Speaking the Same Language: Matching Machine to Human Captions by Adversarial Training

Rakshith Shetty[1] Marcus Rohrbach[2,3] Lisa Anne Hendricks[2]
Mario Fritz[1] Bernt Schiele[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Saarbrücken, Germany
[2]UC Berkeley EECS, CA, United States [3]Facebook AI Research

Abstract

While strong progress has been made in image captioning recently, machine and human captions are still quite distinct. This is primarily due to the deficiencies in the generated word distribution, vocabulary size, and strong bias in the generators towards frequent captions. Furthermore, humans – rightfully so – generate multiple, diverse captions, due to the inherent ambiguity in the captioning task which is not explicitly considered in today's systems.

To address these challenges, we change the training objective of the caption generator from reproducing ground-truth captions to generating a set of captions that is indistinguishable from human written captions. Instead of handcrafting such a learning target, we employ adversarial training in combination with an approximate Gumbel sampler to implicitly match the generated distribution to the human one. While our method achieves comparable performance to the state-of-the-art in terms of the correctness of the captions, we generate a set of diverse captions that are significantly less biased and better match the global uni-, bi- and tri-gram distributions of the human captions.

1. Introduction

Image captioning systems have a variety of applications ranging from media retrieval and tagging to assistance for the visually impaired. In particular, models which combine state-of-the-art image representations based on deep convolutional networks and deep recurrent language models have led to ever increasing performance on evaluation metrics such as CIDEr [39] and METEOR [8] as can be seen e.g. on the COCO image Caption challenge leaderboard [6].

Despite these advances, it is often easy for humans to differentiate between machine and human captions – particularly when observing multiple captions for a single image.

Ours: a person on skis jumping over a ramp

Ours: a skier is making a turn on a course

Ours: a cross country skier makes his way through the snow

Ours: a skier is headed down a steep slope

Baseline: a man riding skis down a snow covered slope

Figure 1: Four images from the test set related to skiing, with captions from our model and a baseline. Baseline describes all four images with a generic caption, whereas our model produces diverse and more image specific captions.

As we analyze in this paper, this is likely due to artifacts and deficiencies in the statistics of the generated captions, which is more apparent when observing multiple samples. Specifically, we observe that state-of-the-art systems frequently "reveal themselves" by generating a different word distribution and using smaller vocabulary. Further scrutiny reveals that generalization from the training set is still challenging and generation is biased to frequent fragments and captions.

Also, today's systems are evaluated to produce a single caption. Yet, multiple potentially distinct captions are typically correct for a single image – a property that is reflected in human ground-truth. This diversity is not equally reproduced by state-of-the-art caption generators [40, 23].

Therefore, our goal is to make image captions less distinguishable from human ones – similar in the spirit to a Turing

Test. We also embrace the ambiguity of the task and extend our investigation to predicting sets of captions for a single image and evaluating their quality, particularly in terms of the diversity in the generated set. In contrast, popular approaches to image captioning are trained with an objective to reproduce the captions as provided by the ground-truth.

Instead of relying on handcrafting loss-functions to achieve our goal, we propose an adversarial training mechanism for image captioning. For this we build on Generative Adversarial Networks (GANs) [14], which have been successfully used to generate mainly continuous data distributions such as images [9, 30], although exceptions exist [27]. In contrast to images, captions are discrete, which poses a challenge when trying to backpropagate through the generation step. To overcome this obstacle, we use a Gumbel sampler [20, 28] that allows for end-to-end training.

We address the problem of caption set generation for images and discuss metrics to measure the caption diversity and compare it to human ground-truth. We contribute a novel solution to this problem using an adversarial formulation. The evaluation of our model shows that accuracy of generated captions is on par to the state-of-the-art, but we greatly increase the diversity of the caption sets and better match the ground-truth statistics in several measures. Qualitatively, our model produces more diverse captions across images containing similar content (Figure 1) and when sampling multiple captions for an image (see supplementary)[1].

2. Related Work

Image Description. Early captioning models rely on first recognizing visual elements, such as objects, attributes, and activities, and then generating a sentence using language models such as a template model [13], n-gram model [22], or statistical machine translation [34]. Advances in deep learning have led to end-to-end trainable models that combine deep convolutional networks to extract visual features and recurrent networks to generate sentences [11, 41, 21].

Though modern description models are capable of producing coherent sentences which accurately describe an image, they tend to produce generic sentences which are replicated from the train set [10]. Furthermore, an image can correspond to many valid descriptions. However, at test time, sentences generated with methods such as beam search are generally very similar. [40, 23] focus on increasing sentence diversity by integrating a diversity promoting heuristic into beam search. [42] attempts to increase the diversity in caption generation by training an ensemble of caption generators each specializing in different portions of the training set. In contrast, we focus on improving diversity of generated captions using a single model. Our method achieves this by learning a corresponding model using a different training loss as opposed to after training has completed. We note that generating diverse sentences is also a challenge in visual question generation, see concurrent work [19], and in language-only dialogue generation studied in the linguistic community, see e.g. [23, 24].

