Group-based Distinctive Image Captioning with Memory Attention

Jiuniu Wang[1,2,3], Wenjia Xu[2,3], Qingzhong Wang[1,4], Antoni B. Chan[1,*]
[1] Department of Computer Science, City University of Hong Kong
[2] Aerospace Information Research Institute, Chinese Academy of Sciences
[3] University of Chinese Academy of Sciences
[4] Baidu Research
{jiuniwang2-c,qingzwang2-c}@my.cityu.edu.hk, xuwenjia16@mails.ucas.ac.cn, abchan@cityu.edu.hk

ABSTRACT

Describing images using natural language is widely known as image captioning, which has made consistent progress due to the development of computer vision and natural language generation techniques. Though conventional captioning models achieve high accuracy based on popular metrics, i.e., BLEU, CIDEr, and SPICE, the ability of captions to distinguish the target image from other similar images is under-explored. To generate distinctive captions, a few pioneers employ contrastive learning or re-weighted the ground-truth captions, which focuses on one single input image. However, the relationships between objects in a similar image group (e.g., items or properties within the same album or fine-grained events) are neglected. In this paper, we improve the distinctiveness of image captions using a Group-based Distinctive Captioning Model (GdisCap), which compares each image with other images in one similar group and highlights the uniqueness of each image. In particular, we propose a group-based memory attention (GMA) module, which stores object features that are unique among the image group (i.e., with low similarity to objects in other images). These unique object features are highlighted when generating captions, resulting in more distinctive captions. Furthermore, the distinctive words in the ground-truth captions are selected to supervise the language decoder and GMA. Finally, we propose a new evaluation metric, distinctive word rate (DisWordRate) to measure the distinctiveness of captions. Quantitative results indicate that the proposed method significantly improves the distinctiveness of several baseline models, and achieves the state-of-the-art performance on both accuracy and distinctiveness. Results of a user study agree with the quantitative evaluation and demonstrate the rationality of the new metric DisWordRate.

CCS CONCEPTS

• **Computing methodologies → Computer vision**.

KEYWORDS

Image Caption, Distinctiveness, Memory Attention, Similar Image

ACM Reference Format:
Jiuniu Wang[1,2,3], Wenjia Xu[2,3], Qingzhong Wang[1,4], Antoni B. Chan[1,]. 2021. Group-based Distinctive Image Captioning with Memory Attention. In *Proceedings of the 29th ACM International Conference on Multimedia (MM ’21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475215

1 INTRODUCTION

The task of image captioning has drawn much attention from both the computer vision and natural language generation communities, and consistent progress has been made due to the development of vision and language techniques [1, 20, 30, 38]. The fluent and accurate descriptions have aided many applications such as summarizing photo albums, tagging images on the internet, etc. An exciting application has been adopted on mobile phones, which speaks out what the cellphone camera sees[1], to act as a tool describing the world for visually-impaired people. However, as pointed out in [19, 31, 35], traditional captioning models that optimize the cross-entropy loss or reinforcement reward may lead to over-generic image captions. Although the automatic image caption generators can accurately describe the image, they generate generic captions for images with similar semantic meaning, lacking the intrinsic human ability to describe the unique details of images to distinguish these images from others. For instance, as shown in Figure 1, simply mentioning the traffic light without explaining the specific meaning (e.g., the color of the traffic light) cannot help visually-impaired people to make a decision whether or not to cross the street. We argue that a model that describes the distinctive contents of each image is more likely to highlight the truly useful information. In this paper, we consider the goal of endowing image captioning models to generate distinctive captions. Here we refer to the *distinctiveness* of a caption as its ability to *describe the unique objects or context of the target image so as to distinguish it from other semantically similar images.*

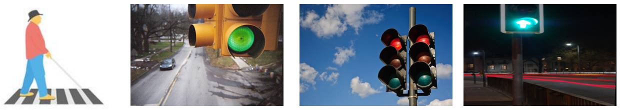
Most of the existing image captioning models aim to generate a caption that best describes the semantics of a target image. Distinctiveness, on the other hand, requires the caption to best match the target image among similar images, i.e., describing the distinctive parts of the target image. Some efforts have been made to generate *diverse* captions to enrich the concepts by employing conditional GAN [7, 25], VAE [13, 32] or reinforcement learning [36, 37]. However, improving the diversity cannot guarantee distinctiveness,

[*] Corresponding author.

[1] https://www.apple.com/accessibility/vision/

MM ’21, October 20–24, 2021, Virtual Event, China

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475215



Ours	A green traffic light on the side of a street.	A red traffic light with blue sky in the background.	A green traffic light on a city street at night.
DiscCap [18]	A traffic light on the side of a street.	A traffic light with a blue sky on it.	A traffic light on the side of a street.
CIDErBtwCap [32]	A traffic light on the side of a road.	Two traffic lights on a pole in the sky.	A traffic light on a city street at night.

Figure 1: Our model generates distinctive captions that can distinguish the target image from other similar images. Compared to current distinctive image captioning models [18, 31], our captions can specify the important details, e.g., the color and the environment of the traffic light, which can help a visually-impaired person to cross the street.

since rephrasing the expressions or enriching the vocabulary do not necessarily introduce novel and distinctive information. Several methods proposed to improve the distinctiveness by contrastive learning [8, 16, 18], where they either aggregate the contrastive image features with target image feature, or apply contrastive loss to suppress the estimated conditional probabilities of mismatched image-caption pairs [8, 18]. However, the distractors are either a group of images with scene graph that partially overlaps with the target image [16], or randomly selected unmatched image-caption pairs [8, 18], which are easy to distinguish. [17, 29] introduce self-retrieval reward with contrastive loss, which requires the generated captions to retrieve the target image in a visual-language space. However, weighing too much on image retrieval could lead a model to repeat the distinctive words [31], which may hurt caption quality.

In this work, we consider the hard negatives, i.e., similar images that generally share similar semantics with the target image, and push the captions to clearly show the difference between these images. For instance, as shown in Figure 1, the generated captions should specify the different aspects of the target image (e.g., different light colors) compared with other images that share similar semantics (e.g., images of traffic lights). Recently, the developments of transformer-based model and attention mechanism improve the accuracy of image captions. In this paper, we propose a plug-and-play distinctive memory attention module to extend the transformer-based captioning models, where we put high attention on distinctive objects detected in the target image – object features in the target image with low similarity to object features of the similar images are considered more distinctive, and thus receive higher attention.

