# Context-aware Captions from Context-agnostic Supervision

Ramakrishna Vedantam[1]     Samy Bengio[2]     Kevin Murphy[2]     Devi Parikh[3]     Gal Chechik[2]

[1]Virginia Tech     [3]Georgia Institute of Technology     [2]Google

[1]vrama91@vt.edu    [3]parikh@gatech.edu    [2]{bengio,kpmurphy,gal}@google.com

## Abstract

*We introduce an inference technique to produce discriminative context-aware image captions (captions that describe differences between images or visual concepts) using only generic context-agnostic training data (captions that describe a concept or an image in isolation). For example, given images and captions of "siamese cat" and "tiger cat", we generate language that describes the "siamese cat" in a way that distinguishes it from "tiger cat". Our key novelty is that we show how to do joint inference over a language model that is context-agnostic and a listener which distinguishes closely-related concepts. We first apply our technique to a justification task, namely to describe why an image contains a particular fine-grained category as opposed to another closely-related category of the CUB-200-2011 dataset. We then study discriminative image captioning to generate language that uniquely refers to one of two semantically-similar images in the COCO dataset. Evaluations with discriminative ground truth for justification and human studies for discriminative image captioning reveal that our approach outperforms baseline generative and speaker-listener approaches for discrimination.*

## 1. Introduction

Language is the primary modality for communicating, and representing knowledge. To convey relevant information, we often use language in a way that takes into account context. For example, instead of describing a situation in a "literal" way, one might pragmatically emphasize selected aspects in order to be persuasive, impactful or effective. Consider the target image at the bottom left in Fig. 1. A literal description "An airplane is flying in the sky" conveys the semantics of the image, but would be inadequate if the goal was to disambiguate this image from the distractor image (bottom right). For this purpose, a more pragmatic description would be, "A large passenger jet flying through a blue sky". This description is aware of context, namely, that the distractor image also has an airplane. People use such pragmatic considerations continuously, and effortlessly in
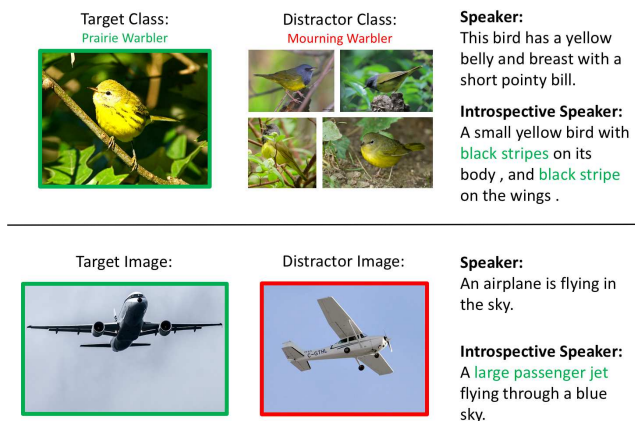


Figure 1: An illustration of two tasks requiring pragmatic reasoning explored in this paper. 1) *justification*: Given an image of a bird, a target (ground-truth) class (green), and a distractor class (red), describe the target image to explain why it belongs to the target class, and not the distractor class. The distractor class images are only shown for illustration, and are not provided to the algorithm. 2) *discriminative image captioning*: Given two similar images, produce a sentence to identify a target image (green) from the distractor image (red). Our introspective speaker model improves over a context-free speaker.

teaching, conversation and discussions.

In this vein, it is desirable to endow machines with pragmatic reasoning. One approach would be to collect training data of language used in context, for example, discriminative ground truth utterances from people describing images in context of other images, or justifications explaining why an image contains a target class as opposed to a distractor class (Fig. 1). Unfortunately, collecting such data has a prohibitive cost, since the space of objects in possible contexts is often too large. Furthermore, in some cases the context in which we wish to be pragmatic may be unknown *apriori*. For example, a free-form conversation agent may have to respond in a context-aware or discriminative fashion depending upon the history of a conversation. Such scenarios also arise in human-robot interaction, as in the case where, a robot may need to reason about which spoon a person is asking for. Thus, in this paper, we focus on deriving pragmatic (context-aware) behavior given access only to generic (context-agnostic) ground truth.

We study two qualitatively different real-world vision tasks that require pragmatic reasoning. The first is *justification*, where the model needs to justify why an image corresponds to one fine-grained object category, as opposed to a closely related, yet undepicted category. Justification is a task that is important for hobbyists, and domain experts: ornithologists and botanists often need to explain why an image depicts particular species as opposed to a closely-related species. Another potential application for justification is in machine teaching, where an algorithm instructs non-expert humans about new concepts.

Our second task is *discriminative image captioning*, where the goal is to generate a sentence that describes an image in context of other semantically similar images. This task is not only grounded in pragmatics, but is also interesting as a scene understanding task to check fine-grained image understanding. It also has potential applications to human robot interaction.

Recent work by Andreas and Klein [1] derives pragmatic behaviour in neural language models using only context-free data. While we are motivated by similar considerations, the key algorithmic novelty of our work over [1] is a unified inference procedure which leads to more efficient search for discriminative sentences (Sec. 5). Our approach is based on the realization that one may simply re-use the sampling distribution from the generative model, instead of training a separate model to assess discriminativeness [1]. This also has important implications for practitioners, since one can easily adapt existing context-free captioning models for context-aware captioning without additional training. Furthermore, while [1] was applied to an abstract scenes dataset [43], we apply our model to two qualitatively different real-image datasets: the fine-grained birds dataset CUB-200-2011 [38], and the COCO [21] dataset which contains real-life scenes with common objects.

In summary, the key contributions of this paper are:

- A novel inference procedure that models an introspective speaker (IS), allowing a speaker (S) (say a generic image captioning model) to reason about pragmatic behavior without additional training.
- Two new tasks for studying discriminative behaviour, and pragmatics, grounded in vision: *justification*, and *discriminative image captioning*.
- A new dataset (CUB-Justify) to evaluate justification systems on fine-grained bird images with 5 captions for 3161 (image, target class, distractor class) triplets.
- Our evaluations on CUB-Justify, and human evaluation on COCO show that our approach outperforms baseline approaches at inducing discrimination.

