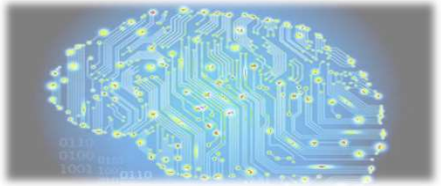
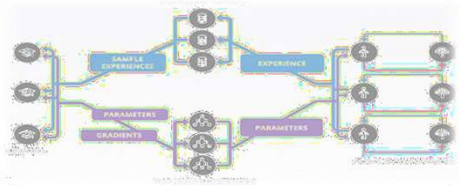
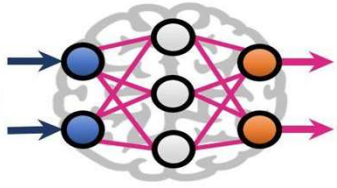
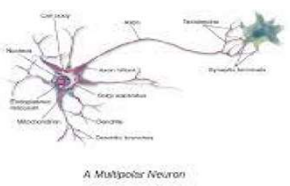
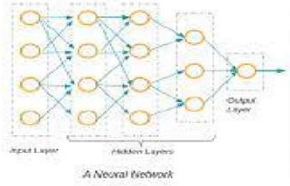







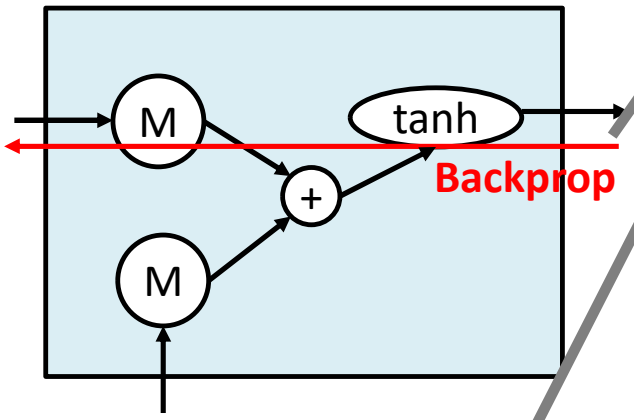
Long Short-Term Memory Network

LSTM 네트워크 구현 및 실험

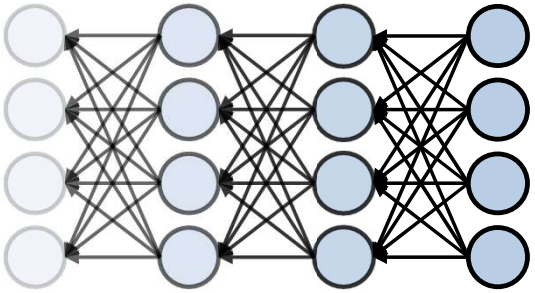
Geonil Yun,
rjsdlf45@gmail.com
2021. 05. 11.

Vanilla RNN

RNN problem

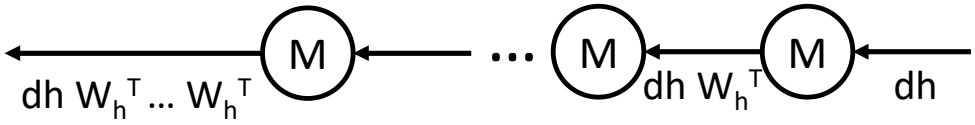


gradient vanishing in BPTT



Back Propagation Through Time

① matrix multiply operation



- Same W_h^T is used in each calculation of gradient
- It causes gradient vanishing or gradient exploding

② hyperbolic tangent (tanh)

- Derivative value of tanh is 0.0 to 1.0
- Gradients get smaller in each back propagation process
- => gradient vanishing problem

