Recurrent Neural Network (RNN)

COSC 6336: Natural Language Processing
Spring 2020

Lecturer : Niloofar Safi Samghabadi

Some content on these slides was borrowed from Andrew Ng's online course

This Week

- ★ **Understand a Recurrent Neural Network**
  - Forward Propagation in RNN
  - Backpropagation Through Time (BPTT)
  - Different Types of RNN
  - Vanishing and Exploding Gradient Problem
- ★ **Long Short-Term Memory (LSTM)**
- ★ **Gated Recurrent Unit (GRU)**
- ★ **Bidirectional RNN**

Sequence Data

★ Speech Recognition

Sequence Data

★ Speech Recognition

X | Y

★ Translation

How are you? → Wie geht es Ihnen?

“English” “German”

Sequence Data

X Y

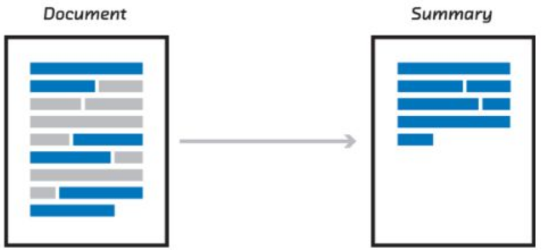
★ Speech Recognition

★ Translation

How are you? → Wie geht es Ihnen?

“English” **“German”**

★ Summarization



Sequence Data

★ Sentiment Classification

X **Y**

"Not to my taste, will skip and watch another movie!" → 

Sequence Data

X Y

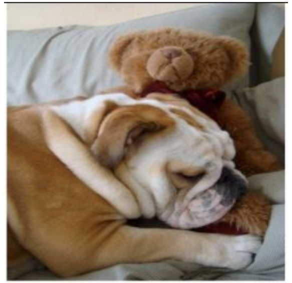
★ Sentiment Classification

"Not to my taste, will skip and watch another movie!" → 

★ Music Generation

Genre → 

Sequence Data

	X	Y
★ Sentiment Classification		
	"Not to my taste, will skip and watch another movie!" →	
★ Music Generation		
	Genre →	
★ Image Captioning		
	 →	A bulldog resting on and sleeping with a teddy bear.

Word Representation

X: Michael Jordan is a professor at Berkeley.

$w^{<1>}$ $w^{<2>}$ ...... $w^{<t>}$ ...... $w^{<7>}$

Vocabulary

a	0
...	
Berkeley	1023
...	
is	4673
...	
Jordan	5378
...	
Michael	8002
...	

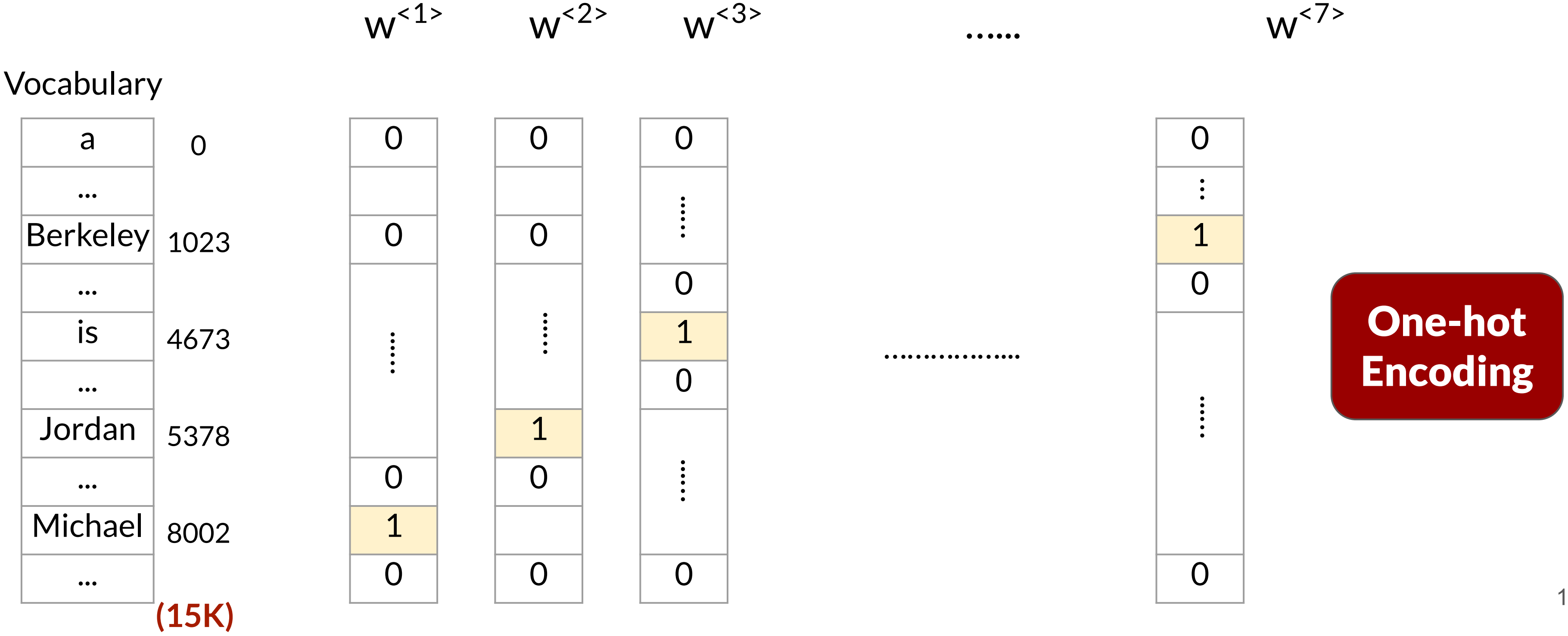
(15K)

