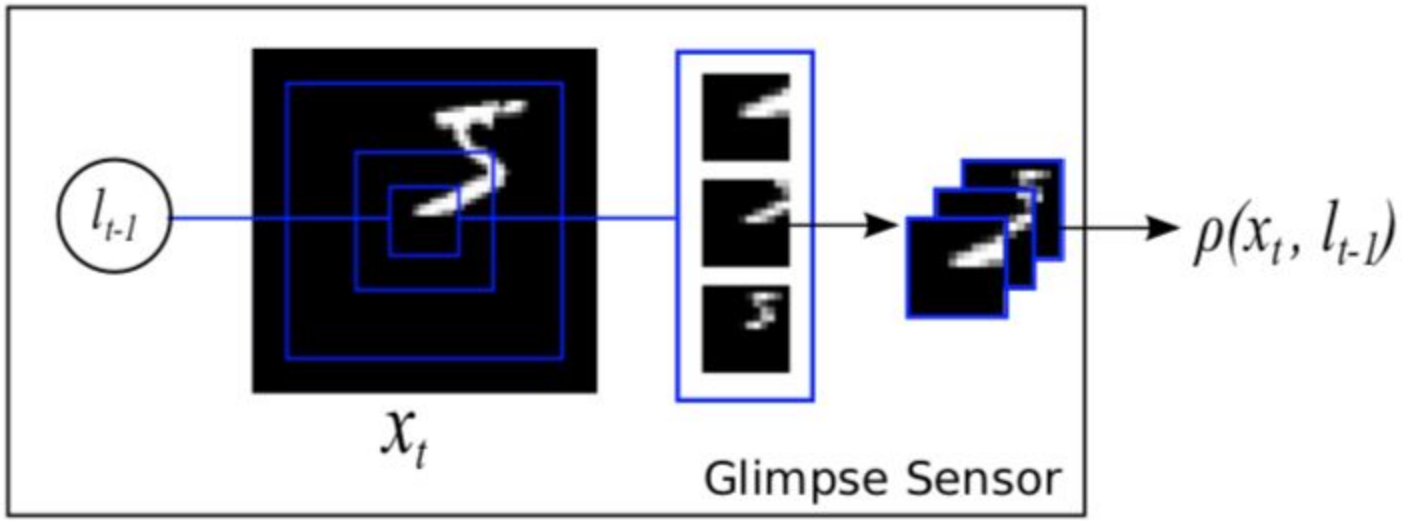
Recurrent Models of Visual Attention

Journal Club: 16th April 2018

- A network architecture which can glimpse at data and decide what parts are useful:
  - Focus on parts of an image
  - Process timeframes in video
  - Move an agent which interacts with a scene
- Uses a retina-like sensor to interpret ROIs - multi res
- Remove the “sliding window” protocol that we all tend to use when processing images

