Describing like Humans: on Diversity in Image Captioning

**ABSTRACT**

Recently, the state-of-the-art models for image captioning have overtaken human performance based on the most popular metrics, such as BLEU, METEOR, ROUGE and CIDEr. Does this mean we have solved the task of image captioning? The above metrics only measure the similarity of the generated caption to the human annotations, which reflects its accuracy. However, an image contains many concepts and multiple levels of detail, and thus there is a variety of captions that express different concepts and details that might be interesting for different humans. Therefore only evaluating accuracy is not sufficient for measuring the performance of captioning models – the diversity of the generated captions should also be considered. In this paper, we proposed a new metric for measuring diversity of image captions, which is derived from latent semantic analysis and kernelized to use CIDEr similarity. We conduct extensive experiments to re-evaluate recent captioning models in the context of both diversity and accuracy. We find that there is still a large gap between the model and human performance in terms of both accuracy and diversity, and that models that have optimized accuracy (CIDEr) have low diversity. We also show that balancing the cross-entropy loss and CIDEr reward in reinforcement learning during training can effectively control the tradeoff between diversity and accuracy of the generated captions.

This paper will be presented in the IEEE Conference on Computer Vision and Pattern Recognition (CVPR2019), June 16-20, 2019, Long Beach, CA, America.

**All are welcome!**