Partial Off-policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning

Jiahe Shi Yali Li* Shengjin Wang
Department of Eletronic Engineering, Tsinghua University, Beijing 100084, China
shijh18@mails.tsinghua.edu.cn {liyali13, wgsgj}@tsinghua.edu.cn

Abstract

Human-oriented image captioning with both high diversity and accuracy is a challenging task in vision+language modeling. The reinforcement learning (RL) based frameworks promote the accuracy of image captioning, yet seriously hurt the diversity. In contrast, other methods based on variational auto-encoder (VAE) or generative adversarial network (GAN) can produce diverse yet less accurate captions. In this work, we devote our attention to promote the diversity of RL-based image captioning. To be specific, we devise a partial off-policy learning scheme to balance accuracy and diversity. First, we keep the model exposed to varied candidate captions by sampling from the initial state before RL launched. Second, a novel criterion named max-CIDEr is proposed to serve as the reward for promoting diversity. We combine the above-mentioned off-policy strategy with the on-policy one to moderate the exploration effect, further balancing the diversity and accuracy for human-like image captioning. Experiments show that our method locates the closest to human performance in the diversity-accuracy space, and achieves the highest Pearson correlation as 0.337 with human performance.

1. Introduction

Image captioning is a challenging task in the field of computer vision and natural language processing. It requires not only extracting semantic information from images but also understanding and reorganizing such information in the form of natural language. To describe like humans, image captioning models should be capable of producing diverse and accurate captions. Besides generating correct captions, several different captions about the visual content should also be provided. From such a perspective, image caption models are supposed to generate human-oriented predictions by balancing accuracy and diversity.

Recent image captioning methods focus more on accuracy with deep reinforcement learning (RL). In particular, on-policy RL is adopted in [26, 23] to reduce the exposure bias and acquire sentence-level supervision. These methods are proven to benefit the accuracy performance on multiple metrics [22, 31, 9, 18, 1, 45]. However, they are prone to generate common sentences, resulting in poor diversity [21]. Some other works focus on maintaining diversity [27, 10, 35, 38, 3, 4]. Based on VAE or GAN, the diverse captions can be obtained. Yet the reported fair accuracy is acquired under the selection of oracle or consensus re-ranking [11] process. When considering the entire posterior, there will be noticeable inaccurate cases predicted by these models. As demonstrated in Fig. 1, we sample 5 times according to the modeled posterior to evaluate the quality on both accuracy and diversity. Though the on-policy RL trained model [34] generates captions with no faults, it fails to produce distinct sentences in other forms. The diverse captioning model [35] can provide varied predictions, but incorrect descriptions exist within the outputs. In a word, there are obvious performance gaps on either diversity or accuracy for existing image captioning methods.

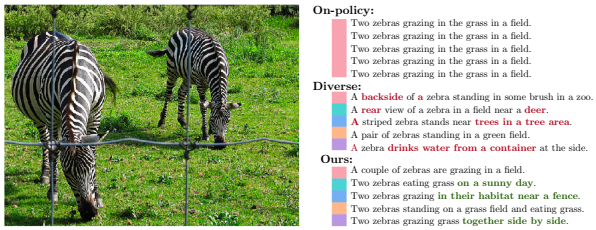


Figure 1. Typical outputs from different models by sampling 5 times according to the posteriors, where the captions generated by on-policy trained model are correct but unvaried (**top**), and captions generated by models like VAE are diverse but less accurate (**middle**). We aim at balancing the two aspects to mimic generating human-oriented captions (**bottom**).

In this paper, we motivate to balance the accuracy and diversity of image captioning models. To favor accuracy, we train the image captioning model based on deep reinforcement learning. We investigate why current RL-based methods fail to generate diverse captions. We discover that the

*corresponding author

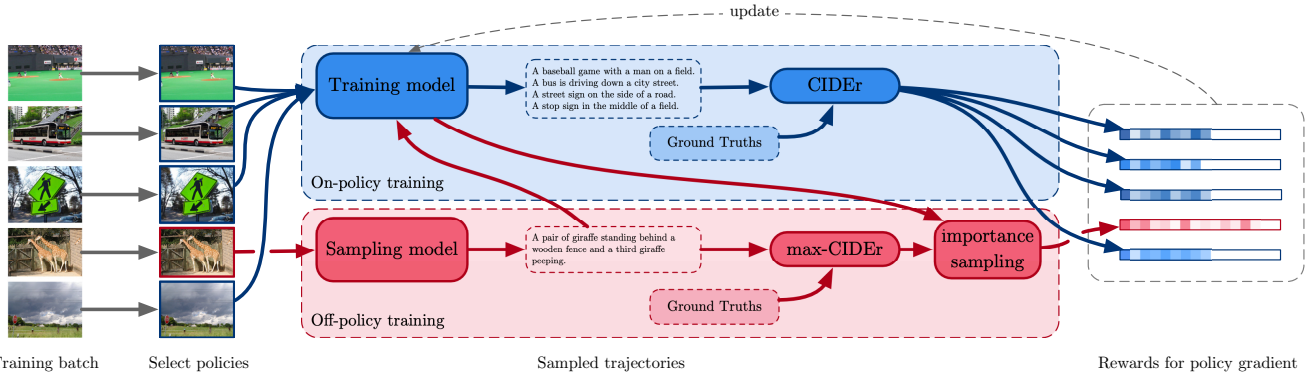


Figure 2. High-level overview for the proposed partial off-policy learning scheme. Training samples are allocated to different training strategies to balance accuracy promotion and diversity preservation.

