CONCRETE IMAGE CAPTIONING BY INTEGRATING CONTENT SENSITIVE AND GLOBAL DISCRIMINATIVE OBJECTIVE

Jie Wu[1], *Tianshui Chen*[1,2], *Hefeng Wu*[1,3]*, *Zhi Yang*[1], *Qing Wang*[1,2], *and Liang Lin*[1,2]

[1]Sun Yat-Sen University [2]DarkMatter AI Research [3]Guangdong University of Foreign Studies

wujie23@mail2.sysu.edu.cn, {tianshuichen, wuhefeng, ericwangqing}@gmail.com,
issyz@mail.sysu.edu.cn, linliang@ieee.org

ABSTRACT

Current methods for image captioning tend to generate sentences that are generally overly rigid and composed of some most frequent words/phrases, leading to inaccurate and indistinguishable descriptions. This is primarily due to the uneven word distribution of the ground truth captions that encourages to generate high frequent words/phrases while suppressing the less frequent but more concrete ones (see Figure 1). In this work, we propose a new Content Sensitive and Global Discriminative objective, which is formulated as two constraints on top of a reference model to facilitate generating concrete and discriminative image captions. More specifically, the content sensitive constraint is designed to place greater focus on the less frequent and more concrete words/phrases, thus facilitating the generation of sentences that better describe visual details of the given images. To further improve the discriminability, the global discriminative constraint is designed to pull the generated sentence to better discern the corresponding image from others. We evaluate the proposed method on the widely used MS-COCO dataset, where it achieves superior performance over existing competing methods. We also conduct self-retrieval experiments to demonstrate the discriminability of the proposed method.

Index Terms— Concrete image captioning, Content sensitive constraint, Global discriminative constraint, Self-retrieval

1. INTRODUCTION

Image captioning, i.e., automatically generating descriptive sentences of images, has received increasing attention in the fields of vision and language in recent years. Compared with other image semantic analysis tasks such as image tagging or object detection, it provides a more comprehensive and deeper understanding of the images and benefits a wide range of applications including image retrieval/classification [1], image editing, and scene graph generation [2].

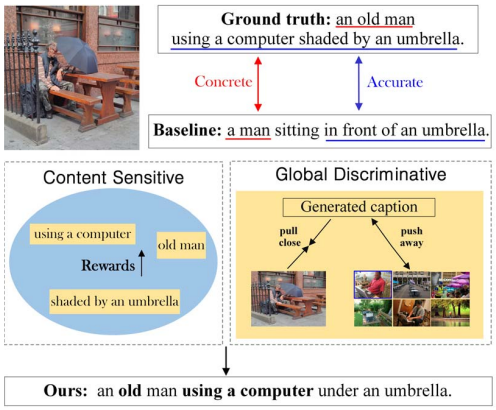

Fig. 1. An example of how our proposed method aids generating more concrete and accurate description.

With the advancement of deep learning, existing methods [3, 4] generally adopt neural network based encoder-decoder architecture [5] and resort to reinforcement learning technology [6, 7] for optimization to address this task. Despite acknowledged successes, the captions generated by these methods [4, 8] are often overly rigid and tend to replicate the words/phrases that frequently occur in the training set; thus, these captions can hardly accurately and concretely describe the given images. This is primarily due to the uneven word distribution that encourages to generate high frequent words/phrases while suppressing the less frequent but more concrete ones. For example, given a scene as shown in Figure 1, existing methods are inclined to generate the common phrase "a man". However, the less frequent phrase "an old man" is a more accurate and concrete choice. Another reason is that some images may share similar contents and these methods tend to focus more on these contents, thus leading to similar or even exactly the same descriptions. Recently, some works have resorted to adversarial learning [9, 10] or ranking loss [11] to enable generating diverse and discriminative captions. However, these methods focus more on diversity and may not

* This work was supported in part by National Natural Science Foundation of China (NSFC) under Grant No. 61876045, 61622214, 61402120, and 61836012, in part by State Key Development Program under Grant No. 2016YFB1001004, in part by the National Key Research and Development Program of China under Grant No. 2018YFC0830103, and in part by National High Level Talents Special Support Plan (Ten Thousand Talents Program). Corresponding author is Hefeng Wu.

