Joint Optimization for Cooperative Image Captioning

Gilad Vered
Bar-Ilan University

Gal Oren
Bar-Ilan University

Yuval Atzmon
Bar-Ilan University, NVIDIA

Gal Chechik
Bar-Ilan University, NVIDIA

gal.chechik@biu.ac.il

Abstract

When describing images with natural language, descriptions can be made more informative if tuned for downstream tasks. This can be achieved by training two networks: a "speaker" that generates sentences given an image and a "listener" that uses them to perform a task. Unfortunately, training multiple networks jointly to communicate, faces two major challenges. First, the descriptions generated by a speaker network are discrete and stochastic, making optimization very hard and inefficient. Second, joint training usually causes the vocabulary used during communication to drift and diverge from natural language.

*To address these challenges, we present an effective optimization technique based on partial-sampling from a multinomial distribution combined with straight-through gradient updates, which we name **PSST** for Partial-Sampling Straight-Through. We then show that the generated descriptions can be kept close to natural by constraining them to be similar to human descriptions. Together, this approach creates descriptions that are both more discriminative and more natural than previous approaches. Evaluations on the COCO benchmark show that PSST improve the recall@10 from 60% to 86% maintaining comparable language naturalness. Human evaluations show that it also increases naturalness while keeping the discriminative power of generated captions.*

1. Introduction

Describing images with natural language is a key step for developing automated systems that communicate with people. The complementary part of this human-machine communication involves networks that can understand natural descriptions of images. Both of these tasks have been studied intensively, but mostly as two separate problems, *image captioning* and *image retrieval*. It is natural to "close the loop" and seek to jointly train networks to cooperatively communicate about visual content in natural language.

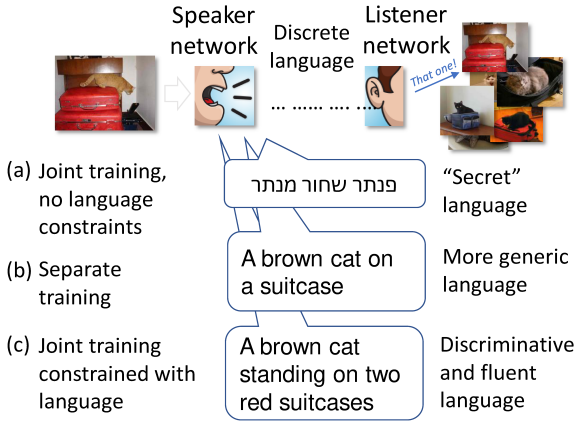

Figure 1: Challenges in training agents to communicate about an image. **(a)** When a speaker network is trained jointly with a listener network, the communicated language may drift away from natural language, unless constrained, yielding language that no longer maps to standard English terms. **(b)**: When the two network agents are trained separately, descriptions become less specific, because agents cannot expect the other side to "understand" subtleties. **(c)**: Training both networks jointly while keeping communication close to natural language can yield descriptions that are more discriminative while maintaining intelligibility.

Training multiple networks to communicate has been studied in the context of visual dialogues [11, 12]. There, a sequence of sentences is passed back-and-forth between learning agents. Here, we take a step back and focus on a *single* transmission between a "speaker network" and a "listener network". We seek to develop the building blocks of trainable communication by training both speaker and listener to communicate effectively with natural language.

What should be the properties of such natural communication? Good visual descriptions should obey two competing objectives. First, a description should be natural and fluent, using well-formed and meaningful sentences, so they can be communicated to people. Second, a description should be image-specific and informative, capturing the rel-

evant elements of an image that make it unique. Aiming to address these two objectives, several studies trained models that consists of a speaker and a listener networks [9, 30, 31] with corresponding losses to achieve both goals.

Unfortunately, training a speaker network together with a listener network faces multiple challenges, primarily language-drift and optimization. First, when the listener and speaker can tune their communication, the resulting language typically **drifts away**, losing its original semantic meaning and becoming confusing when communicated to people. For instance, networks may assign a new meaning (blue) to a common word (red), or code highly specific information within a single symbol ("field" means "giraffes near a tree") [24].

Second, training end-to-end speaker-listener systems requires to optimize through an intermediate communicating layer which is discrete and stochastic. Standard backpropagation of gradients cannot be applied to such layers [4] and alternative methods are often complex or slow to converge [38, 44]. Because of these limitations, previous discriminatie captioning approaches like in [31, 39] avoided end-to-end training or obtained limited quality captions [9].

The current paper addresses these two challenges. First, we show that keeping the discriminative captions close to human-generated captions, is sufficient for maintaining fluent and well-formed language while providing enough flexibility such that captions are discriminative. Second, we develop a new effective optimization procedure for jointly training a cooperative speaker-listener network. It is based on partial-sampling from a multinomial distribution combined with straight-through (ST) gradient updates, which we name **PSST** for *Partial-Sampling Straight-through*. It can be applied to a multinomial model or with ST Gumbel Softmax. PSST is very simple to implement and robustly outperformed all baselines it was compared with.

This paper makes the following novel contributions (1) A new and simple partial-sampling procedure for optimizing through discrete stochastic layers, directly applicable to generating discriminative language. (2) New state-of-the-art results on MS COCO discriminative captioning, improving recall from ~60% to ~86% for similar naturalness evaluated using CIDEr and human evaluation. (3) Systematic evaluation of all the leading approaches for optimization through stochastic layers, using a unified captioning benchmark. (4) An evaluation scheme that explicitly quantifies the full curve of naturalness-vs-discriminability, instead of a one-dimensional metric.

