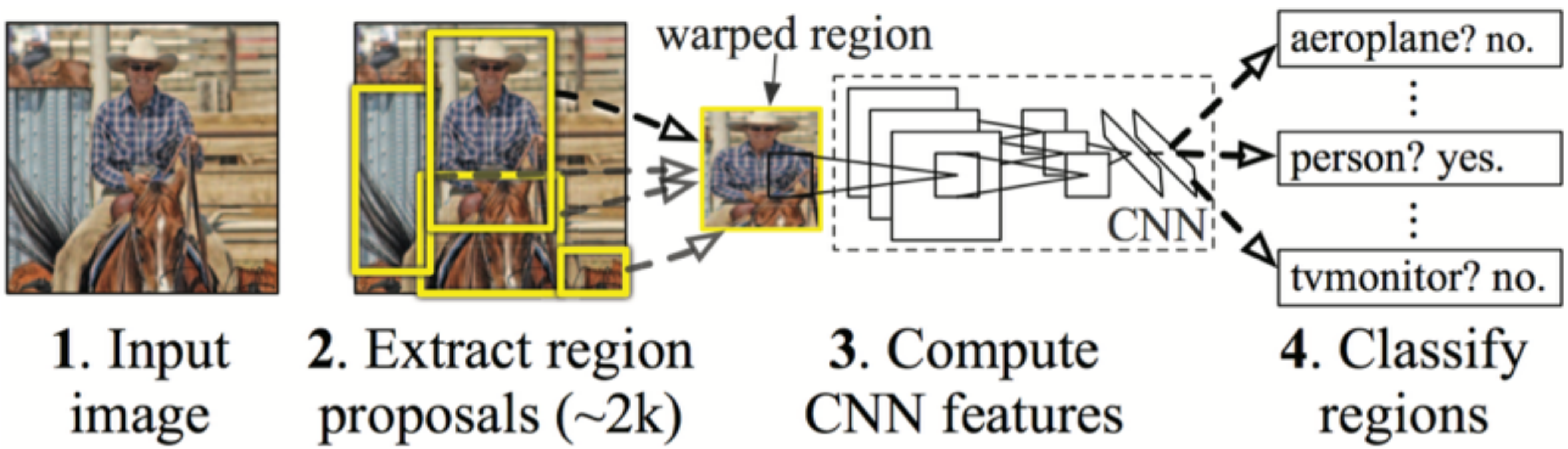
Visual Attention With Neural Networks

Main Paper: Recurrent Models of Visual Attention

Presentation by Matthew Shepherd

Mnih, V., Heess, N., Graves, A., & kavukcuoglu, K. (2014). Recurrent Models of Visual Attention. Advances in Neural Information …, 2204–2212.

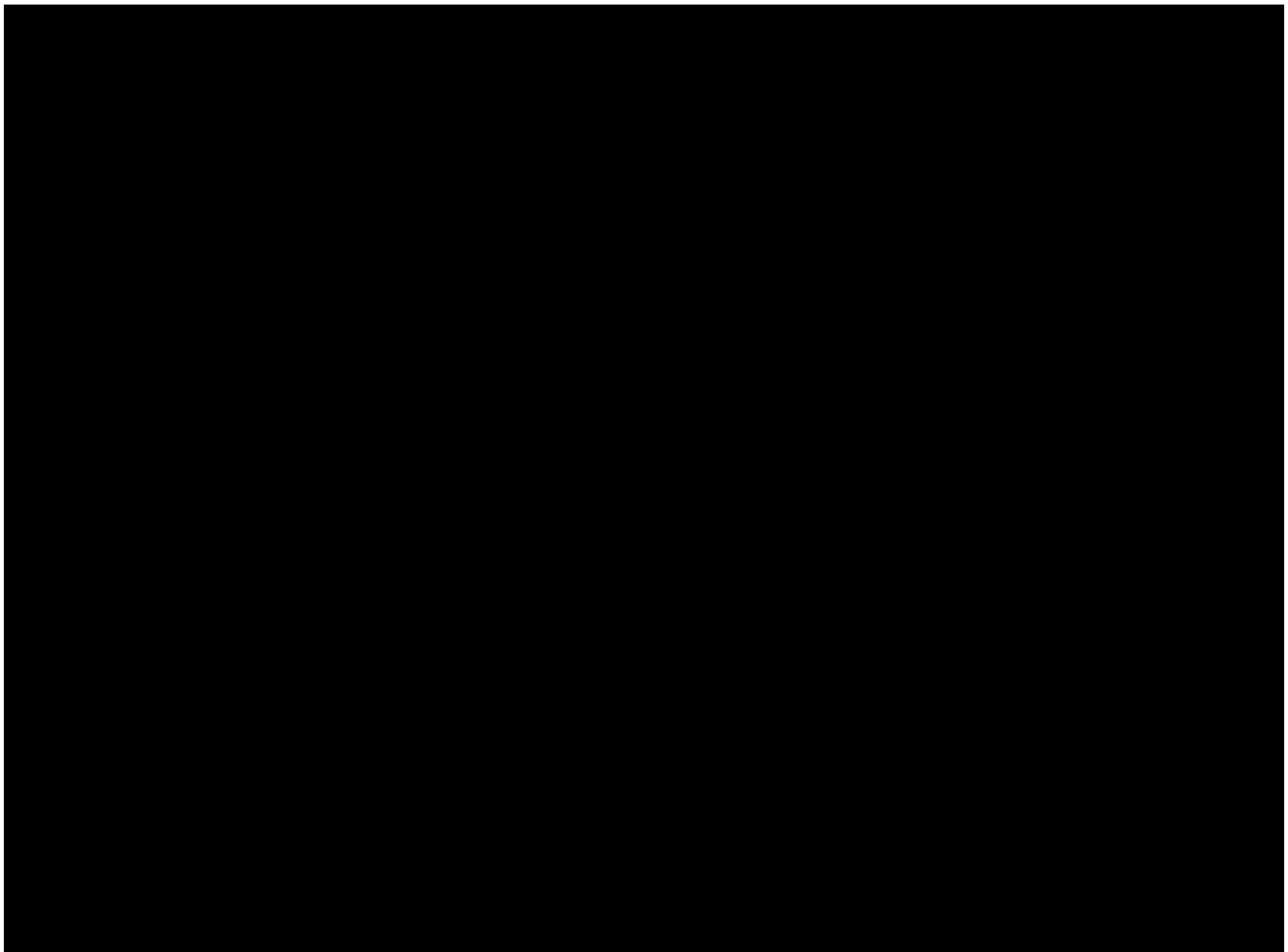
Full image processing is computationally expensive



