Recurrent Models of Visual Attention: A Case Study on MNIST Dataset

Beijing Fang
Department of Civil and Environmental Engineering
bang@connect.ust.hk

Huangshi Tian
Department of Computer Science and Engineering
htianaa@connect.ust.hk

Yunfei Yang
Department of Mathematics
yunfei.yang@connect.ust.hk

1 Introduction

In spite of the excellent recognition accuracy of neural network-based architectures in image classification and object detection tasks, the computation costs of them are always high in both training and testing time, especially for the large images with high resolution. Unlike the traditional neural networks, human eyes do not tend to process the whole image at once. Instead,human only focus on the selective parts and get the important information if needed, then combine the information from different parts over time to build the representation of the scene for perception.This idea can be applied in the computer vision,such as image classification, by using recurrent neural networks which focus on the most informative parts of the image and followed by reinforcement learning with rewards toward minimizing misclassification errors.

2 Model

The model we applied for image classification is the Recurrent Attention Model(RAM). The model structure is shown in Figure 1. The general idea of RAM is that the agent can partially observe the environment via a bandwidth-limited sensor, and actively control how to deploy its sensor resources, such as choose sensor location. Then the agent integrates the information over time to find the most effective way to act and deploy its sensor. At each step, the agent receives a scalar reward and the goal of the agent is to maximize the total sum of such rewards [2].

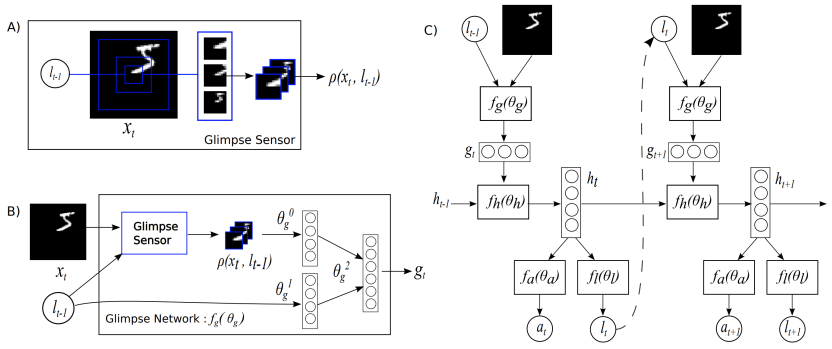

Figure 1: Diagram of RAM [2]

30th Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain.

- Glimpse Sensor: Glimpse Sensor takes a full-sized image and a location, outputs the retina-like representation $\rho(x_t, l_t - 1)$ the image x_t around the given location $l_t - 1$ which contains multiple resolution patches.
- Glimpse Network: Glimpse Network takes as the inputs the retina representation$\rho(x_t, l_t - 1)$ and glimpse location $l_t - 1$,and maps them into a hidden space using independent linear layers parameterized by θ_g^0 and θ_g^1 respectively using rectified units followed by another linear layer θ_g^2 to combine the information from both components. Finally it outputs a glimpse representation g_t.
- Recurrent Neural Network (RNN): Overall, the model is an RNN. The core network takes as the input the glimpse representation g_t at each step and history internal state $h_t - 1$ then outputs a transition to a new state h_t, which is then mapped to action a_t by an action network $f(a)\theta(a)$ and a new location l_t by a location network $f(l)\theta(t)$. The location is to give an attention at next step, while the action, for image classification, gives a prediction based on current information. The prediction result, then, is used to generate the reward point, which is used to train these networks using Reinforcement Learning.
- Loss Function: The reward could be based on classification accuracy (e.g. r_t = 1 for correct classification and r_t = 0 otherwise). In reinforcement learning, the loss can be finite-sum reward or discounted infinite reward. Cross-entropy loss for prediction at each time step is an alternative choice. It's also interested to figure out the difference function between these two loss and whether it's a good idea to use their combination.