## 2. Related Work

**Pragmatics:** The study of pragmatics – how context influences usage of language, stems from the foundational work of Grice [13] who analyzed how cooperative multi-agent linguistic agents could model each others' behavior to achieve a common objective. Consequently, a lot of pragmatics literature has studied higher-level behavior in agents including conversational implicature [5] and the Gricean maxims [37]. These works aim to derive pragmatic behavior given minimal assumptions on individual agents and typically use hand-tuned lexicons and rules. More recently, there have been exciting developments on applying reinforcement learning (RL) techniques to these problems [25, 7, 19], requiring less manual tuning.

We are also interested in deriving pragmatic behavior, but our focus is on scaling context-sensitive behavior to vision tasks. Other works model ideas from pragmatics to learn language via games played online [39] or for human-robot collaboration [32]. In a similar spirit, here we are interested in applying ideas from pragmatics to build systems that can provide justifications (Sec. 4.1) and provide discriminative image captions (Sec. 4.2).

Most relevant to our work is the recent work on deriving pragmatic behavior in abstract scenes made with clipart, by Andreas, and Klein [1]. Unlike their technique, our proposed approach does not require training a second listener model and supports more efficient inference (Sec. 3.3). More details are provided in Sec. 3.1.

**Beyond Image Captioning:** Image captioning, the task of generating natural language description for an image, has seen quick progress [10, 11, 36, 40]. Recently, research has shifted beyond image captioning, addressing tasks like visual question answering [2, 12, 23, 42], referring expression generation [18, 24, 26, 30], and fill-in-the-blanks [41]. In a similar spirit, the two tasks we introduce here, *justification*, and *discriminative image captioning*, can be viewed as "beyond image captioning" tasks. Sadovnik *et al.* [29] first studied a discriminative image description task, with the goal of distinguishing one image from a set of images. Their approach incorporates cues such as discriminability and saliency, and uses hand-designed rules for constructing sentences. In contrast, we develop inference techniques to induce discriminative behavior in neural models. The reference game from [1] can also be seen as a discriminative image captioning task on abstract scenes made from clipart, while we are interested in the domain of real images. The work on generating referring expressions by Mao *et al.* [24] generates discriminative captions which refer to particular objects in an image given context-aware supervision. Our work is different in the sense that we address an instance of pragmatic reasoning in the common case where context-dependent data is not available for training.

**Rationales:** Several works have studied how machines can understand human rationales, including enriching classification by asking explanations from humans [9], and incorporating human rationales in active learning [6, 27]. In

contrast, we focus on machines providing justifications to humans. This could potentially allow machines to teach new concepts to humans (machine teaching). Other recent work [14] looks at post-hoc explanations for classification decisions. Instead of explaining why a model thinks an image is a particular class, [14] describes why an image is of a class predicted by the classifier. Unlike this task, our justification task requires reasoning about explicit context from the distractor class. Further, we are not interested in providing rationalizations for classification decisions but in explaining the differences between confusing concepts to humans. We show a comparison to [14] in [33], demonstrating the importance of context for justification.

**Beam Search with Modified Objectives:** Beam search is an approximate, greedy technique for inference in sequential models. We perform beam search on a modified objective for our introspective speaker model to induce discrimination. This is similar in spirit to recent works on inducing diversity in beam search [35], and maximum mutual information inference for sequence-to-sequence models [20].

## 3. Approach

We describe our approach for inducing context-aware language for: 1) *justification*, where the context is another class, and 2) *discriminative image captioning*, where the context is a semantically similar image. For clarity, we first describe the formulation for justification, and then discuss a modification for discriminative image captioning.

In the justification task (Fig. 1 top), we wish to produce a sentence $s$, comprised of a sequence of words $\{s_i\}$, based on a given image $I$ of a target concept $c_t$ in the context of a distractor concept $c_d$. The produced justification should capture aspects of the image that discriminate between the target, and the distractor concepts. Note that images of the distractor class are not provided to the algorithm.

We first train a generic context-agnostic image captioning model (from here on referred to as speaker) using training data from Reed *et al*. [28] who collected captions describing bird images on the CUB-200-2011 [38] dataset. We condition the model on $c_t$ in addition to the image. That is, we model $p(s|I, c_t)$. This not only helps produce better sentences (providing the model access to more information), but is also the cornerstone of our approach for discrimination (Sec. 3.2). Our language models are recurrent neural networks which represent the state-of-the-art for language modeling across a range of popular tasks like image captioning [36, 40], machine translation [3] *etc*.

### 3.1. Reasoning Speaker

To induce discrimination in the utterances from a language model, it is natural to consider using a generator, or speaker, which models $p(s|I, c_t)$ in conjunction with a listener function $f(s, c_t, c_d)$ that scores how discriminative an

utterance $s$ is. The task of a pragmatic reasoning speaker $RS$, then, is to select utterances which are good sentences as per the generative model $p$, and are discriminative per $f$:

$$RS(I, c_t, c_d) = \arg\max_s \lambda p(s|I, c_t) + (1-\lambda)f(s, c_t, c_d)$$
$$(1)$$

where $0 \leq \lambda \leq 1$ controls the tradeoff between linguistic adequacy of the sentence, and discriminativeness.

A similar reasoning speaker model forms the core of the approach of [1], where $p$, and $f$ are implemented using multi-layer perceptrons (MLPs). As noted in [1], selecting utterances from such a reasoning speaker poses several challenges. First, exact inference in this model over the exponentially large space of sentences is intractable. Second, in general one would not expect the discriminator function $f$ to factorize across words, making joint optimization of the reasoning speaker objective difficult. Thus, Andreas, and Klein [1] adopt a sampling based strategy, where $p$ is considered as the proposal distribution whose samples are ranked by a linear combination of $p$, and $f$ (Eq. 1). Importantly, this distribution is over full sentences, hence the effectiveness of this formulation depends heavily on the distribution captured by $p$, since the search over the space of all strings is solely based on the speaker. This is inefficient, especially when there is a mismatch in the statistics of the context-free (generative), and the unknown context-aware (discriminative) sentence distributions. In such cases, one must resort to drawing many samples to find good discriminative sentences.

