Visual Attention Models of Object Counting
Jack Lindsey, Steven Jiang
Stanford University

Background
- Industrial feed-forward convolutional neural networks perform well on traditional image recognition tasks
- However, they’re computationally expensive and are unreliable when image inputs have imperfect resolution outside of the focus area for object recognition
- Need better network architecture; look to biological models of retinal motion for designs

Methods
- We train on a database of microscope cell images synthesized by SIMCEP and developed by Lehmussola et. al.
- Benefits for object counting include:
  - Cells are similar enough in appearance to train network architectures for counting
  - Cells vary in shape and color to prevent naïve approaches, like integrating non-background mass
- We use 5000 128 x 128 images, split into 4000 for training and 1000 for testing, and divided evenly among count classes
- To test effectiveness of glimpse network on basic cases, we:
  - Fig. 1. Sample input. Varying blurriness, sizes, and shape ensure RAM integrity. Replicates imperfection of human vision. We have three panels layered as input: the entire image blurred, a window with less blur, and a sub-window at full resolution to mimic retinal focus

Network Formulation
- We use a simple reward function to reinforcement learn, where y is our training prediction and c is the correct count class for each image over successive iterations
- True state of the environment remains unobserved—the retinal sensor can only focus on one area at a time
- Layer focus with 8 x 8 full resolution window and 32 x 32 \frac{1}{4} resolution layer around focus point, with rest of image at 1/16 resolution, repeat N = (1, ..., 7) times
- Convolutional on initial input and previous location, recurrent on updating parameters and creating next location and classification iteration

- Fig. 3. A visual representation of the RAM we build on—credits to Mnih et. al. for original architecture and diagram. A glimpse network—in—processes input images and previous location iterations into hidden layers that output the count classification and the next location iteration with current internal state

Final Results
- RAM soundly beats CNN—accuracy averages 65-70%, fixing optimal CNN parameters with heuristic estimate
- Experimental flaw—can only know location if predict correctly and vice versa, so accuracy occasionally remains constant at 20%

Acknowledgements
We thank Steven Hansen and Professor James McClelland, Department of Psychology, for their guidance and the inspiration behind this project.

Fig. 1. Sample input.
Varying blurriness, sizes, and shape ensure RAM integrity. Replicates imperfection of human vision. We have three panels layered as input: the entire image blurred, a window with less blur, and a sub-window at full resolution to mimic retinal focus.