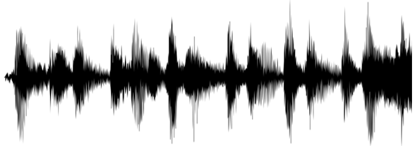
NLP

Week-3

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.

Motivating example

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

Named Entity Recognition
Find out input and output.

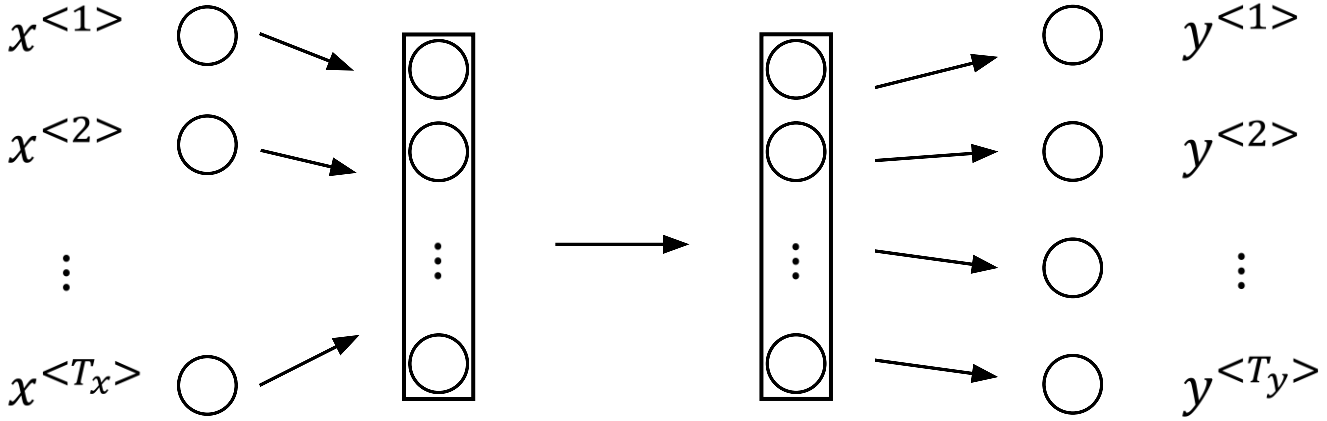
Representing words

Use Dictionary for the feature representation of words

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

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

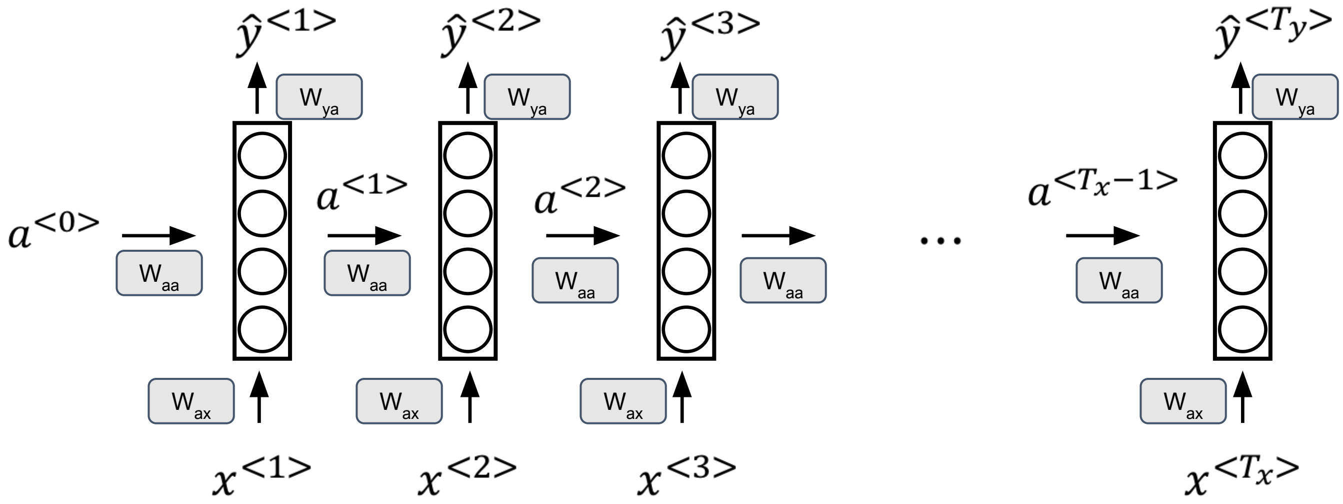
Why not a standard network?