In summary, the contributions of this paper are three-fold:

(1) We propose a Group-based Distinctive Captioning Model (GdisCap), which builds memory vectors from object regions that are weighted by distinctiveness among the image group, and then generates distinctive captions for the images in the group.
(2) To enforce the weighted memory to contain distinctive object information, we further propose two distinctive losses, where the supervision is the distinctive words occurring in the ground-truth (GT) captions.
(3) We conduct extensive experiments and user studies, demonstrating that the proposed model is able to generate distinctive captions. In addition, our model also highlights the unique regions of each image, which is more interpretable.

2 RELATED WORK

Image captioning bridges two domains—images and texts. Classical approaches usually extract image representations using Convolutional Neural Network (CNN), then feed them into Recurrent Neural Network (RNN) and output sequences of words [14, 20, 30]. Recent advances mainly focus on improving the image encoder and the language decoder. For instance, Anderson et al. [1] proposes bottom-up features, which are extracted by a pre-trained Faster-RCNN [23] and a top-down attention LSTM, where an object is attended in each step when predicting captions. Apart from using RNNs as the language decoder, Aneja et al. [2], Wang and Chan [33, 34] utilize CNNs since LSTMs cannot be trained in a parallel manner. Cornia et al. [6], Li et al. [15] adopt transformer-based networks with multi-head attention to generate captions, which mitigate the long-term dependency problem in LSTMs and significantly improves the performance of image captioning. Recent advances usually optimize the network with a standard training procedure, where they pre-train the model with word-level cross-entropy loss (XE) and fine-tune with reinforcement learning (RL). However, as pointed out in [7, 8, 31], training with XE and RL may encourage the model to predict an “average” caption that is close to all ground-truth (GT) captions, thus resulting in over-generic captions that lacks distinctiveness.

More relevant to our work are the recent works on group-based image captioning [4, 16, 28], where a group of images is utilized as context when generating captions. Vedantam et al. [28] generates sentences that describe an image in the context of other images from closely related categories. Chen et al. [4] summarizes the unique information of the target images contrasting to other reference images, and Li et al. [16] emphasizes both the relevance and diversity. Our work is different in the sense that we simultaneously generate captions for each image in a similar group, and highlight the difference among them by focusing on the distinctive image regions. Both Chen et al. [4], Vedantam et al. [28] extract one image feature from the FC layer for each image, where all the semantics and objects are mixed up. While our model focuses on the object-level features and explicitly finds the unique objects that share less similarity with the context images, leading to fine-grained and concrete distinctiveness.

Distinctive image captioning aims to overcome the problem of generic image captioning, by describing sufficient details of the target image to distinguish it from other images. Dai and Lin [8] promotes the distinctiveness of image caption by contrastive learning. The model is trained to give a high probability to the GT image-caption pair and lower the probability of the randomly sampled negative pair. Liu et al. [17], Luo et al. [18], Vered et al. [29] takes the same idea that the generated caption should be similar to the target image rather than other distractor images in a batch, and applies caption-image retrieval to optimize the contrastive loss.

However, the distractor images are randomly sampled in a batch, which can be easily distinguished from the target images. In our work, we consider *hard negatives* that share similar semantics with the target image, and push the captions to contain more details and clearly show the difference between these images. More recently, [31] proposes to give higher weight to the distinctive GT captions during model training. Chen et al. [4] models the diversity and relevance among positive and negative image pairs in a language model, with the help of a visual parsing tree [3]. In contrast to these works, our work compares a group of images with a similar context, and highlights the unique object regions in each image to distinguish them from each other. That is, our model infers which object-level features in each image are unique among all images in the group. Our model is applicable to most of the transformer-based captioning models.

Attention mechanism applies visual attention to different image regions when predicting words at each time step, which has been widely explored in image captioning [5, 11, 21, 38, 41]. For instance, You et al. [41] adopts semantic attention to focus on the semantic attributes in image. Anderson et al. [1] exploits object-level attention with bottom-up attention, then associates the output sequences with salient image regions via top-down mechanism. More recently, self-attention networks introduced by Transformers [27] are widely adapted in both language and vision tasks [9, 22, 26, 39, 40]. Guo et al. [11] normalizes the self-attention module in the transformer to solve the internal covariate shift. Huang et al. [12] weights the attention information by a context-guided gate. These works focus on learning self-attention between every word token or image region in one image. Li et al. [16] migrates the idea of self-attention to visual features from different images, and averages the group *image-level* vectors with self-attention to detect prominent features. In contrast, in our work, we take a further step by learning memory attention among the R-CNN *object-level* features [1] extracted from similar images, to highlight the prominent features that convey distinguishing semantics in the similar image group.

3 METHODOLOGY

We present the framework of our proposed Group-based Distinctive Captioning model (GdisCap) in Figure 2. Our model aims at generating distinctive captions for each image within a group of semantically similar images. Given an image group with $K+1$ images, denoted as $\{I_0, I_1, \dots, I_K\}$, GdisCap generates distinctive captions for each image. Different from the conventional image captioning task, the generated captions should describe both the salient content in the target image, and also highlight the uniqueness of the target image (e.g., I_0) compared to other K images (e.g., I_1 to I_K) in the same group. Specifically, during training, each image in the group is treated equally, and we use each image as a target. In Figure 2, we show an example where I_0 is the target image.

To achieve the goal of distinctive image captioning, we first construct similar image groups, then we employ the proposed Group-based Memory Attention (GMA) to extract the distinctive object features. Finally, we design two distinctive losses to further encourage generating distinctive words.

3.1 Similar image group

Similar image group was first introduced in [31] to evaluate the distinctiveness of the image captions. For training, our model handles several similar image groups as one batch, simultaneously using each image in the group as a target image. Here, we dynamically construct the similar image groups during training as follows:

(1) To construct a similar image group, we first randomly select one image as the target image I_0, and then retrieve its K nearest images through a semantic similarity metric, measured by the visual-semantic retrieval model VSE++ [10], as in [31]. In detail, given the target image I_0, we use VSE++ to retrieve those captions that well describe I_0 among all human-annotated captions, and then the corresponding images of those captions are the nearest images.

