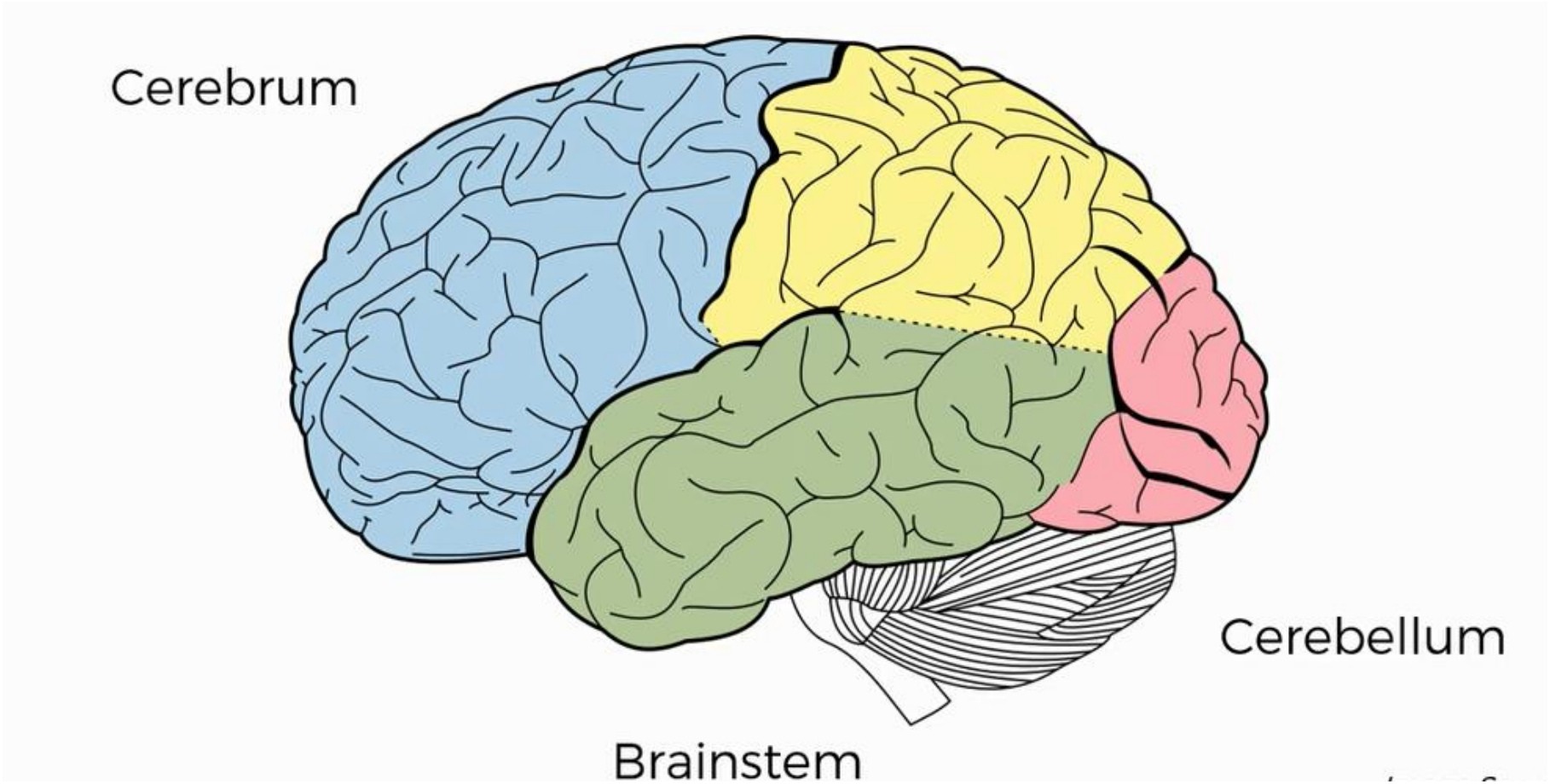
Recurrent Neural Network


Cerebrum
Cerebellum
Brainstem

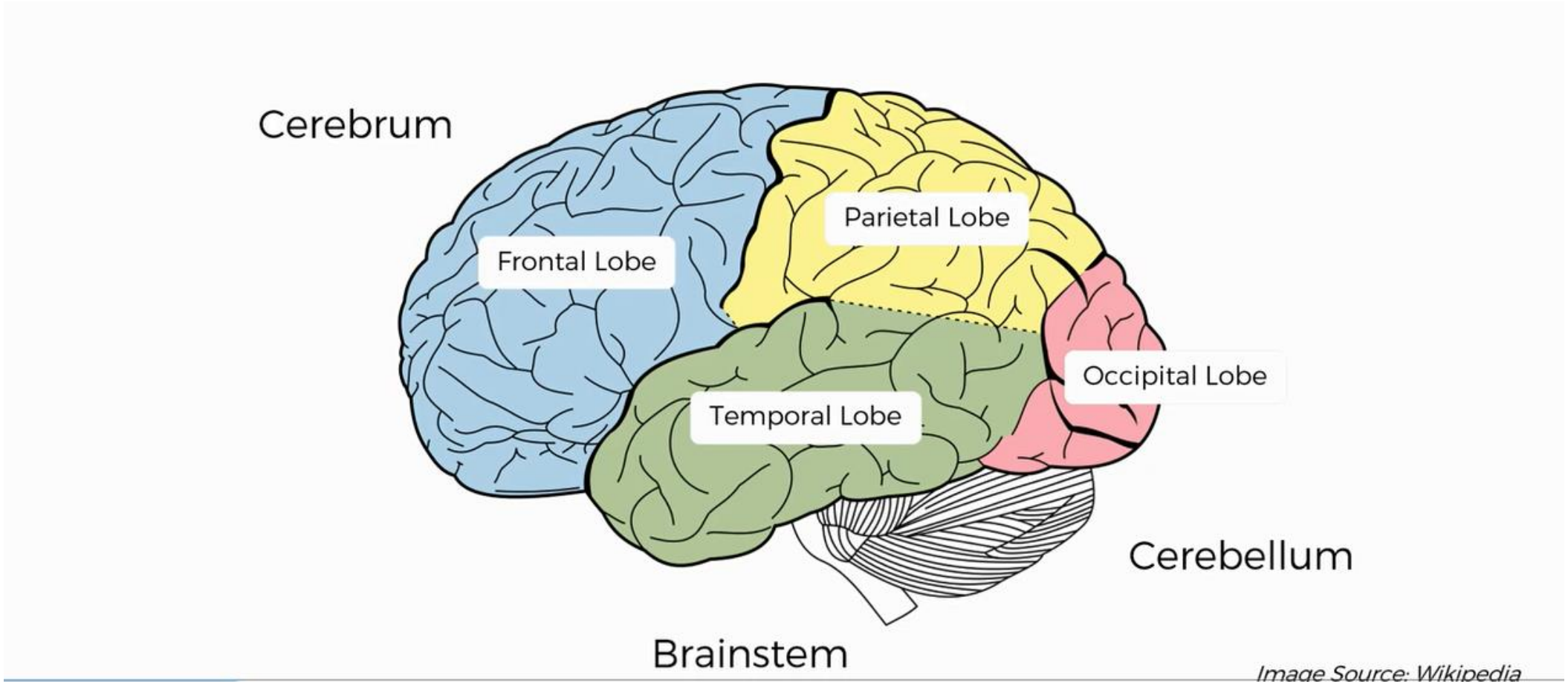


Image Source: Wikipedia

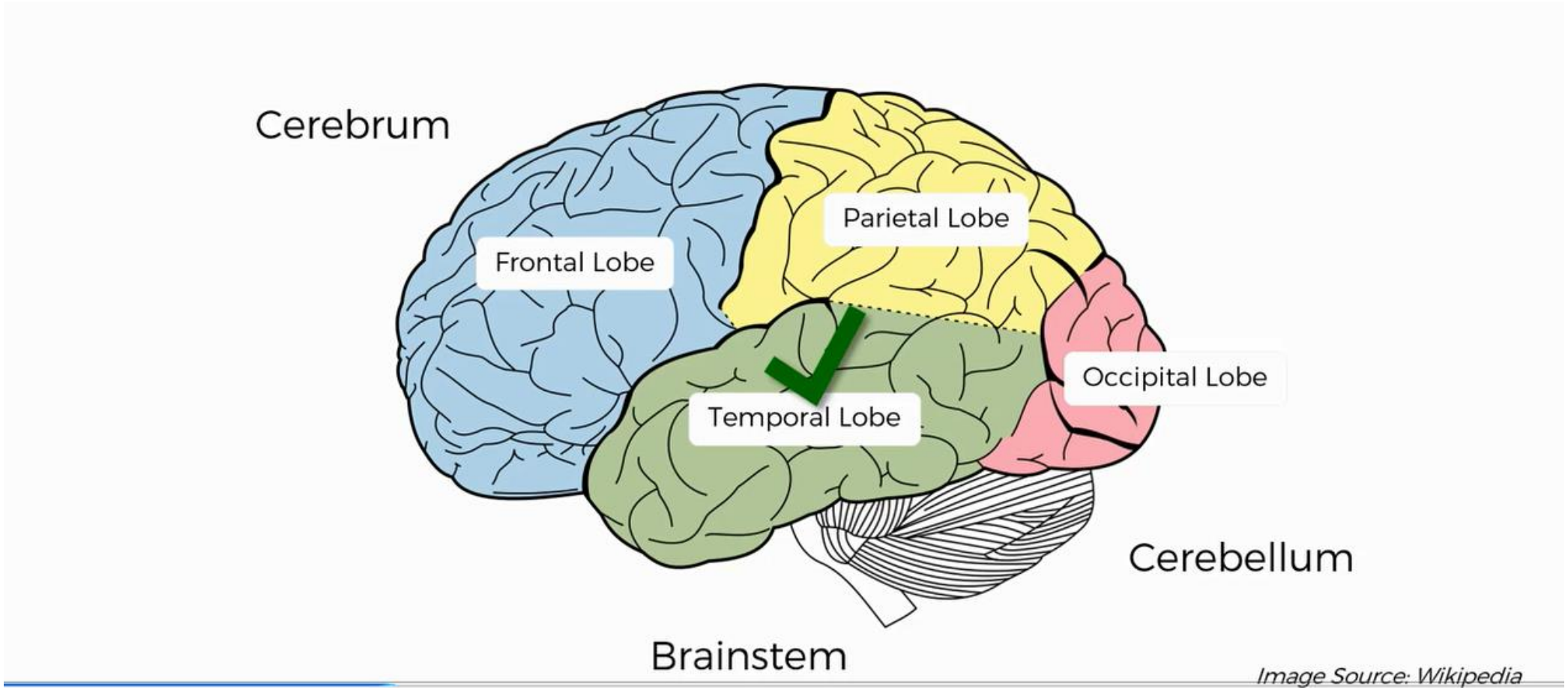
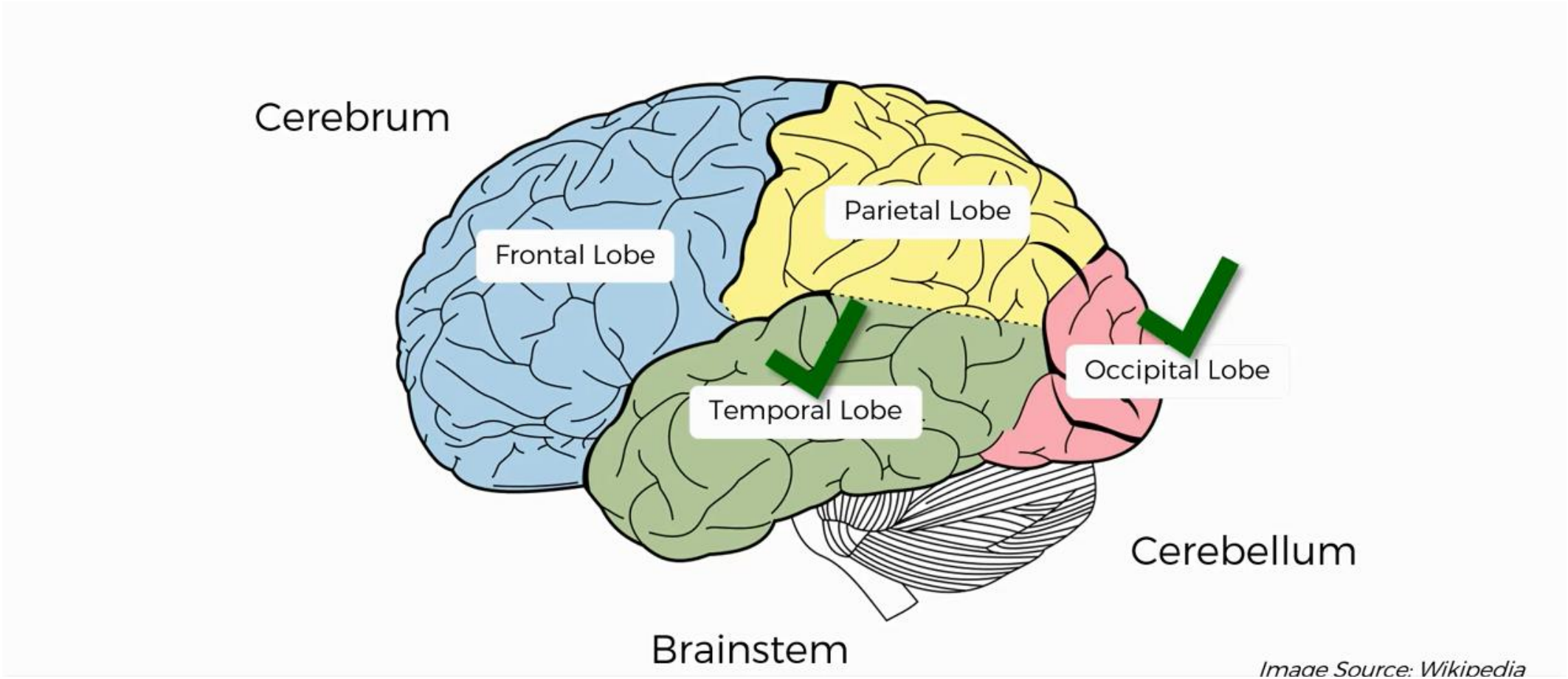


Image Source: Wikipedia


Cerebrum
Frontal Lobe
Parietal Lobe
Temporal Lobe
Occipital Lobe
Cerebellum
Brainstem
Image Source: Wikipedia