When training recurrent description models, the most common method is to predict a word w_t conditioned on an image and all previous *ground truth* words. At test time, each word is predicted conditioned on an image and previously *predicted* words. Consequently, at test time predicted words may be conditioned on words that were incorrectly predicted by the model. By only training on ground truth words, the model suffers from *exposure bias* [31] and cannot effectively learn to recover when it predicts an incorrect word during training. To avoid this, [4] proposes a scheduled sampling training scheme which begins by training with ground truth words, but then slowly conditions generated words on words previously produced by the model. However, [17] shows that the scheduled sampling algorithm is inconsistent and the optimal solution under this objective does not converge to the true data distribution. Taking a different direction, [31] proposes to address the exposure bias by gradually mixing a sequence level loss (BLEU score) using REINFORCE rule with the standard maximum likelihood training. Several other works have followed this up with using reinforcement learning based approaches to directly optimize the evaluation metrics like BLEU, METEOR and CIDER [33, 25]. However, optimizing the evaluation metrics does not directly address the diversity of the generated captions. Since all current evaluation metrics use n-gram matching to score the captions, captions using more frequent n-grams are likely to achieve better scores than ones using rarer and more diverse n-grams.

In this work, we formulate our caption generator as a generative adversarial network. We design a discriminator that explicitly encourages generated captions to be diverse and indistinguishable from human captions. The generator is trained with an adversarial loss with this discriminator. Consequently, our model generates captions that better reflect the way humans describe images while maintaining similar correctness as determined by a human evaluation.

Generative Adversarial Networks. The Generative Adversarial Networks (GANs) [14] framework learns generative models without explicitly defining a loss from a target distribution. Instead, GANs learn a generator using a loss from a discriminator which tries to differentiate real and generated samples, where the generated samples come from the generator. When training to generate real images, GANs have shown encouraging results [9, 30]. In all these works the target distribution is continuous. In contrast our target, a sequence of words, is discrete. Applying GANs to discrete sequences is challenging as it is unclear how to best back-propagate the loss through the sampling mechanism.

[1] https://goo.gl/3yRVnq

A few works have looked at generating discrete distributions using GANs. [27] aim to generate a semantic image segmentation with discrete semantic labels at each pixel. [46] uses REINFORCE trick to train an unconditional text generator using the GAN framework but diversity of the generated text is not considered.

Most similar to our work are concurrent works which use GANs for dialogue generation [24] and image caption generation [7]. While [24, 46, 7] rely on the reinforcement rule [43] to handle backpropagation through the discrete samples, we use the Gumbel Softmax [20]. See Section 3.1 for further discussion. [24] aims to generate a diverse dialogue of multiple sentences while we aim to produce diverse sentences for a single image. Additionally, [24] uses both the adversarial and the maximum likelihood loss in each step of generator training. We however train the generator with only adversarial loss after pre-training. Concurrent work [7] also applies GANs to diversify generated image captions. Apart from using the gumbel softmax as discussed above, our work differs from [7] in the discriminator design and quantitative evaluation of the generator diversity.

3. Adversarial Caption Generator

The image captioning task can be formulated as follows: given an input image x the generator G produces a caption, $G(x) = [w_0, \ldots, w_{n-1}]$, describing the contents of the image. There is an inherent ambiguity in the task, with multiple possible correct captions for an image, which is also reflected in diverse captions written by human annotators (we quantify this in Table 4). However, most image captioning architectures ignore this diversity during training. The standard approach to model $G(x)$ is to use a recurrent language model conditioned on the input image x [11, 41], and train it using a maximum likelihood (ML) loss considering every image–caption pair as an independent sample. This ignores the diversity in the human captions and results in models that tend to produce generic and commonly occurring captions from the training set, as we will show in Section 5.3.

We propose to address this by explicitly training the generator G to produce multiple diverse captions for an input image using the adversarial framework [14]. In adversarial frameworks, a generative model is trained by pairing it with adversarial discriminator which tries to distinguish the generated samples from true data samples. The generator is trained with the objective to fool the discriminator, which is optimal when G exactly matches the data distribution. This is well-suited for our goal because, with an appropriate discriminator network we could coax the generator to capture the diversity in the human written captions, without having to explicitly design a loss function for it.

To enable adversarial training, we introduce a second network, $D(x, s)$, which takes as input an image x and a caption set $S_p = \{s_1, \ldots, s_p\}$ and classifies it as either real or fake. Providing a set of captions per image as input to the discriminator allows it to factor in the diversity in the caption set during the classification. The discriminator can penalize the generator for producing very similar or repeated captions and thus encourage the diversity in the generator.

