y_g, y_r \sim G_{\theta_1}, G_{\theta_2}} [\log(1 - D_\phi^2(\langle y^+, y_r, \tilde{y} \rangle | x))] \quad (29)$$

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

y_g and y_r are the caption candidates produced by the caption generation model G_{θ_1} and the caption re-ranking model G_{θ_2} , \tilde{y} is a caption from candidate caption pool P .

A. Discriminator Update

The objective of the discriminative model D_ϕ is to maximize the probability of correctly discriminating between the ground truth captions and the generated ones. In this paper, we propose a multi-level ranking objective that learns the multi-level difference between the ground-truth caption and the candidate captions. By fixing the two generators, we optimize the discriminative model D_ϕ by maximizing the objective function \mathcal{L} defined in Eq. (27) with respect to ϕ :

$$\phi^* = \operatorname{argmax}_{\phi} (\mathcal{L}_1 + \mathcal{L}_2) \quad (30)$$

where \mathcal{L}_1 and \mathcal{L}_2 are defined in Eq. (28)-Eq. (29). Such an optimization problem is typically solved by gradient descent since D_ϕ is differentiable with respect to ϕ .

B. Generator Update

Each time when the discriminator D_ϕ has been learned, we are ready to update the caption generation model G_{θ_1} and the caption re-ranking model G_{θ_2} on the basis of the returned rewards (ranking scores) by the discriminator D_ϕ .

1) *Updating Caption Generation Model*: Given the input image x , the decoding process of the target caption $y_g = [\hat{w}_q, \hat{w}_2, \dots, \hat{w}_T]$ can be treated as a sequence of decision making by policy $P_{\theta_1}(w_t | \hat{w}_{1:t-1}, x)$. However, since D_ϕ^1 only provides the reward for a complete caption, the lack of intermediate reward makes the generation shortsighted. We hence use Monte Carlo (MC) roll-out policy [36] to tackle this problem. With the gold caption y^+ for image x , the long-term reward of the generated image caption y_g is computed by the reward function $R(y_g)$, which is defined as:

$$R(y_g) = D_\phi^1(\langle y^+, y_g, y^- \rangle | x) \quad (31)$$

However, the logarithm may lead to instability of training [37]. We thus follow [38] with the reward advantage function:

$$R(y_g) = 2 * D_\phi^1(\langle y^+, y_g, y^- \rangle | x) - 1 \quad (32)$$

Since the optimization process is non-differential, we employ the reinforcement learning, i.e., policy gradient algorithm [39] to update the parameters of G_{θ_1} . The objective function of the caption generation model G_{θ_1} is computed as below:

$$J_{\theta_1}(y_g | x) = \sum_{t=2}^T \sum_{w_t} P_{\theta_1}(w_t | \hat{w}_{1:t-1}, x) R([\hat{w}_{1:t-1}; w_t]) \quad (33)$$

where P_{θ_1} is defined in Eq. (20). Based on the policy gradient theorem [39], we compute the gradient of the objective function Eq.(33) with respect to parameter θ_1 as:

$$\begin{aligned} \nabla_{\theta_1} J_{\theta_1}(y_g | x) \\ \simeq \sum_{t=1}^T \sum_{w_t} R([\hat{w}_{1:t-1}; w_t]) \nabla_{\theta_1} \log P_{\theta_1}(w_t | \hat{w}_{1:t-1}, x) \end{aligned} \quad (34)$$

Hence, the parameters θ_1 of G_{θ_1} can be solved by gradient descent.

Algorithm 1 The EnsCaption With MiniMax Game

-
- 1) **Input:** Generators G_{θ_1} , G_{θ_2} , and discriminator D_ϕ ; The training set $\{x_{1:N}, y_{1:N}^+\}$;
 - 2) Initialize models G_{θ_1} , G_{θ_2} , D_ϕ with random weights, and pre-train them on the training set;
 - 3) **Repeat**
 - 4) **For** G_1 -step **do**:
 - a) Generate M_1 captions for each image x with G_{θ_1} ;
 - b) Update G_{θ_1} via gradient descent based on Eq.(34).
 - 5) **End for**
 - 6) **For** G_2 -step **do**:
 - a) Generate M_1 captions with G_{θ_1} ;
 - b) Extract k captions with pre-retrieval model;
 - c) Retrieve the best caption y_r from candidate pool P (M_1 and k captions) with G_{θ_2} ;
 - d) Update G_{θ_2} via gradient descent based on Eq.(36).
 - 7) **End for**
 - 8) **For** D -step **do**:
 - a) Use G_{θ_1} and G_{θ_2} to generate and re-rank the candidate captions;
 - b) Update discriminator D_ϕ according to Eq.(30).
 - 9) **End for**
-

2) *Updating Caption Re-Ranking Model*: We train $G_{\theta_2|\theta_1}$ to obtain a competitive image caption y_r that has a high ranking score from D_ϕ^2 . Formally, we ameliorate G_{θ_2} with the objective function as below:

$$J_{\theta_2}(y_r|x) = \mathbb{E}_{y_g, y_r \sim G_{\theta_1}, G_{\theta_2|\theta_1}} [\log(1 - D_\phi^2(\langle y^+, y_r, \tilde{y} \rangle | x))] \quad (35)$$

where the discriminator function D_ϕ^2 is defined in Eq. (25). We compute the gradient of objective function Eq. (35) with respect to θ_2 based on policy gradient theorem [39]:

$$\nabla_{\theta_2} J_{\theta_2}(y_r|x) \simeq -\nabla_{\theta_2} \log P_{\theta_2}(y_r|x) \log D_\phi^2(\langle y^+, y_r, \tilde{y} \rangle | x) \quad (36)$$

We summarize the adversarial training process of EnsCaption in Algorithm 1. At the beginning of the training, we pre-train the generation-based model G_{θ_1} , the retrieval-based model G_{θ_2} , and the discriminator D_ϕ on the training set. After the pre-training, the two generators (G_{θ_1} and G_{θ_2}) and the discriminator D_ϕ are trained alternatively.