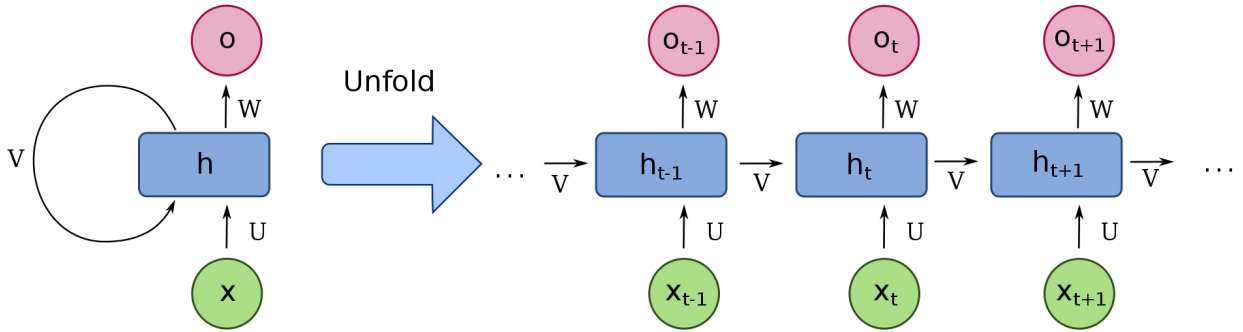

l_{t-1}
x_t
Glimpse Sensor
$\rho(x_t, l_{t-1})$

RNNs

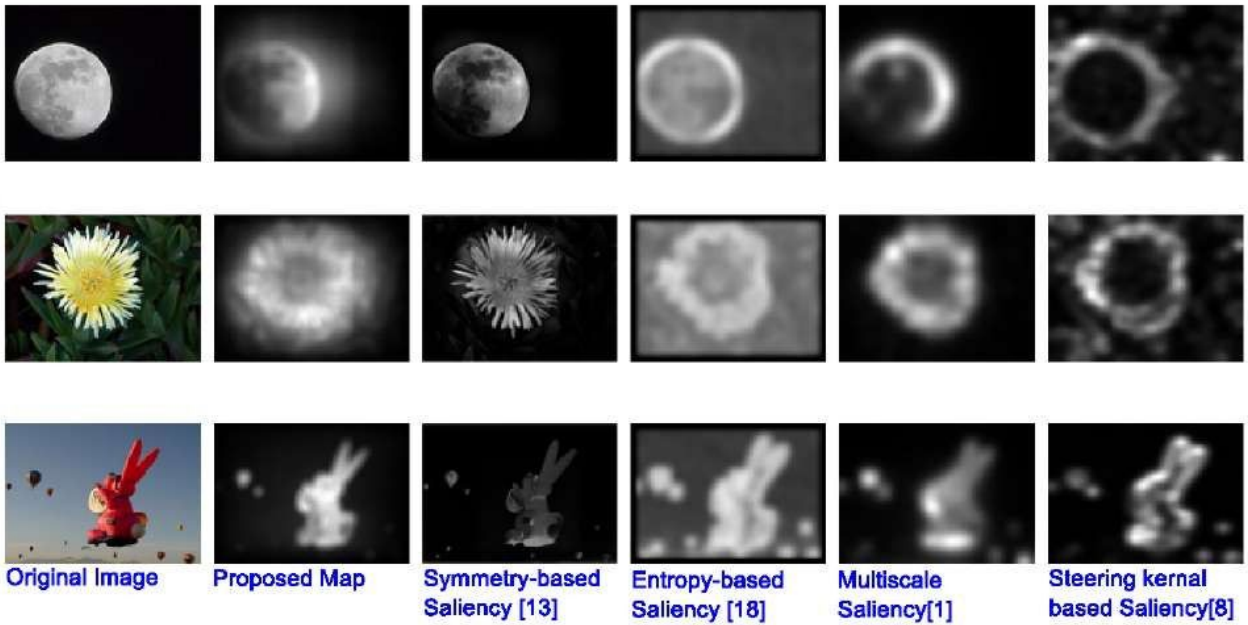
Originally designed to model sequences of data e.g. sentences

Short term memory (h)

- Allowing them to look into the past (for a predetermined time)
- Updated based in previous time point and x (input)



Saliency Maps



http://what-when-how.com/pattern-recognition-and-image-analysis/a-visual-saliency-map-based-on-random-sub-window-means-pattern-recognition-and-image-analysis/

Recurrent attention model (RAM)

Think of this is an agent looking at a scene

- C1: Can only see a subset of the scene at a time (bandwidth limited sensor)
- C2: It can only extract info only in a local region or narrow frequency band
- C3: Agent can choose where it want to go next (sensor resources)