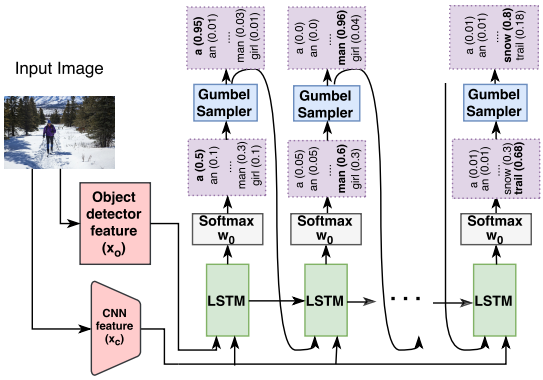

Figure 2: Caption generator model. Deep visual features are input to an LSTM to generate a sentence. A Gumbel sampler is used to obtain *soft* samples from the softmax distribution, allowing for backpropagation through the samples.

Specifically, the discriminator is trained to classify the captions drawn from the reference captions set, $R(x) = \{r_0, \cdots, r_{k-1}\}$, as real while classifying the captions produced by the generator, $G(x)$, as fake. The generator G can now be trained using an adversarial objective, i.e. G is trained to fool the discriminator to classify $G(x)$ as real.

3.1. Caption generator

We use a near state-of-the art caption generator model based on [36]. It uses the standard encoder-decoder framework with two stages: the encoder model which extracts feature vectors from the input image and the decoder which translates these features into a word sequence.

Image features. Images are encoded as activations from a pre-trained convolutional neural network (CNN). Captioning models also benefit from augmenting the CNN features with explicit object detection features [36]. Accordingly, we extract a feature vector containing the probability of occurrence of an object and provide it as input to the generator.

Language Model. Our decoder shown in Figure 2, is adopted from a Long-Short Term Memory (LSTM) based language model architecture presented in [36] for image captioning. It consists of a three-layered LSTM network with residual connections between the layers. The LSTM network takes two features as input. First is the object detection feature, x_o, which is input to the LSTM at only $0th$ time step and shares the input matrix with the word vectors. Second is the global image CNN feature, x_c, and is input to the LSTM at all time-steps through its own input matrix.

The softmax layer at the output of the generator produces

a probability distribution over the vocabulary at each step.

$$y_t = \text{LSTM}(w_{t-1}, x_c, y_{t-1}, c_{t-1}) \quad (1)$$

$$p(w_t|w_{t-1}, x) = \text{softmax}\left[\beta W_d * y_t\right], \quad (2)$$

where c_t is the LSTM cell state at time t and β is a scalar parameter which controls the peakyness of the distribution. Parameter β allows us to control how large a hypothesis space the generator explores during adversarial training. An additional uniform random noise vector z, is input to the LSTM in adversarial training to allow the generator to use the noise to produce diversity.

Discreteness Problem. To produce captions from the generator we could simply sample from this distribution $p(w_t|w_{t-1}, x)$, recursively feeding back the previously sampled word at each step, until we sample the END token. One can generate multiple sentences by sampling and pick the sentence with the highest probability as done in [12]. Alternatively we could also use greedy search approaches like beam-search. However, directly providing these discrete samples as input to the discriminator does not allow for backpropagation through them as they are discontinuous. Alternatives to overcome this are the reinforce rule/trick [43], using the softmax distribution, or using the Gumbel-Softmax approximation [20, 28].

Using policy gradient algorithms with the reinforce rule/trick [43] allows estimation of gradients through discrete samples [16, 2, 46, 24]. However, learning using reinforce trick can be unstable due to high variance [38] and some mechanisms to make learning more stable, like estimating the action-value for intermediate states by generating multiple possible sentence completions (e.g used in [46, 7]), can be computationally intensive.

Another option is to input the softmax distribution to the discriminator instead of samples. We experimented with this, but found that the discriminator easily distinguishes between the softmax distribution produced by the generator and the sharp reference samples, and the GAN training fails.

The last option, which we rely on in this work, it to use a continuous relaxation of the samples encoded as one-hot vectors using the Gumbel-Softmax approximation proposed in [20] and [28]. This continuous relaxation combined with the re-parametrization of the sampling process allows backpropagation through samples from a categorical distribution. The main benefit of this approach is that it plugs into the model as a differentiable node and does not need any additional steps to estimate the gradients. Whereas most previous methods to applying GAN to discrete output generators use policy gradient algorithms, we show that Gumbel-Softmax approximation can also be used successfully in this setting. An empirical comparison between the two approaches can be found in [20].

We use straight-through variation of the Gumbel-Softmax approximation [20] at the output of our generator to sample words during the adversarial training.