Regions are can be selected intelligently but time still scales with the size of the image

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. … And Pattern Recognition …, 580–587. http://doi.org/10.1109/CVPR.2014.81

Humans focus on specific regions in their FOV



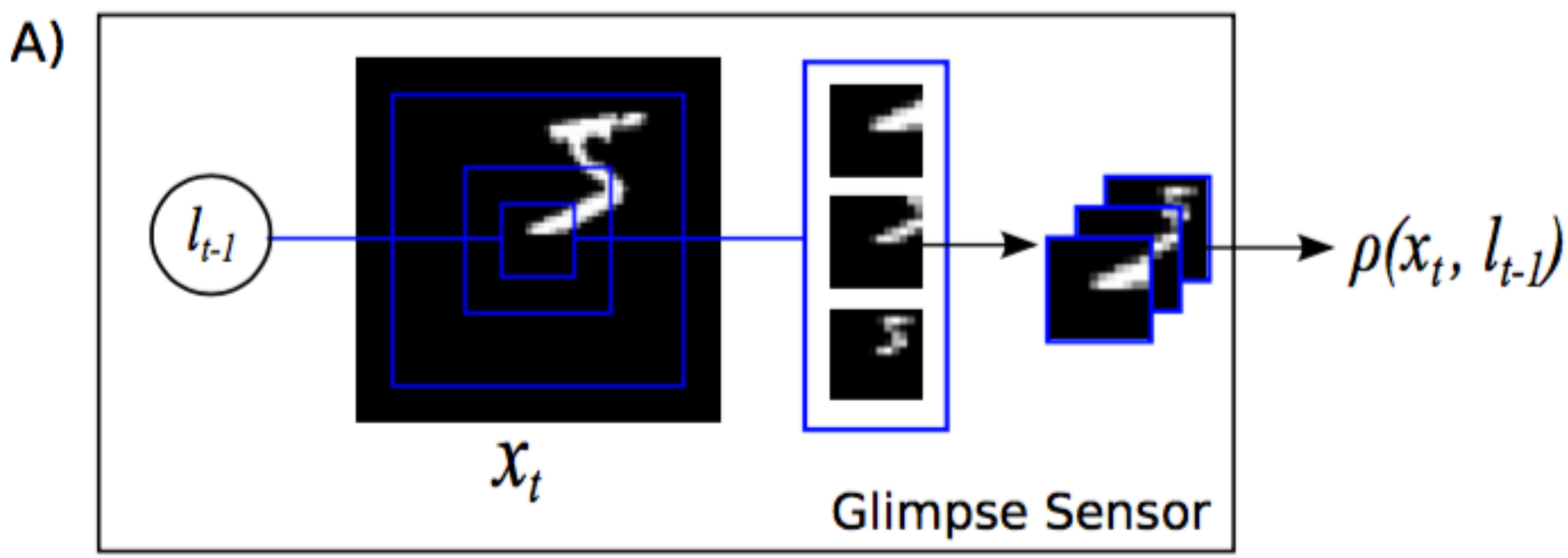
https://www.youtube.com/watch?v=vJG698U2Mvo

“Vision as a sequential decision task”

The model sequentially chooses small windows of information on the data

Integrates information from all past windows to make its next decision.

