



Group-Based Distinctive Image Captioning with Memory Difference Encoding and Attention

Jiuniu Wang¹ · Wenjia Xu² · Qingzhong Wang¹ · Antoni B. Chan¹

Received: 3 April 2023 / Accepted: 27 July 2024
© The Author(s) 2024

Abstract

Recent advances in image captioning have focused on enhancing accuracy by substantially increasing the dataset and model size. While conventional captioning models exhibit high performance on established metrics such as BLEU, CIDEr, and SPICE, the capability of captions to distinguish the target image from other similar images is under-explored. To generate distinctive captions, a few pioneers employed contrastive learning or re-weighted the ground-truth captions. However, these approaches often overlook the relationships among objects in a similar image group (e.g., items or properties within the same album or fine-grained events). In this paper, we introduce a novel approach to enhance the distinctiveness of image captions, namely Group-based Differential Distinctive Captioning Method, which visually compares each image with other images in one similar group and highlights the uniqueness of each image. In particular, we introduce a Group-based Differential Memory Attention (GDMA) module, designed to identify and emphasize object features in an image that are uniquely distinguishable within its image group, i.e., those exhibiting low similarity with objects in other images. This mechanism ensures that such unique object features are prioritized during caption generation for the image, thereby enhancing the distinctiveness of the resulting captions. To further refine this process, we select distinctive words from the ground-truth captions to guide both the language decoder and the GDMA module. Additionally, we propose a new evaluation metric, the Distinctive Word Rate (DisWordRate), to quantitatively assess caption distinctiveness. Quantitative results indicate that the proposed method significantly improves the distinctiveness of several baseline models, and achieves state-of-the-art performance on distinctiveness while not excessively sacrificing accuracy. Moreover, the results of our user study are consistent with the quantitative evaluation and demonstrate the rationality of the new metric DisWordRate.

Keywords Image caption · Vision and language · Distinctiveness · Memory attention

1 Introduction

Acquiring knowledge through multiple sensory modalities, such as vision and language, is gaining considerable interest and facilitating the development of multimodal applications. Among them, the task of image captioning has drawn much attention from both the computer vision and natural language generation communities, and steady progress has been made due to the development of vision and language techniques (Anderson et al., 2018; Hu et al., 2022; Kuo & Kira, 2023; Li et al., 2023; Xu et al., 2023). Image captioning has aided many applications, ranging from summarizing photo albums to tagging images online. A particularly noteworthy application is a mobile app¹ that vocalizes the content captured by a smartphone camera, serving as an invaluable

Communicated by Vittorio Murino.

✉ Antoni B. Chan
abchan@cityu.edu.hk

Jiuniu Wang
jiuniwang2-c@my.cityu.edu.hk

Wenjia Xu
xuwenjia@bupt.edu.cn

Qingzhong Wang
qingzwang@outlook.com

¹ Department of Computer Science, City University of Hong Kong, Kowloon, Hong Kong, China

² State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, China

¹ <https://www.apple.com/accessibility/vision/>.

resource for visually impaired people by describing the world around them.

Automatic image caption generators, while accurate, often produce generic captions for semantically similar images, missing unique details that differentiate one image from another. For example, as depicted in Fig. 1, a description of a traffic light without specifying its color provides insufficient information for visually impaired individuals to decide about crossing the street. A model that describes the distinctive contents of each image is more likely to highlight the truly useful information. We define the *distinctiveness* of a caption by its capacity to *identify and articulate the unique objects or context of the target image, thereby differentiating it from semantically similar images*. This paper aims to enhance image captioning models with the capability to produce distinctive captions.

Existing image captioning models predominantly focus on generating captions that accurately reflect 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. Research has highlighted that traditional captioning approaches, which often rely on optimizing cross-entropy loss or reinforcement rewards (typically the CIDEr score), tend to produce overly generic captions (Luo & Shakhnarovich, 2019; Wang et al., 2020a; Wang & Chan, 2019). Some efforts have been made to generate *diverse* captions to enrich the concepts by employing conditional GAN (Dai et al., 2017; Shetty et al., 2017), VAE (Jain et al., 2017; Wang et al., 2017) or reinforcement learning (Wang & Chan, 2020; Wang et al., 2020b). Several methods are proposed to improve the distinctiveness by contrastive learning (Luo et al., 2018; Dai & Lin, 2017; Li et al., 2020), where they either aggregate the contrastive image features with the target image feature, or apply the contrastive loss to suppress the estimated conditional probabilities of mismatched image-caption pairs (Dai & Lin, 2017; Luo et al., 2018). However, the distractors are either a group of images with scene graphs that partially overlaps with the target image (Li et al., 2020), or randomly selected unmatched image-caption pairs (Dai & Lin, 2017; Luo et al., 2018), which are easy to distinguish. Liu et al. (2018) and Vered et al. (2019) 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 (Wang et al., 2020a), which hurts caption quality.

In this work, to generate distinctive captions, we consider the hard negatives, i.e., *similar images* that generally share similar semantics with the target image, and push the generated captions to clearly show the difference between the target image and these hard negative images. For instance, as shown in Fig. 1, the generated captions should specify the dif-

ferent aspects of the target image (e.g., different light colors and context) compared with other images that share similar semantics (i.e., images of traffic lights). To this end, we propose a differential distinctive memory attention module that puts high attention on distinctive objects detected in the target image but not in the similar images, and low attention on objects that are common among the target and similar images. Specifically, object features in the target image with low similarity to object features in similar images are considered more distinctive, and thus receive a higher attention value. Our proposed attention mechanism is a plug-and-play module that works to extend existing transformer-based captioning models. Moreover, we further propose two loss functions to facilitate the training, which encourages the model to focus its captions on the distinct image regions: (1) the memory classification loss predicts the distinctive words from the image features; (2) the weighted distinctive loss encourages the captioning model to predict distinctive words describing the unique image regions and gives higher weights to the distinctive words that are highly related to the image.

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

(1) We propose a Group-based Differential Distinctive Captioning Method (DifDisCap), which constructs memory features from object regions, weighted by their distinctiveness within an image group, to generate captions that uniquely describe each image in the group. Specifically, our model employs a memory difference encoding technique designed to accentuate the feature differences between the target image and its corresponding group of similar images.

(2) To ensure the weighted memory contains distinctive object information, we introduce two novel distinctive loss functions. These functions are supervised by the occurrence of distinctive words found in the ground-truth captions, thereby reinforcing the emphasis on unique object details within the images.

(3) We have carried out comprehensive experiments and user studies, which demonstrate that our proposed model is able to generate distinctive captions. Furthermore, our model emphasizes the unique regions of each image, enhancing the interpretability.

The preliminary conference version of our work has been published in Wang et al. (2021). This journal article extends our preliminary work in four aspects. First, we propose a memory difference encoding to emphasize the feature difference between the target image and the similar image group. Second, we propose Indicated Training (IndTrain) to apply our memory attention only to those distinctive GT captions, which makes distinctive training more effective. Third, in order to emphasize those distinctive words highly related to the target image and discards the unrelated words, we measure the text-image *relatedness* with a pretrained multi-modal network [i.e., CLIP (Radford et al., 2021)] and weight the distinctive word loss (Wang et al., 2021) accord-

ing to this *relatedness*. Moreover, using the newly introduced DifDisCap method, we present new state-of-the-art results on several baseline models.

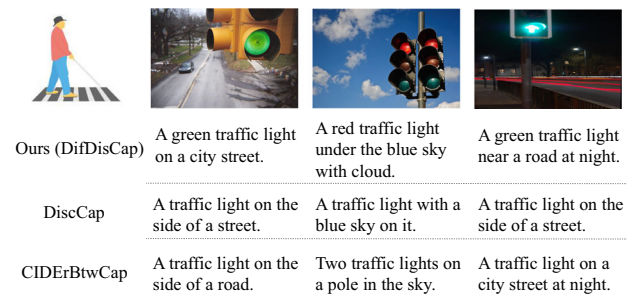
The remainder of the paper is organized as follows. In Sect. 2, we present related works, including image captioning models and metrics. In Sect. 3, we introduce our Group-based Differential Distinctive Captioning method. The experimental setting and quantitative results are presented in Sect. 4, and the user study and qualitative results are presented in Sects. 5 and 6. Finally, we conclude the paper in Sect. 7.

2 Related Work

2.1 Image Captioning

Image captioning bridges two domains—images and texts. Classical approaches usually extract image representations using a convolutional neural network (CNN), then feed them into a recurrent neural network (RNN) and output sequences of words (Karpathy and Fei-Fei, 2015; Mao et al., 2015; Vinyals et al., 2015). Recent advances mainly focus on improving the image encoder and the language decoder. For instance, Anderson et al. (2018) propose bottom-up features, which are extracted by a pre-trained Fast R-CNN (Ren et al., 2015) 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, some works (Aneja et al., 2018; Wang & Chan, 2018a,b) utilize CNNs since LSTMs cannot be trained in a parallel manner. More recently, some approaches (Cornia et al., 2020; Li et al., 2019) adopt transformer-based networks with multi-head attention to generate captions, which mitigates the long-term dependency problem in LSTMs and significantly improves the performance of image captioning. Recent advances usually optimize the network with a two-stage training procedure, where they pre-train the model with word-level cross-entropy loss (XE) and then fine-tune with reinforcement learning (RL) using the CIDEr score (Vedantam et al., 2015) as the reward. Also, some work (Wang et al., 2020c) introduces similar images to improve the accuracy of the generated captions. However, as pointed out in Wang et al. (2020a), Dai and Lin (2017), Dai et al. (2017), 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 lack distinctiveness. In contrast, our work gives higher attention to the image regions that are different from other similar images, leading to more distinctive captions. We further propose weighted distinctive loss to encourage the model to predict distinctive words.

The above models typically focus on improving the accuracy of generated captions. Recently, various works aim to expand on traditional image captioning by better utilizing



Ours (DifDisCap)	A green traffic light on a city street.	A red traffic light under the blue sky with cloud.	A green traffic light near a road at night.
DiscCap	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	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.