At each time step, the agent receives an award

The Model

- Sensor
- Internal State
- Action
- Reward

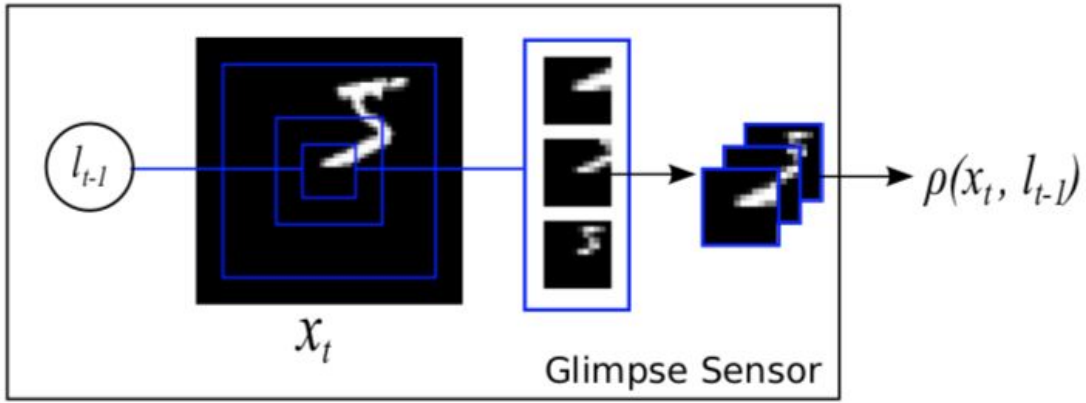
The Model

Sensor

Internal State

Action

Reward



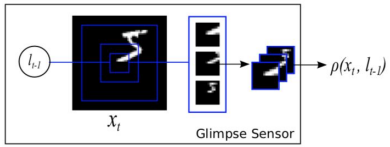
The Model

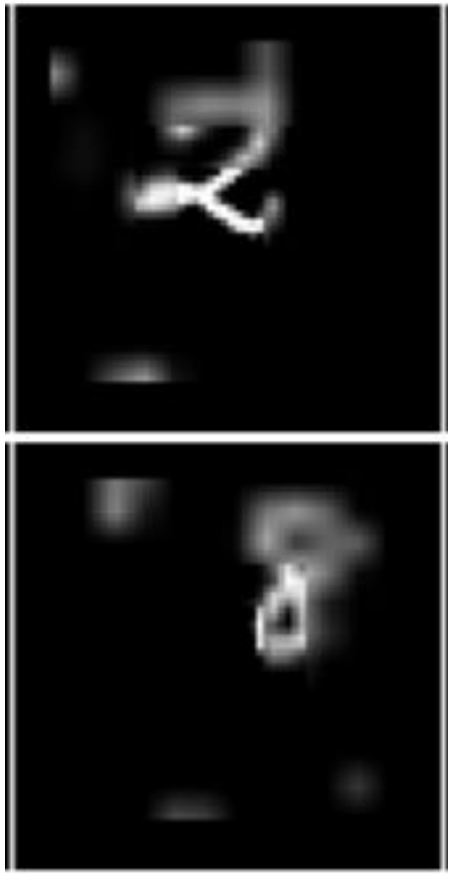
Sensor

Internal State

Action

Reward





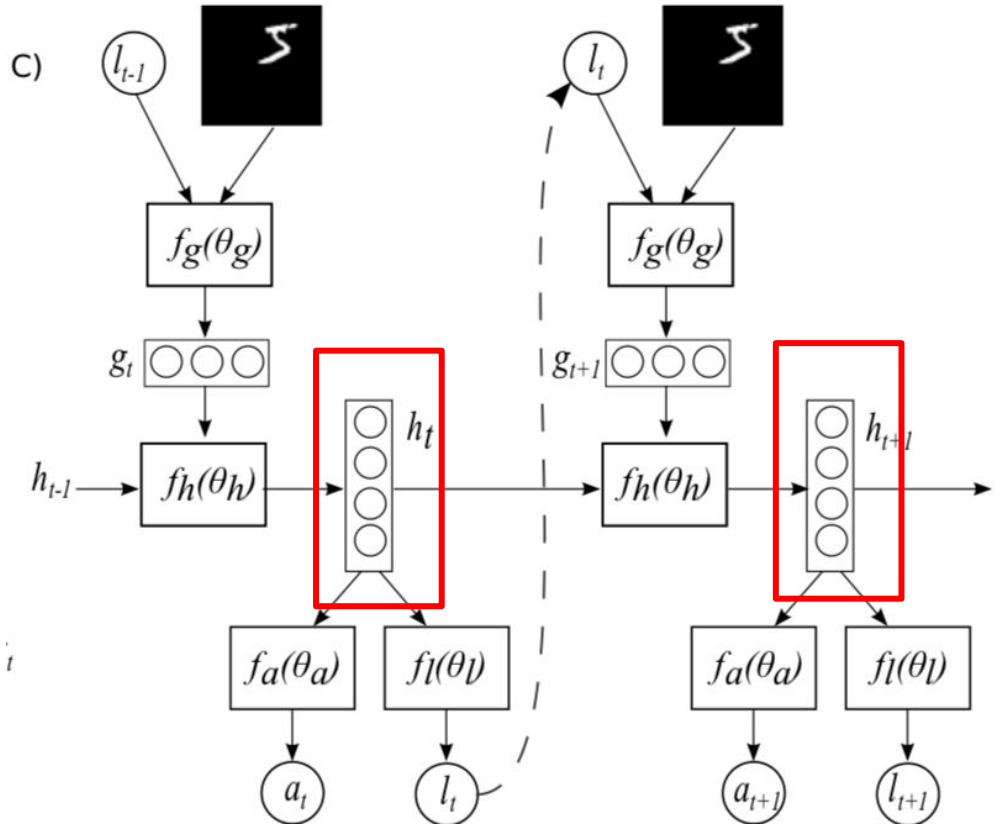