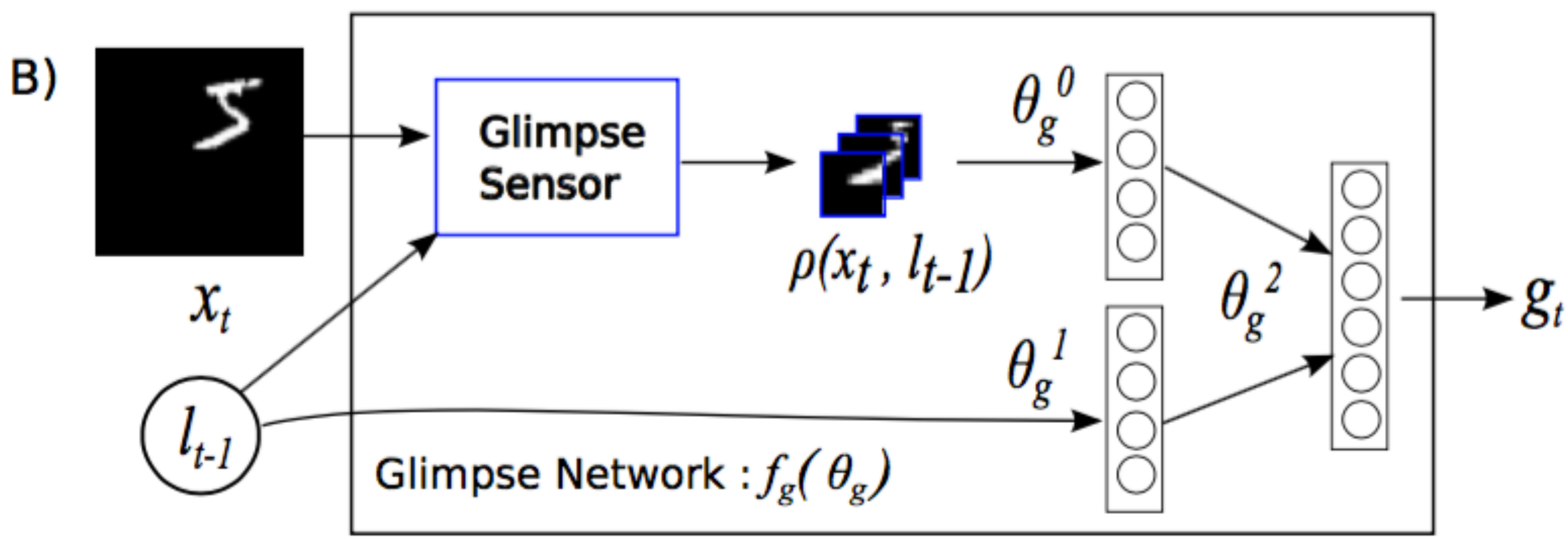
The sensor only provides limited information about the scene, X_t, focused at a location, l_{t-1}



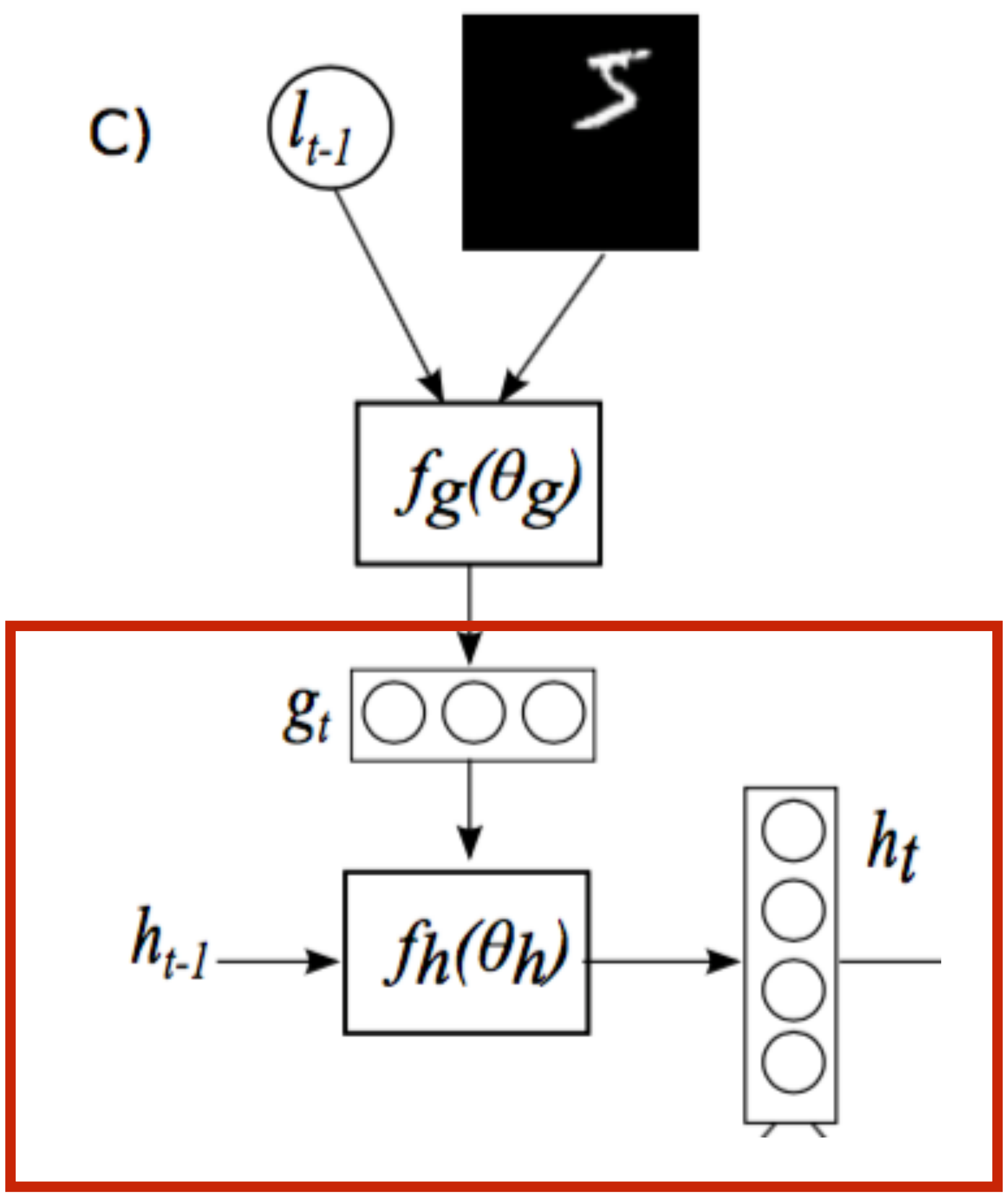
Referred to as a "Glimpse"

