CS224U Literature Review

Julia Gong, Suvir Mirchandani, Benjamin Newman

May 2019

1 Introduction

In our project, we are generally interested in investigating the role that grounded pragmatic models can play in evaluating natural language generation models. In particular, we hope to focus on evaluation metrics for generated color descriptions (*captions*). In this vein, we have divided our literature review into three components. First, we investigate the literature on grounded language understanding in the color domain and look at the insights regarding pragmatics that we can take away from these models. Next, we review current methods for evaluating captioning models, and in the process, put in context what we might want an evaluation metric to capture. Finally, we consider the literature related to the concept of cycle consistency. This concept, which in this context refers to the fact that relevant features of a color or image ought to be extracted from a faithful caption, serves as as a possible inspiration for how such a pragmatically-informed evaluation metric might be implemented. In what follows, we consider each of these sections separately and then combine the ideas from these places in our future work.

2 Grounded Language Understanding for Image Captioning and Color Description

Image captioning systems benefit from pragmatic reasoning, especially in the context of reference games. A reference game involves a listener and a speaker who are presented with the same set of referents (images). The speaker is assigned a target referent and produces a description which is shown to the listener, whose goal it is to identify the speaker's target. Recursive pragmatic reasoning is thus important for both the speaker and the listener.

As a useful stepping stone to developing methods for pragmatic captioning and evaluation, we review the color description problem. While color descriptions can be complex and compositional, the feature space of colors is much smaller than that for arbitrary images. This allows the tendencies of captioning models to be probed and visualized.

Munroe (2010) collected a dataset of colors and descriptions from an open online survey, and McMahan and Stone (2015) offer a cleansed version of the dataset that is appropriate for training and testing color captioning models. Monroe, Goodman, and Potts (2016) present an long short-term memory (LTSM) recurrent neural network model for generating color captions given a color's HSL (hue-saturation-lightness) value. Their model uses Fourier-basis features of the HSL value to capture periodicity in color space. For each token in a given description, there is an LTSM layer which accepts the color's feature representation concatenated with an embedding of the output from the previous layer (the first layer uses an embedding of the start token). Each LSTM layer passes thorugh a fully-connected layer, followed by a softmax to predict the next token.

The model captures important aspects of color semantics, such as nonconvexity in description denotations. (On the other hand, in their Bayesian generative model LUX, McMahan and Stone (2015) assume color categories are convex regions. The prime benefit of nonconvexity

is for descriptors such as *greenish*, which have stronger conditional likelihood for colors *near* green but not green itself). Furthermore, the model is able to capture the compositionality of certain modifiers and construct descriptions that are unseen in training. For evaluating their system, Monroe et al. (2016) use three metrics: perplexity (to measure how highly the model's probability distribution ranks descriptions in the dataset), the Akaike information criterion (to measure the fit of the model while penalizing model complexity), and accuracy (to measure how often model predictions exactly match the dataset). Their RNN system with Fourier features beats LUX in all three metrics.

This system deals with describing colors in isolation—not in reference games—and thus focus on the semantic complexity of captions rather than pragmatic reasoning in a color reference game. In contrast, Andreas and Klein (2016) model pragmatic reasoning for the image description problem (a speaker tries to describe an image), and Monroe, Hawkins, et al. (2017) focus on the flip side of the color description problem described above (a listener tries to identify a color from a description, given a set from which to choose).

Andreas and Klein (2016) present a description generator based on a reasoning speaker model, which itself is built upon neural models for a literal listener and literal speaker. This approach is rooted in the Rational Speech Acts Model (Goodman and Frank, 2016).

The literal listener model produces a distribution over the referent choices. The description space is large, so the listener's distribution is approximated by sampling descriptions given by a literal speaker (which yields descriptions for a single referent in isolation). Finally, the reasoning speaker computes a weighted joint probability for a sentence being produced by the literal speaker and correctly interpreted by the literal listener.

