Introduction to Deep Learning
Chapter 6: Recurrent Neural Networks

Pao-Ann Hsiung

National Chung Cheng University

Contents

Contents

Why Sequence Models?

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.

Motivating Example: Named Entity Recognition

x:	Harry	Potter	and	Hermione	Granger	invented	a	new	spell.	
	$x^{\langle 1 \rangle}$	$x^{\langle 2 \rangle}$	$x^{\langle 3 \rangle}$	…………………….		$x^{\langle t \rangle}$			$x^{\langle 9 \rangle}$	
	1	1	0	1	1	0	0	0	0	
y:	$y^{\langle 1 \rangle}$	$y^{\langle 2 \rangle}$	$y^{\langle 3 \rangle}$	…………………….		$y^{\langle t \rangle}$			$y^{\langle 9 \rangle}$	

$x^{(i)\langle t \rangle}$, $y^{(i)\langle t \rangle}$: t^{th} element in the input or output of the i^{th} training sample

$T_x^{(i)} = 9$, $T_y^{(i)} = 9$: lengths of input and output sequences in the i^{th} training example

Representing words

浪費很多空間沒有效率。

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

$x^{<1>}$ $\quad x^{<2>}$ $\quad x^{<3>}$ $\quad \ldots \quad$ $x^{<9>}$

Vocabulary

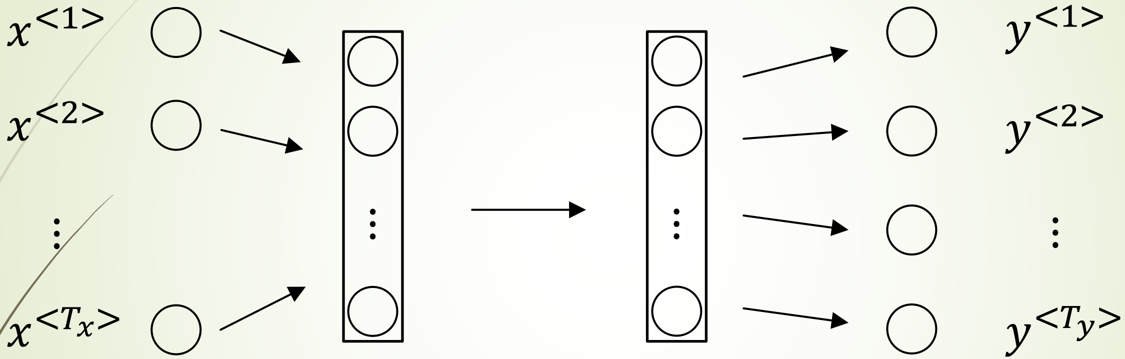
$$\begin{bmatrix} a \\ aaron \\ \vdots \\ and \\ \vdots \\ harry \\ \vdots \\ potter \\ \vdots \\ zulu \end{bmatrix} \quad \begin{array}{c} 1 \\ 2 \\ \vdots \\ 367 \\ \vdots \\ 4075 \\ \vdots \\ 6{,}830 \\ \vdots \\ 10{,}000 \end{array} \qquad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{array}{c} \\ \\ \\ \\ \\ \leftarrow 4075 \\ \\ \\ \\ \end{array} \qquad \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix} \begin{array}{c} \\ \\ \\ \\ \\ \leftarrow 6830 \\ \\ \\ \\ \end{array}$$

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

One-hot encoding

Contents

Why not a standard network model?