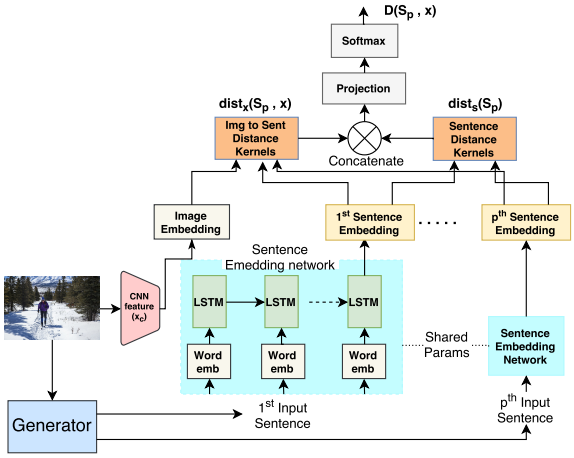

Figure 3: Discriminator Network. Caption set sampled from the generator is used to compute image to sentence $(dist_x(S_p, x))$ and sentence-to-sentence $(dist_s(S_p))$ distances. They are used to score the set as real/fake.

3.2. Discriminator model

The discriminator network, D takes an image x, represented using CNN feature x_c, and a set of captions $S_p = \{s_1, \ldots, s_p\}$ as input and classifies S_p as either real or fake. Ideally, we want D to base this decision on two criteria: a) do $s_i \in S_p$ describe the image correctly ? b) is the set S_p is diverse enough to match the diversity in human captions ?

To enable this, we use two separate distance measuring kernels in our discriminator network as shown in Figure 3. The first kernel computes the distances between the image x and each sentence in S_p. The second kernel computes the distances between the sentences in S_p. The architecture of these distance measuring kernels is based on the minibatch discriminator presented in [35]. However, unlike [35], we only compute distances between captions corresponding to the same image and not over the entire minibatch.

Input captions are encoded into a fixed size sentence embedding vector using an LSTM encoder to obtain vectors $f(s_i) \in \mathbb{R}^M$. The image feature, x_c, is also embedded into a smaller image embedding vector $f(x_c) \in \mathbb{R}^M$. The distances between $f(s_i), i \in \{1, \ldots, p\}$ are computed as

$$K_i = T_s \cdot f(s_i) \quad (3)$$

$$c_l(s_i, s_j) = exp\left(-\|K_{i,l} - K_{j,l}\|_{L_1}\right) \quad (4)$$

$$d_l(s_i) = \sum_{j=1}^{p} c_l(s_i, s_j) \quad (5)$$

$$dist_s(S_p) = [d_1(s_1), ..., d_O(s_1), ..., d_O(s_p)] \in \mathbb{R}^{p \times O} \quad (6)$$

where T_s is a $M \times N \times O$ dimensional tensor and O is the number of different $M \times N$ distance kernels to use.

Distances between $f(s_i), i \in 1, \ldots, p$ and $f(x_c)$ are obtained with similar procedure as above, but using a different tensor T_x of dimensions $M \times N \times O$ to yield $dist_x(S_p, x) \in \mathbb{R}^{p \times O}$. These two distance vectors capture the two aspects we want our discriminator to focus on. $dist_x(S_p, x)$ captures how well S_p matches the image x and $dist_s(S_p)$ captures the diversity in S_p. The two distance vectors are concatenated and multiplied with a output matrix followed by softmax to yield the discriminator output probability, $D(S_p, x)$, for S_p to be drawn from reference captions.

3.3. Adversarial Training

In adversarial training both the generator and the discriminator are trained alternatively for n_g and n_d steps respectively. The discriminator tries to classify $S_p^r \in R(x)$ as real and $S_p^g \in G(x)$ as fake. In addition to this, we found it important to also train the discriminator to classify few reference captions drawn from a random image as fake, i.e. $S_p^f \in R(y), y \neq x$. This forces the discriminator to learn to match images and captions, and not just rely on diversity statistics of the caption set. The complete loss function of the discriminator is defined by

$$L(D) = -\log\left(D(S_p^r, x)\right) - \log\left(1 - D(S_p^g, x)\right) - \log\left(1 - D(S_p^f, x)\right) \quad (7)$$

The training objective of the generator is to fool the discriminator into classifying $S_p^g \in G(x)$ as real. We found helpful to additionally use the feature matching loss [35]. This loss trains the generator to match activations induced by the generated and true data at some intermediate layer of the discriminator. In our case we use an l_2 loss to match the expected value of distance vectors $dist_s(S_p)$ and $dist_x(S_p, x)$ between real and generated data. The generator loss function is given by

$$\begin{aligned} L(G) = & -\log\left(D(S_p^g, x)\right) \\ & + \|\mathbb{E}\left[dist_s(S_p^g)\right] - \mathbb{E}\left[dist_s(S_p^r)\right]\|_2 \\ & + \|\mathbb{E}\left[dist_x(S_p^g, x)\right] - \mathbb{E}\left[dist_x(S_p^r, x)\right]\|_2, \end{aligned} \quad (8)$$

where the expectation is over a training mini-batch.

4. Experimental Setup

We conduct all our experiments on the MS-COCO dataset [5]. The training set consists of 83k images with five human captions each. We use the publicly available test split of 5000 images [21] for all our experiments. Section 5.4 uses a validation split of 5000 images.

