# Adversarial Semantic Alignment for Improved Image Captions

Pierre Dognin*, Igor Melnyk*, Youssef Mroueh*, Jerret Ross* & Tom Sercu*
IBM Research, Yorktown Heights, NY
{pdognin,mroueh,rossja}@us.ibm.com, {igor.melnyk,tom.sercu1}@ibm.com

## Abstract

*In this paper we study image captioning as a conditional GAN training, proposing both a context-aware LSTM captioner and co-attentive discriminator, which enforces semantic alignment between images and captions. We empirically focus on the viability of two training methods: Self-critical Sequence Training (SCST) and Gumbel Straight-Through (ST) and demonstrate that SCST shows more stable gradient behavior and improved results over Gumbel ST, even without accessing discriminator gradients directly. We also address the problem of automatic evaluation for captioning models and introduce a new semantic score, and show its correlation to human judgement. As an evaluation paradigm, we argue that an important criterion for a captioner is the ability to generalize to compositions of objects that do not usually co-occur together. To this end, we introduce a small captioned Out of Context (OOC) test set. The OOC set, combined with our semantic score, are the proposed new diagnosis tools for the captioning community. When evaluated on OOC and MS-COCO benchmarks, we show that SCST-based training has a strong performance in both semantic score and human evaluation, promising to be a valuable new approach for efficient discrete GAN training.*

## 1. Introduction

Significant progress has been made on the task of generating image descriptions using neural image captioning. Early systems were traditionally trained using cross-entropy (CE) loss minimization [27, 11, 28]. Later, reinforcement learning techniques [22, 23, 14] based on policy gradient methods, e.g., REINFORCE, were introduced to directly optimize the $n$-gram matching metrics such as CIDEr [26], BLEU4 [20] or SPICE [1]. Along a similar idea, [23] introduced Self-critical Sequence Training (SCST), a light-weight variant of REINFORCE, which produced state of the art image captioning results using CIDEr as an optimization metric. Although optimizing the above automatic metrics might be a promising direction to take, these metrics unfortunately miss an essential part of the semantic alignment between image and caption. They do not provide a way to promote naturalness of the language, e.g., as measured by a Turing test, so that the machine-generated text becomes indistinguishable from the text created by humans.

To address the problem of diversity and naturalness, image captioning has recently been explored in the framework of GANs [6]. The main idea is to train a discriminator to detect a signal on the misalignment between an image and a generated sentence, while the generator (captioner) can use this signal to improve its text generation mechanism to better align the caption with a given image. Due to the discrete nature of text generation, GAN training remains challenging and has been generally tackled with either reinforcement learning techniques [29, 3, 8, 21, 4] or by using the Gumbel softmax relaxation [9], for example, as in [25, 12].

Despite these impressive advances, image captioning is far from being a solved task. It still is a challenge to satisfactory bridge a semantic gap between image and caption, and to produce diverse, creative and human-like captions. The current captioning systems also suffer from a dataset bias: the models overfit to common objects co-occurring in common context, and they struggle to generalize to scenes where the same objects appear in unseen contexts. Although the recent advances of applying GANs for image captioning to promote human-like captions is a very promising direction, the discrete nature of the text generation process makes it challenging to train such systems. The results in [4, 25] are encouraging but the proposed solutions are still complex and computationally expensive. Moreover, the recent work of [2] showed that the task of text generation for the current discrete GAN models is still challenging, many times producing unsatisfactory results, and therefore requires new approaches and methods. Finally, evaluation of image captioning using automated metrics such as CIDEr, BLEU4, etc. is still unsatisfactory since simple $n$-gram matching, that does not reference the image, remains inadequate and sometimes misleading for scoring diverse and descriptive captions.

In this paper, we propose to address the above issues by

*Equal Contributions. Authors in alphabetical order.

accomplishing the following three main objectives: **1)** *Architectural and algorithmic improvements*: We propose a novel GAN-based framework for image captioning that enables better language composition and more accurate compositional alignment of image and text (Section 2.1), as well as a light-weight and efficient approach for discrete GAN training based on SCST (Section 2.2). **2)** *Automated scoring metric*: We propose the *semantic score*, which enables quantitative automatic evaluation of caption quality and its alignment to the image across multiple models (Section 3). **3)** *Diagnostic dataset*: Finally, we introduce the Out of Context (OOC) test set which is a quick and useful diagnostic tool for checking a model's generalization to out of context scenes (Section 3).

## 2. Adversarial Caption Generation

In this Section we present our novel captioner and the discriminator models. We employ SCST for discrete GAN optimization and compare it with the approach based on the Gumbel trick. Our experiments (Section 4) show that SCST obtains better results, even though it does not directly access the discriminator gradients.

### 2.1. Compositional Captioner and Discriminator

Here we introduce an image captioning model with attention that we call *context aware captioning* based on [15]. This allows the captioner to compose sentences based on fragments of observed visual scenes in the training. Furthermore, we introduce a discriminator that scores the alignment between images and captions based on a co-attention model [16]. This gives the generator a signal on the semantic alignment and the compositional nature of visual scenes and language. We show in Section 4 that we obtain better results across evaluation metrics when using this co-attentive discriminator.

**Context Aware Captioner $G_\theta$.** For caption generation, we use an LSTM with visual attention [28, 23] together with a visual sentinel [15] to give the LSTM a choice of attending to visual or textual cues. While [15] feeds only an average image feature to the LSTM at each step, we feed a mixture of image and visual sentinel features $\hat{c}_{t-1}$ from the previous step to make the LSTM aware of the last attention context, as seen in Figure 1. We call it *Context Aware* Attention. This simple modification gives significant gains, as the captioner is now aware of the contextual information used in the past. As reported in Table 1, a captioner with an adaptive visual sentinel [15] gives 99.7 CIDEr vs. 103.3 for our Context Aware Attention on COCO validation set.

**Co-attention Pooling Discriminator $D_\eta$.** The task of the discriminator is to score the similarity between an image and a caption. Previous works jointly embed the modalities at the similarity computation level, which we call *late joint embedding*, see Figure 2 (a). Instead, we propose to jointly embed image and caption in earlier stages using a co-attention model [16, 5] and compute similarity on the attentive pooled representation. We call this a *Co-attention* discriminator (Figure 2 (b)) and provide architectural details below.