0
0
⋮
0
1
0

(15K)

Word Representation

X: Michael Jordan is a professor at Berkeley.



Word Representation

★ What if we have a word which is **NOT** in the vocabulary?

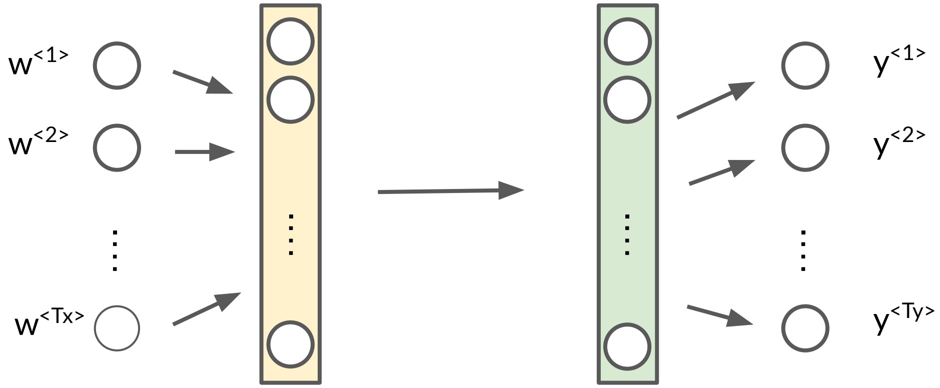
Word Representation

★ **What if we have a word which is NOT in the vocabulary?**

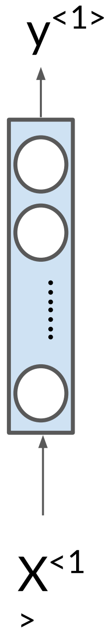
- Unknown token → <UNK>

Recurrent Neural Network

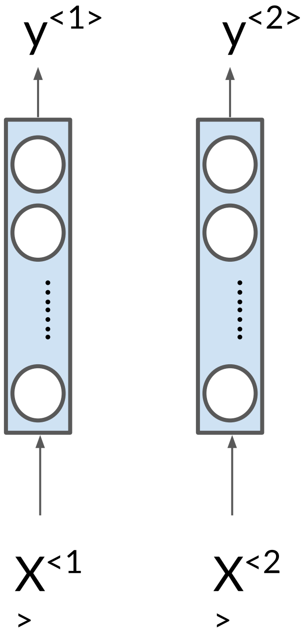
- ★ **Why not a standard network?**
  - Inputs, outputs might have different lengths
  - Doesn't share features across different positions of text.