3 Experiments

3.1 Dataset

We evaluated the Recurrent Attention Model on several image classification tasks. Two datasets are used in the experiments: Translated-MNIST and Cluttered-MNIST. Both datasets are based on MNIST [1] images, each generated with a certain type of transformation on the original images. For Translated-MNIST, we first paded the image from 28x28 to 60x60 with black background, and then translated the digit along both x- and y-axis with a randomly chosen amount. Figure 2 shows several sample digits from this dataset. For Cluttered-MNIST, we also performed the same aforementioned padding and translation. Following that, we randomly clipped four 8x8 patches from other images and place them on the image as 'clutters'. Four sample images from the dataset are displayed in Figure 3.

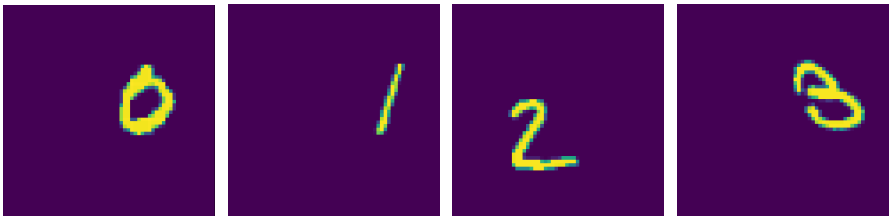

Figure 2: Sample digits in Translated-MNIST.

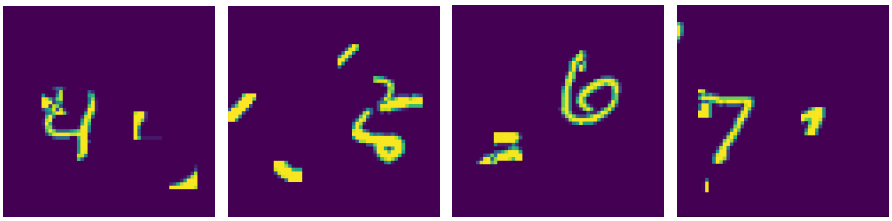

Figure 3: Sample digits in Cluttered-MNIST.

3.2 Training

In the experiments, the core RNN network has hidden size 256 and the Gaussian policy standard deviation is set to 0.17. Since our tasks are image classification, the reward at the last time step is 1 if the agent classified correctly and 0 otherwise. The rewards for all other timesteps are 0. There are three main hyper-parameters of RAM: the number of glimpse, the retina size and the scale number in retina. To study their effect on the RAM, we trained several RAMs with different hyper-parameters. All RAMs were trained using Adam optimizer with learning rate 0.0003 with minibatches of size 100. We trained each RAM 100 epochs and recorded the model with highest validation accuracy, which was used to test the capacity of RAM.

To demonstrate the capability of RAM, we also trained a CNN on Translated-MNIST dataset. The network constitutes two convolution layers and two fully connected layers. The convolution is followed by ReLU as activation and we also performed dropout to the output of the second convolution layer.

3.3 Result

Table 1 presents the results of our experiments. The highest classification accuracy for translated MNIST and cluttered MNIST achieved by our RAMs are 96.87% and 93.01% respectively, which are compatible with CNN. However, the accuracies of most RAMs are lower than the CNN baseline, which is different from the results in the original paper. Besides, our results also show that increasing the number of glimpse and scale do not increase the test accuracy much, but increasing the retina size has a significant impact on the performance of RAM.

Table 1: Evaluation on RAM

Dataset	Model	Accuracy	Epoch
Translated	CNN	96.66%	96
Translated	RAM, 6 glimpse, 8×8, 2 scale	79.98%	57
Translated	RAM, 6 glimpse, 8×8, 3 scale	81.22%	96
Translated	RAM, 6 glimpse, 12×12, 3 scale	94.35%	92
Translated	RAM, 8 glimpse, 12×12, 3 scale	92.52%	63
Translated	RAM, 6 glimpse, 16×16, 3 scale	96.87%	69
Cluttered	CNN	93.50%	96
Cluttered	RAM, 6 glimpse, 12×12, 3 scale	85.65%	72
Cluttered	RAM, 6 glimpse, 12×12, 4 scale	85.11%	52
Cluttered	RAM, 8 glimpse, 12×12, 3 scale	86.07%	67
Cluttered	RAM, 6 glimpse, 16×16, 3 scale	91.26%	95
Cluttered	RAM, 6 glimpse, 20×20, 3 scale	93.01%	96

In Figure 4, we also visualize several glimpses from translated MNIST task to view their locations in detail. The locations of glimpses clearly show that most glimpses focus on the most informative parts of the translated images which mostly around the objects, instead of focusing on the center or empty parts.