Given a sentence $w$ composed of a sequence of words $(w_1, \dots w_T)$, the discriminator embeds each word using the LSTM (state dimension $m=512$) to get $H = [h_1, \dots h_T]^\top$ for $H \in \mathbb{R}^{T\times m}$, where $h_t, c_t = \text{LSTM}(h_{t-1}, c_{t-1}, w_t)$. For image $I$, we extract features $(I_1, \dots I_C)$, where $C = 14\times 14 = 196$ (number of crops) and also embed them as $I = [WI_1, \dots WI_C]^\top \in \mathbb{R}^{C\times m}$, where $W \in \mathbb{R}^{m\times d_I}$, and $d_I = 2048$, our image feature size. Following [16], we then compute a correlation $Y$ between image and text using bilinear projection $Q \in \mathbb{R}^{m\times m}$, $Y = \tanh(IQH^\top) \in \mathbb{R}^{C\times T}$. Matrix $Y$ is used to compute co-attention weights of

| Attention Model | CE | RL |
|---|---|---|
| Att2All [23] | 98.5 | 115.7 |
| Sentinel [15] | 99.7 | |
| Context Aware (ours) | 103.3 | 118.6 |

Table 1: CIDEr performance of captioning systems given various attention mechanisms, Att2All [23], sentinel attention [4] and Context Aware attention on COCO validation set. Models are built using cross-entropy (CE) and SCST [23] (RL). Context aware attention brings large gains in CIDEr for both CE and RL trained models.

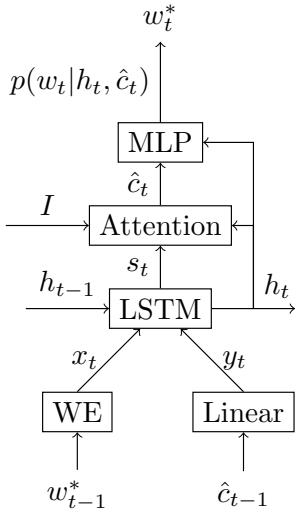
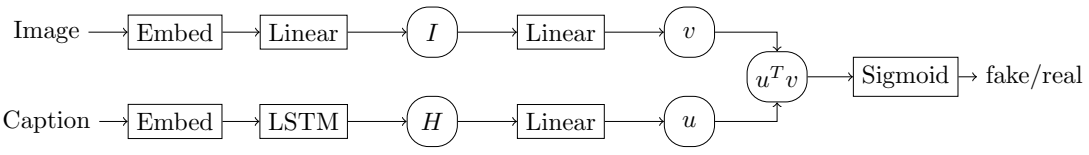
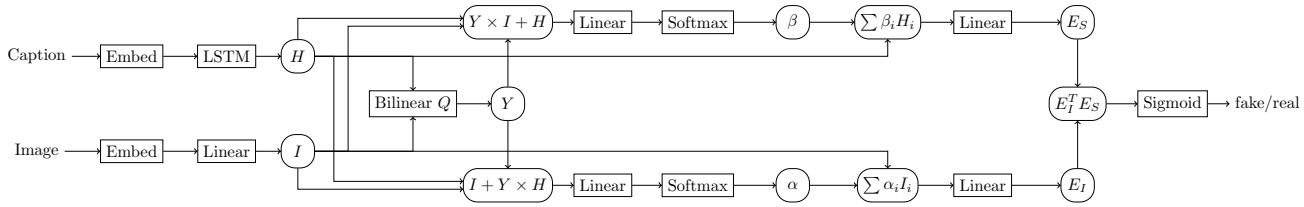



Figure 1: Context Aware Captioner. At each step $t$, the textual information $w_{t-1}^*$, and the mixture of image features and visual sentinel $\hat{c}_{t-1}$ from previous step $t-1$ are fed to the LSTM to make it aware of past attentional contexts.




(a) Joint-Embedding Discriminator (Joint-Emb) [4]




(b) Proposed Co-Attention Discriminator (Co-att)

Figure 2: Discriminator architectures. **(a)** Joint-Embedding Discriminator from [4]. **(b)** Our proposed $D_\eta$. By jointly embedding the image and caption with a co-attention model, we give the discriminator the ability to modulate the image features depending on the caption and vice versa.

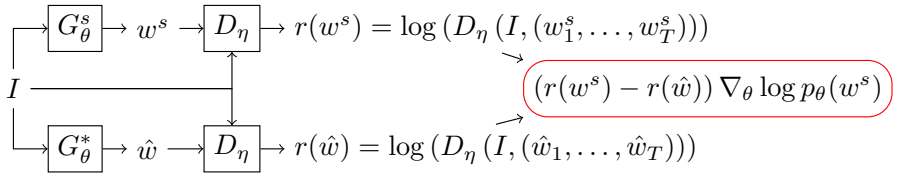



Figure 3: SCST Training of GAN-captioning.

one modality conditioned on another:

$$\alpha = \text{Softmax}(\text{Linear}(\tanh(IW_I + YHW_{Ih}))) \in \mathbb{R}^C,$$
$$\beta = \text{Softmax}(\text{Linear}(\tanh(HW_h + Y^\top IW_{hI}))) \in \mathbb{R}^T,$$

where all new matrices are in $\mathbb{R}^{m \times m}$. The above weights are used then to combine the word and image features: $E_I = U_I \left(\sum_{i=1}^{C} \alpha_i W I_i\right)$ and $E_S = V_S \left(\sum_{j=1}^{T} \beta_j h_j\right)$ for $U_I, V_S \in \mathbb{R}^{m \times m}$. Finally, the image-caption score is computed as $D_\eta(I, w) = \text{Sigmoid}(\text{E}_\text{I}^\top \text{E}_\text{S})$ ($\eta$, discriminator parameters). In Section 4 we compare $D_\eta$ with the late joint embedding approach of [4, 25], where $E_I$ is the average spatial pooling of CNN features and $E_S$ the last state of LSTM. We refer to this discriminator as *Joint-Emb* and to ours as *Co-att* (see Figure 2).

