GRU

Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation

황주훈

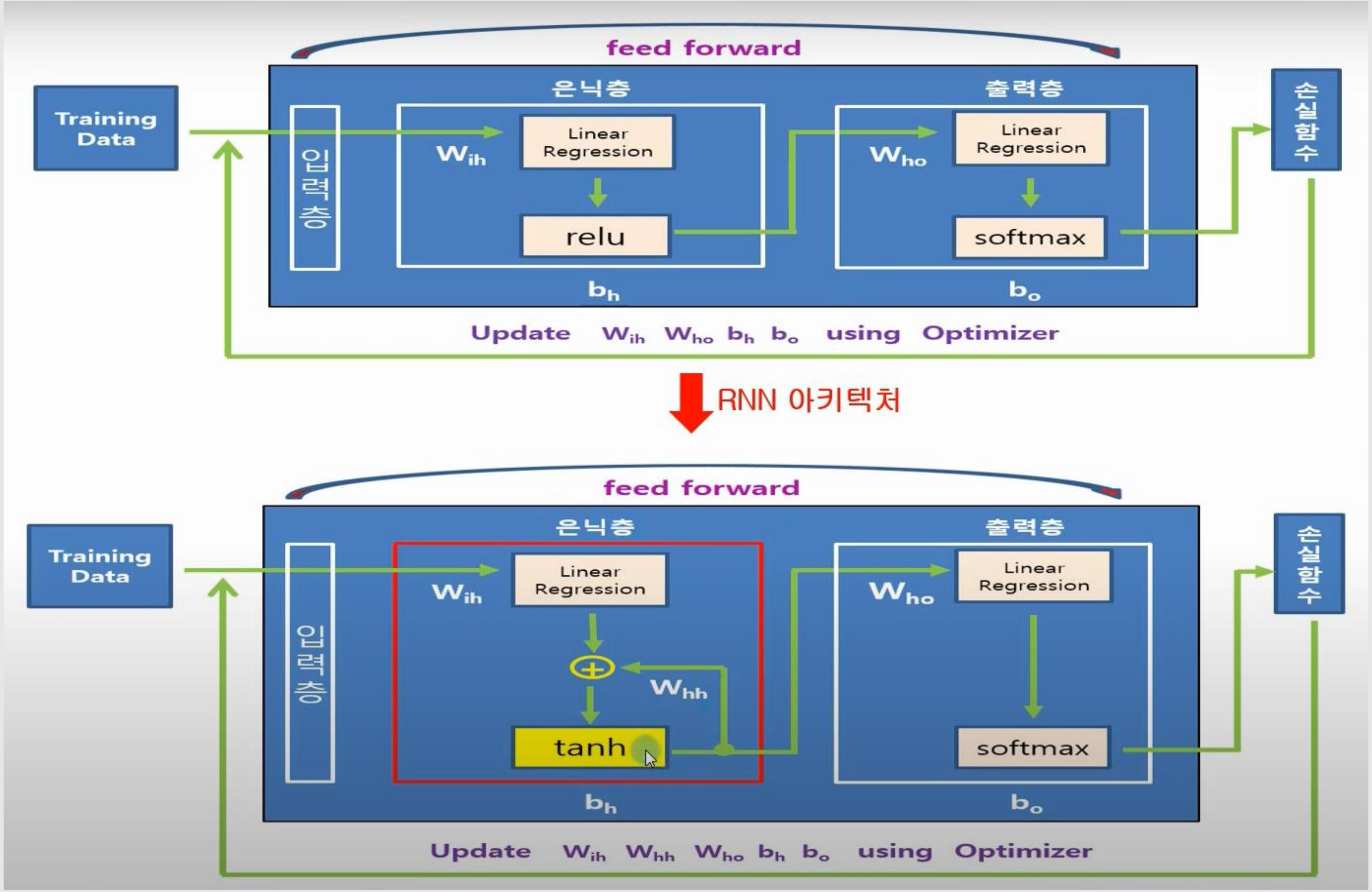
MADE BY HENDO

INDEX

01. 배경지식

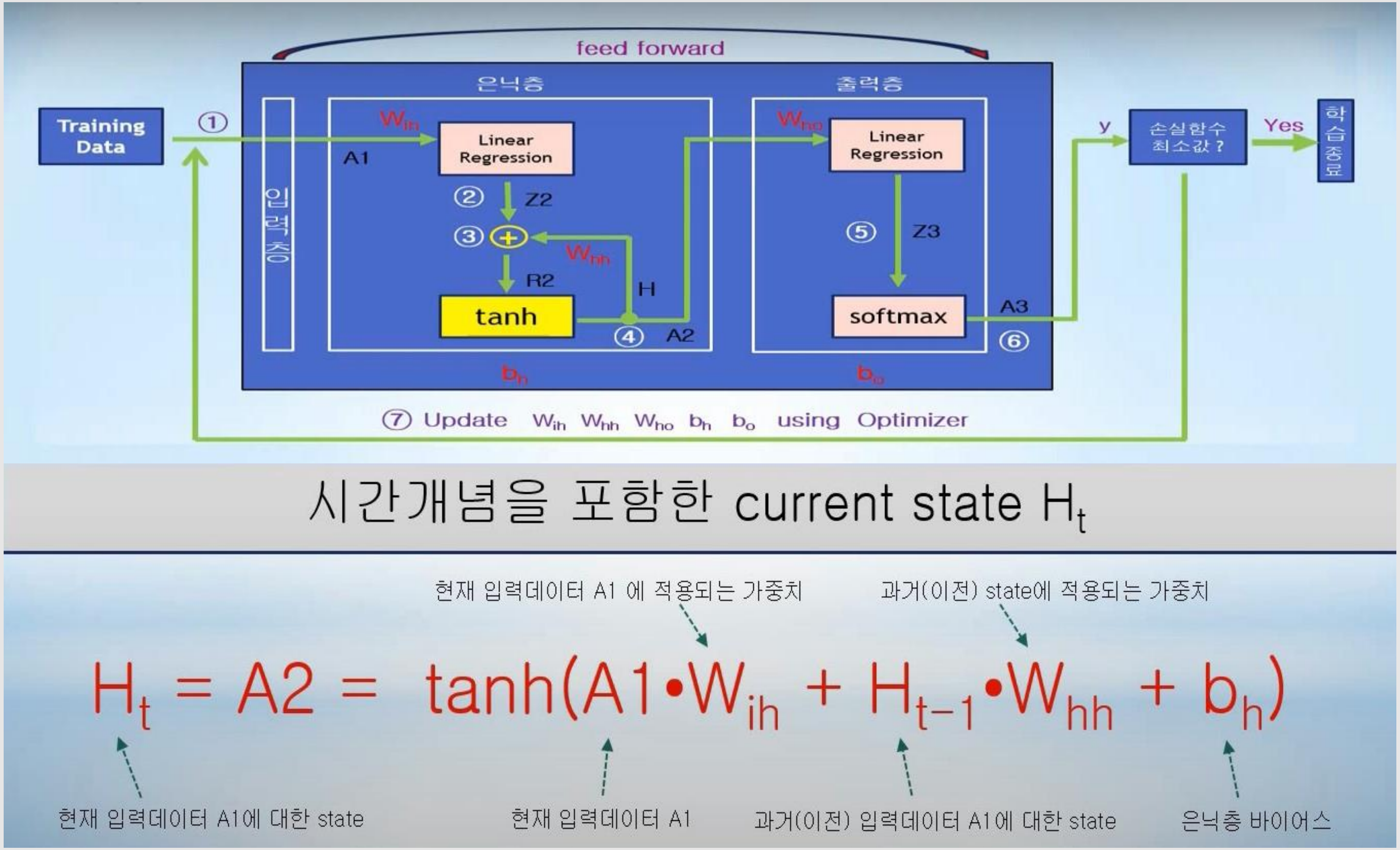
MADE BY HENDO

01. 배경지식 (RNN)

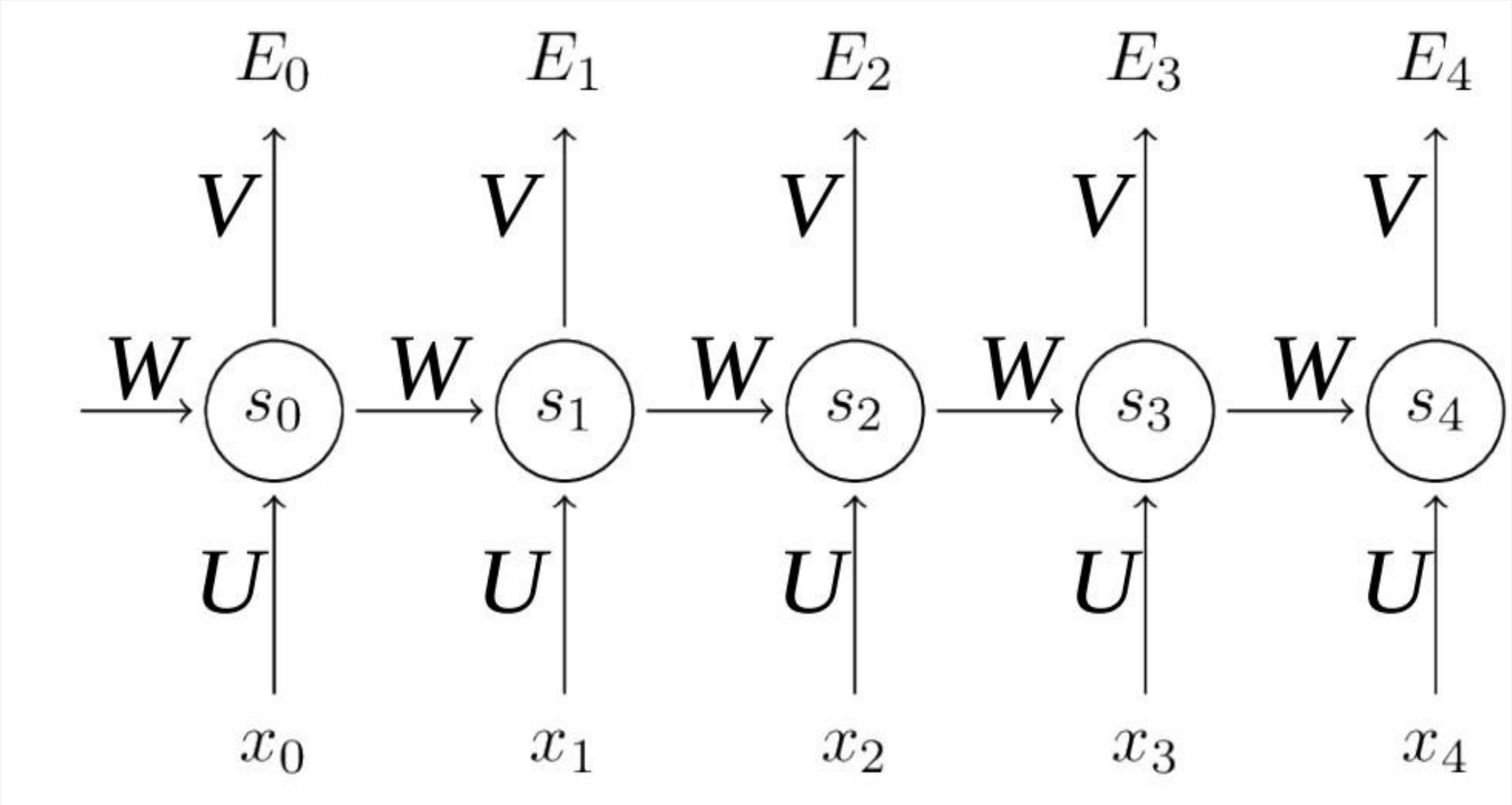


feed forward
Training Data
은닉층
출력층
입력층
Linear Regression
relu
Linear Regression
softmax
W_{ih}
W_{ho}
b_h
b_o
손실함수
Update W_{ih} W_{ho} b_h b_o using Optimizer
RNN 아키텍처
feed forward
Training Data
은닉층
출력층
입력층
Linear Regression
tanh
W_{hh}
Linear Regression
softmax
W_{ih}
W_{ho}
b_h
b_o
손실함수
Update W_{ih} W_{hh} W_{ho} b_h b_o using Optimizer

01. 배경지식 (RNN)

