Generating Diverse and Accurate Visual Captions by Comparative Adversarial Learning

Dianqi Li[1*], **Qiuyuan Huang**[2], **Xiaodong He**[3*], **Lei Zhang**[2], **Ming-Ting Sun**[1]

[1]University of Washington, [2]Microsoft Research, [3]JD AI Research

{dianqili, mts}@uw.edu, xiaodong.he@jd.com, {leizhang, qihua}@microsoft.com

Abstract

We study how to generate captions that are not only accurate in describing an image but also diverse across different images. The problem is both fundamental and interesting, as most machine-generated captions, despite phenomenal research progresses in the past several years, are expressed in a very monotonic and featureless format. While such captions are normally accurate, they often lack important characteristics in human languages - distinctiveness for each image and diversity across different images. To address this problem, we propose a novel conditional generative adversarial network for generating diverse captions across images. Instead of estimating the quality of a caption solely on one image, the proposed comparative adversarial learning framework better assesses the quality of captions by comparing a set of captions within the image-caption joint space. By contrasting with human-written captions and image-mismatched captions, the caption generator effectively exploits the inherent characteristics of human languages, and generates more diverse captions. We show that our proposed network is capable of producing accurate and diverse captions across images.

Introduction

Image caption generation has attracted great attentions due to its wide applications in many fields, such as semantic image search, image commenting in social chat bot, and assistance to visually impaired people. Benefiting from recent advancements of deep learning, most existing works employ convolutional neural networks (CNNs) and deep recurrent language models, and have achieved great performance improvement on automatic evaluation metrics, such as BLEU (Papineni et al. 2002), CIDEr (Vedantam, Lawrence Zitnick, and Parikh 2015), etc.

Despite such successes, machine-generated captions are often in a generic format and can be easily differentiated from human-written captions, which tend to be more descriptive and diverse. As most state-of-the-art image caption algorithms are learning-based, to best match with the ground truth captions, such algorithms often produce high-frequency n-gram patterns or common expressions. As a result, the generated image captions receive high scores on

Images	Generated Captions	Binary Score	Comparative Score
	MLE: a man wearing a suit and tie	0.15	0.45
	G-GAN: a young man in business gear poses for the camera	0.18	0.59
	CAL (ours): a man with glasses wearing a striped tie and black suit	0.19	0.72
GT: a man with glasses and his eyes closed dressed in a black shirt and a necktie			
	MLE: a cow standing on the side of a street	0.50	0.73
	G-GAN: a brown cow standing in a city	0.49	0.68
	CAL (ours): a large cow walking on a side street in front of a door	0.52	0.82
GT: there is a cow on the sidewalk standing in front of a door			

Figure 1: Captions generated by MLE, conditional GANs (G-GAN) with binomial scores and our comparative adversarial learning network (CAL) with comparative scores. The shown scores are evaluated by the discriminators in G-GAN and CAL, respectively. The proposed adversarial framework estimates comparative scores by comparing a collection of captions, leading to more accurate and discriminative rewards for caption generator.

automatic evaluation metrics, yet lack a significant characteristic in human language - diversity across different images. From human perspectives, as demonstrated in (Jas and Parikh 2015), each image possesses its own specificity, and accordingly its related captions should acquire its distinctiveness, leading to diverse captions for different images. In general, distinctive descriptions are often pursued by human, who can easily distinguish a specific image among a group of similar images. In this work, our goal is to generate diverse and accurate captions which are similar to human-written descriptions.

Recent success of Generative Adversarial Networks (GANs) (Mirza and Osindero 2014) provides a possible way to generate diverse captions (Dai et al. 2017; Shetty et al. 2017). In this setting, a caption generator and a discriminator are jointly trained by a binomial distribution, which estimates the relevance and quality of the captions to the image. However, due to the large variability of natural language, a binary predictor is usually incapable of representing the

richness and diversity of captions. To ensure semantic relevance, a regularization term for distinguishing mismatched captions must be included during training.

In contrast to assigning an absolute score to a caption for one image, we noticed that it is relatively easier to distinguish the qualities of two captions by comparison. Motivated by this, we propose a comparative adversarial learning (CAL) network to learn human-like captions. Specifically, contrary to an absolute binary score for one caption, the quality of the caption is assessed relatively by comparing it with other captions in the image-caption space. In adversarial learning, the proposed discriminator ranks the human references, which are more specific and distinctive, higher than generic captions that have high-frequency n-gram patterns or common expressions. Consequently, with the guides from the discriminator, the generator effectively learns to generate more specific and distinctive captions, hence increases the diversity across the corpus. In summary, our main contributions lie in three aspects:

- We propose a novel comparative adversarial learning network, which is capable of generating more diverse and better captions across images by comparing different captions.

- By suppressing the scores of image-mismatched captions, especially for those from similar images, the proposed comparative learning framework can inherently ensure semantic relevance without involving an regularization term for mismatched captions.

- To effectively measure the caption diversity across images, we propose a new metric based on the semantic variance from caption embedding features. Additionally, experimental results clearly demonstrate the effectiveness of the proposed framework in terms of diversity and quality.