LSTM Concept

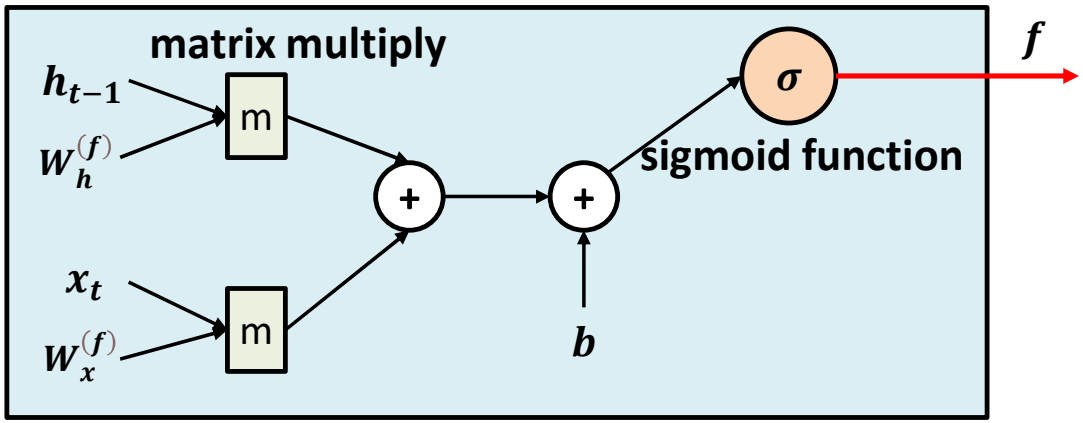
LSTM Concept

Gates : adjust data flow by using sigmoid function (0.0~1.0)
Cell state(c) : only use in LSTM layer

- Forget gate(f) : Forget cell state (c_{t-1}:previous cell state)
- Input gate(i) : Input cell state (g : tanh(xW_x+hW_h+b))
- Output gate(o) : output cell state (tanh(c))