Problems:

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

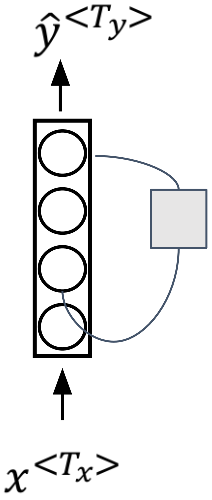
Recurrent Neural Networks



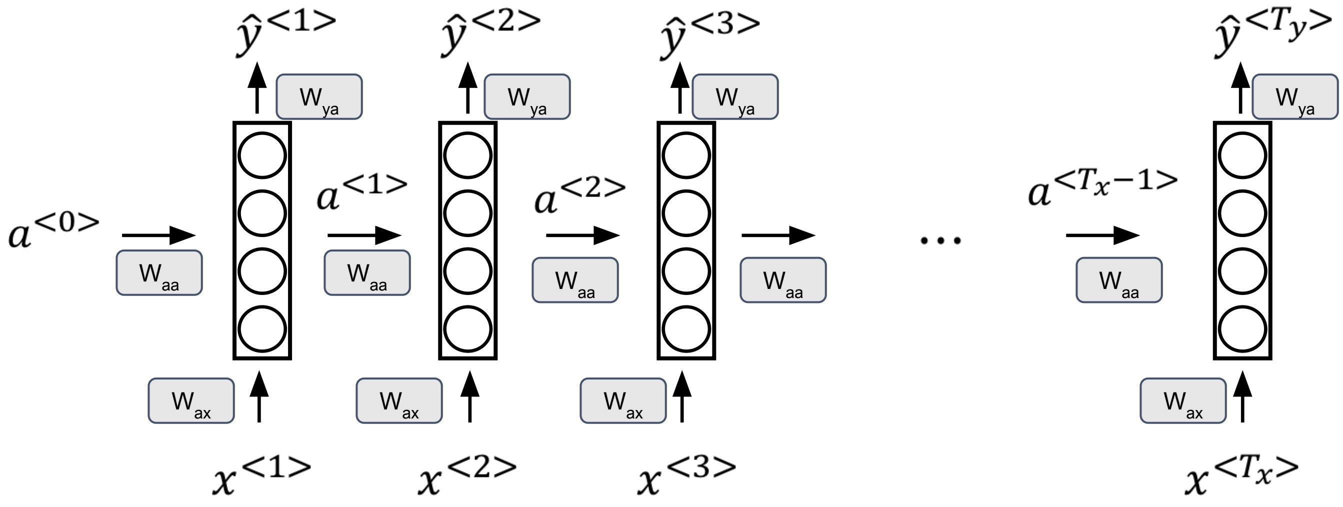
He said, “Teddy Roosevelt was a great President.”

He said, “Teddy bears are on sale!”

Any prediction of words uses the information preceding it. It does not uses the information from the latter word in the sequence



