MATH 6380p Final Project: Reinforcement Learning for Image Classification with Recurrent Attention Models

Dong Qian and Kejing Yin
Department of Computer Science, Hong Kong Baptist University
{dongqian, cskjyin}@comp.hkbu.edu.hk

1. Introduction

- Convolutional neural networks have advanced the state-of-the-art performance on computer vision tasks.
- The excellent performance comes at a higher computational cost.
- The attention module in the human visual system naturally allows for acquiring relevant information from different fixations to build up an internal representation of the scene.
- Recurrent visual attention model uses a recurrent neural network as its core network which processes inputs sequentially, attending to different locations at a time, and incrementally updates current information based on past information.

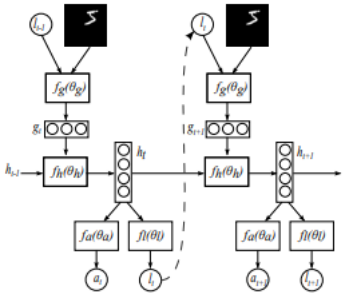


Fig. 1: The core network $f_h(\theta_h)$ is a RNN, which produces a dynamic state h_t through the glimpse representation g_t and previous internal state h_{t-1}. The location network $f_l(\theta_l)$ and action network $f_a(\theta_a)$ use h_t to produce the next location l_t to attend to and the action a_t, respectively.

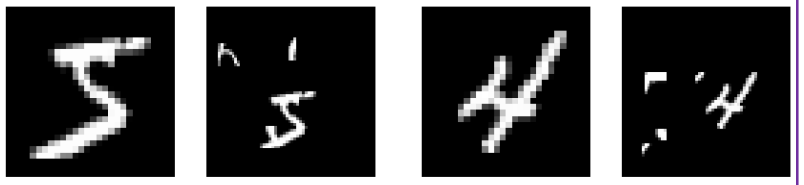
2. Datasets

MNIST

- 28 x 28 hand-written digits.

Cluttered MNIST

- Generated based on the original MNIST dataset.
- Enlarged to 60x60 with additional noise and translation added.
- Four pieces of clutters added to each data sample.
- Two examples: (left: MNIST, right: cluttered translated MNIST)



3. Experiments

- Conduct experiments of **Image Classification** task.
- Model parameters are determined by the highest accuracy score on validation set.
- We evaluate the performance by accuracy score, and report the results on the held-out testing set.

Baselines

- **FC:** Fully connected neural networks with two hidden layers, each with 256 hidden units of ReLU activation function.
- **ConvLayer:** One convolutional layer followed by one fully connected layer with 256 hidden units of ReLU activation function.
- **ConvNet:** CNN model with two convolutional layers, each followed by a max-pooling layer. Two fully connected layers are used for prediction.
- Dropout with p=0.5 is added before the final fully connected layer.

4. Results

(a) 28x28 MNIST dataset

Model	Accuracy
Baseline-FC	**98.37%**
RAM, 2 glimpses, 8x8, 1 scale	97.10%
RAM, 3 glimpses, 8x8, 1 scale	96.49%
RAM, 4 glimpses, 8x8, 1 scale	97.15%
RAM, 5 glimpses, 8x8, 1 scale	96.77%
RAM, 6 glimpses, 8x8, 1 scale	98.32%

(b) 60x60 cluttered translated MNIST

Model	Accuracy
Baseline-FC	68.37%
Baseline-ConvLayer	83.35%
Baseline-ConvNet	**92.07%**
RAM, 4 glimpses, 12x12, 3 scales	85.54%
RAM, 5 glimpses, 12x12, 3 scales	88.12%
RAM, 6 glimpses, 12x12, 3 scales	70.02%

- Increasing the number of glimpses almost improves the performance of RAM on the MNIST dataset until it reaches its maximum with 6 glimpses.
- The RAM model is more robust than the baselines. RAM with 5 glimpses reaches 88.12% accuracy while the FC baseline drops to only 68.37%.
- Fine-tuning hyperparameters of RAM is a bit difficult due to many different components.

5. Visualization

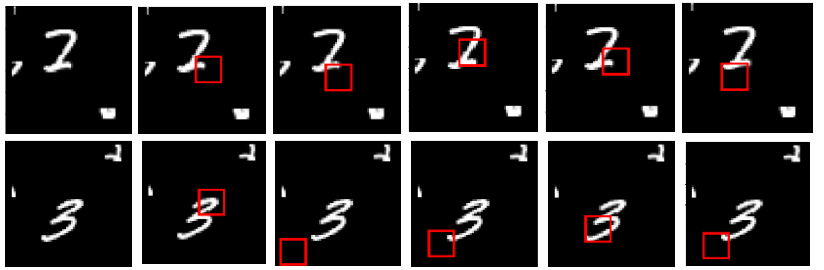


Fig. 2: Visualization of the glimpses

The five glimpses are chosen by the network. The rectangle in red shows the full resolution glimpse, avoiding the computation in empty or cluttered parts of the input images and directly focusing on the areas of interest. But some of them fail.

6. Conclusion and Discussion

- The experiments show that RAM is comparable to a convolutional architecture on the original and cluttered MNIST dataset.
- The attention mechanism makes it easier to learn representations of the relevant parts of the input images and ignore the cluttered parts.
- Both the number of model parameters and the amount of computational cost could be controlled independently of the size of the input images.
- Policy gradient-based reinforcement learning takes too long to convergence.
- The core network can be augmented with Long Short-term Memory (LSTM) that alleviates gradient vanishing issue of the RNN for learning long-term dependencies.

7. References

Volodymyr Mnih, Nicolas Heess, Alex Graves and Koray Kavukcuoglu. "Recurrent Models of Visual Attention." *Advances in Neural Information Processing Systems*, pp: 2204-2212, 2014.