VIII. EXPERIMENTAL SETUP

A. Dataset

1) *MSCOCO*: We use the widely used MSCOCO 2014 image captions [6] to estimate the performance of the proposed EnsCaption model. MSCOCO contains 82,783 training images, 40,504 validation images, and 40,775 testing images, each of which is associated with five human-written captions. We adopt the commonly used Karpathy split [6], using 113,287 samples for training, 5,000 samples for validation, and 5,000 samples for testing. This setting has been widely adopted in previous studies.

2) *Flickr30k*: Flickr30k dataset contains 31,783 images in total. Each image has five ground truth text descriptions. Following the same data splitting as in [6], we adopt 29,000 samples for training, 1,000 samples for validation, and 1,000 samples for testing.

B. Baseline Methods

In this work, we compare EnsCaption with state-of-the-art image captioning methods, and several recent strong competitors are described below:

1) *Adaptive Model*: This model presents an adaptive attention mechanism with a visual sentinel for image captioning [43].

2) *SCST: Att2in/Att2all Model*: This model uses a self-critical sequence training (SCST) optimization method, which adopts the output of the test time inference to normalize the rewards [26].

3) *ATT-FCN Model*: This model proposes semantic attention to select semantic concept proposals and fuse them into the recurrent neural networks [50].

4) *Up-Down Model*: This model introduces a top-down attention network with two LSTM layers for image captioning, where [33].

5) *StackCap Model*: This model presents a coarse-to-fine prediction method for image captioning, where the first LSTM layer is a top-down visual attention network and the second LSTM layer is a language model [20].

6) *Text-Guided Attention (TextAtt) Model*: This model uses related captions to learn visual attention for image captioning [19]. For each image, several associated captions are retrieved from training data, and they are used to learn attention on visual features.

7) *SCN-LSTM Model*: This model first detects the semantic concepts from the image, and then uses the learned probability of each concept to compose the parameters in the LSTM network [51].

8) *Self-Retrieval-SR-PL Model*: This model uses a self-retrieval module to guide the training of the image captioning model, encouraging the model to generate discriminative captions [52].

9) *Convolutional Image Captioning (CNN+Att) Model*: This model presents a convolutional architecture for the image captioning model without employing the LSTM generator [44]. It uses an attention network to leverage spatial image features.

10) *Group-Based Image Captioning (GroupCap) Model*: This model jointly learns the structure relevance and diversity among groups images [21]. A visual tree parser is constructed to learn the structured semantic correlations within individual images. Then, the authors model the relevance and diversity among the images by leveraging the correlations among their tree structures.

11) *Neural Baby Talk (NBT) Model*: This model reconciles the slot filling model with the image captioning model, which first generates a hybrid template with slot locations, and then fills in the slots with word [45].

12) *Up-Down+HIP Model*: This model applies the Hierarchy Parsing (HIP) into the Up-Down model [33] for boosting image encoder in captioning [46].

13) *Actor-Critic Model*: This model employs the actor-critic algorithm for image captioning, which adopts an n -step reformulated advantage function [47].

14) *Object Relation Transformer (ORT) Model*: This model uses an object relation Transformer to explicitly incorporate the spatial relationship between the detected objects for image captioning [48].

15) *LBPF Model*: This model adopts the look-back approach to encode the visual features from the past, and use the predict forward method to encode the future features [49].

16) *IDGAN Model*: This model develops an interactive dual generative adversarial network for image captioning [28], which trains two generators and two discriminators via adversarial training.

C. Implementation Details

Following previous work [33], we learn the image features with the final convolutional layer of ResNet-101, resulting in a $14 \times 14 \times 2048$ feature map. In this way, the decoder is capable of attending to certain locations of each image by choosing a subset of the annotated feature vectors. We set the number of hidden units in the LSTM caption encoder as 512.

We initialize the parameters of the LSTM network by using Gaussian distribution $\mathcal{N}(0, 0.01)$, while other weight parameters are initialized with Gaussian distribution $\mathcal{N}(0, 0.02)$. The number of hidden units in LSTM⁽¹⁾ is set to be 1,000. The number of hidden units in LSTM⁽²⁾ is set to be 512. We set the dimension of the word embedding to be 512. Dropout (dropout rate = 0.2) and L_2 regularization (weight decay = 0.001) are applied to tackle the overfitting issue. We utilize a beam size of 5 to generate captions. Adam stochastic gradient descent optimization algorithm is employed for training the generator and the discriminator models.