(2) Due to the distribution inequality, the images sharing similar semantic meaning will form clusters in the VSE++ space. The images in the cluster center may be close to many other images, and to prevent them from dominating the training, the $K+1$ images that are used to create one similar image group are removed from the image pool, so they will not be selected when constructing other groups. In this way, each image will belong to only one group, with no duplicate images appearing in one epoch.[2]

Each data split (training, validation, test) is divided into similar image groups independently. For each training epoch, we generate new similar image groups to encourage training set diversity.

3.2 Group-based Distinctive Captioning Model

Here we introduce the group-based distinctive captioning model (GdisCap), and how we incorporate the Group-based Memory Attention (GMA) module that encourages the model to generate distinctive captions. Notably, the GMA module can serve as a plug-and-play tool for distinctive captioning, which can be applied to most existing transformer-based image captioning models.

3.2.1 *Transformer-based Image Captioning.*

Our captioning model is build on a transformer-based architecture [6], as illustrated in Figure 2 (left). The model can be divided into two parts: an image *Encoder* that processes input image features, and a caption *Decoder* that predicts the output caption word by word. In transformer-based architecture, *Encoder* and *Decoder* are both composed of several multi-head attention and MLP layers.

Here we take the bottom-up features [1] extracted by Faster R-CNN [23] as the input. Given an image I, let $X = \{x^i\}_{i=1}^N$ denotes the object features, where N is the number of region proposals and $x^i \in R^d$ is the feature vector for the i-th proposal. The output of the l-th encoder layer is calculated as follows:

$$O_l^{att} = \mathbf{LN}\left(X_{l-1} + \mathbf{MH}\left(\mathbf{W}_\mathrm{q} X_{l-1}, \mathbf{W}_\mathrm{k} X_{l-1}, \mathbf{W}_\mathrm{v} X_{l-1}\right)\right), \quad (1)$$

$$X_l = \mathbf{LN}\left(O_l^{att} + \mathbf{MLP}\left(O_l^{att}\right)\right), \quad (2)$$

where $\mathbf{LN}(\cdot)$ denotes layer normalization, $\mathbf{MLP}(\cdot)$ denotes a multilayer perceptron, and $\mathbf{MH}(\cdot)$ represents the multi-head attention layer. $\mathbf{W}_\mathrm{q}, \mathbf{W}_\mathrm{k}, \mathbf{W}_\mathrm{v}$ are learnable parameters.

The *Encoder* turns X into memory vectors $M = \{m^i\}_{i=1}^N$, where $m^i \in R^{d_m}$ encodes the information from the i-th object proposal x^i,

[2] When almost all images are selected, the remaining images are not similar enough to construct groups. We regard them as target images one by one, and find similar images from the whole image pool.

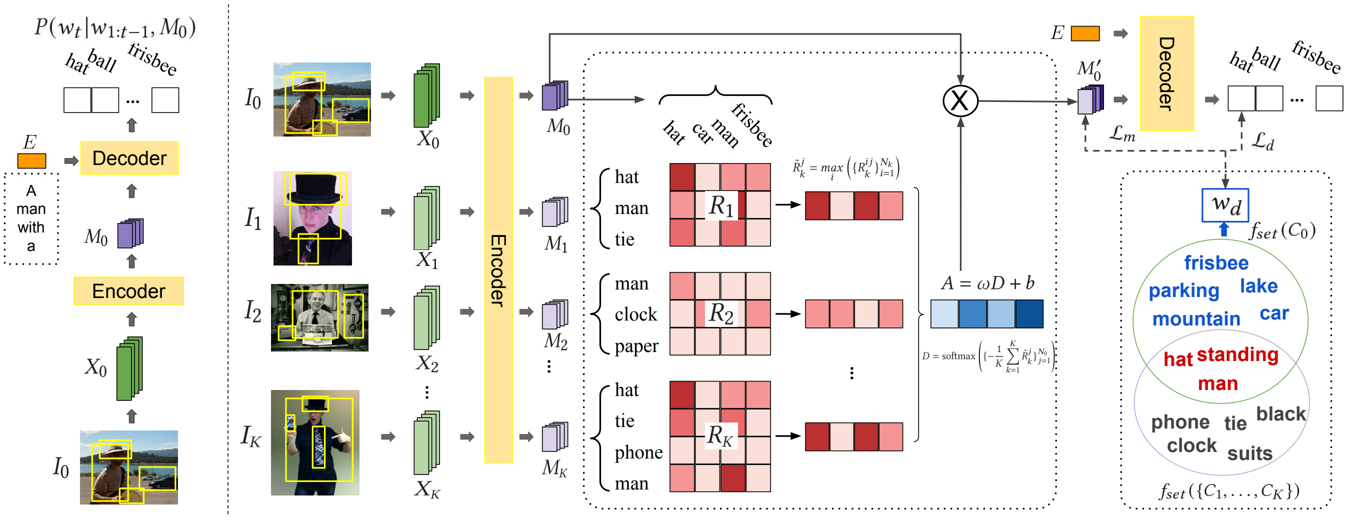


Figure 2: Left: the standard transformer-based captioning model, where the target image features X_0 are the region-based visual features extracted via RoI pooling from Faster R-CNN. Right: our Group-based Distinctive Captioning Model (GdisCap), which consists of a group-based memory attention (GMA) module that weights the memory vectors according to their similarity with other similar images. Our model takes a group of images as input, and outputs one caption for each image. Only one target memory, one decoder and one output caption are shown here to reduce clutter.

and is affected by other objects in the multi-head attention layers, which contains both single object features and the relationships among objects. According to the memory M and the embedding E of the previous word sequence $\{w_1, \ldots, w_{t-1}\}$, the *Decoder* generates the v-dimensional word probability vector $P_t = P(w_t|w_{1:t-1}, M)$ at each time step t, where v is the size of vocabulary.

3.2.2 *Group-based Memory Attention.*

The goal of the group-based memory attention is to highlight the distinctive features of the target image that do not appear in other similar images. For instance, in Figure 2 (right), the concept of *man* and *hat* that appear in the target image I_0 are also shared in other similar images, but *frisbee* and *cars* are unique for I_0 and can distinguish I_0 from other images. However, the standard captioning model in Figure 2 (left) cannot highlight those objects, since each memory vector m_0^i for different image regions are treated equally when fed into the *Decoder*.

