LSTM (Long Short-Term Memory)

Byeongjin Kang
qudwlskbj@gmail.com
NLP Team
2024/03/26

TRAIN AND TEST

Contents

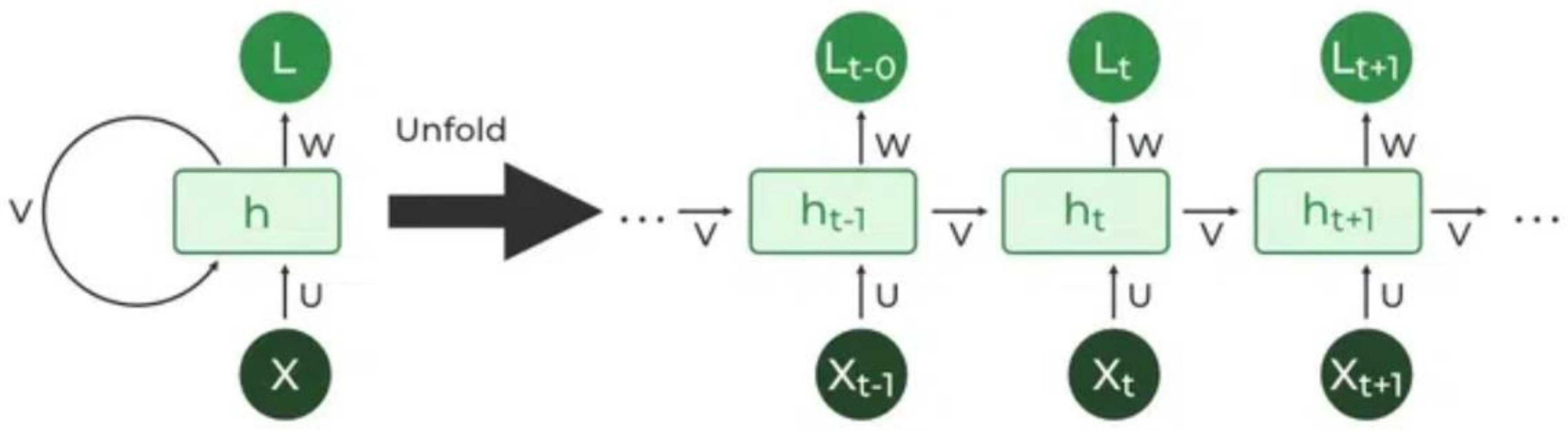
- Problems in RNN?
- What is LSTM?
- Paper review - *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*

TRAIN AND TEST

RNN

RNN (Recurrent Neural Networks)

RNN

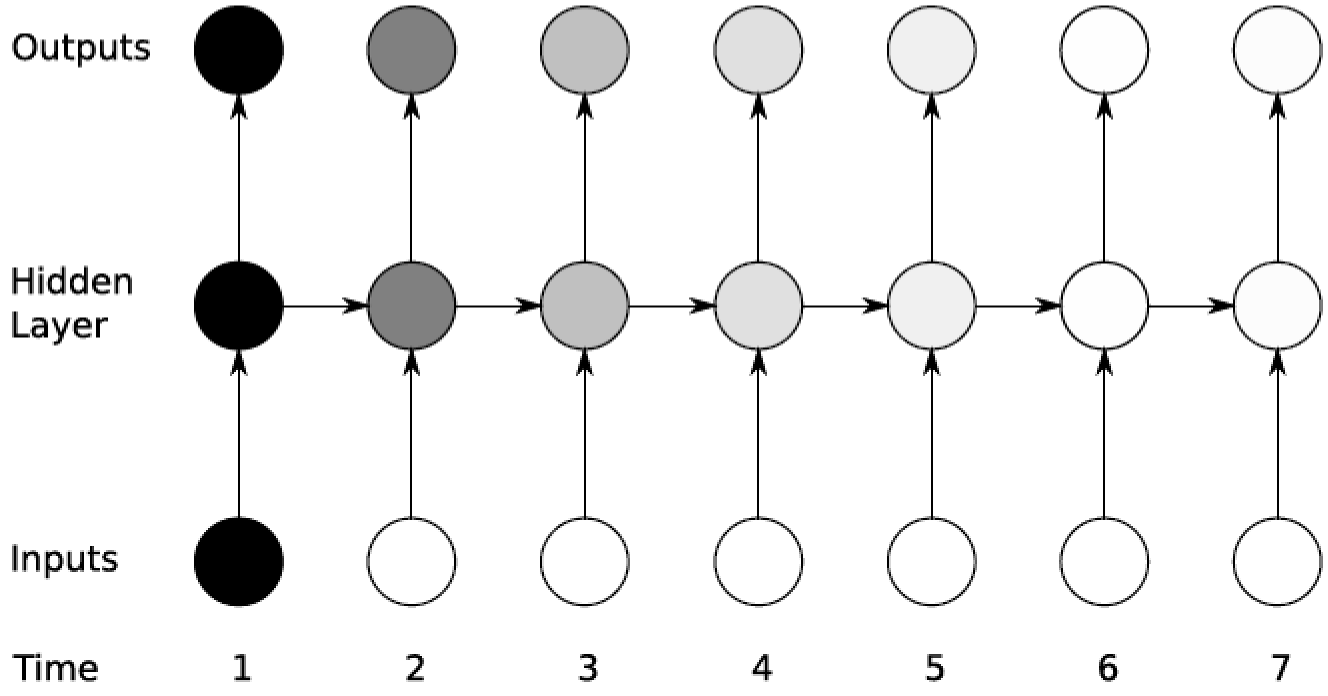


TRAIN AND TEST

Problems in RNN

Tom was watching TV in his room. Mary came into the room. Mary said hi to $\boxed{?}$

Problems in RNN



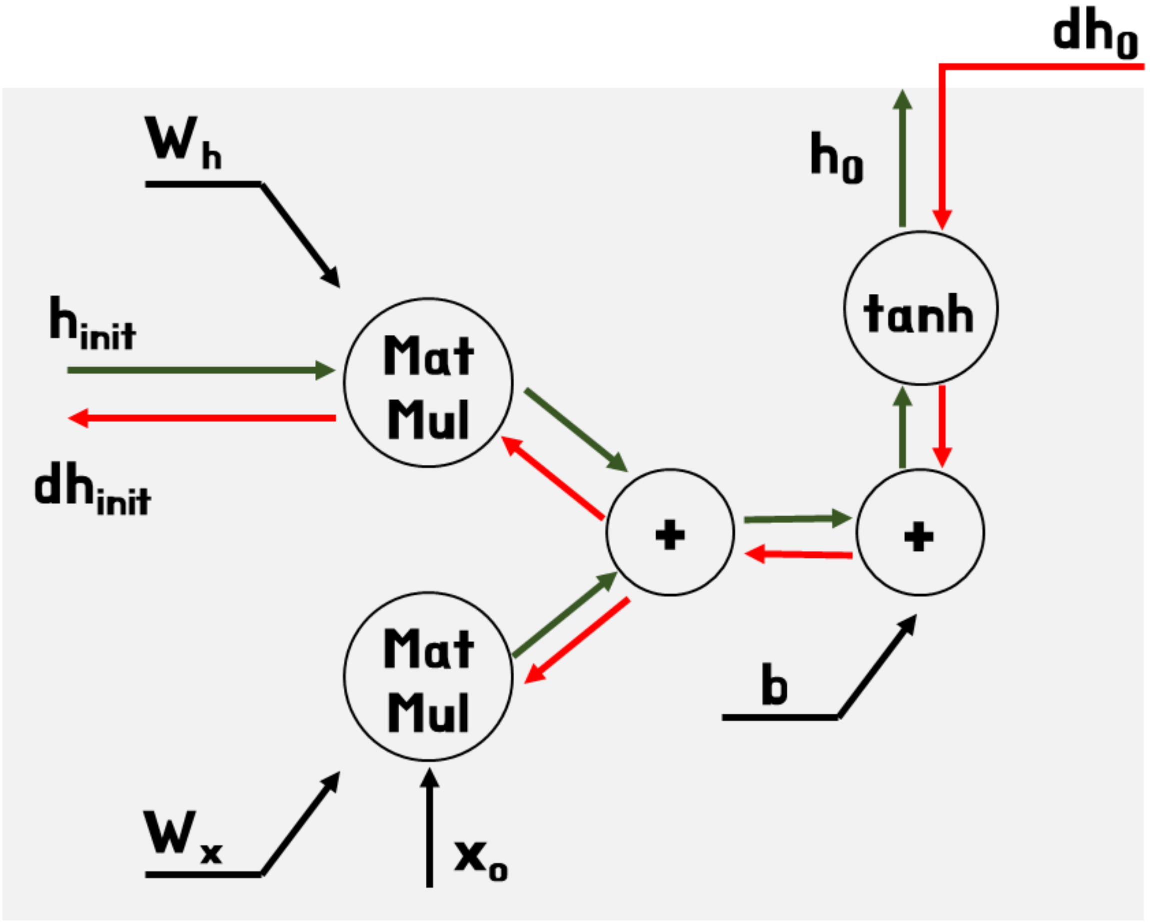
장기 의존성 문제 (the problem of Long-Term Dependencies)