For image feature extraction, we use activations from *res5c* layer of the 152-layered *ResNet* [15] convolutional neural network (CNN) pre-trained on ImageNet. The input images are scaled to 448×448 dimensions for *ResNet* feature extraction. Additionally we use features from the VGG network [37] in our ablation study in Section 5.4. Following [36], we additionally extract 80-dimensional object detection features using a Faster Region-Based Convolutional Neural Network (RCNN) [32] trained on the 80 object categories in the COCO dataset. The CNN features are input to both the generator (at x_p) and the discriminator. Object detection features are input only to the generator at the x_i input and is used in all the generator models reported here.

4.1. Insights in Training the GAN

As is well known [3], we found GAN training to be sensitive to hyper-parameters. Here we discuss some settings which helped stabilize the training of our models.

We found it necessary to pre-train the generator using standard maximum likelihood training. Without pre-training, the generator gets stuck producing incoherent sentences made of random word sequences. We also found pre-training the discriminator on classifying correct image-caption pairs against random image-caption pairs helpful to achieve stable GAN training. We train the discriminator for 5 iterations for every generator update. We also periodically monitor the classification accuracy of the discriminator and train it further if it drops below 75%. This prevents the generator from updating using a bad discriminator.

Without the feature matching term in the generator loss, the GAN training was found to be unstable and needed additional maximum likelihood update to stabilize it. This was also reported in [24]. However with the feature matching loss, training is stable and the ML update is not needed.

A good range of values for the Gumbel temperature was found to be $(0.1, 0.8)$. Beyond this range training was unstable, but within this range the results were not sensitive to it. We use a fixed temperature setting of 0.5 in the experiments reported here. The softmax scaling factor, β in (2), is set to value 3.0 for training of all the adversarial models reported here. The sampling results are also with $\beta = 3.0$.

5. Results

We conduct experiments to evaluate our adversarial caption generator w.r.t. two aspects: how human-like the generated captions are and how accurately they describe the contents of the image. Using diversity statistics and word usage statistics as a proxy for measuring how closely the generated captions mirror the distribution of the human reference captions, we show that the adversarial model is more human-like than the baseline. Using human evaluation and automatic metrics we also show that the captions generated by the adversarial model performs similar to the baseline model in terms of correctness of the caption.

Henceforth, *Base* and *Adv* refer to the baseline and adversarial models, respectively. Suffixes *bs* and *samp* indicate decoding using beamsearch and sampling respectively.

5.1. Measuring if captions are human-like

Diversity. We analyze n-gram usage statistics, compare vocabulary sizes and other diversity metrics presented below to understand and measure the gaps between human written captions and the automatic methods and show that the adversarial training helps bridge some of these gaps.

To measure the corpus level diversity of the generated captions we use:

- *Vocabulary Size* - number of unique words used in all generated captions
- *% Novel Sentences* - percentage of generated captions not seen in the training set.

To measure diversity in a set of captions, S_p, corresponding to a single image we use:

- *Div-1* - ratio of number of unique unigrams in S_p to number of words in S_p. Higher is more diverse.
- *Div-2* - ratio of number of unique bigrams in S_p to number of words in S_p. Higher is more diverse.
- *mBleu* - Bleu score is computed between each caption in S_p against the rest. Mean of these p Bleu scores is the mBleu score. Lower values indicate more diversity.

Correctness. Just generating diverse captions is not useful if they do not correctly describe the content of an image. To measure the correctness of the generated captions we use two automatic evaluation metrics Meteor [8] and SPICE [1]. However since it is known that the automatic metrics do not always correlate very well with human judgments of the correctness, we also report results from human evaluations comparing the baseline model to our adversarial model.

5.2. Comparing caption accuracy

Table 1 presents the comparison of our adversarial model to the baseline model. Both the baseline and the adversarial models use *ResNet* features. The beamsearch results are with beam size 5 and sampling results are with taking the best of 5 samples. Here the best caption is obtained by ranking the captions as per probability assigned by the model.

Table 1 also shows the metrics from some recent methods from the image captioning literature. The purpose of this comparison is to illustrate that we use a strong baseline and that our baseline model is competitive to recent published work, as seen from the Meteor and Spice metrics.

Comparing baseline and adversarial models in Table 1 the adversarial model does worse in-terms of Meteor scores and overall spice metrics. When we look at Spice scores on individual categories shown in Table 2 we see that adversarial models excel at counting relative to the baseline and describing the size of an object correctly.

However, it is well known that automatic metrics do not always correlate with human judgments on correctness of a caption. A primary reason the adversarial models do poorly on automatic metrics is that they produce significantly more unique sentences using a much larger vocabulary and rarer n-grams, as shown in Section 5.3. Thus, they are less likely to do well on metrics relying on n-gram matches.