2. Discriminative captioning

In our setup of discriminative captioning, two networks cooperate to communicate the content of a given image (Figure 1). The first network, the **speaker**, is given an image and produces a series of discrete tokens that describe the image *in natural language*. Each token is represented by a 1-hot vector from a predefined vocabulary. The second network, the **listener**, takes this series of tokens and uses it to find the input image among a set of distractor images.

In this setup, the speaker network is trained to focus on the unique features of an image that would allow the listener to detect it among distractors. Unlike [21, 39], the distractor images are not available to the speaker as an explicit context. Importantly, The two networks share a common goal: communicate such that the listener identifies the image that the speaker described. Their interaction therefore defines a cooperative game, which is fundamentally different from GAN-based adversarial approaches [9].

We address this task by training the speaker network (parametrized by ϕ) jointly with the listener network (parametrized by θ). When considering the objective function of this joint optimization, it must contain two complementing components. First, as a **discriminability loss** l^{disc}, the objective contains the loss suffered by the listener when detecting the target image I among distractors using the produced sentence $\mathbf{w}$. Second, since natural language is far from optimal for this task, the networks can find other communication schemes that drift away from natural language. To keep the communication interpretable to people, we add a second component to the objective, a **naturalness loss** l^{nat}. It is aimed to ensure that the produced sentence $\mathbf{w}$ is natural, by keeping it similar to human-created captions for that image I. Together, the optimization problem is

$$\min_{\phi,\theta} \lambda l^{disc}(\mathbf{w}, I) + (1-\lambda) l^{nat}(\mathbf{w}). \tag{1}$$

Here ϕ are the parameters of the speaker network and θ are the parameters of the listener network. For the naturalness loss, we use CIDEr [40], $l^{nat}(\mathbf{w}) = -CIDEr(\mathbf{w})$. For the discriminative loss l^{disc}, we use the sum of two hinge losses: one for selecting the correct image among a batch of distractor images, and a second for selecting the correct caption among a batch distractor captions as in [13]:

$$\begin{aligned} l^{disc}(\mathbf{w}, I) = &\max\left[0, 1 - \Phi_\theta(\mathbf{w}, I) + \Phi_\theta(\mathbf{w}', I)\right] \\ &+ \max\left[0, 1 - \Phi_\theta(\mathbf{w}, I) + \Phi_\theta(\mathbf{w}, I')\right], \end{aligned} \tag{2}$$

where $\mathbf{w}'$ is the hardest negative caption among candidate captions, I' is the hardest negative image and Φ is the cosine similarity over the embedding of the image and captions. Instead of fixing a single value of λ, we compute the full curve that captures the trade-off between discriminative and natural descriptions, obtained by optimizing the model with varying values of λ.

3. Related work

Image captioning has been studied intensively since encoder-decoder models were introduced [45], aiming to make captions more natural, diverse and distinctive.

Naturalness. Several efforts to improve caption naturalness use conditional GANs, since using the adversarial discriminator alleviates the obstacle of defining a language-naturalness loss [7, 9, 37]. As an alternative to trained losses, information theory can also be used to select description terms that are natural [6].

Diversity. Several techniques were designed to improve the diversity of captions generated for a given image [8, 27, 37, 42]. [8] used a hierarchical compositional model over captions, [27] used a discriminator that compares a human-generated, unrelated and generated captions and [43] suggested a metric to evaluate captions diversity.

Discriminability. Generating captions that describe an image in a distinct way is key for effective captioning. Such captions allow discriminating an image from other similar images. Also, early captioning models suffered from poor generalization, often producing over-generic captions, and making captions more discriminative may alleviate that problem. In [39, 2], distractor images were used at inference time to create a distinctive caption. [19] recently described a dataset that contains pairs of closely similar images, that can be used as hard-negatives for evaluating image retrieval tasks. In [10] captions are made more distinct using contrastive learning, where the estimated conditional probability for caption-image pair is required to be higher than the reference model to positive pair and lower for negative pair. [30] trained a discriminative captioning model with REINFORCE over CIDEr rewards, using a self-retrieval module to select hard negatives and CIDEr reward. Most relevant to the current paper is [31]. They use a pre-trained listener network to increase discriminate power of captions. However, to avoid language drift, the listener was kept fixed, rather than trained jointly with the speaker.

Several studies characterized the (non-human-readable) languages that are learned when agents communicate in visual tasks [5, 24, 25, 26]. The current paper purposefully focuses on keeping the language close to natural, rather than study properties or emergent language.

4. Optimizing discrete stochastic layers

Joint training of two networks communicating through a language layer is equivalent to training a network that has an intermediate layer that is discrete and stochastic. We first define formally the learning setup and then describe existing optimization approaches for this setup.

In our model (Figure 2), caption generation is treated as a stochastic process. At each step, $t = 0, \ldots, T$ the caption generator (the speaker) outputs a distribution over a vocabulary of words $p_\phi(w_t|I, w_0, ..., w_{t-1})$. This distribution depends on the input image I and the previous terms in the sentence and is parametrized by the deterministic parameters ϕ. We therefore treat the output of the speaker network $s_\phi(I)$ as a random sequence W with a distribution $p_\phi(\mathbf{w}|I)$ over all word sequences $\mathbf{w}$. From that distribution, one specific sequence is sampled and passed to the listener. Given this sampled word sequence $\mathbf{w} = w_0, \ldots, w_T$, the listener network, parametrized by θ, makes a prediction $\hat{y} = f_\theta(\mathbf{w})$ = $f_\theta(s_\phi(I))$ and suffers a loss $l(y, \hat{y}; \theta)$. Our goal is to propagate the gradient of that loss, first to update the parameters of the listener θ and then through the stochastic layer to update the parameters of the speaker ϕ.