TRAIN AND TEST

Problems in RNN

$$h_t = \tanh(h_{t-1}W_h + x_tW_x + b)$$

Computational graph



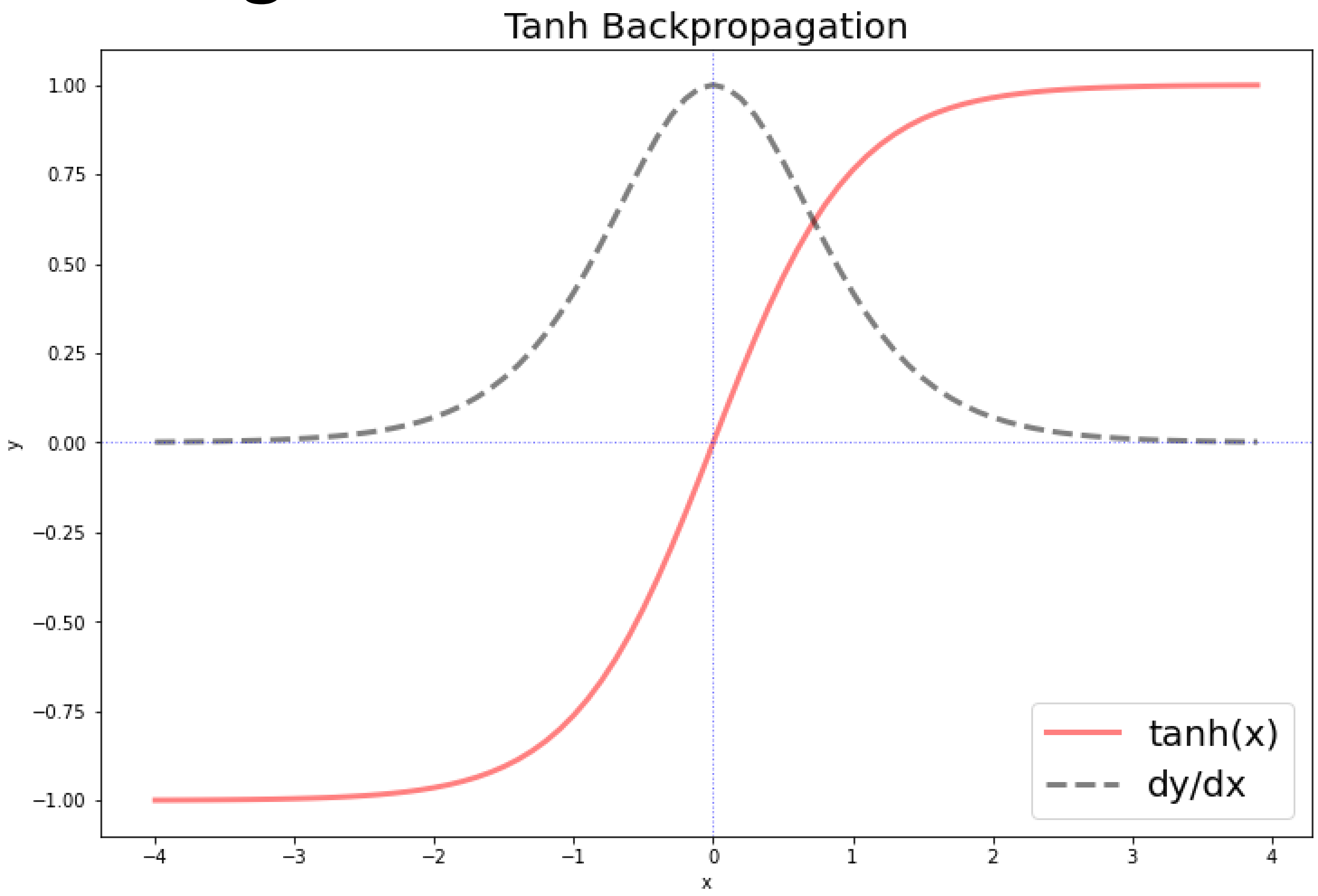
TRAIN AND TEST

Problems in RNN

$$y = \tanh(x) \xrightarrow[\text{미분}]{} \frac{\partial y}{\partial x} = 1 - y^2$$

Problems in RNN

Gradient Vanishing



TRAIN AND TEST

Problems in RNN

Sigmoid function?

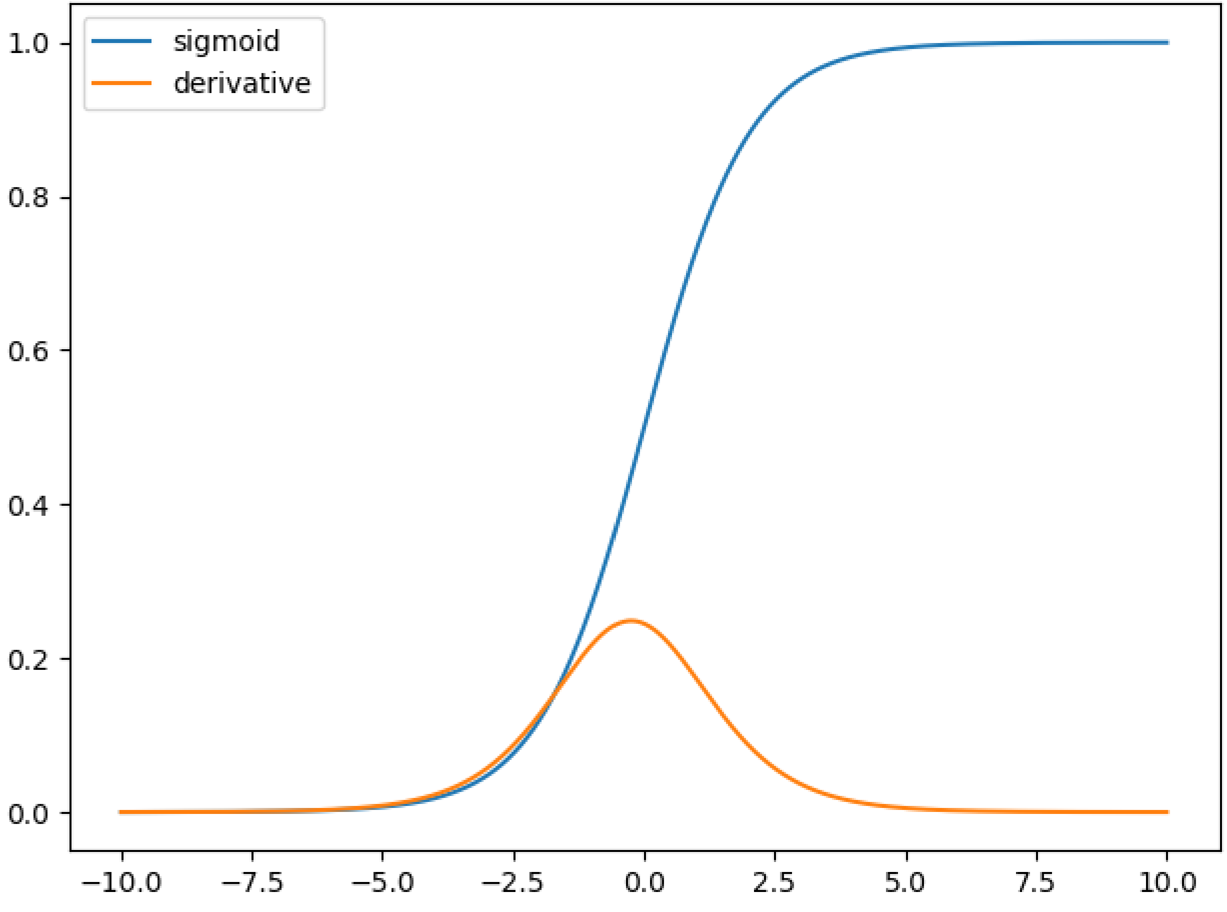
$$f(x) = \frac{1}{1 + e^{-x}}$$

TRAIN AND TEST

Problems in RNN

Sigmoid function?

$$f(x) = \frac{1}{1 + e^{-x}}$$



TRAIN AND TEST

Problems in RNN

ReLU function?

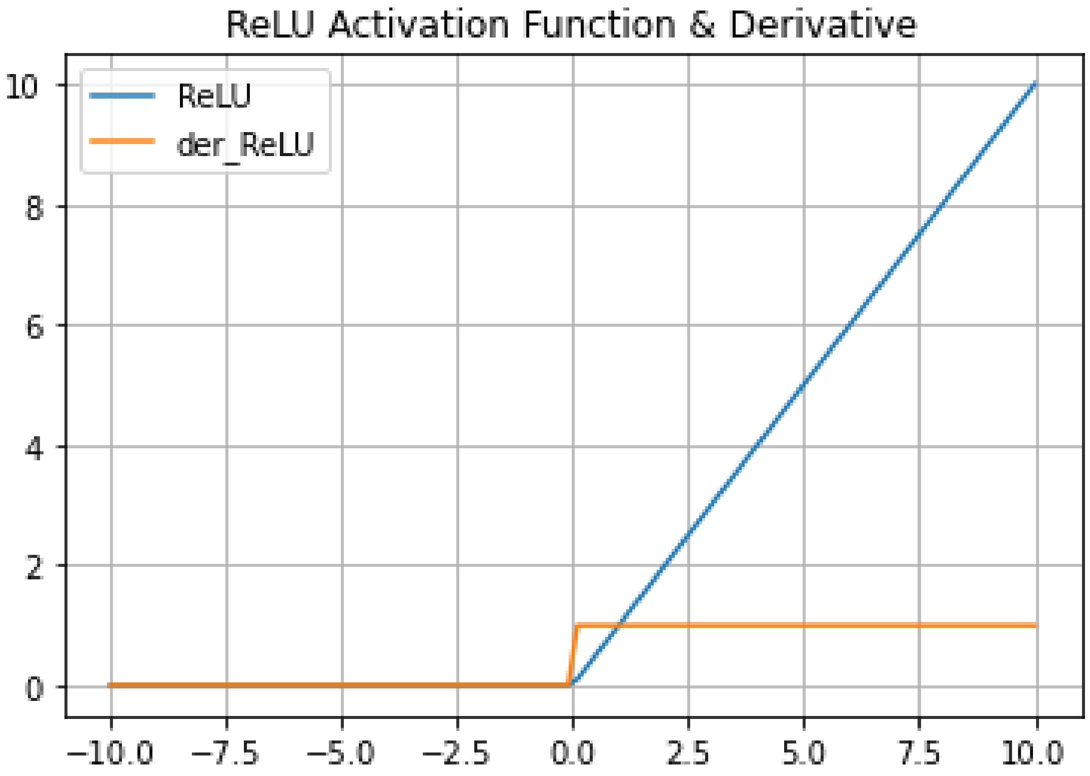
$$\sigma(x) = \begin{cases} max(0, x) & , x >= 0 \\ 0 & , x < 0 \end{cases}$$

TRAIN AND TEST

Problems in RNN

ReLU function?