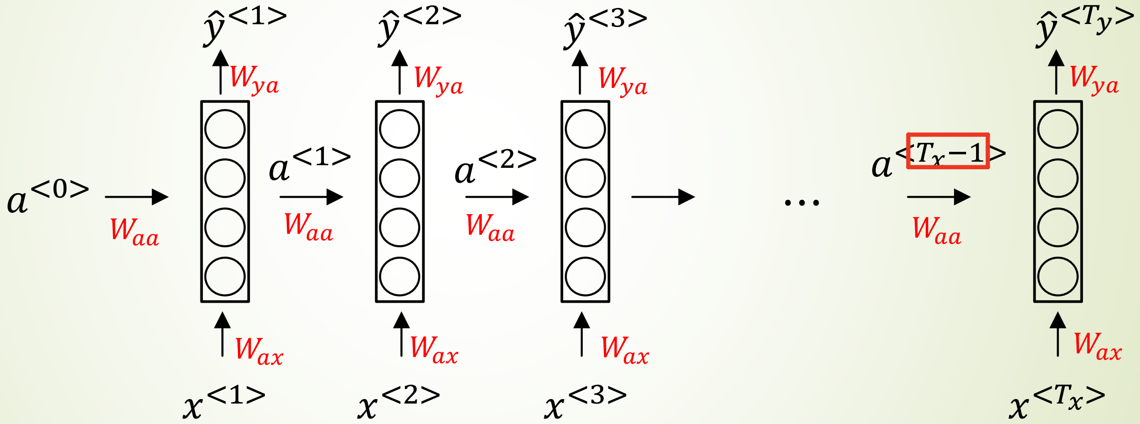
Problems:

- Inputs, outputs can be different lengths in different examples.
- Doesn't share features learned across different positions of text.

Recurrent Neural Networks

- RNN uses information from the previous inputs

Randomly initialized or Vector Zero



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

He said, "Teddy bears are on sale!"

Forward Propagation for RNNs

$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

$$a^{<t>} = g(W_a[a^{<t-1>}, x^{<t>}] + b_a) \text{ where } W_a = [W_{aa}W_{ax}], [a, x] = \begin{bmatrix} a \\ x \end{bmatrix}$$

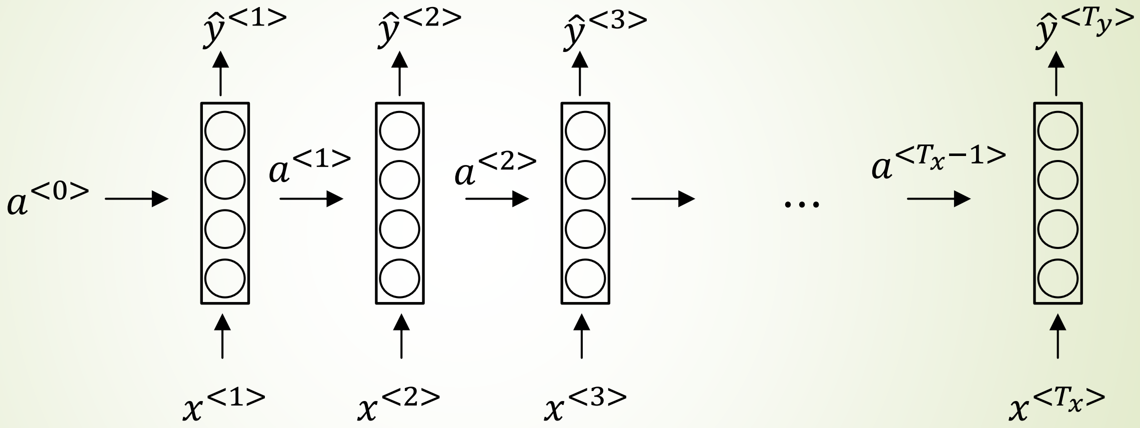
- g() could be tanh or ReLu

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

$$\hat{y}^{<t>} = g(W_a a^{<t>} + b_y)$$

- g() could be Sigmoid or Softmax

Backpropagation through time



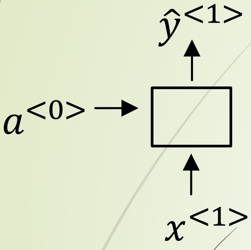
$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = -y^{<t>} \log \hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>})$$

$$\mathcal{L}(\hat{y}, y) = \sum_{t=1}^{T_y} \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

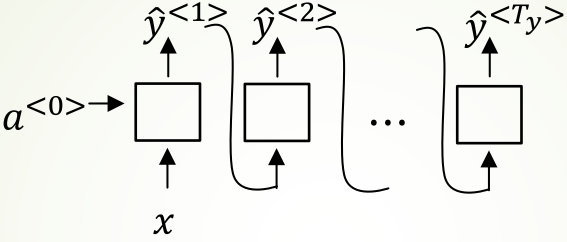
Different Types of RNN

- French-English Translation: many-to-many
  - Voulez vous chanter avec moi? (5 words)
  - Would you like to size with me? (7 words)
  - $T_x \neq T_y$
- Sentiment Classification: many-to-one
  - x = text, y = 1,…,5 (stars) or 0/1 (negative/positive)
- Standard: One-to-one
- Music Generation: One-to-many
  - x = music style (genre), y = music

