Attention Model 101

Shaofan Lai

Papers

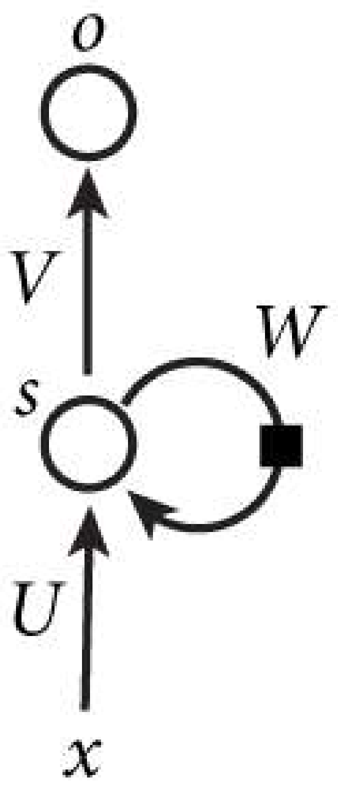
- **Survey**: *Survey on the attention based RNN model and its applications in computer vision*
- **Toy Model**: *Recurrent Models of Visual Attention (NIPS2014)*
- **Image Caption**: *Show, Attend and Tell: Neural Image Caption*
- *Generation with Visual Attention*
- **Action Recognition**: *Action Recognition using Visual Attention (ICLR 2016)*
- **A Structure**: *Spatial Transformer Networks (NIPS2015)*

Contents

- RNN 101
- What is attention
- Different kinds of attention models
- Application
- Pros and cons

RNN 101

RNN 101



$$x \in \mathcal{R}^{T \times D}$$

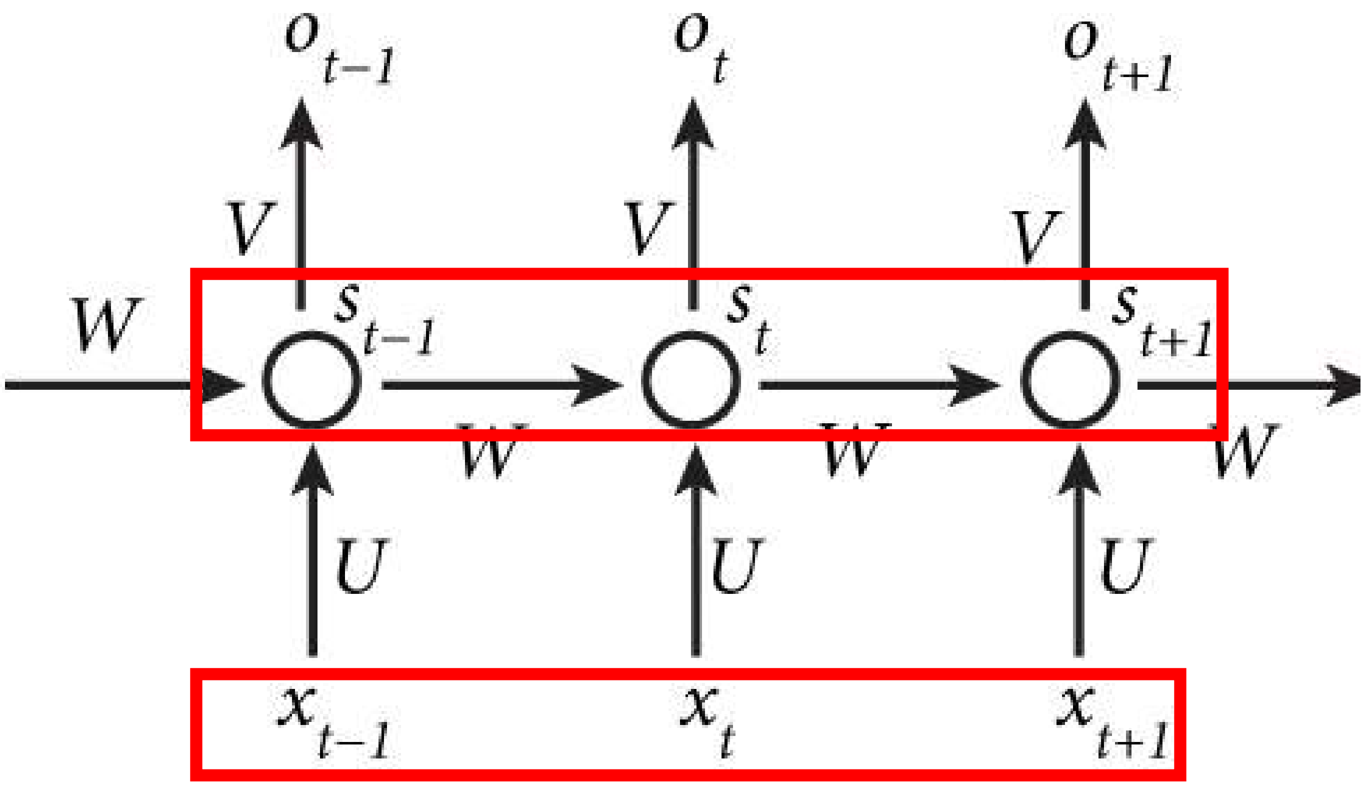
$$x_1, x_2, ..., x_T \in \mathcal{R}^D$$

