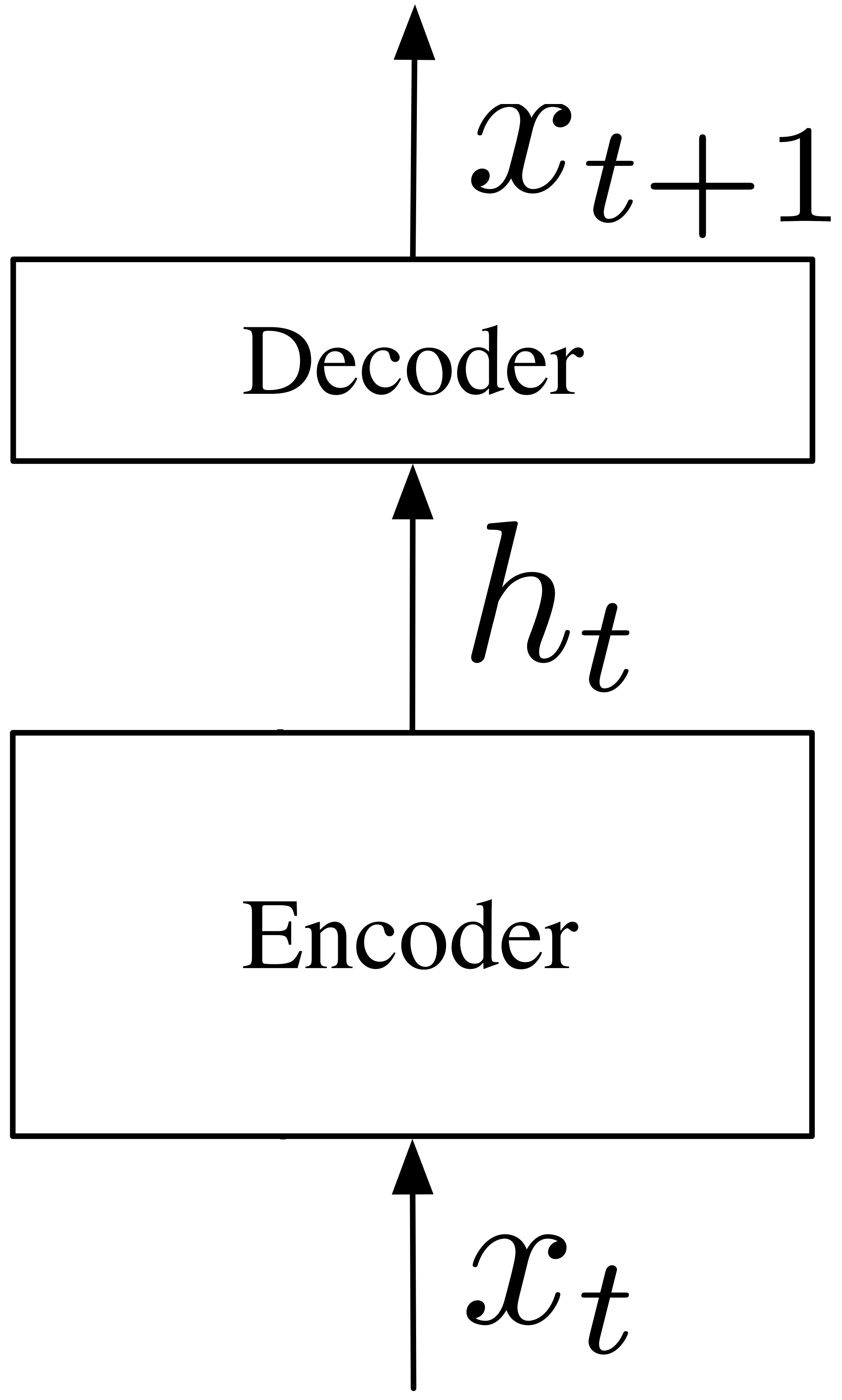
Recurrent Neural Networks

Elman Network

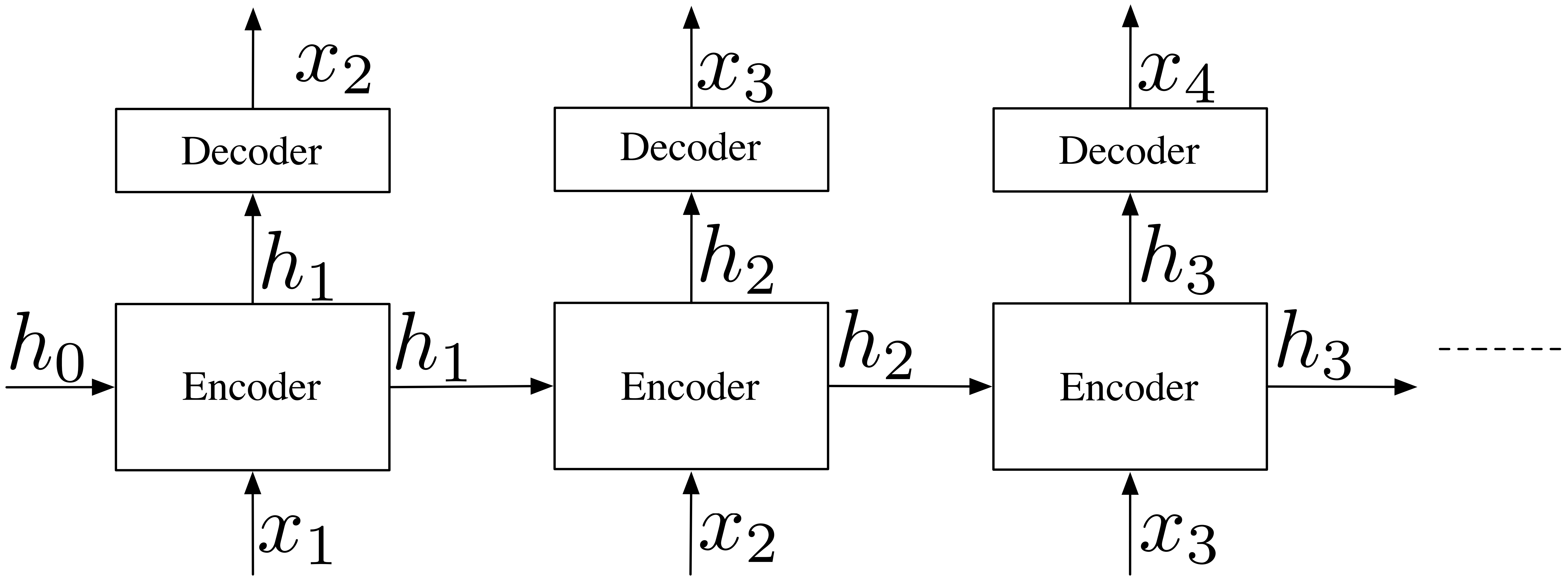
$$h_t = \sigma(W_1 x_t + W_2 h_{t-1})$$

$$x_{t+1} = \rho(W_3 h_t)$$

Can be viewed as a non-linear IIR Filter



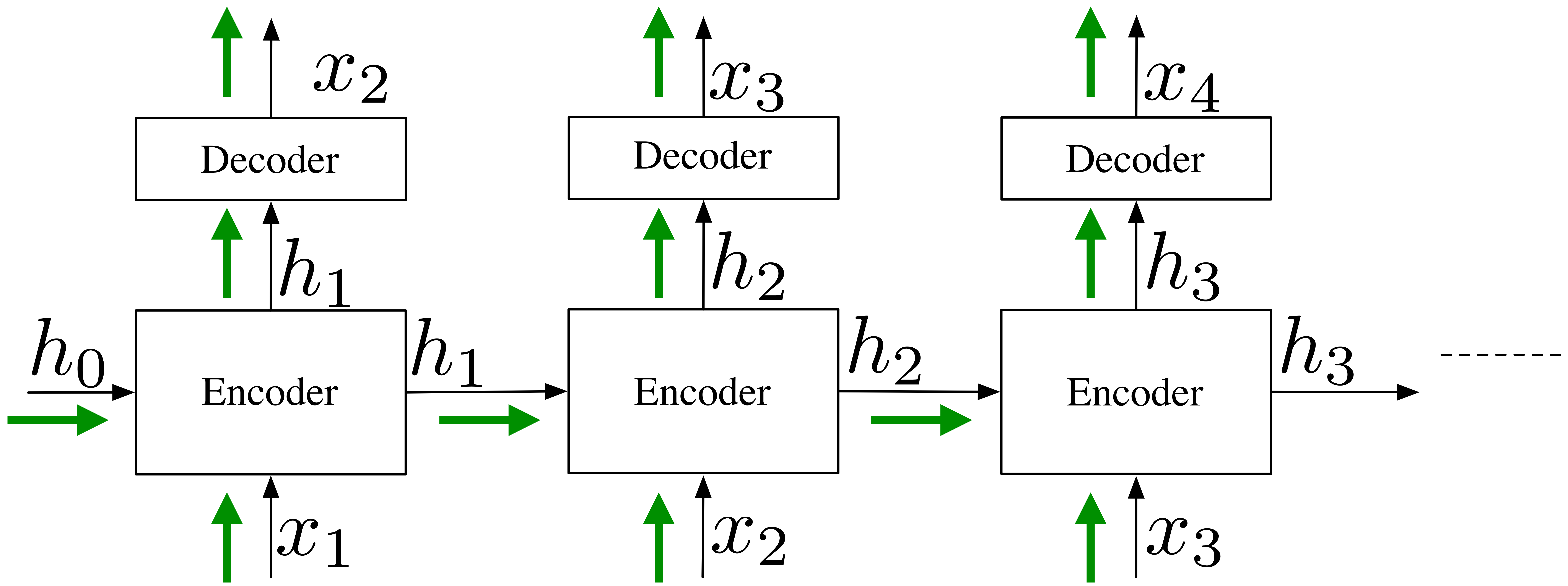
Elman Network



$$h_t = \sigma(W_1 x_t + W_2 h_{t-1})$$

$$x_{t+1} = \rho(W_3 h_t)$$

