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, {leizhang, qihua}@microsoft.com, xiaodong.he@jd.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.

1 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 trained by maximum likelihood estimation (MLE) [6, 9, 22, 24] or reinforcement learning [2, 12, 13, 14, 18, 17], and have achieved great performance improvement on automatic evaluation metrics, such as BLEU [16], CIDEr [21], etc.

Despite such successes, machine-generated captions can still 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 automatic evaluation metrics, yet lack a significant characteristic in human language - diversity across different images. However, as demonstrated in [8], distinctive descriptions are often pursued by human, who can easily distinguish a specific image among a group of similar images. Therefore, diverse and descriptive captions across images are essential to the goal of generating human-like captions.

Recent success of Generative Adversarial Networks (GANs) [15] provides a possible way to generate diverse captions [4, 19]. 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.

*The work was done while Dianqi Li and Xiaodong He were at Microsoft Research.

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

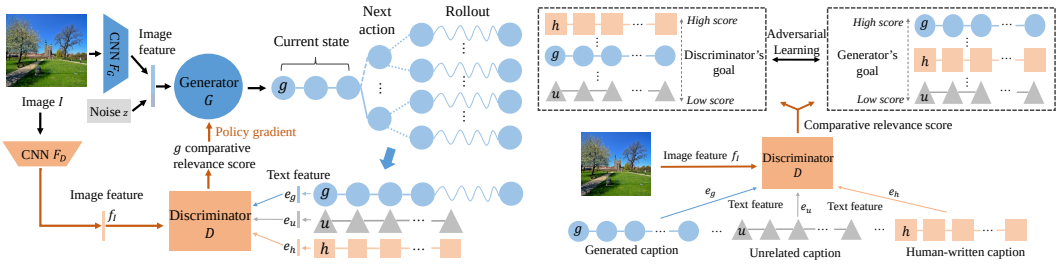

Figure 1: (a) The proposed Comparative Adversarial Learning Network. (b) The training objectives.

However, due to the large variability of natural language, a binary predictor is usually incapable of representing the richness and diversity of captions. Moreover, to ensure semantic relevance in this binary setting, a regularization term for 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. Meanwhile, 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. Our work can be highlighted in three aspects: (i) We propose a novel comparative adversarial learning network to generate more diverse and better captions. (ii) The proposed model can inherently ensure semantic relevance without involving an extra regularization term for mismatched captions. (iii) To effectively measure the caption diversity across images, we propose a new metric based on the semantic variance from caption embedding features.

2 Comparative Adversarial Learning Network

As shown in Fig. 1, 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 and optimize the loss function $\mathcal{L}$:

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

in which θ and ϕ are trainable parameters in caption generator G and cr-discriminator D, respectively. Given a reference image I, the generator G_θ outputs a sentence g as the caption for I. Meanwhile, the cr-discriminator 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. Then G_θ is trained to maximize the cr-score of g and generate human-like descriptions trying to confuse the cr-discriminator D_ϕ.

Comparative Relevance Discriminator Compared to **G-GAN** [4] which uses an absolute binary discriminator solely on each caption, the proposed **CAL** network measures an overall image-text quality of caption c by comparing a set of captions $\mathcal{C}^c$ given image I:

$$D_\phi(c|I, \mathcal{C}^c) = \frac{exp(\gamma S(e_c, f_I)}{\sum_{c' \in \mathcal{C}^c} exp(\gamma S(e_{c'}, f_I))}, \quad where \quad S(e_c, f_I) = \frac{e_c^T f_I}{\|e_c\| \|f_I\|} \tag{2}$$

$\mathcal{C}^c$ 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. γ is an empirical parameter defined by validation experiment.

$D_\phi(c|I, \mathcal{C}^c))$ 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

Model	BLEU4	METEOR	ROUGE	CIDEr	SPICE
Human	0.190	0.240	0.465	0.861	**0.208**
MLE [6]	**0.297**	**0.252**	**0.519**	**0.921**	0.175
G-GAN [4]	0.208	0.224	0.467	0.705	0.156
CAL (ours)	0.213	0.225	0.472	0.721	0.161

Table 1: Metric performances from different models on the MSCOCO test set.