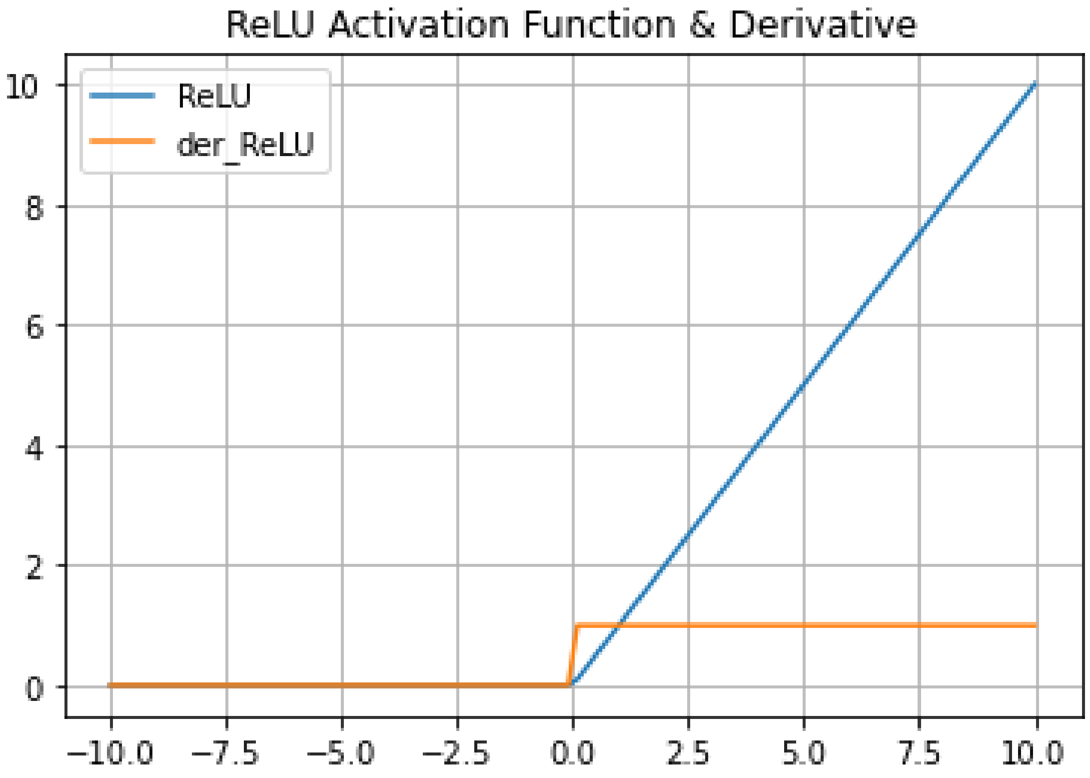
$$\sigma(x) = \begin{cases} max(0, x) & , x >= 0 \\ 0 & , x < 0 \end{cases}$$



Problems in RNN

ReLU function?

$$\sigma(x) = \begin{cases} max(0, x) & , x >= 0 \\ 0 & , x < 0 \end{cases}$$

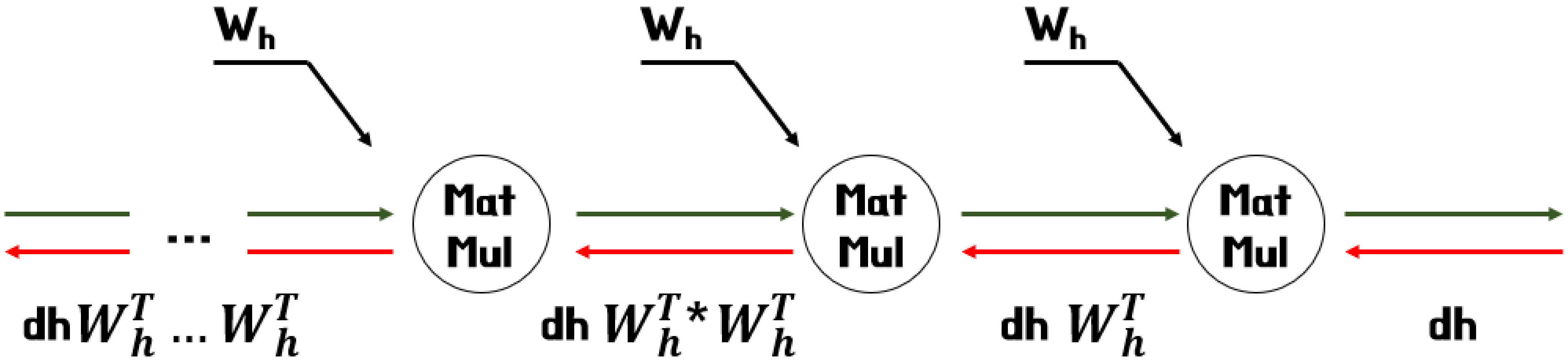


Q. Why don't we use ReLU in RNN?

TRAIN AND TEST

Problems in RNN

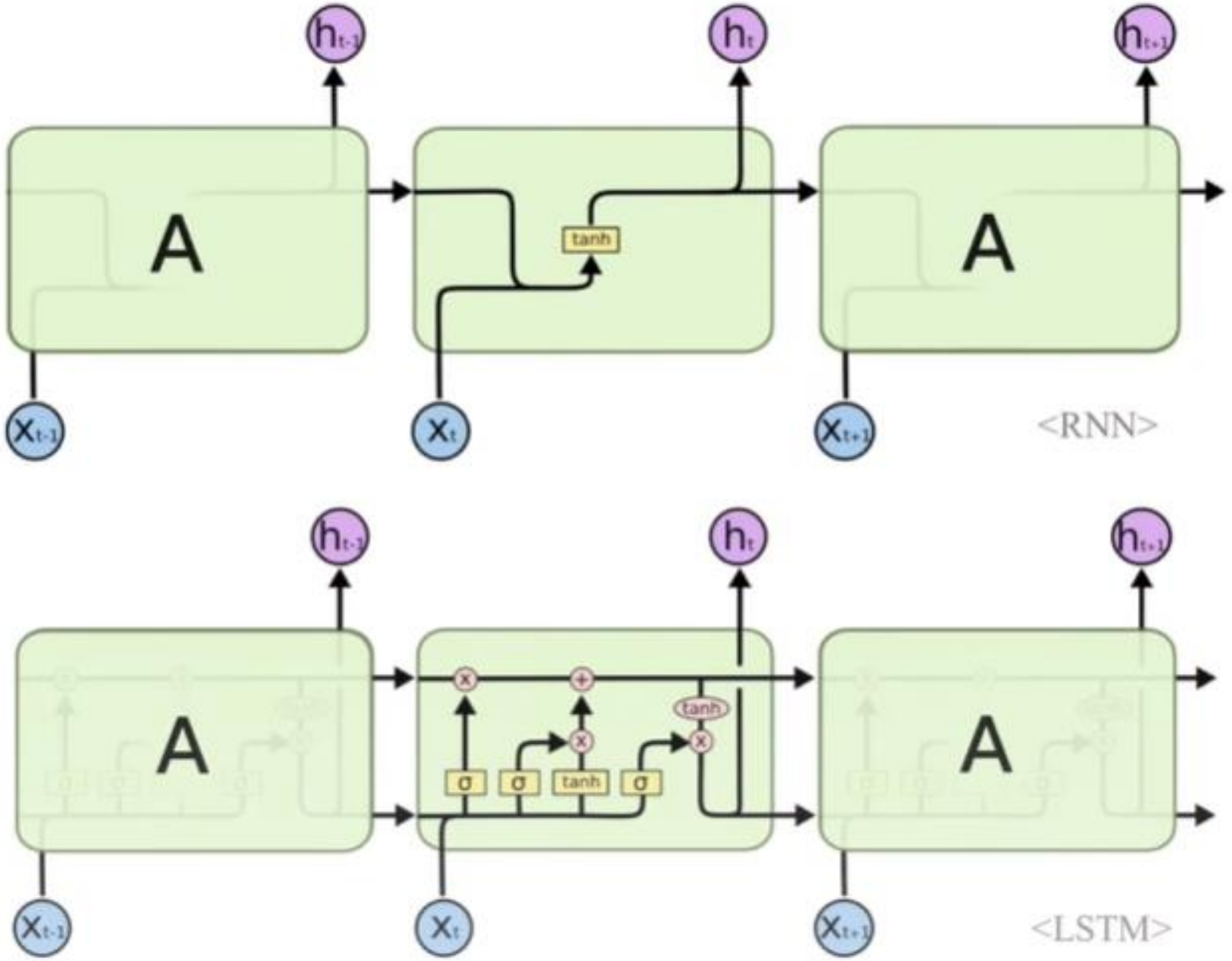
Gradient Exploding



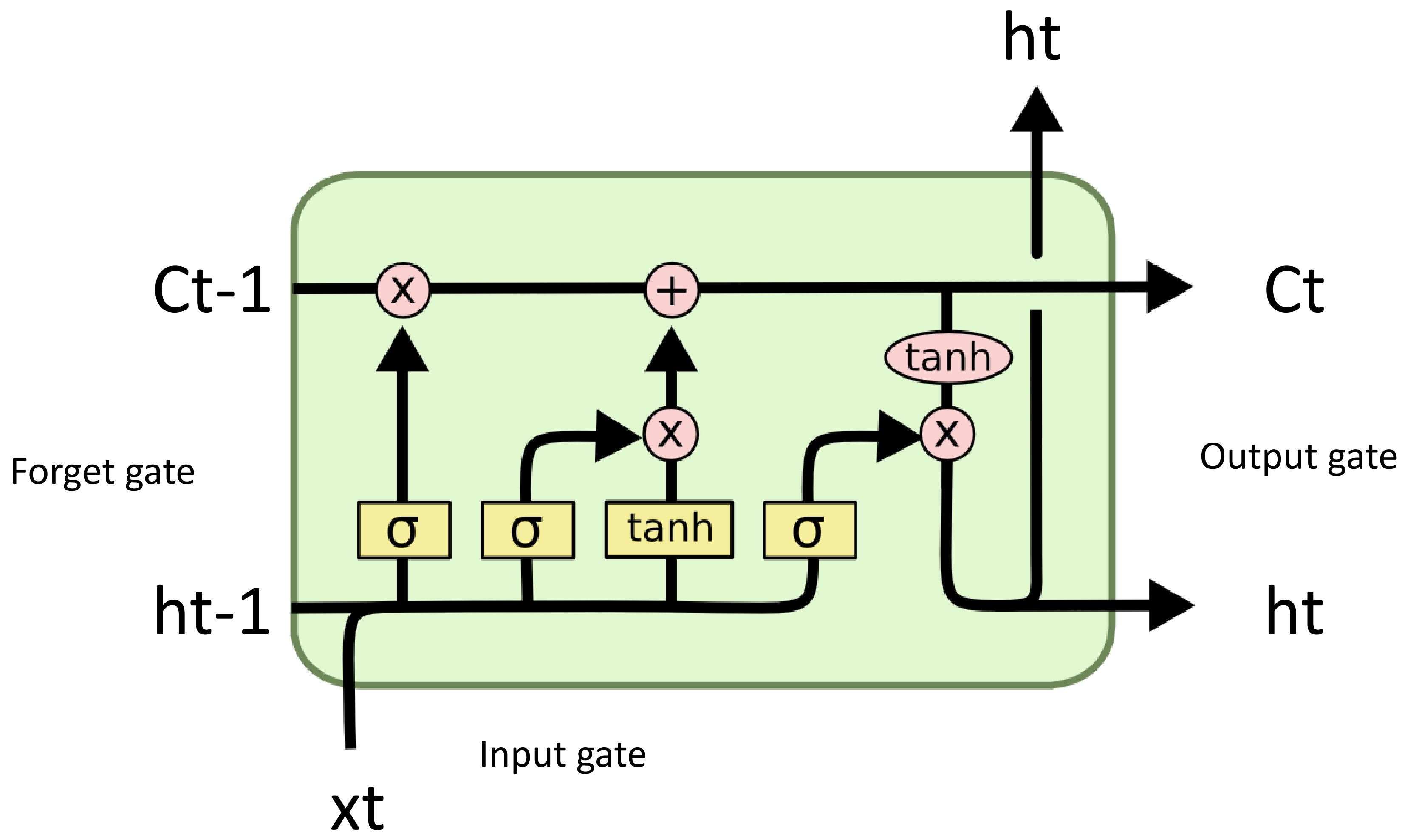
LSTM

LSTM (Long Short-Term Meomory)