978-1-5386-9552-4/19/$31.00 ©2019 IEEE
DOI 10.1109/ICME.2019.00227

well balance the diversity and accuracy.

In this work, we design a novel Content Sensitive and Global Discriminative objective for training the captioning model. Specifically, we first design a content sensitive constraint that pays more attention to the more concrete words/phrases. As suggested in [12], the concrete words/phrases generally occur less frequently, because they merely describe some distinct and detailed contents of some specific images. Thus, we implement this constraint by assigning higher rewards to the less frequent words. This can well address the uneven word distribution issue, and help to capture more concrete visual details of the images. On the other hand, we devise the global discriminative constraint to encourage the generated caption to better describe the corresponding image than other similar ones. To this, we adopt the ranking loss that pulls the generated caption to match the corresponding image while pushing the caption away from other similar images. The two constraints are built on top of a reference model to facilitate generating discriminative captions and simultaneously improving the accuracy.

The main contributions of this work are threefolds: 1) We design a new content sensitive and global discriminative objective for the generated caption, which encourages generating more concrete and discriminative image descriptions. 2) We conduct extensive experiments on the widely used MS-COCO dataset and set a new state-of-the-art on this dataset. In particular, our method improves the CIDEr by 0.7 on the Karpathy test split and by 2.1 on the online test server compared with the previous best-performing methods. 3) We also perform self-retrieval experiments [13] and demonstrate that our proposed method exhibits superior discriminability over existing leading and baseline methods.

2. RELATED WORK

Recent advances in image captioning have benefited from the encoder-decoder pipeline that adopted CNNs to encode semantic image representation and RNNs to decode the representation into a descriptive sentence [5, 14]. Early works [14] introduced Maximum likelihood estimation (MLE) loss that maximized the likelihood of the ground-truth word at timestep t to optimize the captioning models, but they suffered from exposure bias problem, leading to poor captioning performance [15]. To address this issues, recent works [15, 3] introduced the RL technique for sequence level training.

To address these issue, recent works [15, 3] introduced the reinforcement learning technique for sequence level training. For example, [15] defined the sequence level metric used at test time, such as BLEU [16] or ROUGE [17], thus leading to notable performance improvement during testing. Although these methods achieved impressive successes in the past several years, they tend to generate overly rigid sentences that are usually composed of the most frequent words/phrases, leading to non-concrete and indistinguishable descriptions. Most recently, since diversity and discriminability were also considered as important properties for the generated captions [10, 11], a series of efforts were dedicated to exploring generating diverse and discriminative descriptions. To achieve this, recent methods resorted to adversarial learning [10, 9] to generate humanoid and natural captions or adopted the contractive objectives [11, 13] to increase the discriminability of the generated captions. However, these methods suffered from a serious performance drop on metrics such as CIDEr [12] as they primarily focused on diversity and discriminability.

3. METHODOLOGY

Currently, advanced and typical image captioning methods adopt the encoder-decoder pipeline and generally resort to reinforcement learning (RL) technology for optimization. In this work, we also utilize this encoder-decoder pipeline [14] as our reference model. During training, the sequential word generation process is formulated as a sequential decision-making problem, and the RL technology is introduced to learn a policy network for decision making. Currently, the reward is defined based on the CIDEr score [12] because it can well measure the quality of the generated captions. However, this reward is susceptible to the uneven word distribution that encourages generating highly frequent words/phrases while suppressing the less frequent but more concrete ones.

To address the issues, we develop a content sensitive and global discriminative objective, which is formulated as two constraints on top of the above-described reference model, as shown in Figure 2. Specifically, rather than treating all words equally, the content sensitive constraint provides higher rewards for the less frequent but concrete words via word-level reward reweighting mechanism. In this way, the model may pay more attention to these words and thus alleviate the strong bias of the generated words/phrases. To encourage the generated captions to well describe the corresponding images, we further develop the global discriminative constraint that pulls the generated caption to match the corresponding image while pushing the caption away from other similar images via a ranking loss. The two constraints are formulated as two additional rewards, and thus the whole reward can be defined as

$$R = R_{\mathrm{C}} + R_{\mathrm{CS}} + R_{\mathrm{GD}}, \tag{1}$$

where R_{C} is the original reward defined based on the CIDEr score, R_{CS} and R_{GD} are the two rewards defined according to the content sensitive and global discriminative constraints, respectively. We will introduce these two rewards in detail in the following.