At time t

$$s_t = f(W s_{t-1} + U x_t)$$

$$o_t = f(V s_t)$$

RNN 101



$x \in \mathcal{R}^{T \times D}$

$x_1, x_2, ..., x_T \in \mathcal{R}^D$

At time t

$$s_t = f(Ws_{t-1} + Ux_t)$$

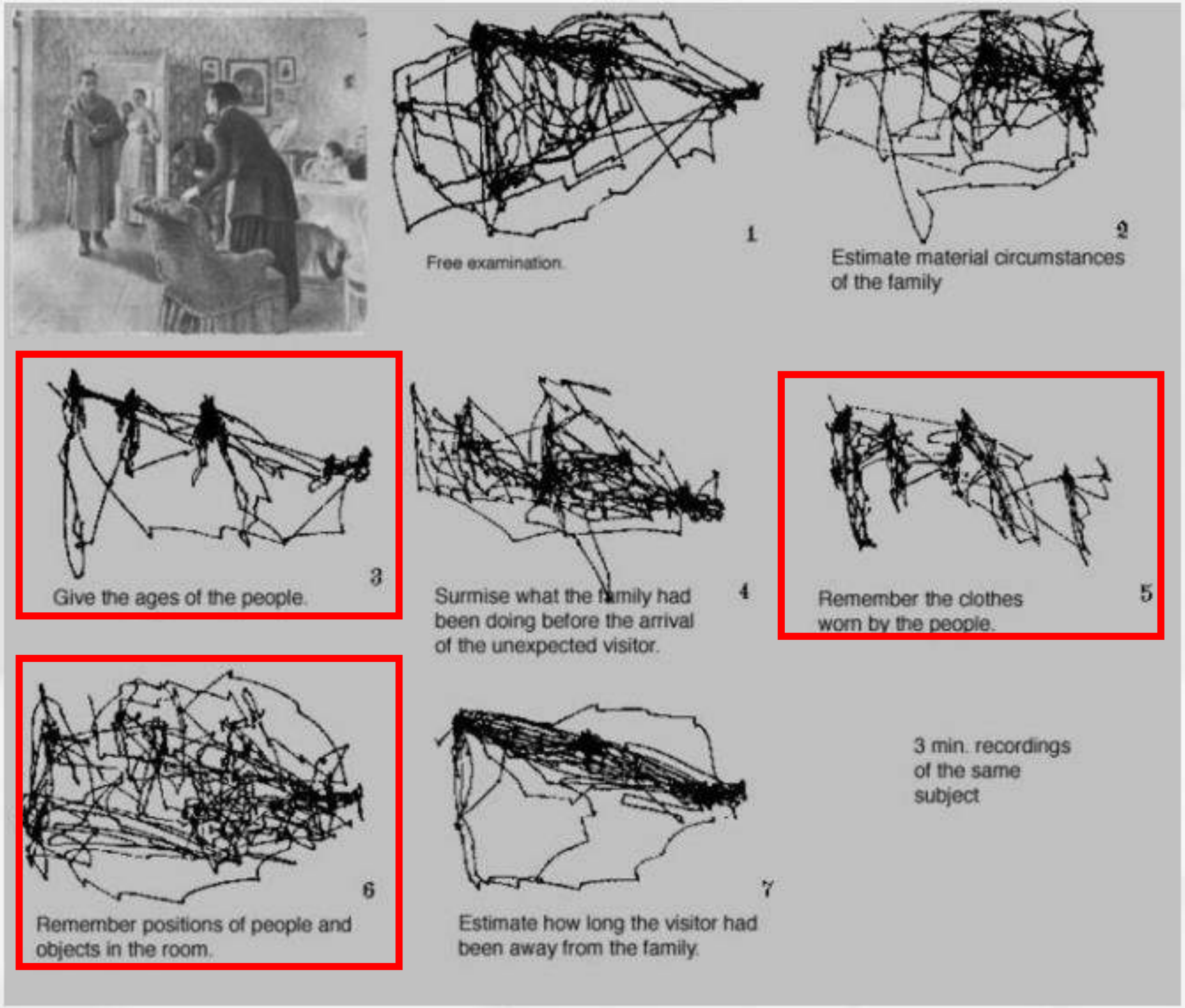
$$o_t = f(Vs_t)$$

What is attention

What is attention

- In psychology, limited by the processing bottlenecks, humans tend to **selectively concentrate on a part of the information**, and at the same time ignore other perceivable information. The above mechanism is usually called attention [3].

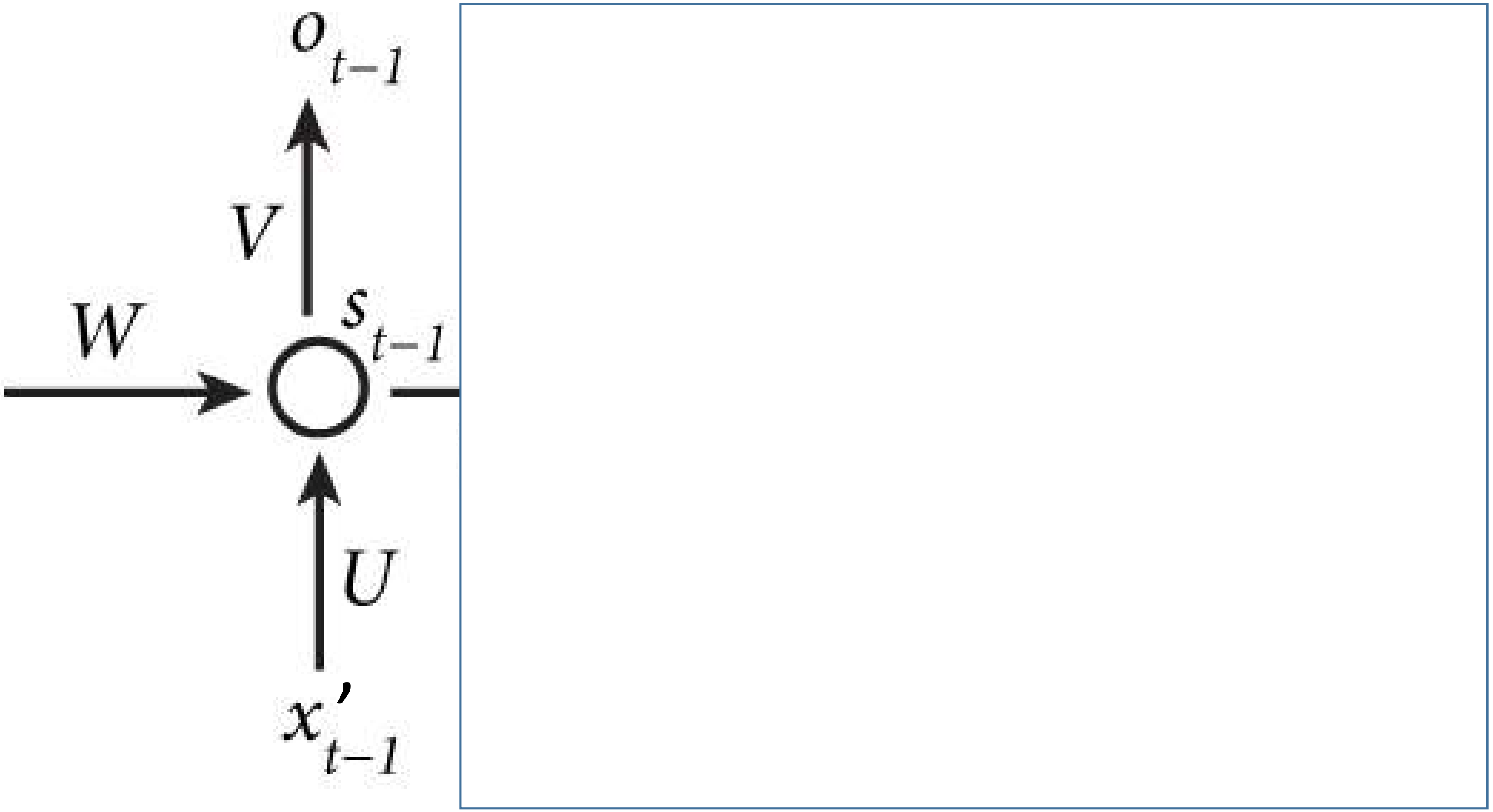
Top-down factors: The task has a strong influence on where you attend and look