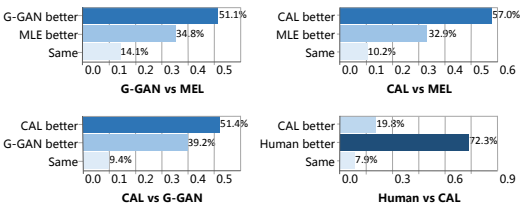

Figure 2: Human evaluation results by comparing model pairs.

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 (Fig. 1(b)). Hence, the objective function to be maximized for discriminator can be defined as:

$$\mathop{\mathbb{E}}_{h,I\sim\mathcal{P}_h(I)}\Big[\log D_\phi(h|I,\mathcal{C}^h)\Big] + \mathop{\mathbb{E}}_{g,I\sim G_\theta(I,z)}[\log(1-D_\phi(g|I,\mathcal{C}^g)] \tag{3}$$

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.

Caption Generator G_θ Our caption generator G_θ is based on the standard encoder-decoder architecture (Fig. 1(a)). 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 [20]. The gradient for updating generator G_θ during adversarial training can be formulated as:

$$\mathop{\mathbb{E}}_{g,I\sim G_\theta(I,z)}\sum_{t=1}^{T}\nabla_\theta\pi_\theta(g_t|I,g_{0:t-1})\cdot D_{\theta,\phi}(g_{0:t}|I,C^{g_{k,t}}) \tag{4}$$

where $g_{0:t}$ is a partial sentence belonging to g at generating time step t. π_θ is the word probability when generating token g_t at time step t. By connecting the generator with the cr-score, the goal is to maximize the expected reward, encouraging the discriminator to acknowledge the generated captions with higher cr-scores.

3 Experiments

Implementation details To test the effectiveness of the proposed Comparative Adversarial Learning (CAL) network, we use LSTM-R [6] and *G-GAN* [4] as our MLE and adversarial baseline model, respectively. To make a fair comparison, all image features for generators and discriminators are extracted by *ResNet-152* [7]. Both adversarial models take random vectors as extra input. All text-decoders in generators and text-encoders in discriminators are implemented by LSTMs. During testing, the generated captions are sampled based on policy in adversarial models or by greedy in the MLE model. We conduct all experiments on the MSCOCO image caption dataset [11].

Accuracy We first evaluate the generated captions from different models on five automatic metrics: BLEU4 [16], METEOR [3], ROUGE_L [10], CIDEr-D [21] and SPICE[1]. 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 reported in previous works [4] [19], automatic evaluation metrics overly focus on n-grams matching with ground truth captions and ignore other important factors in human language. As a result, captions written with variant expressions usually receive lower scores than those largely fitting with annotations. Thus, the automatic metrics only partially reflect the caption correctness.

To align the criterion with human, we also provide the results from human evaluations, in which the subjects are asked to choose best caption when comparing two captions given the corresponding image. We received more than 9000 responses in total and the results are summarized in Fig. 2. 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

Table 2: Diversity evaluations across various image categories. All denotes all the categories.*

Category \ Model	MLE [6]	G-GAN [4]	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**

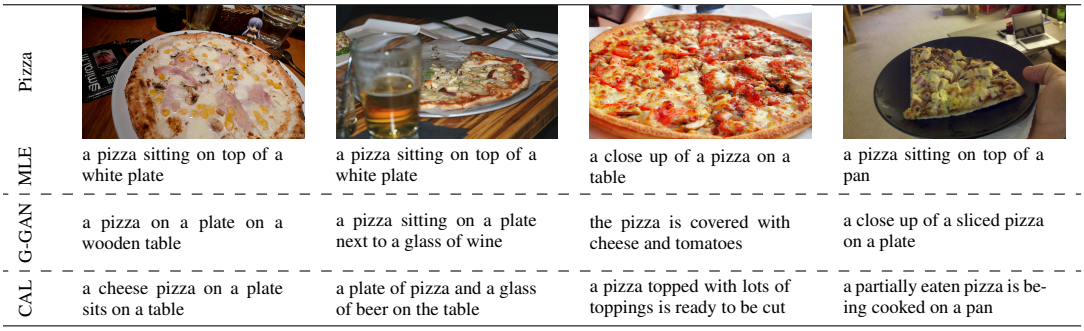

Figure 3: Qualitative results of diverse captions across images.