Related Work

Diverse Image Captioning

Most image captioning systems use an encoder-decoder framework which shares a similar idea as sequence learning (Sutskever, Vinyals, and Le 2014; Gehring et al. 2017; Vaswani et al. 2017; Peng et al. 2018; Peng et al. 2019). Typically, the networks are trained by maximum likelihood estimation (MLE) (Vinyals et al. 2015; Karpathy and Fei-Fei 2015; Xu et al. 2015; Gan et al. 2017) or reinforcement learning (Rennie et al. 2017; Ren et al. 2017; Liu et al. 2017; Anderson et al. 2018; Luo et al. 2018; Liu et al. 2018). Although such methods achieve outstanding performances on conventional evaluation metrics, such as BLEU, CIDEr, etc., the generated captions usually consist of high-frequency n-gram patterns but lose the diversity across images and thus are unnatural to human. To remedy this weakness, diverse beam search and ensemble methods (Vijayakumar et al. 2016; Wang et al. 2016) have been proposed. (Wang, Schwing, and Lazebnik 2017; Chatterjee and Schwing 2018) work on diverse image captioning by using variational auto-encoders. To achieve better caption diversity, (Dai et al. 2017; Shetty et al. 2017) incorporate generative adversarial nets (GANs) into image cap-
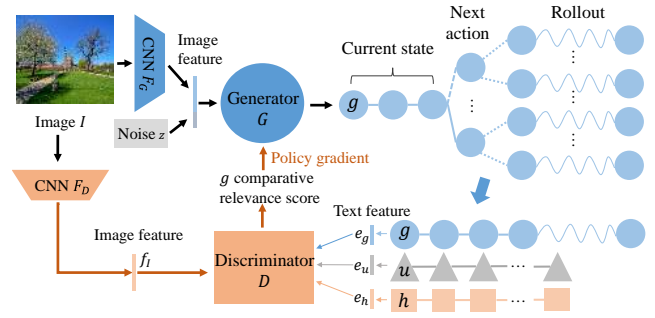

Figure 2: Comparative Adversarial Learning Network. The discriminator D is trained over comparative relevance scores for each image by comparing a generated caption g, a human-written caption h, and unrelated captions u. The generator G is optimized by policy gradient where the reward is estimated by the expectation of the comparative relevance score g over K rollout simulation captions.

tioning systems, with a binary-based discriminator. However, in sequence adversarial training, a binary-based discriminator is easily trained much stronger than the generator (Che et al. 2017; Guo et al. 2018; Lin et al. 2017), resulting in less distinguishable rewards or gradient vanishing problems for the generator (Figure 1). To generate captions with correct semantic relevance, (Dai et al. 2017; Shetty et al. 2017) must train the binary discriminator under an additional regularization.

In this work, we propose a comparative adversarial learning framework that explicitly estimates the quality of captions in a more discriminative way, which in turn helps the generator to produce more diverse captions while maintaining the caption correctness without the additional discriminator regularization.

Diversity Metrics

Automatic evaluation metrics such as BLEU, CIDEr-D, etc., have been widely applied for evaluating the quality of generated captions. Nonetheless, the evaluation of diversity across captions is still an open problem. Human language, inherited immense complexity and sophisticated interpretation, poses a thorny problem for developing standard criterion. (Li et al. 2015; Vijayakumar et al. 2016; Jain, Zhang, and Schwing 2017) measure the degree of diversity by analyzing distinct n-grams or word usages for generated sentences with respect to ground truths. This reflects an inventiveness for generated sentences, but not a diversity aspect among all the generated sentences. To estimate the caption diversity at the token level, (Shetty et al. 2017; Wang, Schwing, and Lazebnik 2017; Deshpande et al. 2018) inspect n-gram usage statistics and the size of vocabulary in all generate captions. However, the diversity of sentences is not only represented by various word or phrase usages, but also variant long-term sentence patterns and even implications of sentences. A simple n-gram statistics is unable to assess the diversity at the sentence level. In this paper, we propose a novel diversity metric based on semantic sentence features which compensate the defects of previous methods.

Comparative Adversarial Learning Network

As shown in Figure 2, the proposed Comparative Adversarial Learning (**CAL**) Network consists of a caption generator G and a comparative relevance discriminator (**cr-discriminator**) D. The two subnetworks play a min-max game as follow:

$$\min_\theta \max_\phi \mathcal{L}(G_\theta, D_\phi), \tag{1}$$

in which $\mathcal{L}$ is an overall loss function, while θ and ϕ are trainable parameters in G and D, respectively. Given a reference image I, the generator G_θ outputs a sentence g as the corresponding caption. D_ϕ aims at correctly estimating the comparative relevance score (**cr-score**) of g with respect to human-written caption h within the image-caption joint space. G_θ is trained to maximize the cr-score of g and generate human-like descriptions trying to confuse D_ϕ. We will elaborate each subnetwork in the following sections.

Caption Generator

Our caption generator G_θ is based on the standard encoder-decoder architecture (Vinyals et al. 2015). The captioning image encoder model F_G first extracts a fixed dimensional feature from image I using a CNN. Then a text decoder, implemented by a long short-term memory (LSTM) network, interprets the encoded feature $F_G(I)$ into a word sequence $g_{0:T} = (g_0, ..., g_T)$ to describe image I, where g_t is a token in time step t and T is the maximum time step. To produce captions with more variations, the encoded feature $F_G(I)$ can be concatenated with a random vector z. The notation of z will be ignored in the rest parts for simplicity. In time step t generation, the next token g_t can be sampled by:

$$g_t \sim \pi_\theta(g_t|I, g_{0:t-1}), \quad t \in (1, T) \tag{2}$$

π_θ is a word distribution, determined by inputs and θ, over all the words in vocabulary V. By sequentially sampling or greedy decoding words according to π_θ, a complete caption $g_{1:T}$ can be generated by captioner G_θ. In comparative adversarial training, G_θ expects to produce better captions with higher cr-scores. However, unlike cross-entropy loss in the MLE method, the cr-score of $g_{1:T}$ estimated by discriminator D_ϕ is based on discrete tokens, whose gradients cannot be directly employed for G_θ through back-propagation. Therefore, we adopt a common technique - Policy Gradient method (Sutton et al. 2000) to solve this gradient issue. The details will be discussed in Section 3.3.