Gates

Forget gate $f = \sigma(x_t W_x^{(f)} + h_{t-1} W_h^{(f)} + b)$

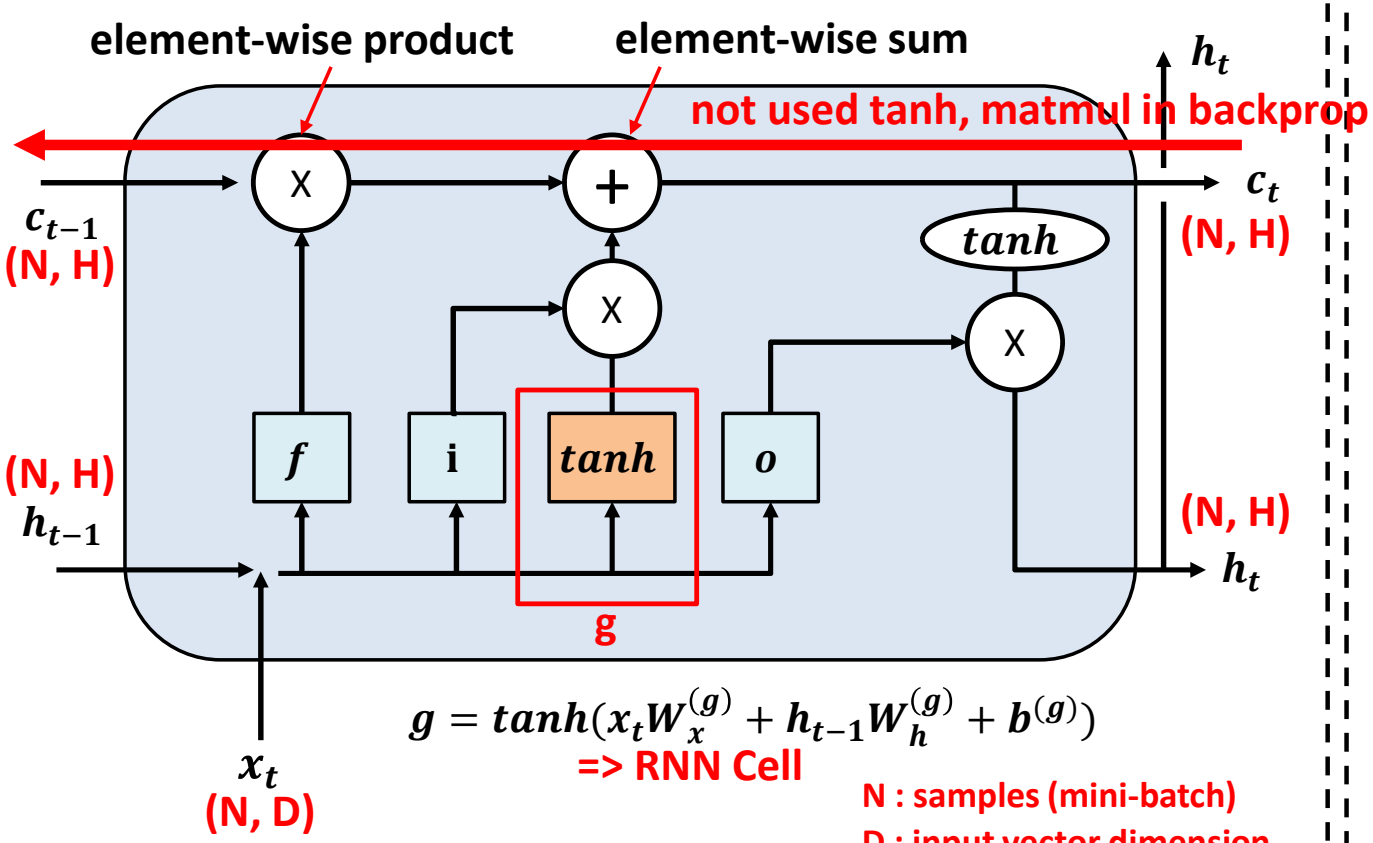


- input, output gates have same architecture
- each gate has its own parameters(Wx, Wh, b)

$$f = \sigma(x_t W_x^{(f)} + h_{t-1} W_h^{(f)} + b^{(f)})$$
$$\mathbf{i} = \sigma(x_t W_x^{(i)} + h_{t-1} W_h^{(i)} + b^{(i)})$$
$$o = \sigma(x_t W_x^{(o)} + h_{t-1} W_h^{(o)} + b^{(o)})$$

LSTM Architecture

LSTM Architecture



- Weights and bias (parameters)
  - $[\ W_x^{(f)}, W_h^{(f)}, b^{(f)}\ ]$ for forget gate
  - $[\ W_x^{(i)}, W_h^{(i)}, b^{(i)}\ ]$ for input gate
  - $[\ W_x^{(o)}, W_h^{(o)}, b^{(o)}\ ]$ for output gate
  - $[\ W_x^{(g)}, W_h^{(g)}, b^{(g)}\ ]$ for g cell(RNN)

$$f = \sigma(x_t W_x^{(f)} + h_{t-1} W_h^{(f)} + b^{(f)})$$

$$\mathbf{i} = \sigma(x_t W_x^{(i)} + h_{t-1} W_h^{(i)} + b^{(i)})$$

$$o = \sigma(x_t W_x^{(o)} + h_{t-1} W_h^{(o)} + b^{(o)})$$

$$g = tanh(x_t W_x^{(g)} + h_{t-1} W_h^{(g)} + b^{(g)})$$

$$c_t = f \odot c_{t-1} + g \odot i$$

$$h_t = tanh(c_t) \odot o$$

LSTM (for speed)

Affine transformation

$$A = x_t W_x + h_{t-1} W_h + b$$

(N, 4H)