Fig. 1 Our model generates distinctive captions that can distinguish the target image from other similar images. Compared to current distinctive image captioning models such as DiscCap (Luo et al., 2018) and CIDErBtwCap (Wang et al., 2020a), our captions can specify the important details, e.g., the color and the context of the traffic light, which can help a visually-impaired person to cross the street

cross-domain and linguistic knowledge, linking words to objects in the image, and addressing dataset bias. Zhao et al. (2020) and Yuan et al. (2022) propose cross-domain image captioning models that are trained on a source domain and generalized to other domains, to alleviate the demands for massive data in target domains. Moreover, Chen et al. (2022) propose to adapt the linguistic knowledge from large pre-trained language models such as GPT (Radford et al., 2018) to image captioning models. LEMON (Hu et al., 2022), mPLUG (Xu et al., 2023) and BLIP-2 (Li et al., 2023) leverage the visual and semantic information from the vision-language pretrained model to boost the performance of image captioning. Apart from cross-domain adaptation, other image captioning models aim to ground objects in images (Huang et al., 2020; Zhou et al., 2020) and 3D scenes (Cai et al., 2022). Jiang et al. (2022) propose to implicitly link the words in captions and the informative regions on images with a cluster-based grounding model. Furthermore, Kuo and Kira (2022) combine attribute detection with image captioning to achieve accurate attention localization. Besides, understanding and quantifying the social biases in image captioning, e.g., gender bias (Hirota et al., 2022), racial bias (Zhao et al., 2021) and emotional bias (Mohamed et al., 2022), can inspire new directions for mitigating the biases found in image captioning datasets and evoke models with less bias.

More relevant to our work are the recent works on group-based image captioning (Li et al., 2020; Vedantam et al., 2017; Chen et al., 2018), where a group of images is utilized as context when generating captions. Vedantam et al. (2017) generate sentences that describe an image in the context of other images from closely related categories. Chen et al. (2018) summarize the unique information of the target images contrasting to other reference images, and Li et al. (2020) emphasize 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 and object-level features. Both Chen et al. (2018) and Vedantam et al. (2017) 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.

To construct groups of images that share similar semantics, our method initially involves randomly selecting an image as the target. Subsequently, we retrieve its nearest images using a visual-semantic retrieval model. Visual-semantic retrieval models, predominantly based on a one-to-one mapping of instances into a shared embedding space, are well-suited to retrieve images with similar characteristics. One widely-adopted method involves maximizing the correlation between related instances within a shared embedding space, e.g., using canonical correlation analysis to maximize the correlation between images and text (Rasiwasia et al., 2010; Yan & Mikołajczyk, 2015). Another popular approach is based on triplet ranking, which aims to ensure that the distance between positive image-text pairs is smaller than that between negative pairs. Drawing inspiration from hard negative mining, VSE++ (Faghri et al., 2018) leverages maximum violating negative pairs to enhance performance. More recently, CLIP (Radford et al., 2021) introduced a visual-language model employing a contrastive learning objective across various image-text pairs, while ALIGN (Jia et al., 2021) expands this methodology by incorporating noisy text descriptions on a larger scale. Our work uses both VSE++ and CLIP for the construction of similar image groups and compares their performance in the experiments.

Different from traditional image captioning tasks that are trained and tested on datasets from the same domain, zero-shot image captioning aims to develop robust models that can generate captions for images from unseen environments or contain new concepts. Recent advances achieve zero-shot image captioning by combining large Vision-Language Models (VLMs) and Large Language Models (LLMs) (Fei et al., 2023; Tewel et al., 2022; Zeng et al., 2023). ZeroCap (Tewel et al., 2022) employs CLIP (Radford et al., 2021) together with the GPT-2 (Radford et al., 2019) language model to generate a textual description of a given image at inference time, without any further training or tuning step. To tackle the object hallucination problem in ZeroCap, ViECap (Fei et al., 2023) proposes a transferable decoding model that incorporates entity-aware decoding and prompts to guide LLMs' attention toward the visual entities present in the image. A retrieval-augmented zero-shot captioning model (Kim et al., 2023) utilizes external contextual knowledge complementary to the knowledge in the original model, and consequently helps the captioner to achieve higher accuracy. To rigorously assess the capabilities of image captioning models in a zero-shot context, Kim et al. (2023) introduced a comprehensive

evaluation dataset, comprising 26,000 image-caption pairs for testing without accompanying training data. Our paper primarily aims to improve the distinctiveness of supervised image captioning models by guiding model attention toward unique objects within images. However, the core concept of this paper has the potential to be combined with VLM-based zero-shot image captioners, enhancing the distinctiveness of captions.

2.2 Distinctive and Diverse Image Captioning

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 (2017) promote the distinctiveness of an image caption by contrastive learning. The model is trained to give a high probability to the GT image-caption pair and a low probability to a randomly sampled negative pair. Luo et al. (2018) and Liu et al. (2018) take 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 contrast, in our work, we consider *hard negative images* 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, Wang et al. (2020a) propose to give higher weight to the distinctive GT captions during model training. Chen et al. (2018) model the diversity and relevance among positive and negative image pairs in a language model, with the help of a visual parsing tree (Chen et al., 2017a). 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.

Additionally, several works aim to improve the diversity of generated captions, where the model can generate a set of different captions for the same image. One group of works is based on conditional GANs (Dai et al., 2017; Shetty et al., 2017) and auto-encoders (Aneja et al., 2019; Mahajan & Roth, 2020). However, promoting the variability of generated captions may not improve the distinctiveness (Wang et al., 2022). For instance, using synonyms and changing the word orders in generated captions encourage diversity in syntax and word usage, but do not introduce distinctiveness information.

Our work is relevant to the object detection models that associate object regions in images with semantic labels.

Deep learning detectors can be classified into two distinct categories: (1) two-stage detectors, which encompass both region proposal and bounding box regression modules (He et al., 2015), e.g., Faster R-CNN (Ren et al., 2015); (2) one-stage detectors that divide the image into regions and predict bounding boxes and probabilities for each region simultaneously, e.g., YOLO (Redmon et al., 2016). Our work leverages spatial image features derived from Faster R-CNN to construct the memory features for object regions. These vectors are weighted based on their distinctiveness within the image group, enabling the generation of distinctive captions.

2.3 Metrics for Image Captioning

In recent years, many metrics have been proposed to assess the performance of a captioning model, most of which evaluate the fluency and accuracy, e.g., BLEU (Papineni et al., 2002), CIDEr (Vedantam et al., 2015), SPICE (Anderson et al., 2016), and METEOR (Denkowski & Lavie, 2014). However, these traditional metrics normally evaluate the word-level and phrase-level similarity between generated captions and the GT captions, instead of considering the semantic similarity (Stefanini et al., 2022). Furthermore, the captioning models trained with reinforcement learning to optimize these metrics (Cornia et al., 2020; Luo et al., 2018) tend to generate over-generic captions instead of pointing out the distinctive details in each image (Wang et al., 2022). Some related works propose diversity metrics to evaluate the corpus-level diversity. For instance, Van Miltenburg et al. (2018) propose a metric to quantify the number of unique words in the captions, and further calculates the number of unique bigram or unigrams appearing in the generated captions. Wang et al. (2017) measure the percentage of novel sentences that do not appear in the training set. Wang et al. (2020b) employ latent semantic analysis to quantify the semantic diversity of generated captions. These metrics measure the variability of generated words and phrases, but cannot tell if the generated captions can distinguish the target image from other similar ones, i.e., the distinctiveness.

The first metric for distinctiveness was the retrieval method, which employs a pretrained semantic-visual embeddings model VSE++ (Faghri et al., 2018) to retrieve the target image with the generated captions and reports the Recall at k . Ideally, a distinctive caption should retrieve the correct image as the first item in the retrieval list. Furthermore, Wang et al. (2022) consider that distinct captions are less similar to other captions, where the similarity is measured by the CIDEr between generated captions and the GT captions of other similar images. While these metrics only consider the sentence level distinctiveness, we argue that captions describing the unique details of target images usually contain distinct words. In this paper, we propose two novel metrics that consider both word-level and sentence-level distinctness.

Recent advances in image captioning evaluation metrics have focused on zero-shot and non-reference captioning metrics. BERTScore (Zhang et al., 2019) introduces an automated metric for text generation that calculates similarity scores by utilizing contextual embeddings to compare tokens in candidate sentences against those in reference sentences. Extending this concept, ViLBERTScore (Lee et al., 2020) enhances image caption evaluation by incorporating both textual and visual information, generating image-conditioned embeddings for each token. To address the limitation of BERTScore, UMIC (Lee et al., 2021), PAC-S (Sarto et al., 2023), and CLIP-S (Hessel et al., 2021) propose methods to evaluate image captions without reference captions. Furthermore, to specifically assess zero-shot captioning models, V-METEOR (Demirel & Cinbis, 2022) was proposed to evaluate the visual and textual content of generated sentences independently. While the aforementioned metrics aim to evaluate captions without references, our paper diverges in focus, aiming to assess the distinctiveness of the generated captions.

2.4 Attention Mechanisms

Attention mechanisms apply visual attention to different image regions when predicting words at each time step, and have been widely utilized in image captioning (Xu et al., 2015; You et al., 2016; Chen et al., 2017b; Guo et al., 2020; Pan et al., 2020). For instance, You et al. (2016) adopt semantic attention to focus on the semantic attributes in the image. Anderson et al. (2018) exploit object-level attention with bottom-up attention, then associates the output sequences with salient image regions via a top-down mechanism. More recently, self-attention networks introduced by Transformers (Vaswani et al., 2017) are widely adapted in both language and vision tasks (Dosovitskiy et al., 2020; Luo et al., 2018; Ramachandran et al., 2019; Su et al., 2020; Ye et al., 2019; Yang et al., 2020). Guo et al. (2020) normalize the self-attention module in the transformer to solve the internal covariate shift. Huang et al. (2019) weight 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. (2020) migrate 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 proposing learnable memory attention that highlights prominent *object-level* R-CNN features with distinct semantics among similar images.