### 3.2. Introspective Speaker

Our approach for incorporating contextual behavior is based on a simple modification to the listener $f$ (Eq. 1). Given the generator $p$, we construct a listener module that wants to discriminate between $c_t$, and $c_d$, using the following log-likelihood ratio:

$$f(s, c_t, c_d) = \log \frac{p(s|c_t, I)}{p(s|c_d, I)}. \qquad (2)$$

This listener only depends on a generative model, $p(s|c, I)$, for the two classes $c_t$, and $c_d$. We name it "introspector" to emphasize that this step re-uses the generative model, and does not need to train an explicit listener model. Substituting the introspector into Eq. 1 induces the following introspective speaker model for discrimination:

$$\underbrace{\Delta(I, c_t, c_d)}_{\text{introspective speaker}} = \arg\max_s \lambda \underbrace{\log p(s|c_t, I)}_{\text{speaker}}$$
$$+ (1-\lambda) \underbrace{\log \frac{p(s|c_t, I)}{p(s|c_d, I)}}_{\text{introspector}} , \quad (3)$$

with $\lambda$ that trades-off the weight given to generation, and introspection (similar to Eq. 1). In general, we expect this

approach to provide sensible results when $c_t$, and $c_d$ are similar. That is, we expect humans to describe similar concepts in similar ways, hence $p(s|c_t, I)$ should not be too different from $p(s|c_d, I)$. Thus, the introspector is less likely to overpower the speaker in Eq. 3 in such cases (for a given $\lambda$). Note that for sufficiently different concepts the speaker alone is likely to be sufficient for discrimination. That is, describing the concept in isolation is likely to be enough to discriminate against a different or unrelated concept.

A careful inspection of the introspective speaker model reveals two desirable properties over previous work [1]. First, the introspector model does not need training, since it only depends on $p$, the original generative model. Thus, existing language models can be readily re-used to produce context-aware outputs by conditioning on $c_d$. We demonstrate empirical validation of this in Sec. 5. This would help scale this approach to scenarios where it is not known apriori which concepts need to be discriminated, in contrast to approaches which train a separate listener module. Second, it leads to a unified, and efficient inference for the introspective speaker (Eq. 3), which we describe next.

### 3.3. Emitter-Suppressor (ES) Beam Search for RNNs

We now describe a search algorithm for implementing the maximization in Eq. 3, which we call *emitter-suppressor* (ES) beam search. We use the beam search [22] algorithm, which is a heuristic graph-search algorithm commonly used for inference in Recurrent Neural Networks [15, 35].

We first factorize the posterior log-probability terms in the introspective speaker equation (Eq. 3) $p(s|c_t, I) = \prod_{\tau=1}^{T} p(s_\tau|s_{1:\tau-1}, c_t, I)$, denoting $s_{1:T} = \{s_\tau\}_{\tau=1}^{T}$ ($s_{1:0}$ corresponds to a null string). $T$ is the length of the sentence. We then combine terms from Eq. 3, yielding the following emitter-suppressor objective for the introspective speaker:

$$\Delta(I, c_t, c_d) = \arg\max_s \sum_{\tau=1}^{T} \log \frac{\overbrace{p(s_\tau|s_{1:\tau-1}, c_t, I)}^{\text{emitter}}}{\underbrace{p(s_\tau|s_{1:\tau-1}, c_d, I)^{1-\lambda}}_{\text{suppressor}}}.$$

(4)

The emitter (numerator in Eq. 4) is the generative model conditioned on the target concept $c_t$, deciding which token to select at a given timestep. The suppressor (the denominator in Eq. 4) is conditioned on the distractor concept $c_d$, providing signals to the emitter on which tokens to avoid. This is intuitive – to be discriminative, we want to emit words that match $c_t$, but avoid emitting words that match $c_d$.

We maximize the emitter-suppressor objective (Eq. 4) using beam search. Vanilla beam search, as typically used in language models, prunes the output space at every timestep keeping the top-B (usually incomplete) sentences with highest log-probabilities so far (speaker in Eq. 3). Instead,
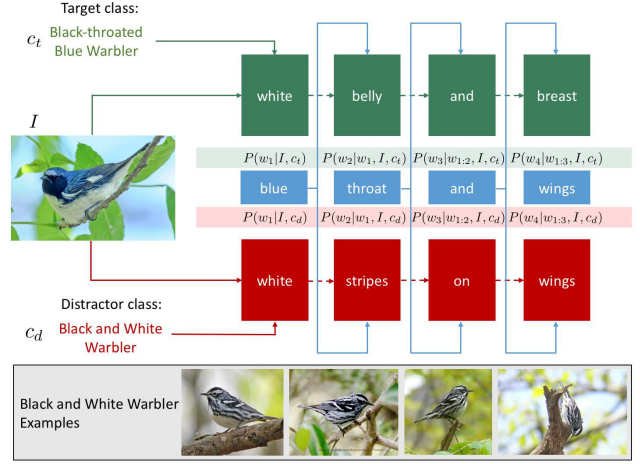


Figure 2: Emitter-suppressor beam search for beam size 1, for distinguishing an image of "black-throated blue warbler" from the distractor class "black and white warbler". Green: A language model $p(s|c_t, I)$ produces a caption "white belly and breast ... ". Red: When feeding the distractor class to the language model, since the two birds share the attribute white belly, which appears in the image, the term "white" is highly suppressed. Blue: Picking likely words for the emitter, and unlikely for the suppressor yields a discriminative caption "blue throat ..". Note that emitter, and suppressor share history (the previouly generated words).

we run beam search to keep the top-B sentences with highest ES ratio in Eq. 4. Fig. 2 illustrates this ES beam search for a beam size of 1.

