Recurrent Neural Network (RNN)

Tuan Nguyen - AI4E

Outline

- Motivation for RNN
- Introduction to RNN
- The structure RNN
- Deep RNN
- RNN application

Image classification

convolution + nonlinearity

max pooling

convolution + pooling layers

vec

fully connected layers

bird p_{bird}

sunset p_{sunset}

dog p_{dog}

cat p_{cat}

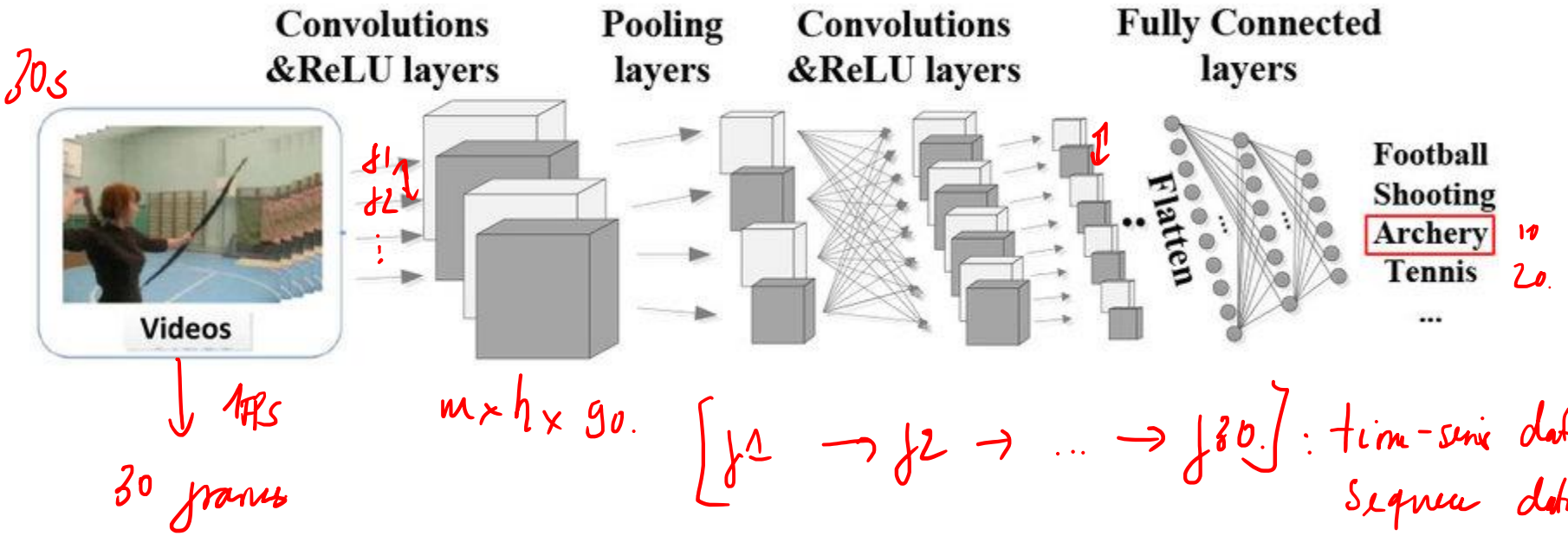
...

Nx binary classification

Video classification

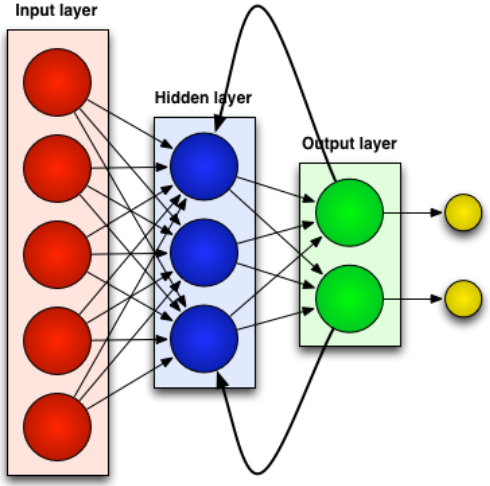
tôi yêu Việt Nam

https://www.youtube.com/watch?v=BmoVg3BCwF0&feature=emb_title

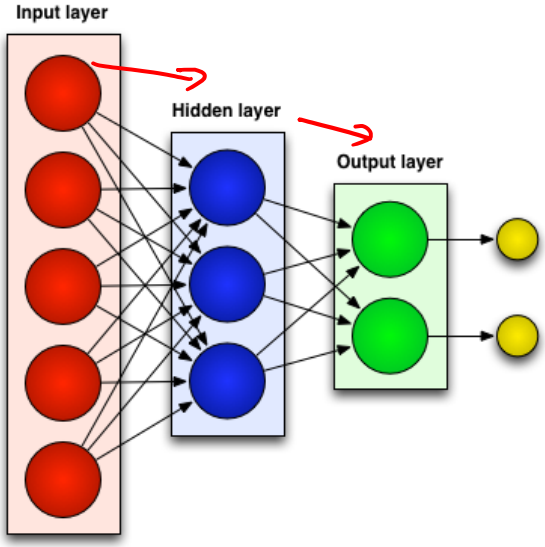


Neural network

Recurrent Neural Network (RNN)

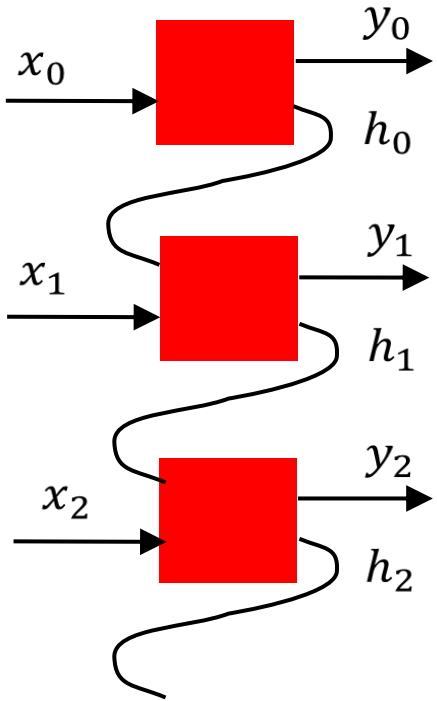


Feed forward neural network



Recurrent Neural Network

Usually drawn as:



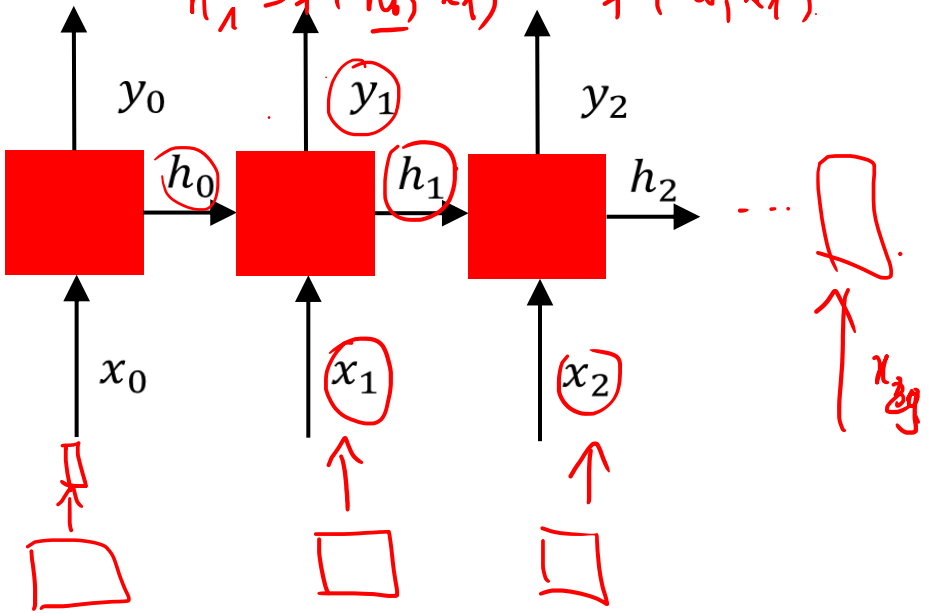
- Input can vector

$h_0 = f(x_0)$

$h_1 = f(h_0, x_1)$

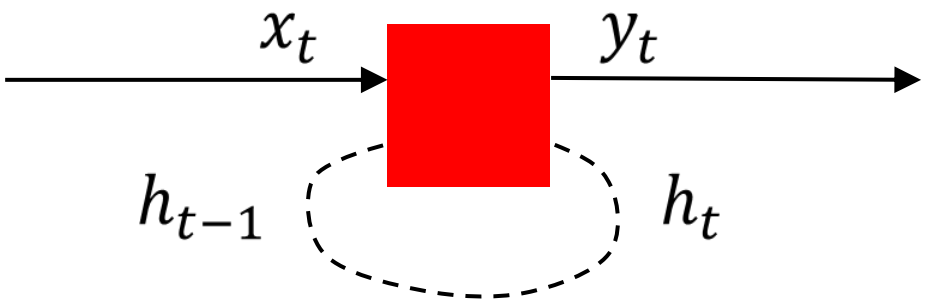
$h_2 = f(h_1, x_2)$

$h_1 = f(h_0, x_1) = f(x_0, x_1)$

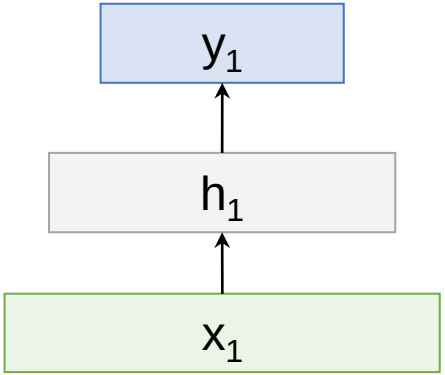


Recurrent neural network

Recurrent networks introduce cycles and a notion of time.