Forward Propagation

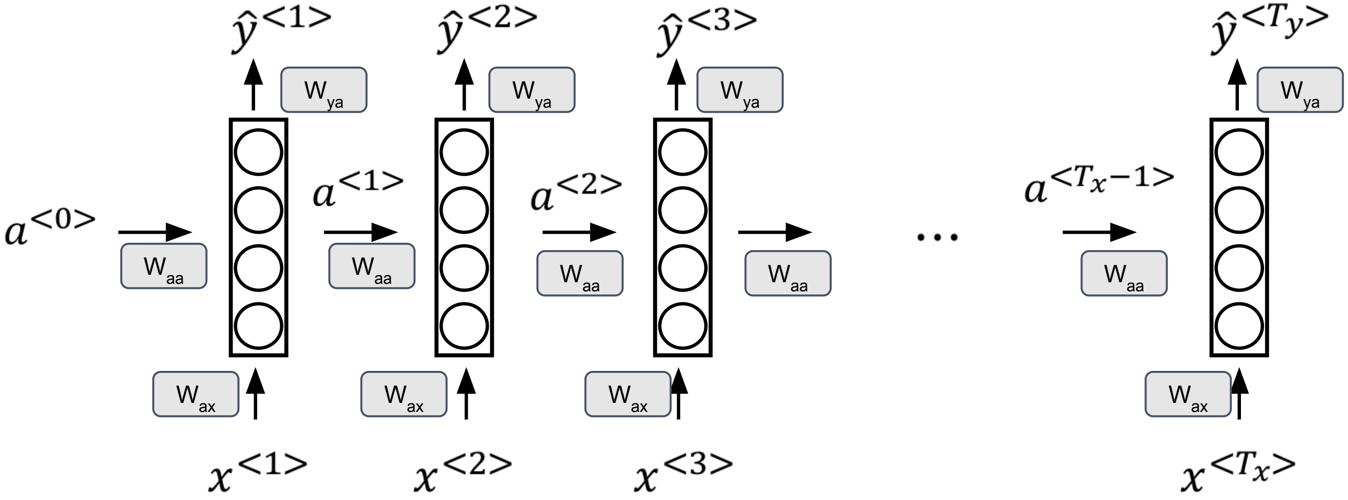


$$a^{<t>} = g(W_{aa}a^{<t-1>} + W_{ax}x^{<t>} + b_a)$$

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

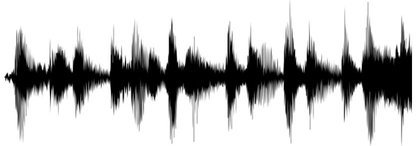
$$a^{<t>} = g(W_a\,[a^{<t-1>}, x^{<t>}] + b_a)$$