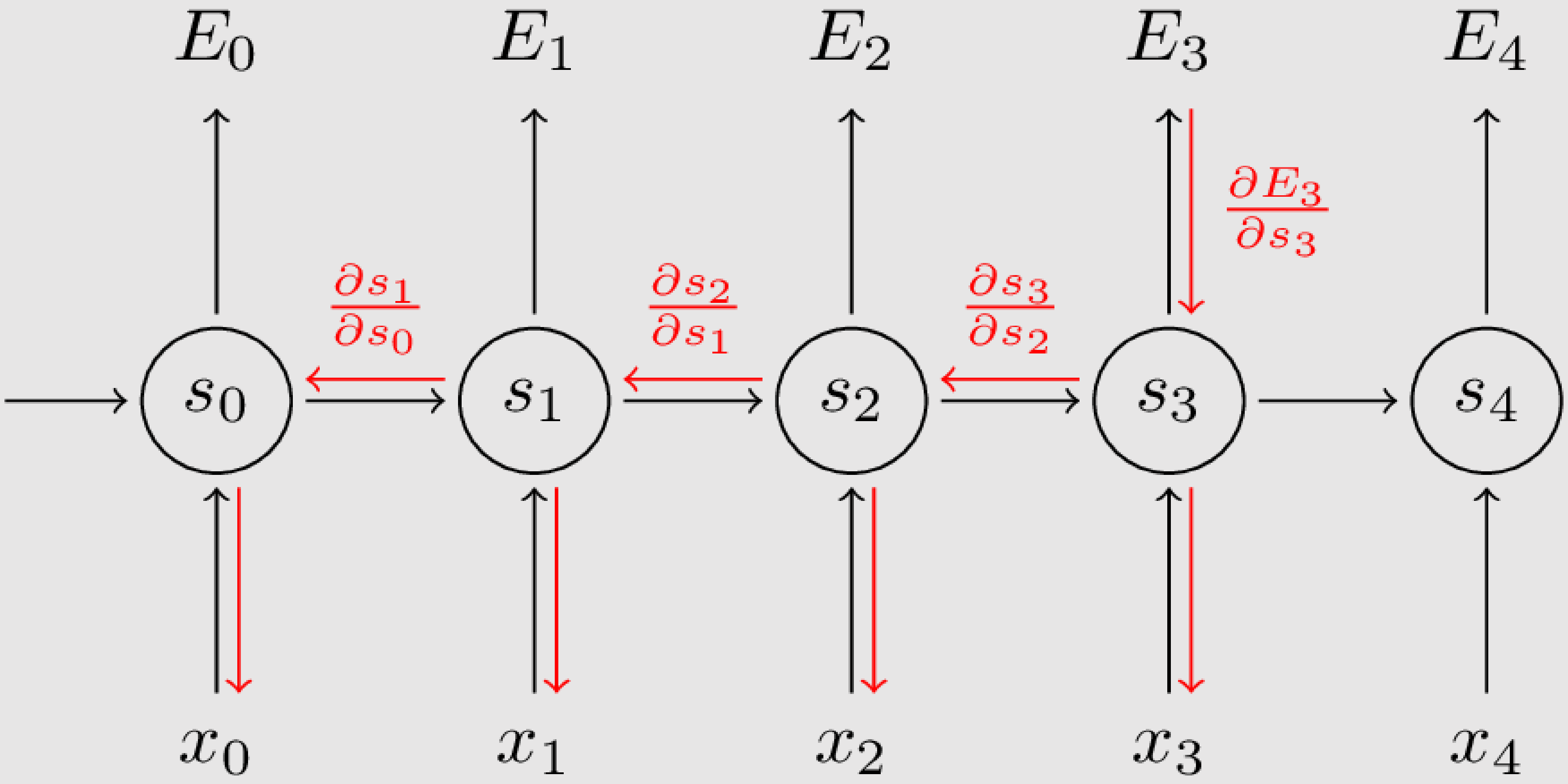

feed forward
은닉층
출력층
Training Data
①
입력층
A1
Linear Regression
② Z2
③ +
W_{hh}
R2
H
tanh
④ A2
W_{ho}
Linear Regression
⑤ Z3
softmax
A3
⑥
b_h
b_o
y
손실함수 최소값 ?
Yes
학습종료
⑦ Update W_{ih} W_{hh} W_{ho} b_h b_o using Optimizer
시간개념을 포함한 current state H_t
현재 입력데이터 A1 에 적용되는 가중치
과거(이전) state에 적용되는 가중치
$H_t = A2 = \tanh(A1 \cdot W_{ih} + H_{t-1} \cdot W_{hh} + b_h)$
현재 입력데이터 A1에 대한 state
현재 입력데이터 A1
과거(이전) 입력데이터 A1에 대한 state
은닉층 바이어스

01. 배경지식 (RNN)



E=Error
E=Y - ŷ

01. 배경지식 (RNN)

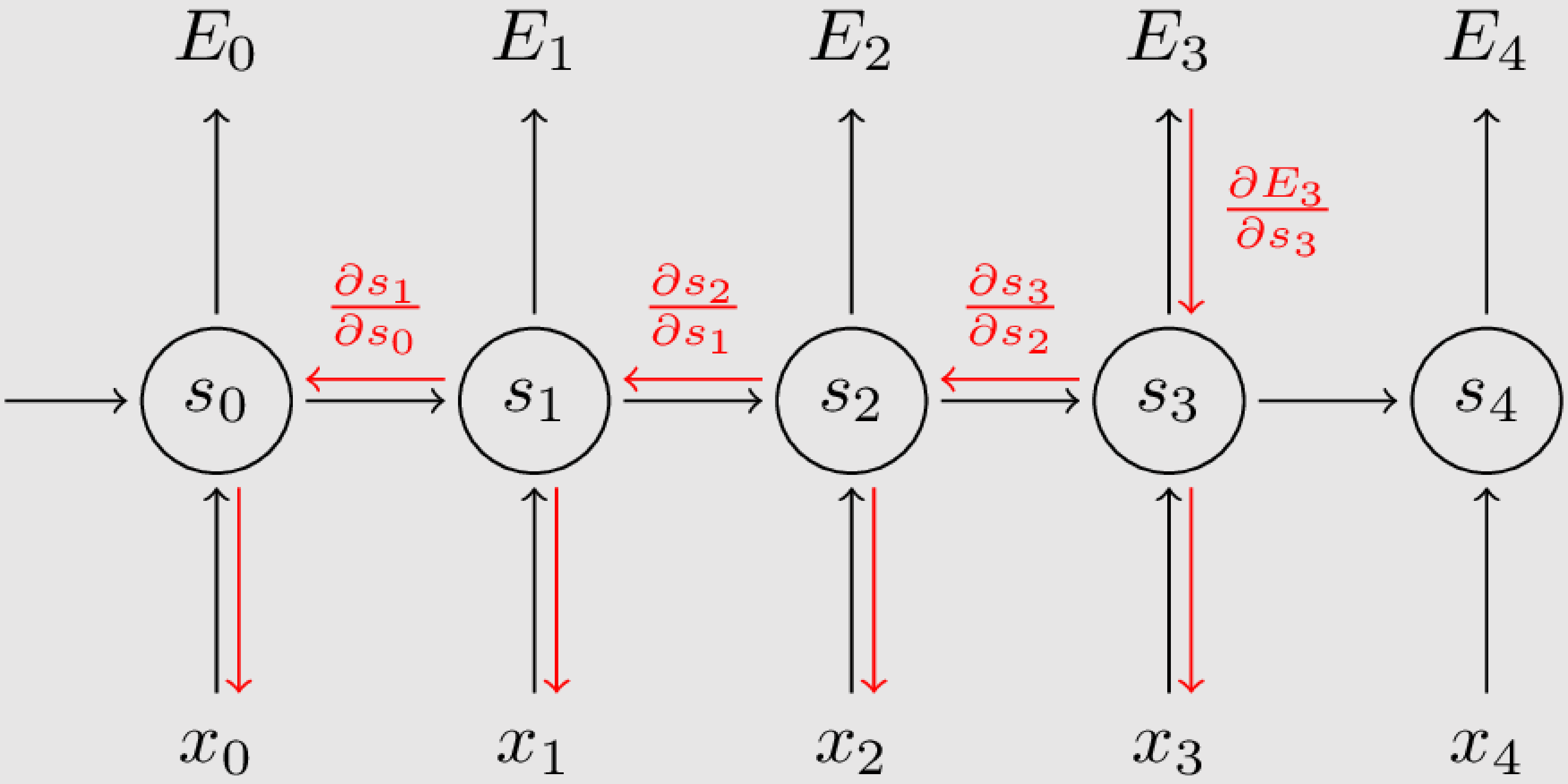


$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

01. 배경지식 (RNN)



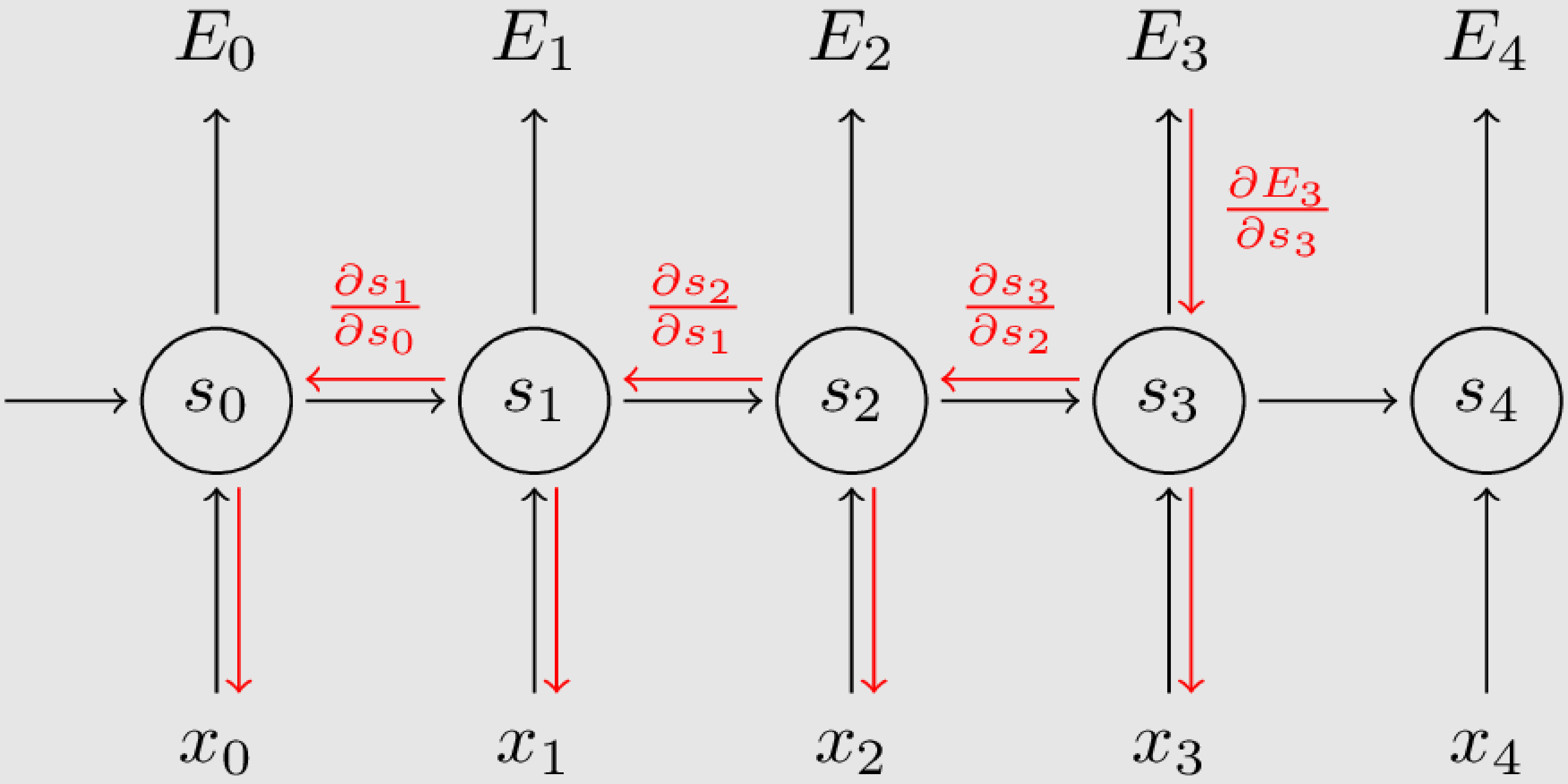
$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

01. 배경지식 (RNN)



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

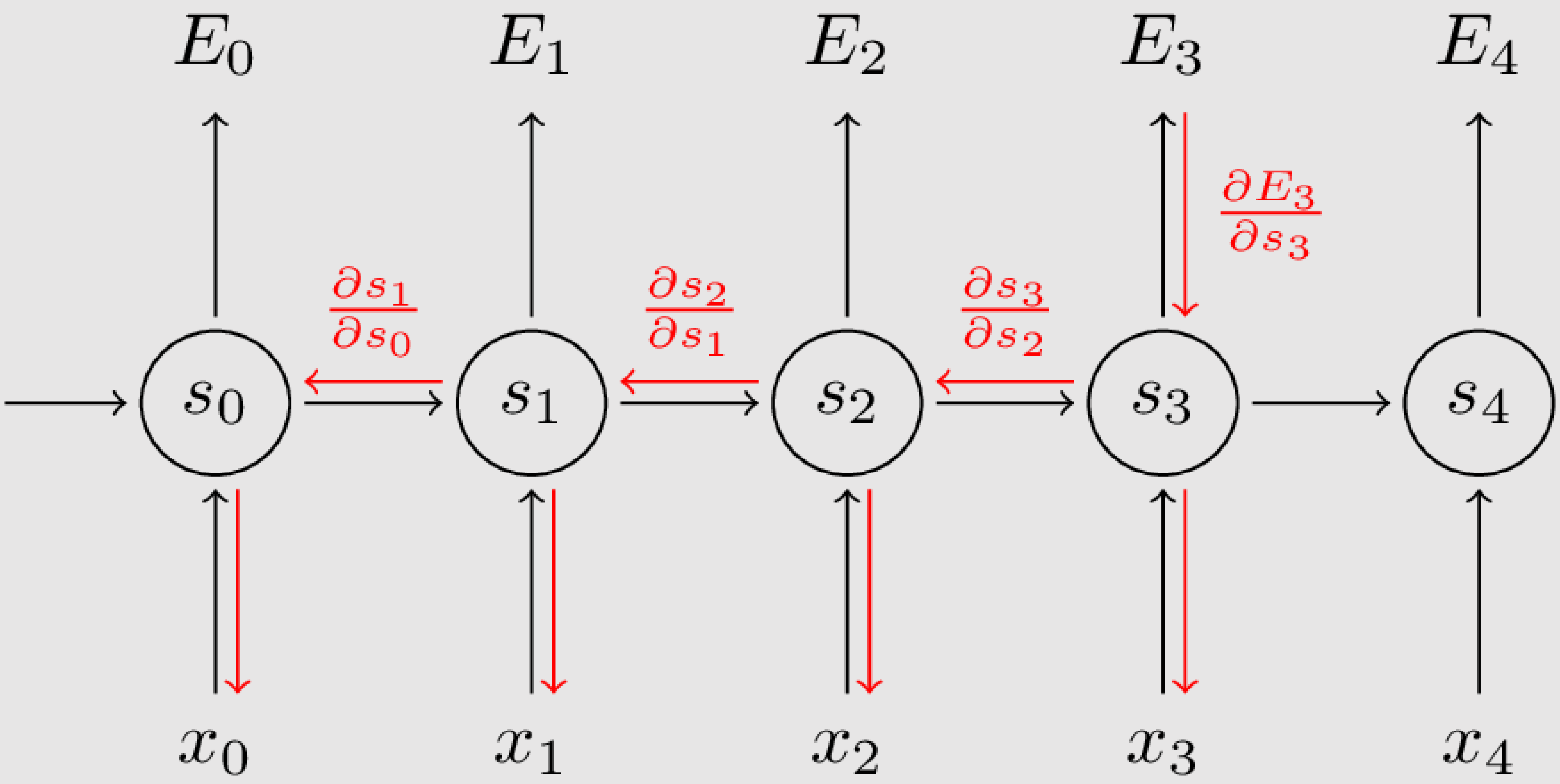
$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

01. 배경지식 (RNN)



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s} + \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s} + \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

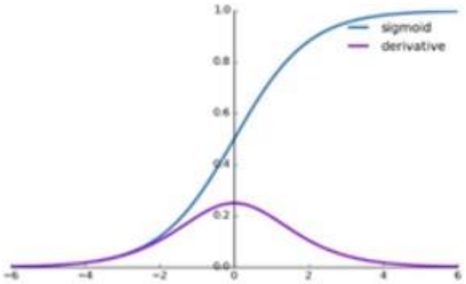
$$W = W - \text{learning\_rate} * \frac{\partial E}{\partial W}$$

01. 배경지식 (RNN)

왜 tanh를 쓰는가?

