Machine Learning Engineer Course
Day 30

- LSTM -

DIVE INTO CODE

Thursday November 11, 2021

DIOP Mouhamed

Agenda

1. Check-in
2. Quick Review
3. LSTM
4. Sample code
5. To do by next class
6. Check-out

Check-in

3 minutes Please post the following point to Zoom chat.

Q. Have you started working on your graduation project ?

Quick Review (RNN)

- Recurrent Neural Network (RNN)

LSTM (Gated RNN)

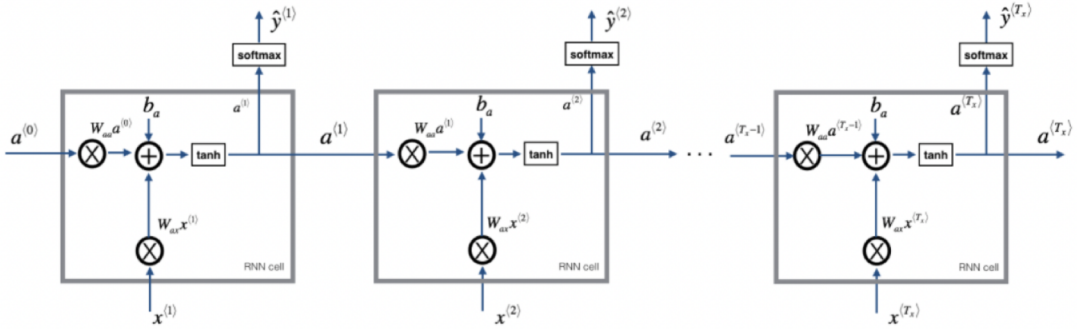
What is RNN?

（Background）
At each time step, the RNN was able to add the activated value from the previous time step as information. Specifically, this value is multiplied by the transition matrix from the hidden state to the hidden state and added together.

When the singular value of this transition matrix is other than 1 (absolute value), there is a problem of gradient vanishing or gradient explosion.

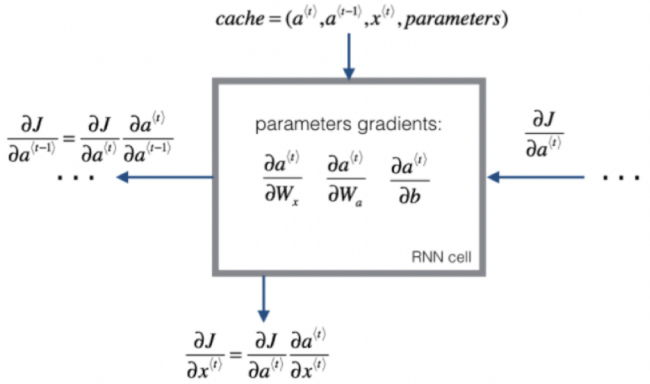
LSTM (Gated RNN)

RNN Forward Pass



[Andrew Ng, Sequential Models Course, Deep Learning Specialization]

RNN Backward Pass



$$a^{\langle t \rangle} = \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)$$

$$\frac{\partial \tanh(x)}{\partial x} = 1 - \tanh(x)^2$$

$$\frac{\partial a^{\langle t \rangle}}{\partial W_{ax}} = (1 - \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)^2)\, x^{\langle t \rangle T}$$

$$\frac{\partial a^{\langle t \rangle}}{\partial W_{aa}} = (1 - \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)^2)\, a^{\langle t-1 \rangle T}$$

$$\frac{\partial a^{\langle t \rangle}}{\partial b} = \sum_{batch} (1 - \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)^2)$$

$$\frac{\partial a^{\langle t \rangle}}{\partial x^{\langle t \rangle}} = W_{ax}{}^{T} . (1 - \tanh(W_{ax}x^{\langle t \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)^2)$$

$$\frac{\partial a^{\langle t \rangle}}{\partial a^{\langle t-1 \rangle}} = W_{aa}{}^{T} . (1 - \tanh(W_{ax}x^{\langle t-1 \rangle} + W_{aa}a^{\langle t-1 \rangle} + b)^2)$$