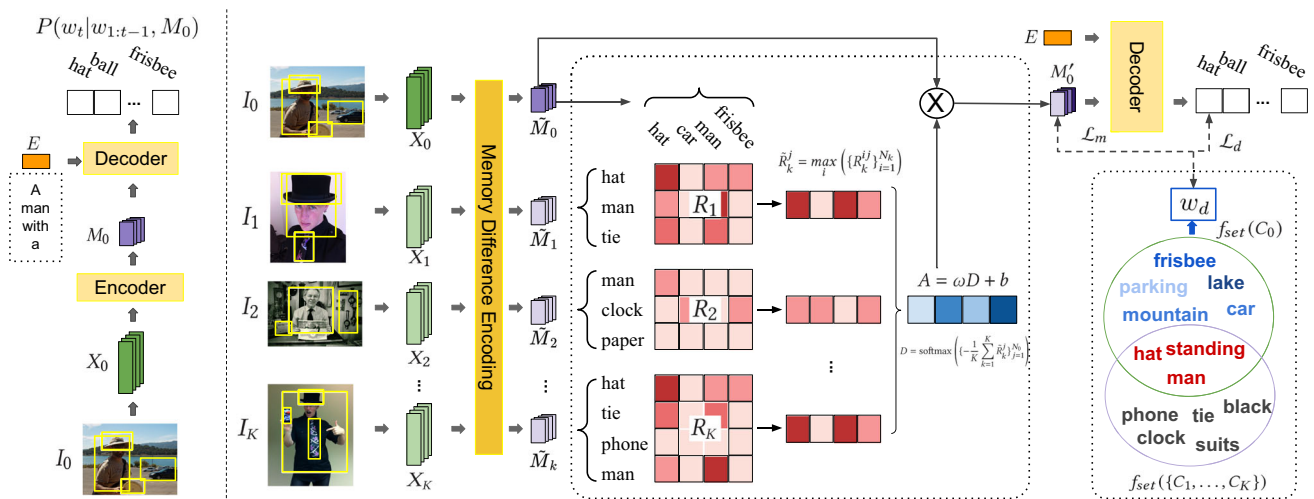


Fig. 2 Left: the standard transformer-based captioning model, where the target image features X_0 are the region-based visual features extracted via RoI pooling from Fast R-CNN. Right: our Group-based Differential Distinctive Captioning method (DifDisCap), which consists of a group-based differential memory attention (GDMA) module that weights the memory features according to their similarity with

other similar images. The words in blue are distinctive words w_d , and the words with higher relatedness are marked in the darker color. Our model takes a group of images as input, and outputs one caption for each image. Only one target memory M'_0 , one decoder, and one output caption are shown here to reduce clutter (Color figure online)

3 Methodology

We present the framework of our proposed Group-based Differential Distinctive Captioning method (DifDisCap) in Fig. 2. Our model aims to generate distinctive captions for each image within a group of semantically similar images. Given an image group with $K + 1$ images, denoted as $\{I_0, I_1, \dots, I_K\}$, DifDisCap 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 (i.e., I_0) compared to other K images (i.e., 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 iteratively. In Fig. 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, comprising semantically similar images, then we employ the proposed Group-based Differential Memory Attention (GDMA) module 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 groups were first introduced in Wang et al. (2020a) 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 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 the visual-semantic retrieval model VSE++ (Faghri et al., 2018), as in Wang et al. (2020a). 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 similar images.

(2) Due to the non-uniform distribution of training images, the images sharing similar semantic meanings 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 in that, so they will not be selected when constructing other groups in the epoch. 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.

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.

3.2 Group-Based Differential Distinctive Captioning Method

Here we introduce the group-based differential distinctive captioning method (DifDisCap), and how we incorporate the Group-based Differential Memory Attention (GDMA) module that encourages the model to generate distinctive captions. Notably, the GDMA can serve as a plug-and-play module 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 built on a transformer-based architecture (Cornia et al., 2020), as illustrated in Fig. 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 architectures, the *Encoder* and *Decoder* are both composed of several multi-head attention and MLP layers.

In our work, we take the bottom-up features (Anderson et al., 2018) extracted by Fast R-CNN (Ren et al., 2015) as the input. Given an image I , let $X = \{x^i\}_{i=1}^N$ denote the object features, where N is the number of region proposals and $x^i \in \mathbb{R}^d$ is the feature vector for the i -th proposal. The output of the l -th encoder layer is calculated as follows:

$$O_l^{att} = \text{LN}(X_{l-1} + \text{MH}(\mathbf{W}_q X_{l-1}, \mathbf{W}_k X_{l-1}, \mathbf{W}_v X_{l-1})), \quad (1)$$

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

where $\text{LN}(\cdot)$ denotes layer normalization, $\text{MLP}(\cdot)$ denotes a multi-layer perceptron, and $\text{MH}(\cdot)$ represents the multi-head attention layer. $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v$ are learnable parameters.

The *Encoder* turns features X into memory features $M = \{m^i\}_{i=1}^N$, where $m^i \in \mathbb{R}^{d_m}$ encodes the information from the i -th object proposal x^i , 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 $\{w_1, \dots, w_{t-1}\}$, 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 Differential Memory Attention (GDMA)

The goal of group-based differential memory attention is to highlight the distinctive features of the target image that do not appear in other similar images. For instance, in Fig. 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

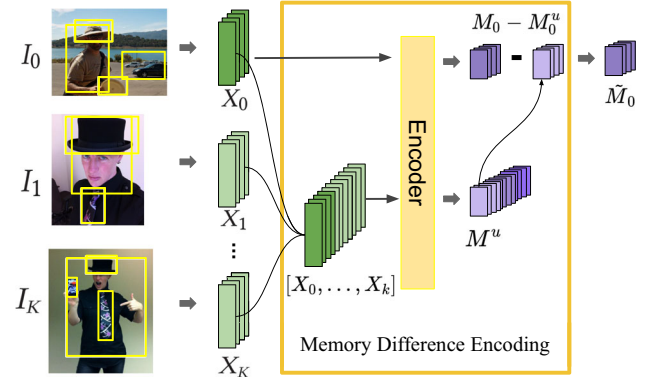


Fig. 3 The architecture of the Memory Difference Encoding module. The union memory vector M^u encodes the fused information from images I_0 to I_K , and \tilde{M}_0 encodes the difference between the target image I_0 and other similar images

from other images. However, the standard captioning model in Fig. 2(left) cannot highlight those objects, since each memory vector m_0^i for different image regions is treated equally when fed into the *Decoder*. Here we propose memory difference encoding and distinctive attention to highlight the distinct regions by assigning higher weights to their corresponding memory features.

Memory Difference Encoding Existing methods usually predict the caption \hat{c} only according to the image I_0 , while our model also consider $\{I_0, \dots, I_K\}$. The memory features M_0 is obtained via a feature encoder f_{en} :

$$M_0 = f_{en}(X_0). \quad (3)$$

In this work, we need the memory features \tilde{M}_0 for image I_0 to contain the difference between I_0 and other similar images (i.e., I_1, \dots, I_K). We therefore propose a Memory Difference Encoding module as shown in Fig. 3. Instead of simply encoding X_0 to X_K separately, we first concatenate the $K + 1$ images feature (i.e., X_0 to X_K) and apply the encoder to obtain a union memory M^u ,

$$M^u = [M_0^u, \dots, M_K^u] = f_{en}([X_0, \dots, X_K]). \quad (4)$$

The union memory vector M_0^u encodes the fused information of image X_0 and other similar images feature (X_1 to X_K), where M_0^u is the part of M^u corresponding to X_0 . Next, we extract the difference between M_0 and the union memory vector M_0^u as M_0^d :

$$\tilde{M}_0 = M_0 - M_0^u, \quad (5)$$

and employ the difference encoding \tilde{M}_0 to generate the caption for I_0 .

Computing distinctive attention In this work, we aim to give higher 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 the group-based differential memory attention (GDMA) module [see Fig. 2(right)], where the attention weight for each object region is obtained by calculating the distinctiveness of its memory vector \tilde{m}_0^i . Then we encourage the model to generate distinctive words associated with the unique object regions. Specifically, the GDMA produces distinctive attention $A = \{a_i\}_{i=1}^{N_0} \in \mathbb{R}^{N_0}$ for differential memory features $\tilde{M}_0 = \{\tilde{m}_0^1, \dots, \tilde{m}_0^{N_0}\}$. When generating captions, instead of using \tilde{M}_0 , a weighted target memory is fed into the decoder:

$$M'_0 = \{a_i \cdot \tilde{m}_0^i\}_{i=1}^{N_0}. \quad (6)$$

To compute the distinctive attention, we need to compare the objects in the target image with those in similar images. As shown in Fig. 2(right), the target image I_0 and its similar images $\{I_k\}_{k=1}^K$ are transferred into memory features $\tilde{M}_0 = \{\tilde{m}_0^j\}_{j=1}^{N_0}$ and $\tilde{M}_k = \{\tilde{m}_k^i\}_{i=1}^{N_k}$ ($k = 1, \dots, K$), via the image encoder, where N_k denotes the number of objects in the k -th image. The GDMA first measures the similarity $R_k \in \mathbb{R}^{N_k \times N_0}$ of each target memory vector \tilde{m}_0^j and each memory vector \tilde{m}_k^i in similar images via cosine similarity:

$$R_k^{ij} = \cos(\tilde{m}_k^i, \tilde{m}_0^j), \quad (7)$$

where $\tilde{m}_0^j \in \mathbb{R}^{d_m}$ is the j -th vector in \tilde{M}_0 (e.g., memory features for *hat*, *car*, *man* and *frisbee* in Fig. 2), and \tilde{m}_k^i is the i -th vector in \tilde{M}_k (e.g., memory features for *hat*, *man*, and *tie* from \tilde{M}_k in Fig. 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 Fig. 2(right), *hat* is less distinctive since it occurs in multiple images, while *car* is a unique object that only appear in target image. To summarize the similarity matrix, we compute an object-image similarity map $\tilde{R}_k \in \mathbb{R}^{N_0}$ as

$$\tilde{R}_k^j = \max_{i \in \{1, \dots, N_k\}} R_k^{ij}, \quad (8)$$

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 we define the raw distinctiveness score for the j -th region in I_0 , corresponding to memory vector \tilde{m}_0^j , as the negative average of its object-image similarity maps with the similar images,

$$d_j = -\frac{1}{K} \sum_{k=1}^K \tilde{R}_k^j, \quad j \in \{1, \dots, N_0\}. \quad (9)$$

Here, higher scores indicate higher distinctiveness, i.e., lower average similarity to other similar images. The raw distinctiveness scores are normalized by applying the softmax function, yielding the final distinctiveness scores $D \in \mathbb{R}^{N_0}$ for the memory features $\{\tilde{m}_0^j\}$,

$$D = [D_1, \dots, D_{N_0}] = \text{softmax}([d_1, \dots, d_{N_0}]), \quad (10)$$

Note that the values of D range in $[0, 1]$ due to the softmax function.

