

Bài 9: Mạng hồi quy

Nội dung

1. Bài toán dự đoán chuỗi
2. Mạng hồi quy thông thường
3. Lan truyền ngược theo thời gian (BPTT)
4. Mạng LSTM và GRU
5. Một số áp dụng

Bài toán dự đoán chuỗi

Bài toán dự đoán chuỗi

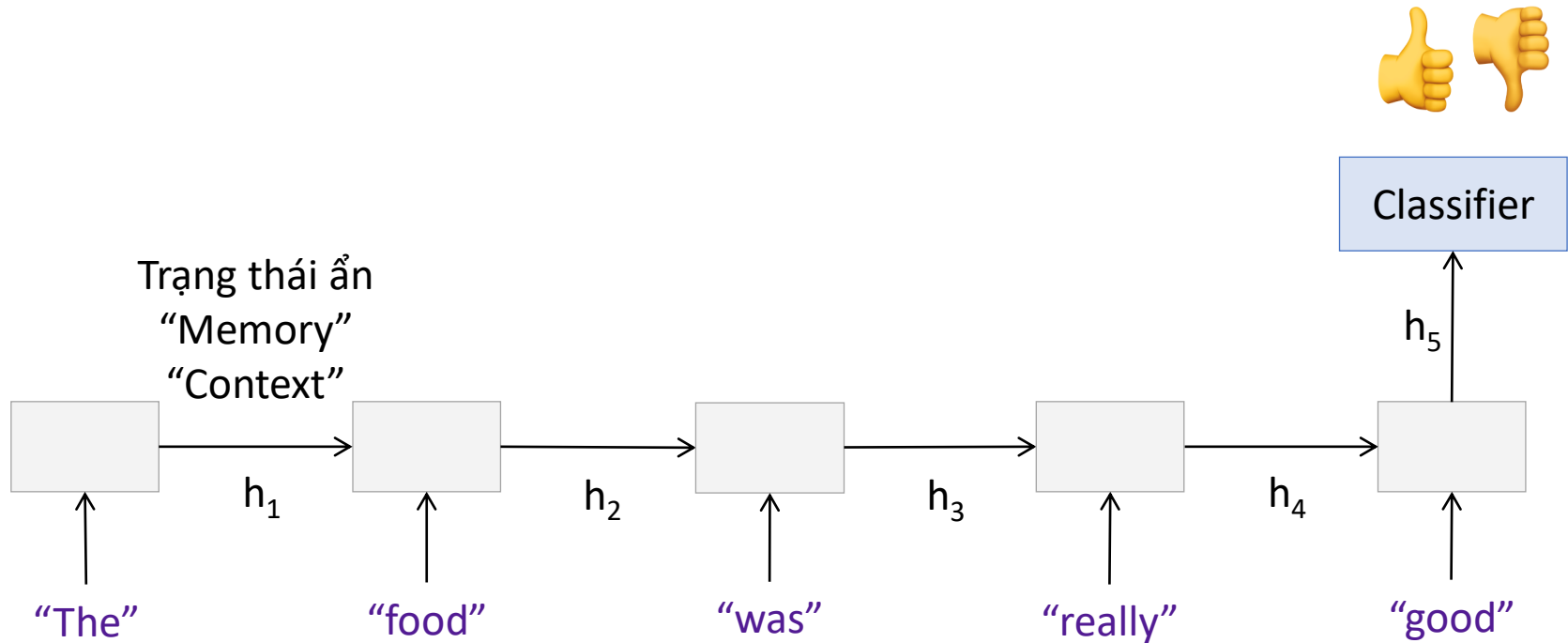
- Trước giờ, ta chỉ tập trung vào vấn đề dự đoán với đầu vào và đầu ra kích thước cố định
- Chuyện gì sẽ xảy ra nếu đầu vào và đầu ra là một chuỗi có kích thước thay đổi?

Phân lớp văn bản

- Phân loại sắc thái (sentiment): phân loại bình luận một nhà hàng hay một bộ phim hay một sản phẩm là tích cực hay tiêu cực
 - “The food was really good” - “Thức ăn rất ngon”
 - “Máy hút bụi bị hỏng trong vòng hai tuần”
 - “Bộ phim có những phần buồn tẻ, nhưng tổng thể là rất đáng xem”
- Cần dùng đặc trưng gì và mô hình phân loại gì để giải quyết bài toán này?

Phân loại sắc thái

- “The food was really good”



Recurrent Neural Network (RNN)

Mô hình ngôn ngữ



RNN Bible
@RNN_Bible

Random bible verses generated using Recurrent Neural Networks (char-rnn).

Joined May 2015

Tweets **2,197** Following **1** Followers **485**

Tweets **Tweets & replies**

RNN Bible @RNN_Bible · 20 Jun 2016
24:11 Thus saith the LORD of hosts; Ask now this stones are for the righteous and the children of Israel.
1 2 3

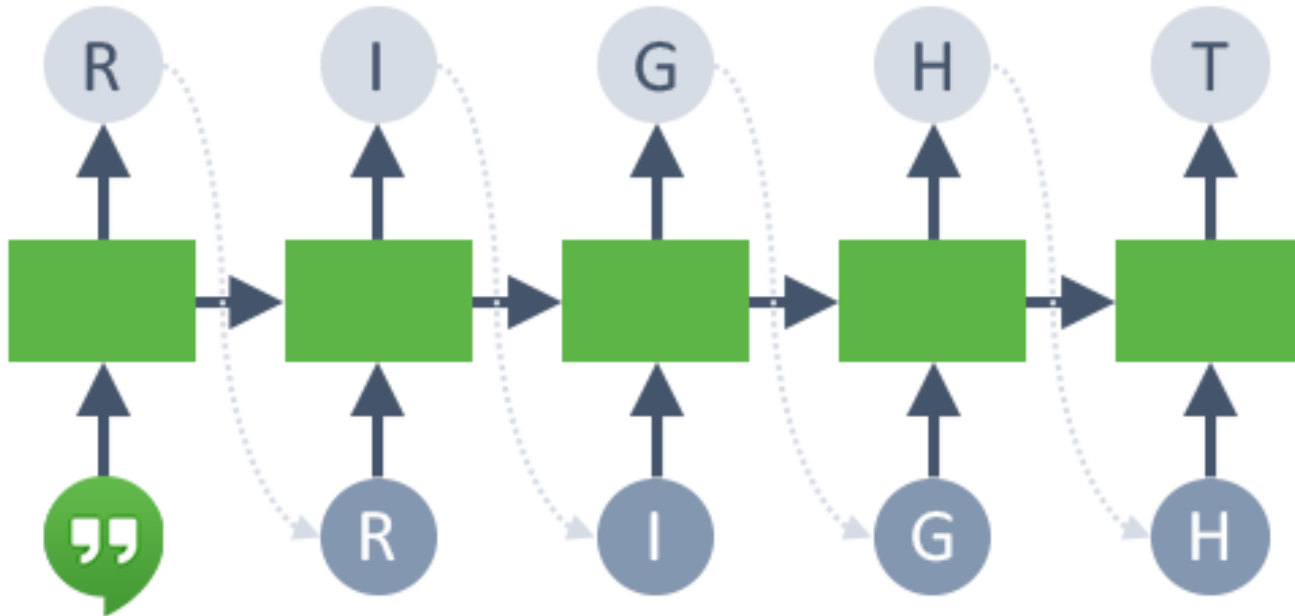
RNN Bible @RNN_Bible · 19 Jun 2016
24:16 And they took up twelve stones out of the city of David, and discomfit Jordan.
1

RNN Bible @RNN_Bible · 19 Jun 2016
3:20 And the LORD shall send a proverb against the LORD thy God, and shalt not each laugh.
1 5 3

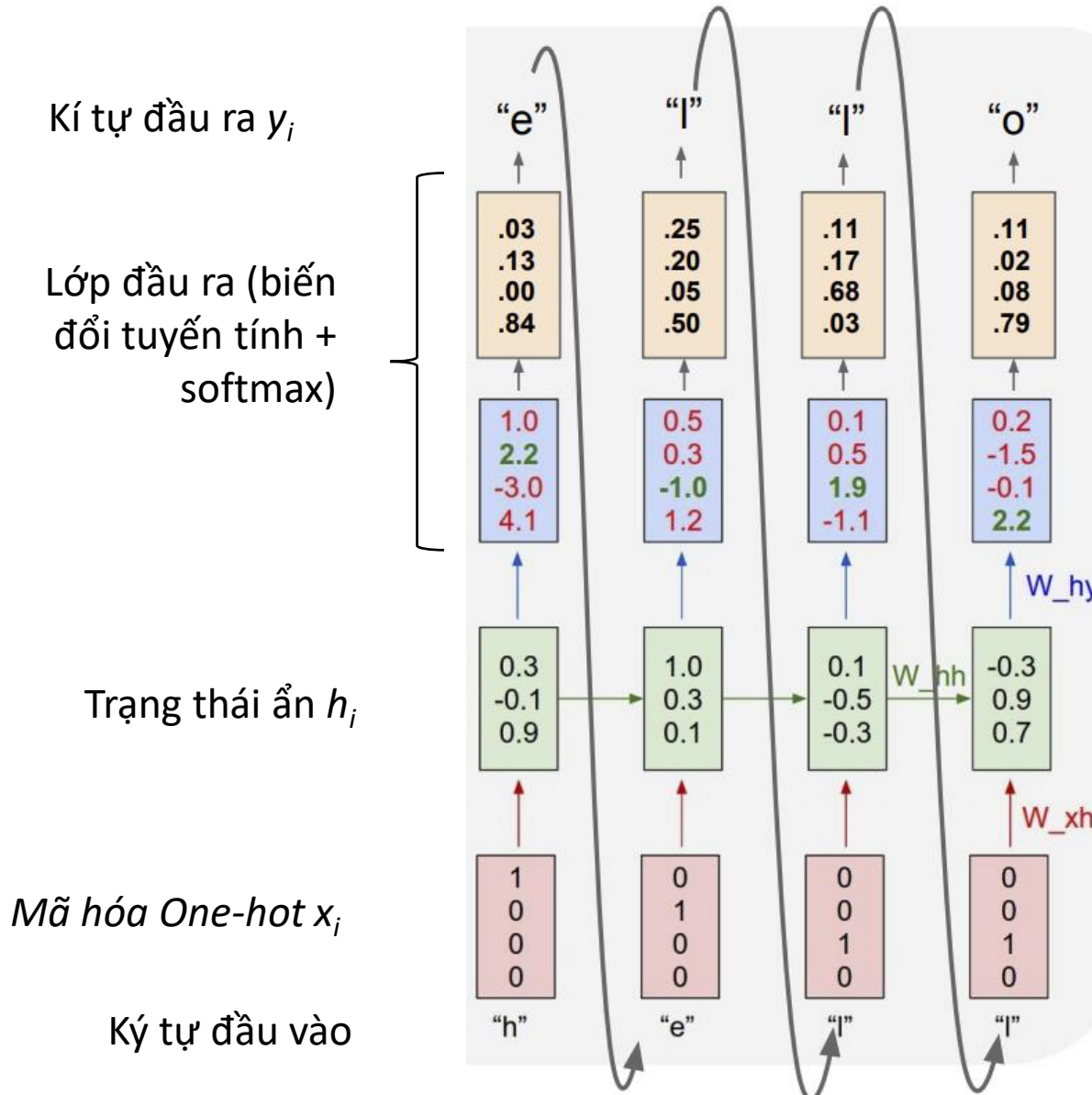
RNN Bible @RNN_Bible · 19 Jun 2016
23:2 And the vision of the breaking thereof shall be in rubbick, and they shall take away the stones out of the land.
1

Mô hình ngôn ngữ

- Character RNN



Character RNN



$$\begin{aligned}
 & p(y_1, y_2, \dots, y_n) \\
 &= \prod_{i=1}^n p(y_i | y_1, \dots, y_{i-1}) \\
 &\approx \prod_{i=1}^n P_W(y_i | h_i)
 \end{aligned}$$

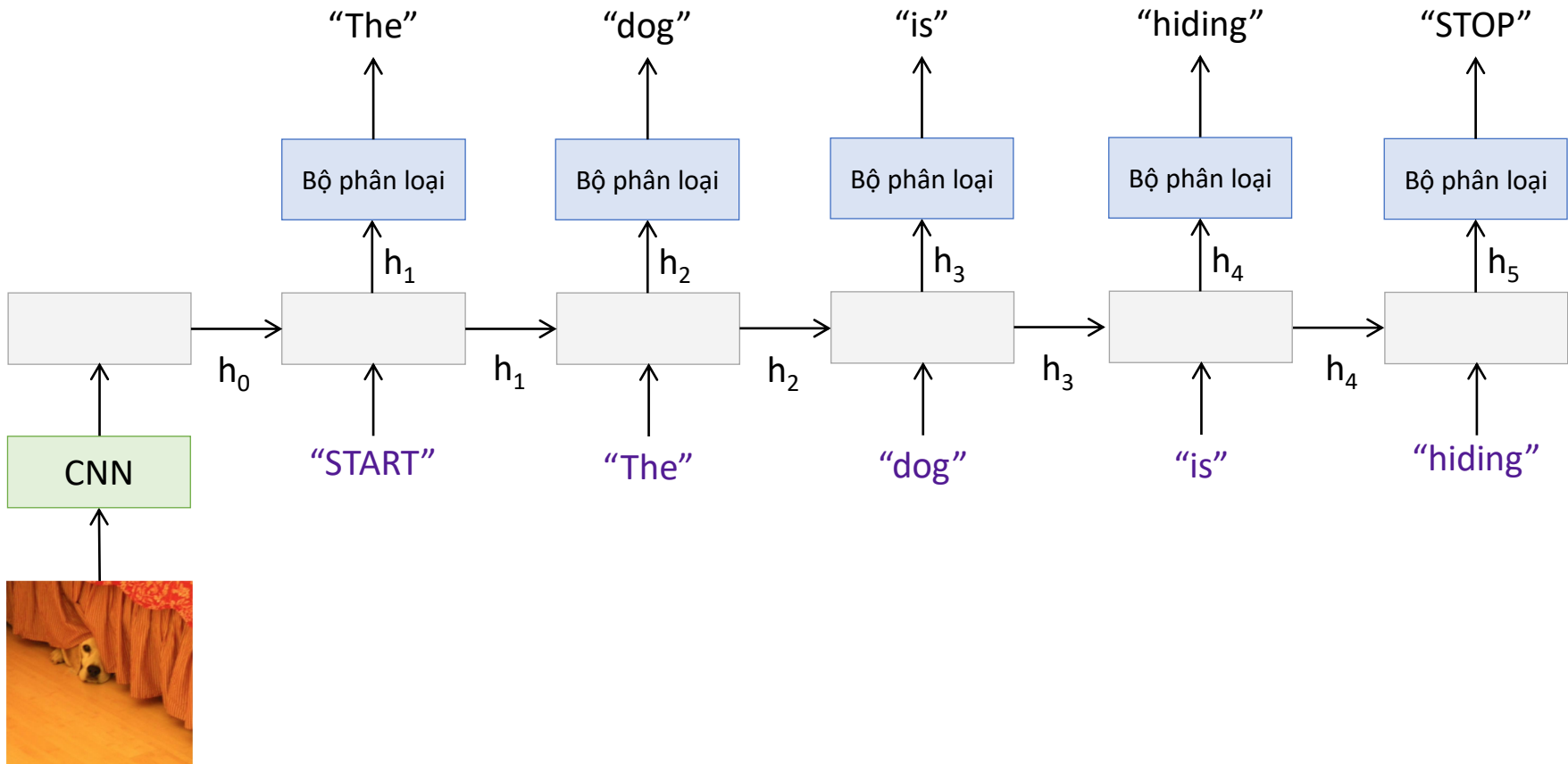
Sinh mô tả bức ảnh

- Cho một bức ảnh, cần sinh ra một câu mô tả nội dung bức ảnh



“The dog is hiding”

