Visual Attention Models of Object Counting

Jack Lindsey, Stanford University, jacklindsey@stanford.edu
Steven Jiang, Stanford University, syjiang@stanford.edu

***Abstract*— We develop a sequential learning model using a recurrent neural network architecture and reinforcement learning to recognize and count objects in images. Simple feedforward neural networks perform well on this task when trained using backpropagation; however, convolutional neural networks are computationally expensive and results are less certain when the image input has imperfect resolution outside of the focus area (like input to the human visual system). We build an architecture based on visual attention models implemented in the literature, such as by Mnih et. al., applied to the task of counting objects in images. Our model scans an image for distinguishing features based on models of the retina, layering blurriness on the image to simulate focus on a particular area of the image. Our recurrent neural network sequentially incorporates the data from inputs to produce an accurate count of the objects in the entire image, and reinforcement learning dictates glimpse locations for each successive iterative step. We achieve accuracy of 66.7 percent, and we prove that the glimpse model tracks discrete object figures in images through the relationship between the number of glimpses required to yield accurate estimates and the number of objects present.**

I. INTRODUCTION AND RELATED WORK

Feedforward convolutional neural networks have achieved some success when applied to problems of image recognition and object discrimination. As it stands, convolutional neural networks have made impressive gains in traditional image recognition tasks, such as telling objects apart and classifying images based on familiar features from training data.

However, several problems persist with convolutional neural networks on more general image recognition tasks. Firstly, convolutional neural networks are computationally expensive to train, and they require large amounts of manually labeled data to train on, usually involving tens of thousands of manually labeled images and several GPU cores to train to convergence. However, models of human visual attention from neuroscience literature suggest that human eyes scan and glimpse throughout an image[5], fixating on significant features or objects that are distinguished from the background or other less significant features. Recurrent learning algorithms based on this type of visual attention offer potentially significant optimization above implementing vanilla recurrent neural networks. This biologically-inspired approach allows the algorithm to integrate information obtained from processing smaller and more salient focus points within a large image, ideally making the training and prediction processes more efficient.

Recurrent neural networks can be used to implement this attention model by iterating through the internal network states in a fashion parallel to image processing in the brain when humans initially encounter and recognize images. In particular, the problem of object counting has mostly been approached from the perspective of convolutional neural networks, and it presents a different set of challenges than traditional image recognition. We build upon previous recurrent attention model (RAM) efforts and apply the architecture, equipped with reinforcement learning, to object counting. We believe that sequentially incorporating data from multiple glimpses could outperform a feedforward approach in cases where objects are sparse or clustered, or otherwise unevenly distributed. Our goals are to compare the efficacy of two different approaches to the counting problem and analyze the performance of the attention model in more detail. In particular, we evaluate neural networks' abilities to closely replicate the pattern of retinal glimpses human eyes take upon encountering images with multiple objects.

II. METHODS

A. Data and General Design

Training and testing images were generated using SIMCEP [2], a publicly available tool that synthesizes realistic portrayals of arbitrary numbers of randomly placed biological cells. The tool was developed by Lehmussola et. al. based upon real datasets of fluorescent cell microscopy data. This particular dataset presents several advantages for the task of object counting. First, the cells produced are similar

*This paper is not sponsored by any organization.

enough in appearance to allow decent performance by a trained network, but they vary in shape enough to prevent the network from simply deploying naive methods to maximize the reward function, like integrating the total mass of non-background material in an image. Second, the images are vary significantly in color and resolution, well representing many of the core features of realistic images. We generated a total of five thousand 64 x 64 pixel images, one thousand in each count class for counts ranging from one to five. The first eight hundred in each class for used for training, and the rest for testing. The images in our dataset are all uniformly sized. We choose to represent the counting problem as a classification problem, where probabilities are assigned to each of five classes and a prediction is made by choosing the count class with the highest probability in the distribution. We train the network by comparing predictions with each image's true object count. We choose count classes from one to five to optimize our evaluation of the effectiveness of the glimpse model in the RAM, creating a clear distinction in how the glimpse network performs on images with few objects and images with a greater number of objects.

B. Convolutional Feedforward Network

First, we implement a feedforward network with convolution layers. The network has two spatial convolution layers with max pooling to reduce the dimensionality of the input, followed by three standard linear hidden layers with nonlinear activation functions. The network is representative of feedforward networks commonly deployed for many applications.

C. Recurrent Attention Model

Second, we train a model that sequentially chooses a predetermined number of glimpse locations within each image and uses the images resulting from these fixations as the sequential inputs to a recurrent neural network.