One-step delay



t = 1
Feed forward model

- They are designed to process sequences of data $x_1, \dots, x_n$ and can produce sequences of outputs $y_1, \dots, y_m$.

RNN Formula

We can process a sequence of vectors **x** by applying a recurrence formula at every time step:

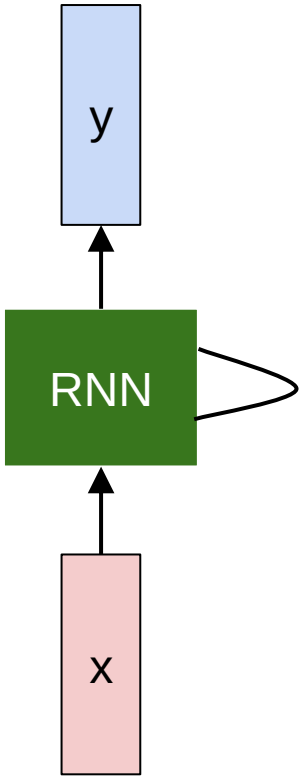
$$h_t = f_W(h_{t-1}, x_t)$$

new state

some function with parameters W

old state

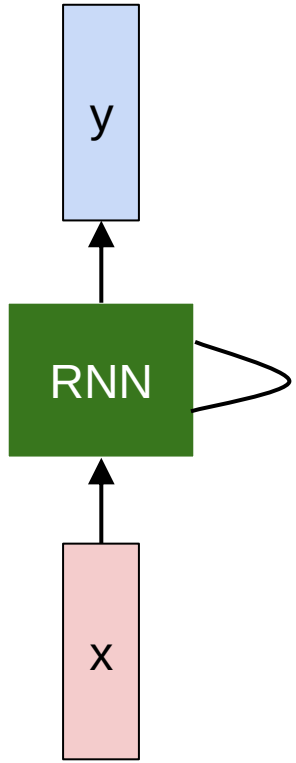
input vector at some time step



RNN Formula

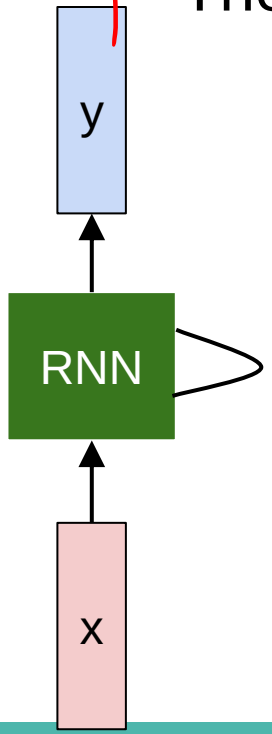
We can process a sequence of vectors **x** by applying a recurrence formula at every time step:

$$h_t = f_W(h_{t-1}, x_t)$$



Notice: the same function and the same set of parameters are used at every time step.

RNN Formula



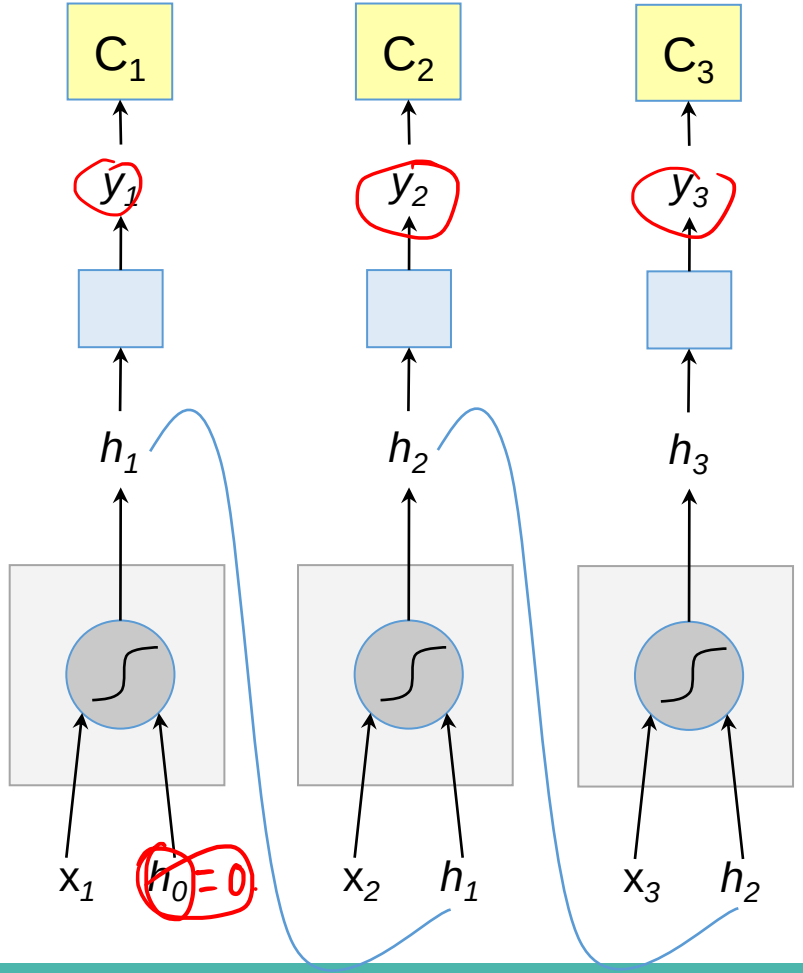
The state consists of a single *"hidden"* vector **h**:

$$h_t = f_W(h_{t-1}, x_t)$$

$$h_t = \tanh(W_{hh} h_{t-1} + W_{xh} x_t)$$

$$y_t = W_{hy} h_t$$

Forward

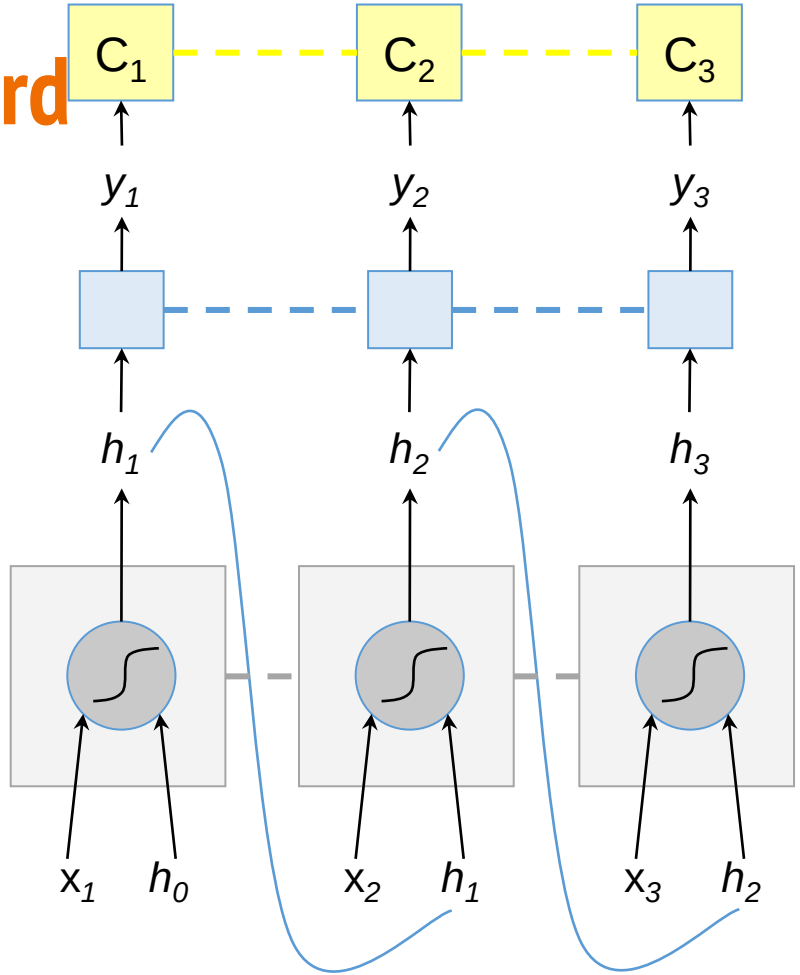


$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

Forward



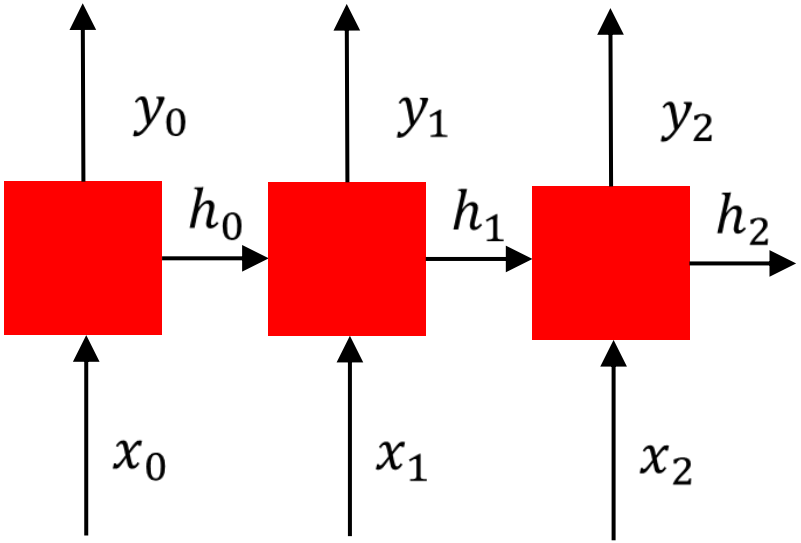
$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