We first pre-train the caption generation model G_{θ_1} , the caption re-ranking model G_{θ_2} , and the discriminator D_ϕ on the training data. During adversarial training, the number of generated captions by G_{θ_1} is $M_1 = 10$ and the number of pre-retrieved captions is $k = 10$. Similar to [53], we apply several additional strategies to stabilize our adversarial training process. First, we apply the batch norm layers in both the discriminator and generator models, except the output of the generators and input to the discriminator. In this way, we can alleviate the sample oscillation and model instability problems. Second, we apply ReLU to implement the activation function ϱ_1 in the generators (G_{θ_1} and G_{θ_2}) and use Leaky ReLU to implement the activation function ϱ_2 in the discriminator D_ϕ . Third, in order to keep a balance of the adversarial training, we set $G_1 - \text{step} = G_2 - \text{step} = 3$ and set $D - \text{step} = 1$.

D. Automatic Evaluation Measures

We utilize the official evaluation measures of MSCOCO Image Captioning Challenge, which are most widely adopted in previous studies [6], [7], [54], including BLEU-N (N=1,2,3,4) [55], METEOR [56], ROUGE [57], CIDEr [58].

These measures estimate the overlap between the n -gram existence in ground truth image captions and generated text descriptions.

IX. EXPERIMENTAL RESULTS

A. Quantitative Evaluation

We first verify the performance of EnsCaption quantitatively. The automatic evaluation results on the MSCOCO dataset are reported in Table I. The results are calculated using the MSCOCO captioning evaluation tool [59]. From the results, we can observe that EnsCaption substantially and consistently surpasses the compared models by a large margin on six out of seven automatic evaluation measures. The improvement from EnsCaption is statistically significant over the compared models (t-test, p-value < 0.05). Specifically, EnsCaption achieves higher scores on all evaluation metrics than the Up-Down model that adopts the same basic CNN-LSTM backbone as ours. The key superiority of EnsCaption comes from its capability of exploiting the benefit of the retrieval-based and the generation-based methods via adversarial learning for image captioning.

We also evaluate EnsCaption on Flickr30k that is a widely used benchmark in image captioning. The experimental results are reported in Table II. We observe that EnsCaption also substantially outperforms the compared methods by a remarkable margin on all the automatic evaluation metrics (improve 3.2% on BLEU-1, 4.2% on BLEU-2, 3.8% on BLEU-3, 7.1% on BLEU-4, 9.8% on METEOR).

B. Ablation Study

To investigate the effectiveness of different parts of *EnsCaption* for image captioning, we also conduct ablation study of *EnsCaption* in terms of removing the discriminative model D_ϕ (w/o D_ϕ), the caption generation model (w/o G_{θ_1}), the caption re-ranking model (w/o G_{θ_2}), the copy mechanism (w/o copy), and the intra-decoder attention (w/o ID-attention), respectively.

The quantitative ablation results are demonstrated in Table III. Generally, all the factors contribute a improvement to *EnsCaption*. From Table III, we can see that the automatic evaluation scores decrease sharply when discarding the discriminative ranking model on all the evaluation metrics. This is because that the adversarial training takes advantage of the generation-based and retrieval-based image captioning approaches, leading to a better image captioning ensemble. The caption generation model (G_{θ_1}) contributes great improvement to our model by providing the generated caption candidates for the caption re-ranking model to produce the best captions. The copying mechanism also has a non-negligible impact on the performance of EnsCaption. This is within our expectation since the copying mechanism can naturally incorporate the retrieved candidate captions into the decoding process, enriching the informativeness and fluency of the generated captions. In addition, the guidance captions highlight relevant regions and suppress unimportant ones, enabling the generation-based model to produce more detailed and accurate captions. It is no surprise that combining all the

TABLE I
AUTOMATIC EVALUATION RESULTS OF ENSCAPTION AND COMPARED MODELS ON MSCOCO KARPATY TEST SPLIT

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDEr |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| Soft-Attention [18] | 70.7 | 49.2 | 34.4 | 24.3 | 23.9 | - | - |
| Hard-Attention [18] | 71.8 | 50.4 | 35.7 | 25.0 | 23.0 | - | - |
| VAE [40] | 72.0 | 52.0 | 37.0 | 28.0 | 24.0 | - | 90.0 |
| Google NICv2 [41] | - | - | - | 32.1 | 25.7 | - | 99.8 |
| Attributes-CNN [42] | 74.0 | 56.0 | 42.0 | 31.0 | 26.0 | - | 94.0 |
| SCA-CNN [22] | 71.9 | 54.8 | 41.1 | 31.1 | 25.0 | - | - |
| CNN _L +RNN [16] | 72.3 | 55.3 | 41.3 | 30.6 | 26.0 | - | 94.0 |
| PG-SPIDER-TAG [25] | 75.4 | 59.1 | 44.5 | 33.2 | 25.7 | 55.0 | 101.3 |
| Adaptive [43] | 74.2 | 58.0 | 43.9 | 33.2 | 26.6 | 54.9 | 108.5 |
| SCST:Att2in [26] | 76.9 | 60.2 | 45.1 | 33.3 | 26.3 | 55.3 | 111.4 |
| SCST:Att2all [26] | 77.4 | 60.9 | 46.0 | 34.1 | 26.7 | 55.7 | 114.0 |
| Up-Down [33] | 79.8 | 63.4 | 48.4 | 36.3 | 27.7 | 56.9 | 120.1 |
| StackCap [20] | 78.4 | 62.5 | 47.9 | 36.1 | 27.4 | 56.9 | 120.4 |
| TextAtt+ResNet [19] | 74.9 | 58.1 | 43.7 | 32.6 | 25.7 | - | 102.4 |
| CNN+Att [44] | 71.1 | 53.8 | 39.4 | 28.7 | 24.4 | 52.2 | 91.2 |
| GroupCap [21] | 74.4 | 58.1 | 44.3 | 33.8 | 26.2 | - | - |
| NBT [45] | 75.5 | - | - | 34.7 | 27.1 | - | 107.2 |
| Up-Down-HIP [46] | - | - | - | 38.2 | 28.4 | 58.3 | 127.2 |
| Actor-Critic [47] | 77.9 | 61.5 | 46.7 | 34.9 | 26.9 | 56.2 | 115.2 |
| ORT [48] | 80.5 | - | - | 38.6 | 28.7 | 58.4 | 128.3 |
| LBPF [49] | 80.5 | - | - | 38.3 | 28.5 | 58.4 | 127.6 |
| IDGAN [28] | 81.3 | 65.4 | 50.7 | 38.5 | 28.5 | 58.8 | 123.5 |
| EnsCaption (ours) | 81.7 | 65.3 | 51.1 | 39.2 | 29.4 | 59.0 | 125.5 |