## 2.2. Adversarial Training

In this Section we describe the details of the adversarial training of the discriminator and the captioner.

**Training $D_\eta$.** Our discriminator $D_\eta$ is not only trained to distinguish real captions from fake (generated), but also to detect when images are coupled with random unrelated real sentences, thus forcing it to check not only the sentence composition but also the semantic relationship between image and caption. To accomplish this, we solve the following optimization problem: $\max_\eta \mathcal{L}_D(\eta)$, where the loss $\mathcal{L}_D(\eta)$

$$\mathbb{E}_{I,w \in S(I)} \log D_\eta(I, w) + \frac{1}{2}\mathbb{E}_{I,w^s \sim p_\theta(.|I)} \log\left(1 - D_\eta(I, w^s)\right) + \frac{1}{2}\mathbb{E}_{I,w' \notin S(I)} \log\left(1 - D_\eta(I, w')\right), \quad (1)$$

where $w$ is the real sentence, $w^s$ is sampled from generator $G_\theta$ (fake caption), and $w'$ is a real but randomly picked caption.

**Training $G_\theta$.** The generator is optimized to solve $\max_\theta \mathcal{L}_G(\theta)$, where $\mathcal{L}_G(\theta) = \mathbb{E}_I \log D_\eta(I, G_\theta(I))$. The main difficulty is the discrete, non-differentiables nature of the problem. We propose to solve this issue by adopting SCST [23], a light-weight variant of the policy gradient method, and compare it to the Gumbel relaxation approach of [9].

**Training $G_\theta$ with SCST.** SCST [23] is a REINFORCE variant that uses the reward under the decoding algorithm as baseline. In this work, the decoding algorithm is a "greedy max", selecting at each step the most probable word from $\arg\max p_\theta(.|h_t)$. For a given image, a single sample $w^s$ of the generator is used to estimate the full sequence reward, $\mathcal{L}_G^I(\theta) = \log(D(I, w^s))$ where $w^s \sim p_\theta(.|I)$. Using SCST,

the gradient is estimated as follows:

$$\nabla_\theta \mathcal{L}_G^I(\theta) \approx (\log D_\eta(I, w^s) - \underbrace{\log D_\eta(I, \hat{w})}_{\text{Baseline}}) \nabla_\theta \log p_\theta(w^s|I)$$
$$= \left(\log \frac{D_\eta(I, w^s)}{D_\eta(I, \hat{w})}\right) \nabla_\theta \log p_\theta(w^s|I),$$

where $\hat{w}$ is obtained using *greedy max* (see Figure 3). Note that the baseline does not change the expectation of the gradient but reduces the variance of the estimate.

Also, observe that the GAN training can be regularized with any NLP metric $r_{\text{NLP}}$ (such as CIDEr) to enforce closeness of the generated captions to the provided ground truth on the $n$-gram level; the gradient then becomes:

$$\left(\log \frac{D_\eta(I, w^s)}{D_\eta(I, \hat{w})} + \lambda \left(r_{\text{NLP}}(w^s) - r_{\text{NLP}}(\hat{w})\right)\right) \nabla_\theta \log p_\theta(w^s|I).$$

There are two main advantages of SCST over other policy gradient methods used in the sequential GAN context: **1)** The reward in SCST can be global at the sentence level and the training still succeeds. In other policy gradient methods, e.g., [4, 14], the reward needs to be defined at each word generation with the *full* sentence sampling, so that the discriminator needs to be evaluated $T$ times (sentence length). **2)** In [4, 14, 8], many Monte-Carlo rollouts are needed to reduce variance of gradients, requiring many forward-passes through the generator. In contrast, due to a strong baseline, only a single sample estimate is enough in SCST.

**Training $G_\theta$ : Gumbel Trick.** An alternative way to deal with the discreteness of the generator is by using Gumbel re-parameterization [9]. Define the soft samples $y_t^j$, for $t = 1, \dots T$ (sentence length) and $j = 1, \dots K$ (vocabulary size) such that: $y_t^j = \text{Softmax}\left(\frac{1}{\tau}(\text{logits}_\theta(j|h_t, I) + g_j)\right)$, where $g_j$ are samples from Gumbel distribution, $\tau$ is a temperature parameter. We experiment with the Gumbel Soft and Gumbel Straight-Through (Gumbel ST) approaches, recently used in [25, 12].

For *Gumbel Soft*, we use the soft samples $y_t$ as LSTM input $w_{t+1}^s$ at the next time step and in $D_\eta$:

$$\nabla_\theta \mathcal{L}_G^I(\theta) = \nabla_\theta \log(D_\eta(I, (y_1, \dots y_T))).$$

For *Gumbel ST*, we define one-hot encodings $\mathcal{O}_t = \text{OneHot}(\arg\max_j y_t^j)$ and approximate the gradients $\partial \mathcal{O}_t^j / \partial y_t^{j'} = \delta_{jj'}$. To sample from $G_\theta$ we use the hard $\mathcal{O}_t$ as LSTM input $w_{t+1}^s$ at the next time step and in $D_\eta$, hence the gradient becomes:

$$\nabla_\theta \mathcal{L}_G^I(\theta) = \nabla_\theta \log(D_\eta(I, (\mathcal{O}_1, \dots \mathcal{O}_T))).$$

Observe that this loss can be additionally regularized with Feature Matching (FM) as follows:

$$\mathcal{L}_G^I(\theta) = \log(D_\eta(I, (y_1, \dots y_T)))$$
$$- \lambda_F^I \left(||E_I(w_1^*, \dots w_T^*) - E_I(y_1, \dots y_T)||^2\right)$$
$$- \lambda_F^S \left(||E_{S=(w_1^*, \dots w_T^*)}(I) - E_{S=(y_1, \dots y_T)}(I)||^2\right), \quad (2)$$

where $(w_1^*, \dots w_T^*)$ is the ground truth caption corresponding to image $I$, and $E_I$ and $E_S$ are co-attention image and sentence embeddings (as defined in Section 2.1). Feature matching enables us to incorporate more granular information from discriminator representations of the ground truth caption, similar to how SCST reward can be regularized with CIDEr, computed with a set of baseline captions.

