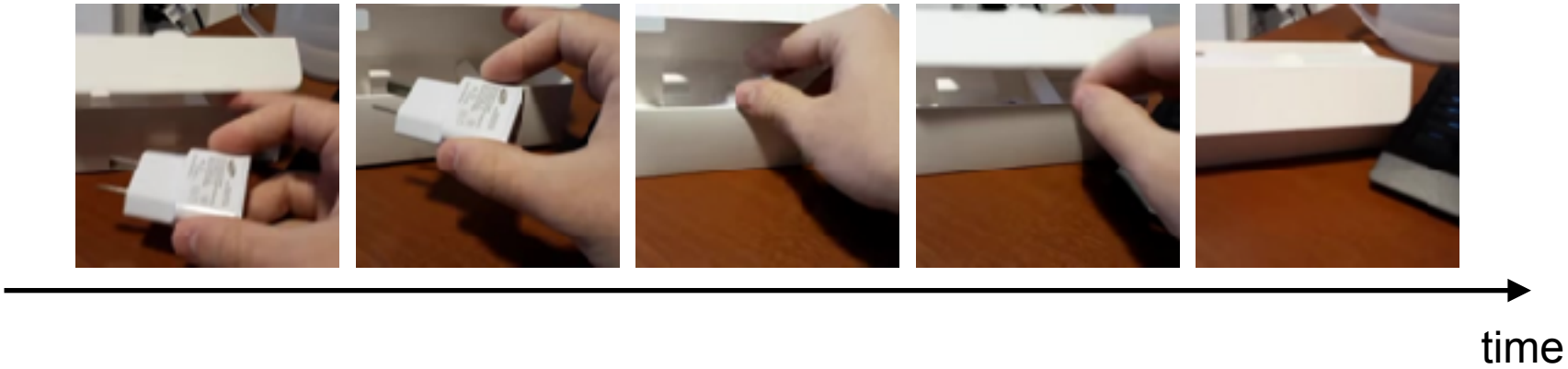
Video Modeling

How can we model temporal information in the video?

Input Video

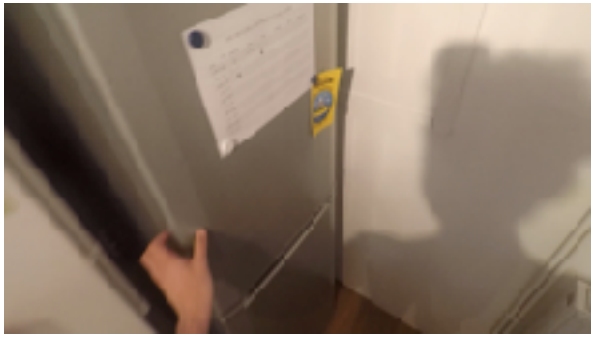


Video Classification

Given a video, we want to classify it into one of the human action categories.

Cartwheeling

Braiding Hair



Opening a Fridge

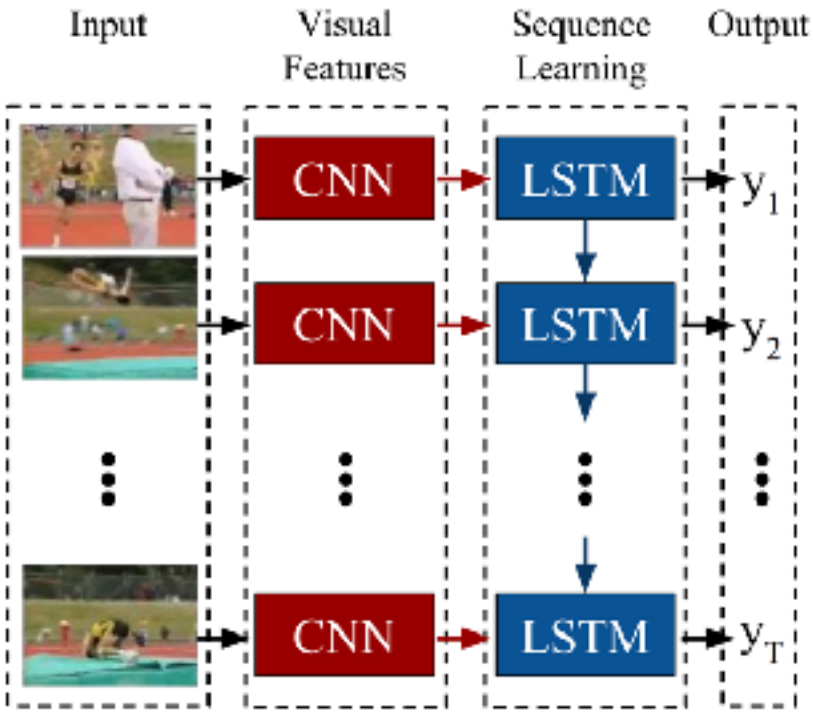
Long-term Recurrent Convolutional Networks for Visual Recognition and Description

CVPR 2015

Jeff Donahue, Lisa Anne Hendricks, Marcus Rohrbach,
Subhashini Venugopalan, Sergio Guadarrama,
Kate Saenko, Trevor Darrell

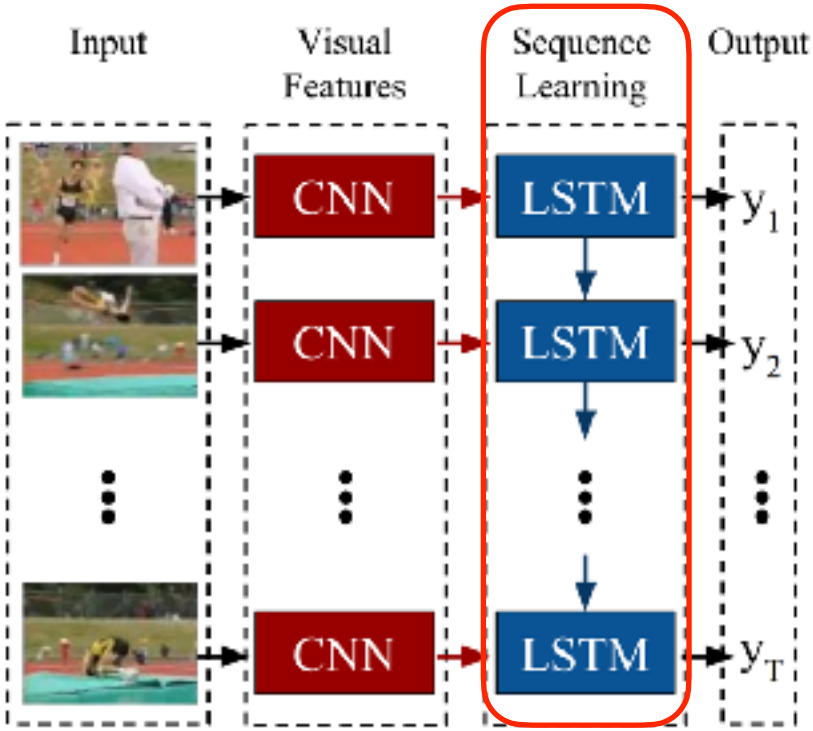
LRCN Model

- The paper proposes a long-term recurrent convolutional network (LRCN).
- The proposed model enables learning visual dependencies in space and time.



LRCN Model

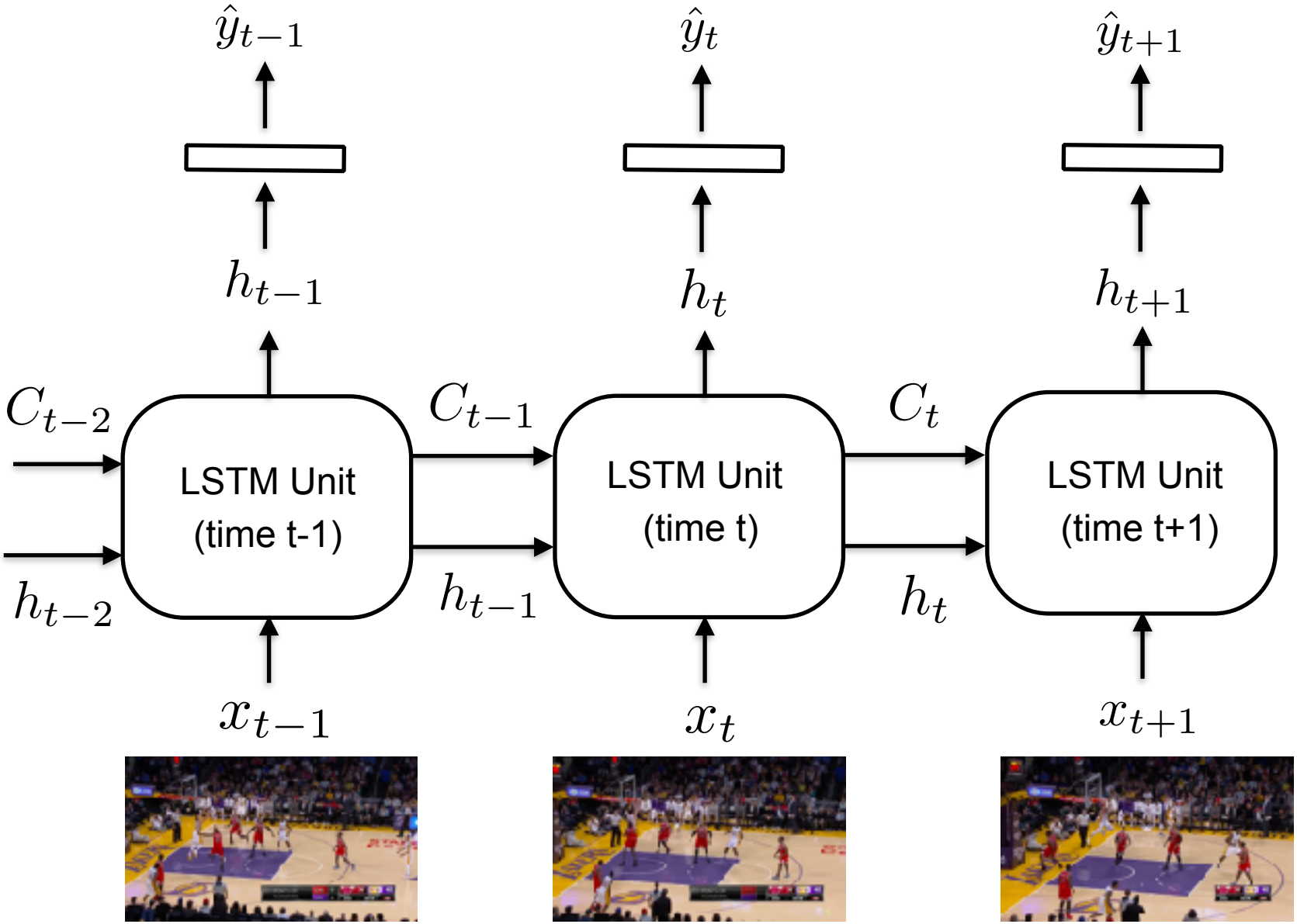
- The paper proposes a long-term recurrent convolutional network (LRCN).
- The proposed model enables learning visual dependencies in space and time.