The Model

Sensor

Internal State

Action

Reward



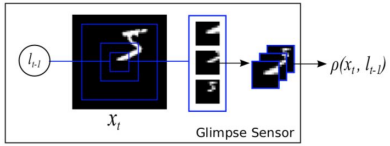
The Model

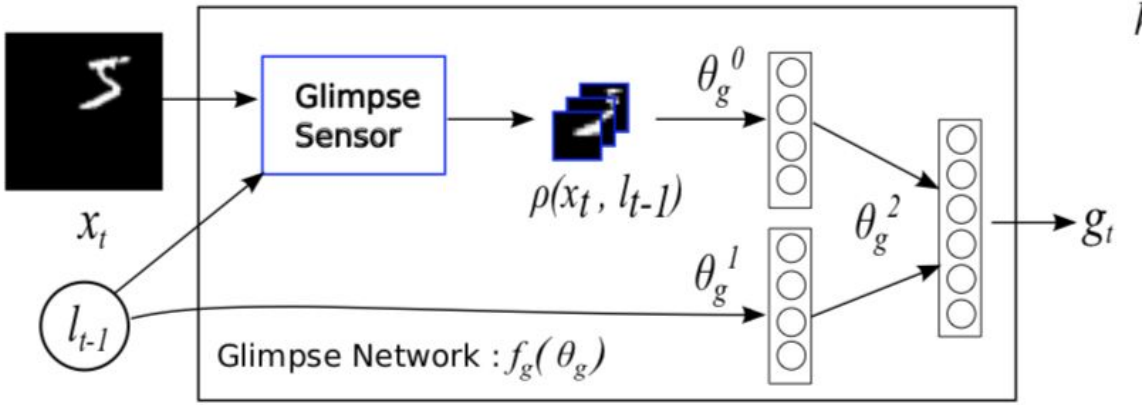
Sensor

Internal State

Action

Reward





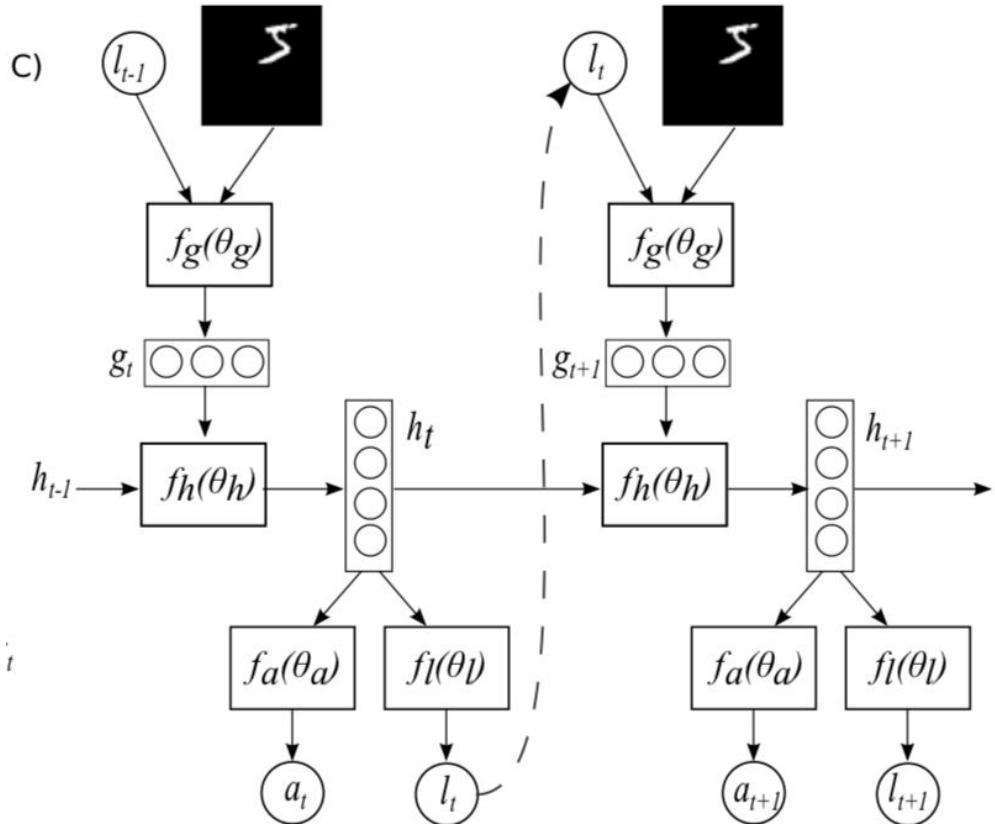
The Model

Sensor

Internal State

Action

Reward



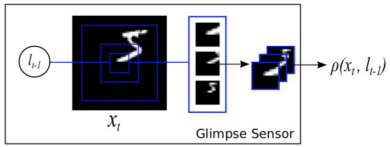
The Model