## 3. Evaluation: Semantic Score and OOC Set

**Semantic Score.** Traditional automatic language metrics, such as CIDEr or BLEU4, are inadequate for evaluating GAN-based image caption models. As an early alternative, [4, 25] used GAN discriminator for evaluation, but this is not a fair comparison across models since the GAN generator was trained to maximize the discriminator likelihood. In order to enable automatic evaluation across models we propose the *semantic score*. Analogous to "Inception Score" [24] for image generation, leveraging a large pretrained classification network, the semantic score relies on a powerful model, trained with supervision, to heuristically evaluate caption quality and its alignment to the image. In Section 4 we show that our semantic score correlates well with human judgement across metrics, algorithms and test sets.

The semantic score is based on a Canonical Correlation Analysis (CCA) retrieval model [18] which brings the image into the scoring loop by training on the combination of COCO [13] and SBU [19] (∼1M images), ensuring a larger exposure of the score to diverse visual scenes and captions, and lowering the COCO dataset bias. The semantic score is a cosine similarity in CCA space based on a 15k dimension image embedding from resnet-101 [7], and a sentence embedding computed using a Hierarchical Kernel Sentence Embedding [18] based on word2vec [17]:

$$s(x, y) = \frac{\langle \Sigma U^\top x, V^\top y \rangle}{\|\Sigma U^\top x\|_2 \, \|V^\top y^*\|_2},$$

where $x$ and $y$ are caption and image embedding vectors, respectively; $U$, $\Sigma$, and $V$ are matrices obtained from CCA as described in details in [18]. Note that the use of word2vec allows the computation of scores for captions whose words fall outside of the COCO vocabulary. The computed score can be interpreted as a likelihood of the image given a caption, it also penalizes the sentences which mention non-existent attributes or objects. See Table 4 in Appendix A for examples.

**Out of Context Set (OOC).** An important property of the captioner is the ability to generalize to images with objects

| | COCO Test Set | | | | | | | | OOC (Out of Context) | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CIDEr | | METEOR | | Semantic Score | | Vocabulary Coverage | | CIDEr | | METEOR | | Semantic Score | | Vocabulary Coverage | |
| CE | 101.6 | ±0.4 | 0.260 | ±.001 | 0.186 | ±.001 | 9.2 | ±0.1 | 42.2 | ±0.6 | 0.169 | ±.001 | 0.118 | ±.001 | 2.8 | ±0.1 |
| CIDEr-RL | **116.1** | ±0.2 | 0.269 | ±.000 | 0.184 | ±.001 | 5.1 | ±0.1 | 45.0 | ±0.6 | 0.170 | ±.003 | 0.117 | ±.002 | 2.1 | ±0.0 |
| $GAN_1$(SCST, Co-att, $\log(D)$) | 97.5 | ±0.8 | 0.256 | ±.001 | 0.190 | ±.000 | 11.0 | ±0.1 | 41.0 | ±1.6 | 0.168 | ±.003 | **0.124** | ±.000 | 3.2 | ±0.1 |
| $GAN_2$(SCST, Co-att, $\log(D)+5\times$CIDEr) | 111.1 | ±0.7 | **0.271** | ±.002 | **0.192** | ±.000 | 7.3 | ±0.2 | **45.8** | ±0.9 | **0.173** | ±.001 | 0.122 | ±.002 | 2.8 | ±0.1 |
| $GAN_3$(SCST, Joint-Emb, $\log(D)$) | 97.1 | ±1.2 | 0.256 | ±.002 | 0.188 | ±.000 | 11.2 | ±0.1 | 41.8 | ±1.6 | 0.167 | ±.002 | 0.122 | ±.001 | 3.3 | ±0.0 |
| $GAN_4$(SCST, Joint-Emb, $\log(D)+5\times$CIDEr) | 108.2 | ±4.9 | 0.267 | ±.004 | 0.190 | ±.000 | 8.3 | ±1.6 | 45.4 | ±1.4 | 0.173 | ±.002 | 0.122 | ±.003 | 2.8 | ±0.2 |
| $GAN_5$(Gumbel Soft, Co-att, $\log(D)$) | 93.6 | ±3.3 | 0.253 | ±.007 | 0.187 | ±.002 | 11.1 | ±1.2 | 38.3 | ±3.7 | 0.164 | ±.006 | 0.121 | ±.004 | 3.3 | ±0.3 |
| $GAN_6$(Gumbel ST, Co-att, $\log(D)$) | 95.4 | ±1.5 | 0.249 | ±.004 | 0.184 | ±.003 | 10.1 | ±0.9 | 38.5 | ±1.9 | 0.161 | ±.005 | 0.116 | ±.004 | 3.0 | ±0.2 |
| $GAN_7$(Gumbel ST, Co-att, $\log(D)$+FM) | 92.1 | ±5.4 | 0.243 | ±.011 | 0.175 | ±.006 | 8.6 | ±0.8 | 36.8 | ±2.3 | 0.157 | ±.006 | 0.110 | ±.005 | 2.5 | ±0.2 |
| CE* – *denotes *non-attentional* models | 87.6 | ±1.2 | 0.242 | ±.001 | 0.175 | ±.002 | 9.9 | ±0.8 | 32.0 | ±0.4 | 0.152 | ±.002 | 0.103 | ±.002 | 2.6 | ±.1 |
| CIDEr-RL* | 100.4 | ±7.9 | 0.253 | ±.006 | 0.173 | ±.002 | 6.8 | ±1.4 | 33.4 | ±1.4 | 0.154 | ±.003 | 0.101 | ±.003 | 2.1 | ±.2 |
| $GAN_1$*(SCST, Co-att, $\log(D)$) | 89.7 | ±0.9 | 0.246 | ±.001 | 0.184 | ±.001 | **13.2** | ±0.2 | 30.8 | ±1.0 | 0.155 | ±.003 | 0.111 | ±.001 | 3.4 | ±0.1 |
| $GAN_2$*(SCST, Co-att, $\log(D)+5\times$CIDEr) | 103.1 | ±0.5 | 0.261 | ±.001 | 0.183 | ±.001 | 7.1 | ±0.2 | 33.7 | ±1.9 | 0.157 | ±.001 | 0.108 | ±.001 | 2.7 | ±0.1 |
| $GAN_3$*(SCST, Joint-Emb, $\log(D)$) | 90.7 | ±0.1 | 0.248 | ±.001 | 0.181 | ±.001 | 12.9 | ±0.1 | 30.8 | ±2.1 | 0.153 | ±.002 | 0.108 | ±.001 | **3.5** | ±0.1 |
| $GAN_4$*(SCST, Joint-Emb, $\log(D)+5\times$CIDEr) | 102.7 | ±0.4 | 0.260 | ±.001 | 0.182 | ±.001 | 7.7 | ±0.1 | 33.3 | ±2.4 | 0.157 | ±.004 | 0.106 | ±.000 | 2.7 | ±0.1 |
| G-GAN [4] from Table 1 | 79.5 | – | 0.224 | – | – | – | – | – | – | – | – | – | – | – | – | – |

