Towards Diverse and Natural Image Descriptions via a Conditional GAN

Bo Dai\textsuperscript{1}  Sanja Fidler\textsuperscript{23}  Raquel Urtasun\textsuperscript{234}  Dahua Lin\textsuperscript{1}
\textsuperscript{1}Department of Information Engineering, The Chinese University of Hong Kong
\textsuperscript{2}University of Toronto  \textsuperscript{3}Vector Institute  \textsuperscript{4}Uber Advanced Technologies Group
db014@ie.cuhk.edu.hk  fidler@cs.toronto.edu  urtasun@cs.toronto.edu  dhlin@ie.cuhk.edu.hk

Abstract

Despite the substantial progress in recent years, the image captioning techniques are still far from being perfect. Sentences produced by existing methods, e.g. those based on RNNs, are often overly rigid and lacking in variability. This issue is related to a learning principle widely used in practice, that is, to maximize the likelihood of training samples. This principle encourages high resemblance to the “ground-truth” captions, while suppressing other reasonable descriptions. Conventional evaluation metrics, e.g. BLEU and METEOR, also favor such restrictive methods. In this paper, we explore an alternative approach, with the aim to improve the naturalness and diversity – two essential properties of human expression. Specifically, we propose a new framework based on Conditional Generative Adversarial Networks (CGAN), which jointly learns a generator to produce descriptions conditioned on images and an evaluator to assess how well a description fits the visual content. It is noteworthy that training a sequence generator is nontrivial. We overcome the difficulty by Policy Gradient, a strategy stemming from Reinforcement Learning, which allows the generator to receive early feedback along the way. We tested our method on two large datasets, where it performed competitively against real people in our user study and outperformed other methods on various tasks.

1. Introduction

Generating descriptions of images has been an important task in computer vision. Compared to other forms of semantic summary, e.g. object tagging, linguistic descriptions are often richer, more comprehensive, and a more natural way to convey image content. Along with the recent surge of deep learning technologies, there has been remarkable progress in image captioning over the past few years [11, 29–32]. Latest studies on this topic often adopt a combination of an LSTM or its variant and a CNN. The former is to produce the word sequences while the latter is to capture the visual features of the images.

The advance in image captioning has been marked as a prominent success of AI\textsuperscript{1}. It has been reported [29, 30] that with certain metrics, like BLEU [23] or CIDEr [28], state-of-the-art techniques have already surpassed human’s performance. A natural question to ask is then: \textit{has the problem of generating image descriptions been solved?} Let us take a step back, and look at a sample of the current results. Figure 1 shows two vivid scenes together with three sentences produced by the Encoder-and-Decoder model [29] (marked as “G-MLE”), a state-of-the-art caption generator. Though faithfully describing the content of the images, these sentences feel rigid, dry, and lacking in vitality.

This is not surprising. Our brief survey (see Section 2)

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure1}
\caption{This figure shows two images with descriptions generated by humans, an LSTM net trained with our GAN-based framework (G-GAN), and an LSTM net trained with MLE (G-MLE). The last two columns compare the metric values of BLEU-3 and E-GAN, the evaluator trained using our method. As we can see, the sentences generated by G-GAN are more natural and demonstrate higher variability, as compared to those by G-MLE. Also, the E-GAN metrics are more consistent with human’s evaluations, while BLEU only favors those that significantly overlap with the training samples in detailed wording.}
\end{figure}

\textsuperscript{1}ARTIFICIAL INTELLIGENCE AND LIFE IN 2030, https://ai100.stanford.edu/2016-report
shows that existing efforts primarily focus on *fidelity*, while other essential qualities of human languages, e.g. *naturalness* and *diversity*, have received less attention. More specifically, mainstream captioning models, including those based on LSTMs [8], are mostly trained with the (conditional) maximum likelihood objective. This objective encourages the use of the n-grams that appeared in the training samples. Consequently, the generated sentences will bear high resemblance to training sentences in *detailed wording*, with very limited variability in expression [4]. Moreover, conventional evaluation metrics, such as BLEU [23], METEOR [14], ROUGE [18], and CIDEr [28], tend to favor this “safe” but restricted way. Under these metrics, sentences that contain matched n-grams would get substantially higher scores than those using variant expressions [1]. This issue is manifested by the fact that human descriptions get considerably lower scores.