Indicated Training In the training set, each target image I_0 has several ground truth (GT) captions. Not all these GT captions are distinctive, and thus attending to unique distinctive features may confuse the training process. We therefore use a distinctive metric [e.g., CIDErBtw (Wang et al., 2020a)] to divide the ground truth captions into *distinctive* ones and *common* ones.

For those GT captions indicated as *distinctive*, the distinctive attention weights $A = [a_1, \dots, a_{N_0}]$ for the target memory features in (6) are calculated as:

$$A = \omega D + b, \quad (11)$$

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 values in A are non-negative.

For those GT captions indicated as *common*, we let $A = \mathbf{1}$ so that the memory features are all considered equally in Eq. 6.

3.3 Loss Functions

Two typical loss functions for training image captioning are cross-entropy loss and reinforcement loss. The reinforcement loss uses the average CIDEr with the GT captions of the image for supervision, which may encourage the generated captions to mimic “average” GT caption, 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 weighted distinctive loss (WeiDisLoss) and memory classification loss (MemClsLoss) for training our GDMA module.

3.3.1 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

$$\mathcal{L}_{xe} = - \sum_{t=1}^T \log P(w_t | w_{1:t-1}, M'_0), \quad (12)$$

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 features M'_0 , as generated by the caption *Decoder*.

3.3.2 Reinforcement Learning Loss

Following Rennie et al. (2017), we apply reinforcement learning to further improve the accuracy of our trained network using the loss:

$$\mathcal{L}_r = -E_{\hat{c} \sim p(c|I)} \left[\frac{1}{d_c} \sum_{i=1}^{d_c} g(\hat{c}, C_0^i) \right], \quad (13)$$

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.

3.3.3 Weighted Distinctive Loss (WeiDisLoss)

We propose a weighted distinctive loss to encourage the caption model to focus on the distinctive words that appear in captions C_0 of the target image, but not in captions $\{C_1, \dots, C_K\}$ of similar images. We define the distinctive word set Ω for I_0 as

$$\Omega = f_{set}(C_0) - f_{set}(\{C_1, \dots, C_K\}), \quad (14)$$

where $f_{set}(\cdot)$ denotes the function that converts the sentence into a word set, and “ $-$ ” here means set subtraction.

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

$$\mathcal{L}_d = - \sum_{t=1}^T \sum_{i=1}^{|\Omega|} \lambda_{\omega_i} \log P(w_t = \omega_i | w_{1:t-1}, M'_0), \quad (15)$$

where ω_i denotes the i -th distinctive word in Ω , and $P(w_t = \omega_i | w_{1:t-1}, M'_0)$ denotes the probability of predicting word ω_i as the t -th word in sentence. $|\Omega|$ is the number of words in Ω , and T is the length of the sentence.

In practice, due to the personalized language preference of each annotator, not every word in Ω is highly related to

the image I_0 , and those unrelated words would distract the captioning model. We therefore apply a weight λ_{ω_k} to each term in the WeiDisLoss in (15). The weight λ_{ω_k} measures the *relatedness* of the k -th distinct word ω_k with the target image I_0 . In detail, ω_k is placed into a sentence c_{ω_k} with the template “this picture includes ω_k ”, and the relatedness λ_{ω_k} is calculated as

$$\hat{\lambda}_{\omega_k} = \theta(c_{\omega_k}) \cdot \phi(I_0), \quad (16)$$

$$\lambda_{\omega_k} = \frac{\hat{\lambda}_{\omega_k}}{\max_k \hat{\lambda}_{\omega_k}} \quad (17)$$

where $\theta(\cdot)$ and $\phi(\cdot)$ denote the sentence embedding and image embedding of a multimodal embedding model [e.g., CLIP (Radford et al., 2021)]. Thus, distinctive words that are more related to the target image (according to the embedding model) will have higher weights in the loss.

Since CLIP is weak at identifying small or particular objects, employing CLIP to assess the relevance of these distinct words to the target image might result in lower weighting for smaller objects. This concern can be partially addressed by combining with other loss functions that focus on small objects. For example, memory classification loss introduced in Sect. 3.3.4 encourages the model to pay attention to all objects denoted by distinct words. Additionally, both the cross-entropy loss and the reinforcement learning loss encourage the model to generate captions following human supervision, regardless of the object sizes within these captions.

3.3.4 Memory Classification Loss (MemClsLoss)

In order to generate distinctive captions, the *Decoder* requires the GDMA to produce memory contents containing distinctive concepts. However, the supervision of the GDMA through the *Decoder* could be too weak, which may allow the GDMA 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 GDMA, we introduce an *auxiliary classification task* that predicts the distinctive words from the weighted memory features M'_0 of the GDMA,

$$P_M = f_{MC}(M'_0), \quad (18)$$

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

$$\mathcal{L}_m = - \sum_{k=1}^{|\Omega|} \lambda_{\omega_k} \log(P_{M, \omega_k}), \quad (19)$$

Algorithm 1 The training procedure of DifDisCap in each step

Require: A similar image group I_0, \dots, I_K with captions C_0, \dots, C_K

Ensure: The final loss \mathcal{L} of this similar image group to optimize the *Encoder and Decoder*

- 1: Encode the image group $\{I_0, \dots, I_K\}$ into $\{M_0, \dots, M_K\}$, where the memory of the k -th image is $M_k = \{m_k^i\}_{i=1}^{N_k}$
- 2: Encode the whole image group into $[M_0^u, \dots, M_K^u] = f_{en}([X_0, \dots, X_K])$
- 3: **for** $k \leftarrow 0$ to K **do**
- 4: Calculate the memory difference $\tilde{M}_k = M_k - M_k^u$.
- 5: Calculate the distinctive attention A as in (11) for the distinctive ground-truth captions, where $A = \{a_i\}_{i=1}^{N_k} \in \mathbb{R}^{N_k}$. Let $A = \mathbf{1}$ for the common ground-truth captions.
- 6: Calculate weighted target memory: $M'_k = \{a_i \cdot \tilde{m}_k^i\}_{i=1}^{N_k}$
- 7: Calculate distinctive word set $\Omega_k = f_{set}(C_k) - f_{set}(\{C_j\}_{j \neq k})$
- 8: Decode M'_k as probability of generated words $\{P_i\}_{i=1}^T$
- 9: Classify M'_k as possible words $P_M \leftarrow f_{MC}(M'_k)$
- 10: Calculate each loss for the k -th image, including:
 - 11: Cross-entropy loss \mathcal{L}_{xe} with P_t and C_k ,
 - 12: Reinforcement learning loss \mathcal{L}_c with P_t and C_k ,
 - 13: Weighted distinctive word loss \mathcal{L}_d with P_t and Ω_k ,
 - 14: Memory classification loss \mathcal{L}_m with P_M and Ω_k .
- 15: Get the loss for the k -th image $\mathcal{L}_k = \alpha_c \mathcal{L}_{xe} + \alpha_r \mathcal{L}_r + \alpha_d \mathcal{L}_d + \alpha_m \mathcal{L}_m$
- 16: **end for**
- 17: Accumulate the loss $\mathcal{L} = \sum_{k=0}^K \mathcal{L}_k$

where P_{M, ω_k} is the predicted probability of the k -th distinctive word.

3.3.5 The Final Loss

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

$$\mathcal{L} = \alpha_c \mathcal{L}_{xe} + \alpha_r \mathcal{L}_r + \alpha_d \mathcal{L}_d + \alpha_m \mathcal{L}_m, \quad (20)$$

where $\{\alpha_c, \alpha_r, \alpha_d, \alpha_m\}$ are hyper-parameters for their respective losses. The training procedure has two stages following Luo et al. (2018). 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. We show the details for processing one image group in Algorithm 1.

4 Experiments

In this section, we first introduce the implementation details and dataset preparation, then we quantitatively evaluate the

effectiveness of our model through an ablation study and a comparison with other state-of-the-art models.

4.1 Implementation Details

Following Anderson et al. (2018), we use the spatial image features extracted from Fast R-CNN (Ren et al., 2015) with dimension $d = 2048$. Each image usually contains around 50 object region proposals, i.e. $N_0 \approx 50$. Each object proposal has a corresponding memory vector with dimension $d_m = 512$. We set $K = 5$ for constructing similar image groups. The values in learned distinctive attention A are mostly in the range of (0.4, 0.9). To verify the effectiveness of our model, we conduct experiments on four baseline methods [i.e., Transformer (Luo et al., 2018), M²Transformer (Cornia et al., 2020), Transformer + SCST (Luo et al., 2018) and M²Transformer + SCST (Cornia et al., 2020)]. Our experimental settings (e.g., data preprocessing and vocabulary construction) follow these baseline models. The models with DifDisCap have similar training parameters with their baseline models, where Transformer based models have 55M parameters and M²Transformer based models have 42M parameters. Furthermore, we train 15 epochs for cross-entropy loss, and an additional 15 epochs for reinforcement learning loss if with SCST. Transformer based models are with batch size 10, and M²Transformer based models are with batch size 50. We apply our GDMA module to the Transformer model and all three layers in M²Transformer 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,³ which contains 12,387 images, and each image has 5 human annotations. Following Anderson et al. (2018), we utilize the Karpathy split (Karpathy and Fei-Fei, 2015) to divide the dataset into three sets—5000 images for validation, 5000 images for testing, and the rest for training. When constructing similar image groups for the test set, we adopt the same group split as Wang et al. (2020a) for a fair comparison.