Training the parameters of the *listener* network poses no special problems. The function f_θ implemented by the listener is deterministic and differentiable (almost everywhere), hence gradients of the losses can be propagated in the standard way. This is also true for propagating the gradients back through the sequence of terms in a sentence using standard "back-propagation through time".

Unfortunately, for the *speaker* network, parameter tuning is harder because this network emits discrete terms in a stochastic way, making the speaker network non-differentiable. Computation in stochastic neural networks can be formalized using stochastic-computation graphs (SCGs) [36]. In our case, we view the computation graph as including a single stochastic computing node, corresponding to the random sequence W. We think about the listener network as providing the speaker with a loss $l_\theta(\mathbf{w})$ for every (sampled) sentence $\mathbf{w}$. Our goal is to minimize the expected loss $\min_\phi L(\theta, \phi)) = \min_\phi E_{\phi(\mathbf{w})}\left[l_\theta(\mathbf{w})\right]$. The gradient of this objective w.r.t. the speaker parameters ϕ is $\nabla_\phi \int p_\phi(\mathbf{w}) l_\theta(\mathbf{w}) d\mathbf{w} = \int \nabla_\phi p_\phi(\mathbf{w}) l_\theta(\mathbf{w}) d\mathbf{w}$. Since this gradient does not have a form of an expectation, it cannot be directly estimated efficiently by sampling.

Before describing our approach to estimate these gradients, we briefly describe the two main existing approaches to this problem: Score-function estimators and Straight-through Gumbel softmax.

Score-function estimators [14, 15], and specifically the REINFORCE algorithm [44], are often described in the context of reinforcement learning. There, an agent aims to maximize its reward by choosing the best action for a given state according to a policy. In our context, the state is determined by the input image and the preceding words, the actions correspond to the set of words that can be emitted at a given time step, and the reward is (minus) the loss imposed by the listener. REINFORCE yields an unbiased estimator of the gradient, but its variance tends to be large. Several variance reduction techniques were proposed [16, 17, 33, 38]. but due to their complexity, their adoption is still limited.

ST Gumbel Softmax [20, 32], the second main approach to optimize the stochastic discrete layer, consists of three components. (1) To handle stochasticity, the computation graph is reparameterized, allowing to propagate gradients through deterministic paths [23, 32, 35]. (2) A Gumbel max process is used for sampling from a pre-determined distri-

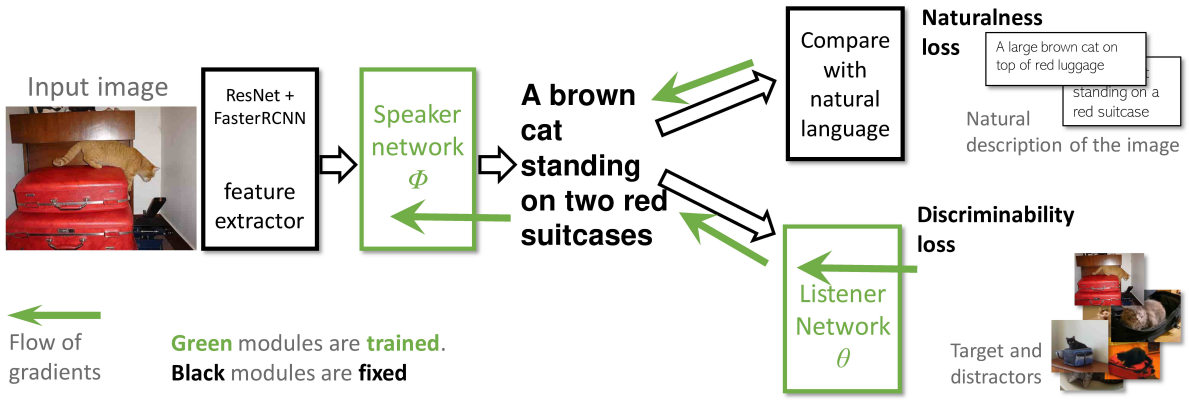

Figure 2: The model architecture. The speaker network and the listener networks are trained jointly, by passing gradients through the text layer. The loss contains two components that are linearly weighted with a hyper parameter tuned on a validation set. **Naturalness loss:** A measure of agreement between a generated caption and a set of predefined, ground-truth captions for that image, using a CIDEr score. Those captions need not be discriminative. **Discriminative loss:** Measures how well a listener can identify the input image among a set of 127 randomly-chosen distractor images.

bution and the Gumbel distribution [18] is relaxed using a Gumbel softmax [20]. (3) A "straight-through trick" (ST) [4] is used: In the backward pass, gradients are computed as if the full continuous distribution was passed. In the forward pass, tokens are sampled from that distribution. More details about these methods for the current context are given in our arXiv report [41].

5. Partially-sampled straight-through

For image captioning, using straight-through with Gumbel softmax approach, as described above, suffers from high variance and bias.

The **variance** is high because the forward pass is stochastic. At each step, the speaker computes the probability distribution $p_\phi(w)$ over the vocabulary V of terms to be emitted, then draws and emits a single term. This adds inherent variance and conveys less information per sample than passing the full continuous distribution. The added variance hurts optimization because presenting the same input to the network leads to different estimates of the gradients. It can be viewed as adding noise to the loss, or as training with noisy labels, which hurts convergence to good minima. The effect of this variance on generated captions is demonstrated in rows 2-3 of Figures 4b,c.

Furthermore, the ST estimator is also **biased**, because the estimates of the gradients are computed as if the full distribution was passed. It would have been preferable to pass the full distribution without sampling, but unfortunately, at test time we must produce discrete word selections to generate specific sentences.