TABLE II
AUTOMATIC EVALUATION RESULTS OF ENSCAPTION AND COMPARED MODELS ON FLICKR30K DATASET

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDEr |
|----------------------------|-------------|-------------|-------------|-------------|-------------|-------------|-------------|
| Soft-Attention [18] | 66.9 | 43.4 | 28.8 | 19.1 | 18.5 | - | - |
| Hard-Attention [18] | 66.7 | 43.9 | 29.6 | 19.9 | 18.5 | - | - |
| VAE [40] | 72.0 | 53.0 | 38.0 | 25.0 | - | - | - |
| Google NIC [41] | 63.0 | 41.0 | 27.0 | - | - | - | - |
| Attributes-CNN [42] | 73.0 | 55.0 | 40.0 | 28.0 | - | - | - |
| SCA-CNN [22] | 68.2 | 49.6 | 35.9 | 25.8 | 22.4 | 50.9 | 66.5 |
| CNN _L +RNN [16] | 73.8 | 56.3 | 41.9 | 30.7 | 21.6 | - | 61.8 |
| Adaptive [43] | 67.7 | 49.4 | 35.4 | 25.1 | 20.4 | - | 53.1 |
| ATT-FCN [50] | 64.7 | 46.0 | 32.4 | 23.0 | 18.9 | - | - |
| SCN-LSTM [51] | 73.5 | 53.0 | 37.7 | 25.7 | 21.0 | - | - |
| Self-retrieval-SR-PL [52] | 72.9 | 54.5 | 40.1 | 29.3 | 21.8 | 49.9 | 65.0 |
| IDGAN [28] | 75.6 | 58.4 | 42.3 | 32.0 | 23.5 | 52.8 | 67.5 |
| EnsCaption (ours) | 76.2 | 58.7 | 43.5 | 32.9 | 24.6 | 54.2 | 69.3 |

factors can obtain the best performance on all the automatic evaluation measures.

We additionally illustrate several representative image captions generated by *EnsCaption* and the ablation models in order to evaluate the *EnsCaption* model from the qualitative perspective. The generated descriptive captions are reported in Table IV. Obviously, the proposed *EnsCaption* model is able to produce relevant and informative textual descriptions for the given images. For example, the sentence “a man riding a bike on the street next to a train on the tracks” generated by *EnsCaption* precisely describes the content of the image. In contrast, the w/o G_{θ_1} and w/o G_{θ_2} models often

fail in such cases. On the other hand, the copy mechanism makes the LSTM decoder be able to enrich the generated captions with retrieved guidance captions. Thus, our model can generate more informative and fluent captions. In addition, the intra-decoder attention keeps track of the previously decoded sequence and prevents the model from generating repeated words or phrases.

C. Human Evaluation



The automatic metrics such as METEOR and CIDEr scores have wide adoption in evaluating image captioning systems.

TABLE III
ABLATION STUDY ON MSCOCO KARPATY TEST SPLIT. HERE, ID-ATTENTION INDICATES INTRA-DECODER ATTENTION

| Methods | BLEU-1 | BLEU-2 | BLEU-3 | BLEU-4 | METEOR | ROUGE | CIDEr |
|--------------------------|-------------|-------------|-------------|-------------|-------------|-------------|--------------|
| EnsCaption (ours) | 81.7 | 65.3 | 51.1 | 39.2 | 29.4 | 59.0 | 125.5 |
| w/o D_ϕ | 78.9 | 63.4 | 48.7 | 36.8 | 27.6 | 57.1 | 121.5 |
| w/o G_{θ_1} | 79.3 | 63.8 | 49.4 | 37.5 | 28.1 | 57.6 | 122.1 |
| w/o G_{θ_2} | 79.9 | 64.2 | 49.7 | 37.9 | 28.3 | 58.0 | 122.7 |
| w/o copy | 81.0 | 64.8 | 50.5 | 38.6 | 29.0 | 58.5 | 124.3 |
| w/o ID-attention | 80.9 | 65.0 | 50.7 | 38.7 | 28.9 | 58.7 | 124.6 |
| w/o copy+ID-attention | 80.7 | 64.6 | 50.2 | 38.4 | 38.6 | 58.3 | 123.9 |

