

# Pragmatically Informative Image Captioning with Character-Level Inference

Reuben Cohn-Gordon , Noah Goodman , and Chris Potts

Stanford University

{reubencg, ngoodman, cgpotts}@stanford.edu

## Abstract

We combine a neural image captioner with a Rational Speech Acts (RSA) model to make a system that is *pragmatically informative*: its objective is to produce captions that are not merely true but also distinguish their inputs from similar images. Previous attempts to combine RSA with neural image captioning require an inference which normalizes over the entire set of possible utterances. This poses a serious problem of efficiency, previously solved by sampling a small subset of possible utterances. We instead solve this problem by implementing a version of RSA which operates at the level of characters (“a”, “b”, “c”, ...) during the unrolling of the caption. We find that the utterance-level effect of referential captions can be obtained with only character-level decisions. Finally, we introduce an automatic method for testing the performance of pragmatic speaker models, and show that our model outperforms a non-pragmatic baseline as well as a word-level RSA captioner.

## 1 Introduction

The success of automatic image captioning (Farhadi et al., 2010; Mitchell et al., 2012; Karpathy and Fei-Fei, 2015; Vinyals et al., 2015) demonstrates compellingly that end-to-end statistical models can align visual information with language. However, high-quality captions are not merely *true*, but also *pragmatically informative* in the sense that they highlight salient properties and help distinguish their inputs from similar images. Captioning systems trained on single images struggle to be pragmatic in this sense, producing either very general or hyper-specific descriptions.

In this paper, we present a neural image captioning system<sup>1</sup> that is a *pragmatic speaker* as defined by the Rational Speech Acts (RSA) model (Frank and Goodman, 2012; Goodman and Stuhlmüller,



$S_0$  caption: the dog is brown  
 $S_1$  caption: the head of a dog

Figure 1: Captions generated by literal ( $S_0$ ) and pragmatic ( $S_1$ ) model for the target image (in green) in the presence of multiple distractors (in red).

2013). Given a set of images, of which one is the *target*, its objective is to generate a natural language expression which identifies the target in this context. For instance, the literal caption in Figure 1 could describe both the target and the top two distractors, whereas the pragmatic caption mentions something that is most salient of the target. Intuitively, the RSA speaker achieves this by reasoning not only about what is true but also about what it’s like to be a listener in this context trying to identify the target.

This core idea underlies much work in referring expression generation (Dale and Reiter, 1995; Monroe and Potts, 2015; Andreas and Klein, 2016; Monroe et al., 2017) and image captioning (Mao et al., 2016a; Vedantam et al., 2017), but these models do not fully confront the fact that the agents must reason about all possible utterances, which is intractable. We fully address this problem by implementing RSA at the level of characters rather than the level of utterances or words: the neural language model emits individual characters, choosing them to balance pragmatic informativeness with overall well-formedness. Thus, the agents reason not about full utterances, but rather only about all possible character choices, a very small space. The result is that the information encoded recurrently in the neural model allows us

<sup>1</sup>The code is available at <https://github.com/reubenharry/Recurrent-RSA>

to obtain global pragmatic effects from local decisions. We show that such character-level RSA speakers are more effective than literal captioning systems at the task of helping a reader identify the target image among close competitors, and outperform word-level RSA captioners in both efficiency and accuracy.

## 2 Bayesian Pragmatics for Captioning

In applying RSA to image captioning, we think of captioning as a kind of reference game. The *speaker* and *listener* are in a shared context consisting of a set of images  $W$ , the speaker is privately assigned a target image  $w^* \in W$ , and the speaker’s goal is to produce a caption that will enable the listener to identify  $w^*$ .  $U$  is the set of possible utterances. In its simplest form, the *literal speaker* is a conditional distribution  $S_0(u|w)$  assigning equal probability to all true utterances  $u \in U$  and 0 to all others. The pragmatic listener  $L_0$  is then defined in terms of this literal agent and a prior  $P(w)$  over possible images:

$$L_0(w|u) \propto \frac{S_0(u|w) * P(w)}{\sum_{w' \in W} S_0(u|w') * P(w')} \quad (1)$$

The pragmatic speaker  $S_1$  is then defined in terms of this pragmatic listener, with the addition of a rationality parameter  $\alpha > 0$  governing how much it takes into account the  $L_0$  distribution when choosing utterances.  $P(u)$  is here taken to be a uniform distribution over  $U$ :

$$S_1(u|w) \propto \frac{L_0(w|u)^\alpha * P(u)}{\sum_{u' \in U} L_0(w|u')^\alpha * P(u')} \quad (2)$$

As a result of this back-and-forth, the  $S_1$  speaker is reasoning not merely about what is true, but rather about a listener reasoning about a literal speaker who reasons about truth.