It is important to consider how the trade-off parameter $\lambda$ affects the produced sentences. For $\lambda = 1$, the model generates descriptions that ignore the context. At the other extreme, low $\lambda$ values are likely to make the produced sentences very different from any sentence in the training set (repeated words, ungrammatical sentences). It is not trivial to assume that there exists a wide enough range of $\lambda$ creating sentences that are both discriminative, and well-formed. However, our results (Sec. 5) indicate that such a range of $\lambda$ exists in practice.

### 3.4. Discriminative Image Captioning

We are given a target image $I_t$, and a distractor $I_d$, that we wish to distinguish, similar to the two classes for the justification task. We construct a speaker (or generator) for this task by training a standard image captioning model. Given this speaker, we construct an emitter-suppressor equation (as in Eq. 4):

$$\Delta(I_t, I_d) = \arg\max_s \sum_{\tau=1}^{T} \log \frac{\overbrace{p(s_\tau|s_{1:\tau-1}, I_t)}^{\text{emitter}}}{\underbrace{p(s_\tau|s_{1:\tau-1}, I_d)^{1-\lambda}}_{\text{suppressor}}}. \quad (5)$$

We re-use the mechanics of emitter-suppressor beam search from Sec. 3.3, conditioning the emitter on the target image $I_t$, and the suppressor on the distractor image $I_d$.

## 4. Experimental Setup

We provide details of the CUB dataset, of our CUB-Justify dataset used for evaluation, and of the speaker-training setup for the justification task. We then discuss the experimental protocols for discriminative image captioning.

### 4.1. Justification

**CUB Dataset:** The Caltech UCSD birds (CUB) dataset [38] contains 11788 images for 200 species of North American birds. Each image in the dataset has been annotated with 5 fine-grained captions by Reed *et al.* [28]. These captions mention various details about the bird ("This is a white spotted bird with a long pointed black beak.") while not mentioning the name of the bird species.

**CUB-Justify Dataset:** We collect a new dataset (CUB-Justify) with ground truth justifications for evaluating justification. We first sample the target, and distractor classes from within a hyper-category created based on the last name of the folk names of the 200 species in CUB. For instance, "rufous hummingbird", and "ruby throated hummingbird" both fall in the hyper-category "hummingbird". We induce 37 such hyper-categories. The largest single hyper-category is "Warbler" with 25 categories. We then select a subset of (approx.) 15 images from the test set of CUB-200-2011 [38] for each of the 200 classes, to form a CUB-Justify test split. We use the rest for speaker training (CUB-Justify train split).

Workers were then shown an image of the "rufous hummingbird", for instance, and a set of 6 other images (from CUB-Justify test split) all belonging to the distractor class "ruby throated hummingbird", to form the visual notion of the distractor class. They were also shown a diagram of the morphology of birds indicating various parts such as tarsus, rump, wingbars *etc.* (similar to Reed *et al.* [28]). The instruction was to describe the target image such that it is not confused with images from the distractor class. Some birds are best distinguished by non-visual cues such as their call, or their migration patterns. Thus, we drop the categories of birds from the original list of triplets which were labeled as too hard to distinguish by the workers. At the end of this process we are left with 3161 triplets with 5 captions each. We split this dataset into 1070 validation (for selecting the best value of $\lambda$), and 2091 test examples respectively. More details on the interface can be found in [33].

**Speaker Training:** We implement a model similar to "Show, Attend, and Tell" from Xu *et al.* [40], modifying the original model to provide the class as input, similar in spirit to [14]. Exact details of our model architecture are given in [33]. We train the model on the CUB-Justify train split. Recall that this just has context-agnostic captions from [28].

To evaluate the quality of our speaker model, we report numbers here using the CIDEr-D metric [34] commonly used for image captioning [14, 17, 36] computed on the

context-agnostic captions from [28]. Our captioning model with both the image, and class as input reaches a validation score of 50.2 CIDEr-D, while the original image-only captioning model reaches a CIDEr-D of 49.1. The scores are in a similar range as existing CUB captioning approaches [14].

**Justification Evaluation:** We measure performance of the (context-aware) justification captions on the CUB-Justify discriminative captions using the CIDEr-D metric. CIDEr-D weighs n-grams by their inverse document frequencies (IDF), giving higher weights to sentences having "content" n-grams ("red beak") than generic n-grams ("this bird") [14]. Further, CIDEr-D captures importance of an n-gram for the image. For instance, it emphasizes "red beak" over, say, "black belly" if "red beak" is used more often in human justifications. We also report METEOR [4] scores for completeness. More detailed discussion on metrics can be found in [33].

### 4.2. Discriminative Image Captioning

**Dataset:** We want to test if reasoning about context with an introspective speaker can help discriminate between pairs of very similar images from the COCO dataset. To construct a set of confusing image pairs, we follow two strategies. First, *easy confusion*: For each image in the validation (test) set, we find its nearest neighbor in the FC7 space of a pre-trained VGG-16 CNN [31], and repeat this process of neighbor finding for 1000 randomly chosen source images. Second, *hard confusion*: To further narrow down to a list of semantically similar confusing images, we then run the speaker model on the nearest neighbor images, and compute word-level overlap (intersection over union) of their generated sentences. We then pick the top 1000 pairs with most overlap. Interestingly, the top 539 pairs had identical captions. This reflects the issue of the output of image captioning models lacking diversity, and seeming templated [8, 36].

**Speaker Training and Evaluation:** We train our generative speaker for use in emitter-suppressor beam search using the model from [36] implemented in the neuraltalk2 project [16]. We use the train/val/test splits from [17]. Our trained and finetuned speaker model achieves a performance of 91 CIDEr-D on the test set. As seen in Eq. 5, no category information is used for this task. We evaluate approaches for discriminative image captioning based on how often they help humans to select the correct image out of the pair of images.