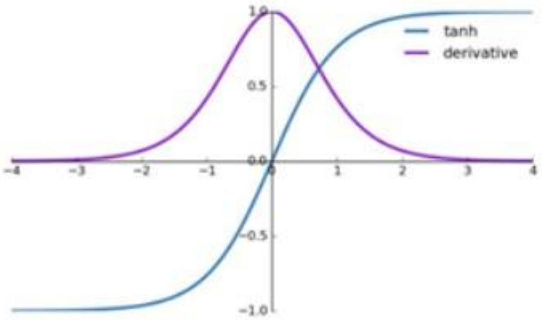
Activation Function

- Sigmoid



Sigmoid	
$f(x)$	$\frac{1}{1+e^{-x}}$ (y: 0 ~ 1)
$\frac{d}{dx}f(x)$	$\frac{1}{1+e^{-x}}\left(1-\frac{1}{1+e^{-x}}\right)$ (y′: 0~0.25)

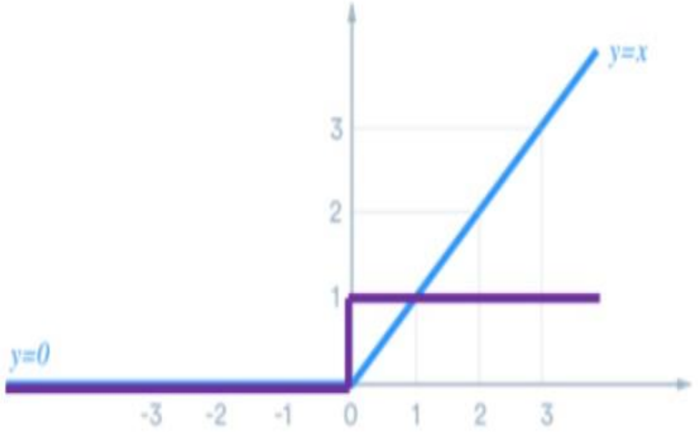
- Tanh



Tanh	
$f(x)$	$\frac{e^x - e^{-x}}{e^x + e^{-x}}$ (y: −1~1)
$\frac{d}{dx}f(x)$	$1-\tanh(x)^2$ (y′: 0~1)

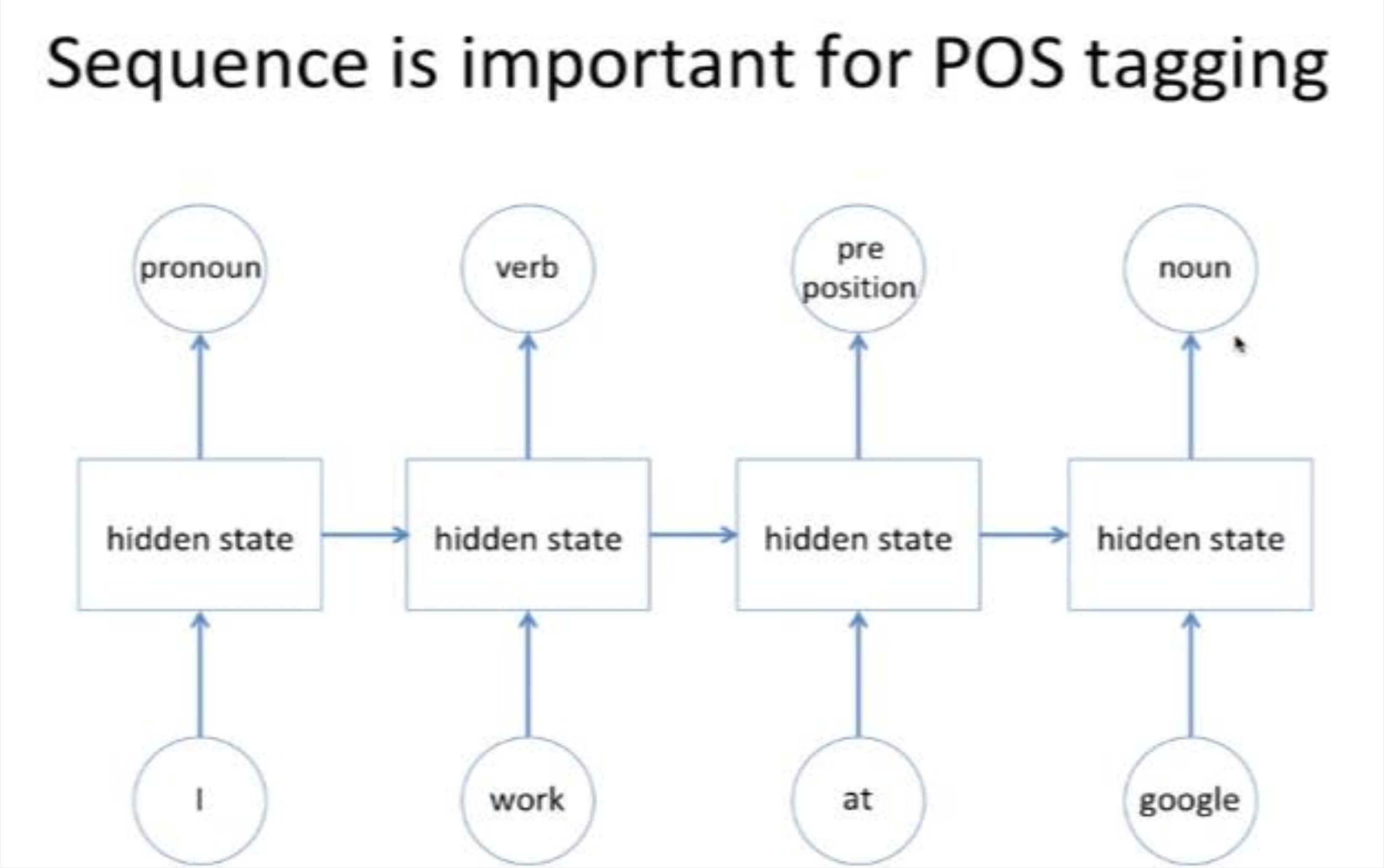
01. 배경지식 (RNN)

왜 tanh를 쓰는가?

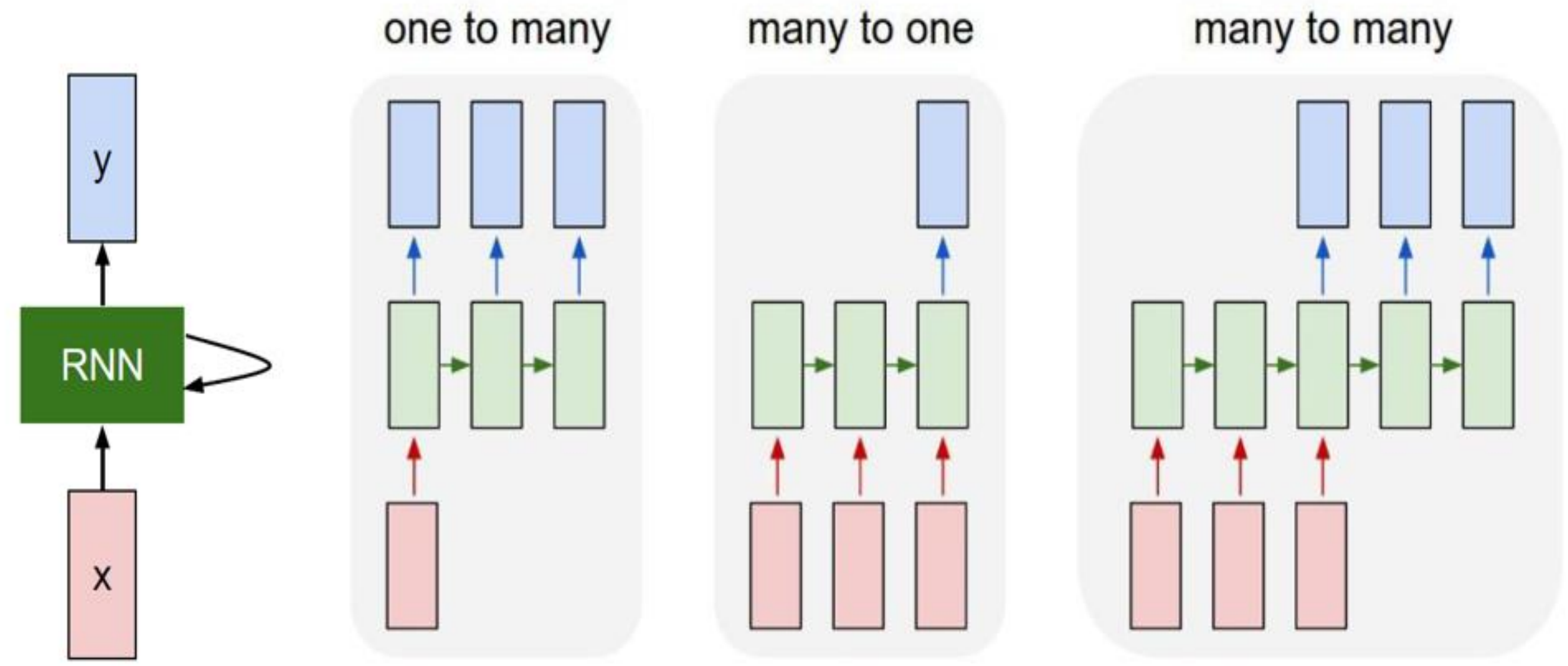


ReLU	
$f(x)$	$\max(0,x)$
$\frac{d}{dx}f(x)$	$\begin{cases} 1 & (x \geq 0) \\ 0 & (x < 0) \end{cases}$

01. 배경지식 (RNN)


Sequence is important for POS tagging
pronoun
verb
pre position
noun
hidden state
hidden state
hidden state
hidden state
I
work
at
google

01. 배경지식 (RNN)



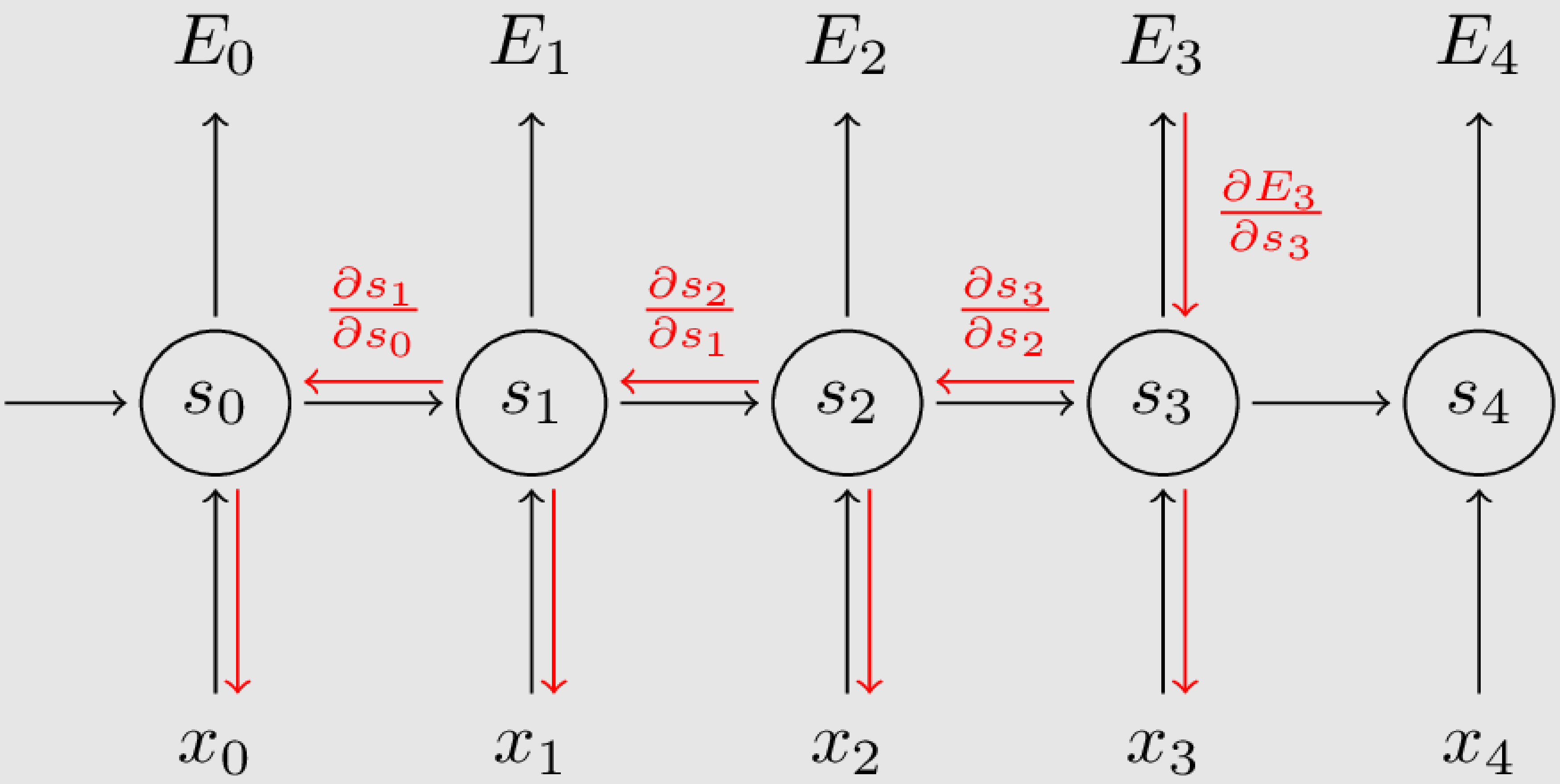
<기본구조>

<사진설명 붙이기>
사진 → 단어들

<감성분석>
단어들 → 감성점수

<번역>
단어들 → 단어들

01. 배경지식 (RNN)