Comparative Relevance Discriminator

(Dai et al. 2017) propose to estimate the semantic relevance, naturalness, and quality of a generated caption by a logistic function over the similarity between the caption and the given image. However, an absolute binary value is very restrictive to evaluate them all, especially the quality of a caption. To evaluate a discriminative score, it is more justifiable to compare a generated caption with other captions, primarily with human-written caption h. Therefore, we formulate a comparative relevance score (**cr-score**) to measure an over-
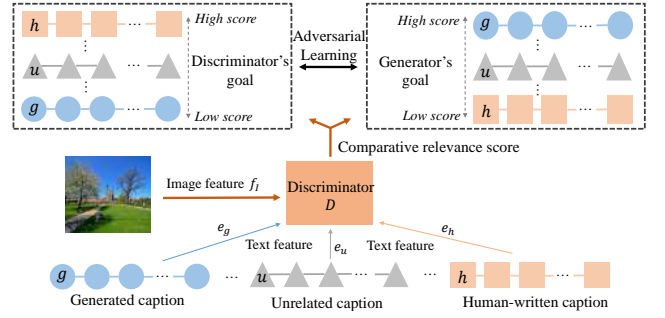

Figure 3: Training objectives in our adversarial learning. While the discriminator desires to judge human-written captions correctly with higher cr-scores, the generator aims to produce captions with higher cr-scores and thus confusing the discriminator.

all image-text quality of caption c by comparing a set of captions $\mathcal{C}^c$ given image I:

$$D_\phi(c|I, \mathcal{C}^c) = \frac{exp(\gamma S(e_c, f_I)}{\sum_{c' \in \mathcal{C}^c} exp(\gamma S(e_{c'}, f_I))} \tag{3}$$

where $\mathcal{C}^c$ denotes a set of captions including c, and the cr-score of c is what we care about here. e_c and f_I are the text feature and image feature extracted by the text encoder and CNN image encoder F_{D_ϕ} in discriminator D_ϕ, respectively. The cosine similarity between e_c and f_I is defined as $S(e_c, f_I) = (e_c^T f_I)/(\|e_c\|\|f_I\|)$. γ is an empirical parameter defined by validation experiment. A higher γ leads $D_\phi(c|I, \mathcal{C}^c)$ towards the caption that better matches with image I. $D_\phi(c|I, \mathcal{C}^c))$ estimates the cr-score of caption c by comparing with other captions in the image-caption joint space - a higher score represents caption c is superior in $\mathcal{C}^c$. To obtain more accurate cr-score for c, it is favorable to include human-written caption h for image I in $\mathcal{C}^c$. In this case, the cr-score of c contains a discrepancy information between caption c and human-written caption h. The discriminator is designed to differentiate generated captions from human-written captions for image I. Specifically, from the discriminator's perspective, a human-written caption desires to receive a higher cr-score, whereas a generated caption should receive a lower cr-score (Figure 3). Hence, the objective function to be maximized for discriminator can be defined as:

$$\mathbb{E}_{h \sim \mathcal{P}_h} \left[\log D_\phi(h|I, \mathcal{C}^h)\right] + \mathbb{E}_{g \sim G_\theta} \left[\log(1 - D_\phi(g|I, \mathcal{C}^g)\right] \tag{4}$$

where $\mathcal{P}_h(I)$ represents human-written caption distribution given image I. Set $\mathcal{C}^h$ and $\mathcal{C}^g$ encloses a human-written caption h, a machine-generated caption g, and other unrelated captions u. In experiments, u can be directly obtained from image-mismatched captions in one mini-batch.

Policy Gradient Optimization for G_θ

In contrast to D_ϕ, the caption generator G_θ attempts to maximize the cr-scores of machine-generated captions and thus fool the discriminator (Figure 3). However, the cr-scores of a generated caption g are assessed by D_ϕ based on a series

of sequential discrete samples, which are non-differentiable during training. We address this problem by a classic policy gradient method (Sutton et al. 2000). Considering in each time step t, the generation of each word g_t is an action of an "agent" G_θ from policy π_θ according to the current state $(I, g_{0:t-1})$. An intermediate reward r for this action is approximated as the expected future reward:

$$Q_\theta(g_t|I, g_{0:t-1}) = \mathop{\mathbb{E}}_{g_{t+1:T}} [r(g_{0:t-1}, g_t, g_{t+1:T}|I)] \quad (5)$$