1) Simulating Visual Focus: We model visual focus by giving the network access to several small, concentric windows that progressively decrease in size and increase in resolution. Sequential inputs take the form of 3 different 64 x 64 images layered on each other (see fig. 1, where these 3 images are placed side by side for comparison). The first layer is a 4 x 4 window around the glimpse location. The second layer is a larger 16 x 16 window around the glimpse location, distorted with 1/4 resolution. This simulates the reduced quality of vision in the human retina away from a fixation point. The last layer is the full image blurred to 1/16 resolution. These three layers are combined into three 3 x N x N inputs to the RAM, where N is the patch size (each pixel has dimension 3, corresponding to RGB values).

The area of high focus is most helpful for counting, while the low-resolution windows allow the program to only make judgments about where to place the focus in the next time step of the network processing. This replicates the parallel processing of input to the retina while discerning and counting objects in a scene, being vaguely aware of objects in the background while sharply focused on discrete objects in a given location.

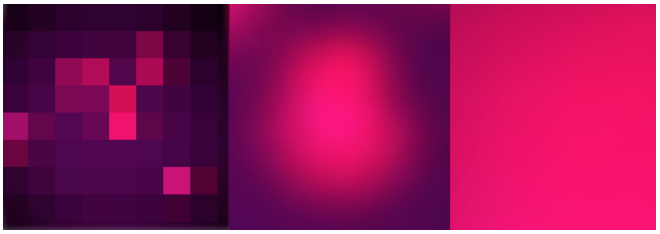

Fig. 1. A representation of the input to the recurrent network. Here, three windows of decreasing size and increasing resolution are shown.

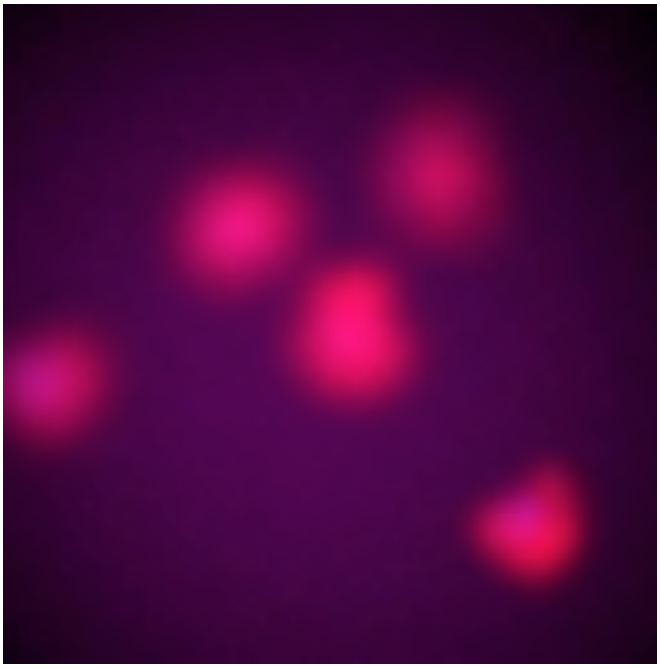

Fig. 2. The original image from which the above windows were obtained.

2) Controlling Focus Trajectory: We implement an algorithm that can choose along what path the algorithm "glimpses" to recognize discrete objects within the attention window. Previous approaches have included learning policies using reinforcement learning to plot glimpse paths [1]. Our network uses essentially the same approach, learning a policy for glimpse locations based on a simple reward function for a given image, described by

$$R = 1\{y^{(t)} = c\}$$

where $y^{(t)}$ is the prediction for an image on iteration t, and c is the true object count for an image. The network outputs both the next location based on the glimpse network and a guess to the object count at every iteration of the network. The new location is used in the next iteration to produce the new input to the network.

3) Integrating Information from Sequential Perspectives: We choose the recurrent neural network as the base architecture because of its ability to accept arbitrarily long sequences of input. Traditional feedforward networks lack this capability. In our training model, we input the images in certain sequences. The neural network accepts these sequences and incorporates them into a global visual understanding of the image from multiple glimpses rather than taking in all the data at once.

D. Unifying the Model

The aforementioned considerations - deciding where to focus and how to integrate the information obtained from these focus points - are integrated into a single network architecture. The model has two central components: the processing component that mixes the input image sequence and the glimpse location, and the recurrent component. The recurrent component mixes the image input and the internal network representation at each iteration, updating the internal network representation at each iteration. The network sequentially integrates the information from each iteration into a final prediction for the object count of an image.