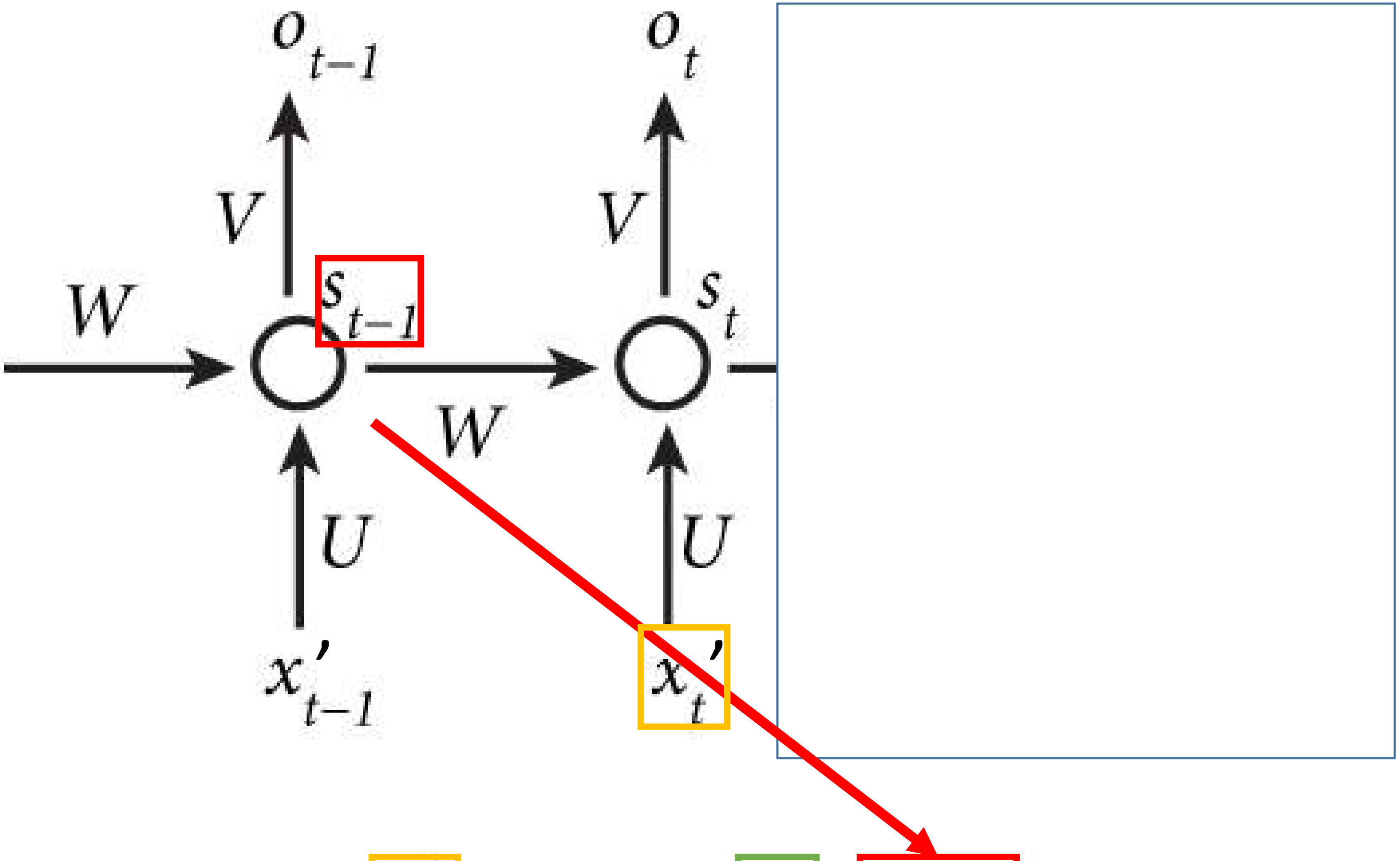
Yarbus, A. L. (1967).

(b) A dog is standing on a hardwood floor.