Table 2: Results for all models mentioned in this work. Scores are reported for both COCO and OOC sets. All results are averaged (± standard deviation) over 4 models trained with different random seeds. See Table 5 in Appendix B for a full set of results.

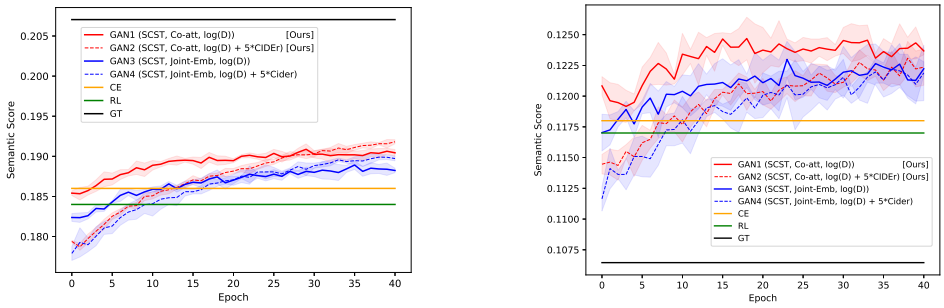
falling outside of their common contexts. In order to test the compositional and generalization properties to out-of-context scenes (see Figure 7 for an example), we expanded the original set of [10] (containing 218 images) to a total of 269 images and collected 5 captions per image on Amazon MTurk. We call the resulting dataset the Out of Context (OOC) set. We note that although the size of OOC set is not large, its main purpose is to be a useful quick diagnostic tool rather than a traditional dataset. The evaluation on OOC is a good indicator of a captioner's generalization: poor performance is a sign that the model is over-fitted to the training context. Improving OOC scores remains an open area for future work, and we plan to release the OOC set as well as the scripts for computing the semantic score.

## 4. Experiments

**Experimental Setup.** We evaluate our proposed method and the baselines on COCO dataset [13] (vocabulary size is 10096) using data splits from [11]: training set of 113k images with 5 captions each, validation and test sets 5k each; as well as on the proposed OOC diagnostic set. Each image is encoded by a resnet-101 [7] without rescaling or cropping, followed by a spatial adaptive max-pooling to ensure a fixed size of 14×14×2048. An attention mask is produced over the 14×14 spatial locations, resulting in a spatially averaged 2048-dimension representation. LSTM hidden state, image, word, and attention embedding dimensions are fixed to 512 for all models. Before the GAN training, all the models are first pretrained with cross entropy (CE) loss. We report standard language evaluation metrics, the proposed semantic score, and the vocabulary coverage (percentage of vocabulary used at generation).

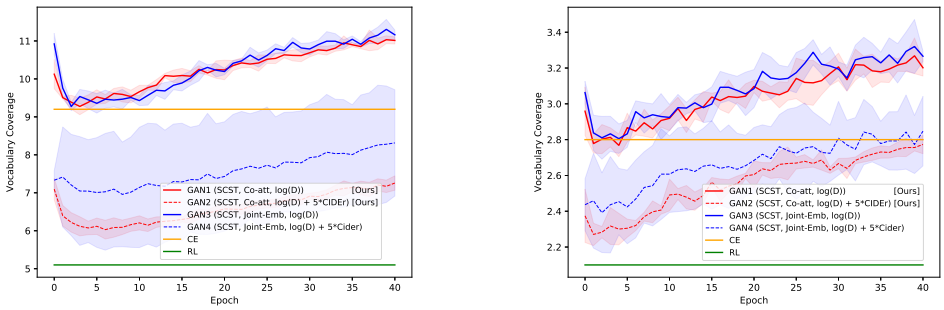
**Experimental Results.** Table 2 presents results for both COCO and OOC datasets for two discriminator architectures (ours Co-att, and baseline Joint-Emb) for all training algorithms (SCST, Gumbel ST, and Gumbel Soft). For reference, we also include results for non-GANs captioners: CE (trained only with cross entropy) and CIDEr-RL (pretrained with CE, followed by SCST to optimize CIDEr), as well as results from non-attentional models. As expected, CIDEr-RL greatly improves the language metrics as compared to the CE model (from 101.6 to 116.1 CIDEr on COCO), but this also leads to a significant drop in the vocabulary coverage (from 9.2% to 5.1% for COCO), indicating that the $n$-gram optimization can lead to vanilla sentences, discouraging style deviations from the ground truth captions. In the table, $GAN_1$ , . . . , $GAN_4$ denote the GAN-based models, where we use SCST training (with $\log(D)$ or $\log(D)+5\times$CIDEr rewards) with either Co-att or Joint-Emb discriminators; and $GAN_5$ , . . . , $GAN_7$ are the models trained with the Gumbel relaxation. From our extensive experiments we observed that SCST provides significantly more stable training of the models and better results as compared to Gumbel approaches, which often become unstable beyond 15 epochs and underperform SCST GANs on many evaluation metrics (see also Section E in Supplement for additional discussion on SCST vs. Gumbel).