For evaluation, Andreas and Klein (2016) use two metrics: fluency and accuracy. Fluency is calculated using human ratings of linguistic quality, and accuracy is the success rate of humans, acting as listeners, selecting the target referent of the speaker model. The reasoning speaker significantly improves on the literal speaker model and a prior contrastive approach by Mao et al. (2016).

At the intersection of the color description problem and image-based reference games, Monroe, Hawkins, et al. (2017) focus on the problem of identifying colors from descriptions, given a particular context of three colors from which to choose. That is, their model of a listener (who, given a color description, chooses a color from the context) depends upon a model of a pragmatic speaker and a literal listener. Pragmatics is important in this task because color descriptions depend heavily on context. For example, in their training dataset, humans tended to be more verbose, use more comparatives and superlatives, use more negatives, and use more specific words when the colors in the context were more similar. Recursively reasoning about possible utterances of a hypothetical speaker given a hypothetical listener makes the model more cognitively realistic.

The authors blend two pragmatic listener models: one based on an RNN representing the literal listener, followed by Bayesian inference to model a pragmatic speaker; and one based on an RNN directly representing a literal speaker. Both pragmatic listeners outperform the literal listener in accuracy, and there is further improvement when both listener-based agents and speaker-based agents are used.

We return to the discussion of pragmatic models after contextualizing recent work in image captioning and cycle consistency.

3 Image Captioning Evaluation Metrics

In a reference game setting, one might want to ensure that the model that generates language produces utterances useful for humans to achieve some task. In a setting where context does

not appear to play as large a role, good evaluation metrics are even more elusive. There are a number of different approaches to this problem, three groups of which will be discussed here: n-gram overlap, evaluation by model, and human evaluation. Each one of these has pros and cons. N-gram overlap is usually easy to compute and fairly intelligible. They do require a large number captions for the best evaluation performance and can usually be tricked into giving high scores to captions that humans would find completely ungrammatical or at least uninformative. One such example is the Consensus-based Image Description Evaluation (CIDEr) metric introduced by Vedantam, Zitnick, and Parikh (2015). Very simply, given a candidate caption and a list of reference captions, CIDEr breaks up the captions into n-grams, computes TF-IDF weightings of the n-grams across the entire dataset and then computes the average cosine similarity between the TF-IDF vectors for the candidate and reference captions. They find that this metric along with other n-gram based ones such as BLEU, ROUGE, and METEOR also perform better when there are more reference captions. As a result, they create two datasets with 50 reference captions for each image.

As a response to some of the shortcomings of n-gram based evaluation metrics and the high cost of human evaluation, Cui et al. (2018) propose a learned "critic" that attempts to distinguish between human and model-generated captions. Their model takes in a candidate caption, a reference caption, and the image the two supposedly refer to and outputs a probability that the the candidate caption is generated by a machine versus by a human. They emphasize the robustness of their model and illustrate that when it makes mistakes, those mistakes can be used as negative examples in later training data to prevent them from occurring in the future. This is compared to n-gram metrics which necessitate changing the metric to avoid these kinds of mistakes. While these models do seem to correlate better with human judgments than n-gram metrics, they are also more opaque and it is harder to tell what exactly they are basing their evaluation off of. One possibility that Cui et al. (2018) brings up is that humans use rare words much more often than state-of-the-art models do, so if an evaluator sees a relatively rare word, it can use that as a shortcut to predict that a caption is human-generated.

A final approach toward evaluating models is to just use humans. This has the obvious benefit that humans are the best judges of language quality, so their judgments will be more precise than n-gram or model-based evaluation. That said, human evaluation does have some potential drawbacks. In addition to the high cost of scaling, Hashimoto, Zhang, and Liang (2019) points out that humans can only tell if a model output is good; they cannot tell if a model can produce a sufficient range of utterances. They claim that a metric has to take into account both of these features which they refer to as *quality* and *diversity*. Measuring quality with human judgments and diversity with the distribution of the model itself, they combine these two scores to create a metric they call Human Unified with Statistical Evaluation (HUSE). This is a standardized metric for exploring the trade-off between model quality and diversity and is very interpretable.