Motivated to move beyond these limitations, we explore an alternative approach in this work. We wish to produce sentences that possess three properties: (1) **Fidelity**: the generated descriptions should reflect the visual content faithfully. Note that we desire the fidelity in *semantics* instead of *wording*. (2) **Naturalness**: the sentences should *feel* like what real people would say when presented with the image. In other words, when these sentences are shown to a real person, she/he would ideally not be able to tell that they are machine-generated. (3) **Diversity**: the generator should be able to produce notably different expressions given an image – just like human beings, different people would describe an image in different ways.

Towards this goal, we develop a new framework on top of the Conditional GAN [22]. GAN has been successfully used in image generation. As reported in previous works [9, 24], they can produce *natural* images nearly indistinguishable from real photos, freely or constrained by conditions. This work studies a different task for the GAN method, namely, generating *natural* descriptions conditioned on a given image. To our best knowledge, this is the first time the GAN method is used for image description.

Applying GANs to text generation is nontrivial. It comes with two significant challenges due to the special nature of linguistic representation. First, in contrast to image generation, where the transformation from the input random vector to the produced image is a deterministic continuous mapping, the process of generating a linguistic description is a *sequential sampling* procedure, which samples a *discrete* token at each step. Such operations are non-**differentiable**, making it difficult to apply back-propagation directly. We tackle this issue via *Policy Gradient*, a classical method originating from reinforcement learning [27]. The basic idea is to consider the production of each word as an *action*, for which the reward comes from the evaluator. By approximating the stochastic policy with a parametric function approximator, we allow gradients to be back-propagated.

Second, in the conventional GAN setting, the generator would receive feedback from the evaluator when an entire sample is produced. For sequence generation, this would lead to several difficulties in training, including *vanishing gradients* and *error propagation*. To mitigate such difficulties, we devise a mechanism that allows the generator to get early feedback. Particularly, when a description is *partly* generated, our framework would calculate an approximated *expected future reward* through Monte Carlo rollouts [34]. Empirically, we found that this significantly improves the efficiency and stability of the training process.

Overall, our contributions can be briefly summarized as follows: (1) We explore an alternative approach to generate image descriptions, which, unlike most of the previous work, encourages not only *fidelity* but also *naturalness* and *diversity*. (2) From a technical standpoint, our approach relies on the conditional GAN method to learn the generator, instead of using MLE, a paradigm widely adopted in state-of-the-art methods. (3) Our framework not only results in a generator that can produce natural and diverse expressions, but also yields a description evaluator at the same time, which, as we will show in our experiments, is substantially more consistent with human evaluation.

## 2. Related Work

**Generation.** Generating descriptions for images has been a long standing topic in computer vision. Early studies mostly adopted *detection-based* approaches. Such methods first detect visual concepts (e.g. object categories, relationships, and attributes) using CRFs [2, 6, 12], SVMs [16], or CNNs [5, 17], then generate descriptions thereon using simple methods, such as sentence templates [12, 16], or by retrieving relevant sentences from existing data [5, 6, 13, 15].

In recent years, the Encoder-and-Decoder paradigm proposed in [29] became increasingly popular. Many state-of-the-art frameworks [21, 29–32, 35] for this task adopt the *maximum likelihood* principle for learning. Such a framework usually works as follows. Given an image , it first derives a feature representation , and then generates the words sequentially, following a Markov process conditioned on . The model parameters are learned via maximum likelihood estimation (MLE), i.e., maximizing the conditional log-likelihood of the training samples, as:

\[
\sum_{(I_i, S_i) \sim \mathcal{D}} \sum_{t=0}^{T_i} \log p \left( w_i^{(t)} | f(I), w_i^{(t-1)}, \ldots, w_i^{(t-n)} \right) \quad (1)
\]

Here, and are the image and the corresponding descriptive sentence of the *i*-th sample, and *n* is the order of the Markov chain – the distribution of the current word depends on *n* preceding words. Along
with the popularity of deep neural networks, latest studies often adopt neural networks for both image representation and language modeling. For example, [30] uses a CNN for deriving the visual features $f(I)$, and an LSTM [8] net to express the sequential relations among words. Despite the evolution of the modeling choices, the maximum likelihood principle remains the predominant learning principle.