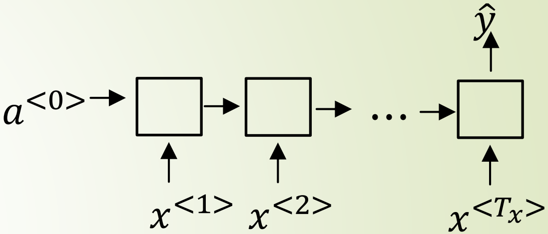
Different Types of RNN



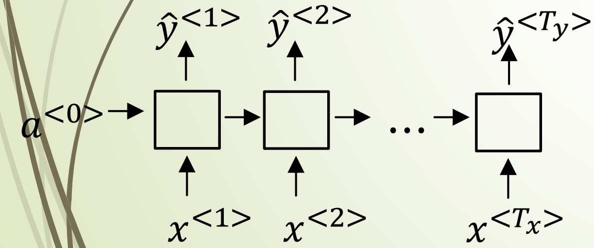
One to one



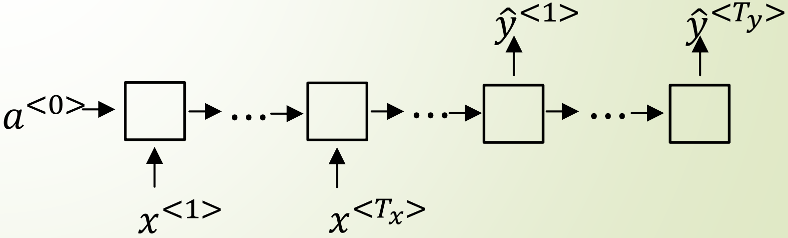
One to many



Many to one



Many to many



Many to many

Contents

What is language modelling?

Speech recognition

The apple and pair salad.

The apple and pear salad.

$$P(\text{The apple and pair salad}) = 3.2 \times 10^{-13}$$

$$P(\text{The apple and pear salad}) = 5.7 \times 10^{-10}$$

Output the sentence (sequence) with a high probability $P\left(y^{<1>}, y^{<2>}, \ldots, y^{<T_y>}\right)$.

Language Modelling with an RNN

Training set: large corpus of english text.

- Tokenize

Cats average 15 hours of sleep a day.

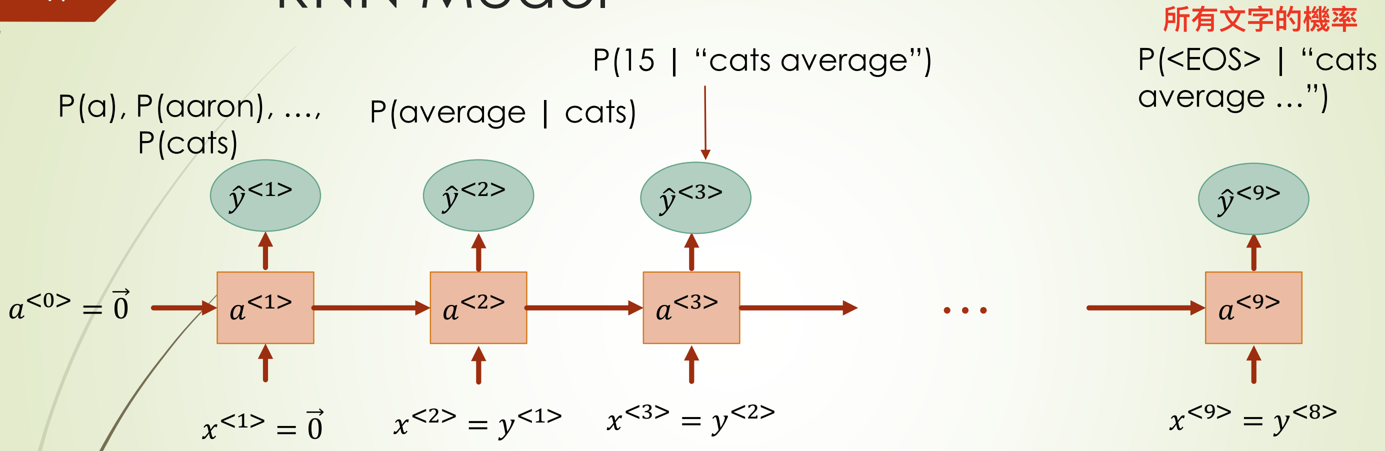
$y^{<1>}$ $y^{<2>}$ $y^{<3>}$ $y^{<8>}$ $y^{<9>}$