The question of diversity in model output is not specifically highlighted by either Vedantam, Zitnick, and Parikh (2015) or Cui et al. (2018). Both CIDEr and the learned metric are focused on approximating the quality of captions. They ask the question: "does this model produce utterances similar to what a human would?" CIDEr somewhat disregards the diversity aspect of generation by positing that a good caption is close to some consensus caption in a TF-IDF vector space. That said, comparing a candidate caption to 50 other reference captions might give a reasonable approximation for diversity as well. In the case of the learned metric, attempting to distinguish model outputs from actual captions can be seen as encouraging a model to closely regurgitate their training data. The reliance on image features is really the only aspect that prevents a model from doing just that. Their nod to the disparity in word frequency distribution between models and humans is an acknowledgement of the diversity

failure of these models. They note that their learned critic can detect if an utterance comes from a human or a model if it uses a rare word. However, it is unclear if a model could be encouraged to use rare words to fool the evaluator or instead it would just learn to assign probabilities to those words only in certain contexts.

One important area of these evaluation metrics that appears to be under-explored is the role that pragmatics (as described in Section 2) might play in evaluating image captions, which we elaborate on in the Section 5 section below.

4 Cycle Consistency as a Framework for Image Captioning System Evaluation

One key reason for difficulties in creating highly effective caption evaluation metrics—and generative models in general—has been the subjectivity and incomparability that comes with evaluating a generated value (caption) that has no well-defined ground truth value. One way to resolve this issue is to translate the output (caption) back into the domain of the original input value (image), for which there is a definitive ground truth (the original image). Motivated by this principle, one increasingly popular framework for building generative models has been the notion of cycle consistency: building models that translate the intended output value back into the input domain and taking the performance of the model to be a direct comparison between the translated output value and the original input value, both of which reside in the same domain.

For the color captioning task in particular, the interpretation of cycle consistency is that if a system can successfully translate a target color into a caption and translate this back into a color, the caption likely preserves the most salient and relevant features of the color needed for a speaker to refer to it and for a listener to pick it out from other colors. Translation between domains without losing relevant features forms the basis for generative architectures, such as the CycleGAN model in computer vision (Zhu et al., 2017).

CycleGANs are designed for image-to-image translation, also known as style transfer from one image domain to another. The primary motivation of CycleGANs comes from the need for image style transfer models that do not rely on paired data of specific corresponding images in the input and target domains, which is expensive and scarce—instead, it only makes use of sets of data, one from each of the domains. Though not immediately apparent, the motivation for this system deeply parallels that of the task of color captioning in reference games. In color captioning, the two domains are colors and natural language descriptions rather than different image domains; however, the motivation of the CycleGAN stems from handling translation tasks with no bijective input-output ground truth pairing, and as we will see, generating data that does not lose the diversity of the true target distribution.

The CycleGAN architecture consists of two translators, each from one domain to the other domain. Each translator's objective is to learn the true distribution of images in its target domain, and thus an optimally learned pair of translators form a bijection between the two domains. The model objective uses the cycle consistency loss that encourages consistency between the two translators when learning the target domain distributions, in addition to the adversarial loss of the conventional generative adversarial model (GAN) that imposes pressure on the generator and discriminator in each of the translators. The CycleGAN outperforms baseline unpaired models such as BiGAN and CoGAN, but does not surpass pix2pix, which is trained on paired data, though the qualitative results come close. The crucial contributions of this study are that (1) cycle consistency encouraged less mode collapse (producing outputs with very low diversity) as opposed to only a GAN+forward (only a translator from input to target domain)

model, allowing the CycleGAN to handle translation between domains with higher image diversity, and (2) CycleGAN is second only to models trained on paired images. The paper's results show that cycle consistency, both as an architecture and as an objective, is promising for handling diverse mapping domains, and that it is the state-of-the-art for translation tasks that have domain distributions instead of paired ground truth data.