To verify this claim, we conduct human evaluations comparing captions from the baseline and the adversarial model. Human evaluators from Amazon Mechanical Turk are shown an image and a caption each from the two models and are asked "Judge which of the two sentences is a better description of the image (w.r.t. correctness and relevance)!". The choices were either of the two sentences or to report that they are the same. Results from this evaluation are presented in Table 3. We can see that both adversarial and baseline models perform similarly, with adversarial models doing slightly better. This shows that despite the poor performance in automatic evaluation metrics, the adversarial models produce captions that are similar, or even slightly better, in accuracy to the baseline model.

Method	**Meteor**	**Spice**
ATT-FCN [45]	0.243	–
MSM [44]	0.251	–
KWL [26]	0.266	**0.194**
Ours Base-bs	**0.272**	0.187
Ours Base-samp	0.265	0.186
Ours Adv-bs	0.239	0.167
Ours Adv-samp	0.236	0.166

Table 1: Meteor and Spice metrics comparing performance of baseline and adversarial models.

Method	**Spice**					
	Color	Attribute	Object	Relation	Count	Size
Base-bs	**0.101**	**0.085**	0.345	0.049	0.025	0.034
Base-samp	0.059	0.069	**0.352**	**0.052**	0.032	0.033
Adv-bs	0.079	0.082	0.318	0.034	**0.080**	0.052
Adv-samp	0.078	0.082	0.316	0.033	0.076	**0.053**

Table 2: Comparing baseline and adversarial models in different categories of Spice metric.

Comparison	Adversarial - Better	Adversarial - Worse
Beamsearch	36.9	34.8
Sampling	35.7	33.2

Table 3: Human evaluation comparing adversarial model vs the baseline model on 482 random samples. Correctness of captions. With agreement of at least 3 out of 5 judges in %. Humans agreed in 89.2% and 86.7% of images in beamsearch and sampling cases respectively.

5.3. Comparing vocabulary statistics

To characterize how well the captions produced by the automatic methods match the statistics of the human written

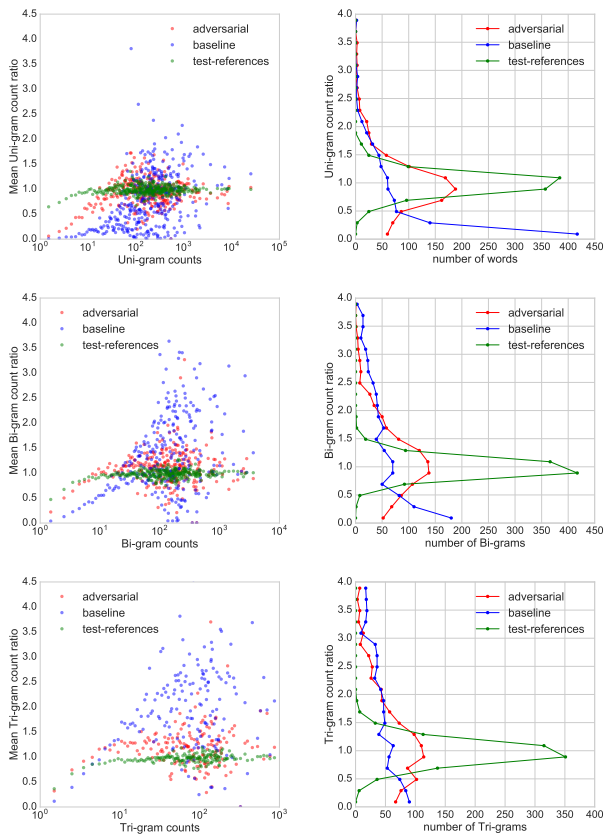

Figure 4: Comparison of n-gram count ratios in generated test-set captions by different models. Left side shows the mean n-gram count-ratios as a function of counts on training set. Right side shows the histogram of the count-ratios.

captions, we look at n-gram usage statistics in the generated captions. Specifically, we compute the ratio of the actual count of an n-gram in the caption set produced by a model to the expected n-gram count based on the training data.

Given that an n-gram occurred m times in the training set we can expect that it occurs $m * |\text{test-set}|/|\text{train-set}|$ times in the test set. However actual counts may vary depending on how different the test set is from the training set. We compute these ratios for reference captions in the test set to get an estimate of the expected variance of the count ratios. The left side of Figure 4 shows the mean count ratios for uni-, bi- and tri-grams in the captions generated on test-set plotted against occurrence counts in the training set. Histogram of these ratios are shown on the right side.

Count ratios for the reference captions from the test-set are shown in green. We see that the n-gram counts match well between the training and test set human captions and the count ratios are spread around 1.0 with a small variance.

The baseline model shows a clear bias towards more frequently occurring n-grams. It consistently overuses more

Method	n	Div-1	Div-2	mBleu-4	Vocab-ulary	% Novel Sentences
Base-bs	1 of 5	–	–	–	756	34.18
	5 of 5	0.28	0.38	0.78	1085	44.27
Base-samp	1 of 5	–	–	–	839	52.04
	5 of 5	0.31	0.44	0.68	1460	55.24
Adv-bs	1 of 5	–	–	–	1508	68.62
	5 of 5	0.34	0.44	0.70	2176	72.53
Adv-samp	1 of 5	–	–	–	1616	73.92
	5 of 5	**0.41**	**0.55**	**0.51**	**2671**	**79.84**
Human captions	1 of 5	–	–	–	3347	92.80
	5 of 5	0.53	0.74	0.20	7253	95.05

Table 4: Diversity Statistics described in Section 5.1. Higher values correspond to more diversity in all except mBleu-4, where lower is better.