## 5. Results

### 5.1. Justification

**Methods and Baselines:** We evaluate the following models: **1.** IS($\lambda$): Introspective speaker from Eq. 3; **2.** IS(1): standard literal speaker, which generates a caption conditioned on the image and target class, but which ignores the
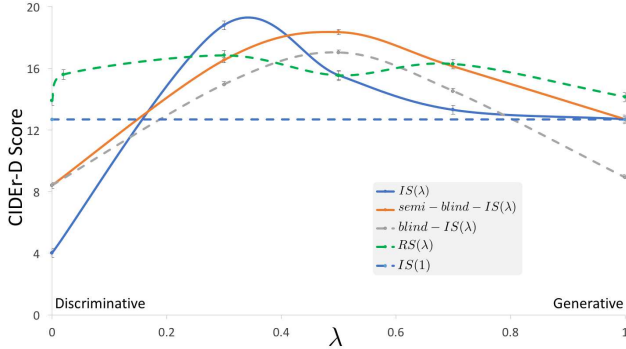
Figure 3: **CUB-Justify validation results:** CIDEr-D vs. $\lambda$ on CUB-Justify validation. Our introspective speaker approaches (IS($\lambda$) and semi-blind-IS($\lambda$)) models perform best, followed by the class-only introspective speaker (blind-IS($\lambda$)). semi-blind-IS($\lambda$) outperforms other methods for a wider range of $\lambda$. All approaches which reason about pragmatics beat the baseline generative approach IS(1). Error bars denote standard error of the mean score estimated across the validation set.

distractor class; **3.** semi-blind-IS($\lambda$): Introspective speaker in which the listener does not have access to the image, but the speaker does; **4.** blind-IS($\lambda$): Introspective speaker without access to image, conditioned only on classes; **5.** RS($\lambda$): Our implementation of Andreas and Klein [1], but using our (more powerful) language model, and Eq. 3 with a listener that models $\frac{p(s|c_t)}{p(s|c_d)}$ (similar to semi-blind-IS($\lambda$)) for ranking samples (as opposed to a trained MLP [1], to keep things comparable). All approaches use 10 beams/samples (which is better than lower values) unless stated otherwise.

**Validation Performance:** Fig. 3 shows the performance on CUB-Justify validation set as a function of $\lambda$, the hyperparameter controlling the tradeoff between the speaker and the introspector (Eq. 3). For the RS($\lambda$) baseline, $\lambda$ stands for the tradeoff between the log-probability of the sentence and the score from the discriminator function for sample re-ranking. A few interesting observations emerge. First, both our IS($\lambda$) and semi-blind-IS($\lambda$) models outperform the baselines for the mid range of $\lambda$ values. IS($\lambda$) model does better overall, but semi-blind-IS($\lambda$) has a more stable performance over a wider range of $\lambda$. This indicates that when conditioned on the image, the introspector has to be highly discriminative (low lambda values) to overcome the signals from the image, since discrimination is between classes.

Second, as $\lambda$ is decreased from 1, most methods improve as the sentences become more discriminative, but then get worse again as $\lambda$ becomes too low. This is likely to happen because when $\lambda$ is too low, the model explores rare tokens and parts of the output space that have not been seen during training, leading to badly-formed sentences (Fig. 4). This effect is stronger for IS($\lambda$) models than for RS($\lambda$), since RS($\lambda$) searches the output space over samples from the generator and only ranks using the joint reasoning speaker objective (Eq. 1). Interestingly, at $\lambda = 1$ (no discrimination), the RS($\lambda$) approach, which samples from the generator, also performs better than other approaches, which use beam

| Approach | CIDEr-D | METEOR |
|---|---|---|
| IS($\lambda$) | $18.4 \pm 0.2$ | 26.5 |
| semi-blind-IS($\lambda$) | $\mathbf{18.5 \pm 0.2}$ | **27.5** |
| RS($\lambda$) | $15.8 \pm 0.2$ | 26.5 |
| IS(1) | $12.3 \pm 0.1$ | 25.3 |
| blind-IS($\lambda$) | $16.1 \pm 0.2$ | 26.8 |

Table 1: **CUB-Justify test results:** CIDEr-D, and METEOR scores (higher the better) computed on test set of CUB-Justify. Each model used the best $\lambda$ selected on the validation set (Fig. 3). Error values are standard error of the mean (SEM is less than 0.05 for METEOR). semi-blind-IS($\lambda$) outperforms other methods.

search to select high log-probability (context-agnostic) sentences. This indicates that in the absence of ground truth justifications, there is indeed a discrepancy between searching for discriminativeness and searching for a highly likely context-agnostic sentence.

We perform more comparisons with the RS($\lambda$) baseline, sweeping over $\{10, 50, 100\}$ samples from the generator for listener reranking (Eq. 1). We find that using 100 samples, RS($\lambda$) gets comparable CIDEr-D scores (18.8) (but lower METEOR scores) than our semi-blind-IS($\lambda$) approach with a beam size of 10. This suggests that our semi-blind-IS($\lambda$) approach is more computationally efficient at exploring the output space because our emitter-suppressor beam search allows us to do joint greedy inference over speaker and introspector, leading to more meaningful local decisions. For completeness, we also trained a listener module discriminatively, and used it as a ranker for RS($\lambda$). We found that this gets to $16.2 \pm 0.3$ CIDEr-D (at $\lambda = 0.5$) on validation, which is lower than IS($\lambda$), showing that the bottleneck for performance is sampling, rather than the discriminativeness of the listener. More details can be found in [33].

**Test Performance:** Table. 1 details the performance of the above models on the test set of CUB-Justify, with each model using its best-performing $\lambda$ on the validation set (Fig. 3). Both introspective-speaker models strongly outperform the baselines, with semi-blind-IS($\lambda$) slightly outperforming the IS($\lambda$) model. This could be due to the performance of semi-blind-IS($\lambda$) being less sensitive to the exact choice of $\lambda$ (from Fig. 3). Among the baselines, the best performing method is the blind-IS($\lambda$) model, presumably because this model does emitter-suppressor beam search, while the other two baseline approaches rely on sampling and regular beam search respectively.