Sinh mô tả bức ảnh



Dịch máy



Translate

Turn off instant translation



Google

En

Correspondances

La Nature est un temple où de vivants piliers
Laissent 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

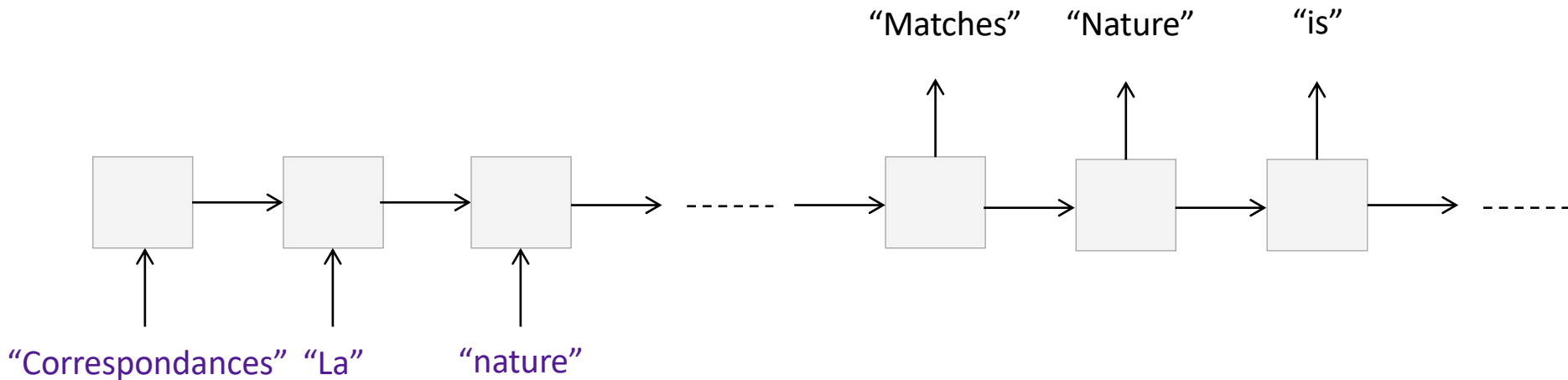


693/5000

<https://translate.google.com/>

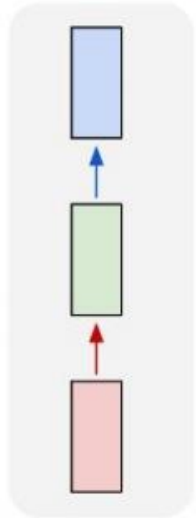
Dịch máy

- Nhiều đầu vào – nhiều đầu ra (hay còn gọi là sequence to sequence)



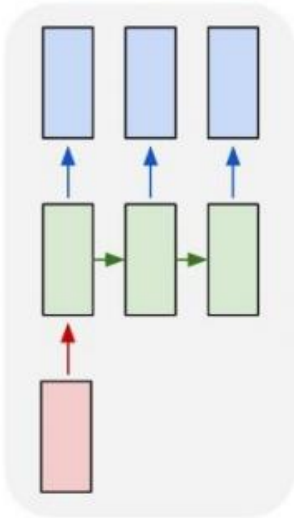
Tổng hợp các loại dự đoán

one to one



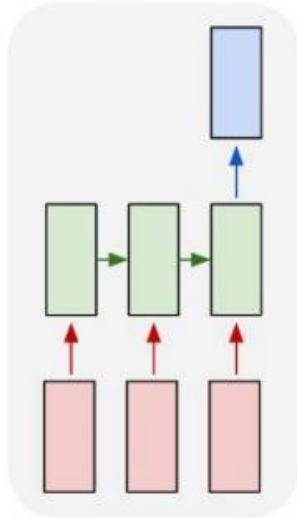
Phân
lớp
ảnh

one to many



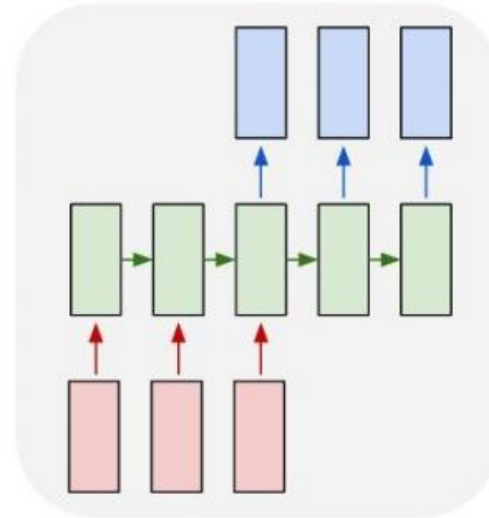
Sinh mô
tả ảnh

many to one



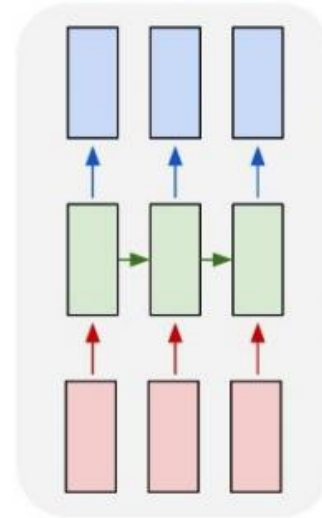
Phân
loại sắc
thái câu

many to many



Dịch máy

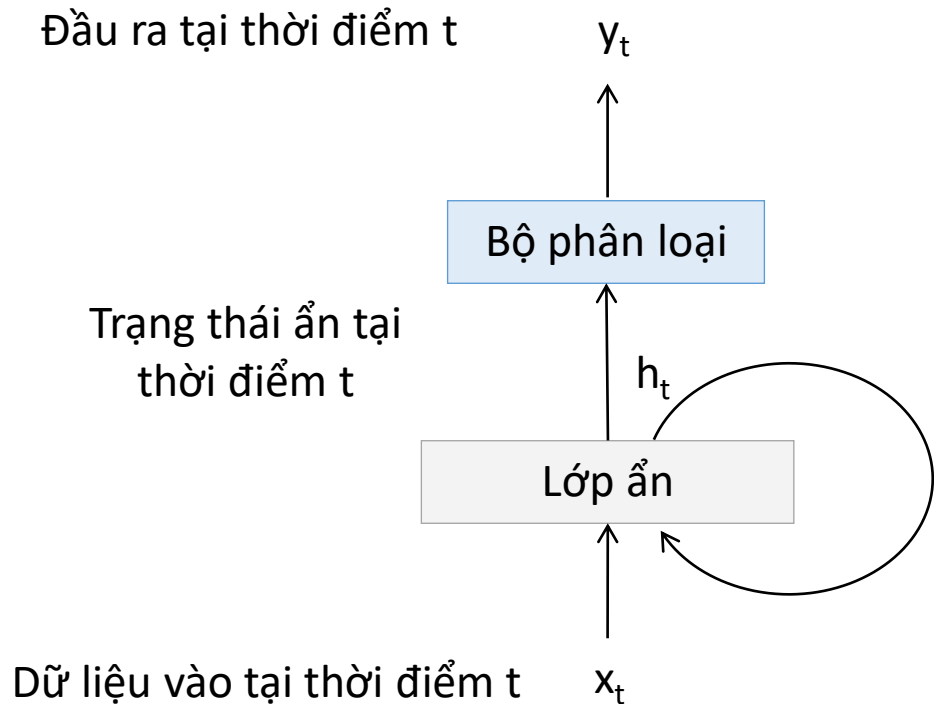
many to many



Phân loại
video
mức
frame

Mạng hồi quy thông thường

Mạng hồi quy Recurrent Neural Network (RNN)

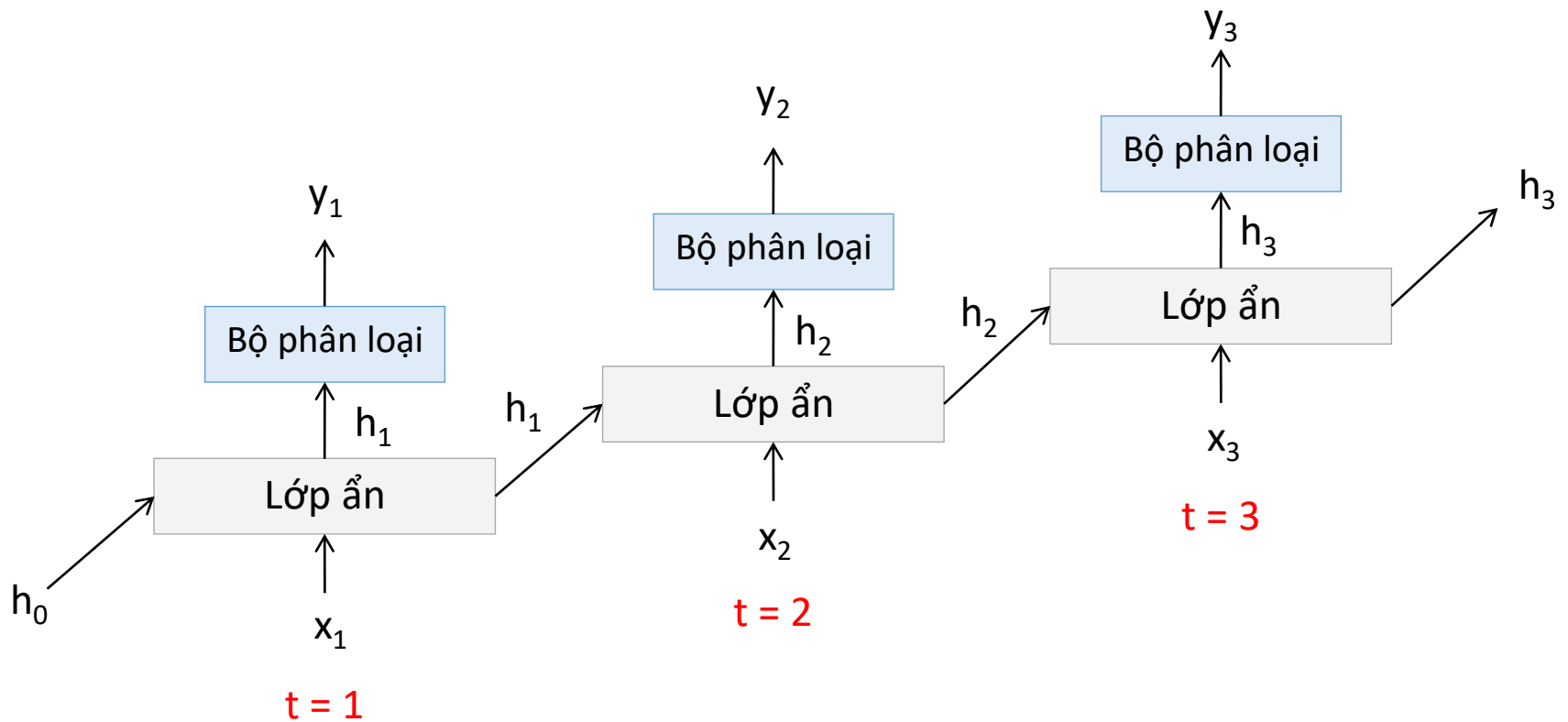


Hồi quy:

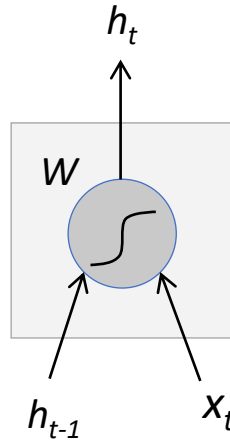
$$h_t = f_W(x_t, h_{t-1})$$

new state function of W input at time t old state

Duỗi (unroll) RNN



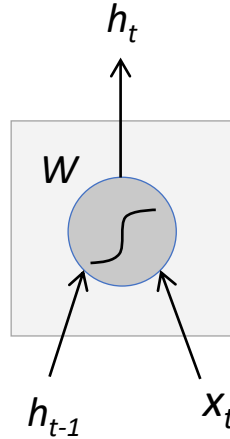
RNN thông thường



$$\begin{aligned} h_t &= f_W(x_t, h_{t-1}) \\ &= \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \end{aligned}$$

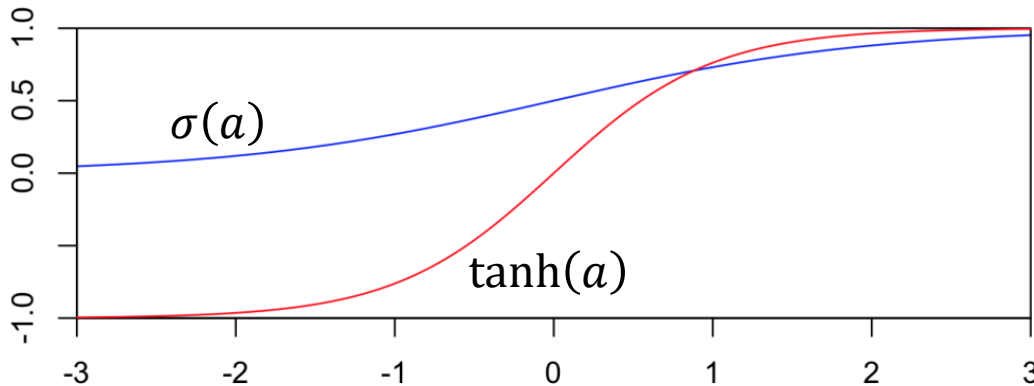
J. Elman, [Finding structure in time](#), Cognitive science 14(2), pp. 179–211, 1990