Training: Back Propagation Through Time

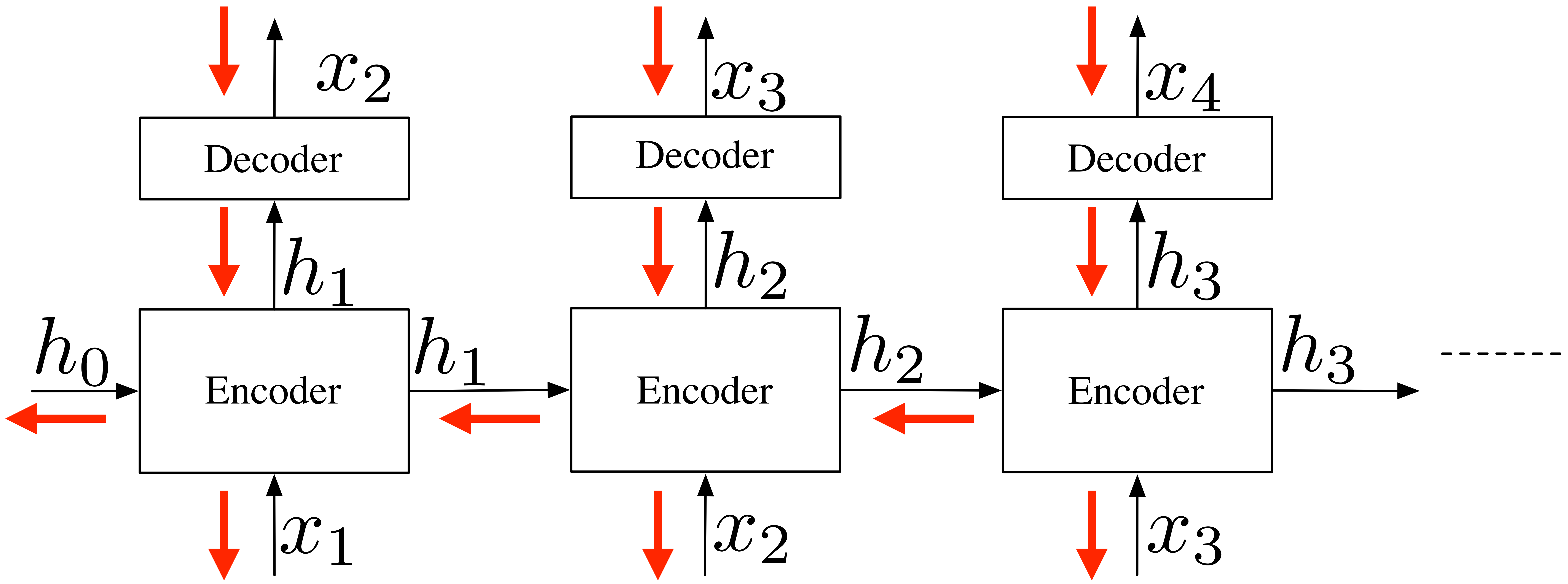


Forward Pass

Pass inputs through the unrolled network

compute the error at every time step

Training: Back Propagation Through Time



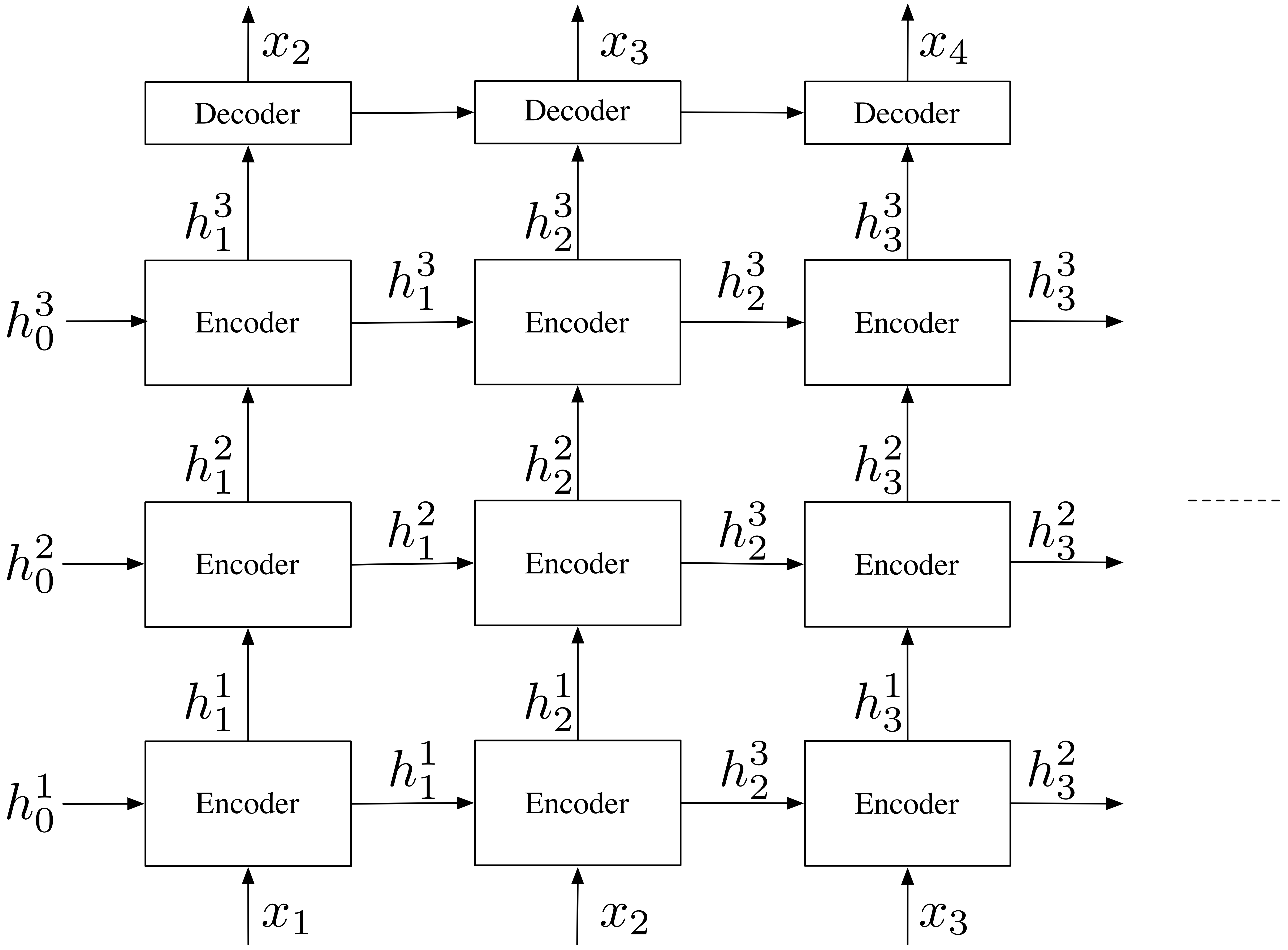
Backward Pass

Compute the contribution of error at every time step

accumulate this gradient while going back in time

