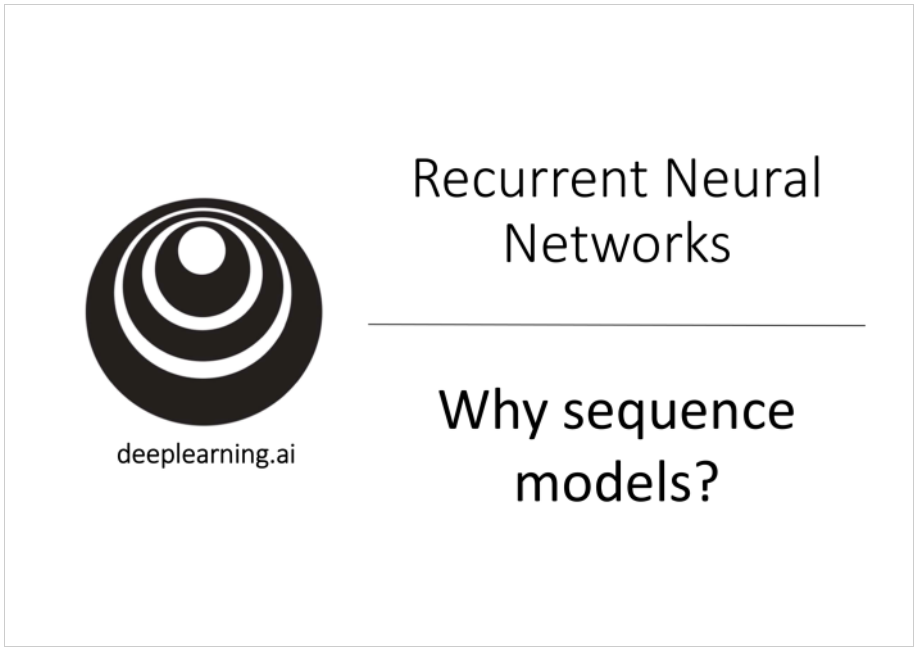
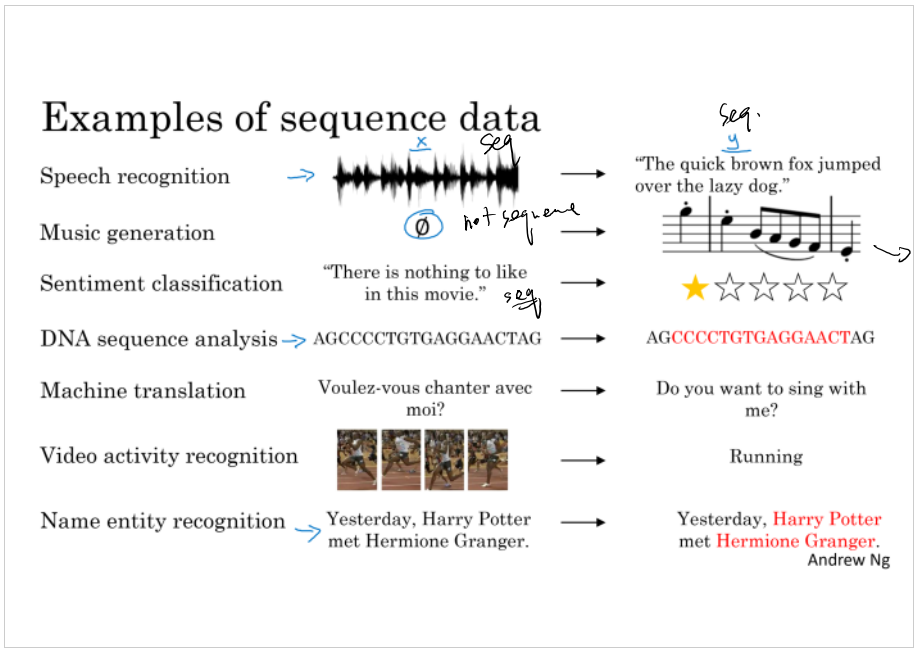
Recurrent Neural Networks

Wednesday, June 6, 2018 3:46 PM

intro





notation

deeplearning.ai

Recurrent Neural Networks

Notation

Motivating example

NLP

where are the peoples name: task NER

application: search engines

x: Harry Potter and Hermione Granger invented a new spell.

9 words

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ..... $x^{<t>}$ ..... $x^{<9>}$

$T_x = 9$

y: 1 1 0 1 1 0 0 0 0

0 = yes, part of name.

label

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ ..... $y^{<9>}$

could be other output representation.