TABLE IV

EXAMPLE CAPTIONS GENERATED BY ENSCAPTION AND ITS ABLATION MODELS FOR QUERY IMAGES IN THE MSCOCO DATASET. HERE, ID-ATTENTION INDICATES INTRA-DECODER ATTENTION

| | | |
|--------------------|---|---|
| |  |  |
| Ground truth | "a man riding a bike past a train traveling along tracks" | "a kitchen is shown with a variety of items on the counters" |
| w/o D_ϕ | "a person riding a bike on the street" | "a kitchen with a variety of cabinets in the window" |
| w/o G_{θ_1} | "a man riding a bike past a blue train with upside down people on it" | "a kitchen with a metal sink and lots of wooden cabinets" |
| w/o G_{θ_2} | "a man riding on the street next to a train" | "a kitchen with a sink and cabinets on the counters" |
| w/o copy | "a boy riding on the street next to a like building" | "a lovely bathroom with two sinks and a mirror" |
| w/o ID-attention | "a person riding a bike down the street on the tracks" | "a kitchen with a sink and items on in the counters" |
| EnsCaption | "a man riding a bike on the street next to a train on the tracks" | "a kitchen with a sink and a variety of items on the counters" |

These metrics attempt to measure text properties such as saliency, correctness and human consensus. However, how well the existing automatic measures are matched with human judgment is still controversial, because the sentences that contain grammatical/discourse errors can still obtain high scores over the automatic evaluation measures. For example, the sentence "a group of people standing next to a fire truck." is on par with the sentence "a group of people standing next to a." which is incomplete. Therefore, we also adopt human evaluation to verify the image captioning models from quantitative perspective. Specifically, we randomly choose 100 images from the MSCOCO test set and ask 10 workers from Amazon Mechanical Turk² to assign scores to generated captions by considering their *Informativeness* (whether the caption is appropriate and natural to a image) and *Fluency* (whether the generated caption is fluent with correct grammatical rules). 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 agreement ratio computed with Fleiss' kappa [60] is 0.46, showing moderate agreement.

In Table V, we show average scores provided by the annotators. According to Table V, *EnsCaption* achieves significantly better human evaluation scores than the strong baseline methods on the MSCOCO dataset. The advantage of our model comes from its capability of combining both retrieval- and

TABLE V

HUMAN EVALUATION RESULTS OF THE CAPTIONS GENERATED BY OUR MODEL AND SEVERAL STRONG BASELINES

| Methods | Informativeness | Fluency |
|--------------------------|-----------------|-------------|
| Adaptive model [43] | 2.65 | 2.82 |
| Up-Down [33] | 3.06 | 3.08 |
| StackCap [20] | 2.98 | 3.05 |
| TextAtt [19] | 2.84 | 3.02 |
| NBT [45] | 3.04 | 3.01 |
| EnsCaption (ours) | 3.13 | 3.17 |
| Gold captions | 4.26 | 4.52 |

generation-based approaches. We also ask the annotators to score the ground-truth captions based on their *Informativeness* and *Fluency*. From the results, we can observe that the human-written captions cannot get perfect scores of 5, which demonstrates the difficulty of evaluating image captioning (text generation) models.

D. Diversity Evaluation

Similar to the previous work [61], we also adopt three evaluation metrics to measure the caption diversity: (1) *novelty* that indicates the percentage of generated captions not seen in the training set, (2) *diversity* that indicates the percentage of distinct captions out of the total number of the generated captions, (3) *vocabulary size* that indicates the number of unique words used in the produced captions.

The diversity evaluation results on the MSCOCO test set are reported in Table VI. Higher evaluation values indicate more diversity of the captions. We can observe that our *EnsCaption* model is able to generate more novel sentences compared to the baseline models. In addition, Table VI also shows the diversity statistics and vocabulary size on the test set, which verifies the effectiveness of our model in producing more diverse captions.

E. Qualitative Analysis

To evaluate the *EnsCaption* model from the qualitative perspective, we choose two exemplary images from the MSCOCO test set and illustrate some generated captions by *EnsCaption* and two baseline methods (i.e., Up-Down and TextAtt) in Table VII. From Table VII, we observe that Up-Down fails to detect the objects in the query image, thus generates uninformative captions. For example, there is "a little girl"

²<https://www.mturk.com/>

TABLE VI

DIVERSITY STATISTICS OF THE CAPTIONS GENERATED BY OUR MODEL AND SEVERAL STRONG BASELINES

| Methods | Novelty (%) | Diversity (%) | Vocab. size |
|--------------------------|-------------|---------------|-------------|
| Adaptive model [43] | 71.6 | 55.8 | 1324 |
| Up-Down [33] | 75.4 | 66.5 | 1501 |
| StackCap [20] | 72.3 | 57.7 | 1320 |
| TextAtt [19] | 74.6 | 67.5 | 1369 |
| NBT [45] | 72.8 | 61.6 | 1485 |
| EnsCaption (ours) | 77.4 | 70.3 | 1562 |



TABLE VII

EXAMPLE CAPTIONS GENERATED BY ENSCAPTION AND THE BASELINE MODELS (UP-DOWN AND TEXTATT)

| | | |
|--------------|---|---|
| |  |  |
| Ground truth | "a man in a red shirt and a red hat is on a motorcycle on a hill side" | "a little girl is getting ready to blow out a candle on a small dessert" |
| Up-Down | "a man is on a motor bike going down the road" | "a group of kids watching a candle on a cake" |
| TextAtt | "a police officer is on a motorcycle with a helmet with street" | "a girl is sitting at a table next to a candle" |
| EnsCaption | "a man in red shirt riding a motor bike on a dirt road on the countryside" | "a little girl is sitting at a table to bowl out a candle on the dessert" |