RNN thông thường



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

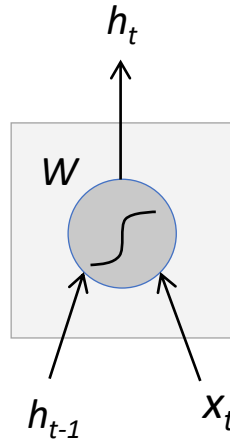
$$= \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$



$$\tanh(a) = \frac{e^a - e^{-a}}{e^a + e^{-a}}$$

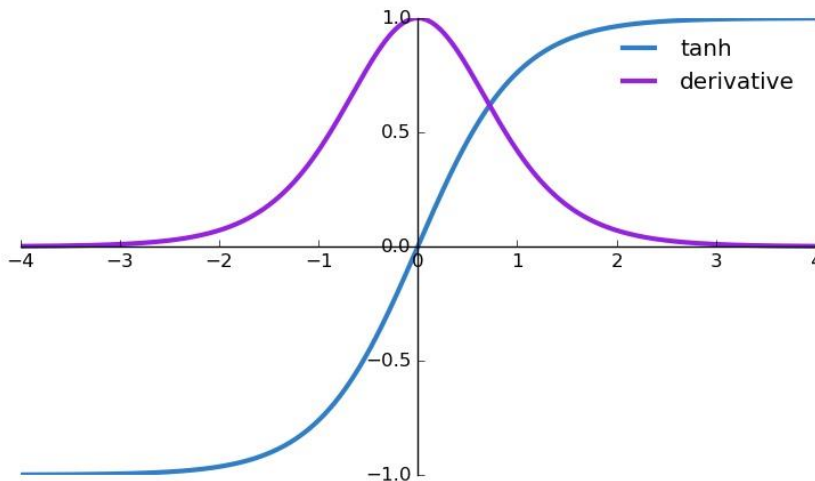
$$= 2\sigma(2a) - 1$$

RNN thông thường



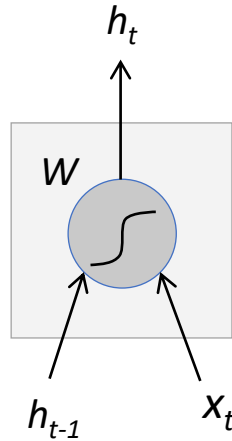
$$h_t = f_W(x_t, h_{t-1})$$

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

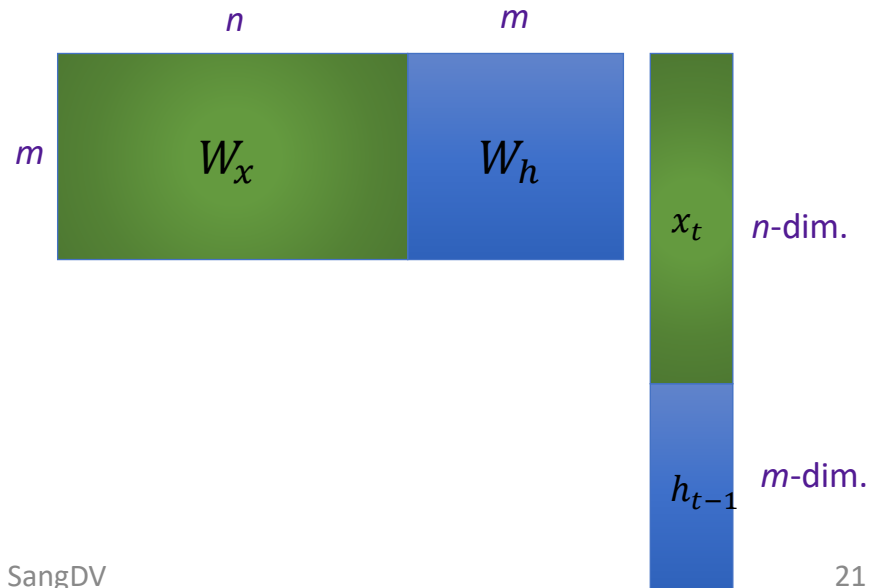


$$\frac{d}{da} \tanh(a) = 1 - \tanh^2(a)$$

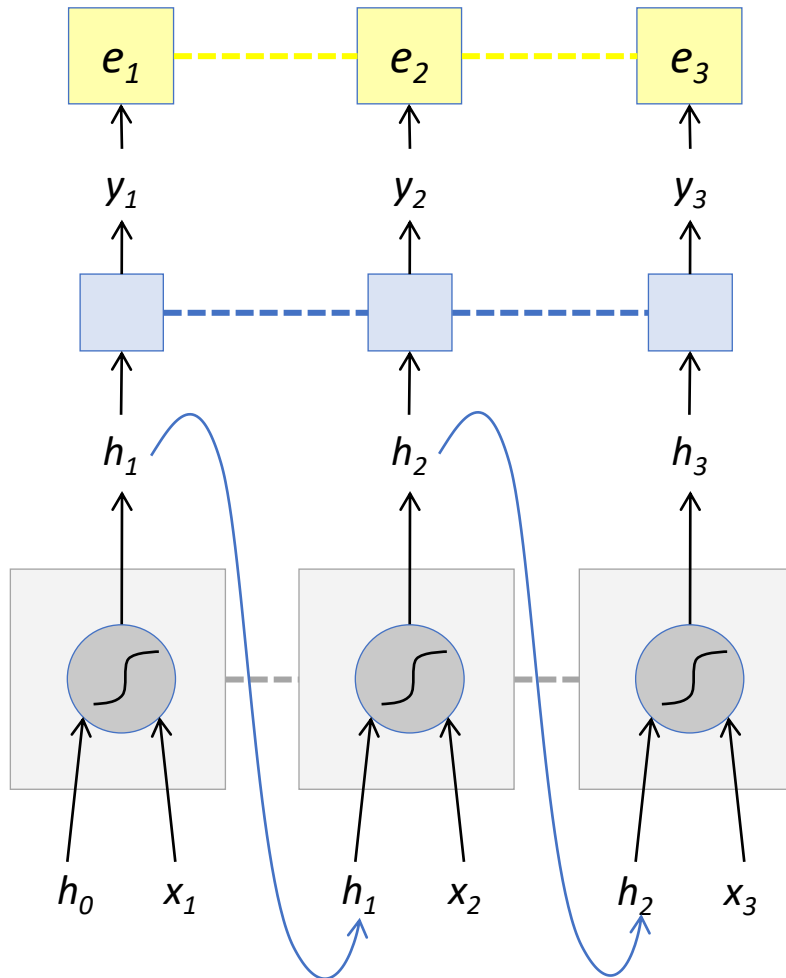
RNN thông thường



$$\begin{aligned}h_t &= f_W(x_t, h_{t-1}) \\&= \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \\&= \tanh(W_x x_t + W_h h_{t-1})\end{aligned}$$



RNN Forward Pass



$$e_t = -\log(y_t(GT_t))$$

$$y_t = \text{softmax}(W_y h_t)$$

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

----- Dùng chung trọng số
(shared weights)

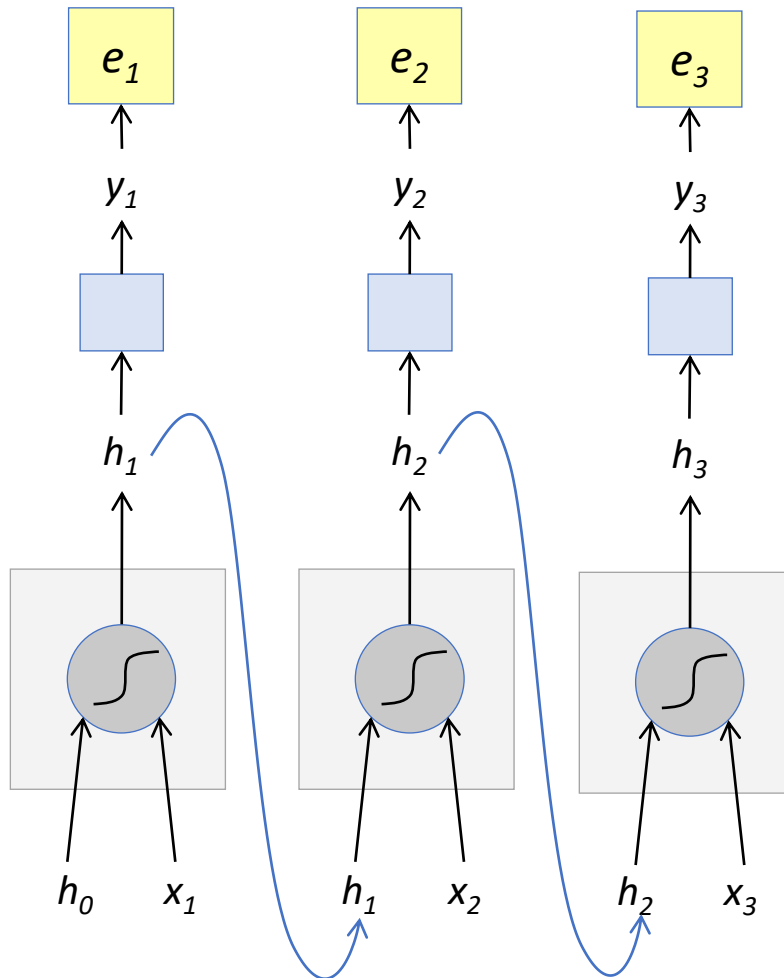
Lan truyền ngược theo thời gian (BPTT)

Lan truyền ngược theo thời gian (BPTT)



- Đây là phương pháp thông dụng nhất để huấn luyện RNNs
- Mạng sau khi duỗi được xem như một mạng nơ-ron *feed-forward* lớn nhận dữ liệu đầu vào là cả chuỗi dữ liệu
- Gradient đối với một trọng số mạng RNN được tính tại mỗi bản sao của nó trong mạng duỗi (unfolded network), sau đó được cộng lại (hoặc tính trung bình) và được sử dụng để cập nhật trọng số mạng.

Tính toán tiến (forward pass) mạng RNN chuỗi

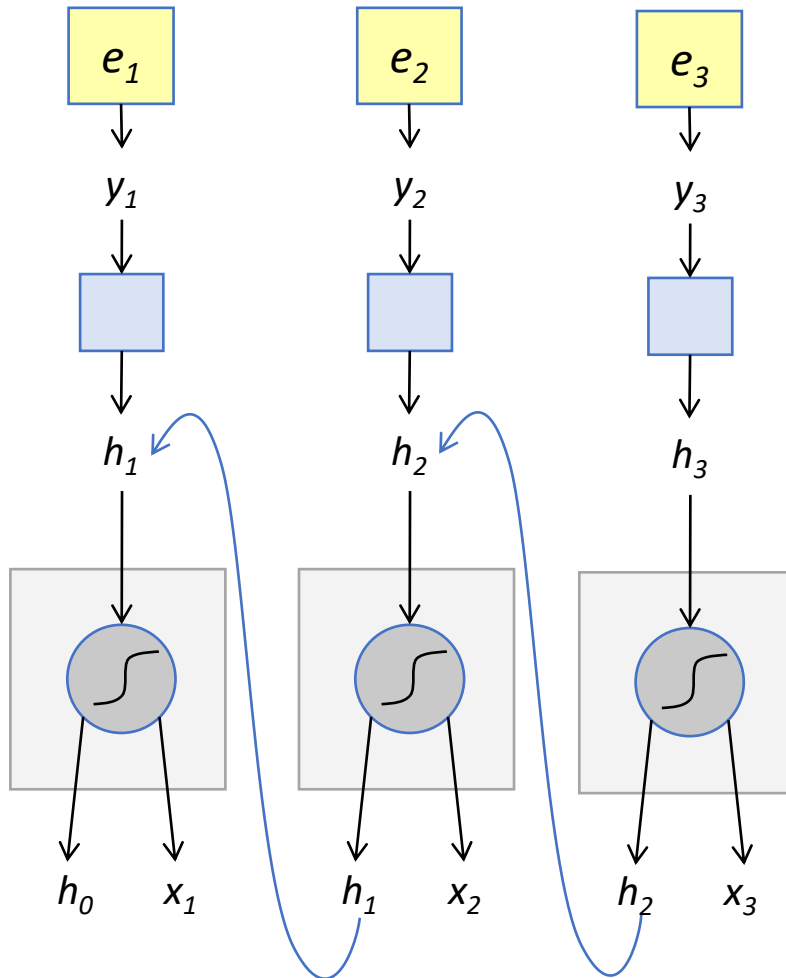


$$e_t = -\log(y_t(GT_t))$$

$$y_t = \text{softmax}(W_y h_t)$$

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

Tính toán tiến (forward pass) mạng RNN duỗi



$$e_t = -\log(y_t(GT_t))$$

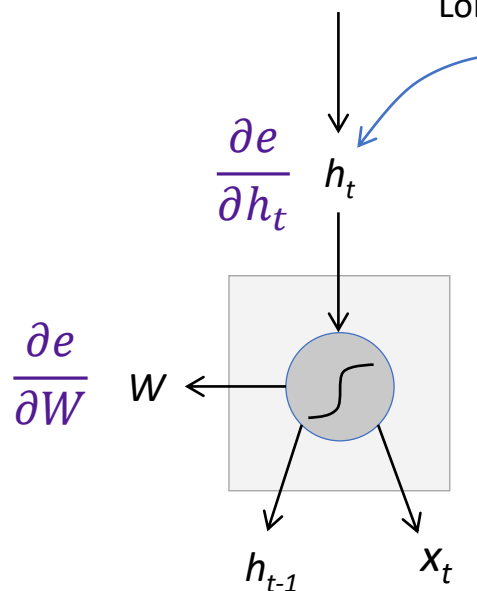
$$y_t = \text{softmax}(W_y h_t)$$

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

Lan truyền ngược mạng RNN

Lỗi từ y_t

Lỗi từ dự đoán ở các bước tương lai



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

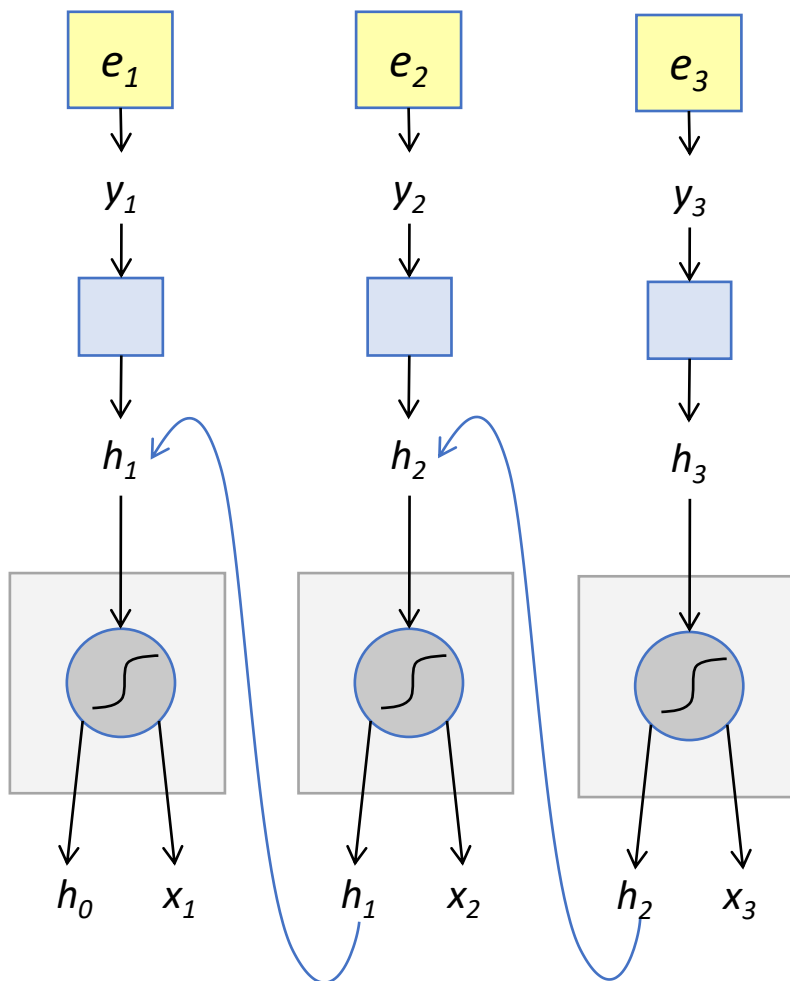
$$\frac{\partial e}{\partial W_h} = \frac{\partial e}{\partial h_t} \odot (1 - \tanh^2(W_x x_t + W_h h_{t-1})) h_{t-1}^T$$