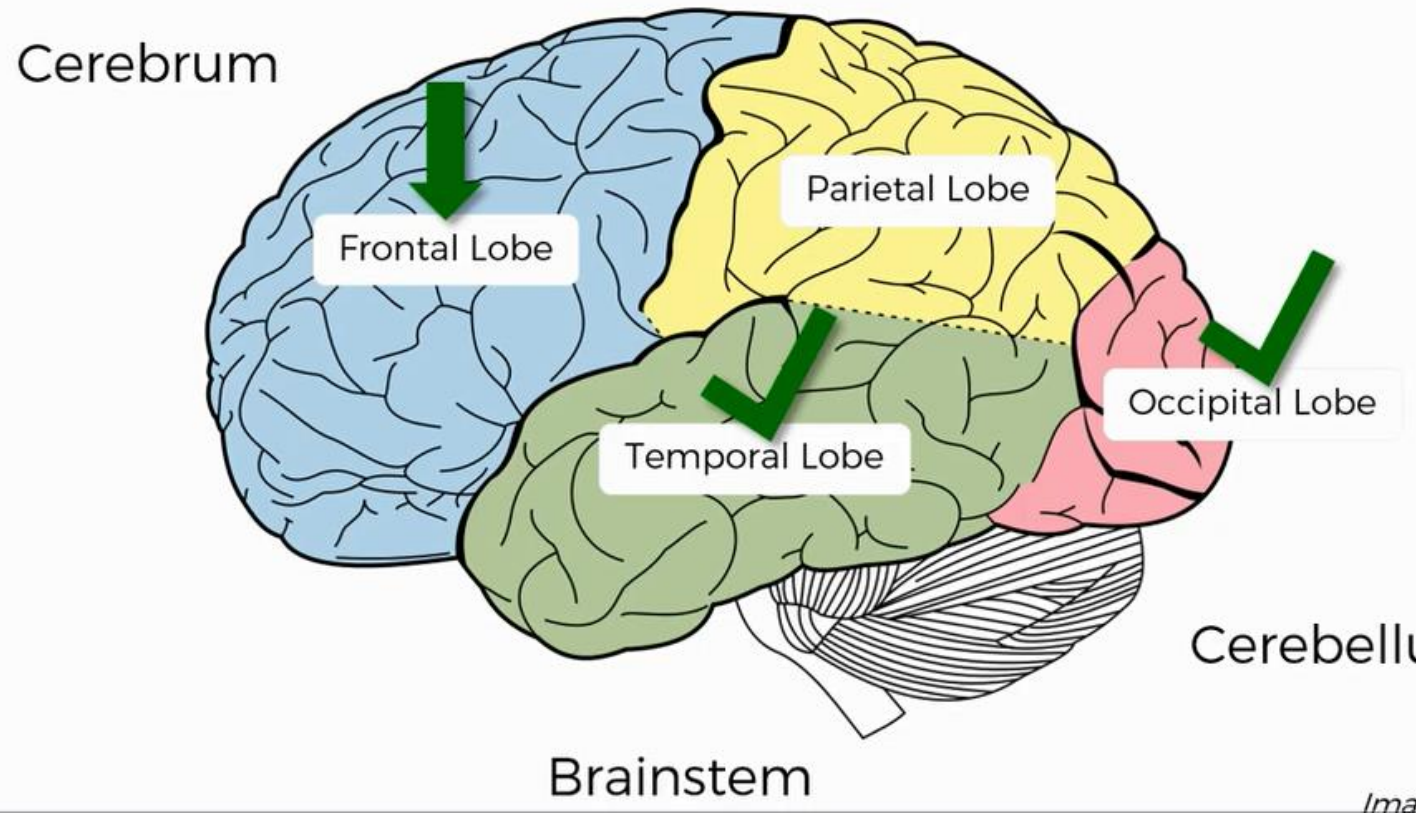
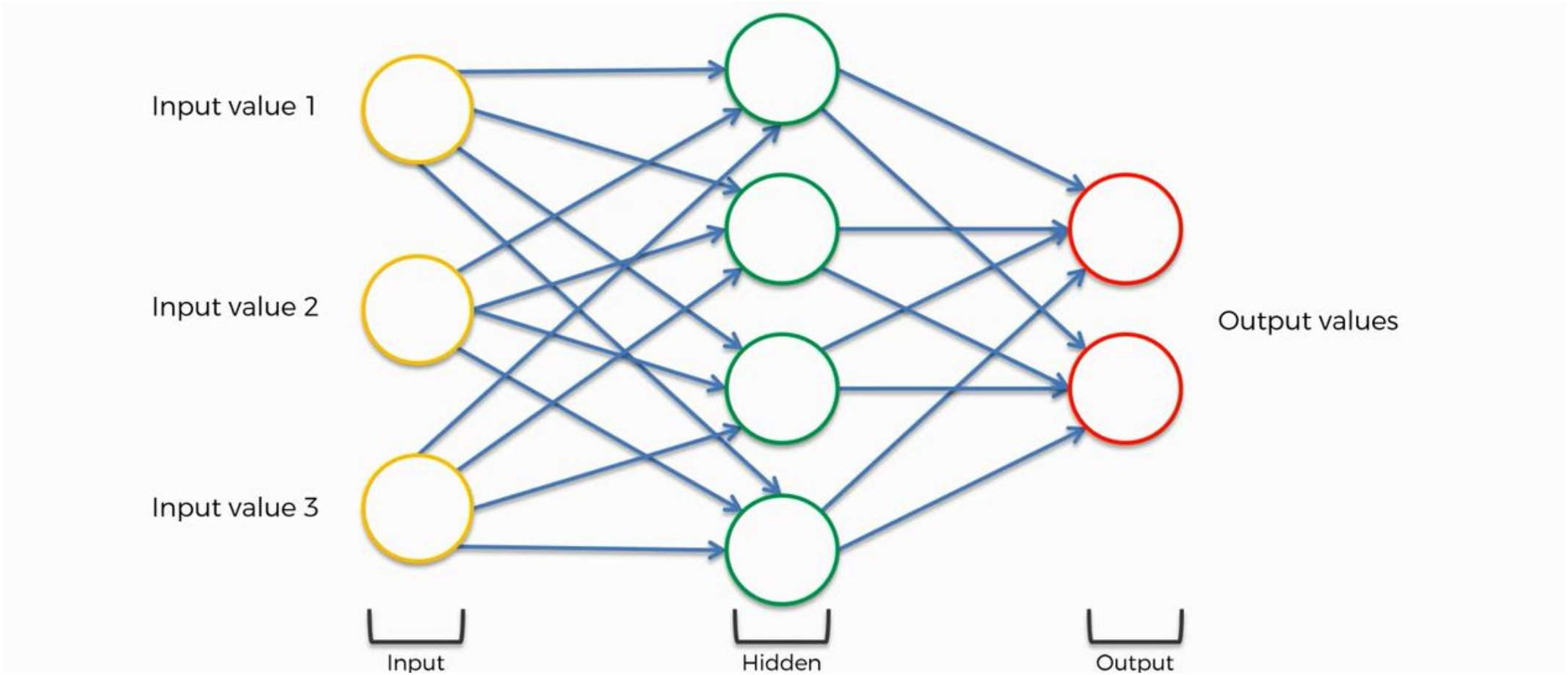
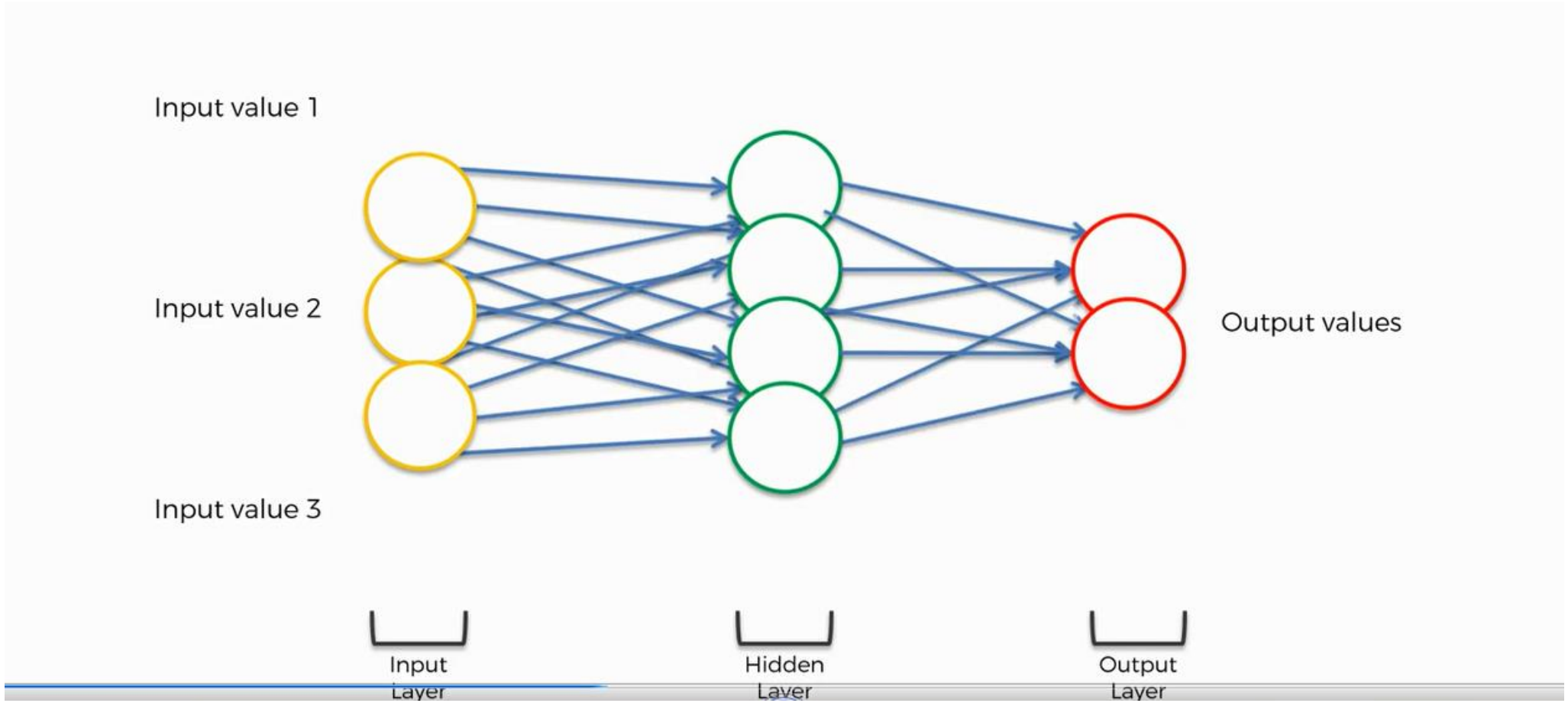
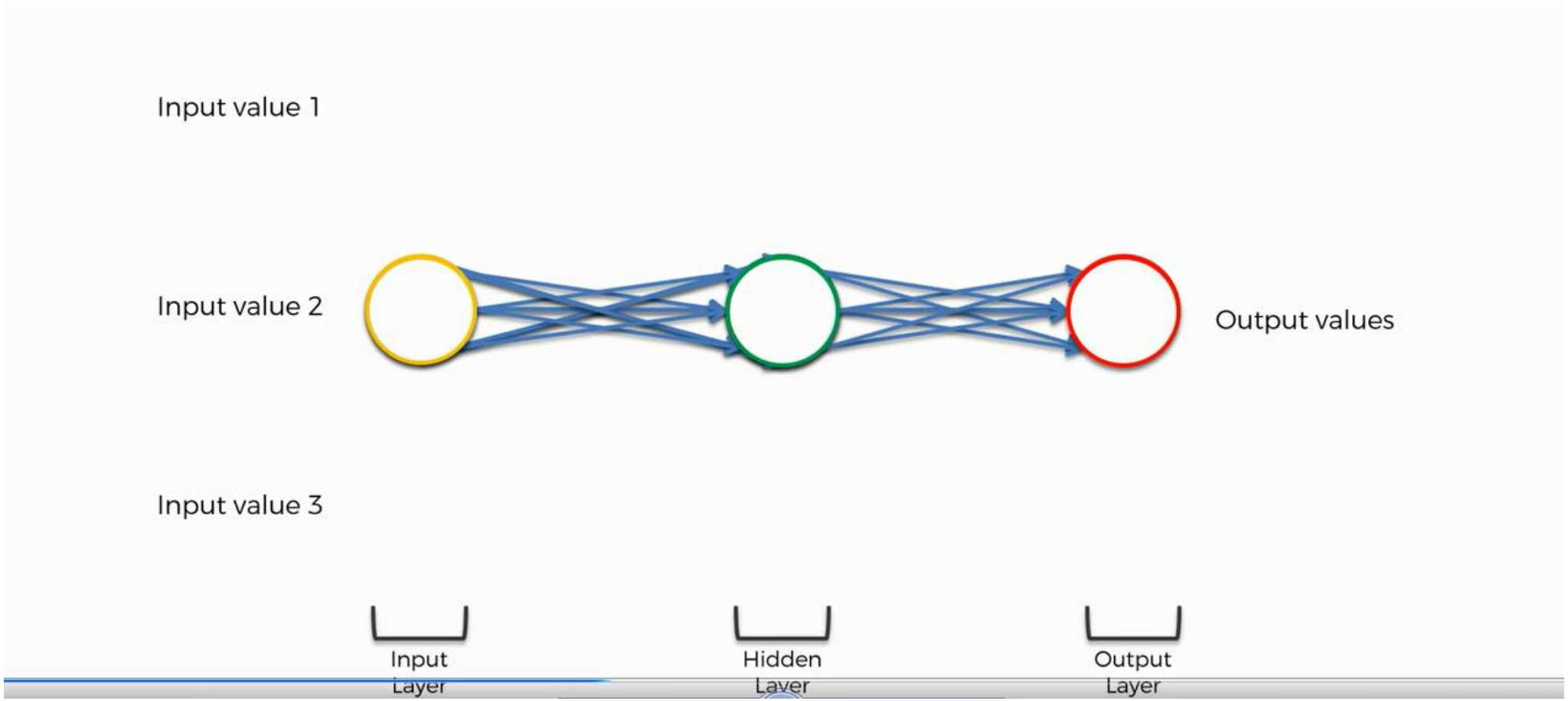
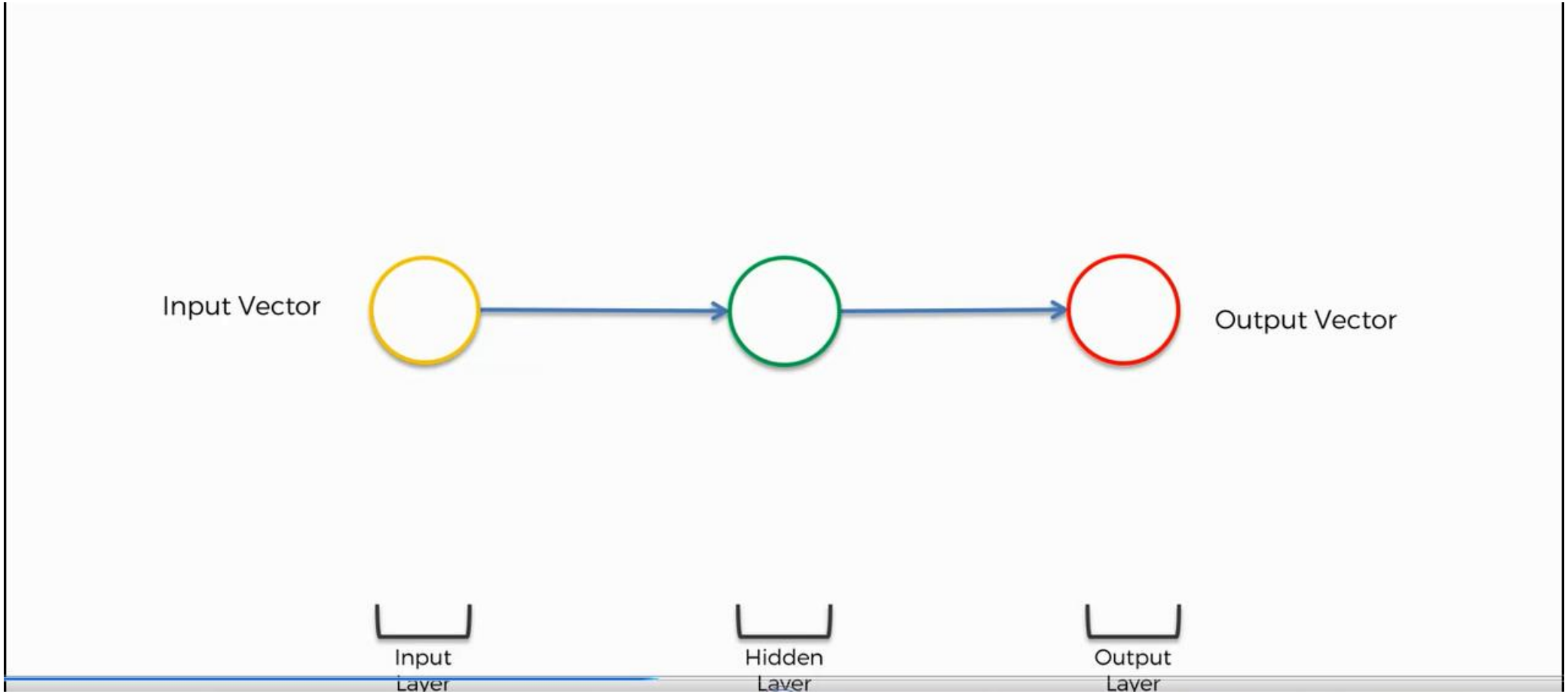


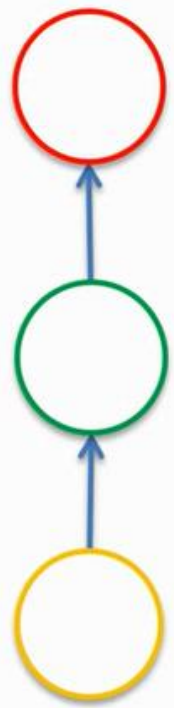
Image Source: Wikipedia

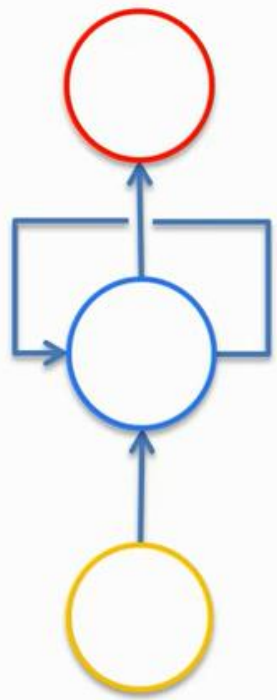

Input value 1
Input value 2
Input value 3
Output values
Input
Hidden
Output

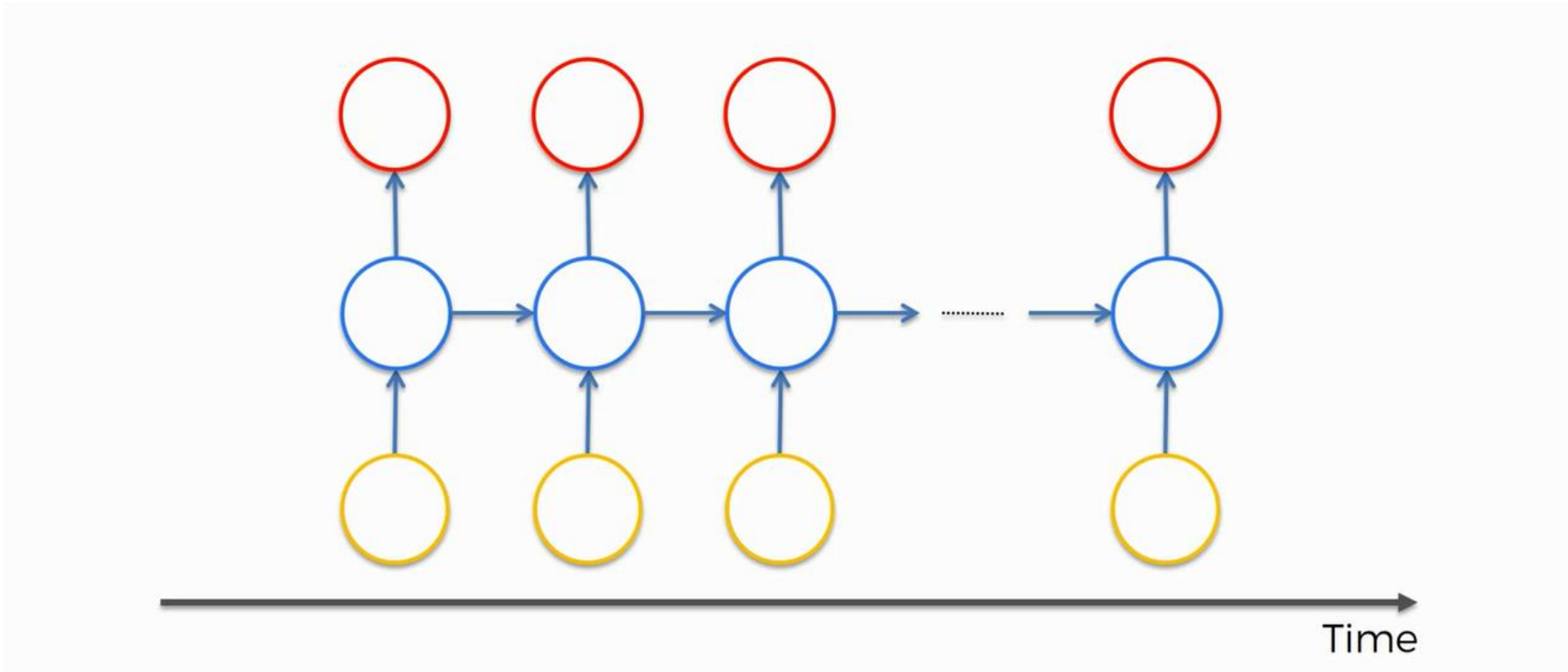

Input value 1
Input value 2
Input value 3
Output values
Input Layer
Hidden Layer
Output Layer

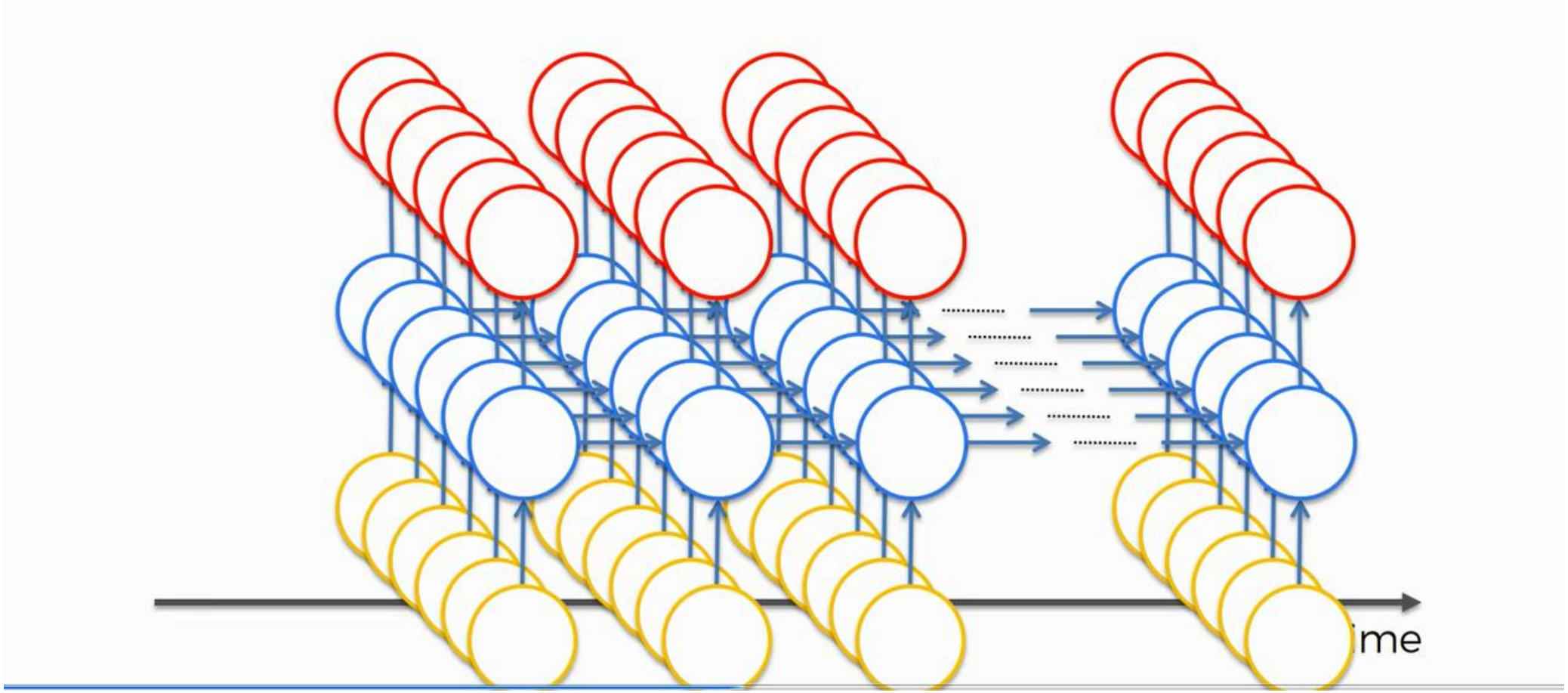

Input value 1
Input value 2
Input value 3
Output values
Input Layer
Hidden Layer
Output Layer