TABLE VIII

TWO EXAMPLES FOR THE FIRST ERROR CATEGORY

| | | |
|--------------|--|--|
| |  |  |
| ground truth | "a elephant drinks from a stream with several other elephants walking in background" | "a group of elephants bathing and playing in the water" |
| EnsCaption | "a herd of elephants walking in a watering hole" | "a herd of elephants walking in a watering hole" |



in the first image rather than "a group of kids". On the other hand, TextAtt generates the caption "a police officer is on a motorcycle with a helmet with street", which is grammatically incorrect. Compared to Up-Down and TextAtt models, the proposed EnsCaption tends to produce more informative and specific captions given the query images. For example, the caption "a man in red shirt riding a motor bike on a dirt road on the countryside" produced by EnsCaption is more precise in describing the status of the man in the image. The advantage of EnsCaption comes from its capability of combining the merits of both retrieval-based and generation-based image captioning methods in a unified framework via adversarial training.

F. Error Analysis

To investigate the limitations of our EnsCaption method, we also carry out an analysis of the errors made by the EnsCaption model. In particular, we randomly select 100 images from the MSCOCO test set whose captions generated by our model have low evaluation scores. We reveal several reasons for the low evaluation scores, which can be divided into two primary categories. **First**, EnsCaption fails to identify the difference between visually similar images, thus generates inappropriate captions that are not tailored to the given images. For example, as shown in Table VIII, the

TABLE IX

TWO EXAMPLES FOR THE SECOND ERROR CATEGORY

| | | |
|--------------|--|---|
| |  |  |
| ground truth | "a fork an apple and orange and an onion sitting in a row" | "a yellow toilet with a red helmet on top of it" |
| EnsCaption | "two apples and a knife sitting on a table" | "a yellow toy sitting on top of a banana" |

proposed EnsCaption model generates the same caption for two different images. This may be because we do not explore the object relationships and the attributes of objects contained in the image, and we cannot differentiate visually similar images. One possible solution is to develop a scene graph method to connect the objects (or entities), their attributes, and their relationships in an image by directed edges. In particular, a spatial graph convolutional network can be employed to encode the objects, attributes, and their relationships into vector representations, which can be seamlessly integrated into the encoder-decoder model. **Second**, EnsCaption fails to detect some salient objects in the images. As shown in Table IX, EnsCaption cannot correctly identify the "orange" and "onion" objects in the image, thus generates object-irrelevant captions. This may be because our model does not explore the fine-grained local features of the query image. One possible solution is to propose certain object detection methods in the future in order to generate better captions for specific images. For example, we could propose a multi-level attention mechanism to explore the coarse global image features and fine-grained local image features.

G. Computational Cost

Since it is difficult to investigate the time complexity of deep learning methods theoretically, we instead analyze the computational cost of deep learning methods. We train and test the methods on a single NVIDIA Tesla P100 GPU. Our EnsCaption model takes about 35 minutes per epoch for the MSCOCO dataset. While most compared baseline models take about 20-25 minutes per epoch on average for the MSCOCO dataset. All methods typically converge within 30 epochs using the early stopping criterion. Generating captions at the testing time is reasonably fast with a throughput of about 4 captions per second using a batch size of 1.

X. CONCLUSION

In this study, we propose *EnsCaption* that aims at enhancing a generation-retrieval ensemble model with a novel dual generator generative adversarial network, allowing for both generation-based and retrieval-based image captioning methods to be mutually enhanced. A generation-based model synthesizes tailored captions for the query image. A retrieval-based method is built on the top of neural ranking models to select the best caption candidate among the pre-retrieved and generated caption candidates. A discriminator learns the multi-level difference between the generated/retrieved captions

and the ground-truth captions. Comprehensive experiments are conducted to measure the effectiveness of *EnsCaption* for image captioning on MSCOCO and Flickr-30K datasets. The evaluation results demonstrate that *EnsCaption* significantly outperforms the compared models by a remarkable margin.

Randomly generated negative samples for the ranking model may create “easy” instances that hinder the *EnsCaption* model from learning more complex latent alignments between images and captions. In the future, we will exploit adversarial training to adaptively create “difficult” and informative negative instances to further improve the ranking process in *EnsCaption* for image captioning. Another possible direction is to develop a scene graph method with a graph convolutional network to encode the objects, attributes, and their relationships into vector representations, which can be integrated into the encoder-decoder model to generate more fine-grained image captions.