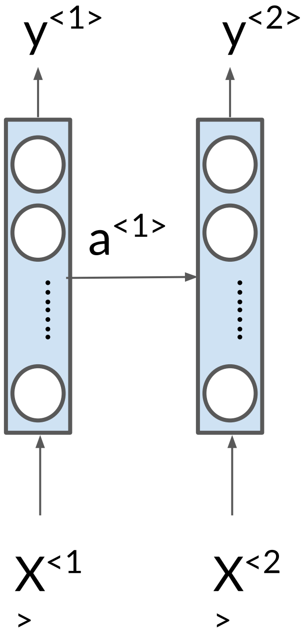
Recurrent Neural Network



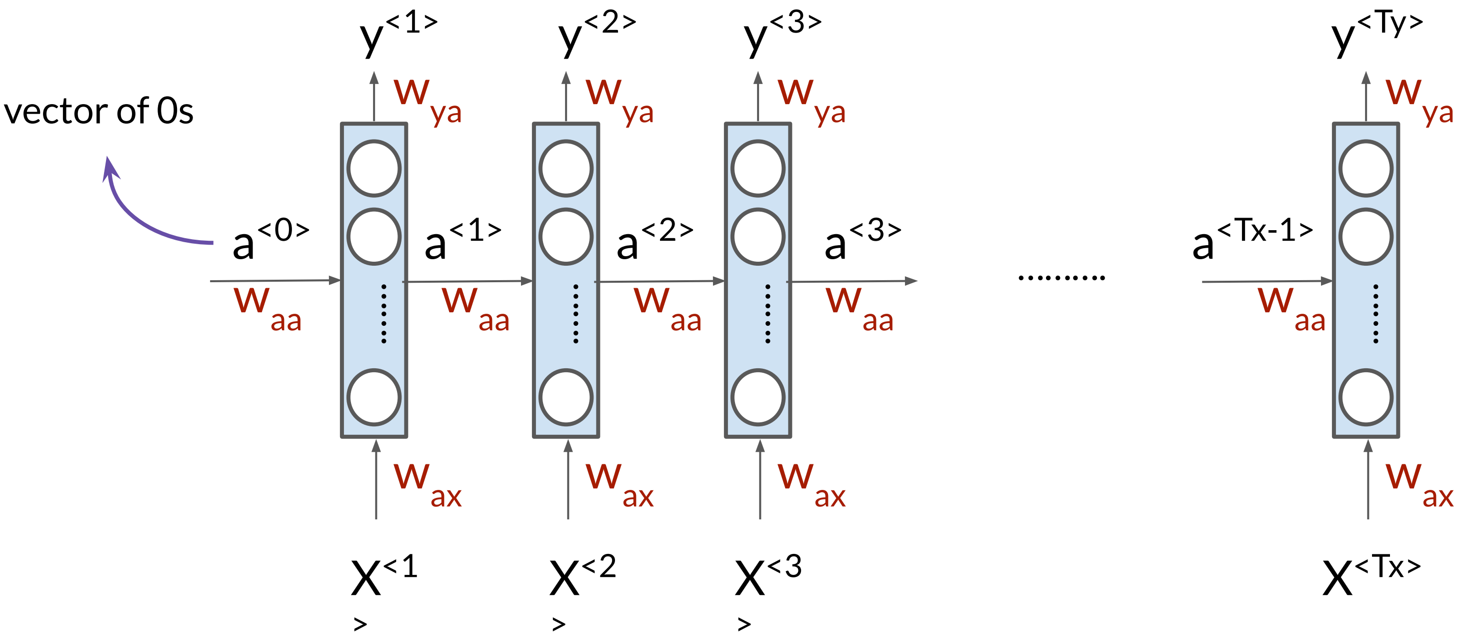
Recurrent Neural Network



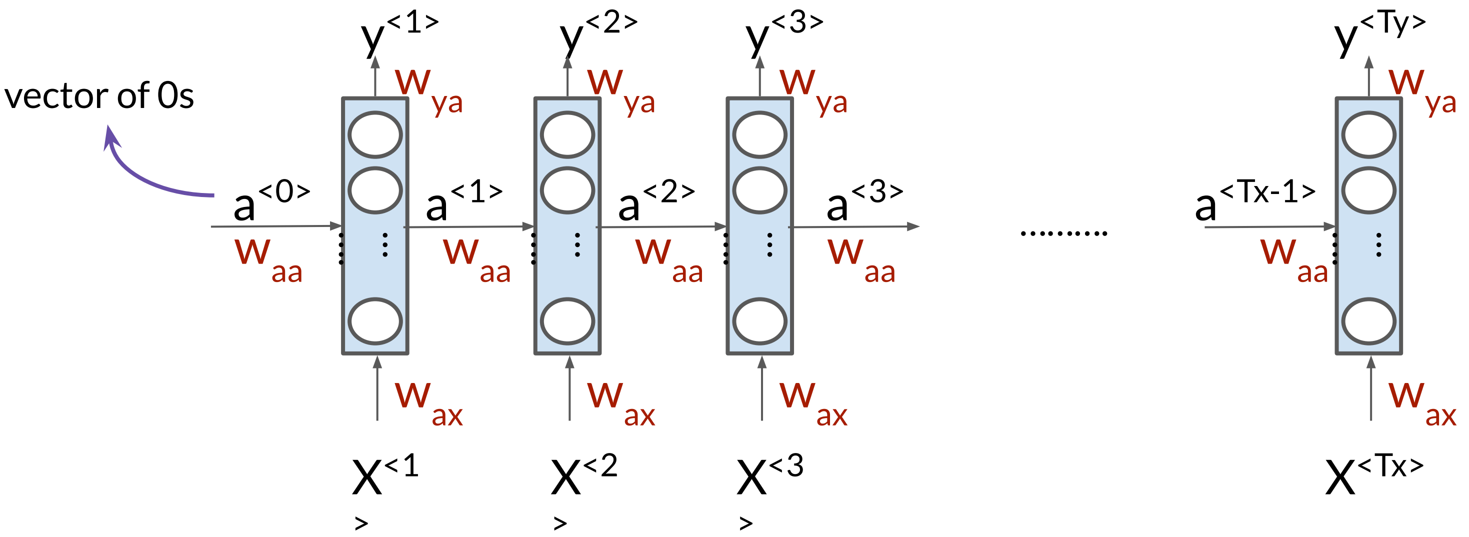
Recurrent Neural Network



Recurrent Neural Network



Recurrent Neural Network