on-policy strategy is easily trapped for a single prediction. Therefore, more exploration is required during the training process to expose the agent to more fair cases. However, traditional exploration strategies treat unexplored cases indiscriminately. Considering the enormous searching space of the generated sentences, such exploration may be inefficient for the task of image captioning. Based on these observations, we propose a novel **partial off-policy** learning scheme to encourage the exploration of new possibilities efficiently. To be specific, we first introduce an off-policy strategy into the image captioning framework, for which a diverse distribution is chosen as behavior policy for exploration. Samples derived from the such a policy are then fed into the model and rewarded by a novel criterion as *max-CIDEr* to encourage recurrence. With such behavior policy, the enormous searching space can be narrowed down to a certain sub-space to facilitate the training process. In practice, we select over the on-policy and above-mentioned off-policy strategies with a certain probability to moderate the exploration effect. Such *partial off-policy* learning scheme allows us to negotiate the trade-off between diversity preservation and accuracy promotion, ultimately encouraging the model to mimic human-like performance.

The main contributions of this paper are: **1)** We propose the off-policy strategy and the novel max-CIDEr reward for RL-based image captioning to promote diversity. **2)** We propose the partial off-policy learning to balance the diversity and accuracy for human-oriented image captioning. Our work is evaluated on MSCOCO dataset [19]. We achieve a significant boost on diversity compared with the on-policy baseline, while acquiring the highest accuracy on all sampled predictions compared with other diverse captioning works. Besides, our method locates closest to the human performance in the diversity-accuracy space and shows the strongest correlation to human evaluation with Pearson correlation as 0.337. Our work is modular and can be applied to most other works for image captioning, making it easy to facilitate such balance in future researches.

2. Related Work

Image Captioning. Inspired by the success of sequence-to-sequence learning [28, 6] under neural machine translation, an encoder-decoder framework [34, 33] has been introduced into image captioning and achieves noteworthy improvements. The framework extracts semantic features of images with an encoder CNN and models the posterior given such features using a decoder LSTM. Within such a framework, there are many modifications to the model structure. [2] proposes to use faster R-CNN [24] as encoder to extract object-level representations. [13, 7] replaces the decoder LSTM with modified transformer [30] structure for better language generation. Different types of attention modules [40, 44, 20, 2, 14] have been designed to further bridge the gap between the visual concept and the lingual one. Furthermore, [43, 42, 41] introduce Graph Convolutional Network to employ scene graphs [15] within the image.

Improve Accuracy with RL. While early researches train the networks using cross-entropy (CE) loss word-by-word, [23] treats the sentence generation process as a sequential decision problem and introduce reinforcement learning (RL). [25] defines a reward using visual-semantic embedding. [26] further proposes self-critical reward which becomes the mainstream RL method for image captioning later on. Technically, CE pre-trained model is used to initialize a reasonable baseline for the RL training process, then CIDEr [31] is chosen as sentence-level evaluation to provide rewards for the sampled sentences. Here, the sampled sentences used to acquire rewards are derived from the same policy with the one to be updated. In other words, the self-critical approach can be classified as the on-policy training strategy. Experiments have shown that models trained with such an approach can obtain high accuracy scores [22, 31, 9, 18, 1, 45], which even exceeds human grades by a significant margin.

Improve Diversity for Image Captioning. To enhance diversity, [38] employs multiple describing models to learn

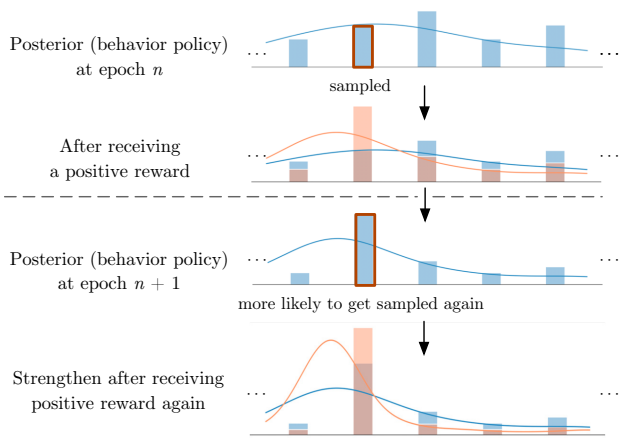


Figure 3. Intuitive illustration of how on-policy training impairs diversity. Trajectories receiving positive rewards will be more likely to be sampled in the following epochs, making the target policy become a unimodal one.

different modalities. [32] proposes diverse beam search (DBS) to encourage different captions during inference. [10] generates different captions by utilizing different part-of-speech. [5] promotes diversity by directly adding diversity evaluations onto optimization target. Recent works address the diverse captioning task by introducing generative models like GAN and VAE. [8] introduces sequence GAN to generate different captions, while [27] further designs a discriminator taking distribution within the caption set into consideration. On the other hand, [35] designs multiple priors for using VAE in caption tasks. [3] learns word-wise latent space and [4] models latent variables for syntactic or lexical domain knowledge. The above-mentioned researches mainly report accuracy over a certain sentence within the sampled caption set. On the contrary, [36] proposes to evaluate the accuracy performance of the entire caption set. To provide comprehensive evaluation over the modeled posterior, we follow such implementation when reporting performance in this paper.

3. Proposed Approach

To balance accuracy and diversity towards human performance, we adopt RL for its significant boosting effect on accuracy and manage to alleviate the impairment effect on diversity. In this section, we start by formulating the limitations of the existing on-policy strategy in Sec. 3.1. We then propose the **partial off-policy** learning scheme. The overall training scheme is demonstrated as Fig. 2, where the off-policy training branch is introduced as a complement to the traditional on-policy training branch.