?

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

Unknown word is represented by **<UNK>**

RNN Model



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

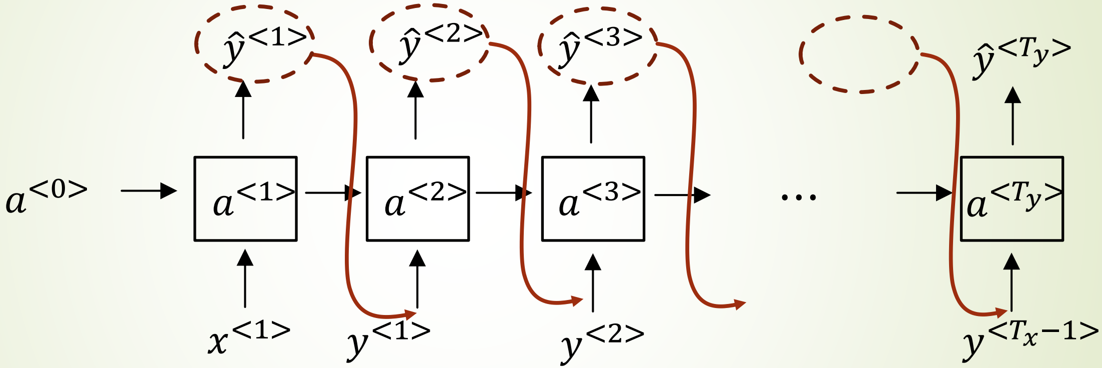
$$\mathcal{L}(\hat{y}^{<t>}, y^{<t>}) = -\sum_i y_i^{<t>} \log \hat{y}_i^{<t>}$$

$$\mathcal{L} = \sum_t \mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>})$$

$$P(y^{<1>}, y^{<2>}, y^{<3>}) = P(y^{<1>}) \times P(y^{<2>} | y^{<1>}) \times P(y^{<3>} | y^{<1>}, y^{<2>})$$

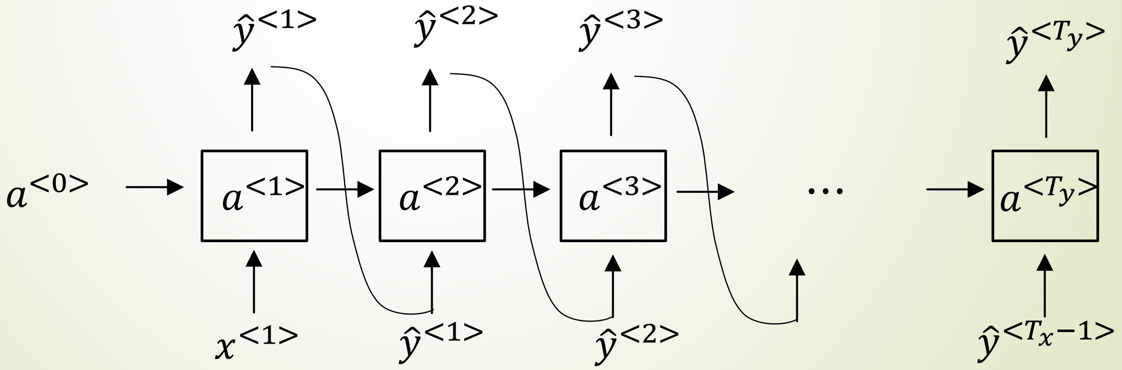
Sampling Novel Sequences

Random sampling: np.random.choice()



Character-level language model

Vocabulary = [a, aaron, …, zulu, <UNK>]
Vocabulary = [a, b, c, …, z, “ ”, ‘.’ , ‘,’ , ‘;’, 0, …, 9, A, … Z]



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.