III. Performance

A. Convolutional Feedforward Network

Preliminary testing yielded a 95.4 percent accuracy rate after eight minutes of stochastic gradient descent

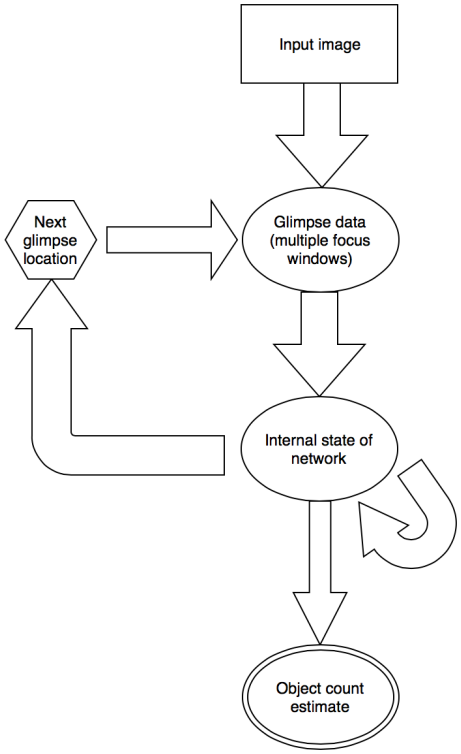

Fig. 3. A simplified depiction of the recurrent model used. At each time step the network outputs an estimate of the object count and an (x,y) coordinate pair specifying the next glimpse location, which is used along with the input image to produce the series of focus windows described elsewhere in this paper.

reinforcement training on the feedforward network described above. It should be noted that only limited time was spend optimizing the parameters of this model.

B. Attention Model

After training the attention model, using three concentric resolution windows and seven total glimpses, for eight minutes, and obtained a final accuracy on the test set of 66.6 percent. This value is a reflection of numerous successful trials. However, it should be noted that on some trials, the recurrent model's accuracy remained flat at 20 percent (no better than random). This is a result of the RAM glimpsing at locations from which it can glean no substantial information.

C. Comparison of Models

IV. Evaluating the Focus Mechanism

An issue that has received relatively little treatment in the literature is the extent to which the recurrent attention model successfully learns to focus on the most salient areas of the image. Our decision to apply the model to object counting arose in part because this

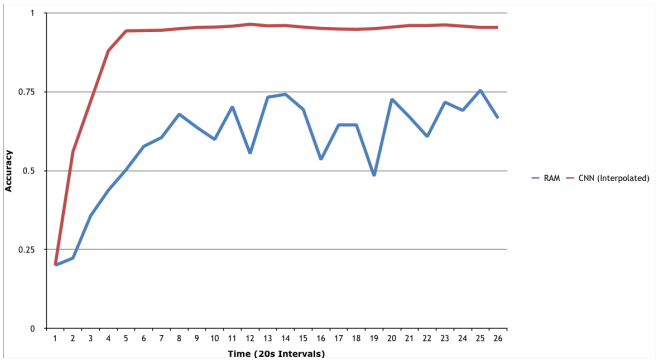

Fig. 4. Comparison of performance of the feedforward network and attention model, controlling for training time.

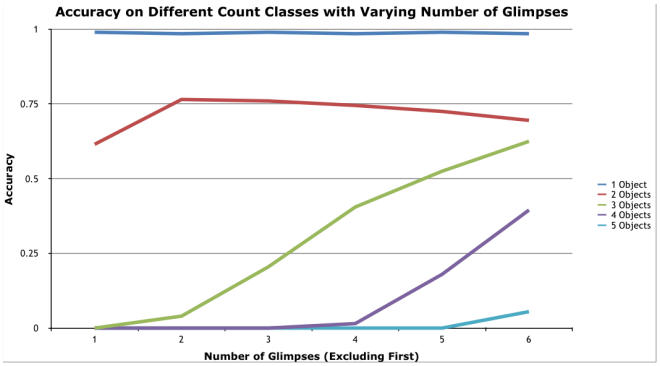

Fig. 5. Demonstration of the efficacy of the glimpse location decision mechanism. Lower count images after only a few glimpses, while higher-count images could not, indicating that glimpses were indeed targeted at the relevant areas of the image (i.e. the locations of the objects) after sufficient training

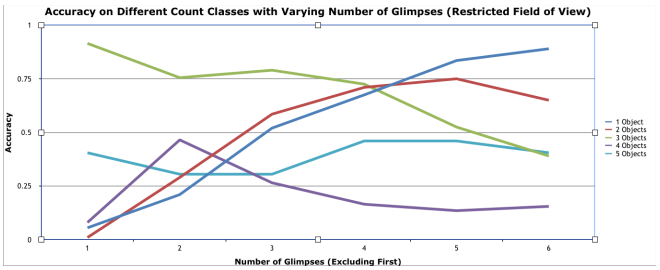

Fig. 6. Repeated version of the previous experiment, but with the model's field of view restricted entirely to the focus window. The effect observed in the original version was not replicated here, indicated that the low-resolution peripheral image data was indeed enabling proper functioning of the attention mechanism

task lends itself well to rigorous evaluation of the algorithm's decisions about where to focus. We tested this behavior as follows. First, we trained the network to estimate counts after seven glimpses. During the testing phase, however, we limited it to N glimpses, with N ranging from 1 to 7, inclusive, as a hyperparameter for the network. In doing so, we could analyze the model's intermediate estimates of the object count after only a few glimpses and evaluate how each additional glimpse affects predictive power for each count class.

