Character-Based Pragmatic Captioning

Reuben Cohn-Gordon, Hiroto Udagawa, Poorvi Bhargava
Stanford University

Introduction

While semantics is key to artificial intelligence, human language use involves another linguistic component: pragmatics, or the ability to resolve ambiguity from context.

Referring expression generation is an example of pragmatic reasoning, in which a speaker is presented with a pair (or more generally a set) of objects, and must generate a linguistic expression which singles out only one of the objects.

Bayesian models of pragmatics involve:
1. a Speaker *S*, modeled as a distribution over utterances given world states,
2. a Listener *L*, modeled as a distribution over world states given utterances.

which can naturally be extended to neural pragmatic caption generation.

Previous work focused on word models for image captioning. Our project explored both character and word models in both end-to-end and modular approaches on a new dataset. Limitations involved challenges in quantitative analysis on caption quality.

Problem Statement

Our model addresses the linguistic challenge of pragmatic caption generation, which aims to generate a caption which serves as a referring expression for one but not the other of a pair (or more generally a set) of images.

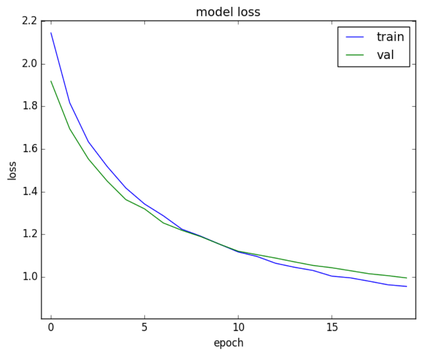

Figure 1. Training and Validation Loss.

Technical Approach

Our model, trained on the Visual Genome dataset, is composed of three parts, derived from the work of Vedantam et al.[3]:

1) $S_0 = f_{S_0}(I) = P(C|I)$
which generates "literal speaker" semantic captions
- pretrained ResNet-50 CNN for forward pass, feature vectors
- a word-based LSTM to generate captions from feature vectors

2) $L = f_L(C, I_t, I_d) = P_{posterior}(I_t|C) = \frac{P_{prior}(I_t) * P(C|I_t)}{\sum_{j \in \{t,d\}} P_{prior}(I_j) * P(C|I_j)}$
a Bayesian method "listener" which tries to differentiate between the target image, I_t, and the distractor image, I_d

3) $S_1 = f_{S_1}(I_t, I_d) = \underset{c}{\text{argmax}}\{\lambda S_0 + (1-\lambda)L\}$, where $0 \leq \lambda \leq 1$
which generates "pragmatic speaker" captions.

Results

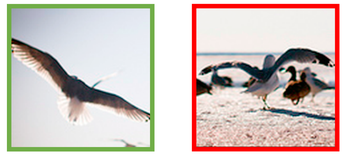

Figure 2.
S_0 Target: "this is a bird"
S_0 Distractor: "this is a bird"
S_1 Target: "the sky is blue"

Figure 3.
S_0 Target: "the shirt is blue"
S_0 Distractor: "the grass is green"
S_1 Target: "red shirt on the man"

Figure 4.
S_0 Target: "this is a table"
S_0 Distractor: "a bunch of bananas"
S_1 Target: "the table is white"

Figure 5.
S_0 Target: "the dog is black"
S_0 Distractor: "this is a building"
S_1 Target: "a black and white dog"

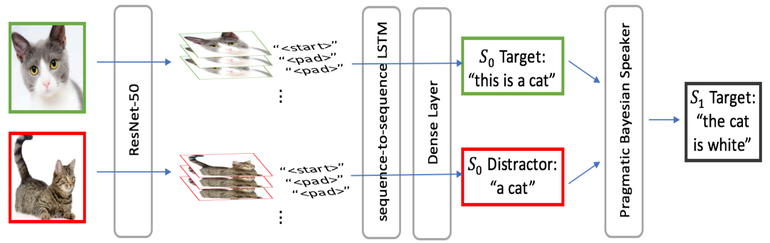

Figure 6. Model Architecture

Dataset

The chosen dataset was Visual Genome, which contains about 100,000 images, with over 4 million region descriptions (~40 region descriptions per image). The average length of each region description was about 5 words.

Discussion

The word model was less challenging to implement, however, a character model, in theory, has several advantages:
- versatility in predictions, to cope with typos
- no need for a pre-trained word embedding
- smaller beam width in beam search

We expected and found that S_1 generated longer, more detailed captions so as to better describe the target image., even when the distractor image was not similar to the target.

As expected, adding beam search produced qualitatively better captions, as did the modular model.

Conclusions and Future Work

Pragmatic image captioning is challenging to evaluate quantitatively. Future work includes building a separate listener model that can take in the generated caption and predict the target image from a pair of images. This can act as a quantitative method for pragmatic caption evaluation. A limitation of such an approach is that both models may interpret parts of the image incorrectly in the same way and give the caption an exaggerated score.

References

[1] N. D. Goodman and M. C. Frank. Pragmatic language inter- pretation as probabilistic inference. *Trends in Cognitive Sci- ences*, 20(11):818–829, 2016.
[2] J. Mao, J. Huang, A. Toshev, O. Camburu, A. L. Yuille, and K. Murphy. Generation and comprehension of unambiguous object descriptions. pages 11–20, 2016.
[3] R. Vedantam, S. Bengio, K. Murphy, D. Parikh, and G. Chechik. Context-aware captions from context-agnostic supervision. *arXiv preprint arXiv:1701.02870*, 2017.
[4] K.Xu,J.Ba,R.Kiros,K.Cho,A.Courville,R.Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural im- age caption generation with visual attention. In *International Conference on Machine Learning*, pages 2048–2057, 2015.