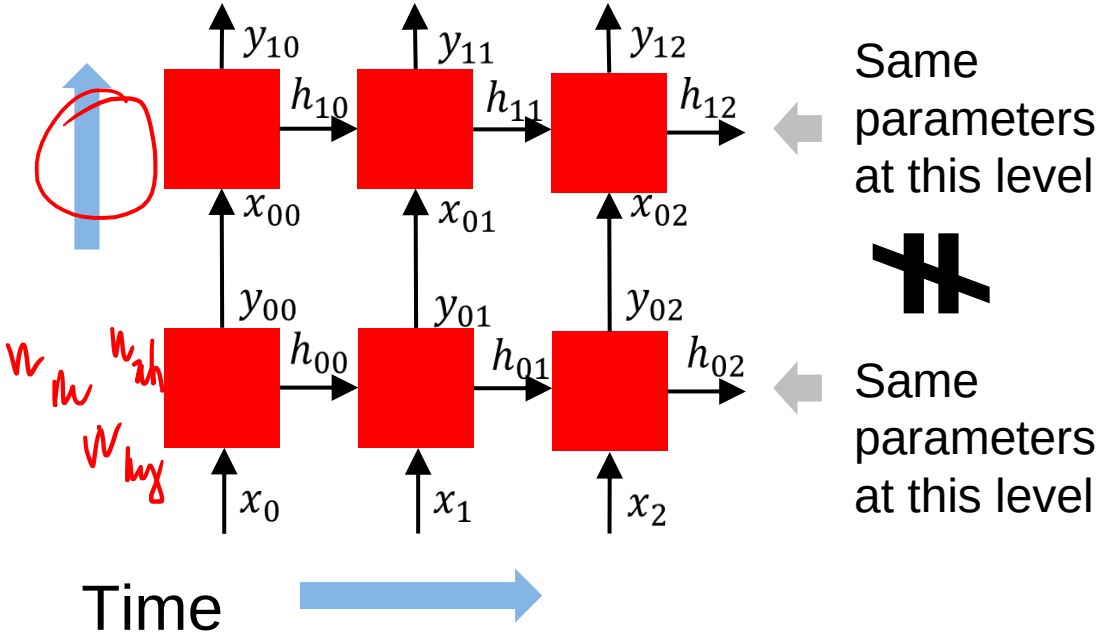
$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

- - - - indicates shared weights

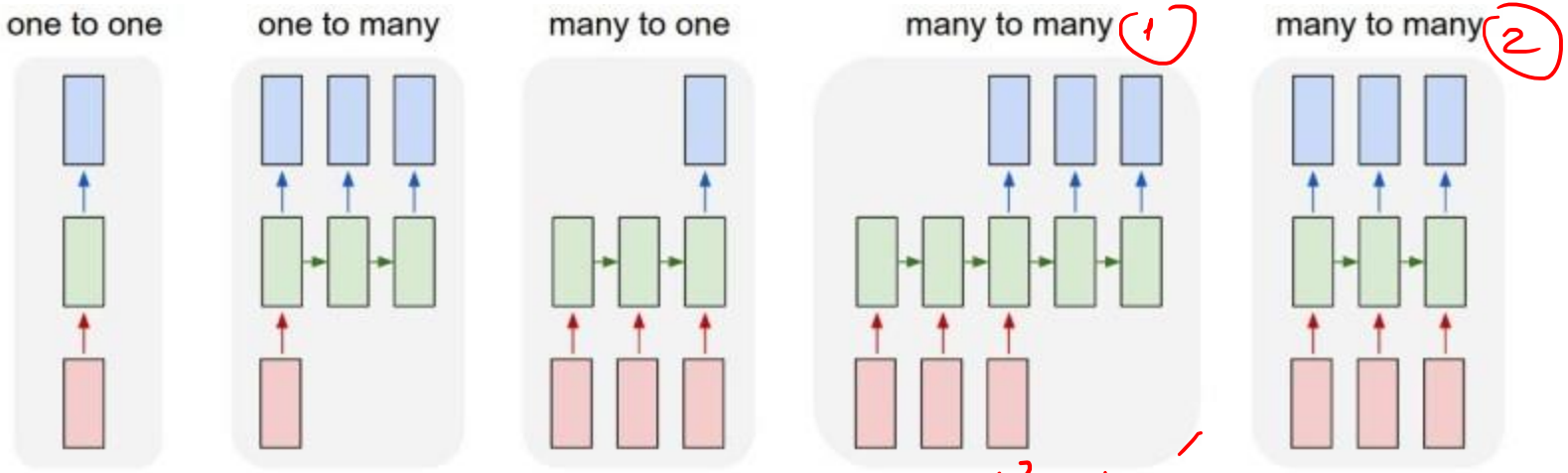
Deep RNN


y_0
y_1
y_2
h_0
h_1
h_2
x_0
x_1
x_2


y_{10}
y_{11}
y_{12}
h_{10}
h_{11}
h_{12}
x_{00}
x_{01}
x_{02}
y_{00}
y_{01}
y_{02}
h_{00}
h_{01}
h_{02}
x_0
x_1
x_2
Same parameters at this level
Same parameters at this level
Time

one: 1 ảnh, 1 từ.
many: video, câu.

Recurrent neural network problem



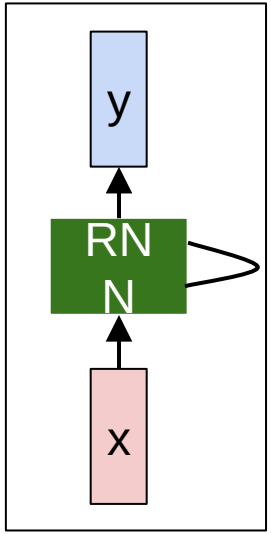
deep fake:
ảnh → video
image captioning:
ảnh → Capti

- mô tả video.
- dịch.

Character-level language model example

Vocabulary:
[h,e,l,o]

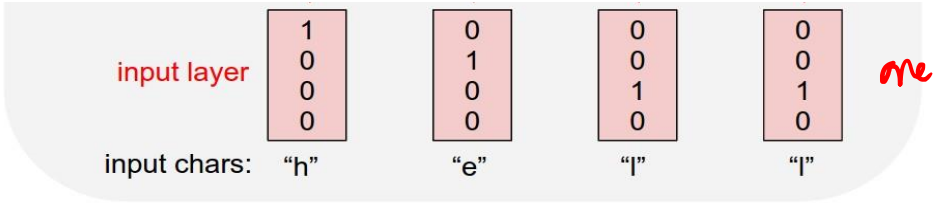
Example training sequence:
“hello”



Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”

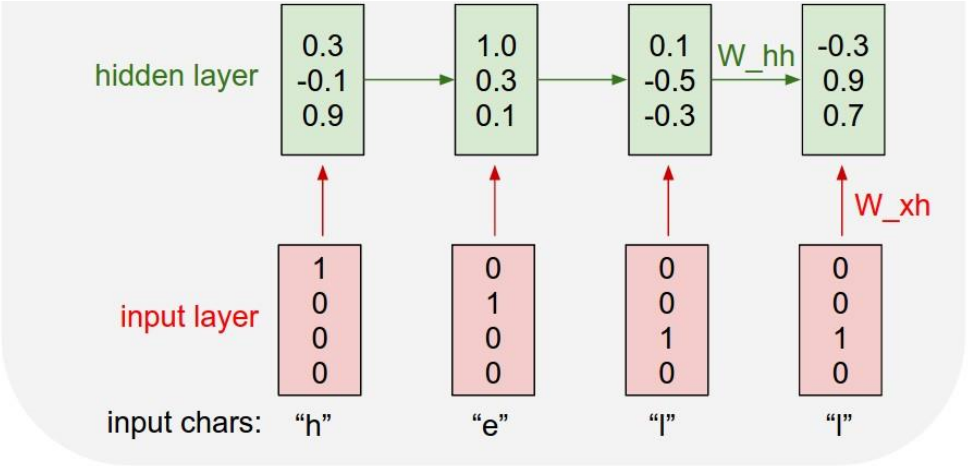


Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”