As illustrated in Figure 2, when similar images are presented, the sentences generated by such a model often contain repeated patterns [3]. This is not surprising – under the MLE principle, the joint probability of a sentence is, to a large extent, determined by whether it contains the frequent n-grams from the training set. Therefore, the model trained in this way will tend to produce such n-grams. In particular, when the generator yields a few of words that match the prefix of a frequent n-gram, the remaining words of that n-gram will likely be produced following the Markov chain.

**Evaluation.** Along with the development of the generation methods, various evaluation metrics have been proposed to assess the quality of the generated sentences. Classical metrics include BLEU [23] and ROUGE [18], which respectively focuses on the precision and recall of n-grams. Beyond them, METEOR [14] uses a combination of both the precision and the recall of n-grams. CIDEr [38] uses weighted statistics over n-grams. As we can see, such metrics mostly rely on matching n-grams with the “ground-truths”. As a result, sentences that contain frequent n-grams will get higher scores as compared to those using variant expressions, as shown in Figure 3. Recently, a new metric SPICE [1] was proposed. Instead of matching between n-grams, it focuses on those linguistic entities that reflect visual concepts (e.g. objects and relationships). However, other qualities, e.g. the naturalness of the expressions, are not considered in this metric.

**Our Alternative Way.** Previous approaches, including both generation methods and evaluation metrics, primarily focus on the resemblance to the training samples. While this is a safe way to generate plausible descriptions, it is limited. For example, when presented an image, different people would probably give different descriptions that do not overlap much in the wording patterns. This diversity in expression is an essential property of human languages, which, however, is often overlooked in previous works (both generation and evaluation). In this work, we explore an alternative approach – instead of emphasizing n-gram matching, we aim to improve the naturalness and diversity, i.e. generating sentences that feel like what real people would say, rather than focusing on word-by-word matching. Specifically, our approach jointly trains a generator $G$ and an evaluator $E$ in an adversarial way, where $G$ is to produce natural descriptions, while $E$ is to distinguish irrelevant or artificial descriptions from natural ones.

From a technical standpoint, our approach is based on the conditional GAN approach. GANs [7] and conditional GANs [22] are popular formulations for learning generators. For computer vision, GAN was originally introduced to generate images [24]. In a recent work [34], a text generator based on the GAN method was proposed. Note that this is an unconstrained generator that does not take into account any conditions. Hence, it can not be directly used for generating descriptions for images – in this task, the relevance of the generated text to the given image is essential. To our best knowledge, this is the first study that explores the use of conditional GAN in generating image descriptions.

### 3. Framework

We propose a new framework for generating image descriptions based on the conditional GAN [22] method, which consists of a generator $G$, and an evaluator $E$. Given
an image \(I\), the former is for generating natural and semantically relevant descriptions; while the latter is for evaluating how well a sentence or paragraph describes \(I\). We start with generating single sentences as descriptions, and then extend our framework to paragraph generation.

### 3.1. Overall Formulation

Our framework contains a generator \(G\) and an evaluator \(E\), whose structures are respectively shown in Figure 4 (a) and (b). It is worth noting that our framework is orthogonal to works that focus on architectural designs of the \(G\) and the \(E\). Their structures are not restricted to the ones introduced in this paper. In our framework, given an image \(I\), the generator \(G\) takes two inputs: an image feature \(f(I)\) derived from a convolutional neural network (CNN) and a random vector \(z\). In particular, we follow the setting in NeuralTalk\(^2\), adopting VGG16 [26] as the CNN architecture. The random vector \(z\) allows the generator to produce different descriptions given an image. One can control the diversity by tuning the variance of \(z\). With both \(f(I)\) and \(z\) as the initial conditions, the generator relies on an LSTM [8] net as a decoder, which generates a sentence, word by word. Particularly, the LSTM net assumes a sequence of latent states \((s_0, s_1, \ldots)\). At each step \(t\), a word \(w_t\) is drawn from the conditional distribution \(p(w_t|s_t)\).