The action reward r can be any metric, including the cr-score from D_ϕ. Unfortunately, the discriminator cannot provide a score unless a complete sentence is generated. The lack of intermediate rewards will result in a gradient vanishing problem. To imitate an accurate intermediate reward, following (Yu et al. 2017), we deploy a K-times Monte Carlo rollout process conditioned on the current caption generator G_θ to explore the rest unknown words $g_{t+1:T}$. Then the intermediate reward for action g_t can be approximated by the expected cr-score over K rollout simulation captions:

$$Q_{\theta,\phi}(g_t|I, g_{0:t-1}) \simeq \frac{1}{K} \sum_{k=1}^{K} D_\phi(g_{k,t}|I, C^{g_k,t}) \quad (6)$$

where $g_{k,t} = g_{0:t-1}, g_t \oplus \bar{g}_{t+1,T}$, and $\bar{g}_{t+1,T}$ is sampled from G_θ by the rollout method. Besides, $t \in (1, T-1)$, as g_0 is a start token and G_θ receives an accurate reward once generating a full sequence. $C^{g_k,t}$ contains a simulated caption $g_{k,t}$, a human-written caption h and other unmatched descriptions u, corresponding to the image I. To train the generator, the objective is to optimize the policy and adjust the generator to receive a maximum long-term reward - higher cr-scores for generated captions in each time step (Fig. 3). In the end, the gradient for updating generator G_θ can be finalized by:

$$\mathop{\mathbb{E}}_{g \sim G_\theta} \sum_{t=1}^{T} \nabla_\theta \pi_\theta(g_t|I, g_{0:t-1}) \cdot Q_{\theta,\phi}(g_t|I, g_{0:t-1}) \quad (7)$$

where g_t is an intermediate token belonging to g at time step t. The goal of the generator is to maximize the expected cr-scores of generated captions.

Comparisons with Previous Models

During discriminator training, (Dai et al. 2017; Shetty et al. 2017) introduce a regularization term to learn image-caption matching by minimizing binary scores of mismatched captions u (last term in the below equation):

$$\mathop{\mathbb{E}}_{h,g,u} \log D_b(h|I) + \log(1 - D_b(g|I)) + \log(1 - D_b(u|I)) \quad (8)$$

where D_b is a binary discriminator. However, the cr-discriminator D_ϕ can naturally learn such image-caption matching by placing mismatched captions in the comparison set C^h with true captions h and generated captions g. Specifically, by enlarging the cr-score of the matched image-caption pair (h, I) in set C^h, D_ϕ can consistently distinguish its corresponding caption from others, and suppress the scores for mismatched descriptions u (Equation 3). D_ϕ can in turn help the caption generator G_θ produce diverse

Model	BLEU4	METEOR	ROUGE	CIDEr	SPICE
Human	0.190	0.240	0.465	0.861	**0.208**
MLE	**0.297**	**0.252**	**0.519**	**0.921**	0.175
G-GAN	0.208	0.224	0.467	0.705	0.156
CAL (ours)	0.213	0.225	0.472	0.721	0.161

Table 1: Performance comparisons on MSCOCO test set. In human result, a sentence randomly sampled from ground-truth annotations is evaluated by the rest annotations for each image.

captions for corresponding images, ensuring semantic relevances of generated captions. Meanwhile, the binary discriminator D_b separates the decisions on g and h. The proposed network simply combines the two separate decisions into a single ranking process. The cr-score of the generated captions are estimated by contrasting human-written sentences subject to image I. This can assist the cr-score to comprise more informative guidances, including both naturalness and quality from ground truths, benefiting the training of the caption generator G_θ.

Experiments

Models. To test the effectiveness of the proposed Comparative Adversarial Learning (CAL) network, we compare two baseline models:
(1) *MLE*: We use LSTM-R (Gan et al. 2017) based on the mainstream CNN-LSTM architecture as our MLE baseline model. The training follows the standard MLE method.
(2) *Adversarial models*: We use G-GAN (Dai et al. 2017) as the baseline model for diverse image captioning (**G** represents the generator). The corresponding discriminator D_b outputs a binary score in $[0, 1]$ through a logistic function over the dot product between image and text features (Equation 8).

To make a fair comparison, all image features for generators and discriminators are extracted by *ResNet-152* (He et al. 2016) (we reimplement *G-GAN* by using *ResNet-152* network as image encoders). All text-decoders in generators and text-encoders in discriminators are implemented by LSTMs. More details related to the networks are included in Appendices A.1

Training. Before adversarial training, the caption generator G_θ in both adversarial models is pretrained by the standard MLE method (Vinyals et al. 2015) (Karpathy and Fei-Fei 2015) for 20 epochs, and the cr-discriminator is pretrained according to Equation 4 for 10 epochs. During the experiment, we found the generator pretraining is necessary, otherwise it will encounter mode collapse problem and generate nonsense captions. On the other hand, pretraining discriminator helps more stable training later. In the adversarial learning stage, two sub-networks are trained jointly, in which every one generator iteration is followed by 5 discriminator iterations. We set the learning rate to 0.0005 and the batch size to 64. In each mini-batch with 64 image-caption
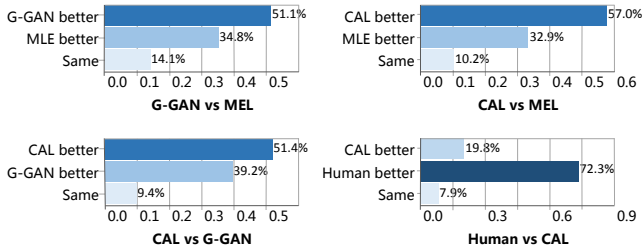
Figure 4: Human evaluation results by comparing model pairs. The majority of respondents agree that our proposed CAL generates better captions than the two baselines. The numbers in the figure represent the ratio of total survey cases.