Note that we did not run our main experiments on the online MSCOCO test set since we need to evaluate distinctiveness using the below metrics, which are not available from the MSCOCO submission server. Please see the Appendix for

³ <https://cocodataset.org/#download>.

the evaluation on the online MSCOCO test set using standard metrics, e.g., CIDEr.

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 (Wang et al., 2020a) 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 values, e.g., a random caption that has lower overlap with the GT captions will be considered 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 \hat{c} with GT captions of images in the same group as

$$s_k = \frac{1}{d_c} \sum_{i=1}^{d_c} g(\hat{c}, C_k^i), \quad (21)$$

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 sort the CIDEr values $\{s_k\}_{k=0}^K$ in descending order, and use the rank r of s_0 to measure the distinctiveness of \hat{c} . The best rank is $r = 1$, indicating the generated caption \hat{c} is more similar to its GT captions the 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:

$$\text{DisWordRate} = \max_i \frac{|\Omega \cap \hat{c} \cap C_0^i|}{|\Omega \cap C_0^i|}, \quad i = 1, \dots, d_c, \quad (22)$$

where Ω is the set of distinctive words for image I_0 , d_c is the number of sentence in C_0 , $|\Omega \cap \hat{c} \cap C_0^i|$ represents the number of words in Ω that appear in both \hat{c} and C_0^i . Thus, DisWordRate reflects the percentage of distinctive words in the generated captions.

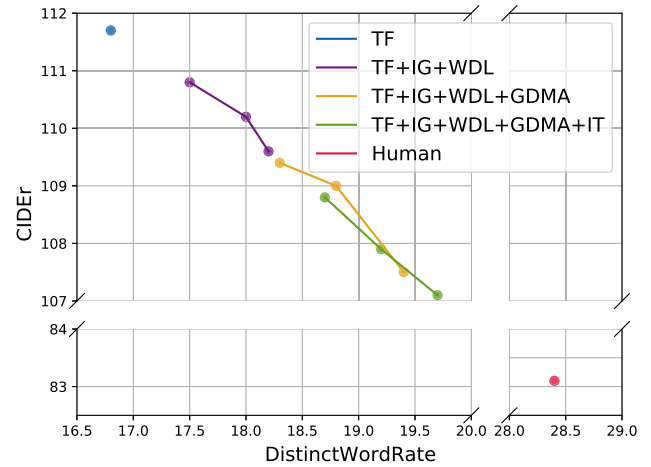


Fig. 4 The trade-off between accuracy (CIDEr) and distinctiveness (DisWordRate): human-annotated GT captions (Human), baseline Transformer model (TF) (Luo et al., 2018), and three variants of our model using various components: image group based training (Image-Group, IG), weighted distinctive loss (WeiDisLoss, WDL), group-based differential memory attention (GDMA), and Indicated Training (Ind-Train, IT). For our models, we show three training stages (at different epochs), which demonstrates the trade-off between accuracy and distinctiveness during training

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 the ablation studies of our model.

4.3.1 Comparison with the State-of-the-Art

We compare our model with two groups of state-of-the-art models: (a) FC (Rennie et al., 2017), Att2in (Rennie et al., 2017), UpDown (Anderson et al., 2018), AoANet (Huang et al., 2019), BLIP-2 (OPT-2.7B) (Li et al., 2023) that aim to generate captions with high accuracy; (b) DiscCap (Luo et al., 2018), CL-Cap (Dai & Lin, 2017), and CIDErBtwCap (Wang et al., 2020a) 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 17.3 percent (from 16.8 to 19.7%) and reduce the CIDErBtw by 5.5 percent (from 74.8 to 70.7) for Transformer, while only sacrificing the CIDEr by 4.1 percent. Second, in terms of distinctiveness, models trained with cross-entropy loss tend to perform better than models trained with SCST (Rennie et al., 2017). For example, we achieve the highest DisWordRate with Transformer + DifDisCap at 19.7%. M²Transformer + DifDisCap also achieves higher DisWordRate than M²Transformer +

Table 1 Comparison of caption distinctiveness and accuracy on MSCOCO test split: **DisWordRate**, **CIDErRank**, and **CIDErBtw** measure the distinctiveness, while **CIDEr** and **BLEU-3** and **BLEU-4** measure the accuracy

Method	DisWordRate (%) \uparrow	CIDErRank \downarrow	CIDErBtw \downarrow	CIDEr \uparrow	BLEU3 \uparrow	BLEU4 \uparrow
Transformer (Luo et al., 2018)	16.8	2.47	74.8	111.7	45.1	34.0
+ DifDisCap (ours)	19.7	2.40	70.7	107.1	43.3	32.6
M ² -Transformer (Cornia et al., 2020)*	16.4	2.52	76.8	111.8	45.2	34.7
+ DifDisCap (ours)	18.8	2.43	72.4	109.8	44.3	33.5
Transformer + SCST (Luo et al., 2018)	14.7	2.38	83.2	127.6	51.3	38.9
+ DifDisCap (ours)	17.0	2.34	81.5	126.9	50.6	38.4
M ² -Transformer + SCST (Cornia et al., 2020)*	17.3	2.38	82.9	128.9	50.6	38.7
+ DifDisCap (ours)	18.7	2.30	80.1	127.2	49.9	38.2
FC (Rennie et al., 2017)	6.5	3.03	89.7	102.7	43.2	31.2
Att2in (Rennie et al., 2017)	10.8	2.65	88.0	116.7	48.0	35.5
UpDown (Anderson et al., 2018)	12.9	2.55	86.7	121.5	49.2	36.8
AoANet (Huang et al., 2019)*	14.6	2.47	87.2	128.6	50.4	38.2
BLIP-2 (Li et al., 2023)*	14.9	2.43	84.6	123.4	48.4	37.1
DiscCap (Luo et al., 2018)	14.0	2.48	89.2	120.1	48.5	36.1
CL-Cap (Dai & Lin, 2017)	14.2	2.54	81.3	114.2	46.0	35.3
CIDErBtwCap (Wang et al., 2020a)	15.9	2.39	82.7	127.8	51.0	38.5

\uparrow and \downarrow show whether higher or lower scores are better according to each metric. We apply our model on four baseline models: Transformer (Luo et al., 2018) and M²-Transformer (Cornia et al., 2020) trained only with cross-entropy loss, Transformer + SCST (Luo et al., 2018) and M²-Transformer + SCST (Cornia et al., 2020) trained with reinforcement learning. *Denotes we train the model from scratch with officially released code

Bold values indicate the best result among the compared methods

Table 2 Ablation study on four components

Method	DisWordRate(%) \uparrow	CIDErRank \downarrow	CIDErBtw \downarrow	CIDEr \uparrow	BLEU3 \uparrow	BLEU4 \uparrow
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
+ WeiDisLoss	17.3	2.48	74.3	110.0	45.0	34.4
+ GDMA	18.6	2.45	72.7	110.1	45.1	34.6
+ IndTrain	18.8	2.43	72.4	109.8	44.3	33.5
M ² Transformer+SCST	17.3	2.38	82.9	128.9	50.6	38.7
+ ImageGroup	17.4	2.36	82.3	127.8	50.4	38.5
+ WeiDisLoss	18.3	2.31	81.0	128.2	50.2	38.4
+ GDMA	18.4	2.31	80.4	127.6	49.9	38.3
+ IndTrain	18.7	2.30	80.1	127.2	49.9	38.2

We train two baselines (i.e., M²Transformer and M²Transformer+SCST), and gradually add four components of our full model: ImageGroup (image group based training), WeiDisLoss (weighted distinctive loss), GDMA (group-based differential memory attention), and IndTrain (Indicated Training)

Bold values indicate the best result among the compared methods

SCST + DifDisCap. Third, compared with state-of-the-art models that improve the accuracy of generated captions, our model M²Transformer + SCST + DifDisCap achieves comparable accuracy, while also significantly improving distinctiveness. For instance, we obtain comparable CIDEr with AoANet (127.2 vs 128.6), and attain significantly higher DisWordRate, i.e., 14.6% (AoANet) vs. 18.7% (ours). When compared to other models that focus on distinctiveness, M²Transformer + SCST + DifDisCap achieves higher distinctiveness by a large margin—we gain DisWordRate by 18.7% compared with 15.9% (CIDErBtwCap), and we obtain much lower CIDErBtw by 80.1 vs. 89.2 (DiscCap).

4.3.2 Trade-Off Between accuracy and Distinctiveness

We next study the accuracy (CIDEr) versus the distinctiveness (DisWordRate) for the models in the ablation study. For comparison, we also compute the CIDEr for the human GT captions, which is the average CIDEr score between one GT caption and the remaining GT captions. The results in Fig. 4 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 the CIDEr of 83.1. Compared to the baselines, our work achieves results more similar to human performance.

4.4 Ablation Study

4.4.1 Ablation Study of the Proposed Model

To measure the influence of each component in our DifDisCap, we design an ablation study where we train the baselines M²Transformer and M²Transformer + SCST with various components of our DifDisCap. Four variants of DifDisCap are trained by gradually adding the components, i.e., ImageGroup (image group based training), WeiDisLoss (weighted distinctive loss), GDMA (group-based differential memory attention), and IndTrain (Indicated Training), 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 Wang and Chan (2019), 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 M²Transformer increases the DisWordRate by 15 percent (from 16.4 to 18.8%), while only sacrificing 1.8 percent of the CIDEr value (from 111.8 to 109.8). Similarly, applying the proposed modules on M²Transformer+SCST boosts the distinctiveness consistently. Notably, applying the weighted distinctive loss (WeiDisLoss) improves the DisWordRate from 17.3 to 18.3%, which demonstrates the importance of highlighting the distinct words in the GT captions. Moreover, adding the group-based differential memory attention (GDMA) and Indicated Training (IndTrain) increases the DisWordRate to 18.7%, which demonstrates that highlighting the distinct visual information in the target images helps the model to generate distinctive captions.

