Deep Learning: Theory and Practice

Recurrent Neural Networks

28-03-2019

Established 1911

INDIAN INSTITUTE OF SCIENCE
भारतीय विज्ञान संस्थान

Introduction

- The standard DNN/CNN paradigms
  - (x,y) - ordered pair of data vectors/images (x) and target (y)
- Moving to sequence data
  - (x(t),y(t)) where this could be sequence to sequence mapping task.
  - (x(t),y) where this could be a sequence to vector mapping task.

Introduction

- Difference between CNNs/DNNs
  - (x(t),y(t)) where this could be sequence to sequence mapping task.
    - Input features / output targets are correlated in time.
    - Unlike standard models where each pair is independent.
    - Need to model dependencies in the sequence over time.

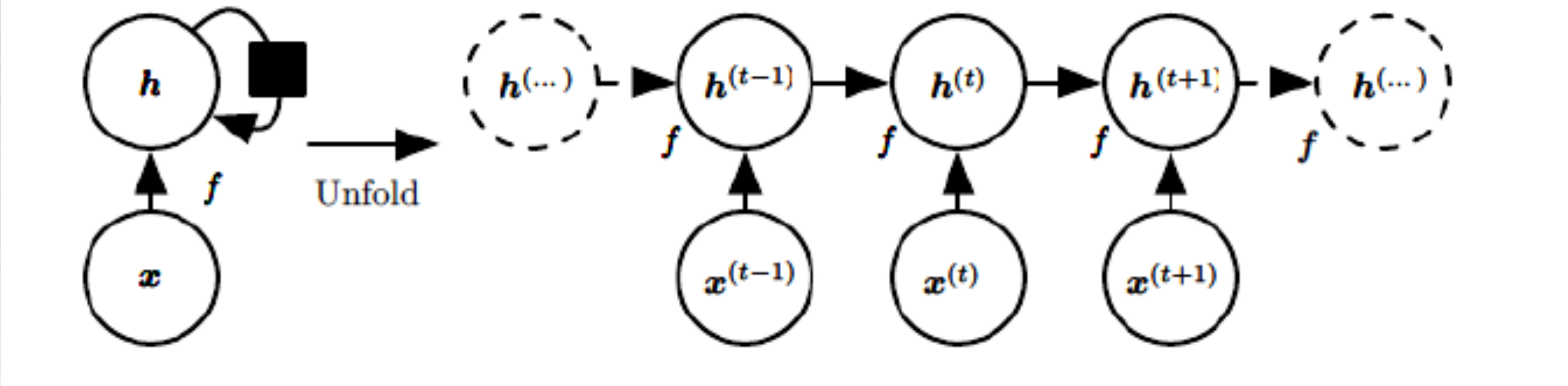
Introduction to Recurrent Networks

$$s^{(t)} = f(s^{(t-1)}; \theta),$$

$$\begin{aligned} s^{(3)} &= f(s^{(2)}; \theta) \\ &= f(f(s^{(1)}; \theta); \theta) \end{aligned}$$

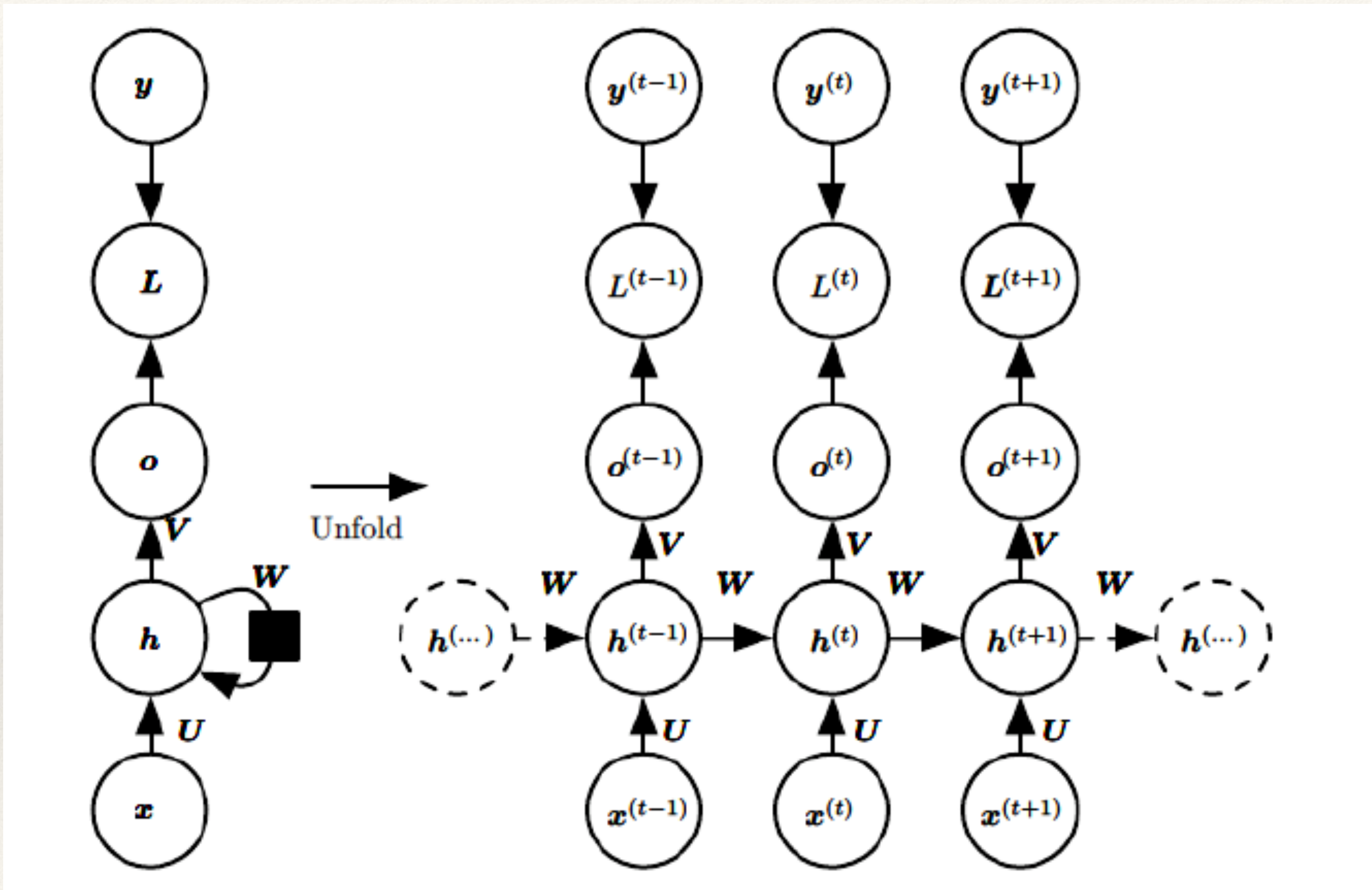
$$s^{(t)} = f(s^{(t-1)}, x^{(t)}; \theta),$$

$$h^{(t)} = f(h^{(t-1)}, x^{(t)}; \theta),$$



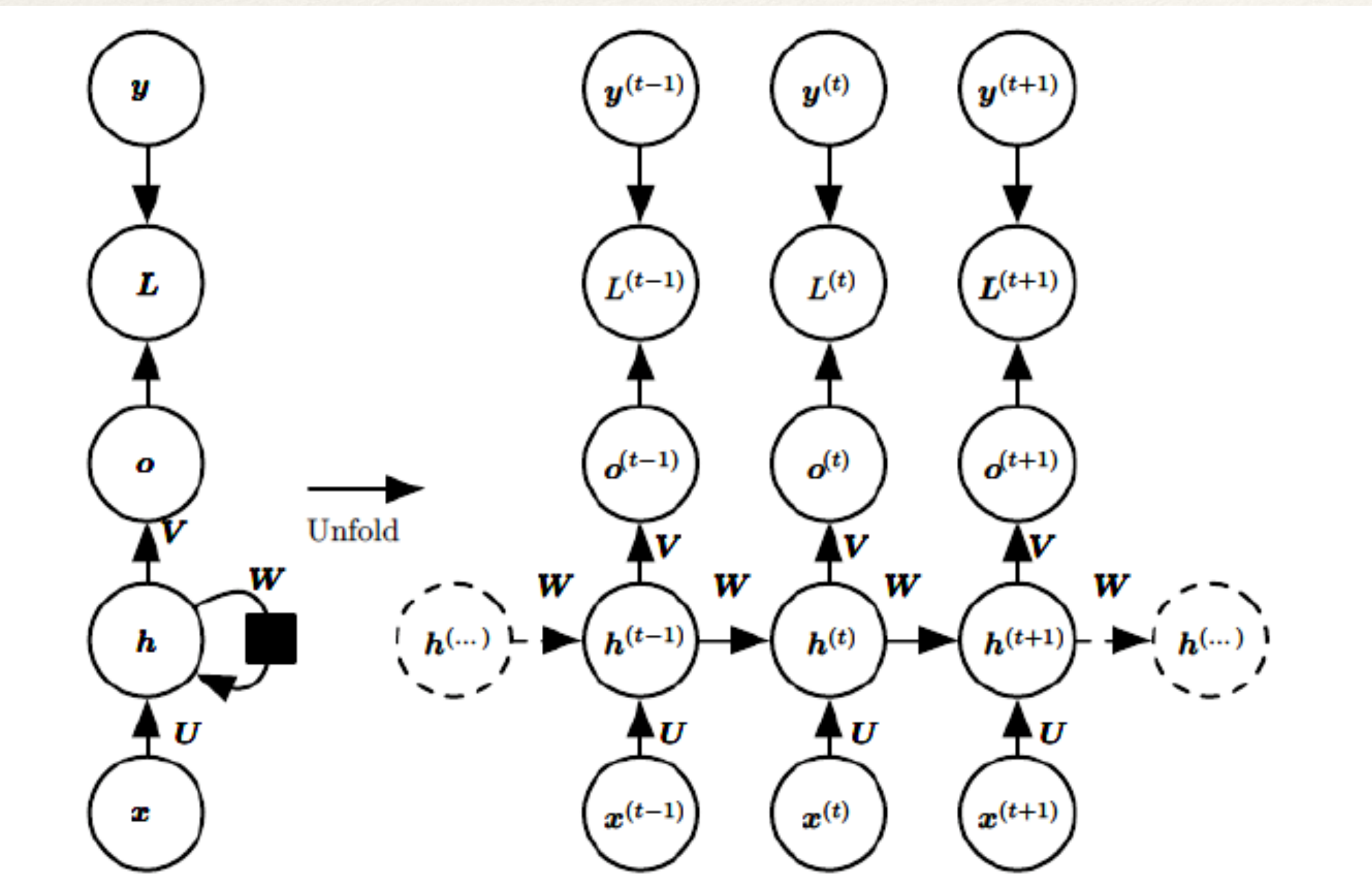
"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