In this work, we aim to give more attention to the distinctive image regions when generating captions. Hence, the model will describe the distinctive aspects of the input image instead of only describing the most salient regions. To this end, we propose a Group-based Memory Attention (GMA) module (see Figure 2 (right)), where the attention weight for each object region is obtained by calculating the distinctiveness of its memory vector m_0^i. Then we encourage the model to generate distinctive words associated with the unique object regions. Specifically, the GMA produces distinctive attention $A = \{a_i\}_{i=1}^{N_0} \in R^{N_0}$ for memory vectors M_0. When generating captions, instead of using M_0, a weighted target memory is fed into the decoder:

$$M_0' = \{a_i \cdot m_0^i\}_{i=1}^{N_0}. \tag{3}$$

Computing distinctive attention. To compute the distinctive attention, we need to compare the objects in the target image with those in the similar images. As shown in Figure 2 (right), the target image I_0 and its similar images $\{I_k\}_{k=1}^K$ are transferred into memory vectors $M_0 = \{m_0^j\}_{j=1}^{N_0}$ and $M_k = \{m_k^i\}_{i=1}^{N_k}$, $k = 1 \ldots K$, via the image encoder, where N_k denotes the number of objects in the k-th image. The GMA first measures the similarity $R_k \in R^{N_k \times N_0}$ of each target memory vector m_0^j and each memory vector m_k^i in similar images via cosine similarity:

$$R_k^{ij} = \cos(m_k^i, m_0^j)\,, \tag{4}$$

where $m_0^j \in R^{d_m}$ is the j-th vector in M_0 (e.g., memory vectors for *hat*, *car*, *man* and *frisbee* in Figure 2), and m_k^i is the i-th vector in M_k (e.g., memory vectors for *hat*, *man*, and *tie* from M_k in Figure 2).

The similarity matrix reflects how common an object is – a common object that occurs in many images is not distinctive for the target image. For example, as shown in Figure 2 (right), hat is less distinctive since it occurs in multiple images, while car is an unique object that only appear in target image. To summarize the similarity matrix, we compute an object-image similarity map $\tilde{R}_k \in R^{N_0}$ as

$$\tilde{R}_k^j = \max_i \left(\{R_k^{ij}\}_{i=1}^{N_k}\right), \tag{5}$$

where $\tilde{R}_k^j$ is the similarity of the best matching object-region in image I_k to region j in the target image I_0.

We assume that objects with higher similar scores are less distinctive. Hence, the distinctiveness scores $D \in R^{N_0}$ for each memory vector m_0^i are computed by softmax of the average of object-image similarity maps:

$$D = \text{softmax}\left(\{-\frac{1}{K}\sum_{k=1}^{K}\tilde{R}_k^j\}_{j=1}^{N_0}\right), \tag{6}$$

where higher scores indicates higher distinctiveness, i.e., lower average similarity to other similar images. Note that the values of D are in range $(0, 1)$ due to the softmax function. Finally, the distinctive attention A for each target memory vector in (3) is calculated as:

$$A = \omega D + b\,, \tag{7}$$

where ω and b are two learnable parameters. The bias term b controls the minimum value of A, i.e., the base attention level for all regions, while ω controls the amount of attention increase due to the distinctiveness. We clip ω and b to be non-negative, so that the attention value in A is non-negative.

3.3 Loss functions

Two typical loss functions for training image captioning are cross-entropy loss and reinforcement loss. Since these two losses only use the GT captions of the image for supervision, they may encourage the generated captions to mimic GT captions, resulting in over-genericness, i.e., lack of distinctiveness. To address this issue, we take a step further to define distinctive words and explicitly encourage the model to learn more from these words. In this section, we first review the two typical loss functions used in captioning models, and then present our proposed Distinctive word loss (DisLoss) and Memory Classification loss (MemCls) for training our GMA module.

Cross-Entropy Loss. Given the i-th GT caption of image I_0, $C_0^i = \{w_t\}_{t=1}^T$, the cross-entropy loss is

$$\mathcal{L}_{xe} = -\sum_{t=1}^{T} \log P(w_t|w_{1:t-1}, M_0')\,, \tag{8}$$

where $P(w_t|w_{1:t-1}, M_0)$ denotes the predicted probability of the word w_t conditioning on the previous words $w_{1:t-1}$ and the weighted memory vectors M_0', as generated by the caption *Decoder*.

Reinforcement learning loss. Following [24], we apply reinforcement learning to further improve the accuracy of our trained network using the loss:

$$\mathcal{L}_r = -E_{\hat{c}\sim p(c|I)}\left[\frac{1}{d_c}\sum_{i=1}^{d_c} g(\hat{c}, C_0^i)\right], \tag{9}$$

where $g(\hat{c}, C_0^i)$ is the CIDEr value between the predicted caption $\hat{c}$ and the i-th GT C_0^i, and d_c denotes the number of GT captions.

Distinctive word loss (DisLoss). In this work, we focus on the distinctive words w_d that appear in captions C_0 of target image but not in captions $\{C_1, \ldots, C_K\}$ of similar images. We define the distinctive word set as

$$w_d = f_{set}(C_0) - f_{set}(\{C_1, \ldots, C_K\})\,, \tag{10}$$

where $f_{set}(\cdot)$ denotes the function that converts the sentence into word set, and "−" means set subtraction.

In the training phase, we explicitly encourage the model to predict the distinctive words in w_d by optimize the distinctive loss $\mathcal{L}_d$,

$$\mathcal{L}_d = -\sum_{t=1}^{T}\sum_{i=1}^{u} \log P(w_t = w_d^i|w_{1:t-1}, M_0')\,, \tag{11}$$

where w_d^i denotes the i-th distinctive word in w_d, and $P(w_t = w_d^i|w_{1:t-1}, M_0')$ denotes the probability of predicting word w_d^i as the t-th word in sentence. u is the number of words in w_d, and T is the length of the sentence.

