Project Proposal: Training Recurrent Attention Models on Handwritten word Images

Name: Yadnyesh Patil, Roll No. : 193050067
Name: Rohit, Roll No. : 144054003
Name: Gaurav, Roll No : 19V980006
Name: Shivam Dixit, Roll No : 193050012

Humans focus on different parts of images to recognize the objects and read the text. Processing the complete image for recognition through Convolutional Neural Networks (CNNs) involves operations at undesirable parts of image which can be avoided using attention models [1]. Recurrent Attention Models have been used for reading MNIST digit images [2]. The partial information is made available by applying different bandwidth limited glimpse sensors (for multiple resolutions focussed at the same location) on the image (refer left and top right part of Figure 2). The RAM model (refer Figure 1) acts as an agent that learns to attend the relevant parts in image, obtained via glimpses, sequentially through a Recurrent Neural Network (RNN) to read a digit. The overall problem is posed as partially observed markov decision problem (POMDP) [2].

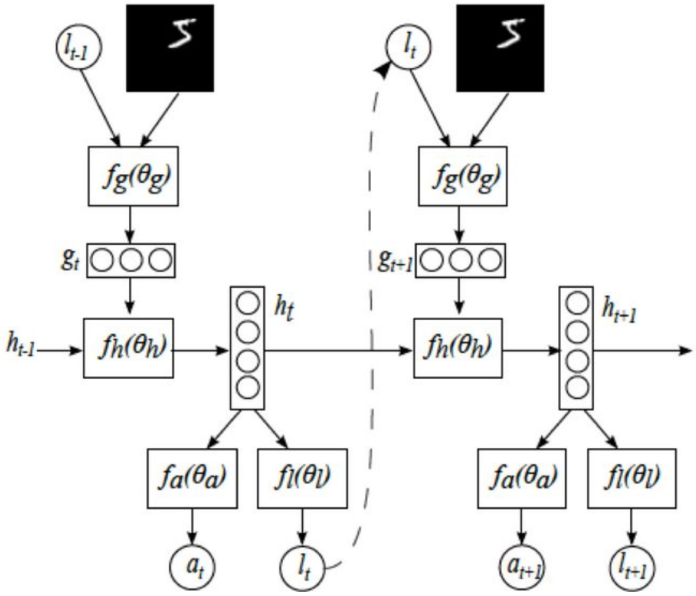

Figure 1: RAM Model

The training as well as evaluation starts with a glimpse at a random location in the image. The RNN then 1) sequentially regresses the location “l_t” for the next glimpse in the image via location network trained via RL techniques (e.g. Policy gradient methods [3], as Q-Learning is not possible for large space of states and actions i.e. image glimpses and locations) and 2) classifies the sequence of (fixed number of) glimpses as the character “a_t” through action network trained via cross entropy loss at the end of sequence.

Policy gradient methods learn policy parameter based on the gradient of some scalar performance measure J with respect to the policy parameter. These methods seek to maximize performance, so their updates approximate gradient ascent in J. If we have a score function f(x) that can tell us how much reward we’re going to have, then for a point x in order to have higher score in next iteration we have to look for the gradient of score function which is approximated as:

$$\nabla_\theta E_x [f(x)] = E_x [f(x) \nabla_\theta log p(x)]$$

which is the same as the direction of maximizing loglikelihood of sample point x, with magnitude of f(x) itself.

We propose the following series of experiments for our project:

Experiment 0: Single fisheye (like) glimpse for multiple-resolutions.

The present work (on RAM [2]) exploits multiple glimpses each consisting of ‘k’ different patches (of different resolutions) around the common center (shown in different colors in left of Figure 2). The larger patches are then resized to the size of smallest to save compute (Refer top right of Figure 2). This includes redundancy because the central part of the glimpse is present in all the patches with different resolutions. As an initial contribution, we would like to make it a single variable resolution patch where resolution decreases as we move away from the center of the glimpse. The overall result would be a fisheye (like) glimpse for multiple-resolutions as shown in Figure 2.

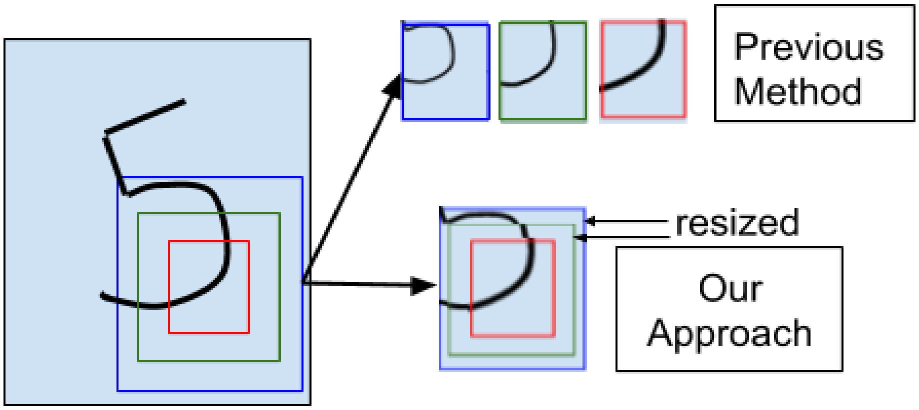

Figure 2: Our Approach: Single fisheye (like) glimpse (bottom right) for multiple-resolutions in original image (left). Method used in RAM [2] (top right).

Experiment 1: Reading MNIST images with variable number of glimpses.

The present work explores constant number of glimpses (best results for 8 glimpses) to recognize a digit. Intuitively lesser number of glimpses are required to read certain digits as compared to others. So we propose to train the RAM Model based on different RL algorithms and reward schemes in order to let the agent decide the number of glimpses it needs to read the digit, which can vary for different digits or even for different examples of the same digit.

Experiment 2: Reading EMNIST images.

We would also like to work on EMNIST dataset [4], which contain handwritten characters as well as digits as an intermediate step of our work.

Experiment 3: Reading word images.

Finally, we would like to read word images with reward schemes that penalizes the delay in reading as well as reinforces behaviour for reading with minimum number of glimpses.

References:

[1] Denil, M., Bazzani, L., Larochelle, H. and de Freitas, N., 2012. Learning where to attend with deep architectures for image tracking. *Neural computation*, *24*(8), pp.2151-2184.
[2] Mnih, V., Heess, N. and Graves, A., 2014. Recurrent models of visual attention. In *Advances in neural information processing systems* (pp. 2204-2212).
[3] Sutton, R.S., McAllester, D.A., Singh, S.P. and Mansour, Y., 2000. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems* (pp. 1057-1063).
[4] Cohen, G., Afshar, S., Tapson, J. and van Schaik, A., 2017. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373*.