**Qualitative Results:** We next showcase some qualitative results that demonstrate 1) aspects of pragmatics, and 2) context dependence captured by our best-performing semi-blind-IS($\lambda$) model. Fig. 4 demonstrates how sentences uttered by the introspective speaker change with $\lambda$. At $\lambda = 1$ the sentence describes the image well, but is oblivious of the context (distractor class). The sentence "A small sized bird has a very long and pointed bill." is discriminative of hummingbirds against other birds, but not among hummingbirds (many of which tend to have long beaks/bills). At $\lambda = 0.7$,
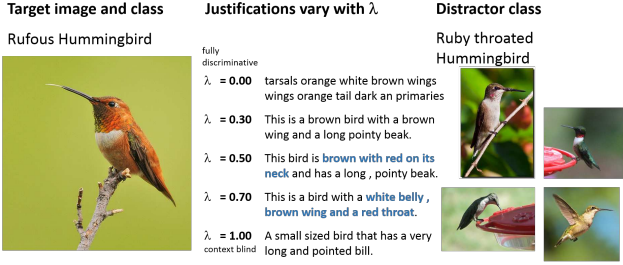
Figure 4: **The effect of context weight**: An image of a "Rufous Hummingbird" in the context of another hummingbird type. A generative (context-blind) description describes the bird as having a long beak, but this feature is not discriminative. When taking into account the context, intermediate $\lambda$ values yield descriptions that highlight that the Rufous is brown with a red throat. For $\lambda = 0$, the model does not force sentences to be well formed.
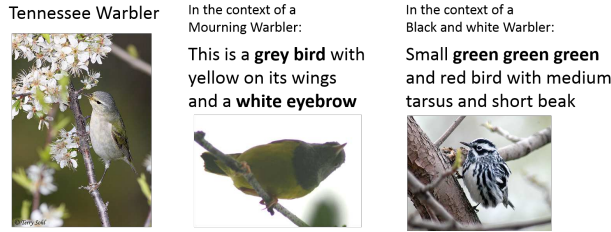


Figure 5: **The effect of context class**: An image of a "Tennessee Warbler", which has light green wings, and a white eyebrow. When described in the context of a mourning warbler, which has a green hue, the description highlights that the target bird has a white eyebrow. When described in the context of the "Black and White Warbler", the description highlights that the target bird has green color.

and $\lambda = 0.5$, the model captures discriminative features such as the "red neck", "white belly", and "red throat". Interestingly, at $\lambda = 0.7$ the model avoids saying "long beak", a feature shared by both birds. Next, Fig. 5 demonstrates how the selected utterances change based on the context. A limitation of our approach is that, since the model never sees discriminative training data, in some cases it produces repeated words ("green green green") when encouraged to be discriminative at inference time.

Finally, Fig. 6 illustrates the importance of visual reasoning for the justification task. Fine-grained species often have large intra-class variances which a *blind* approach to justification would ignore. Thus, a good justification approach needs to be grounded in the image signal to pick the discriminative cues appropriate for the given instance.

## 5.2. Discriminative Image Captioning

As explained in Sec. 4.2 we create two sets of semantically similar target, and distractor images: *easy confusion* based on FC7 features alone, and *hard confusion* based on both FC7, and sentences generated from the speaker (image captioning model). We are interested in understanding if emitter-suppressor inference helps identify the target image better than the generative speaker baseline. Thus the two approaches are speaker (S) (baseline), and introspective speaker (IS) (our approach). We use $\lambda = 0.3$ based on
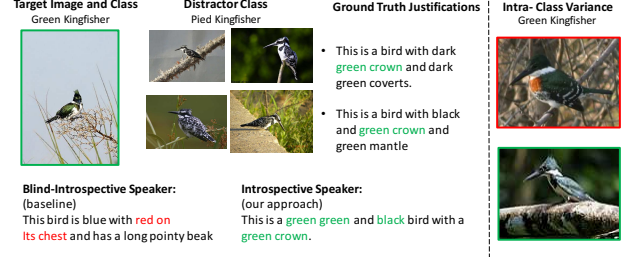


Figure 6: **The importance of visual signal for justification in fine-grained categories**. Given the image of a green kingfisher (left), a blind-IS($\lambda$) model says the bird has "red on its chest", which is inaccurate for this image, and a "long pointy beak", which is not a discriminative feature for this context. At the same time, the semi-blind-IS($\lambda$) model mentions the "green crown", and avoids uttering "red chest". Given the complicated intra-category invariances in bird categories (right), it is intuitive that the image signal is important for justification.

| Approach | *easy confusion* (%) | *hard confusion* (%) |
|---|---|---|
| S (baseline) | 74.6 | 52.5 |
| IS (ours) | **89.0** | **74.1** |

Table 2: % of image pairs that are correctly discriminated by humans, based on descriptions in COCO. Introspective speaker (IS) is better at pointing to the target image given a confusing distractor image across both easy, and hard data splits than a speaker (S). Standard error is below the precision we report numbers at.

our results on the CUB dataset. We run all approaches at a beam size of 2 (typically best for COCO [16]).

**Human Studies:** We setup a two annotation forced choice (2AFC) study where we show a caption to raters asking them to "pick an image that the sentence is more likely to be describing.". Each target distractor image pair is tested against the generated captions. We check the fraction of times a method caused the target image to be picked by a human. A discriminative image captioning method is considered better if it enables humans to identify the target image more often. Results of the study are summarized in Table. 2. We find that our approach outperforms the baseline speaker (S) on the *easy confusion* as well as the *hard confusion* splits. However, the gains from our approach are larger on the *hard confusion* split, which is intuitive.

**Qualitative Results:** The qualitative results from our COCO experiments are shown in Fig. 7. The target image, when successfully identified, is shown with a green border. We show examples where our model identifies the target image better in the first two rows, and some failure cases in the third row. Notice how the model is able to modify its utterances to account for context, and pragmatics, when going from $\lambda = 1$ (speaker) to $\lambda = 0.3$ (introspective speaker). Note that the sentences typically respect grammatical constructs despite being forced to be discriminative.