3.1. Problem Formulation

The caption generation procedure can be formulated as a sequential decision-making process from the perspective of RL [25, 26]. The fully generated sentence $s = \{w_1, w_2, ..., w_T\}$ can be viewed as a *trajectory*. The optimization target is to minimize the negative expected sentence-level reward $\mathbf{R}(s; I)$:

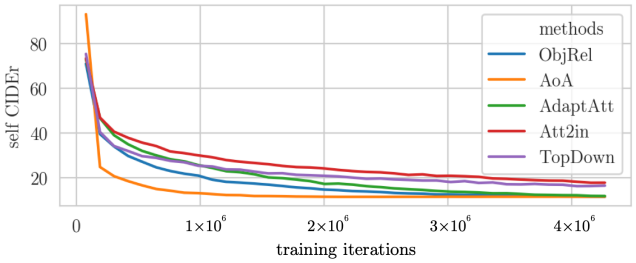


Figure 4. Diversity of predicted sentences declined significantly as on-policy RL training proceeds.

$$L(\theta) = -\mathbb{E}_{s\sim p_\theta}[\mathbf{R}(s; I)] \tag{1}$$

where p_θ is the target policy, *i.e.* modeled posterior to be trained. Since it is expensive to calculate the expectation, $L(\theta)$ is usually estimated according to a single trajectory:

$$\hat{L}(\theta) = -R(s, I), \quad s \sim p_\theta \tag{2}$$

We can then optimize the non-differentiable reward using REINFORCE policy gradient [39]:

$$\nabla_\theta \hat{L}(\theta) = -R(s, I)\nabla_\theta \log p_\theta(s|I), \quad s \sim p_\theta \tag{3}$$

Note that the baseline term is omitted for clarity here and after. Eqn.(3) is the foundation of current RL-based training schemes for image captioning. It implies on-policy strategy, *i.e.* the sampled sentence s is derived from the same distribution p_θ with the one to be trained.

According to Eqn.(3), the target policy can be updated over each trajectory from the sentence space. However, the model may practically learn from only a small number of trajectories. Such a phenomenon hinders the model from searching for other potentially good candidates. We illustrate the issue in Fig. 3. If a sampled sentence s^* receives a high reward, it will be increased in probability and relatively suppress other cases in p_θ. In the subsequent epochs, s^* will be more likely sampled according to p_θ and further strengthen such tendency. Ultimately, the posterior will gradually become unimodal resulting in the deficiency of diversity, as illustrated in Fig. 4.

To maintain diversity while optimizing accuracy using RL, we need to **1)** expose the model to a variety of possible samples during training and **2)** encourage the model to output with higher probability if the sample is decent enough. We address the first problem in Sec. 3.2 by introducing an off-policy sampling model for image captioning. The second issue is handled in Sec. 3.3 by the proposed *max-CIDEr*

reward optimization. Moreover, to balance between diversity and accuracy, we propose to adopt the off-policy strategy with a certain probability, which is interpreted as *partial off-policy* learning in Sec. 3.4.

3.2. Off-policy Strategy for Image Captioning

Since the existing on-policy strategy applied in image captioning emphasizes accuracy too much, we propose to first strengthen diversity before considering the diversity-accuracy trade-off. Technically, a different behavior policy is introduced to provide varied trajectories for the model during the training process, *i.e.*, the model will update in an off-policy manner. Mathematically, gradients of off-policy strategy can be estimated as:

$$\nabla_\theta \hat{L}(\theta) = -R(s, I)\nabla_\theta \log p_\theta(s|I) \cdot \frac{p_\theta(s|I)}{q(s|I)}, \quad s \sim q \tag{4}$$

where q denotes the behavior policy. $\frac{p_\theta(s|I)}{q(s|I)}$ is the importance sampling ratio term ensuring unbiased estimation. By adopting q, the model is exposed to decent cases outside the current local optimum, breaking the loop shown in Fig. 3 and avoiding establishing unimodal posterior.

Considering the enormous sentence space, we need to narrow the scope to some fair candidates when selecting q. We solve such a requirement by deploying a sampling model to provide a reasonable q. Since the initial model when RL launched shows considerable diversity as reported in Fig. 4, it is used as the sampling model here:

$$q \leftarrow p_{\theta=\theta_0} \tag{5}$$

where θ_0 denotes the parameter of the initial model. Eqn.(5) ensures the rudimentary capability to generate potentially competitive captions while maintaining necessary diversity as required by Eqn.(4).

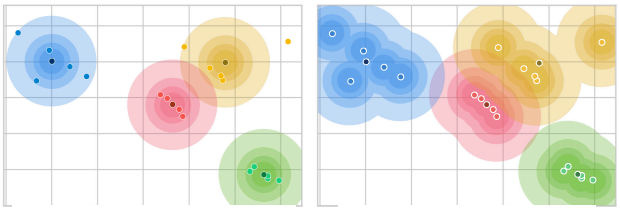
3.3. Max-CIDEr Optimization

In this section, we discuss how to fully utilize varied caption samples provided by Sec. 3.2. To start with, we first discuss the limitation of the widely used reward *CIDEr*. According to [31], original CIDEr score can be represented as

$$\text{CIDEr}(s, I) = \frac{1}{4}\sum_{n=1}^{4} \text{CIDEr}_n(s, I) \tag{6}$$

where $\text{CIDEr}_n(s)$ is derived by the cosine similarity of the clipped n-gram tf-idf vectors $\mathbf{g}_n$ between candidate prediction s and ground truths $\mathcal{G}(I)$:

$$\begin{aligned} \text{CIDEr}_n(s, I) &= \frac{1}{|\mathcal{G}(I)|}\sum_{j\in\mathcal{G}(I)} \cos\langle \mathbf{g}_n(s), \mathbf{g}_n(j)\rangle \\ &= \mathbf{e}_n(s) \cdot \left(\frac{1}{|\mathcal{G}(I)|}\sum_{j\in\mathcal{G}(I)} \mathbf{e}_n(j)\right) \\ &= \mathbf{e}_n(s) \cdot \bar{\mathbf{e}}_n(\mathcal{G}(I)) \end{aligned} \tag{7}$$



a) rewards using original CIDEr **b)** rewards using max-CIDEr

Figure 5. Schematic of distribution of **a)** original CIDEr reward and **b)** proposed max-CIDEr reward. Max-CIDEr encourages more diverse output comparing with the original CIDEr.

Here, $\mathbf{e}_n = \mathbf{g}_n/\|\mathbf{g}_n\|$ is the unit vector representing the direction of $\mathbf{g}_n$. As Eqn.(7) suggests, the model can only receive guidance from a single vector $\bar{\mathbf{e}}_n$ representing the merged semantics. To provide an intuitive impression, we sample several groups of ground truths from the MSCOCO dataset [19] and visualize the corresponding tf-idf vectors by PCA in Fig. 5. The contours within the diagram act as simple illustrations for the reward. As we can see from Fig. 5**a**, different sampled captions for an image are encouraged to approach the same merged semantics (darker points) to pursue higher CIDEr rewards, which results in diversity declining despite various training samples provided.

Thus, to promote diversity for accuracy-diversity balance, we optimize a novel reward *max-CIDEr* defined as:

$$\text{max-CIDEr}(s, I) = \max_{j\in\mathcal{G}(I)}\left(\frac{1}{4}\sum_{n=1}^{4}\cos\langle \mathbf{g}_n(s), \mathbf{g}_n(j)\rangle\right) \tag{8}$$

Eqn.(8) suggests that a sentence is considered as a good candidate as long as it is similar enough with any of the annotations rather than all of them. Thus, a sampled sentence may be *pushed up* under looser conditions. As illustrated in Fig. 5**b**, max-CIDEr is able to provide different guidance information for each trajectory during training. Moreover, by adopting the proposed max-CIDEr reward, ground truths rather than the merged semantics are counted as optimal, which is consistent with human cognition.

3.4. Partial Off-policy Learning

The traditional on-policy learning scheme for image captioning acquires poor diversity performance on account of insufficient exploration. On the contrary, the proposed off-policy strategy has more exploration effect because of the diverse behavior policy [29]. However, due to the imperfection of the sampling model, the deployment of off-policy training solely may bring about low accuracy performance, since the training model needs to update policies over much more trajectories while some of which are of low qualities.

To alleviate such an issue, we moderate the exploration effect of the proposed off-policy strategy by combining it

with the traditional on-policy one of image captioning. Inspired by the ϵ-greedy algorithm [29], we introduce a hyper-parameter ϵ to control such balance. Specifically, we perform strategy selection over the proposed off-policy approach and the traditional on-policy one for each image in the training set. The learning scheme is named as *partial off-policy learning* and is illustrated in Fig. 2, which can be formulated as pseudo-code in Algo. 1. The partial off-policy learning scheme preserves the advantage of traditional on-policy strategy for image captioning to effectively exploit and guarantee accuracy. It also introduces exploration to maintain diversity using off-policy training with max-CIDEr. With ϵ well-tuned, the proposed method can balance between exploration and exploitation, and consequently derive a posterior with both accuracy and diversity.

Algo 1: Partial off-policy learning for image captioning

input : pretrained model parameters θ_0, training dataset $\mathcal{I}$, balance coefficient ϵ
output: optimized model posterior p_θ

initialize p_θ with θ_0;
initialize q with θ_0;
while *not reach maximum epochs* **do**
  for $I \in \mathcal{I}$ **do**
    sample $\varepsilon \sim U(0,1)$;
    if $\varepsilon > \epsilon$ **then**
      sample a caption s according to p_θ;
      $R(s,I) \leftarrow \text{CIDEr}(s,I)$;
      estimate gradient $\nabla_\theta \hat{L}(\theta)$ using Eqn.(3);
    else
      sample a caption s according to q;
      $R(s,I) \leftarrow \text{max-CIDEr}(s,I)$;
      estimate gradient $\nabla_\theta \hat{L}(\theta)$ using Eqn.(4);
    update p_θ using $\nabla_\theta \hat{L}(\theta)$;
  end
end

4. Experiment

In this section, we first briefly introduce the dataset, evaluation metrics, and experimental settings, then report extensive results to illustrate effectiveness.

4.1. Dataset and Metrics

Dataset. MSCOCO [19] is the most popular benchmark for the task of image captioning. The dataset contains 82,783 training images, 40,504 validation images and 40,775 test images. Each image is associated with 5 human-annotated ground truths as references. In the experiment, we follow the widely adopted split in [16] with 113,287 images used for training, 5,000 for validation, and 5,000 for testing. Rare words appearing less than 5 times within the training set are replaced with a <UNK> token, resulting in the final vocabulary consisting of 9,487 words.

