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

- Extend **Recurrent Attention Model (RAM) of Mnih et. al.** to recognize and count objects in images with **reinforcement learning** and an attention mechanism[1]
- Produce **accurate counts for the number of objects** in an image **by scanning and layering blurriness** to simulate focus
- Use **glimpse network** to track locations for detecting objects Evaluate accuracy of generated retinal glimpses by measuring against object count accuracy in images

Methods

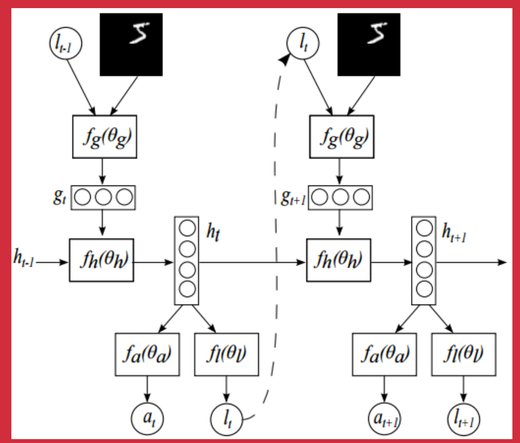
- We train on a database of microscrope cell images synthesized by **SIMCEP** and developed by Lehmussola et. al. [2] Benefits for object counting include:
- Cells are similar enough in appearance to train network architectures for counting
- Cells vary in shape and color to prevent naïve approaches, 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

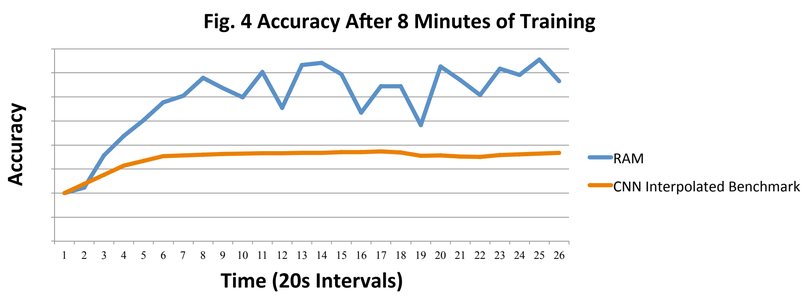
$$R = 1\{y = c\}$$

- 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 ¼ 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

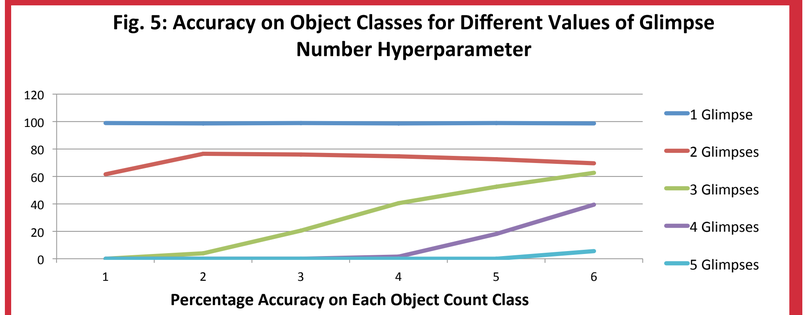


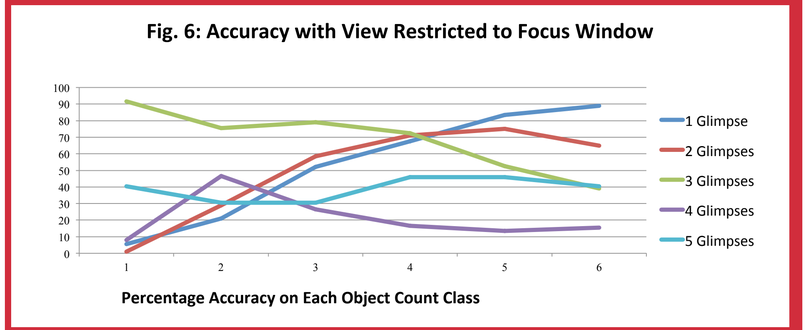
- Fig. 3. A visual representation of the RAM we build on—credits to Mnih et. al. for original architecture and diagram. A glimpse network—fg--processes input images and previous location iterations into hidden layers that output the count classification and the next location iteration with current internal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





- Experimental proof glimpse networks work (fig. 5)—low counts predicted accurately with few glimpses, and there is strong correlation b/w glimpse # and object count
- Control: when tested against input without access to entire field of vision (fig. 6), the glimpse network and model's predictive power fail to show correlation

Acknowledgements

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

We credit [4] for the code framework for the base RAM model and [1] for developing the original design. We credit [2] for the data we draw on for our problem [3] for the mathematical framework of the gradient function for our reward function.

1. *Volodymyr Mnih, Nicolas Heess, Alex Graves, Koray Kavukcuoglu*, Recurrent Models of Visual Attention,NIPS 2014.
2. *A. Lehmussola, P. Ruusuvuori, J. Selinummi, H. Huttunen, and O. Yli-Harja*. **Computational framework for simulating fluorescence microscope images with cell populations.** IEEE Trans. Med. Imaging, 26(7):1010–1016, 2007.
3. *Williams, Ronald J.*, Simple statistical gradient-following algorithms for connectionist reinforcement learning, Machine learning 8.3-4 (1992): 229-256.
4. *Leonard, Nicholas.* Recurrent Model of Visual Attention, Torch Documentation 2015