It can noticed that SCST GAN models outperform CE and CIDEr-RL captioners on semantic score and vocabulary coverage for both COCO and OOC sets. The CIDEr regularization of SCST GAN additionally improves CIDEr and METEOR scores, and also results in the improvement of the semantic score (at the cost of some vocabulary coverage loss) as seen for $GAN_1$ vs. $GAN_2$ and $GAN_3$ vs. $GAN_4$. We also see that SCST GANs using our Co-att discriminator




(a) COCO Test (b) OOC

Figure 4: Evolution of semantic scores over training epochs for COCO Test and OOC datasets. Our Co-att models achieve consistently higher scores than CE, RL and Joint-Emb models [4].



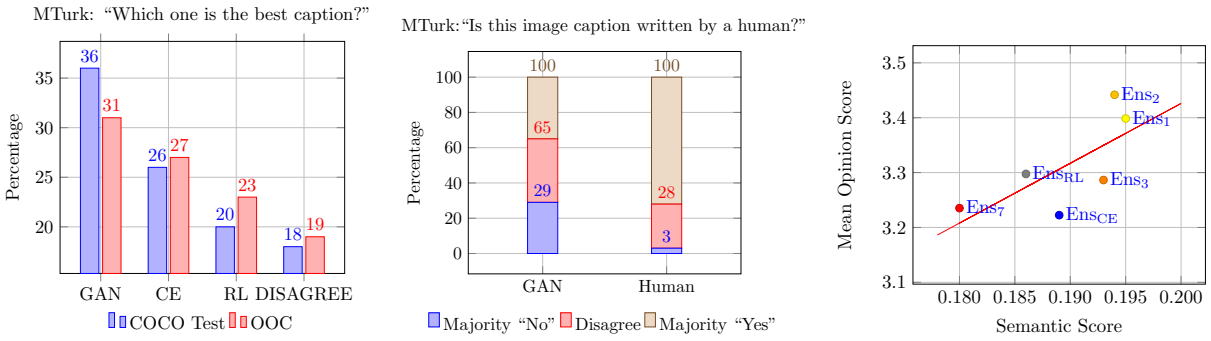

(a) COCO Test (b) OOC

Figure 5: Evolution of vocabulary coverage over training epochs for COCO and OOC datasets. As training progresses, we see a correlation between vocabulary coverage and semantic scores for all models. Models without CIDEr-regularized SCST GAN rewards achieve best vocabulary coverage.

| | | COCO Test Set | | | | OOC (Out of Context) | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | CIDEr | METEOR | Semantic Score | Vocabulary Coverage | CIDEr | METEOR | Semantic Score | Vocabulary Coverage |
| (CE and RL Baselines) | $\text{Ens}_{\text{CE}}$(CE) | 105.8 | 0.266 | 0.189 | 8.4 | 44.8 | 0.172 | 0.122 | 2.6 |
| | $\text{Ens}_{\text{RL}}$(CIDEr-RL) | **118.9** | 0.273 | 0.186 | 5.0 | 48.8 | 0.175 | 0.122 | 2.1 |
| (SCST, Co-att, $*$) | $\text{Ens}_1(\text{GAN}_1)$ | 102.6 | 0.262 | **0.195** | 9.9 | 44.8 | 0.172 | **0.129** | **3.0** |
| | $\text{Ens}_2(\text{GAN}_2)$ | 115.1 | **0.277** | 0.194 | 7.0 | 48.3 | 0.176 | 0.127 | 2.7 |
| | $\text{Ens}_{12}(\text{GAN}_1,\text{GAN}_2)$ | 113.2 | 0.274 | **0.195** | 7.3 | 49.9 | 0.178 | **0.129** | 2.6 |
| (SCST, Joint-Emb, $*$) | $\text{Ens}_3(\text{GAN}_3)$ | 109.8 | 0.270 | 0.193 | 8.5 | 48.5 | 0.175 | 0.127 | 2.8 |
| | $\text{Ens}_4(\text{GAN}_4)$ | 113.0 | 0.274 | 0.193 | 7.6 | 48.0 | 0.178 | 0.127 | 2.7 |
| | $\text{Ens}_{34}(\text{GAN}_3,\text{GAN}_4)$ | 111.1 | 0.271 | 0.193 | 8.1 | **50.1** | 0.177 | 0.127 | 2.8 |
| (Gumbel $*$, Co-att, $*$) | $\text{Ens}_5(\text{GAN}_5)$ | 100.1 | 0.259 | 0.191 | **10.0** | 43.1 | 0.170 | 0.127 | **3.0** |
| | $\text{Ens}_6(\text{GAN}_6)$ | 99.6 | 0.253 | 0.187 | 9.3 | 41.0 | 0.165 | 0.122 | 2.8 |
| | $\text{Ens}_7(\text{GAN}_7)$ | 100.2 | 0.254 | 0.180 | 7.8 | 38.9 | 0.164 | 0.113 | 2.3 |
| | $\text{Ens}_{567}(\text{GAN}_5,\text{GAN}_6,\text{GAN}_7)$ | 103.2 | 0.258 | 0.188 | 8.7 | 41.8 | 0.164 | 0.121 | 2.7 |
| (SCST+Gumbel Soft, Co-att, $*$) | $\text{Ens}_{125}(\text{GAN}_1,\text{GAN}_2,\text{GAN}_5)$ | 112.4 | 0.273 | 0.195 | 7.7 | 49.8 | **0.179** | 0.129 | 2.7 |

Table 3: Ensembling results for some GANs from Table 2 for COCO and OOC sets. See Table 6 in Appendix B for complete set of results including BLEU4 and ROUGEL.

outperform their Joint-Emb [4] counterparts on *every* metric except vocabulary coverage (for COCO). We conclude that $\text{GAN}_2$, a CIDEr-regularized SCST with Co-att discriminator, is the model with the best overall performance on COCO and OOC sets. For baselining, we also reproduced results from [4] with non-attentional generators (same architecture as in




(a) Best Caption (b) Turing Test (c) MOS COCO

Figure 6: Human evaluations of $\text{Ens}_{\text{CE}}$, $\text{Ens}_{\text{RL}}$ and several GAN ensembles on COCO and OOC sets. (a) A distribution of preferences for the best caption determined by the majority of 5 human evaluators; here GAN label indicates the $\text{Ens}_2$ model. (b) Turing test on detecting the human-written versus GAN-generated captions on COCO. We assign "yes/no' with at least 4 out of 5, disagree otherwise. (c) Mean opinion score vs. Semantic score on COCO test images.