[Andrew Ng, Sequential Models Course, Deep Learning Specialization]

LSTM (Gated RNN)

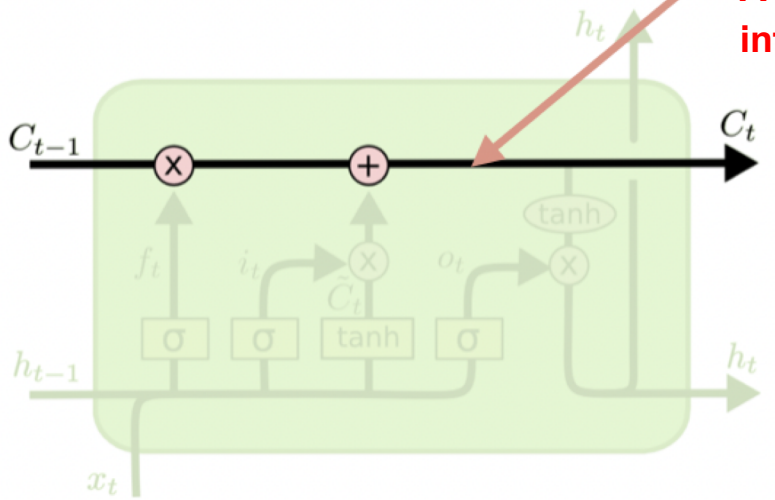
What is LSTM ?

Long Short-Term Memory (1997)

https://www.bioinf.jku.at/publications/older/2604.pdf

As a way to solve the problem of vanishing gradients
It was proposed by Sepp Hochreiter and Jürgen Schmidhuber in 1997.

Point.
A mechanism to retain and control cell information has been added.



http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM (Gated RNN)