retina-like representation

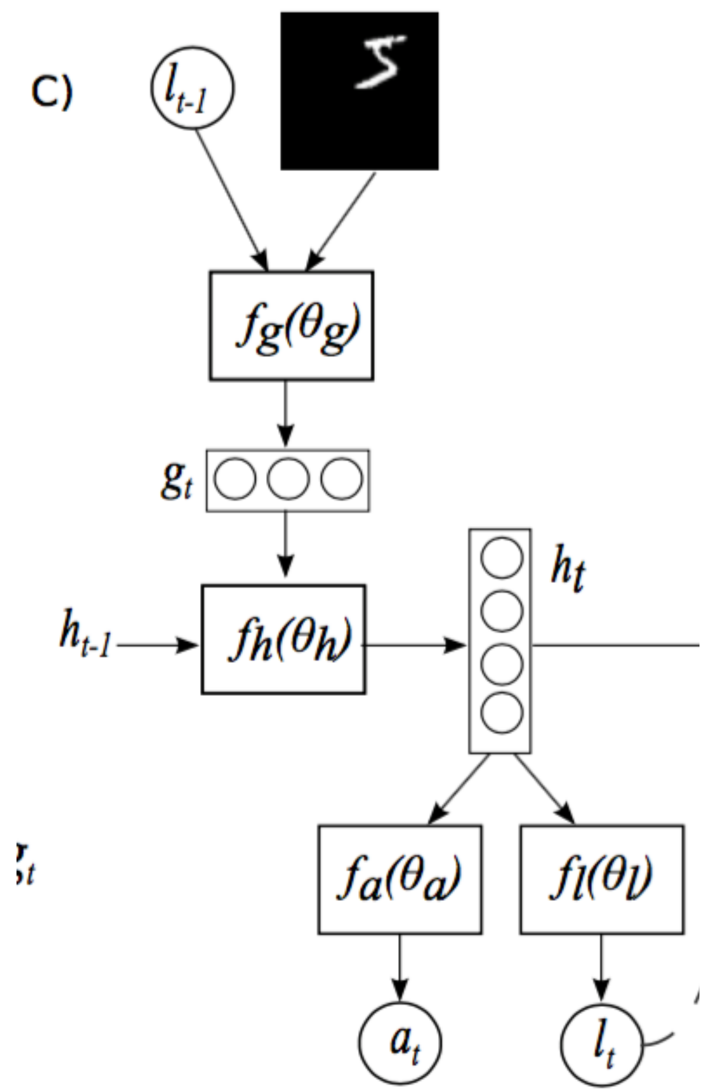
The “glimpse network” incorporates sensor and location information into a single vector