Category	MLE	G-GAN	CAL (ours)	Human
Bathroom	2.733	6.145	6.501	**9.066**
Computer	3.710	6.012	7.228	**8.943**
Pizza	3.837	5.779	6.805	**9.117**
Building	4.019	5.940	6.088	**9.344**
Cat	4.196	5.225	6.473	**9.155**
Car	4.968	5.910	6.661	**8.741**
Daily supply	5.056	6.204	7.330	**9.075**
*All**	6.947	7.759	8.812	**9.465**

Table 2: Diversity evaluations across various image categories. All^* denotes all the images in MSCOCO test set.

pairs, all other 63 captions that are not corresponding to the correct image are used as the unrelated captions during training. The rollout number K is empirically set to 16, and γ is set to 10. During testing, the generated captions are sampled based on policy and the one with the best cr-score is chosen for evaluation.

Dataset. We conduct all experiments on the MSCOCO dataset (Lin et al. 2014). MSCOCO contains 123,287 images, each being annotated with at least 5 human-written captions. All our experiments are based on the public split method from (Karpathy and Fei-Fei 2015): 5000 images for both validation and testing, and the rest for training.

Evaluation. We evaluate the generated captions based on both the correctness and diversity metrics, which guarantee the generation quality in each aspect. While the correctness of the generated captions is measured by common captioning metrics, the diversity across various images is evaluated by the proposed metric based on caption embedding features.

Consider each image is annotated by one caption, whose embedding feature is extracted by a same text encoder. Ideally, all embedding features are identical and the feature variance is zero if all the images have same captions. Conversely, a large variance would present if all the captions were distinct. Thus, the variance across embedding features reflects the diversity of captions on a semantic-level. To measure the variance, all the text embedding features can be concatenated into a feature matrix $A \in \mathbb{R}^{m \times n}$, where m is the number of captions and n is the dimensions of the embedding feature. To estimate the correlation σ_i in each dimension, the covariance matrix $M \in \mathbb{R}^{n \times n}$ of A can be computed. Then, σ_i can be obtained by singular value decomposition (SVD): $M = U\Sigma V^T$, where $\Sigma = diag(\sigma_0, ..., \sigma_{n-1})$; U and V^T are $m \times m$ and $n \times n$ unitary matrix.

Finally, we use l_1-norm $\hat{\sigma} = \sum_{i=0}^{n-1} |\sigma_i|$ to evaluate an overall variance in all dimensions among caption embedding features. A large variance $\hat{\sigma}$ suggests the embedding features of captions are less akin or correlated, representing more distinctive expressions and larger diversity among image captions.

Accuracy. We first evaluate the generated captions from different models on five automatic metrics: BLEU4 (Papineni et al. 2002), METEOR (Banerjee and Lavie 2005), ROUGE_L (Lin 2004), CIDEr-D (Vedantam, Lawrence Zitnick, and Parikh 2015) and SPICE (Anderson et al. 2016). As can be seen in Table 1, although our method CAL slightly outperforms the baseline G-GAN, the standard MLE model yields remarkably better results, even outperforms human. However, as discussed by (Dai et al. 2017; Shetty et al. 2017), these evaluation metrics overly focus on n-grams matching with ground truth captions and ignore other important human language factors such as diversity. The captions, written with variant expressions, have fewer n-grams matched with ground truths. As a result, captions with novel expressions from the human and adversarial models receive lower scores on these metrics. These metrics particularly represent the quality of pattern matching, instead of the overall quality from human perspective.

Human Evaluation. To correlate with human judgments on the correctness of captions, we conducted human evaluation experiments on Amazon Mechanical Turk. Specifically, we randomly sampled 300 images from test set. Then, given an image with two generated captions from different models, subjects are asked to choose one caption that best describes the image. We received more than 9000 responses in total and the results are summarized in Figure 4. It can be seen that the majority of people consider the captions from G-GAN and especially our CAL better than those from the standard MLE method. This illustrates that despite both adversarial models perform poorly on automatic metrics, the generated captions are of higher quality in terms of human views. In the comparison between CAL and G-GAN, our model can generates more human-like captions that receive more acknowledgements. This demonstrates that, by exploiting more comparative relevance information against ground truth and other captions instead of solely on image, the proposed CAL effectively improves the caption generator and achieves better captions. We include failure analysis in Appendices A.3.

Diversity. To compare the capabilities of generating diverse expressions, we measure the variances of captions from different models across images. All the embedding features are

Bathroom			
MLE	a bathroom with a toilet and a sink	a bathroom with a toilet and a sink	a bathroom with a sink and a mirror	a bathroom with a sink and a toilet
G-GAN	a bathroom with a white toilet and tiled walls	a restroom with a toilet sink and shower	a bathroom with a white bathtub and two sinks and a mirror	a pink restroom with a toilet inside of it
CAL	a toilet sits inside of a bathroom next to a wall	a narrow bathroom with a toilet sink and a shower with dirty walls	a clean bathroom with a large sink bathtub and a mirror	a pink bathroom with a sink toilet and mirror
Pizza			
MLE	a pizza sitting on top of a white plate	a pizza sitting on top of a white plate	a close up of a pizza on a table	a pizza sitting on top of a pan
G-GAN	a pizza on a plate on a wooden table	a pizza sitting on a plate next to a glass of wine	the pizza is covered with cheese and tomatoes	a close up of a sliced pizza on a plate
CAL	a cheese pizza on a plate sits on a table	a plate of pizza and a glass of beer on the table	a pizza topped with lots of toppings is ready to be cut	a partially eaten pizza is being cooked on a pan
Car			
MLE	a green truck parked in a parking lot	a black truck is parked in a parking lot	a group of buses driving down a street	a city bus stopped at a bus stop
G-GAN	a green garbage truck in a business district	an antique black car sitting in a parking lot	a city street filled with taxis and buses	people are waiting in line as the bus travel down the road
CAL	a large green truck driving past a tall building	an old style truck parked in a parking space near a building	the city buses are driving through the traffic	people gather to a street where a bus get ready to board

Figure 5: Qualitative results illustrate that adversarial models, especially our proposed CAL, can generate more diverse descriptions.

extracted using the same text-encoder in our framework. Besides estimating the variance across all the images in the test set, we also inspect the variances inside different image categories. Particularly, We use the K-means method to cluster the input image features, and select six clusters with high-frequency topics. All the results are summarized in Table 2. It can be seen that despite the MLE method performs well on automatic metrics, the variance of captions is relatively lower across different images. As shown in Figure 5, the MLE model often generates similar expressions and meanings within one category, even if the images are distinct.