The evaluator \(E\) is also a neural network, with an architecture similar to \(G\) but operating in a different way. Given an input \(I\) and a descriptive sentence \(S = (w_0, w_1, \ldots)\), it embeds them into vectors \(f(I)\) and \(h(S)\) of the same dimension, respectively via a CNN and an LSTM net. Then the quality of the description, i.e. how well it describes \(I\), is measured by the dot product of the embedded vectors, as

\[
r_\eta(I, S) = \sigma \left( \langle f(I, \eta_I), h(S, \eta_S) \rangle \right).
\]

Here, \(\eta = (\eta_I, \eta_S)\) denotes the evaluator parameters, and \(\sigma\) is a logistic function that turns the dot product into a probability value in \([0, 1]\). Note that while the CNN and the LSTM net in \(E\) have the same structure as those in \(G\), their parameters are not tied with each other.

For this framework, the learning objective of \(G\) is to generate descriptions that are natural, i.e. indistinguishable from what humans would say when presented with the same image; while the objective of \(E\) is to distinguish between artificial descriptions (i.e. those from \(G\)) and the real ones (i.e. those from the training set). This can be formalized into a minimax problem as follows:

\[
\min_\theta \max_\eta \mathcal{L}(G_\theta, E_\eta).
\]

Here, \(G_\theta\) and \(E_\eta\) are a generator with parameter \(\theta\) and an evaluator with parameter \(\eta\). The objective function \(\mathcal{L}\) is:

\[
\mathbb{E}_{S \sim \mathcal{P}_I} \left[ \log r_\eta(I, S) \right] + \mathbb{E}_{z \sim \mathcal{N}_0} \left[ \log (1 - r_\eta(I, G_\theta(I, z))) \right].
\]

Here, \(\mathcal{P}_I\) denotes the descriptive sentences for \(I\) provided in the training set, \(\mathcal{N}_0\) denotes a standard normal distribution, and \(G_\theta(I, z)\) denotes the sentence generated with \(I\) and \(z\). The overall learning procedure alternates between the updating of \(G\) and \(E\), until they reach an equilibrium.

This formulation reflects an essentially different philosophy in how to train a description generator as opposed to those based on MLE. As mentioned, our approach aims at the semantical relevance and naturalness, i.e. whether the generated descriptions feel like what human would say, while the latter focuses more on word-by-word patterns.

### 3.2. Training \(G\): Policy Gradient & Early Feedback

As mentioned, unlike in conventional GAN settings, the production of sentences is a discrete sampling process, which is nondifferentiable. A question thus naturally arises - how can we back-propagate the feedback from \(E\) under such a formulation? We tackle this issue via Policy Gradient [27], a technique originating from reinforcement learning. The basic idea is to consider a sentence as a sequence of actions, where each word \(w_t\) is an action. The choices of such “actions” are governed by a policy \(\pi_\eta\).

With this interpretation, the generative procedure works as follows. It begins with an empty sentence, denoted by \(S_{1:0}\), as the initial state. At each step \(t\), the policy \(\pi_\eta\) takes the conditions \(f(I), z\), and the preceding words \(S_{1:t-1}\) as inputs, and yields a conditional distribution \(\pi_\eta(w_t|f(I), z, S_{1:t-1})\) over the extended vocabulary.

\[\text{Figure 4: The structures of the generator } G \text{ for both single sentences and paragraphs, and the evaluator } E \text{ for single sentences.}\]
namely all words plus an indicator of sentence end, denoted by $e$. This computation is done by moving forward along the LSTM net by one step. From this conditional distribution, an action $w_t$ will be sampled. If $w_t = e$, the sentence will be terminated, otherwise $w_t$ will be appended to the end. The *reward* of this sequence of actions $S$ is $r_{\eta}(I, S)$, the score given by the evaluator $E$.

Now, we have defined an action space, a policy, and a reward function, and it seems that we are ready to apply the reinforcement learning method. However, there is a serious issue here—a sentence can only be evaluated when it is completely generated. Following the argument in [24] we extend Eq (4) to consider three types of descriptions for each training image $I$: (1) $S_I$: the set of descriptions for $I$ provided by human, (2) $S_G$: those from the generator $G_\theta$, and (3) $S_{I,1}$: the human descriptions for different images, which is uniformly sampled from all descriptions that are not associated with the given image $I$. To increase the scores for the descriptions in $S_I$ while suppressing those in the others, we use a joint objective formulated as:

$$\max_{\eta} \mathcal{L}_E(\eta) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_E(I_i; \eta).$$  \hfill (7)