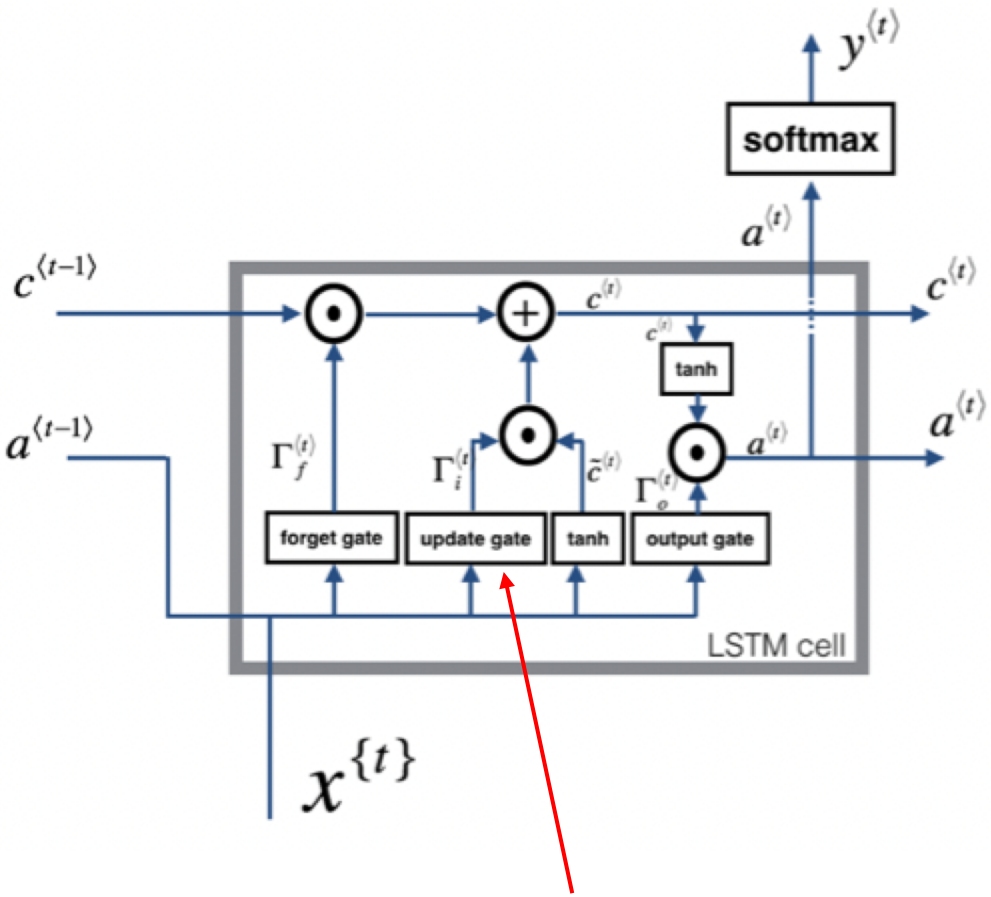
LSTM Networks

A feature of LSTM networks compared to RNNs is that they possess a unit to store the state and a control mechanism called a gate, which is described below.

Input gate：Controls whether or not information is written (write)
Output gate：Controls whether or not information is output (read).
Forget gate：Control whether information is deleted or not (forget)

LSTM (Gated RNN)

LSTM Cell



- Input gate:
  - At 0.0, nothing is written.
  - At 1.0, write all
  - Between 0.0~1.0, partial writing
- Output gate:
  - When 0.0, nothing is output.
  - At 1.0, output everything.
  - Partial output between 0.0~1.0
- Forget gate:
  - At 0.0, delete everything.
  - At 1.0, do not delete anything.
  - Between 0.0~1.0, delete some

The Input gate is also sometimes called the Update gate.

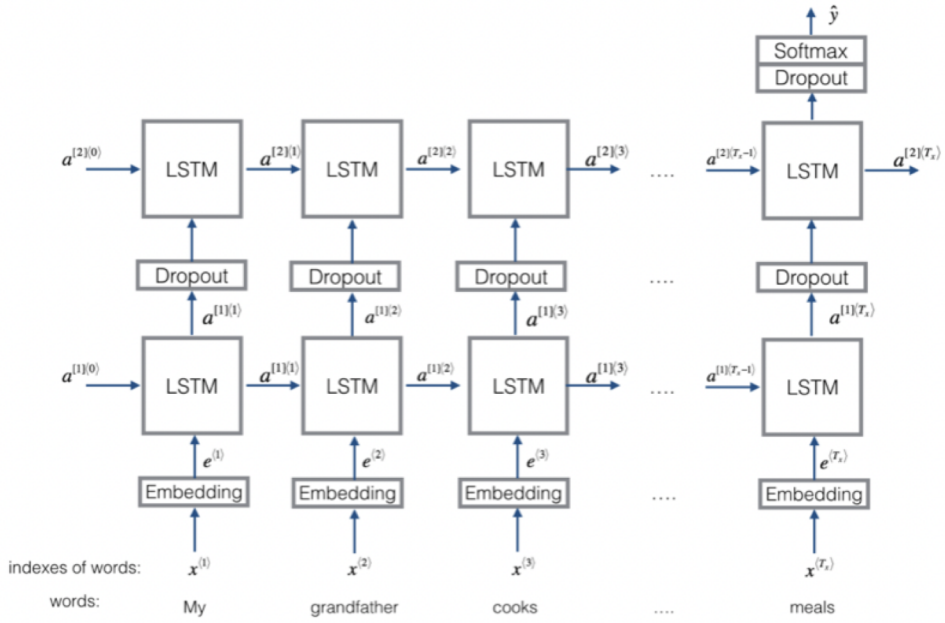
LSTM (Gated RNN)

LSTM Networks

According to the structure described above, the LSTM will have two parts: a long-term memory unit called the cell state and a short-term memory unit called the hidden state. These units propagate the error backwards through the time steps and layers. The essential device that prevents error decay is **the positive (additive) part of long-term memory** (not the multiplicative one).

Gates are responsible for reading and writing information into and out of cells, just like computer memory. What makes them different from memory is that gates learn when to allow input, output, and deletion.