$$a^{<t>} = g_1(W_{aa}a^{<t-1>} + W_{ax}X^{<t>} + b_a)$$

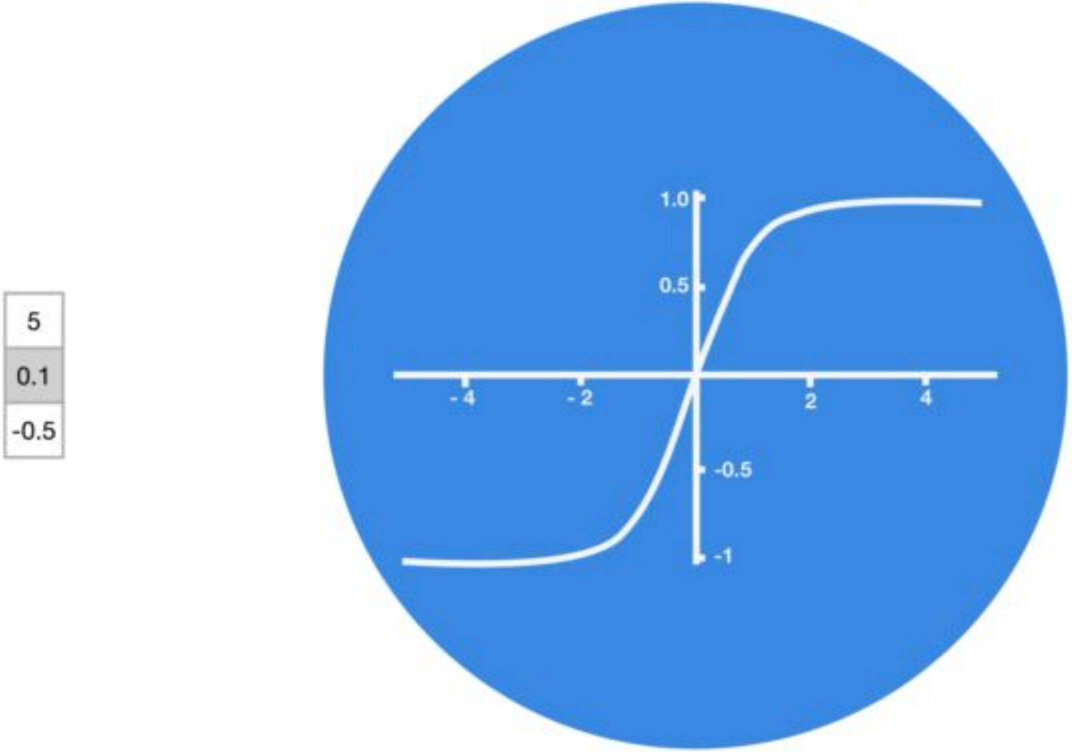
$$y^{<t>} = g_2(W_{ya}a^{<t>} + b_y)$$

?

$$g_1 = tanh / ReLU$$

$$g_2 = sigmoid$$

Tanh Activation



Source: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

Why Tanh?