To illustrate, consider the pair of images 2a and 2b in Figure 2. Suppose that  $U = \{bus, red\ bus\}$ . Then the literal speaker  $S_0$  is equally likely to produce *bus* and *red bus* when the left image 2a is the target. However,  $L_0$  breaks this symmetry; because *red bus* is false of the right bus,  $L_0(2a|bus) = \frac{1}{3}$  and  $L_0(2b|bus) = \frac{2}{3}$ . The  $S_1$  speaker therefore ends up favoring *red bus* when trying to convey 2a, so that  $S_1(red\ bus|2a) = \frac{3}{4}$  and  $S_1(bus|2a) = \frac{1}{4}$ .



$S_0$  caption: a double decker bus  
 $S_1$  caption: a red double decker bus

Figure 2: Captions for the target image (in green).

## 3 Applying Bayesian Pragmatics to a Neural Semantics

To apply the RSA model to image captioning, we first train a neural model with a CNN-RNN architecture (Karpathy and Fei-Fei, 2015; Vinyals et al., 2015). The trained model can be considered an  $S_0$ -style distribution  $P(\text{caption}|\text{image})$  on top of which further listeners and speakers can be built. (Unlike the idealized  $S_0$  described above, a neural  $S_0$  will assign some probability to untrue utterances.)

The main challenge for this application is that the space of utterances (captions)  $U$  will be very large for any suitable captioning system, making the calculation of  $S_1$  intractable due to its normalization over all utterances. The question, therefore, is how best to approximate this inference. The solution employed by Monroe et al. (2017) and Andreas and Klein (2016) is to sample a small subset of probable utterances from the  $S_0$ , as an approximate prior upon which exact inference can be performed. While tractable, this approach has the shortcoming of only considering a small part of the true prior, which potentially decreases the extent to which pragmatic reasoning will be able to apply. In particular, if a useful caption never appears in the sampled prior, it cannot appear in the posterior.

### 3.1 Step-Wise Inference

Inspired by the success of the “emitter-suppressor” method of Vedantam et al. (2017), we propose an incremental version of RSA. Rather than performing a single inference over utterances, we perform an inference *for each step of the unrolling of the utterance*.

We use a character-level LSTM, which defines a distribution over characters  $P(u|pc, \text{image})$ , where *pc* (“partial caption”) is a string of char-

acters constituting the caption so far and  $u$  is the next character of the caption. This is now our  $S_0$ : given a partially generated caption and an image, it returns a distribution over which character should next be added to the caption. The advantage of using a character-level LSTM over a word-level one is that  $U$  is much smaller for the former ( $\approx 30$  vs.  $\approx 20,000$ ), making the ensuing RSA model much more efficient.

We use this  $S_0$  to define an  $L_0$  which takes a partial caption and a new character, and returns a distribution over images. The  $S_1$ , in turn, given a target image  $w^*$ , performs an inference over the set of possible characters to determine which is best with respect to the listener choosing  $w^*$ .

At timestep  $t$  of the unrolling, the listener  $L_0$  takes as its prior over images the  $L_0$  posterior from timestep  $(t - 1)$ . The idea is that as we proceed with the unrolling, the  $L_0$  priors on which image is being referred to may change, which in turn should affect the speaker's actions. For instance, the speaker, having made the listener strongly in favor of the target image, is less compelled to continue being pragmatic.

### 3.2 Model Definition

In our incremental RSA, speaker models take both a target image and a partial caption  $pc$ . Thus,  $S_0$  is a neurally trained conditional distribution  $S_0^t(u|w, pc_t)$ , where  $t$  is the current timestep of the unrolling and  $u$  is a character.

We define the  $L_0^t$  in terms of the  $S_0^t$  as follows, where  $ip$  is a distribution over images representing the  $L_0$  prior:

$$L_0^t(w|u, ip_t, pc_t) \propto S_0^t(u|w, pc_t) * ip_t(w) \quad (3)$$

Given an  $S_0^t$  and  $L_0^t$ , we define  $S_1^t$  and  $L_1^t$  as:

$$S_1^t(u|w, ip_t, pc_t) \propto S_0^t(u|w, pc_t) * L_0^t(w|u, ip_t, pc_t)^\alpha \quad (4)$$

$$L_1^t(w|u, ip_t, pc_t) \propto L_0^t(w|u, ip_t, pc_t) * S_0^t(u|w, pc_t) \quad (5)$$

**Unrolling** To perform greedy unrolling (though in practice we use a beam search) for either  $S_0$  or  $S_1$ , we initialize the state as a partial caption  $pc_0$  consisting of only the start token and a uniform prior over the images  $ip_0$ . Then, for  $t > 0$ , we use our incremental speaker model  $S_0$  or  $S_1$  to

generate a distribution over the subsequent character  $S^t(u|w, ip_t, pc_t)$ , and add the character  $u$  with highest probability density to  $pc_t$ , giving us  $pc_{t+1}$ . We then run our listener model  $L_1$  on  $u$ , to obtain a distribution  $ip_{t+1} = L_1^t(w|u, ip_t, pc_t)$  over images that the  $L_0$  can use at the next timestep.