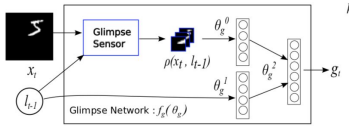
Sensor

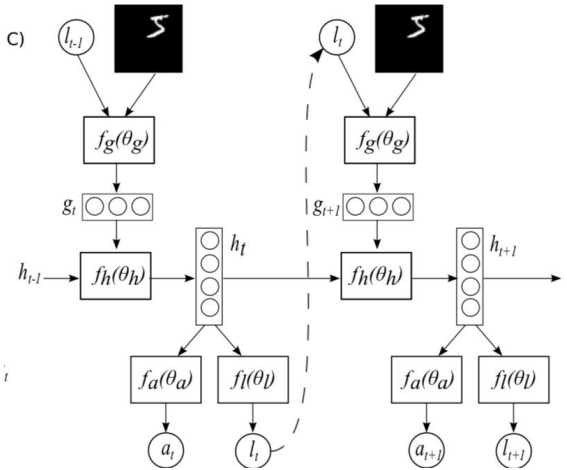
Internal State

Action

Reward







Experiments

Glimpse network: 2 fully connected layers (128 neurons in each)

Location network (f_l): two component Gaussian with fixed variance (x,y)

Action network (f_a): linear softmax classifier (256 neurons)

(a) Translated MNIST inputs.