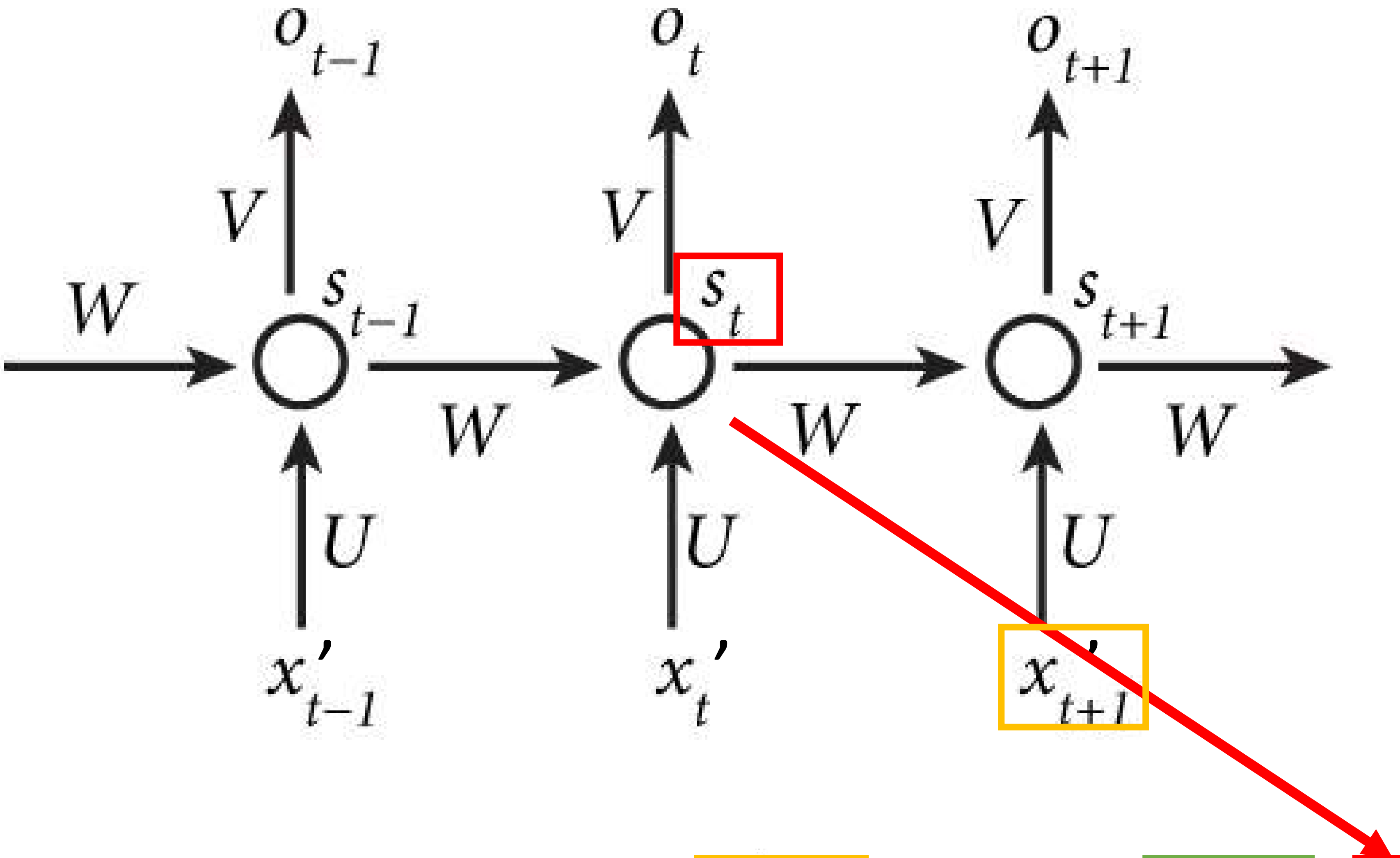
Different kinds of attention models



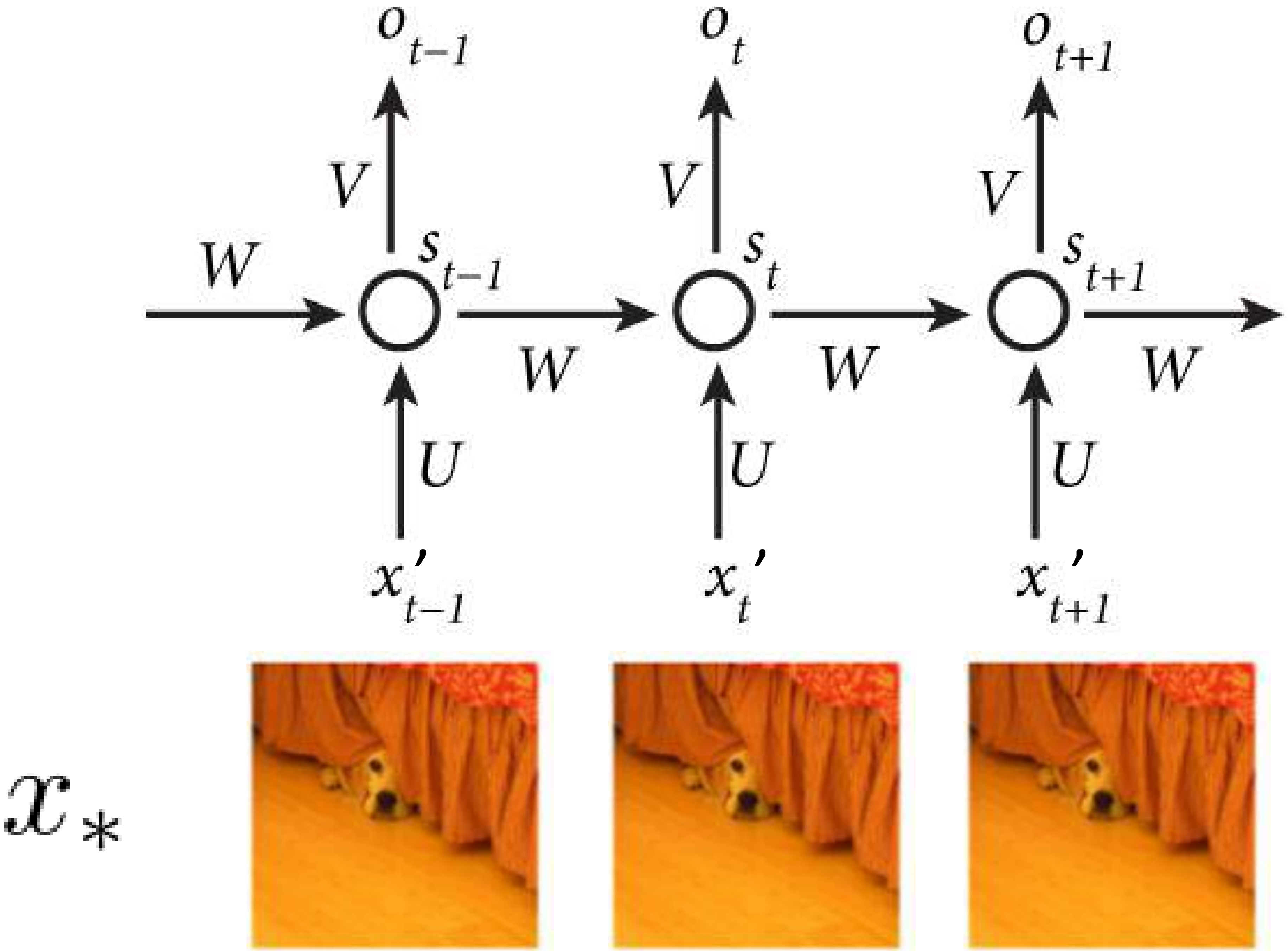
o t−1
V
W
s t−1
U
x' t−1



$$x'_t = f_{att}(x_t, s_{t-1})$$



$$x'_{t+1} = f_{att}(x_{t+1}, s_t)$$



o_{t-1}
o_t
o_{t+1}
V
V
V
W
s_{t-1}
s_t
s_{t+1}
W
W
W
U
U
U
x'_{t-1}
x'_t
x'_{t+1}
x_*

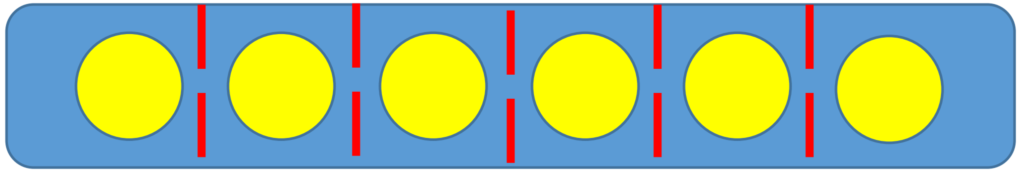
Different kinds of attention models

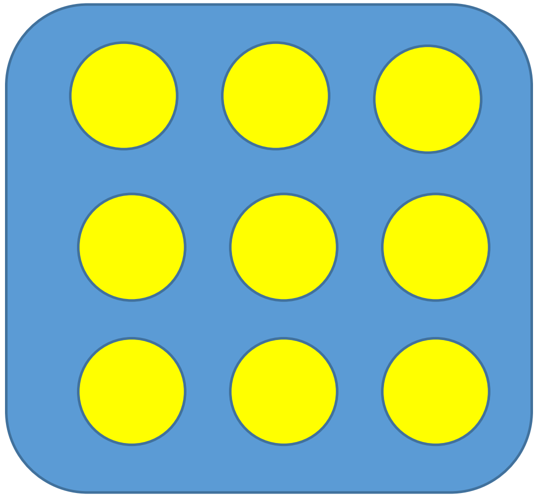
	Item-wise	Location-wise
Hard	Item-wise Hard Attention	Location-wise Hard Attention
Soft	Item-wise Soft Attention	Location-wise Soft Attention