update the parameters

Deep RNNs


x_2
x_3
x_4
Decoder
Decoder
Decoder
h_1^3
h_2^3
h_3^3
h_0^3
Encoder
h_1^3
Encoder
h_2^3
Encoder
h_3^3
h_1^2
h_2^2
h_3^2
h_0^2
Encoder
h_1^2
Encoder
h_2^3
Encoder
h_3^2
h_1^1
h_2^1
h_3^1
h_0^1
Encoder
h_1^1
Encoder
h_2^3
Encoder
h_3^2
x_1
x_2
x_3

Shortcoming of Elman Nets

Exploding gradients

Vanishing gradients

Unable to capture long-term dependencies

Training is somewhat brittle

Long Short-Term Memory (LSTM)

recently gained a lot of popularity

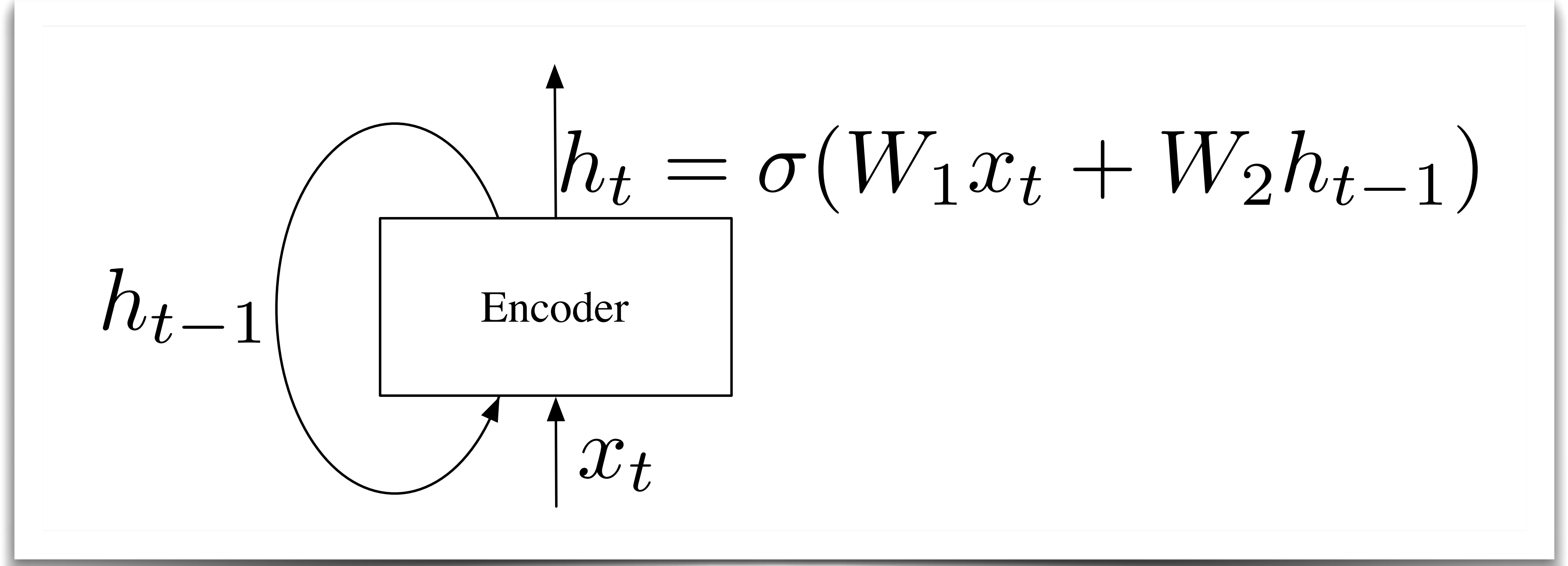
have explicit memory “cells” to store short-term activations

the presence of additional gates partly alleviates the vanishing gradient problem

multi-layer versions shown to work quite well on tasks which have“medium term” dependencies

