

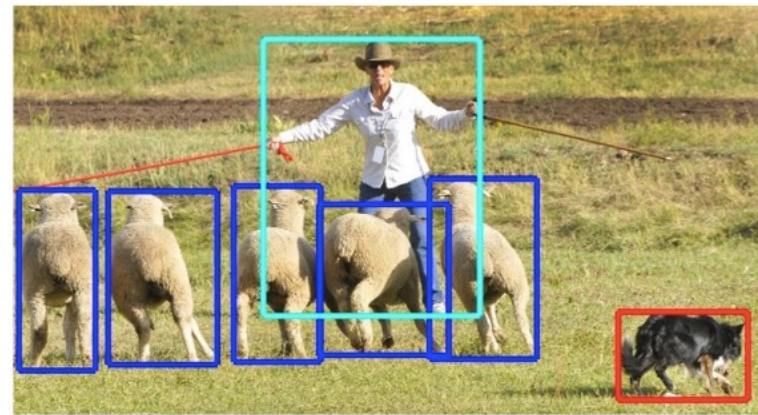
# Recurrent Neural Networks

Xiaolong Wang

# Previous classes



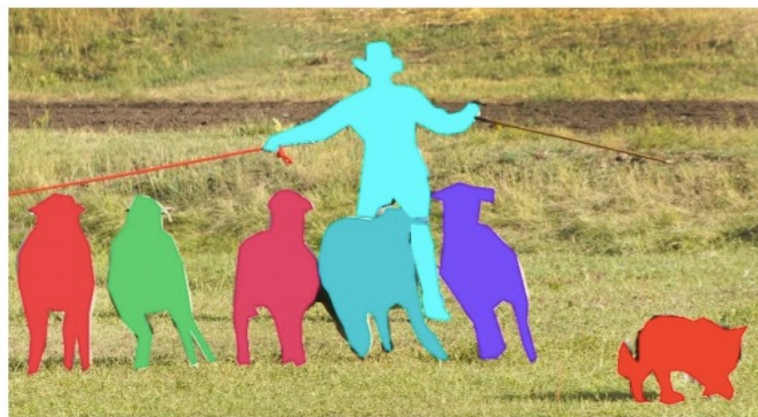
image classification



object detection

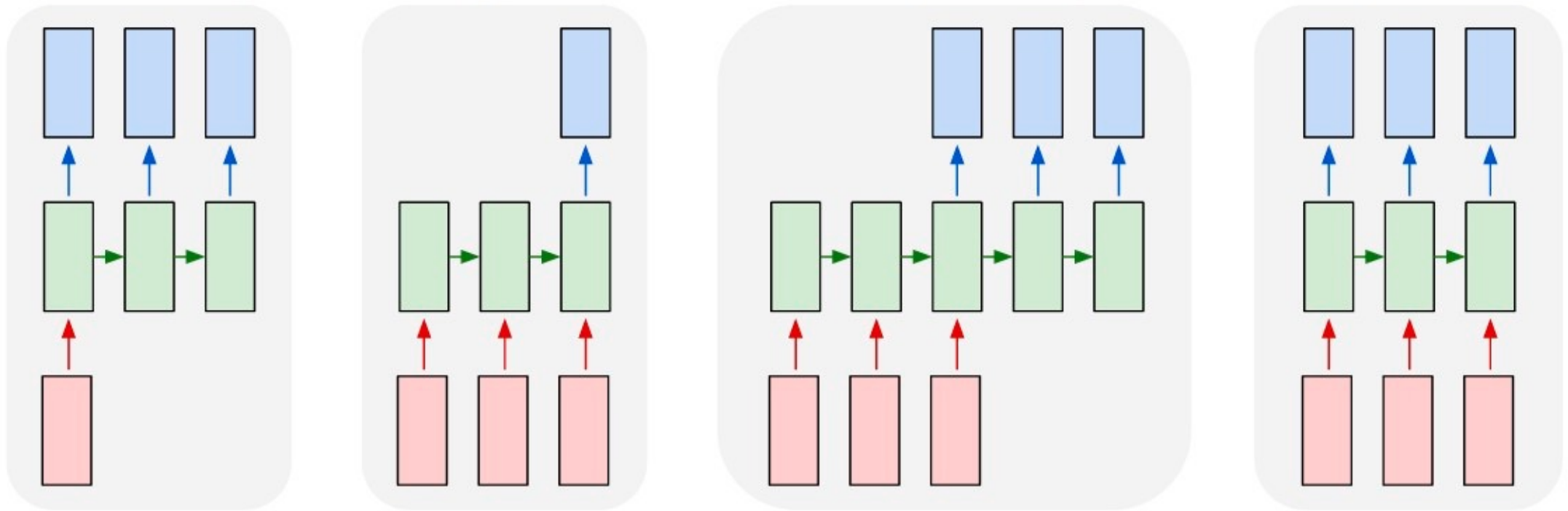


semantic segmentation



instance segmentation

# This Class: Recurrent Neural Networks



# This Class: Recurrent Neural Networks

- The Basic RNN
- LSTM
- Application in language and vision tasks

# The Basic RNN

# Sequential prediction tasks

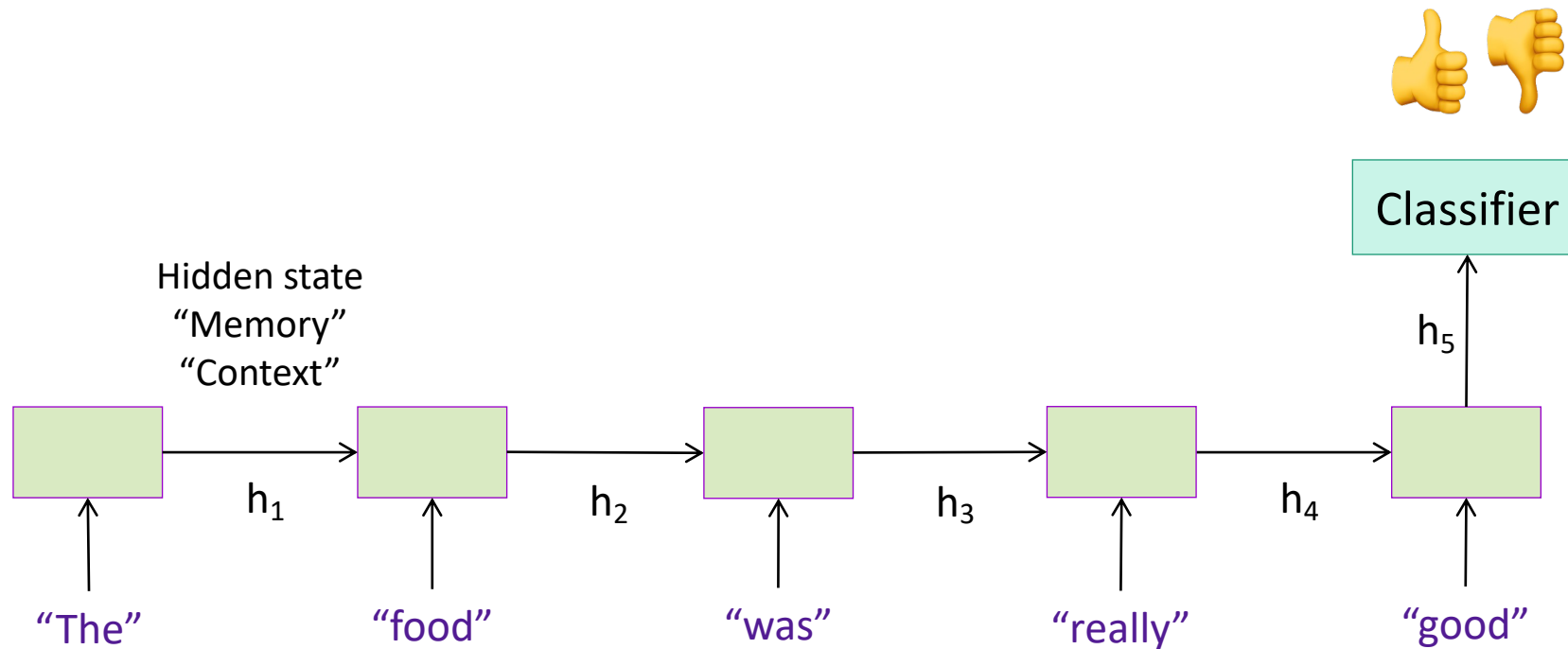
- So far, we focused mainly on prediction problems with fixed-size inputs and outputs
- But what if the input and/or output is a variable-length sequence?

# Task 1: Sentiment classification

- Goal: classify a text sequence (e.g., restaurant, movie or product review, Tweet) as having positive or negative sentiment
  - “The food was really good”
  - “The vacuum cleaner broke within two weeks”
  - “The movie had slow parts, but overall was worth watching”

# Task 1: Sentiment classification

Recurrent model:





# Task 2: Machine translation

The screenshot shows the Google Translate web interface. At the top, the Google logo is on the left, and the word "Translate" is in red. To the right of "Translate" is a link "Turn off instant translation" and a star icon. Below this, there are two language selection menus. The first menu has "English", "Spanish", "French", and "Detect language" with a dropdown arrow. The second menu has "English", "Spanish", and "Arabic" with a dropdown arrow. A blue "Translate" button is to the right of the second menu. Below the menus, there are two text boxes. The left box is titled "Correspondances" and contains a French poem by Charles Baudelaire. The right box is titled "Matches" and contains the English translation of the poem. At the bottom of the left box, there are icons for a speaker, a microphone, and a keyboard, along with the text "693/5000". At the bottom of the right box, there are icons for a star, a document, a speaker, and a share icon, along with a pencil icon.

Google

Translate Turn off instant translation

English Spanish French Detect language

English Spanish Arabic Translate

**Correspondances**

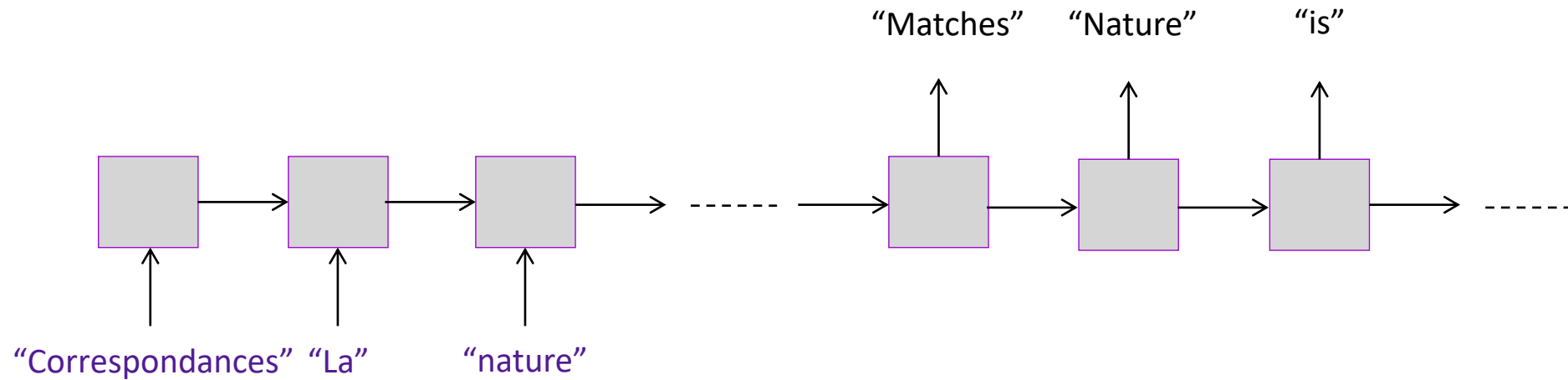
La Nature est un temple où de vivants piliers  
Laisserent parfois sortir de confuses paroles;  
L'homme y passe à travers des forêts de symboles  
Qui l'observent avec des regards familiers.  
Comme de longs échos qui de loin se confondent  
Dans une ténébreuse et profonde unité,  
Vaste comme la nuit et comme la clarté,  
Les parfums, les couleurs et les sons se répondent.  
Il est des parfums frais comme des chairs d'enfants,  
Doux comme les hautbois, verts comme les prairies,  
— Et d'autres, corrompus, riches et triomphants,  
Ayant l'expansion des choses infinies,  
Comme l'ambre, le musc, le benjoin et l'encens,  
Qui chantent les transports de l'esprit et des sens.  
— Charles Baudelaire

**Matches**

Nature is a temple where living pillars  
Sometimes let out confused words;  
Man goes through symbol forests  
Which observe him with familiar eyes.  
Like long echoes that by far merge  
In a dark and deep unity,  
As vast as the night and as clarity,  
The perfumes, the colors and the sounds answer each  
other.  
There are fresh perfumes like children's flesh,  
Sweet like oboes, green like meadows,  
- And others, corrupt, rich and triumphant,  
Having the expansion of infinite things,  
Like amber, musk, benzoin and incense,  
Who sing the transports of the mind and the senses.  
- Charles Baudelaire