## 6. Discussion

Describing absence of concepts and inducing comparative language are exciting directions for future work on jus-

**S: Speaker, a generic image captioning model.**   **IS: Introspective Speaker, with our discriminative inference.**



S: A man and a woman playing a video game
IS: A man is sitting on a couch with a remote

S: A baseball player swinging a bat at a ball
IS: Kids playing baseball in a field with a fence

S: A giraffe standing next to a tree in a forest
IS: A giraffe standing next to a wooden fence

S: A herd of sheep standing on top of a grass covered field.
IS: A herd of sheep grazing in a lush green field

S: A train traveling down tracks next to a forest
IS: A red train is on the tracks in the woods

S: A man riding a motorcycle on the street.
IS: Black and white photograph of people crossing the street

S: A street sign on a pole on a city street
IS: A police officer directing traffic at a bus stop

S: A bus is parked on the side of the road
IS: A blue bus parked in front of a bus

S: A bride and groom cutting their wedding cake
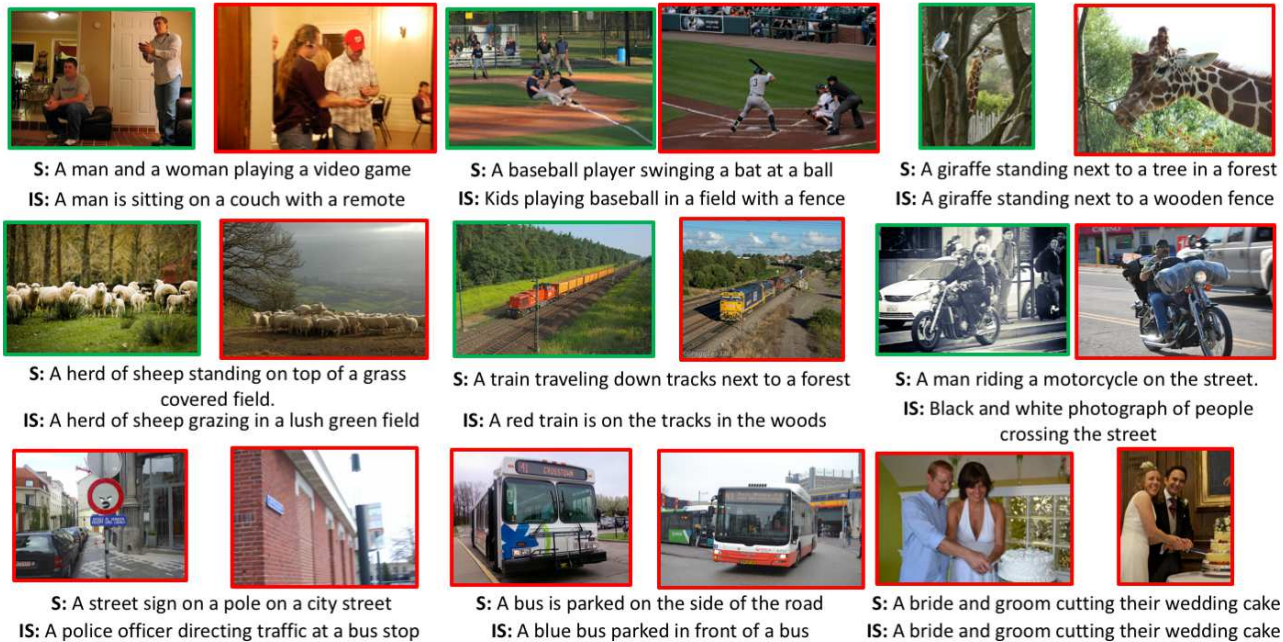IS: A bride and groom cutting their wedding cake

Figure 7: Pairs of images whose captions generated by a generic captioning speaker baseline (S) are identical. We apply our introspective speaker (IS) technique to distinguish the image on the left from the image on the right in each pair. The target image (left) is shown with a green border when the IS generated sentence is able to identify it correctly. Notice how the introspective speaker often refers more unambiguously to the target image. For example, for the sheep image (middle left), the IS generated sentence mentions that the sheep are grazing in a lush green field. In the bottom row we show some failure examples. The bottom left example is interesting, where the model calls the stop sign a policeman. In some cases (the wedding cake image), where the distributions captured by the emitter, and supressor RNN's are identical, our IS approach produces the same sentence as the baseline (S).

tification. For instance, when justifying why an image is a lion and not a tiger, it would be useful to be able to say "because it does not have stripes.", or "because it has a more hair on its face." Beyond pragmatics, the justification task also has interesting relations to human learning. Indeed, we all experience that we learn better when someone takes time out to justify or explain their point of view. One can imagine such justifications being helpful for "machine teaching", where a teacher (machine) can provide justifications to a human learner explaining the rationale for an image belonging to a particular fine-grained category as opposed to a different, possibly mistaken, or confusing fine-grained category.

There are some fundamental limitations to inducing context-aware captions from context-agnostic supervision. For instance, if two distinct concepts are very similar, human-generated context-free descriptions may be identical, and our model (as well as baselines) would fail to extract any discriminative signal. Indeed, it is hard to address such situations without context-aware ground truth.

We believe modeling higher-order reasoning (such as pragmatics) by reusing the sampling distribution from language models can be a powerful tool. It may be applicable to other higher-order reasoning, without necessarily setting up policy gradient estimators on reward functions. Indeed, our inference objective can also be formulated for training. However, initial experiments on this did not yeild significant performance improvements.

## 7. Conclusion

We introduce a novel technique for deriving pragmatic language from recurrent neural network language models, namely, an image-captioning model that takes into account the context of a distractor class or a distractor image. Our technique can be used at inference time to better discriminate between concepts, without having seen discriminative training data. We study two tasks in the vision, and language domain which require pragmatic reasoning: *justification* – explaining why an image belongs to one category as opposed to another, and *discriminative image captioning* – describing an image so that one can distinguish it from a closely related image. Our experiments demonstrate the strength of our method over generative baselines, as well as adaptations of previous work to our setting. We will make the code, and datasets available online.