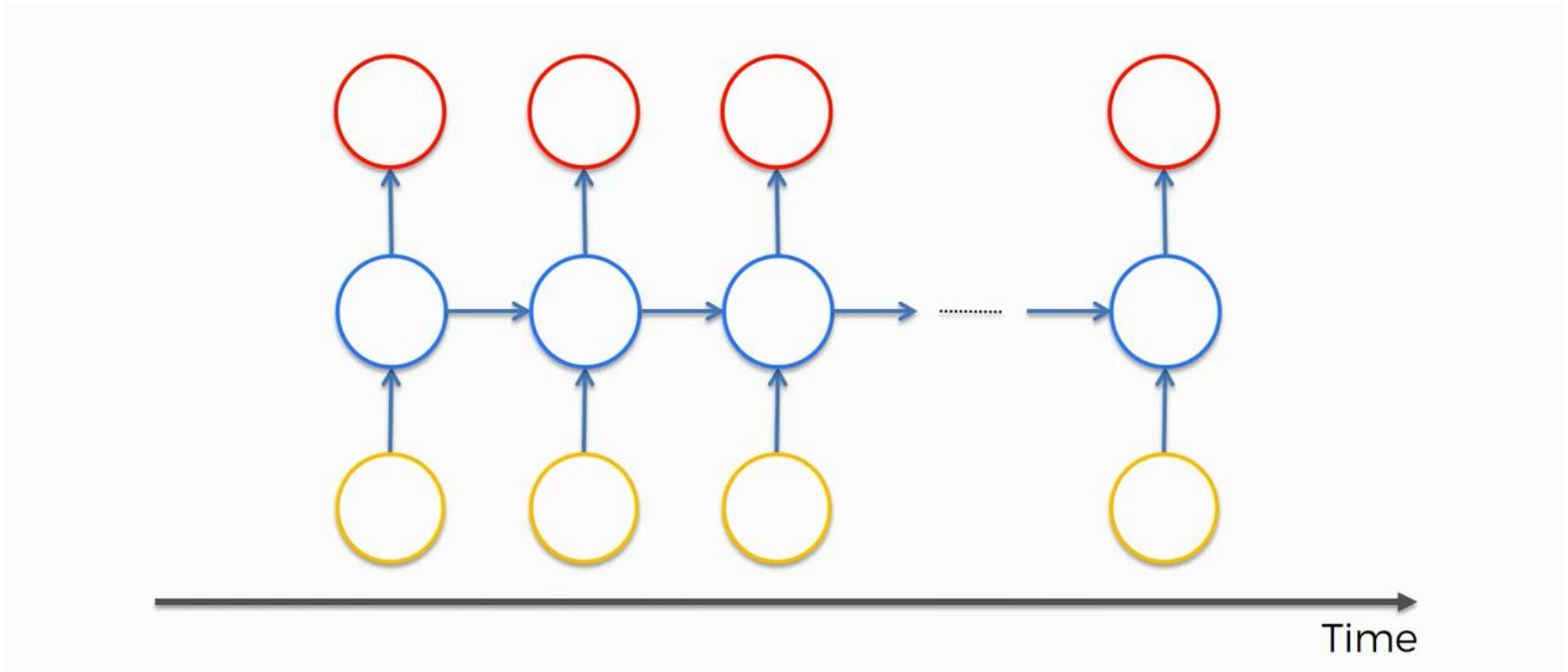

Input Vector
Output Vector
Input Layer
Hidden Layer
Output Layer





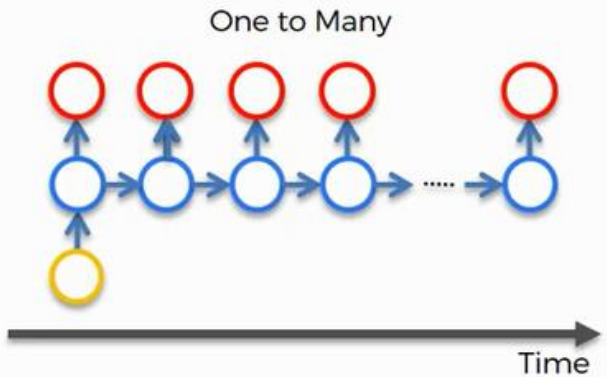

Time


ime


Time

"black and white dog jumps over bar."

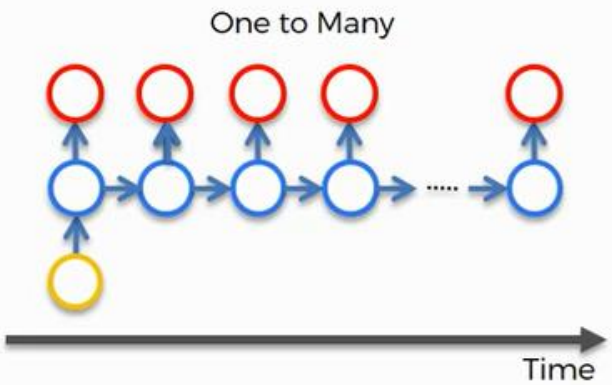
karpathy.github.io

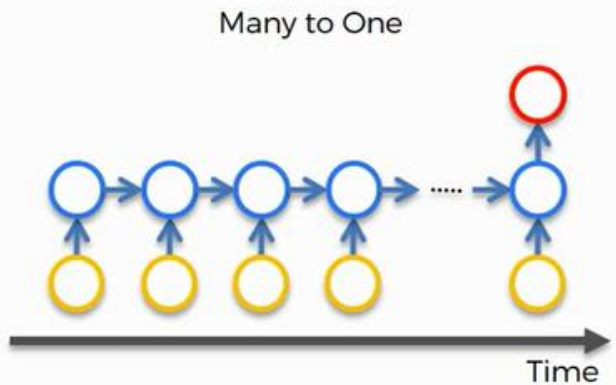


Reference: karpathy.github.io

"black and white dog jumps over bar."

karpathy.github.io



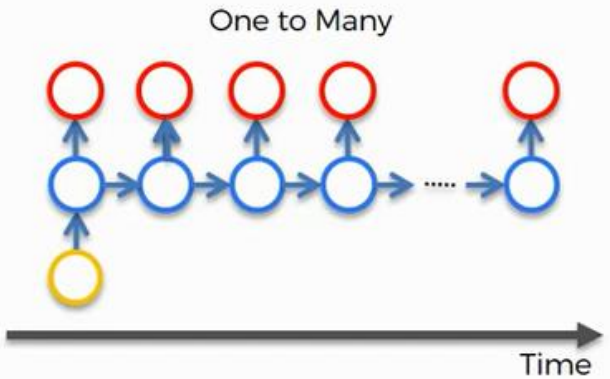


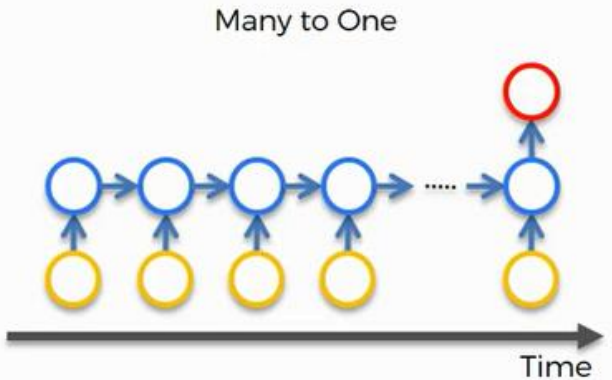
dev.havenondemand.com

Reference: karpathy.github.io

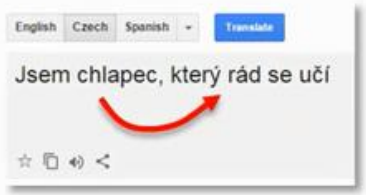
"black and white dog jumps over bar."

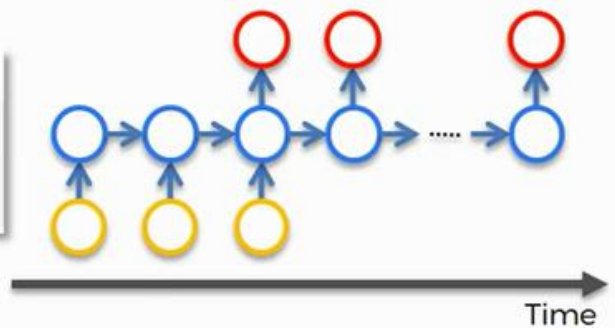
karpathy.github.io





dev.havenondemand.com

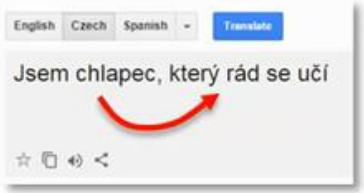


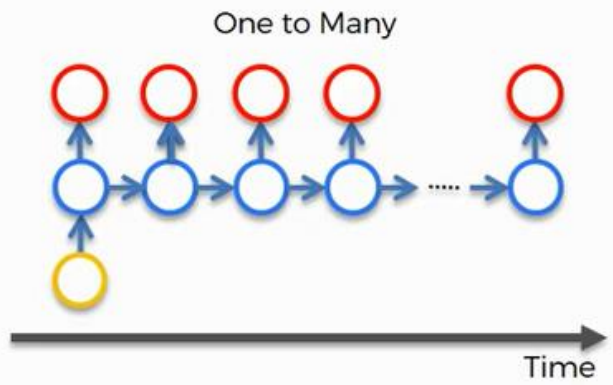


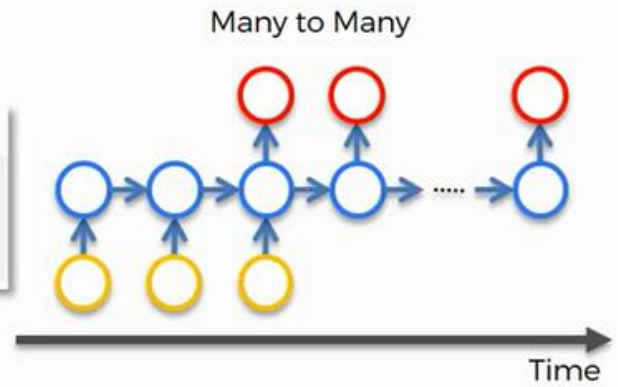
Reference: karpathy.github.io

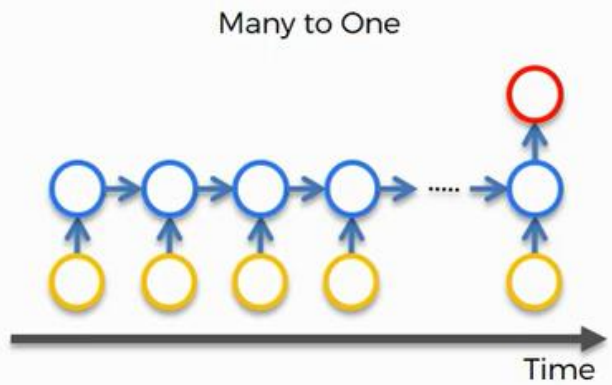
"black and white dog jumps over bar."