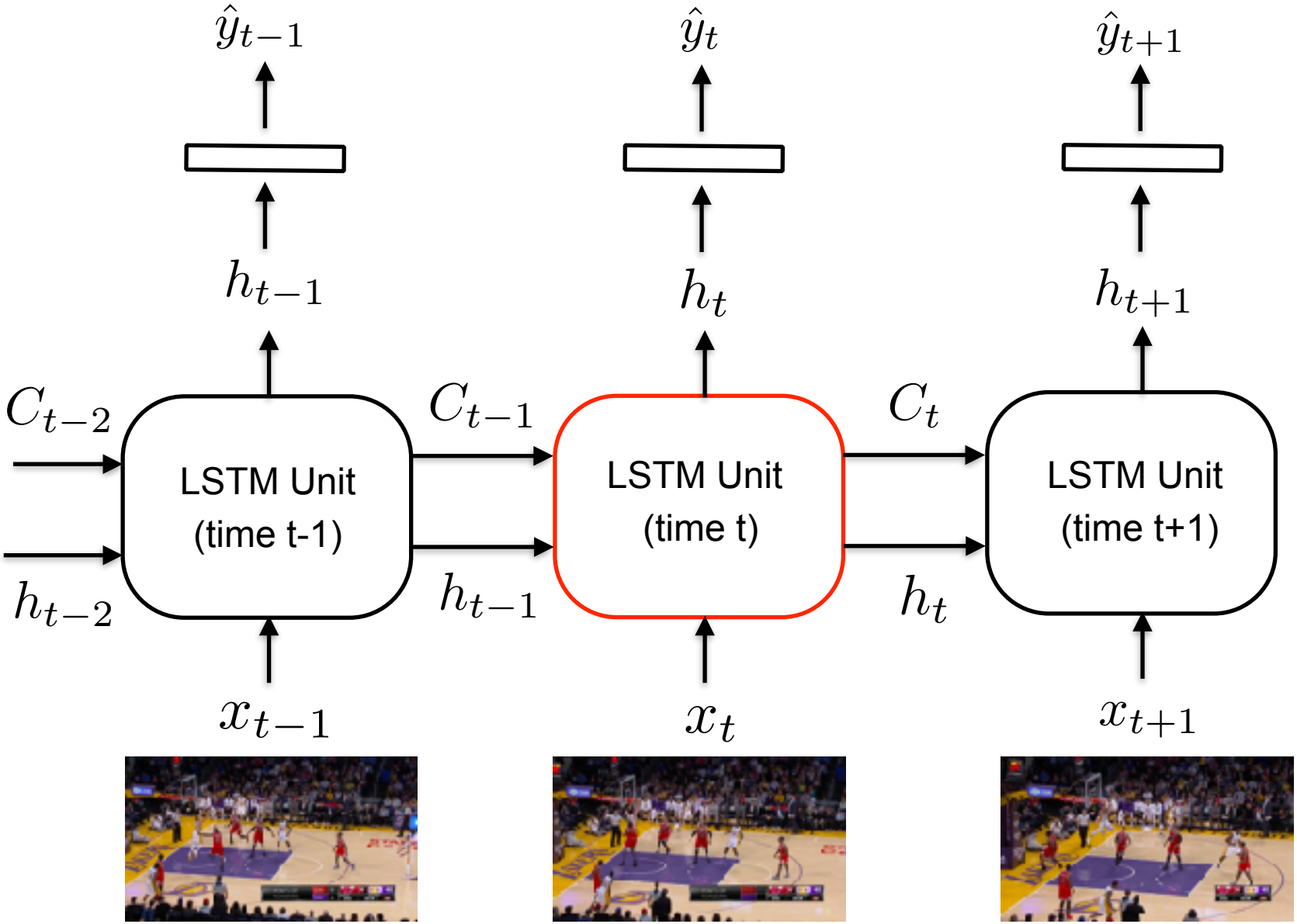
Long Short Term Memory Network

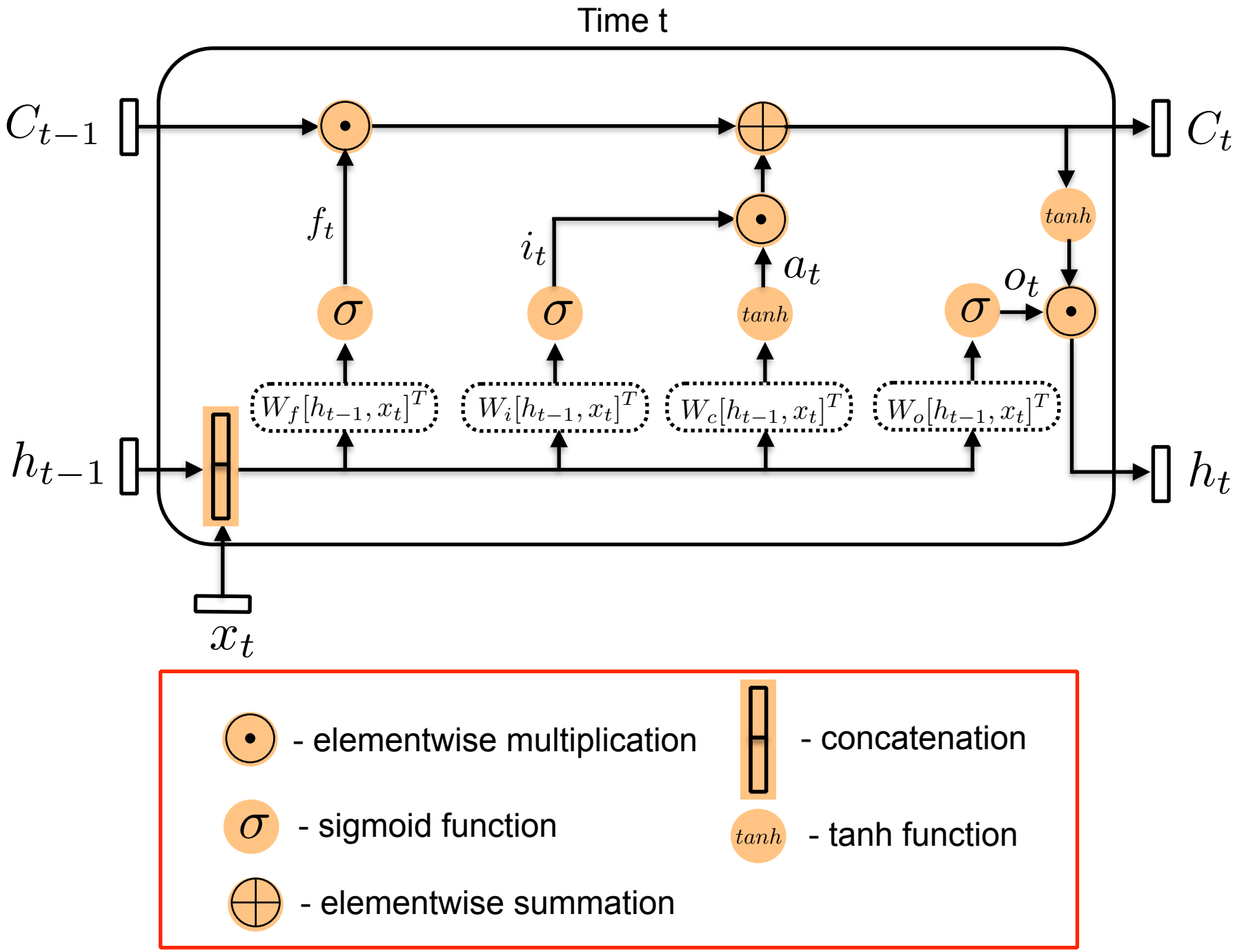
- A recurrent neural network that enables sequence modeling (e.g. videos, text, etc).
- This is achieved via a memory mechanism that allows the network to remember what has happened in the past.
- In contrast, standard CNNs process each input (e.g., video frame) independently thus, forgetting what has happened before.

Long Short Term Memory Network

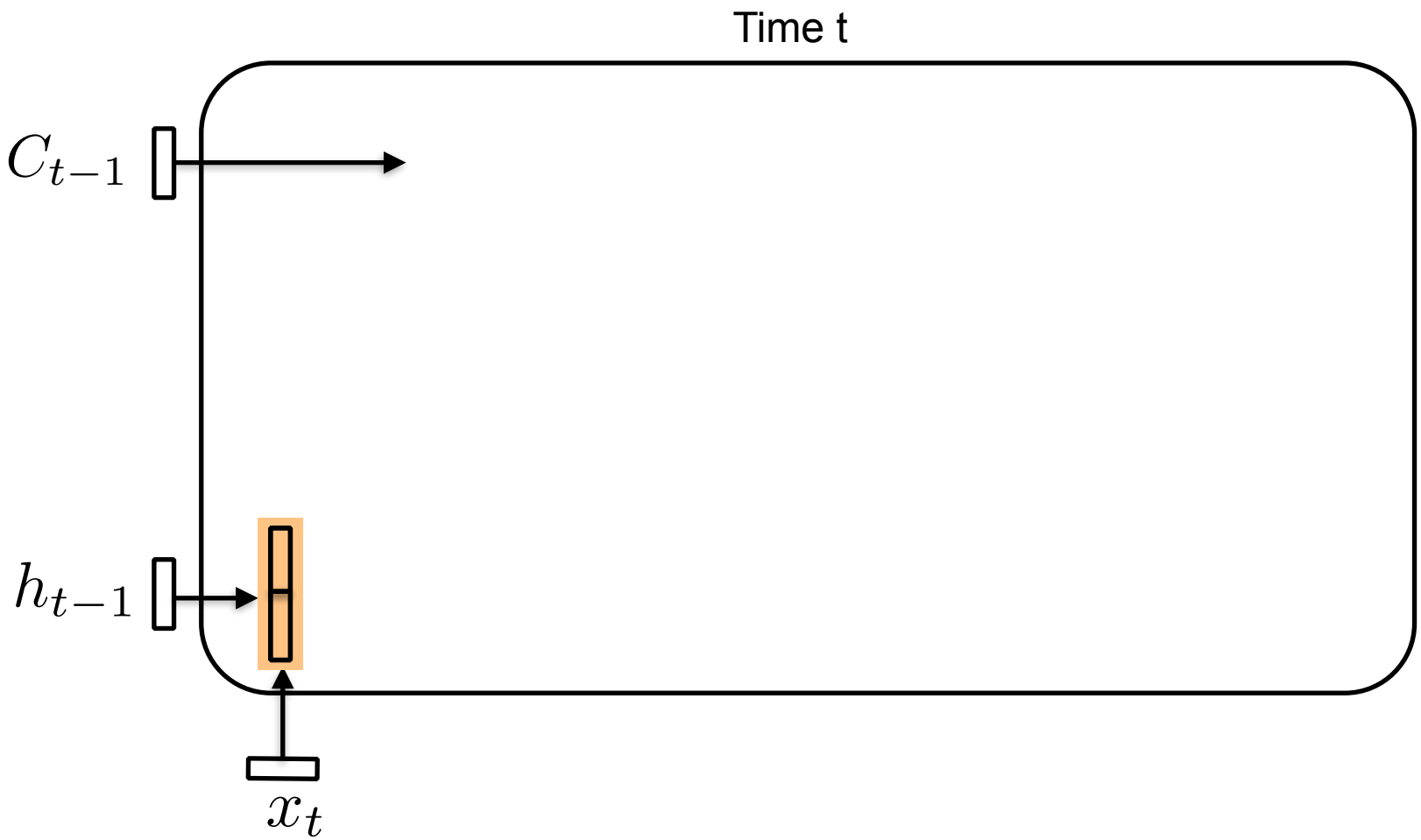


Long Short Term Memory Network

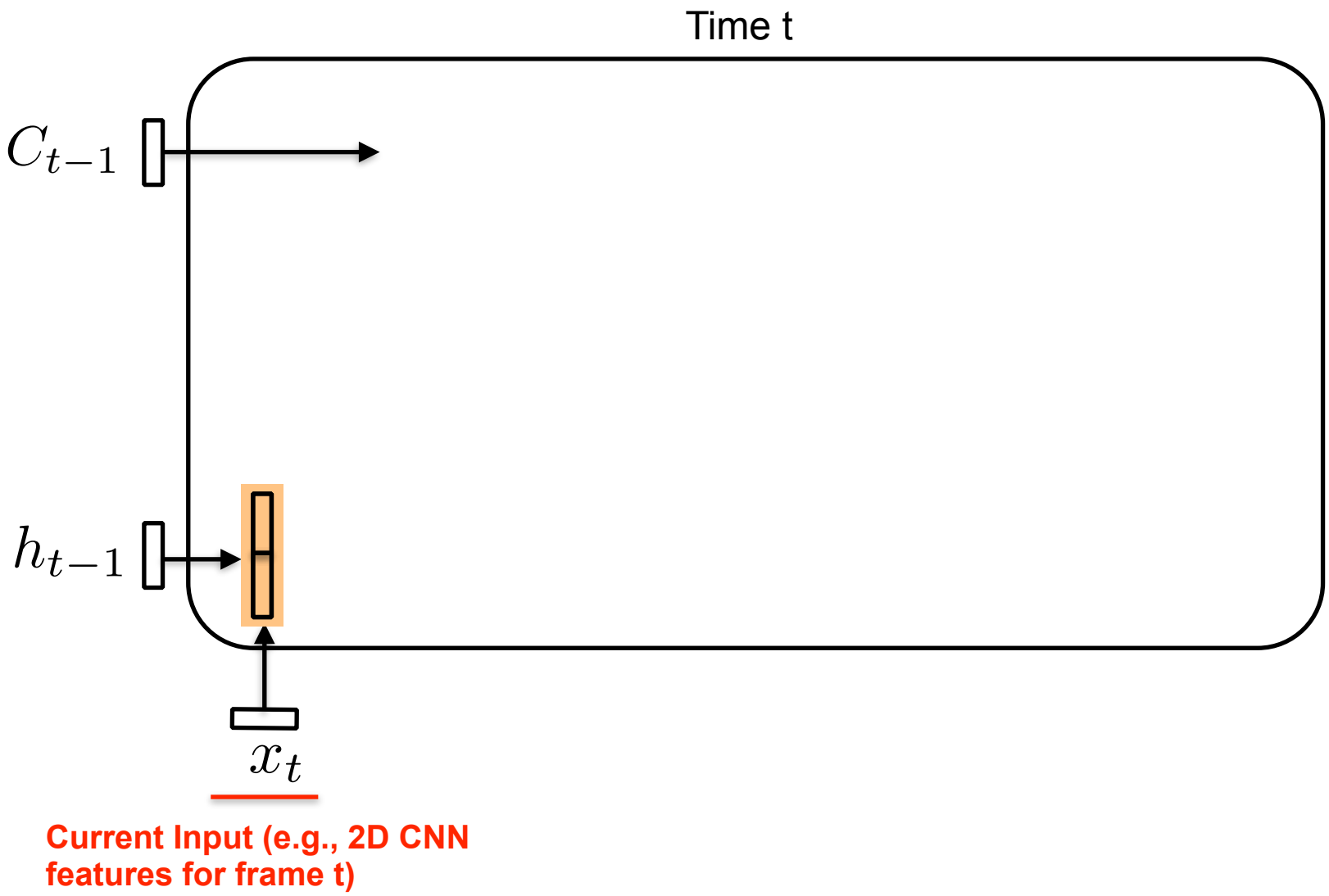



Time t
C_{t-1}
C_t
f_t
i_t
a_t
o_t
σ
σ
$tanh$
σ
$tanh$
$W_f[h_{t-1}, x_t]^T$
$W_i[h_{t-1}, x_t]^T$
$W_c[h_{t-1}, x_t]^T$
$W_o[h_{t-1}, x_t]^T$
h_{t-1}
h_t
x_t
- elementwise multiplication
- concatenation
σ - sigmoid function
$tanh$ - tanh function
- elementwise summation