Recurrent Networks



"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

Recurrent Networks



$$
\begin{aligned}
\boldsymbol{a}^{(t)} &= \boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)} \\
\boldsymbol{h}^{(t)} &= \tanh(\boldsymbol{a}^{(t)}) \\
\boldsymbol{o}^{(t)} &= \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)} \\
\hat{\boldsymbol{y}}^{(t)} &= \operatorname{softmax}(\boldsymbol{o}^{(t)})
\end{aligned}
$$

$$
\begin{aligned}
& L\left(\{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(\tau)}\}, \{\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(\tau)}\}\right) \\
&= \sum_t L^{(t)} \\
&= -\sum_t \log p_{\text{model}}\left(y^{(t)} \mid \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(t)}\}\right)
\end{aligned}
$$

"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

Back Propagation in RNNs

$$
\begin{aligned}
\boldsymbol{a}^{(t)} &= \boldsymbol{b} + \boldsymbol{W}\boldsymbol{h}^{(t-1)} + \boldsymbol{U}\boldsymbol{x}^{(t)} \\
\boldsymbol{h}^{(t)} &= \tanh(\boldsymbol{a}^{(t)}) \\
\boldsymbol{o}^{(t)} &= \boldsymbol{c} + \boldsymbol{V}\boldsymbol{h}^{(t)} \\
\hat{\boldsymbol{y}}^{(t)} &= \operatorname{softmax}(\boldsymbol{o}^{(t)})
\end{aligned}
$$

Model Parameters

$\boldsymbol{U}$, $\boldsymbol{V}$, $\boldsymbol{W}$, $\boldsymbol{b}$ and $\boldsymbol{c}$

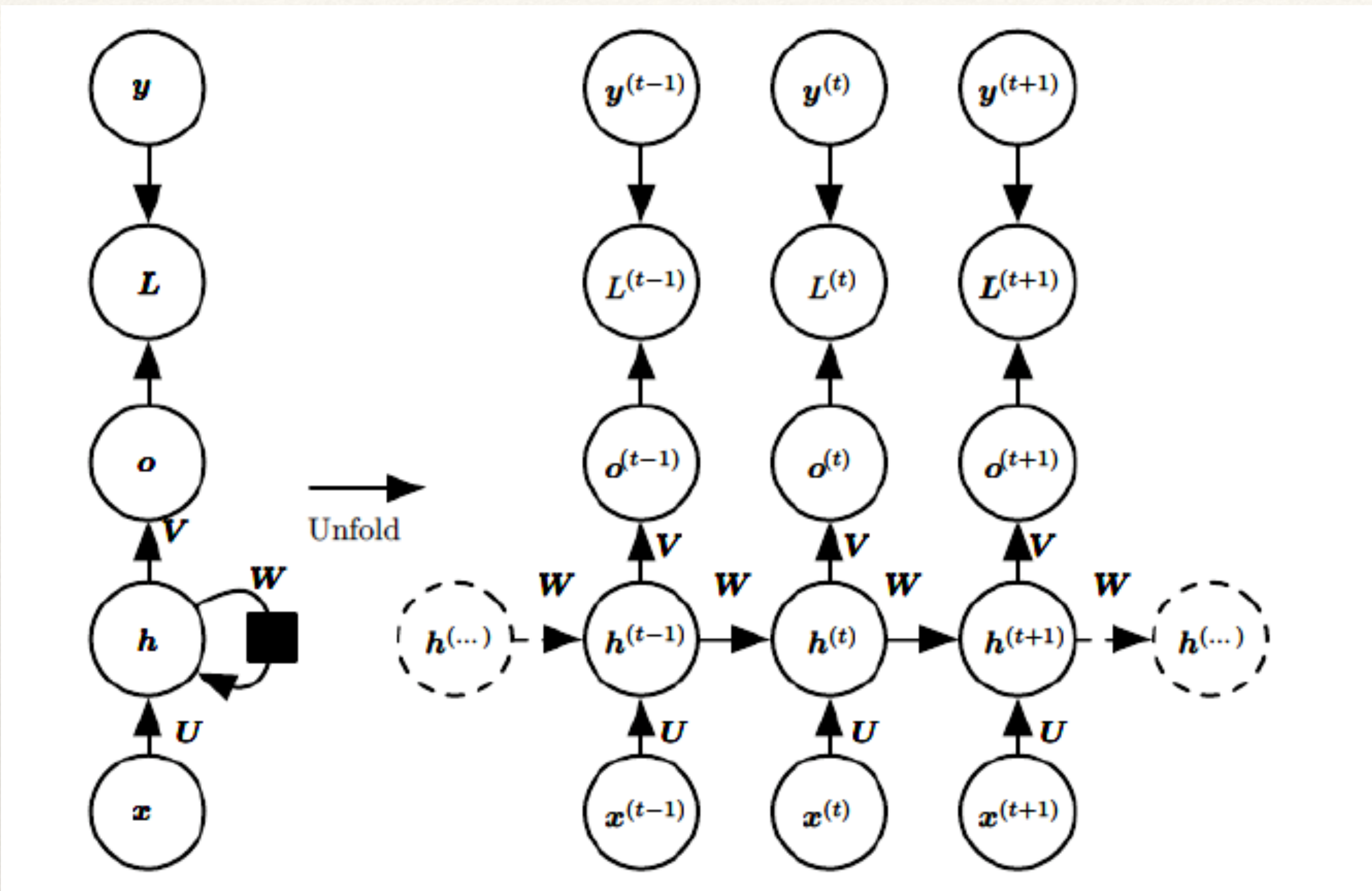
$$
\begin{aligned}
& L\left(\{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(\tau)}\}, \{\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(\tau)}\}\right) \\
&= \sum_t L^{(t)} \\
&= -\sum_t \log p_{\text{model}}\left(y^{(t)} \mid \{\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(t)}\}\right)
\end{aligned}
$$

Gradient Descent

$$\frac{\partial L}{\partial L^{(t)}} = 1.$$

$$(\nabla_{\boldsymbol{o}^{(t)}} L)_i = \frac{\partial L}{\partial o_i^{(t)}} = \frac{\partial L}{\partial L^{(t)}} \frac{\partial L^{(t)}}{\partial o_i^{(t)}} = \hat{y}_i^{(t)} - \mathbf{1}_{i,y^{(t)}}$$

Recurrent Networks

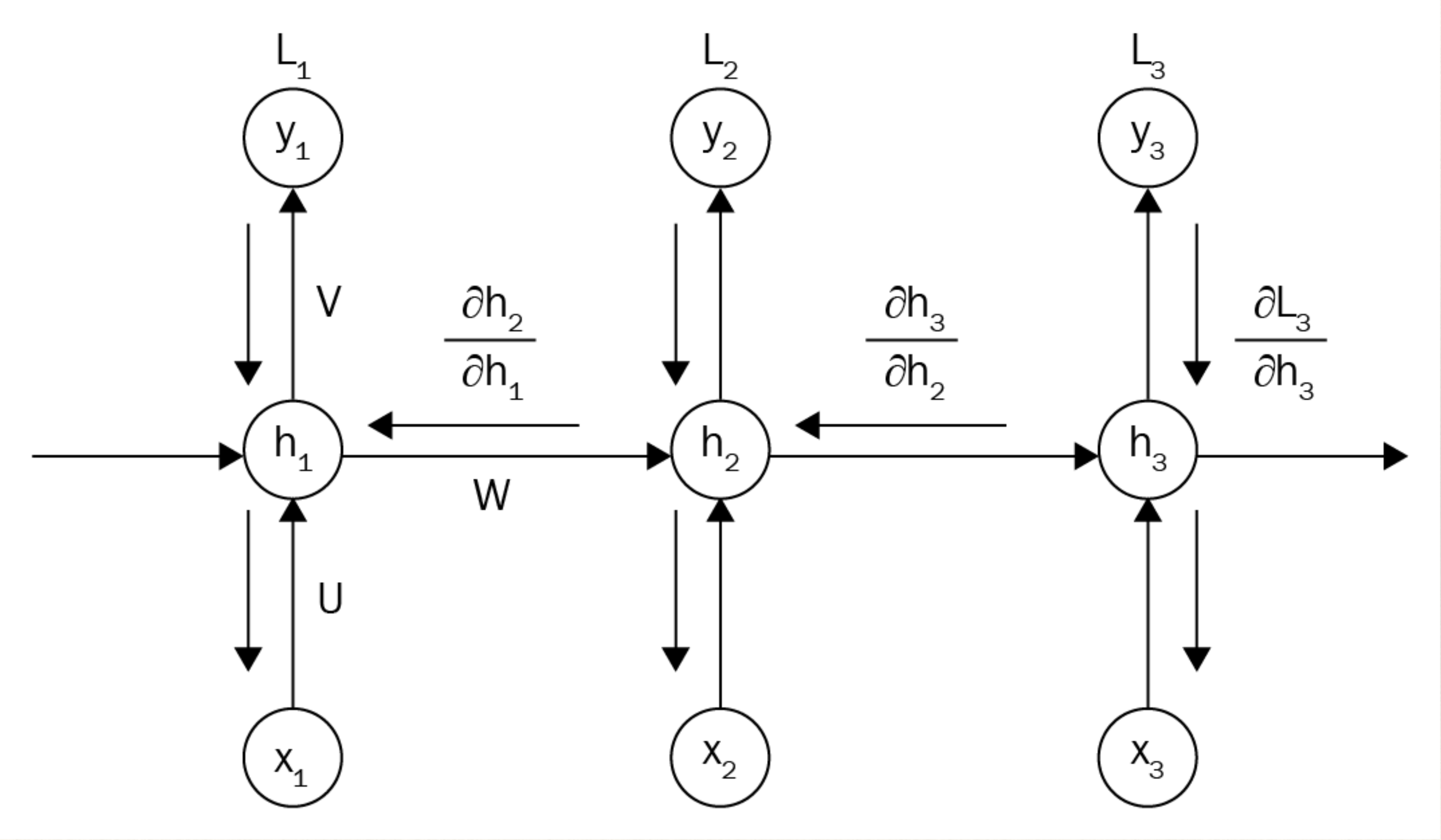


$$(\nabla_{\boldsymbol{o}^{(t)}} L)_i = \frac{\partial L}{\partial o_i^{(t)}} = \frac{\partial L}{\partial L^{(t)}} \frac{\partial L^{(t)}}{\partial o_i^{(t)}} = \hat{y}_i^{(t)} - \mathbf{1}_{i,y^{(t)}}$$