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

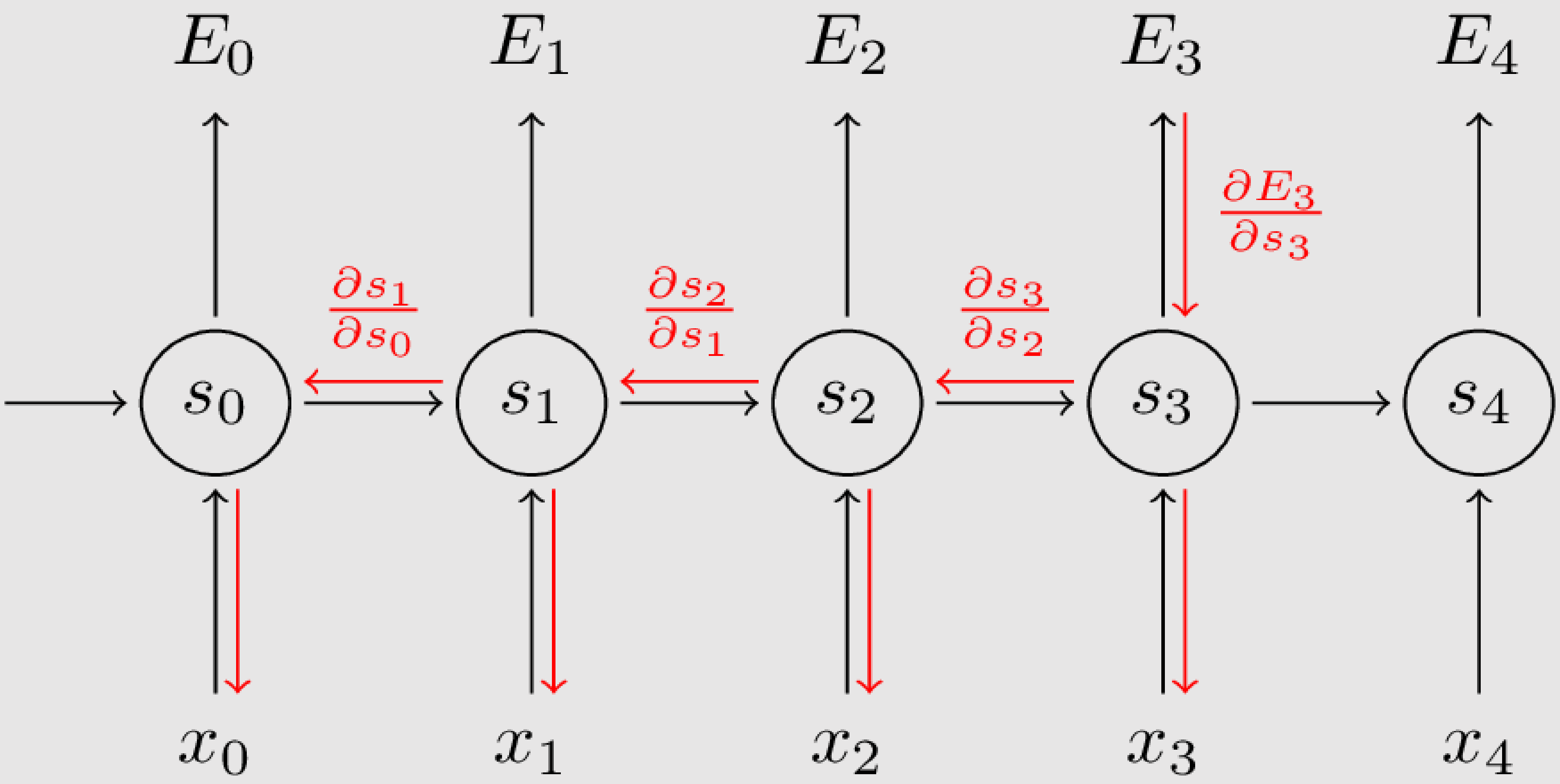
$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

$$W = W - learning\_rate * \frac{\partial E}{\partial W}$$

1보다 미분값이 클경우
Gradient Exploding

01. 배경지식 (RNN)



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

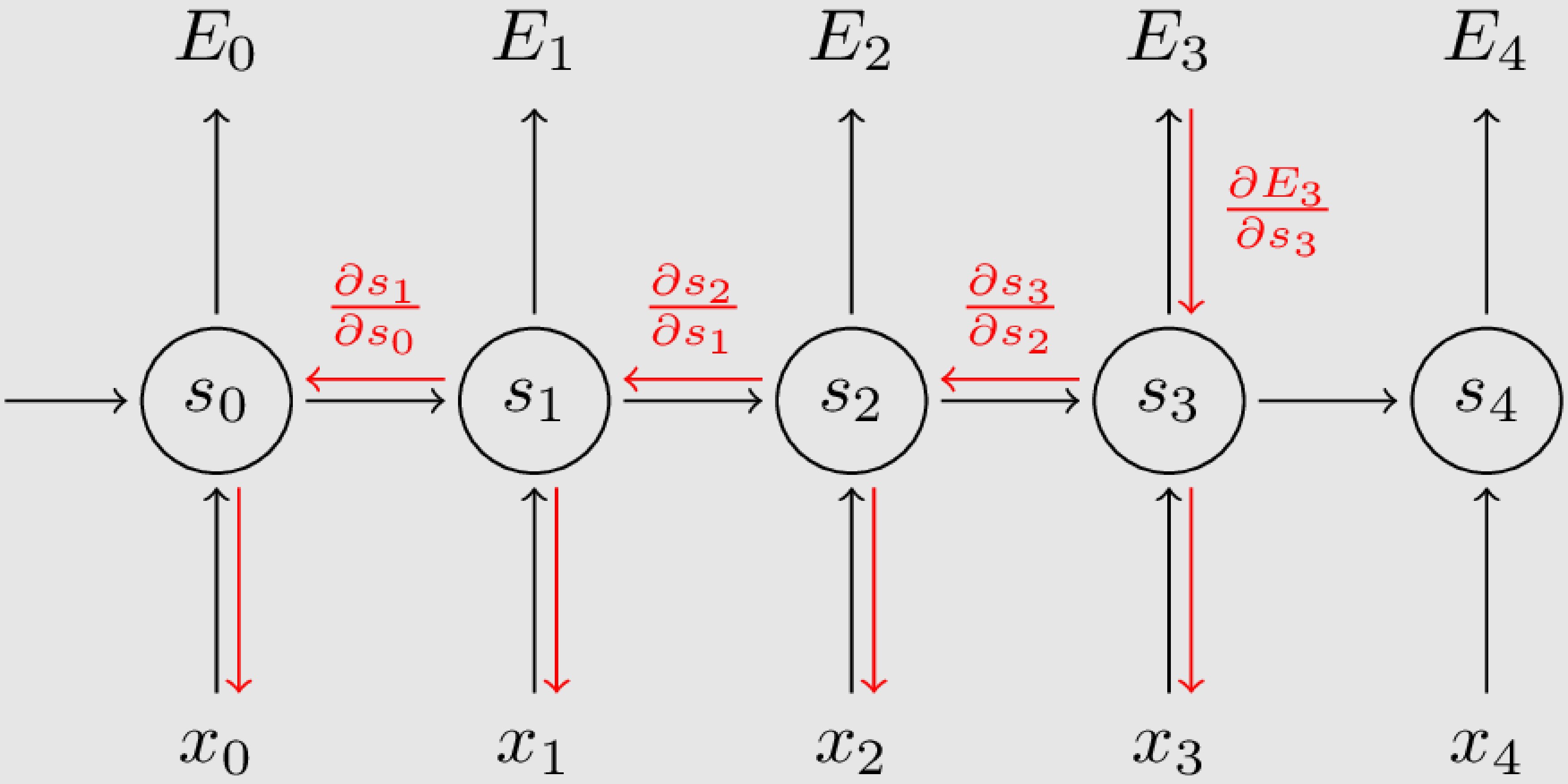
$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

$$W = W - learning\_rate * \frac{\partial E}{\partial W}$$

1보다 미분값이 클경우
Gradient Exploding → Gradient cliffing

01. 배경지식 (RNN)



$$s_t = \tanh(Ux_t + Ws_{t-1})$$

$$\hat{y}_t = softmax(Vs_t)$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial W_s}$$

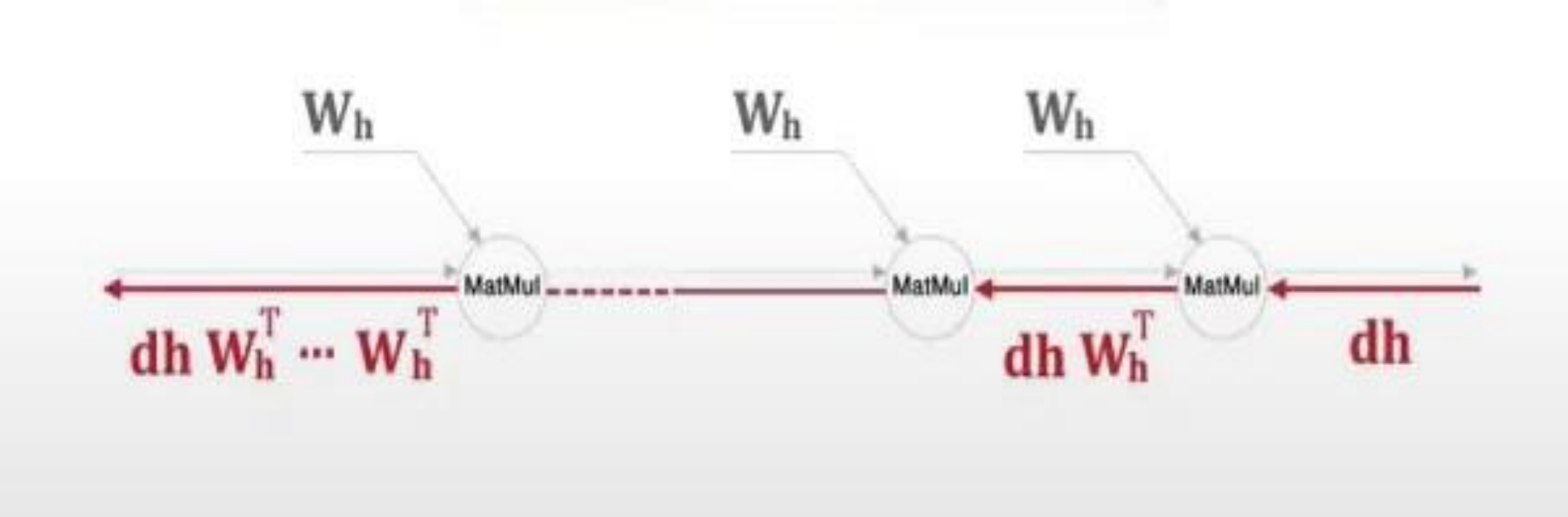
$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial W_s}$$

$$\frac{\partial E_3}{\partial W_s} = \frac{\partial E_3}{\partial \bar{y}_3}\frac{\partial \bar{y}_3}{\partial \bar{s}_3}\frac{\partial \bar{s}_3}{\partial \bar{s}_2}\frac{\partial \bar{s}_2}{\partial \bar{s}_1}\frac{\partial \bar{s}_1}{\partial W_s}$$