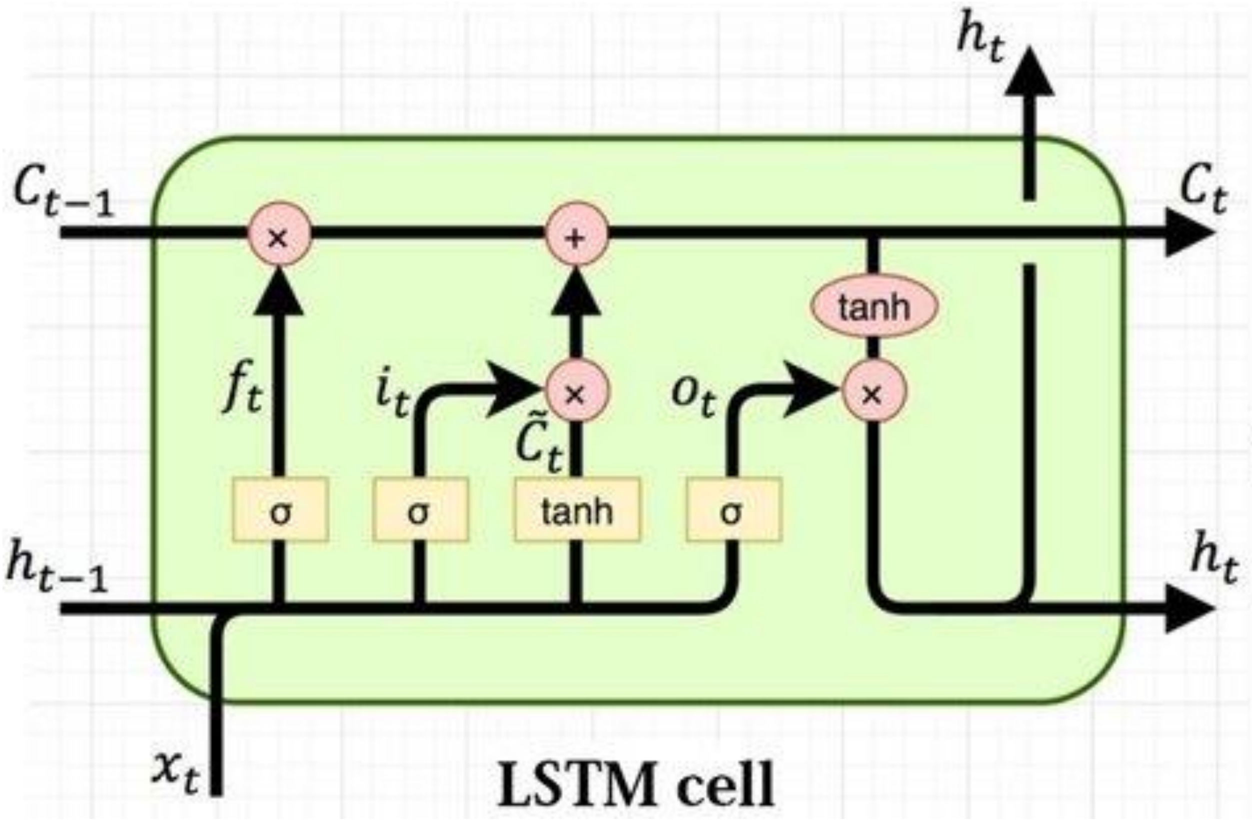
LSTM



LSTM


ht
Ct-1
x
+
Ct
tanh
x
x
Forget gate
Output gate
σ
σ
tanh
σ
ht-1
ht
Input gate
xt

LSTM



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

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

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

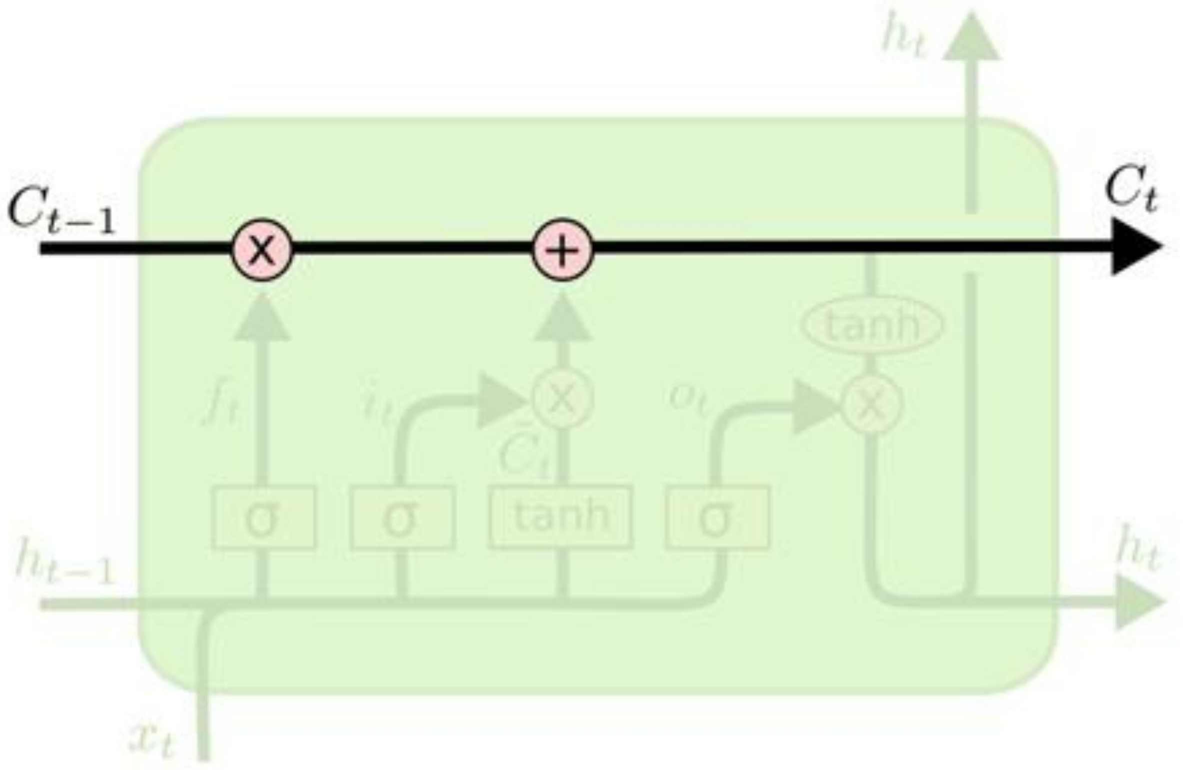
$$\tilde{C}_t = \tanh\left(x_t U^g + h_{t-1} W^g\right)$$