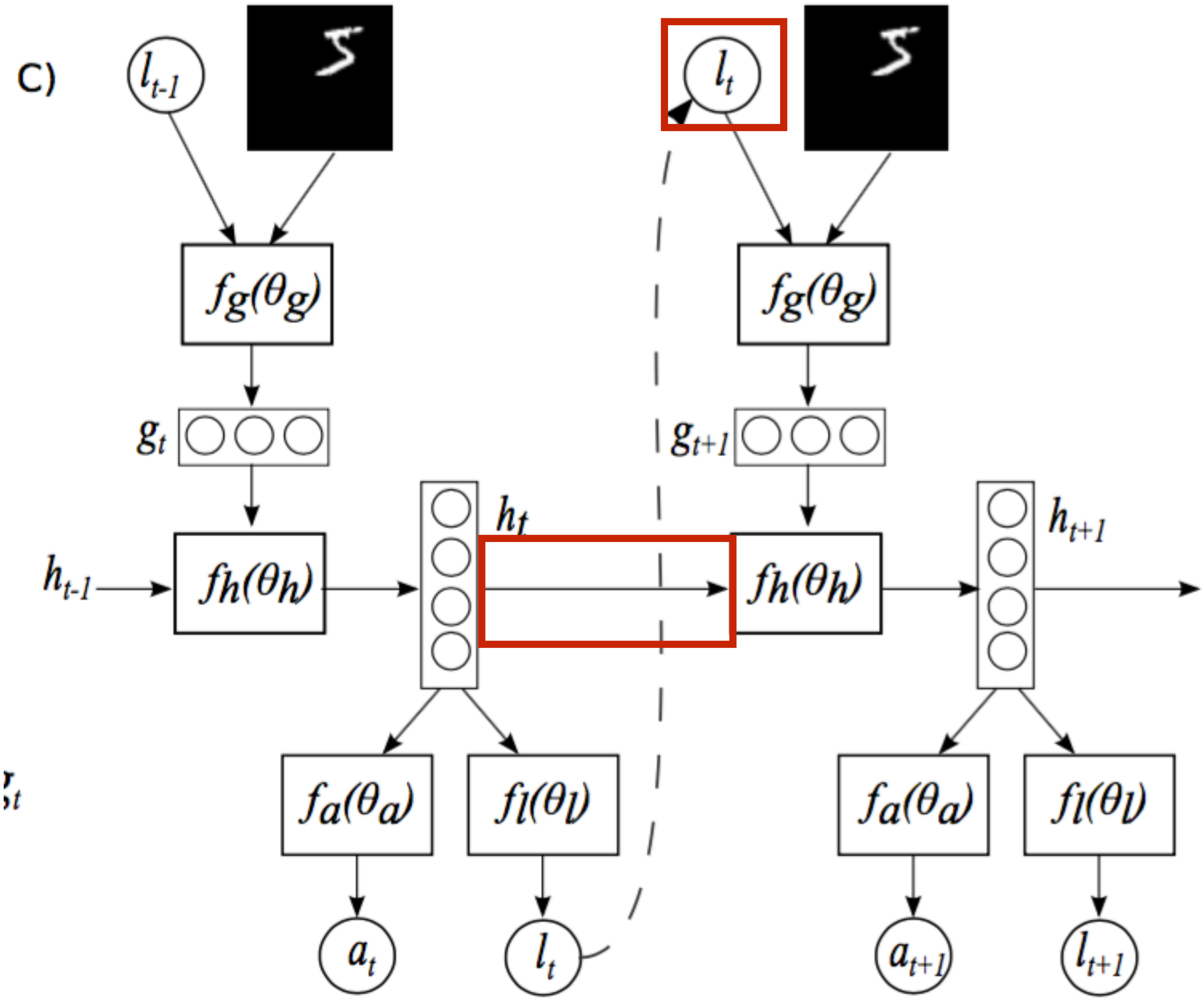
The core network $f_h(\Theta_h)$ incorporates the sensor information as well as past information



The output of the core network is then used to deploy the sensor and make a classification



The network is recurrent



Partially Observed Markov Decision Problem (POMDP)

- Glimpse can be seen as a partial view of the state
- Network must learn a policy

$$\pi((l_t, a_t)|s_{1:t}; \theta)$$

- The policy is determined by the NN

- State history is encapsulated by the hidden state of the network

So how do we train it?

$$J(\theta) = \mathbb{E}_{p(s_{1:T};\theta)}[\sum r_t]$$

$$\nabla_\theta J = \sum_{t=1}^{T} \mathbb{E}_{p(s_{1:T};\theta)}\left[\nabla_\theta \log \pi(u_t|s_{1:t};\theta)R\right] \approx \frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T} \boxed{\nabla_\theta \log \pi(u_t^i|s_{1:t}^i;\theta)} R^i$$

The REINFORCE rule

Additional training details

Useful for determining f_l but f_a can be determined more directly by minimizing cross-entropy loss.

A baseline value, b_t, is added to the gradient approximation to reduce variance

$$\frac{1}{M}\sum_{i=1}^{M}\sum_{t=1}^{T}\nabla_\theta \log \pi(u_t^i|s_{1:t}^i;\theta)\left(R_t^i - b_t\right)$$

Experiments

RAM performs well on translated MNIST

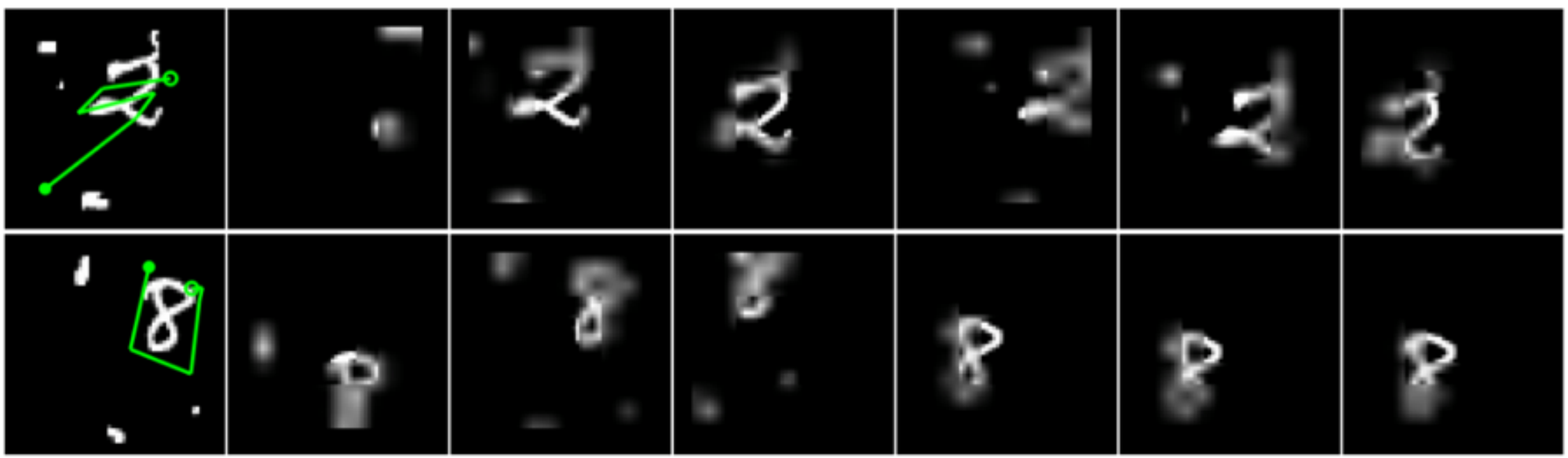
Model	Error
FC, 2 layers (64 hiddens each)	6.42%
FC, 2 layers (256 hiddens each)	2.63%
Convolutional, 2 layers	1.62%
RAM, 4 glimpses, 12×12, 3 scales	1.54%
RAM, 6 glimpses, 12×12, 3 scales	**1.22%**
RAM, 8 glimpses, 12×12, 3 scales	**1.2%**