In contrast to the MLE model, both adversarial models, especially our proposed CAL, can generate more diverse captions with respects to distinct images. G-GAN uses a binary discriminator to separates the decisions on machine-generated and human-written captions. Compared to G-GAN, our network trained by comparative learning binds the information of human-written captions which possess highest diversity characteristics as indicated in Table 2. The comparative learning also encourages the distinctiveness of the generated captions by suppressing the cr-scores of mismatched captions, especially for those from akin images. These allow our caption generator to produce more descriptive captions for different images. As expected, the variance of captions from our model is larger than that from G-GAN across all the images. Similar trends can be observed inside different categories. We also notice that the word usage in our CAL is more diverse and akin to that of human than other baselines (Appendices A.2). These suggest that our proposed CAL has better generative capability than the baseline G-GAN and helps bridge the gap between machine-generated and human-written captions. Figure 6 shows that

z_1	a man holding a rope while skateboarding a small wave	a baseball game is in progress with the large crowd watching	twin skiers stand on skis on a snowy hillside	a crowd of people are having a meal and drinking beer at a table
z_2	a person is trying to flip while skiing water	a person in jersey holding a baseball bat in swing position	two people in ski gear standing on a snowy slope	a group of people are eating outside at a restaurant
z_3	there is a male surfer that is riding a wave	a batter catcher and umpire during a baseball game	two skiers on a thick snow covered mountain	many people are sitting on tables at a restaurant

Figure 6: Qualitative results illustrate the images captioned by CAL with different random vector z.

Adversarial Model				Diversity
bs	samp	noise	compa	
✓				7.078
	✓			7.331
	✓	✓		7.784
	✓	✓	✓	8.845

Table 3: Ablation study of caption diversity of our adversarial model. *bs* and *samp* indicate beam search and sampling decoding respectively. *noise* denotes adding noise vectors in decoding. *compa* represents our discriminator with comparative learning.

Model	R@1	R@3	R@5	R@10
MLE	3.03	8.67	12.75	20.54
G-GAN w/o reg.	14.24	30.28	40.11	56.13
G-GAN	16.07	33.66	43.80	59.74
CAL (ours)	**18.81**	**36.56**	**46.84**	**62.57**

Table 4: Caption-image matching comparison evaluated on MSCOCO test set. Captions are all self-generated by each model. *G-GAN w/o reg.* denotes the G-GAN model without the regularization term in the discriminator. The recall ratio is calculated by ranking discriminator's scores based on caption-image pairs.

the CAL model is able to generate diverse captions for each image.

Ablation. We study the diversity effect of each component in our network on MSCOCO val set. The results are summarized in Table 3, where we can see that the sampling decoding and noise vectors bring a certain amount of diversity. The proposed comparative relevance discriminator compares different captions and maximizes the scores of the generated captions among a set of references, resulting in an even larger diversity gain.

Network Effectiveness. We further investigate the effectiveness of adversarial models by a caption-image matching experiment (Dai et al. 2017; Dai and Lin 2017). Specifically, if all the generated captions have enough diversity and the adversarial discriminator is good enough to distinguish related and unrelated image-caption pairs, the corresponding image could be easily retrieved by the discriminator when given its own caption. For each adversarial model, we can use its generated caption as a query to rank all images, based on the similarity scores from corresponding discriminators. Then a recall ratio can be calculated by inspecting the top-k resulting images in the ranked list. Since the MLE model is not an adversarial model, we use the discriminator from G-GAN to retrieve the generated captions, providing a baseline for comparison.

Table 4 shows the performance comparison of different models. Although captions from MLE commonly describe images well, they are less diverse for different images, resulting in a poor retrieval performance. Meanwhile, our proposed CAL outperforms all the other models, including the adversarial model G-GAN. This further demonstrates that CAL can produce more diverse captions for all images. In Table 4, it is noteworthy that the G-GAN model needs an regularization term to sustain better semantic relevance of captions. Without such regularization, our CAL model still improves discernibility on caption-image pairs. It proves that the cr-discriminator in our proposed network can provide more accurate rewards during adversarial training, leading to a better caption generator.

Conclusions

We presented a comparative adversarial learning network for generating diverse captions across images. A novel comparative learning schema is proposed for the discriminator, which better assesses the quality of captions by comparing with other captions. Thus more caption properties including correctness, naturalness, and diversity can be taken into consideration. This in turn benefits the caption generator to effectively exploit inherent characteristics inside human languages and generate more diverse captions. We also proposed a new caption diversity metric in the semantic level across images. Experimental results clearly demonstrate that our proposed method generates better captions in terms of both accuracy and diversity across images.

References

[Anderson et al. 2016] Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *ECCV*.

[Anderson et al. 2018] Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*.

[Banerjee and Lavie 2005] Banerjee, S., and Lavie, A. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *ACL Workshop*.