3.1. Content Sensitive Constraint

The content sensitive constraint is expected to assign higher rewards to the words/phrases that concretely describe the

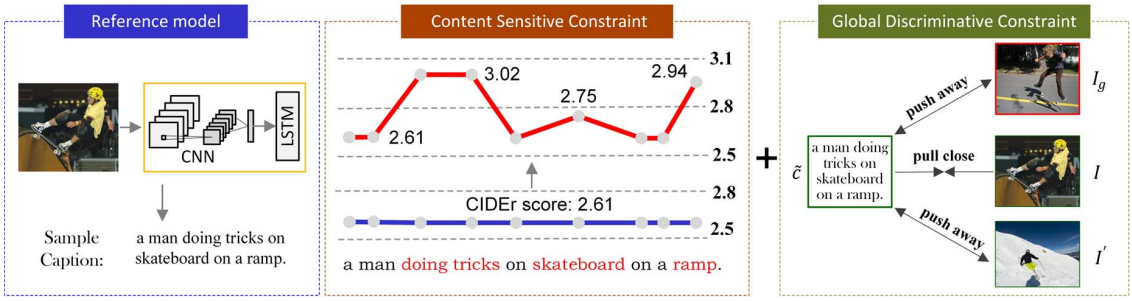

Fig. 2. An illustration of the proposed objective. It consists of a content sensitive constraint and a global discriminative constraint that are formulated on top of a reference model to encourage generating more accurate and concrete descriptions.

visual contents of given images. We find that these words/phrases generally occur less frequently in the dataset because they describe the distinct and detailed contents of some specific images. Thus, we simply assign higher rewards to these less frequent words/phrases. To this end, we introduce the Term Frequency Inverse Document Frequency (TF-IDF) score to compute the frequency of each n-gram phrase and adopt a two-stage mechanism to select the less frequent but concrete words. This mechanism first selects the less frequent n-gram phrases according to the computed TF-IDF scores, and then based on these selected phrases, it further adopts the 1-gram score to filter out the frequently occurring and common words, such as "a", "on", etc. We describe this the selection mechanism in detail in the following.

In the first stage, we follow [12] to compute a TF-IDF weight for each n-gram phrase w_k in the candidate sentence c:

$$g_{\omega_k}(c) = \frac{n_{\omega_k}(c)}{\sum_{\omega\in\Omega} n_{\omega}(c)} \log\Big(\frac{|\mathcal{I}|}{\sum_{I_p\in\mathcal{I}} \min(1, \sum_q n_{\omega_k}(s_{pq}))}\Big), \tag{2}$$

where Ω is the vocabulary of all n-grams and $\mathcal{I}$ is the set of all images in the dataset. $n_{\omega}(c)$ denotes the number of times the n-gram ω occurs in the reference sentence c. s_{pq} is the q-th sentence for image I_p. The TF-IDF weight $g_{\omega_k}(c)$ reflects the saliency of the n-gram ω_k in the dataset, and a higher weight indicates that this n-gram occurs less frequently across all images in the dataset. Thus, we introduce a threshold λ to select the n-gram phrases with TF-IDF weights higher than λ.

Note that not all the words in the selected n-gram phrases are informative, particularly some articles and conjunctions. For example, "in the grass" is a less frequent n-gram, but the article "the" and the preposition "in" occur frequently in the dataset and are usually less relevant to the image content. Thus, in the second stage, we utilize the TF-IDF scores to exclude these 1-gram words using another threshold η. The reward for the content sensitive constraint can be defined as

$$R_{\mathrm{CS}}(w_t^s) = \sum_{w_t^s\in\omega_k} \sum_j \frac{\min(g_{\omega_k}(c), g_{\omega_k}(s_j)) \cdot g_{\omega_k}(s_j)}{\|g_{\omega_k}(c)\| \|g_{\omega_k}(s_j)\|} \quad \text{if } g_{\omega_k}(c) > \lambda\,; \quad g_{w_t^s}(c) > \eta, \tag{3}$$

where $g_{w_t^s}(c)$ denotes the 1-gram TF-IDF weight for the word w_t^s, and $g_{\omega_k}(s_j)$ denotes the TF-IDF weight for the phrase ω_k in the reference sentence.