We propose a simple-to-implement procedure we call *partial-sampling straight-through* (PSST). During training, we pass the full continuous distribution for a fraction ρ of the terms and pass a sampled one-hot for the remaining $1-\rho$. More formally, at each step, the speaker computes the probability distribution $p_\phi(w_t)$ over the vocabulary V of terms to be emitted. Then, we randomly draw a binary value. With probability ρ, the speaker passes the full multinomial distribution $p_\phi(w_t)$. With probability 1-ρ, it samples a value from that distribution and emits the 1-hot vector corresponding to that term.

As a result, for ρ of the terms, the stochastic and discrete units are practically replaced by a deterministic continuous variable, In the extreme case of $\rho = 0$, the speaker always operates as a sampler, and optimization can be viewed as a multinomial version of the binary ST estimator of [4]. In the other extreme case of $\rho = 1$, the speaker operates as a deterministic mapper, and outputs a set of dense multinomial distributions.

This approach has several advantages. First, for ρ of the terms, the estimator of the gradient is exact, because computation is deterministic, therefore reducing the overall bias and variance of gradient estimation. At the same time, for $1-\rho$ of training images, the downstream listener network does experiences stochastic variations, with terms represented as 1-hot vectors, and learns to classify them correctly. This allows it to correctly handle one-hot samples that are observed during the test phase. We find empirically that this approach is highly effective and robust with respect to the value of ρ.

Partial sampling takes advantage of the cooperative nature of the speaker-listener relations. Unlike GAN training (e.g. [9]), where the generator works hard not to reveal any information that may give away its generated captions, the speaker in cooperative games has an explicit aims to convey as much information as possible to the listener. Specifically,

during training, it is allowed to represent generated captions as continuous distributions, which look very different than human-created captions, and would be easily discriminated by GANs. More generally, the fundamental differences in the "game matrix" of communicating agents, cooperative vs competitive, are important to consider when developing joint optimization schemes.

The above discussion describes how we optimize $l^{disc}(\mathbf{w})$ with PSST Multinomial. Unforuntately, to optimize $l^{nat}(\mathbf{w})$, one cannot use PSST because CIDEr requires sparse descriptors as input. Instead, one can use any of the standard methods described in Section 4. In practice, we used REINFORCE because preliminary experiments showed that its performance was comparable to the other approaches.

6. Experiments

We evaluate our approach with two image captioning benchmark datasets: COCO [29] and Flickr30k [46], and compared to seven baselines.

6.1 Datasets. COCO has $\sim 123K$ images annotated with 5 human-generated captions. For a fair comparison with previous work, we used the same data split as in [31, 39], assigning $\sim 113K$, $5K$ and $5K$ images for training, validation and test splits, and using 9487 words [31].

Flickr30K has $\sim 31K$ images, annotated with 5 human-generated captions for a total of $\sim 159K$ captions. We used the split of [22] assigning $29K$, $\sim 1K$ and $1K$ images for train, validation and test splits. The vocabulary contains words that appeared more than 5 times in the annotated captions, with a total of $7K$ words. Captions were clipped to a maximum length of 16.

6.2 Compared Methods. **(1)** PSST MULTINOMIAL. The method of section 5. **(2)** ST MULTINOMIAL. As (1) with $\rho = 0$ (always sampling) **(3)** LUO *et al*. 2018 [31]. Speaker was trained using REINFORCE and a "frozen" pre-trained listener. **(4)** REINFORCE [44]. The speaker and listener were trained alternately. **(5)** ST GUMBEL SOFTMAX [20]. During back-propagation gradients flow through the noisy distribution of Gumbel softmax. During forward pass, tokens are sampled from that distribution. **(6)** ST Gumbel softmax where the temperature was annealed using the schedule and hyper parameters of [20], $\tau = \max(0.5, e^{-rt})$. **(7)** SR-PL [30]. As (3) but using unlabeled data to be part of the mini batch as hard negatives. **(8)** PSST GUMBEL SOFTMAX. Similar to PSST Multinomial, but applying partial sampling to the Gumbel-softmax distribution. **(9)** G-GAN [9]. A conditional GAN with a generator trained with policy gradient and early feedback.

6.3 Implementation details are provided [41]. In general, we followed previously-published evaluation protocols and used published hyper parameters whenever available. Our code is available at http://github.com/vgilad/CooperativeImageCaptioning.

6.4 Automated evaluation metrics.

Naturalness of generated captions was quantified by standard linguistic metrics: CIDEr [40], BLEU4 [34], METEOR [3], ROUGH [28] and SPICE [1].

Discriminability of generated captions was quantified by the recall of the listener network. Specifically, at *test* time, given an input image, the listener receives four inputs: the caption generated by the speaker, the input image, 4999 distractor captions and 4999 distractor images. The listener ranks all images based on their compatibility with the caption (measured using the cosine similarity between the image representation and the caption). Based on this ranking, we compute the recall@K, the average detection rate at the top K. Namely, an image is considered detected if the score of the input image is ranked within the top-K scores. We report below recall@1, @5 and @10.

Balancing discriminability with naturalness. During training, we trade-off discriminability vs naturalness by testing multiple values of the parameter λ of Eq. 1, specifically λ in $\{10, 5, 2.5, 1.6, 1, 0.5\} \times 10^{-3}$ (values are small to offset the different scales of the two losses in Eq. 1).