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

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

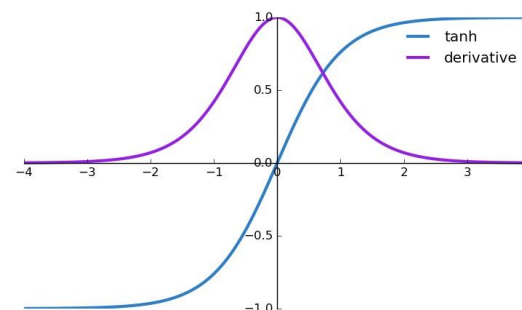
Lan truyền ngược tới các bước sớm hơn

Lan truyền ngược mạng RNN



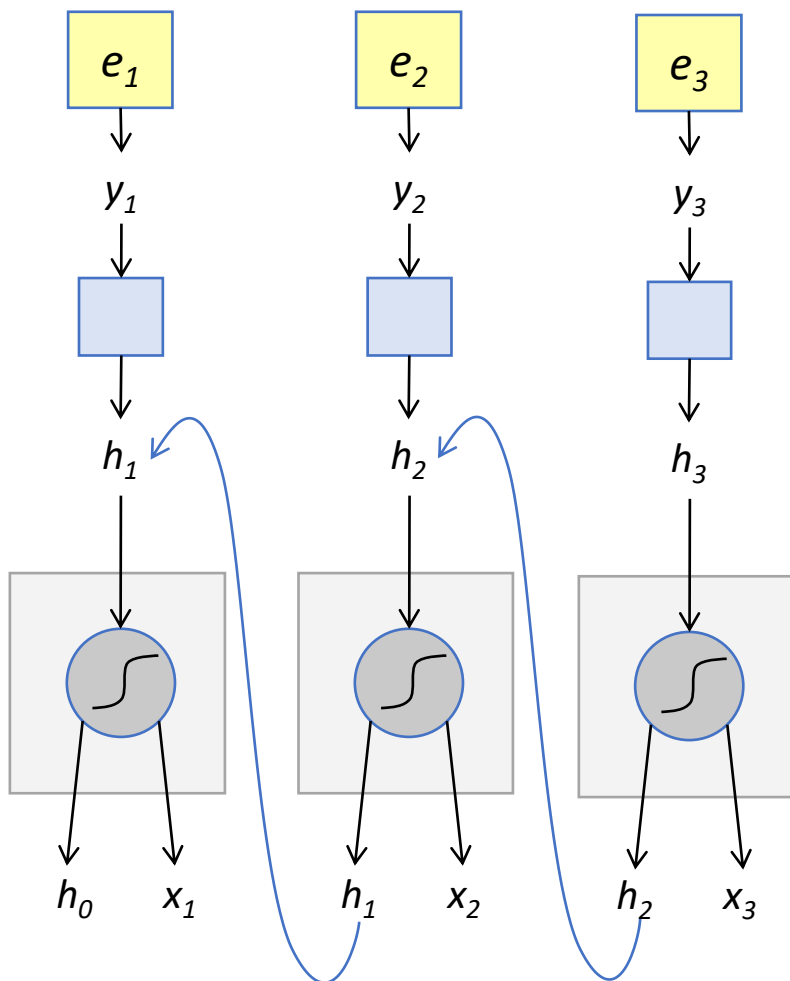
$$\frac{\partial e}{\partial h_{t-1}} = W_h^T (1 - \tanh^2(W_x x_t + W_h h_{t-1})) \odot \frac{\partial e}{\partial h_t}$$

Giá trị hàm tanh lớn sẽ tương ứng với gradient nhỏ (vùng bão hòa)



Xét $\frac{\partial e_n}{\partial h_k}$ với $k \ll n$

Lan truyền ngược mạng RNN



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

Gradient sẽ triệt tiêu nếu
giá trị riêng lớn nhất của
 W_h nhỏ hơn 1

Xét $\frac{\partial e_n}{\partial h_k}$ với $k \ll n$

Chi tiết xem tại khóa cs224n

- Recall: $\mathbf{h}^{(t)} = \sigma \left(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1 \right)$
- What if σ were the identity function, $\sigma(x) = x$?

$$\begin{aligned} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} &= \text{diag} \left(\sigma' \left(\mathbf{W}_h \mathbf{h}^{(t-1)} + \mathbf{W}_x \mathbf{x}^{(t)} + \mathbf{b}_1 \right) \right) \mathbf{W}_h && \text{(chain rule)} \\ &= \mathbf{I} \mathbf{W}_h = \mathbf{W}_h \end{aligned}$$

- Consider the gradient of the loss $J^{(i)}(\theta)$ on step i , with respect to the hidden state $\mathbf{h}^{(j)}$ on some previous step j . Let $\ell = i - j$

$$\begin{aligned} \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(j)}} &= \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \prod_{j < t \leq i} \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} && \text{(chain rule)} \\ &= \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \prod_{j < t \leq i} \mathbf{W}_h = \frac{\partial J^{(i)}(\theta)}{\partial \mathbf{h}^{(i)}} \mathbf{W}_h^\ell && \text{(value of } \frac{\partial \mathbf{h}^{(t)}}{\partial \mathbf{h}^{(t-1)}} \text{)} \end{aligned}$$

If \mathbf{W}_h is “small”, then this term gets exponentially problematic as ℓ becomes large


Chi tiết xem tại khóa cs224n

- What's wrong with W_h^ℓ ?
- Consider if the eigenvalues of W_h are all less than 1:

$\lambda_1, \lambda_2, \dots, \lambda_n < 1$
 q_1, q_2, \dots, q_n (eigenvectors)

sufficient but
not necessary
- We can write $\frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} W_h^\ell$ using the eigenvectors of W_h as a basis:

$$\frac{\partial J^{(i)}(\theta)}{\partial h^{(i)}} W_h^\ell = \sum_{i=1}^n c_i \lambda_i^\ell q_i \approx 0 \text{ (for large } \ell)$$

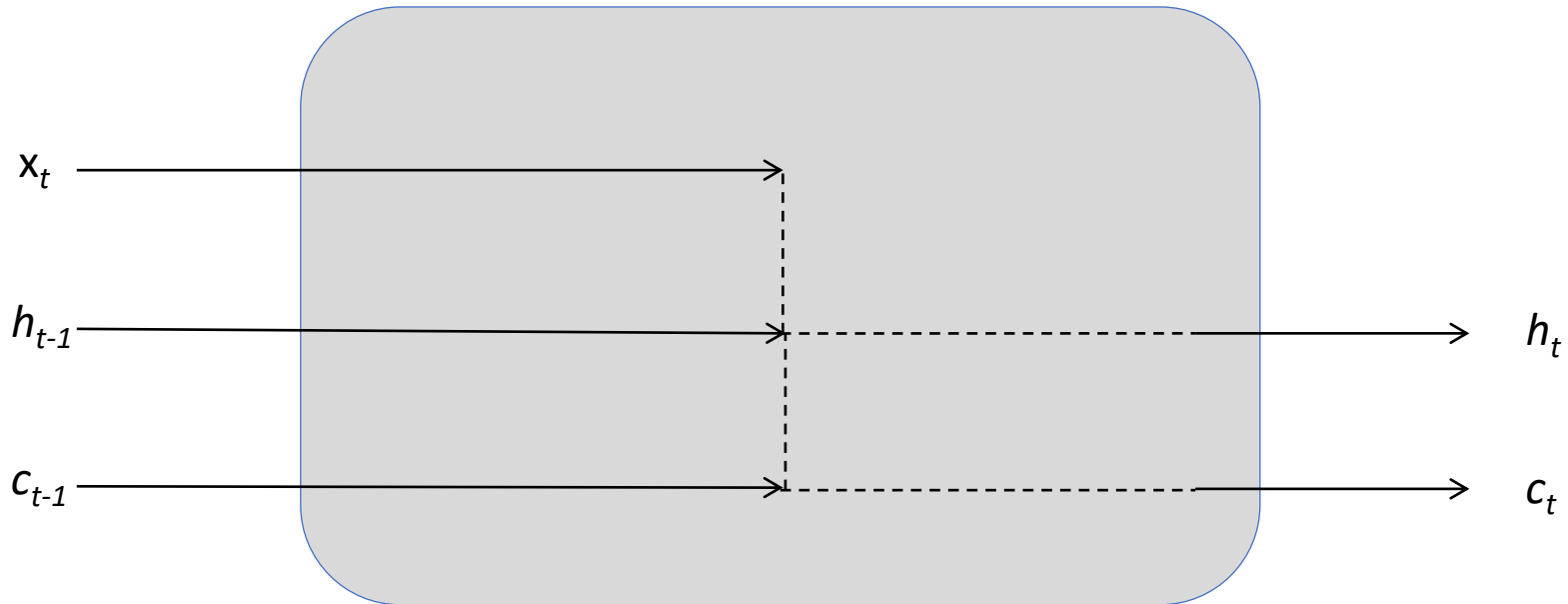


 Approaches 0 as ℓ grows
 so gradient vanishes
- What about nonlinear activations σ (i.e., what we use?)
 - Pretty much the same thing, except the proof requires $\lambda_i < \gamma$ for some γ dependent on dimensionality and σ

Mạng LSTM và GRU

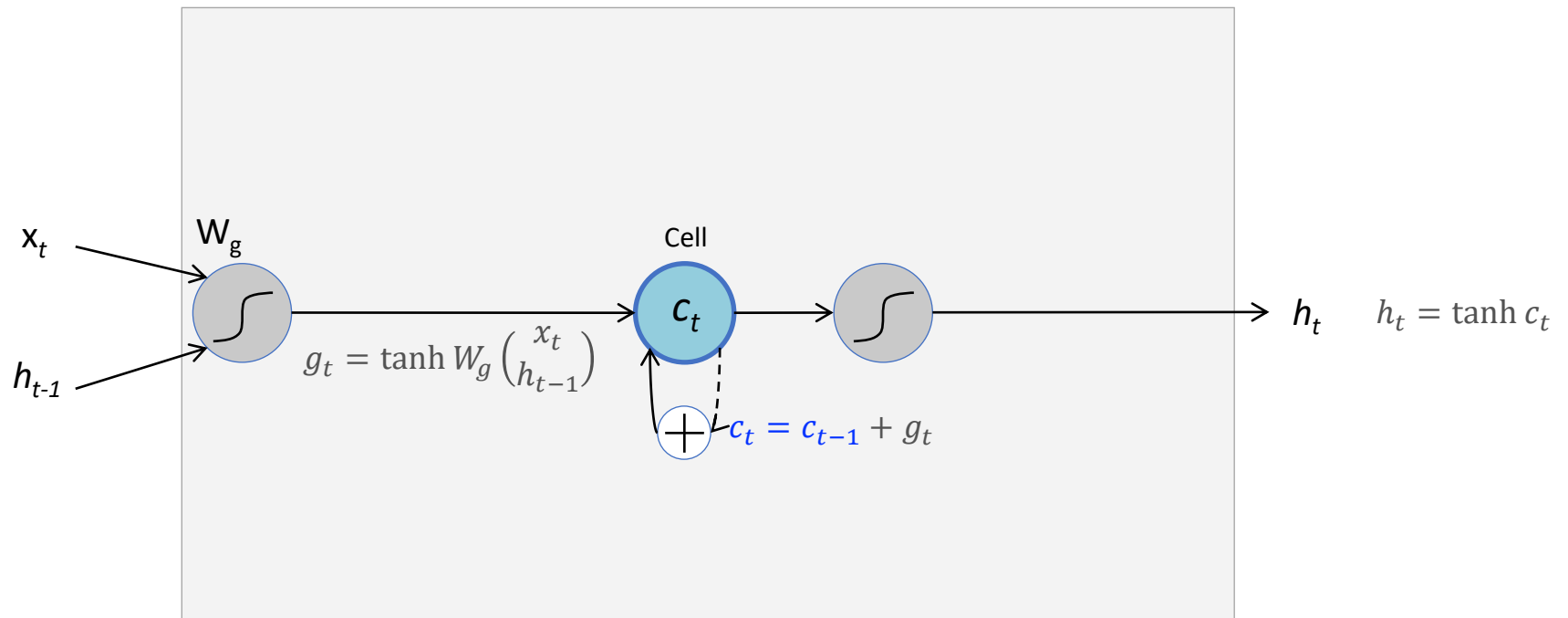
Long Short-Term Memory (LSTM)