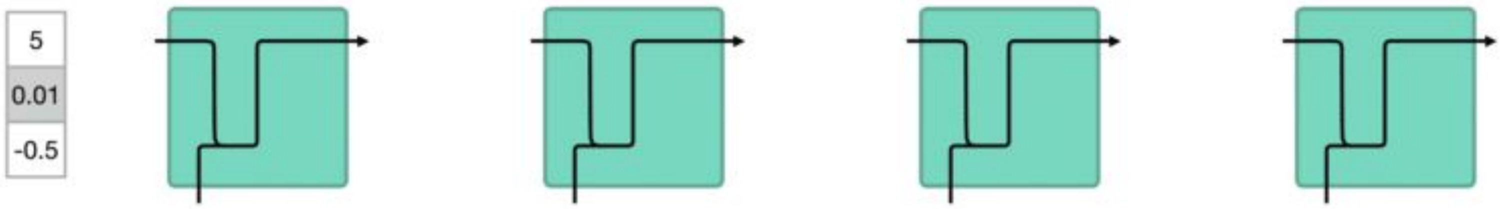


Figure 1: Vector transformation without tanh

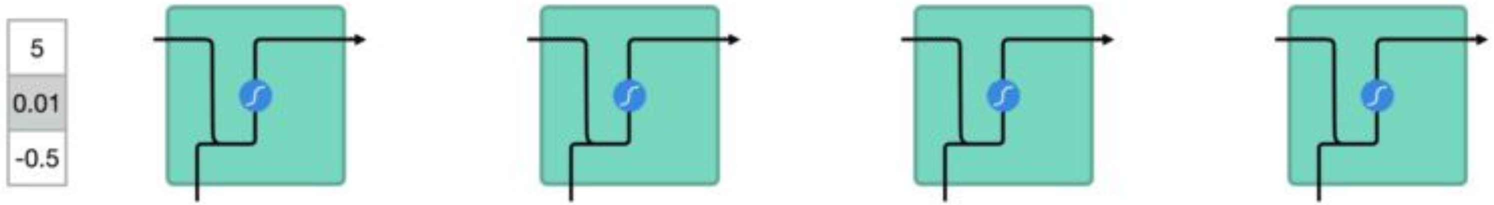


Figure 2: Vector transformation with tanh

Source: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

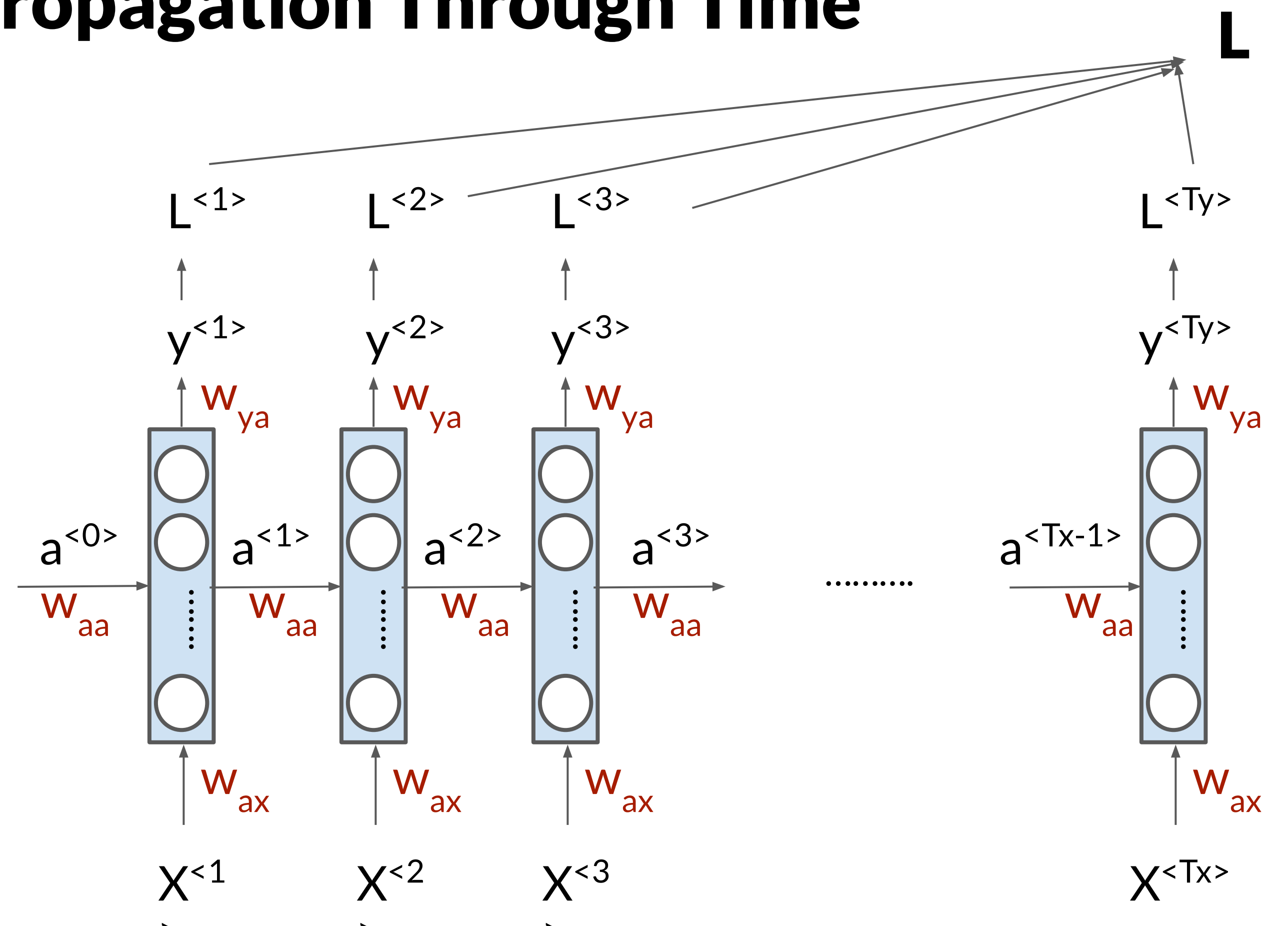
Backpropagation Through Time

$$L^{<t>}(y^{<t>} - y_{true}^{<t>}) = -y_{true}^{<t>} logy^{<t>} - (1 - y_{true}^{<t>} log(1 - y^{<t>}))$$