$T_y = 9$

$X^{(i)<t>}$

$T_x^{(i)} = 9$ lenght of sequence

$y^{(i)<t>}$

$T_y^{(i)}$

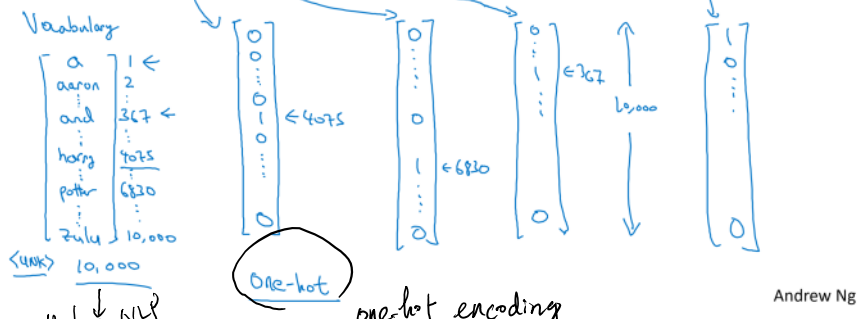
(i) ith training example

Andrew Ng

Representing words

x: Harry Potter and Hermione Granger invented a new spell.

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<9>}$



Andrew Ng

How to represent individual words?

① Come up with a vocab / a dictionary.

later: what if encountering words not in your vocab? create a category "unknown"

small for NLP

Commercial applications → 30,000 - 50,000 (100,000)

Dictionary size is common.

one-hot encoding

Representing words

x: Harry Potter and Hermione Granger invented a new spell.

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<9>}$

And = 367
Invented = 4700
A = 1
New = 5976
Spell = 8376
Harry = 4075
Potter = 6830
Hermione = 4200
Gran... = 4000

Andrew Ng

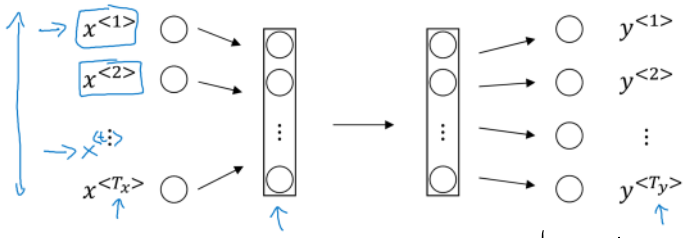
rnnmodel

deeplearning.ai

Recurrent Neural Networks

Recurrent Neural Network Model

Why not a standard network?



Problems: Why this does not work well?

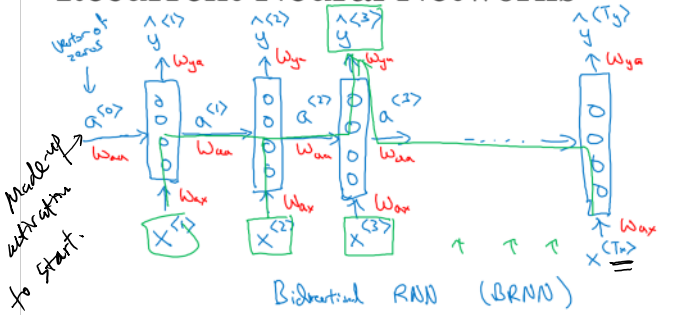
- Inputs, outputs can be different lengths in different examples. → maybe padding, if there's a max length.
- Doesn't share features learned across different positions of text. "Harry" appearing at a different position.

Andrew Ng

为何简单标准网络不适合 HMM...

Recurrent Neural Networks

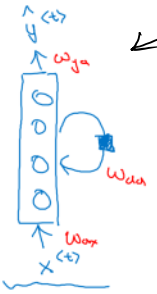
$T_x = T_y$



He said, "Teddy Roosevelt was a great President."

He said, "Teddy bears are on sale!"

params



W_{ya}
W_{aa}
W_{ax}

same parameters

Andrew Ng

What is a recurrent neural network.

another way to draw it — "harder to interpret"

Problem: Does not use the word after to predict.