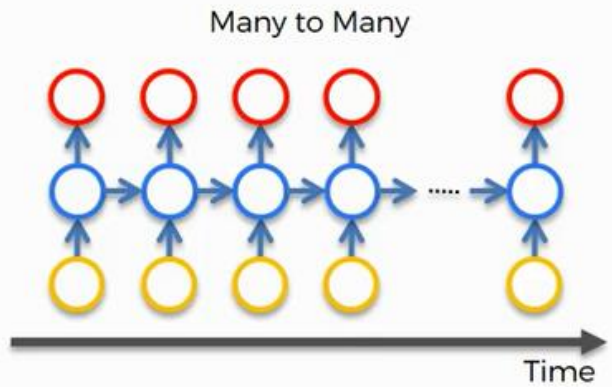
karpathy.github.io







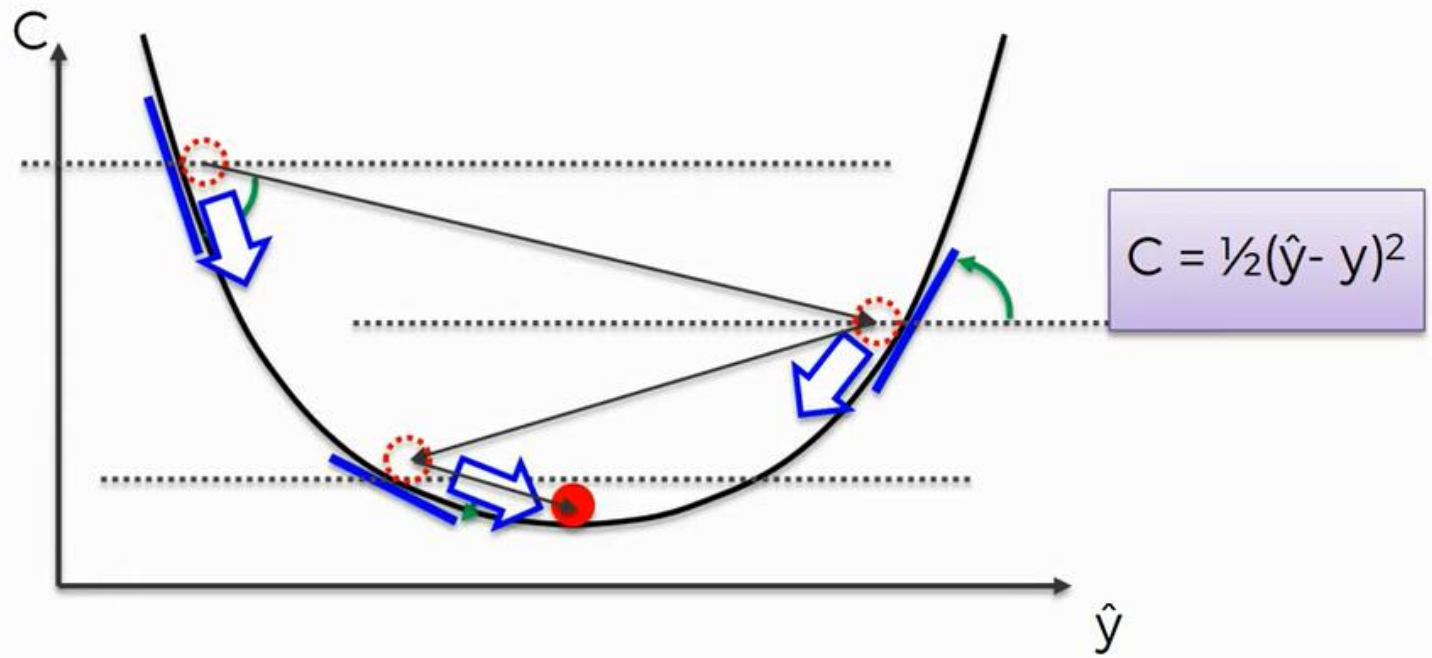


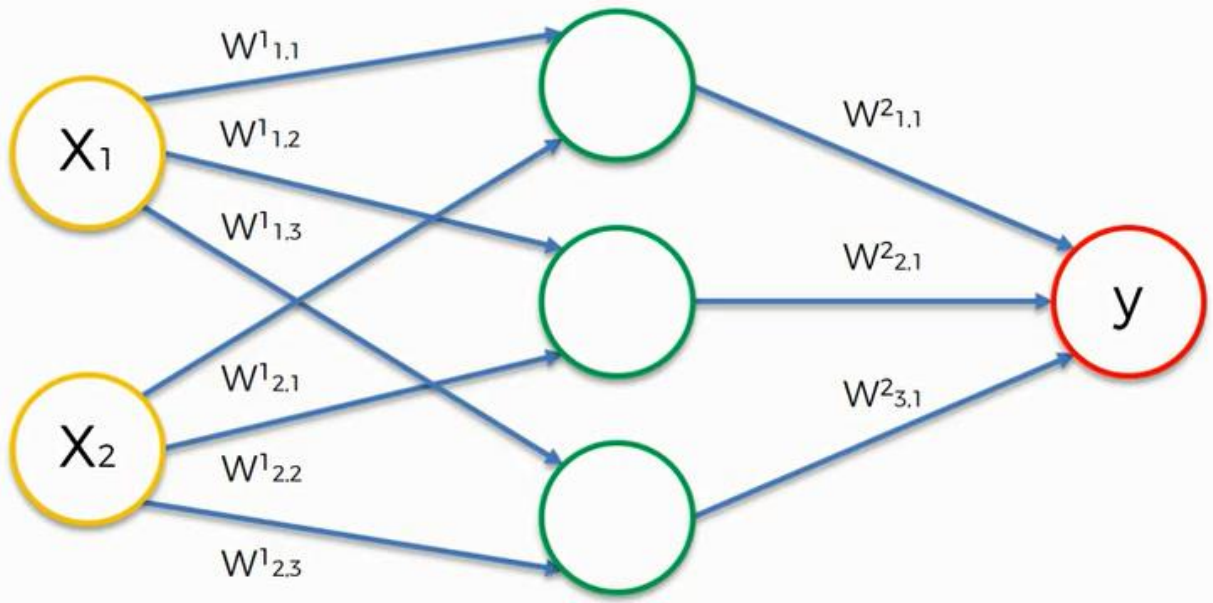


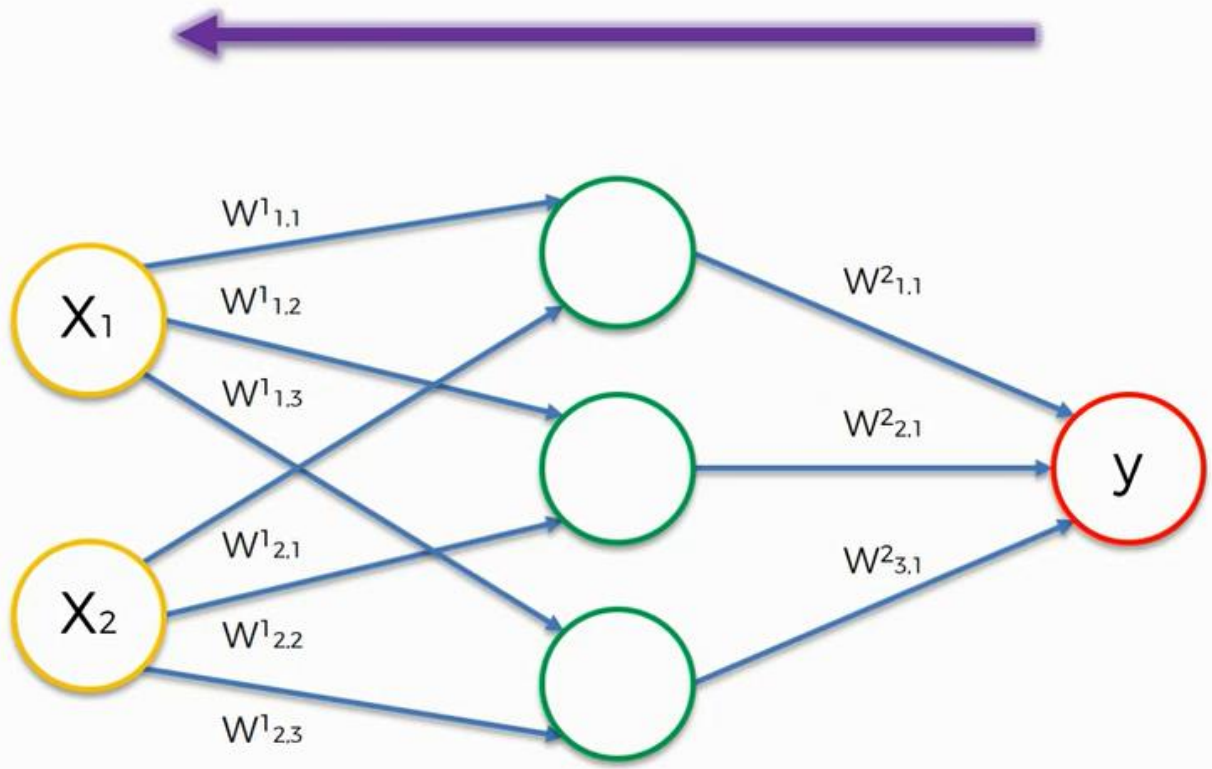
dev.havenondemand.com

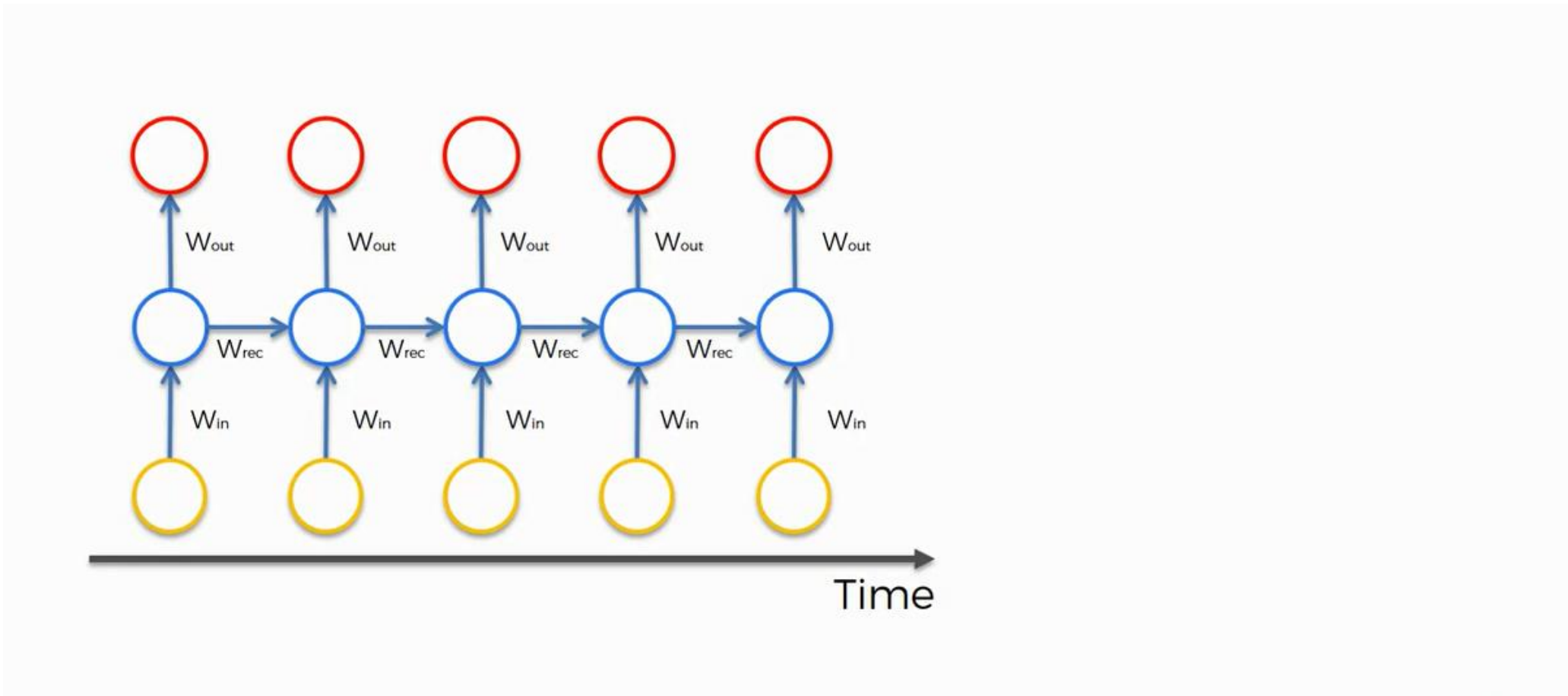
Reference: karpathy.github.io

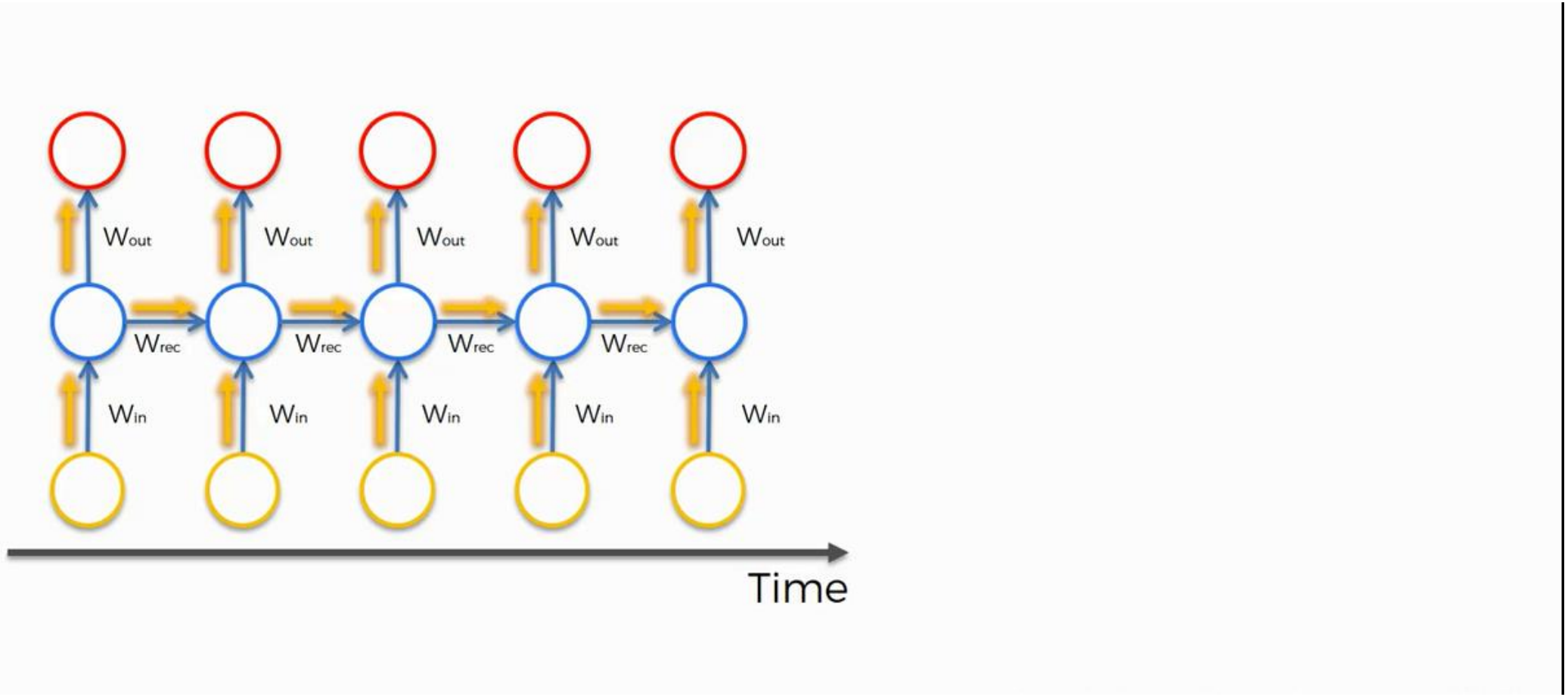
Vanishing Gradient

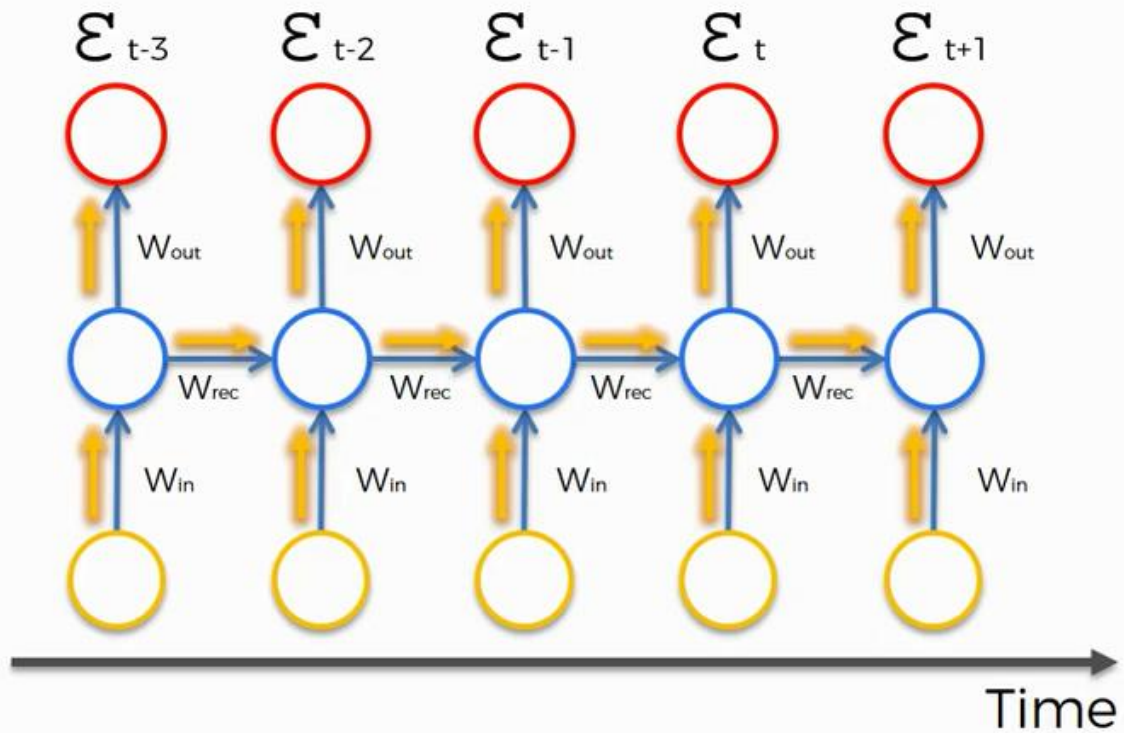

C
ŷ
C = ½(ŷ- y)²

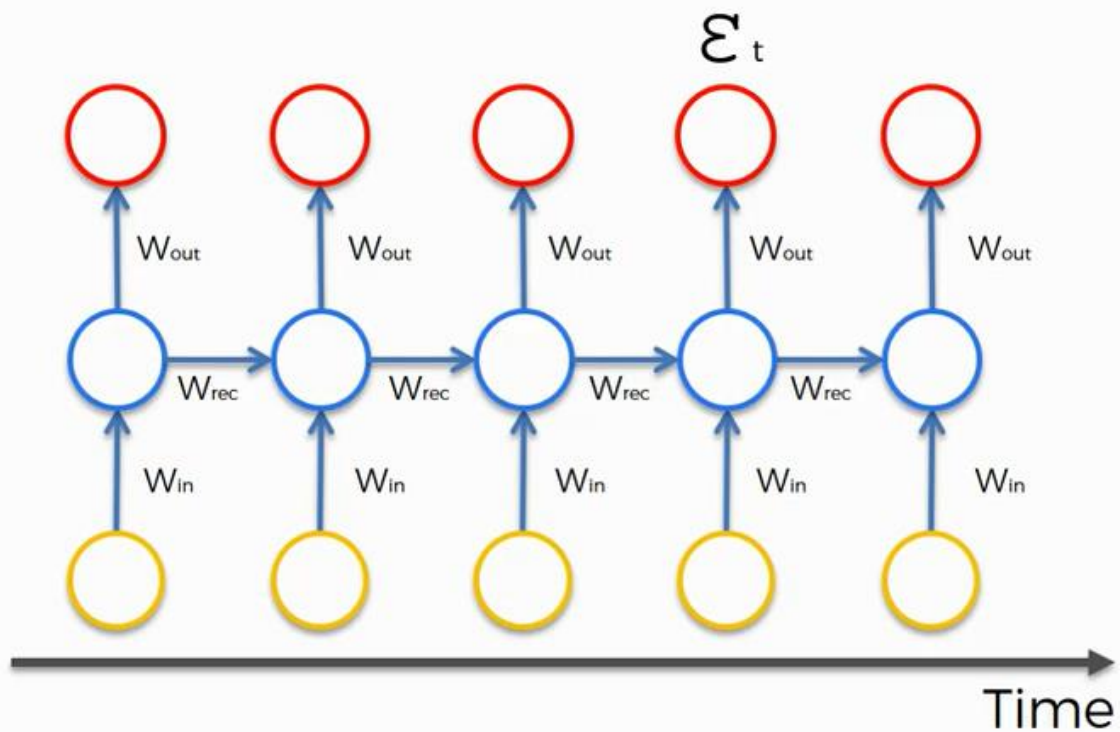

$W^1_{1,1}$
$W^1_{1,2}$
$W^1_{1,3}$
$W^1_{2,1}$
$W^1_{2,2}$
$W^1_{2,3}$
$W^2_{1,1}$
$W^2_{2,1}$
$W^2_{3,1}$
X_1
X_2
y

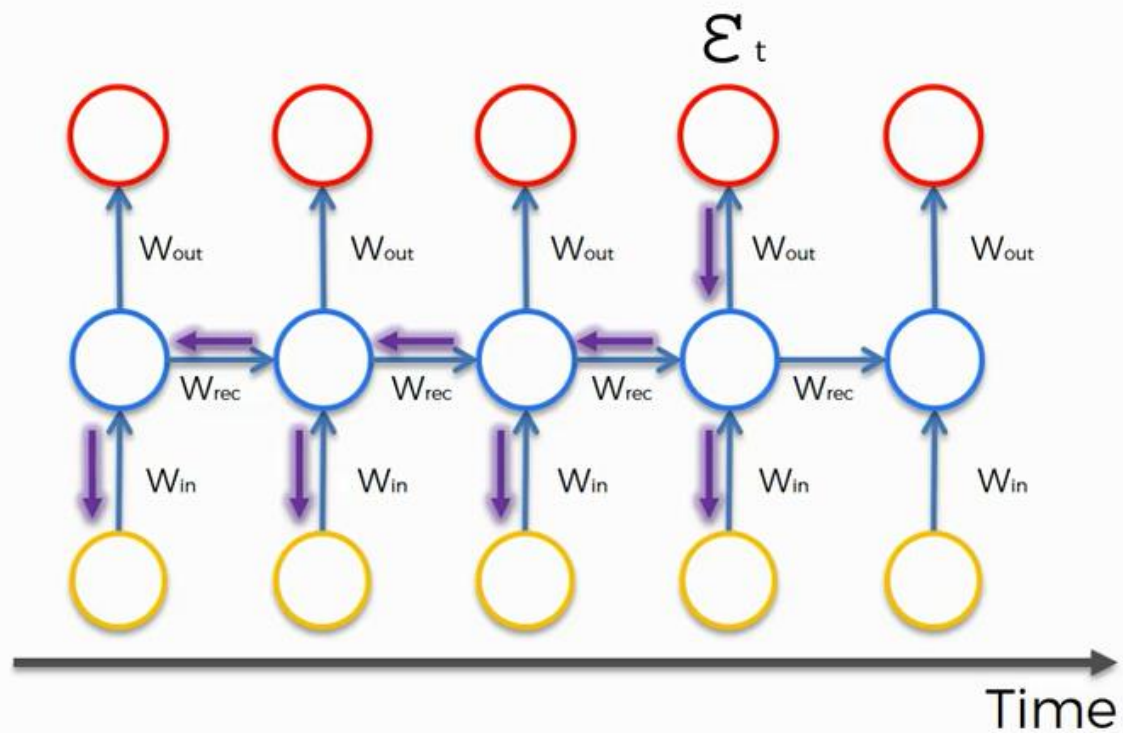

$W^1_{1,1}$
$W^1_{1,2}$
$W^1_{1,3}$
$W^1_{2,1}$
$W^1_{2,2}$
$W^1_{2,3}$
$W^2_{1,1}$
$W^2_{2,1}$
$W^2_{3,1}$
X_1
X_2
y

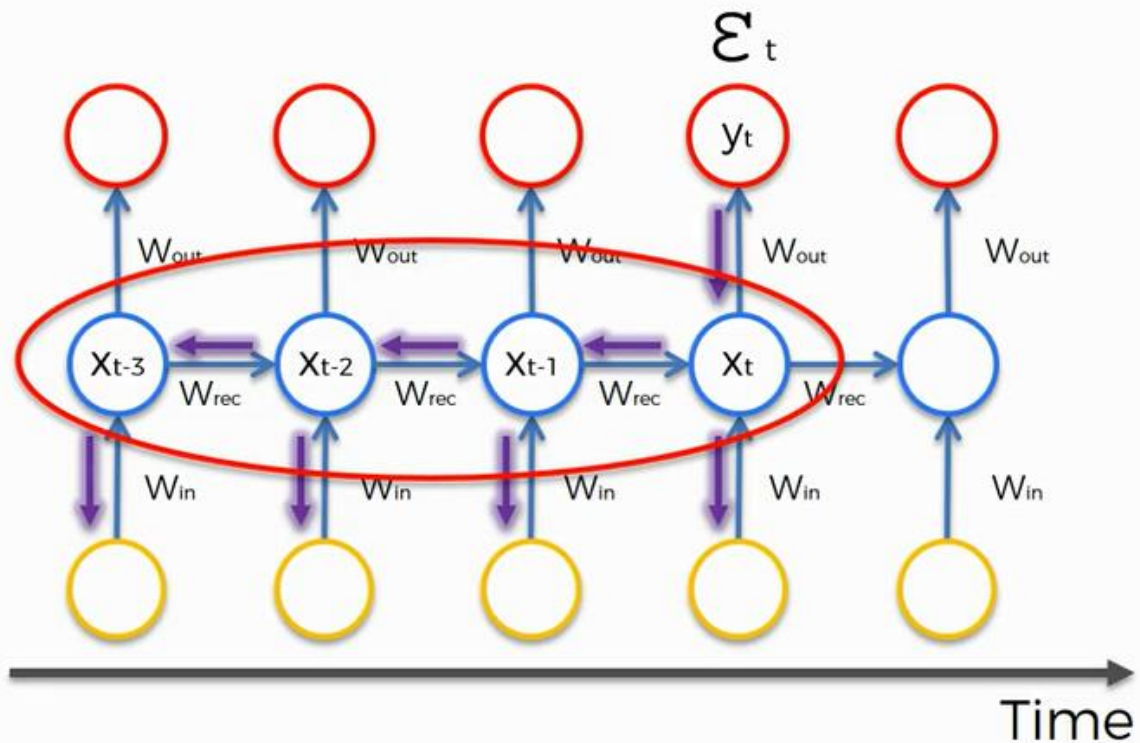

Wout
Wout
Wout
Wout
Wout
Wrec
Wrec
Wrec
Wrec
Win
Win
Win
Win
Win
Time

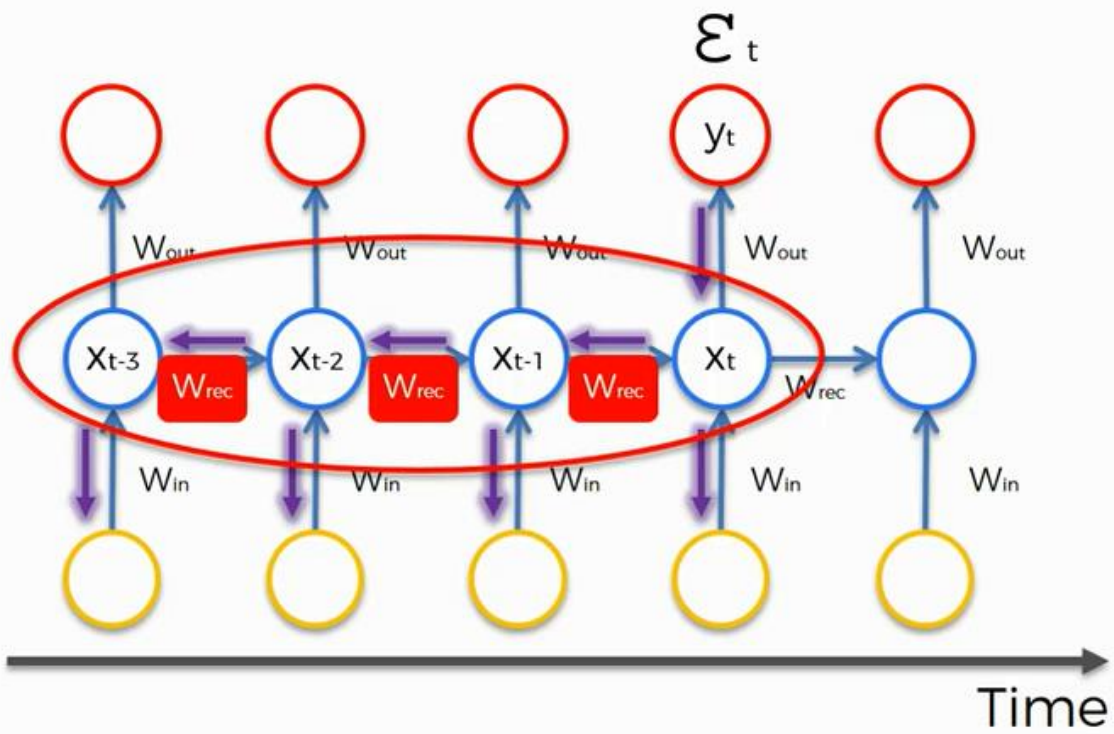

Wout
Wout
Wout
Wout
Wout
Wrec
Wrec
Wrec
Wrec
Win
Win
Win
Win
Win
Time


ε t-3
ε t-2
ε t-1
ε t
ε t+1
Wout
Wrec
Win
Time


$\mathcal{E}_t$
W_{out}
W_{out}
W_{out}
W_{out}
W_{out}
W_{rec}
W_{rec}
W_{rec}
W_{rec}
W_{in}
W_{in}
W_{in}
W_{in}
W_{in}
Time