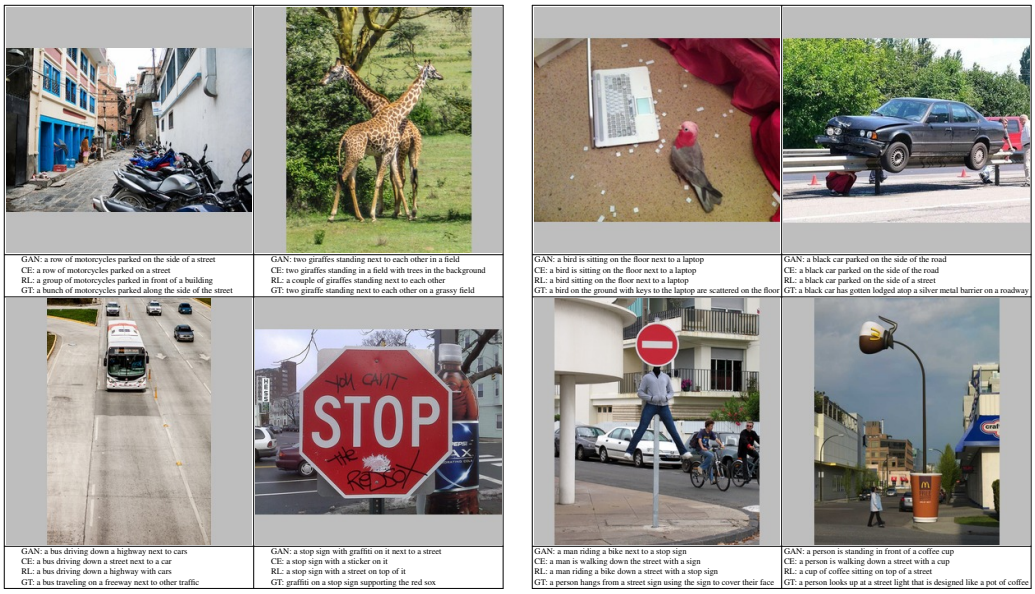
[4]). Non-attentional models are behind in all metrics, except for vocabulary coverage on both datasets. Interestingly, Co-att discriminators still provide better semantic scores than Joint-Emb despite non-attentional generators.

Figures 4 and 5 show the evolution of semantic scores and vocabulary coverage over the training epochs for $\text{GAN}_1, \ldots \text{GAN}_4$, CE, CIDEr-RL and ground truth (GT) captions. Semantic scores increase steadily for all cost functions and discriminator architectures as the training sees more data. In Figure 4 (a), GAN models improve steadily over CE and RL, ultimately surpassing both of them mid-training. Moreover, Co-att GANs achieve higher semantic scores across the epochs than Joint-Emb GANs. For CIDEr-regularized SCST GANs, the same trend is observed but with a faster rate since the models start off worse than CE and RL. For OOC in Figure 4 (b), we see the same trend: Co-att GANs outperforming the other approaches. For COCO, GT semantic score is higher than the other models while the opposite is true for OOC. This may be caused by the vocabulary mismatch between OOC and the combination of COCO and SBU. Figures 4 and 5 show that the semantic score improvement of GAN-trained models correlates well with the vocabulary coverage increase for both COCO and OOC.

**Ensemble Models.** Table 3 presents results for *ensemble models*, where a caption is generated by first averaging the softmax scores from 4 different models before word selection. $\text{Ens}_{\text{CE}}$ and $\text{Ens}_{\text{RL}}$ ensemble CE and CIDEr-RL models. Similarly, $\text{Ens}_1, \ldots \text{Ens}_7$ ensemble models from $\text{GAN}_1, \ldots \text{GAN}_7$ respectively ($\text{Ens}_{ijk}$ denotes an ensemble of $\text{GAN}_i$, $\text{GAN}_j$, and $\text{GAN}_k$). As compared to individual models, the ensembles show improved results on *all* metrics. Ensembling SCST GANs provides the best results, reinforcing the conclusion that SCST is a superior method for a stable sequence GAN training. For comparison, we also computed SPICE [1] scores on COCO dataset: $\text{Ens}_{\text{CE}}$ 19.69, $\text{Ens}_{12}$ 20.64 (GAN Co-attention) and $\text{Ens}_{34}$ 20.46 (GAN Joint embedding from [4]), showing that SCST GAN training additionally improves the SPICE metric. Finally, we observe that underperformance of GANs over CIDEr-RL in terms of CIDEr is expected and explained by the fact that in GAN the objective is to make the sentences more descriptive and human-like, deviating from the vanilla ground truth captions, and this can potentially sacrifice the CIDEr performance. The generated captions are evaluated using the proposed semantic score which showed a good correlation with human judgment; see Figure 6 (c) for more details.

**Gradient Analysis.** Throughout the extensive experiments, the SCST showed to be a more stable approach for training discrete GAN nodels, achieving better results compared to Gumbel relaxation approaches. Figure 8 compares gradient behaviors during training for both techniques, showing that the SCST gradients have smaller average norm and variance across minibatches, confirming our conclusion.

**Human Evaluation.** To validate the benefits of the semantic score, we also evaluate the image/caption pairs generated by several GAN ensembles, $\text{Ens}_{\text{CE}}$ and $\text{Ens}_{\text{RL}}$ on Amazon MTurk. For a given image, 5 workers are asked to rate each caption on the scale 1 to 5 (from which we computed mean opinion score (MOS)) as well as to select the best caption overall (additional details are given in Appendix D). Figure 6 (a) shows that GAN ensemble $\text{Ens}_2$ scored higher than CE and CIDEr-RL on a majority vote, confirming that GAN training significantly improves perceived quality of the captions as compared to a more vanilla CE or RL-based captions. Figure 6 (b) gives Turing test results where the workers are asked if a given caption is human or machine-generated. Here, our GANs again performed well, demonstrating a good capacity at fooling humans. In Figure 6 (c) we show that




(a) COCO Test (b) OOC

Figure 7: Examples of captions for our proposed model on COCO and OOC sets.