Forward propagation and backpropagation

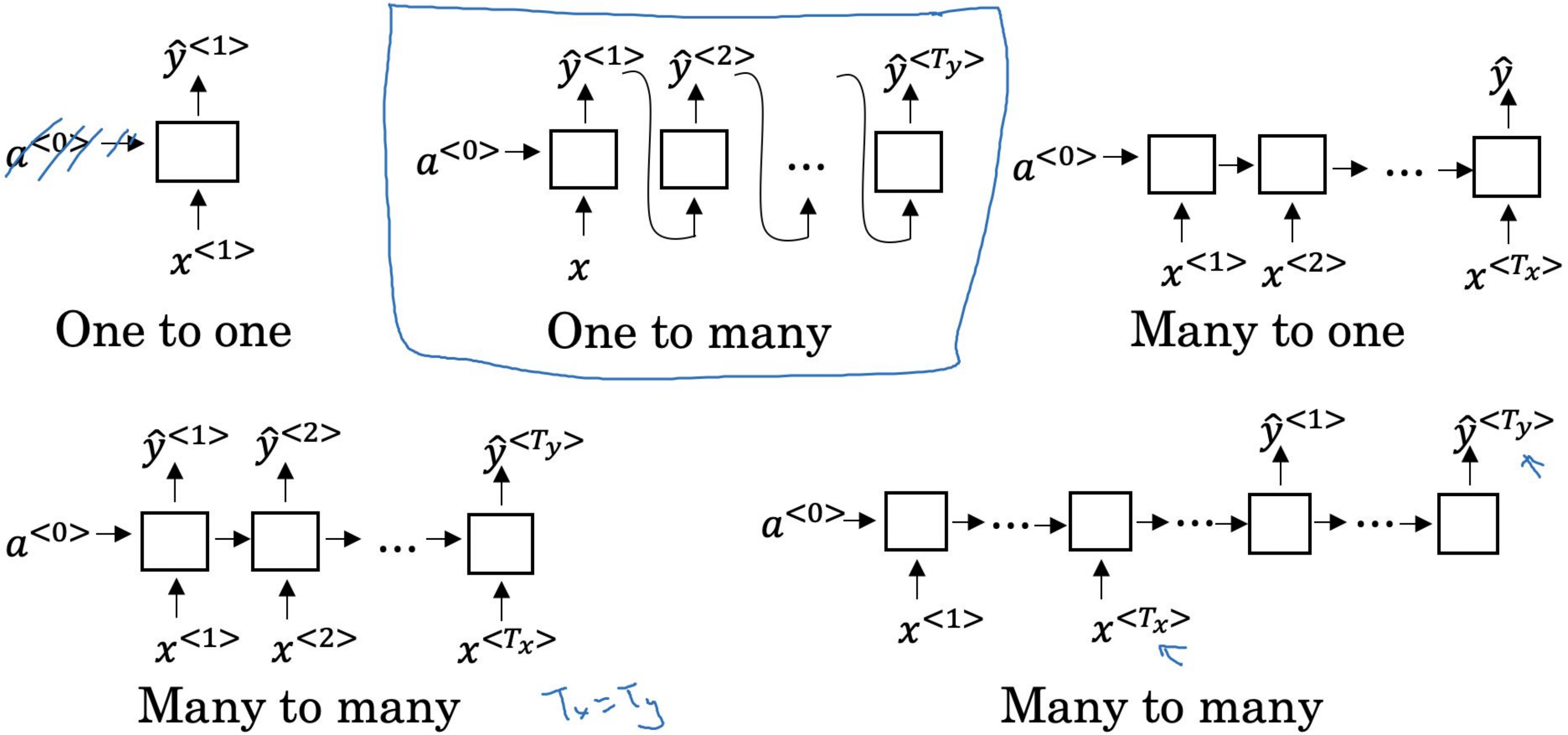


$$\mathcal{L}^{<t>}(\hat{y}^{<t>}, y^{<t>}) = \quad -y^{<t>} \log\hat{y}^{<t>} - (1 - y^{<t>}) \log(1 - \hat{y}^{<t>})$$

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.

Examples of RNN types



What is language modelling?

Speech recognition

The apple and pair salad.

The apple and pear salad.

$$P(\text{The apple and pair salad}) =$$

$$P(\text{The apple and pear salad}) =$$

Language model tells what is the probability of the sentence being correct. It estimates the probability of the given sequence of words