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

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

LSTM

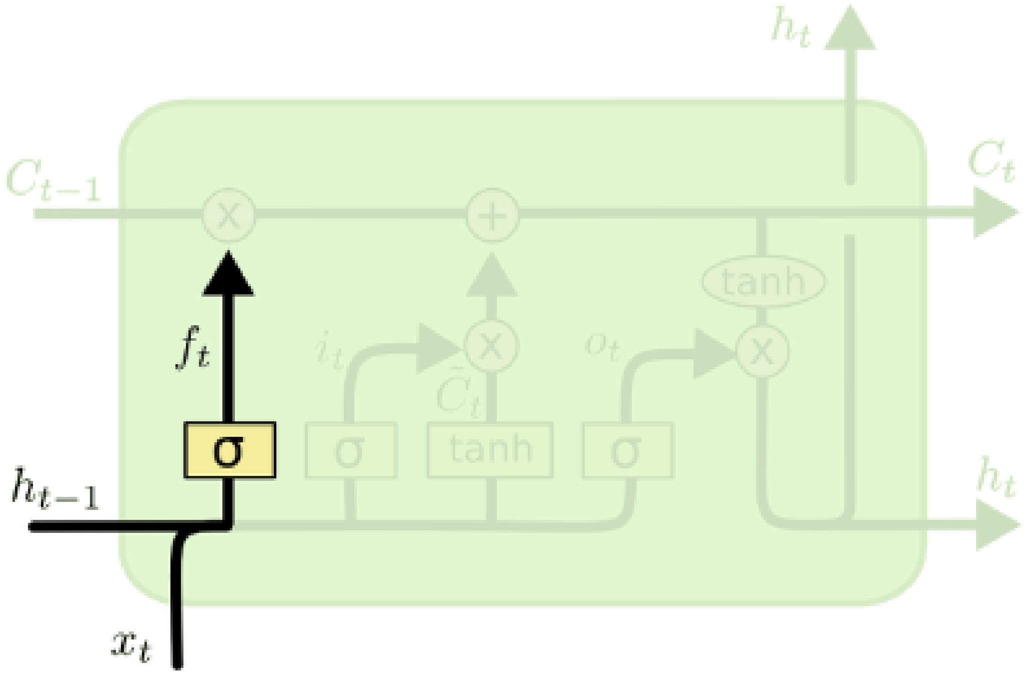
Cell State



Prevent Vanishing gradient

LSTM

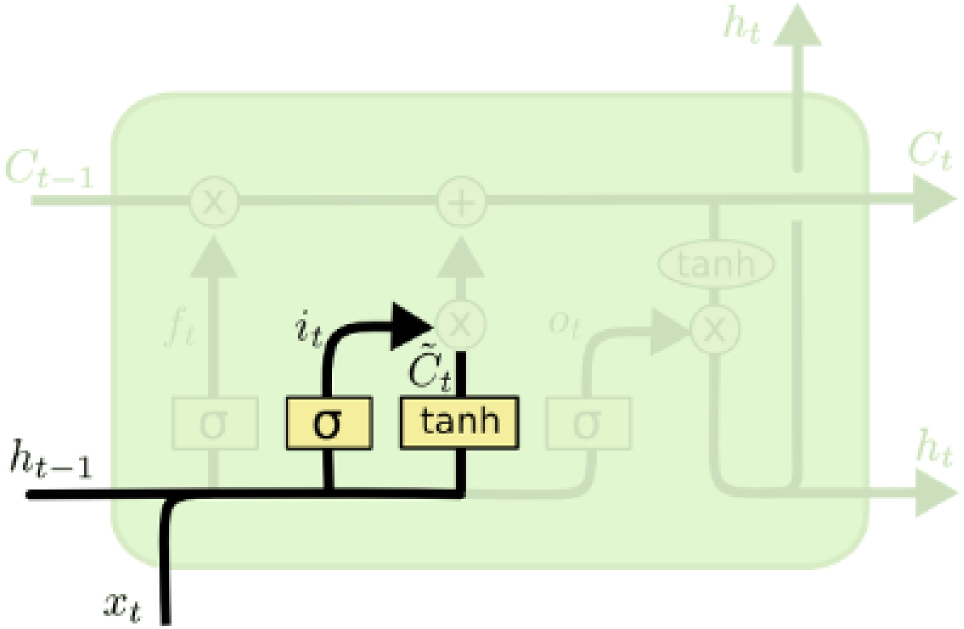
Forget State



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

LSTM

Input State



About Input(x)

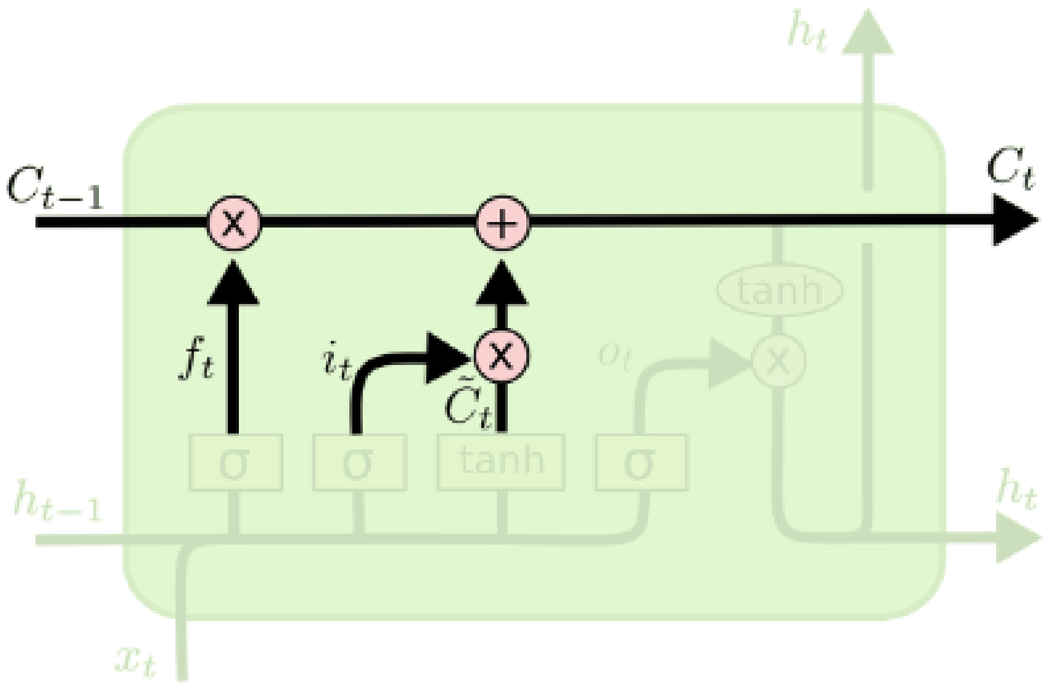
$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

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

TRAIN AND TEST

LSTM

State Update

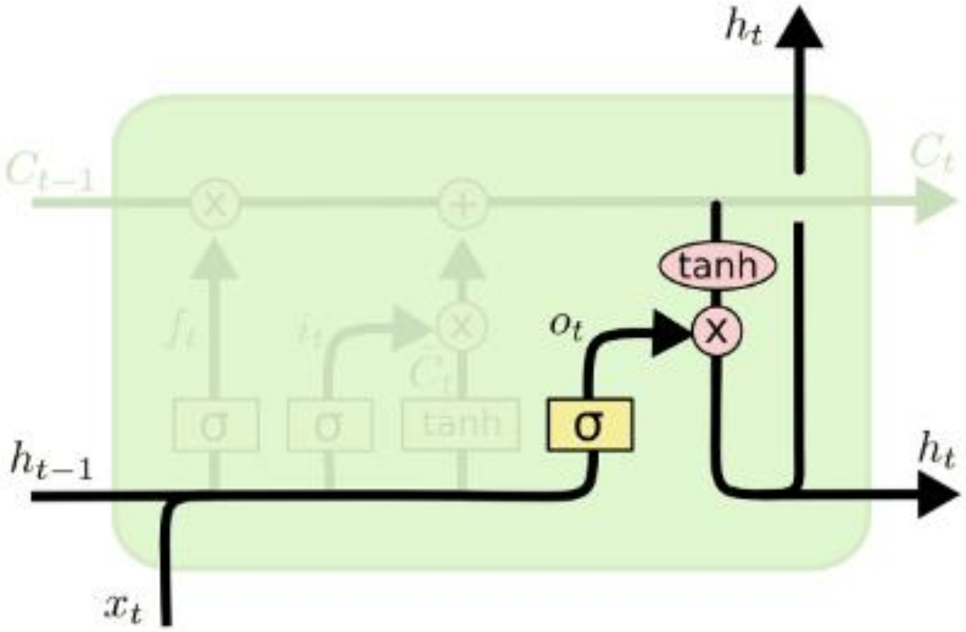


$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$

TRAIN AND TEST

LSTM

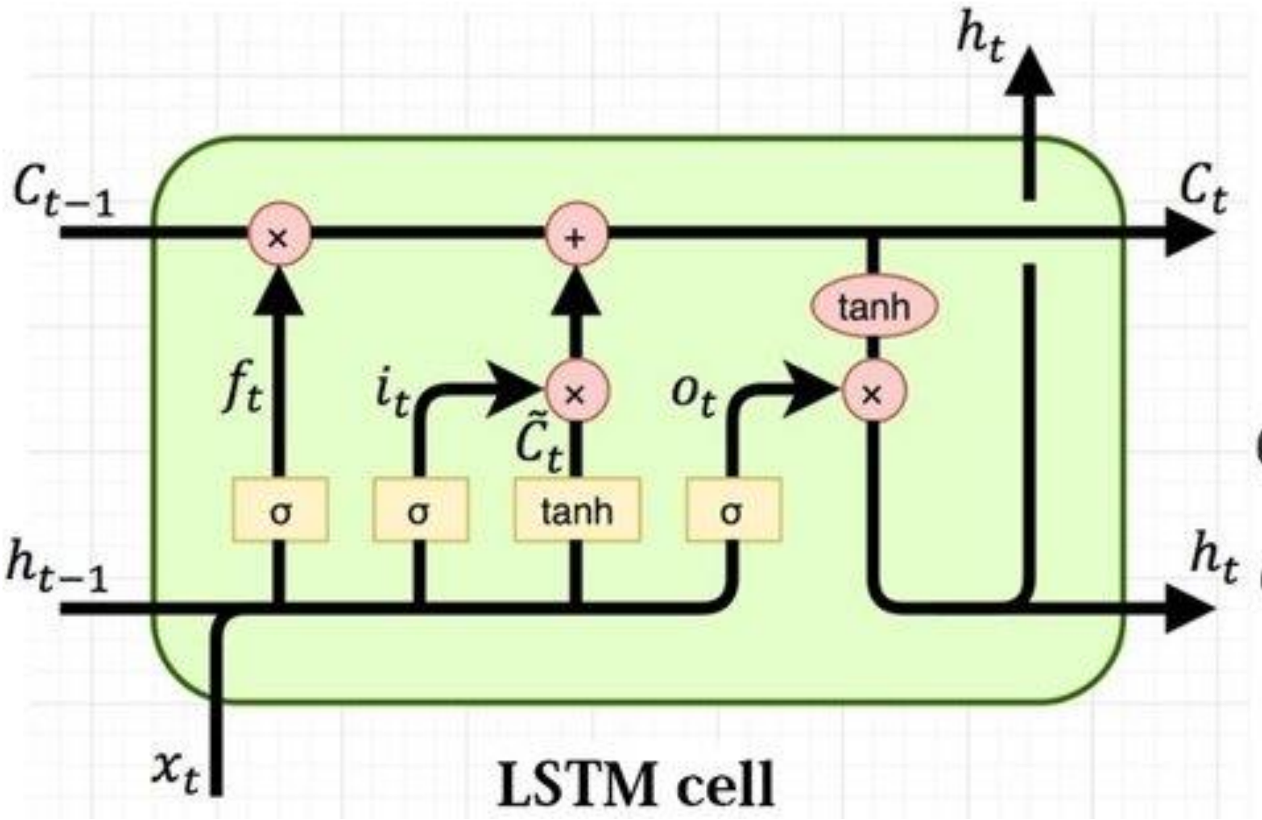
Output Gate



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

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

LSTM



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

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

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

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

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

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

TRAIN AND TEST

Paper review

Paper review - *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling*

Paper review

Acoustic modeling?

the process of establishing statistical representations for the feature vector sequences computed from the speech waveform

Acoustic Model은 입력으로 character나 phoneme등을 받아서, 혹은 Text Analysis부분에서 만들어진 linguistic feature들을 받아서, acoustic feature를 생성해주는 부분을 말함.