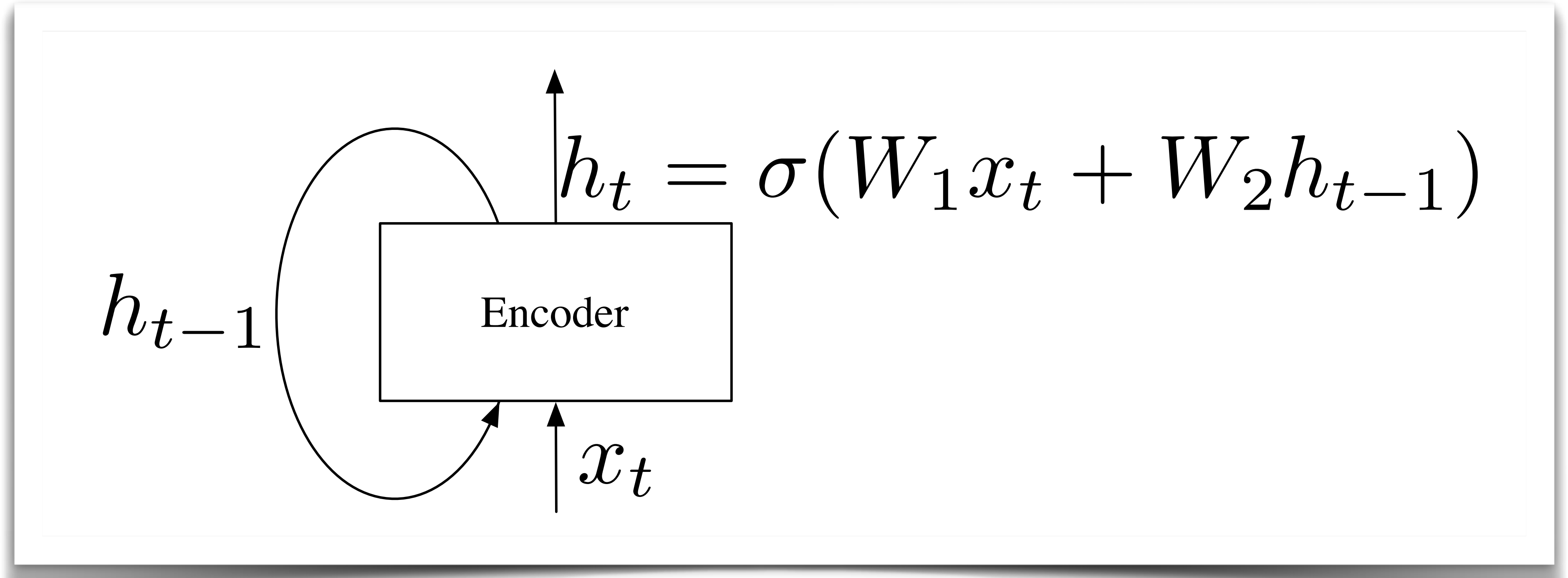
Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)

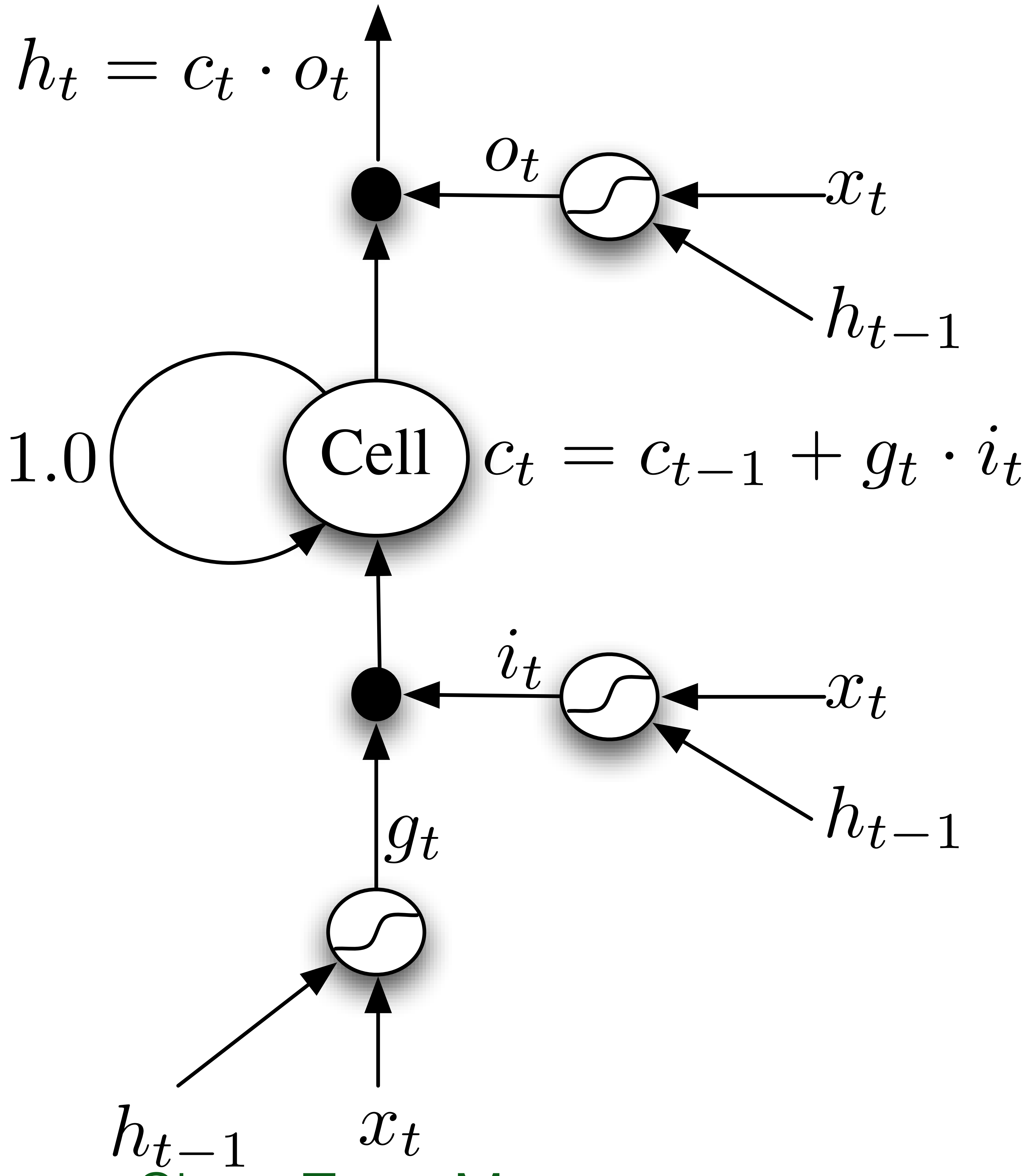


Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)