Item-wise vs Location-wise

$$x'_t = f_{att}(x_t, s_{t-1})$$

- Item-wise:
  - A sequence of items
- Location-wise:
  - A feature map/picture/image/frame
- **Spatial connection / Shuffle immutable**





Hard vs Soft

$$x'_t = f_{att}(x_t, s_{t-1})$$

- Hard:
  - Discretely **sampling**
  - **Non-differential**
  - Learning by Reinforcement learning
- Soft:
  - Linear combination/Masking/Weights
  - **Differential**

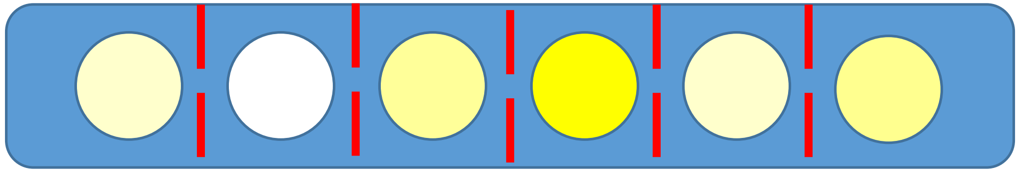
Item-wise Soft Attention

$$x'_t = f_{att}(x_t, s_{t-1})$$

$$e_t = g(x_t, s_{t-1}; \theta) \quad (e_t, x_t \in \mathcal{R}^D)$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{i=1}^{D} exp(e_{ti})}$$

$$x'_t = \sum_{j=1}^{D} \alpha_{tj} x_{tj}$$



Item-wise Hard Attention

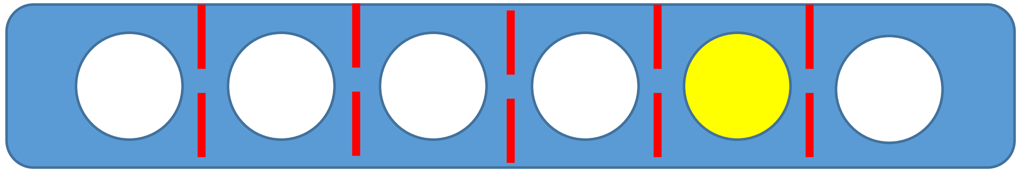
$$x'_t = f_{att}(x_t, s_{t-1})$$

$$e_t = g(x_t, s_{t-1}; \theta) \quad (e_t, x_t \in \mathcal{R}^D)$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{i=1}^{D} exp(e_{ti})}$$

$$\mathcal{L} \sim \mathcal{C}(D, \{\alpha_{tj}\}_{j=1}^{D})$$

$$x'_t = x_{t\mathcal{L}}$$



Location-wise Soft Attention

$$x'_t = f_{att}(x_t, s_{t-1})$$

$$e_t = g(x_t, s_{t-1}; \theta) \quad (e_t, x_t \in \mathcal{R}^D) \rightarrow \mathcal{R}^{H \times W}$$

$$\alpha_{tj} = \frac{exp(e_{tj})}{\sum_{i=1}^{D} exp(e_{ti})}$$

$$x'_t = \sum_{j=1}^{D} \alpha_{tj} x_{tj}$$

Location-wise Hard Attention

$$x'_t = f_{att}(x_t, s_{t-1})$$

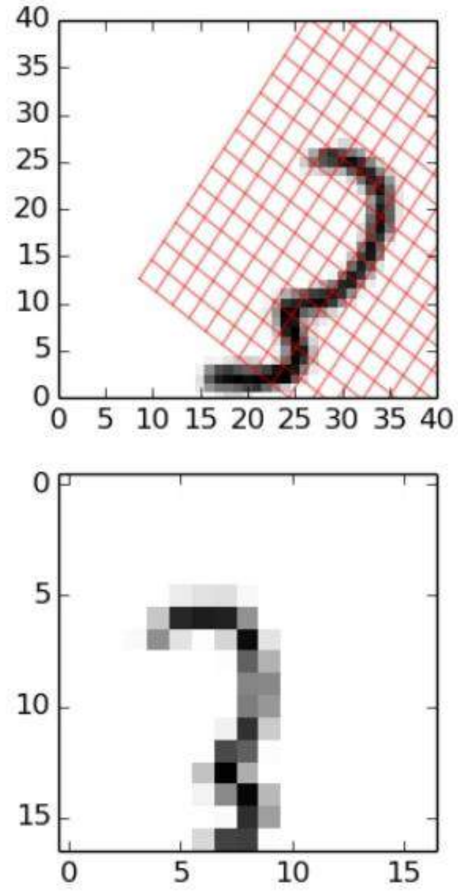
$$[X, Y] = g(x_t, s_{t-1}; \theta) \quad (X, Y \in \mathcal{R})$$
$$[X', Y'] \sim \mathcal{N}([X, Y], \sigma)$$
$$x'_t = f_{crop}(x_t, [X, Y, h, w])$$

(Another) Location-wise Soft Attention

- *Spatial Transformer Networks* (NIPS2015)