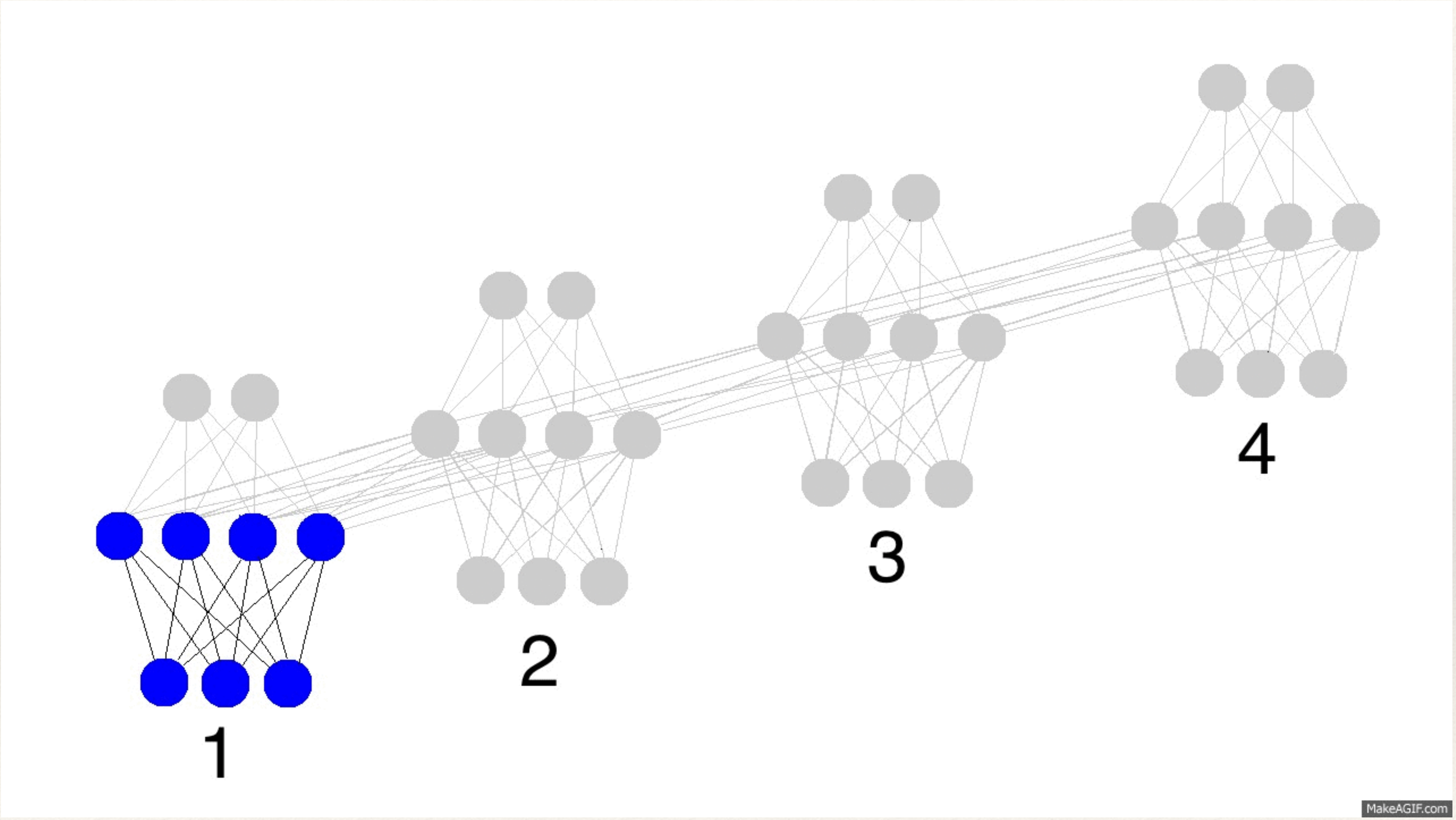
$$\nabla_{\boldsymbol{h}^{(\tau)}} L = \boldsymbol{V}^\top \nabla_{\boldsymbol{o}^{(\tau)}} L.$$

$$\begin{aligned}
\nabla_{\boldsymbol{h}^{(t)}} L &= \left(\frac{\partial \boldsymbol{h}^{(t+1)}}{\partial \boldsymbol{h}^{(t)}}\right)^\top (\nabla_{\boldsymbol{h}^{(t+1)}} L) + \left(\frac{\partial \boldsymbol{o}^{(t)}}{\partial \boldsymbol{h}^{(t)}}\right)^\top (\nabla_{\boldsymbol{o}^{(t)}} L) \\
&= \boldsymbol{W}^\top (\nabla_{\boldsymbol{h}^{(t+1)}} L) \operatorname{diag}\left(1 - \left(\boldsymbol{h}^{(t+1)}\right)^2\right) + \boldsymbol{V}^\top (\nabla_{\boldsymbol{o}^{(t)}} L)
\end{aligned}$$