3.2. Global Discriminative Constraint

The global discriminative constraint is designed to pull the generated captions to better match the corresponding image than all the others. To this end, we first introduce a score function $s(I, c)$ that measures the similarity of an image I and sentence c. We will introduce this score function in the Experiment Setting section. Then, given an input image I and its generated caption $\tilde{c}$, it is expected that the score $s(I, \tilde{c})$ is higher than score $s(I_a, \tilde{c})$ for any image I_a taken from $\mathcal{I}$. Here, as it is impractical to generate captions for all images during training, we approximate this target by enabling $s(I, \tilde{c})$ to be higher than $s(I_g, \tilde{c})$, in which I_g is the image most similar to I, formulated as

$$R_{\mathrm{H}}(I_g, \tilde{c}) = -[\epsilon + s(I_g, \tilde{c}) - s(I, \tilde{c})]_+, \tag{4}$$

where $[x]_+$ is a ramp function defined by $\max(0, x)$. To obtain the most similar image I_g for each image I, we extract the image feature using ResNet-101 [20] pretrained on the ImageNet dataset and compute the Euclidean distance between features of I and all other images. The image with the smallest distance is selected as I_g. We can retrieve the most similar image for each image before training; thus, this process does not incur additional cost.

This reward can help improve the discriminability from a global perspective, and it merely takes one reference image (i.e., the most similar image) into consideration. Actually, some other images exist that also share very similar content with the given image. Taking these images into account can further improve the discriminability. Inspired by [11], we introduce another ranking target defined on the minibatch during training:

$$R_{\mathrm{B}}(I, \tilde{c}) = -\max_{c'}[\epsilon + s(I, c') - s(I, \tilde{c})]_+ - \max_{I'}[\epsilon + s(I', \tilde{c}) - s(I, \tilde{c})]_+, \tag{5}$$

where (I, c') and $(I', \tilde{c})$ are mismatching pairs on minibatch.

Model	BLEU4	BLEU3	BLEU2	BLEU1	ROUGEL	METEOR	SPICE	CIDEr
AdaAtt [18]	33.2	44.5	59.1	74.2	-	26.6	-	108.5
TD-ATT [19]	34.0	45.6	60.3	76.5	55.5	26.3	-	111.6
Rennie [3]	34.2	-	-	-	55.7	26.7	-	114.0
Stack-cap [4]	36.1	47.9	62.5	78.6	56.9	27.4	20.9	120.4
Up-down [8]	**36.3**	-	-	**79.8**	56.9	27.7	21.4	120.1
Ours (TDA+CS-GD)	36.1	**48.0**	**62.6**	78.8	**57.1**	**27.8**	**21.6**	**121.1**

Table 1. Performance of our proposed and existing state-of-the-art methods on Karpathy test splits. We report our results that use the more advanced TDA baseline. - indicates that the corresponding values are not available.

Finally, we simply sum the two term to obtain the global discriminative reward:

$$R_{\mathrm{GD}}(I, I_g, \tilde{c}) = R_{\mathrm{H}}(I_g, \tilde{c}) + R_{\mathrm{B}}(I, \tilde{c}). \tag{6}$$

3.3. Optimization

At the training stage, we aim to minimize the following negative expected reward, formulated as

$$\mathcal{L}(\theta) = -\mathbb{E}_{c^s \sim p_\theta}[\sum_{t=1}^{T} R(w_t^s)], \tag{7}$$

where $c^s = \{w_1^s, w_2^s, \ldots, w_T^s\}$ is a sentence sampled from the model characterized by distribution p_θ [3] and w_t^s is the word at time step t. In practice, we utilize a sample mechanism to approximate the expectation and introduce the REINFORCE algorithm [6] to compute the gradients, formulated as

$$\begin{aligned} \nabla\mathcal{L}(\theta) &= -\mathbb{E}_{c^s \sim p_\theta}[\sum_{t=1}^{T} R(w_t^s)\nabla\log(p_\theta(w_t^s))] \\ &= -\frac{1}{M}\sum_{m=1}^{M}\sum_{t=1}^{T} R(w_{mt}^s)\nabla\log(p_\theta(w_{mt}^s)), \end{aligned} \tag{8}$$

where M is the number of sampled sentences. The gradient estimated by the above approximation is of high variance, making the model extremely difficult to converge. To address this issue, we follow [19] to introduce a reference sentence to obtain an unbiased low-variance gradient estimation.