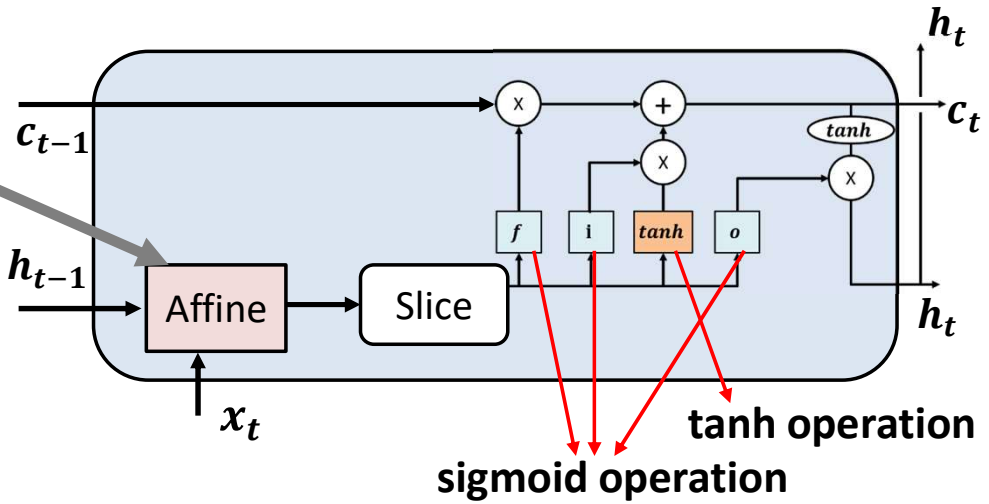
(D, 4H)

$W_x^{(f)}$ $W_x^{(i)}$ $W_x^{(o)}$ $W_x^{(g)}$

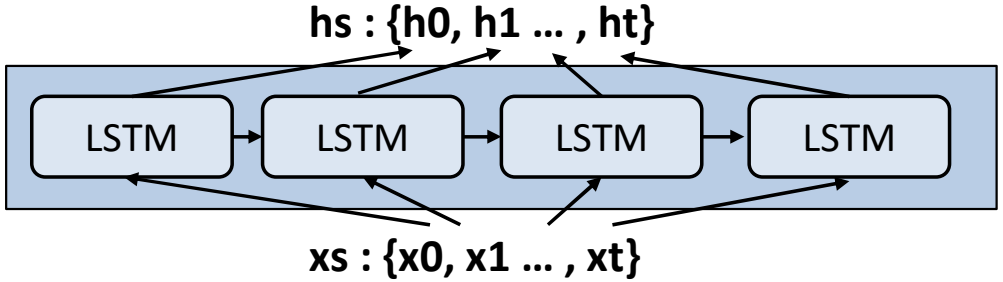
(H, 4H)

$W_h^{(f)}$ $W_h^{(i)}$ $W_h^{(o)}$ $W_h^{(g)}$

LSTM Architecture (for calculate speed)



LSTM for time series



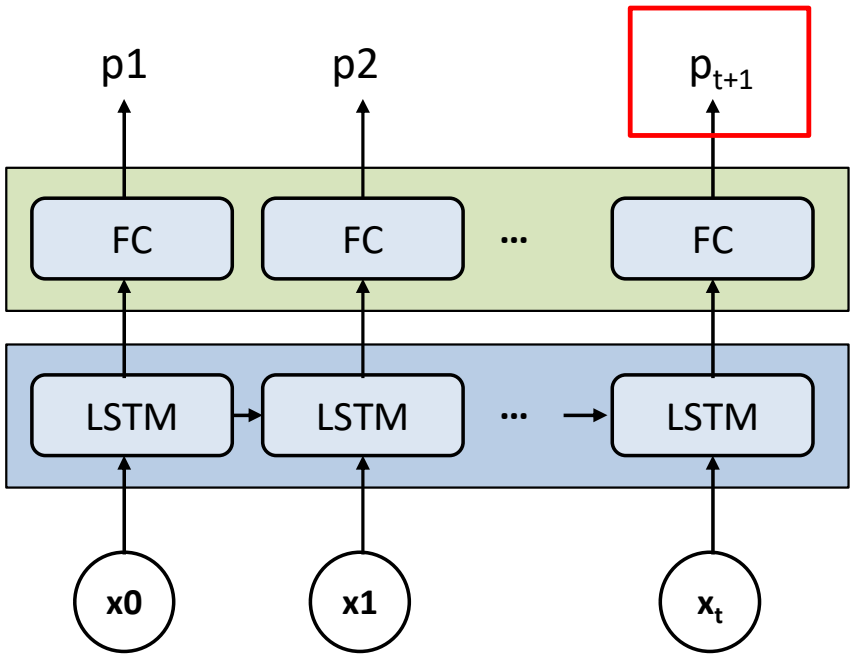
view with code

https://colab.research.google.com/drive/1AkSI-SjhVzMSA0TOWsVQ35tI5mdB_kxE?usp=sharing