- Sử dụng thêm “cell” có bộ nhớ để tránh hiện tượng triệt tiêu gradient

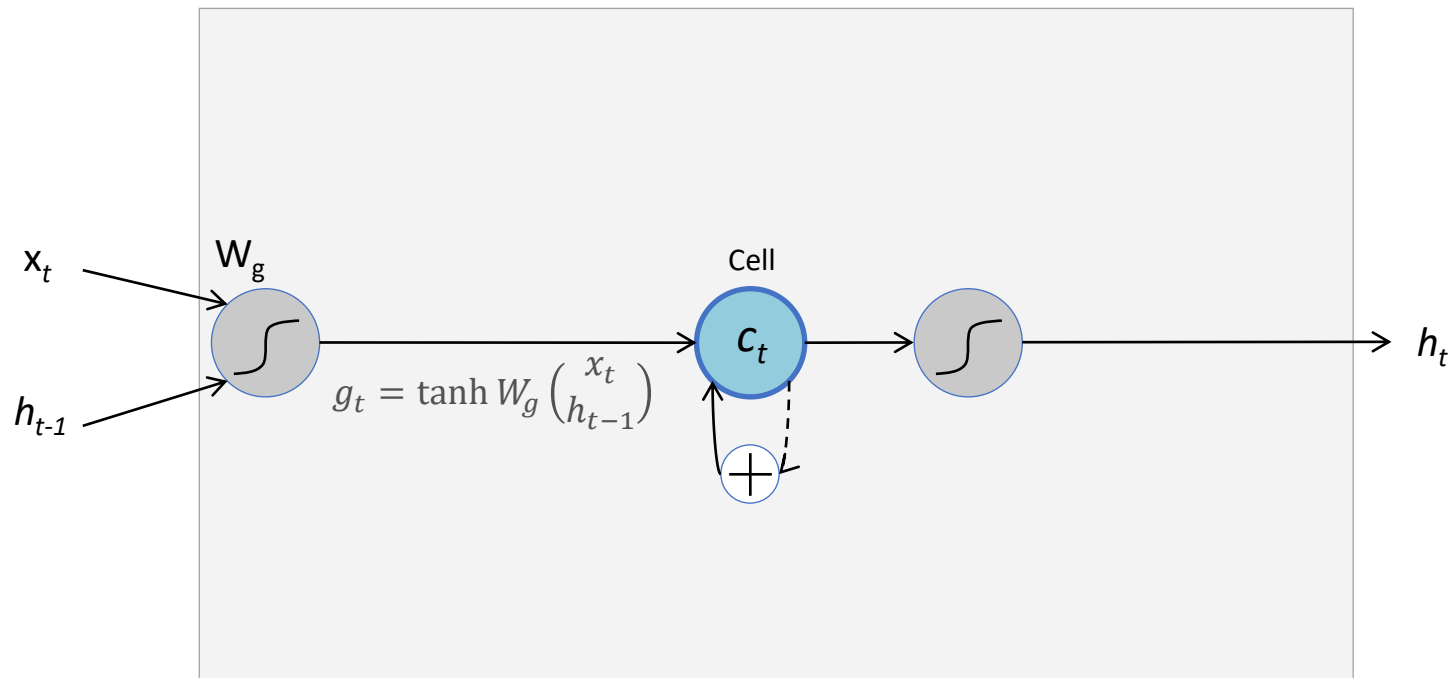


S. Hochreiter and J. Schmidhuber, [Long short-term memory](#), Neural Computation 9 (8), pp. 1735–1780, 1997

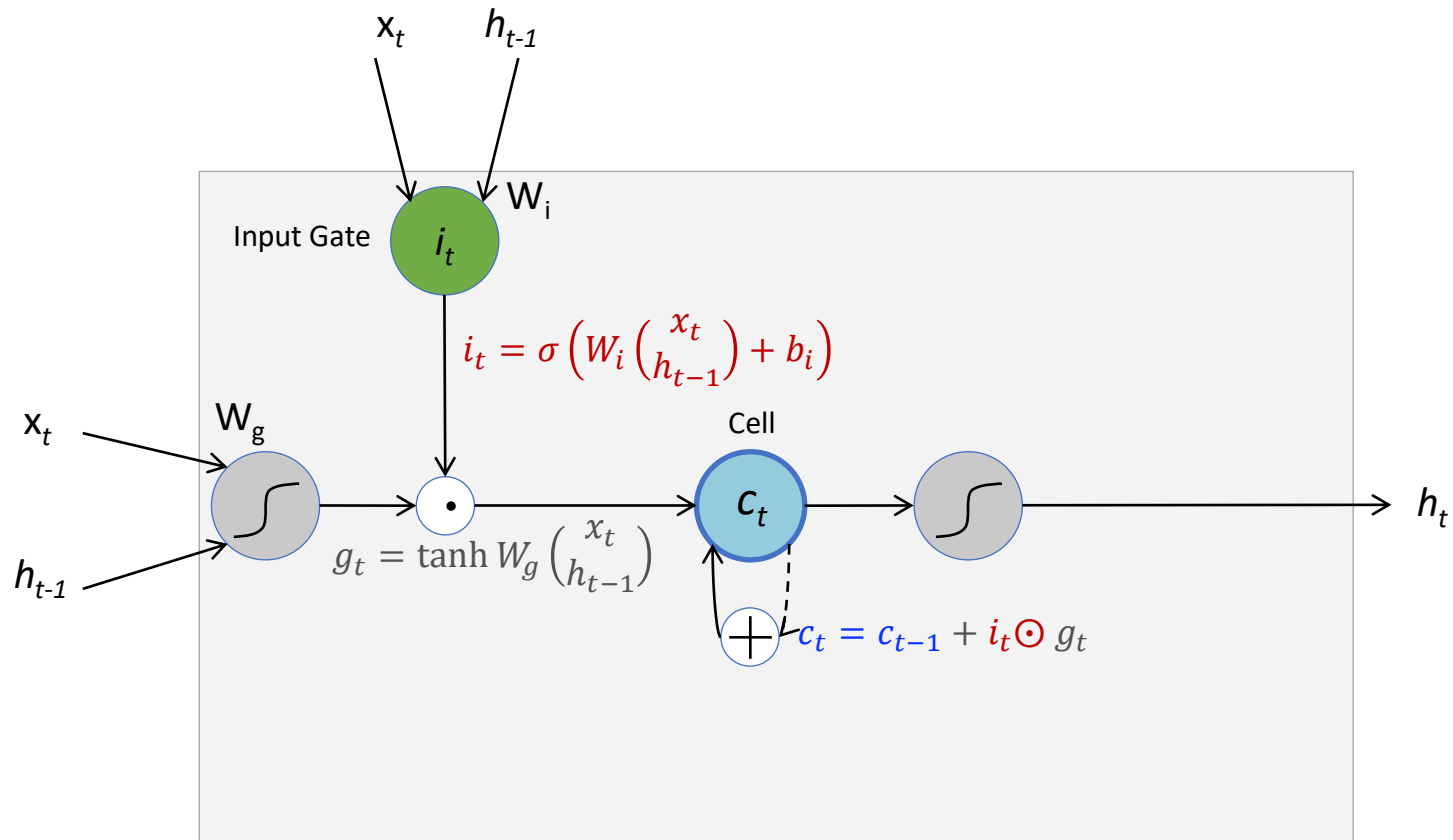
LSTM Cell



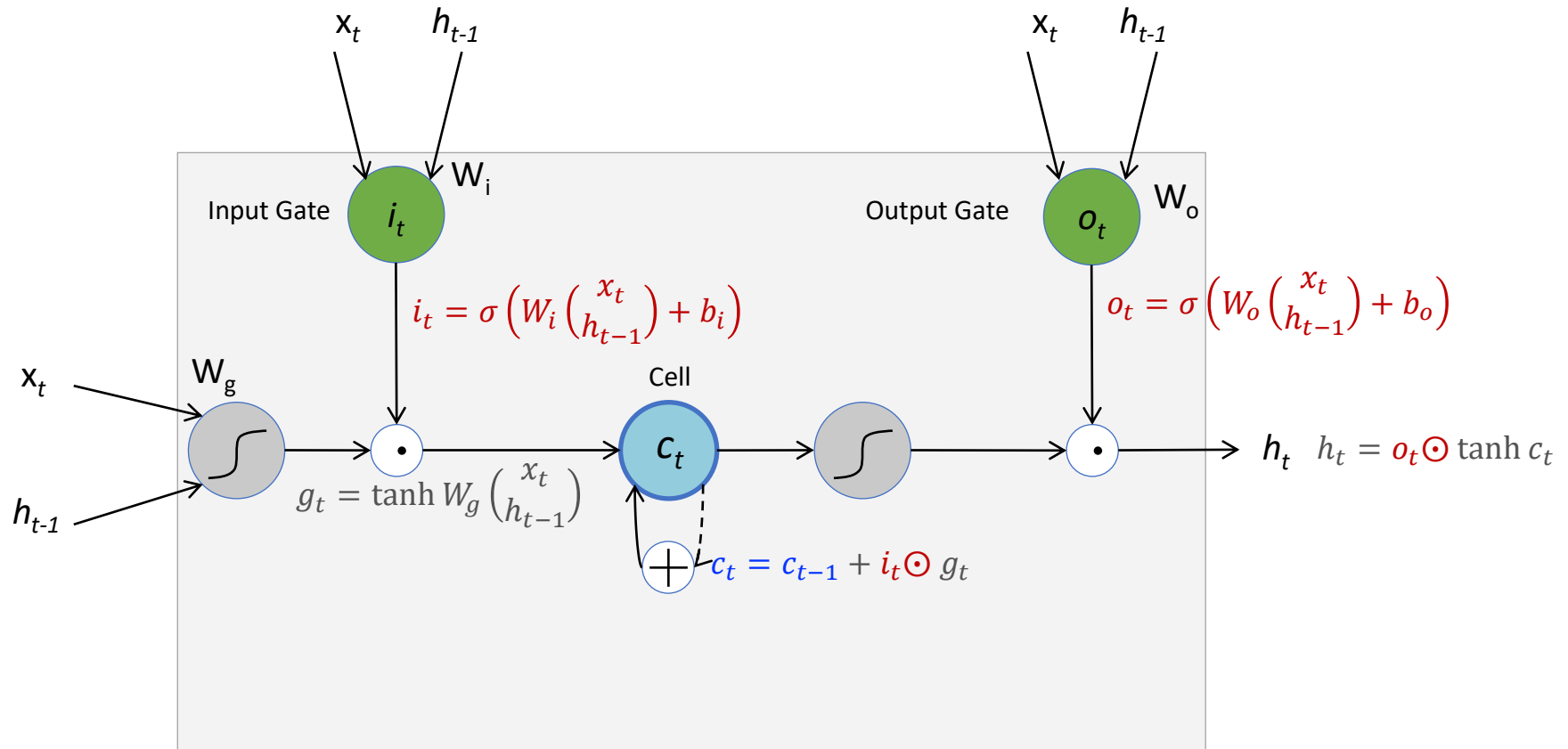
LSTM Cell



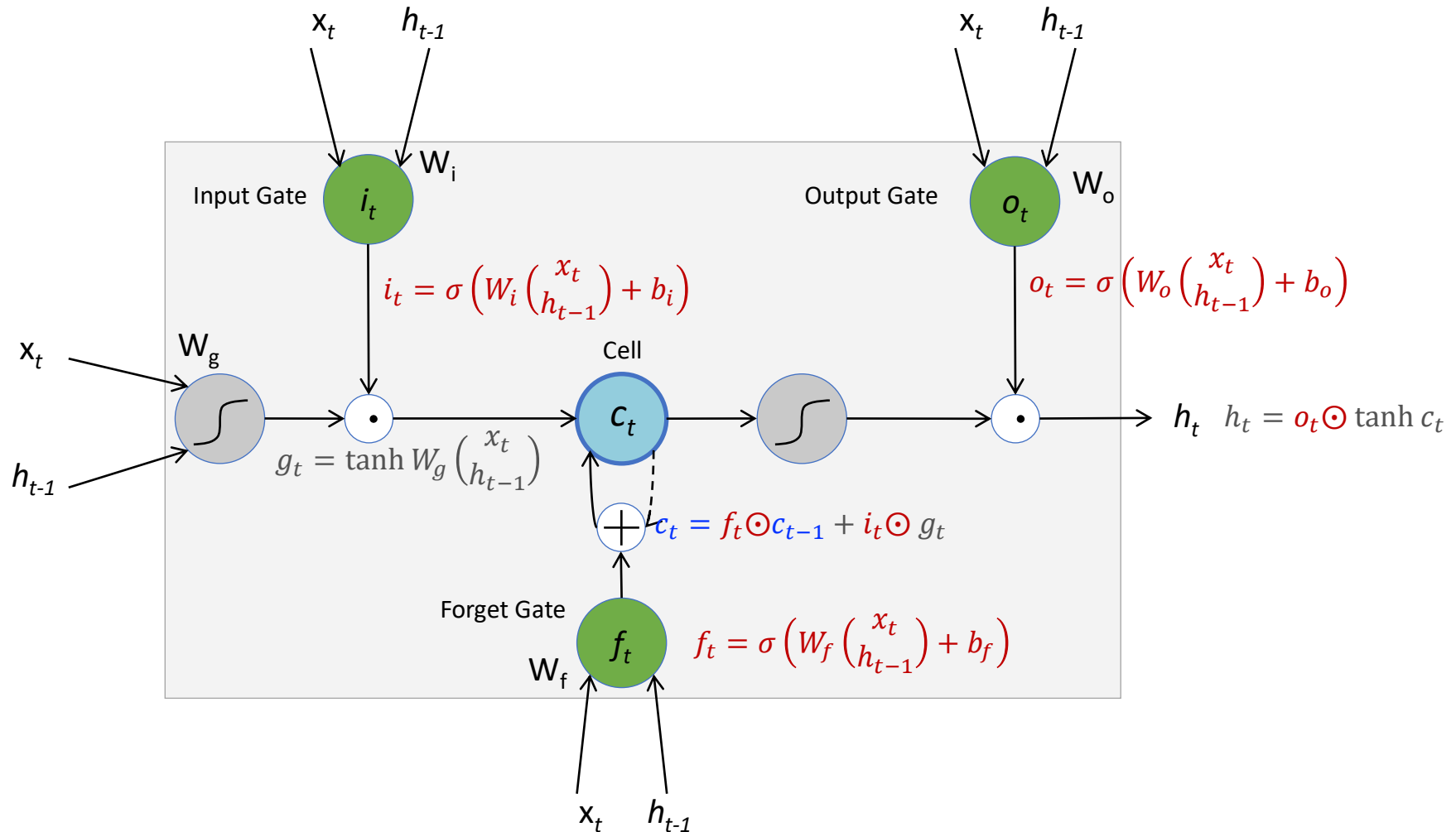
LSTM Cell



LSTM Cell



LSTM Cell

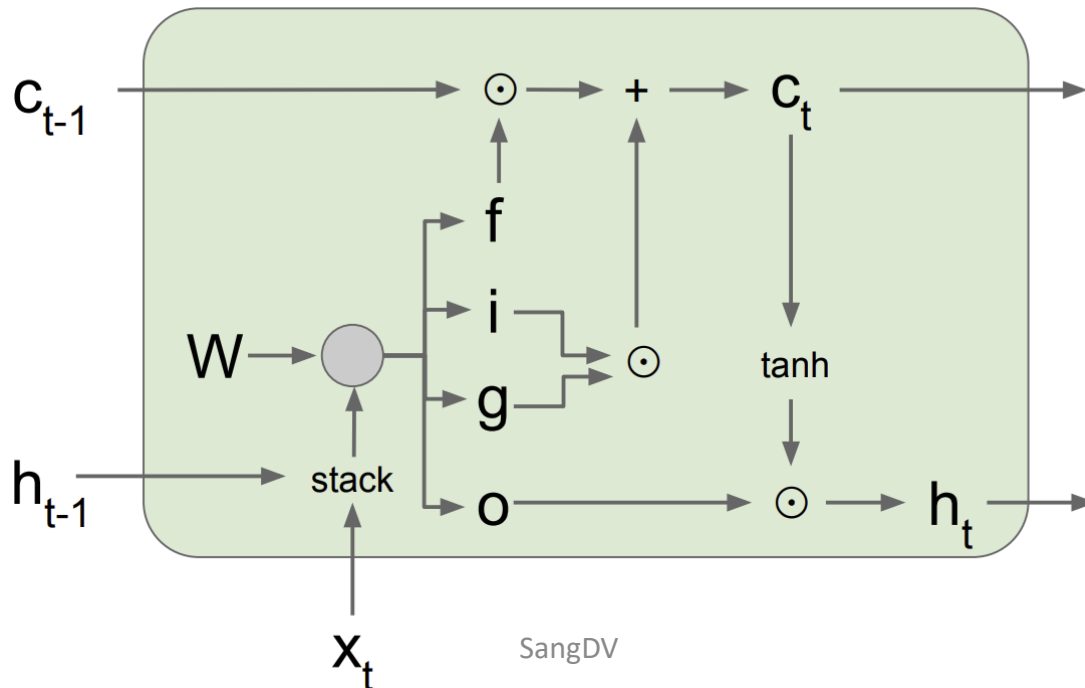


LSTM Forward Pass Summary

- $$\begin{pmatrix} g_t \\ i_t \\ f_t \\ o_t \end{pmatrix} = \begin{pmatrix} \tanh \\ \sigma \\ \sigma \\ \sigma \end{pmatrix} \begin{pmatrix} W_g \\ W_i \\ W_f \\ W_o \end{pmatrix} \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