Accuracy Metrics. There are various influential criteria to evaluate accuracy of generated image captions including BLEU [22], METEOR [9] and CIDEr [31]. To provide a comprehensive evaluation on generating *multiple* accurate captions, we follow the implementation in [36] to evaluate the averaged scores within a Monte-Carlo sampled caption set according to the modeled posterior for each image. Since such scores are traditionally calculated on a single prediction for each image to derive corresponding scores, our implementation results in relatively lower scores compared with the published ones.

Diversity Metrics. We report three types of diversity metrics for evaluation: **1)** Unique Sentence Ratio (Uni.), which is the average ratio of distinct sentences in sampled sets; **2)** mBLEU-4, which is the averaged BLEU-4 score of each prediction with the rest captions in the sampled set counted as references; and **3)** self-CIDEr [36], which calculates singular vector decomposition (SVD) over autocorrelation matrices of the sampled caption set using CIDEr as the kernel. Note that higher scores indicate diverse outputs except for mBLEU-4, which is the lower the better.

4.2. Implementation Details

Model Structure. We implement the proposed method based on the Top-down model [2] with ResNet-101 [12] set as the backbone network for the encoder faster R-CNN [24]. The number of object regions per image is set to range from 10 to 100 adaptively, where feature of each object is a 2,048-dimensional vector. Words are first encoded as one-hot vectors and then embedded as 512-dimensional vectors before fed into the decoder. The sizes of hidden layers within each LSTM in the decoder are all set as 512.

Experiment Settings. Adam [17] is used as the optimizer during training. Parameters of the encoder are pre-trained according to [2] and fixed during training to save GPU memory usage. We set batch size as 16 and pre-train the model for 35 epochs using cross-entropy (CE) loss. The pre-trained parameters are used to provide a fair initialization for the subsequent RL training process. Learning rate is initialized as 5×10^{-4} and decays by 0.8 every 3 epochs during the CE training phase, and is fixed as 3×10^{-5} during RL training. We use the cosine distance between each ground truths and the merged semantics as the baseline term in policy gradient to reduce estimation variance. We conduct experiments on different settings of the hyper-parameter ϵ and find that 0.1 works fine for human-oriented performance.

4.3. Performance

We provide comprehensive evaluations on both accuracy and diversity performance to illustrate the effective-

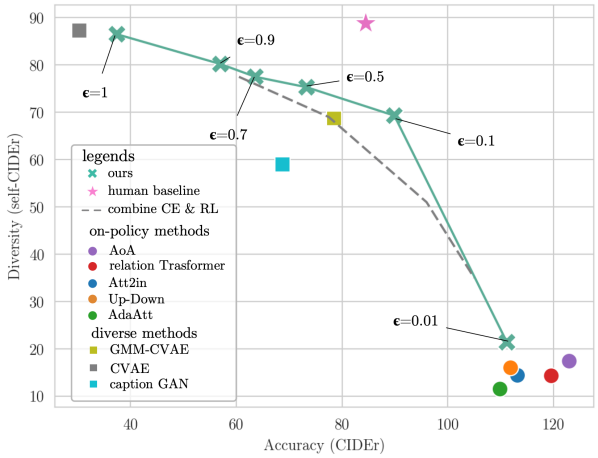


Figure 6. Performance of different works considering both diversity and accuracy. Different ϵ enables models to negotiate the trade-off between diversity and accuracy. With ϵ set as 0.1, our method locates closest to human performance.

ness of the proposed learning scheme. To evaluate the modeled posterior directly, we do not incorporate selection processes like oracle and consensus re-ranking [11], or decoding strategies like beam search or diverse beam search[32]. All the sentences are derived by directly sampled from the modeled posterior. Before we present the results, we need to quantify our goal, *i.e.* the scores of human-annotated captions. For diversity metrics, we regard the multiple ground truths provided by the dataset MSCOCO as sampled results from human posterior and calculated diversity scores. For accuracy evaluations, we compute the metrics in a leave-one-out manner following the implementation of [36, 37].

Fig.6 demonstrates the evaluation on both diversity (self-CIDEr) and accuracy (CIDEr) performance. From the figure, we can see that the scatter for some diverse captioning methods (*e.g.* CVAE [35]) distributes in the top-left portion, with high diversity (self-CIDEr) yet low accuracy (CIDEr). In contrast, the on-policy RL-based methods (*e.g.*, Up-Down [2], AoA [14], and relation Transformer [13]) are in the bottom-right portion, with high accuracy (CIDEr) yet low diversity (CIDEr). By setting the hyper-parameter ϵ as 0.1, our method balances the two aspects and acquires the closest performance to humans. The comparative results validate that the proposed partial off-policy learning scheme can enable the model to approximate a human-like posterior. We also compare our method with the one to combine CE and RL for balance [36], shown as the dashed line in Fig. 6. The curve acquired by our methods forms an upper envelope above such approach, indicating that we achieve a better balance between diversity and accuracy.

Comparison with Works to Promote Accuracy. We first compare with the recent works to promote accuracy in Table 1. Most of them are based on on-policy learning. It can be seen from the table that the proposed method acquires performance gains on multiple diversity metrics by a significant margin. To be specific, the state-of-the-art models achieve satisfactory performance on accuracy while performing poorly on diversity metrics. For example, AoA [14] obtains mBLEU-4 of 98.2, self-CIDEr of 17.4, Unique Sentence Ratio of 17.8. In contrast, our method with $\epsilon = 0.1$ achieves the mBLEU-4 of 54.0, self-CIDEr of 69.3, and Unique Sentence Ratio of 92.0. Recall that the objective of image captioning is supposed to be mimicking humans. Overall, we achieve close performance comparing with the human baseline on both accuracy and diversity.

