Recurrent Attention Model on Image Classification

MIAO, Lizhang (SID: 20296174), NG, Ka Chung (SID: 20097188), LO, Yi-su (SID: 20399988)

The Hong Kong University of Science and Technology

Motivation

It is fascinating and heuristic that human vision does not process a whole scene entirely at once. Instead, human eyes selectively focus on parts of an image, and then the brain combine information from several fixations over time. Is it possible we can use this idea to boost the performance of neural network models, which attempt to mimic human brains, on image classification tasks?

In this project, we try to follow the method and the experiment design of the paper below. An approach of **Recurrent Attention Model** (RAM) based on the idea above is proposed and examined.

Mnih, Volodymyr, Nicolas Heess, and Alex Graves. "Recurrent models of visual attention." Advances in neural information processing systems. 2014.

Method

At each time step t, the steps below are carried out.

1. **Glimpse sensor** ρ: For a given location l_{t-1} and input image x_t, the band-limited sensor extracts 6 square patches centered at l_{t-1} from x_t to form a retina-like representation $\rho(x_t, l_{t-1})$. The first patch ('retina') is 8×8, and the successive patches have twice the width of the previous.
2. **Glimpse network** f_g: Through two fully-connected layers

$$\text{Linear}(x) = Wx_t + b \quad \text{and} \quad \text{Rect}(x) = \max\{x, 0\},$$

the output is

$$g_t = f_g(x_t, l_{t-1}) = \text{Rect}\left[\,\text{Linear}(h_g) + \text{Linear}(h_l)\,\right],$$

where

$$h_g = \text{Rect}\left[\,\text{Linear}(\rho(x_t, l_{t-1}))\,\right] \quad \text{and} \quad h_l = \text{Rect}\left[\,\text{Linear}(l_{t-1}))\,\right].$$

3. **Core network** f_h: The output is $h_t = f_h(h_{t-1}, g_t) = \text{Rect}\left[\,\text{Linear}(h_{t-1}) + \text{Linear}(g_t)\,\right]$.
4. **Location network** f_l: The output is $l_t = f_l(h_t) = \text{Linear}(h_t)$.
5. **Action network** f_a: Make the classification decision only at the last time step by

$$f_a(h_t) = \exp\left(\,\text{Linear}(h_t)/Z\,\right),$$

where Z is a normalizing constant. The reward at the last time step is 1 if the classification result is correct; otherwise, it is 0.

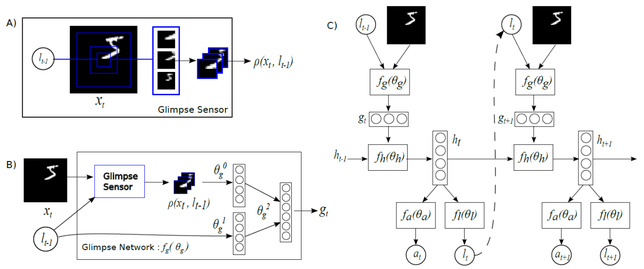

Figure: A flowchart illustrating the RAM training in the paper.

Experiments

In our experiments, we evaluate the effectiveness of RAM by classifying the following two varieties of MNIST. MNIST is a dataset of images of handwritten digits which contains a training set of 50,000 images. The size of every image is 20×20 pixels.

1. **Translated MNIST**: The image of every digit is placed in a random location of a lager blank image of 60×60 pixels.
2. **Cluttered Translated MNIST**: The image of every digit is firstly translated in the same way as above, and then a randomly selected part of the larger image is replaced with a 8×8 patch cut from the other digit image.

We consider both tasks above more challenging than the classification on original MNIST, especially the second one.

Results of Experiment 1

The following figures present the experiment results on Translated MNIST.

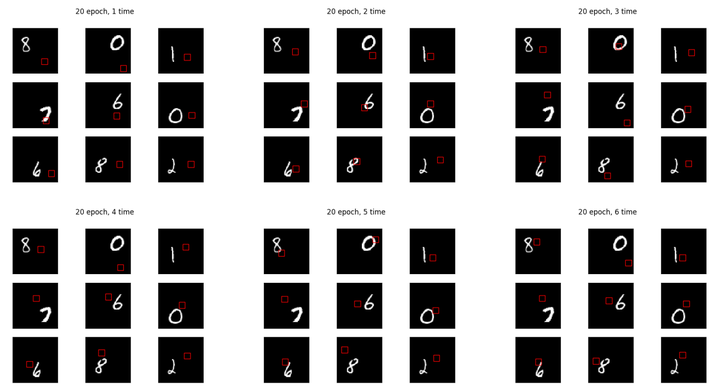

Figure: Some examples in the experiment on Translated MNIST. The red square on every image is the 'retina' chosen by the model.

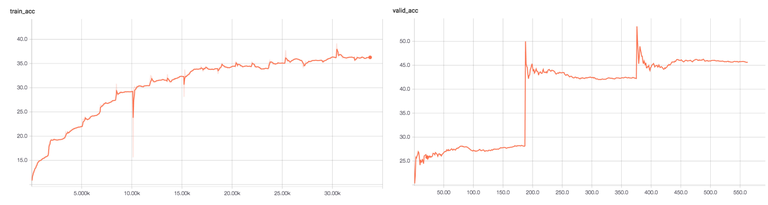

Figure: The accurate rates of classification in training (left) and validation (right).

Results of Experiment 2

The following figures present the experiment results on Cluttered Translated MNIST.

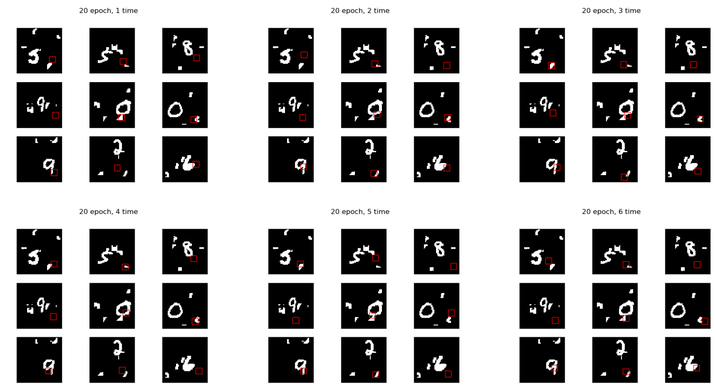

Figure: Some examples in the experiment on Cluttered Translated MNIST. The red square on every image is the 'retina' chosen by the model.

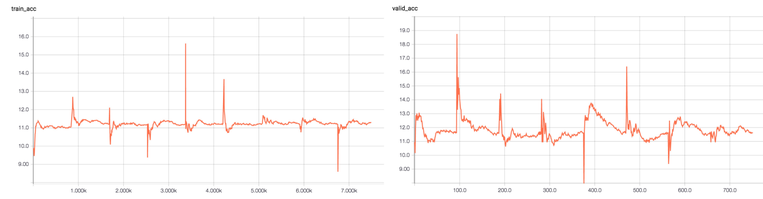

Figure: The accurate rates of classification in training (left) and validation (right).

Conclusion

- Unfortunately, the accurate rates of both experiments are much worst than the results of the paper, but they are still better than wild guess (in which the accuracy rate should be near 1/10). It is probably because our model is not fine-tuned with the number of time steps (glimpses), patches sizes, or the other parameters. Or, we can further follow the authors' suggestion that adding one more convolutional layer could be very helpful.
- The classification on Cluttered Translated MNIST is obviously more difficult than the Translated MNIST.
- The code for two experiments can be downloaded at `http://bit.ly/2xBOg0N`.