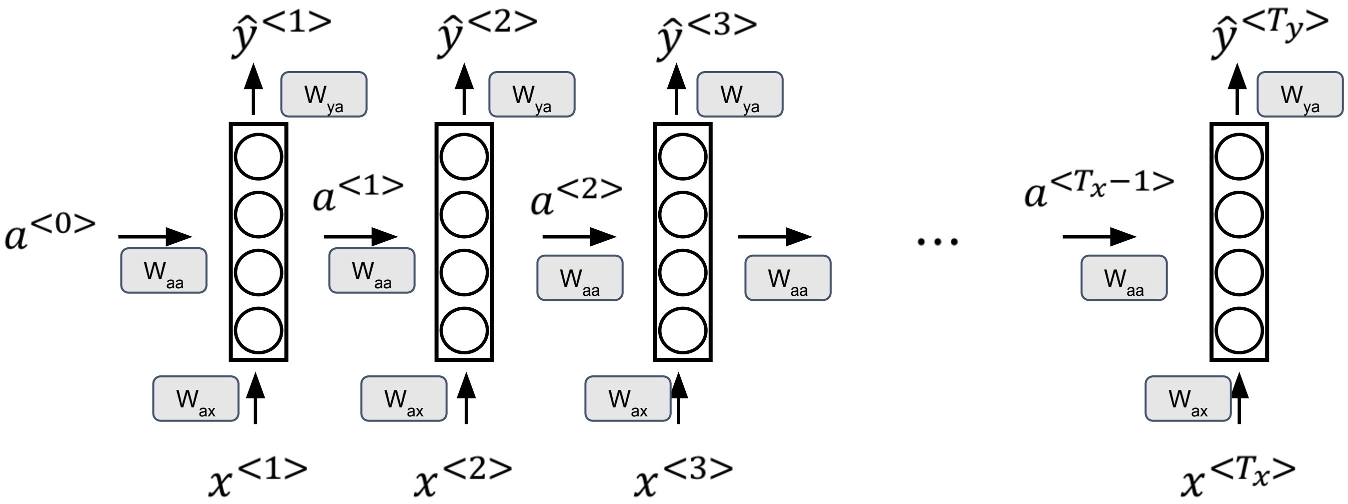
Language modelling with an RNN

Training set: large corpus of english text.

Cats average 15 hours of sleep a day.

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

RNN model



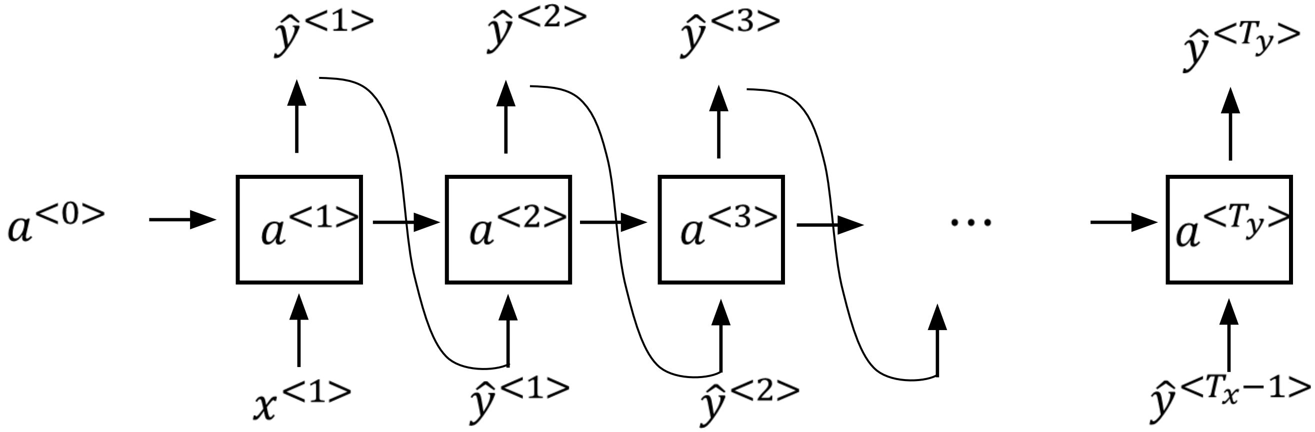
Cats average 15 hours of sleep a day. <EOS>

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

At every step of RNN, we look at some set of preceding words.

Character-level language model

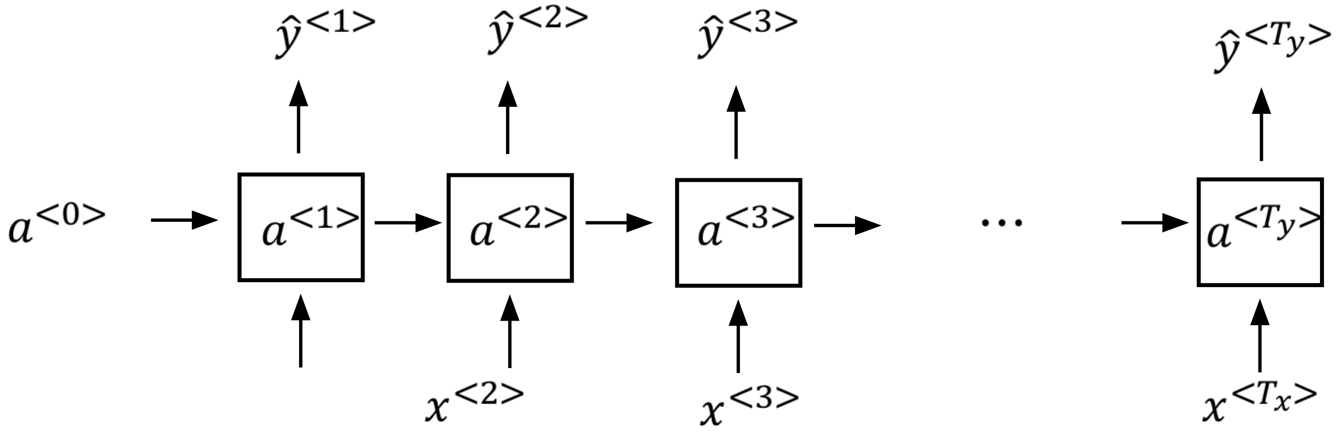
Vocabulary = [a, aaron, ..., zulu, <UNK>]