TRAIN AND TEST

Paper review

Abstract

rately than conventional RNNs. In this paper, we explore LSTM RNN architectures for large scale acoustic modeling in speech recognition. We recently showed that LSTM RNNs are more

optimization on a large cluster of machines. We show that a two-layer deep LSTM RNN where each LSTM layer has a linear recurrent projection layer can exceed state-of-the-art speech recognition performance. This architecture makes more effec-

TRAIN AND TEST

Paper review

LSTM RNN

- Converge quickly
- Effective use of model parameters
- Outperforms a deep feed forward neural network having an order of magnitude more parameters

TRAIN AND TEST

Paper review

Introduction

- DNNs provide only limited temporal modeling fixed-sized sliding window of acoustic frames
- RNNs mechanism exploit dynamically changing contextual window
- Deep BLSTM RNNs have recently been shown to perform better than DNNs in the hybrid speech recognition approach

“Two-layer deep LSTM RNN”

TRAIN AND TEST

Paper review

Conventional LSTM + Peephole connection

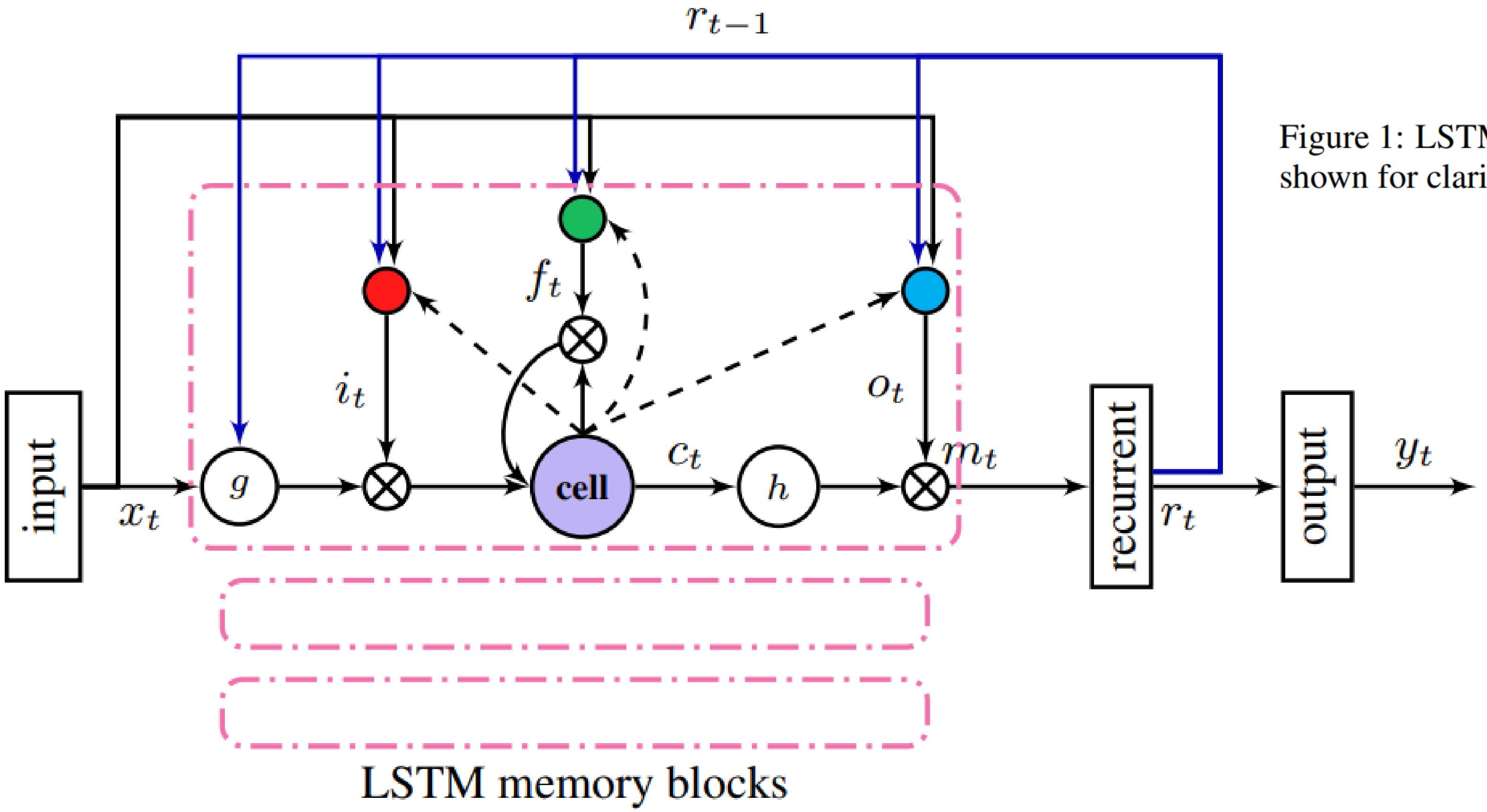
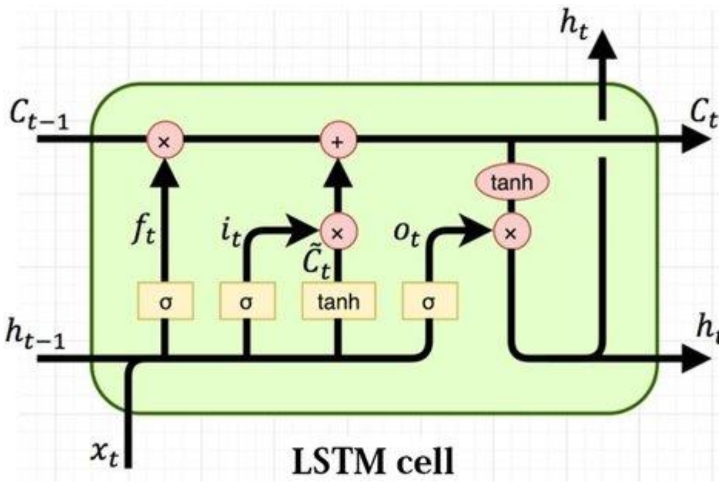


Figure 1: LSTMP RNN architecture. A single memory block is shown for clarity.



TRAIN AND TEST

Paper review

Conventional LSTM

$$i_t = \sigma(W_{ix}x_t + W_{im}m_{t-1} + W_{ic}c_{t-1} + b_i) \tag{1}$$

$$f_t = \sigma(W_{fx}x_t + W_{fm}m_{t-1} + W_{fc}c_{t-1} + b_f) \tag{2}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot g(W_{cx}x_t + W_{cm}m_{t-1} + b_c) \tag{3}$$

$$o_t = \sigma(W_{ox}x_t + W_{om}m_{t-1} + W_{oc}c_t + b_o) \tag{4}$$

$$m_t = o_t \odot h(c_t) \tag{5}$$

$$y_t = \phi(W_{ym}m_t + b_y) \tag{6}$$

Paper review

Deep LSTM

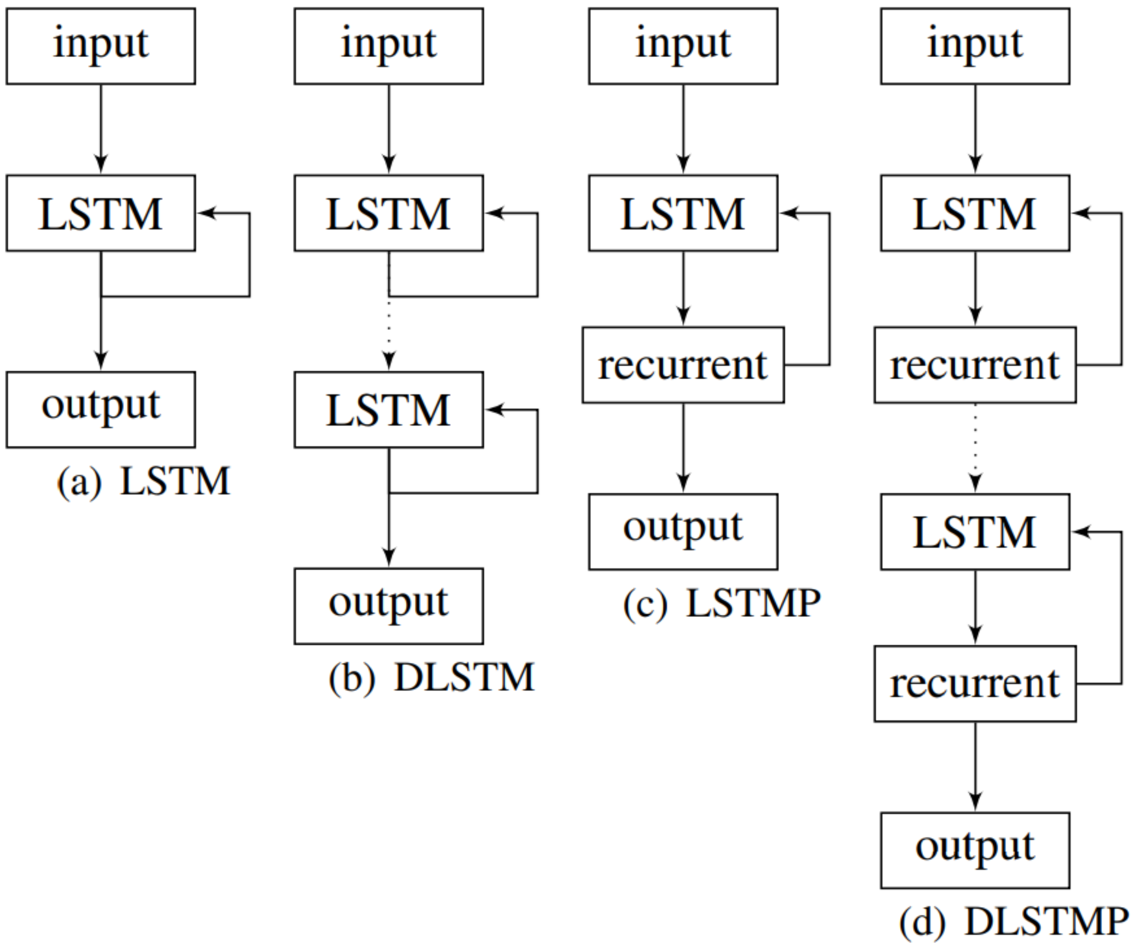


Figure 2: LSTM RNN architectures.

TRAIN AND TEST

Paper review

Deep LSTM

It has been argued that deep layers in RNNs allow the network to learn at different time scales over the input [20].