[Chatterjee and Schwing 2018] Chatterjee, M., and Schwing, A. G. 2018. Diverse and coherent paragraph generation from images. *arXiv preprint arXiv:1809.00681 2.*

[Che et al. 2017] Che, T.; Li, Y.; Zhang, R.; Hjelm, R. D.; Li, W.; Song, Y.; and Bengio, Y. 2017. Maximum-likelihood augmented discrete generative adversarial networks.

[Dai and Lin 2017] Dai, B., and Lin, D. 2017. Contrastive learning for image captioning. In *NIPS*, 898–907.

[Dai et al. 2017] Dai, B.; Fidler, S.; Urtasun, R.; and Lin, D. 2017. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, 2989–2998.

[Deshpande et al. 2018] Deshpande, A.; Aneja, J.; Wang, L.; Schwing, A.; and Forsyth, D. A. 2018. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589.*

[Gan et al. 2017] Gan, Z.; Gan, C.; He, X.; Pu, Y.; Tran, K.; Gao, J.; Carin, L.; and Deng, L. 2017. Semantic compositional networks for visual captioning. In *CVPR*.

[Gehring et al. 2017] Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; and Dauphin, Y. N. 2017. Convolutional sequence to sequence learning. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1243–1252. JMLR. org.

[Guo et al. 2018] Guo, J.; Lu, S.; Cai, H.; Zhang, W.; Yu, Y.; and Wang, J. 2018. Long text generation via adversarial training with leaked information. *AAAI*.

[He et al. 2016] He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

[Jain, Zhang, and Schwing 2017] Jain, U.; Zhang, Z.; and Schwing, A. 2017. Creativity: Generating diverse questions using variational autoencoders. In *CVPR*.

[Jas and Parikh 2015] Jas, M., and Parikh, D. 2015. Image specificity. In *CVPR*, 2727–2736.

[Karpathy and Fei-Fei 2015] Karpathy, A., and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 3128–3137.

[Li et al. 2015] Li, J.; Galley, M.; Brockett, C.; Gao, J.; and Dolan, B. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055.*

[Lin et al. 2014] Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

[Lin et al. 2017] Lin, K.; Li, D.; He, X.; Zhang, Z.; and Sun, M.-T. 2017. Adversarial ranking for language generation. In *NIPS*, 3155–3165.

[Lin 2004] Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL Workshop*.

[Liu et al. 2017] Liu, S.; Zhu, Z.; Ye, N.; Guadarrama, S.; and Murphy, K. 2017. Improved image captioning via policy gradient optimization of spider. In *ICCV*, volume 3.

[Liu et al. 2018] Liu, X.; Li, H.; Shao, J.; Chen, D.; and Wang, X. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *arXiv preprint arXiv:1803.08314.*

[Luo et al. 2018] Luo, R.; Price, B.; Cohen, S.; and Shakhnarovich, G. 2018. Discriminability objective for training descriptive captions. In *CVPR*, 6964–6974.

[Mirza and Osindero 2014] Mirza, M., and Osindero, S. 2014. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784.*

[Papineni et al. 2002] Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

[Peng et al. 2018] Peng, H.; Schwartz, R.; Thomson, S.; and Smith, N. A. 2018. Rational recurrences. In *In Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[Peng et al. 2019] Peng, H.; Parikh, A. P.; Faruqui, M.; Dhingra, B.; and Das, D. 2019. Text generation with exemplar-based adaptive decoding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

[Ren et al. 2017] Ren, Z.; Wang, X.; Zhang, N.; Lv, X.; and Li, L.-J. 2017. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*.

[Rennie et al. 2017] Rennie, S. J.; Marcheret, E.; Mroueh, Y.; Ross, J.; and Goel, V. 2017. Self-critical sequence training for image captioning. In *CVPR*.

[Shetty et al. 2017] Shetty, R.; Rohrbach, M.; Hendricks, L. A.; Fritz, M.; and Schiele, B. 2017. Speaking the same language: Matching machine to human captions by adversarial training. In *ICCV*.

[Sutskever, Vinyals, and Le 2014] Sutskever, I.; Vinyals, O.; and Le, Q. V. 2014. Sequence to sequence learning with neural networks. In *NIPS*.

[Sutton et al. 2000] Sutton, R. S.; McAllester, D. A.; Singh, S. P.; and Mansour, Y. 2000. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*.

[Vaswani et al. 2017] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, 5998–6008.

[Vedantam, Lawrence Zitnick, and Parikh 2015] Vedantam, R.; Lawrence Zitnick, C.; and Parikh, D. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.

[Vijayakumar et al. 2016] Vijayakumar, A. K.; Cogswell, M.; Selvaraju, R. R.; Sun, Q.; Lee, S.; Crandall, D.; and Batra, D. 2016. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*.

[Vinyals et al. 2015] Vinyals, O.; Toshev, A.; Bengio, S.; and Erhan, D. 2015. Show and tell: A neural image caption generator. In *CVPR*, 3156–3164.

[Wang et al. 2016] Wang, Z.; Wu, F.; Lu, W.; Xiao, J.; Li, X.; Zhang, Z.; and Zhuang, Y. 2016. Diverse image captioning via grouptalk. In *IJCAI*.

[Wang, Schwing, and Lazebnik 2017] Wang, L.; Schwing, A.; and Lazebnik, S. 2017. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NIPS*, 5756–5766.

[Xu et al. 2015] Xu, K.; Ba, J.; Kiros, R.; Cho, K.; Courville, A.; Salakhutdinov, R.; Zemel, R. S.; and Bengio, Y. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*.