Even more closely related to color captioning are the tasks of image captioning (image-to-text) and image generation (text-to-image), which have seen promising usage of cycle consistency for producing informative image captions. Hagiwara, Mukuta, and Harada (2019) tackles the image captioning task under the premise that a caption generated from an image represents said image faithfully if the image reconstructed from the caption closely resembles the original image. Similar in flavor to Zhu et al. (2017), the model utilizes two generator-discriminator pairs: one from images to captions and one from captions to images. The text encoder used in the image generator was a bi-directional LSTM, and the cycle consistency enforced for images involve comparison on the level of pixels as well as features, which are extracted using a VGG16 model. For cycle consistency between captions, the generated text's vector is used for comparisons. For evaluation, both quantitative (automatically generated metrics such as inception score) and qualitative (human ratings of faithfulness) metrics are used. The captioning model that uses a paired dataset rather than the unpaired dataset from Zhu et al. (2017) performs slightly better than previous work (without cycle consistency) on most datasets, but the models that utilized unpaired data does not show significant improvement from previous work. Also of note is the authors' observation that translating from text to image is a more difficult task than the reverse operation due to the lower amount of information in textual data. The biggest improvement that can be made on this model in terms of image captioning is thus improving the text-to-image GAN, which appears to be the weak link in the cycle consistency framework. The same information imbalance in Hagiwara, Mukuta, and Harada (2019) may not be as present in the color captioning framework, however, as colors have a much smaller feature space than images. In any case, the cycle consistency framework enhances the performance of GAN models on the image captioning task.

While the work of Hagiwara, Mukuta, and Harada (2019) involves applying cycle consistency to image captioning, Gorti and Ma (2018) uses cycle consistency for image generation. The model translates captions into images, and those images back into natural language for evaluation. Similar to Hagiwara, Mukuta, and Harada (2019), Gorti and Ma (2018) use a dual translator framework; however, their slightly different approach divides into two stages: the stage 1 generator maps from text to images, which produces a fuzzy first image that is then fed as input into the second generator that upsamples the input and creates a higher fidelity output image. Each of the generators then has a corresponding discriminator. They generate Skip Thought embeddings for the dataset captions as part of data pre-processing, and the objective function minimizes the distance between generated embeddings and ground truth embeddings. An interesting improvement on the work in Hagiwara, Mukuta, and Harada (2019) might be to incorporate the text-to-image architecture of Gorti and Ma (2018) in the text-to-image branch of this GAN, as this model has a higher quantitative inception score than cycle-free models. One very relevant result from this paper also highlights the key insight from Zhu et al. (2017): cycle consistency again allows the model to drastically avoid mode collapse in comparison to models without cycle consistency, thereby increasing diversity of the GAN output. In addition, the authors briefly mention the idea of comparing the semantic differences between the generated and target captions using the Skip Thought embeddings from the model as a future research direction, which keys into the objective of our project.

5 Future Work

A potential area of exploration we see in current evaluation metrics is that of incorporating pragmatics and cycle consistency into image caption evaluation. We could argue that it is only necessary for a model to produce high quality and diverse outputs as long as those outputs help the model use language to achieve some task as a human would. What is the goal in these image captioning tasks? In grounded, reference game tasks like the ones described in Section 2, a good utterance is one that allows a partner to select the correct image. In a less grounded task, like ones where annotators are presented with individual pictures, there does not appear to be as strong of a communicative goal. Instead, what we might want is for someone reading the caption to be able to capture and recreate some important features of the image. Vedantam, Zitnick, and Parikh (2015) referenced this when they were describing how annotators performed their caption similarity task: "Interestingly, even though we do not say that the sentences are image descriptions, some workers commented that they were imagining the scene to make the choice." Cui et al. (2018) also reference the idea that captions and images should share features as motivation for including the image in their critic model—they want to ensure that their metric evaluates if a model "focuses on the important aspects of the image." In short, having an evaluator act like a literal listener in an grounded language task, trying to imagine an image is an under-explored area of the image captioning literature.