RAM performs very well on cluttered translated images

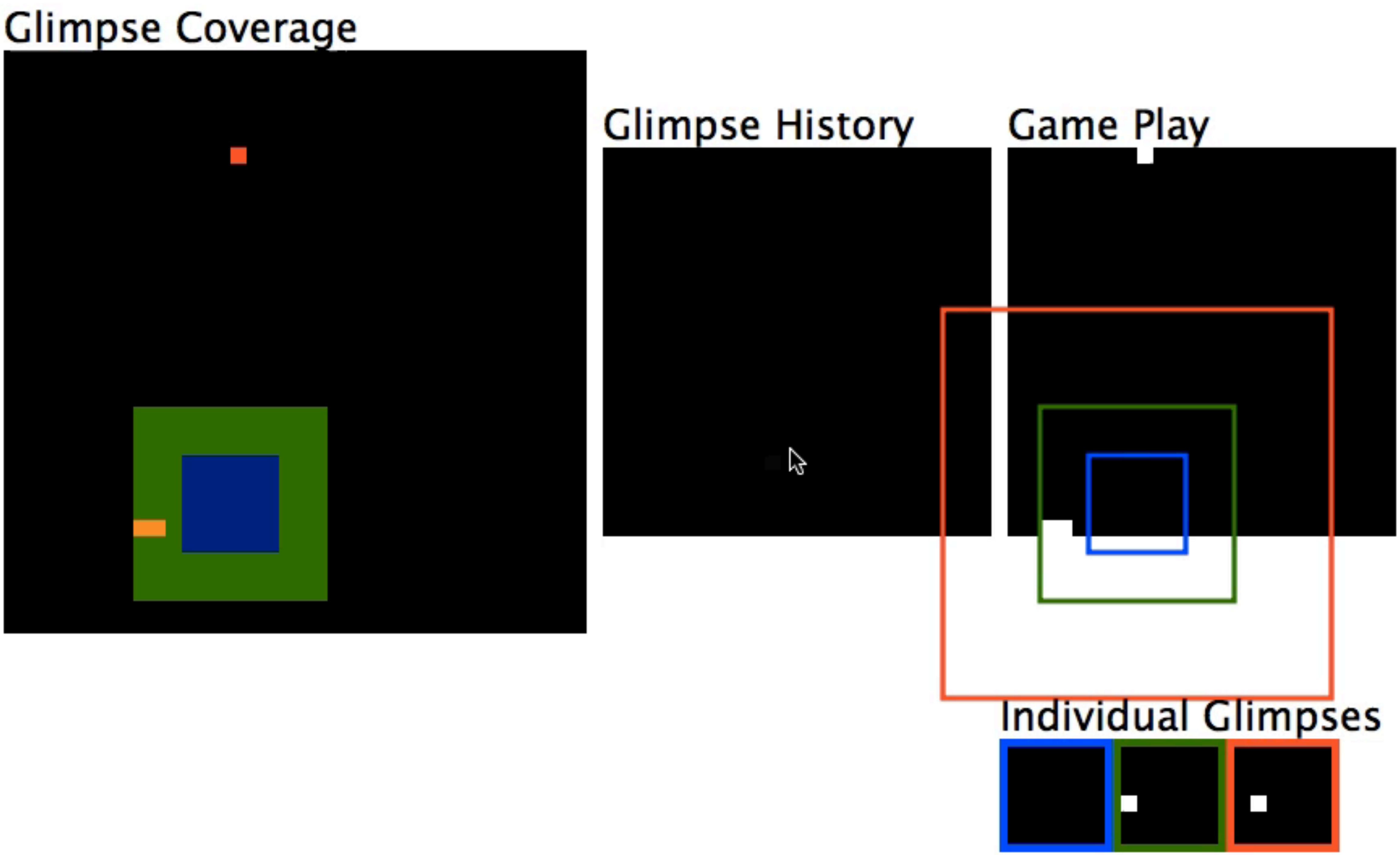
(b) 100x100 Cluttered Translated MNIST

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%

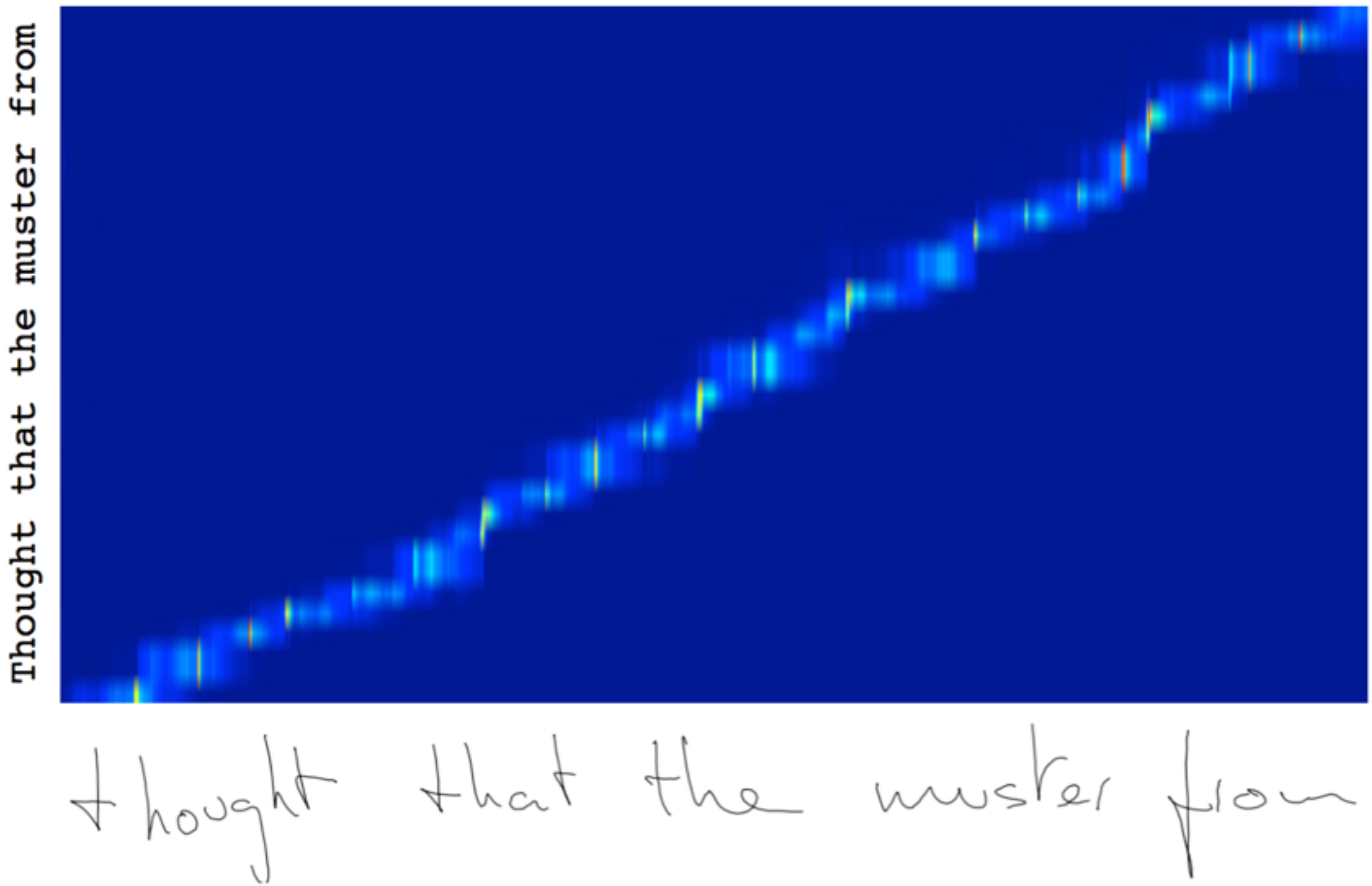