[Yu et al. 2017] Yu, L.; Zhang, W.; Wang, J.; and Yu, Y. 2017. Seqgan: Sequence generative adversarial nets with policy gradient. In *AAAI*, 2852–2858.

Appendices

Implementation Details.

Following (Karpathy and Fei-Fei 2015), we convert all the captions to lowercases and remove its non-alphabet characters. We also discard the tokens with frequency less than 5 in the training dataset, resulting in a vocabulary size of 8,791. Both image encoders F_{G_θ} and F_{D_ϕ} in the generator and discriminator are implemented using *ResNet* (He et al. 2016) with 152 layers, separately. The image activations in the *pool5* layer are extracted, yielding 2048-dimensional image features. Noise vector z with 100-dimensions is sampled from a uniform distribution. All the image features are projected to 512 dimensions by fully connected layers. The text-decoder in the generator and the text-encoder in the discriminator are all implemented using LSTMs with 512 hidden nodes. We use the last hidden activations from the text-encoder as text feature, which shares the same dimension with the projected image feature.

Diversity visualization

As distinctive image contents are described by specific words, caption diversity across images could be visualized by diverse word usages. For this purpose, we can inspect the word usage frequency at each position t of the generated captions:

$$p(w_t) = \mathbb{E}_{w_{t-1}} p(w_t|w_{t-1})p(w_{t-1}), \quad t \in (1,T)$$

where w_0 only takes one word which is the "START" token and thus $p(w_0 = \text{"START"}) = 1$, and the image notation I is ignored for simplification. The above probability can be estimated by the Markov Chain method over the generated vocabulary distribution (Equation 2). However, as the word space is too large, it is practically difficult to calculate the frequency distribution for all the words. Instead, we use a sampling method to approximate the frequency of the observed words in the generated captions:

$$p(w_t) \simeq count(w_t)/count(\bar{w}_t) \quad t \in (1,T)$$

where $count(\bar{w}_t)$ denotes the total count of all the observed words at position t. Ideally, diverse word usage implies that each word in the vocabulary is used less repeatedly in caption generation across different images, leading to lower $p(w_t)$ for each word.

To visualize the word usage frequency, we sampled 300 images from the test set and visualize the statistics in Figure 7. Here we chose 300 but not more images for the sake of visualization clarity. As can be seen, the MLE-generated captions usually pick up fewer content words such as "sitting", "riding", and "standing", regardless of different and distinctive image contents. Meanwhile, its corresponding * regions are much narrower than those in other models, meaning that it rarely uses other words in the vocabulary. In contrast, although our CAL model also uses the same function words (e.g. "a", "the", "of", "is", etc.), most of the used content words are less identical and contribute to much wider * regions. We also find that our CAL uses more adjectives and adverbs in generated captions.

By comparison, we can conclude that the word frequency distribution of our CAL is more akin to that of Human. This demonstrates that our CAL model has more diverse word usages than the baseline G-GAN, resulting in more distinctive captions across images.

Failure analysis

For some images with complex content, we find the CAL-generated captions are imprecise or defective in describing their contents. One possible reason is that if a complicate image does not have a focused topic, its ground truths are normally divergent for different aspects of the image. As a result, during comparative adversarial learning, the caption generator can not simultaneously capture all the opposed details from ground truths. Additionally, to generate more descriptive and diverse captions for images, our framework bears risks that involving some incorrect details. Figure 8 shows some failure examples from our method. We will consider these problems in our future study.
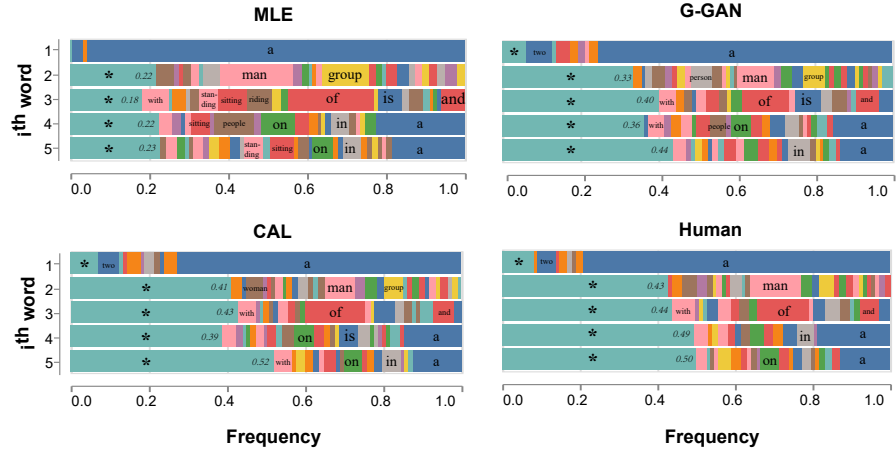
Figure 7: Visualization of word diversity produced by different models. For each subgraph, the ith row represents the ith word's frequency distribution in all the generated captions. Different colors denote different words, and the width of each region is proportional to the frequency of the corresponding word. We only plot the first five words for easy readability. We mark the words of high frequency in the figure. The words of low frequency ($< 0.5\%$) are merged into *the others* (denoted as *****) category. Larger proportion of ***** means more chances of using diverse words (or long tail words) in the vocabulary. The decimal in each ***** region denotes its proportion value for easy comparison.

a police officer directing traffic at a bus stop

a display store of various old speckled benches

Figure 8: Failure examples from the proposed network.