Table 3 Comparison of using different features (VSE++ and CLIP) for constructing similar image groups for generating distinctive captions

Method	Cluster Feature	DisWordRate ↑	CIDErRank↓	CIDErBtw↓	CIDEr↑	BLEU3↑	BLEU4↑
Transformer+SCST+DifDisCap	VSE++	17.0	2.34	81.5	126.9	50.6	38.4
	CLIP	17.1	2.33	81.3	126.8	50.7	38.5
M ² -Transformer+SCST+DifDisCap	VSE++	18.7	2.30	80.1	127.2	49.9	38.2
	CLIP	18.6	2.29	80.0	127.0	49.6	37.9

Bold values indicate the best result among the compared methods

Table 4 Comparison of different kinds of region feature, Bottom-up (Anderson et al., 2018), ViLBERT (Lu et al., 2019), and RegionCLIP (Zhong et al., 2022)

Method	Region feature	DisWordRate ↑	CIDErRank↓	CIDErBtw↓	CIDEr↑	BLEU3↑	BLEU4↑
Transformer + SCST + DifDisCap	Bottom-up	17.0	2.34	81.5	126.9	50.6	38.4
	ViLBERT	16.8	2.26	82.2	127.5	51.0	38.6
	RegionCLIP	17.3	2.20	81.7	128.1	51.1	38.7
M ² -Transformer + SCST + DifDisCap	Bottom-up	18.7	2.30	80.1	127.2	49.9	38.2
	ViLBERT	18.3	2.24	82.1	127.8	50.3	38.4
	RegionCLIP	18.9	2.22	80.2	128.2	50.8	38.8

Bold values indicate the best result among the compared methods

Table 5 Comparison of model performance on additional metrics: BERTScore (Zhang et al., 2019), CLIP-S (Hessel et al., 2021), and PAC-S (Sarto et al., 2023)

Method	BERTScore↑	CLIP-S↑	PAC-S↑	DisWordRate↑	CIDErRank↓	CIDErBtw↓	CIDEr↑	BLEU3↑	BLEU4↑
Transformer + SCST	0.346	0.593	0.792	14.7	2.38	83.2	127.6	51.3	38.9
+ DifDisCap	0.343	0.595	0.793	17.0	2.34	81.5	126.9	50.6	38.4
M ² -Transformer + SCST	0.343	0.602	0.798	17.3	2.38	82.9	128.9	50.6	38.7
+ DifDisCap	0.342	0.603	0.800	18.7	2.30	80.1	127.2	49.9	38.2

Bold values indicate the best result among the compared methods

4.4.2 Similar Image Group Construction

We compare the performance of using VSE++ (Faghri et al., 2018) and CLIP (Radford et al., 2021) to create similar image groups and conduct experiments on two base models, Transformer+SCST+DifDisCap and M²Transformer+SCST+DifDisCap. First, the average overlap of the similar image groups from VSE++ and CLIP was 22.6%. Second, as shown in Table 3, the models based on different image groups (VSE++ or CLIP) obtain similar final results on the accuracy metrics such as CIDEr and BLEU. The grouping using CLIP could lead to slightly better distinctive metrics (i.e., DisWordRate, CIDErRank, and CIDErBtw). These results indicate that although VSE++ and CLIP may create different image groups, both groupings are sufficient and have limited effect on the quality of the final model.

4.4.3 Image Region Features

We evaluate the influence of image features by replacing the bottom-up features (Anderson et al., 2018) extracted by Faster R-CNN (Ren et al., 2015) with two alternatives, ViLBERT (Lu et al., 2019) and RegionCLIP (Zhong et al., 2022) features. ViLBERT (Lu et al., 2019) learns task-agnostic joint representations of image content and natural language by extending the BERT architecture to a multi-modal two-stream model with interactions using co-attentional transformer layers. RegionCLIP (Zhong et al., 2022) learns region-level visual representations by aligning image regions with template captions in the feature space.

As shown in Table 4, the RegionCLIP features achieve superior performance, exemplified by a DisWordRate of 18.9% and a CIDEr score of 128.2 on the M²Transformer+SCST+DifDisCap model, outperforming other features. This advantage is attributed to its training on a larger dataset, specifically GoogleCC 3 M (Sharma et al., 2018). In contrast, the ViLBERT features, while offering improved accuracy, result in reduced distinctiveness compared to the Bottom-up feature. For instance, on the same model configuration, the CIDEr score marginally increases from 127.2 to 127.8, whereas the DisWordRate decreases from 18.7 to 18.3%.

4.4.4 Additional Metrics

We also evaluate model performance using additional metrics, BertScore (Zhang et al., 2019), CLIP-S (Hessel et al., 2021), and PAC-S (Sarto et al., 2023), in Table 5. These metrics evaluate the semantic similarity between generated captions and target images. Our DifDisCap leads to slightly higher scores on some no-reference metrics (i.e., CLIP-S and PAC-S) since these metrics measure the similarity between generated captions and target images' visual representation directly. For instance, the CLIP-S score from 0.593 to 0.595

Table 6 The performance on different group size K

Method	K	DisWordRate \uparrow	CIDErRank \downarrow	CIDErBtw \downarrow	CIDEr \uparrow	BLEU3 \uparrow	BLEU4 \uparrow
M ² Transformer + SCST	–	17.3	2.38	82.9	128.9	50.6	38.7
M ² Transformer + SCST + DifDisCap	1	17.6	2.36	81.6	128.3	50.4	38.8
	3	18.2	2.35	81.3	127.8	50.1	38.4
	5	18.7	2.30	80.1	127.2	49.9	38.2
	7	18.4	2.32	79.9	126.9	49.8	38.1
	10	18.6	2.33	80.9	127.4	50.0	38.2

We experiment on M²Transformer + SCST + DifDisCap model, and also include the performance of M²Transformer + SCST model
 Bold values indicate the best result among the compared methods

Table 7 Comparison of constructing similar image groups with static size and dynamic size

Method	Group	DisWordRate \uparrow	CIDErRank \downarrow	CIDErBtw \downarrow	CIDEr \uparrow	BLEU3 \uparrow	BLEU4 \uparrow
TF + SCST + DifDisCap	Static	17.0	2.34	81.5	126.9	50.6	38.4
	Dynamic	17.2	2.32	80.7	126.4	50.3	38.3
M ² TF + SCST + DifDisCap	Static	18.7	2.30	80.1	127.2	49.9	38.2
	Dynamic	18.8	2.28	79.6	126.8	49.7	38.1

Bold values indicate the best result among the compared methods

Table 8 The performance of our DifDisCap on Flickr30k dataset

Method	DisWordRate \uparrow	CIDErRank \downarrow	CIDErBtw \downarrow	CIDEr \uparrow	BLEU3 \uparrow	BLEU4 \uparrow
FC	7.4	2.95	32.8	45.8	30.4	19.7
UpDown	8.1	2.62	40.0	56.6	38.7	27.3
Transformer + SCST + DifDisCap	9.4	2.47	45.7	62.6	39.1	28.5
	11.7	2.28	42.0	61.9	38.5	27.9
M ² Transformer + SCST + DifDisCap	10.9	2.52	43.4	63.9	39.8	29.2
	12.1	2.17	40.2	62.2	38.9	28.5

Bold values indicate the best result among the compared methods

and the PAC-S score from 0.792 to 0.793 by adding DifDisCap to the Transformer+SCST model. Our DifDisCap leads to a slightly worse BERTScore since it still measures the similarity between generated captions and reference captions.

4.4.5 Size of Similar Image Group

Here we also discuss the effect of group size hyper-parameter K . As shown in Table 6, the results for different K values show that there is a trade-off between the accuracy and distinctiveness (e.g., DisWordRate increases from 17.6 to 18.7%, while CIDEr score decreases from 128.3 to 127.2 when setting K from 1 to 5). We use $K = 5$ for the best distinctiveness as default. In all, these experiments show that our DifDisCap method is quite robust.

The images in the data are not guaranteed to be evenly distributed, with some images having more similar images than

others. Thus, we conduct an experiment using dynamic group sizes, where larger groups are formed for groups with higher similarities. We use a similar grouping strategy as described in Sect. 3.1, but modify it to build dynamic group sizes from $(K + 3)$ to $(K - 1)$. Specifically, we first evaluate the similarity between each image and its K nearest neighbors, and then sort the image pool according to the average similarities, so that the images with higher similarity are first. The sorted list is traversed, selecting the next available image as the target image I_0 and a group is formed using the without-replacement method. Since the images that are selected first have higher similarities to other images, we set the group size of the first one-fifth of groups as $K + 3$. For each subsequent fifth, the group size is decremented by one. This procedure ensures that the average group size of the dynamic grouping is $K + 1$ for fair comparison with the static method.

As shown in Table 7, the performance gap between the static group and the dynamic group is not significant. In detail, the models based on dynamic groups have slightly better distinctiveness while slightly worse accuracy, e.g., DisWordRate from 17.0 to 17.2% while CIDEr score from 126.9 to 126.4 on Transformer+SCST+DifDisCap model. The similar group construction affects the trade-off between distinctiveness and accuracy—if the average similarity in a group is higher, the generated captions tend to be slightly more distinctive while also slightly less accurate.

4.4.6 Performance on Flickr30k dataset

In order to evaluate the generality of our model, we further perform experiments on another widely used captioning dataset Flickr30k. We adapt our DifDisCap on two models, Transformer+SCST and M²Transformer+SCST, which are trained on Flickr30k. As shown in Table 8, the DifDisCap method enhances the distinctiveness of generated captions with minimal impact on accuracy. For instance, upon integrating DifDisCap with the M²Transformer+SCST model, there is a notable increase in DisWordRate (from 10.9 to 12.1%) and a slight decrease in CIDEr score (from 66.9 to 65.2) on the Flickr30k dataset. These results mirror the trends observed on the MSCOCO dataset, indicating a consistent performance across different datasets.