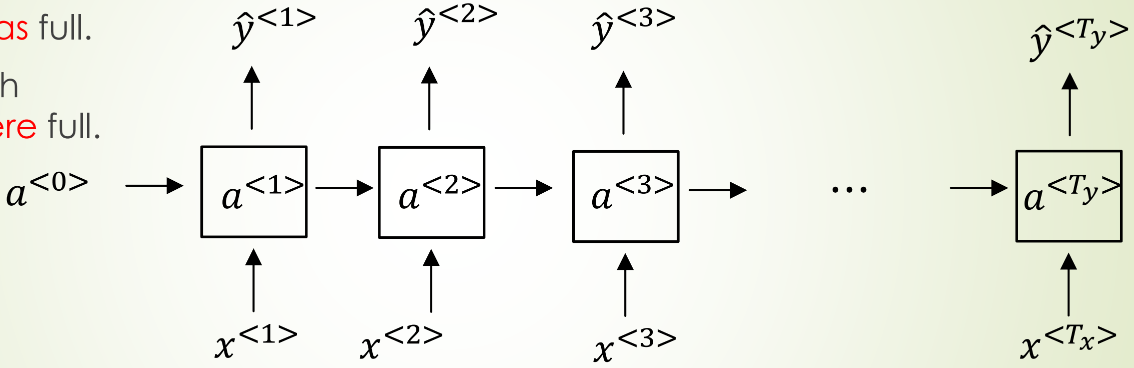
And subject of this thou art another this fold.

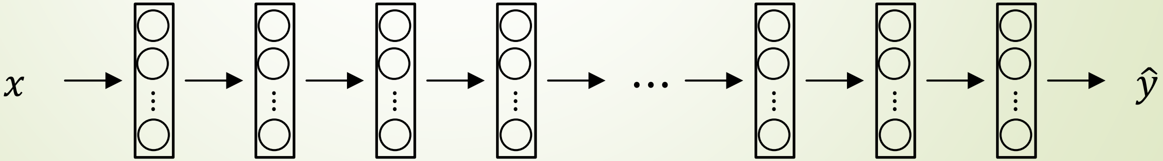
When besser be my love to me see sabl's.

For whose are ruse of mine eyes heaves.

Vanishing Gradients with RNNs

- The cat, which already … , was full.
- The cats, which already …, were full.



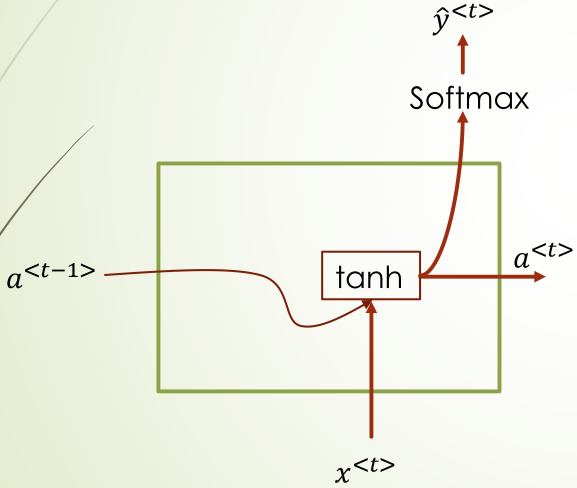


Exploding gradients easy to detect (NaNs), use gradient clipping (re-scale)
Vanishing gradients much harder to detect and solve.