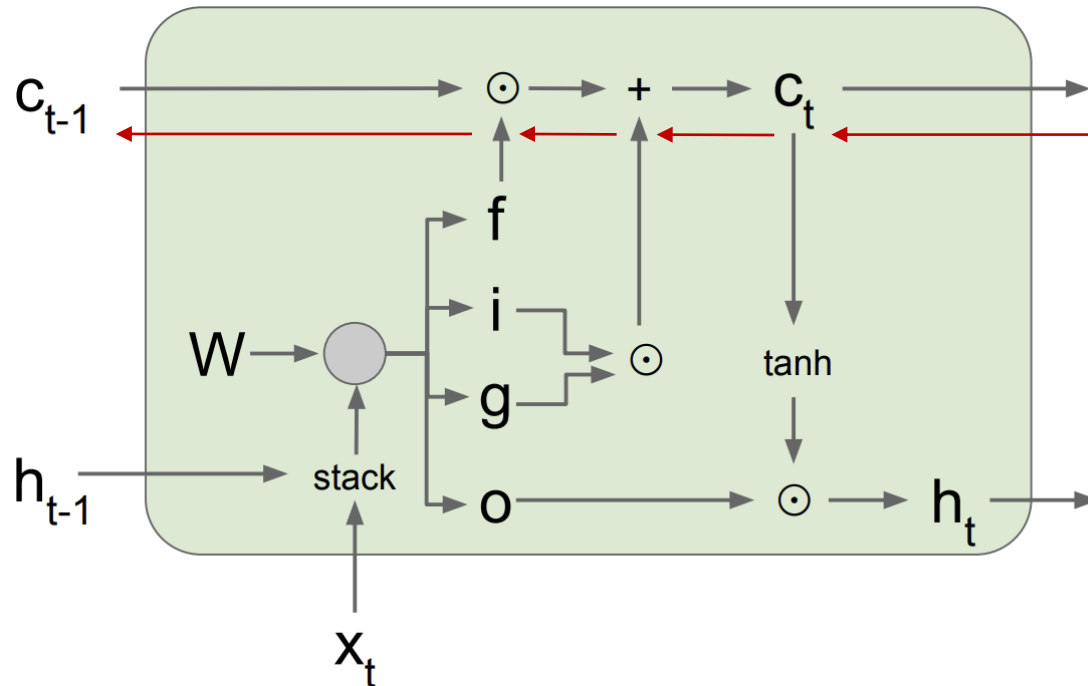
- $c_t = f_t \odot c_{t-1} + i_t \odot g_t$

- $h_t = o_t \odot \tanh c_t$



Lan truyền ngược LSTM

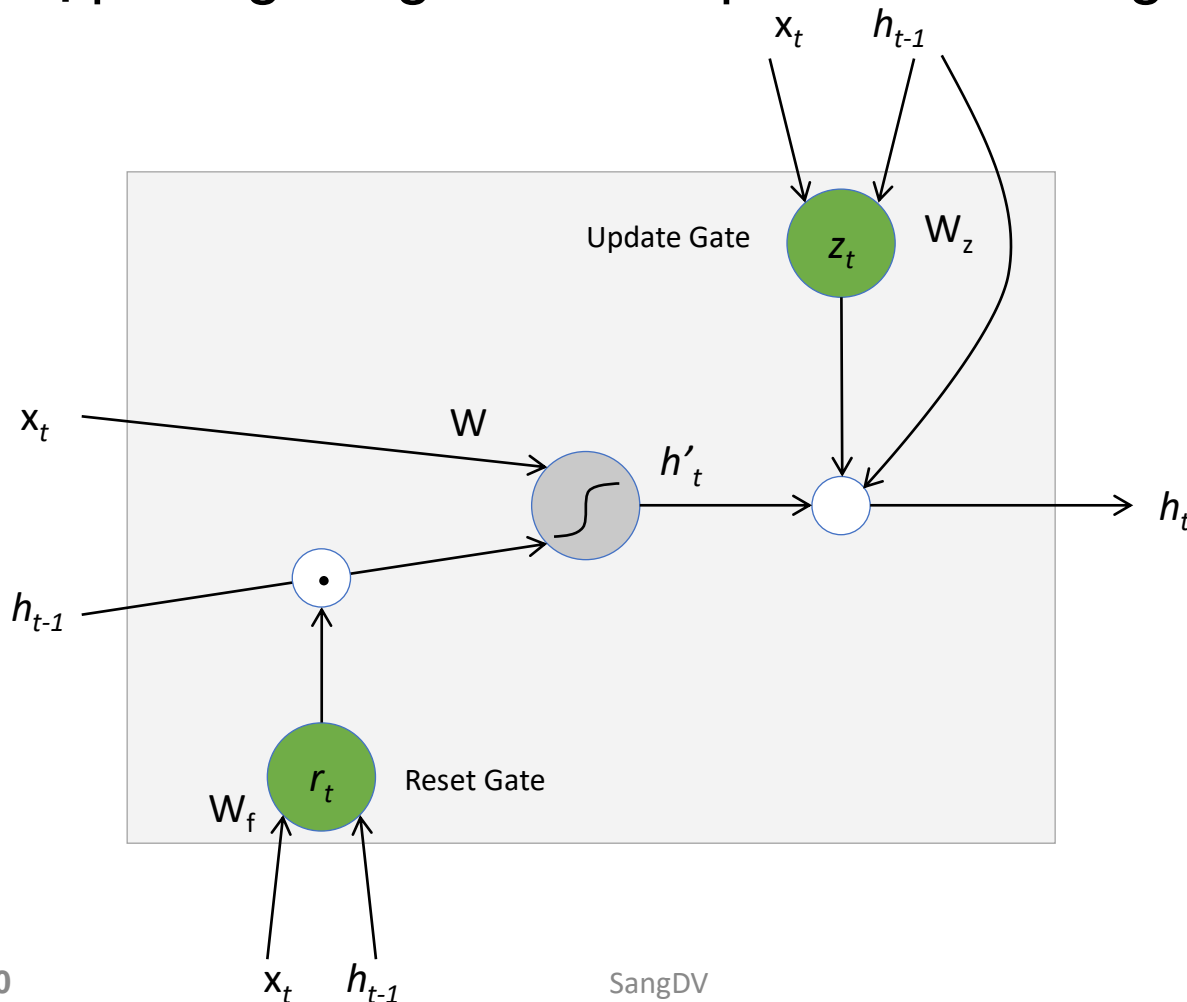
- Luồng gradient từ c_t tới c_{t-1} chỉ lan truyền ngược qua phép cộng và nhân từng phần tử, không đi qua phép nhân ma trận và hàm tanh



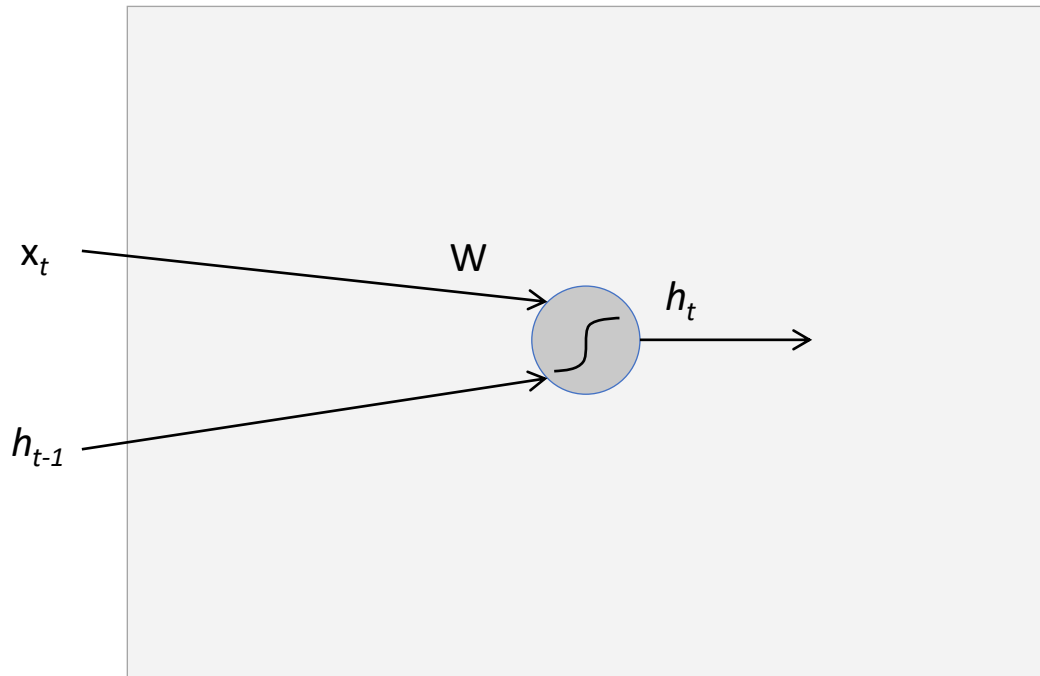
For complete details: [Illustrated LSTM Forward and Backward Pass](#)

Gated Recurrent Unit (GRU)

- Không dùng “cell state” riêng biệt, ghép chung với hidden state
- Kết hợp cổng “forget” và “output” thành cổng “update”

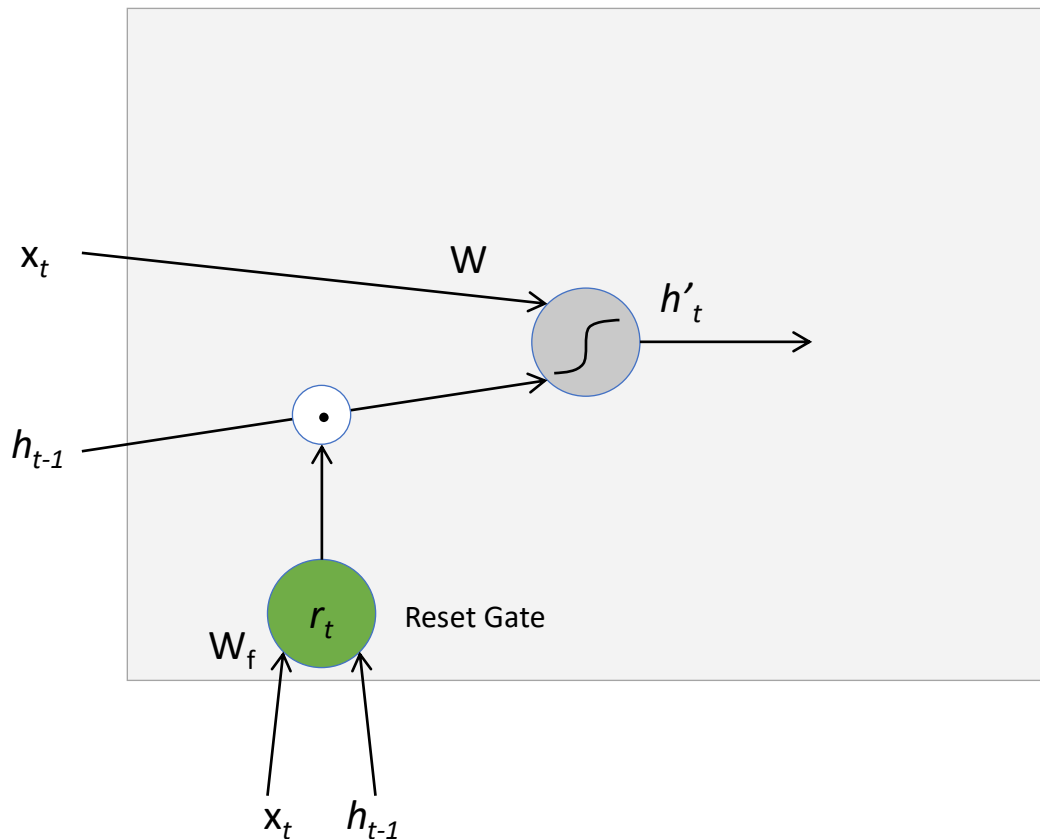


Gated Recurrent Unit (GRU)