	accuracy metrics			diversity metrics		
	C↑	M↑	B-4↑	mB-4↓	s-C↑	Uni.↑
Att2in [26]	113.2	27.0	34.8	98.5	14.4	14.8
Up-Down [2]	111.9	27.4	**35.3**	98.3	16.0	16.3
AdaAtt [20]	109.9	26.3	33.0	98.8	11.5	11.5
ReTrans [13]	119.6	29.3	28.0	98.5	14.3	14.2
AoA [14]	**123.0**	**29.7**	29.4	98.2	17.4	17.8
Ours ($\epsilon = 0.9$)	57.3	19.9	15.0	**27.3**	**80.6**	**99.6**
Ours ($\epsilon = 0.1$)	89.9	24.5	26.5	54.0	69.3	92.0
Ours ($\epsilon = 0.01$)	111.9	26.8	33.7	96.8	21.4	22.7
Human Performance	84.5	24.4	12.8	7.7	88.8	100.0

Table 1. Performances on Karpathy's test split of MSCOCO dataset compared with other on-policy-learning-based works.

	accuracy metrics			diversity metrics		
	C↑	M↑	B-4↑	mB-4↓	s-C↑	Uni.↑
CVAE[35]	30.3	15.0	6.8	**14.1**	**87.3**	**99.9**
GMM-CVAE[35]	78.5	21.7	18.9	45.6	70.7	90.9
CapGAN[27]	68.7	22.1	15.8	76.9	59.0	78.0
Ours ($\epsilon = 0.1$)	**89.9**	**24.5**	**26.5**	54.0	69.3	92.0
Human Performance	84.5	24.4	12.8	7.7	88.8	100.0

Table 2. Performances on Karpathy's test split of MSCOCO dataset compared with works for diverse captioning.

Comparison with Works to Promote Diversity. For comprehensive illustration, we present evaluations for diverse captioning models using generative frameworks like GAN[27] or VAE[35]. Our method outperforms the GAN-based framework in both accuracy and diversity. As for [35] using VAE framework, it obtains decent performance on diversity evaluations, but acquires accuracy scores far lower than the human baseline. For example, sampled results of CVAE[35] obtains averaged CIDEr, METEOR, and BLEU-4 of 30.3, 15.0, and 6.8 respectively over the sampled caption set. In contrast, our method achieves performance comparable with humans with the hyper-parameter ϵ set as 0.1, while preserving considerable diversity.

Correlation Analysis with Human Performance. Fig. 6 presents the close correlation between our method and human performance intuitively. To provide quantified evidence for such correlation, we calculate Pearson's correlation coefficient ρ between model and human performances over the validation set. In Table 3, we use Pearson's correlation coefficient on complex numbers to take multiple aspects into consideration. For each image in the dataset, we compose a complex number reflecting the comprehensive performance of the corresponding captions, where the real part is a certain accuracy score and the imaginary part