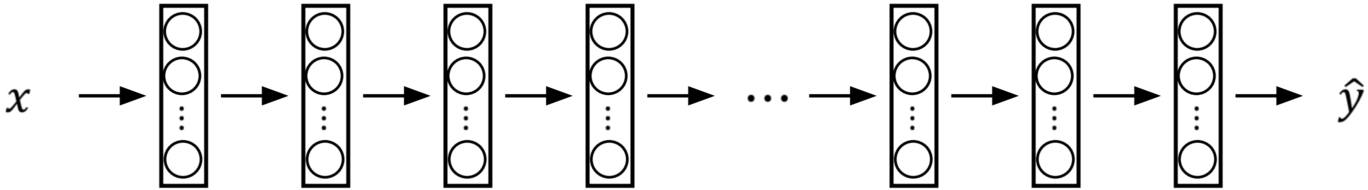
No need to worry about <UNK> word token

Ends up in much longer sequences
Computationally Expensive

Vanishing gradients with RNNs



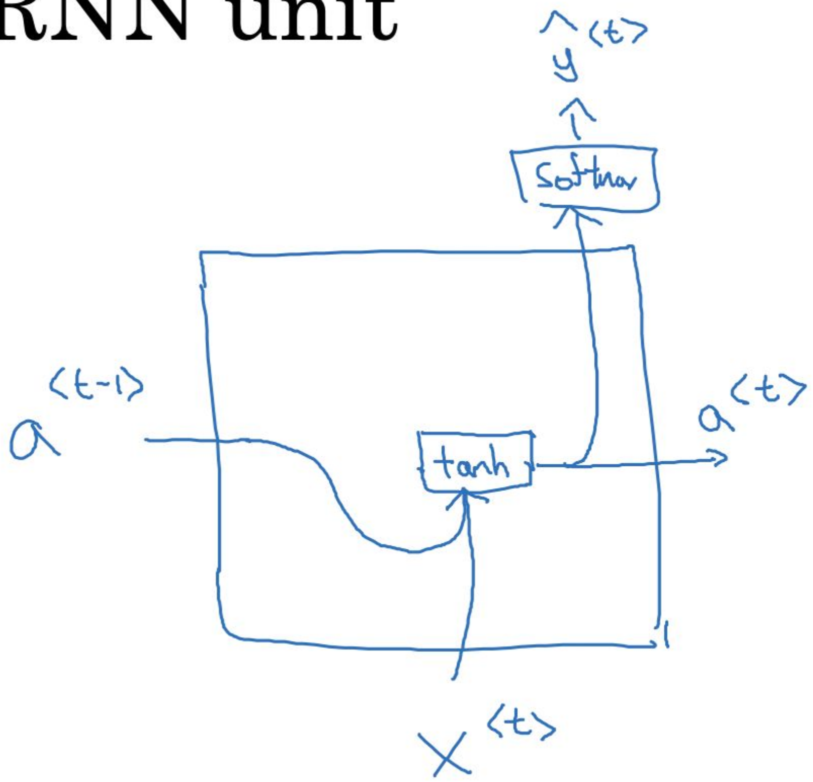
Earlier words determines what word comes later.
Output is influenced by the input which is very early in the sequence.
This makes RNN unable to catch long range dimensions.