Contents

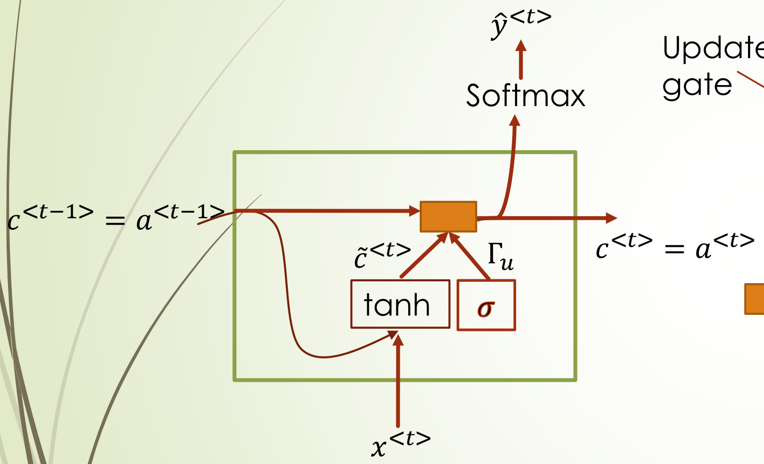
- Introduction
- Recurrent Neural Networks (RNN)
- Language Model and Sequence Generation
- Gated Recurrent Unit (GRU)
- Long Short Term Memory (LSTM)
- Bidirectional RNN
- Deep RNN

RNN Unit



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

GRU (simplified)



- C = memory cell, $c^{<t>} = a^{<t>}$
- $\tilde{c}^{<t>} = \boxed{\tanh(W_c[c^{<t-1>}, x^{<t>}] + b_c})$
- $\Gamma_u = \sigma(W_u[c^{<t-1>}, x^{<t>}] + b_u)$ (mostly 0 or 1)
- $c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + (1 - \Gamma_u) * c^{<t-1>}$

Element-wise multiplication

- Γ_u is often very close to 0, so the memory cell is maintained.

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

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

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

[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]

Full GRU

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

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

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

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

The cat, which ate already, was full.

Contents

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

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

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

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

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

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

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

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

LSTM in pictures

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

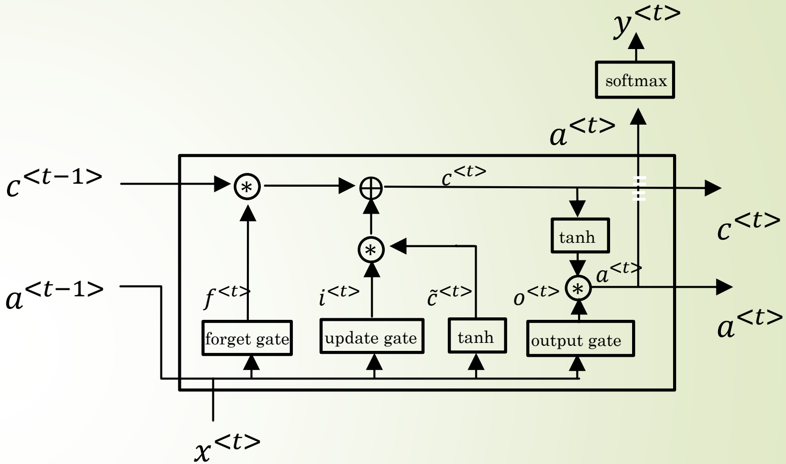
$$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$$

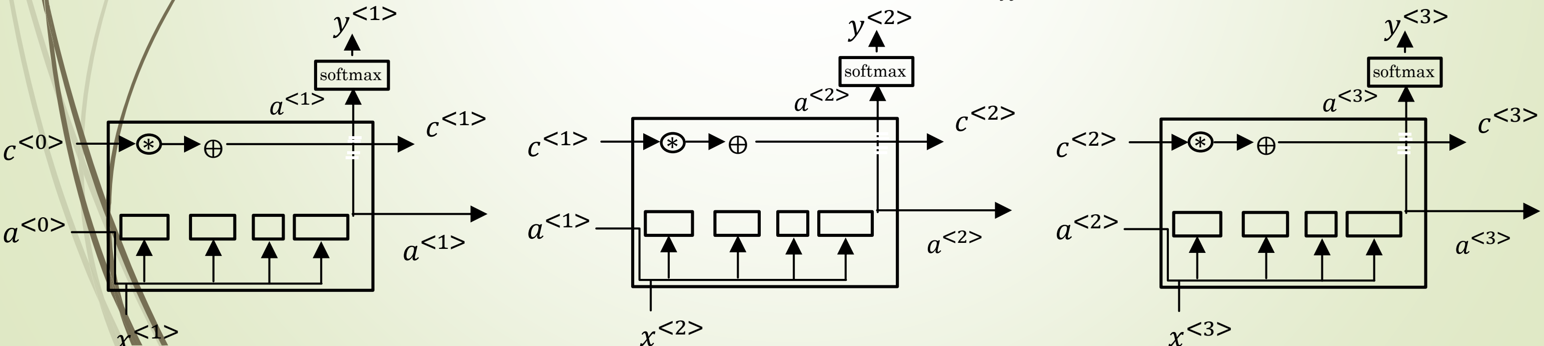
$$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$$