693/5000

# Task 2: Machine translation



# Task 3: Image caption generation



*A cat sitting on a suitcase on the floor*



*A cat is sitting on a tree branch*



*A dog is running in the grass with a frisbee*



*A white teddy bear sitting in the grass*



*Two people walking on the beach with surfboards*



*A tennis player in action on the court*

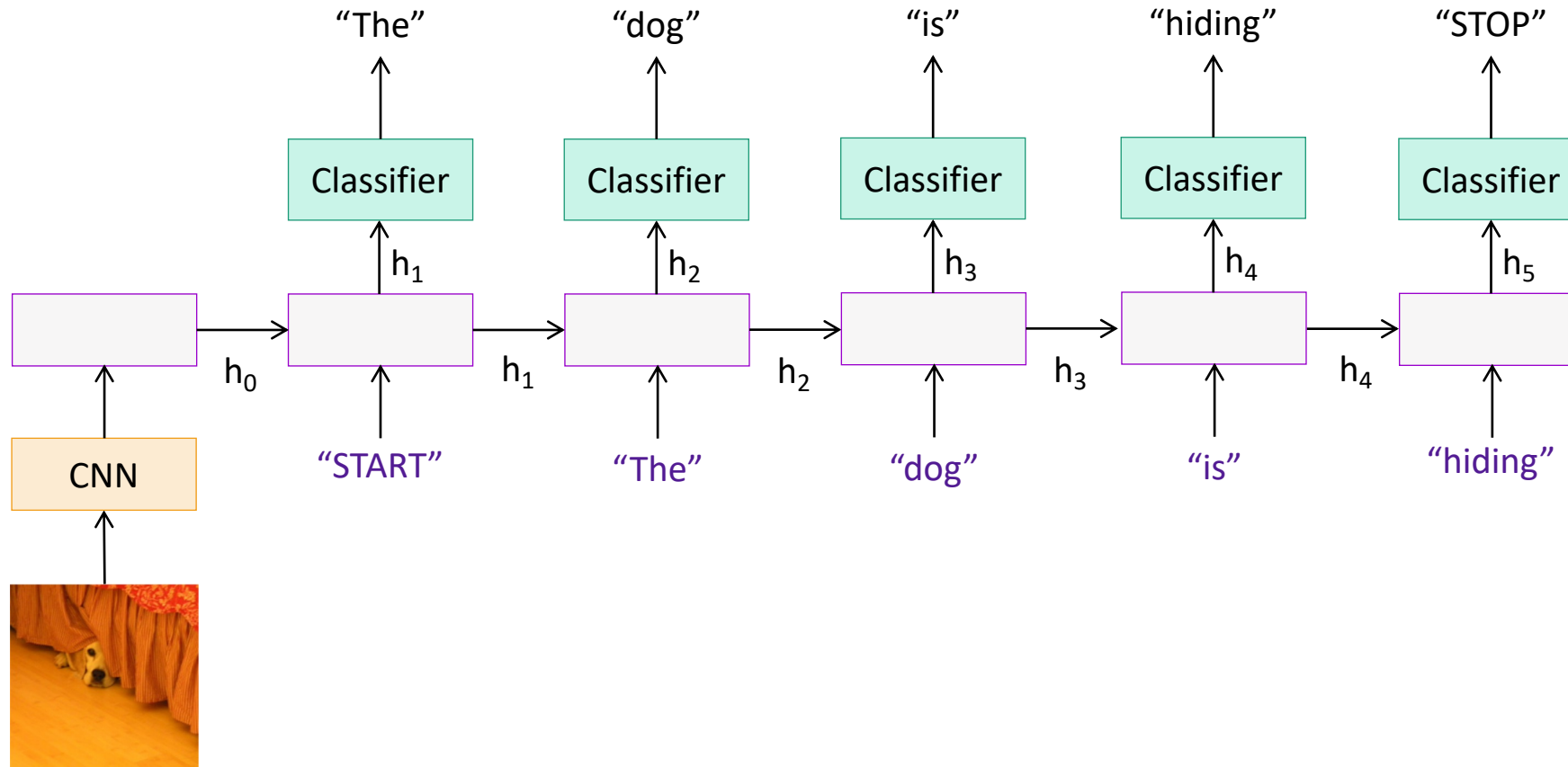


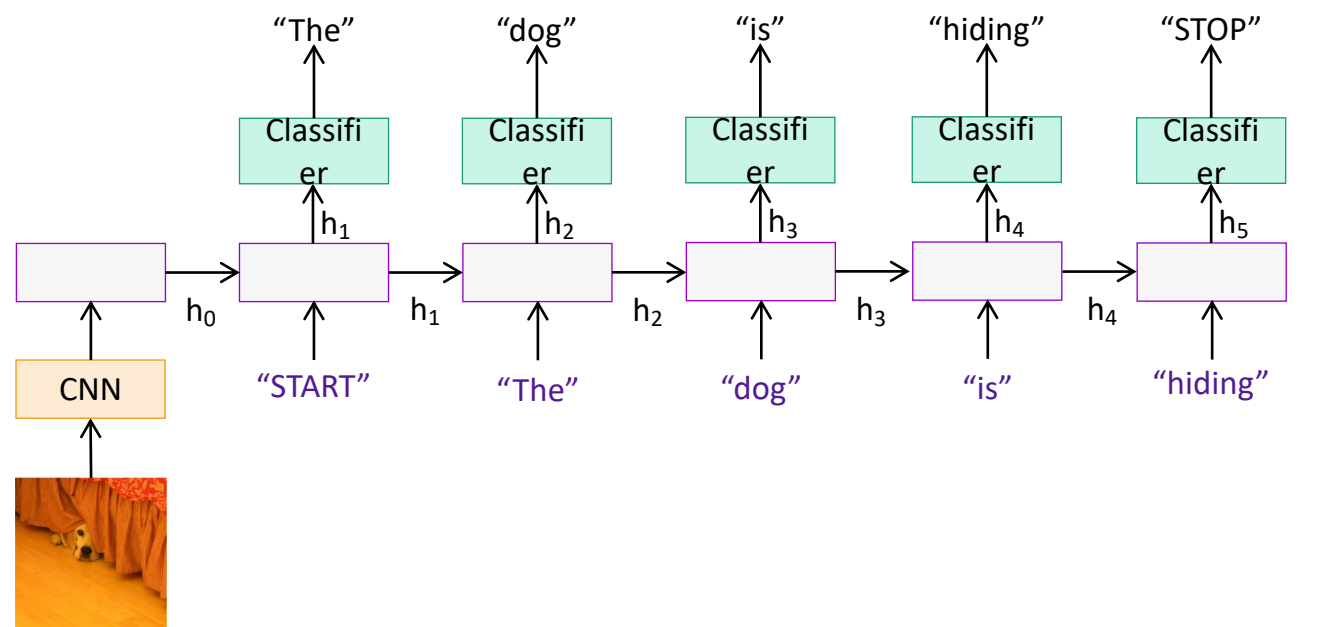
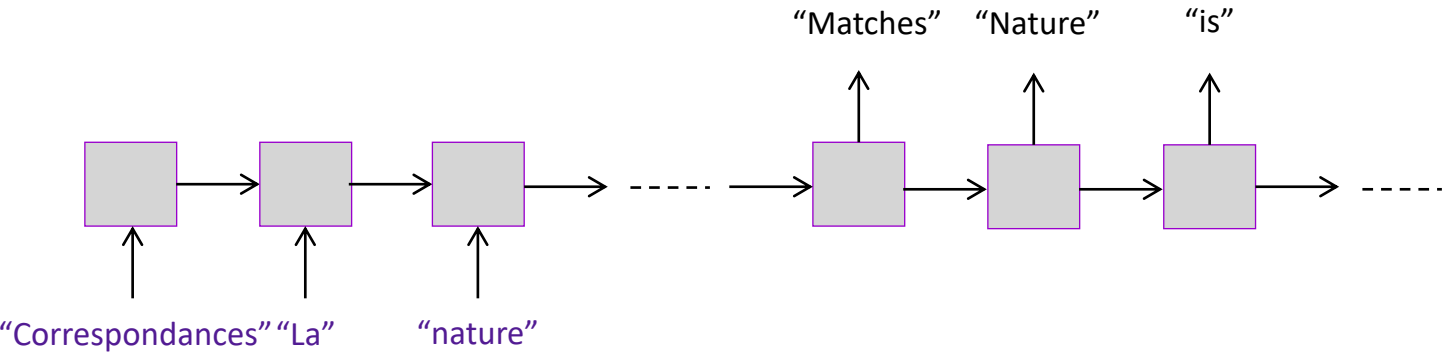
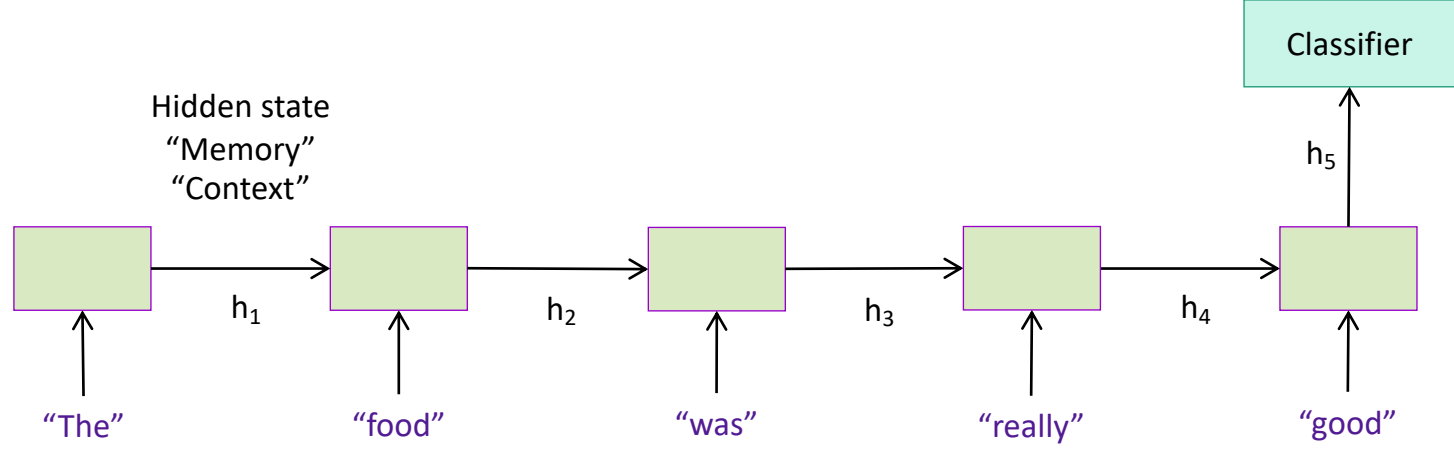
*Two giraffes standing in a grassy field*



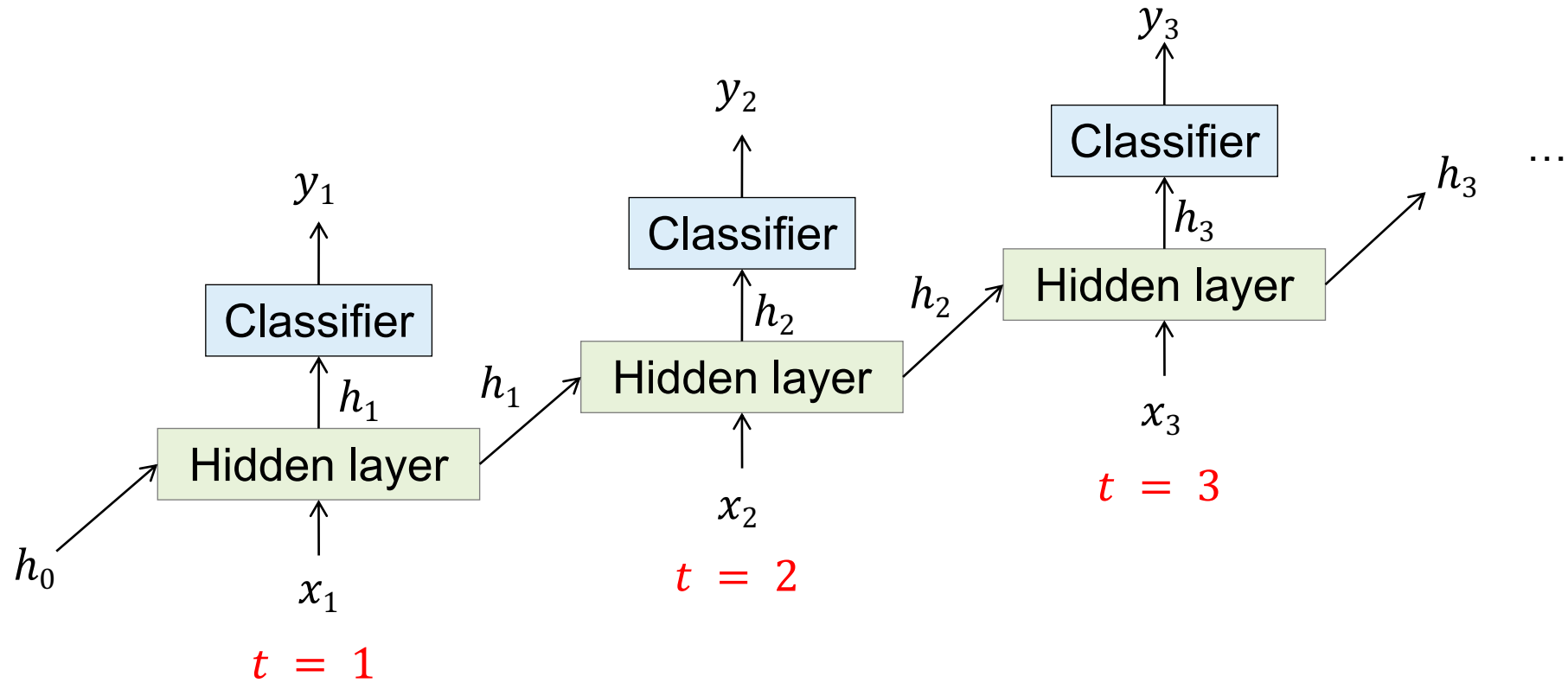
*A man riding a dirt bike on a dirt track*