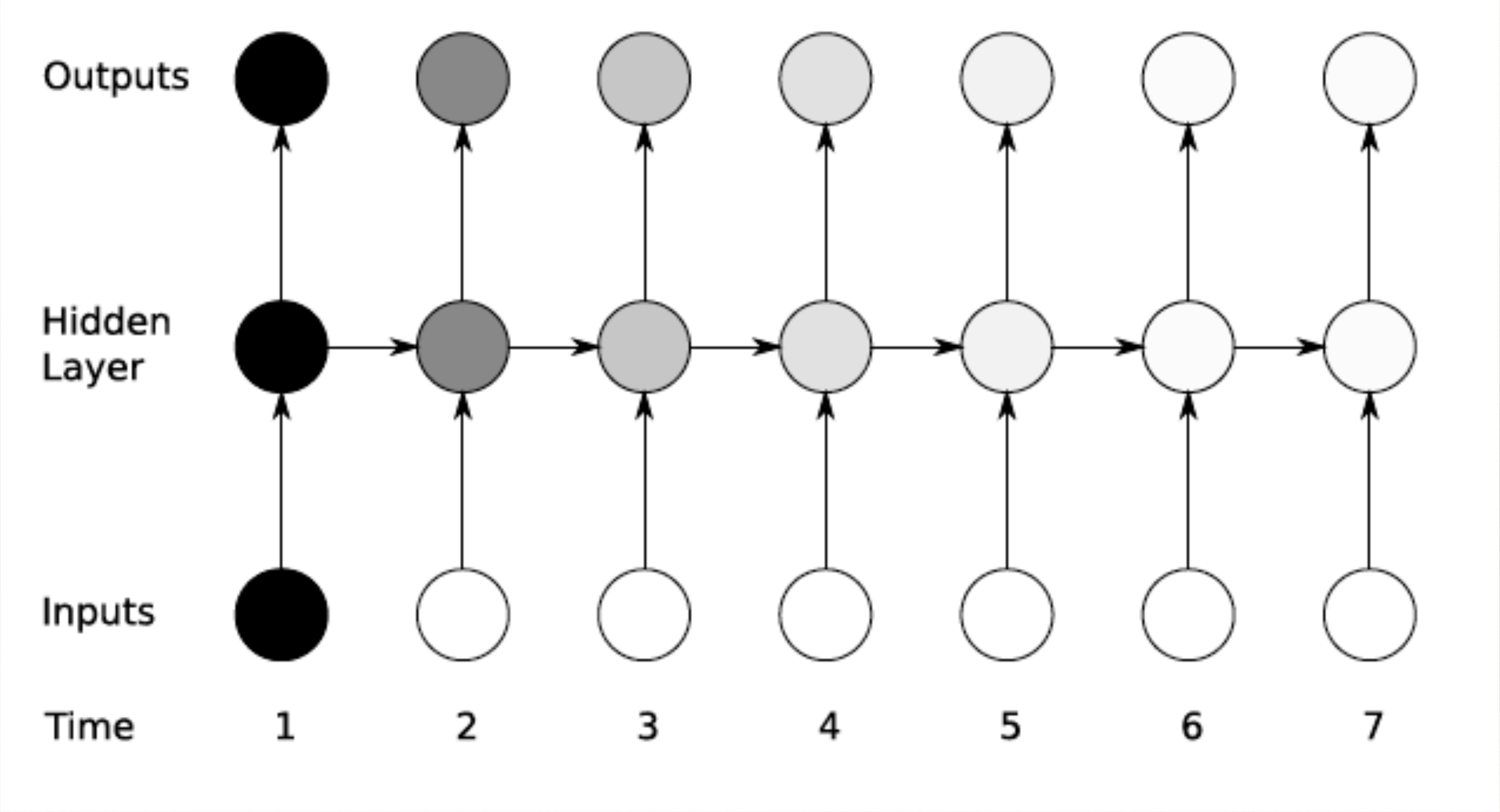
Back Propagation Through Time



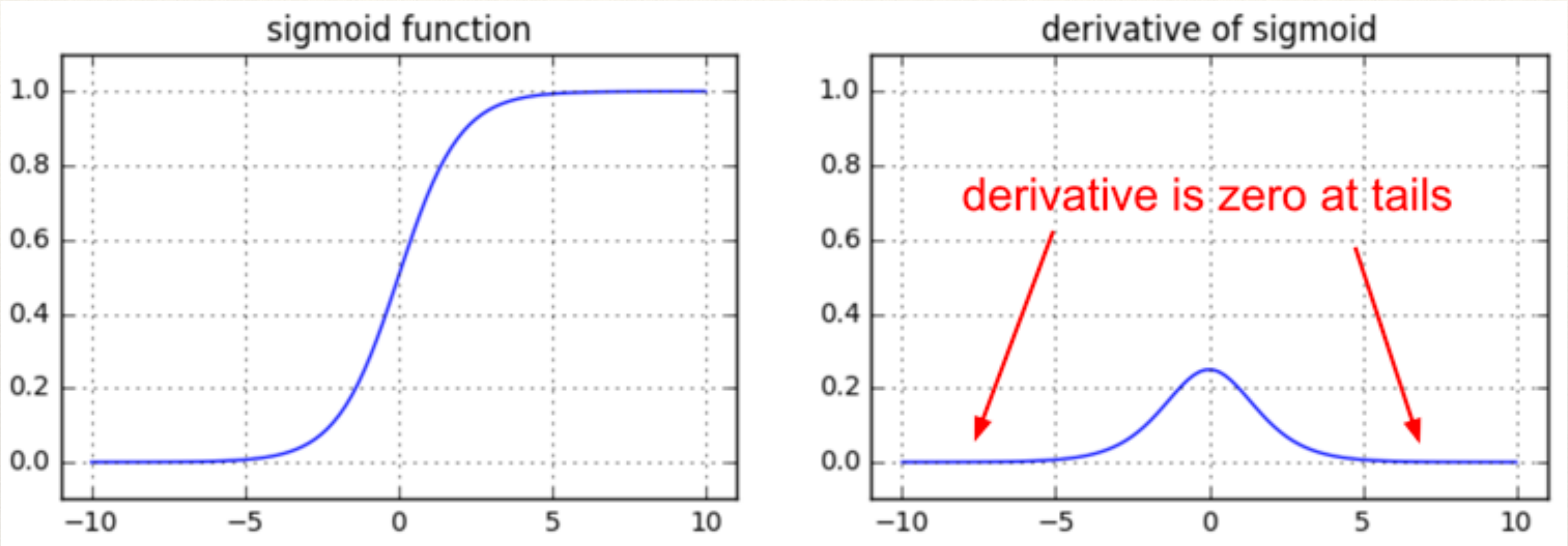
Back Propagation Through Time



Long-term Dependency Issues

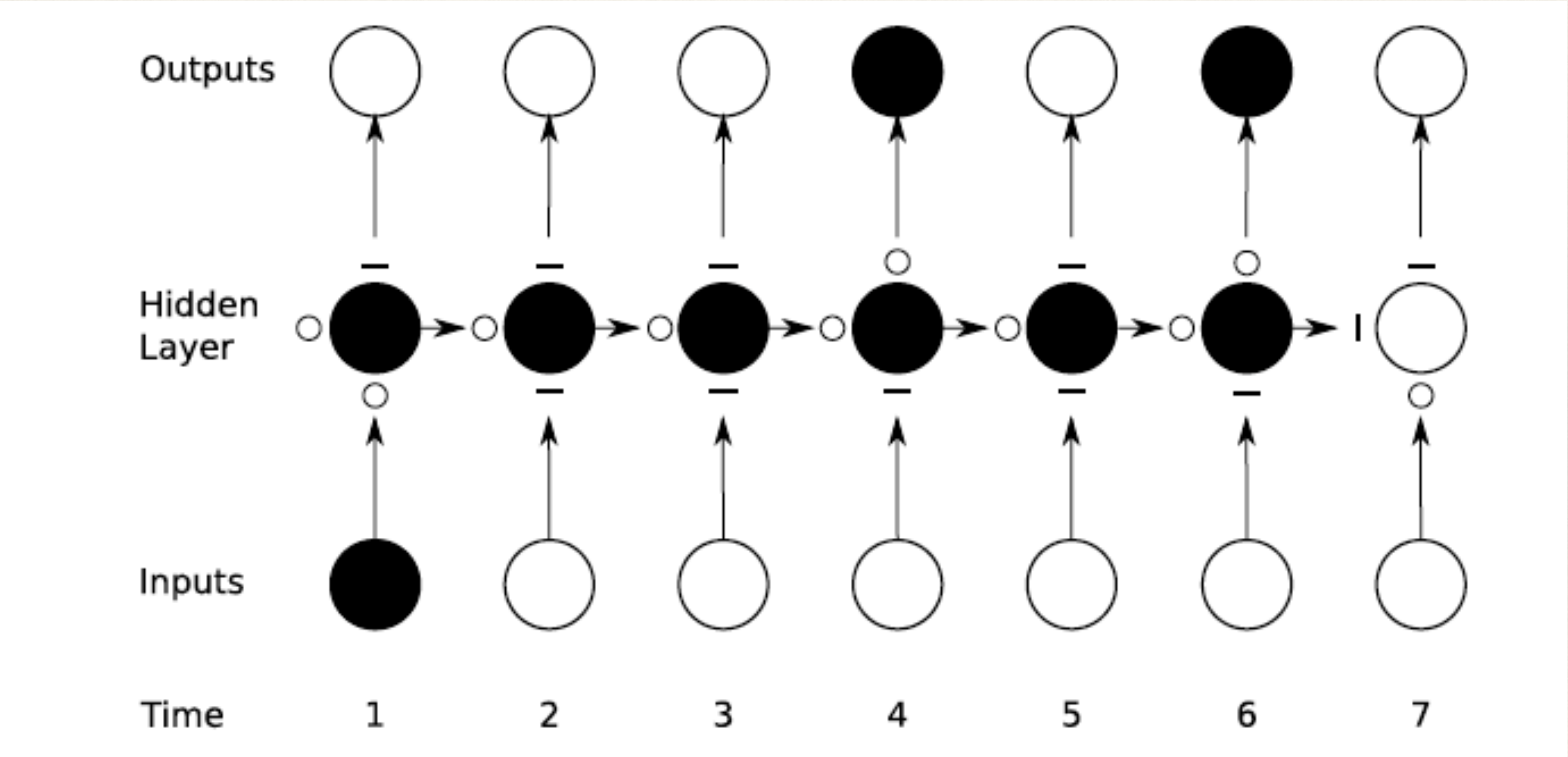


Vanishing/Exploding Gradients

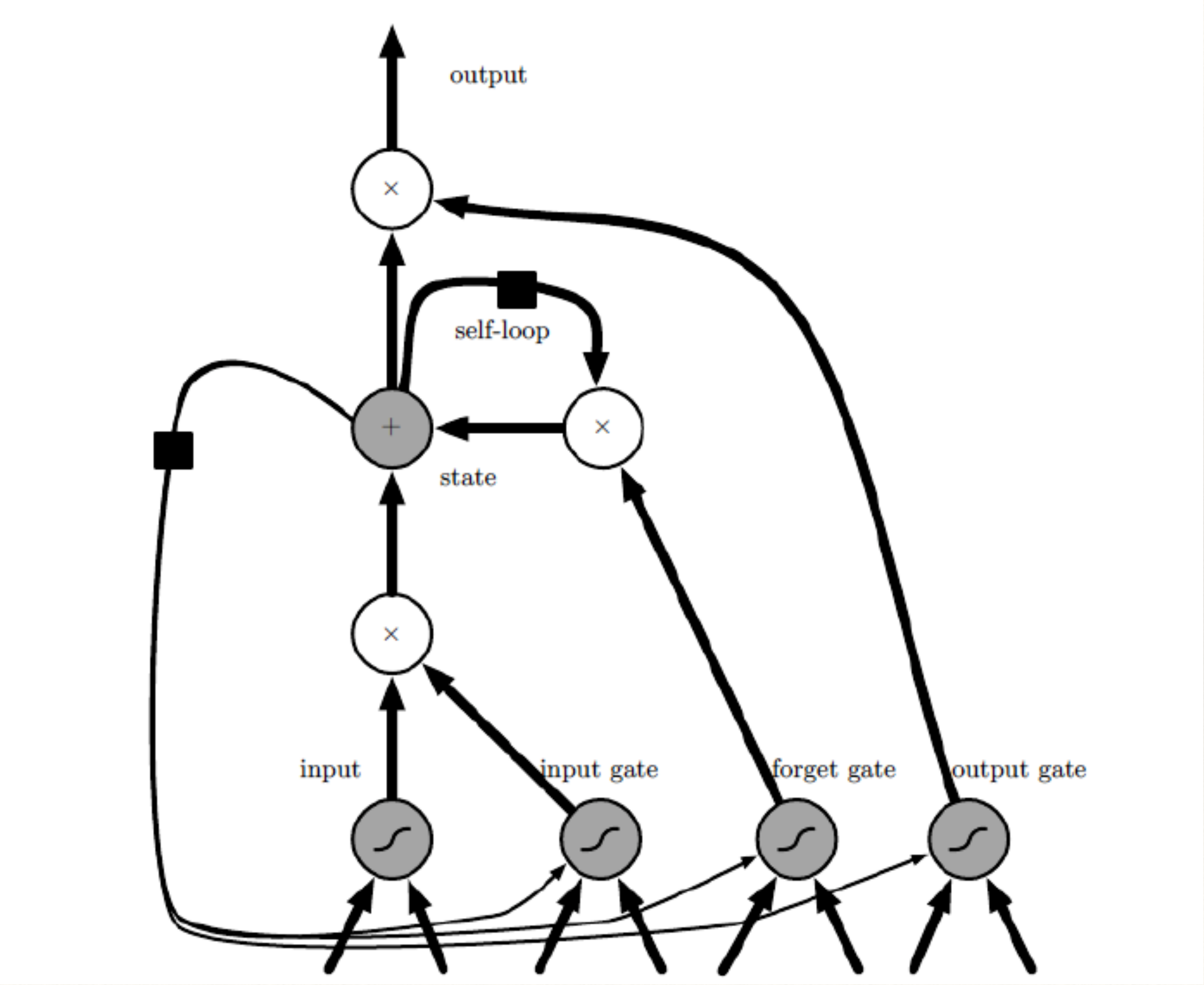


- Gradients either vanish or explode
  - Initial frames may not contribute to gradient computations or may contribute too much.