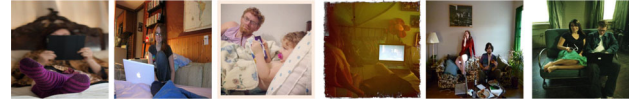
5 User Study

To 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 Wang et al. (2020a), Luo et al. (2018). Each test is a tuple $(I_0, I_1, \dots, I_K, \hat{c})$, which includes a similar image group and a caption generated by a model describing a randomly-selected 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. An answer is regarded as a hit if the user selects the image that was used to generate the caption, 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. In Fig. 5, we display the interface for our user study.

We compare our DifDisCap model with five other models, i.e., DiscCap (Luo et al., 2018), CIDErBtwCap (Wang et al., 2020a), M²Transformer + SCST (Cornia et al., 2020), BLIP-2 (Li et al., 2023) and GdisCap (Wang et al., 2021). The results are shown in Table 9, where our model achieves the highest caption-image retrieval accuracy—69.5 compared to

Following is one caption that describes one of the six images:

A man and a baby laying in a bed reading a book.



Please choose which image do you think the following caption is describing.

Fig. 5 The user study interface. We display a group of six similar images, and a caption generated from one image by an image captioning model. The users are asked to select the image that they think the caption is describing

Table 9 User study results for caption-image retrieval

Method	Accuracy
DiscCap (Luo et al., 2018)	48.1
CIDErBtwCap (Wang et al., 2020a)	58.7
M ² Transformer+SCST (Cornia et al., 2020)	61.9
BLIP-2 (Li et al., 2023)	63.6
GdisCap (Wang et al., 2021)	68.2
DifDisCap (Ours)	69.5

Our model produces captions with significantly higher retrieval accuracy (2-sample z-test on proportions, $p < 0.01$)

Bold value indicates the best result among the compared methods

BLIP-2 with 63.6 and GdisCap with 68.2. 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 for generating distinctive captions based on similar image groups. Second, we visualize the group-based memory attention calculated by our model to highlight the distinct objects. We compare our DifDisCap with four other models, BLIP-2 (Li et al., 2023), M²Transformer (Cornia et al., 2020), DiscCap (Luo et al., 2018), and CIDErBtwCap (Wang et al., 2020a), which are the best-competing methods on distinctiveness.

Figure 6(left) displays the captions generated by the five 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. BLIP-2 tends to generate simple sentences, like “a dog sitting on a wall”, which lacks details. M²Transformer and DiscCap generate captions that only mention the most salient objects in the image, using less distinctive words. For instance, in Fig. 6 (column 1), our DifDisCap generates

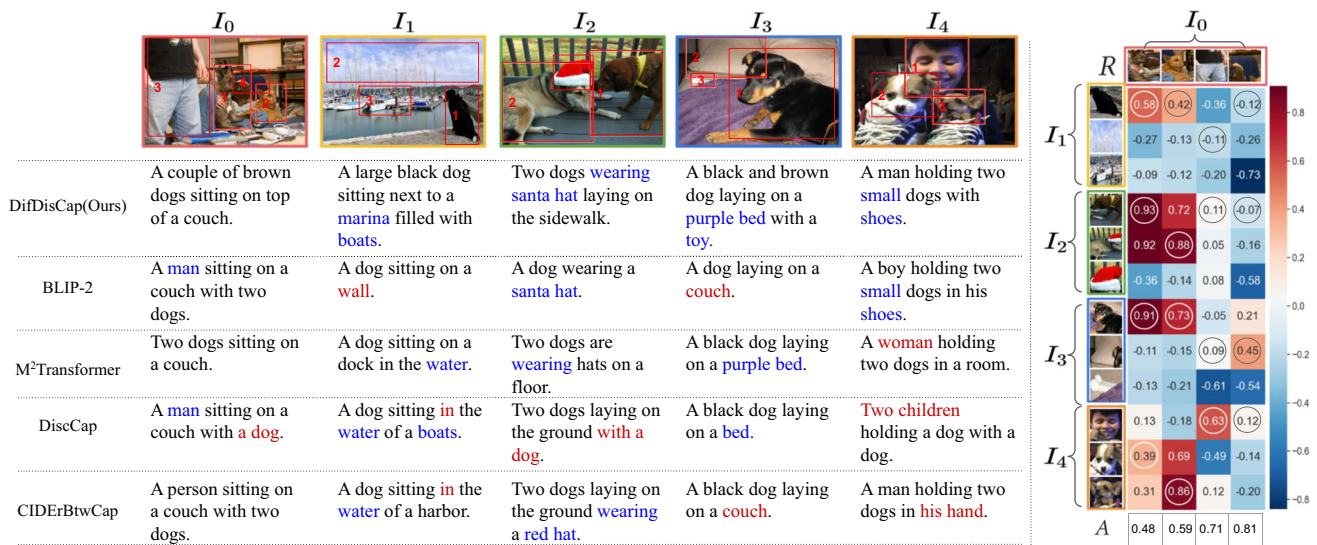


Fig. 6 Qualitative results. Left: Captions for one similar image group with five images from the test set. We compare our model with four state-of-the-art methods, BLIP-2 (Li et al., 2023), M²Transformer (Cornia et al., 2020), DiscCap (Luo et al., 2018), and CIDErBtwCap (Wang et al., 2020a). The blue words indicate the distinctive words Ω that appear in GT captions of the target image, but not in captions of similar images. The red words denote the mistakes in the generated captions.

captions “a large black dog sitting next to a marina filled with boats”, compared to the simpler caption “a dog sitting on a dock in the water” from M²Transformer. Similarly, in Fig. 6 (column 3), our DifDisCap describes the most distinctive property of the target image, the “santa hat”, compared to DiscCap which only provides “two dogs laying on the ground”. The lack of distinctiveness from M²Transformer 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 a higher attention value to the unique details and objects that appear in the image. For example, in Fig. 6 (column 3), DifDisCap describes the distinctive “santa hat”, while CIDErBtwCap mentions it as a “red hat”.

Remarkably, DifDisCap is more aware of the locations of objects in the image and the relationships among them. For example, in Fig. 6 (column 5), our caption specifies the “a man holding two small dogs with shoes”, compared with CIDErBtwCap, which wrongly describes “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, Fig. 6(right) displays the similarity matrix R and distinctive attention A for the 2nd image as the target image. The object regions with the 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 Differential Distinctive Captioning Method (DifDisCap) that compares the objects in the target image to objects in semantically similar images and highlights the unique image regions. Moreover, we developed the weighted distinctive loss to train the proposed model. The weighted distinctive loss includes the following two components: 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 on human judgment.

As seen in Fig. 4, there is still an apparent gap between human and model performance on the distinctive image captioning task. In the future, we will continue leveraging this gap, focusing on both generation and evaluation. Meanwhile, our method could be adapted to other tasks (e.g., distinctive video captioning and contrastive learning in similar image groups).

Acknowledgements This work was supported by a Strategic Research Grant (Proj. No. 7005840) from City University of Hong Kong. It was also supported by the National Natural Science Foundation of China under Grant 62301063 and the Fundamental Research Funds for the Central Universities No. 070323006.

Funding Open access publishing enabled by City University of Hong Kong Library's agreement with Springer Nature

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Anderson, P., Fernando, B., Johnson, M., & Gould, S. (2016). SPICE: Semantic propositional image caption evaluation. In *ECCV*.
- Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., & Zhang, L. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.
- Aneja, J., Agrawal, H., Batra, D., & Schwing, A. (2019). Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4261–4270).
- Aneja, J., Deshpande, A., & Schwing, A. G. (2018). Convolutional image captioning. In *CVPR*.
- Cai, D., Zhao, L., Zhang, J., Sheng, L., & Xu, D. (2022). 3djcg: A unified framework for joint dense captioning and visual grounding on 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16464–16473).
- Chen, J., Guo, H., Yi, K., Li, B., & Elhoseiny, M. (2022). VisualGPT: Data-efficient adaptation of pretrained language models for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18030–18040).
- Chen, F., Ji, R., Su, J., Wu, Y., & Wu, Y. (2017a). Structcap: Structured semantic embedding for image captioning. In *ACM MM*.
- Chen, F., Ji, R., Sun, X., Wu, Y., & Su, J. (2018). Groupcap: Group-based image captioning with structured relevance and diversity constraints. In *CVPR*.
- Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., & Chua, T. S. (2017b). SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *CVPR*.
- Cornia, M., Stefanini, M., Baraldi, L., & Cucchiara, R. (2020). Meshed-memory transformer for image captioning. In *CVPR*.
- Dai, B., & Lin, D. (2017). Contrastive learning for image captioning. In *NeurIPS*.
- Dai, B., Fidler, S., Urtasun, R., & Lin, D. (2017). Towards diverse and natural image descriptions via a conditional GAN. In *ICCV*.
- Demirel, B., & Cinbis, R. G. (2022). Caption generation on scenes with seen and unseen object categories. *Image and Vision Computing*, 124, 104515.
- Denkowski, M., & Lavie, A. (2014). Meteor universal: Language specific translation evaluation for any target language. In *EACL workshop*.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., & Uszkoreit, J. (2020). An image is worth 16 × 16 words: Transformers for image recognition at scale. In *ICLR*.
- Faghri, F., Fleet, D. J., Kiros, J. R., & Fidler, S. (2018). VSE++: Improving visual-semantic embeddings with hard negatives. In *BMVC*.
- Fei, J., Wang, T., Zhang, J., He, Z., Wang, C., & Zheng, F. (2023). Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 3136–3146).
- Guo, L., Liu, J., Zhu, X., Yao, P., Lu, S., & Lu, H. (2020). Normalized and geometry-aware self-attention network for image captioning. In *CVPR*.
- Hessel, J., Holtzman, A., Forbes, M., Bras, R. L., & Choi, Y. (2021). Clipscore: A reference-free evaluation metric for image captioning. In *EMNLP*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9), 1904–1916.
- Hirota, Y., Nakashima, Y., & Garcia, N. (2022). Quantifying societal bias amplification in image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13450–13459).
- Hu, X., Gan, Z., Wang, J., Yang, Z., Liu, Z., Lu, Y., & Wang, L. (2022). Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17980–17989).
- Huang, L., Wang, W., Chen, J., & Wei, X. Y. (2019). Attention on attention for image captioning. In *ICCV*.
- Huang, Y., Chen, J., Ouyang, W., Wan, W., & Xue, Y. (2020). Image captioning with end-to-end attribute detection and subsequent attributes prediction. *IEEE Transactions on Image processing*, 29, 4013–4026.
- Jain, U., Zhang, Z., & Schwing, A. G. (2017). Creativity: Generating diverse questions using variational autoencoders. In *CVPR*.
- Jia, C., Yang, Y., Xia, Y., Chen, Y. T., Parekh, Z., Pham, H., Le, Q., Sung, Y. H., Li, Z., & Duerig, T. (2021). Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, PMLR (pp. 4904–4916).
- Jiang, W., Zhu, M., Fang, Y., Shi, G., Zhao, X., & Liu, Y. (2022). Visual cluster grounding for image captioning. *IEEE Transactions on Image Processing*, 31, 3920.
- Karpathy, A., & Fei-Fei, L. (2015). Deep visual-semantic alignments for generating image descriptions. In *CVPR*.
- Kim, T., Ahn, P., Kim, S., Lee, S., Marsden, M., Sala, A., Kim, S. H., Han, B., Lee, K. M., Lee, H., Bae, K., Wu, X., Gao, Y., Zhang, H., Yang, Y., Guo, W., Lu, J., Oh, Y., Cho, J. W., ... & Sun, M. (2023). NICE: CVPR 2023 challenge on zero-shot image captioning. 2309.01961