Memory classification loss (MemCls). In order to generate distinctive captions, the *Decoder* requires the GMA to produce memory contents containing distinctive concepts. However, the supervision on the GMA through the *Decoder* could be too weak, which may allow the GMA to also produce non-useful information, e.g., highlighting too much background or focusing on small objects that are not mentioned in the GT captions. To improve the distinctive content produced by the GMA, we introduce an *auxiliary classification task* that predicts the distinctive words from the weighted memory vectors M_0' of the GMA,

$$P_M = f_{MC}(M_0')\,, \tag{12}$$

where P_M denotes the word probability vector and f_{MC} is the classifier. To associate the memory vectors with distinctive words, we employ the multi-label classification loss $\mathcal{L}_m$ to train the classifier,

$$\mathcal{L}_m = -\sum_{i=1}^{u} \log(P_{M,w_d^i})\,, \tag{13}$$

where P_{M,w_d^i} is the predicted probability of the i-th distinctive word.

The final loss. The final training loss $\mathcal{L}$ is formulated as

$$\mathcal{L} = \alpha_c\mathcal{L}_{xe} + \alpha_r\mathcal{L}_r + \alpha_d\mathcal{L}_d + \alpha_m\mathcal{L}_m\,, \tag{14}$$

where $\{\alpha_c, \alpha_r, \alpha_d, \alpha_m\}$ are hyper-parameters for their respective losses. The training procedure has two stages. In the first stage, we set $\alpha_c = 1$ and $\alpha_r = 0$, so that the network is mainly trained by cross-entropy loss α_c. In the second stage, we set $\alpha_c = 0$ and $\alpha_r = 1$, so that the parameters are mainly optimized by reinforcement learning loss $\mathcal{L}_r$. We adaptively set $\{\alpha_d, \alpha_m\}$ so that $\alpha_d\mathcal{L}_d$ and $\alpha_m\mathcal{L}_m$ are one quarter of $\mathcal{L}_{xe}$ (or $\mathcal{L}_r$).

During training, each mini-batch comprises several similar image groups, with the loss aggregated over each image as a target in its group. Details for processing one image group are in supplemental.

4 EXPERIMENTS

In this section, we first introduce the implementation details and dataset preparation, then we quantitatively evaluate the effectiveness of our model by an ablation study and a comparison with other state-of-the-art models.

4.1 Implementation details

Following [1], we use the spatial image features extracted from Faster-RCNN [23] with dimension $d = 2048$. Each image usually contain around 50 object region proposals, i.e. $N_k \approx 50$. Each object proposal has a corresponding memory vector with dimension $d_m = 512$. We set $K = 5$ for constructing the similar image groups. The values in distinctive attention A are mostly in the range of $(0.5, 0.9)$. To verify the effectiveness of our model, we conduct experiments on four baseline methods (i.e., Transformer [27], M^{2}Transformer [6], Transformer + SCST [27] and M^{2}Transformer + SCST) [6]. Our experimental settings (e.g., data preprocessing and vocabulary construction) follow these baseline models. We apply our GMA module on Transformer model and all three layers in M^{2}Transformer model. Note that our model is applicable to most of the transformer-based image captioning models, and we choose these four models as the baseline due to their superior performance on accuracy-based metrics.

4.2 Dataset and metrics

4.2.1 Dataset. We conduct the experiments using the most popular dataset—MSCOCO, which contains 12, 387 images, and each image has 5 human annotations. Following [1], we split the dataset into 3 sets—5, 000 images for validation, 5, 000 images for testing and the rest for training. When constructing similar image groups for test set, we adopt the same group split as [31] for a fair comparison.

4.2.2 Metrics. We consider two groups of metrics for evaluation. The first group includes the metrics that evaluate the accuracy of the generated captions, such as CIDEr and BLEU. The second group assesses the distinctiveness of captions. For the latter, CIDErBtw [31] calculates the CIDEr value between generated captions and GT captions of its similar image group. However, CIDErBtw only works when comparing two methods with similar CIDEr value, e.g., a random caption that has lower overlap with the GT captions will be considered as distinctive since it achieves lower CIDErBtw. Hence, we propose two new distinctiveness metrics as follows.

CIDErRank. A distinctive caption $\hat{c}$ (generated from image I_0) should be similar to the target image's GT captions C_0, while different from the GT captions of other images in the same group $\{C_1, \dots, C_K\}$. Here we use CIDEr values $\{s_k\}_{k=0}^{K}$ to indicate the similarity of the caption c with GT captions in images group as

$$s_k = \frac{1}{d_c}\sum_{i=1}^{d_c} g(\hat{c}, C_k^i)\,, \tag{15}$$

where $g(\hat{c}, C_k^i)$ represents the CIDEr value of predicted caption $\hat{c}$ and i-th GT caption in C_k. We use the rank of s_0 in $\{s_k\}_{k=0}^{K}$ to show the distinctiveness of the models as

$$r = f_{rank}\left(s_0, \{s_k\}_{k=0}^{K}\right)\,, \tag{16}$$

where $f_{rank}(\cdot)$ means s_0 is the r-th largest value in $\{s_k\}_{k=0}^{K}$. The best rank is 1, indicating the generated caption $\hat{c}$ is mostly similar to its GT captions and different from other captions, while the worst rank is $K + 1$. Thus, the average r reflects the performance of captioning models, with more distinctive captions having lower CIDErRank.

DisWordRate. We design this metric based on the assumption that using distinctive words should indicate that the generated captions are distinctive. The *distinctive word rate* (DisWordRate) of a generated caption $\hat{c}$ is calculated as:

$$DisWordRate = \max_i \frac{|w_d \cap \hat{c}|}{|w_d \cap C_0^i|}, \quad i = 1, \dots, d_c\,, \tag{17}$$

where d_c is the number of sentence in C_0, $|w_d \cap \hat{c}|$ represents the number of elements in w_d that appear in $\hat{c}$. Thus, DisWordRate reflects the percentage of distinctive words in the generated captions.