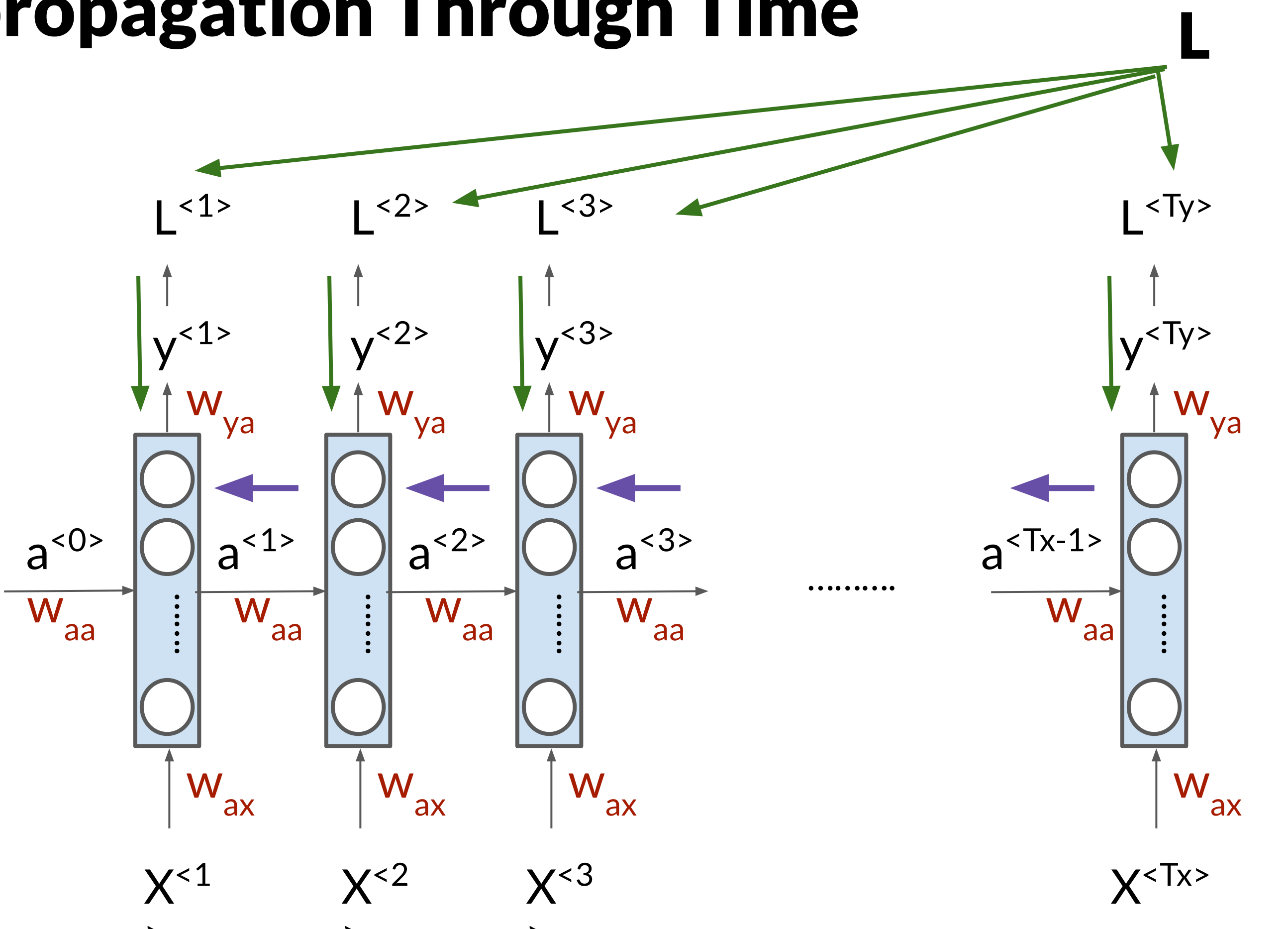
Total loss

$$\boxed{L(y, y_{true}) = \Sigma_{t=1}^{Ty} L^{<t>}(y^{<t>}, y_{true}^{<t>})}$$

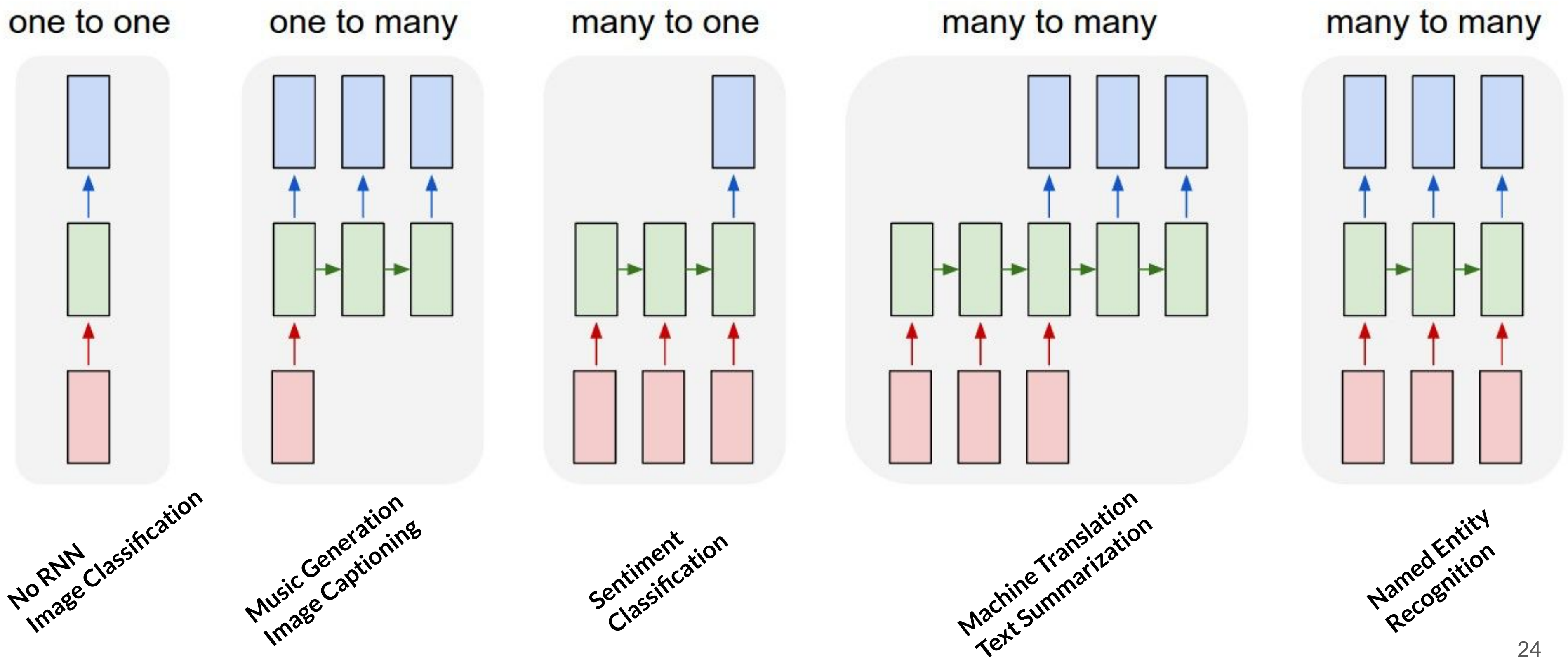
Backpropagation Through Time



Backpropagation Through Time



Different Types of RNN



Source: http://karpathy.github.io/2015/05/21/rnn-effectiveness/

Pytorch Tutorial:

Sentiment Analysis Using RNN