After training and evaluation, the network produces the clear result that images with fewer objects (e.g. one or two) were correctly classified even when the model was restricted to fewer glimpses. Images with more objects required more glimpses to be classified accurately. There exists a close correlation between the number of glimpses a network is allowed and the number of objects that it must count in the image. This suggests that the network succeeds in choosing glimpse locations that closely correspond to an object in the image.

As a control to demonstrate that this effect was in fact do to intelligent glimpse decisions, we repeated this test, but restricting the input to the network to only the smallest, highest-resolution window. In other words, we restricted the model's field of view, removing its access to low-resolution information about the entire image. In this implementation, performance even on low-count images improved with more glimpses, indicating that the glimpse locations were being decided more or less randomly and therefore the program benefited from being allowed as many as possible. This result indicates that the model's access to a low-resolution version of parts of the image outside its focus window was allowing it to choose salient points to focus on.

V. DISCUSSION

The RAM was clearly outperformed by a more traditional convolutional neural network when trained for the same amount of time, and it was unclear whether the performance of the recurrent network would have improved further given additional training time. That said, we have reasons to be optimistic about the recurrent attention model. First, allowing the model additional glimpses would likely, given the trends we observed, result in significantly improved accuracy. Second, other parameter adjustments (e.g. the number or size of focus windows, or other aspects of the network architecture) could optimize the algorithm further – such models have not received nearly as much study as feedforward networks, and thus parameter optimization has not been explored as fully.

The RAM has advantages that make it worth studying further. Once trained, it can classify images more rapidly since it processes less data. Furthermore, it more accurately models human perception, making it a

valuable model for cognitive scientists. Such a model has applications outside of object counting, as the concept of integrating a series of incomplete snapshots could in principle be applied to processing of text, speech, or a number of other tasks.

Even with our current model, however, we successfully demonstrated the viability of a recurrent model for object counting. Most significantly, we have shown the effectiveness of the attention mechanism, proving that the success of these attention models is in fact due to correct glimpse behavior, not simply good predictive capacity given limited information. We believe this is an important result, as it allows one to aim for symbiotic training of the attention mechanism and prediction mechanism. Though the interrelated nature of these two parts of the algorithm gives the model its power, it also has pitfalls. We believe that the occasional no-better-than-random accuracy of the model after training was due to a chicken-and-egg effect, where the model could not learn how to use glimpse information before learning how to glimpse properly, and vice versa. Introducing some stochasticity into the process could alleviate this issue, but more study is needed.

VI. Directions for Future Research

We identified the largest source of inconsistency in the network's output as dependent upon the choice of glimpse hyperparameter. The current architecture of the network depends upon a preset fixed number of glimpses for the network to take before it stops trying to improve predictive accuracy. Our specific improvement in the model is to implement a network that trains itself on how many glimpses to take for each image.

First, we will insert a mechanism that if the RAM glimpses at a location with no information, then it randomly glimpses somewhere else and restarts the iterative process.

Second, we would thus modify the reward function in the reinforcement learning to not only reward the network for correctly identifying object counts but also continuing glimpsing until the network is able to, within a threshold of accuracy, identify this correct object count. This reward function would take the gradient of the probability given to the correct count class as the network keeps glimpsing and stop the network from further glimpsing once the gradient falls below a certain threshold, meaning that the network has become confident in the correct count for an image. This more closely replicates the motion of the human retina[7]– the human eye does not stop glimpsing around a scene after a fixed number of glimpses, but rather, it keeps on scanning an image for prominent locations until it is confident in the information it has collected about a scene.

Third, we would apply our network to non-image data to demonstrate the versatility of a recurrent attention approach to other sensory input, such as (potentially) sound, tactile input, natural language understanding and translation.

VII. Acknowledgments

We would like to thank Professor James McClelland and Steven Hansen in Stanford's Department of Psychology for the initial inspiration behind the research and their tremendous support and guidance in giving us the resources and references for the work. We also are grateful to Nicholas Leonard for providing open source code that aided us in implementing the recurrent attention architecture.

References

[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 frameworkfor simulating uorescence microscope images with cell populations. IEEE Trans. Med. Imaging, 26(7):10101016, 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.

[5] *Bogdan Alexe, Nicolas Heess, Yee Whye Teh, and Vittorio Ferrari*. Searching for objects driven by context. In NIPS, 2012.

[6] *Misha Denil, Loris Bazzani, Hugo Larochelle, and Nando de Freitas.* Learning where to attend with deep architectures for image tracking. Neural Computation, 24(8):21512184, 2012.

[7] *Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson.* Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev, pages 766786, 2006.

[8] *Stefan Mathe and Cristian Sminchisescu*. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In NIPS, 2013.