REFERENCES

- [1] H. Li, G. Li, L. Lin, H. Yu, and Y. Yu, “Context-aware semantic inpainting,” *IEEE Trans. Cybern.*, vol. 49, no. 12, pp. 4398–4411, Dec. 2019.
- [2] M. Zhang, W. Li, and Q. Du, “Diverse region-based CNN for hyperspectral image classification,” *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2623–2634, Jun. 2018.
- [3] C. Scharfenberger, A. Wong, and D. A. Clausi, “Structure-guided statistical textural distinctiveness for salient region detection in natural images,” *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 457–470, Jan. 2015.
- [4] Y. Chen, W. Zou, Y. Tang, X. Li, C. Xu, and N. Komodakis, “SCOM: Spatiotemporal constrained optimization for salient object detection,” *IEEE Trans. Image Process.*, vol. 27, no. 7, pp. 3345–3357, Jul. 2018.
- [5] Y. Gong, L. Wang, M. Hodosh, J. Hockenmaier, and S. Lazebnik, “Improving image-sentence embeddings using large weakly annotated photo collections,” in *Proc. ECCV*, 2014, pp. 529–545.
- [6] A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3128–3137.
- [7] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: A neural image caption generator,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3156–3164.
- [8] Y. Zhang, Y. Tian, Y. Kong, B. Zhong, and Y. Fu, “Residual dense network for image super-resolution,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2472–2481.
- [9] X. Peng, J. Feng, S. Xiao, W.-Y. Yau, J. T. Zhou, and S. Yang, “Structured AutoEncoders for subspace clustering,” *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5076–5086, Oct. 2018.
- [10] X. Peng, H. Zhu, J. Feng, C. Shen, H. Zhang, and J. T. Zhou, “Deep clustering with sample-assignment invariance prior,” *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Dec. 31, 2020, doi: 10.1109/TNNLS.2019.2958324.
- [11] X. Zhao, K.-H. Lin, Y. Fu, Y. Hu, Y. Liu, and T. S. Huang, “Text from corners: A novel approach to detect text and caption in videos,” *IEEE Trans. Image Process.*, vol. 20, no. 3, pp. 790–799, Mar. 2011.
- [12] D. Elliott and F. Keller, “Image description using visual dependency representations,” in *Proc. EMNLP*, 2013, pp. 1292–1302.
- [13] A. Farhadi *et al.*, “Every picture tells a story: Generating sentences from images,” in *Proc. Eur. Conf. Comput. Vis. Springer*, 2010, pp. 15–29.
- [14] V. Ordonez, G. Kulkarni, and T. L. Berg, “Im2text: Describing images using 1 million captioned photographs,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.
- [15] M. Hodosh, P. Young, and J. Hockenmaier, “Framing image description as a ranking task: Data, models and evaluation metrics,” *J. Artif. Intell. Res.*, vol. 47, pp. 853–899, Aug. 2013.
- [16] J. Gu, G. Wang, J. Cai, and T. Chen, “An empirical study of language CNN for image captioning,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1222–1231.
- [17] Y. Xian and Y. Tian, “Self-guiding multimodal LSTM—When we do not have a perfect training dataset for image captioning,” *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5241–5252, May 2019.
- [18] K. Xu *et al.*, “Show, attend and tell: Neural image caption generation with visual attention,” in *Proc. ICML*, 2015, pp. 2048–2057.
- [19] F. Mun, M. Cho, and B. Han, “Text-guided attention model for image captioning,” in *Proc. AAAI*, 2017, pp. 1–7.
- [20] J. Gu, J. Cai, G. Wang, and T. Chen, “Stack-captioning: Coarse-to-fine learning for image captioning,” in *Proc. AAAI*, 2018, pp. 1–8.
- [21] F. Chen, R. Ji, X. Sun, Y. Wu, and J. Su, “GroupCap: Group-based image captioning with structured relevance and diversity constraints,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1345–1353.
- [22] L. Chen *et al.*, “SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5659–5667.
- [23] W. Jiang, L. Ma, X. Chen, H. Zhang, and W. Liu, “Learning to guide decoding for image captioning,” in *Proc. AAAI*, 2018, pp. 1–8.
- [24] W. Zhao *et al.*, “A multi-task learning approach for image captioning,” in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 1205–1211.
- [25] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy, “Optimization of image description metrics using policy gradient methods,” in *Proc. ICCV*, 2017, pp. 873–881.
- [26] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel, “Self-critical sequence training for image captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7008–7024.
- [27] C. Xu, W. Zhao, M. Yang, X. Ao, W. Cheng, and J. Tian, “A unified generation-retrieval framework for image captioning,” in *Proc. 28th ACM Int. Conf. Inf. Knowl. Manage.*, Nov. 2019, pp. 2313–2316.
- [28] J. Liu *et al.*, “Interactive dual generative adversarial networks for image captioning,” in *Proc. AAAI*, 2020, pp. 11588–11595.
- [29] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [30] O. Russakovsky *et al.*, “ImageNet large scale visual recognition challenge,” *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [31] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proc. ICLR*, vol. 2013, pp. 1–12.
- [32] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [33] P. Anderson *et al.*, “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. CVPR*, 2018, pp. 6077–6086.
- [34] R. Paulus, C. Xiong, and R. Socher, “A deep reinforced model for abstractive summarization,” in *Proc. ICLR*, 2018, pp. 1–12.
- [35] J. Zhang, C. Tao, Z. Xu, Q. Xie, W. Chen, and R. Yan, “EnsembleGAN: Adversarial learning for retrieval-generation ensemble model on short-text conversation,” in *Proc. 42nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2019, pp. 435–444.
- [36] L. Yu, W. Zhang, J. Wang, and Y. Yu, “Seqgan: Sequence generative adversarial nets with policy gradient,” in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 1–7.
- [37] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
- [38] J. Wang *et al.*, “IRGAN: A minimax game for unifying generative and discriminative information retrieval models,” in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Aug. 2017, pp. 515–524.
- [39] R. J. Williams, “Simple statistical gradient-following algorithms for connectionist reinforcement learning,” *Mach. Learn.*, vol. 8, nos. 3–4, pp. 229–256, May 1992.
- [40] Y. Pu *et al.*, “Variational autoencoder for deep learning of images, labels and captions,” in *Proc. NIPS*, 2016, pp. 2352–2360.
- [41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan, “Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 652–663, Apr. 2017.
- [42] Q. Wu, C. Shen, L. Liu, A. Dick, and A. Van Den Hengel, “What value do explicit high level concepts have in vision to language problems?” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 203–212.
- [43] J. Lu, C. Xiong, D. Parikh, and R. Socher, “Knowing when to look: Adaptive attention via a visual sentinel for image captioning,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 375–383.
- [44] J. Aneja, A. Deshpande, and A. G. Schwing, “Convolutional image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5561–5570.