Gradient clipping - Solution for exploding gradient.
If Gradient is beyond certain threshold then rescales the values to avoid NaNs.

Exploding gradients.

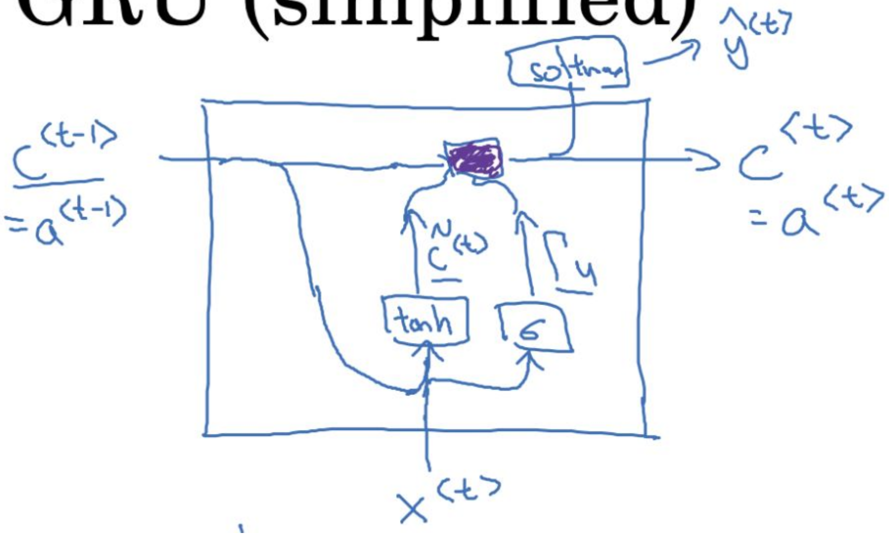
RNN unit



tanh

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

GRU (simplified)



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

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

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

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

The cat, which already ate …, was full.

LSTM units

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

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

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

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

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

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

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

LSTM in pictures

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

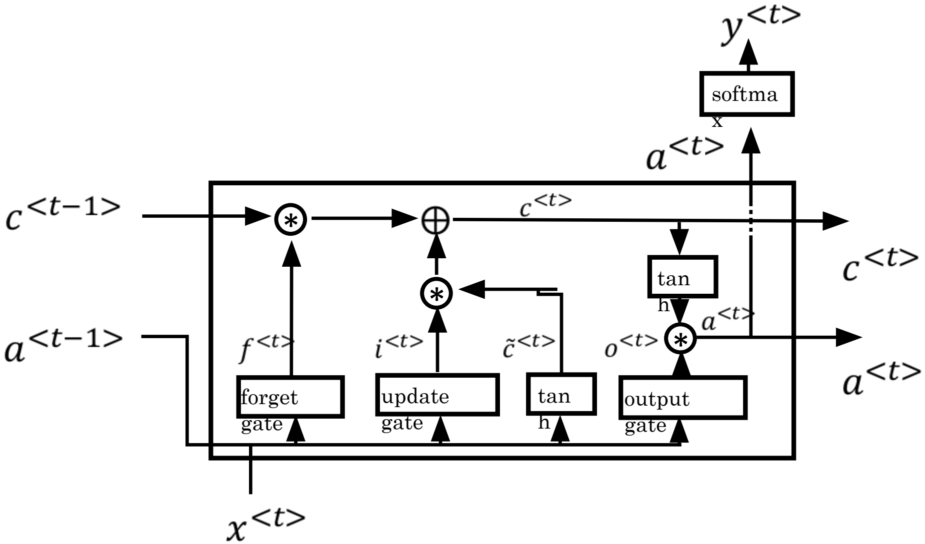
$$\Gamma_u = \sigma(W_u[\, a^{<t-1>}, x^{<t>}] + b_u)$$