params

W_{aa} W_{ax}

Forward Propagation

$a \leftarrow W_{ax} x^{<1>}$

$\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<3>}$ $\hat{y}^{<T_y>}$

从前文的某一部分加过来

to simplify this notation

W_a 的话 W_{aa}, W_{ax} stack 在一起

stack the two matrices

$[W_{aa} \mid W_{ax}] \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix}$

Next: backprop

Simplified RNN notation

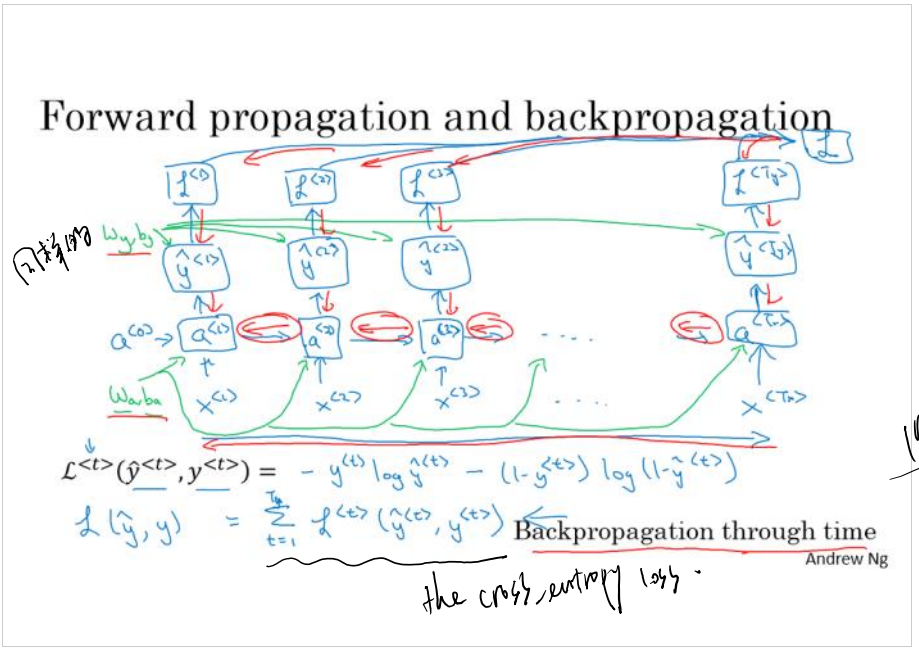
$a^{\langle t \rangle} = g(W_{aa} a^{\langle t-1 \rangle} + W_{ax} x^{\langle t \rangle} + b_a)$

$\hat{y}^{\langle t \rangle} = g(W_{ya} a^{\langle t \rangle} + b_y)$

$a^{\langle t \rangle} = g(W_a [a^{\langle t-1 \rangle}, x^{\langle t \rangle}] + b_a)$

$[W_{aa} \mid W_{ax}] \begin{bmatrix} a^{\langle t-1 \rangle} \\ x^{\langle t \rangle} \end{bmatrix} = W_a$

backprop

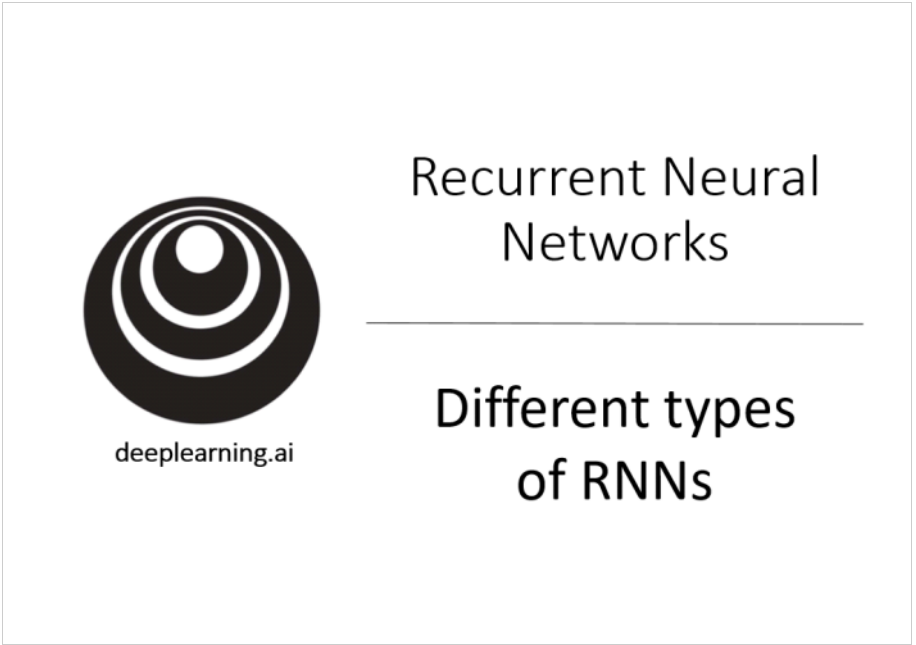


反向传播是为了计算loss

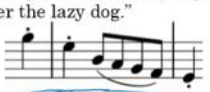
Back propagation through time

rnn different type

Tx need not = Ty the length of input and output need not be equal.