# Task 3: Image caption generation

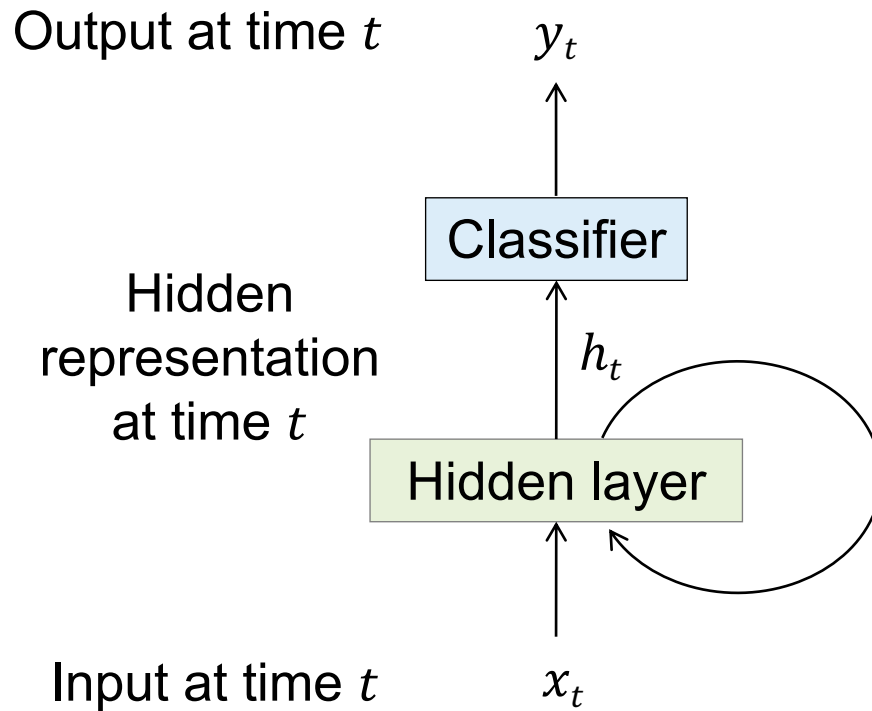




# Recurrent unit



# Recurrent unit



Recurrence:

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

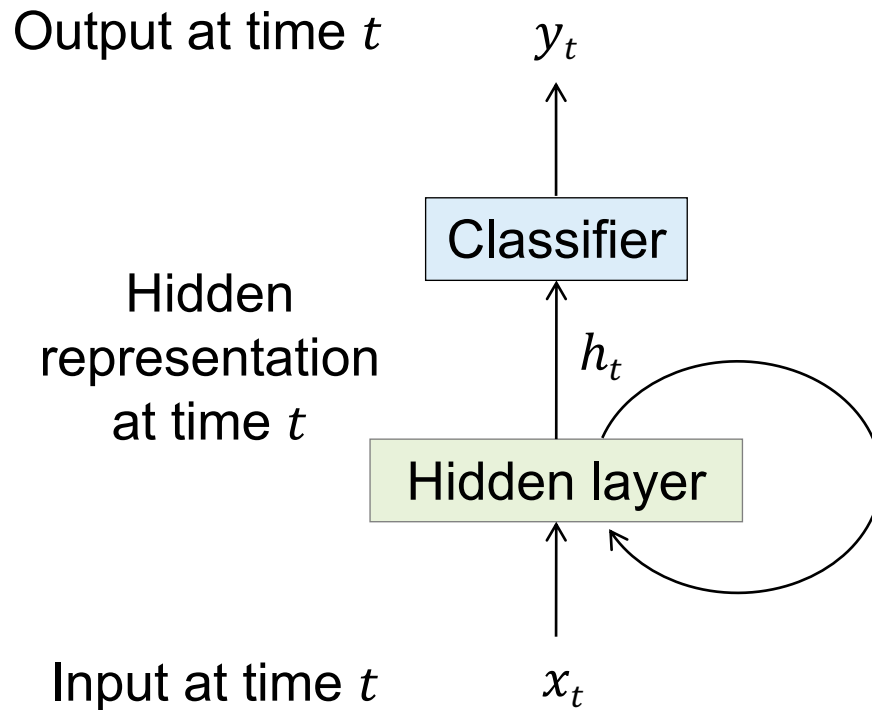
new  
state

function  
of  $W$

input at  
time  $t$

old state

# Recurrent unit



Recurrence:

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

new  
state

function  
of  $W$

input at  
time  $t$

old state

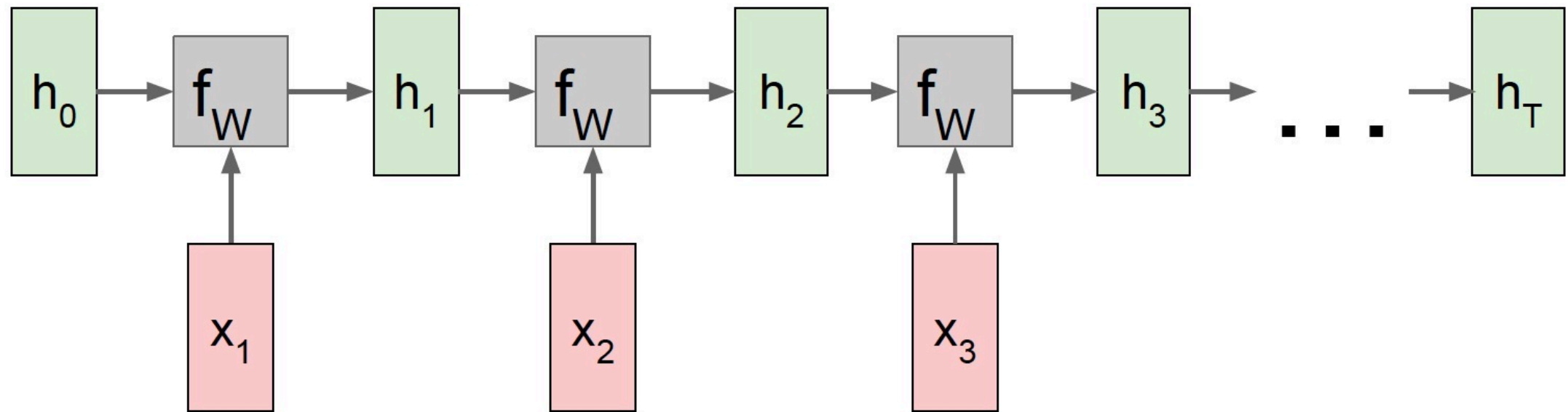


$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t)$$

$$y_t = W_{hy}h_t$$

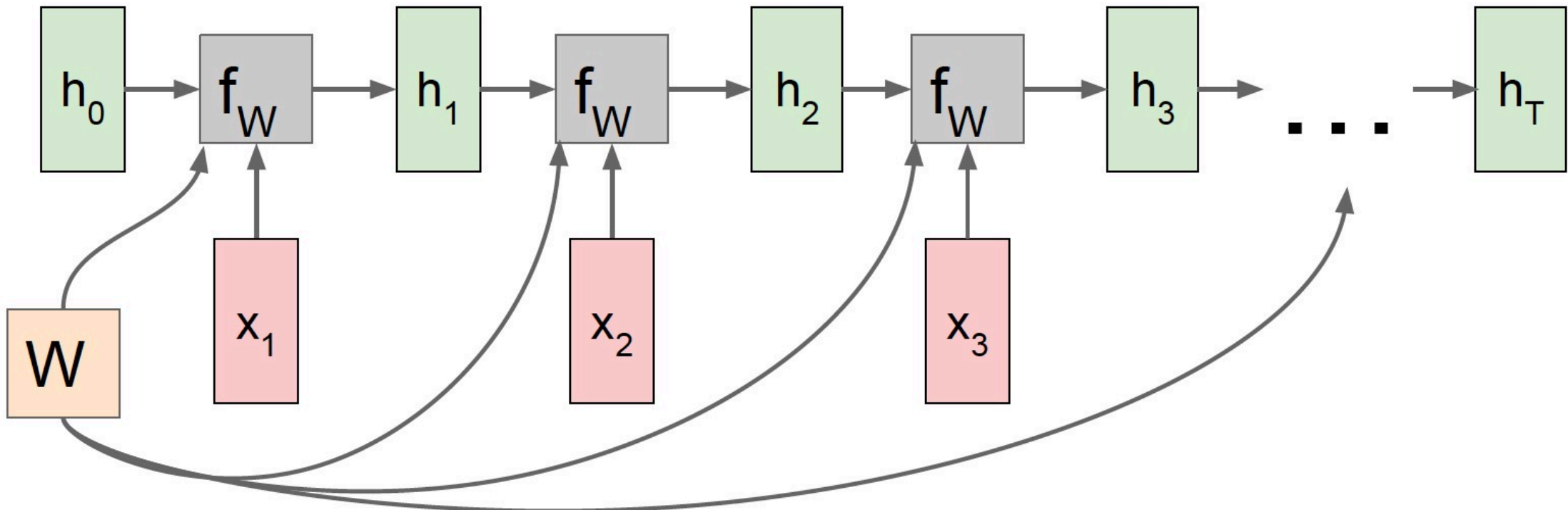


# RNN: Computational Graph

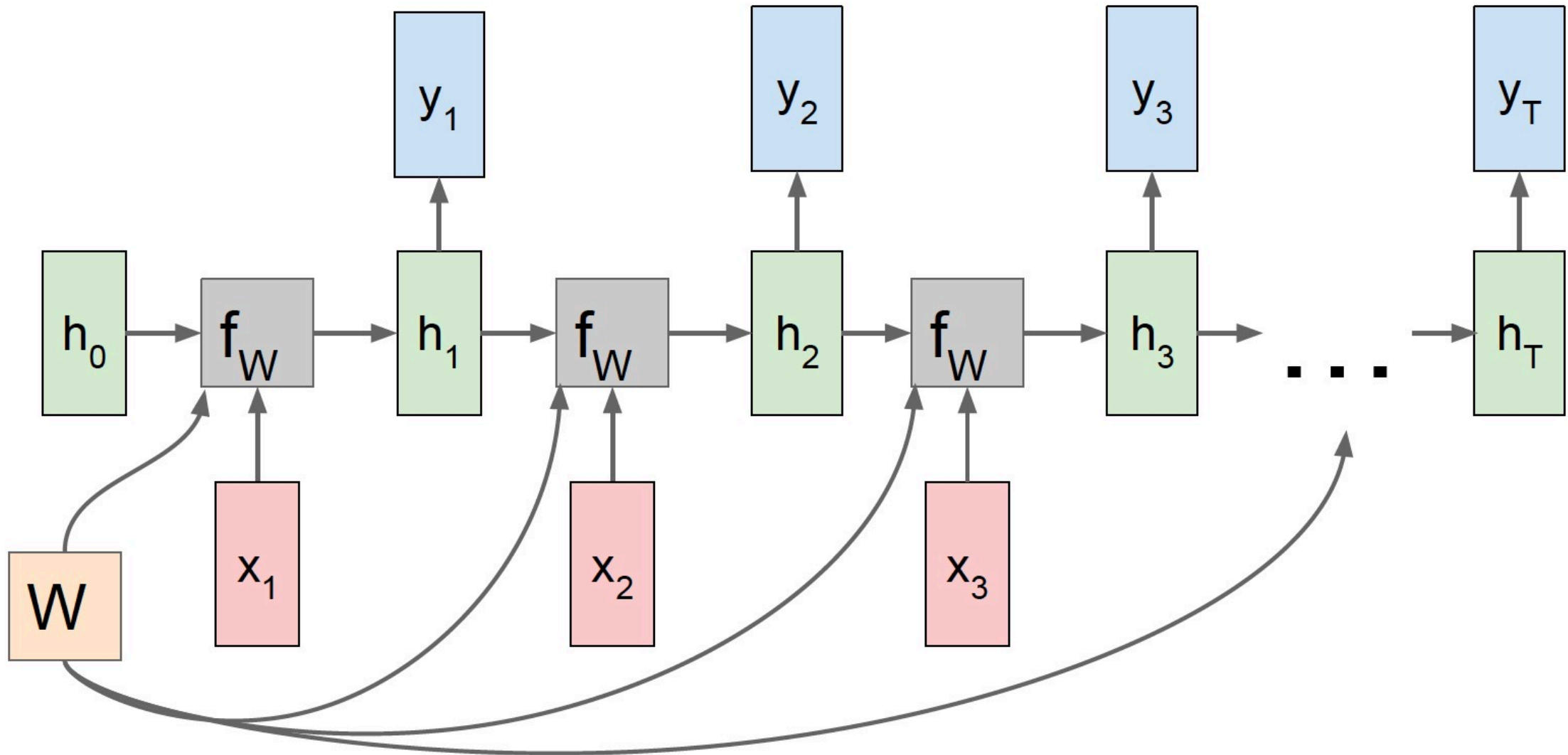


# RNN: Computational Graph

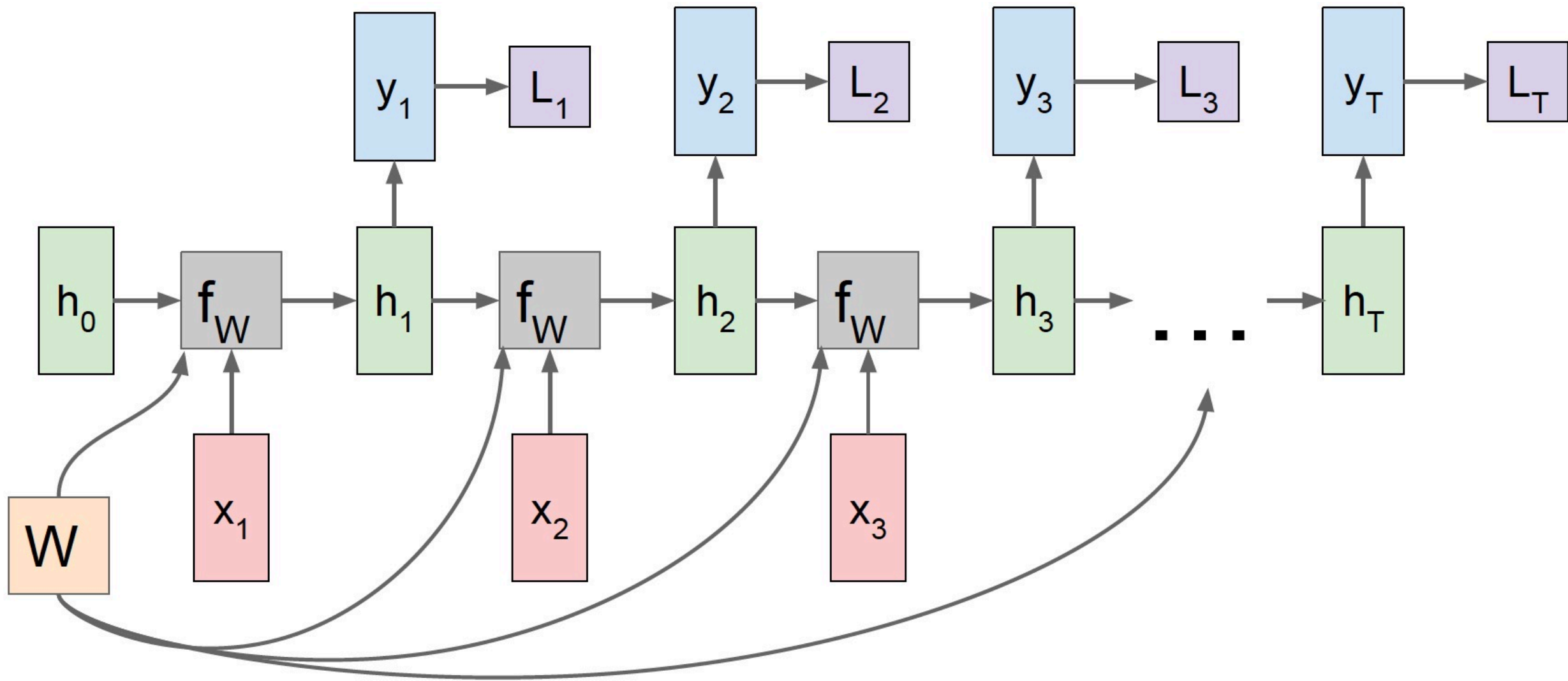
Re-use the same weight in every time step