Long-Short Term Memory



Long Short Term Memory Networks



LSTM Cell

f - sigmoid function
g, h - tanh function

Input Gate

$$a_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} b_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1}$$

$$b_\iota^t = f(a_\iota^t)$$

Forget Gate

$$a_\phi^t = \sum_{i=1}^{I} w_{i\phi} x_i^t + \sum_{h=1}^{H} w_{h\phi} b_h^{t-1} + \sum_{c=1}^{C} w_{c\phi} s_c^{t-1}$$

$$b_\phi^t = f(a_\phi^t)$$

Cell

$$a_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} b_h^{t-1}$$

$$s_c^t = b_\phi^t s_c^{t-1} + b_\iota^t g(a_c^t)$$

Output Gate

$$a_\omega^t = \sum_{i=1}^{I} w_{i\omega} x_i^t + \sum_{h=1}^{H} w_{h\omega} b_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^t$$

$$b_\omega^t = f(a_\omega^t)$$

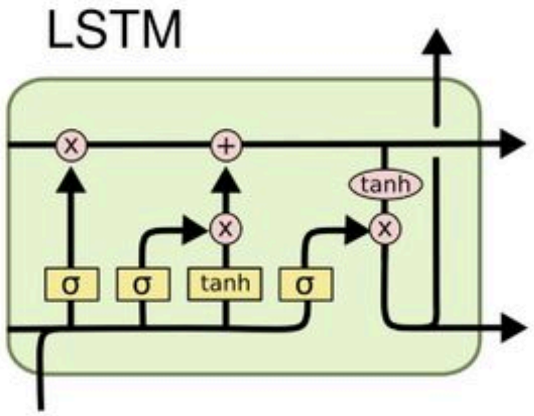
LSTM output

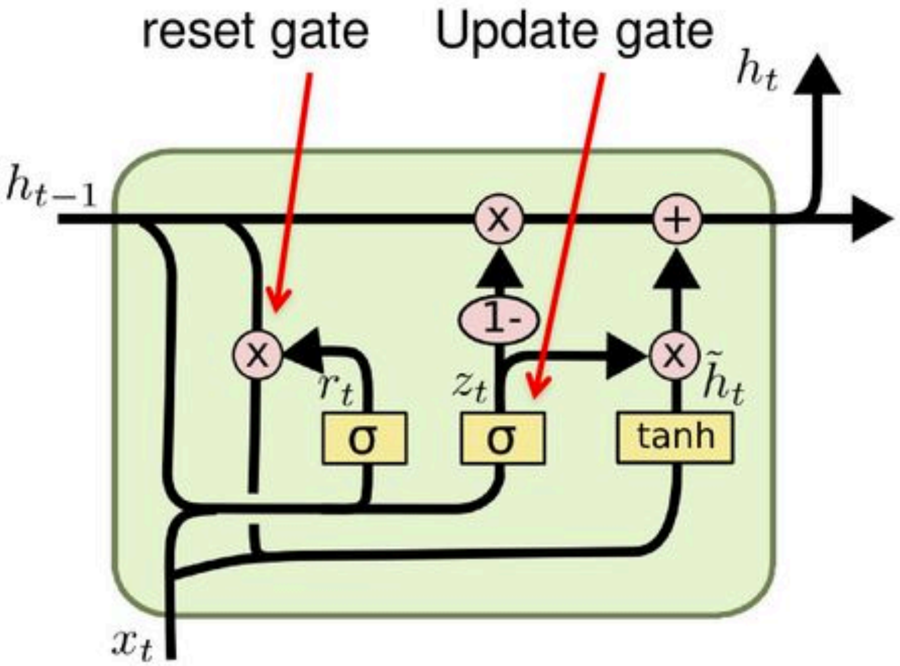
$$b_c^t = b_\omega^t h(s_c^t)$$

Gated Recurrent Units (GRU)

GRU – gated recurrent unit

(more compression)





$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

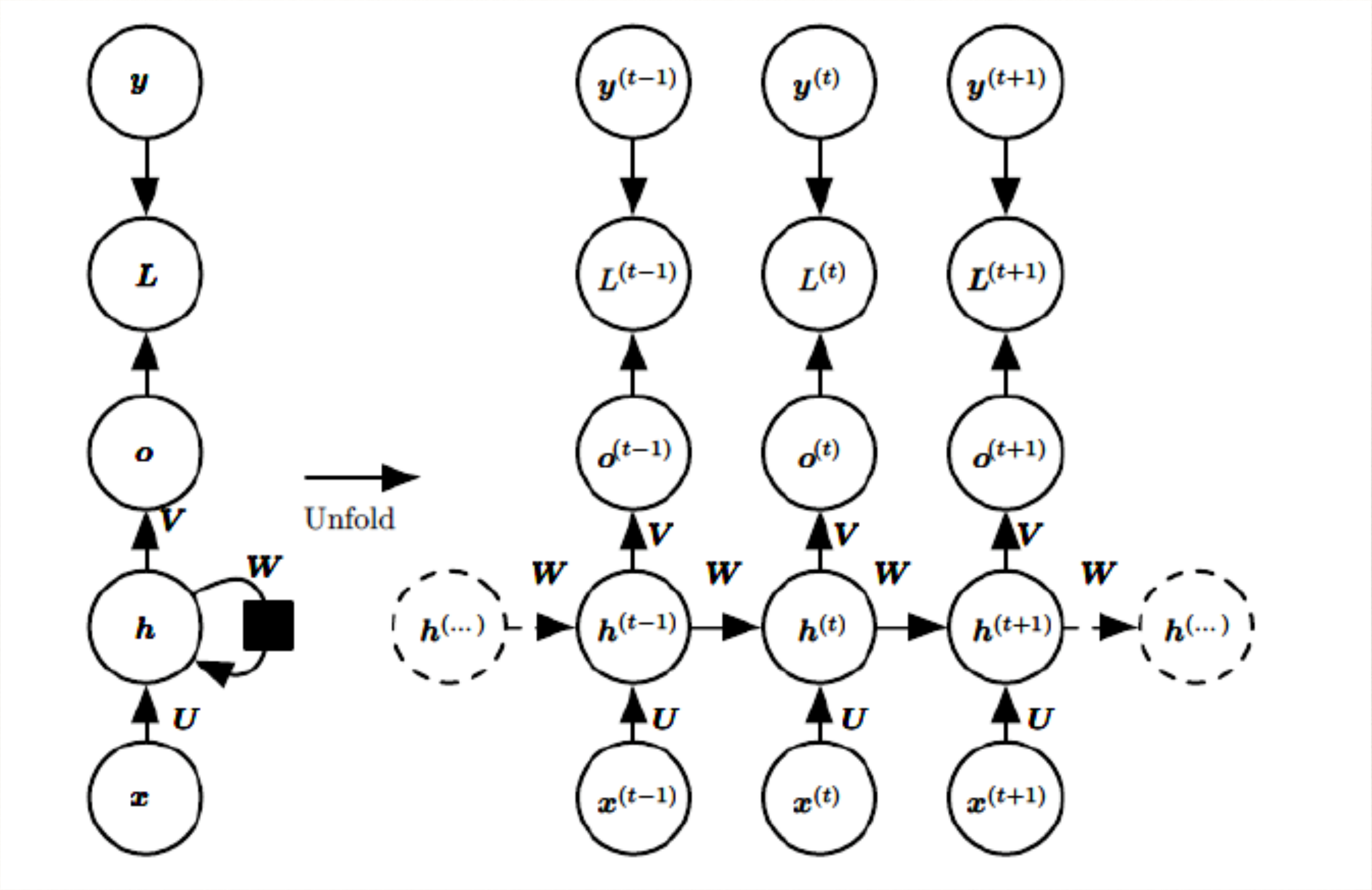
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

It combines the forget and input into a single update gate.
It also merges the cell state and hidden state. This is simpler than LSTM. There are many other variants too.