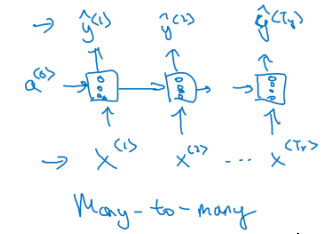


Examples of sequence data

Speech recognition		"The quick brown fox jumped over the lazy dog."
Music generation	∅	
Sentiment classification	"There is nothing to like in this movie."	★☆☆☆☆
DNA sequence analysis	AGCCCCTGTGAGGAACTAG	AGCCCCTGTGAGGAACTAG
Machine translation	Voulez-vous chanter avec moi?	Do you want to sing with me?
Video activity recognition		Running
Name entity recognition	Yesterday, Harry Potter met Hermione Granger.	Yesterday, Harry Potter met Hermione Granger.

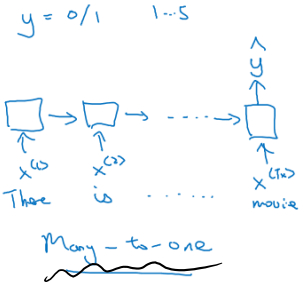
Andrew Ng

Examples of RNN architectures



Many-to-many

Many to many architecture



Many-to-one

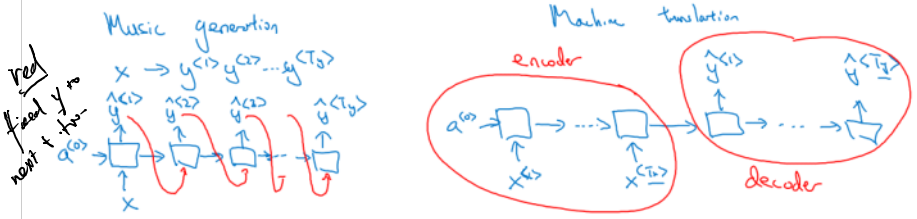
One-to-one

Many to one

example : movie review

Andrew Ng

Examples of RNN architectures

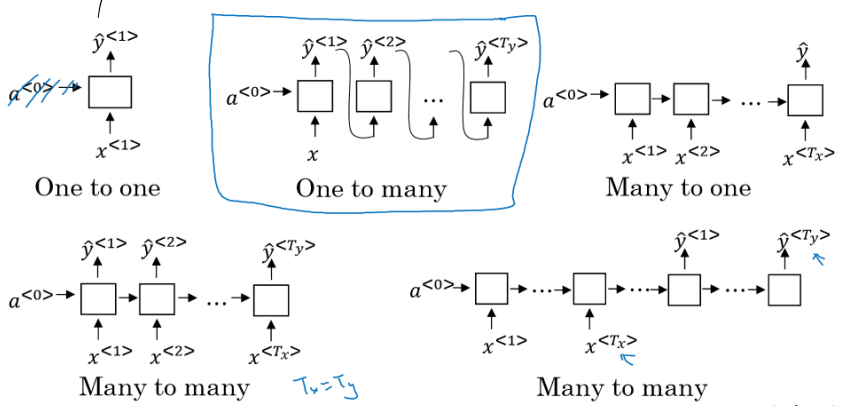


One-to-many ex: music generation $x = \phi$

Many-to-many input & output lengths are different
ex: machine translation sentences of different length

Andrew Ng

Summary of RNN types



Andrew Ng

good summary and graph.

later "attention-based architecture" not in this list.

up next: sequence generation

language model

deeplearning.ai

Recurrent Neural Networks

Language model and sequence generation

What is language modelling?

Speech recognition

The apple and pair salad.

→ The apple and pear salad.

P(The apple and pair salad) = 3.2×10^{-13}

P(The apple and pear salad) = 5.7×10^{-10} more likely.

P(Sentence) = ? $P(y^{<1>}, y^{<2>}, \ldots, y^{<T_y>})$ a particular sequence of word

Andrew Ng

Language modelling with an RNN

Training set: large corpus of english text.