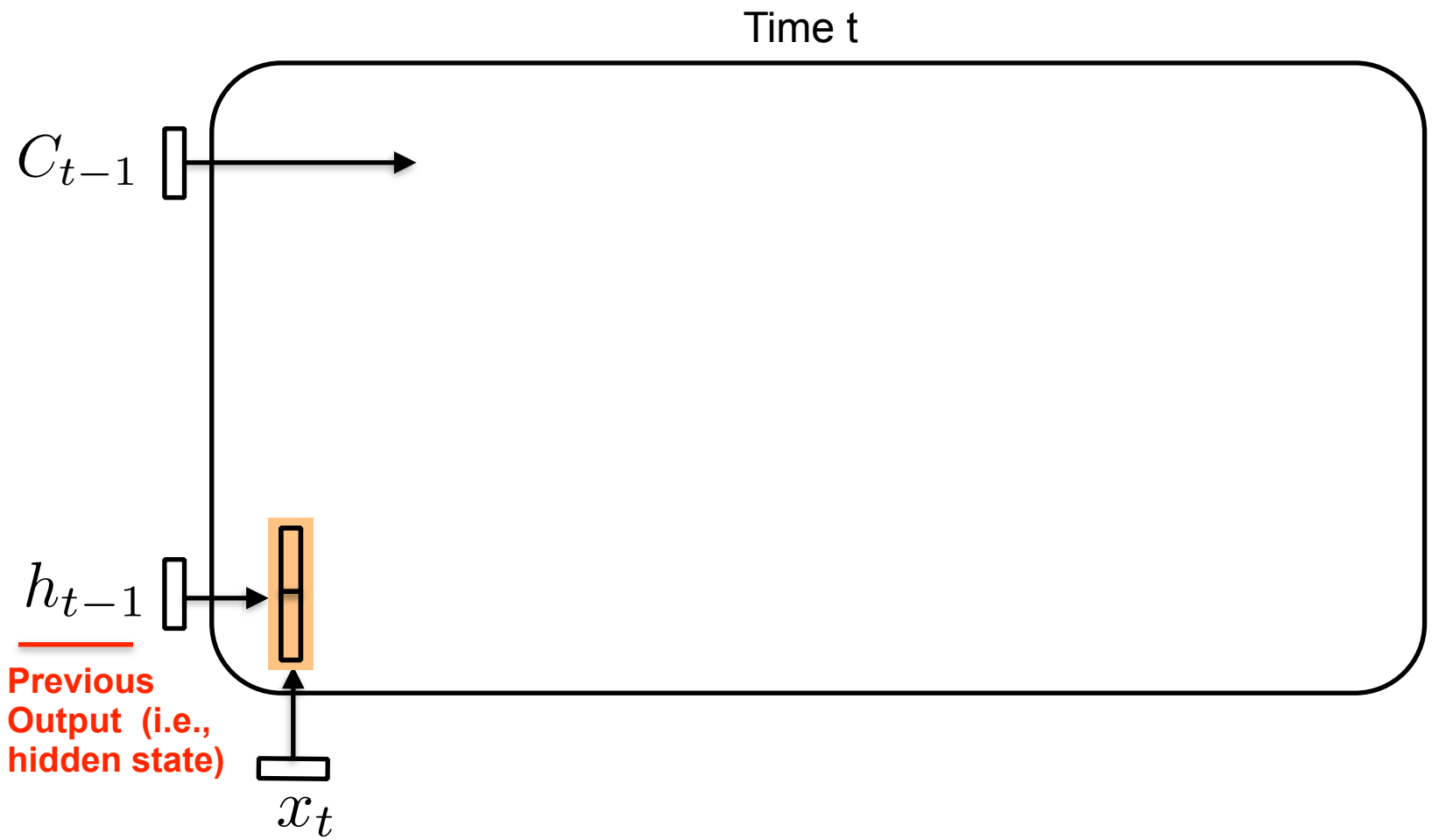
Long Short Term Memory Unit



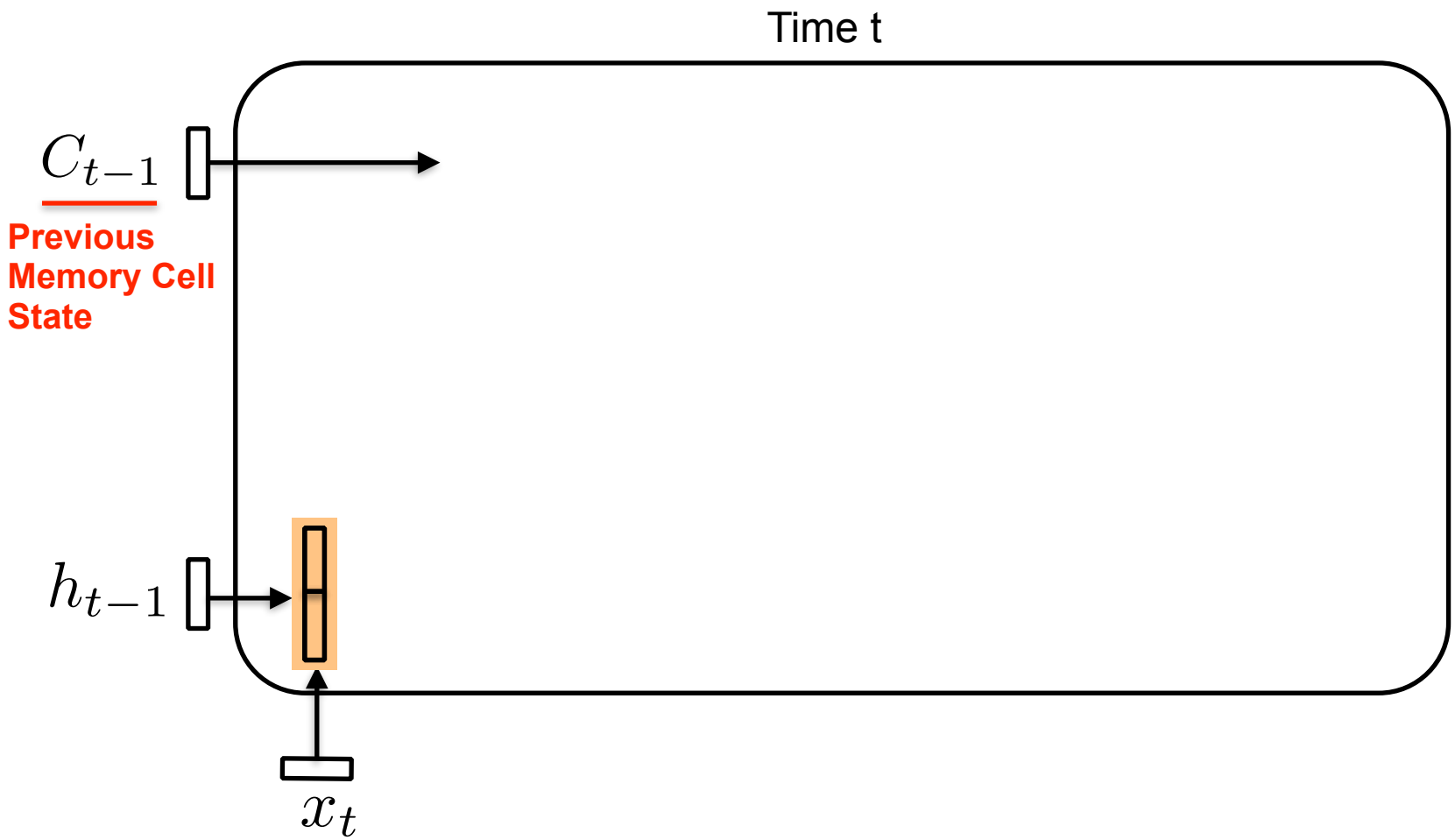
Long Short Term Memory Unit



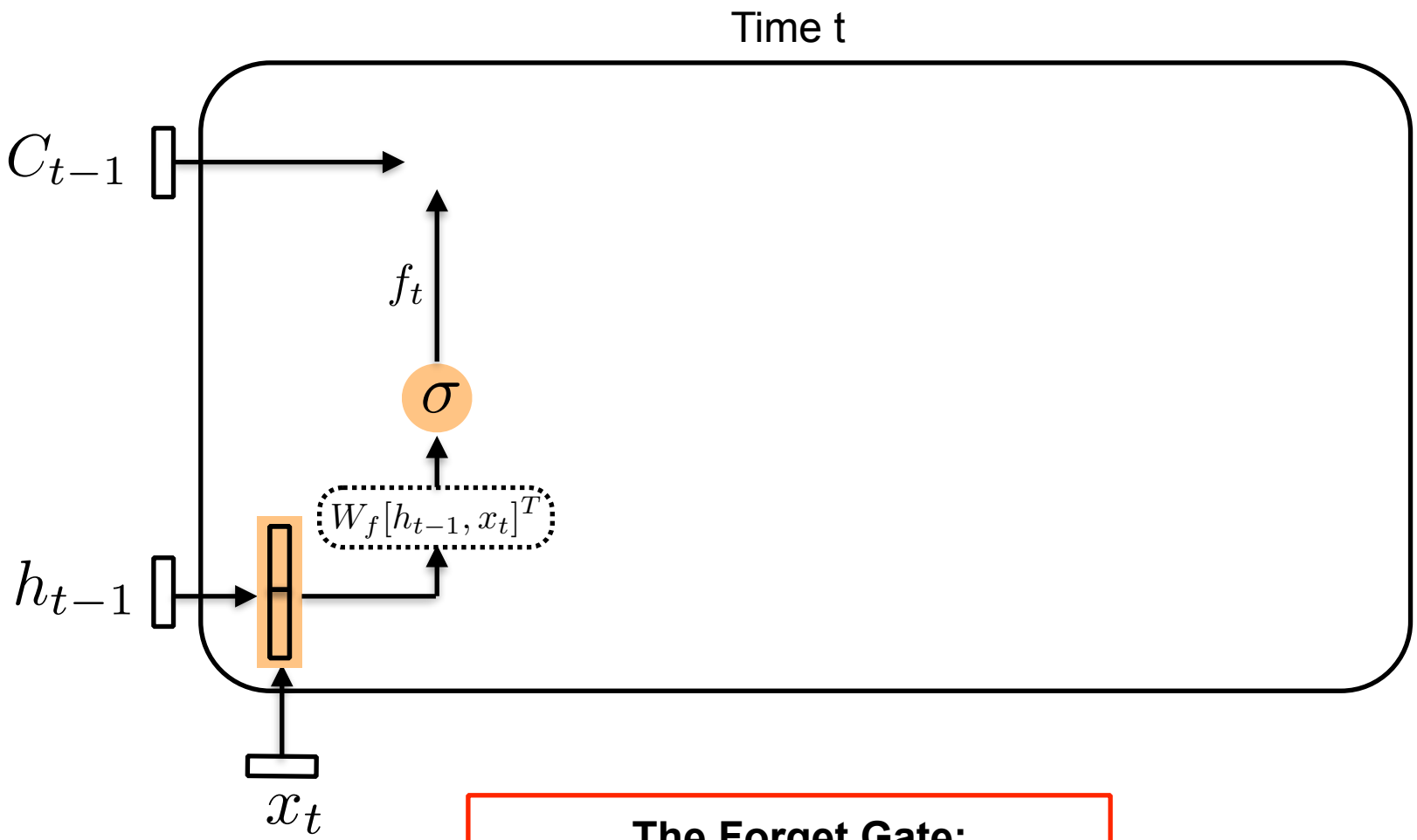
Long Short Term Memory Unit



Long Short Term Memory Unit



Long Short Term Memory Unit



<u>The Forget Gate:</u>

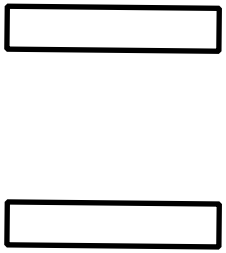
$$f_t = \sigma(W_f[h_{t-1}, x_t]^T)$$

Gating Mechanisms

Gates control what information should be added / retained.