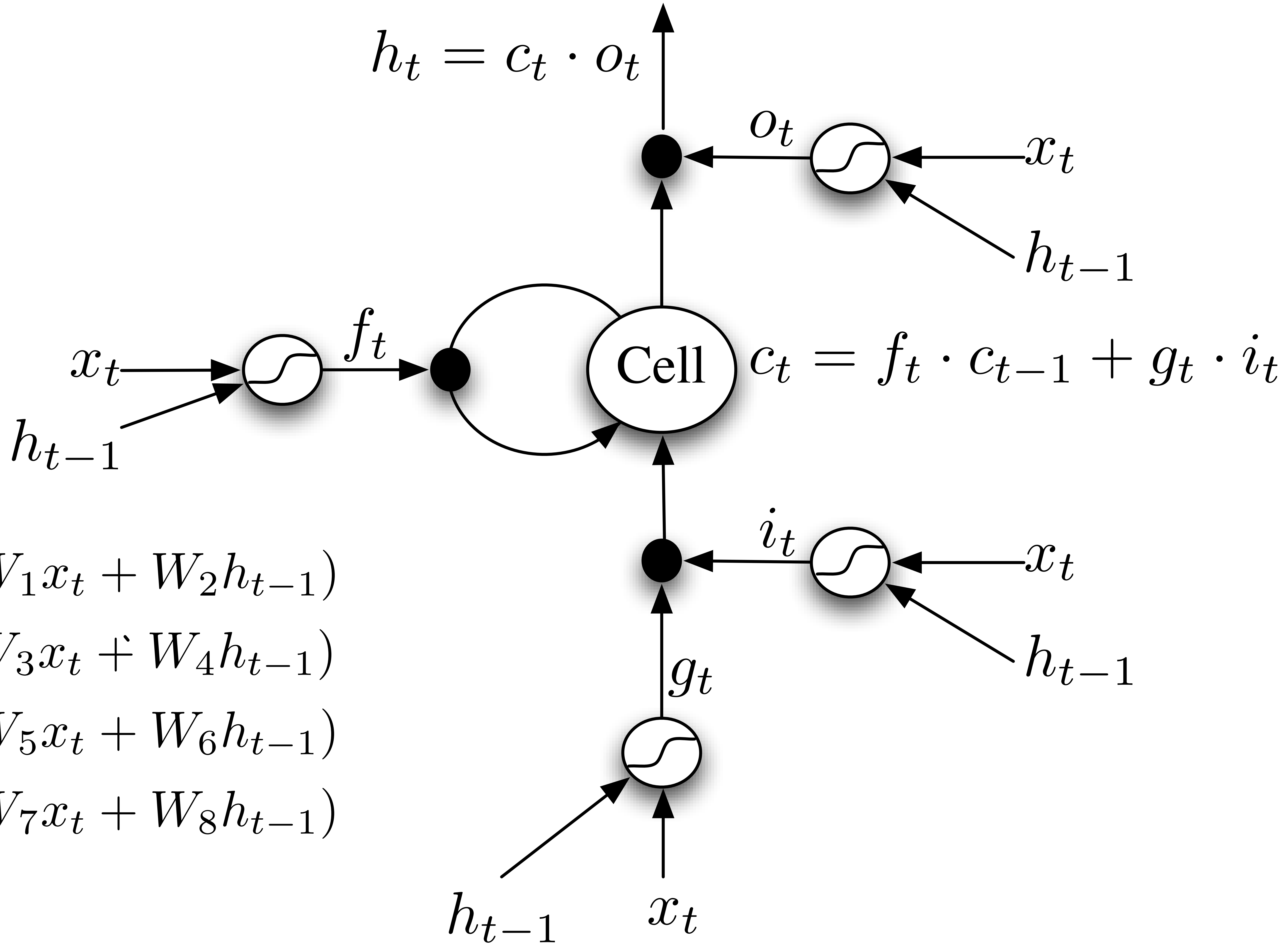


$$g_t = \sigma(W_1 x_t + W_2 h_{t-1})$$
$$i_t = \sigma(W_3 x_t + W_4 h_{t-1})$$
$$o_t = \sigma(W_5 x_t + W_6 h_{t-1})$$



Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)