views. Meanwhile, the comparison between CAL and G-GAN suggests that the captions generated from our model receive more acknowledgements when comparing with those provided by baselines. This demonstrates that, by exploiting more comparative relevance information against ground truth and other captions instead of solely on one image, the proposed CAL effectively improves the caption generator and achieves better captions.

Diversity Previous works evaluate the generated caption diversity by analyzing n-grams or word usages statistics [5, 19, 23]. We argue that 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. Therefore, we propose a novel diversity metric based on the embedding features of sentences. Specifically, we calculate the variance of all generated captions based on the embedding features, which reflects the diversity of captions on a semantic-level. All the caption embedding features are extracted using the same text-encoder in our framework. The detailed formulations can be found in the supplementary materials. During the experiment, we cluster the similar images into different categories, and calculate the variance of generated captions within each category.

As can be seen in Table 2, despite the MLE method performs well on automatic metrics, the variance of captions is relatively lower across different images. Fig. 3 shows some qualitative results for different categories. We find that 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. 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. More qualitative results are included in the supplementary materials.

4 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

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

[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 *CVPR*, 2018.

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

[4] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *ICCV*, pages 2989–2998, 2017.

[5] Aditya Deshpande, Jyoti Aneja, Liwei Wang, Alexander Schwing, and David A Forsyth. Diverse and controllable image captioning with part-of-speech guidance. *arXiv preprint arXiv:1805.12589*, 2018.

[6] Zhe Gan, Chuang Gan, Xiaodong He, Yunchen Pu, Kenneth Tran, Jianfeng Gao, Lawrence Carin, and Li Deng. Semantic compositional networks for visual captioning. In *CVPR*, 2017.

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

[8] Mainak Jas and Devi Parikh. Image specificity. In *CVPR*, pages 2727–2736, 2015.

[9] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, pages 3128–3137, 2015.

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

[11] 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 *ECCV*, 2014.

[12] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, volume 3, 2017.

[13] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. *arXiv preprint arXiv:1803.08314*, 2018.

[14] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *CVPR*, pages 6964–6974, 2018.

[15] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

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

[17] Zhou Ren, Xiaoyu Wang, Ning Zhang, Xutao Lv, and Li-Jia Li. Deep reinforcement learning-based image captioning with embedding reward. In *CVPR*, 2017.

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

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

[20] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In *NIPS*, 2000.

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

[22] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, pages 3156–3164, 2015.

[23] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *NIPS*, pages 5756–5766, 2017.

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

Supplementary Material

Diversity Metric

To demonstrate diversity across various images, we propose a novel metric based on the embedding features of sentences. Consider each image is annotated by one caption, whose embedding feature is extracted by a same text encoder. Ideally, all embedding features are identical if all the images have same captions. As a result, the variance among all the caption feature vectors would be zero. 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. If we sketch the m caption vectors in an n-dimensional space, the $m \times n$-dimensional matrix A can be enclosed by a hyperellipse in $\mathbb{R}^n$. In each orthogonal direction i, the principle semiaxis of hyperellipse can be measured by a scale factor σ_i, on behalf of the standard variance in this axis. Correspondingly, the variance of captions in each dimension i can be approximated by σ_i, where $i \in [0, n-1]$. To estimate σ_i, the correlation in these n-dimensions can be computed by the covariance matrix $M \in \mathbb{R}^{n \times n}$ of A. 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, respectively.

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.

Caption Diversity across Images

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