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

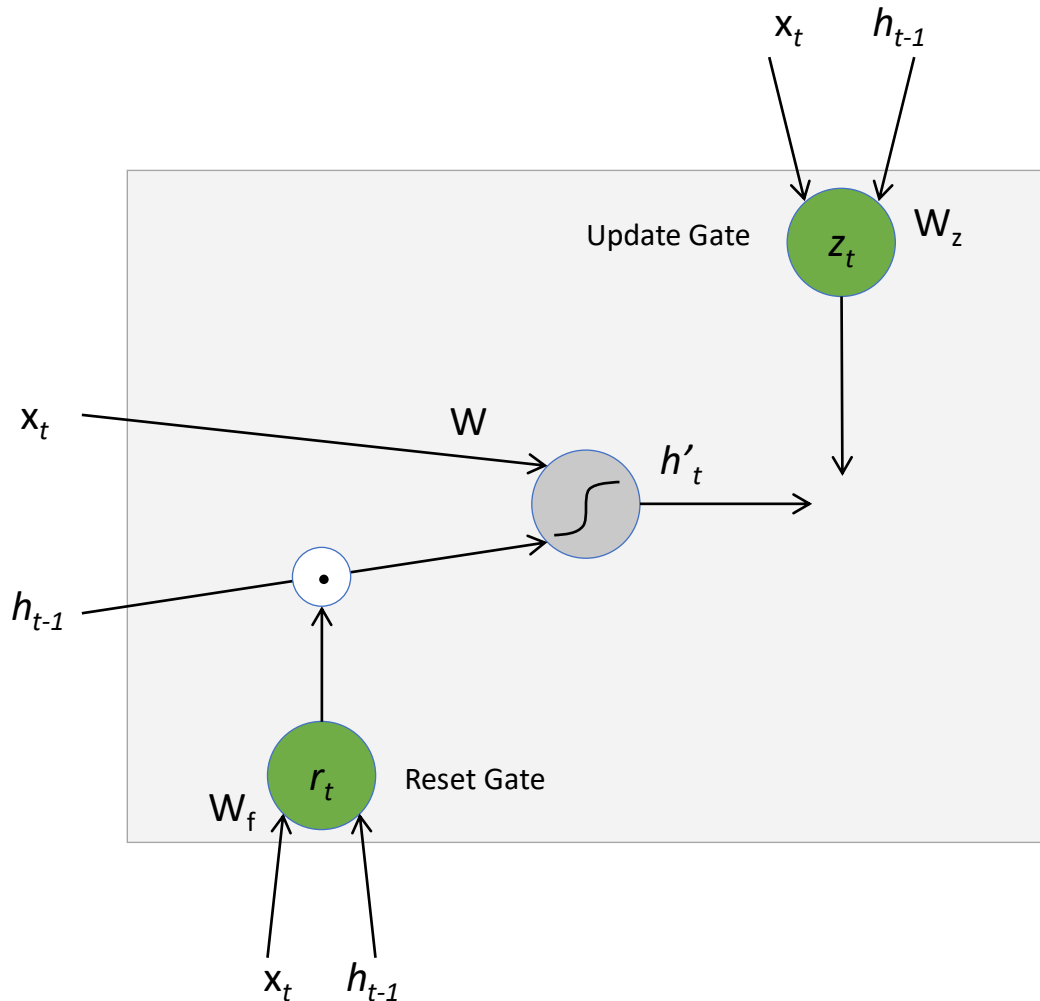
Gated Recurrent Unit (GRU)



$$r_t = \sigma \left(W_r \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_r \right)$$

$$h'_t = \tanh W \left(r_t \odot h_{t-1} \right)$$

Gated Recurrent Unit (GRU)

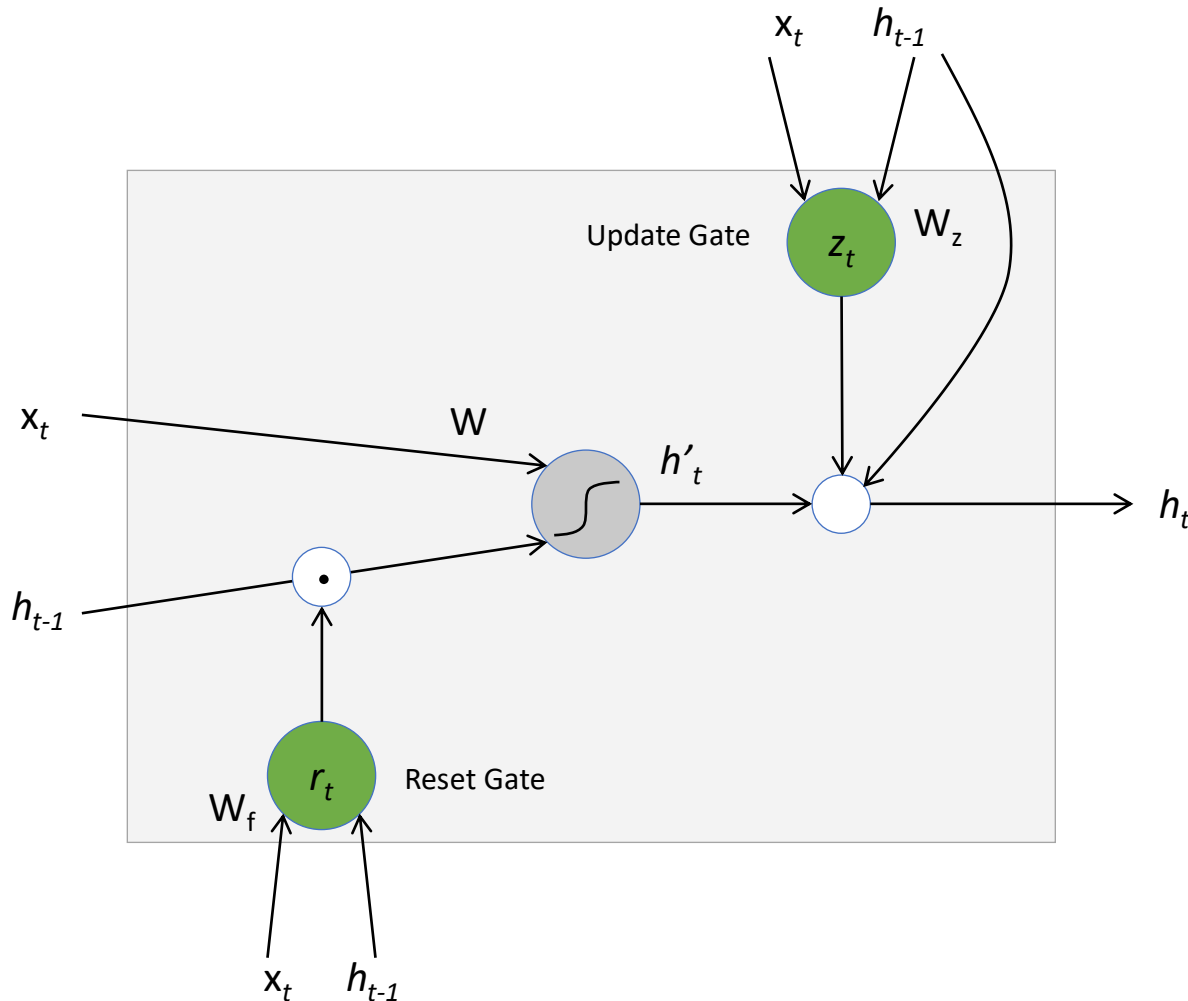


$$r_t = \sigma \left(W_r \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_r \right)$$

$$h'_t = \tanh W \begin{pmatrix} x_t \\ r_t \odot h_{t-1} \end{pmatrix}$$

$$z_t = \sigma \left(W_z \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_z \right)$$

Gated Recurrent Unit (GRU)



$$r_t = \sigma \left(W_r \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_t \right)$$

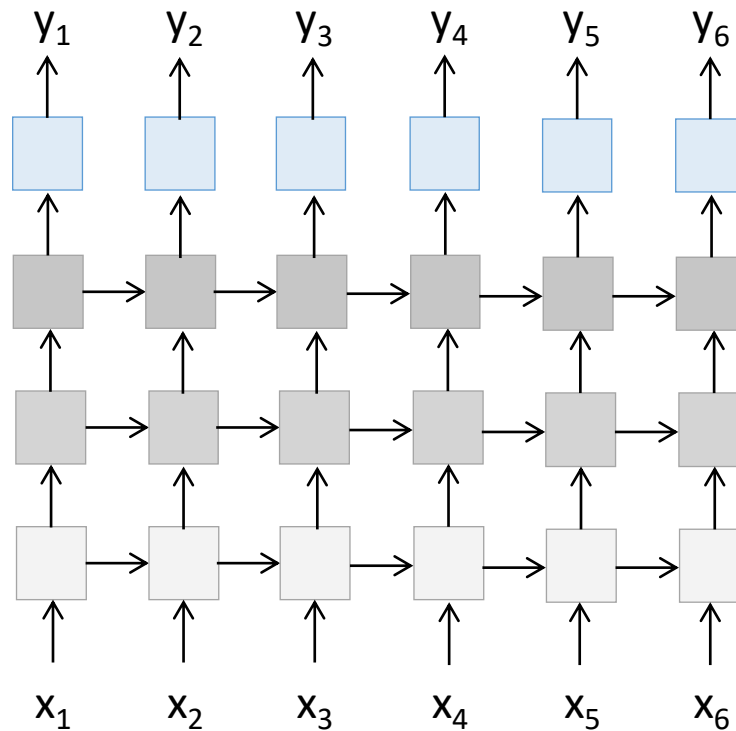
$$h'_t = \tanh W \left(r_t \odot h_{t-1} \right)$$

$$z_t = \sigma \left(W_z \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_z \right)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot h'_t$$

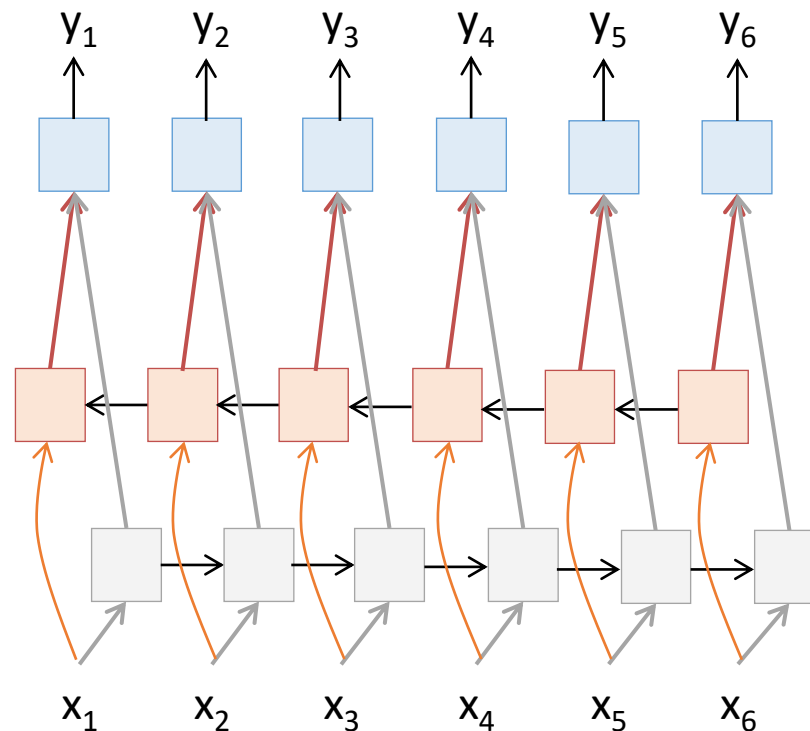
RNNs nhiều lớp

- Có thể thiết kế RNNs với nhiều lớp ẩn



RNNs hai chiều

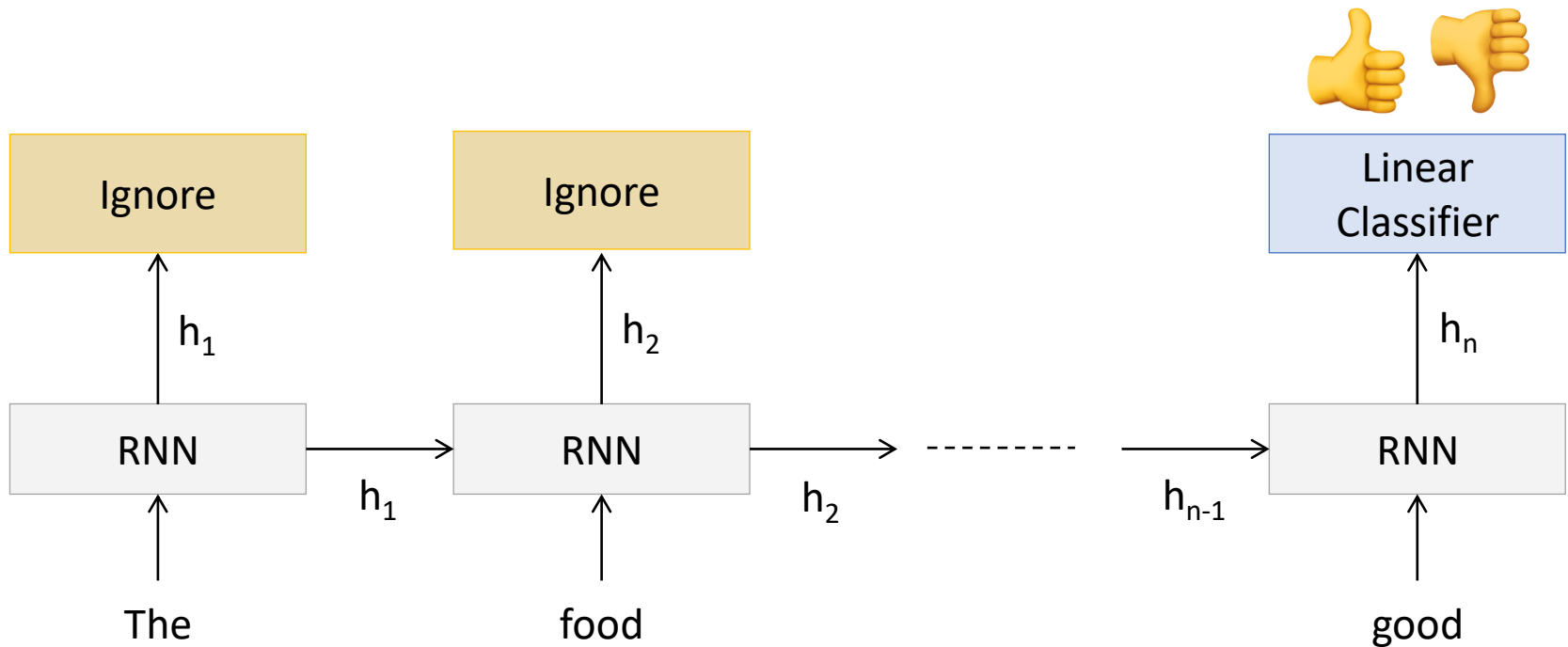
- RNNs có thể xử lý chuỗi đầu vào theo chiều ngược vào chiều xuôi