4. EXPERIMENTS

4.1. Experiment Settings

Datasets and Metrics. MS-COCO [23] is a widely used benchmark for image captioning. This dataset contains 123,287 images, with each image annotating 5 sentences. In this work, we follow the Karpathy split [24] that divides the dataset into a training set of 113,287 images, a validation set of 5,000 images, and a test set of 5,000 images for evaluation. We also submit our results to the online MS-COCO test server for fair comparisons with the published methods. We evaluate our method, the baseline methods and other competitors on widely used metrics including BLEU [16], CIDEr [12], SPICE [25], ROUGEL [17], and METEOR [26].

Implementation details. We utilize two typical and advanced methods as our reference models, i.e., Show-Tell (ST) [5] and Top-Down Attention (TDA) [8]. Both baseline methods adopt the encoder-decoder pipeline, and we follow existing methods [3] to use ResNet-101 [20] for image encoding and an LSTM with hidden state size of 512 for decoding captions. We exactly follow work [3] to train the models using the MLE loss for the first 20 epochs and then switch to the RL loss to continue training. During inference, we use beam search with a size of 3 to decode the captions.

	Evaluation Metrics					Diversity	
	BLEU4	ROUGEL	METEOR	SPICE	CIDEr	Number	Ave.Len.
ST	32.8	54.7	25.7	19.1	103.1	2713	9.20
ST+CS	**33.1**	**55.0**	25.8	19.0	107.2	2765	9.22
ST+GD	32.9	54.8	25.8	19.0	104.0	3040	9.28
ST+CS-GD	33.0	54.9	**25.9**	**19.3**	**107.7**	**3140**	**9.29**
TDA	36.1	57.1	27.5	21.0	117.0	3589	9.33
TDA+CS	**36.3**	**57.2**	**27.8**	21.4	121.0	3448	9.41
TDA+GD	36.1	57.1	27.6	21.3	117.9	3612	9.52
TDA+CS-GD	36.1	57.1	**27.8**	**21.6**	**121.1**	**3797**	**9.56**

Table 3. Performance of different baseline models on evaluation metrics and diversity.

4.2. Comparison with State-of-the-art methods

We first present the comparison results on the Karpathy's test split [24] of MS-COCO in Table 1. As shown, the previous best-performing methods are Stack-cap [4] and Up-down [8] that obtain the CIDEr scores of 120.4 and 120.1, respectively. Our method outperforms these competitors on nearly all metrics, e.g., improving the CIDEr score to 121.1.

For more comprehensive comparisons, we also submit our result to the online MS-COCO test server for evaluation, and we present the results of our method and those of the published leading competitors in Table 2. Our method still outperforms these methods by a sizable margin, e.g., improving the CIDEr (c5) score by 2.1 compared with the previous best (i.e., Stack-cap [4]). Some methods also report the results of ensembling several models [8, 3]. By simply ensembling four models, our method achieves competitive performance compared with existing state-of-the-art methods [8].

4.3. Ablation Study

As the proposed objective is formulated on existing reference models (i.e., ST and TDA), we emphasize the comparison with these models to demonstrate its effectiveness in Table 3. As shown, our method exhibits notable improvements on all metrics, e.g., improving the CIDEr scores by 4.6 if using the ST baseline and by 4.1 if using the TDA baseline. Meanwhile, our method also encourages generating more diverse and concrete descriptions than these baselines. Specifically,