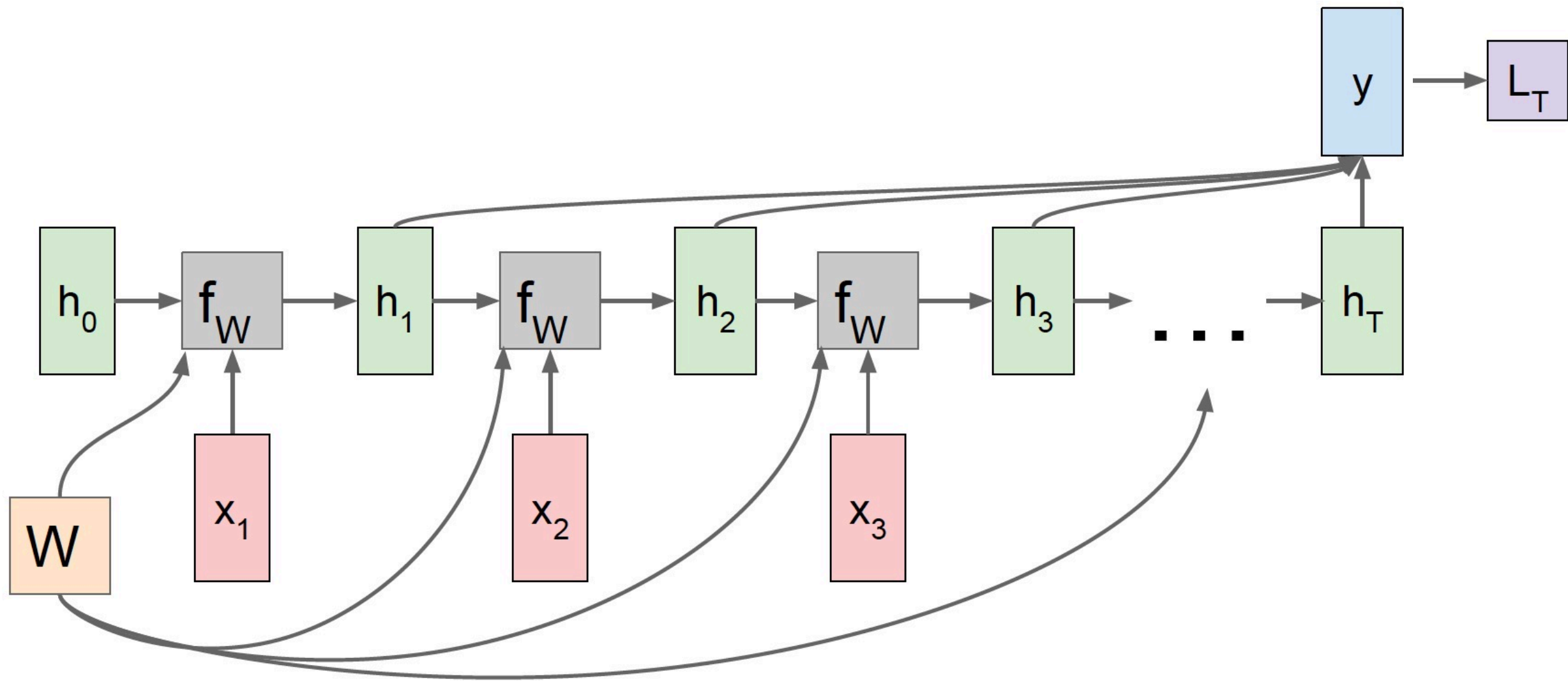
# RNN: Computational Graph



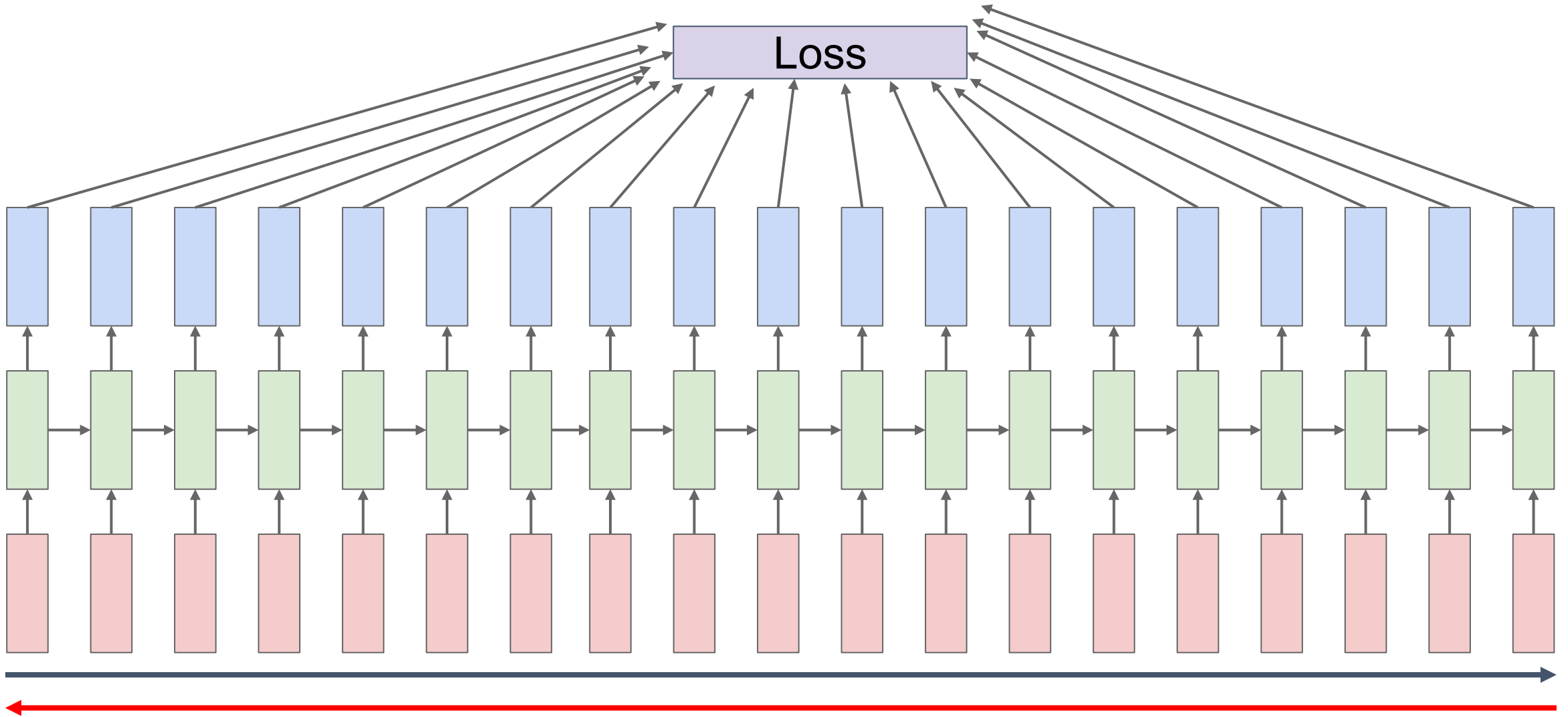
# RNN: Computational Graph



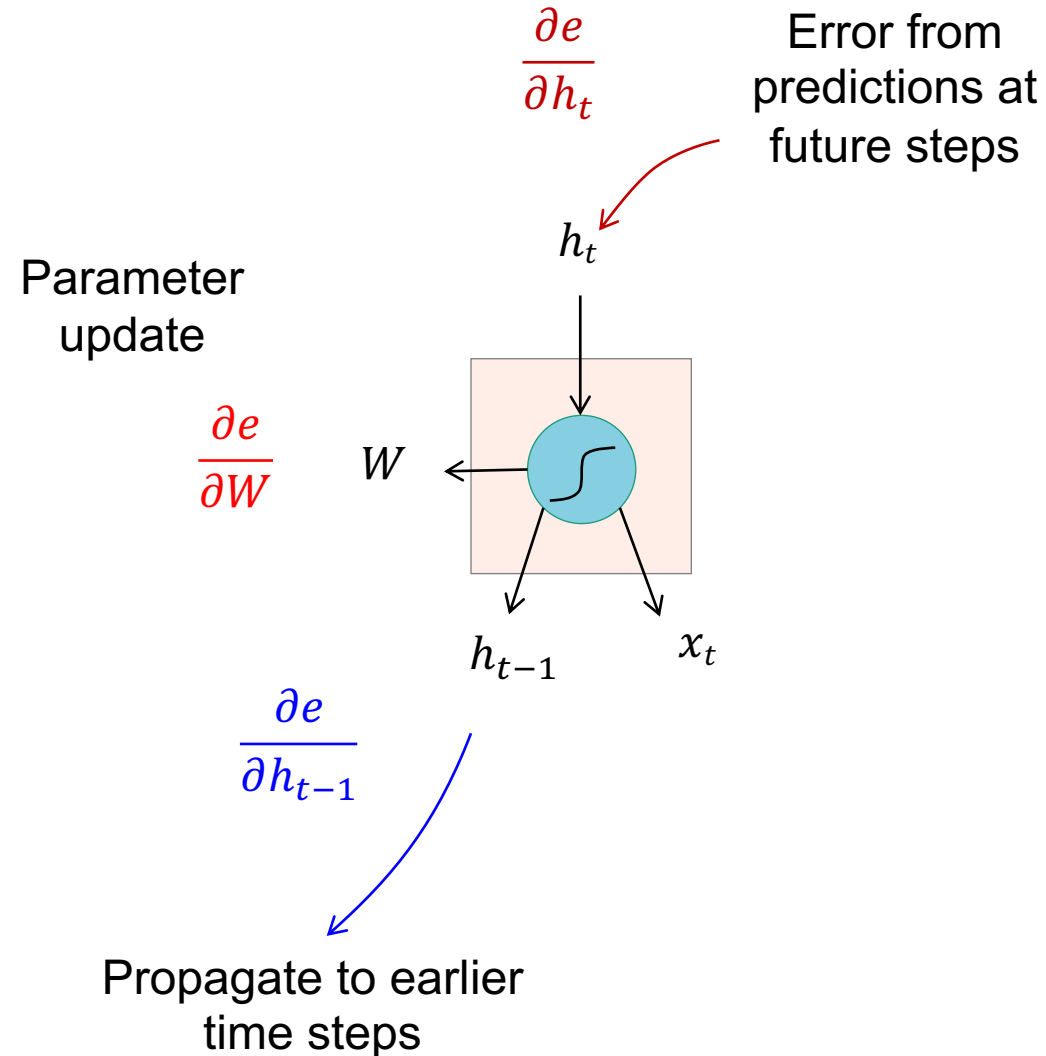
# RNN: Computational Graph



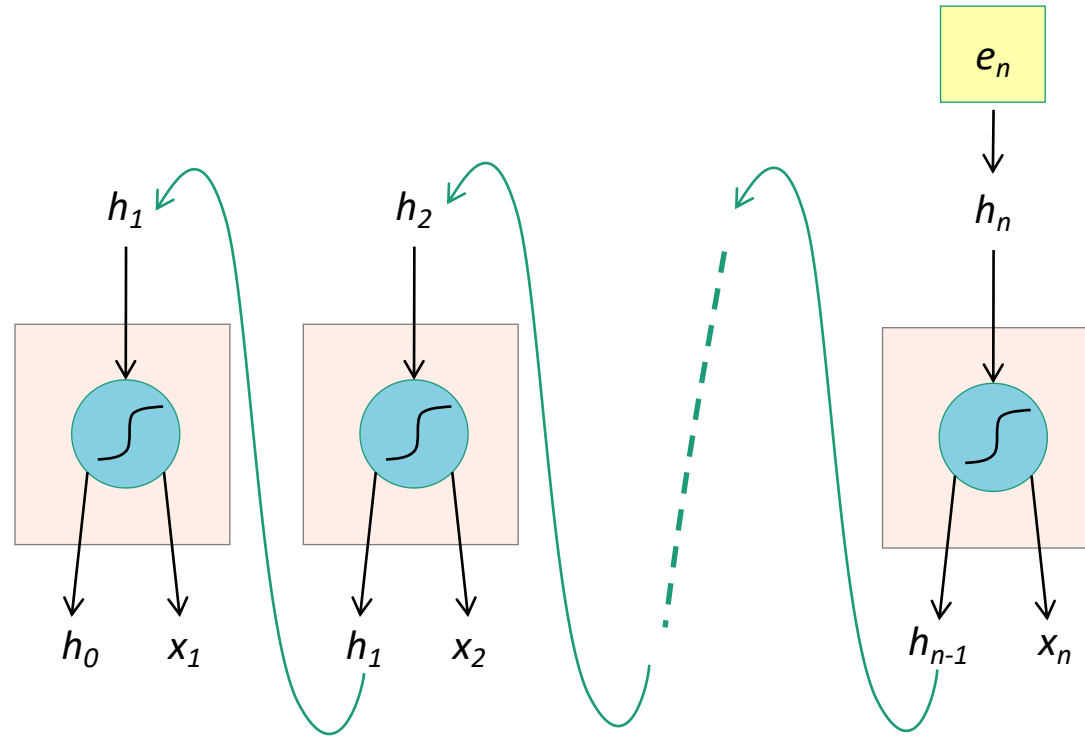
# Backpropagation through time



# Backpropagation through time (BPTT)

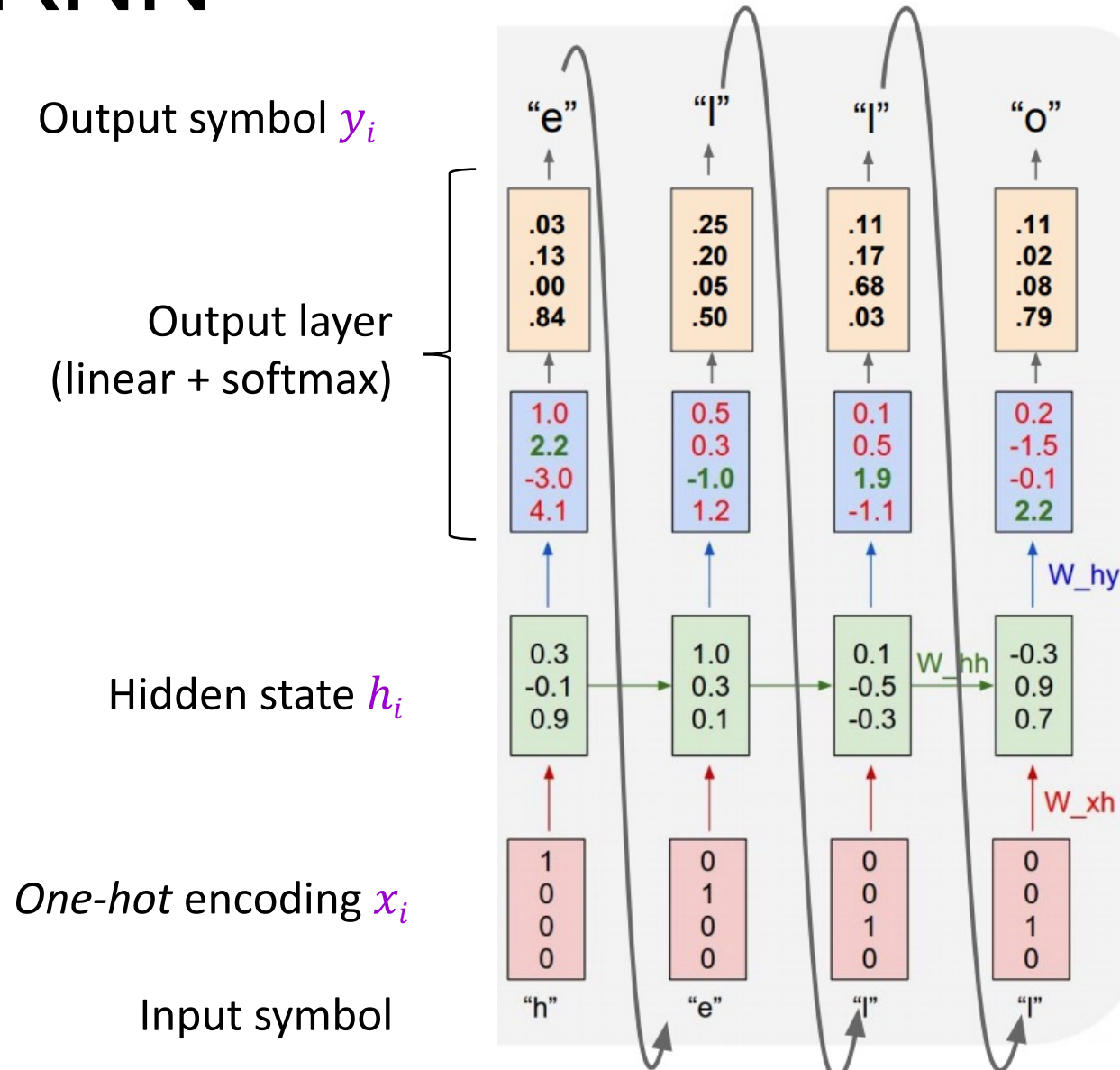


# Backpropagation through time (BPTT)





# Char-RNN



# Char-RNN

100th  
iteration

```
tyntd-iafhatawiaoihrdemot lytdws e ,tfti, astai f ogoh eoase rrranbyne 'nhthnee e  
plia tklrqd t o idoe ns,smtt h ne etie h,hregtrs niglike,aoaenns lng
```

train more

300th  
iteration

```
"Tmont thithey" fomesscerliund  
Keushey. Thom here  
sheulke, anmerenith ol sivh I lalterthend Bleipile shuwv fil on aseterlome  
coaniogennc Phe lism thond hon at. MeiDimorotion in ther thize."
```

train more

700th  
iteration

```
Aftair fall unsuch that the hall for Prince Velzonski's that me of  
her hearly, and behs to so arwage fiving were to it beloge, pavu say falling misfort  
how, and Gogition is so overelical and offer.
```

train more

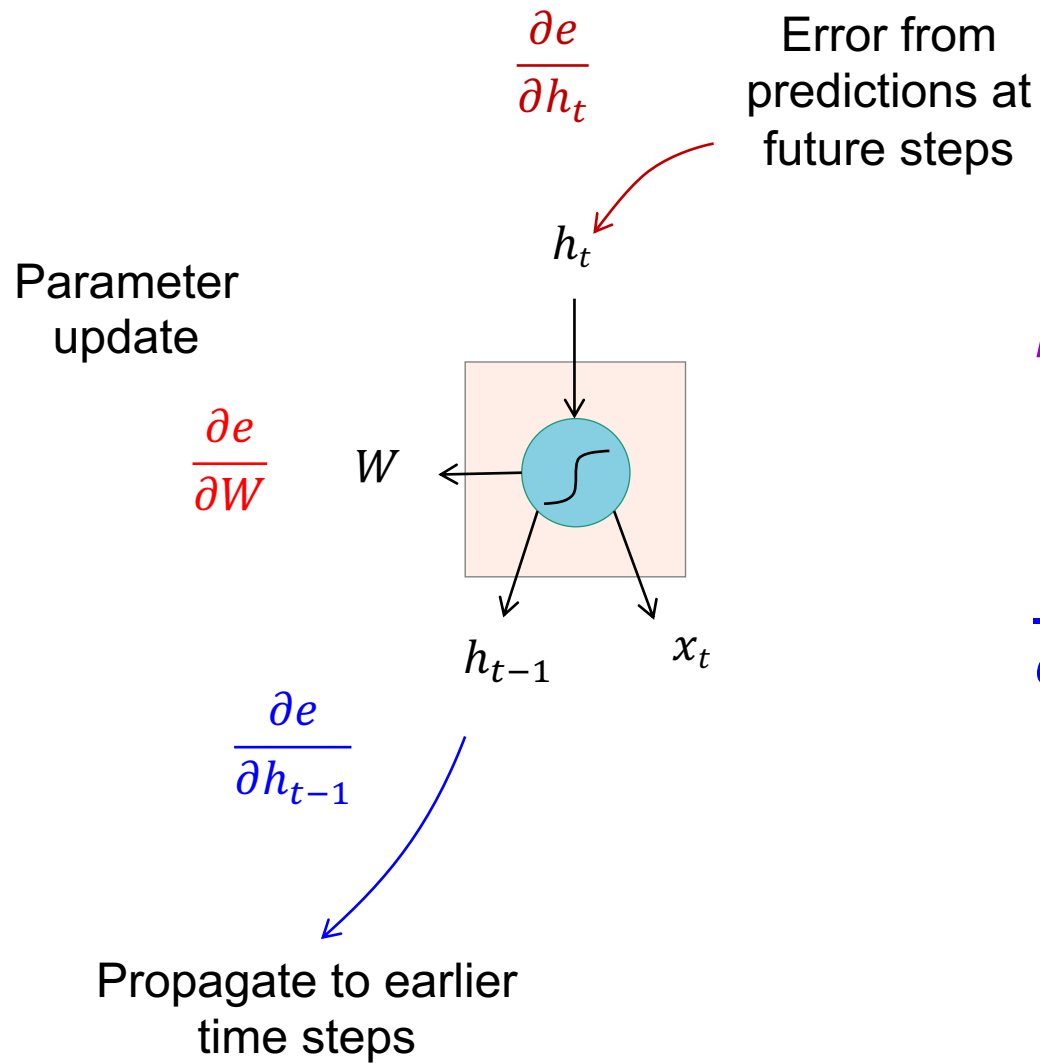
2000th  
iteration

```
"Why do what that day," replied Natasha, and wishing to himself the fact the  
princess, Princess Mary was easier, fed in had oftened him.  
Pierre aking his soul came to the packs and drove up his father-in-law women.
```

5-min break

Long short-term memory (LSTM)