7. Results

We first evaluate the naturalness and discriminability of PSST on COCO. Figure 3 depicts recall@10 as a function of five naturalness scores: CIDEr, BLEU4, METEOR, ROUGE and SPICE. For each method, we trained a series of models, each with a different value of the trade-off parameter λ (the weight of l^{disc} in Eq. 1). With high values of λ, models generate captions that are more discriminative, at the expense language quality, while models trained with low λ generate highly natural captions but with lower discriminability. The values of language metrics are provided in Table 1, for a fixed recall value. PSST Multinomial achieved best scores across all five metrics. Values of Recall for a fixed CIDEr value are reported in Table 2. Here as well, PSST Multinomial outperforms other approaches.

The effect of **joint training** and of **partial sampling** are both significant. All methods that train networks jointly consistently improve over separate training (red curve). Broadly speaking, all three methods, REINFORCE, ST Gumbel softmax and ST Multinomial achieve comparable scores for high naturalness (BLEU4>0.3 or CIDEr>1.1). Second, PSST Multinomial (blue curve) provides a significant further improvement over all baselines. The underlying reason for this improvement is that baseline approaches have high variance because instead of *deterministically* transmitting the full distribution over words, they sample a single word from the distribution and transmit it. In PSST, a frac-

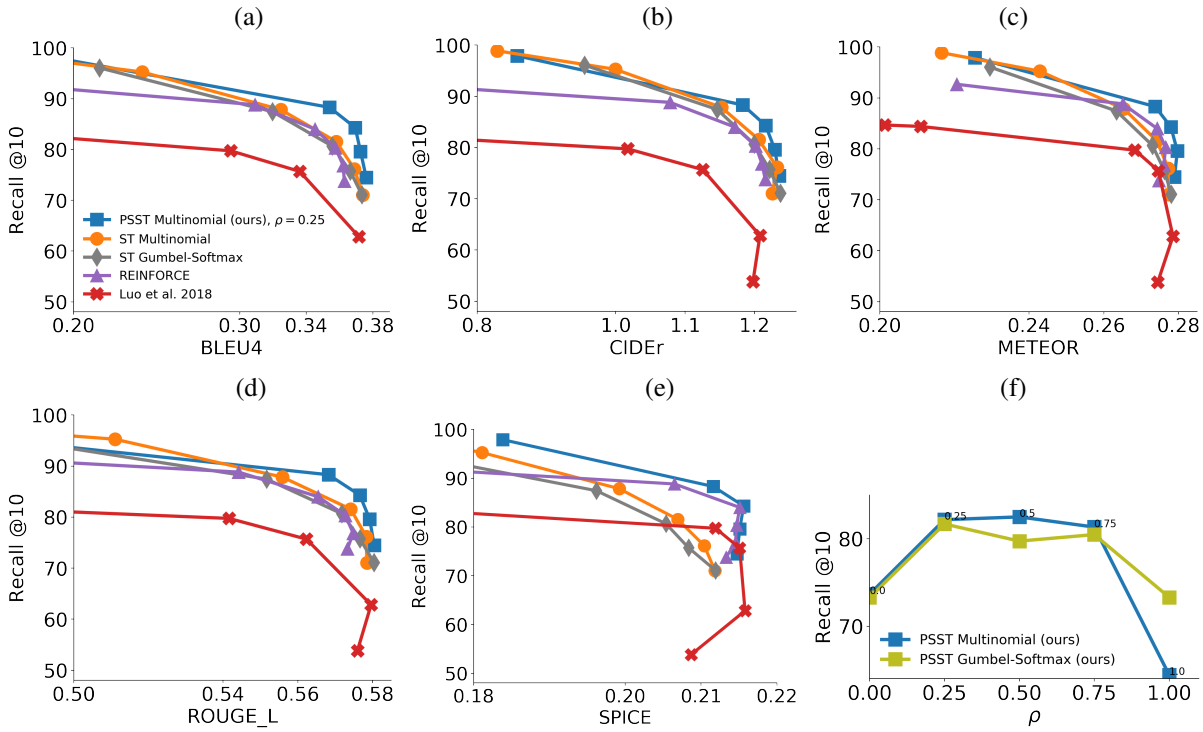

Figure 3: **Discriminability-Naturalness curves for COCO**. **(a-e)** Each panel traces recall@10 as a function of naturalness, as evaluated using five language metrics: (a) BLEU4 (b) CIDEr (c) METEOR (d) ROUGE (e) SPICE. Within each panel, each curve corresponds to an optimization method. Markers on each curve correspond to models trained with different values of the trade-off parameter λ. **(f) The effect of partial sampling rate (ρ) on recall rate**. Recall was extracted from panel (b) by interpolation and selecting scores for CIDEr=1.22. Partial sampling ($0<\rho< 1$) is superior to full- or zero- sampling.

tion ρ of steps are deterministic, passing the full distribution as a vector. **For those steps, PSST variance is zero**, hence PSST reduces the variance by a factor of ρ. Similarly, PSST zeros the bias when it passes the full distribution without sampling, and therefore also cuts the bias by a factor of ρ.

Recall@5=80%	C	B	M	R	S
REINFORCE	0.902	0.251	0.247	0.505	0.189
ST Gumbel Softmax	1.087	0.288	0.253	0.528	0.187
ST Multinomial	1.106	0.300	0.259	0.542	0.194
PSST Gumbel Softmax (ours)	1.109	0.320	0.263	0.541	0.205
PSST Multinomial (ours)	**1.119**	**0.322**	**0.264**	**0.544**	**0.206**

Table 1: **Naturalness for a fixed recall rate on COCO**. CIDEr, BLEU4, METEOR, ROUGE, SPICE, for R@5=80%. Compared methods without joint training are not shown in this table because their best recall was much lower: R@5=72 for [31], R@5=66.4 for [30].