	BLEU1		BLEU2		BLEU3		BLEU4		METEOR		ROUGE-L		CIDEr	
	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40	c5	c40
AdaAtt [18]	74.8	92.0	58.4	84.5	44.4	74.4	33.6	63.7	26.4	35.9	55.0	70.5	104.2	105.9
SPIDEr [21]	75.4	91.8	59.1	84.1	44.5	73.8	33.2	62.4	25.7	34.0	55.0	69.5	101.3	103.2
TD-ATT [19]	75.7	91.3	59.1	83.6	44.1	72.6	32.4	60.9	25.9	34.2	54.7	68.9	105.9	109.0
AC [22]	77.8	92.9	61.2	85.5	45.9	74.5	33.7	62.5	26.4	34.4	55.4	69.1	110.2	112.1
Stack-cap [4]	77.8	93.2	61.6	86.1	46.8	76.0	34.9	64.6	27.0	35.6	56.2	70.6	114.8	118.3
Ours (TDA+CS-GD)	**78.7**	**93.7**	**62.6**	**86.9**	**47.8**	**77.1**	**35.9**	**65.8**	**27.5**	**36.2**	**56.9**	**71.6**	**116.9**	**119.5**
Rennie† [3]	78.1	93.1	61.9	86.0	47.0	75.9	35.2	64.5	27.0	35.5	56.3	70.7	114.7	116.7
Up-down† [8]	**80.2**	**95.2**	**64.1**	**88.8**	**49.1**	**79.4**	**36.9**	**68.5**	27.6	**36.7**	**57.1**	**72.4**	**117.9**	**120.5**
Ours (TDA+CS-GD)†	79.0	94.0	63.0	87.4	48.2	77.7	36.3	66.6	**27.7**	36.6	**57.1**	71.9	**117.9**	120.4

Table 2. Performance of our proposed and existing state-of-the-art methods on the online MS-COCO test server. We report our results that use the more advanced TDA baseline. † indicates the results of ensemble models.

Fig. 3. Captions generated by the TDA baseline and Our TDA+CS-GD.

upon the 5,000 test images, the ST baseline can merely generate 2,713 different sentences, and our method can increase the number to 3,140, a relative increase of 15.7%. A similar phenomenon can be observed if using the TDA baseline. Note that the two baselines are different, and we can achieve consistent improvement on both baselines, suggesting that our proposed objective is capable of adapting to various caption models.

Qualitative Comparison. To provide more direct comparisons, we also visualize some sentences generated by our proposed and the baseline methods. As shown in Figure 3, the baseline usually ignores the detailed contents (e.g., the samples at first row) and tends to describe the shared content for similar images (e.g., the samples at second row). In contrast, our method can better capture more concrete details, facilitating generating more detailed and distinguishable captions.

To give a deeper understanding of the proposed method, we further conduct ablative studies to assess the actual contributions of the CS and GD constraints.

Contribution of CS Constraint. We first evaluate the contribution of the CS constraint by merely incorporating this constraint in the reference model. As shown in Table 3, it leads to a clear performance improvement, e.g., the CIDEr score improvements of 4.1 and 4.0 using the two reference models, respectively.

Contribution of GD Constraint. Similarly, we merely incorporate the GD constraint in the reference model to evaluate its contribution. It can also improve most of the evaluation metrics. As shown in Table 3, the CIDEr score both increases by 0.9 if using the ST baseline or the TDA baseline. On the other hand, the GD constraint can significantly increase the diversity of generated captions. Specifically, it increases the number of different sentences from 2,713 to 3,040 if using the ST baseline and from 3,589 to 3,612 if using the TDA baseline.

4.4. Evaluation on Self-Retrieval

In this section, we follow previous work [13] to conduct self-retrieval comparisons to evaluate the discriminability of the proposed method. Specifically, it first randomly selects 5,000 images $\{I_1, I_2, \ldots, I_{5000}\}$ from the MS-COCO test set and adopts the captioning models to generate the corresponding 5,000 sentences $\{c_1, c_2, \ldots, c_{5000}\}$. Then, for each sample i, we use c_i as a query and compute the probabilities conditioned on each image, i.e., $\{p(c_i|I_1), p(c_i|I_2), \ldots, p(c_i|I_{5000})\}$. We consider an image to be top-K recalled if the conditional probability $p(c_i|I_i)$ is within the top K highest probabilities, and we define the Recall@K as the fraction of samples that are top-K recalled with respect to all samples. A high Recall@K indicates that the images are easily retrieved, and thus the generated captions are more discriminative. Work [11] also integrates a ranking loss to improve the discriminability of the generated captions, and we implement this loss upon the ST and TDA baselines for fair comparison. As shown in Table 4, our method consistently outperforms this method over all metrics.