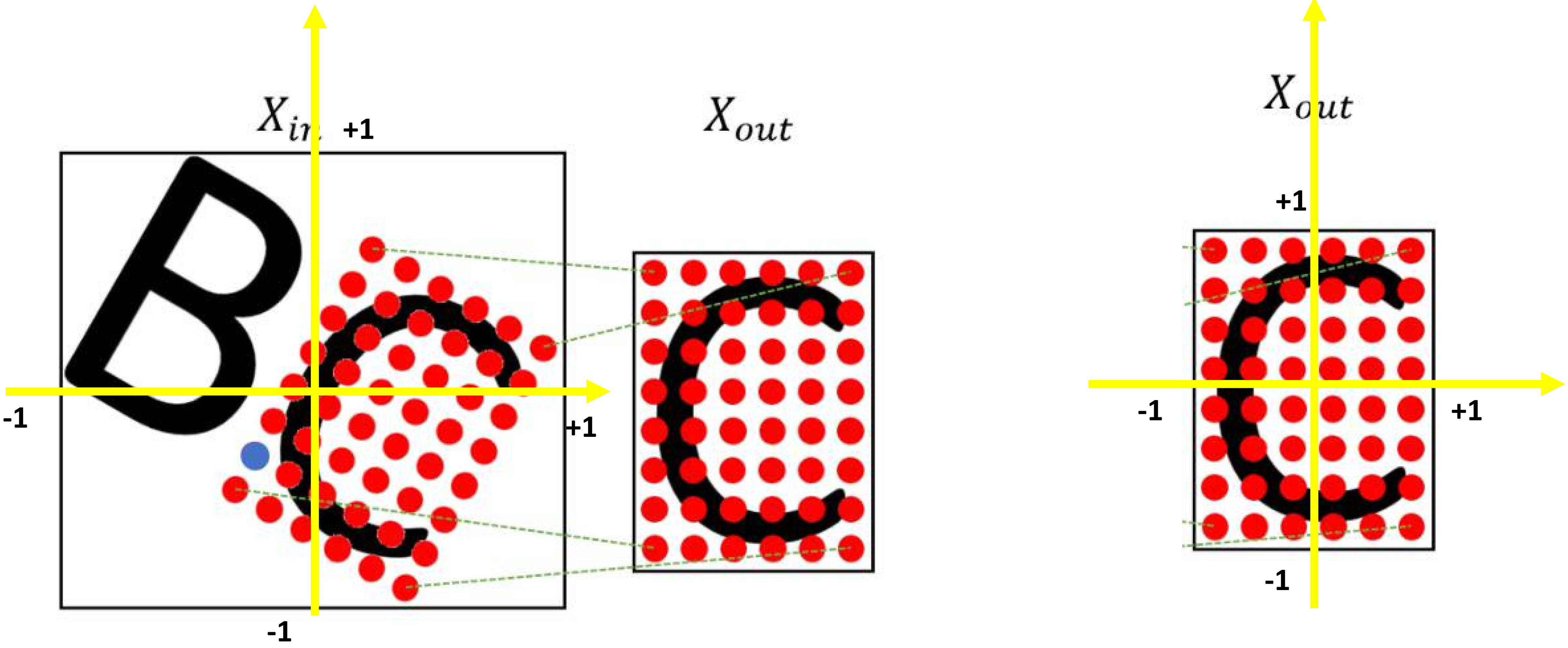
$$A_j = g(x_t, s_{t-1}; \theta)$$

$$A_j = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix}$$



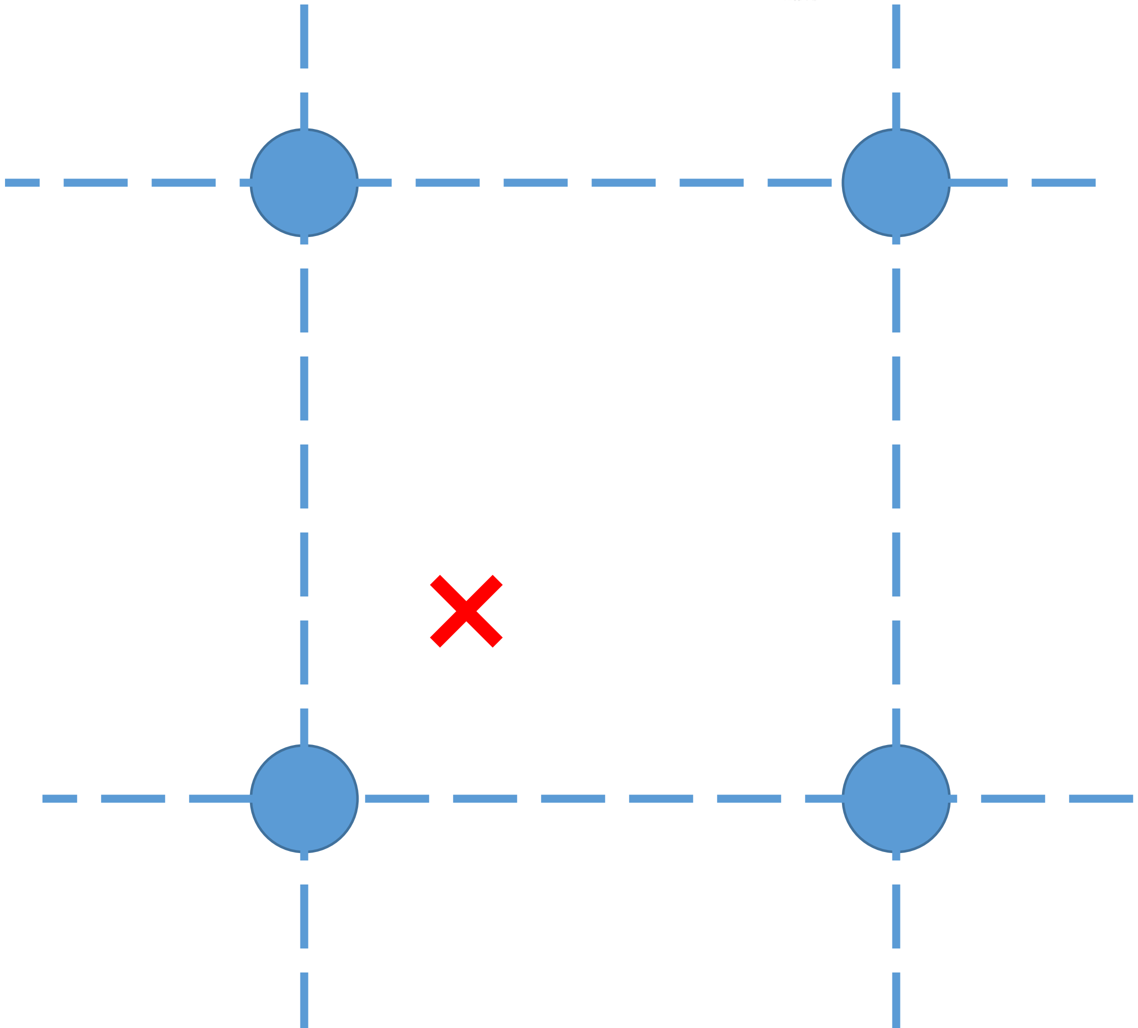
Read this way

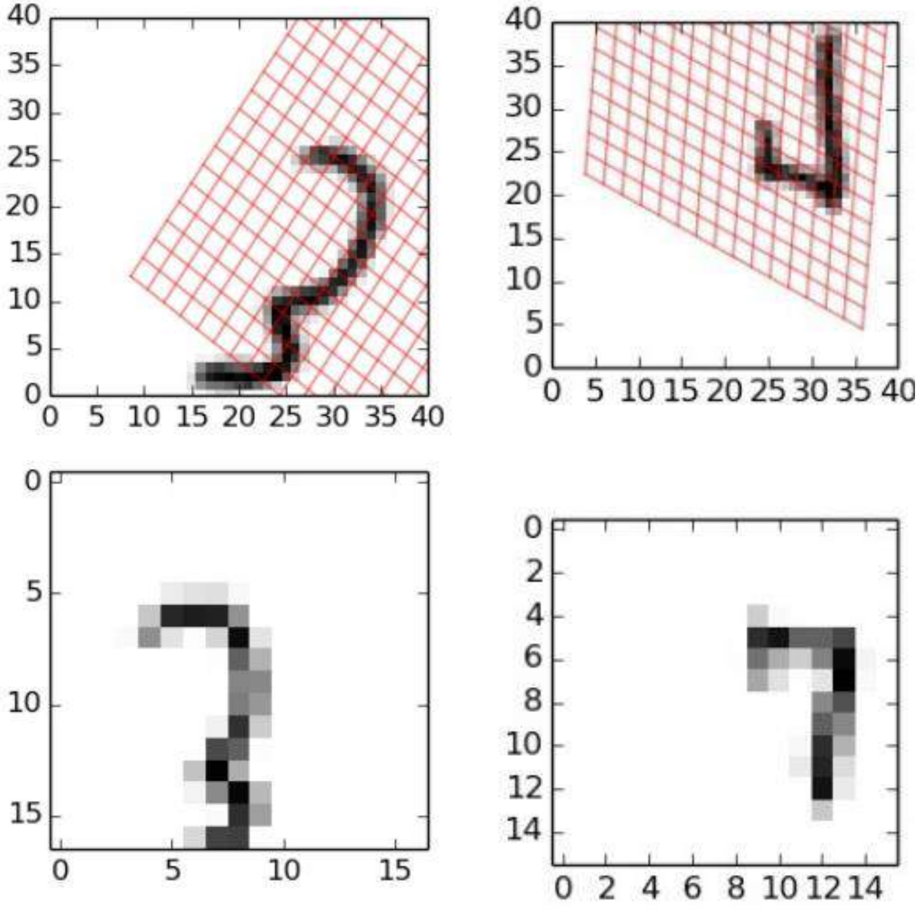
$$S_i = \begin{pmatrix} x_i^S \\ y_i^S \end{pmatrix} = \tau_{A_j}(G_i) = A_j \begin{pmatrix} x_i^{X_{out}} \\ y_i^{X_{out}} \\ 1 \end{pmatrix} = \begin{bmatrix} a_{1,1} & a_{1,2} & a_{1,3} \\ a_{2,1} & a_{2,2} & a_{2,3} \end{bmatrix} \begin{pmatrix} x_i^{X_{out}} \\ y_i^{X_{out}} \\ 1 \end{pmatrix}$$



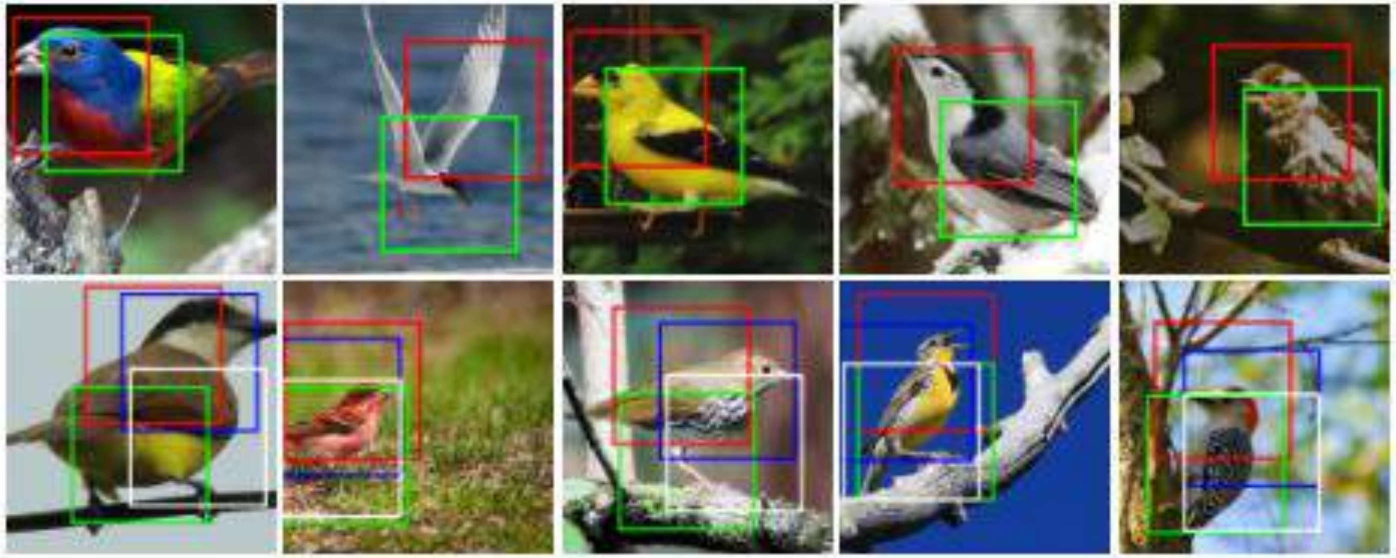
$$X^q_{out,i} = \sum_u^{U_{in}} \sum_v^{V_{in}} X^q_{in,u,v} \max\left(0, 1 - |x^S_i - v|\right) \max\left(0, 1 - |y^S_i - u|\right)$$

$$\forall i \in [1, 2, \ldots, U_{out}V_{out}] \quad \forall q \in [1, 2, \ldots, Q]$$