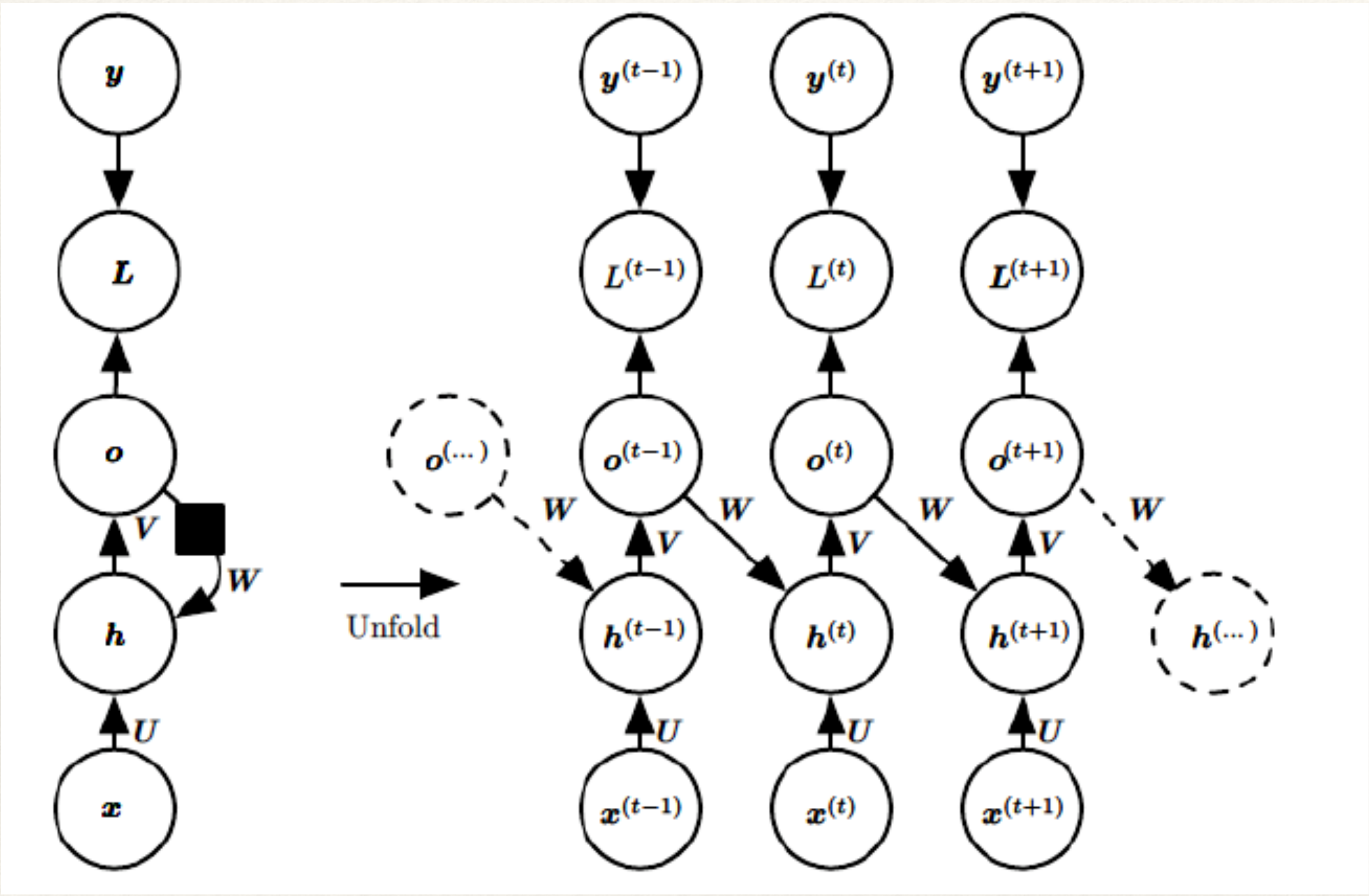
X,*: element-wise multiply

Standard Recurrent Networks



"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

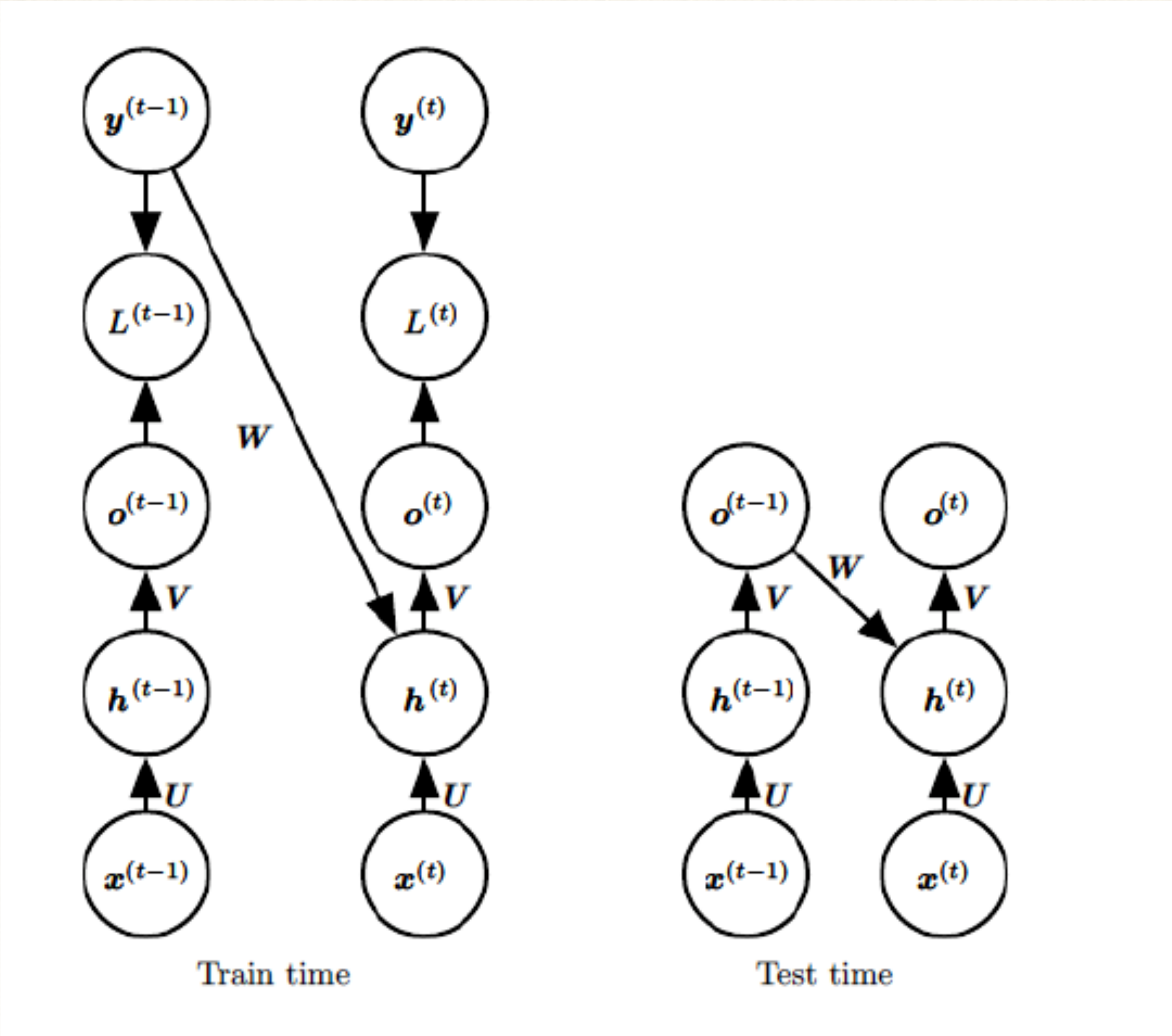
Other Recurrent Networks



Teacher Forcing Networks