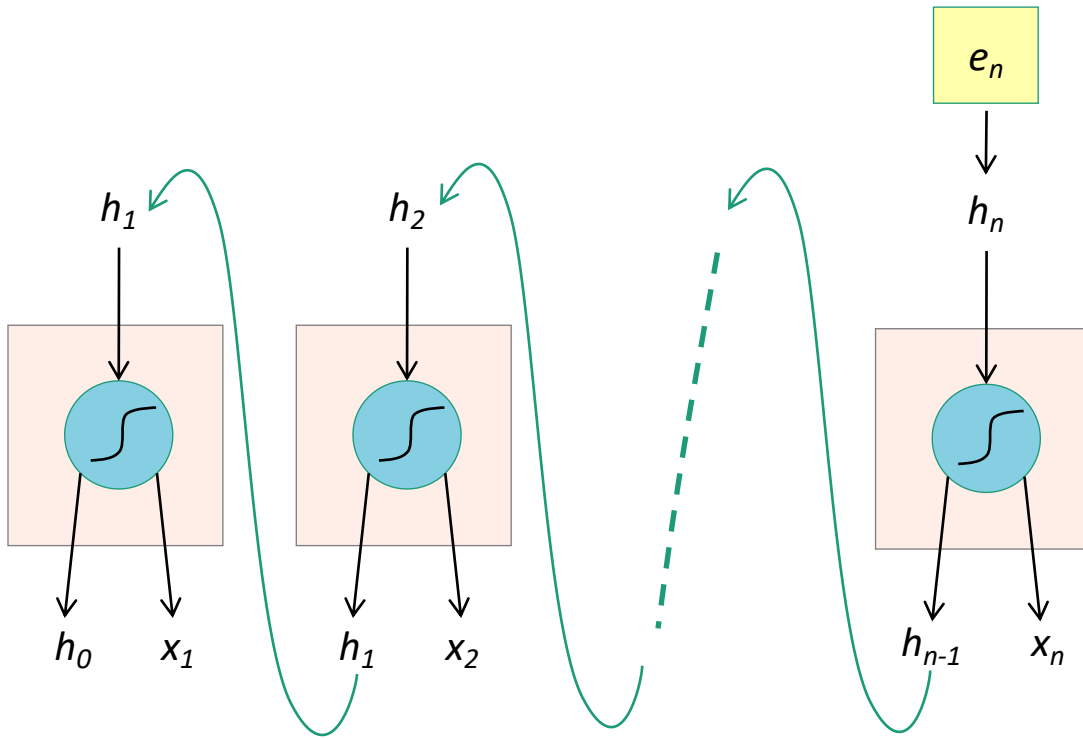
# Vanishing Gradients in RNNs



$$h_t = \tanh(W_x x_t + W_h h_{t-1})$$

$$\frac{\partial e}{\partial h_{t-1}} = W_h^T (1 - \tanh^2(W_x x_t + W_h h_{t-1})) \odot \frac{\partial e}{\partial h_t}$$

# Vanishing Gradients in RNNs



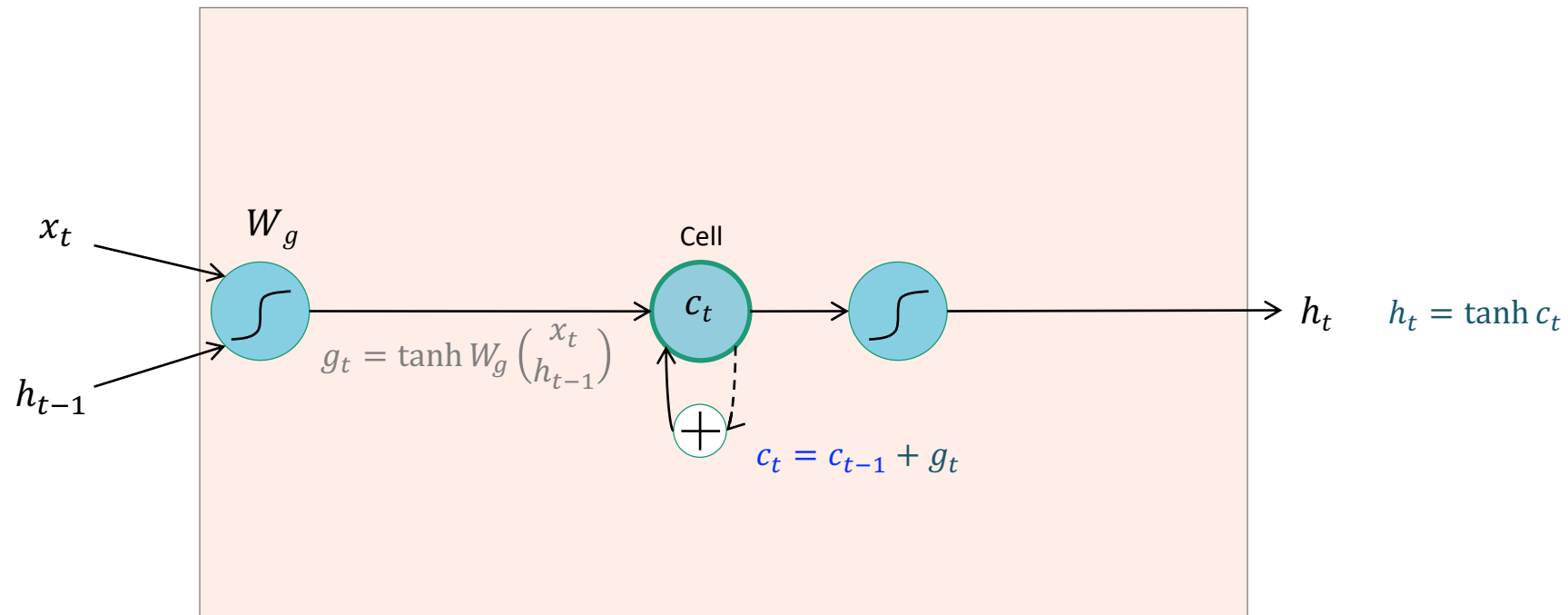
$$\frac{\partial e}{\partial h_{t-1}} = W_h^T (1 - \tanh^2(W_x x_t + W_h h_{t-1})) \odot \frac{\partial e}{\partial h_t}$$

Computing gradient for  $h_0$  involves many multiplications by  $W_h^T$  (and rescalings between 0 and 1)

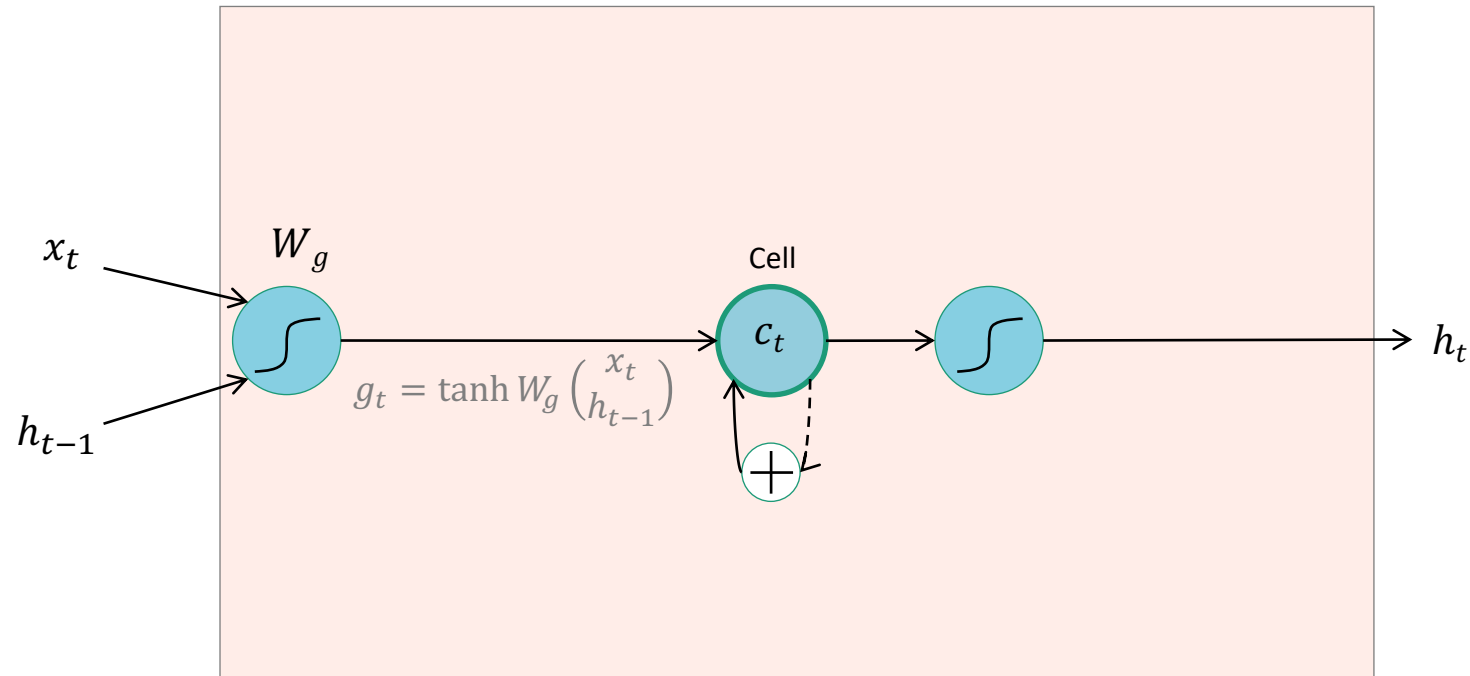
Gradients will *vanish* if largest singular value of  $W_h$  is less than 1 and *explode* if it's greater than 1

# Long short-term memory (LSTM) cell

- Add a *memory cell* that is not subject to matrix multiplication or squishing, thereby avoiding gradient decay