Tokenize — form a vocab

ignore punctuation here

Cats average 15 hours of sleep a day. <EOS>

$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ ... $y^{<8>}$ $y^{<9>}$

$x^{<t>} = y^{<t-1>}$

The Egyptian Mau is a bread of cat. <EOS>

<UNK>

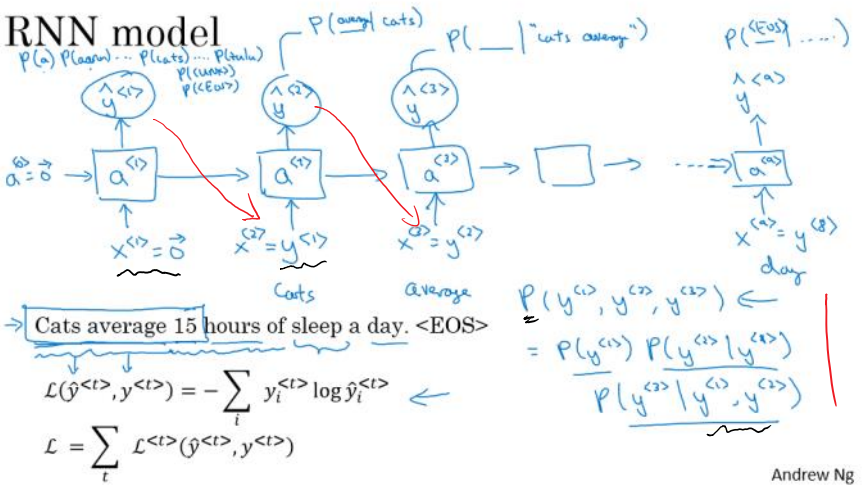
10,000

what if some words are not in vocab?

recode as [unknown]

Andrew Ng

RNN model



Andrew Ng

Model the chance the sentences with RNN

Sample novel sequences.

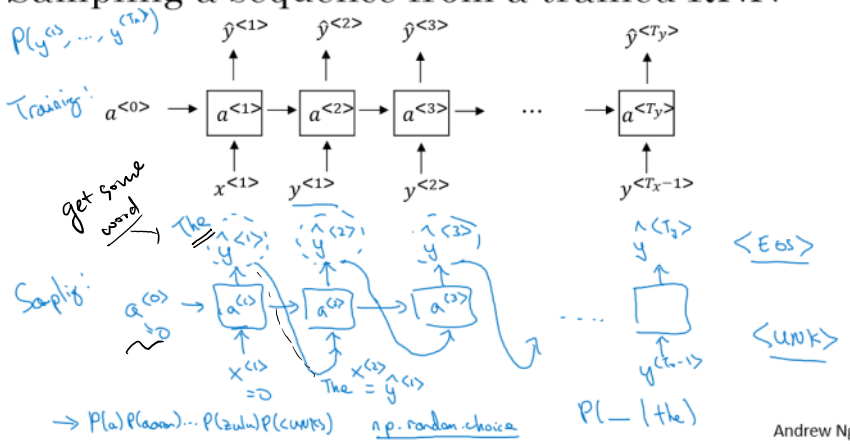
sample sequence

deeplearning.ai

Recurrent Neural Networks

Sampling novel sequences

Sampling a sequence from a trained RNN



Andrew Ng

randomly sample using the softmax distribution

if [EOS] in your vocab, keep sampling until you hit it.

If not, can set the target # of words

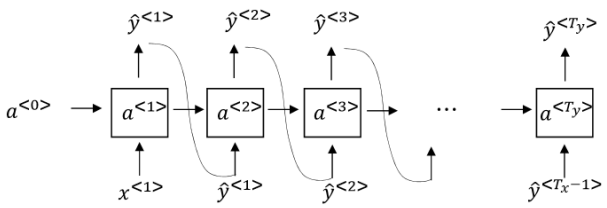
can force out [UNK]

Character-level language model

→ Vocabulary = [a, aaron, ..., zulu, <UNK>]

→ Vocabulary = [a, b, c, ..., z, ␣, ., ,, ;, 0, ..., 9, A, ..., Z]

Cat average ... Mau



Andrew Ng

Can do character-level

Same rationale for character level

add: No need to worry about unknown words.

disad: much longer sequences!

⇒ more computationally expensive

- Usually word level is used

character level

$x^{<1>}$ $\hat{y}^{<1>}$ $\hat{y}^{<2>}$ $\hat{y}^{<T_x-1>}$

Andrew Ng

- Usually word level is used
But in some cases — character level
Rare used for cases you need to deal with lots of unknown words.

Sequence generation

News

President enrique peña nieto, announced sench's sulk former coming football langston paring.

"I was not at all surprised," said hich langston.

"Concussion epidemic", to be examined. ←

The gray football the told some and this has on the uefa icon, should money as.

Shakespeare

The mortal moon hath her eclipse in love.

And subject of this thou art another this fold.

When besser be my love to me see sabl's.

For whose are ruse of mine eyes heaves.