LSTM (Model Structure)

Structure

MSE for Loss = $\frac{1}{2}\sum(y-t)^2$



Data

- **Beijing Air Quality dataset**
  from uci-dataset/00381
  - time series dataset
  - features : 8 (Multivariate dataset)
  - target : pm 2.5 value of next hour

Specific

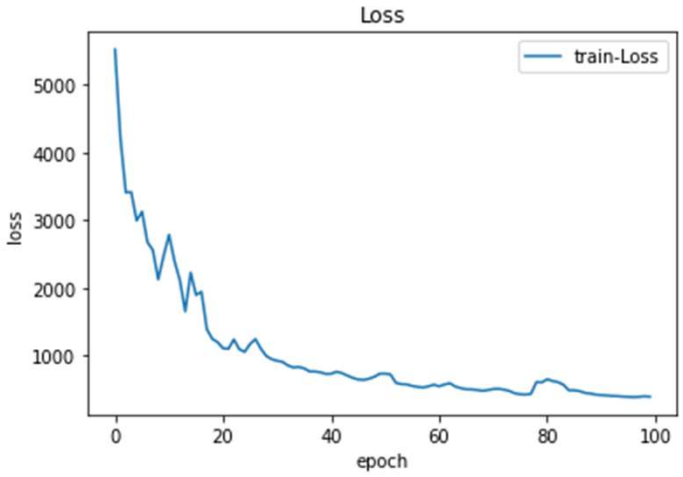
- Weights (initialize)
  - Xavier for LSTM
  - He for FC(Fully-Connected layer)
- Activation function
  - ReLU for FC
- Loss function
  - MSE(Mean Squared Error)

LSTM (Experiment)

Training

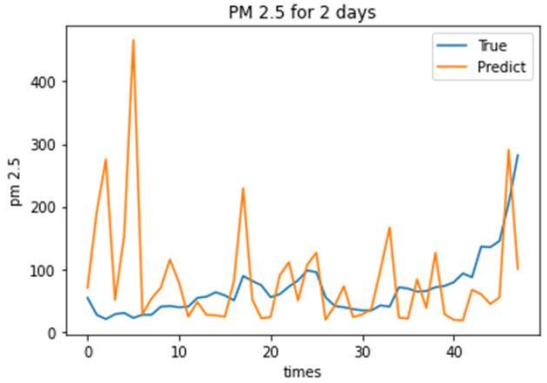
MSE for Loss = $\frac{1}{2}\sum(y-t)^2$

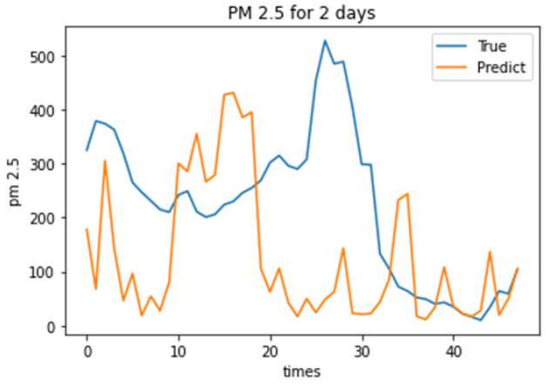
Optimizer : Adam
train data : (320, 24, 8)
batch size : 64
epoch : 100



evaluation

use test data





result

- **working backpropagation**
- **MSE : 8152.33**
- **RMSE : 90.29**

view with code

https://colab.research.google.com/drive/1oNW4HzBzmrjyWcWZbjFMtksP1vTV9IO2?usp=sharing

Thank you