# References

[1] J. Andreas and D. Klein. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*, 2016. 2, 3, 4, 6

[2] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. L. Zitnick, and D. Parikh. Vqa: Visual question answering. In *International Conference on Computer Vision (ICCV)*, 2015. 2

[3] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. *CoRR*, abs/1409.0473, 2014. 3

[4] S. Banerjee and A. Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. pages 65–72, 2005. 5

[5] L. Benotti and D. R. Traum. A computational account of comparative implicatures for a spoken dialogue agent. 2009. 2

[6] A. Biswas and D. Parikh. Simultaneous active learning of classifiers & attributes via relative feedback. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013. 2

[7] A. Das, S. Kottur, J. M. F. Moura, S. Lee, and D. Batra. Learning Cooperative Visual Dialog Agents with Deep Reinforcement Learning. *arXiv:1703.06585 [cs]*, Mar. 2017. arXiv: 1703.06585. 2

[8] J. Devlin, S. Gupta, R. B. Girshick, M. Mitchell, and C. L. Zitnick. Exploring nearest neighbor approaches for image captioning. *CoRR*, abs/1505.04467, 2015. 5

[9] J. Donahue and K. Grauman. Image classification with annotator rationales. 2

[10] 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. *CoRR*, abs/1411.4389, 2014. 2

[11] H. Fang, S. Gupta, F. N. Iandola, R. Srivastava, L. Deng, P. Dollár, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. *CoRR*, abs/1411.4952, 2014. 2

[12] D. Geman, S. Geman, N. Hallonquist, and L. Younes. Visual turing test for computer vision systems. *Proceedings of the National Academy of Sciences*, 112(12):3618–3623, 2015. 2

[13] H. Grice. Logic and conversation. In *Syntax and Semantics*. Academic Press, 1975. 2

[14] L. A. Hendricks, Z. Akata, M. Rohrbach, J. Donahue, B. Schiele, and T. Darrell. Generating visual explanations. In *ECCV*, 2016. 3, 5

[15] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Computation*, 9:1735–1780, 1997. 4

[16] A. Karpathy. Neuraltalk2 image captioning. https://github.com/karpathy/neuraltalk2. 5, 7

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

[18] S. Kazemzadeh, V. Ordonez, M. Matten, and T. L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, 2014. 2

[19] A. Lazaridou, A. Peysakhovich, and M. Baroni. Multi-Agent Cooperation and the Emergence of (Natural) Language. *arXiv:1612.07182 [cs]*, Dec. 2016. arXiv: 1612.07182. 2

[20] J. Li, M. Galley, C. Brockett, J. Gao, and W. B. Dolan. A diversity-promoting objective function for neural conversation models. In *HLT-NAACL*, 2016. 3

[21] T.-Y. Lin, M. Maire, S. J. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 2

[22] B. T. Lowerre and D. R. Reddy. The harpy speech understanding system. *Trends in Speech Recognition*, 1980. 4

[23] M. Malinowski and M. Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. In *NIPS*, 2014. 2

[24] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. 2016. 2

[25] I. Mordatch and P. Abbeel. Emergence of Grounded Compositional Language in Multi-Agent Populations. *arXiv:1703.04908 [cs]*, Mar. 2017. arXiv: 1703.04908. 2

[26] V. K. Nagaraja, V. I. Morariu, and L. S. Davis. Modeling context between objects for referring expression understanding. In *ECCV*, 2016. 2

[27] A. Parkash and D. Parikh. Attributes for classifier feedback. In *European Conference on Computer Vision (ECCV)*, 2012. 2

[28] S. Reed, Z. Akata, B. Schiele, and H. Lee. Learning deep representations of fine-grained visual descriptions. In *IEEE Computer Vision and Pattern Recognition*, 2016. 3, 5

[29] A. Sadovnik, Y.-I. Chiu, N. Snavely, S. Edelman, and T. Chen. Image description with a goal: Building efficient discriminating expressions for images. In *CVPR*, 2012. 2

[30] A. Sadovnik, A. C. Gallagher, and T. Chen. Not everybody's special: Using neighbors in referring expressions with uncertain attributes. In *CVPR Workshops*, 2013. 2

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. 5

[32] S. Tellex, R. A. Knepper, A. Li, D. Rus, and N. Roy. Asking for help using inverse semantics. In *Robotics: Science and Systems*, 2014. 2

[33] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. *CoRR*, abs/1701.02870, 2017. 3, 5, 6

[34] R. Vedantam, C. Lawrence Zitnick, and D. Parikh. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 5

[35] A. K. Vijayakumar, M. Cogswell, R. R. Selvaraju, Q. Sun, S. Lee, D. J. Crandall, and D. Batra. Diverse beam search: Decoding diverse solutions from neural sequence models. *CoRR*, abs/1610.02424, 2016. 3, 4

[36] O. Vinyals, A. Toshev, S. Bengio, and D. Erhan. Show and tell: A neural image caption generator. *CoRR*, abs/1411.4555, 2014. 2, 3, 5

[37] A. Vogel, M. Bodoia, C. Potts, and D. Jurafsky. Emergence of gricean maxims from multi-agent decision theory. In *HLT-NAACL*, 2013. 2

[38] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 2, 3, 5

[39] S. I. Wang, P. Liang, and C. D. Manning. Learning language games through interaction. *CoRR*, abs/1606.02447, 2016. 2

[40] K. Xu, J. Ba, J. 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. 2, 3, 5

[41] L. Yu, E. Park, A. C. Berg, and T. L. Berg. Visual madlibs: Fill in the blank description generation and question answering. In *ICCV*, 2015. 2

[42] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei. Visual7W: Grounded Question Answering in Images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2

[43] C. L. Zitnick and D. Parikh. Bringing semantics into focus using visual abstraction. In *CVPR*, 2013. 2