"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

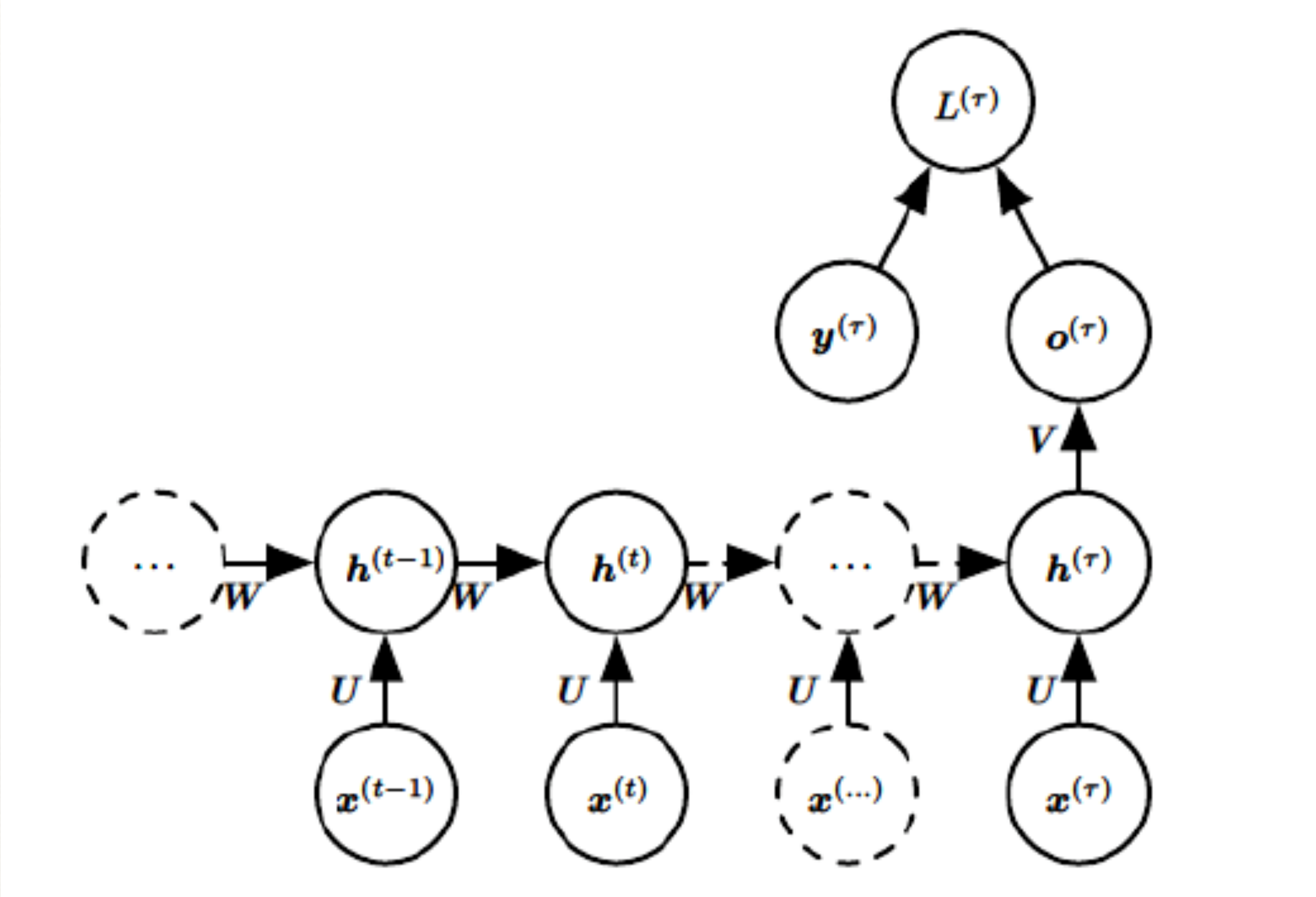
Recurrent Networks



Teacher Forcing Networks

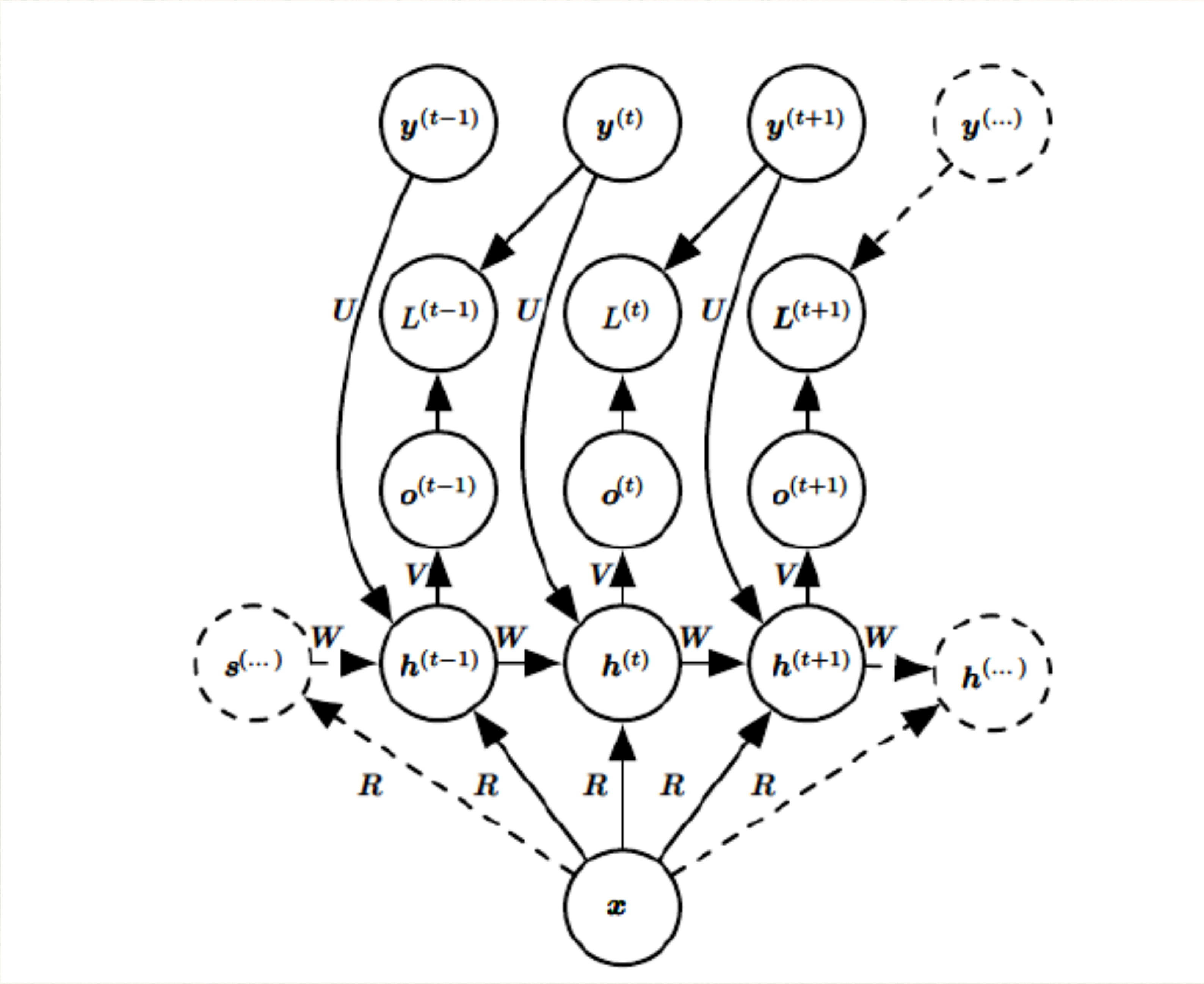
"Deep Learning", Ian Goodfellow, Yoshua Bengio, Aaron Courville