Drawing on the results of Zhu et al. (2017), Hagiwara, Mukuta, and Harada (2019), and Gorti and Ma (2018), there is great potential in using cycle consistency as a framework for developing an evaluation metric for image captioning systems. Cycle consistency can be interpreted as having a speaker (image or color-to-text) and listener (text-to-image or color) perform optimally with one another to achieve consistency in their descriptions and understandings of images or colors. Developing a listener model as an evaluation mechanism for a speaker model, or an image captioning system, is an interesting area of research to probe, especially as cycle consistency is an appropriate tool for this task.

As mentioned previously, cycle consistency improves the performance of image captioning and image generation systems for two key reasons: it is more conducive to handling diverse mapping domains and diversity in its output (preventing mode collapse), and it is useful for translation tasks that have unpaired domain distributions rather than paired ground truth data. As hinted by Gorti and Ma (2018), taking these models one step further and using the model outputs as a means to understand the semantic differences between generated captions, for example, or feature-level differences between images, is a plausible next step that fits well into the model evaluation framework.

The final nuance worth noting is that, while the aforementioned works in this literature review aim to address the underlying objective of producing images and text that are consistent (and thus that can be used to reconstruct the original input), the purpose of these systems has been to generate higher-quality outputs that are then evaluated using paired datasets that have ground truth values. The potential area of expansion on these works that we envision is the usage of cycle consistency and the incorporation of linguistic pragmatics into such a system as an actual evaluation metric for captioning systems, rather than just a tool for improving the systems' performance.

References

[1] Jacob Andreas and Dan Klein. "Reasoning about Pragmatics with Neural Listeners and Speakers". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016,

pp. 1173–1182. DOI: 10.18653/v1/D16-1125. URL: https://www.aclweb.org/anthology/D16-1125.

[2] Yin Cui et al. "Learning to evaluate image captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 5804–5812.

[3] Noah Goodman and Michael Frank. "Pragmatic language interpretation as probabilistic inference". In: *Trends in cognitive sciences* 20.11 (2016), pp. 818–829.

[4] Satya Krishna Gorti and Jeremy Ma. "Text-to-Image-to-Text Translation using Cycle Consistent Adversarial Networks". In: *arXiv preprint arXiv:1808.04538* (2018).

[5] Keisuke Hagiwara, Yusuke Mukuta, and Tatsuya Harada. "End-to-End Learning Using Cycle Consistency for Image-to-Caption Transformations". In: *arXiv preprint arXiv:1903.10118* (2019).

[6] Tatsunori B Hashimoto, Hugh Zhang, and Percy Liang. "Unifying Human and Statistical Evaluation for Natural Language Generation". In: *arXiv preprint arXiv:1904.02792* (2019).

[7] Junhua Mao et al. "Generation and comprehension of unambiguous object descriptions". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 11–20.

[8] Brian McMahan and Matthew Stone. "A Bayesian model of grounded color semantics". In: *Transactions of the Association for Computational Linguistics* 3 (2015), pp. 103–115.

[9] Will Monroe, Noah Goodman, and Christopher Potts. "Learning to Generate Compositional Color Descriptions". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, Nov. 2016, pp. 2243–2248. DOI: 10.18653/v1/D16-1243. URL: https://www.aclweb.org/anthology/D16-1243.

[10] Will Monroe, Robert XD Hawkins, et al. "Colors in context: A pragmatic neural model for grounded language understanding". In: *Transactions of the Association for Computational Linguistics* 5 (2017), pp. 325–338.

[11] Randall Munroe. *Color Survey Results*. May 2010. URL: https://blog.xkcd.com/2010/05/03/color-survey-results/.

[12] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. "Cider: Consensus-based image description evaluation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575.

[13] Jun-Yan Zhu et al. "Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". In: *CoRR* abs/1703.10593 (2017). arXiv: 1703.10593. URL: http://arxiv.org/abs/1703.10593.