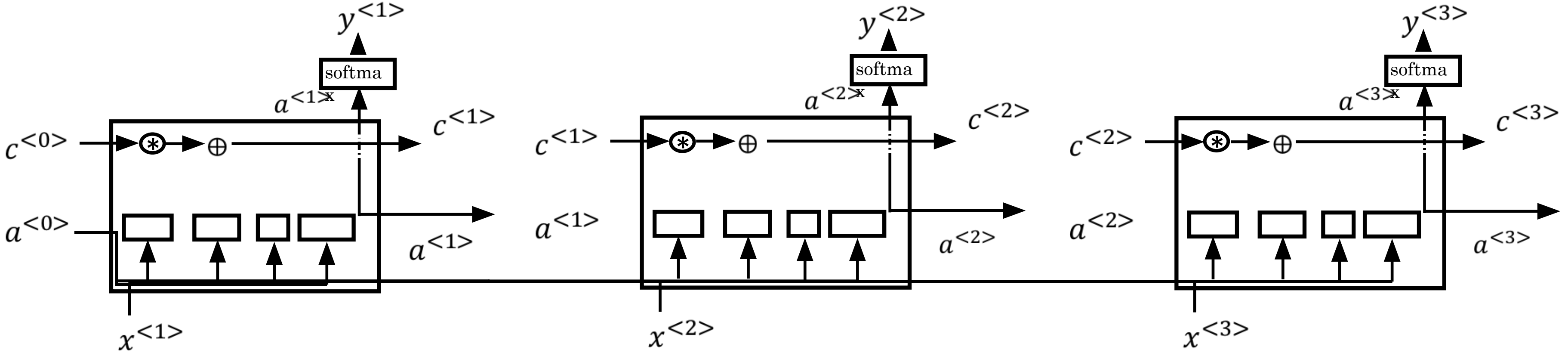
$$\Gamma_f = \sigma(W_f[\, a^{<t-1>}, x^{<t>}] + b_f)$$

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

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

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

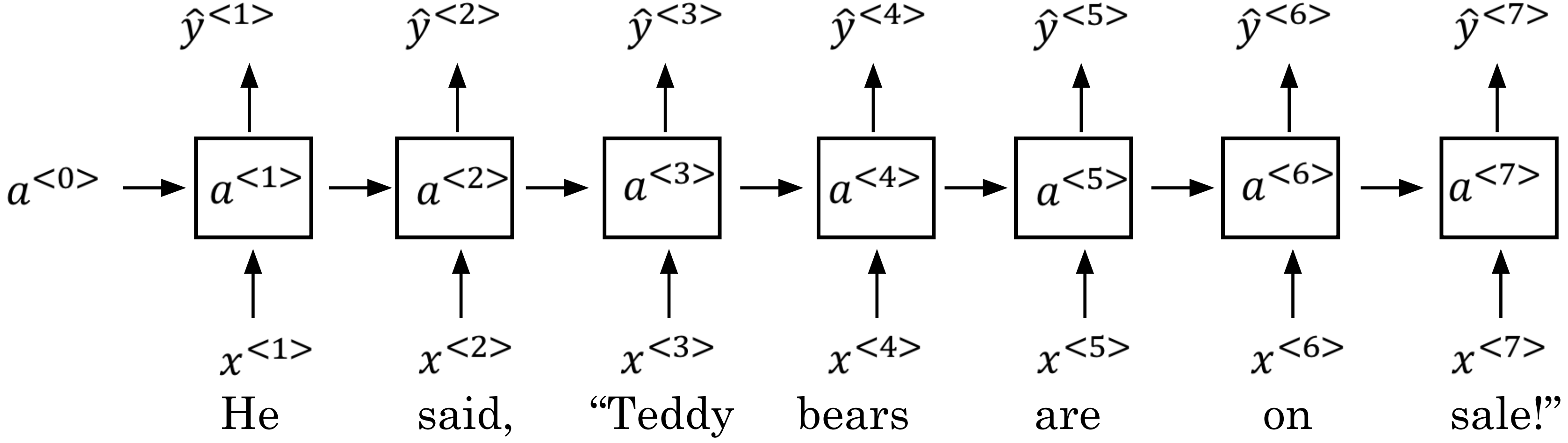




BRNN - Getting information from the future

He said, “Teddy bears are on sale!”

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



Deep RNN example