Recurrent Networks



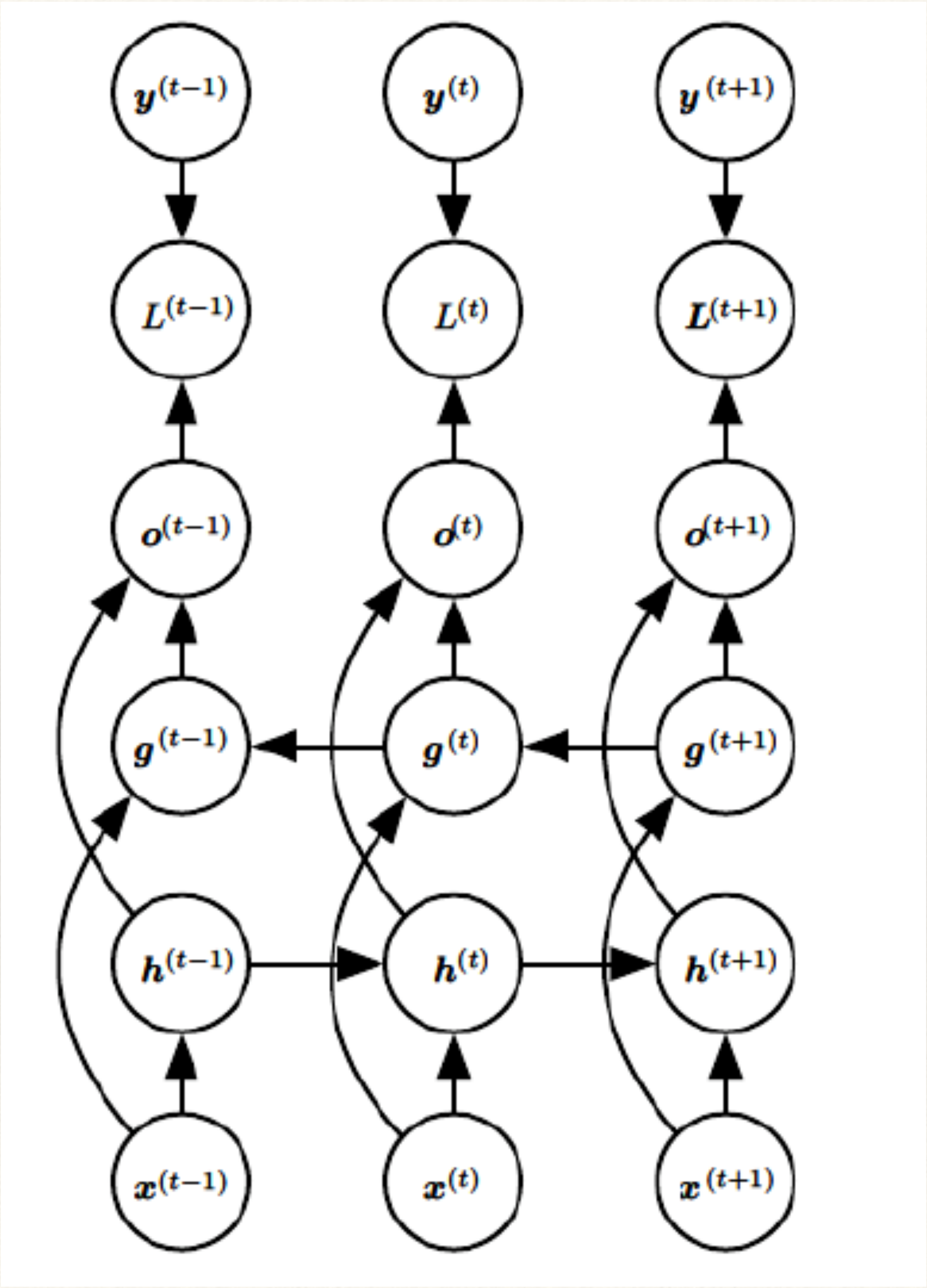
Multiple Input
Single Output

Recurrent Networks



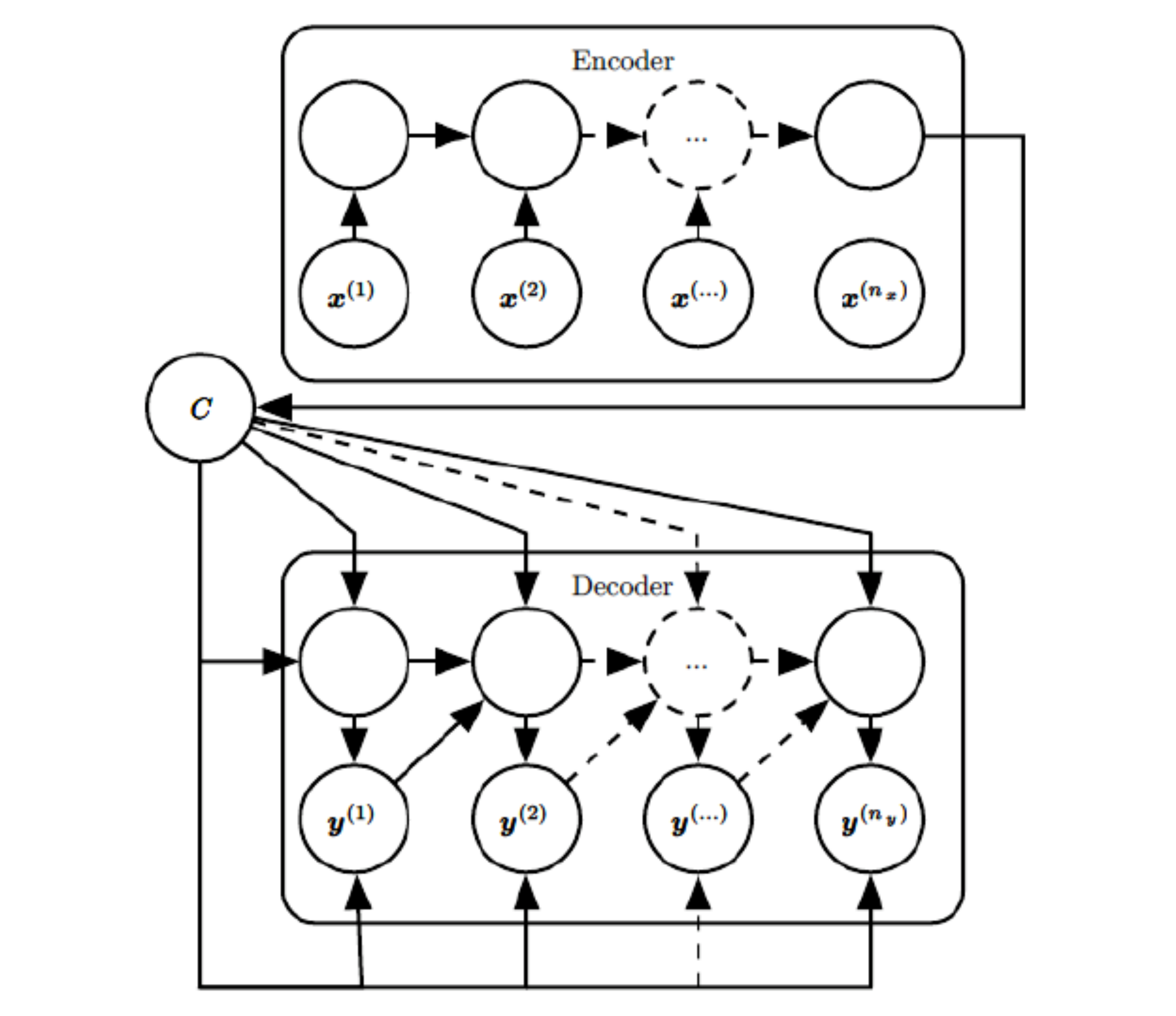
Single Input Multiple Output

Recurrent Networks



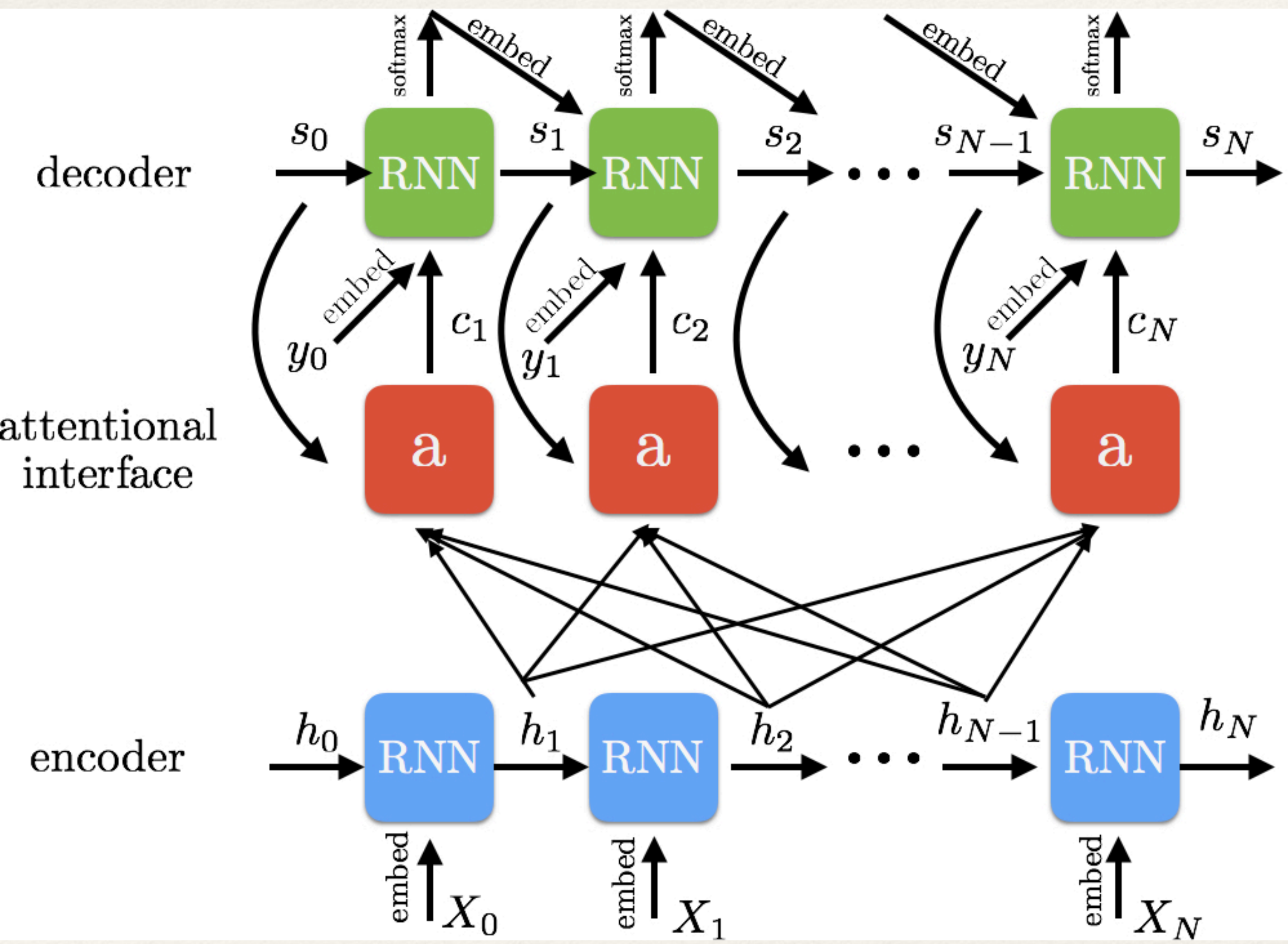
Bi-directional Networks

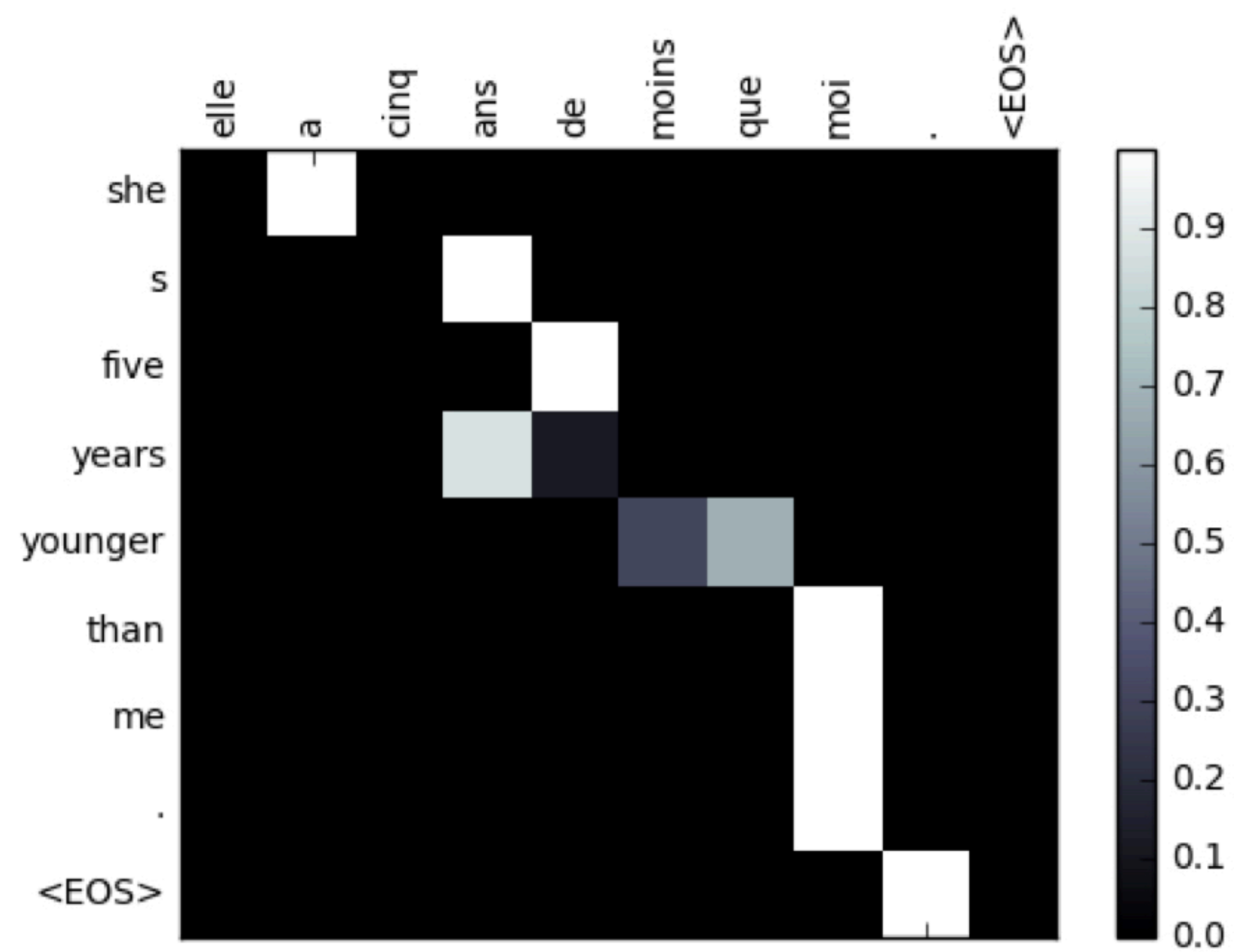
Recurrent Networks



Sequence to Sequence Mapping Networks

Attention Models





Encoder - Decoder Networks with Attention