← Assignment content

Andrew Ng

vanish gradient

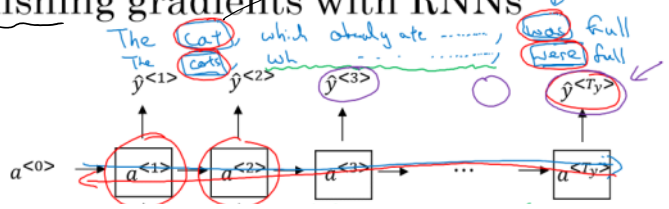
deeplearning.ai

Recurrent Neural Networks

Vanishing gradients with RNNs

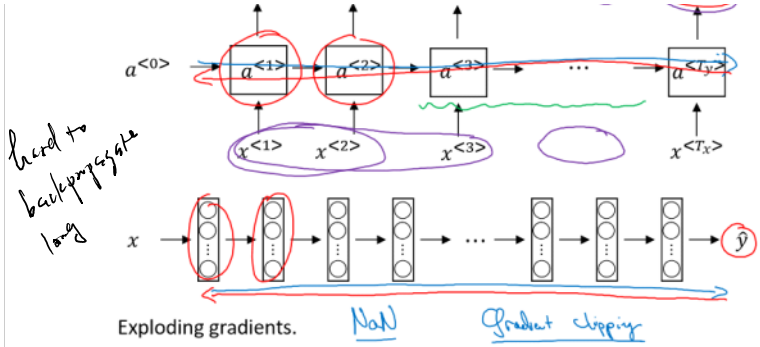
Vanishing gradients with RNNs

cat 和 was/were 的对应关系



Very long term dependency

the RNN introduced above. does not take care of this long term dependency



hard to backpropagate long

100

Vanishing gradient is a bigger problem.

Need to memorize for a very long time.

previous. : ~~of~~ local influence

can't back propogate to many steps before

But exploding gradient is also BAD — easy to observe though. NaN of param

Solution: clip it gradient clipping

Next few videos: focus on vanishing gradient

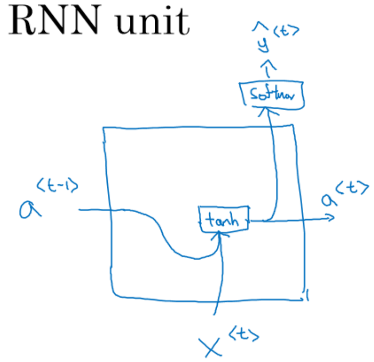
gru

deeplearning.ai

Recurrent Neural Networks

Gated Recurrent Unit (GRU)

RNN unit

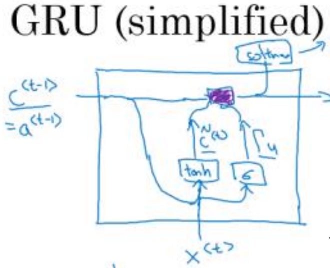


tanh

$$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a)$$

Andrew Ng

GRU (simplified)



c = memory cell

some 1000 dim.

$c^{<t>} = a^{<t>}$ the same value

$\tilde{c}^{<t>} = \tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c)$

element-wise

$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$ "update"

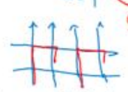
$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1-\Gamma_u) * c^{<t-1>}$

element-wise

$\Gamma_u = 1$ $\Gamma_u = 0$ $\Gamma_u = 0$ $\Gamma_u = 0$ ...... $\Gamma_u = 1$

$c^{<t>} = 1$ ...... $= 1$

→ The cat, which already ate ..., was full.



Gate

$\Gamma_u = 0.00001$

[Cho et al., 2014. On the properties of neural machine translation: Encoder-decoder approaches]
[Chung et al., 2014. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling]

Andrew Ng

感觉是个 weighted average

c: memory cell

Q 那前一个词会很用吗？ 还是说完全是c.

the gate is easy to be set to zero.

sigmoid

gate: decide whether to update [binary]

the gate decide whether to update

if $\Gamma_u = 1$: update
if $\Gamma_u = 0$: not update. still the old c.

$\Gamma_u \in (0,1)$ 吧？

Interesting!

9

Γ_u close to zero

Full GRU

$\tilde{h}$ $\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$

u $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$

r $\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$

h $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1-\Gamma_u) + c^{<t-1>}$

alternative notation

LSTM

Γ_r: relevant

? 得读一读这个GRU 的paper.

GRU one of the most commonly used method to tackle vanishing gradient & model Long Term memory.

The cat, which ate already, was full.