Here, $N$ is the number of training images. The term for each image $I_i$ is given by:

$$\mathcal{L}_E(I_i; \eta) = \mathbb{E}_{S \in S_i} \log r_{\eta}(I, S) + \alpha \cdot \mathbb{E}_{S \in S_G} \log (1 - r_{\eta}(I, S)) + \beta \cdot \mathbb{E}_{S \in S_{I,1}} \log (1 - r_{\eta}(I, S)).$$  \hfill (8)

The second term forces the evaluator to distinguish between the human descriptions and the generated ones, which would in turn provide useful feedbacks to $G_\theta$, pushing it to generate more natural descriptions. The third term, on the other hand, ensures the semantic relevance, by explicitly suppressing mismatched descriptions. The coefficients $\alpha$ and $\beta$ are to balance the contributions of these terms, whose values are empirically determined on the validation set.

### 3.4. Extensions for Generating Paragraphs

We also extend our framework to generate *descriptive paragraphs* by adopting a *Hierarchical LSTM* design. Specifically, our extended design is inspired by [11]. As shown in part (c) of Figure 4, it comprises two LSTM levels—a *sentence-level* LSTM net and a *word-level* LSTM net. Given the conditions $f(I)$ and $z$, to produce a paragraph, it first generates a sequence of vectors based on $f(I)$, each encoding the topics of a sentence. Then for each sentence, it generates the words conditioned on the corresponding topic and the random vector $z$.

For evaluating a paragraph, the evaluator $E$ also adopts a hierarchical design, but reversing the steps. Given an image $I$ and a paragraph $P$, it first embeds each sentence into a vector via a word-level LSTM net, and then embeds the entire paragraph by combining the sentence embeddings via a sentence-level LSTM net. Finally, it computes the score by taking the dot product between the paragraph embedding $p$ and the image representation $f(I)$, and turning it into a probability as $\sigma(p^T f(I))$, where $\sigma$ is the logistic function.
Figure 5: The figure shows the human comparison results between each pair of generators. With names of the generators placed at each side of the comparison, the blue and orange areas respectively indicate percentages of the generator in the left and right being the better one.

<table>
<thead>
<tr>
<th></th>
<th>R@1</th>
<th>R@3</th>
<th>R@5</th>
<th>R@10</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>G-MLE</td>
<td>5.06</td>
<td>12.28</td>
<td>18.24</td>
</tr>
<tr>
<td></td>
<td>G-GAN</td>
<td>14.30</td>
<td>30.88</td>
<td>40.06</td>
</tr>
<tr>
<td>P</td>
<td>G-MLE</td>
<td>9.88</td>
<td>20.12</td>
<td>27.30</td>
</tr>
<tr>
<td></td>
<td>G-GAN</td>
<td>12.04</td>
<td>23.88</td>
<td>30.70</td>
</tr>
</tbody>
</table>

Table 2: The recalls of image rankings for different generators. Here recalls is the ratio of the original image being in the top-k in the ranked lists. The ranks are based on the similarities (S) between a image and a description, estimated by E-GAN, as well as the log-likelihoods (P), computed by different generators.

After pretraining, we fix the sentence-level LSTM net of $G$ and only update the word-level LSTM net of $G$ during the CGAN learning procedure. This can effectively reduce the cost of Monte Carlo rollouts. With a fixed sentence-level LSTM net, the policy gradients for each sentence will be computed separately, following the steps in Sec 3.2. Other parts of the training procedure remain the same.

4. Experiment

Datasets We conducted experiments to test the proposed framework on two datasets: (1) MSCOCO [19], which contains 82,081 training images and 40,137 validation images. (2) Flickr30k [33], which contains 31,783 images in total. We followed the split in [10], which has 1,000 images for validation, 1,000 for testing, and the rest for training. In both datasets, each image has at least 5 ground truth sentences. Note that our experiments involve comparison between human descriptions and model-generated ones. As we have no access to the ground-truth annotations of the testing images in MSCOCO, for this dataset, we use the training set for both training and validation, and the validation set for testing the performance.