Figure 3f illustrates the effect of the sampling-ratio parameter ρ, for *PSST Multinomial* and *PSST Gumbel softmax*. For a fair comparisons, we fixed the CIDEr score at a given value (the maximal value that overlaps all variants), and report the recall@10 on the interpolated discriminability-naturalness curve of Figure 3. For both methods models with ρ = 0 or 1 gave significantly lower

CIDEr=1.2	R@1	R@5	R@10
G-Gan [9] (CIDEr=0.795)	14.3	40.1	55.8
Luo et al. 2018 [31]	20.5	49.1	64.0
Gumbel Temperature Annealing	24.2	56.3	70.9
REINFORCE	31.3	67.0	80.5
ST Gumbel Softmax	31.8	67.3	80.6
ST Multinomial	31.9	67.9	81.8
SR-PL [30] (CIDEr=1.17)	33.0	66.4	80.1
PSST Gumbel Softmax (ours)	37.6	73.0	85.7
PSST Multinomial (ours)	**38.1**	**74.2**	**86.3**

Table 2: **Recall for a fixed CIDEr on COCO**, comparing recall for fixed values CIDEr, as extracted from Figure 3. The metrics are reported on a high CIDEr operating point, showing the strong effect of joint training and the superiority of our approach. For both PSST methods, $\rho = 0.25$ was used. SR-PL [30] used the Karpathy split and CIDEr=1.17. G-Gan [9] used COCO validation split and CIDEr=0.795.

	R@1	R@5	R@10
Separate vs joint training (CIDEr=1.13)			
Luo et al. 2018 [31]	27.7	60.1	74.6
Frozen Speaker (Luo)	32.6	67.8	81.8
Frozen Speaker (MLE)	19.3	46.9	61.3
REINFORCE	38.9	74.0	86.2
PSST Gumbel Softmax (ours)	45.0	78.8	89.4
PSST Multinomial (ours)	**45.3**	**79.4**	**89.9**
With human captions			
PSST Multinomial	21.3	46.9	59.3
Listener trained on GT	25.4	53.9	66.8

Table 3: **Ablation study. Top:** The top section compares the recall of three separate-training baselines, with three joint-training baselines (last three rows). Recall metrics are reported at comparable operating points on the discriminative-vs-natural curves, all at CIDEr=1.13. **Bottom:** Recall values on the validation split, obtained when ablating the speaker network. Feeding human-generated (GT) captions to a listener trained with PSST $\rho = 0.25$ (top) or training a listener with GT captions (bottom).

results then models with ρ between 0.25 to 0.75. This is consistent with the idea that using $\rho < 1$ (some sampling) is necessary for exposing the listener to sparse inputs, so it does not suffer a catastrophic domain shift at test time.

7.1. Ablation study

Comparison with separate training. To quantify the benefit of joint training, we evaluate several separate-training procedure. The top section in Table 3 shows the recall obtained with 2-step training: The speaker model is trained first either (1) as in [31]; (2) training the model of [31] for another 150 epochs; (3) using MLE. It is then kept "frozen" while the listener is trained. We used a lower CIDEr than Table 2, because some baselines did not reach higher CIDEr. PSST is again better for this regime.

Testing listeners with human-generated (GT) captions. The bottom section of Table 3 reports the recall of two models tested with human-generated captions, revealing limitations of the model. First, a listener trained with speaker-generated captions, perform substantially worse when tested with human-generated captions. Even-though human-generated captions are discriminative and people captured the discriminative signals (Table 5 left), the listener fails to use them properly, because it is over-tuned to the speaker-generated ones. Second, training a listener with GT captions performs worse than joint training. Here, the listener fails to learn to capture the discriminative signals in human-generated captions.

7.2. Qualitative results

To get better insight into the captions created by our system, we compare their quality in several ways. First, Figure 4a illustrates the effect of the trade-off parameter λ on the discriminability and naturalness of generated captions.

We then evaluate the benefits of joint training and partial sampling, by comparing PSST Multinomial to ST multinomial (always sampling) and to REINFORCE (separate training). We compare them in two ways. Figure 4b compares caption naturalness at similar recall values and Figure 4c compares discriminability at similar CIDEr values. See details in the caption of Figure 4, and more examples in [41]

7.3. Evaluations on Flickr30

We evaluated PSST and baselines on Flickr30K. This dataset was never used during the development of the method, and evaluations were made after the experiments on COCO were completed. Table 4 lists the naturalness metrics for a fixed recall (90%) on this dataset. PSST is comparable or better than other joint-training approaches.

Recall@5=90%	C	B	M	R	S
REINFORCE	0.431	0.173	0.188	0.435	**0.129**
ST-Gumbel SM	0.484	**0.213**	0.188	**0.455**	0.125
ST Multinomial	0.478	0.212	0.188	0.452	0.124
PSST Gumbel Softmax (ours)	0.485	0.207	0.188	0.447	0.126
PSST Multinomial (ours)	**0.488**	**0.213**	**0.190**	0.448	**0.129**

Table 4: **Evaluation on Flickr30K**. Language quality metrics CIDEr, BLEU4, METEOR, ROUGE, SPICE for R@5=90%. Only methods that reached 90% recall are shown.

7.4. Human evaluations

We evaluated the discriminability and naturalness of various models in a 2-alternative-forced-choice experiment with Amazon Mechanical Turk raters.