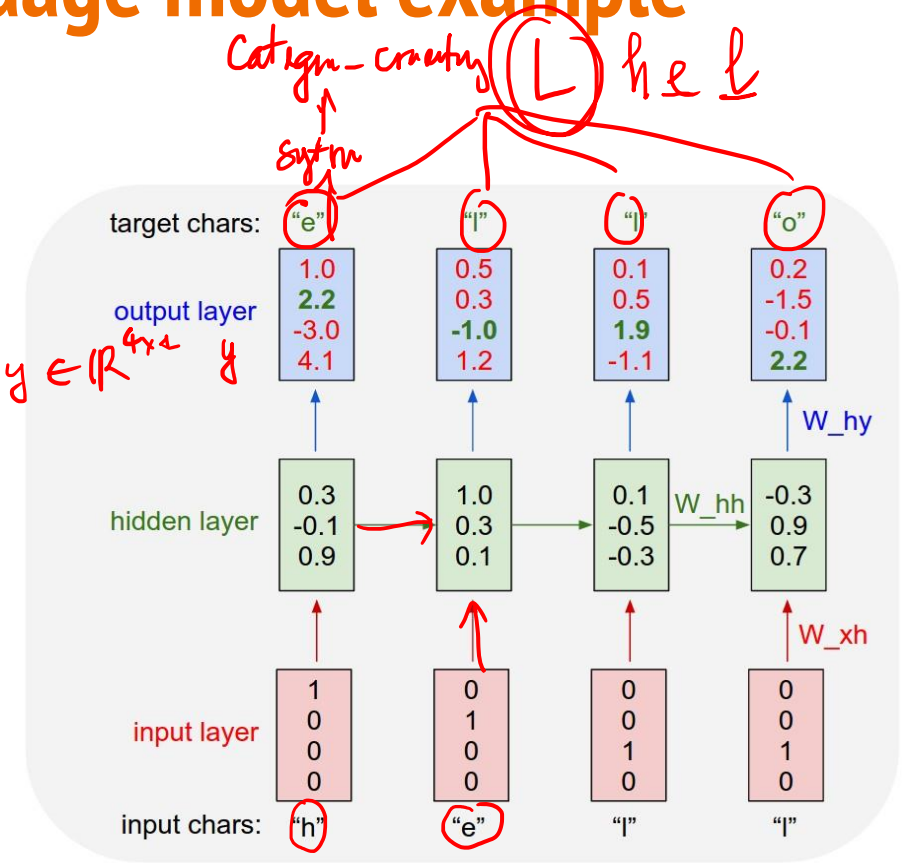
$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$



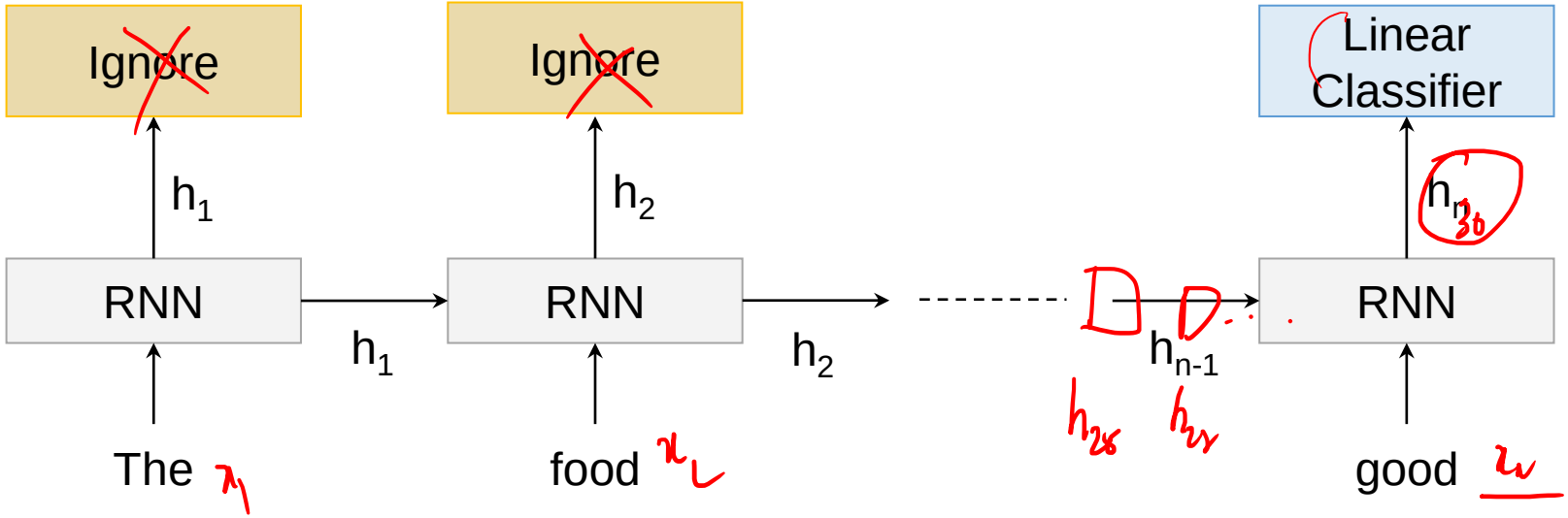
Character-level language model example

Vocabulary:
[h,e,l,o]

Example training sequence:
“hello”



Sentiment Classification



Sentiment Classification

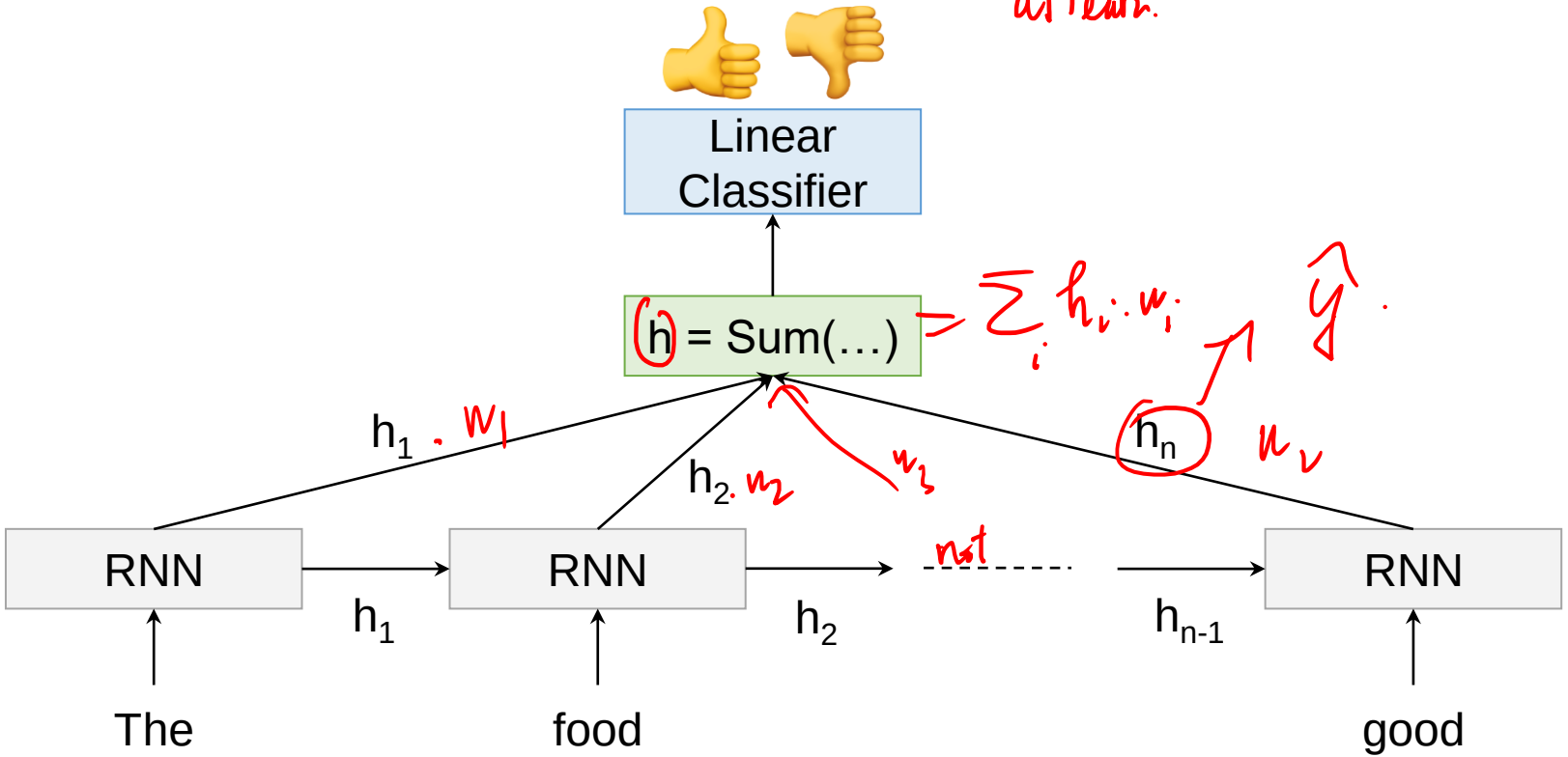


Image Captioning

- Given an image, produce a sentence describing its contents

- Inputs: Image feature (from a CNN)
- Outputs: Multiple words (let's consider one ce)