This incremental approach keeps the inference itself very simple, while placing the complexity of the model in the recurrent nature of the unrolling.<sup>2</sup> While our  $S_0$  is character-level, the same incremental RSA model works for a word-level  $S_0$ , giving rise to a word-level  $S_1$ . We compare character and word  $S_1$ s in section 4.2.

As well as being incremental, these definitions of  $S_1^t$  and  $L_1^t$  differ from the typical RSA described in section 2 in that  $S_1^t$  and  $L_1^t$  draw their priors from  $S_0^t$  and  $L_0^t$  respectively. This generalizes the scheme put forward for  $S_1$  by Andreas and Klein (2016). The motivation is to have Bayesian speakers who are somewhat constrained by the  $S_0$  language model. Without this, other methods are needed to achieve English-like captions, as in Vedantam et al. (2017), where their equivalent of the  $S_1$  is combined in a weighted sum with the  $S_0$ .

## 4 Evaluation

Qualitatively, Figures 1 and 2 show how the  $S_1$  captions are more informative than the  $S_0$ , as a result of pragmatic considerations. To demonstrate the effectiveness of our method quantitatively, we implement an automatic evaluation.

### 4.1 Automatic Evaluation

To evaluate the success of  $S_1$  as compared to  $S_0$ , we define a listener  $L_{eval}(image|caption) \propto P_{S_0}(caption|image)$ , where  $P_{S_0}(caption|image)$  is the total probability of  $S_0$  incrementally generating *caption* given *image*. In other words,  $L_{eval}$  uses Bayes' rule to obtain from  $S_0$  the posterior probability of each image  $w$  given a full caption  $u$ .

The neural  $S_0$  used in the definition of  $L_{eval}$  must be trained on separate data to the neural  $S_0$  used for the  $S_1$  model which produces captions, since otherwise this  $S_1$  production model effectively has access to the system evaluating it. As Mao et al. (2016b) note, "a model might 'com-

<sup>2</sup>The move from standard to incremental RSA can be understood as a switching of the order of two operations; instead of unrolling a character-level distribution into a sentence level one and then applying pragmatics, we apply pragmatics and then unroll. This generalizes to any recursive generation of utterances.

municate’ better with itself using its own language than with others”. In evaluation, we therefore split the training data in half, with one part for training the  $S_0$  used in the caption generation model  $S_1$  and one part for training the  $S_0$  used in the caption evaluation model  $L_{eval}$ .

We say that the caption succeeds as a referring expression if the target has more probability mass under the distribution  $L_{eval}(image|caption)$  than any distractor.

**Dataset** We train our production and evaluation models on separate sets consisting of regions in the Visual Genome dataset (Krishna et al., 2017) and full images in MSCOCO (Chen et al., 2015). Both datasets consist of over 100,000 images of common objects and scenes. MSCOCO provides captions for whole images, while Visual Genome provides captions for regions within images.

Our test sets consist of clusters of 10 images. For a given cluster, we set each image in it as the target, in turn. We use two test sets. Test set 1 (TS1) consists of 100 clusters of images, 10 for each of the 10 most common objects in Visual Genome.<sup>3</sup> Test set 2 (TS2) consists of regions in Visual Genome images whose ground truth captions have high word overlap, an indicator that they are similar. We again select 100 clusters of 10. Both test sets have 1,000 items in total (10 potential target images for each of 100 clusters).

**Captioning System** Our neural image captioning system is a CNN-RNN architecture<sup>4</sup> adapted to use a character-based LSTM for the language model.

**Hyperparameters** We use a beam search with width 10 to produce captions, and a rationality parameter of  $\alpha = 5.0$  for the  $S_1$ .

## 4.2 Results

As shown in Table 1, the character-level  $S_1$  obtains higher accuracy (68% on TS1 and 65.9% on TS2) than the  $S_0$  (48.9% on TS1 and 47.5% on TS2), demonstrating that  $S_1$  is better than  $S_0$  at referring.