LSTM (Gated RNN)



[Andrew Ng, Sequential Models Course, Deep Learning Specialization]

Parameters:

Layer (type)	Output Shape	Param #
input_1 (InputLayer)	(None, 10)	0
embedding_1 (Embedding)	(None, 10, 50)	20000050
lstm_1 (LSTM)	(None, 10, 128)	91648
dropout_1 (Dropout)	(None, 10, 128)	0
lstm_2 (LSTM)	(None, 128)	131584
dropout_2 (Dropout)	(None, 128)	0
dense_1 (Dense)	(None, 5)	645
activation_1 (Activation)	(None, 5)	0

Total params: 20,223,927
Trainable params: 20,223,927
Non-trainable params: 0

LSTM (Gated RNN)

Tuning the LSTM

① When the learning rate is high, perplexity (a measure of how well the prediction candidates are narrowed down by the inverse of probability. It can be calculated without using the correct labels.

② Use softsign instead of tanh (faster calculation and less saturated). tanh converges exponentially, while softsign converges polynomially.

③ If parameters > samples, it is overfitting.

④You can stack the layers.

⑤ Try multiple patterns of epochs to evaluate performance and try early stopping.

⑥ Use RMSProp and AdaGrad (decaying learning rate).

⑦ Use the initial value of Xavier to prevent the variance of the output from being too small or too large.

⑧ Adjustment of softmax sample temperature (temperature of softmax).

LSTM (Gated RNN)

Model samples:

Softmax sample temperature: lower setting will generate more likely predictions, but you'll see more of the same common words again and again. Higher setting will generate less frequent words but you might see more spelling errors.

0.32

the a there the there was there end the tors could startup the of intere the tors to in the the there

in to the tions

in tes a startups is and the the there a tartups the and the the things the and compang the to a star

the best a tors to of the there the the things a startups the is investors to the in the is the tors

in the of the and and you make and the a tions was the the a startups the the tars a startup the thin

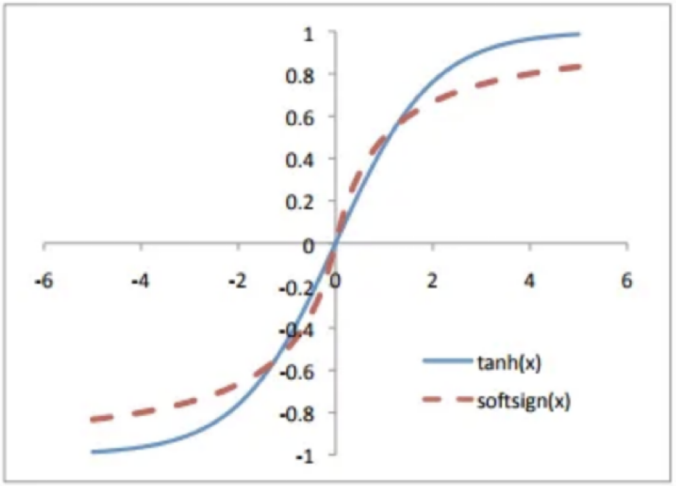
https://cs.stanford.edu/people/karpathy/recurrentjs/