: The man with a straw hat

Image captioning

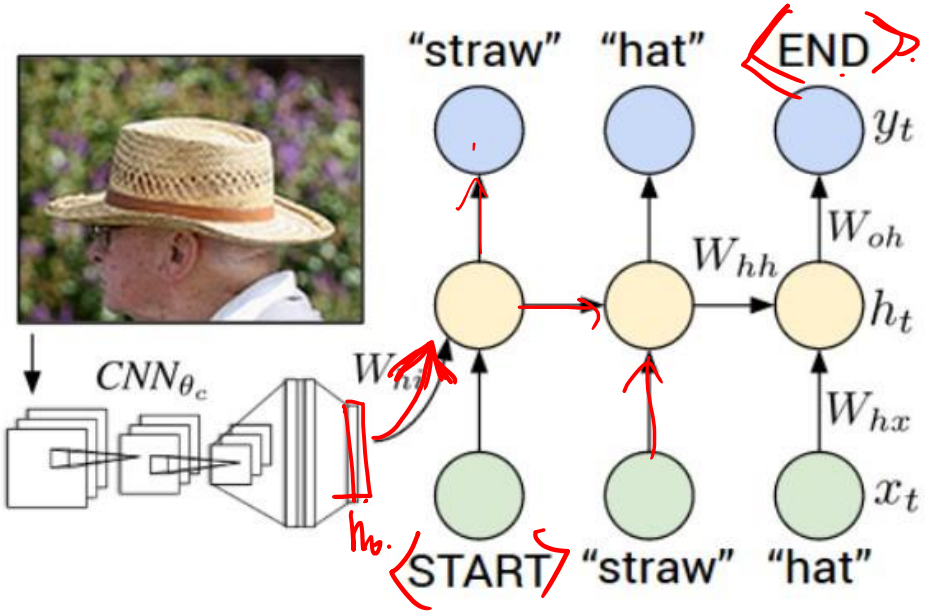
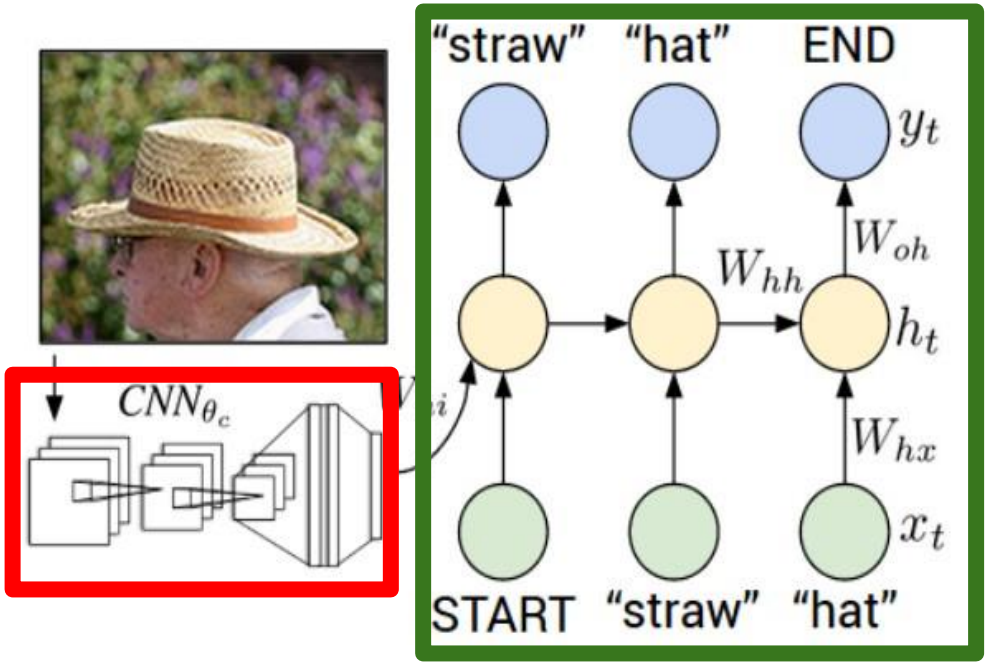


Image captioning

Recurrent Neural Network



Convolutional Neural Network

Image captioning

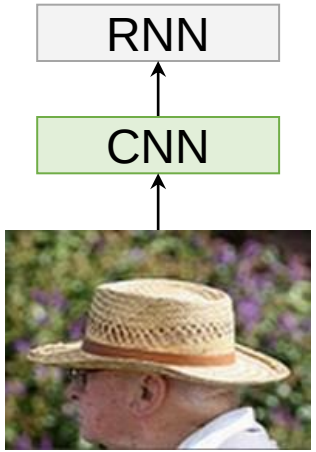


Image captioning

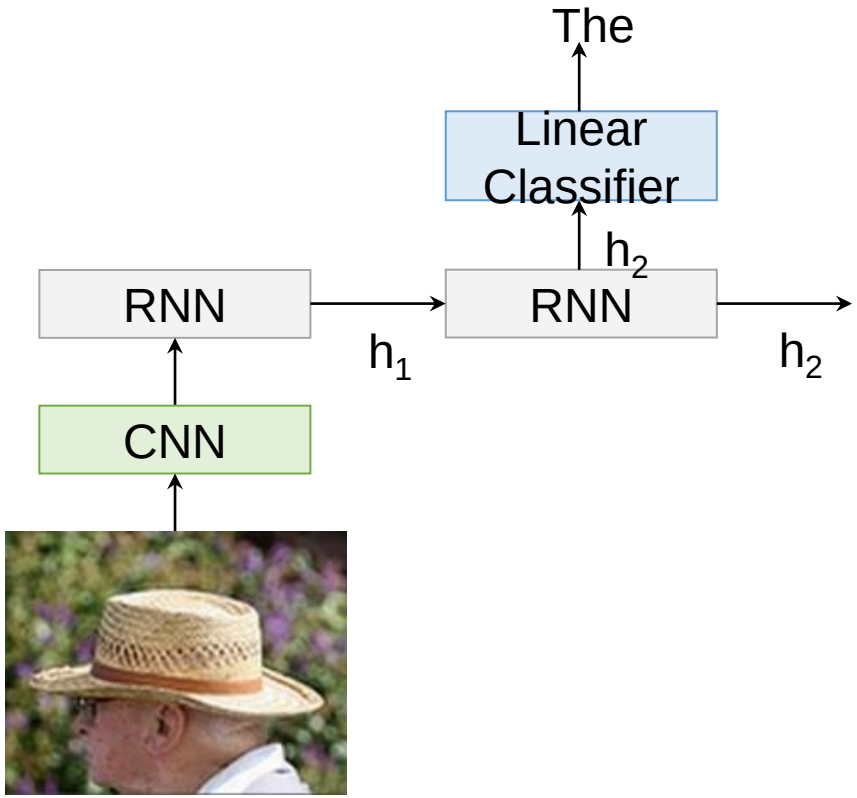
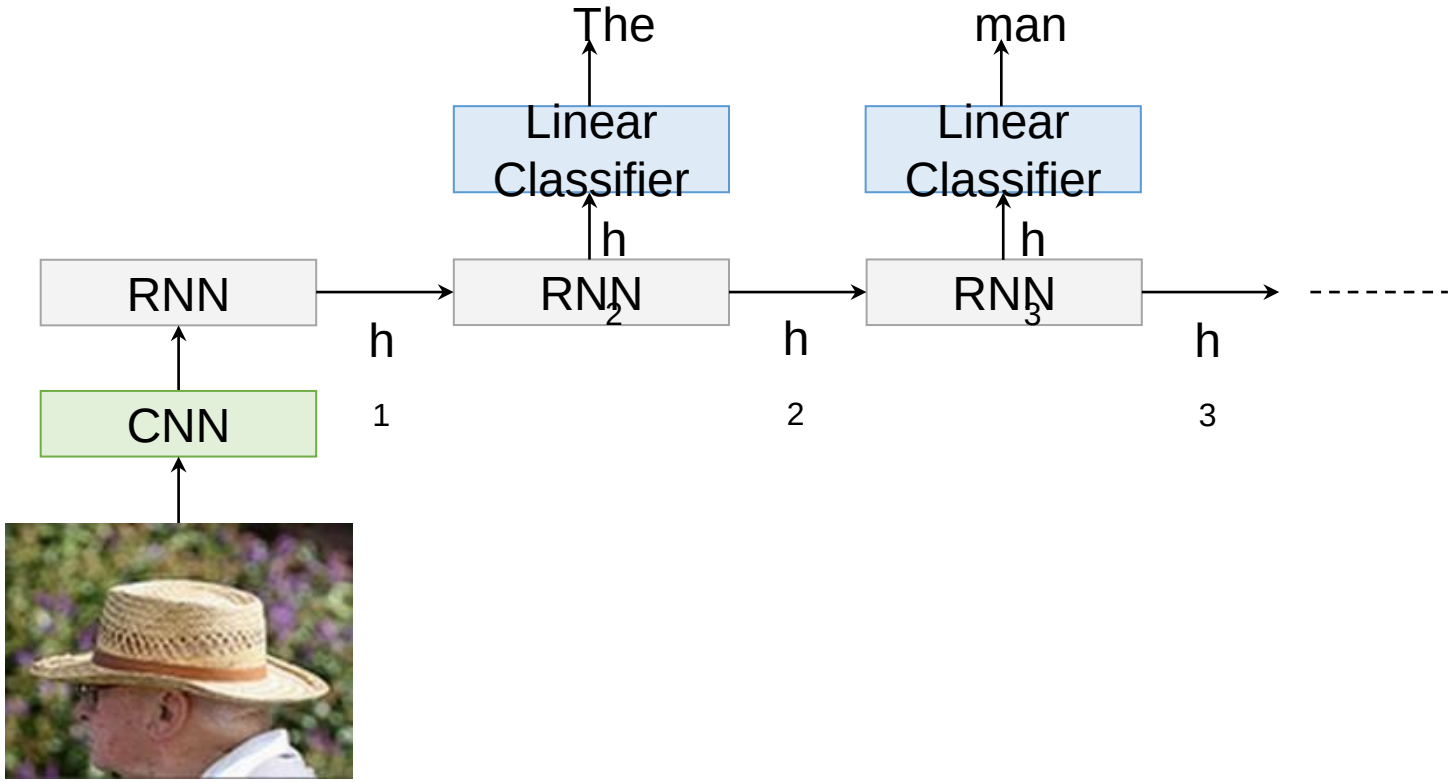