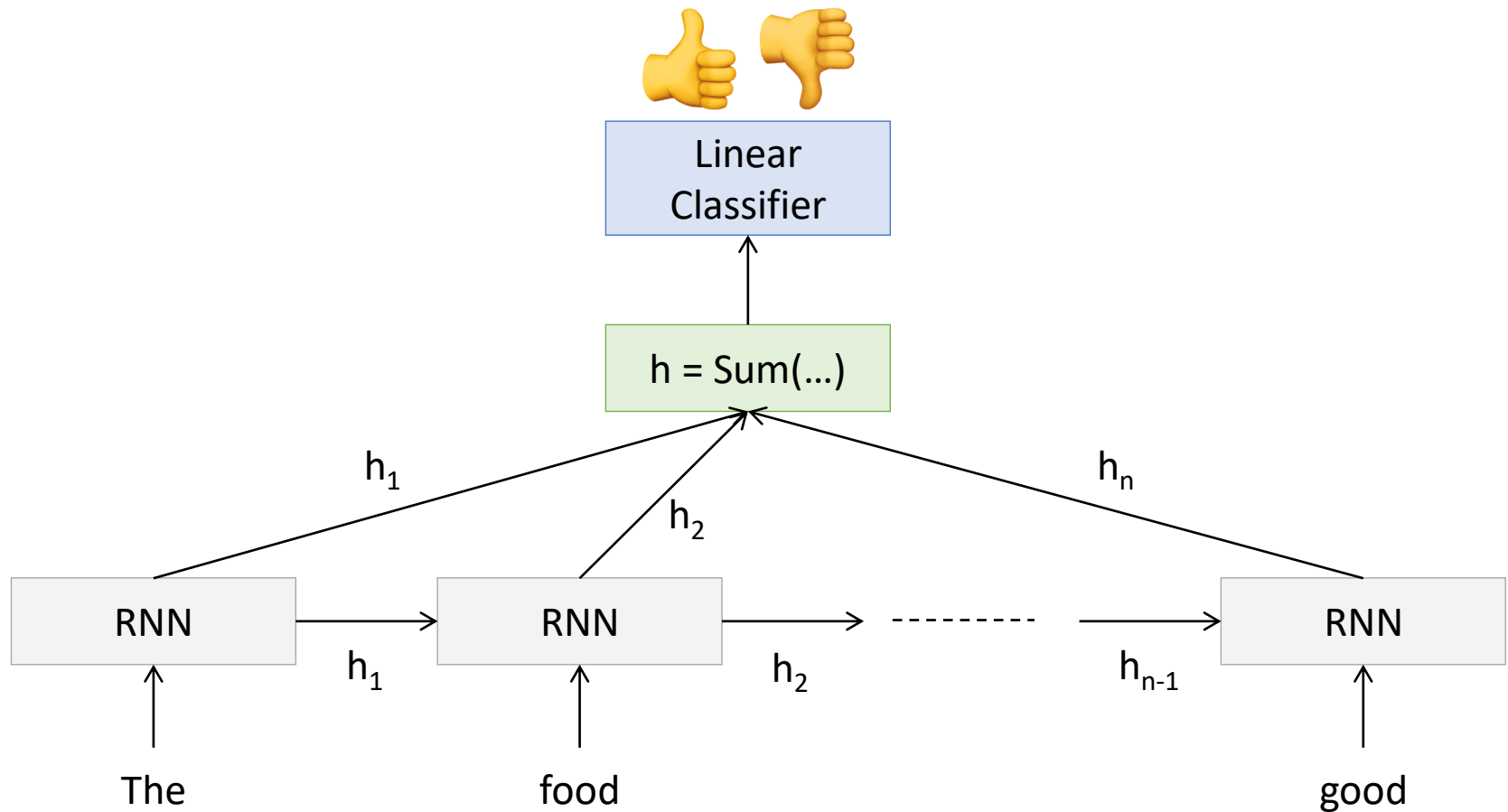
- Phổ biến trong nhận dạng âm thanh

Một số ví dụ ứng dụng

Phân loại chuỗi

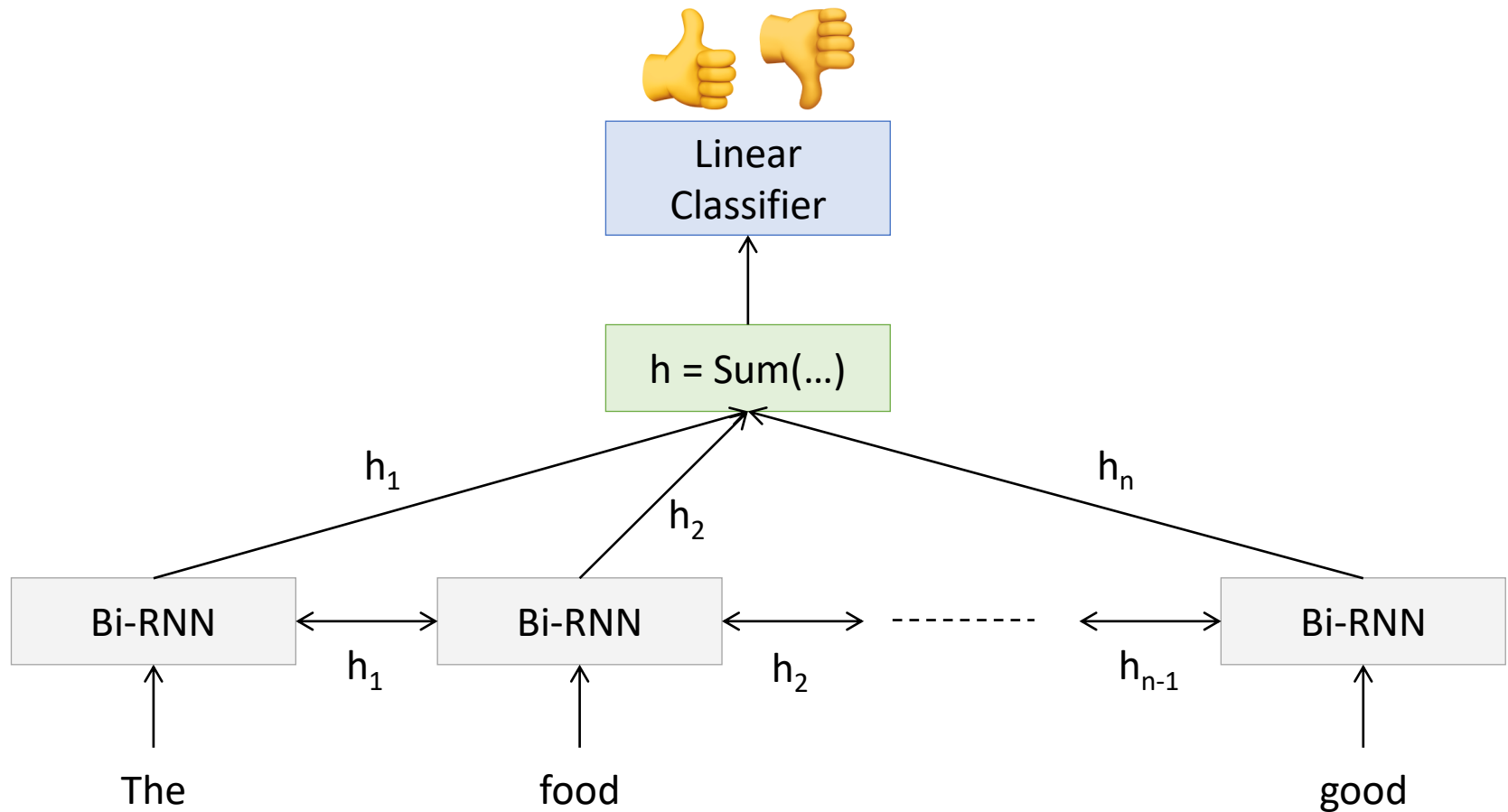


Phân loại chuỗi



<http://deeplearning.net/tutorial/lstm.html>

Phân loại chuỗi



Character 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 nigtiqe,aoaenns lng

↓ train more

**300th
iteration**

"Tmont thithey" fomesscerliund
Keushey. Thom here
sheulke, anmerenith ol sivh I lalterthend Bleipile shuw y 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 ofter.

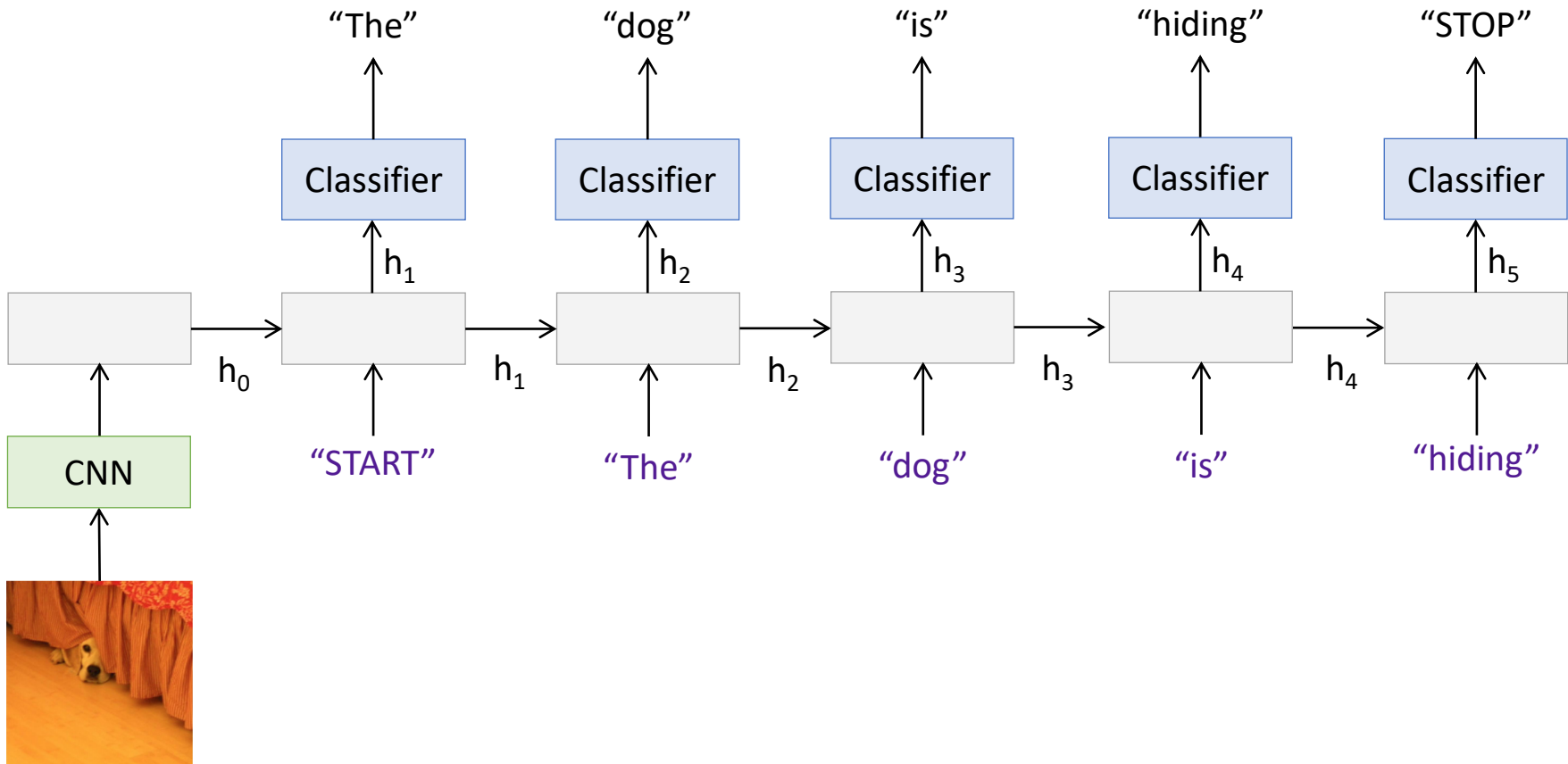
↓ train more

**2000th
iteration**

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

<http://karpathy.github.io/2015/05/21/rnn-effectiveness/>

Image Caption Generation



Sinh mô tả ảnh

A person riding a motorcycle on a dirt road.



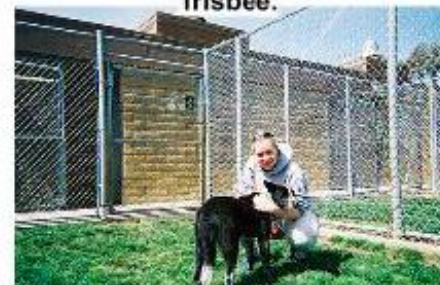
Two dogs play in the grass.



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

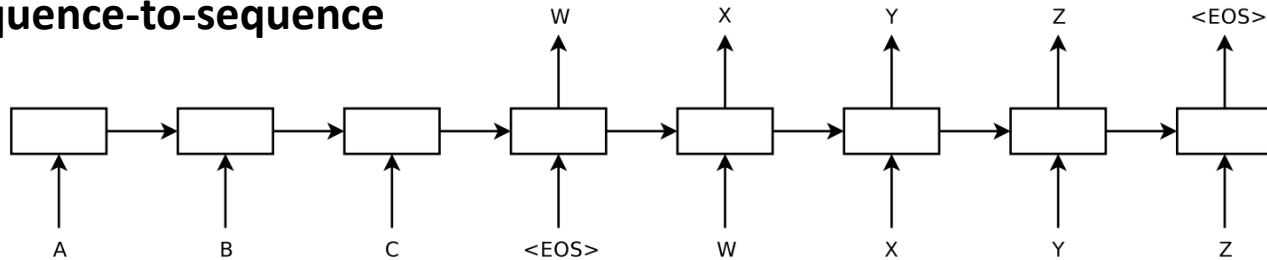
Describes with minor errors

Somewhat related to the image

Unrelated to the image

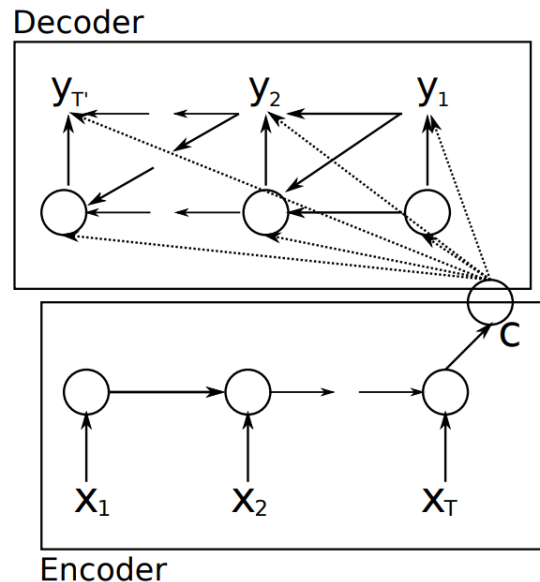
Dịch máy

Sequence-to-sequence



I. Sutskever, O. Vinyals, Q. Le, [Sequence to Sequence Learning with Neural Networks](#), NIPS 2014

Encoder-decoder



K. Cho, B. Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio, [Learning phrase representations using RNN encoder-decoder for statistical machine translation](#), ACL 2014

Tài liệu tham khảo

1. Khóa cs231n của Stanford:

<http://cs231n.stanford.edu>

2. Khóa cs244n của Stanford:

<http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture06-rnnlm.pdf>

<http://web.stanford.edu/class/cs224n/slides/cs224n-2020-lecture07-fancy-rnn.pdf>

3. Training RNNs:

<http://www.cs.toronto.edu/~rgrosse/csc321/lec10.pdf>