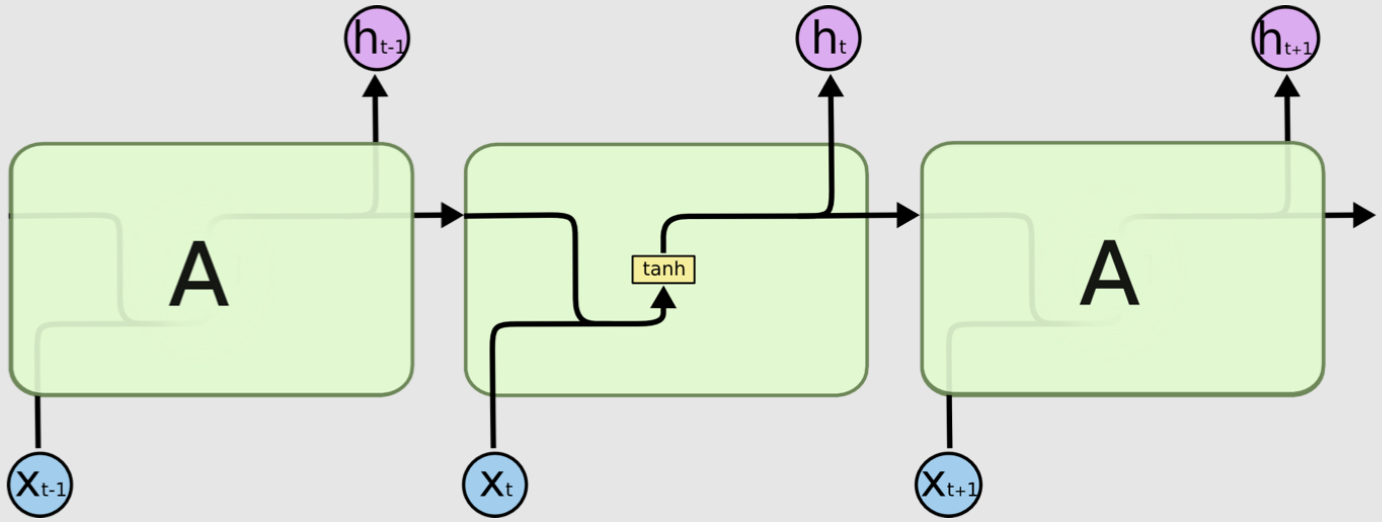
$$W = W - learning\_rate * \frac{\partial E}{\partial W}$$

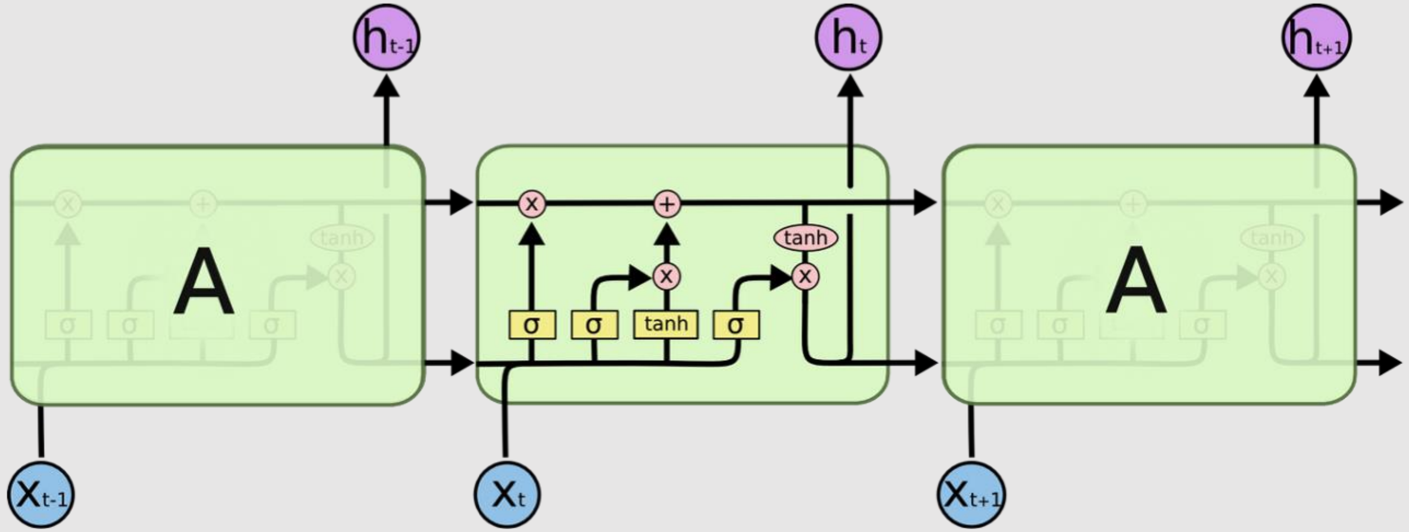
1보다 미분값이 작을경우
Gradient Vanishing

01. 배경지식 (RNN)

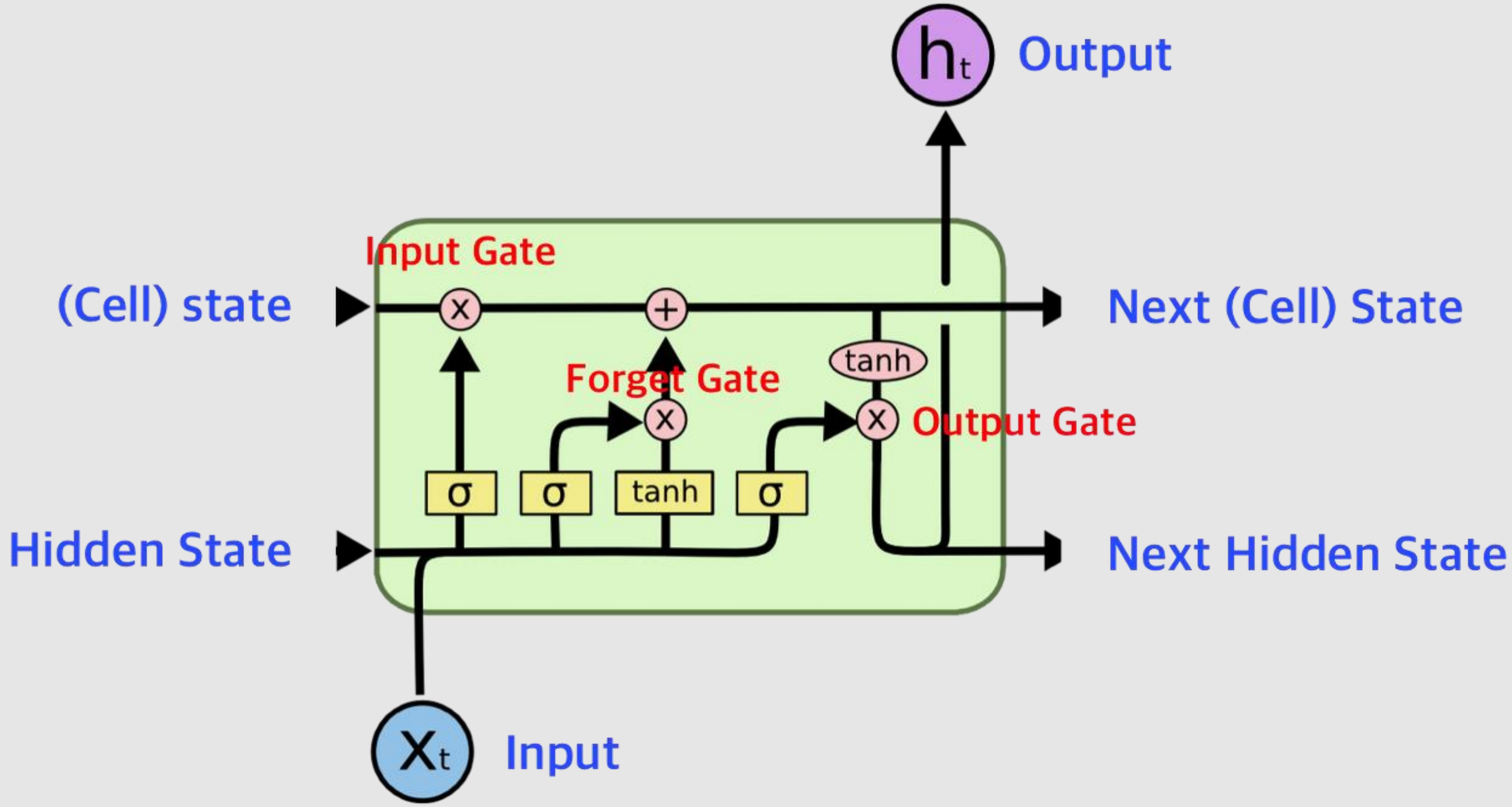


01. 배경지식 (LSTM)

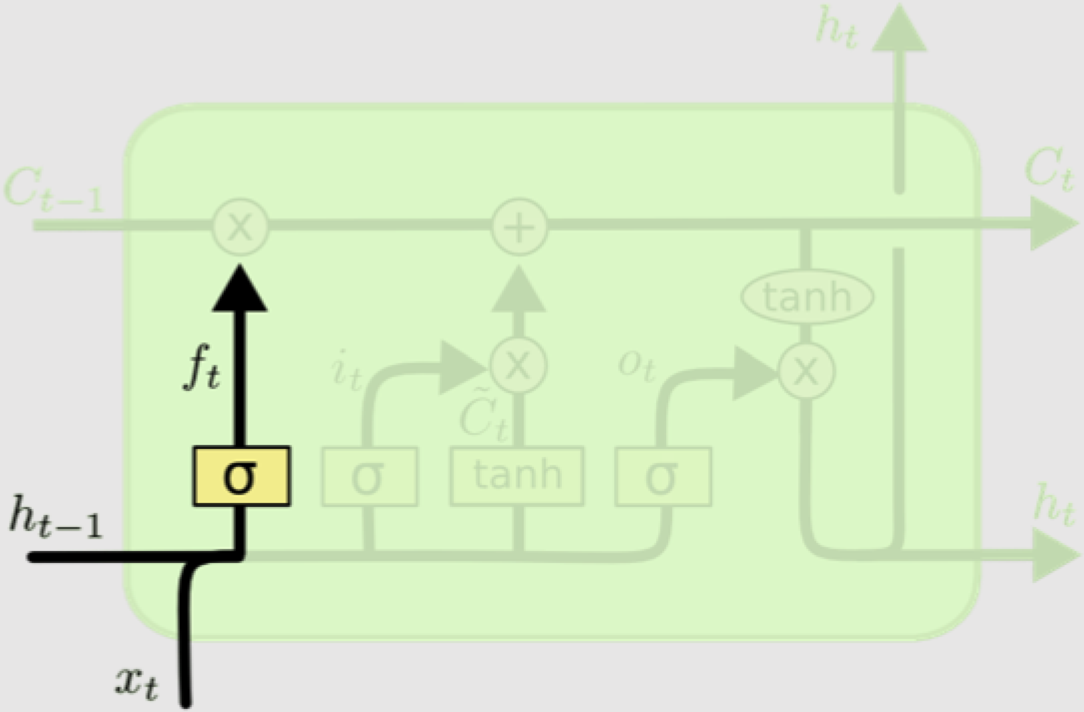




01. 배경지식 (LSTM)

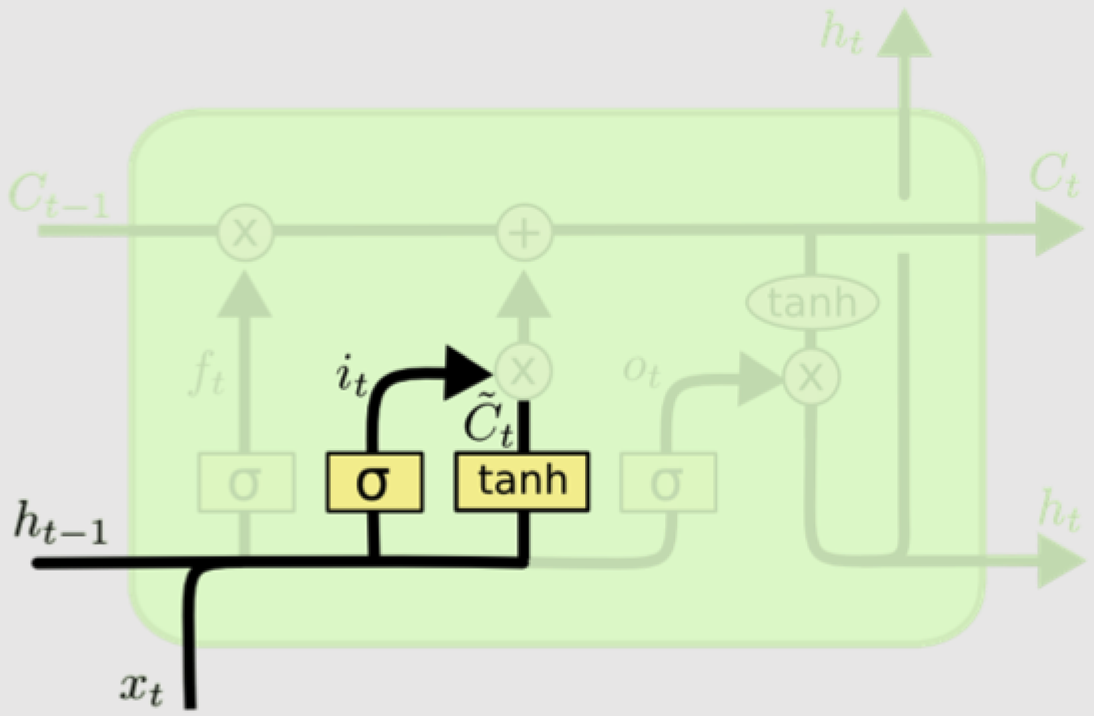

ht
Output
Input Gate
(Cell) state
Next (Cell) State
Forget Gate
tanh
Output Gate
σ
σ
tanh
σ
Hidden State
Next Hidden State
Xt
Input