Image captioning



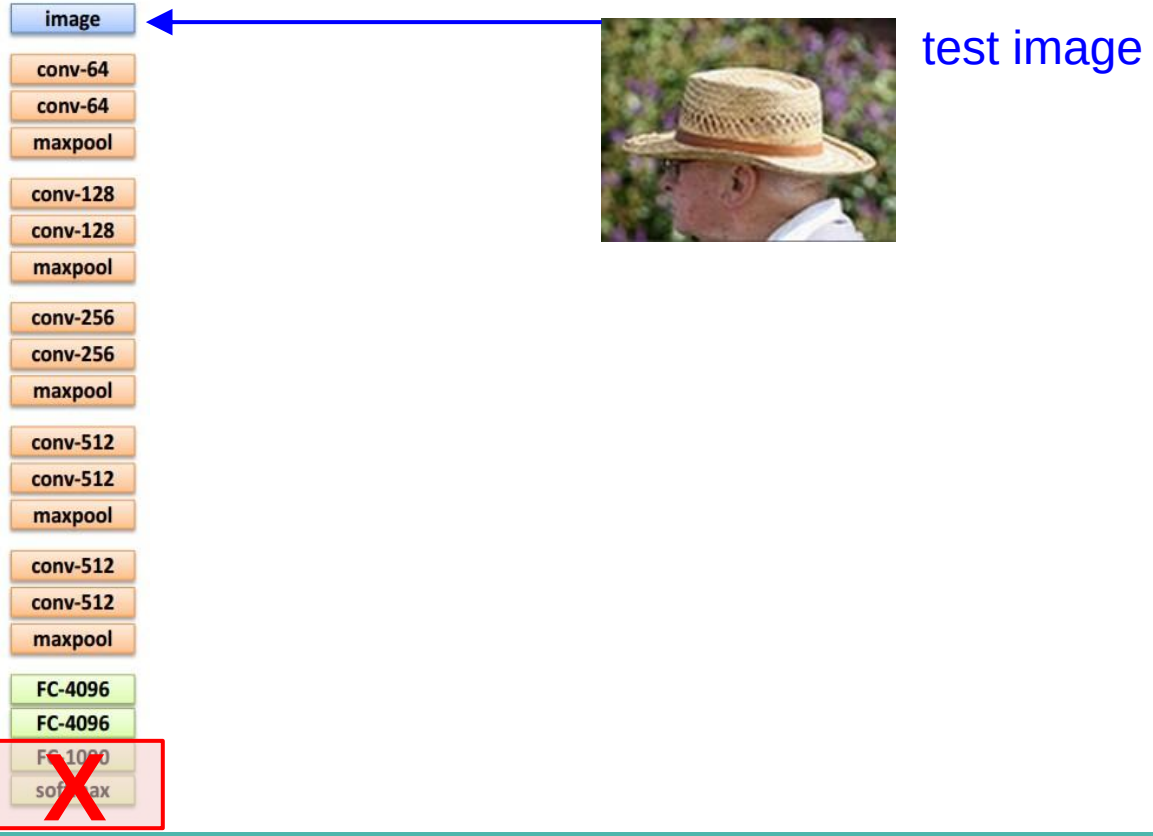


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image

image

conv-64
conv-64
maxpool

conv-128
conv-128
maxpool

conv-256
conv-256
maxpool

conv-512
conv-512
maxpool

conv-512
conv-512
maxpool

FC-4096
FC-4096

test image

x0
<ST
ART
>

<START>

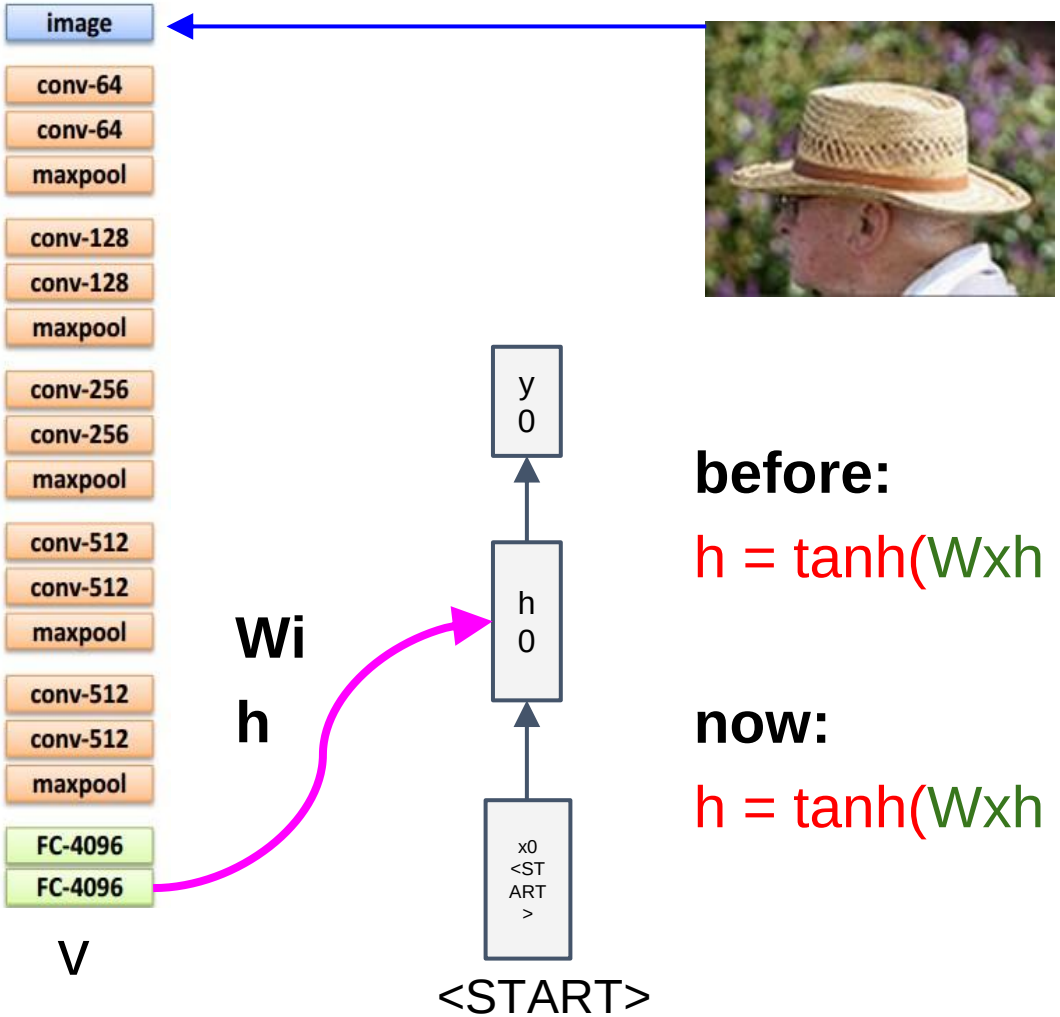


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
V
Wi
h
y
0
h
0
x0
<ST
ART
>
<START>
test image
before:
h = tanh(Wxh * x + Whh * h)
now:
h = tanh(Wxh * x + Whh * h + Wih