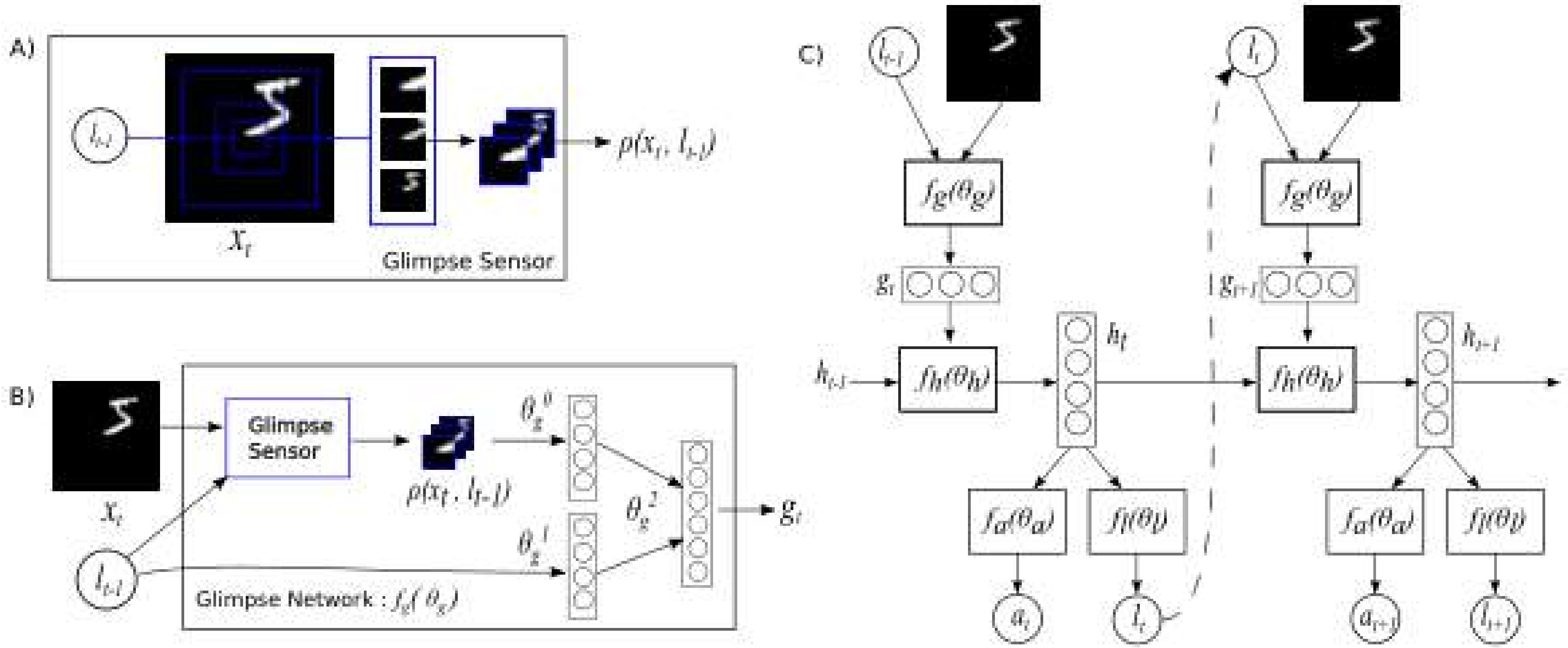


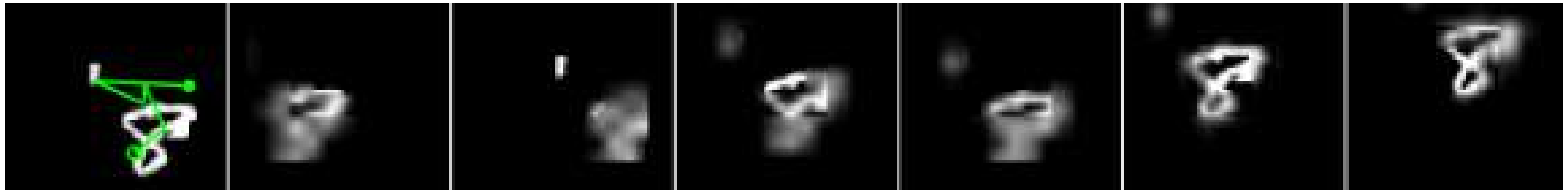
0 5 10 15 20 25 30 35 40
0 5 10 15 20 25 30 35 40
0 5 10 15
0 2 4 6 8 10 12 14



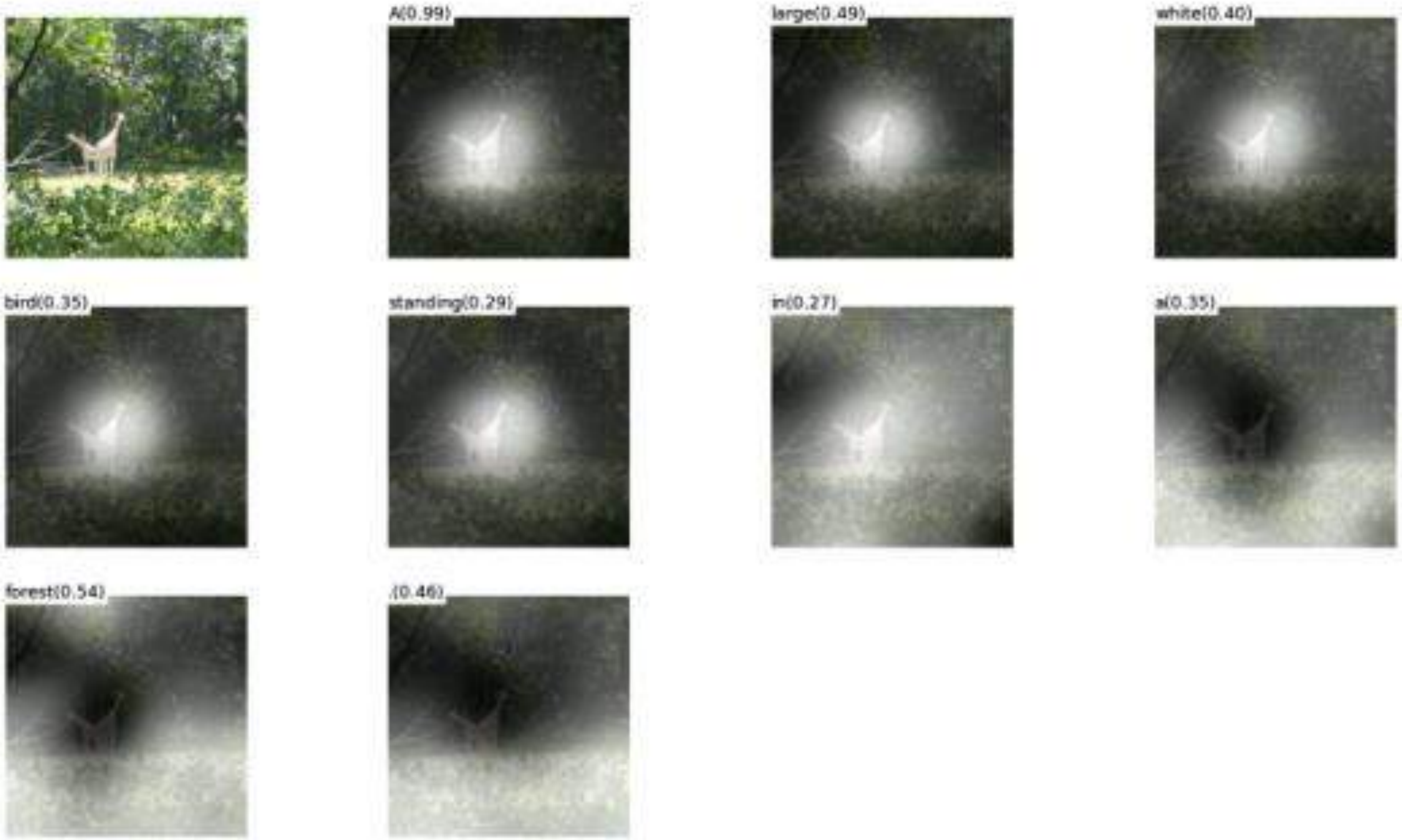
Applications

Glimpses


A)
Glimpse Sensor
B)
Glimpse Sensor
Glimpse Network : $f_g(\theta_g)$
C)



Caption



A(0.99)
large(0.49)
white(0.40)
bird(0.35)
standing(0.29)
in(0.27)
a(0.35)
forest(0.54)
.(0.46)

Video (Sequential data)

(a) Correctly classified as "Pushup"

(f) "soccer juggling"

Pros and cons

- Pros:
  - Learn selectively rather than equally
  - Not limited to computer vision
  - Understandable

- Cons:
  - Hard to train hard attention (RL)
  - Hard to learn $f_{att}(x_t, s_{t-1})$ and classifier simultaneously

Q & A

Thanks!