- Kuo, C. W., & Kira, Z. (2022). Beyond a pre-trained object detector: Cross-modal textual and visual context for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17969–17979).
- Kuo, C. W., & Kira, Z. (2023). HAAV: Hierarchical aggregation of augmented views for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11039–11049).
- Lee, H., Yoon, S., Démoncourt, F., Bui, T., & Jung, K. (2021). UMIC: An unreferenced metric for image captioning via contrastive learning. In *ACL*.
- Lee, H., Yoon, S., Démoncourt, F., Kim, D. S., Bui, T., & Jung, K. (2020). VLBERT: Evaluating image caption using vision-and-language bert. In *Proceedings of the first workshop on evaluation and comparison of NLP systems* (pp. 34–39).
- Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML, PMLR* (pp. 19730–19742).
- Li, Z., Tran, Q., Mai, L., Lin, Z., & Yuille, A. L. (2020). Context-aware group captioning via self-attention and contrastive features. In *CVPR*.
- Li, G., Zhu, L., Liu, P., & Yang, Y. (2019). Entangled transformer for image captioning. In *ICCV*.
- Liu, X., Li, H., Shao, J., Chen, D., & Wang, X. (2018). Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *ECCV*.
- Lu, J., Batra, D., Parikh, D., Lee, S. (2019). ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS* (Vol. 32).
- Luo, R., & Shakhnarovich, G. (2019). Analysis of diversity-accuracy tradeoff in image captioning. In *ICCV Workshop*.
- Luo, R., Price, B., Cohen, S., & Shakhnarovich, G. (2018). Discriminability objective for training descriptive captions. In *CVPR*.
- Mahajan, S., & Roth, S. (2020). Diverse image captioning with context-object split latent spaces. *Advances in Neural Information Processing Systems*, 33, 3613–3624.
- Mao, J., Xu, W., Yang, Y., Wang, J., Huang, Z., & Yuille, A. (2015). Deep captioning with multimodal recurrent neural networks (m-RNN). In *ICLR*.
- Mohamed, Y., Khan, F. F., Haydarov, K., & Elhoseiny, M. (2022). It is okay to not be okay: Overcoming emotional bias in affective image captioning by contrastive data collection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 21263–21272).
- Pan, Y., Yao, T., Li, Y., & Mei, T. (2020). X-linear attention networks for image captioning. In *CVPR*.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). BLEU: A method for automatic evaluation of machine translation. In *ACL*.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., & Krueger, G. (2021). Learning transferable visual models from natural language supervision. In *International conference on machine learning, PMLR* (pp. 8748–8763).
- Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*.
- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., & Shlens, J. (2019). Stand-alone self-attention in vision models. In *NeurIPS*.
- Rasiwasia, N., Costa Pereira, J., Coviello, E., Doyle, G., Lanckriet, G. R., Levy, R., & Vasconcelos, N. (2010). A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM international conference on multimedia* (pp. 251–260).
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 779–788).
- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*.
- Rennie, S. J., Marcheret, E., Mroueh, Y., Ross, J., & Goel, V. (2017). Self-critical sequence training for image captioning. In *CVPR*.
- Sarto, S., Barraco, M., Cornia, M., Baraldi, L., & Cucchiara, R. (2023). Positive-augmented contrastive learning for image and video captioning evaluation. In *CVPR* (pp. 6914–6924).
- Sharma, P., Ding, N., Goodman, S., & Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th annual meeting of the association for computational linguistics (Volume 1: Long Papers)* (pp. 2556–2565).
- Shetty, R., Rohrbach, M., & Hendricks, L. A. (2017). Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*.
- Stefanini, M., Cornia, M., Baraldi, L., Cascianelli, S., Fiameni, G., & Cucchiara, R. (2022). From show to tell: A survey on deep learning-based image captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1), 539.
- Su, W., Zhu, X., Cao, Y., Li, B., Lu, L., Wei, F., & Dai, J. (2020). VL-BERT: Pre-training of generic visual-linguistic representations. In *ICLR*.
- Tewel, Y., Shalev, Y., Schwartz, I., & Wolf, L. (2022). Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 17918–17928).
- Van Miltenburg, E., Elliott, D., & Vossen, P. (2018). Measuring the diversity of automatic image descriptions. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1730–1741).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *NeurIPS*.
- Vedantam, R., Bengio, S., Murphy, K., Parikh, D., & Chechik, G. (2017). Context-aware captions from context-agnostic supervision. In *CVPR*.
- Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015). CIDEr: Consensus-based image description evaluation. In *CVPR* (pp. 4566–4575).
- Vered, G., Oren, G., Atzmon, Y., & Chechik, G. (2019). Joint optimization for cooperative image captioning. In *CVPR*.
- Vinyals, O., Toshev, A., Bengio, S., & Erhan, D. (2015). Show and tell: A neural image caption generator. In *CVPR*.
- Wang, Q., & Chan, A. B. (2018a). CNN + CNN: Convolutional decoders for image captioning. In *CVPR Workshop*.
- Wang, Q., & Chan, A. B. (2018b). Gated hierarchical attention for image captioning. In *ACCV*.
- Wang, Q., & Chan, A. B. (2019). Describing like humans: On diversity in image captioning. In *CVPR*.
- Wang, Q., & Chan, A. B. (2020). Towards diverse and accurate image captions via reinforcing determinantal point process. In *TPAMI*.
- Wang, L., Schwing, A. G., & Lazebnik, S. (2017). Diverse and accurate image description using a variational auto-encoder with an additive Gaussian encoding space. In *NeurIPS*.
- Wang, Q., Wan, J., & Chan, A. B. (2020b). On diversity in image captioning: Metrics and methods. In *IEEE TPAMI*.
- Wang, Q., Wang, J., Chan, A. B., Huang, S., Xiong, H., Li, X., & Dou, D. (2020c). Neighbours matter: Image captioning with similar images. In *31st British machine vision virtual conference (BMVC 2020)*.

- Wang, J., Xu, W., Wang, Q., & Chan, A. B. (2020a). Compare and reweight: Distinctive image captioning using similar images sets. In *ECCV*.
- Wang, J., Xu, W., Wang, Q., & Chan, A. B. (2021). Group-based distinctive image captioning with memory attention. In *Proceedings of the 29th ACM international conference on multimedia* (pp. 5020–5028).
- Wang, J., Xu, W., Wang, Q., & Chan, A. B. (2022). On distinctive image captioning via comparing and reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45, 2088.
- Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., & Bengio, Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.
- Xu, H., Ye, Q., Yan, M., Shi, Y., Ye, J., Xu, Y., Li, C., Bi, B., Qian, Q., Wang, W., Xu, G., Zhang, J., Huang, S., Huang, F., & Zhou, J. (2023). *mPLUG-2: A modularized multi-modal foundation model across text, image and video*.
- Yan, F., & Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3441–3450).
- Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., & Takemura, H. (2020). BERT representations for video question answering. In *WACV*.
- Ye, L., Rochan, M., Liu, Z., & Wang, Y. (2019). Cross-modal self-attention network for referring image segmentation. In *CVPR*.
- You, Q., Jin, H., Wang, Z., Fang, C., & Luo, J. (2016). Image captioning with semantic attention. In *CVPR*.
- Yuan, J., Zhu, S., Huang, S., Zhang, H., Xiao, Y., Li, Z., & Wang, M. (2022). Discriminative style learning for cross-domain image captioning. *IEEE Transactions on Image Processing*, 31, 1723–1736.
- Zeng, Z., Zhang, H., Lu, R., Wang, D., Chen, B., & Wang, Z. (2023). Conzic: Controllable zero-shot image captioning by sampling-based polishing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 23465–23476).
- Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). BERTScore: Evaluating text generation with bert. In *ICLR*.
- Zhao, D., Wang, A., & Russakovsky, O. (2021). Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 14830–14840).
- Zhao, W., Wu, X., & Luo, J. (2020). Cross-domain image captioning via cross-modal retrieval and model adaptation. *IEEE Transactions on Image Processing*, 30, 1180–1192.
- Zhong, Y., Yang, J., Zhang, P., Li, C., Codella, N., Li, L. H., Zhou, L., Dai, X., Yuan, L., Li, Y., & Gao, J. (2022). Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 16793–16803).
- Zhou, Y., Wang, M., Liu, D., Hu, Z., & Zhang, H. (2020). More grounded image captioning by distilling image-text matching model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4777–4786).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.