Forget Gate



$$f_t = \sigma\left(W_f \cdot [h_{t-1}, x_t] \ + \ b_f\right)$$

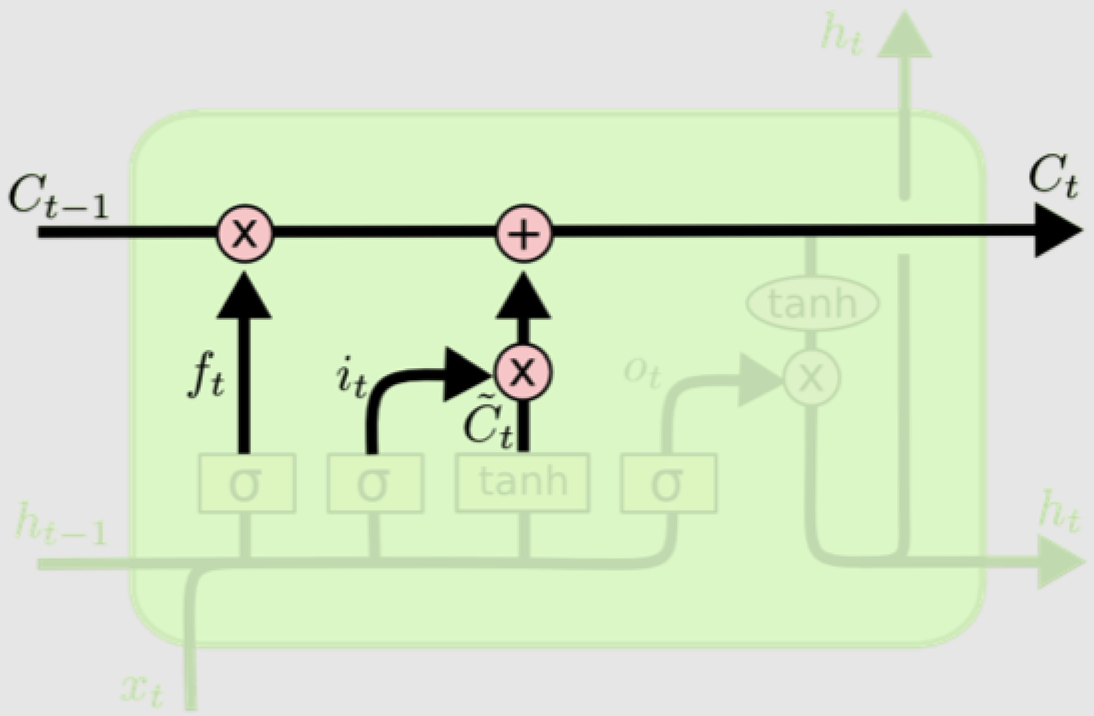
Input Gate



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

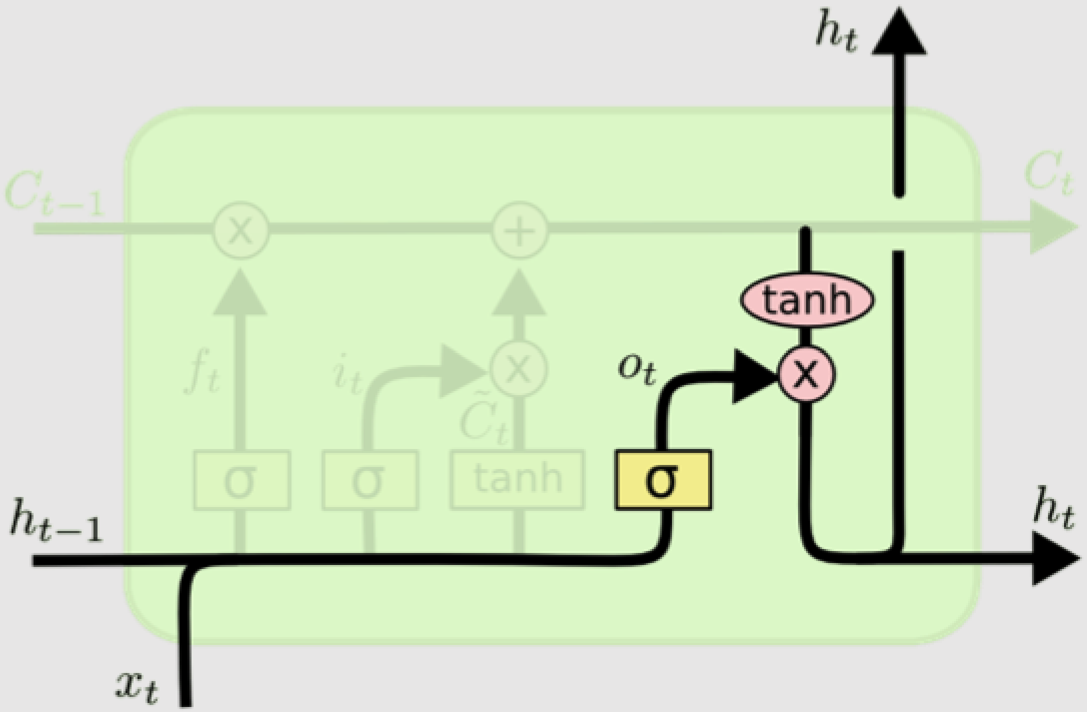
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$

Cell update



$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

Output Gate



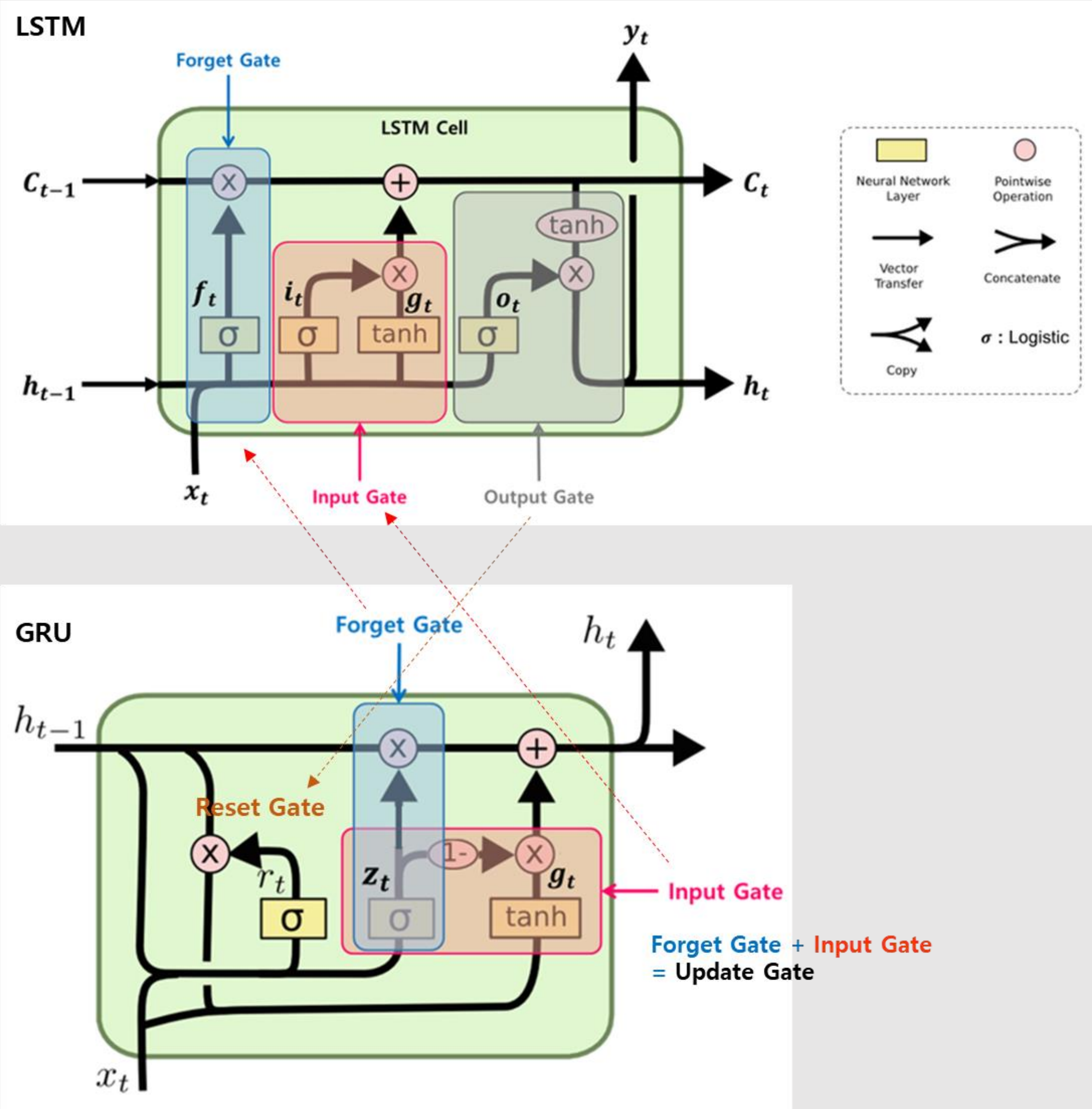
$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