Meaningful policies are learned



RAM performs in a dynamic environment



Other models of attention



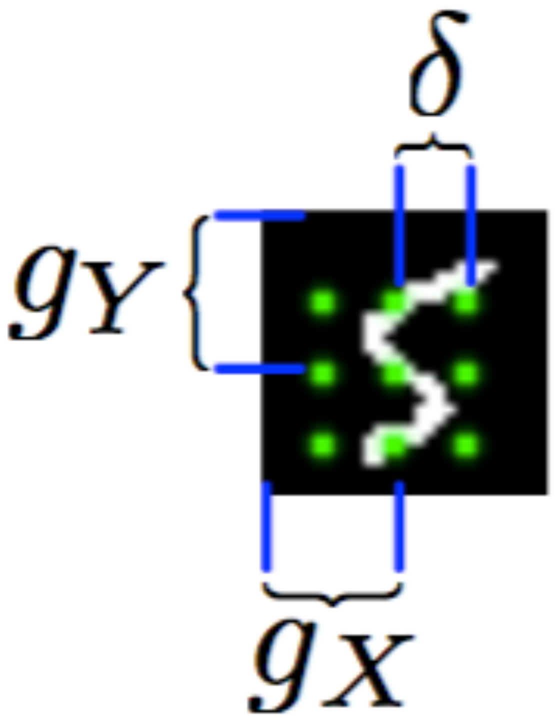
Graves, A. (2014). Generating Sequences with Recurrent Neural Networks, 1–43.