Disadvantages of RNN

- **Gradient Exploding**
  - When W_{aa} getting large
  - Solution:
    - Gradient Clipping

- **Gradient Vanishing**
  - When W_{aa} getting small
  - Solution:
    - Long Short-Term Memory (LSTM)

LSTM vs RNN

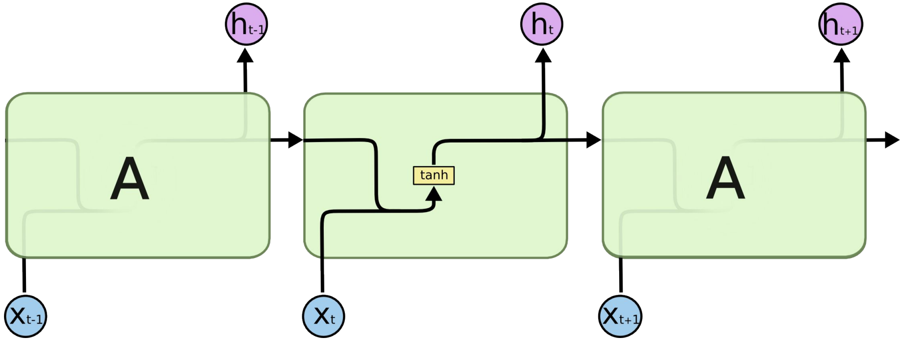


Figure 3: Standard RNN

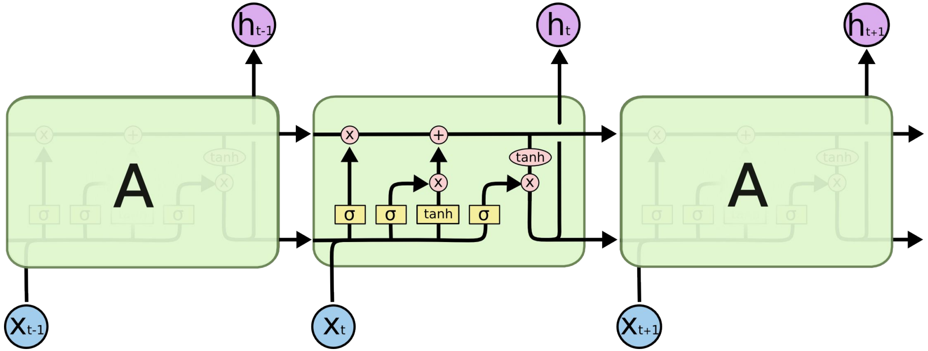
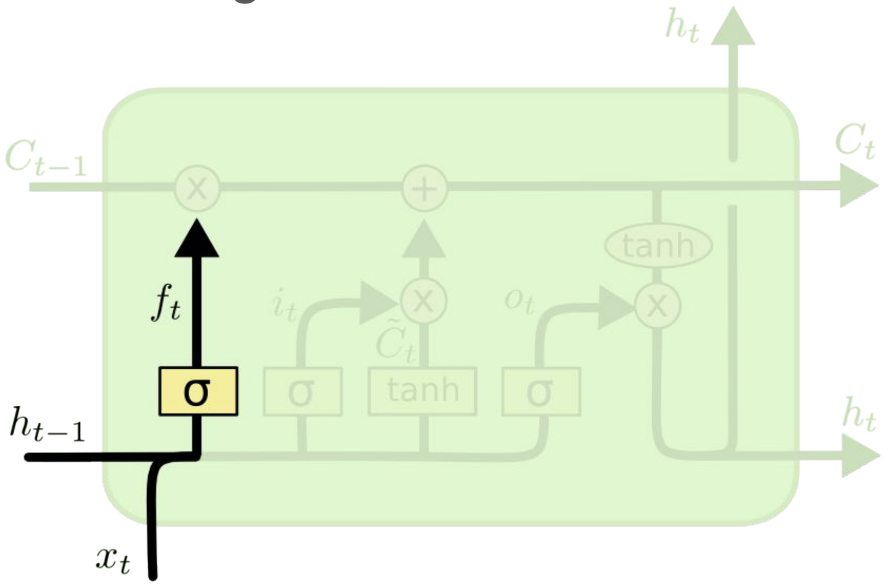