Experimental settings To process the annotations in each dataset, we follow [10] to remove non-alphabet characters, convert all remaining characters to lower-case, and replace all the words that appeared less than 5 times with a special word UNK. As a result, we get a vocabulary of size 9,567 on MSCOCO, and a vocabulary of size 7,000 on Flickr30k. All sentences are truncated to contain at most 16 words during training. We respectively pretrain $G$ using standard MLE [29], for 20 epoches, and $E$ with supervised training based on Eq (8), for 5 epoches. Subsequently, $G$ and $E$ are jointly trained, where each iteration consists of one step of G-update followed by one step of E-update. We set the mini-batch size to 64, the learning rate to 0.0001, and $n = 16$ in Monte Carlo rollouts. When testing, we use beam search based on the expected rewards from E-GAN, instead of the log-likelihoods, which we found empirically leads to better results.

Models We compare three methods for sentence generation: (1) Human: a sentence randomly sampled from ground-truth annotations of each image is used as the output of this method. Other human-provided sentences will be used as the references for metric evaluation. This baseline is tested for the purpose of comparing human-provided and model-generated descriptions. (2) G-MLE: a generator trained based on MLE [29] is used to produce the descriptions. This baseline represents the state-of-the-art of mainstream methods. (3) G-GAN: the same generator trained by our framework proposed in this paper, which is based on the conditional GAN formulations.

For both G-MLE and G-GAN, VGG16 [26] is used as the image encoders. Activations at the fc7 layer, which are of dimension 4096, are used as the image features and fed to the description generators. Note that G-GAN also takes a random vector $z$ as input. Here, $z$ is a 1024-dimensional vector, whose entries are sampled from a standard normal distribution.

Evaluation metrics We consider multiple evaluation metrics, including six conventional metrics BLEU-3 and BLEU-4 [23], METEOR [14], ROUGE-L [18], CIDEr [28], SPICE [1], and two additional metrics relevant to our formulation: E-NGAN and E-GAN. Particularly, E-GAN refers to the evaluator trained using our framework, E-NGAN refers to the evaluator trained according to Eq (8) without updating the generator alternatively. In other words, it is trained to distinguish between human-generated sentences and those generated by an MLE-based model.

Table 1 lists the performances of different generators under these metrics. On both datasets, the sentences produced by G-MLE receive considerably higher scores than those provided by human, on nearly all conventional metrics. This is not surprising. As discussed earlier, such metrics primarily focus on n-gram matching w.r.t. the references, while ignoring other important properties, e.g. naturalness and diversity. These results also clearly suggest that these metrics may not be particularly suited when evaluating the overall quality of the generated sentences. On the contrary, E-GAN regards Human as the best generator, while E-NGAN regards G-GAN as the best one. These two metrics obviously take into account more than just n-gram matching.
<table>
<thead>
<tr>
<th>generator</th>
<th>BLEU-3</th>
<th>BLEU-4</th>
<th>METEOR</th>
<th>ROUGE_L</th>
<th>CIDEr</th>
<th>SPICE</th>
<th>E-NGAN</th>
<th>E-GAN</th>
</tr>
</thead>
<tbody>
<tr>
<td>human</td>
<td>0.290</td>
<td>0.192</td>
<td>0.240</td>
<td>0.465</td>
<td>0.849</td>
<td>0.211</td>
<td>0.527</td>
<td>0.626</td>
</tr>
<tr>
<td>G-MLE</td>
<td>0.393</td>
<td>0.299</td>
<td>0.248</td>
<td>0.527</td>
<td>1.020</td>
<td>0.199</td>
<td>0.464</td>
<td>0.427</td>
</tr>
<tr>
<td>G-GAN</td>
<td>0.305</td>
<td>0.207</td>
<td>0.224</td>
<td>0.475</td>
<td>0.795</td>
<td>0.182</td>
<td>0.528</td>
<td>0.602</td>
</tr>
<tr>
<td>human</td>
<td>0.269</td>
<td>0.185</td>
<td>0.194</td>
<td>0.423</td>
<td>0.627</td>
<td>0.159</td>
<td>0.482</td>
<td>0.464</td>
</tr>
<tr>
<td>G-MLE</td>
<td>0.372</td>
<td>0.305</td>
<td>0.215</td>
<td>0.479</td>
<td>0.767</td>
<td>0.168</td>
<td>0.465</td>
<td>0.439</td>
</tr>
<tr>
<td>G-GAN</td>
<td>0.153</td>
<td>0.088</td>
<td>0.132</td>
<td>0.330</td>
<td>0.202</td>
<td>0.087</td>
<td>0.582</td>
<td>0.456</td>
</tr>
</tbody>
</table>

Table 1: This table lists the performances of different generators on MSCOCO and Flickr30k. On BLEU-{3,4}, METEOR, ROUGE_L, CIDEr, and SPICE, G-MLE is shown to be the best among all generators, surpassing human by a significant margin. While E-NGAN regard G-GAN as the best generator, E-GAN regard human as the best one.