(b) Cluttered Translated MNIST inputs.

Results

(a) 28x28 MNIST

Model	Error
FC, 2 layers (256 hiddens each)	1.69%
Convolutional, 2 layers	1.21%
RAM, 2 glimpses, 8×8, 1 scale	3.79%
RAM, 3 glimpses, 8×8, 1 scale	1.51%
RAM, 4 glimpses, 8×8, 1 scale	1.54%
RAM, 5 glimpses, 8×8, 1 scale	1.34%
RAM, 6 glimpses, 8×8, 1 scale	1.12%
RAM, 7 glimpses, 8×8, 1 scale	**1.07**%

(b) 60x60 Translated MNIST

Model	Error
FC, 2 layers (64 hiddens each)	6.42%
FC, 2 layers (256 hiddens each)	2.63%
Convolutional, 2 layers	1.62%
RAM, 4 glimpses, 12×12, 3 scales	1.54%
RAM, 6 glimpses, 12×12, 3 scales	**1.22**%
RAM, 8 glimpses, 12×12, 3 scales	**1.2**%

Results

(a) 60x60 Cluttered Translated MNIST

Model	Error
FC, 2 layers (64 hiddens each)	28.58%
FC, 2 layers (256 hiddens each)	11.96%
Convolutional, 2 layers	8.09%
RAM, 4 glimpses, 12 × 12, 3 scales	4.96%
RAM, 6 glimpses, 12 × 12, 3 scales	4.08%
RAM, 8 glimpses, 12 × 12, 3 scales	4.04%
RAM, 8 random glimpses	14.4%

(b) 100x100 Cluttered Translated MNIST

Model	Error
Convolutional, 2 layers	14.35%
RAM, 4 glimpses, 12 × 12, 4 scales	9.41%
RAM, 6 glimpses, 12 × 12, 4 scales	8.31%
RAM, 8 glimpses, 12 × 12, 4 scales	8.11%
RAM, 8 random glimpses	28.4%

