Comprehension-guided referring expressions

Ruotian Luo and Gregory Shakhnarovich
{rluo,gregory}ttic.edu

Background Referring expressions are a special case of image captions. Such expressions describe an object or region in the image, with the goal of identifying it uniquely. Thus, in contrast to generic captioning, referring expression generation has a well defined evaluation metric: the ability of the listener to derefence the expression. This implies two related tasks. One is the *comprehension* task (also called natural language object retrieval), namely localizing an object in an image given a referring expression. The other is the *generation* task: generating a discriminative referring expression for an object in an image. Most prior work addresses both tasks by building a sequence model, which can be used generatively for generation or discriminatively for comprehension.

We would like the generated expressions to be both intelligible/fluent and unambiguous to humans. Fluency can be encouraged by using the standard cross entropy loss with respect to human-generated expressions. On the other hand, we adopt a comprehension model as the "discriminator" which tells if the expression can be correctly dereferenced. Note that we can also regard the comprehension model as a "critic" of the "action" made by the generator where the "action" is each generated word. The two components (generator and discriminator) are not adversarial, but collaborative.

Method Our main contribution is the first attempt to integrate automatic referring expression generation with a discriminative comprehension model in a collaborative framework. There are two ways in which we do that. The **generate-and-rerank** method uses comprehension on the fly, similarly to (Andreas and Klein 2016), where they tried to produce unambiguous captions for clip-art images. The generation model generates some candidate expressions and passes them through the comprehension model. The final output expression is the one with highest generation-comprehension score (a combination of perplexity and the discriminative loss from the comprehension). In the **training by proxy** method, the generation and comprehension model are connected and the generation model is trained to lower discriminative comprehension loss (in addition to the cross-entropy loss). Compared to generate-and-rerank method, the training by proxy method doesn't require additional region proposals during test time.

Results We evaluated our methods on four data sets (RefClef, RefCOCO, RefCOCO+, RefCOCOg), and compared the results to those of recently proposed methods, such as (Yu et al. 2016).The training by proxy method achieves competitive results, while in all experiments the generate-and-rerank method is superior to all other methods.

We also performed human evaluation of the generated captions, on 100 images randomly chosen from each split of RefCOCO and RefCOCO+. Subjects clicked on the object which they thought was the most probable match for a generated expression. Each image/expression example was presented to two subjects, with a hit recorded only when both subjects clicked inside the correct region. See Table 1 for results.

Table 1: Human evaluations

	RefCOCO		RefCOCO+	
	Test A	Test B	TestA	TestB
MMI(Yu et al. 2016)	53%	61%	39%	35%
SMIXEC	62%	68%	**46%**	25%
Rerank	**66%**	**75%**	43%	**47%**

Note This work will appear in proceedings of CVPR 2017.

References

Andreas, J., and Klein, D. 2016. Reasoning About Pragmatics with Neural Listeners and Speakers. *1604.00562V1*.

Mao, J.; Huang, J.; Toshev, A.; Camburu, O.; Yuille, A.; and Murphy, K. 2016. Generation and Comprehension of Unambiguous Object Descriptions. *Cvpr* 11–20.

Yu, L.; Poirson, P.; Yang, S.; Berg, A. C.; and Berg, T. L. 2016. Modeling Context in Referring Expressions. In *Eccv*.

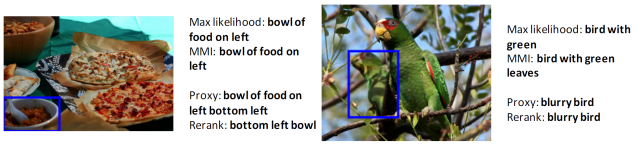

Figure 1: Each image: the top two expressions are generated by baseline models proposed in (Mao et al. 2016); the bottom two expressions are generated by our methods.