DRAW: A Recurrent Neural Network For Image Generation

- Combines an attention mechanism with a sequential variational auto-encoder
- Reading *and* writing are now both sequential tasks

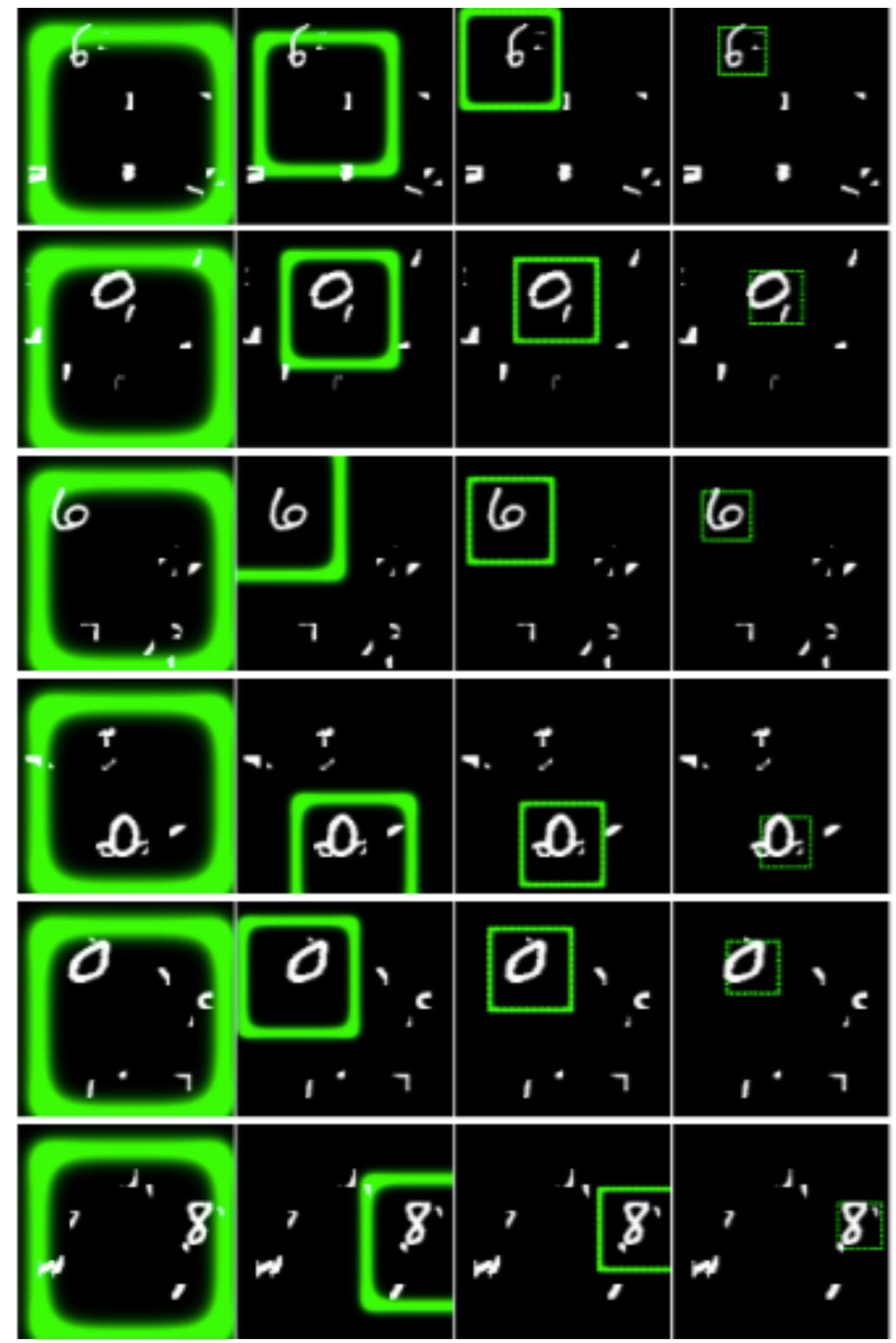
Gregor, K., Danihelka, I., Graves, A., Rezende, D. J., & Wierstra, D. (2015). DRAW: A Recurrent Neural Network For Image Generation , 1–10.

Differentiable RAM



$$\mu_X^i = g_X + (i - N/2 - 0.5)\,\delta$$
$$\mu_Y^j = g_Y + (j - N/2 - 0.5)\,\delta$$

$$(\tilde{g}_X, \tilde{g}_Y, \log\sigma^2, \log\tilde{\delta}, \log\gamma) = W(h^{dec})$$



Differentiable RAM performance

Table 1. Classification test error on 100×100 Cluttered Translated MNIST.

Model	Error
Convolutional, 2 layers	14.35%
RAM, 4 glimpses, 12×12, 4 scales	9.41%
RAM, 8 glimpses, 12×12, 4 scales	8.11%
Differentiable RAM, 4 glimpses, 12×12	4.18%
Differentiable RAM, 8 glimpses, 12×12	**3.36%**

DRAW-ing with attention

https://www.youtube. com/watch?v=Zt-7MI9eKEo