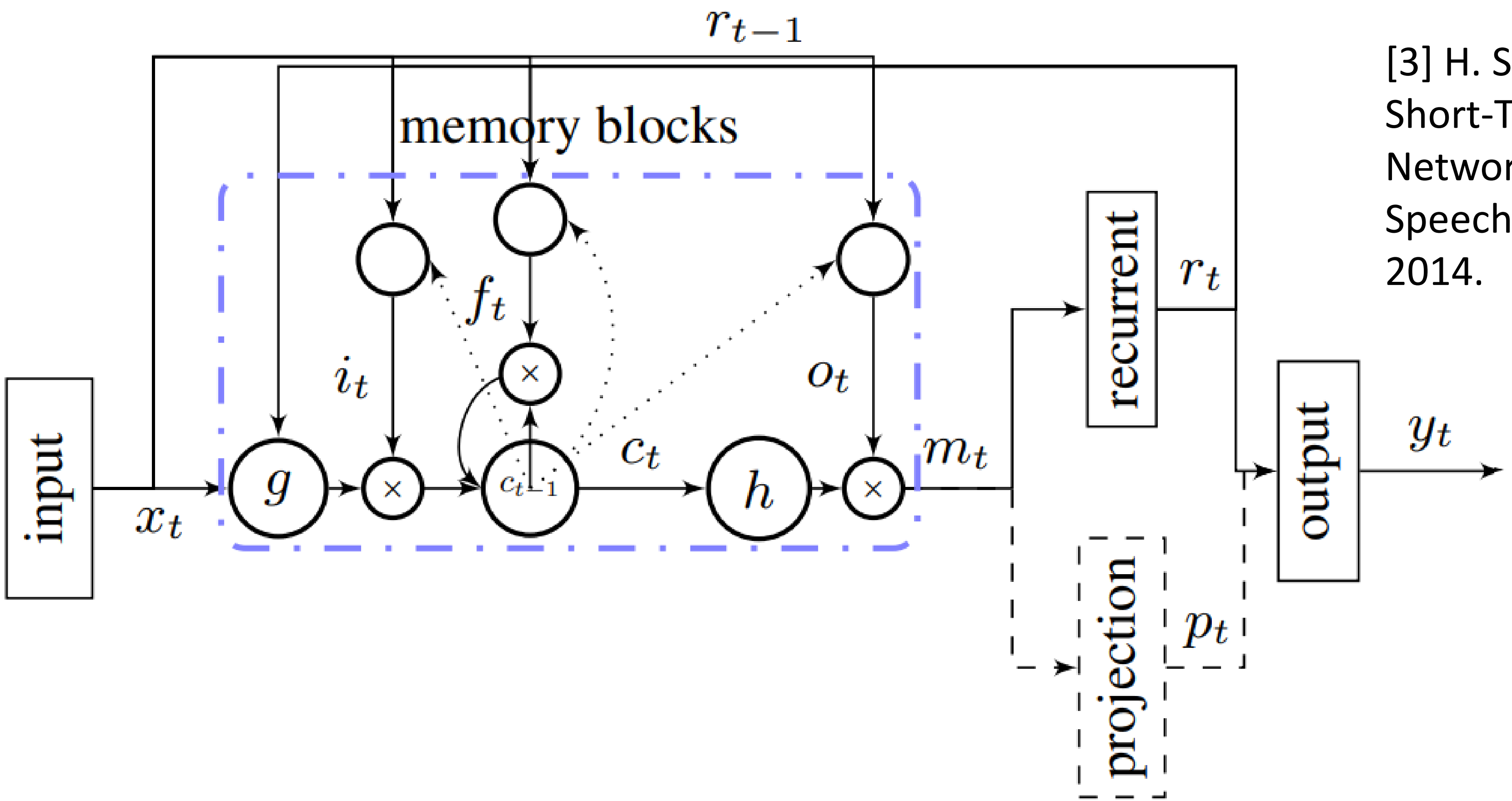
They can make better use of parameters by distributing them over the space through multiple layers.

입력이 각 시간 단계마다 더 많은 non-linear 연산을 하는 효과를 가짐

[20] M. Hermans and B. Schrauwen, "Training and analysing deep recurrent neural networks," in Advances in Neural Information Processing Systems, 2013, pp. 190–198.

Paper review

LSTMP - LSTM with Recurrent Projection Layer



[3] H. Sak, A. Senior, and F. Beaufays, "Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition," ArXiv e-prints, Feb. 2014.

Paper review

LSTMP - LSTM with Recurrent Projection Layer

ture. One of the proposed architectures introduces a recurrent projection layer between the LSTM layer (which itself has no recursion) and the output layer. The other introduces another non-recurrent projection layer to increase the projection layer size without adding more recurrent connections and this decoupling provides more flexibility. We show that the proposed architectures improve the perfor-

TRAIN AND TEST

Paper review

LSTMP - LSTM with Recurrent Projection Layer

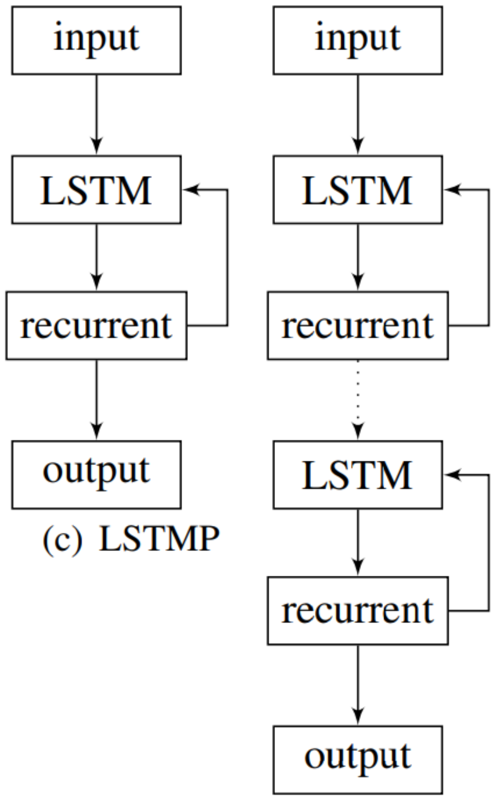
number of units in the non-recurrent projection layer and it allows us to increase the number of units in the projection layers without increasing the number of parameters in the recurrent connections

($n_c \times n_r \times 4$). Note that having two projection layers with regard to output units is effectively equivalent to having a single projection layer with $n_r + n_p$ units.

TRAIN AND TEST

Paper review

Deep LSTMP



(c) LSTMP (d) DLSTMP

LSTMP allows the memory of the model to be increased independently

Increasing memory size make the model prone to overfitting

DNNs generalize benefit (increasing depth make harder to overfit)

With this motivation, we have experimented with deep LSTMP architectures, where the aim is increasing the memory size and generalization power of the model.

Paper review

Distributed Training

We chose to implement the LSTM RNN architectures on multi-core CPU rather than on GPU. The decision was based on CPU's relatively simpler implementation complexity, ease of debugging and the ability to use clusters made from commodity hardware. For matrix operations, we used the Eigen matrix

TRAIN AND TEST

Paper review

Results

The LSTM RNN with five layers approaches the performance
of the best model.

Increasing the number of LSTMP RNN layers seems to alleviate this problem of memorization and to result in better generalization to held-out data.

by having more layers or more memory cells does not give performance improvements

TRAIN AND TEST

Paper review

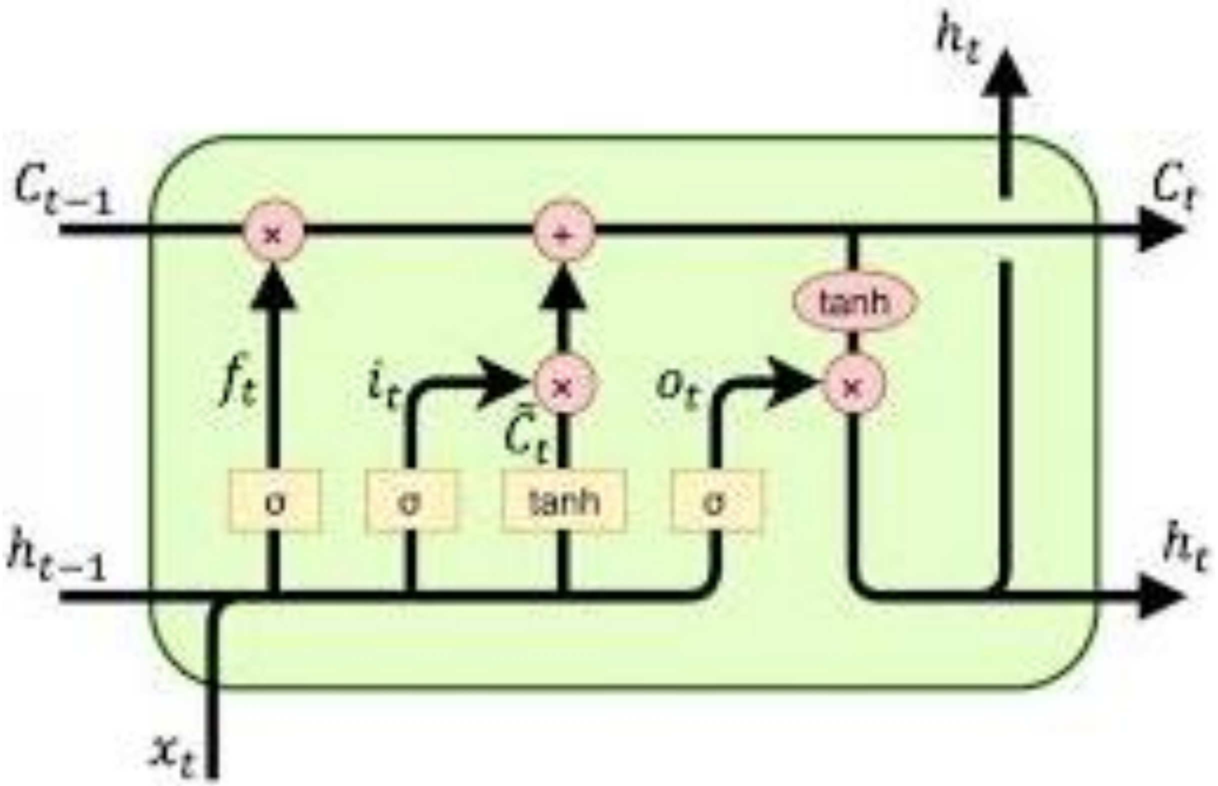
Learning point?