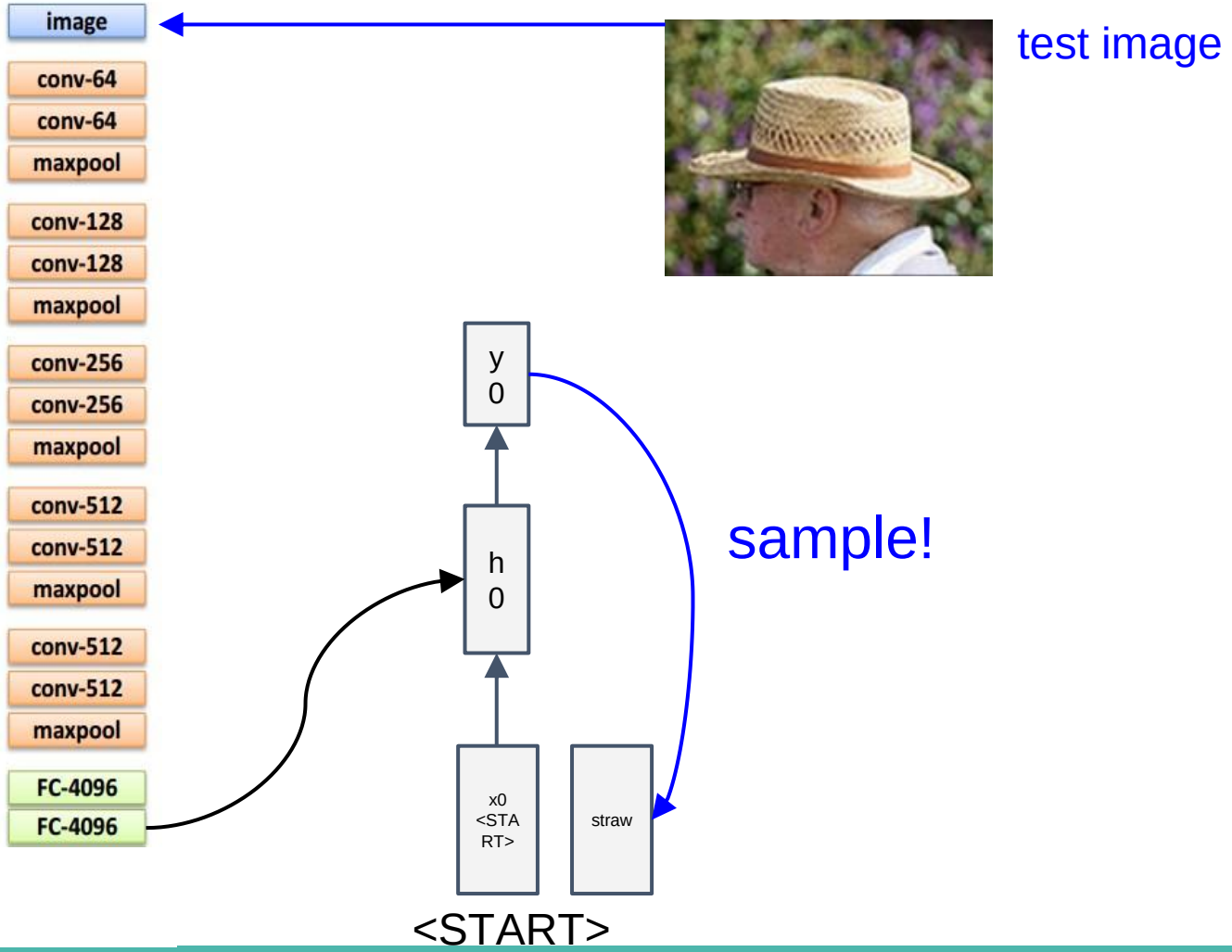


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y
0
h
0
sample!
x0
<STA
RT>
straw
<START>

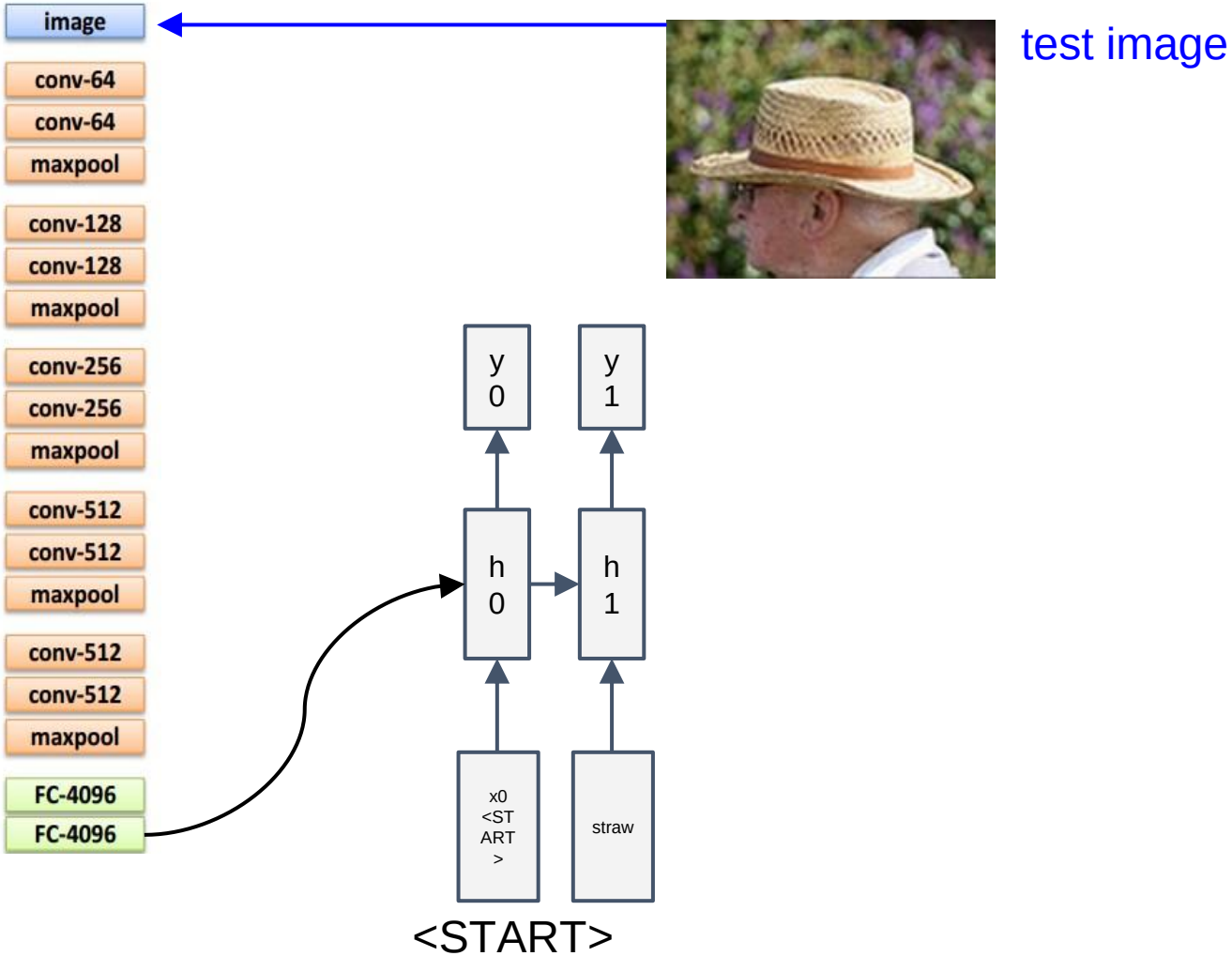


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y 0
y 1
h 0
h 1
x0 <ST ART >
straw
<START>