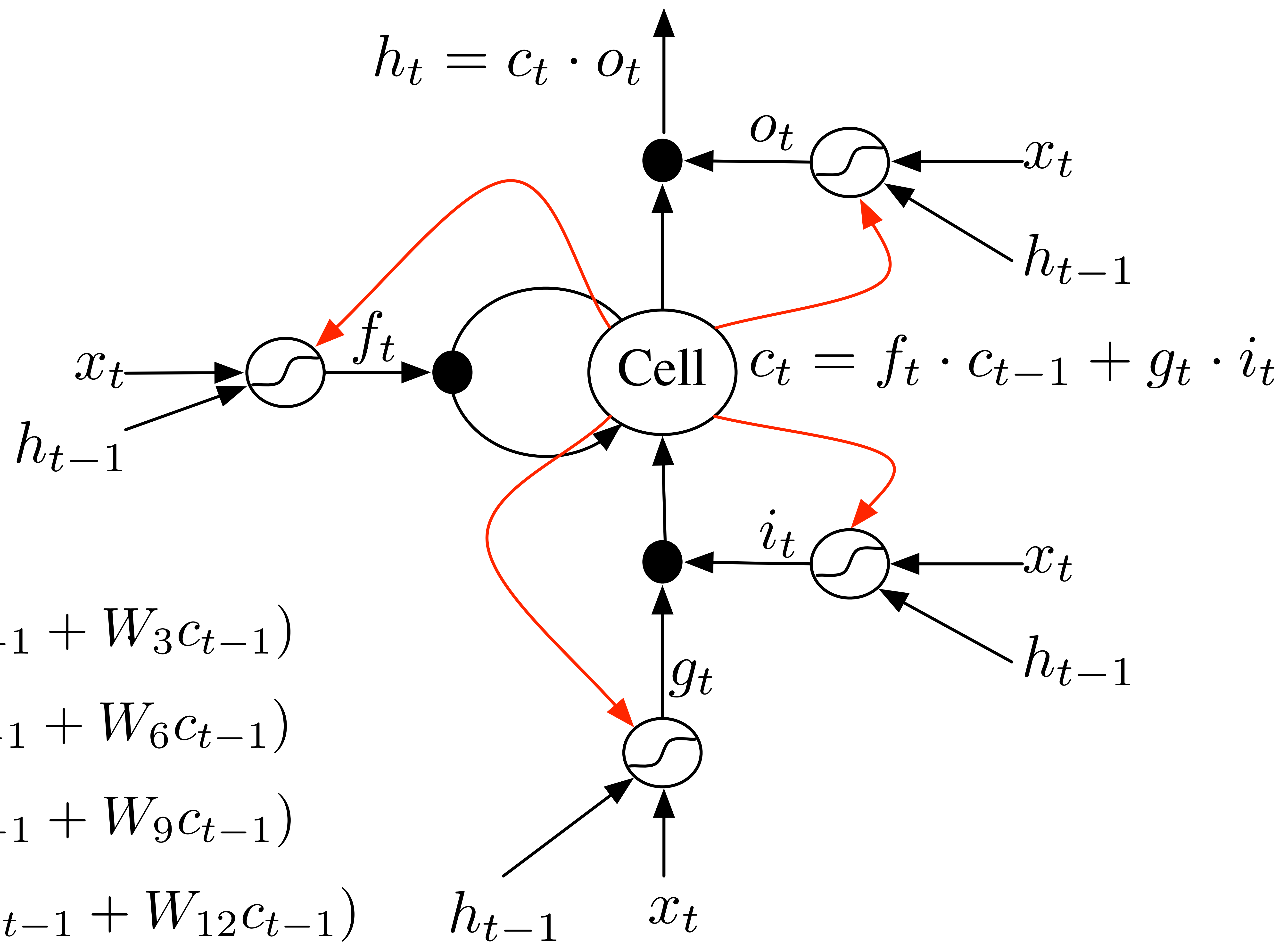


$$g_t = \sigma(W_1 x_t + W_2 h_{t-1})$$
$$i_t = \sigma(W_3 x_t + W_4 h_{t-1})$$
$$o_t = \sigma(W_5 x_t + W_6 h_{t-1})$$
$$f_t = \sigma(W_7 x_t + W_8 h_{t-1})$$

Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)

Peep-Hole Connections



$$g_t = \sigma(W_1 x_t + W_2 h_{t-1} + W_3 c_{t-1})$$

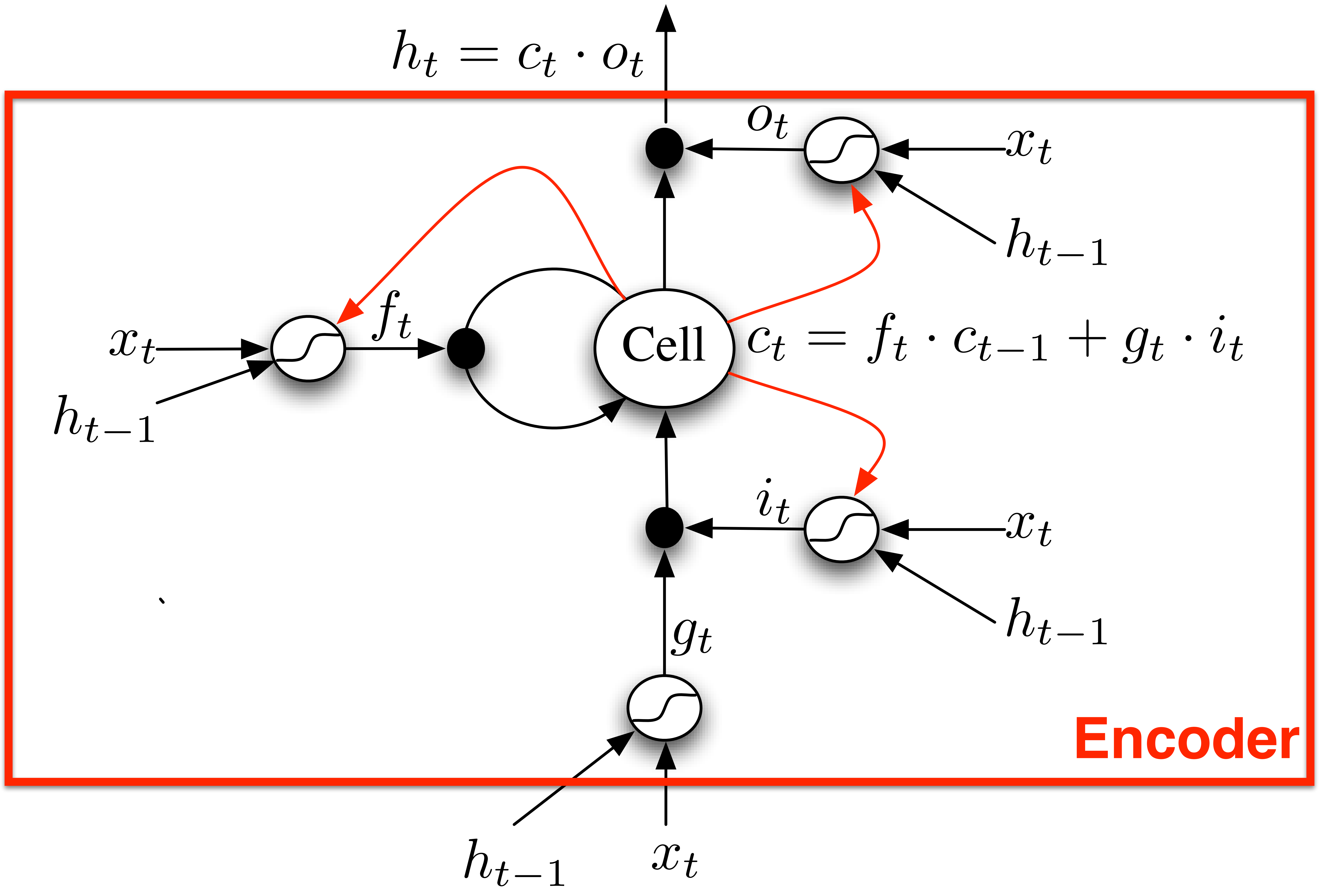
$$i_t = \sigma(W_4 x_t + W_5 h_{t-1} + W_6 c_{t-1})$$