Method	D(%)↑	CR↓	CB↓	C↑	B3↑	B4↑
Transformer [27]	16.8	2.47	74.8	111.7	45.1	34.0
+ GdisCap (ours)	**19.5**	2.42	**70.9**	107.3	43.4	32.7
M^{2}Transformer [6]*	16.4	2.52	76.8	111.8	45.2	34.7
+ GdisCap (ours)	18.7	2.43	72.5	109.8	45.6	34.7
Transformer + SCST [27]	14.7	2.38	83.2	127.6	**51.3**	**38.9**
+ GdisCap (ours)	16.5	2.36	81.7	127.0	50.7	38.4
M^{2}Transformer + SCST [6]*	17.3	2.38	82.9	**128.9**	50.6	38.7
+ GdisCap (ours)	18.5	**2.31**	81.3	127.5	50.0	38.1
FC [24]	6.5	3.03	89.7	102.7	43.2	31.2
Att2in [24]	10.8	2.65	88.0	116.7	48.0	35.5
UpDown [1]	12.9	2.55	86.7	121.5	49.2	36.8
AoANet [12]*	14.6	2.47	87.2	128.6	50.4	38.2
DiscCap [18]	14.0	2.48	89.2	120.1	48.5	36.1
CL-Cap [8]	14.2	2.54	81.3	114.2	46.0	35.3
CIDErBtwCap [31]	15.9	2.39	82.7	127.8	51.0	38.5

Table 1: Comparison of caption distinctiveness and accuracy on MSCOCO test split: DisWordRate (D), CIDErRank (CR), and CIDErBtw (CB) measure the distinctiveness, while CIDEr (C) and BLEU (B3 and B4) measure the accuracy. ↑ and ↓ show whether higher or lower score are better according to each metric. We apply our model on four baseline models: Transformer [27] and M^{2}Transformer [6] trained only with cross entropy loss, Transformer + SCST [27] and M^{2}Transformer + SCST [6] trained with reinforcement learning. ∗ denotes we train the model from scratch with official released code.

Method	D(%)↑	CR↓	CB↓	C↑	B3↑	B4↑
M^{2}Transformer	16.4	2.52	76.8	**111.8**	45.2	34.7
+ ImageGroup	16.9	2.51	75.4	110.7	45.2	34.8
+ DisLoss	17.1	2.48	76.4	110.0	**45.9**	**35.5**
+ GMA	18.4	2.44	73.5	111.2	45.4	34.9
+ MemCls	**18.7**	**2.42**	**72.5**	109.8	45.6	34.7

Table 2: Ablation Study. We train M^{2}Transformer with cross-entropy loss (XE) as the baseline, and gradually add four components of our full model: image group based training, distinctive word loss, group-based memory attention, and memory classification loss.

4.3 Main results

In the following, we present a comparison with the state-of-the-art (distinctive) image captioning models. In addition, we present an ablation study of applying our model to the baseline method.

Comparison with the state-of-the-art. We compare our model with two groups of state-of-the-art models: 1) FC [24], Att2in [24], UpDown [1] and AoANet [12] that aim to generate captions with high accuracy; 2) DiscCap [18], CL-Cap [8], and CIDErBtwCap [31] that generate distinctive captions.

The main experiment results are presented in Table 1, and we make the following observations. First, when applied to four baseline models, our model achieves impressive improvement for the distinctive metrics, while maintaining comparable results on accuracy metrics. For example, we improve the DisWordRate by 16.1 percent (from 16.8% to 19.5%) and reduce the CIDErBtw by 5.2 percent (from 74.8 to 70.9) for Transformer, while only sacrificing the CIDEr by 3.9 percent. Second, in terms of distinctiveness, models trained with cross-entropy loss tend to perform better than models trained with SCST [24]. For example, we achieve the highest DisWordRate with Transformer + GdisCap at 19.5%. M^{2}Transformer + GdisCap also achieves higher DisWordRate than M^{2}Transformer + SCST + GdisCap. Third, compared with state-of-the-art models that improve the accuracy of generated captions, our model M^{2}Transformer + SCST + GdisCap achieves comparable accuracy, while gaining impressive improvement in distinctness – we obtain comparable CIDEr with AoANet (127.5 vs 128.6), and attain significantly higher DisWordRate, i.e., 14.6% (AoANet) vs 18.5% (ours). When compared to other models that focus on distinctiveness, M^{2}Transformer + SCST + GdisCap achieves higher distinctiveness by a large margin – we gain DisWordRate by 18.5% compared with 15.9% (CIDErBtwCap), and we obtain much lower CIDErBtw by 81.3 vs. 89.2 (DiscCap).

Ablation study. To measure the influence of each component in our GdisCap, we design an ablation study where we train the baseline M^{2}Transformer with cross-entropy loss. Four variants of GdisCap are trained by gradually adding the components, i.e., image group based training, distinctive word loss, group-based memory attention (GMA) module, and memory classification loss, to the baseline model. The results are shown in Table 2, and demonstrate that the four additional components improve the distinctive captioning metrics consistently. As pointed out in [35], increasing the distinctiveness of generated captions sacrifices the accuracy metrics such as CIDEr and BLEU, since the distinctive words cannot agree with all the GT captions due to the diversity of human language. Applying our model on top of M^{2}Transformer increases the DisWordRate by 14% (from 16.4% to 18.7%), while only sacrificing 1.8% of the CIDEr value (from 111.8 to 109.8). More ablation studies are in the Supplementary.

Trade-off between accuracy and distinctiveness. The results in Figure 3 demonstrate that improving distinctiveness typically hurts the accuracy, since the distinctive words do not appear in all the GT captions, while CIDEr considers the overlap between the generated captions and all GT captions. This can explain why the human-annotated GT captions, which are considered the upper bound of all models, only achieve CIDEr of 83.1. Compared to the baselines, our work achieves results more similar to human performance.

5 USER STUDY

To fairly evaluate the distinctiveness of our model from the human perspective, we propose a caption-image retrieval user study, which extends the evaluation protocol proposed in [18, 31]. Each test is a tuple $(I_0, I_1, \ldots, I_K, \hat{c})$, which includes a similar image group and a caption generated by a random model describing a random image in the group. The users are instructed to choose the target image that the caption corresponds to. To evaluate one image captioning model, we randomly select 50 tuples with twenty participants for each test. A correct answer is regarded as a hit, and the accuracy scores for twenty participants are averaged to obtain the final retrieval accuracy. A higher retrieval score indicates more distinctiveness of the generated caption, i.e., it can distinguish the target image from other similar images (more details are in the Supplementary).