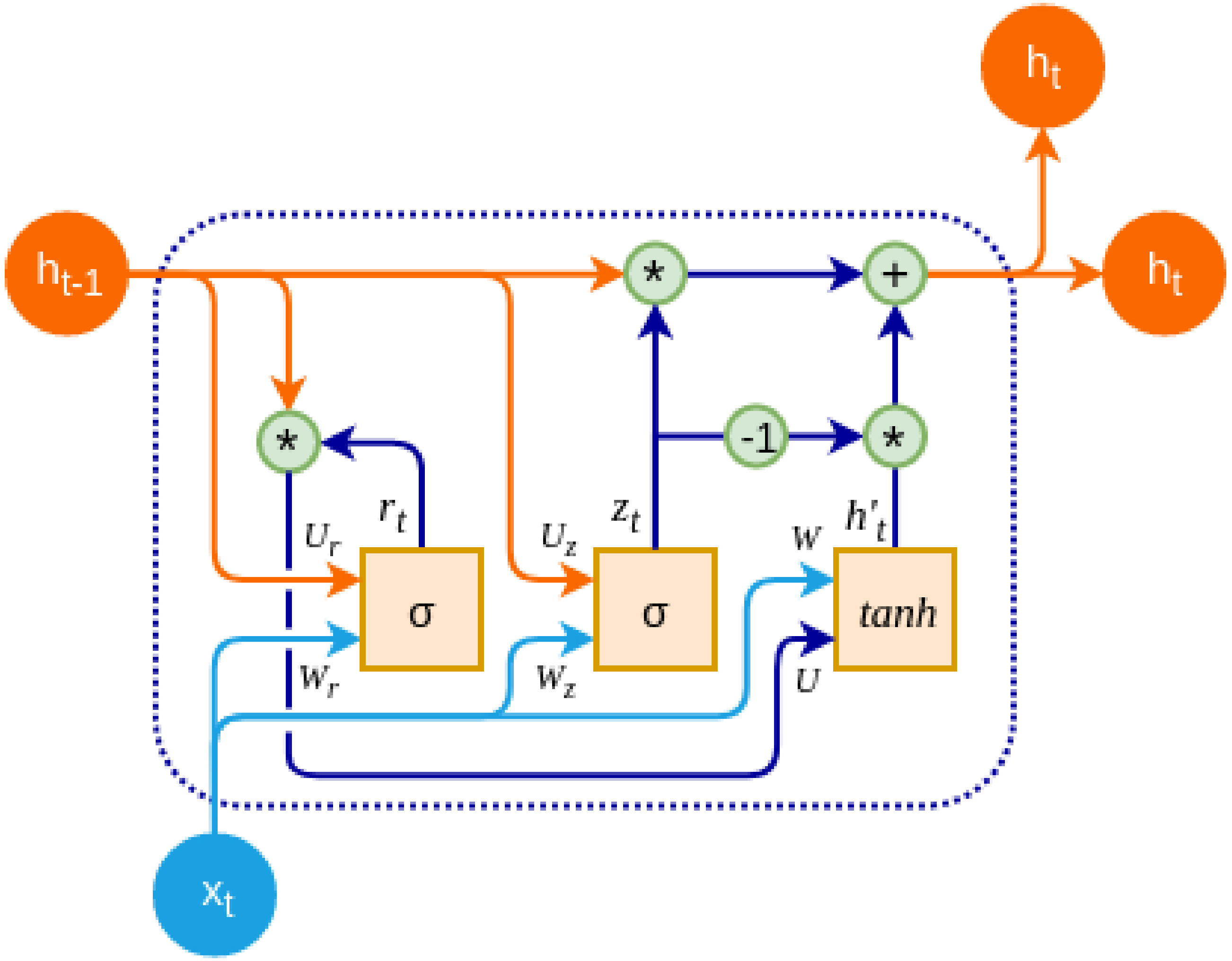
TRAIN AND TEST

GRU

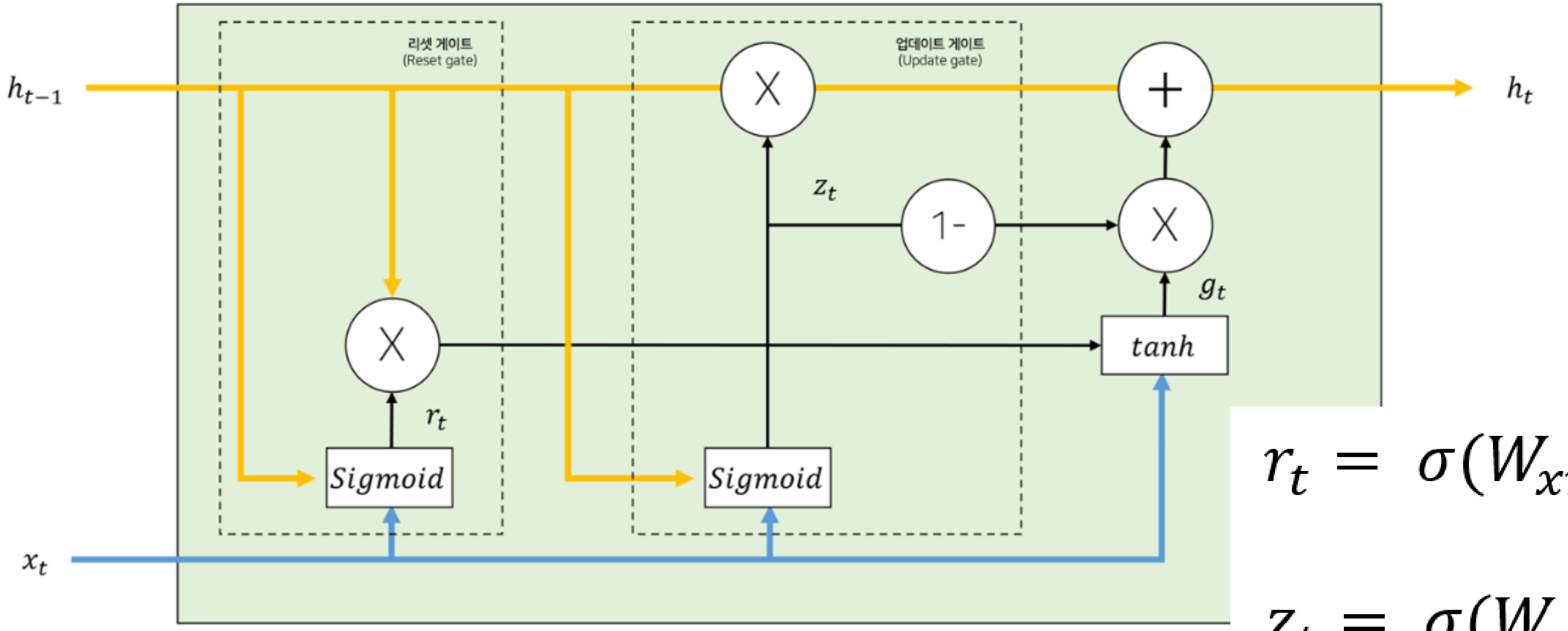
GRU - *Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation*

GRU RNN Encoder–Decoder





GRU RNN Encoder–Decoder



$$r_t = \sigma(W_{xr}x_t + W_{hr}h_{t-1})$$

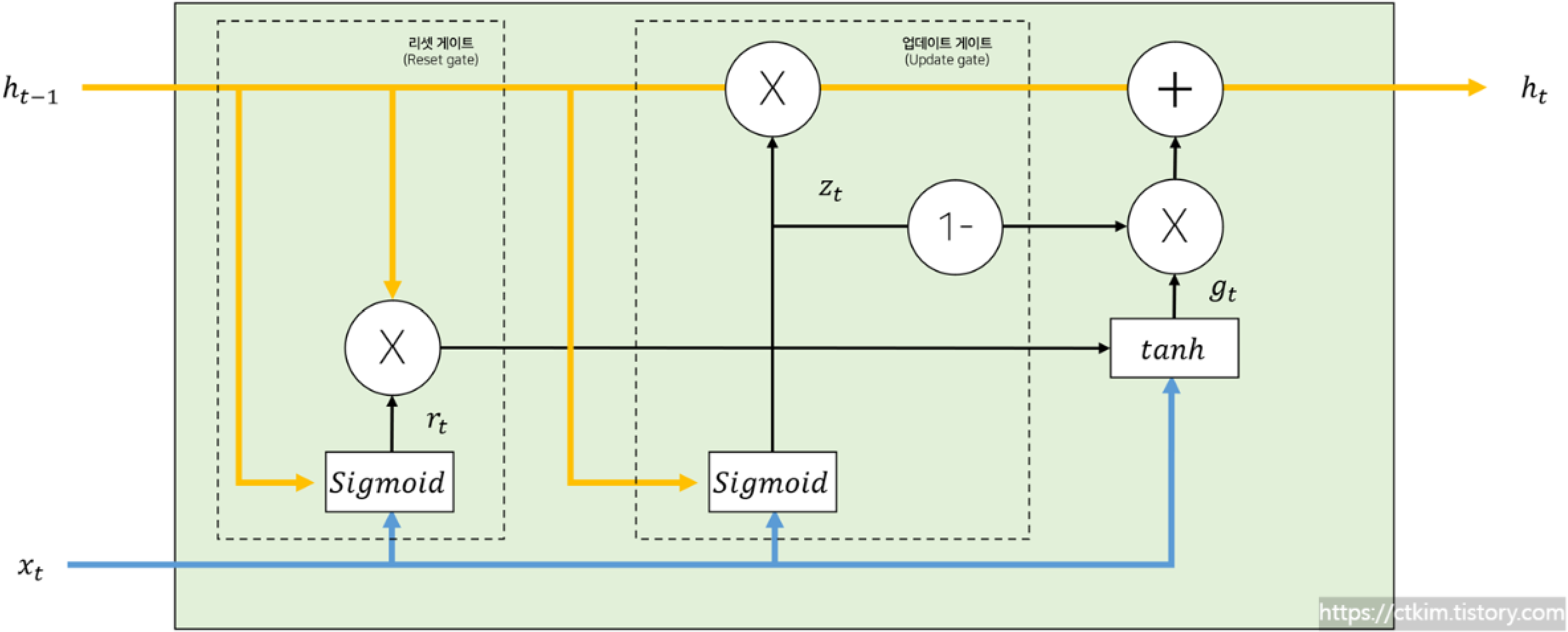
$$z_t = \sigma(W_{xz}x_t + W_{hz}h_{t-1})$$

$$g_t = tanh(W_{hg}(r_t \otimes h_{t-1}) + W_{xg}x_t)$$

$$h_t = (1 - z_t) \otimes g_t + z_t \otimes h_{t-1}$$

GRU RNN Encoder–Decoder

Using Reset gate & Update gate



GRU RNN Encoder–Decoder

Q. Difference between LSTM?

GRU RNN Encoder–Decoder

Q. Difference between LSTM?

- Simple structure
- Less parameters, Less gates
- Performance?

TNT

TRAIN AND TEST