ε t
Wout
Wout
Wout
Wout
Wout
Wrec
Wrec
Wrec
Wrec
Win
Win
Win
Win
Win
Time

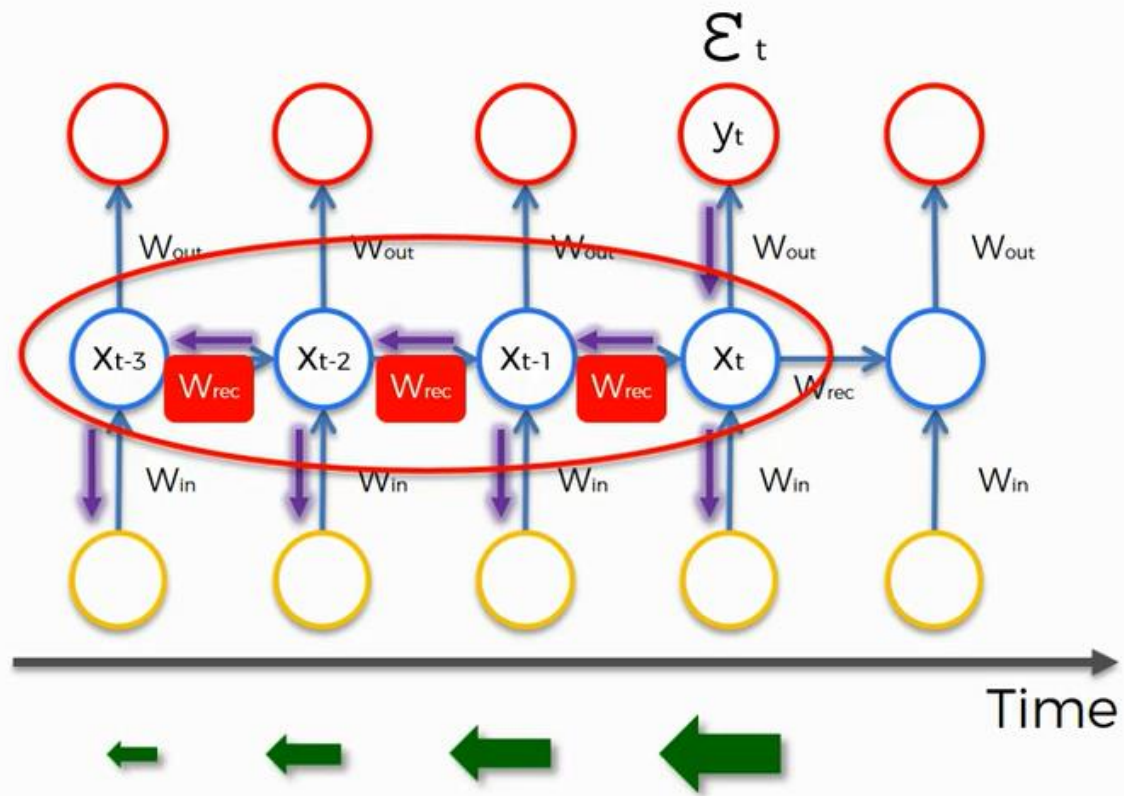

$\mathcal{E}_t$
y_t
W_{out}
W_{out}
W_{out}
W_{out}
W_{out}
x_{t-3}
x_{t-2}
x_{t-1}
x_t
W_{rec}
W_{rec}
W_{rec}
W_{rec}
W_{in}
W_{in}
W_{in}
W_{in}
W_{in}
Time



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

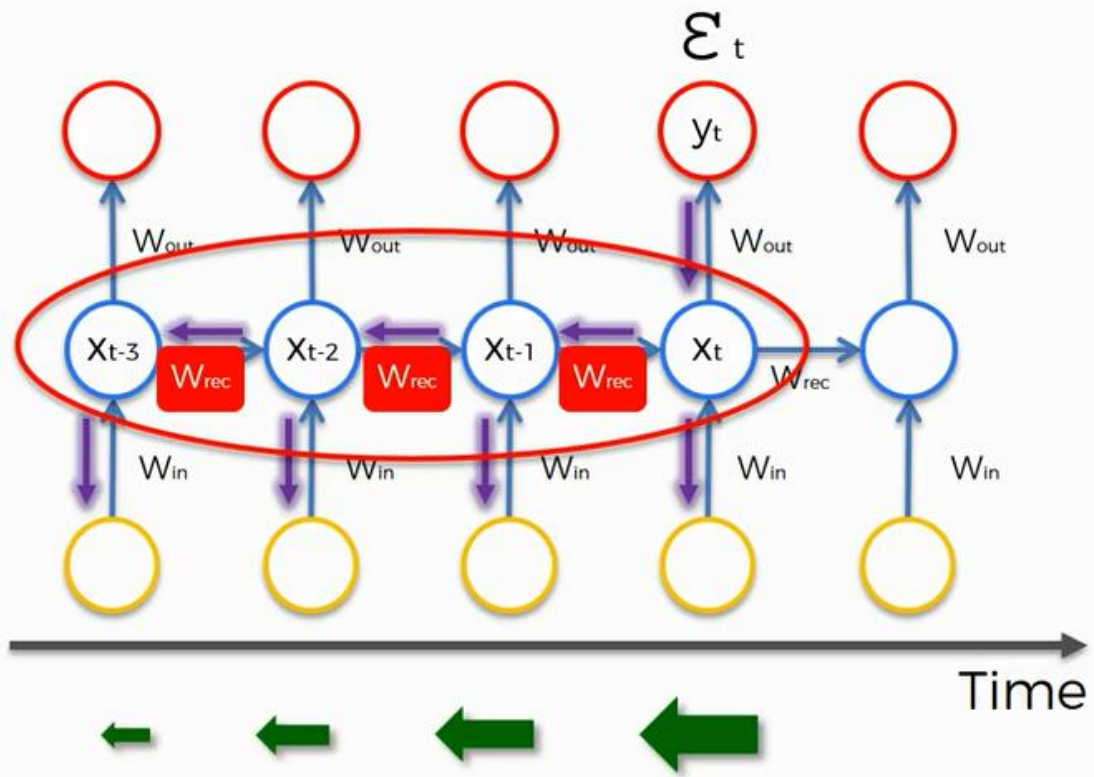
$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \geq i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \geq i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$



$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$

W_{rec} ~ small ⇨ Vanishing

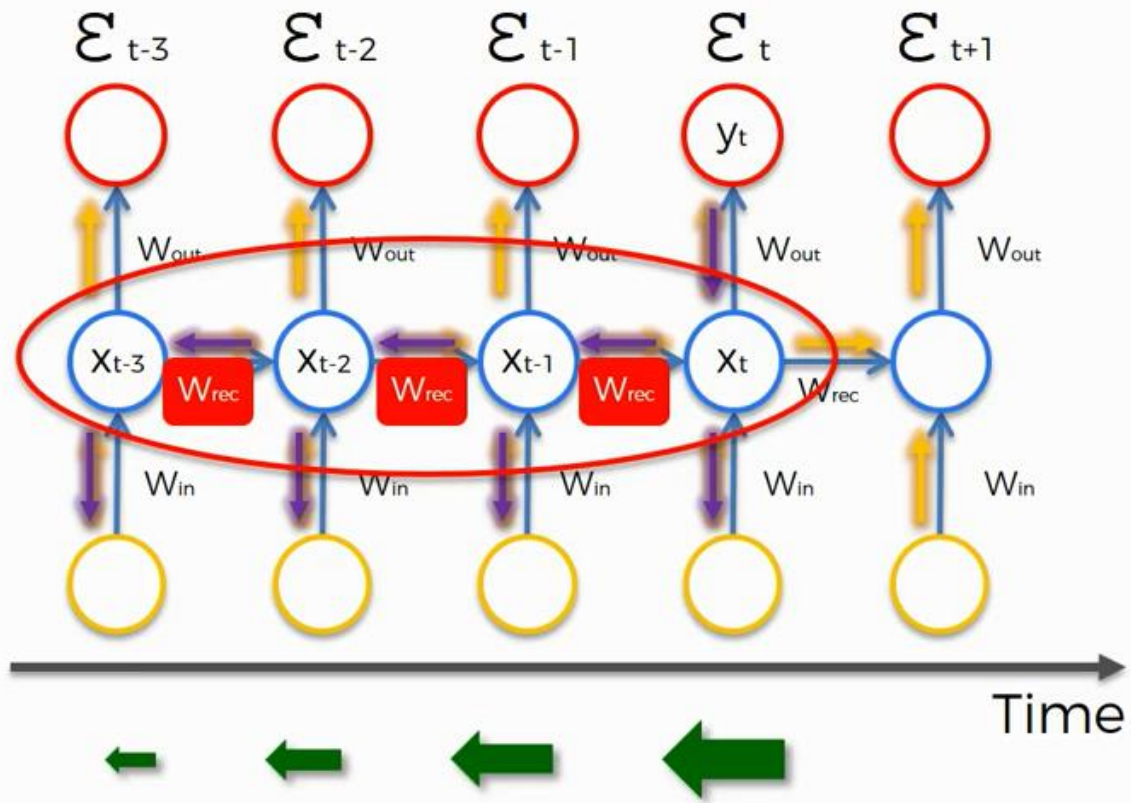
W_{rec} ~ large ⇨ Exploding

Formula Source: Razvan Pascanu et al. (2013)

Solutions:

1. Exploding Gradient
    - Truncated Backpropagation
    - Penalties
    - Gradient Clipping

2. Vanishing Gradient
    - Weight Initialization
    - Echo State Networks
    - Long Short-Term Memory Networks (LSTMs) 

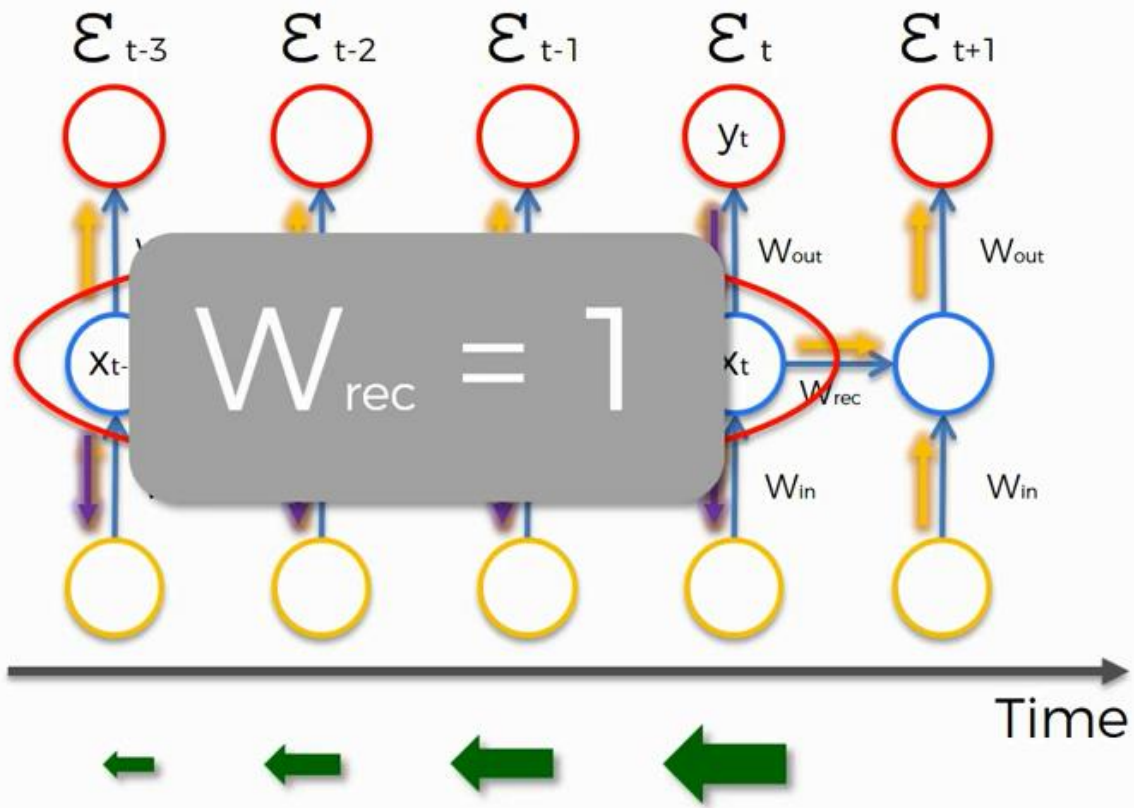


$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$

W_{rec}	< 1	⇨	Vanishing
W_{rec}	> 1	⇨	Exploding

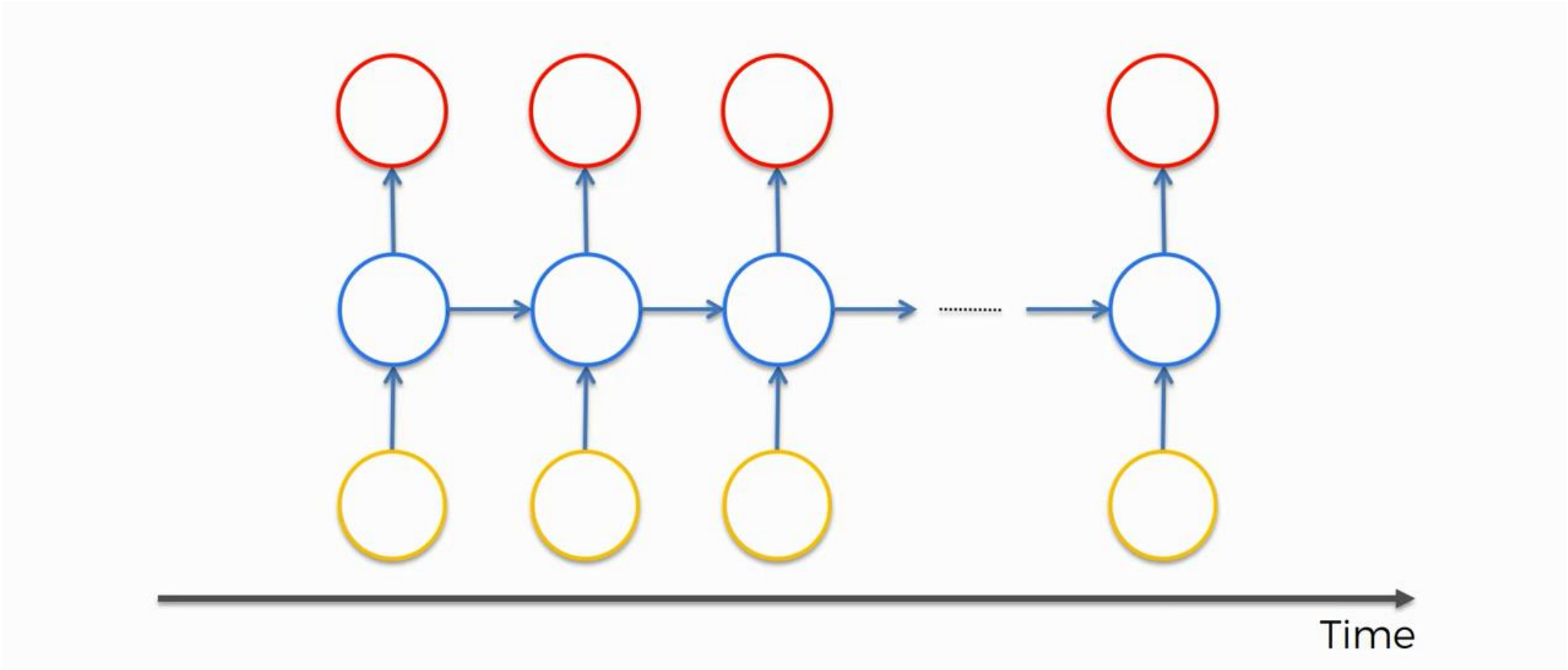


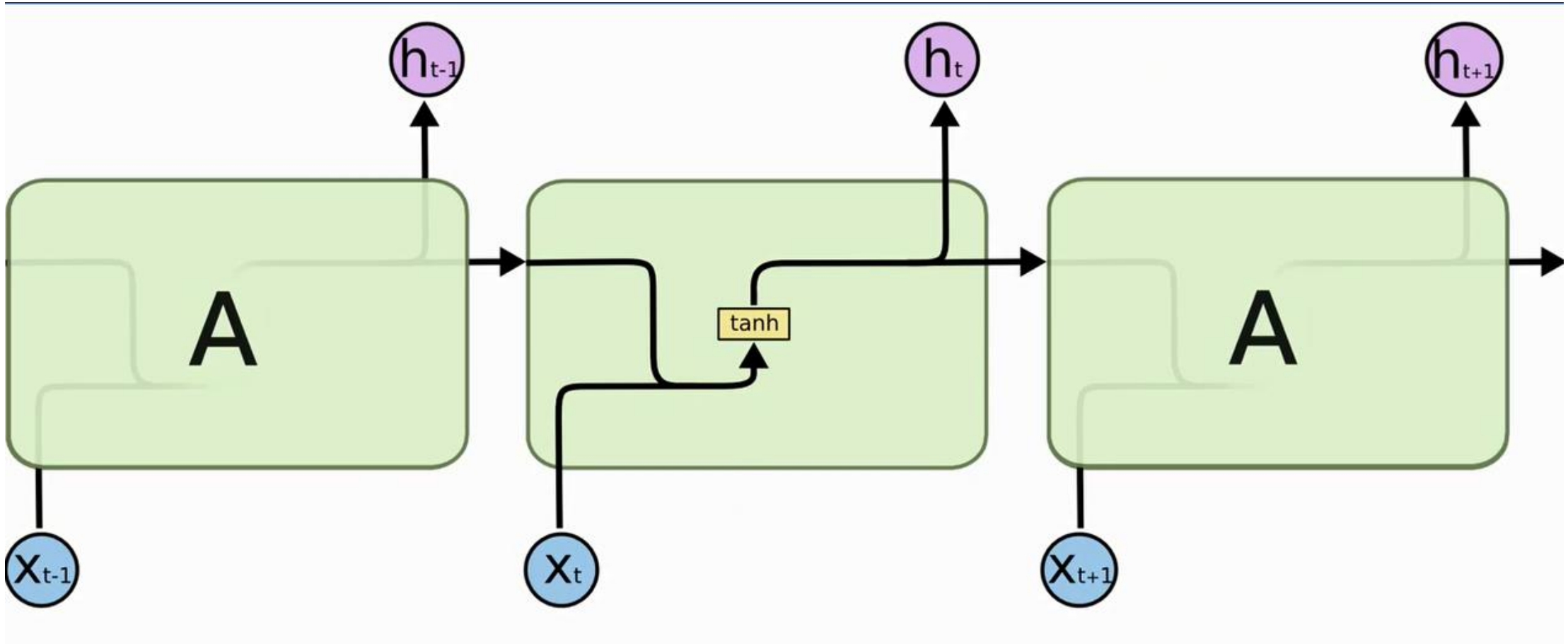
$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \le t \le T} \frac{\partial \mathcal{E}_t}{\partial \theta} \tag{3}$$