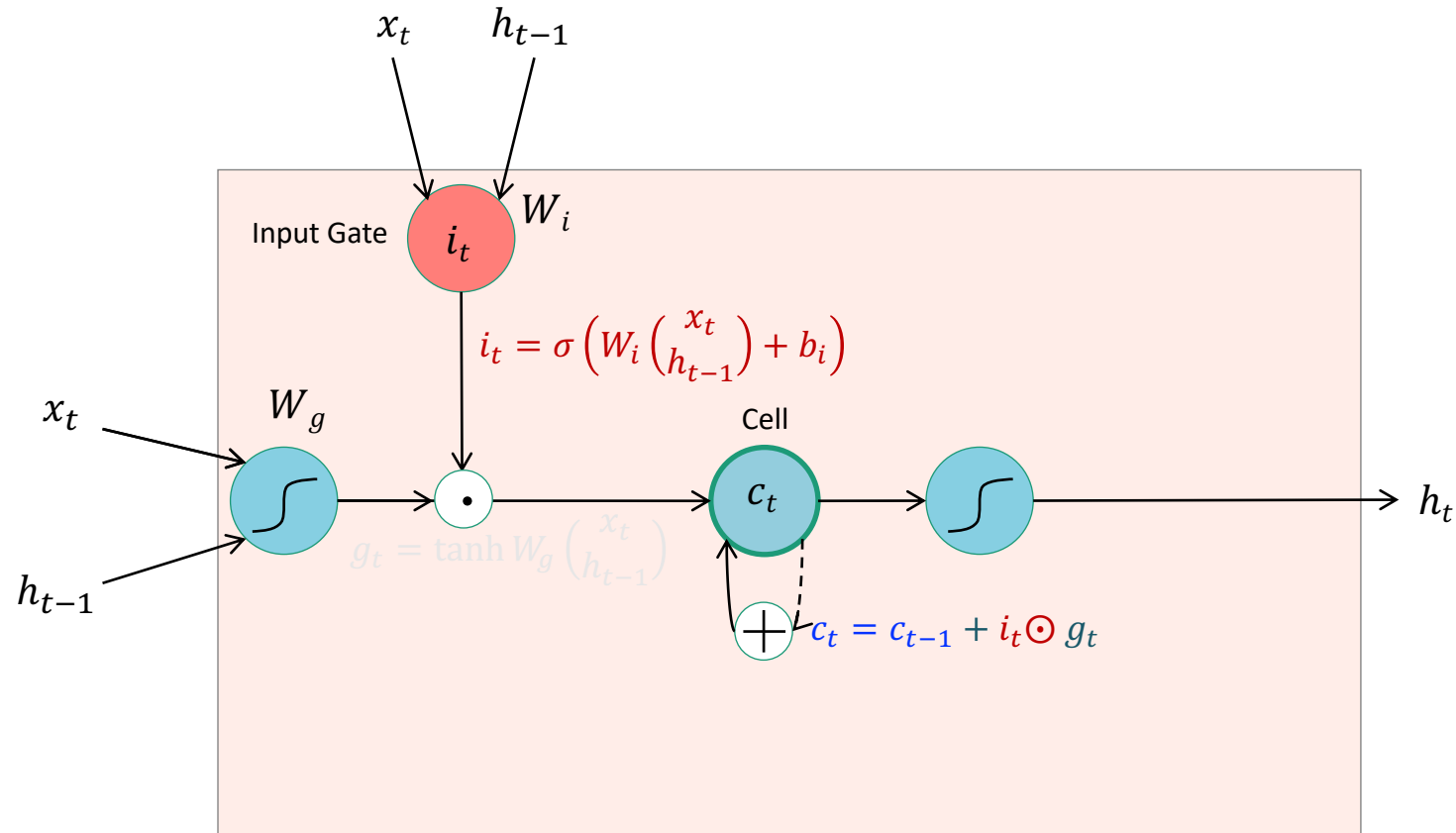


# Long short-term memory (LSTM) cell

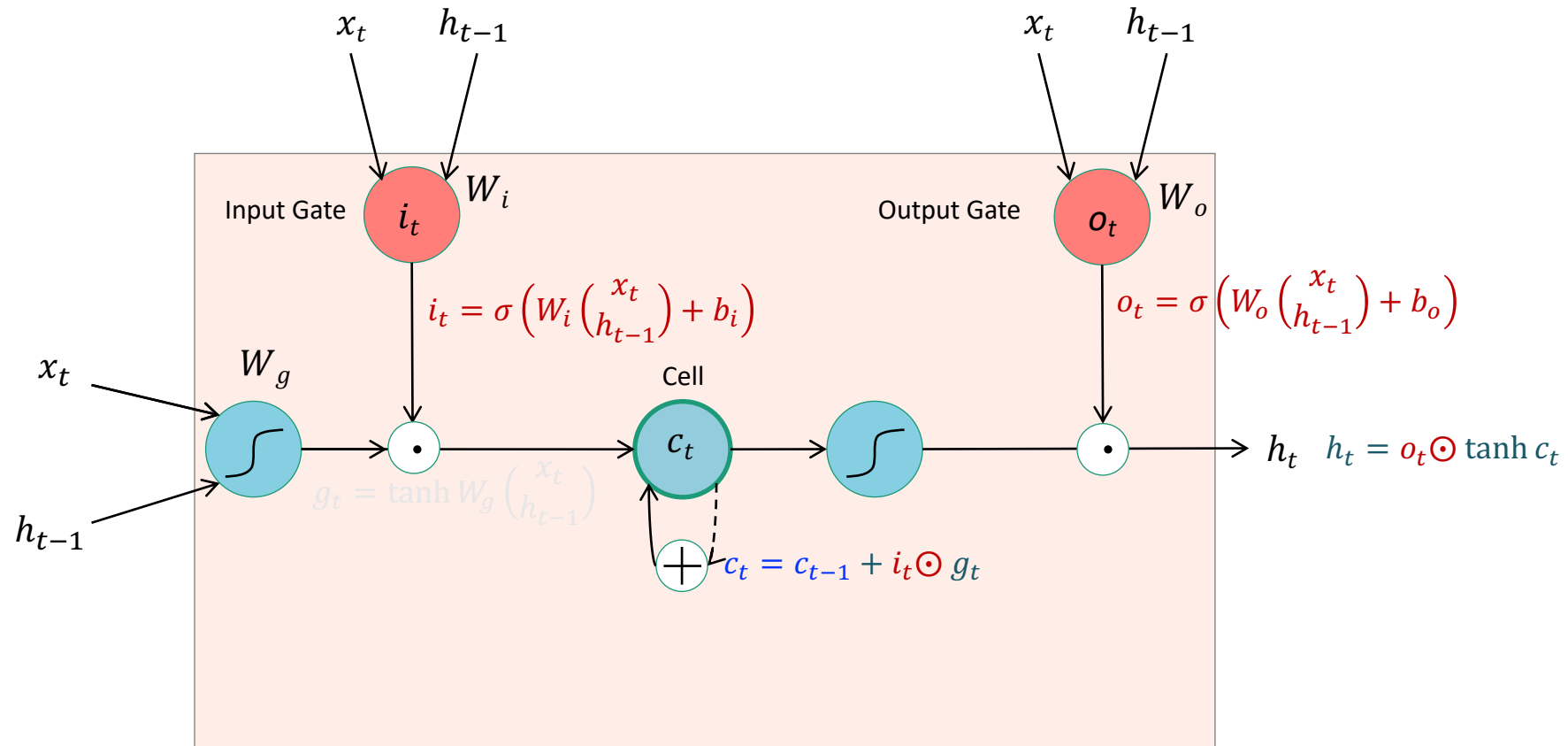




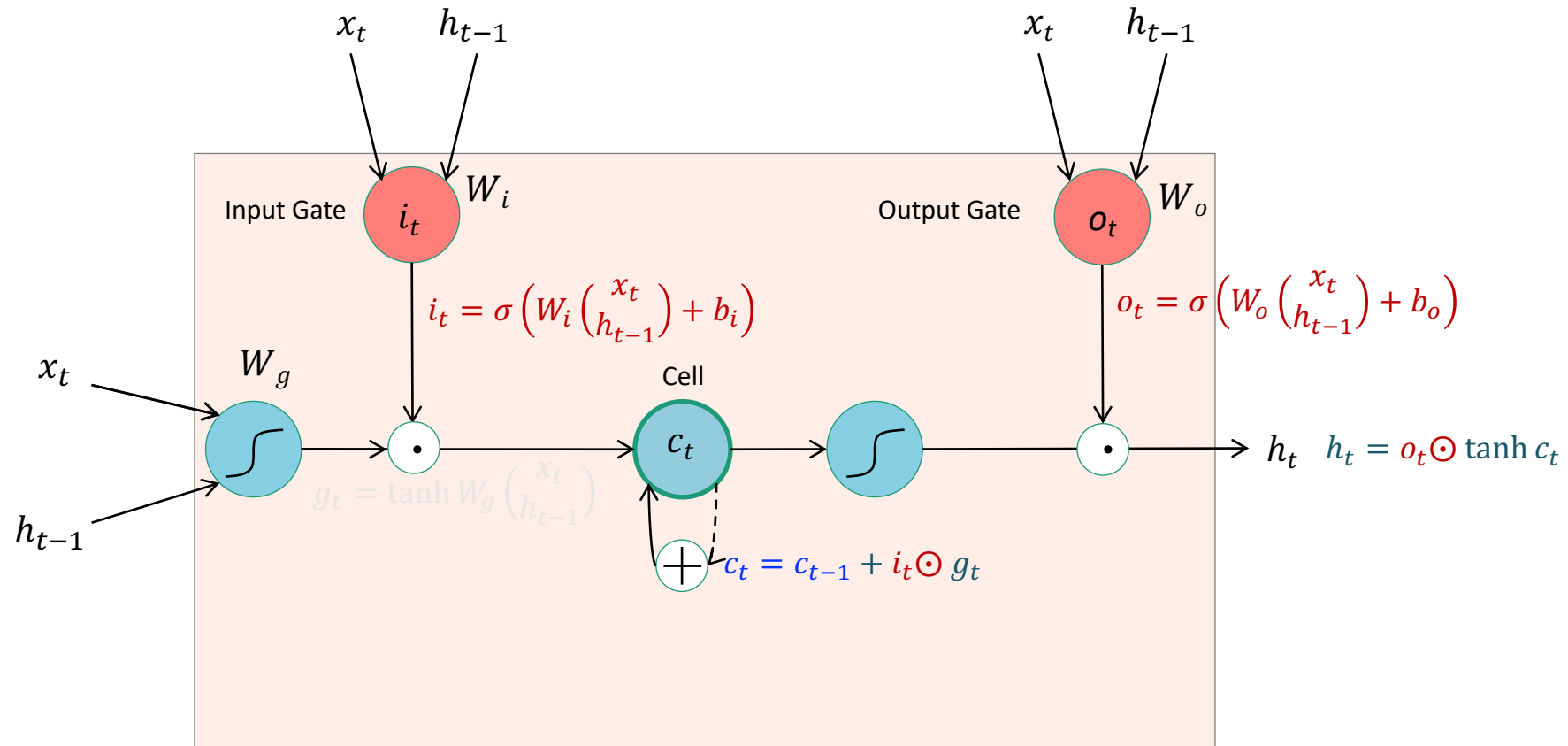
# Long short-term memory (LSTM) cell



# Long short-term memory (LSTM) cell

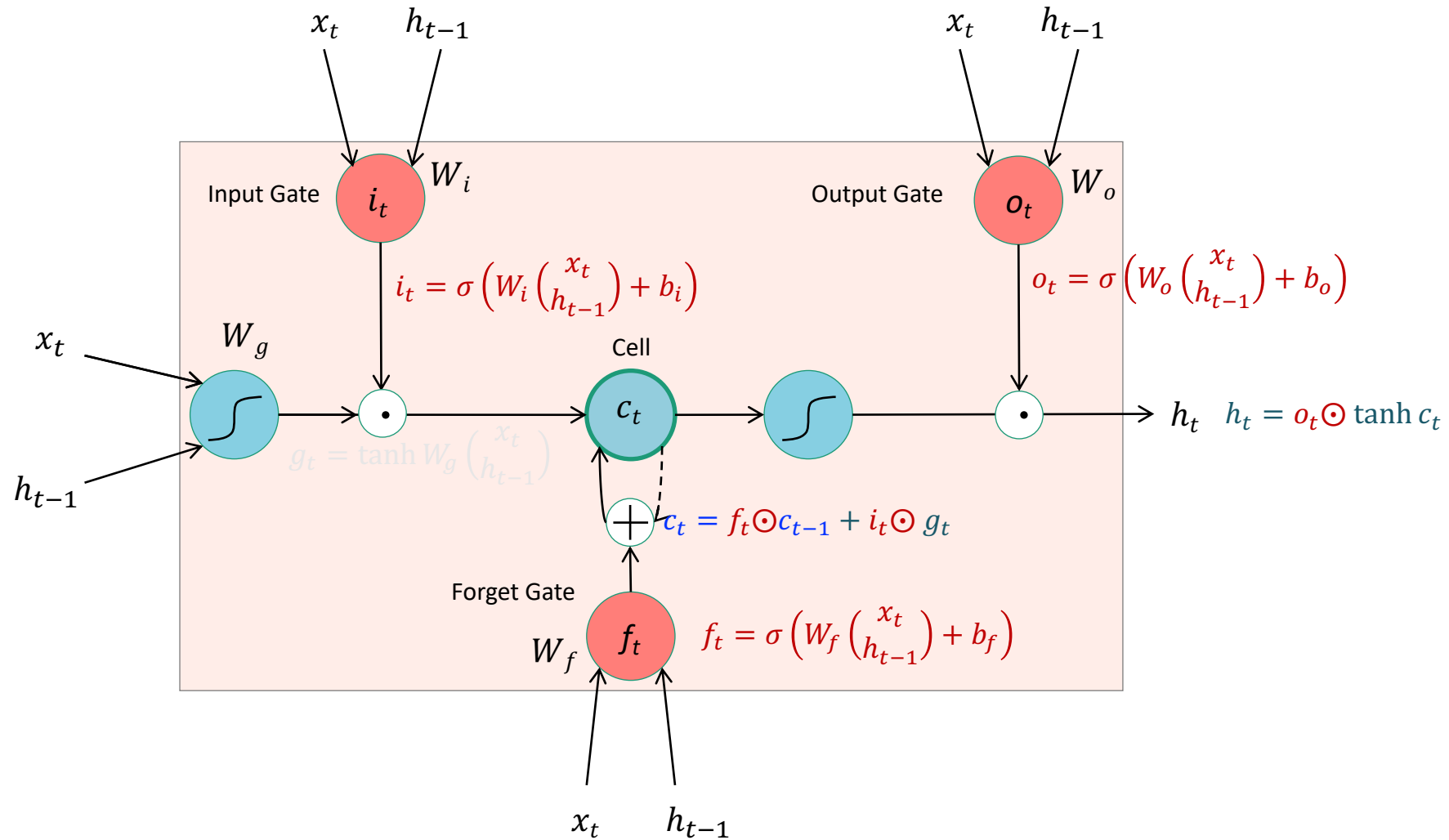


# Long short-term memory (LSTM) cell



The gradients back-propagated from  $c_t$  to  $c_{t-1}$  are maintained. Thus we can do learning for *long-term*

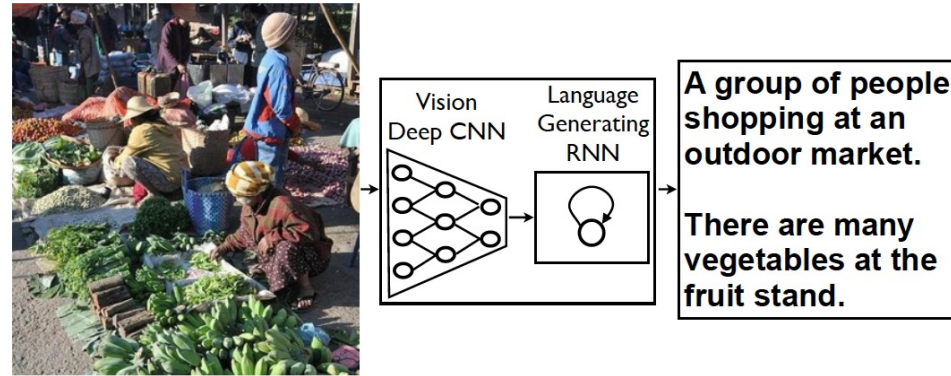
# Adding a forget gate for *short-term*



The gradients back-propagated from  $c_t$  to  $c_{t-1}$  are adjusted by  $f_t$ .