4 Discussion

After conducting the experiments based on the novel visual attention model, we do find some appealing properties of it. Firstly, the number of parameters and the amount of computation can be controlled independently from the size of input image. Secondly, the performance of RAM is comparable or even better than CNN. Besides, we also find some interesting phenomenon from the experiment results, and some are slightly different form the original paper as being discussed as follows:

- The impact of hyper-parameters on model performance: As the result shown in table 1, among the three hyper-parameters: the number of glimpse, the size of glimpse and the number of scales, the size of glimpse has the highest impact on RAM performance. That is,

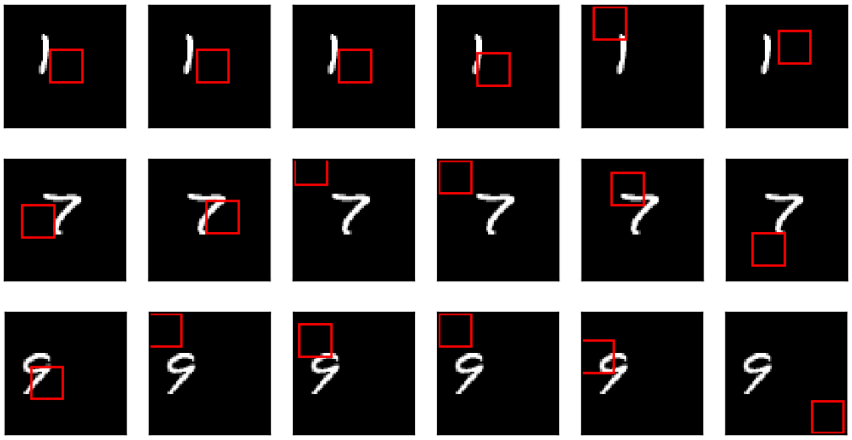

Figure 4: Sample glimpses on digits in Translated-MNIST.Red rectangles denote the location and region of glimpses.

test accuracy increased mostly with the increasing size of glimpse. In both translated MNIST dataset and Clutterd-Translated MNIST dataset, the increase of the number of scale hardly change the test accuracy. It indicates that the RAM networks builded up in this project are insensitive to this parameter, as well as the number of glimpse. Specifically, When the number of glimpse in translated MNIST task increased from 6 to 8,the test accuracy unexpectedly decreased for about 2 percent. The possible explanation for this phenomenon is that, although the number of glimpse increased, the selected glimpse could't always find the most informative parts of the images (as shown in Figure 4). Some may even selected the clutterd parts, inducing additional uncertainties. However, with the increasing size of glimpse,the probability of getting more useful information increased.

- The difference of model performance on different MNIST datasets: Apparently, the translated MNIST task is much easier than the cluttered MNIST task, since the latter induces the additional disturbance. Similar to the original paper, the model performance on clutter MNIST dataset is relatively poorer than the translated MINIST dataset under all different associations of hyper-parameters. However, it should be noticed that, with the increase of the size of glimpse, the gap of test accuracy between these two tasks narrowed. It suggests that, with the fine-tuned parameters, the model performance in cluttered MNIST task can be largely improved and become close to translates MNIST task.
- The comparison between RAM and CNN: Unlike the original paper, the test accuracies of most RAMs are lower than the CNN baseline, except for the one with relatively large size of glimpse. Considering of the time-consuming training and turning process with the time limitation to conduct the experiments, the parameters we tuned may not the optimal parameters as in the original paper. In spite of all these difficulties, the best RAMs still comparable to or even outperformed the CNN, it reveals potential ability of RAM in image classification tasks.

Contribution

In this project, we joint discussed the overall plan and analyzed the results.Individually, Huangshi Tian builded up the RAM and CNN models, Yunfei Yang turned the models, and Beijing Fang organized the report.

References

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. "Gradient-based learning applied to document recognition." Proceedings of the IEEE, 86(11):2278-2324, November 1998.

[2] Mnih V , Heess N , Graves A , et al. Recurrent Models of Visual Attention[J]. 2014.