$$o_t = \sigma(W_7 x_t + W_8 h_{t-1} + W_9 c_{t-1})$$

$$f_t = \sigma(W_{10} x_t + W_{11} h_{t-1} + W_{12} c_{t-1})$$

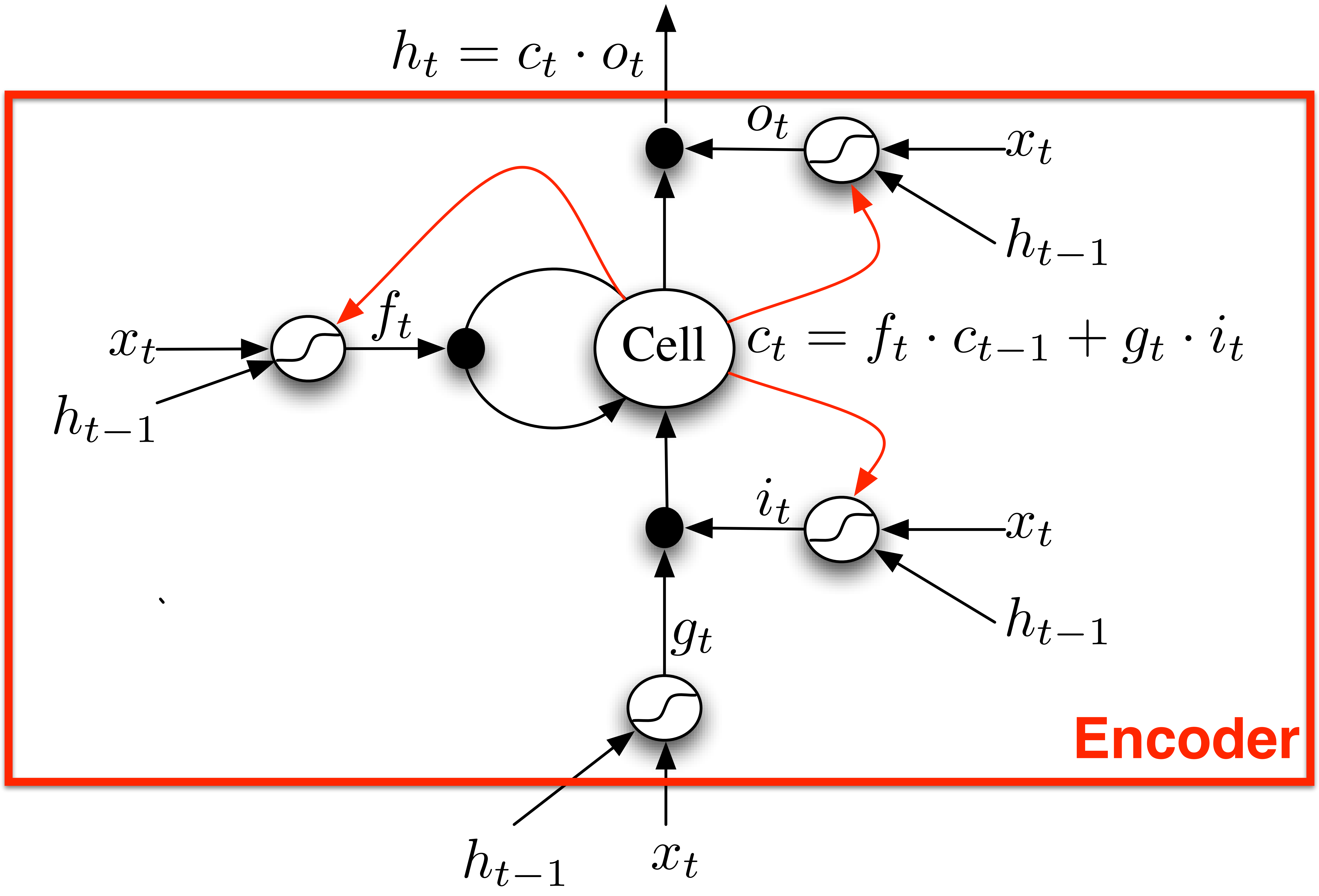
Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)



Hochreiter et.al., 1997: Long Short-Term Memory

Long Short-Term Memory (LSTM)



Hochreiter et.al., 1997: Long Short-Term Memory