Table 5 (left) compares the discriminability of generated captions, for models that share a similar automated naturalness (CIDEr $\approx$ 1.22). Raters were presented with generated caption of the tested model along with a couple of images: the *correct image*, from which the caption was generated, and a second, *distractor image*, that was selected by [31] to be similar to the target. Raters were asked to choose which image is best described by the caption. This task was designed to measure caption discriminative power, regardless of its naturalness, hence we compared models having similar CIDEr but varying recall@10 levels. Results suggest that PSST Multinomial allows raters to detect the correct image slightly better. Table 5(right) compares the naturalness of generated captions, for models that share

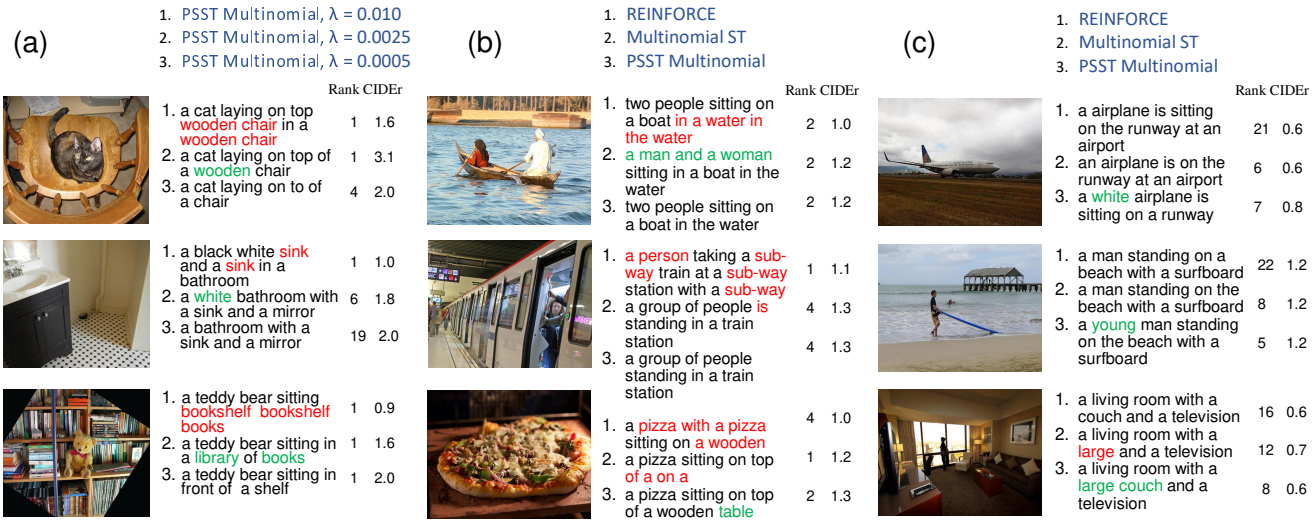

Figure 4: **Qualitative examples**. Red text highlight problematic wording; green text highlights correct grammar with additional discriminative information. Low CIDEr scores are often due to repetitions and missing nouns. (a) **The parameter λ trade-offs discriminability with naturalness**. All captions were created using PSST Multinomial, but with varying values of λ. The top-row caption (high $\lambda = 0.01$), yields more discriminative captions; bottom-row captions (low λ), produce more natural sentences. **(b) The optimization method affects naturalness.** Captions were created using (1) REINFORCE, (2) ST multinomial (3) PSST Multinomial. For each method, λ was chosen to produce a similar recall@10 rate $\approx 80\%$, yielding mean CIDEr scores of 1.017, 1.209, 1.229 respectively. These examples demonstrate that for this fixed recall, models with higher CIDEr tend to produce more natural captions. To provide "typical" images, we selected images with captions whose CIDEr scores was close to the mean CIDEr of each method and at the same time where ranked high. **(c) The optimization method affects discriminability**. For each method, λ was chosen to produce CIDEr ≈ 1.2, yielding an average image retrieval rank of 20, 9, 8 respectively. The examples demonstrate that for a fixed CIDEr score, models with better image retrieval rank tend to produce more discriminative captions. More examples in [41].

a similar automated discriminability (recall@10$\approx$ 80%). Raters found that PSST Multinomial captions were significantly more natural than the reference model and competing model.

8. Conclusion

This paper addresses the problem of building deep models that communicate about their perceived world using human-readable language. We find that training networks jointly while keeping captions similar to human-generated captions, improves both the discriminatibility of captions and their naturalness. Both are further improved using a partial sampling technique, which allows networks to pass more information during training. This optimization approach reduces the variance and bias of gradient estimators, allowing networks to converge to better solutions.

This work can be extended in several natural ways, including having the speaker network and listener network communicate across several rounds (visual dialogues), or introducing communication between multiple agents. We expect our approach to contribute to systems that communicate with people about their perceived environment.

Method	Accuracy	Naturalness
Luo 2018 [31]	72%	-9%
ST Multinomial	69%	reference
PSST Multinomial (ours)	**75%**	**+20%**
Human	85%	-

Table 5: **Human rater evaluation**: **(a) Discriminative power**. Accuracy is computed by asking raters to tell a target image from a distractor image based on a caption. Reported are accuracy of the majority votes among 5 raters over 300 images. **(b) Naturalness** is computed by asking raters to rank two captions for using proper English, with the image they describe. Raters were instructed to pay attention to incoherent singular-plural terms, repeating terms and broken sentences. One caption was generated by ST-multinomial and the second from the evaluated model. All three models were selected to have comparable discriminability, specifically, a recall@10 of $\approx 80\%$.

Acknowledgement: We thank G. Shakhnarovich and Y. Goldberg for insightful discussions. This work was supported by an Israel Science Foundation grant 737/18.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision (ECCV)*, pages 382–398. Springer, 2016.