Andrew Ng

Andrew Ng

lstm

Recurrent Neural Networks

LSTM (long short term memory) unit

deeplearning.ai

GRU and LSTM

GRU

$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$

$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$

$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$

$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

$a^{<t>} = c^{<t>}$

LSTM

$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$

(update) $\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$

(forget) $\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$

(output) $\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$

$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$

$a^{<t>} = \Gamma_o * c^{<t>}$

[Hochreiter & Schmidhuber 1997. Long short-term memory]

Andrew Ng

the paper is difficult to read. more about the vanishing gradient problem

more powerful & general model than GRU.

update gate

output gate

3 Gates slightly more complicated.

LSTM units

GRU	LSTM
$\tilde{c}^{<t>} = \tanh(W_c[\Gamma_r * c^{<t-1>}, x^{<t>}] + b_c)$	$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$
$\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$	$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$
$\Gamma_r = \sigma(W_r[c^{<t-1>}, x^{<t>}] + b_r)$	$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$
$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$	$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$
$a^{<t>} = c^{<t>}$	$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$
	$a^{<t>} = \Gamma_o * c^{<t>}$

[Hochreiter & Schmidhuber 1997. Long short-term memory]

Andrew Ng

LSTM in pictures

$\tilde{c}^{<t>} = \tanh(W_c[a^{<t-1>}, x^{<t>}] + b_c)$

$\Gamma_u = \sigma(W_u[a^{<t-1>}, x^{<t>}] + b_u)$

$\Gamma_f = \sigma(W_f[a^{<t-1>}, x^{<t>}] + b_f)$

$\Gamma_o = \sigma(W_o[a^{<t-1>}, x^{<t>}] + b_o)$

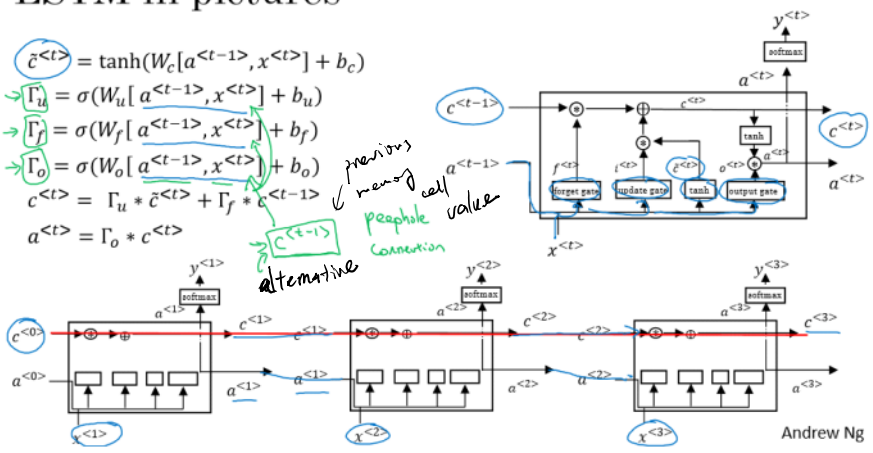
$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>}$

$a^{<t>} = \Gamma_o * c^{<t>}$

previous memory cell value

peephole connection

alternative



Andrew Ng

One variation of LSTM.

Peephole connection

用 $c^{<t-1>}$

good at memorizing things for a long time

GRU or LSTM? No consensus. LSTM comes much earlier.

GRU is treated as a simplified model.

Andrew Ng: GRU simpler. easier to build a bigger model

LSTM is a more historically proven model.

Many teams will use it as the default.