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

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

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



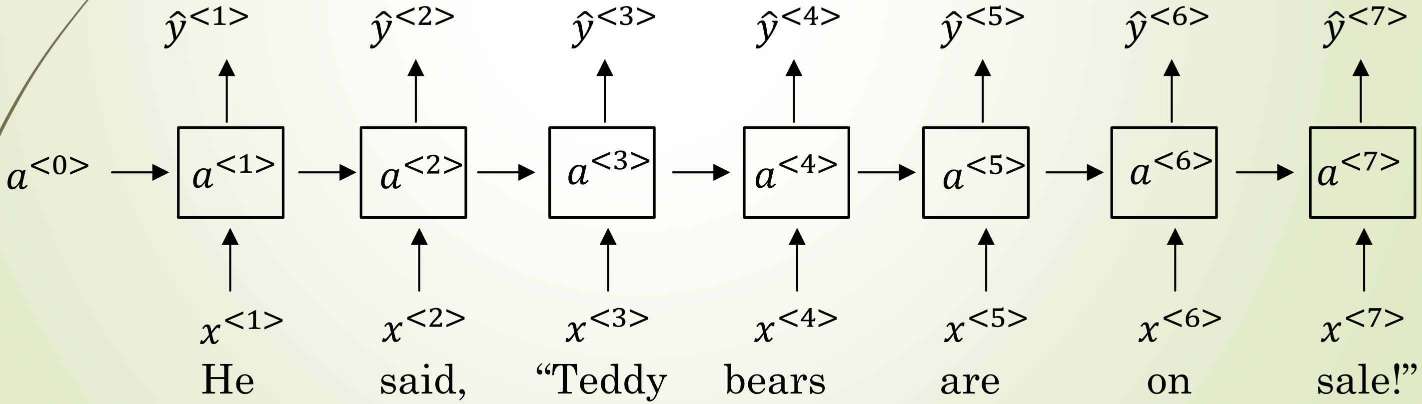


Contents

Single direction RNN

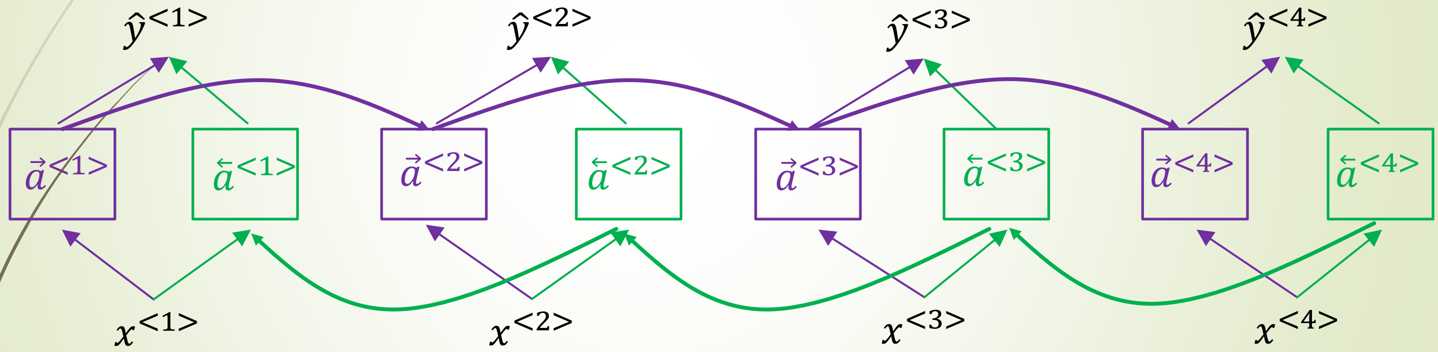
He said, “Teddy bears are on sale!”