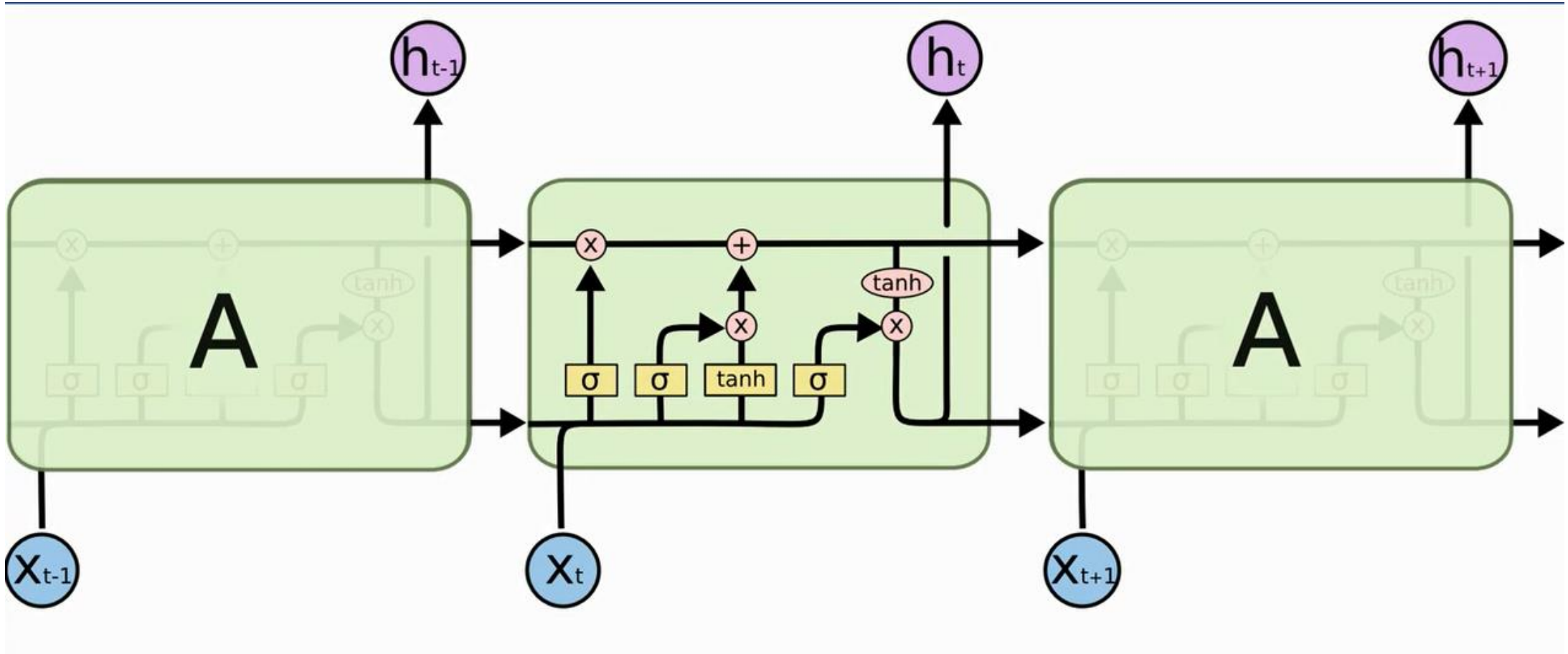
$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \le k \le t} \left(\frac{\partial \mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right) \tag{4}$$

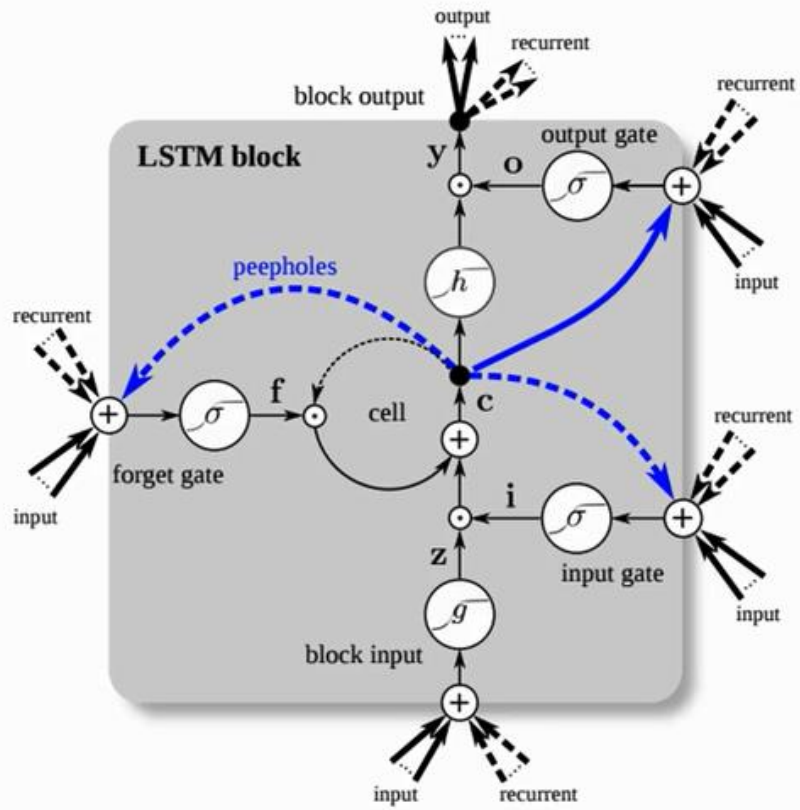
$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \ge i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \ge i > k} \mathbf{W}_{rec}^T diag(\sigma'(\mathbf{x}_{i-1})) \tag{5}$$

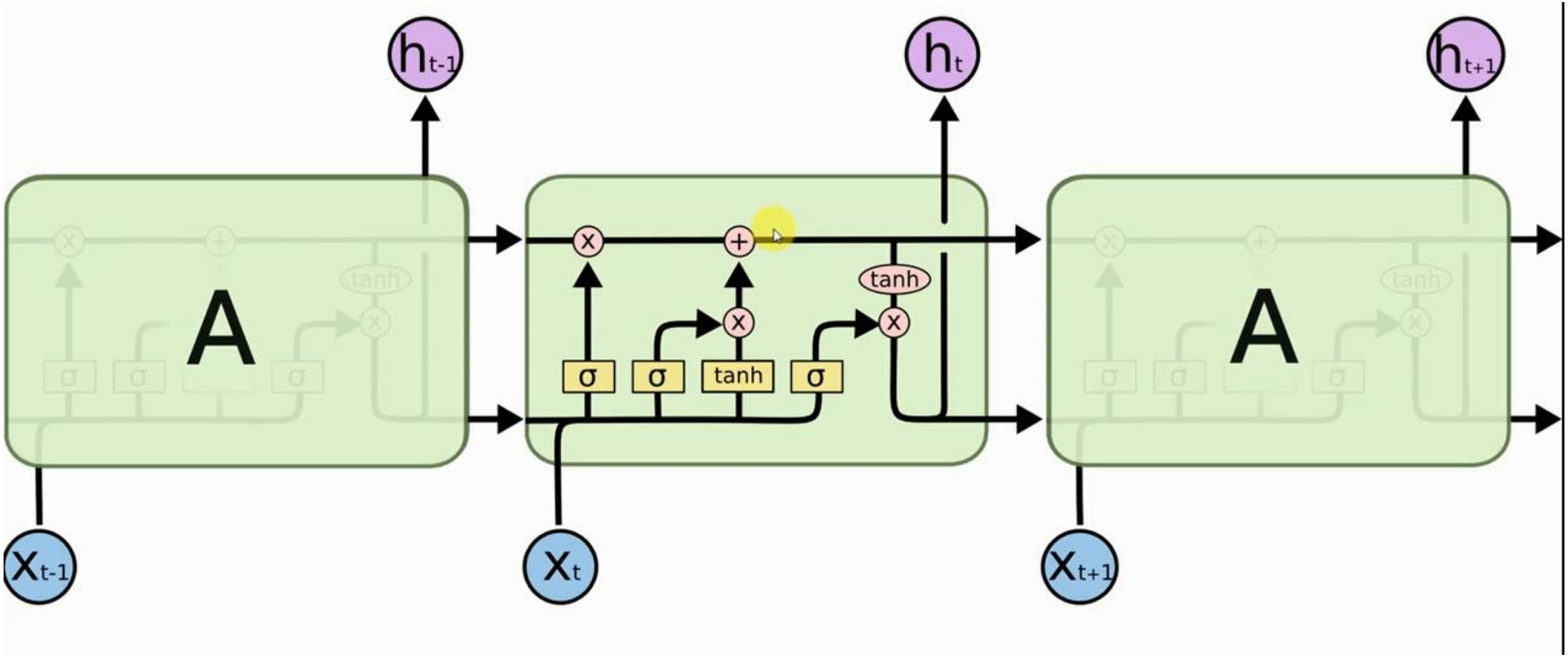
W_{rec}	< 1	⇨	Vanishing
W_{rec}	> 1	⇨	Exploding


Time


h_{t-1}
h_t
h_{t+1}
A
tanh
A
X_{t-1}
X_t
X_{t+1}


h_{t-1}
h_t
h_{t+1}
A
A
tanh
σ
σ
tanh
σ
X_{t-1}
X_t
X_{t+1}


output
recurrent
block output
recurrent
LSTM block
y
output gate
o
peepholes
h
input
recurrent
recurrent
f
cell
c
forget gate
recurrent
input
i
input gate
z
input
g
block input
input
recurrent


h_{t-1}
h_t
h_{t+1}
A
A
tanh
σ
σ
tanh
σ
X_{t-1}
X_t
X_{t+1}

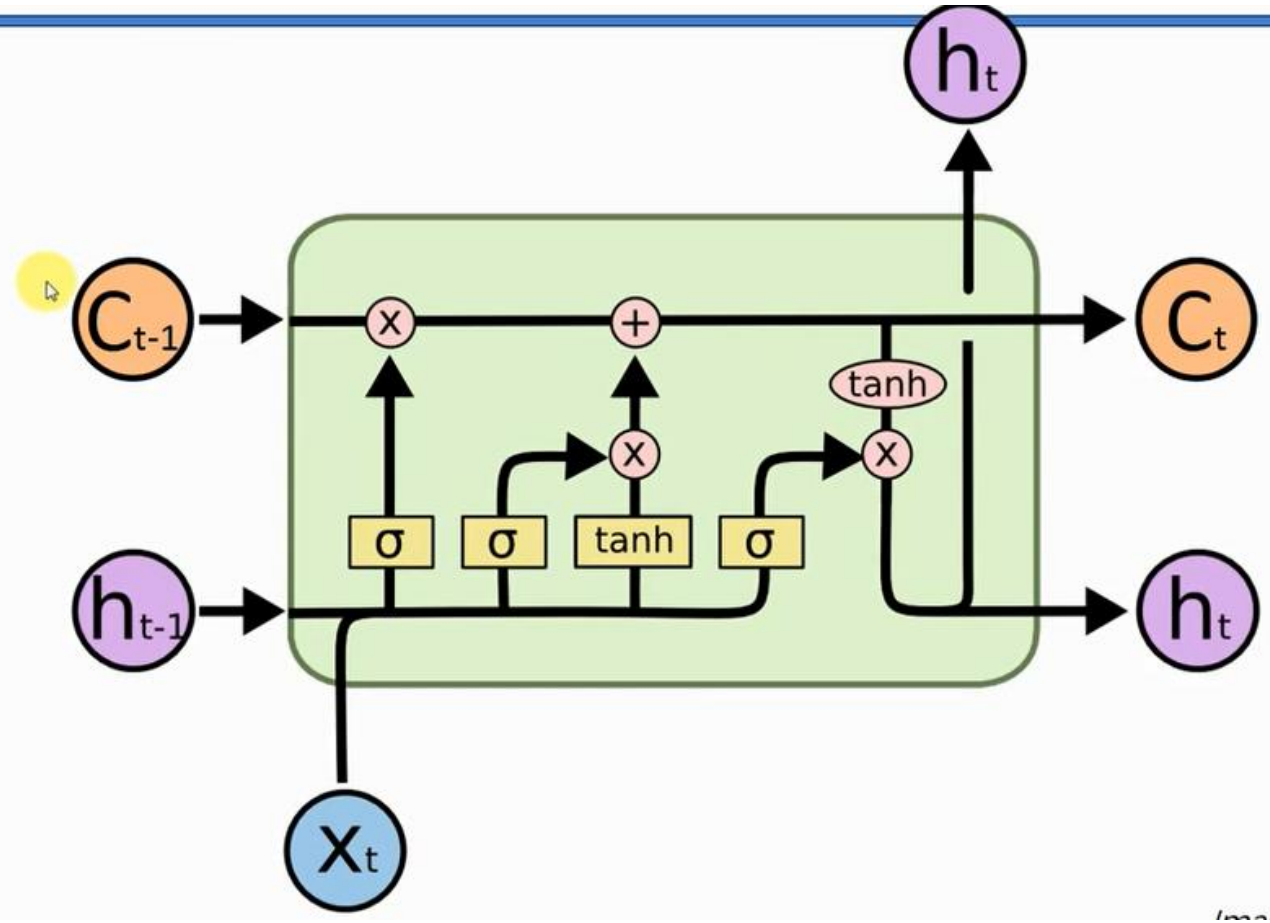
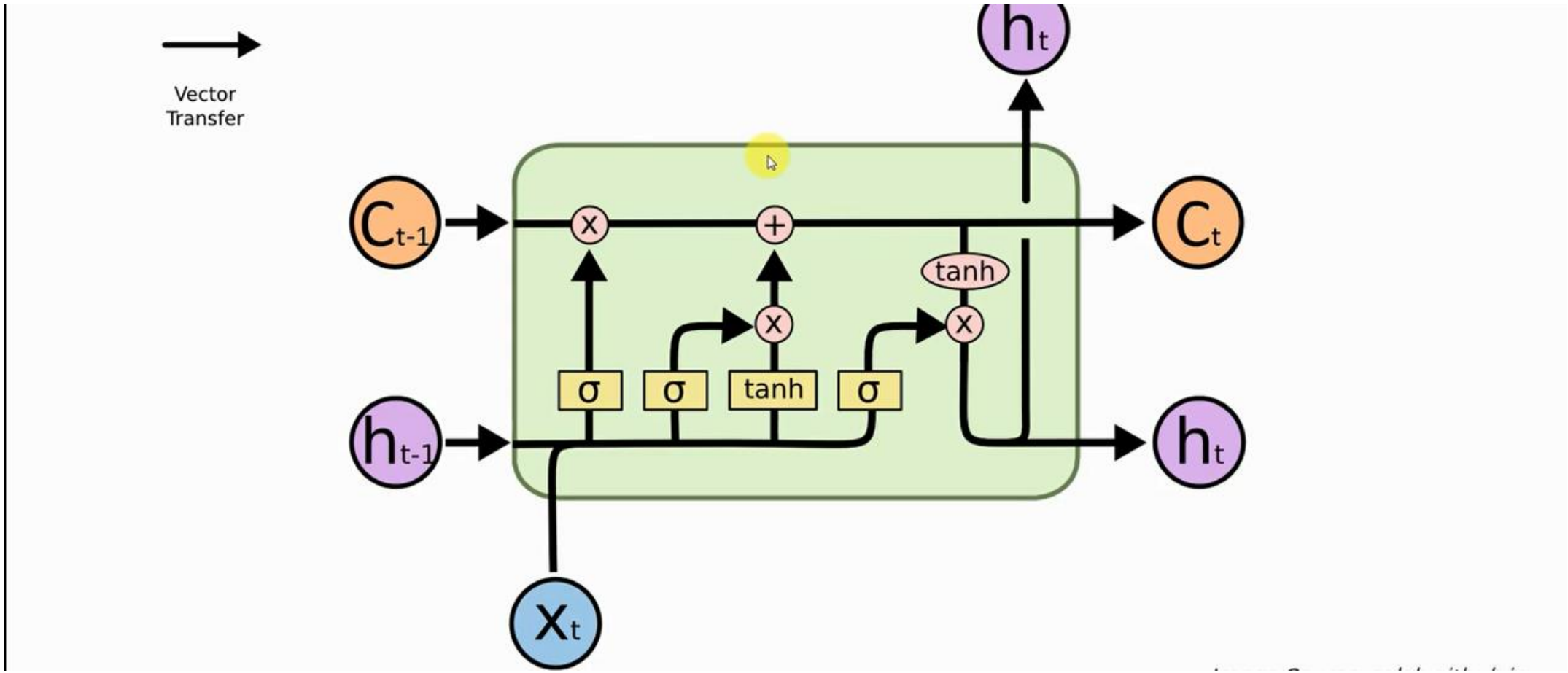