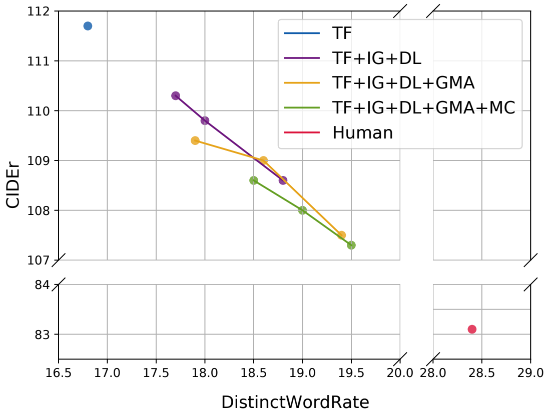


Figure 3: The trade-off between accuracy (CIDEr) and distinctiveness (DisWordRate): human-annotated GT captions (Human), baseline model Transformer (TF) [27], and three variants of our model using different components. IG, DL, MA and MC denote ImageGroup, DisLoss, Group-based Memory Attention and MemCls. For our models, we show three training stages (at different epochs), which demonstrates the trade-off between accuracy and distinctiveness during training.

Method	DiscCap [18]	CIDErBtwCap [31]	M^2+SCST [6]	*Ours*
Accuracy	48.1	58.7	61.9	**68.2**

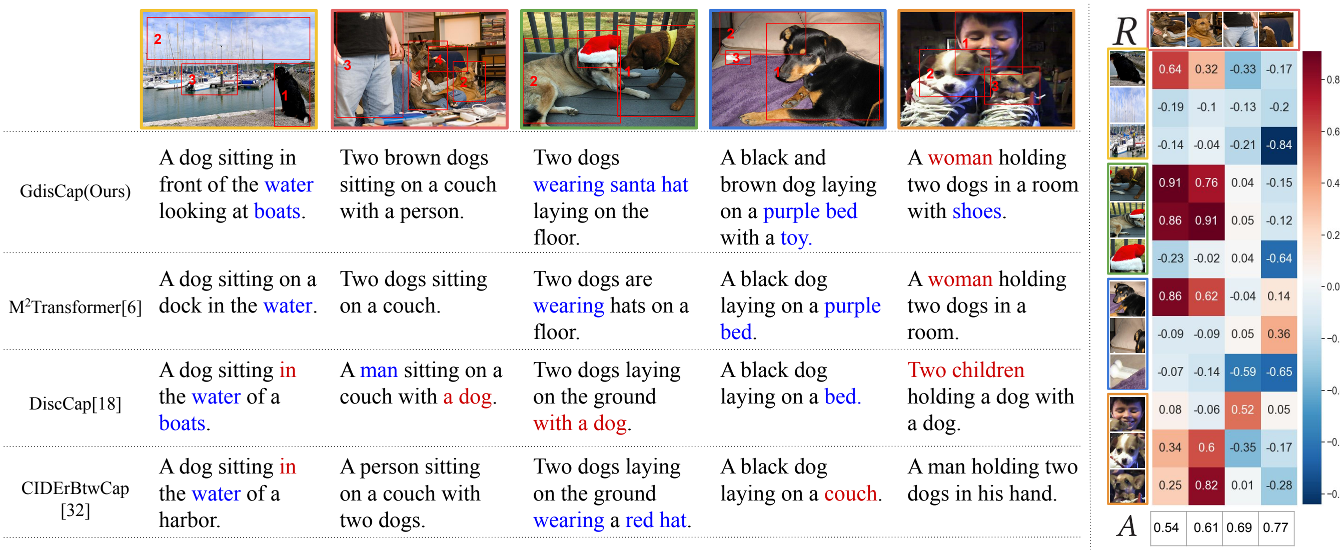
Table 3: User study results for caption-image retrieval. Our model produces captions with significantly higher retrieval accuracy (2-sample z-test on proportions, $p < 0.01$).

We compare our GdisCap model with three competitive models, DiscCap [18], CIDErBtwCap [31], and M^{2}Transformer + SCST [6]. The results are shown in Table 3, where our model achieves the highest caption-image retrieval accuracy – 68.2 compared to M^{2}Transformer + SCST with 61.9. The user study demonstrates that our model generates the most distinctive captions that can distinguish the target image from the other images with similar semantics. The results agree with the DisWordRate and the CIDErRank displayed in Table 1, which indicates that the proposed two metrics are effective evaluations similar to human judgment.

6 QUALITATIVE RESULTS

In this section, we first qualitatively evaluate our model in generating distinctive captions for similar image groups. Second, we visualize the group-based memory attention calculated by our model to highlight the distinct objects. We compare our GdisCap with three other models, M^{2}Transformer [6], DiscCap [18], and CIDErBtwCap [31]. which are the best competing methods on distinctiveness. More examples are found in the supplemental.

Figure 4 (left) displays the captions generated by four models for one similar image group. Overall, all the methods generate accurate captions specifying the salient content in the image. However, their performances on the distinctiveness differ. M^{2}Transformer and DiscCap generate captions that only mention the most salient



GdisCap(Ours)	A dog sitting in front of the water looking at boats.	Two brown dogs sitting on a couch with a person.	Two dogs wearing santa hat laying on the floor.	A black and brown dog laying on a purple bed with a toy.	A woman holding two dogs in a room with shoes.
M^{2}Transformer[6]	A dog sitting on a dock in the water.	Two dogs sitting on a couch.	Two dogs are wearing hats on a floor.	A black dog laying on a purple bed.	A woman holding two dogs in a room.
DiscCap[18]	A dog sitting in the water of a boats.	A man sitting on a couch with a dog.	Two dogs laying on the ground with a dog.	A black dog laying on a bed.	Two children holding a dog with a dog.
CIDErBtwCap [32]	A dog sitting in the water of a harbor.	A person sitting on a couch with two dogs.	Two dogs laying on the ground wearing a red hat.	A black dog laying on a couch.	A man holding two dogs in his hand.

Figure 4: Qualitative results. Left: Captions for one similar image group with five images from the test set. We compare our model with three state-of-the-art methods, M^{2}Transformer [6], DiscCap [18], and CIDErBtwCap [31]. The blue words indicate the distinctive words w_d that appear in GT captions of the target image, while not in captions of similar images. The red words denote the mistakes in the generated captions. Right: Visualization of the similarity matrix R and distinctive attention A. Here we show the second image (in red box) as the target I_0 and display the similarity value between four salient objects in I_0 and the objects in the other four images. The attention A denotes the overall distinctiveness of each object in the image I_0. Objects in the same colored box are from the same image.

objects in the image, using less distinctive words. For instance, in Figure 4 (column 1), our GdisCap generates captions "a dog sitting in front of the water looking at boats", compared to the simpler caption "a dog sitting on a dock in the water" from M^{2}Transformer. Similarly, in Figure 4 (column 3), our GdisCap describes the most distinctive property of the target image, the "santa hat", compared to DiscCap that only provides "two dogs laying on the ground". The lack of distinctiveness from M^{2}Transformer and DiscCap is due to the models being supervised by equally weighted GT captions, which tends to produce generic words that agree with all the supervisory captions.