Memory Cell from
Previous Timestep

$$C_{t-1}$$



$$f_t$$

The Forget Gate

Gating Mechanisms

Gates control what information should be added / retained.

Memory Cell from
Previous Timestep

C_{t-1}

[2.5, -9.3, -0.1, 1.3]

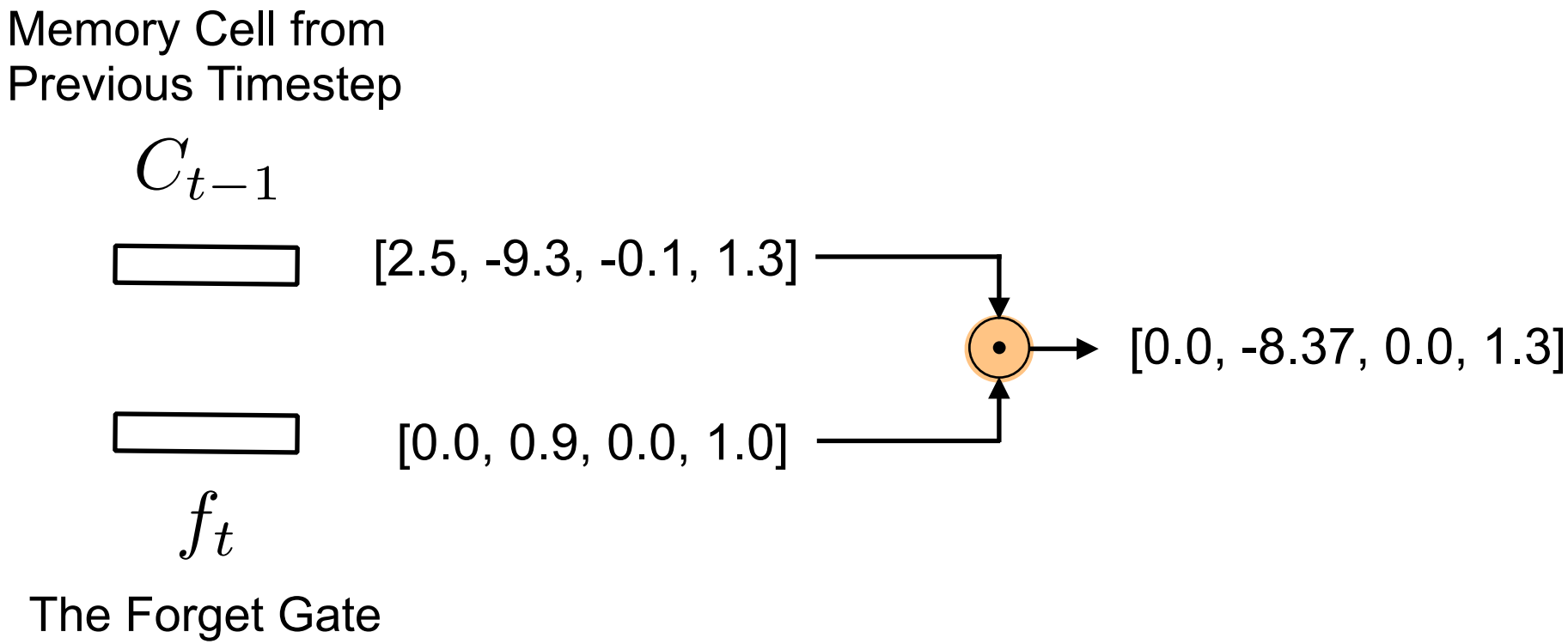
[0.0, 0.9, 0.0, 1.0]

f_t

The Forget Gate

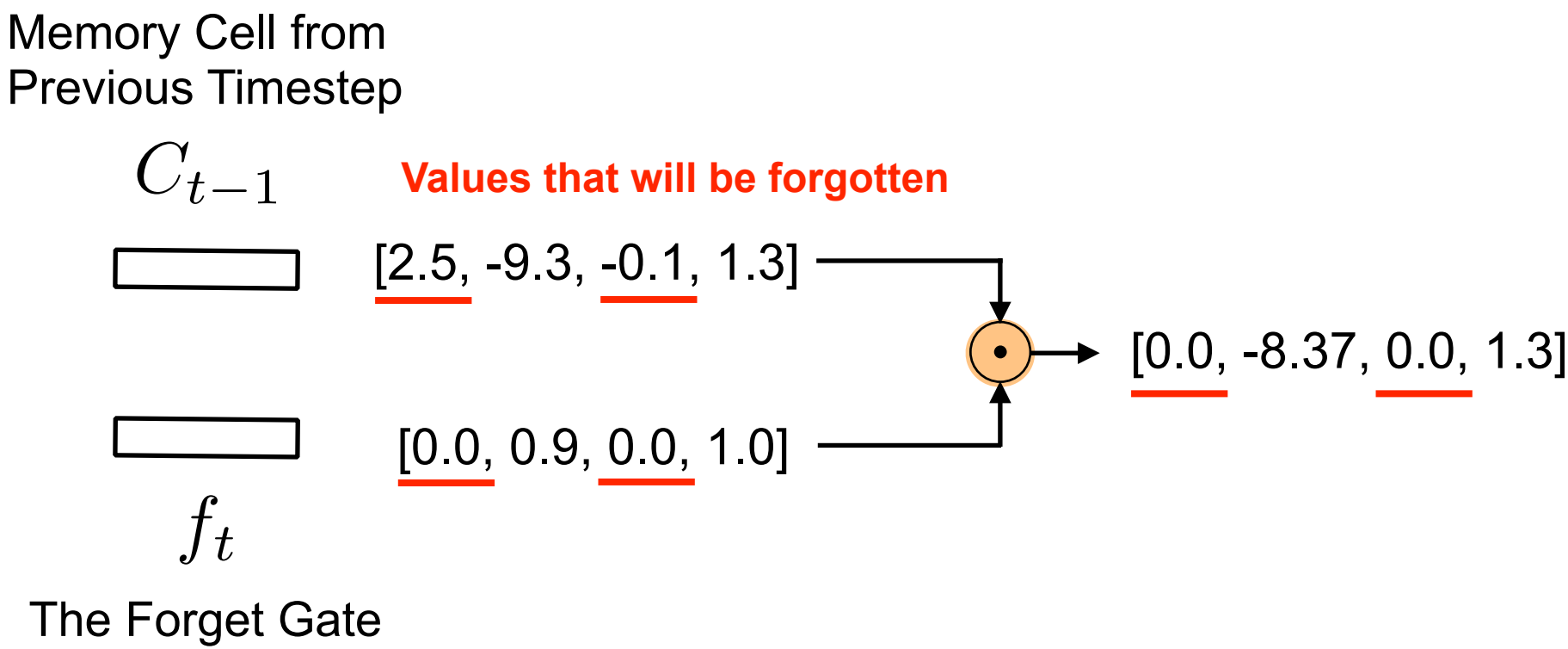
Gating Mechanisms

Gates control what information should be added / retained.



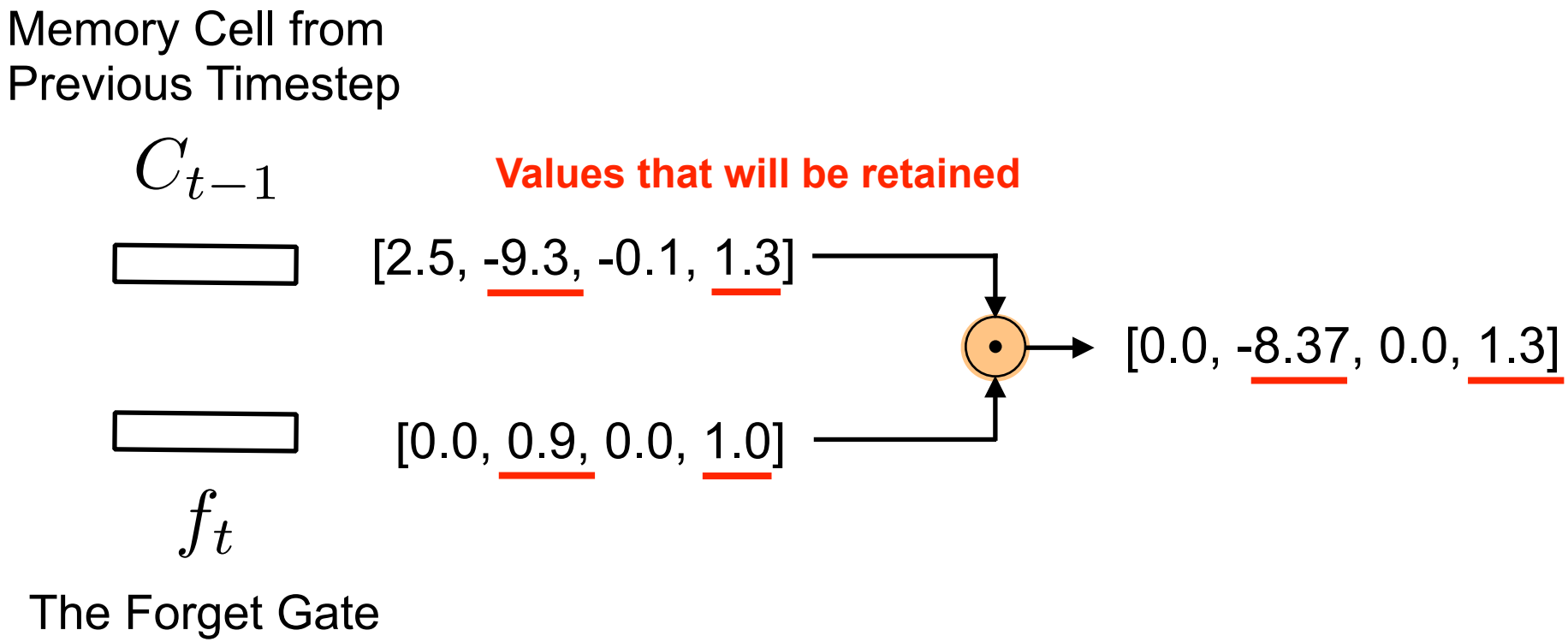
Gating Mechanisms

Gates control what information should be added / retained.

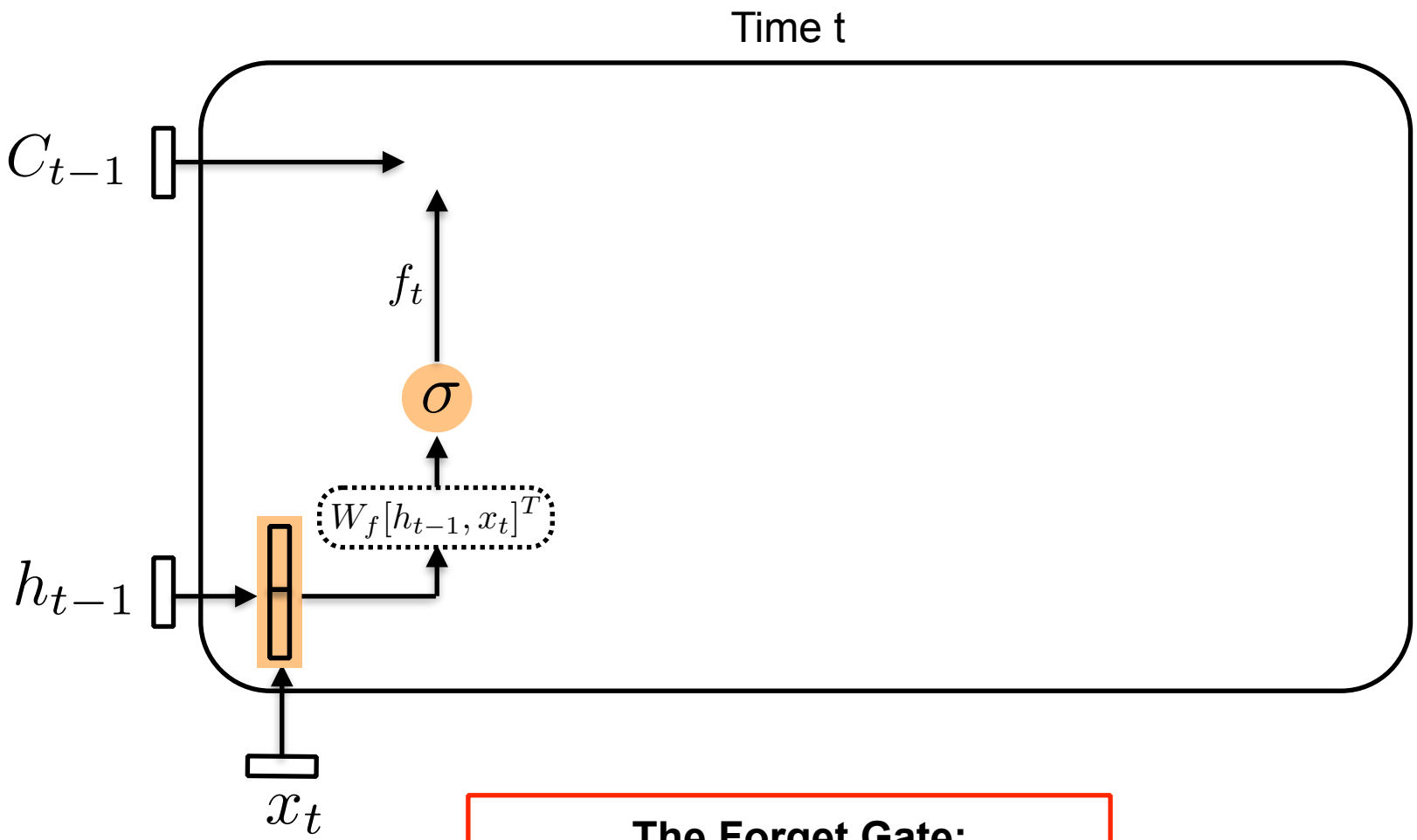


Gating Mechanisms

Gates control what information should be added / retained.