Image Source: colah.github.io


Vector
Transfer
h_t
C_{t-1}
C_t
tanh
σ
σ
tanh
σ
h_{t-1}
h_t
X_t

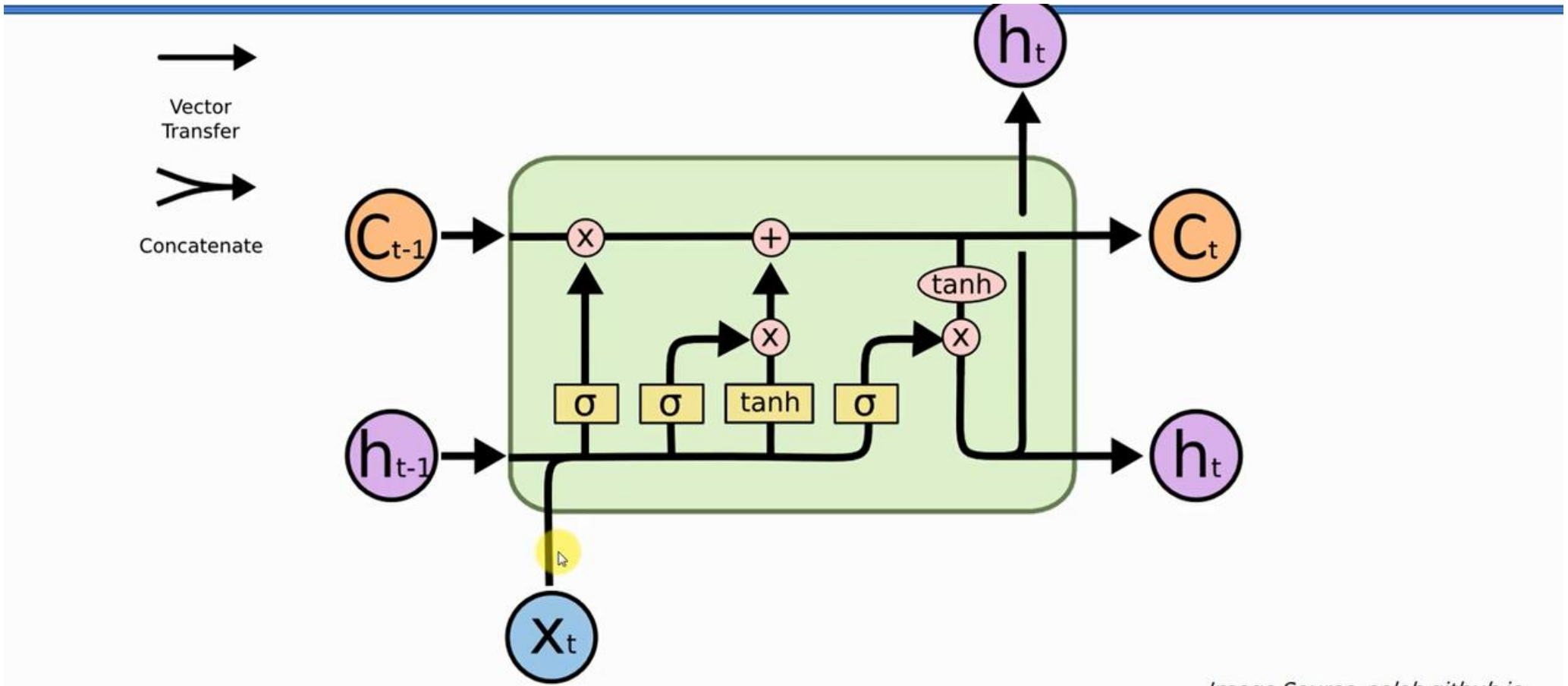


Image Source: colah.github.io

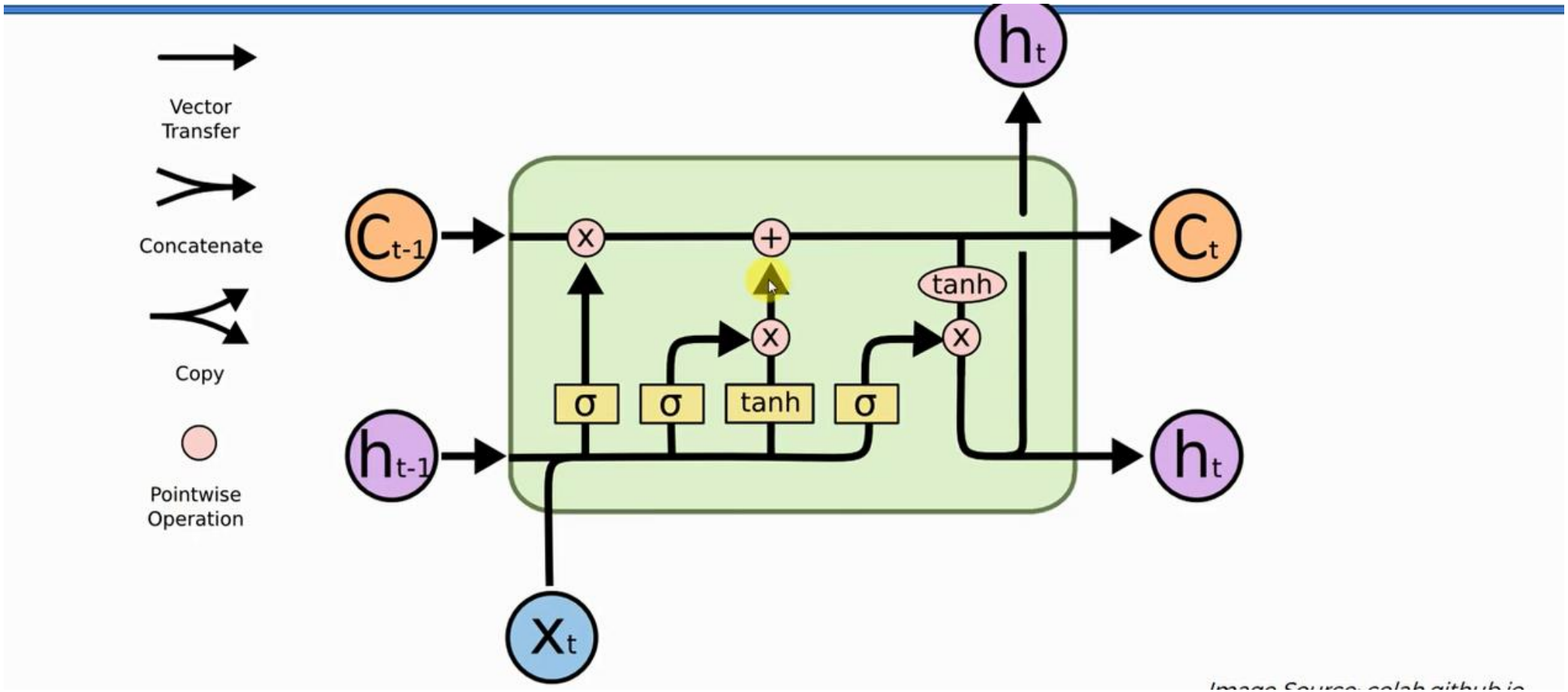


Image Source: colah.github.io


KLINGER

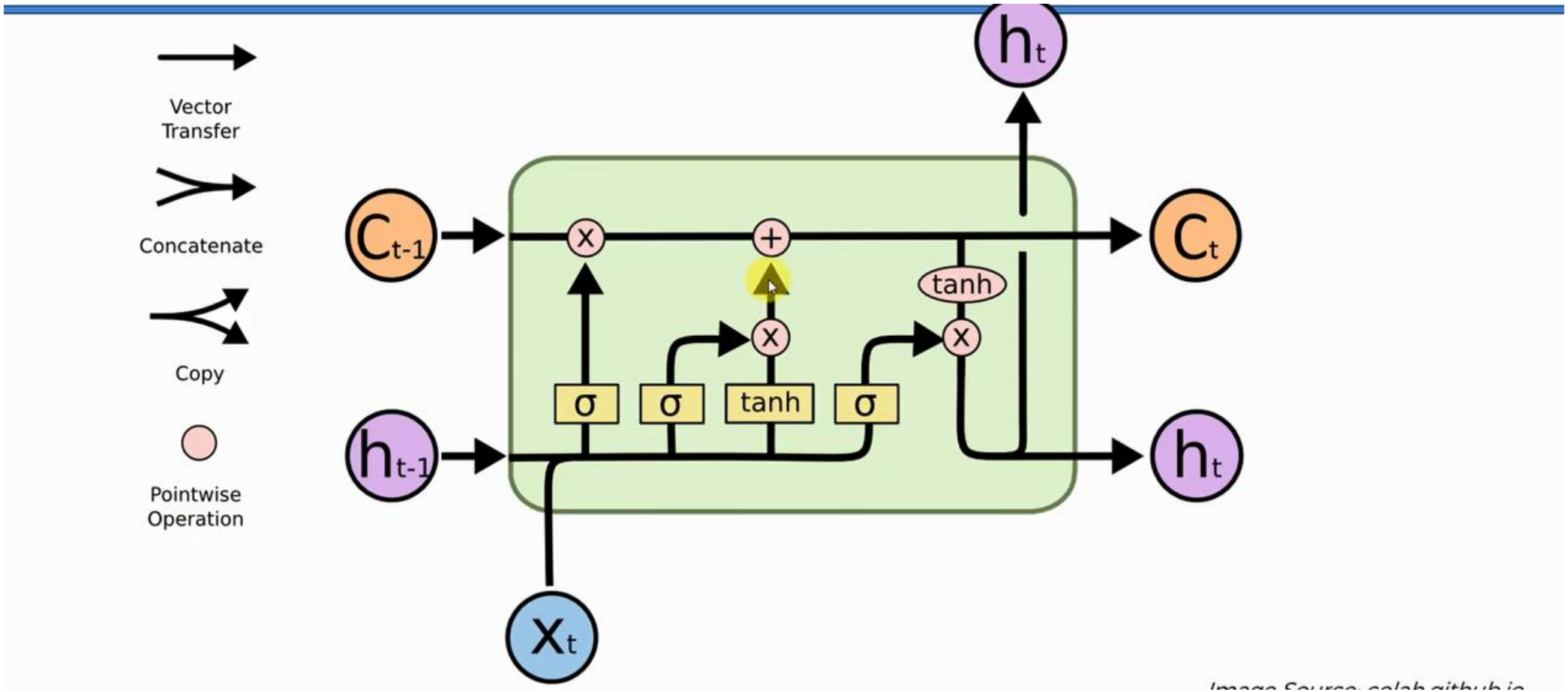


Image Source: colah.github.io

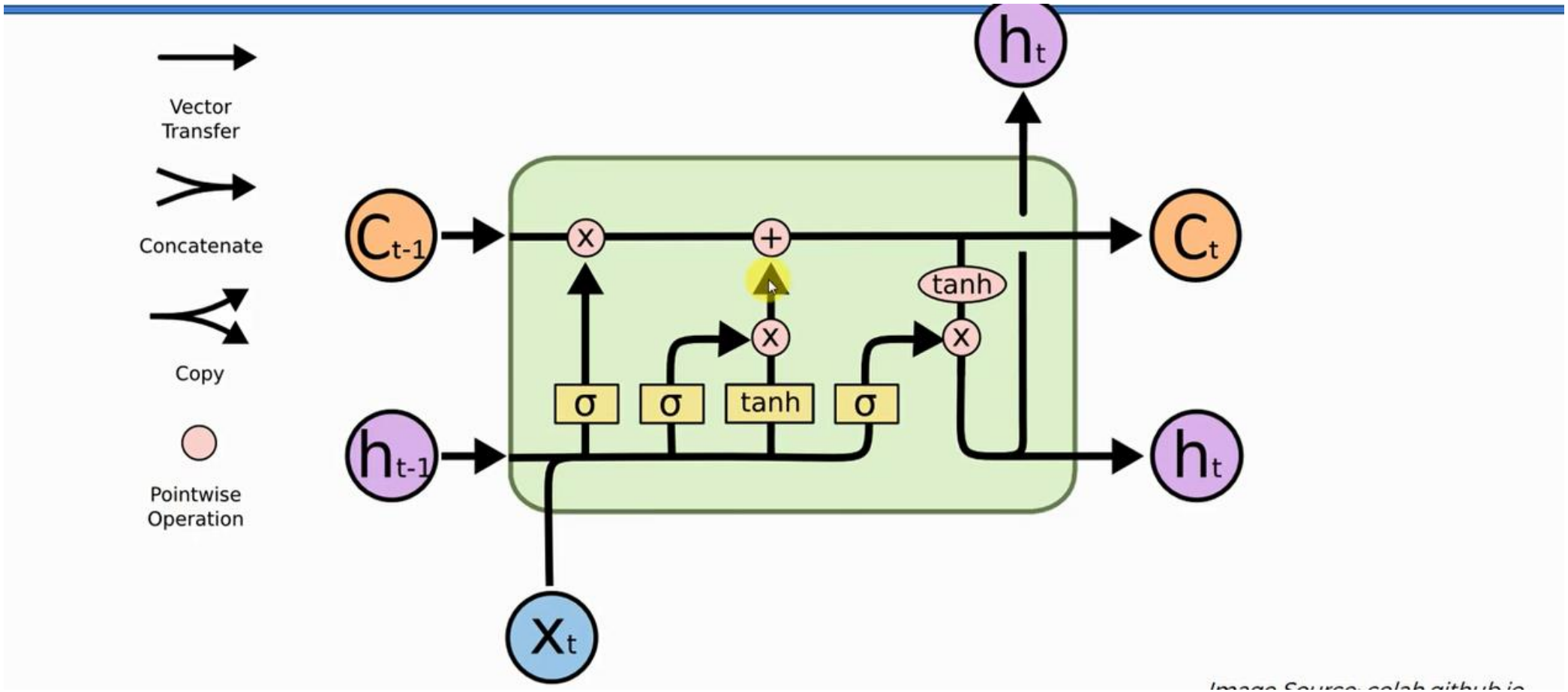


Image Source: colah.github.io

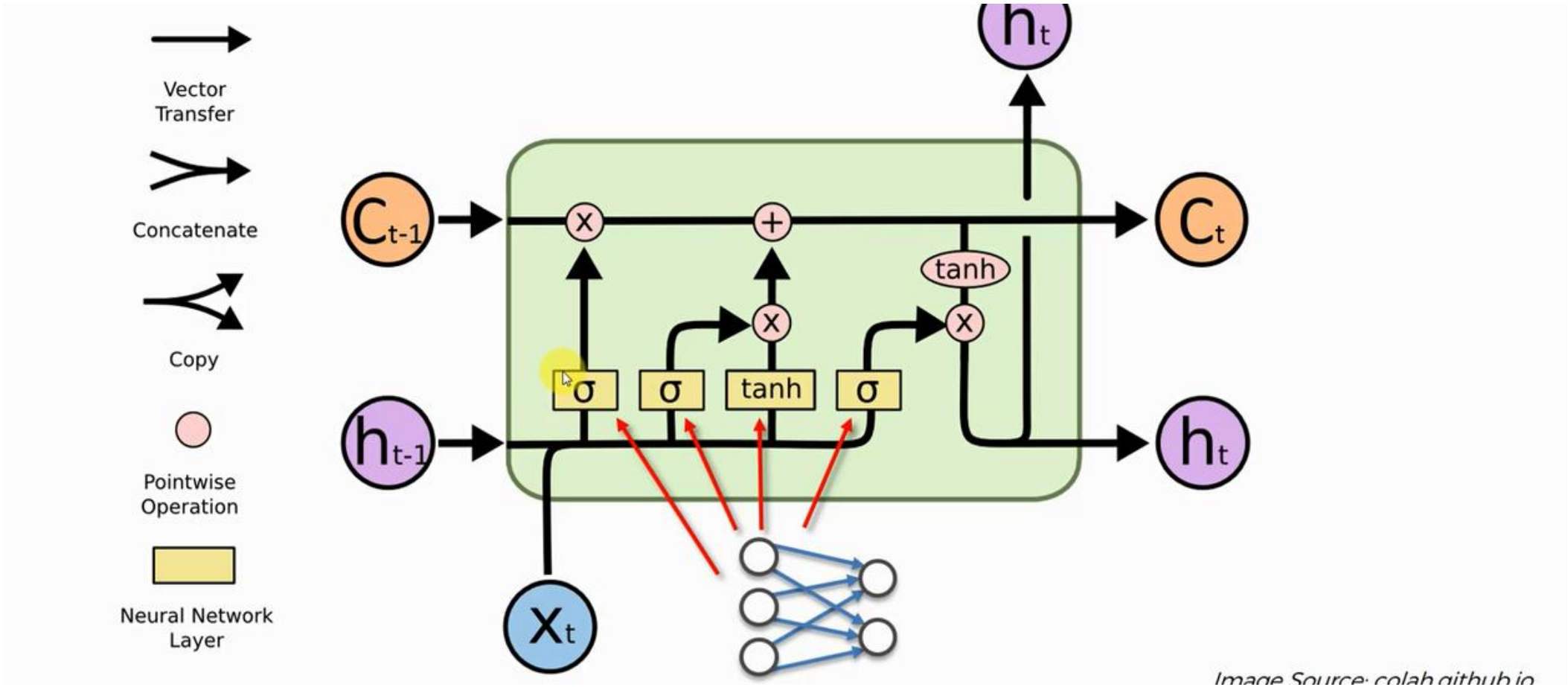


Image Source: colah.github.io

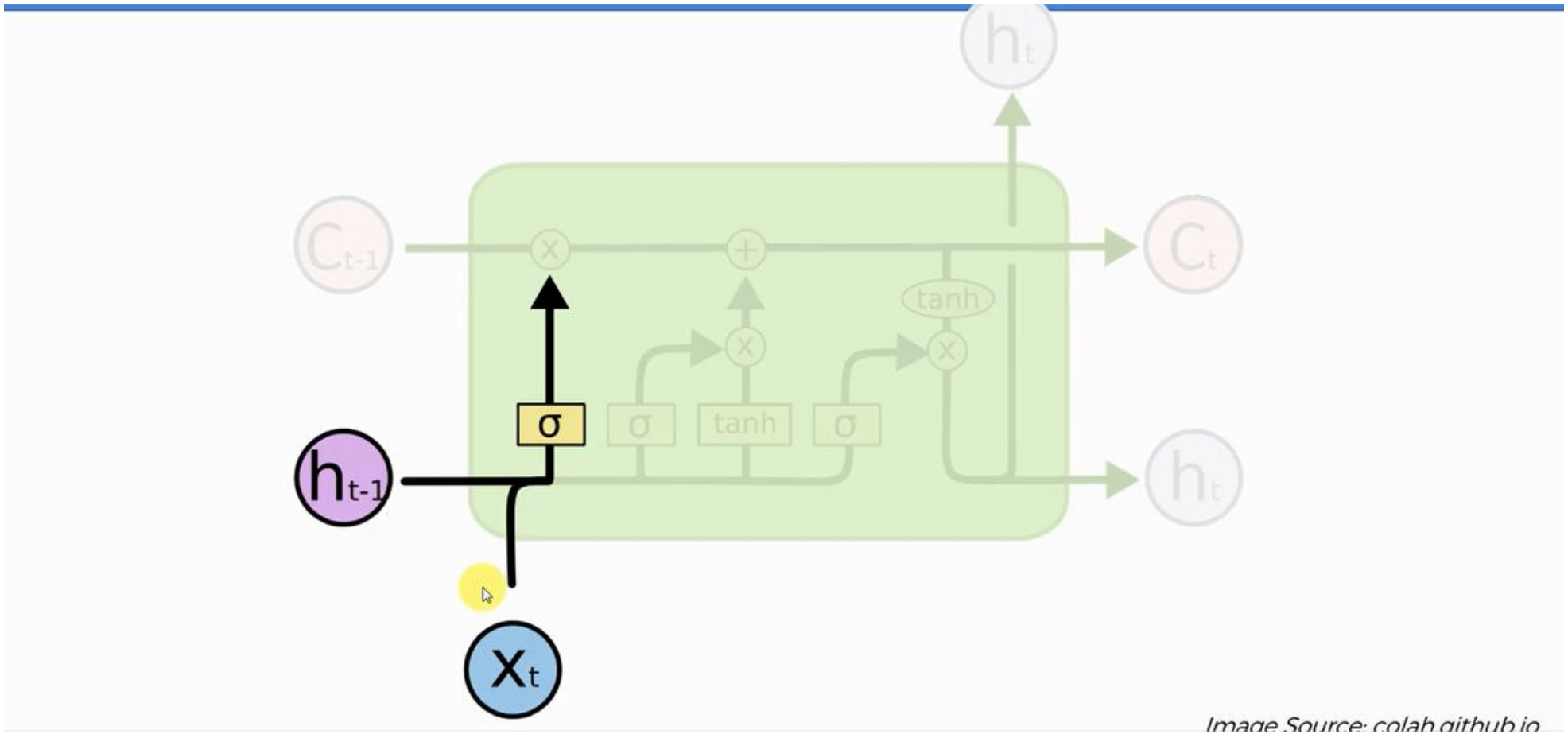


Image Source: colah.github.io

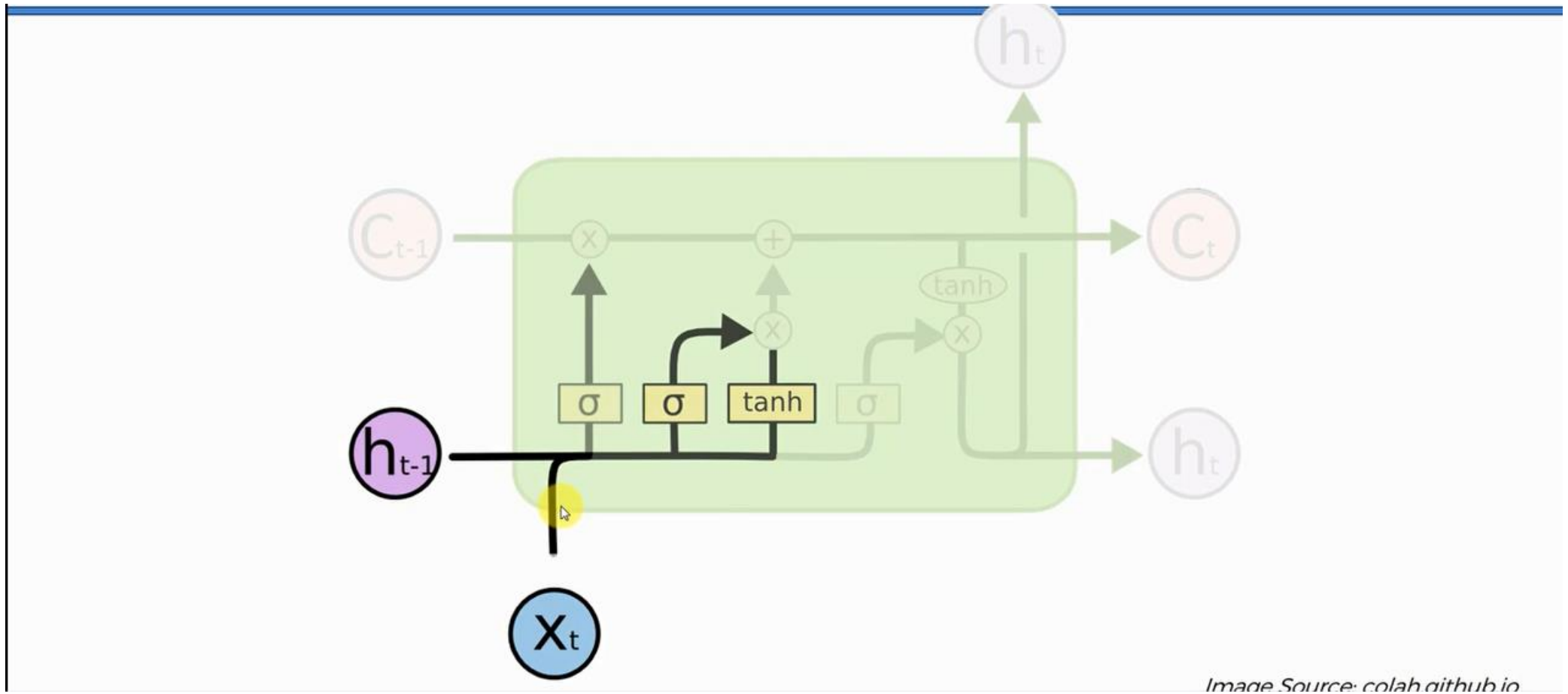


Image Source: colah.github.io

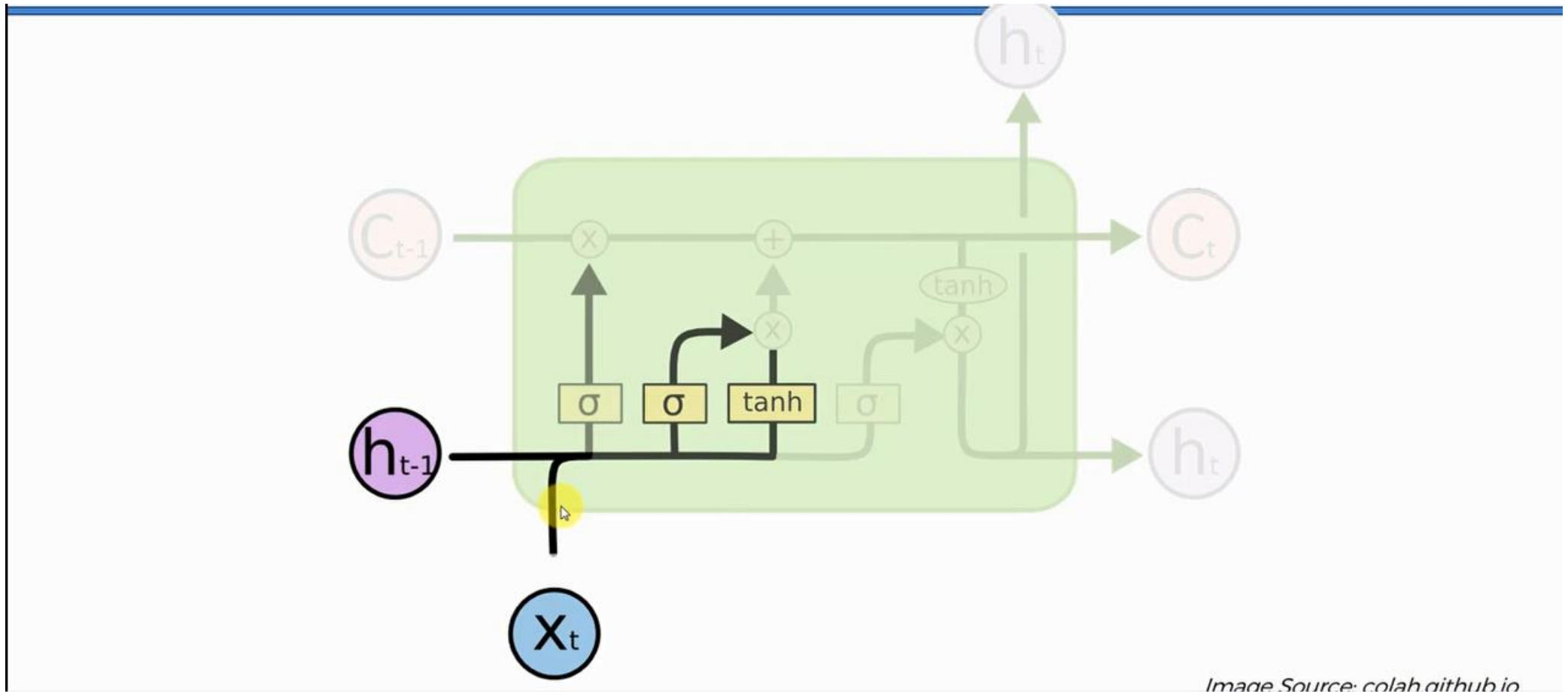


Image Source: colah.github.io