[2] Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *Empirical Methods in Natural Language (EMNLP)*, 2016.

[3] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *association for computational linguistics workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

[4] Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

[5] Diane Bouchacourt and Marco Baroni. How agents see things: On visual representations in an emergent language game. In *Empirical Methods in Natural Language (EMNLP)*, pages 981–985. Association for Computational Linguistics, 2018.

[6] Lior Bracha and Gal Chechik. Informative object annotations: Tell me something i don't know. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12507–12515, 2019.

[7] Chen Chen, Shuai Mu, Wanpeng Xiao, Zexiong Ye, Liesi Wu, Fuming Ma, and Qi Ju. Improving image captioning with conditional generative adversarial nets. In *Association for the Advancement of Artificial Intelligence (AAAI)*, 2019.

[8] Bo Dai, Sanja Fidler, and Dahua Lin. A neural compositional paradigm for image captioning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2018.

[9] Bo Dai, Sanja Fidler, Raquel Urtasun, and Dahua Lin. Towards diverse and natural image descriptions via a conditional gan. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979, 2017.

[10] Bo Dai and Dahua Lin. Contrastive learning for image captioning. In *Advances in Neural Information Processing Systems*, pages 898–907, 2017.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, Jose M. F. Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[12] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. Learning cooperative visual dialog agents with deep reinforcement learning. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2951–2960, 2017.

[13] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: improving visual-semantic embeddings with hard negatives. In *BMVC*, page 12. BMVA Press, 2018.

[14] Michael C Fu. Gradient estimation. *Handbooks in operations research and management science*, 13:575–616, 2006.

[15] Peter W Glynn. Likelihood ratio gradient estimation for stochastic systems. *Communications of the ACM*, 33(10):75–84, 1990.

[16] Will Grathwohl, Dami Choi, Yuhuai Wu, Geoffrey Roeder, and David Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[17] Shixiang Gu, Sergey Levine, Ilya Sutskever, and Andriy Mnih. Muprop: Unbiased backpropagation for stochastic neural networks. In *International Conference on Learning Representations (ICLR)*, 2016.

[18] Emil Julius Gumbel. Statistical theory of extreme values and some practical applications. *NBS Applied Mathematics Series*, 33, 1954.

[19] Hexiang Hu, Ishan Misra, and Laurens van der Maaten. Binary Image Selection (BISON): Interpretable Evaluation of Visual Grounding. *arXiv preprint arXiv:1901.06595*, 2019.

[20] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparametrization with gumbel-softmax. In *International Conference on Learning Representations (ICLR)*, Apr. 2017.

[21] Harsh Jhamtani and Taylor Berg-Kirkpatrick. Learning to describe differences between pairs of similar images. In *Empirical Methods in Natural Language (EMNLP)*, pages 4024–4034. Association for Computational Linguistics, 2018.

[22] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.

[23] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations (ICLR)*, 2014.

[24] Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. Natural language does not emerge 'naturally' in multi-agent dialog. In *Empirical Methods in Natural Language (EMNLP)*, pages 2962–2967. Association for Computational Linguistics, 2017.

[25] Angeliki Lazaridou, Alexander Peysakhovich, and Marco Baroni. Multi-agent cooperation and the emergence of (natural) language. In *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2017.

[26] Jason Lee, Kyunghyun Cho, Jason Weston, and Douwe Kiela. Emergent translation in multi-agent communication. In *International Conference on Learning Representations (ICLR)*, 2018.

[27] Dianqi Li, Qiuyuan Huang, Xiaodong He, Lei Zhang, and Ming-Ting Sun. Generating diverse and accurate visual captions by comparative adversarial learning. *arXiv preprint arXiv:1804.00861*, 2018.

[28] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. *Text Summarization Branches Out*, 2004.

[29] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision (ECCV)*, pages 740–755. Springer, 2014.

[30] Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *European Conference on Computer Vision (ECCV)*, September 2018.

[31] Ruotian Luo, Brian Price, Scott Cohen, and Gregory Shakhnarovich. Discriminability objective for training descriptive captions. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6964–6974, 2018.

[32] Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The Concrete Distribution: A Continuous Relaxation of Discrete Random Variables. In *International Conference on Learning Representations (ICLR)*, 2017.

[33] Andriy Mnih and Danilo Jimenez Rezende. Variational inference for monte carlo objectives. In *International Conference on Machine Learning (ICML)*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 2188–2196. JMLR.org, 2016.

[34] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics, 2002.

[35] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning (ICML)*, volume 32 of *JMLR Workshop and Conference Proceedings*, pages 1278–1286. JMLR.org, 2014.

[36] John Schulman, Nicolas Heess, Théophane Weber, and Pieter Abbeel. Gradient estimation using stochastic computation graphs. In *Advances in Neural Information Processing Systems*, 2015.

[37] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *IEEE International Conference on Computer Vision (ICCV)*, pages 4135–4144, 2017.

[38] George Tucker, Andriy Mnih, Chris J Maddison, John Lawson, and Jascha Sohl-Dickstein. Rebar: Low-variance, unbiased gradient estimates for discrete latent variable models. In *Advances in Neural Information Processing Systems*, pages 2627–2636, 2017.

[39] Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1070–1079. IEEE, 2017.

[40] Ramakrishna Vedantam, Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575, 2015.

[41] Gilad Vered, Gal Oren, Yuval Atzmon, and Gal Chechik. Cooperative image captioning. *arXiv preprint arXiv:1907.11565*, 2019.

[42] Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *arXiv preprint arXiv:1610.02424*, 2016.

[43] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4195–4203, 2019.

[44] Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

[45] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

[46] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.