Adv-bs	a group of friends enjoying a dinner at the restauarant	several cows in their pen at the farm	A dog is trying to get something out of the snow
Base-bs	a group of people sitting around a wooden table	a herd of cattle standing next to each other	a couple of dogs that are in the snow

Figure 5: Some qualitative examples comparing comparing captions generated by the our model to the baseline model.

frequent n-grams (ratio>1.0) from the training set and under-uses less frequent ones (ratio<1.0). This trend is seen in all the three plots, with more frequent tri-grams particularly prone to overuse. It can also be observed in the histogram plots of the count ratios, that the baseline model does a poor job of matching the statistics of the test set.

Our adversarial model does a much better job in matching these statistics. The histogram of the uni-gram count ratios are clearly closer to that of test reference captions. It does not seem to be significantly overusing the popular words, but there is still a trend of under utilizing some of the rarer words. It is however clearly better than the baseline model in this aspect. The improvement is less pronounced with the bi- and tri-grams, but still present.

Another clear benefit from using the adversarial training is observed in terms of diversity in the captions produced by the model. The diversity in terms of both global statistics and per image diversity statistics is much higher in captions produced by the adversarial models compared to the baseline models. This result is presented in Table 4. We can see that the vocabulary size approximately doubles from 1085 in the baseline model to 2176 in the adversarial model us-

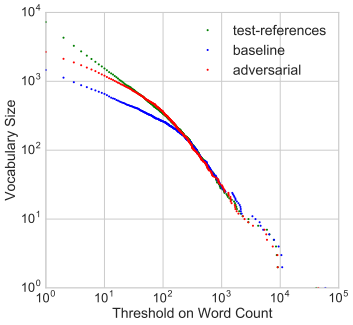

Figure 6: Vocabulary size as a function of word counts.

ing beamsearch. A similar trend is also seen comparing the sampling variants. As expected more diversity is achieved when sampling from the adversarial model instead of using beamsearch with vocabulary size increasing to 2671 in *Adv-samp*. The effect of this increased diversity can be in the qualitative examples shown in Figure 5. More qualitative samples are included in the supplementary material.

We can also see that the adversarial model learns to construct significantly more novel sentences compared to the baseline model with *Adv-bs* producing novel captions 72.53% of the time compared to just 44.27% by the *beam-bs*. All three per-image diversity statistics also improve in the adversarial models indicating that they can produce a more diverse set of captions for any input image.

Table 4 also shows the diversity statistics on the reference captions on the test set. This shows that although adversarial models do considerably better than the baseline, there is still a gap in diversity statistics when compared to the human written captions, especially in vocabulary size.

Finally, Figure 6 plots the vocabulary size as a function of word count threshold, k. We see that the curve for the adversarial model better matches the human written captions compared to the baseline for all values of k. This illustrates that the gains in vocabulary size in adversarial models does not arise from using words with specific frequency, but is instead distributed evenly across word frequencies.

5.4. Ablation Study

We conducted experiments to understand the importance of different components of our architecture. The results are presented in Table 5. The baseline model for this experiment uses VGG [37] features as x_p input and is trained using maximum likelihood loss and is shown in the first row of Table 5. The other four models use adversarial training.

Comparing rows 1 and 2 of Table 5, we see that adversarial training with a discriminator evaluating a single caption does badly. Both the diversity and Meteor score drop compared to the baseline. In this setting the generator can get away with producing one good caption (mode collapse) for an image as the discriminator is unable to penalize the lack of diversity in the generator.

Image Feature	Evalset size (p)	Feature Matching	Meteor	Div-2	Vocab. Size
VGG baseline			0.247	0.44	1367
VGG	1	No	0.179	0.40	812
VGG	5	No	0.197	0.52	1810
VGG	5	yes	0.207	**0.59**	2547
ResNet	5	yes	**0.236**	0.55	**2671**

Table 5: Performance comparison of various configurations of the adversarial caption generator on the validation set.

However, comparing rows 1 and 3, we see that adversarial training using a discriminator evaluating 5 captions simultaneously does much better in terms of Div-2 and vocabulary size. Adding feature matching loss further improves the diversity and also slightly improves accuracy in terms of Meteor score. Thus simultaneously evaluating multiple captions and using feature matching loss allows us to alleviate mode collapse generally observed in GANs.

Upgrading to the *ResNet*[15] increases the Meteor score greatly and slightly increases the vocabulary size. *ResNet* features provide richer visual information which is used by the generator to produce diverse but still correct captions.