But GRU is gaining momentum in recent yrs

bidirectional rnn

deeplearning.ai

Recurrent Neural Networks

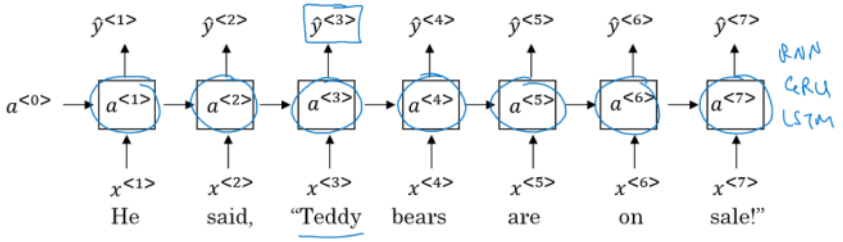
Bidirectional RNN

Getting information from the future

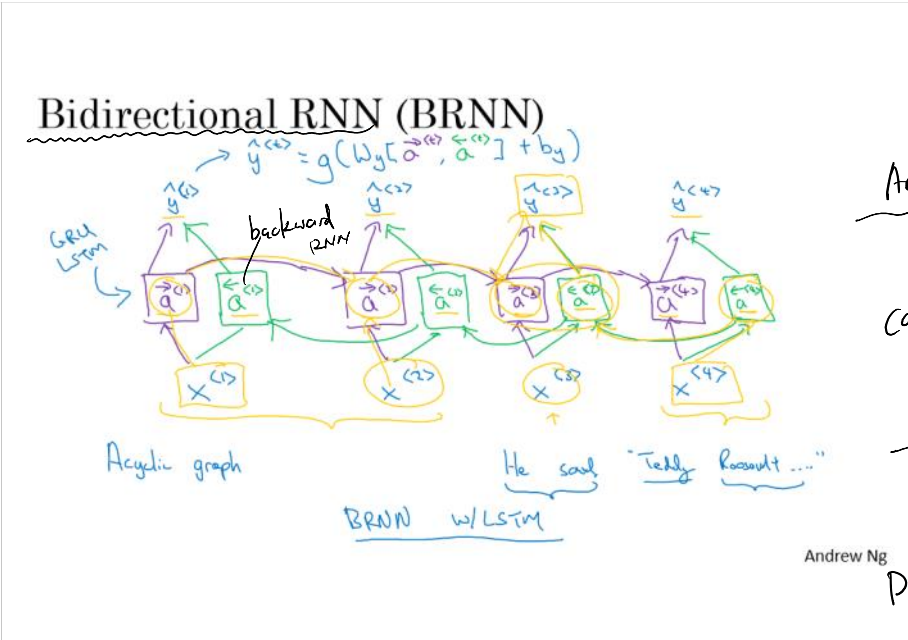
later words help determine earlier words

He said, "Teddy bears are on sale!"

He said, "Teddy Roosevelt was a great President!"



Andrew Ng



Acyclic graph

Can be GRU. LSTM. blocks

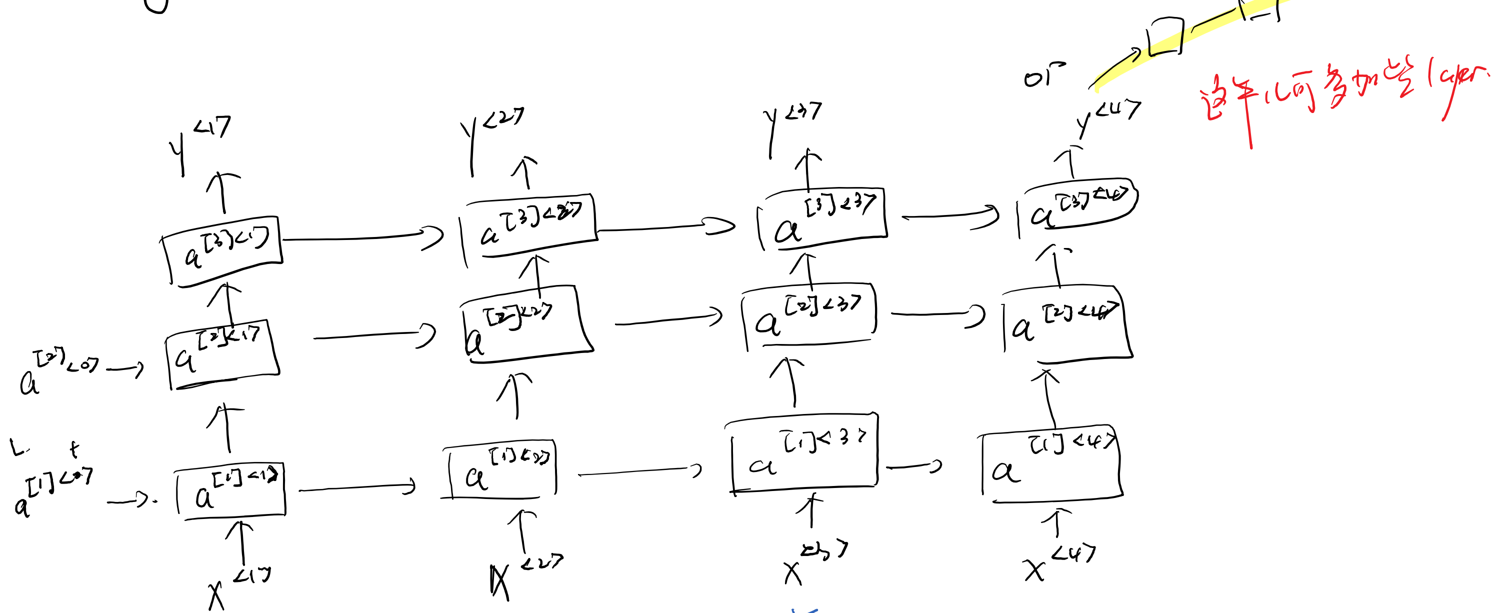
Bidirection + LSTM commonly used

A reasonable first thing to try.

Disadvantage: Need the entire sequence before making prediction

e.g. Speech recognition. Start processing after the person stop talking

Taking all these ideas ⟹ construct Deep version.



RNN 少见许多 layer 的 (3个 layer 已经算多的了)
因为 RNN 本身 model 就比较复杂.