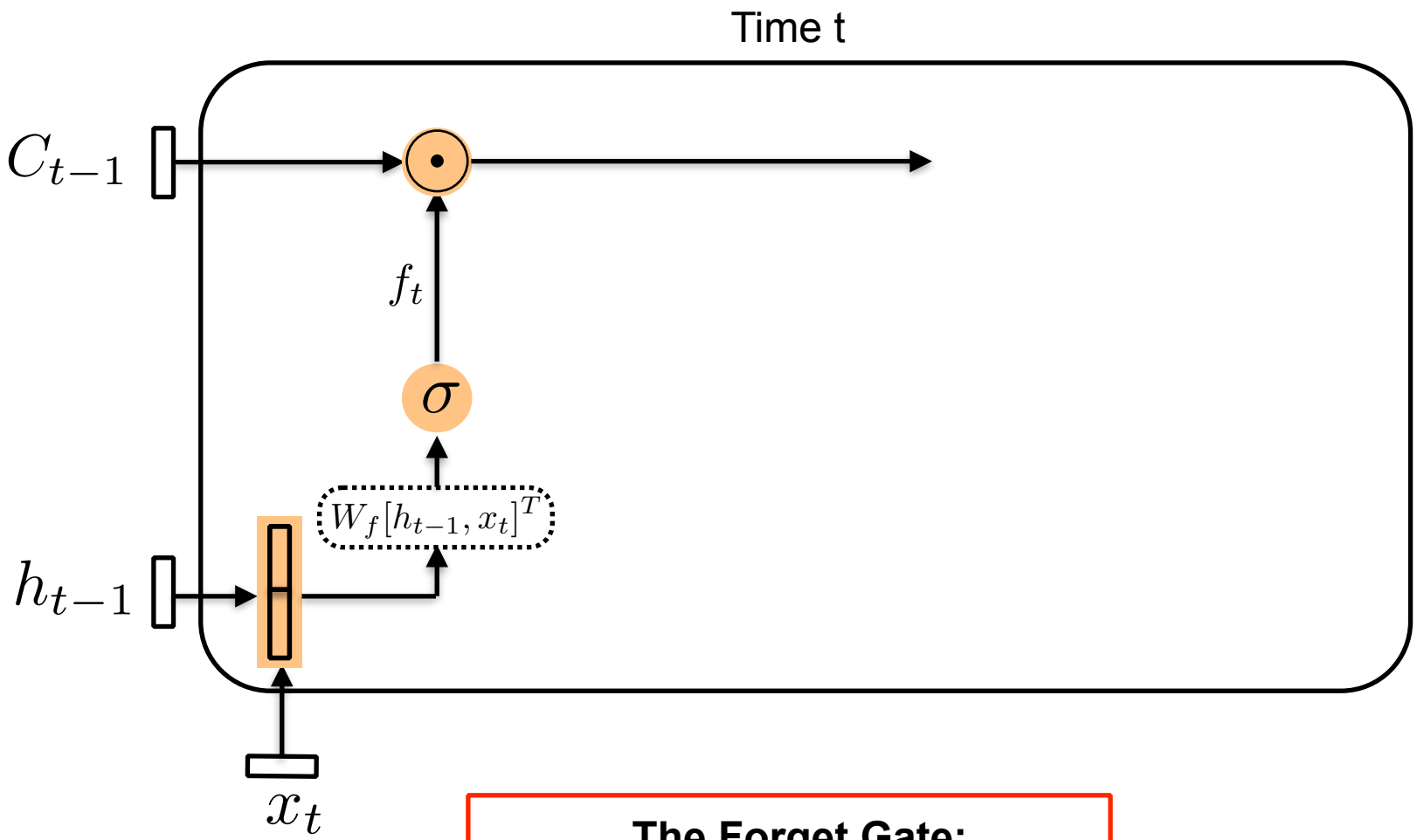
Long Short Term Memory Unit



<u>The Forget Gate:</u>

$$f_t = \sigma(W_f[h_{t-1}, x_t]^T)$$

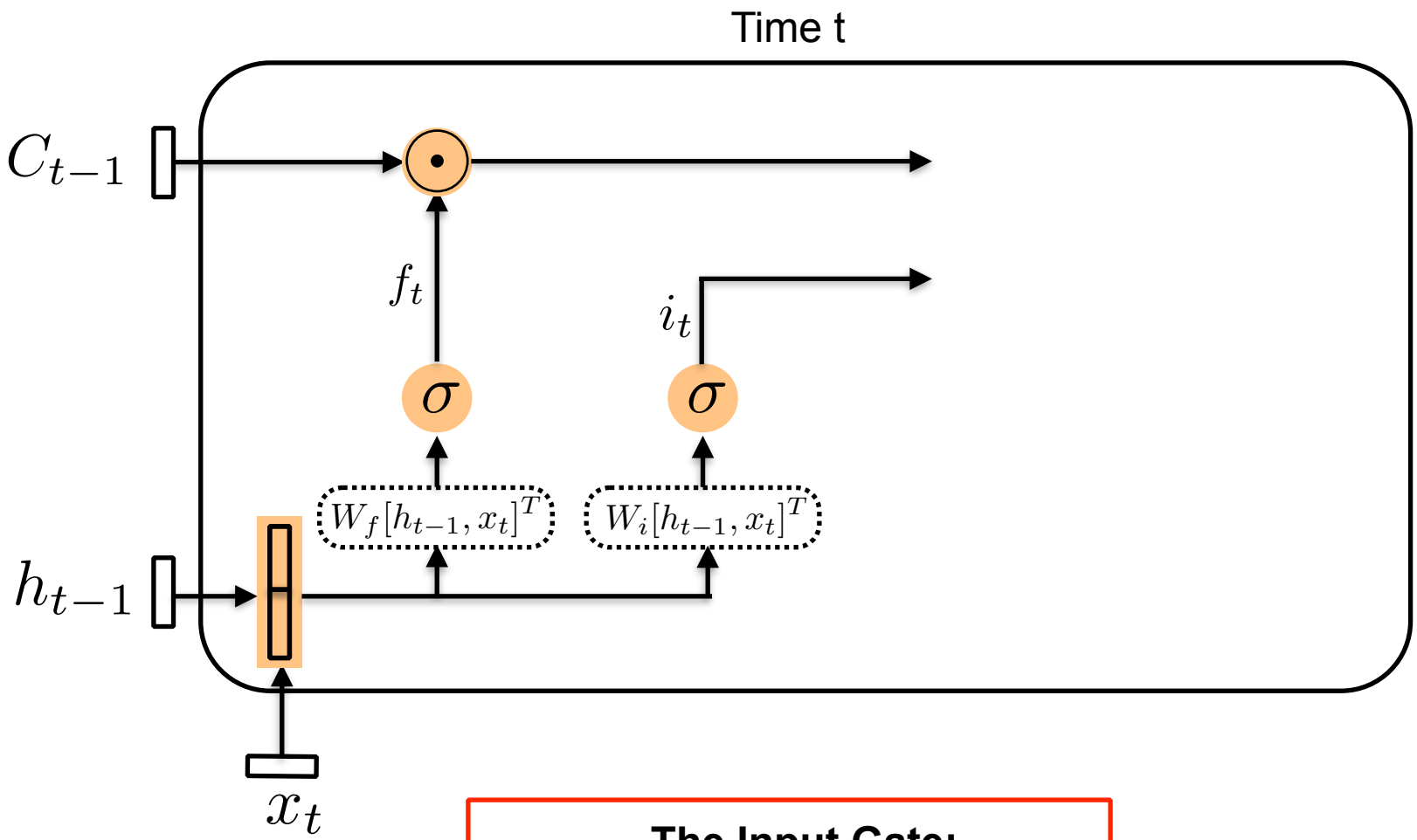
Long Short Term Memory Unit



<u>The Forget Gate:</u>

$$f_t = \sigma(W_f[h_{t-1}, x_t]^T)$$

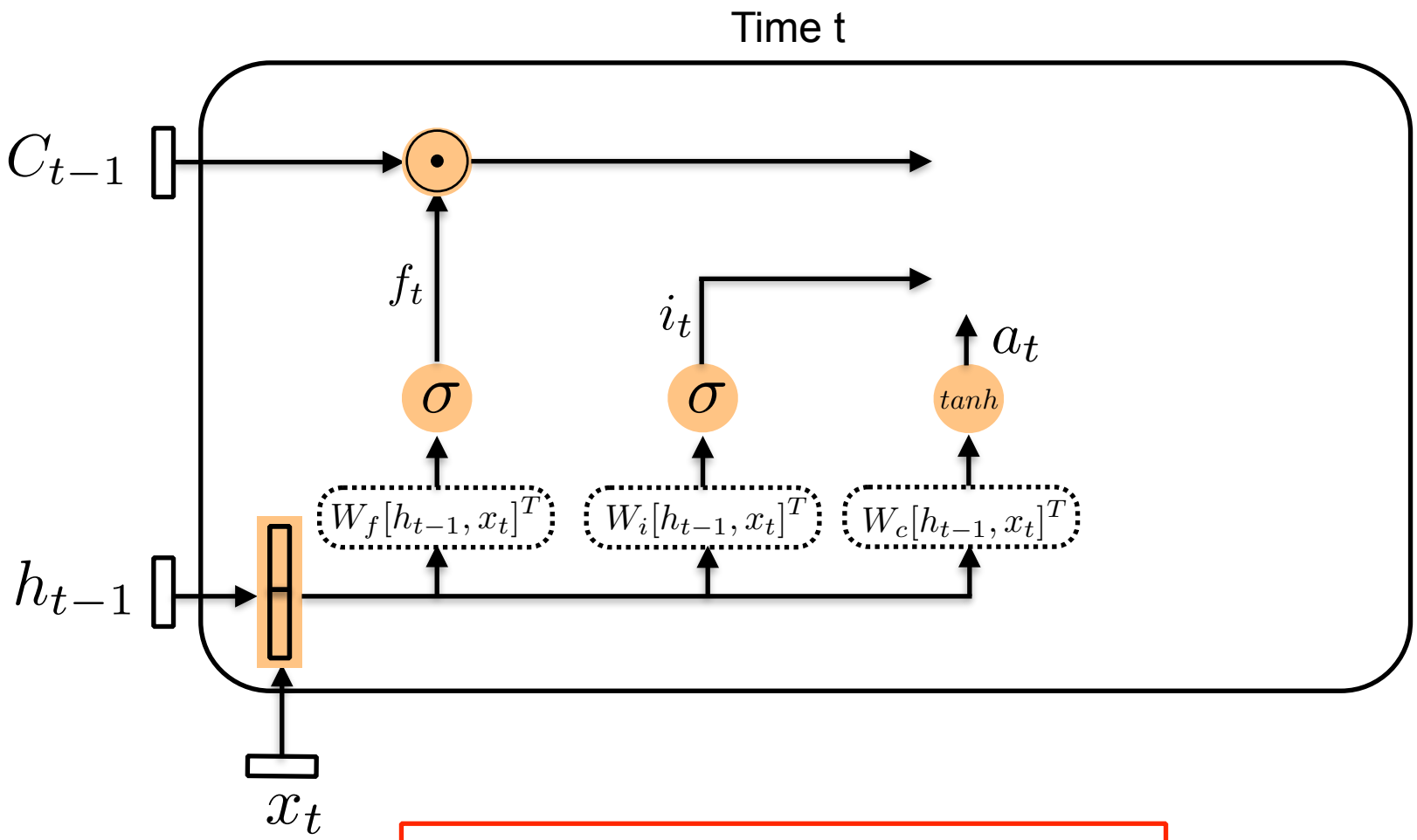
Long Short Term Memory Unit