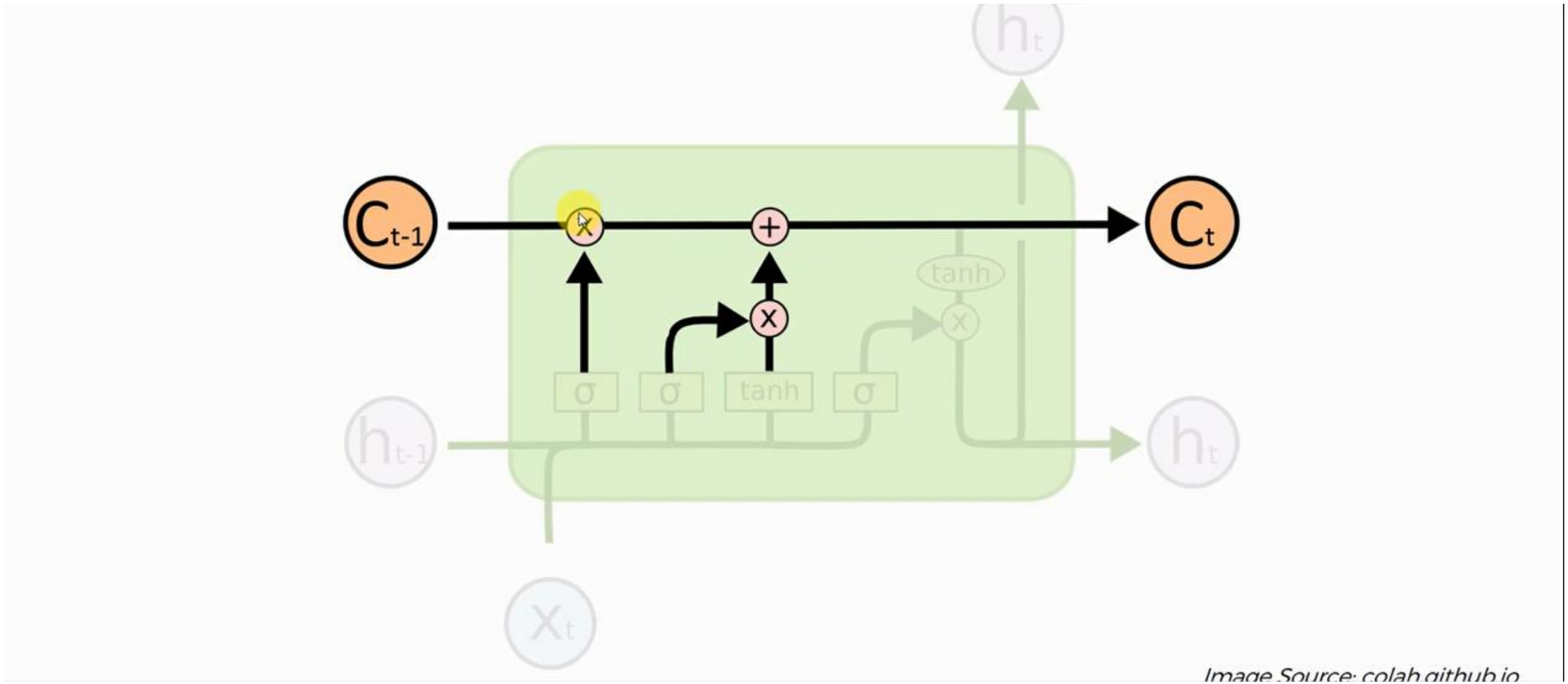


Image Source: colah.github.io

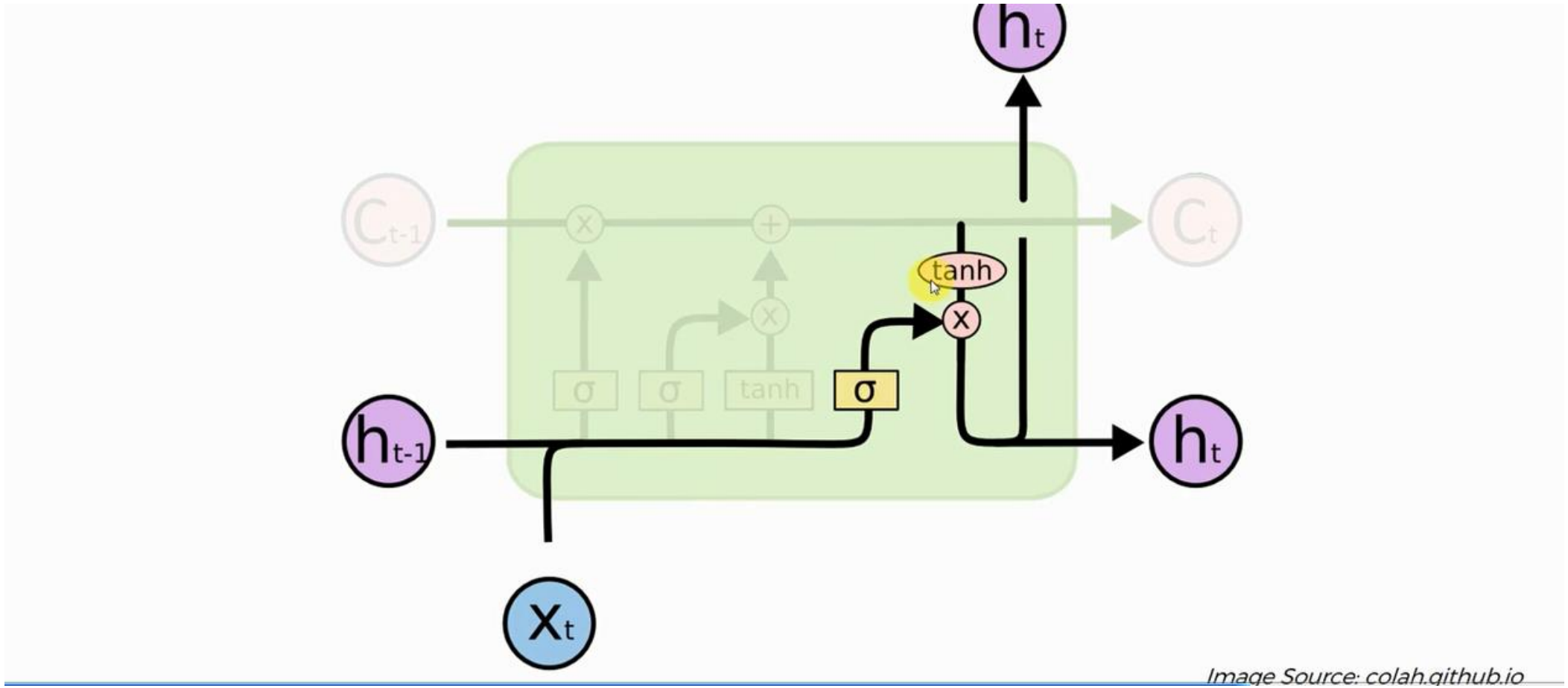
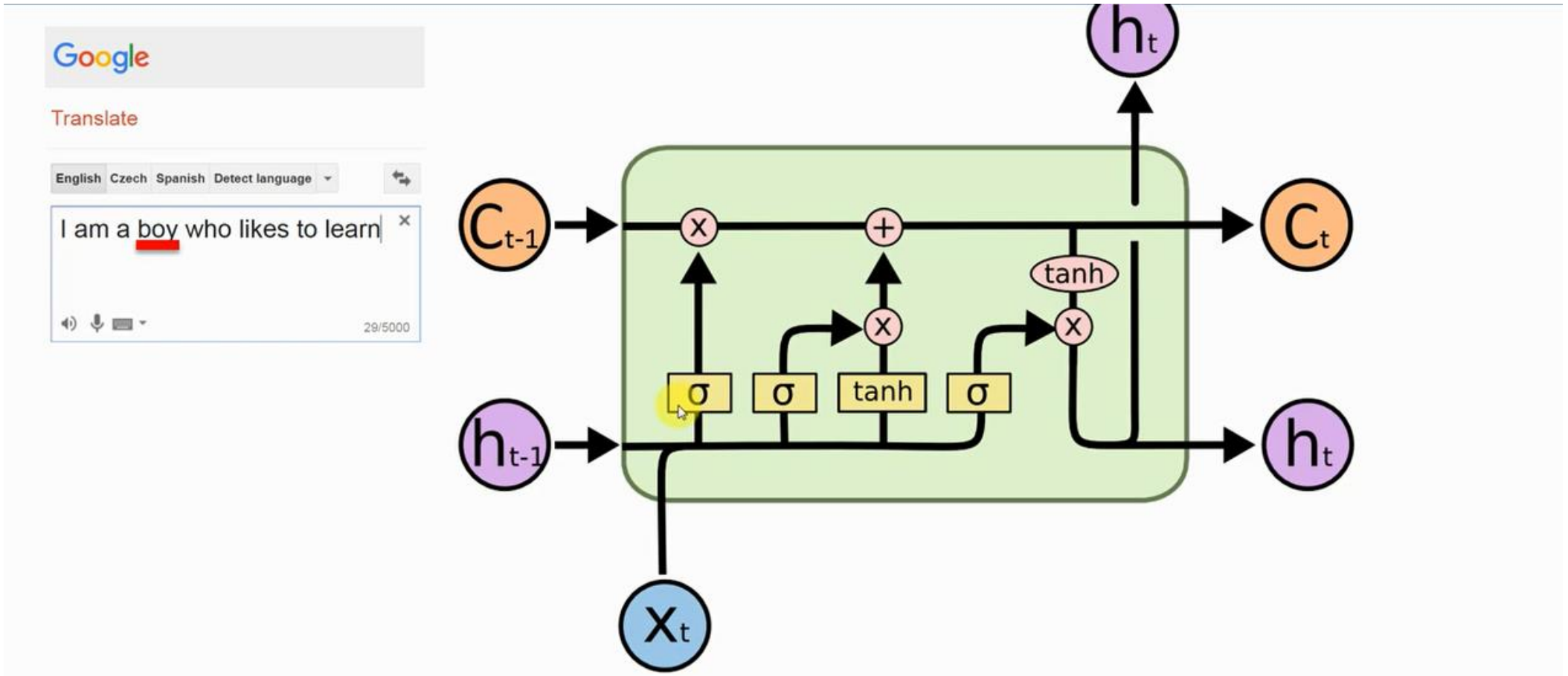
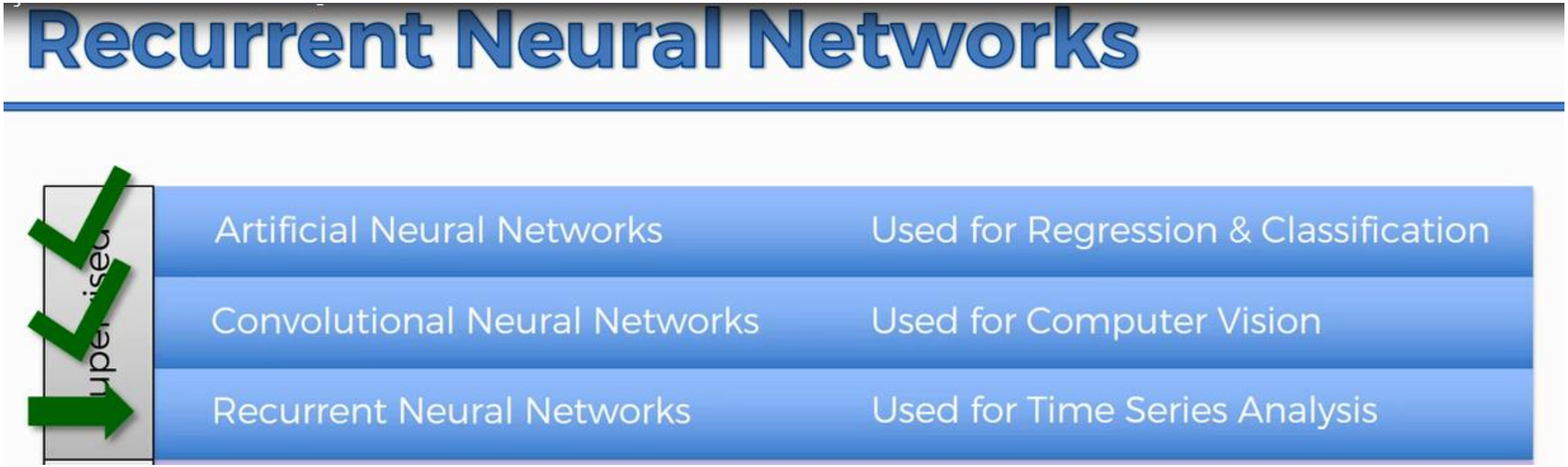
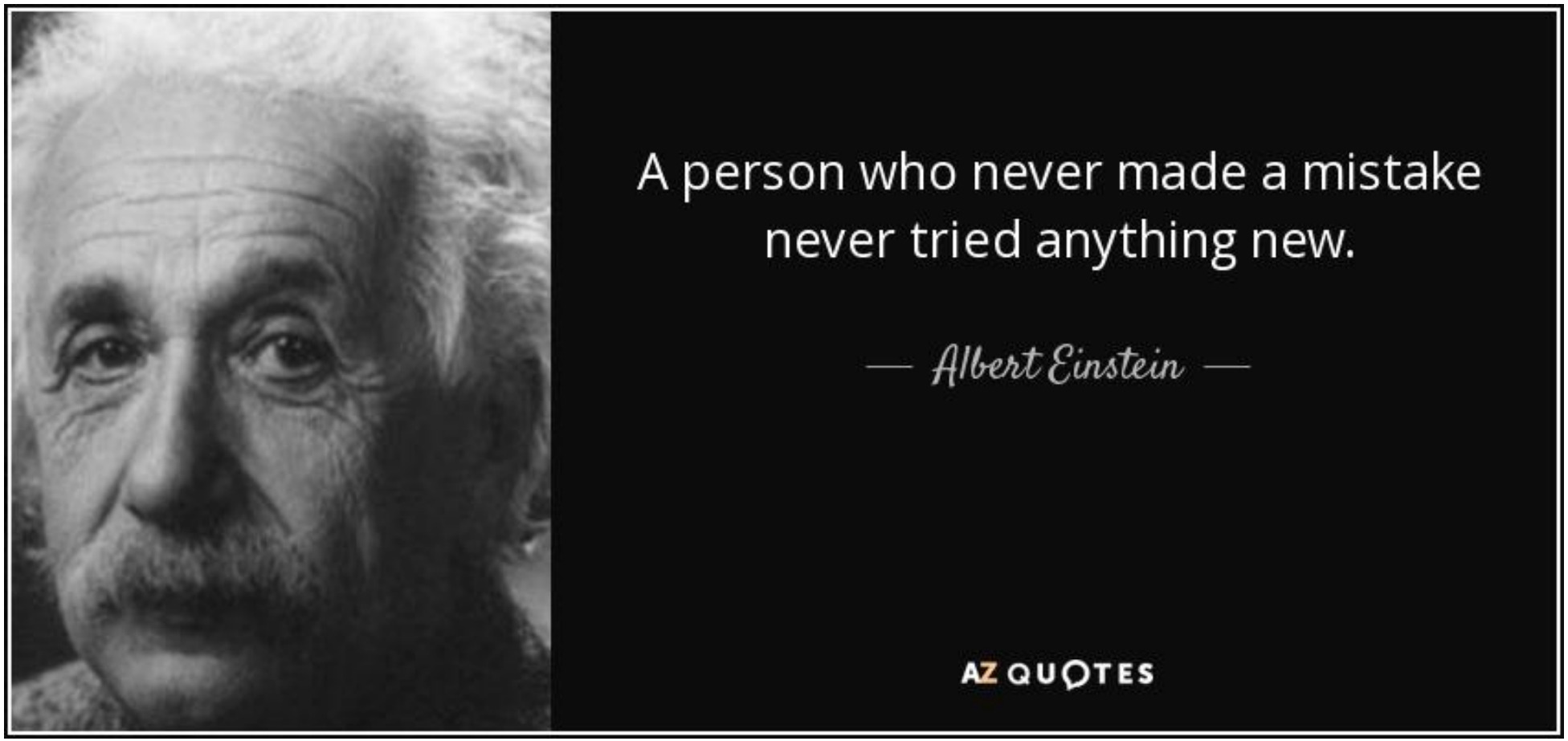


Image Source: colah.github.io


Google
Translate
English Czech Spanish Detect language
I am a boy who likes to learn
29/5000
h_t
C_{t-1}
C_t
tanh
σ
σ
tanh
σ
h_{t-1}
h_t
X_t

Summary


Recurrent Neural Networks
Artificial Neural Networks
Used for Regression & Classification
Convolutional Neural Networks
Used for Computer Vision
Recurrent Neural Networks
Used for Time Series Analysis


A person who never made a mistake never tried anything new.
— Albert Einstein —
AZ QUOTES

Thank You !!!