**User study & qualitative comparison**

To fairly evaluate the quality of the generated sentences as well as how consistent the metrics are with human’s perspective, we conducted a user study. Specifically, we invited 30 human evaluators to compare the outputs of different generators. Each time, a human evaluator would be presented an image with two sentences from different methods and asked to choose the better one. Totally, we collected about 3,000 responses.

The comparative results are shown in Figure 5: From human’s views, G-GAN is better than G-MLE in 61% of all cases. In the comparison between human and models, G-MLE only won in 9% of the cases, while G-GAN won in over 24%. These results clearly suggest that the sentences produced by G-GAN are of considerably higher quality, i.e., being more natural and semantically relevant. The examples in Figure 7 also confirm this assessment. Particularly, we can see when G-MLE is presented with similar images, it tends to generate descriptions that are almost the same. On the contrary, G-GAN describes them with more distinctive and diverse ones. We also varied z to study the capability of G-GAN in giving diverse descriptions while maintaining the semantical relatedness. The qualitative results are listed in Figure 6.

For the evaluation metrics, the assessments provided by E-GAN are the most consistent with human’s evaluation, where the Kendall’s rank correlation coefficient between E-GAN and HE is 0.14, while that for CIDEr and SPICE are -0.30 and -0.25. Also note that E-GAN yields a larger numerical gap between scores of human and those of other generators as compared to E-NGAN, which suggests that adversarial training can improve the discriminative power of the evaluator.

**Evaluation by retrieval**

To compare the semantic relevance, we conducted an experiment using generated descriptions for retrieval. Specifically, we randomly select 5,000 images from the MSCOCO validation set; and for each image, we use the generated description as a query, ranking all 5,000 images according to the similarities between the images and the descriptions, computed by E-GAN, as well as the log-likelihoods. Finally, we compute the recall of the original image that appeared in the top-k ranks. The results for k = 1, 3, 5, 10 are listed in Table 2, where G-GAN is shown to provide more discriminative descriptions, outperforming G-MLE by a large margin across all cases.

**Failure Analysis**

We analyzed failure cases and found that a major kind of errors is the inclusion of incorrect details. e.g. colors (red/yellow hat), and counts (three/four people). A possible cause is that there are only a few samples for each particular detail, and they are not enough to make the generator capture these details reliably. Also, the focus on diversity and overall quality may also encourage the generator to include more details, with the risk of some details being incorrect.
Paragraph Generation We also tested our framework on paragraph generation (See Sec 3.4). We use the dataset provided by [11], which contains 14,575 training images, 2,487 validation images, and 2,489 testing images. Example results are shown in Figure 8. Again, we found that G-GAN can produce diverse and more natural descriptions as compared to G-MLE, which tends to follow similar patterns across sentences.

5. Conclusion

This paper presented an alternative approach to generating image descriptions. Compared to existing methods, which are mostly focused on the match of detailed wording, our approach, instead, aims to improve the overall quality, which involves semantic relevance, naturalness, and diversity. Some of these properties are often overlooked in previous efforts. We proposed a formulation based on conditional GAN that jointly trains a generator G and an evaluator E, and applied Policy Gradient and early feedbacks to tackle the technical challenges in end-to-end training. On both MSCOCO and Flickr30k, the proposed method produced descriptions that are more natural, diverse, and semantically relevant as compared to a state-of-the-art MLE-based model. This is clearly demonstrated in our user studies, qualitative examples, and retrieval applications. Our framework also provides an evaluator that is more consistent with human’s evaluation.

Acknowledgment This work is partially supported by the Big Data Collaboration Research grant from Sense-Time Group (CUHK Agreement No.TS1610626), the General Research Fund (GRF) of Hong Kong (No.14236516) and the Early Career Scheme (ECS) of Hong Kong (No.24204215).
References


[30] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In ICML, volume 14, pages 77–81, 2015. 1, 2, 3