**Advantage of Incremental RSA** We also observe that 66% percent of the times in which the  $S_1$  caption is referentially successful and the  $S_0$

Model	TS1	TS2
Char $S_0$	48.9	47.5
Char $S_1$	<b>68.0</b>	<b>65.9</b>
Word $S_0$	57.6	53.4
Word $S_1$	60.6	57.6

Table 1: Accuracy on both test sets.

caption is not, for a given image, the  $S_1$  caption is not one of the top 50  $S_0$  captions, as generated by the beam search unrolling at  $S_0$ . This means that in these cases the non-incremental RSA method of Andreas and Klein (2016) could not have generated the  $S_1$  caption, if these top 50  $S_0$  captions were the support of the prior over utterances.

**Comparison to Word-Level RSA** We compare the performance of our character-level model to a word-level model.<sup>5</sup> This model is incremental in precisely the way defined in section 3.2, but uses a word-level LSTM so that  $u \in U$  are words and  $U$  is a vocabulary of English. It is evaluated with an  $L_{eval}$  model that also operates on the word level.

Though the word  $S_0$  performs better on both test sets than the character  $S_0$ , the character  $S_1$  outperforms the word  $S_1$ , demonstrating the advantage of a character-level model for pragmatic behavior. We conjecture that the superiority of the character-level model is the result of the increased number of decisions where pragmatics can be taken into account, but leave further examination for future research.

**Variants of the Model** We further explore the effect of two design decisions in the character-level model. First, we consider a variant of  $S_1$  which has a prior over utterances determined by an LSTM language model trained on the full set of captions. This achieves an accuracy of 67.2% on TS1. Second, we consider our standard  $S_1$  but with unrolling such that the  $L_0$  prior is drawn uniformly at each timestep rather than determined by the  $L_0$  posterior at the previous step. This achieves an accuracy of 67.4% on TS1. This suggests that neither this change of  $S_1$  nor  $L_0$  priors has a large effect on the performance of the model.

<sup>3</sup>Namely, *man, person, woman, building, sign, table, bus, window, sky, and tree*.

<sup>4</sup>[https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image\\_captioning](https://github.com/yunjey/pytorch-tutorial/tree/master/tutorials/03-advanced/image_captioning)

<sup>5</sup>Here, we use greedy unrolling, for reasons of efficiency due to the size of  $U$  for the word-level model, and set  $\alpha = 1.0$  from tuning on validation data. For comparison, we note that greedy character-level  $S_1$  achieves an accuracy of 61.2% on TS1.



## 5 Conclusion

We show that incremental RSA at the level of characters improves the ability of the neural image captioner to refer to a target image. The incremental approach is key to combining RSA with language models: as utterances become longer, it becomes exponentially slower, for a fixed  $n$ , to subsample  $n\%$  of the utterance distribution and *then* perform inference (non-incremental approach). Furthermore, character-level RSA yields better results than word-level RSA and is far more efficient.

## Acknowledgments

Many thanks to Hiroto Udagawa and Poorvi Bhargava, who were involved in early versions of this project. This material is based in part upon work supported by the Stanford Data Science Initiative and by the NSF under Grant No. BCS-1456077. This work is also supported by a Sloan Foundation Research Fellowship to Noah Goodman.

## References

- Jacob Andreas and Dan Klein. 2016. [Reasoning about pragmatics with neural listeners and speakers](#). In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182. Association for Computational Linguistics.
- Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. [Microsoft COCO captions: Data collection and evaluation server](#). *arXiv preprint arXiv:1504.00325*.
- Robert Dale and Ehud Reiter. 1995. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19(2):233–263.
- Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European Conference on Computer Vision*, pages 15–29. Springer.
- Michael C. Frank and Noah D. Goodman. 2012. Predicting pragmatic reasoning in language games. *Science*, 336(6084):998.
- Noah D Goodman and Andreas Stuhlmüller. 2013. Knowledge and implicature: Modeling language understanding as social cognition. *Topics in Cognitive Science*, 5(1):173–184.
- Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3128–3137.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. 2016a. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11–20. IEEE.
- Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016b. Generation and comprehension of unambiguous object descriptions. pages 11–20.
- Margaret Mitchell, Jesse Dodge, Amit Goyal, Kota Yamaguchi, Karl Stratos, Xufeng Han, Alyssa Mensch, Alex Berg, Tamara Berg, and Hal Daume III. 2012. [Midge: Generating image descriptions from computer vision detections](#). In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 747–756. Association for Computational Linguistics.
- Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. [Colors in context: A pragmatic neural model for grounded language understanding](#). *Transactions of the Association for Computational Linguistics*, 5:325–338.
- Will Monroe and Christopher Potts. 2015. Learning in the Rational Speech Acts model. In *Proceedings of 20th Amsterdam Colloquium*, Amsterdam. ILLC.
- Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. 2017. [Context-aware captions from context-agnostic supervision](#). *arXiv preprint arXiv:1701.02870*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3156–3164.