Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering

Peter Anderson†, Xiaodong He‡, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould†, Lei Zhang

1Australian National University, 2Microsoft Research, 3University of Adelaide, 4Macquarie University, 5Transitioning to Georgia Tech, 6Now at JD AI Research

1. Visual attention
Visual attention mechanisms learn to focus on image regions that are relevant to the task, requiring:
1. Learned attention function (network), f
2. A set of attention candidates, V
3. Task context representation, h

2. Generation of attention candidates, V

3. Captioning and VQA models

4. Pre-training Faster R-CNN
We pre-train Faster R-CNN on Visual Genome data, using:
• 1600 object classes
• 400 attribute classes

To select k attention candidates, a detection confidence threshold is used

5. Quantitative results

6. Qualitative results

Code, models and pre-trained features available: http://www.panderson.me/up-down-attention

Refer also to our related work: Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge, Poster J21, Wednesday June 20, 10-10:12:00 Poster Session P2-1