CIDErBtwCap, on the other hand, reweights the GT captions according to their distinctiveness, and thus generates captions with more distinctive words. Compared to CIDErBtwCap, where all the objects in the image are attached with the same attention, our method yields more distinctive captions that distinguish the target image from others by attaching higher attention value to the unique details and objects that appear in the image. For example, in Figure 4 (column 3), GdisCap describes the distinctive "santa hat", while CIDErBtwCap mentions it as a "red hat".

Remarkably, GdisCap is more aware of the locations of objects in the image and the relationships among them. For example, in Figure 4 (column 5), our caption specifies the "A woman holding two dogs in a room with shoes", compared with CIDErBtwCap, which wrongly describe "holding two dogs in his hand" when no hands appear on the image. It is interesting because there is no location supervision for different objects, but our model learns the relation solely from the GT captions. Finally, Figure 4 (right) displays the similarity matrix R and distinctive attention A for the 2nd image as the target image. The object regions with highest attention are those with lower similarity to the objects in other images, in this case the "couch" and the "person". The "dogs", which are the common objects among the images, have lower non-zero attention so that they are still described in the caption.

7 CONCLUSION

In this paper, we have investigated a vital property of image captions – distinctiveness, which mimics the human ability to describe the unique details of images, so that the caption can distinguish the image from other semantically similar images. We presented a Group-based Distinctive Captioning Model (GdisCap) that compares the objects in the target image to objects in semantically similar images and highlights the unique image regions. Moreover, we developed two loss functions to train the proposed model: the distinctive word loss encourages the model to generate distinguishing information; the memory classification loss helps the weighted memory attention to contain distinct concepts. We conducted extensive experiments and evaluated the proposed model using multiple metrics, showing that the proposed model outperforms its counterparts quantitatively and qualitatively. Finally, our user study verifies that our model indeed generates distinctive captions based human judgment.

8 ACKNOWLEDGMENTS

This work was supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. CityU 11215820).

REFERENCES

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

[2] Jyoti Aneja, Aditya Deshpande, and Alexander G Schwing. 2018. Convolutional image captioning. In *CVPR*.

[3] F. Chen, R. Ji, J. Su, Y. Wu, and Y. Wu. 2017. StructCap: Structured semantic embedding for image captioning. In *ACM MM*.

[4] Fuhai Chen, Rongrong Ji, Xiaoshuai Sun, Yongjian Wu, and Jinsong Su. 2018. Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *CVPR*.

[5] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*.

[6] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *CVPR*.

[7] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. 2017. Towards diverse and natural image descriptions via a conditional GAN. In *ICCV*.

[8] Bo Dai and Dahua Lin. 2017. Contrastive learning for image captioning. In *NeurIPS*.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

[10] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.

[11] Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu. 2020. Normalized and geometry-aware self-attention network for image captioning. In *CVPR*.

[12] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *ICCV*.

[13] Unnat Jain, Ziyu Zhang, and Alexander G Schwing. 2017. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*.

[14] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

[15] Guang Li, Linchao Zhu, Ping Liu, and Yi Yang. 2019. Entangled transformer for image captioning. In *ICCV*.

[16] Zhuowan Li, Quan Tran, Long Mai, Zhe Lin, and Alan L Yuille. 2020. Context-aware group captioning via self-attention and contrastive features. In *CVPR*.

[17] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*.

[18] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. 2018. Discriminability objective for training descriptive captions. In *CVPR*.

[19] Ruotian Luo and Gregory Shakhnarovich. 2019. Analysis of diversity-accuracy tradeoff in image captioning. In *ICCV Workshop*.

[20] Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, Zhiheng Huang, and Alan Yuille. 2015. Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*.

[21] Yingwei Pan, Ting Yao, Yehao Li, and Tao Mei. 2020. X-linear attention networks for image captioning. In *CVPR*.

[22] Prajit Ramachandran, Niki Parmar, Ashish Vaswani, Irwan Bello, Anselm Levskaya, and Jonathon Shlens. 2019. Stand-alone self-attention in vision models. In *NeurIPS*.

[23] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.

[24] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *CVPR*.

[25] Rakshith Shetty, Marcus Rohrbach, and Lisa Anne Hendricks. 2017. Speaking the same language: Matching machine to human captions by adversarial Training. In *ICCV*.

[26] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *NeurIPS*.

[28] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. Context-aware captions from context-agnostic supervision. In *CVPR*.

[29] Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. 2019. Joint optimization for cooperative image captioning. In *CVPR*.

[30] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *CVPR*.

[31] Jiuniu Wang, Wenjia Xu, Qingzhong Wang, and Antoni B Chan. 2020. Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*.

[32] Liwei Wang, Alexander G Schwing, and Svetlana Lazebnik. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NeurIPS*.

[33] Qingzhong Wang and Antoni B Chan. 2018. CNN + CNN: Convolutional decoders for image captioning. In *CVPR Workshop*.

[34] Qingzhong Wang and Antoni B Chan. 2018. Gated hierarchical attention for image captioning. In *ACCV*.

[35] Qingzhong Wang and Antoni B Chan. 2019. Describing like humans: on diversity in image captioning. In *CVPR*.

[36] Qingzhong Wang and Antoni B Chan. 2020. Towards diverse and accurate image captions via reinforcing determinantal point process. In *TPAMI*.

[37] Qingzhong Wang, Jia Wan, and Antoni B Chan. 2020. On diversity in image captioning: Metrics and methods. In *IEEE TPAMI*.

[38] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[39] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In *WACV*.

[40] Linwei Ye, Mrigank Rochan, Zhi Liu, and Yang Wang. 2019. Cross-modal self-attention network for referring image segmentation. In *CVPR*.

[41] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. 2016. Image captioning with semantic attention. In *CVPR*.