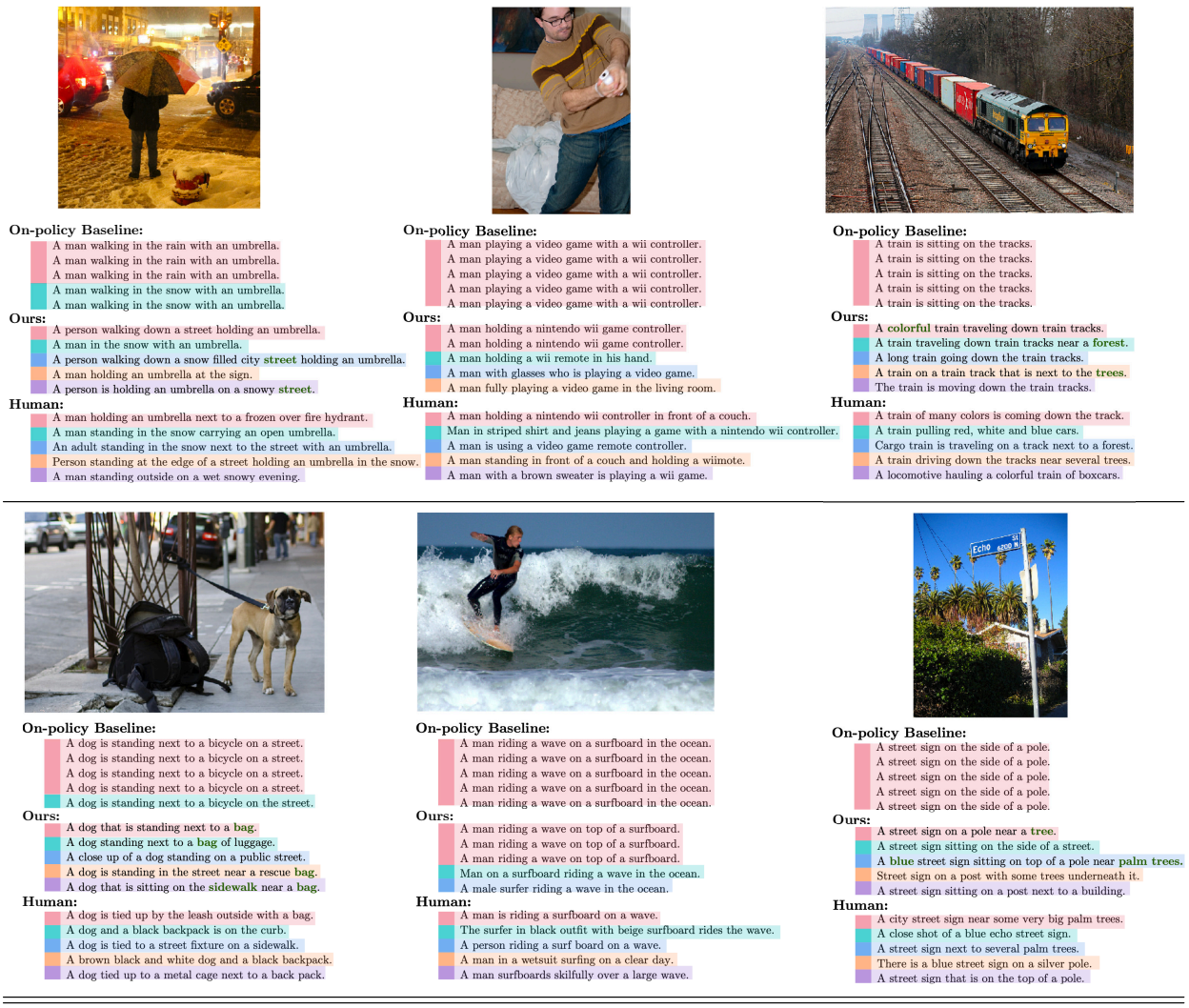


Figure 7. Example captions generated for images in Karpathy's test split of the MSCOCO dataset. We shade every distinct sentence within the sampled caption set using different colors. Our method yields diverse and descriptive outputs as humans do.

methods		Pearson's correlation $\|\rho\|$
accuracy methods	Att2in [26]	0.243
	Up-Down [2]	0.253
	AdaAtt [20]	0.236
	ReTrans [13]	0.295
	AoA [14]	0.289
diversity methods	CVAE [35]	0.200
	GMM-CVAE [35]	0.280
	CapGAN [27]	0.193
Ours ($\epsilon = 0.1$)		**0.337**

Table 3. Comprehensive correlation coefficient between predicted captions and human annotations on Karpathy's test split of MSCOCO dataset.

is a certain diversity score. Consistent with Fig. 6, we use CIDEr for accuracy and self-CIDEr for diversity here. Both accuracy and diversity scores used to compose the complex number are normalized according to the distribution of human performance. The model trained by the proposed partial off-policy learning scheme acquires the highest correlation with human performance, as shown in Table 3.

Qualitative Evaluation. We present several sampled captions according to the modeled posterior in Fig 7 for visualization. Consistent as we implement in Fig. 1, we sample several times using the trained model and observe the generated captions. In Fig. 7, the sample batch size for each image is set as 5 to provide a fair comparison with human performance. For comparison, the generated captions of the on-policy baseline model are quite repetitive. With the proposed partial off-policy learning, the deficiency in diversity is significantly alleviated. Extra semantics (*i.e.* snowy street, palm trees, rescue bags, etc.) are encouraged to be predicted by our method compared with the on-policy

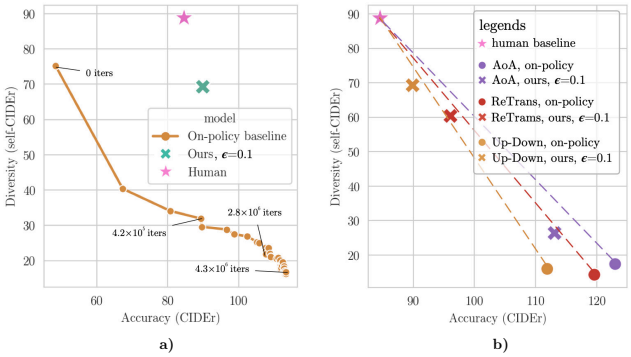


Figure 8. **a)** Result of on-policy baseline on the diversity-accuracy space during training, compared with our method and human performance. **b)** Effect of the proposed partial off-policy learning scheme applied on different model structures.

trained baseline, yielding more descriptive results. Moreover, the generated captions are more close to human annotations with balanced diversity and accuracy.

4.4. Further analysis

Is the diversity improvement caused by insufficient training? We notice that the model would become more accurate but less diverse as the on-policy training proceeds. So a straightforward question is whether our method achieves a diversity-accuracy balance similar to insufficient training under on-policy strategies. We conduct a further experiment to compare the on-policy baseline with the proposed partial off-policy method on both diversity and accuracy metrics. We report the diversity/accuracy metrics of on-policy learning with different training epochs and plot the curves in Fig. 8. Compared to on-policy, the metric point of our method lies above the curve of the on-policy baseline in the diversity-accuracy space, far closer to the human reference. The result indicates that we acquire much better diversity-accuracy balancing than the on-policy baseline, which validates the effectiveness of our method.

Is partial off-policy learning effective on other model structures? Since the proposed learning scheme involves no modifications on model structure, it is supposed to take effect on other image captioning models by simply substituting the on-policy training strategy with the partial off-policy one. We apply the partial off-policy learning scheme on [14, 34, 13] in Fig. 8. The result shows that the proposed learning scheme can promote human-like performance on multiple models. For AoA [14], the partial off-policy learning improves the self-CIDEr by 51.6%. For Up-down [2] and relation Transformer [13], we increase self-CIDEr by about 3 times. In other words, the proposed method is modular and can be easily combined with future works to promote human-oriented caption generation.

How does the off-policy sampling model interact with max-CIDEr? We conduct the ablation study over the modifications proposed in Sec. 3.2 and Sec. 3.3. Results are shown in Table 4.

	CIDEr	self-CIDEr	$\|\|\rho\|\|$
Baseline [2]	**111.9**	16.0	0.253
+ max-CIDEr optimization	111.4	20.7	0.266
+ sampling model	107.2	30.6	0.251
full	89.9	**69.3**	**0.337**
Human Performance	84.5	88.8	1.000

Table 4. Influence of each proposed modification.