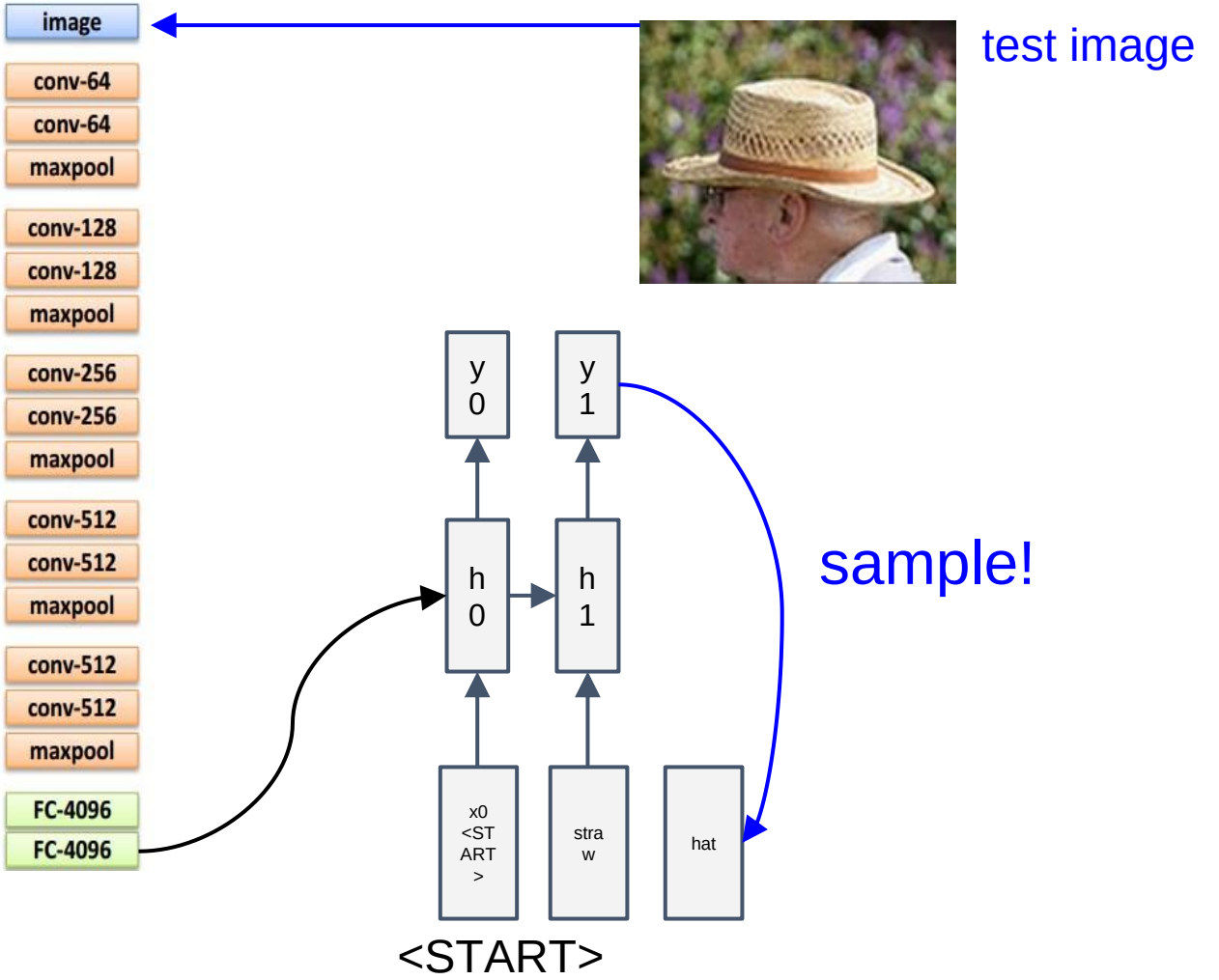


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y 0
y 1
h 0
h 1
sample!
x0 <START>
straw
hat
<START>

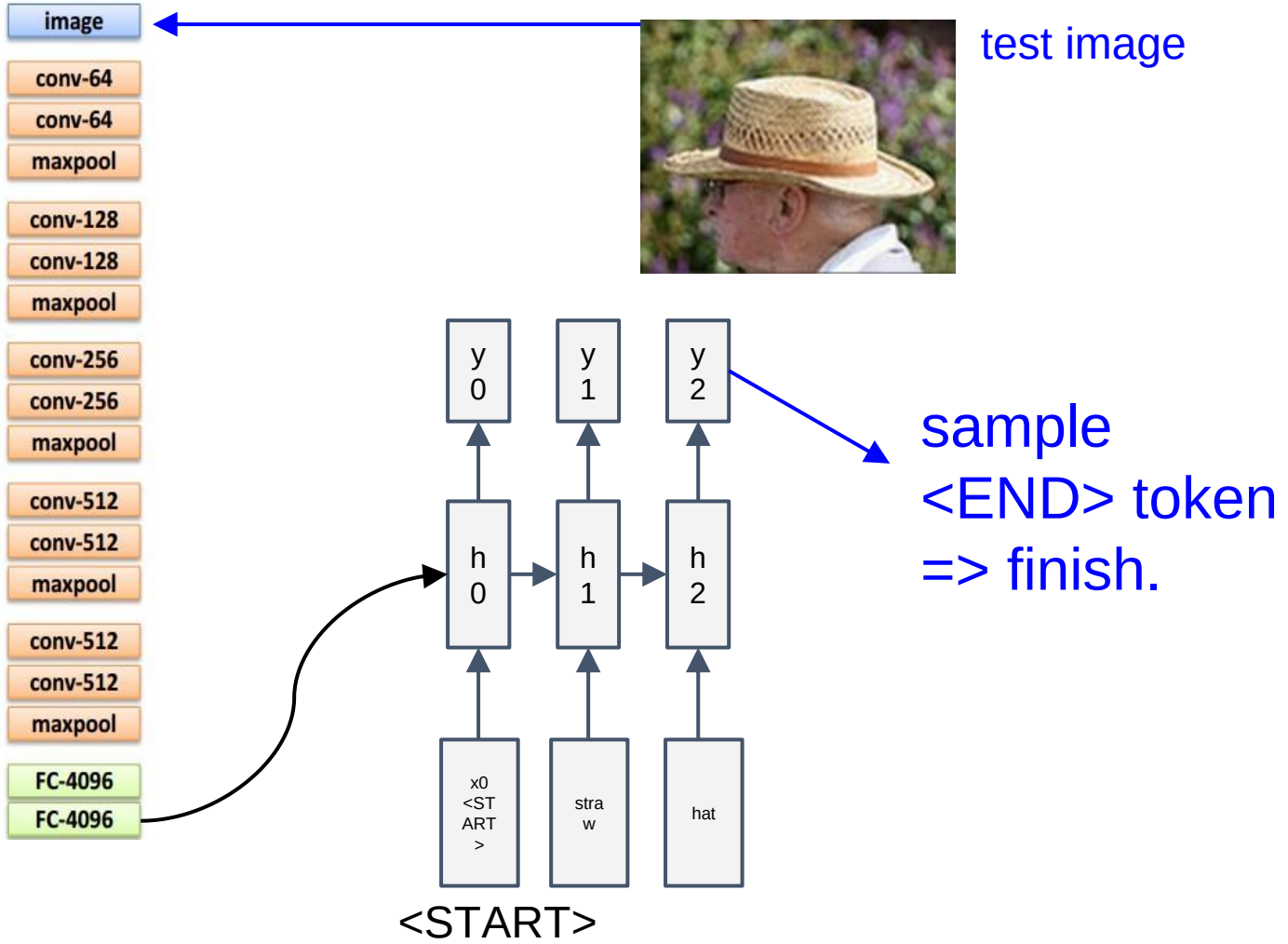


image
conv-64
conv-64
maxpool
conv-128
conv-128
maxpool
conv-256
conv-256
maxpool
conv-512
conv-512
maxpool
conv-512
conv-512
maxpool
FC-4096
FC-4096
test image
y 0
y 1
y 2
h 0
h 1
h 2
x0 <START>
straw
hat
sample <END> token => finish.
<START>

Dataset

a man riding a bike on a dirt path through a forest.
bicyclist raises his fist as he rides on desert dirt trail.
this dirt bike rider is smiling and raising his fist in triumph.
a man riding a bicycle while pumping his fist in the air.
a mountain biker pumps his fist in celebration.

Microsoft COCO
[Tsung-Yi Lin et al. 2014]
mscoco.org

currently:
~120K images
~5 sentences each

Prediction

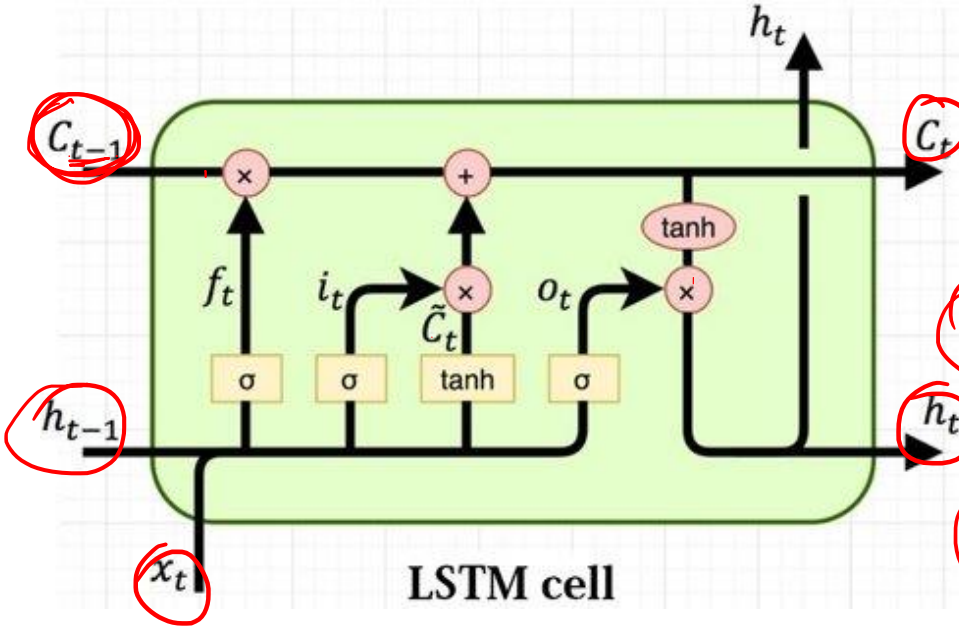
"man in black shirt is playing guitar."

"construction worker in orange safety vest is working on road."

"two young girls are playing with lego toy."

"boy is doing backflip on wakeboard."

LSTM



$$i_t = \sigma\left(x_t U^i + h_{t-1} W^i\right)$$

$$f_t = \sigma\left(x_t U^f + h_{t-1} W^f\right)$$

$$o_t = \sigma\left(x_t U^o + h_{t-1} W^o\right)$$

$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$

$$C_t = \sigma\left(f_t * C_{t-1} + i_t * \tilde{C}_t\right)$$

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

Q&A


@LateNightSeth
THANK YOU!