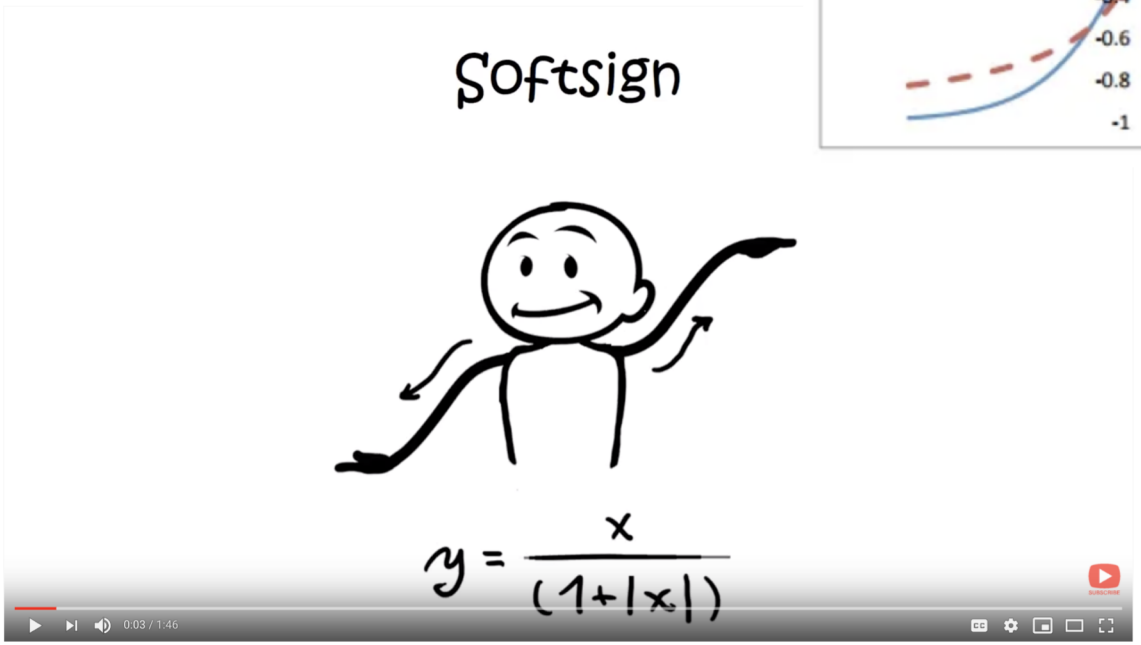
https://cs.stackexchange.com/questions/79241/what-is-temperature-in-lstm-and-neural-networks-generally

LSTM (Gated RNN)

(Extra) Activation function

:D Learn with your body!





https://www.youtube.com/watch?time_continue=47&v=1Du1XScHCww&feature=emb_logo

LSTM (Gated RNN)

Another measure

Countermeasures against gradient divergence

A method called gradient clipping has been proposed.

It keeps the maximum value (or threshold) of all parameters used in the neural network, and when the norm of the gradient Δw |Δw| > Δmax, it restricts it as follows (when using the threshold, Δmax is used as threshold)

$$\Delta w \leftarrow \Delta w \frac{\Delta_{max}}{|\Delta w|}$$

The direction of the gradient remains the same, only the magnitude can be changed.
This allows us to adjust the parameter update step so that it is not too large.

LSTM (Gated RNN)

Developmental Topics

The structure of RNNs and LSTMs is such that predictions from the current timestep are based not only on inputs from this timestep, but also on outputs from previous timesteps (through hidden and cell state paths).

The problem with these architectures is that they only use information from the forward timestep of the entire sequence to make predictions. Even if there is **important information in the backward timestep that can contribute to the prediction at the current timestep**, it cannot be used.

A method that addresses this problem is the Bi-directional RNN (Bi-directional RNN). This is an architecture that allows information from bi-directional time steps to contribute to prediction instead of uni-directional.

Sample code

How to solve problems
“LSTM”

[Problem 1] Execution of various methods
[Problem 2] (Advance Problem) Comparison between multiple datasets
[Problem 3] Explanation of other classes

Sprint 23 – LSTM

Explanation about this Sprint is given but please try it on your own first.

Sprint 23 – LSTM

Please work on your own after class and submit your assignments on DIVER.

Sprint 23 –LSTM

A Sample Code of this Sprint is given but please try it on your own.

Sprint 23 – LSTM

Please work on your own after class and submit your assignments on DIVER.

ToDo by next class

Next class will be Zoom: Thursday 11November 2021 19:30 ~ 20:30

🚨ToDo: Seq2Seq
https://diver.diveintocode.jp/curriculums/2011

Check-out

3 minutes Please post the following point to Zoom chat.

Q. Current feelings and reflections
(joy, anger, sorrow, anticipation, nervousness, etc.)

Thank You For Your Attention