As we can see, either optimizing max-CIDEr or introducing diverse behavior policy via sampling model can promote diversity. The self-CIDEr is improved from 16.0 to 20.7 with max-CIDEr optimization. It is improved to 30.6 with the diverse behavior strategy. Yet the proposed approaches present a more significant balance effect when adopted together. This is because that they are introduced to address issues on different dimensions as discussed in Sec. 3.1. Consequently, they strengthen the effect of each other and facilitate the acquisition of human-like performance.

Is partial off-policy learning complementary to sampling methods? Our method is a training-side method and aims at deriving a decent posterior directly. Theoretically, it can be further strengthened by sampling methods *e.g.* DBS [32]. We conduct experiments to evaluate the performances of DBS applied to the on-policy baseline and our methods. DBS acquires Pearson's correlation as **0.319** over the on-policy baseline while our method achieves the correlation as **0.337** without DBS. The combination of our method and DBS will boost the score up to **0.403**, suggesting that improvements obtained by partial off-policy are complementary to those provided by better sampling.

5. Conclusion

We present a novel learning scheme named *partial off-policy* for image captioning, encouraging human-like performance on both accuracy and diversity. We introduce sampling model and max-CIDEr reward. Such an off-policy strategy is then combined with traditional on-policy via a strategy selection procedure for accuracy-diversity balance. Our method locates closest in the diversity-accuracy space and achieves the highest correlation with human performance in a comprehensive perspective.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant No. 61771288, Cross-Media Intelligent Technology Project of Beijing National Research Center for Information Science and Technology (BNRist) under Grant No. BNR2019TD01022 and the research fund under Grant No. 2019GQG0001 from the Institute for Guo Qiang, Tsinghua University.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 1, 2

[2] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2, 5, 6, 7, 8

[3] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. Sequential latent spaces for modeling the intention during diverse image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4261–4270, 2019. 1, 3

[4] Fuhai Chen, Rongrong Ji, Jiayi Ji, Xiaoshuai Sun, Baochang Zhang, Xuri Ge, Yongjian Wu, Feiyue Huang, and Yan Wang. Variational structured semantic inference for diverse image captioning. In *Advances in Neural Information Processing Systems*, pages 1931–1941, 2019. 1, 3

[5] Jia Chen and Qin Jin. Better captioning with sequence-level exploration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10890–10899, 2020. 3

[6] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014. 2

[7] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10578–10587, 2020. 2

[8] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2970–2979, 2017. 3

[9] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 1, 2, 5

[10] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander G Schwing, and David Forsyth. Fast, diverse and accurate image captioning guided by part-of-speech. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10695–10704, 2019. 1, 3

[11] Jacob Devlin, Saurabh Gupta, Ross Girshick, Margaret Mitchell, and C Lawrence Zitnick. Exploring nearest neighbor approaches for image captioning. *arXiv preprint arXiv:1505.04467*, 2015. 1, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 5

[13] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, pages 11137–11147, 2019. 2, 6, 7, 8

[14] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 2, 6, 7, 8

[15] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015. 2

[16] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 5

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[18] Chin Yew Lin. Rouge: A package for automatic evaluation of summaries. In *In Proceedings of the Workshop on Text Summarization Branches Out (WAS 2004)*, 2004. 1, 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 2, 4, 5

[20] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 2, 6, 7

[21] Ruotian Luo and Gregory Shakhnarovich. Analysis of diversity-accuracy tradeoff in image captioning. *arXiv preprint arXiv:2002.11848*, 2020. 1

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002. 1, 2, 5

[23] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*, 2015. 1, 2

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015. 2, 5

[25] Z. Ren, X. Y. Wang, N. Zhang, X. T. Lv, and L. J. Li. Deep reinforcement learning-based image captioning with embedding reward. *30th Ieee Conference on Computer Vision and Pattern Recognition (Cvpr 2017)*, pages 1151–1159, 2017. 2, 3

[26] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference*

on Computer Vision and Pattern Recognition, pages 7008–7024, 2017. 1, 2, 3, 6, 7

[27] 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 *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 1, 3, 6, 7

[28] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014. 2

[29] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. 4, 5

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[31] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 1, 2, 4, 5

[32] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 3, 6, 8

[33] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015. 2

[34] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):652–663, 2016. 1, 2, 8

[35] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017. 1, 3, 6, 7

[36] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019. 3, 5, 6

[37] Qingzhong Wang, Jia Wan, and Antoni B Chan. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 6

[38] Zhuhao Wang, Fei Wu, Weiming Lu, Jun Xiao, Xi Li, Zitong Zhang, and Yueting Zhuang. Diverse image captioning via grouptalk. In *IJCAI*, pages 2957–2964, 2016. 1, 2

[39] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992. 3

[40] 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 *International conference on machine learning*, pages 2048–2057, 2015. 2

[41] Jianwei Yang, Jiasen Lu, Stefan Lee, Dhruv Batra, and Devi Parikh. Graph r-cnn for scene graph generation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018. 2

[42] Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai. Auto-encoding scene graphs for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10685–10694, 2019. 2

[43] Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei. Exploring visual relationship for image captioning. In *Proceedings of the European conference on computer vision (ECCV)*, pages 684–699, 2018. 2

[44] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4651–4659, 2016. 2

[45] Zhihao Zhu, Zhan Xue, and Zejian Yuan. Topic-guided attention for image captioning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2615–2619. IEEE, 2018. 1, 2