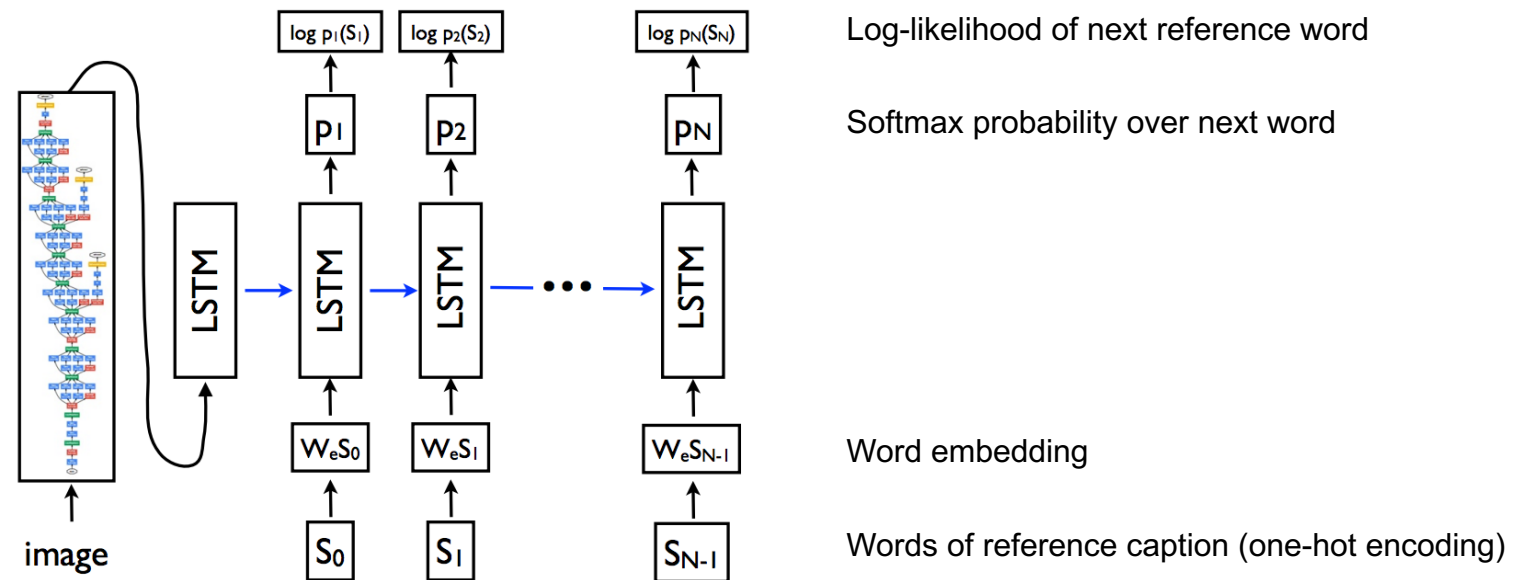
Application in language and vision tasks

# Image caption generation



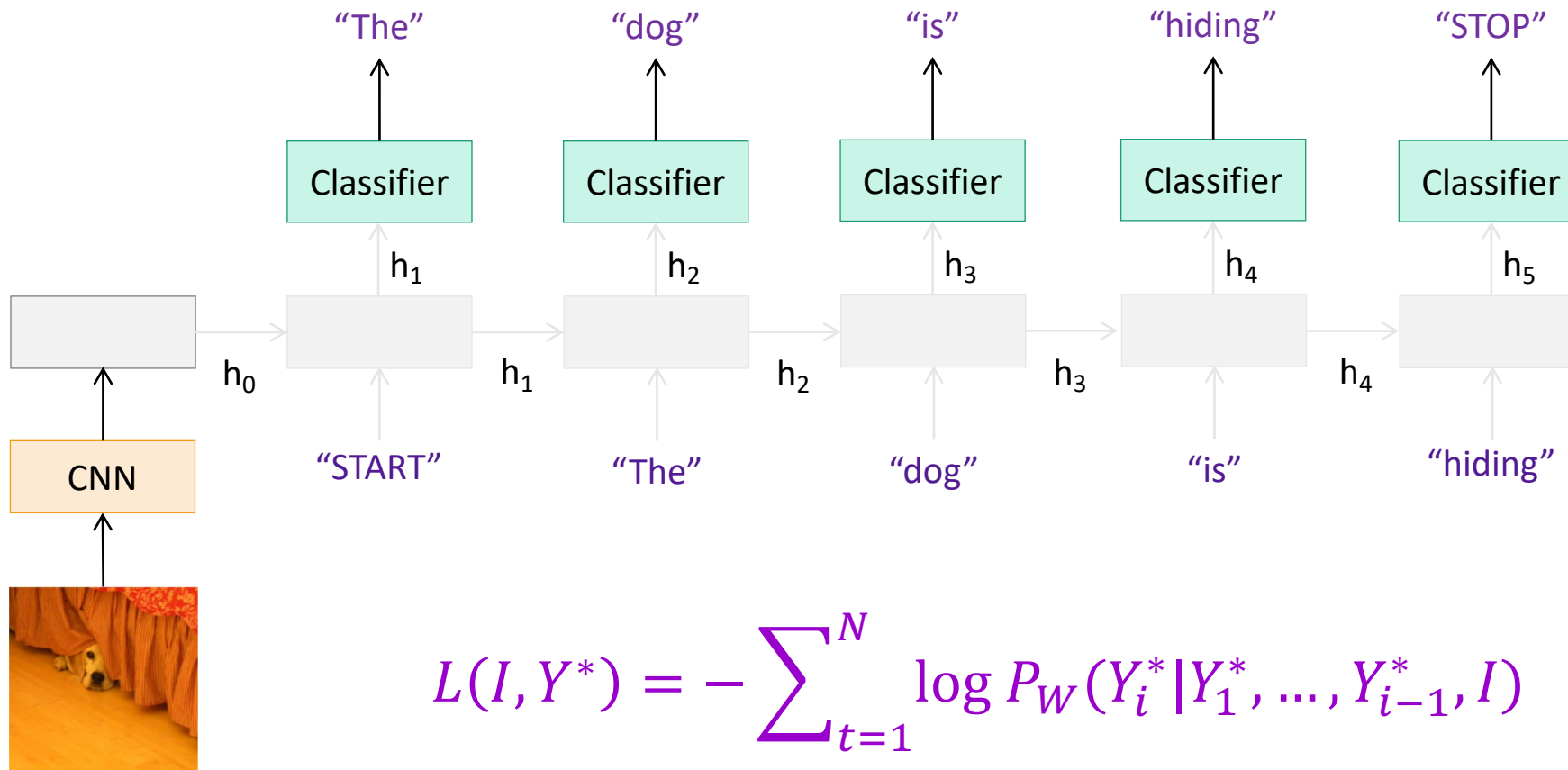
Training time

- Maximize likelihood of reference captions



# Image caption generation

- Minimize negative log-likelihood of the ground truth caption  $Y^* = (Y_1^*, \dots, Y_N^*)$  given image  $I$ :



$$L(I, Y^*) = - \sum_{t=1}^N \log P_W(Y_t^* | Y_1^*, \dots, Y_{t-1}^*, I)$$

# Image caption generation

A person riding a motorcycle on a dirt road.



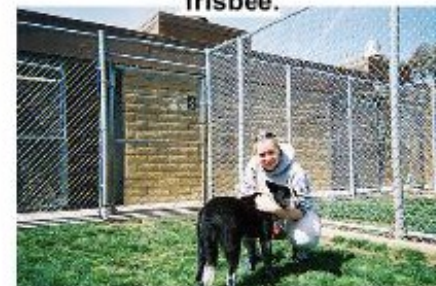
Two dogs play in the grass.



A skateboarder does a trick on a ramp.



A dog is jumping to catch a frisbee.



A group of young people playing a game of frisbee.



Two hockey players are fighting over the puck.



A little girl in a pink hat is blowing bubbles.



A refrigerator filled with lots of food and drinks.



A herd of elephants walking across a dry grass field.



A close up of a cat laying on a couch.



A red motorcycle parked on the side of the road.



A yellow school bus parked in a parking lot.



Describes without errors

Describes with minor errors

Somewhat related to the image

Unrelated to the image



# Visual Question Answering (VQA)



Q: What endangered animal is featured on the truck?

A: **A bald eagle.**

A: A sparrow.

A: A humming bird.

A: A raven.



Q: Where will the driver go if turning right?

A: **Onto 24 3/4 Rd.**

A: Onto 25 3/4 Rd.

A: Onto 23 3/4 Rd.

A: Onto Main Street.



Q: When was the picture taken?

A: **During a wedding.**

A: During a bar mitzvah.

A: During a funeral.

A: During a Sunday church service



Q: Who is under the umbrella?

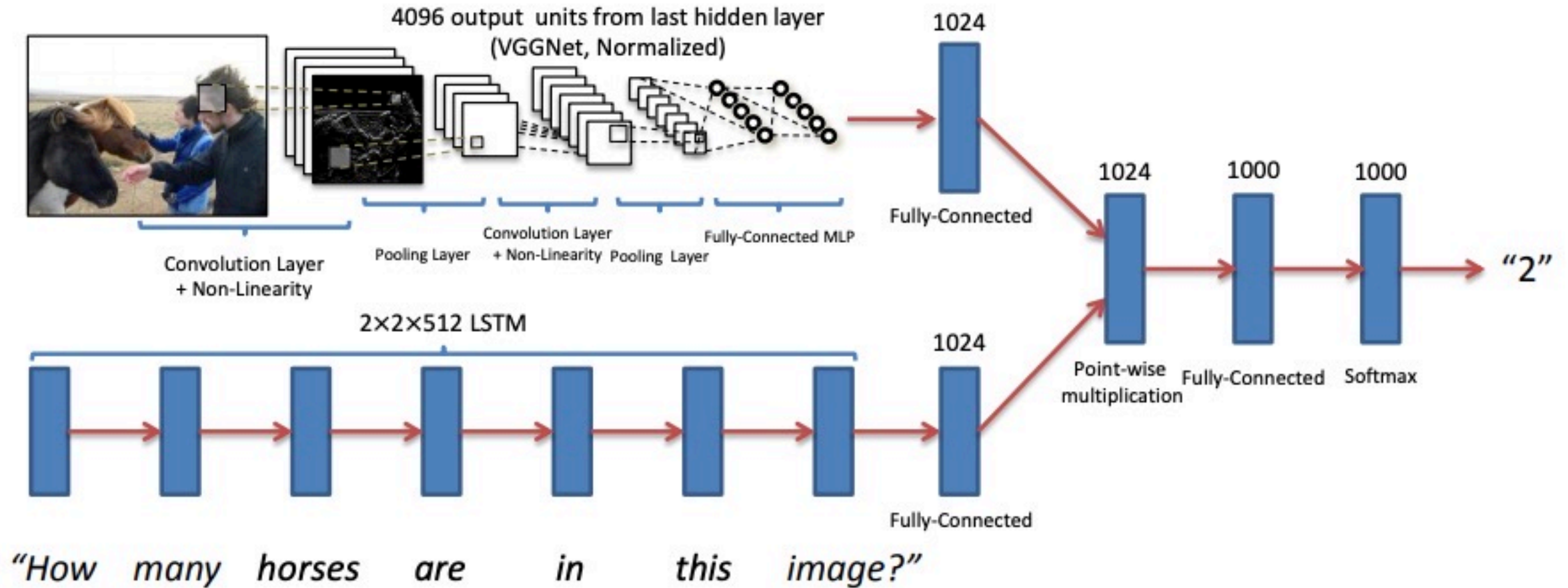
A: **Two women.**

A: A child.

A: An old man.

A: A husband and a wife.

# Visual Question Answering (VQA)



# Visual Language Navigation: Go to the living room


Agent encodes instructions in language and uses an RNN to generate a series of movements as the visual input changes after each move.

## Instruction

Turn right and head towards the *kitchen*. Then turn left, pass a *table* and enter the *hallway*. Walk down the hallway and turn into the *entry way* to your right *without doors*. Stop in front of the *toilet*.

 Initial Position

 Target Position

 Demonstration Path A

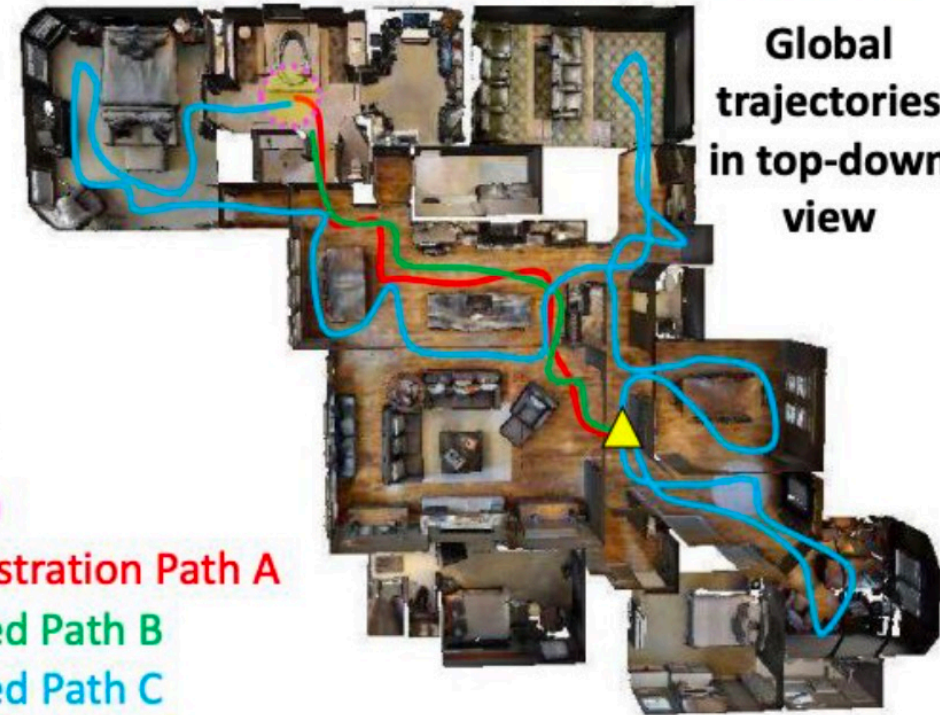
 Executed Path B

 Executed Path C

Local  
visual  
scene



Global  
trajectories  
in top-down  
view



# Summary

- The Basic RNN
- LSTM
- Application in language and vision tasks

# Next Class

Video Understanding

Temporal and 3D Convolution

Visual Prediction