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

02.GRU Model

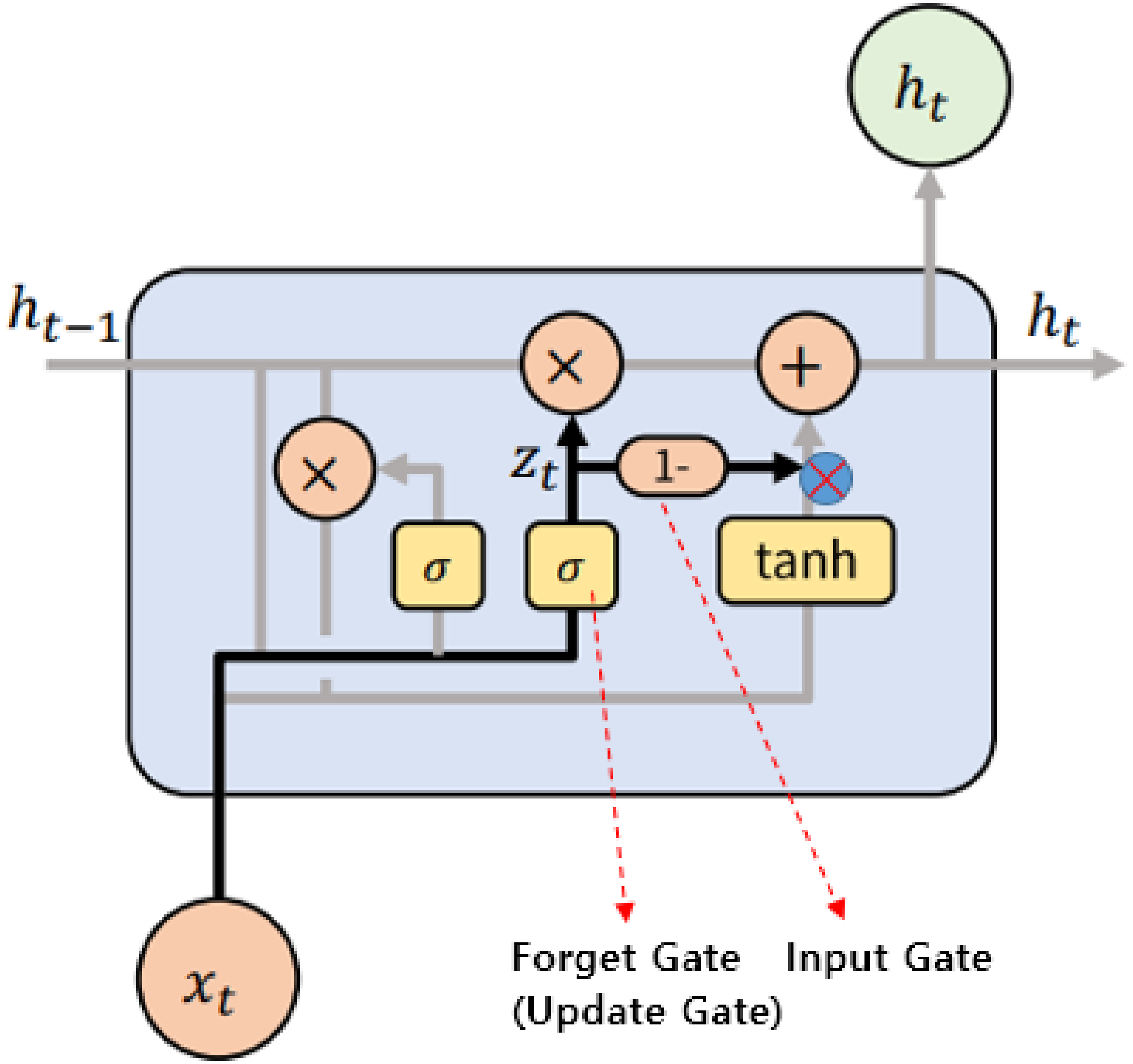
MADE BY HENDO

03. Model



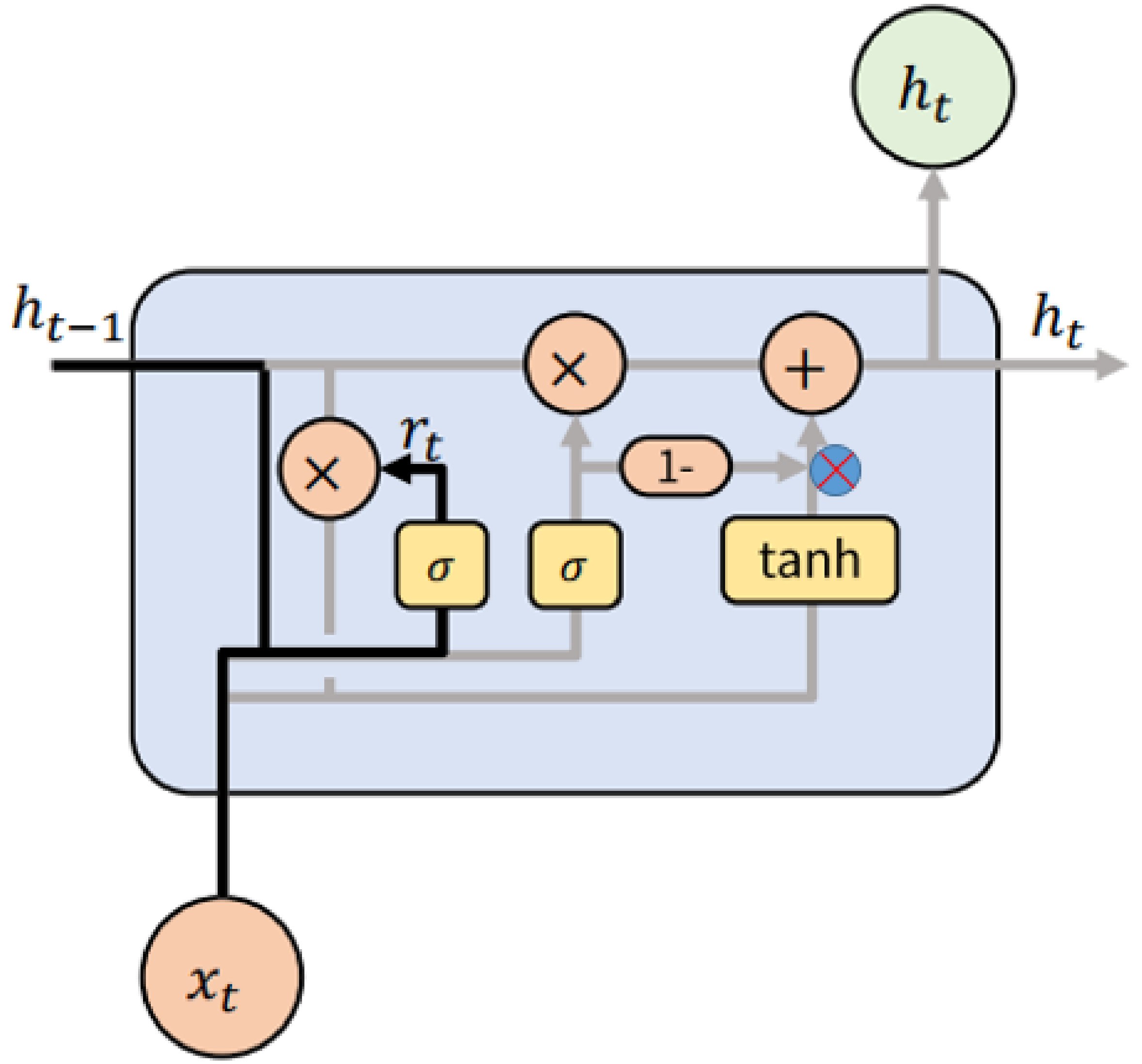
LSTM
Forget Gate
LSTM Cell
y_t
C_{t-1}
C_t
f_t
i_t
g_t
o_t
σ
tanh
h_{t-1}
h_t
x_t
Input Gate
Output Gate
Neural Network Layer
Pointwise Operation
Vector Transfer
Concatenate
Copy
σ : Logistic
GRU
Forget Gate
h_t
h_{t-1}
Reset Gate
r_t
z_t
1-
g_t
Input Gate
x_t
Forget Gate + Input Gate
= Update Gate

03. Model



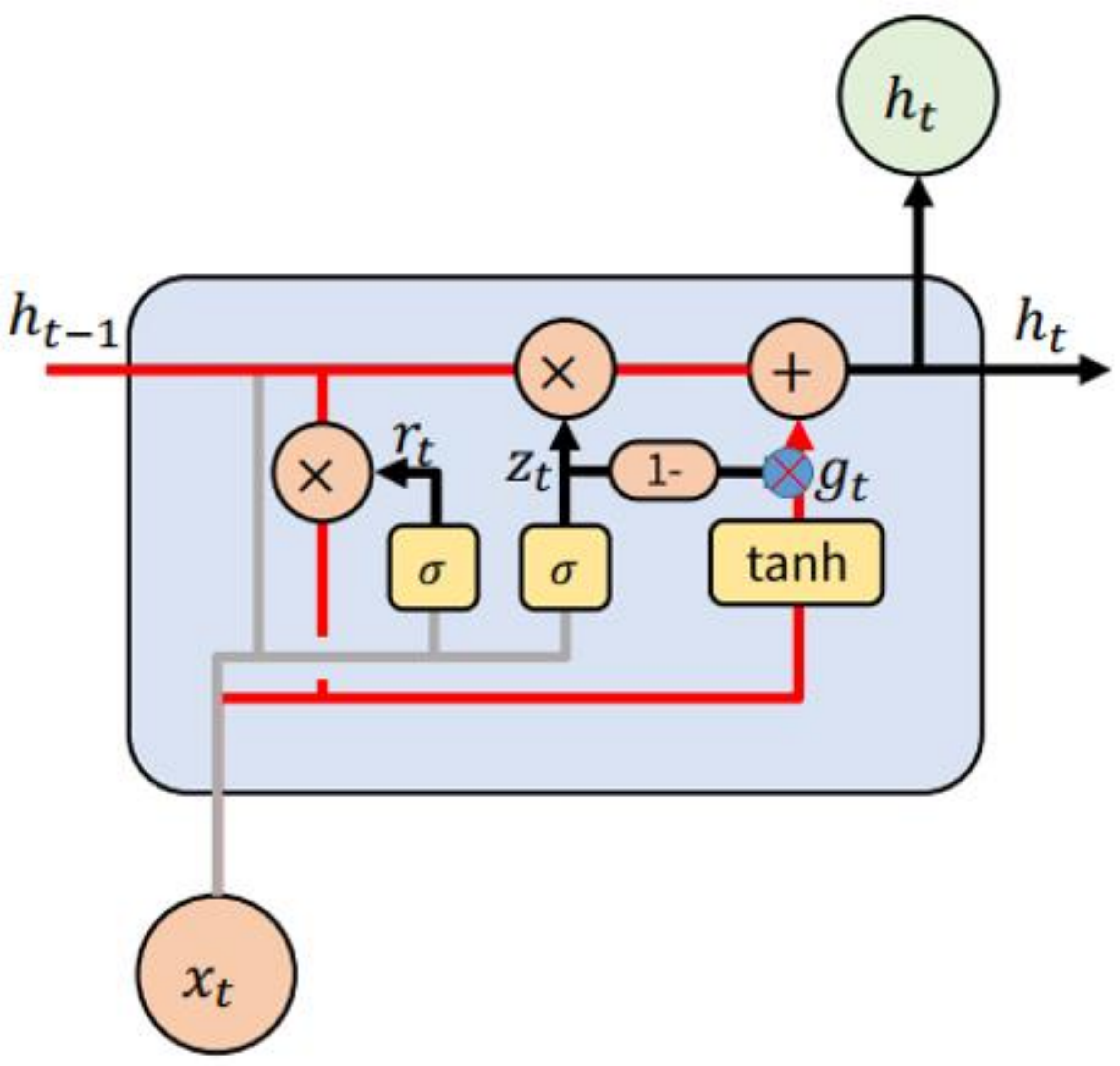
Gate	Equation
Reset Gate	$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)$
Forget Gate	$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)$
Input Gate	$1 - z_t$
Hidden State	$g_t = tanh(W_{xg}x_t + W_{hg}(r_t \odot h_{t-1}) + b_g)$ $h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g_t$

03. Model



Gate	Equation
Reset Gate	$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)$
Forget Gate	$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)$
Input Gate	$1 - z_t$
Hidden State	$g_t = tanh(W_{xg}x_t + W_{hg}(r_t \odot h_{t-1}) + b_g)$
	$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g_t$

03. Model



Gate	Equation
Reset Gate	$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1} + b_r)$
Forget Gate	$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1} + b_z)$
Input Gate	$1 - z_t$
Hidden State	$g_t = tanh(W_{xg}x_t + W_{hg}(r_t \odot h_{t-1}) + b_g)$
	$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot g_t$

03. Model

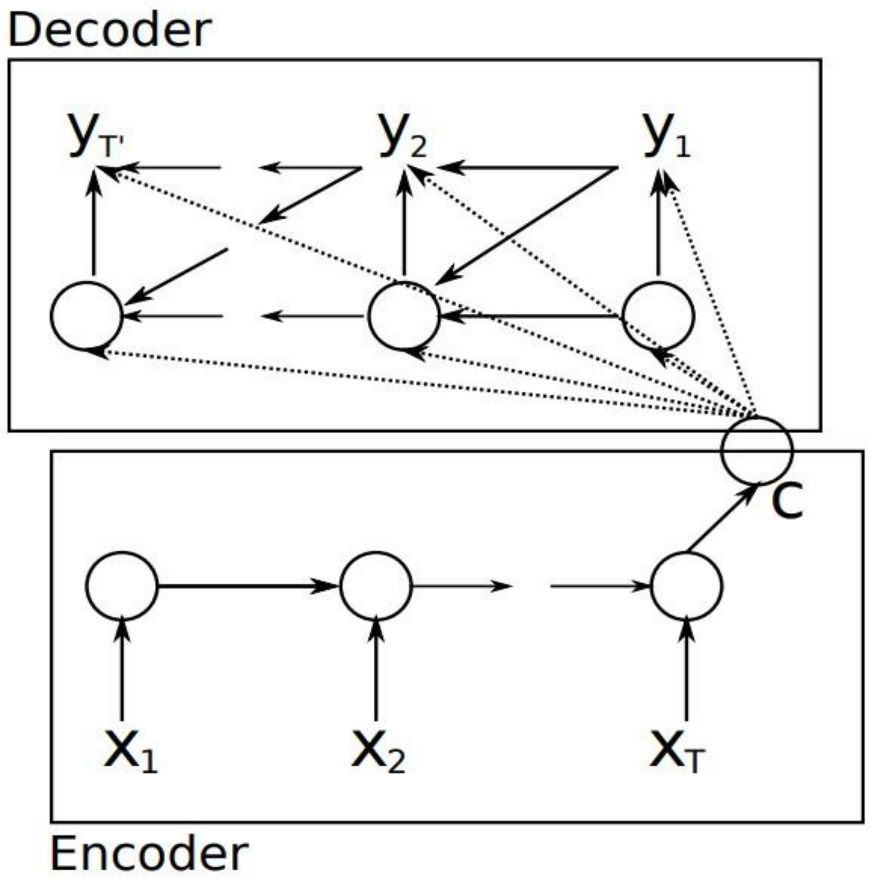


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

$$p(y_1, \ldots, y_{T'} \mid x_1, \ldots, x_T),$$

Encoder

$$\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right),$$

03. Model

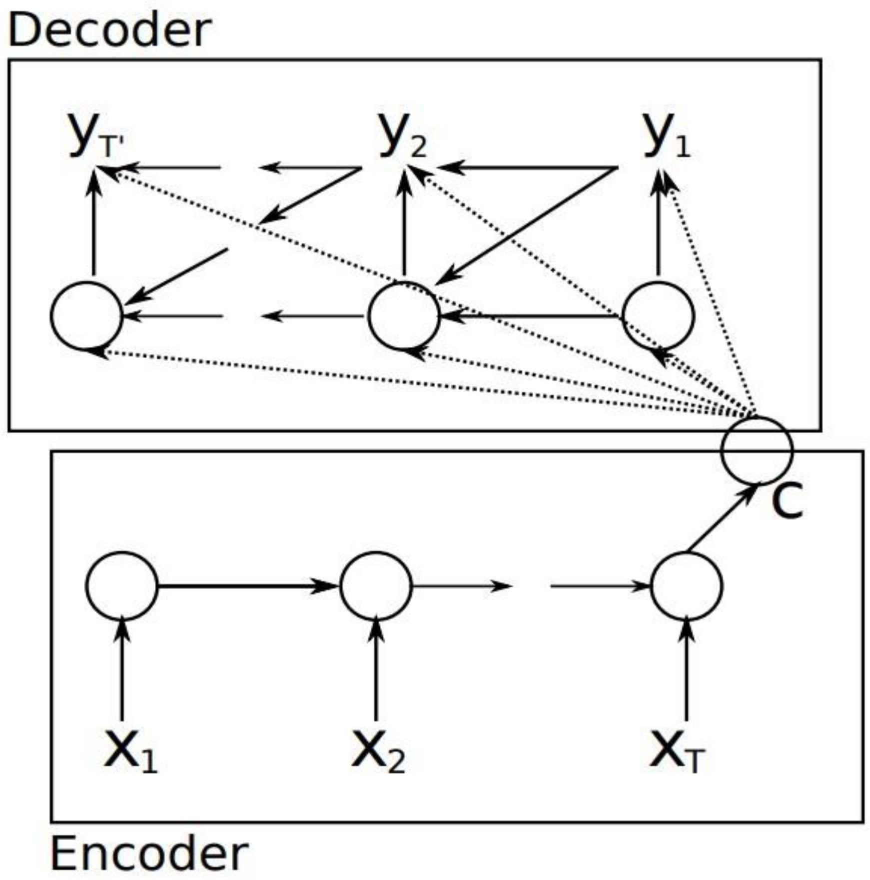


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

$$p(y_1, \ldots, y_{T'} \mid x_1, \ldots, x_T),$$

Encoder → $\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, x_t\right),$

Decoder → $\mathbf{h}_{\langle t \rangle} = f\left(\mathbf{h}_{\langle t-1 \rangle}, y_{t-1}, \mathbf{c}\right),$