<u>The Input Gate:</u>

$$i_t = \sigma(W_i[h_{t-1}, x_t]^T)$$

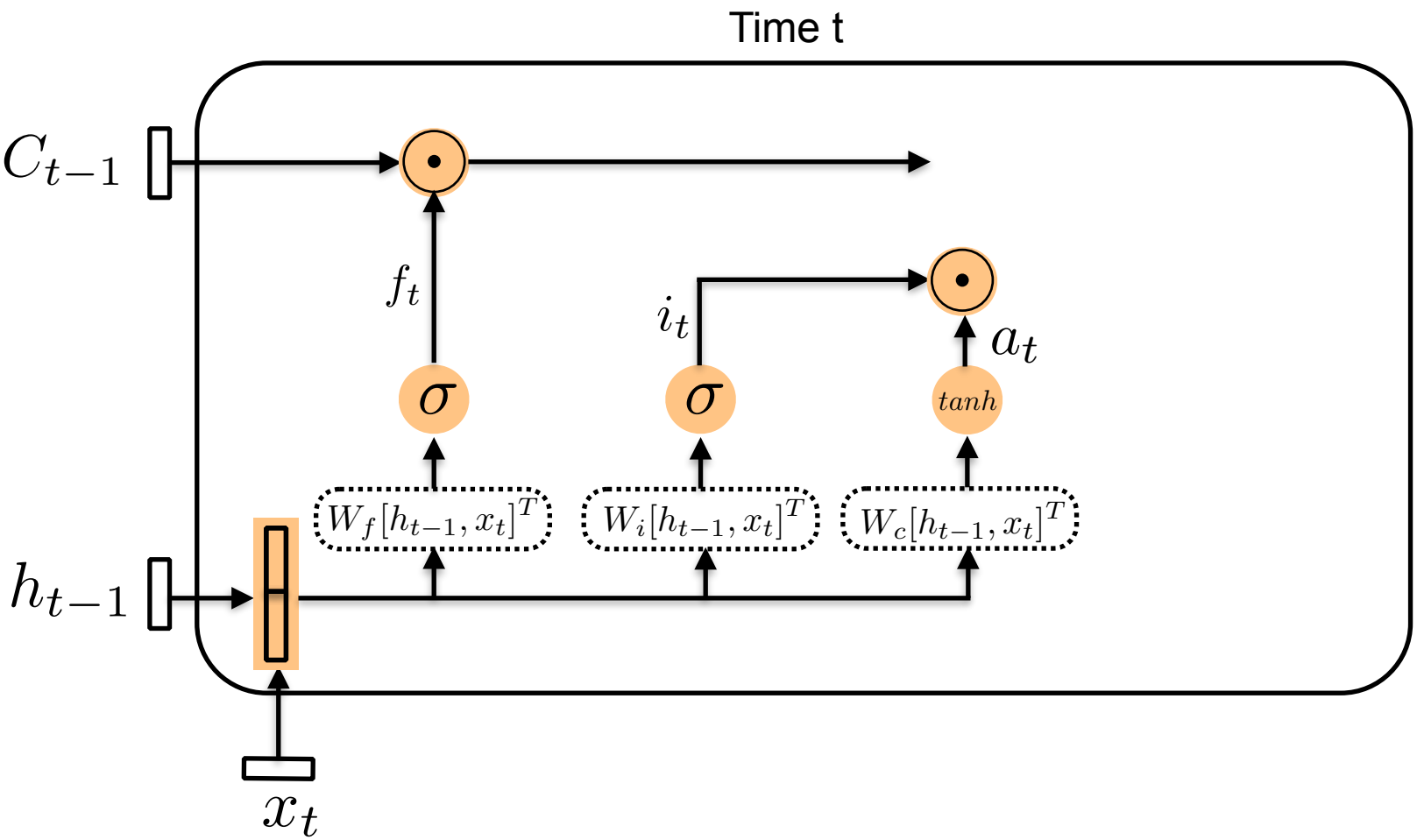
Long Short Term Memory Unit



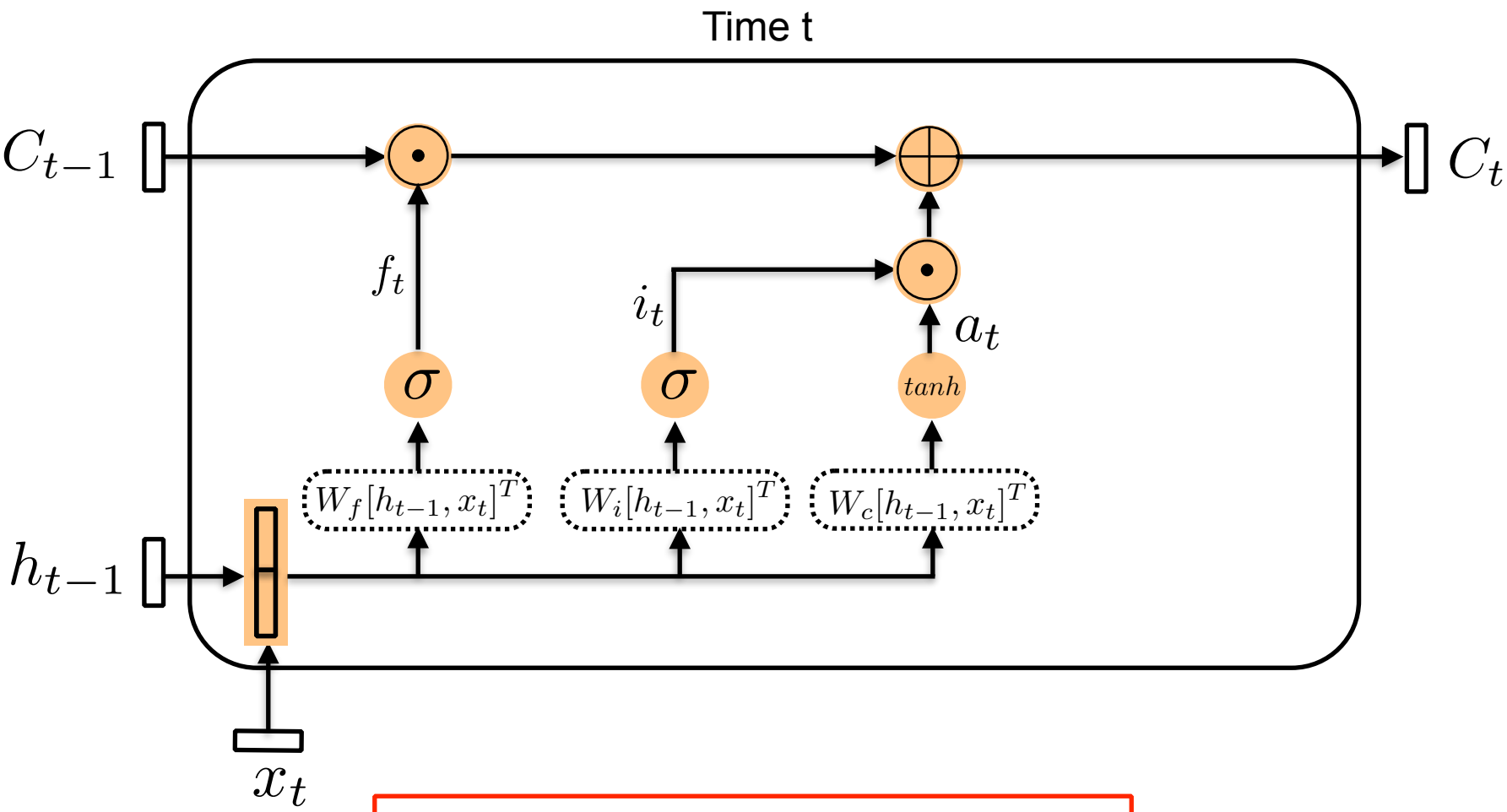
New Candidate Cell Values:

$$a_t = tanh(W_c[h_{t-1}, x_t]^T)$$

Long Short Term Memory Unit



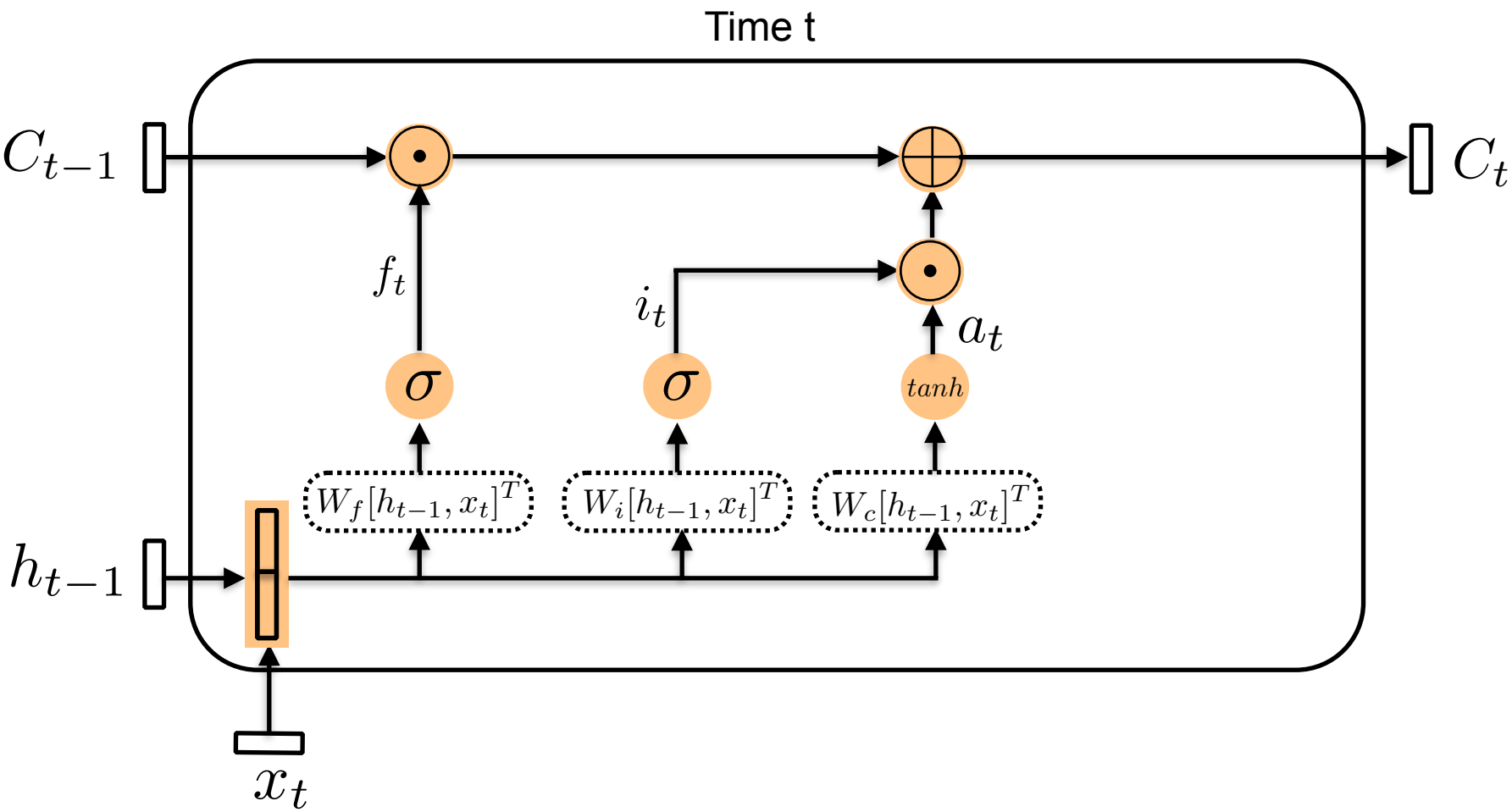
Long Short Term Memory Unit



Memory Cell Update:

$$C_t = f_t \odot C_{t-1} + i_t \odot a_t$$

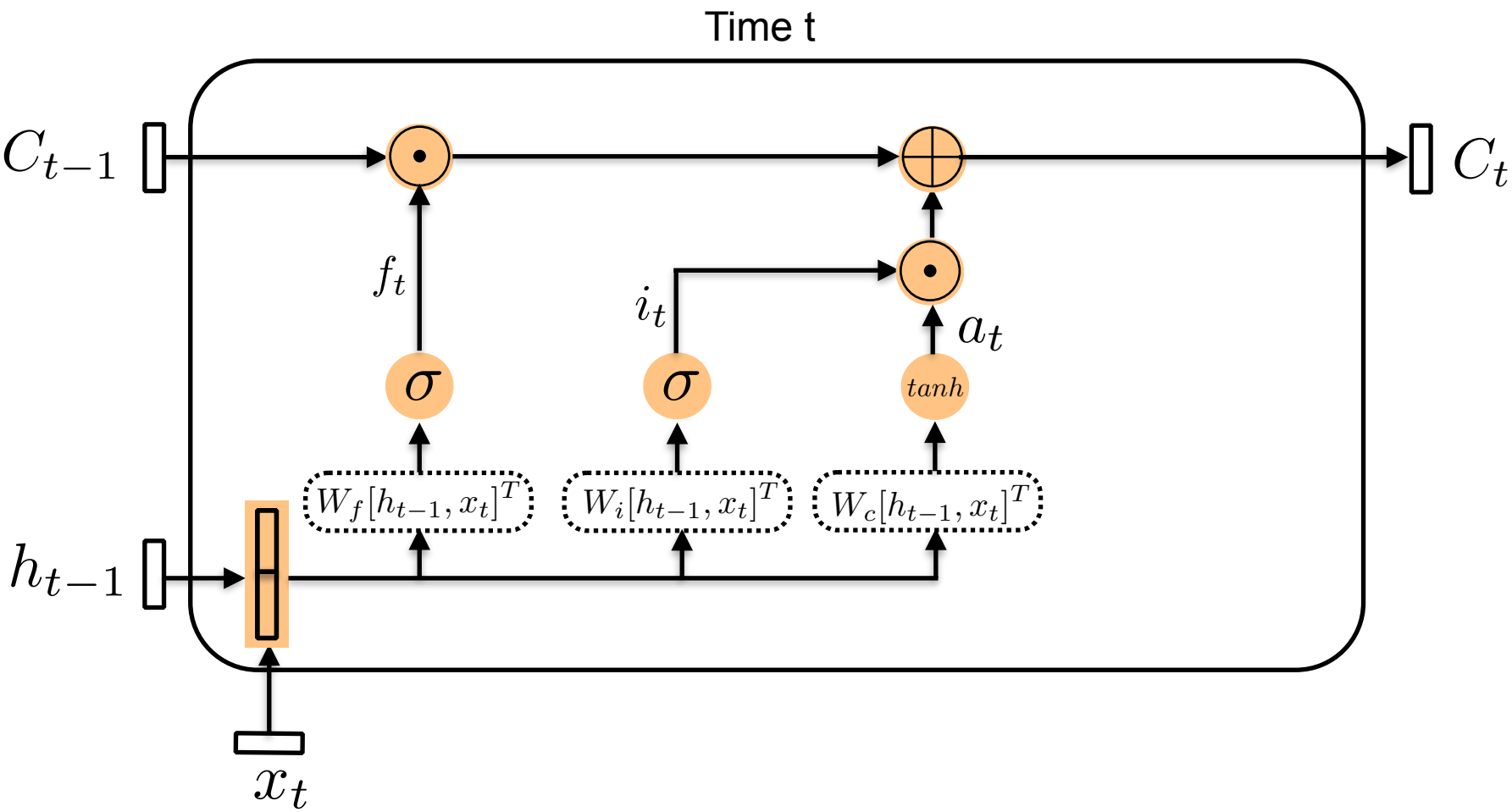
Long Short Term Memory Unit



Memory Cell Update:

$$C_t = \underbrace{f_t \odot C_{t-1}}_{\text{What to forget}} + i_t \odot a_t$$

Long Short Term Memory Unit

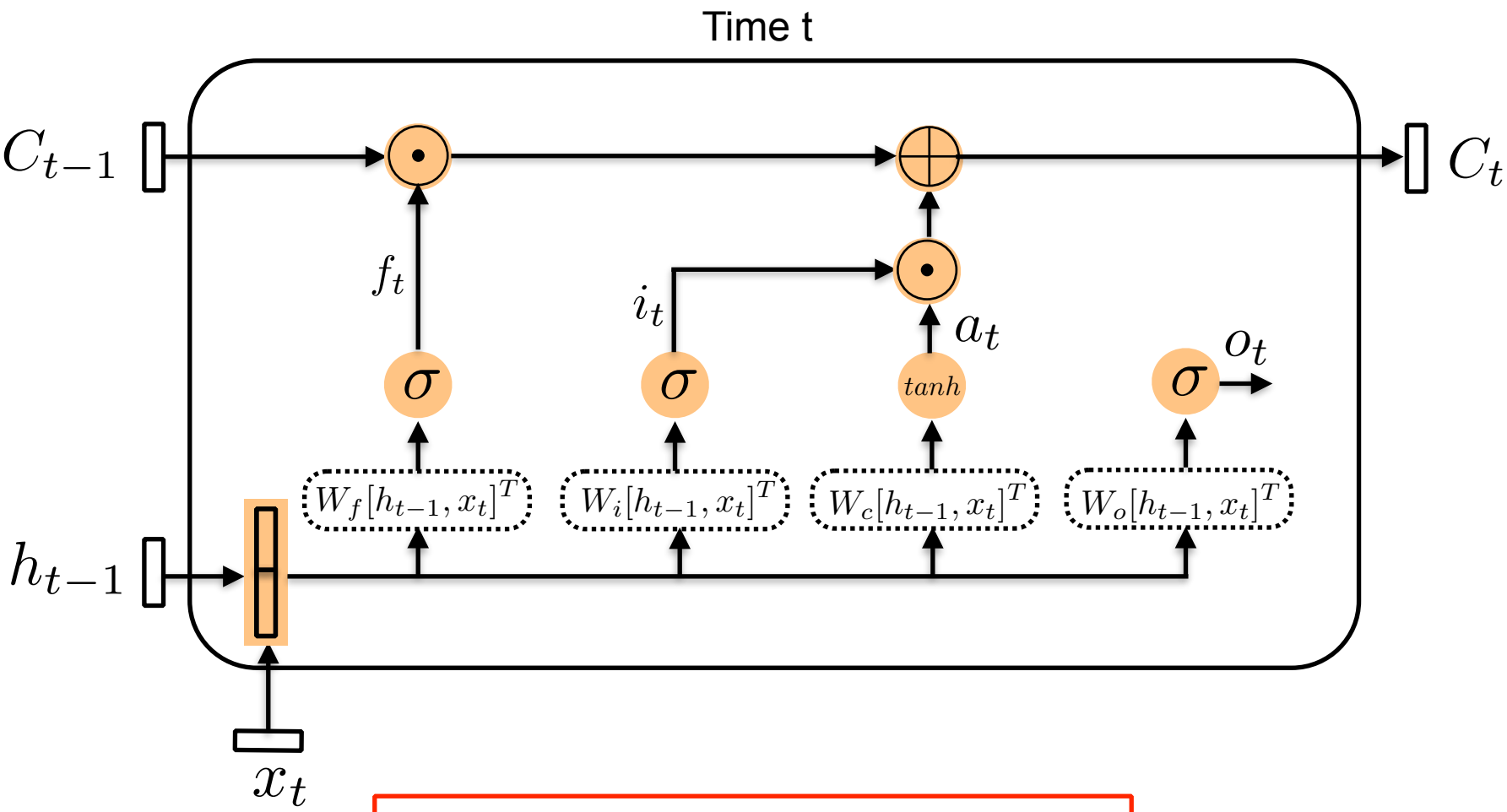


Memory Cell Update:

$$C_t = f_t \odot C_{t-1} + \underline{i_t \odot a_t}$$

What new information to add

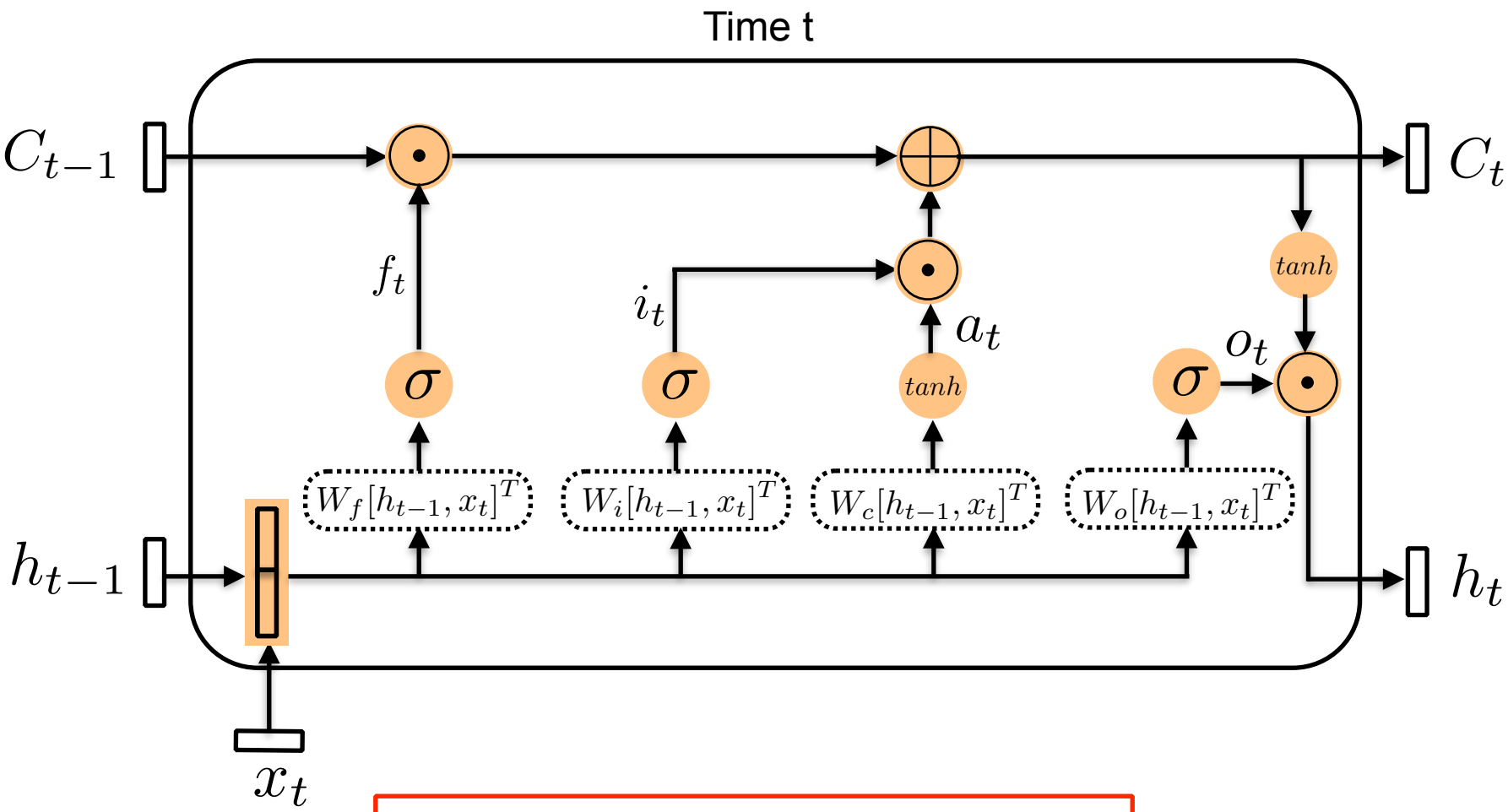
Long Short Term Memory Unit



<u>The Output Gate:</u>

$$o_t = \sigma(W_o[h_{t-1}, x_t]^T)$$

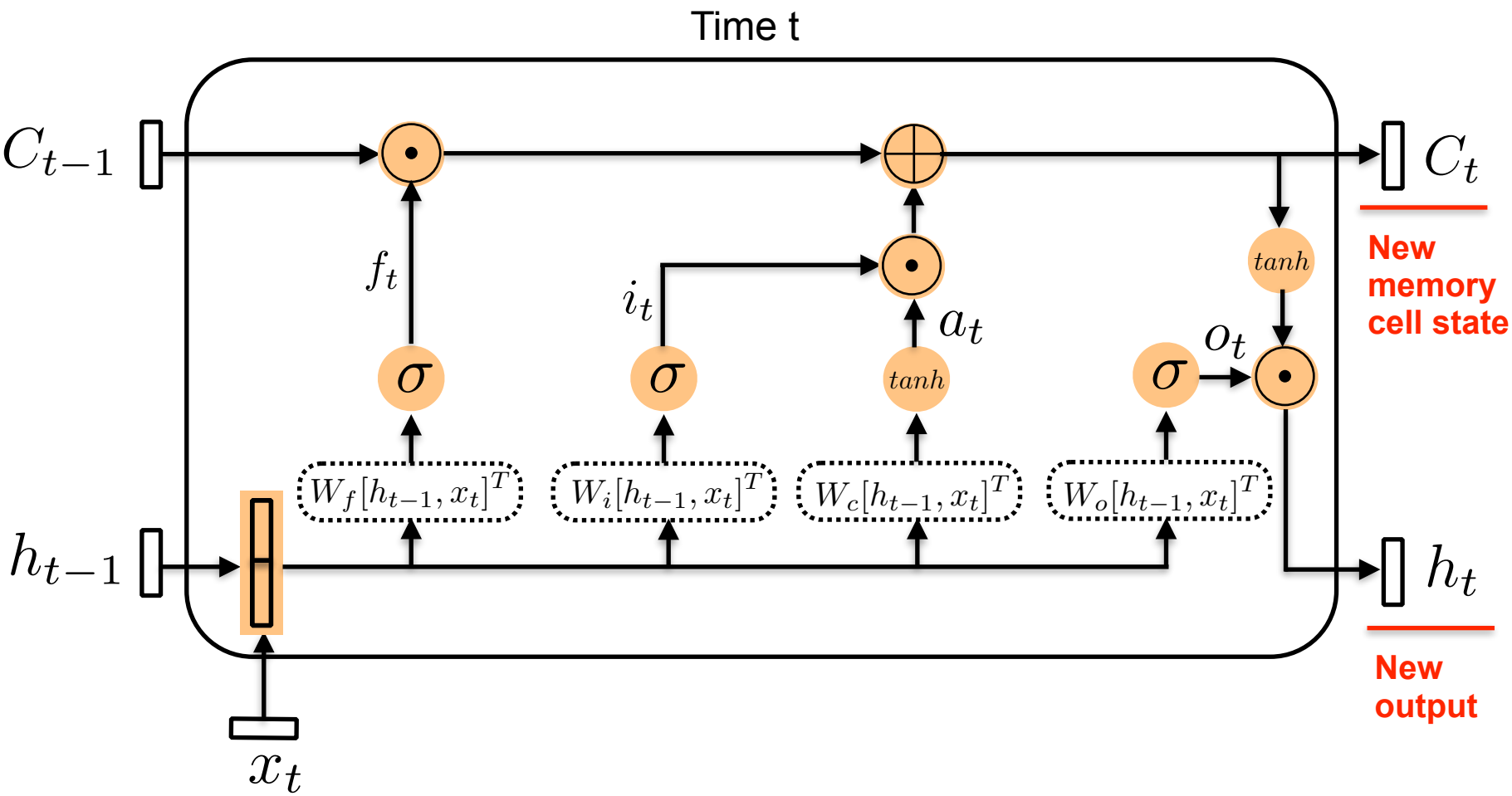
Long Short Term Memory Unit



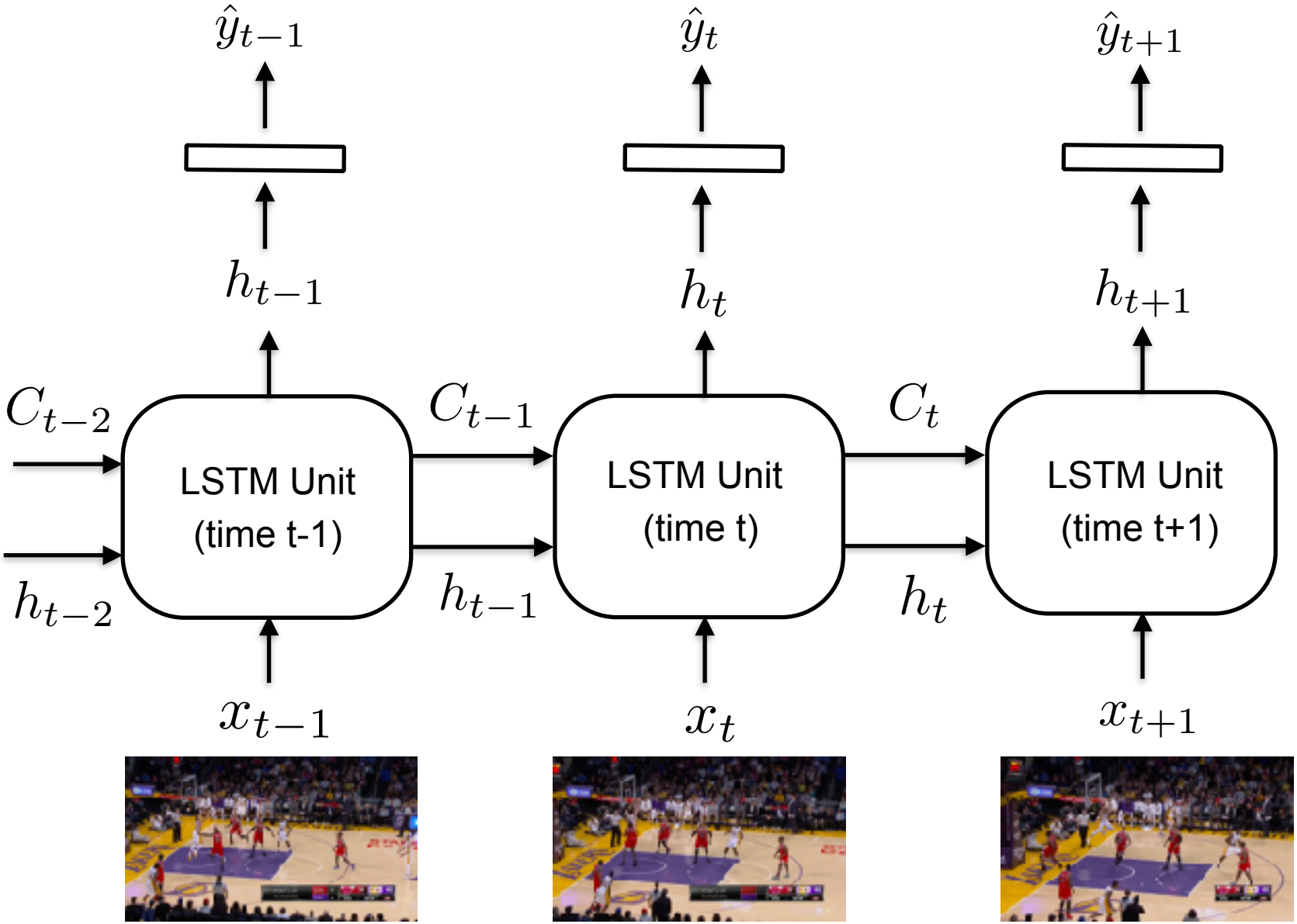
Output:

$$h_t = o_t \odot tanh(C_t)$$

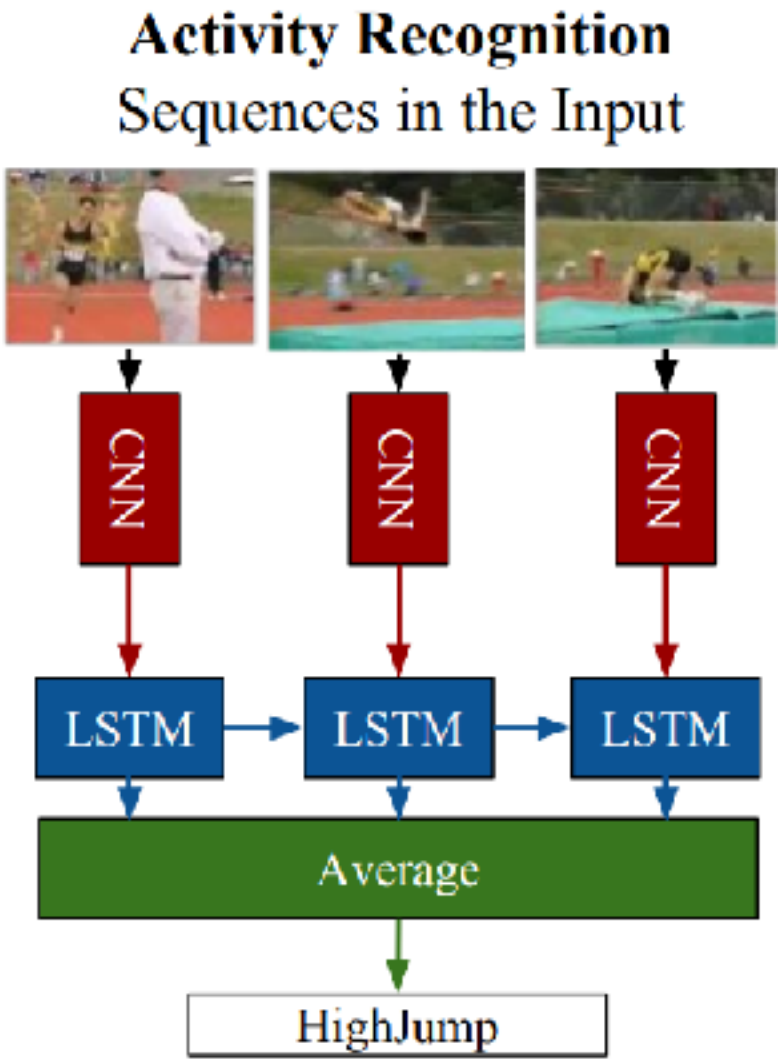
Long Short Term Memory Unit



Long Short Term Memory Network

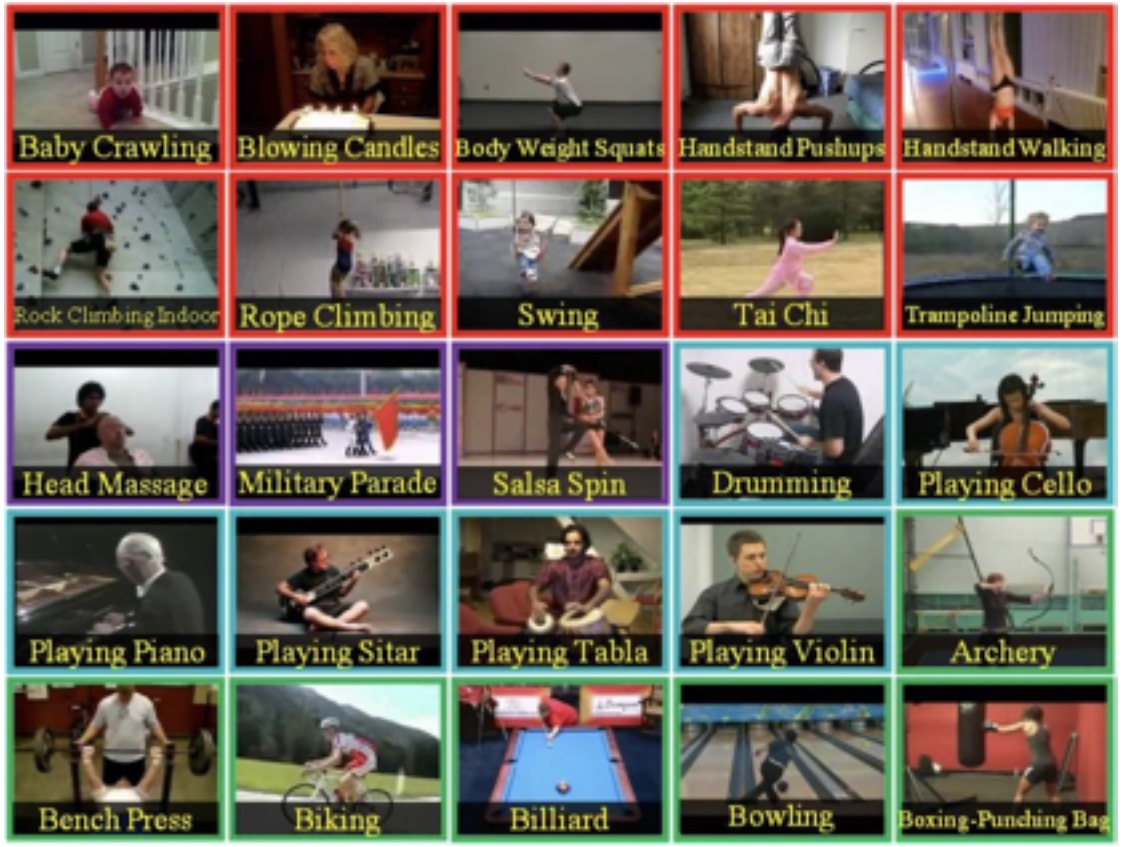


LRCN for Action Recognition



UCF-101 Dataset

- UCF-101 consists of 13,320 videos belonging to 101 action categories.



Action Recognition Results

- Performance is evaluated using action recognition accuracy.

Model	Single Input Type	
	RGB	Flow
Single frame	67.37	74.37
LRCN-fc_6	**68.20**	**77.28**

Ablation Study

- The authors investigate the performance gap between between LRCN and a single-frame baseline.

Label	Δ	Label	Δ
BoxingPunchingBag	40.82	BoxingSpeedBag	-16.22
HighJump	29.73	Mixing	-15.56
JumpRope	28.95	Knitting	-14.71
CricketShot	28.57	Typing	-13.95
Basketball	28.57	Skiing	-12.50
WallPushups	25.71	BaseballPitch	-11.63
Nunchucks	22.86	BrushingTeeth	-11.11
ApplyEyeMakeup	22.73	Skijet	-10.71
HeadMassage	21.95	Haircut	-9.10
Drumming	17.78	TennisSwing	-8.16

Summary

- One of the first approaches to integrate CNNs and LSTMs for visual sequence modeling.
- The entire system can be trained end-to-end.
- The gains from temporal modeling are somewhat limited.

Discussion Points

- The CNN + LSTM architecture was not as successful as we had hoped it would be. Why?

- Is motion information useful on benchmarks like UCF-101? If not, why the results are so much better with the optical flow modality?