We also notice that the generator learns to ignore the input noise. This is because there is sufficient stochasticity in the generation process due to sequential sampling of words and thus the generator doesn't need the additional noise input to increase output diversity. Similar observation was reported in other conditional GAN architectures [18, 29]

6. Conclusions

We have presented an adversarial caption generator model which is explicitly trained to generate diverse captions for images. We achieve this by utilizing a discriminator network designed to promote diversity and use the adversarial learning framework to train our generator. Results show that our adversarial model produces captions which are diverse and match the statistics of human generated captions significantly better than the baseline model. The adversarial model also uses larger vocabulary and is able to produce significantly more novel captions. The increased diversity is achieved while preserving accuracy of the generated captions, as shown through a human evaluation.

Acknowledgements

This research was supported by the German Research Foundation (DFG CRC 1223) and by the Berkeley Artificial Intelligence Research (BAIR) Lab.

References

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Cmoputer Vision (ECCV)*, 2016. 6

[2] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016. 4

[3] M. Arjovsky and L. Bottou. Towards principled methods for training generative adversarial networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017. 5

[4] S. Bengio, O. Vinyals, N. Jaitly, and N. Shazeer. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[5] X. Chen, T.-Y. L. Hao Fang, R. Vedantam, S. Gupta, P. Dollr, and C. L. Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arxiv:1504.00325*, 2015. 5

[6] COCO. Microsoft COCO Image Captioning Challenge. https://competitions.codalab.org/competitions/3221#results, 2017. 1

[7] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional gan. 2017. 3, 4

[8] M. Denkowski and A. Lavie. Meteor universal: Language specific translation evaluation for any target language. *ACL 2014*, 2014. 1, 6

[9] E. L. Denton, S. Chintala, R. Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 2

[10] J. Devlin, H. Cheng, H. Fang, S. Gupta, L. Deng, X. He, G. Zweig, and M. Mitchell. Language models for image captioning: The quirks and what works. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015. 2

[11] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 3

[12] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell. Long-term recurrent convolutional networks for visual recognition and description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016. 4

[13] A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth. Every picture tells a story: Generating sentences from images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010. 2

[14] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NIPS)*, 2014. 2, 3

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 5, 8

[16] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 4

[17] F. Huszar. How (not) to train your generative model: Scheduled sampling, likelihood, adversary? *arXiv preprint arXiv:1511.05101*, 2015. 2

[18] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 8

[19] U. Jain, Z. Zhang, and A. Schwing. Creativity: Generating diverse questions using variational autoencoders. 2017. 2

[20] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 2, 3, 4

[21] A. Karpathy and L. Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 5

[22] G. Kulkarni, V. Premraj, V. Ordonez, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg. Babytalk: Understanding and generating simple image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(12), 2013. 2

[23] J. Li, M. Galley, C. Brockett, J. Gao, and B. Dolan. A diversity-promoting objective function for neural conversation models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2016. 1, 2

[24] J. Li, W. Monroe, T. Shi, A. Ritter, and D. Jurafsky. Adversarial learning for neural dialogue generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017. 2, 3, 4, 5

[25] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Optimization of image description metrics using policy gradient methods. *arXiv preprint arXiv:1612.00370*, 2016. 2

[26] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6

[27] P. Luc, C. Couprie, S. Chintala, and J. Verbeek. Semantic segmentation using adversarial networks. In *Advances in Neural Information Processing Systems Workshops (NIPS Workshops)*, 2016. 2, 3

[28] C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 2, 4

[29] M. Mathieu, C. Couprie, and Y. LeCun. Deep multi-scale video prediction beyond mean square error. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 8

[30] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 2

[31] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2016. 2

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2015. 5

[33] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[34] M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele. Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2013. 2

[35] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in Neural Information Processing Systems (NIPS)*, 2016. 4, 5

[36] R. Shetty, H. R-Tavakoli, and J. Laaksonen. Exploiting scene context for image captioning. In *ACMMM Vision and Language Integration Meets Multimedia Fusion Workshop*, 2016. 3, 5

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. 5, 8

[38] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998. 4

[39] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1

[40] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016. 1, 2

[41] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 3

[42] Z. Wang, F. Wu, W. Lu, J. Xiao, X. Li, Z. Zhang, and Y. Zhuang. Diverse image captioning via grouptalk. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2016. 2

[43] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4), 1992. 3, 4

[44] T. Yao, Y. Pan, Y. Li, Z. Qiu, and T. Mei. Boosting image captioning with attributes. *arXiv preprint arXiv:1611.01646*, 2016. 6

[45] Q. You, H. Jin, Z. Wang, C. Fang, and J. Luo. Image captioning with semantic attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[46] L. Yu, W. Zhang, J. Wang, and Y. Yu. SeqGAN: sequence generative adversarial nets with policy gradient. *Proceedings of the Conference on Artificial Intelligence (AAAI)*, 2016. 3, 4