- [45] J. Lu, J. Yang, D. Batra, and D. Parikh, "Neural baby talk," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7219–7228.
- [46] T. Yao, Y. Pan, Y. Li, and T. Mei, "Hierarchy parsing for image captioning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2621–2629.
- [47] J. Gao, S. Wang, S. Ma, and W. Gao, "Self-critical N-Step training for image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6300–6308.
- [48] S. Herdade, A. Kappeler, K. Boakye, and J. Soares, "Image captioning: Transforming objects into words," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11135–11145.
- [49] Y. Qin, J. Du, Y. Zhang, and H. Lu, "Look back and predict forward in image captioning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8367–8375.
- [50] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo, "Image captioning with semantic attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4651–4659.
- [51] Z. Gan *et al.*, "Semantic compositional networks for visual captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5630–5639.
- [52] X. Liu, H. Li, J. Shao, D. Chen, and X. Wang, "Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 338–354.
- [53] Y. Yu, Z. Gong, P. Zhong, and J. Shan, "Unsupervised representation learning with deep convolutional neural network for remote sensing images," in *Proc. Int. Conf. Image Graph.* New York, NY, USA: Springer, 2017, pp. 97–108.
- [54] J. Donahue *et al.*, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
- [55] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: A method for automatic evaluation of machine translation," in *Proc. 40th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, 2001, pp. 311–318.
- [56] A. Lavie and A. Agarwal, "METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments," in *Proc. 2nd Workshop Stat. Mach. Transl. (StatMT)*, 2007, pp. 228–231.
- [57] C.-Y. Lin and E. Hovy, "Manual and automatic evaluation of summaries," in *Proc. Workshop Autom. Summarization (ACL)*, vol. 8, 2002, pp. 45–51.
- [58] R. Vedantam, C. L. Zitnick, and D. Parikh, "CIDEr: Consensus-based image description evaluation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4566–4575.
- [59] T.-Y. Lin *et al.*, "Microsoft coco: Common objects in context," in *Proc. ECCV*, 2014, pp. 740–755.
- [60] J. L. Fleiss, "Measuring nominal scale agreement among many raters," *Psychol. Bull.*, vol. 76, no. 5, p. 378, 1971.
- [61] A. Lindh, R. J. Ross, A. Mahalunkar, G. Salton, and J. D. Kelleher, "Generating diverse and meaningful captions," in *Proc. Int. Conf. Artif. Neural Netw.* New York, NY, USA: Springer, 2018, pp. 176–187.

Min Yang received the Ph.D. degree from the University of Hong Kong in 2017. She is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. Her research interests include natural language processing, multi-modal learning, and information retrieval.

Junhao Liu received the B.S. degree from the South China University of Technology in 2019. He is currently pursuing the master's degree with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include natural language processing and multi-modal learning.

Ying Shen received the Ph.D. degree from the University of Paris Ouest Nanterre La Défense, France, and the Erasmus Mundus master's degree from the University of Franche-Comté, France, and the University of Wolverhampton, England. She is currently an Associate Professor with the School of Intelligent Systems Engineering, Sun Yat-sen University. Her research interests include medical informatics, natural language processing, and machine learning.

Zhou Zhao (Member, IEEE) received the B.S. and Ph.D. degrees in computer science from the Hong Kong University of Science and Technology (HKUST), in 2010 and 2015, respectively. He is currently an Associate Professor with the College of Computer Science, Zhejiang University. His research interests include machine learning, data mining, and information retrieval.

Xiaojun Chen (Member, IEEE) received the Ph.D. degree from the Harbin Institute of Technology in 2011. He is currently an associate Professor with the College of Computer Science and Software, Shenzhen University. His research interests include machine learning, clustering, feature selection, and massive data mining.

Qingyao Wu (Member, IEEE) received the B.S. degree in software engineering from the South China University of Technology, China, in 2007, and the M.S. and Ph.D. degrees in computer science from the Harbin Institute of Technology, China, in 2009 and 2013, respectively. He is currently a Professor with the School of Software Engineering, South China University of Technology. His research interests include transfer learning, multi-label learning, and weak supervised learning.

Chengming Li (Member, IEEE) received the B.S. and M.S. degree in computer application technology from the Dalian University of Technology in 2009 and 2011, respectively, and Ph.D. degree from the Graduate School of Information Science and Electrical Engineering, Kyushu University, in 2015. He is currently an Associate Professor with the Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. His research interests include data mining, network security, and big data.