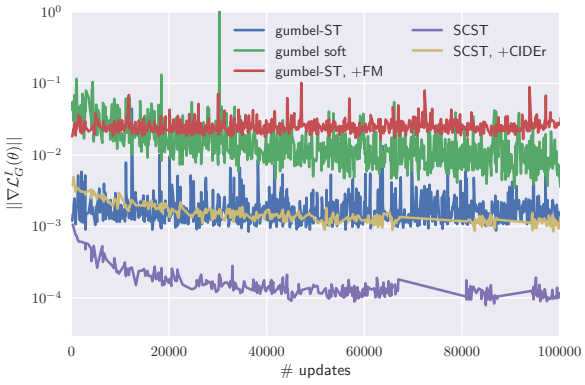



Figure 8: $L_2$ norm of the gradient with respect to the logits during training of $G_\theta$ with different training strategies. The plots show a minibatch-mean during the training; the variance of each curve gives a good idea of the gradient stability between minibatches. We can see that SCST with pure discriminator reward has the lowest gradient norm.

MOS of human evaluations correlates well with our semantic score (see Table 7 in Appendix D for all scores). There is an overall trend (depicted with a red regression line for better visualization), where models that have higher semantic score are generally favored more by human evaluators. For example, Co-att SCST GANs $\text{Ens}_1$ and $\text{Ens}_2$ score the best semantic (resp. 0.195 and 0.194) and MOS scores (resp. 3.398 and 3.442) on COCO. We can see that the semantic score is able to capture semantic alignments pertinent to humans, validating it as a viable alternative to automatic language metrics and a proxy to human evaluation.

Finally, in Figure 7 we present a few examples of the captions for COCO and OOC sets. As compared to the traditional COCO dataset, the OOC images are difficult and illustrate the challenge for the automatic captioning in such settings. The difficulty is not only to correctly recognize the objects in the scene but also to compose a proper description, which is challenging even for humans (see row denoted by GT), as it takes more words to describe such unusual images.

## 5. Conclusion

In conlusion, we summarize the main messages from our study: **1)** SCST training for sequence GAN is a promissing new approach that outperforms the Gumbel relaxation in terms of stability of training and the overall performance. **2)** The modeling part in the captioner is crucial for generalization to out-of-context: we demonstrate that the non-attention captioners and discriminators – while still widely used – fail at generalizing to out of context, hinting at a memorization of the training set. Attentive captioners and discriminators succeed at composing on unseen visual scenes, as was demonstrated with our newly introduced OOC diagnostic set. **3)** Human evaluation is still the *gold standard* for assessing the quality of GAN captioning. We showed that the introduced semantic score correlates well with the human judgement and can be a valuable addition to the existing evaluation toolbox for image captioning.

## References


[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 1, 7

[2] M. Caccia, L. Caccia, W. Fedus, H. Larochelle, J. Pineau, and L. Charlin. Language gans falling short. *arXiv preprint arXiv:1811.02549*, 2018. 1

[3] T. Che, Y. Li, R. Zhang, D. R. Hjelm, W. Li, Y. Song, and Y. Bengio. Maximum-likelihood augmented discrete generative adversarial networks. *arXiv:1702.07983*, 2017. 1

[4] B. Dai, D. Lin, R. Urtasun, and S. Fidler. Towards diverse and natural image descriptions via a conditional GAN. *ICCV*, 2017. 1, 2, 3, 4, 5, 6, 7, 11, 13

[5] C. N. dos Santos, M. Tan, B. Xiang, and B. Zhou. Attentive pooling networks. *Arxiv*, abs/1602.03609, 2016. 2

[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *NIPS*, 2014. 1

[7] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 5

[8] R. D. Hjelm, A. P. Jacob, T. Che, K. Cho, and Y. Bengio. Boundary-seeking generative adversarial networks. *arXiv:1702.08431*, 2017. 1, 4

[9] E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 1, 3, 4

[10] M. Jinchoi, A. Torralba, and A. S. Willsky. Context models and out-of-context objects, 2012. 5

[11] A. Karpathy and F.-F. Li. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015. 1, 5

[12] M. J. Kusner and J. M. Hernández-Lobato. Gans for sequences of discrete elements with the gumbel-softmax distribution. *arXiv:1611.04051*, 2016. 1, 4

[13] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: common objects in context. *EECV*, 2014. 4, 5

[14] S. Liu, Z. Zhu, N. Ye, S. Guadarrama, and K. Murphy. Improved image captioning via policy gradient optimization of spider. In *ICCV*, 2017. 1, 4

[15] J. Lu, C. Xiong, D. Parikh, and R. Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *CVPR*, 2017. 2

[16] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 2016. 2

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *ArXiv*, 2013. 4

[18] Y. Mroueh, E. Marcheret, and V. Goel. Multimodal retrieval with asymmetrically weighted CCA and hierarchical kernel sentence embedding. *ArXiv*, 2016. 4, 10

[19] V. Ordonez, G. Kulkarni, and T. L. Berg. Im2text: Describing images using 1 million captioned photographs. In *NIPS*, 2011. 4

[20] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu. Bleu: A method for automatic evaluation of machine translation. In *ACL*, 2002. 1

[21] S. Rajeswar, S. Subramanian, F. Dutil, C. Pal, and A. Courville. Adversarial generation of natural language. *arXiv:1705.10929*, 2017. 1

[22] M. Ranzato, S. Chopra, M. Auli, and W. Zaremba. Sequence level training with recurrent neural networks. *ICLR*, 2015. 1

[23] S. J. Rennie, E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. Self-critical sequence training for image captioning. In *CVPR*, 2017. 1, 2, 3

[24] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *NIPS*, 2016. 4

[25] R. Shetty, M. Rohrbach, L. A. Hendricks, M. Fritz, and B. Schiele. Speaking the same language: Matching machine to human captions by adversarial training. *ICCV*, 2017. 1, 3, 4

[26] R. Vedantam, C. L. Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, 2015. 1

[27] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CVPR*, 2015. 1

[28] 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*, 2015. 1, 2

[29] L. Yu, W. Zhang, J. Wang, and Y. Yu. Seqgan: Sequence generative adversarial nets with policy gradient. *CoRR*, abs/1609.05473, 2016. 1