03. Model

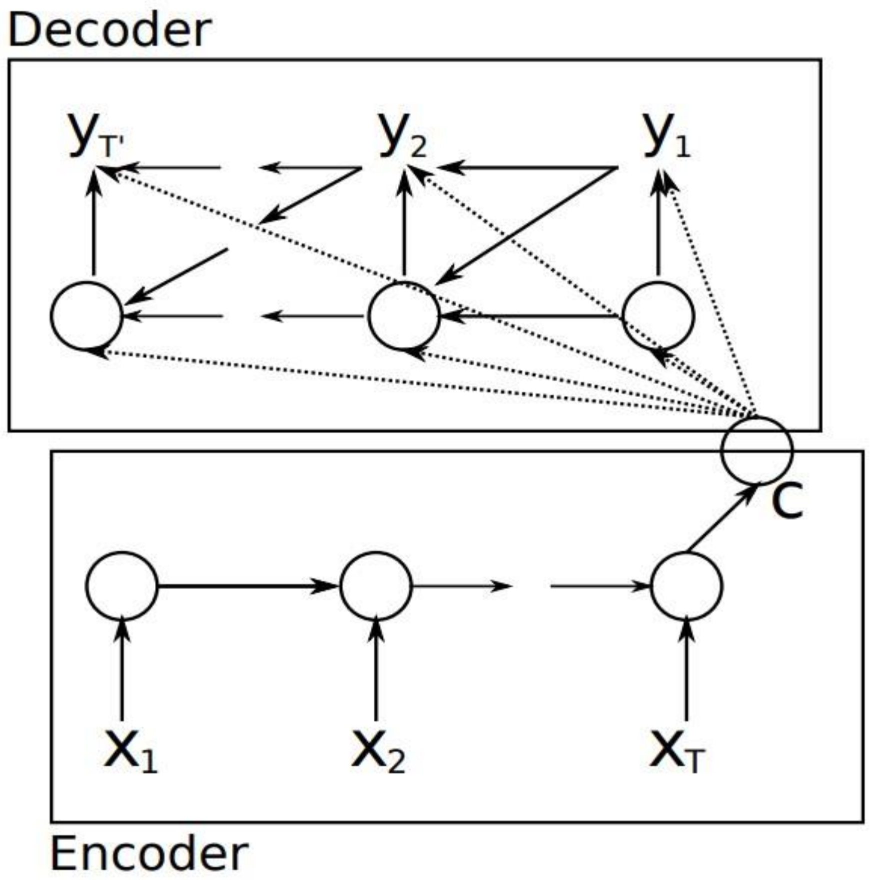
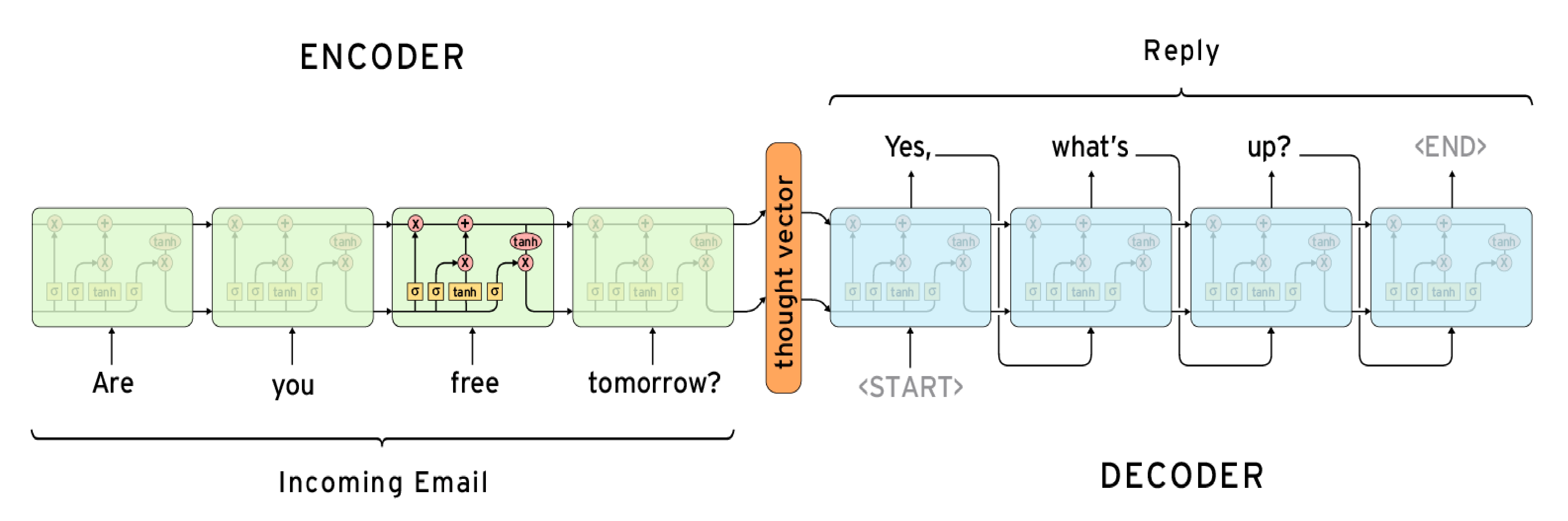


Figure 1: An illustration of the proposed RNN Encoder–Decoder.

$$P(y_t | y_{t-1}, y_{t-2}, \ldots, y_1, \mathbf{c}) = g\left(\mathbf{h}_{\langle t \rangle}, y_{t-1}, \mathbf{c}\right).$$

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{y}_n \mid \mathbf{x}_n),$$

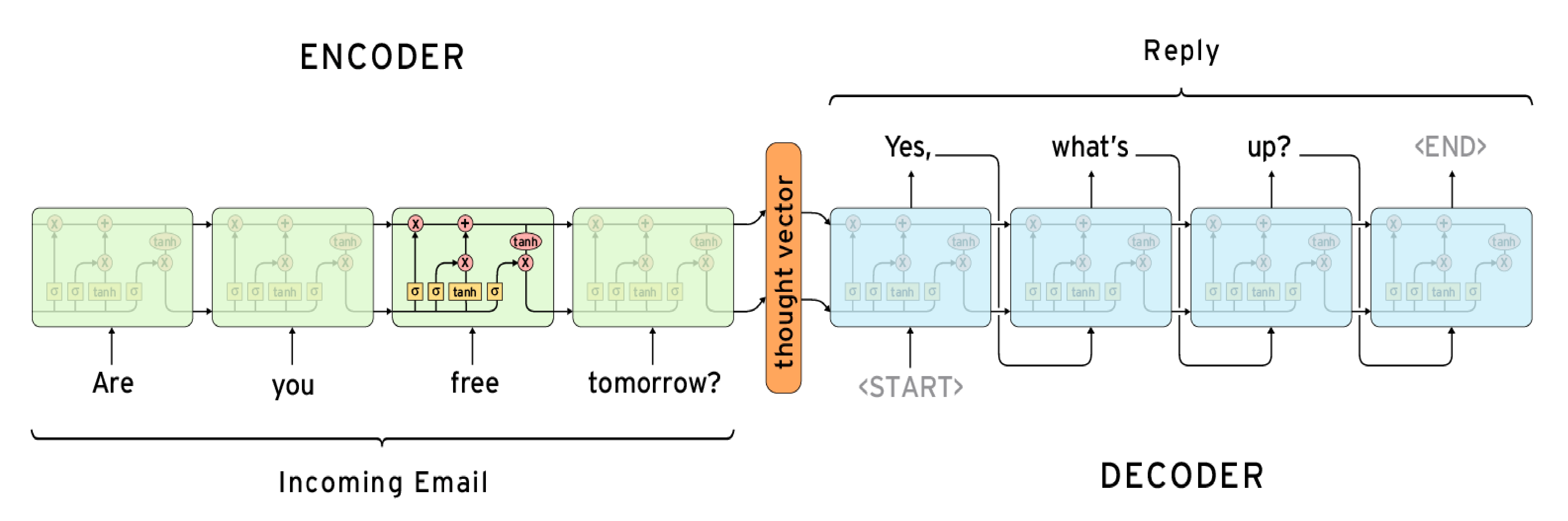
03. Model



학습목표

$$\max_{\boldsymbol{\theta}} \frac{1}{N} \sum_{n=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{y}_n \mid \mathbf{x}_n),$$

03. Model



학습목표 →

$$\max_{\theta} \frac{1}{4}\left(\log P_{(Are)} + \log P_{(you)} + \log P_{(free)} + \log P_{(tomorrow)}\right)$$

03. Model-smt

$$p(e|f)$$

F가 주어졌을때 e가 나올 확률

$$p(e|f) \propto p(f|e)p(e)$$

비례관계

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

가장 높은 확률로 나오는 표현을 골라 번역 e를 찾는다

03. Model-활용

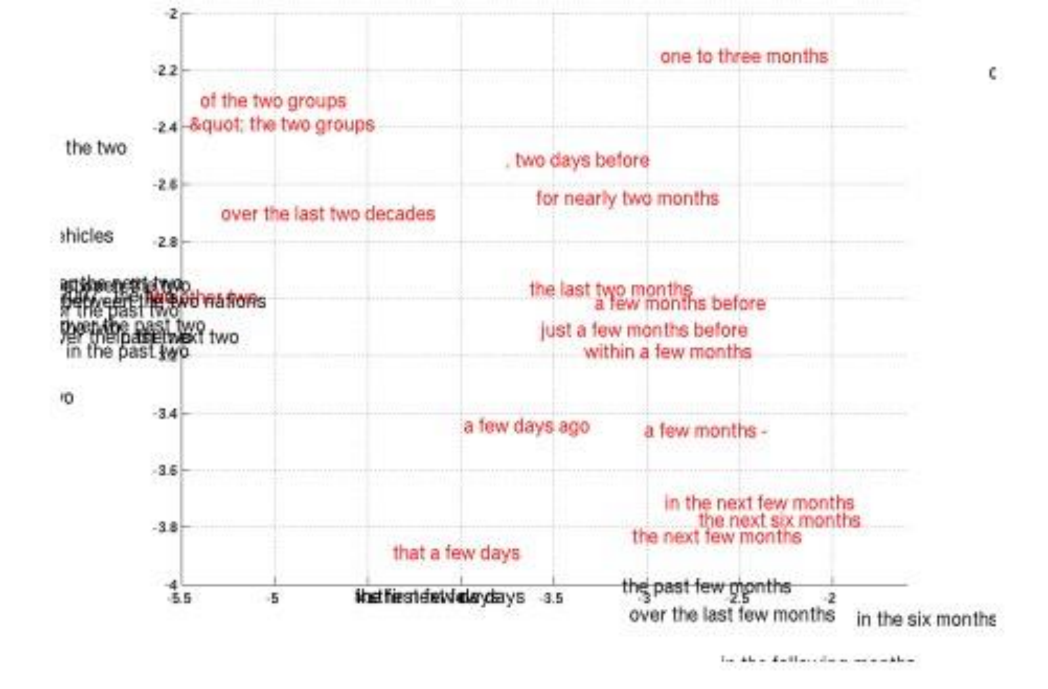
$$\log p(\mathbf{f} \mid \mathbf{e}) = \sum_{n=1}^{N} w_n f_n(\mathbf{f}, \mathbf{e}) + \log Z(\mathbf{e}), \quad (9)$$

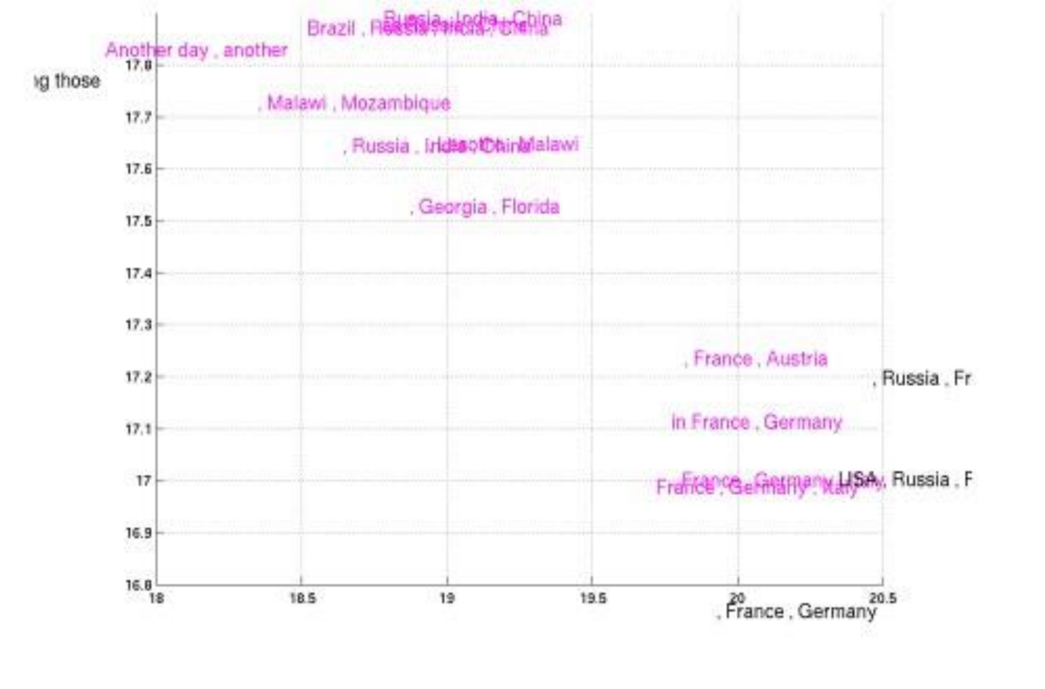
03. Result & Conclusion

03. Model-활용

Models	BLEU	
	dev	test
Baseline	30.64	33.30
RNN	31.20	33.87
CSLM + RNN	31.48	34.64
CSLM + RNN + WP	31.50	34.54

03. Model-활용


one to three months
of the two groups
" the two groups
the two
, two days before
for nearly two months
over the last two decades
the last two months
a few months before
just a few months before
within a few months
in the past two
a few days ago
a few months -
in the next few months
the next six months
the next few months
that a few days
the past few months
over the last few months
in the six months


Another day , another
Brazil ,
, Malawi , Mozambique
, Georgia , Florida
, France , Austria
, Russia , Fr
In France , Germany
, France , Germany

THANK YOU!

MADE BY HENDO