He said, “Teddy Roosevelt was a great President!”



Bidirectional RNN (BRNN)

$$\hat{y}^{<t>} = g(W_y[\vec{a}^{<t>}, \overleftarrow{a}^{<t>}] + b_y)$$

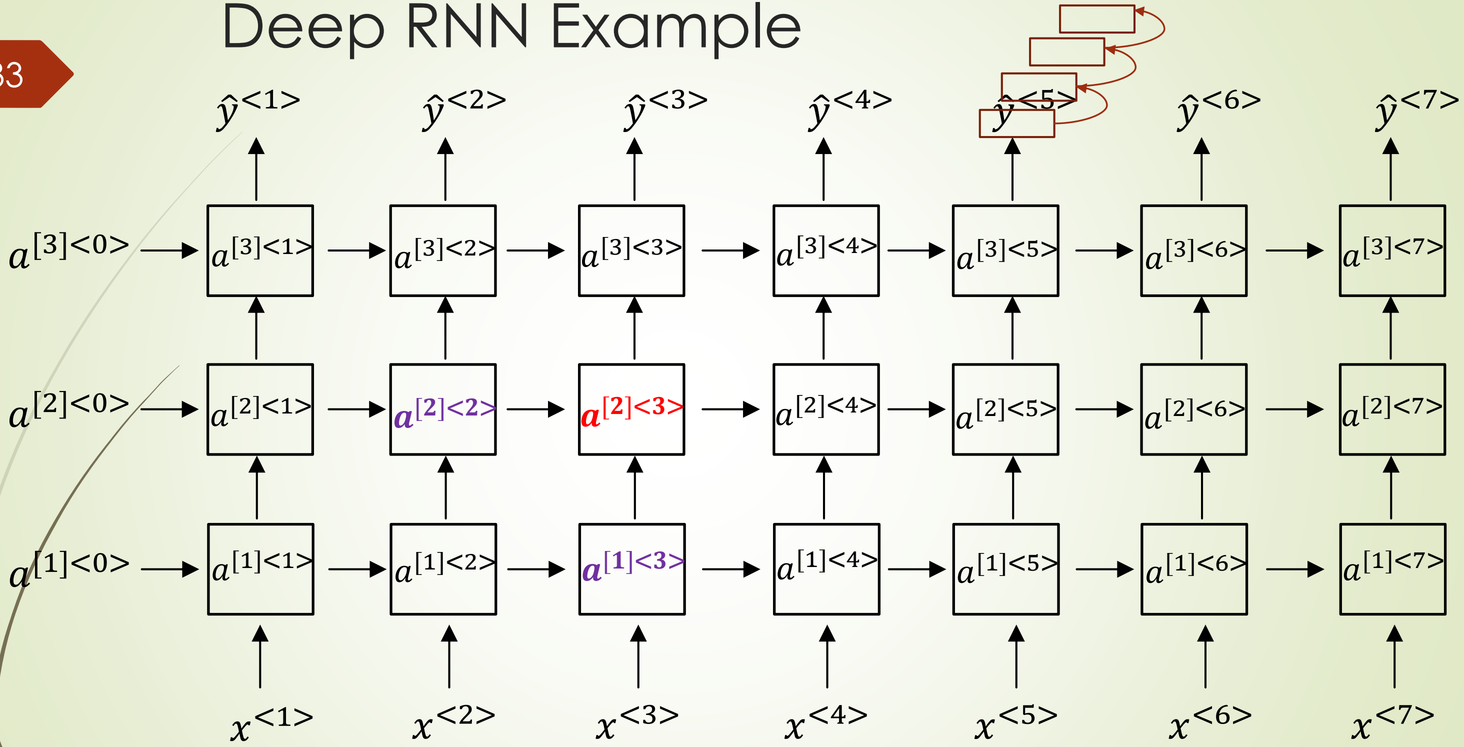


He said, "Teddy Roosevelt was a great President!"

Blocks can be GRU or LSTM units

Contents

Deep RNN Example



$$a^{[2]<3>} = g(W_a^{[2]}[a^{[2]<2>}, a^{[1]<3>}] + b_a^{[2]})$$