We also compare our method with those that merely incorporate CS constraint or GD constraint, and the baseline in Table 4. It can be seen that our method exhibits a notable improvement compared with the baseline. Additionally, the methods that merely incorporate either the CS or the GD constraint also lead to an improvement in the retrieval performance. These results clearly demonstrate that our proposed objective and both of the constraints can help improve discriminability.

	Performancess on Self-retrieval		
	R@1	R@5	R@10
ST+[11]	60.24	86.33	93.18
ST	50.84	78.54	87.58
ST+CS	53.60	82.24	90.38
ST+GD	61.74	87.12	93.94
ST+CS-GD	**62.08**	**88.24**	**94.36**
TDA+[11]	73.60	93.04	96.52
TDA	66.40	88.46	94.26
TDA+CS	68.82	90.90	95.80
TDA+GD	74.53	93.67	97.03
TDA+CS-GD	**76.24**	**94.50**	**97.90**

Table 4. The performances on self-retrieval experiment.

5. CONCLUSIONS

Aiming at generating concrete and discriminative captions, this work proposes a content sensitive and global discriminative objective, which can be simply formulated as two constraints on top of a reference model. Specifically, the content sensitive constraint is designed to focus more on the less frequent word and thus enable describing more detailed and concrete contents of the input image, and the global discriminative constraint is utilized to pull the generated caption to better describe the corresponding image, thus improving their discriminability. Extensive experiments on the MS-COCO dataset demonstrate the superior performance of the proposed method over existing leading and baseline methods. We also conduct self-retrieval experiments that verify the discriminability of our proposed method.

6. REFERENCES

[1] Tianshui Chen, Liang Lin, Riquan Chen, Yang Wu, and Xiaonan Luo, "Knowledge-embedded representation learning for fine-grained image recognition," in *IJCAI*, 2018.

[2] Tianshui Chen, Weihao Yu, Riquan Chen, and Liang Lin, "Knowledge-embedded routing network for scene graph generation," in *CVPR*, 2019.

[3] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jarret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *CVPR*, 2017.

[4] Jiuxiang Gu, Jianfei Cai, Gang Wang, and Tsuhan Chen, "Stack-captioning: Coarse-to-fine learning for image captioning," in *AAAI*, 2018.

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

[6] "Simple statistical gradient-following algorithms for connectionist reinforcement learning," *Machine Learning*, vol. 8, no. 3-4, pp. 229–256, 1992.

[7] Tianshui Chen, Zhouxia Wang, Guanbin Li, and Liang Lin, "Recurrent attentional reinforcement learning for multi-label image recognition," in *AAAI*, 2018.

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

[9] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele, "Speaking the same language: Matching machine to human captions by adversarial training," in *ICCV*, 2017.

[10] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin, "Towards diverse and natural image descriptions via a conditional gan," in *ICCV*, 2017.

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

[12] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *CVPR*, 2015.

[13] Bo Dai and Dahua Lin, "Contrastive learning for image captioning," in *NIPS*, 2017.

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

[15] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba, "Sequence level training with recurrent neural networks," *arXiv preprint arXiv:1511.06732*, 2015.

[16] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002.

[17] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *ACL workshop*, 2004.

[18] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *CVPR*, 2017.

[19] Hui Chen, Guiguang Ding, Sicheng Zhao, and Jungong Han, "Temporal-difference learning with sampling baseline for image captioning," in *AAAI*, 2018.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[21] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy, "Improved image captioning via policy gradient optimization of spider," in *ICCV*, 2017.

[22] Li Zhang, Flood Sung, Feng Liu, Tao Xiang, Shaogang Gong, Yongxin Yang, and Timothy M Hospedales, "Actor-critic sequence training for image captioning," *arXiv preprint arXiv:1706.09601*, 2017.

[23] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick, "Microsoft coco captions: Data collection and evaluation server," *arXiv preprint arXiv:1504.00325*, 2015.

[24] Andrej Karpathy and Li Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," in *CVPR*, 2015.

[25] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *ECCV*, 2016.

[26] Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *ACL workshop*, 2005.