Results

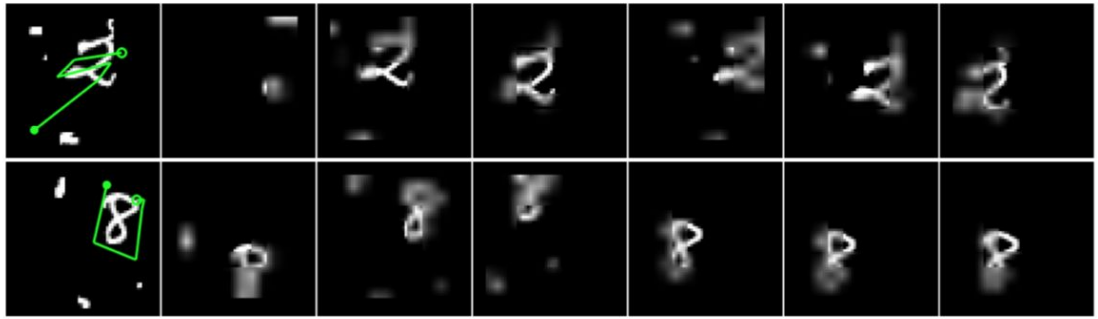
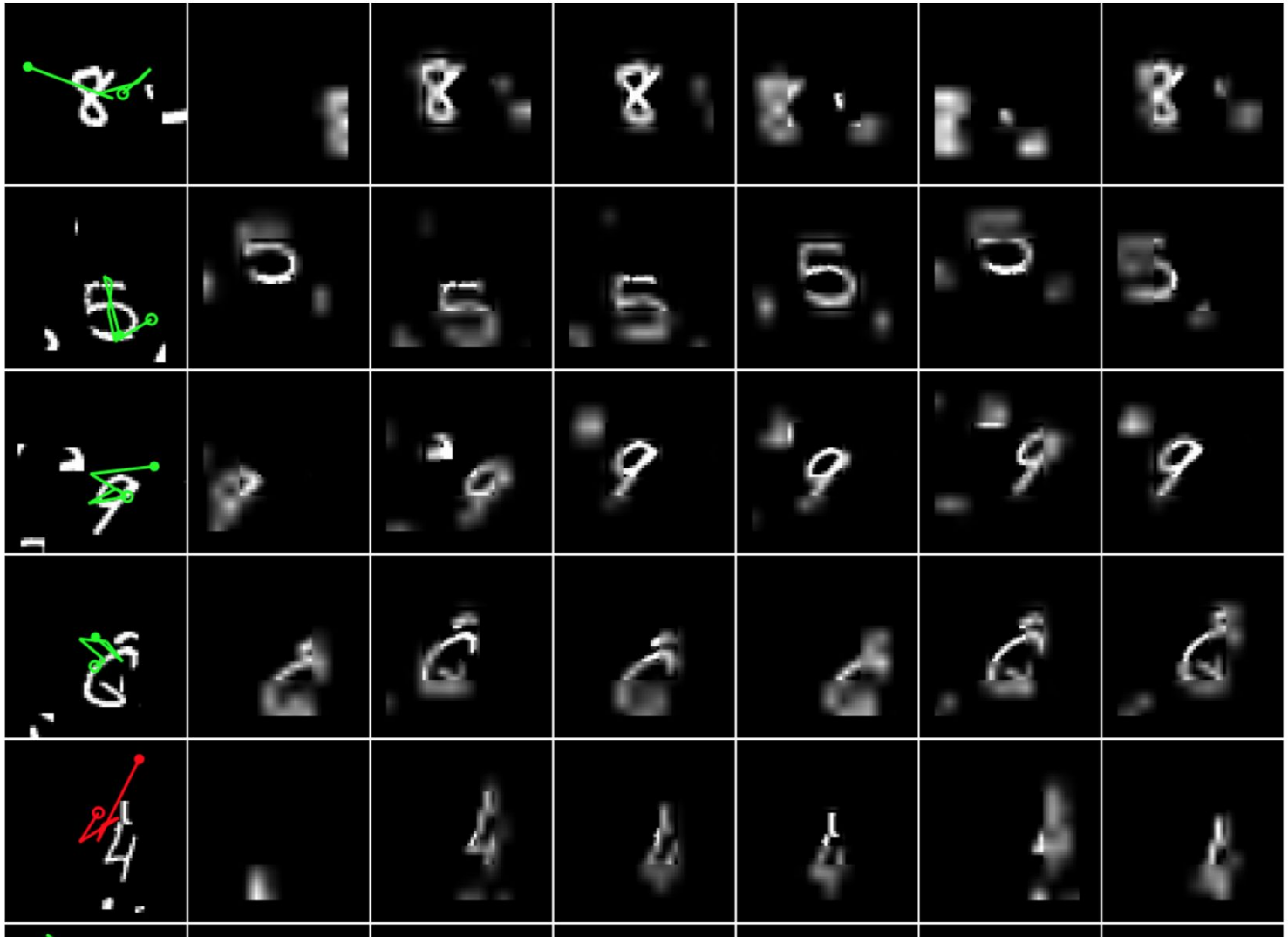


Figure 3: Examples of the learned policy on 60×60 cluttered-translated MNIST task. Column 1: The input image with glimpse path overlaid in green. Columns 2-7: The six glimpses the network chooses. The center of each image shows the full resolution glimpse, the outer low resolution areas are obtained by upscaling the low resolution glimpses back to full image size. The glimpse paths clearly show that the learned policy avoids computation in empty or noisy parts of the input space and directly explores the area around the object of interest.



http://www.cs.toronto.edu/~vmnih/docs/attention.mov