Figure 4: LSTM

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

LSTM

★ **Forget Gate**



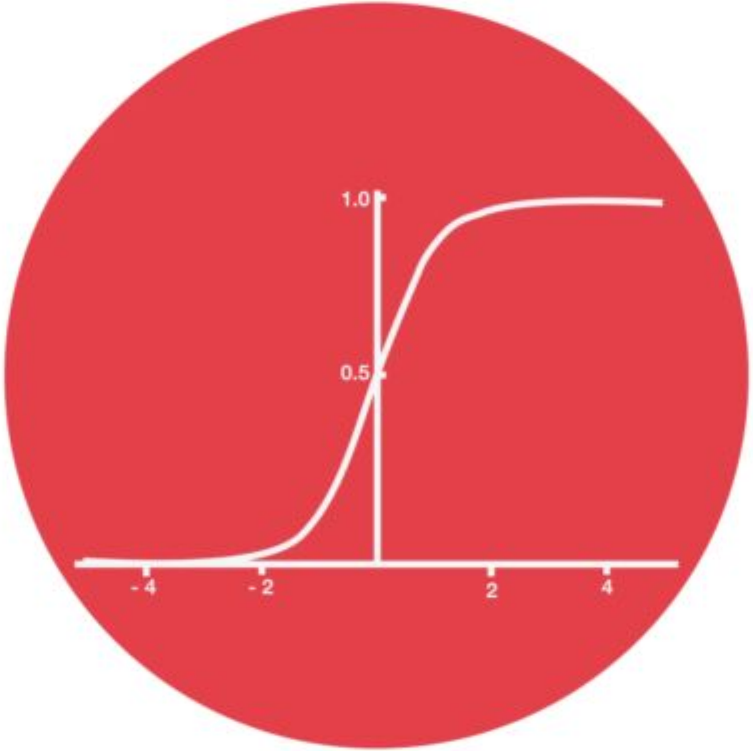
?

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

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

Sigmoid Activation

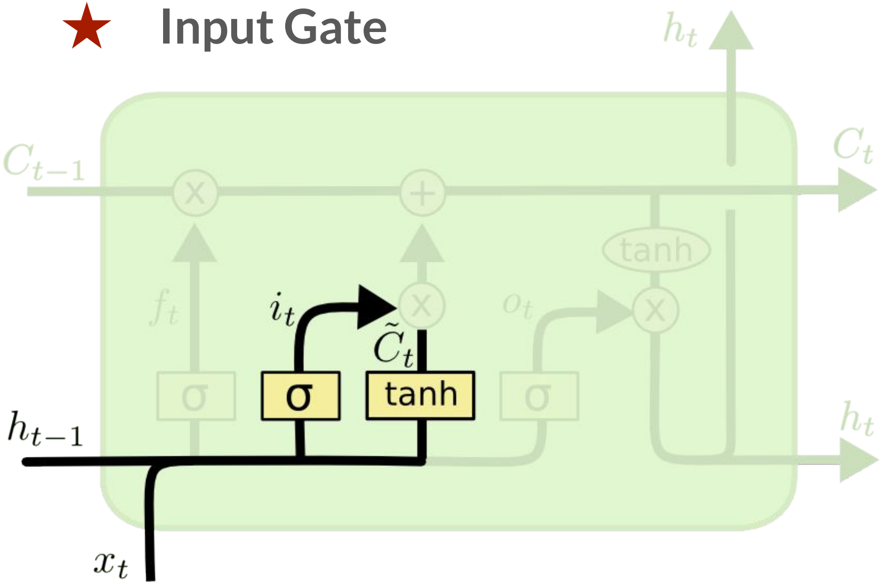
5
0.1
-0.5



Source: https://towardsdatascience.com/illustrated-guide-to-lstms-and-gru-s-a-step-by-step-explanation-44e9eb85bf21

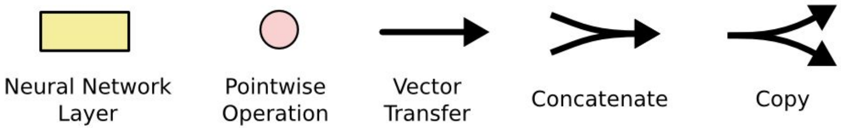
LSTM

★ **Input Gate**



$$i_t = \sigma\left(W_i \cdot [h_{t-1}, x_t] \ + \ b_i\right)$$

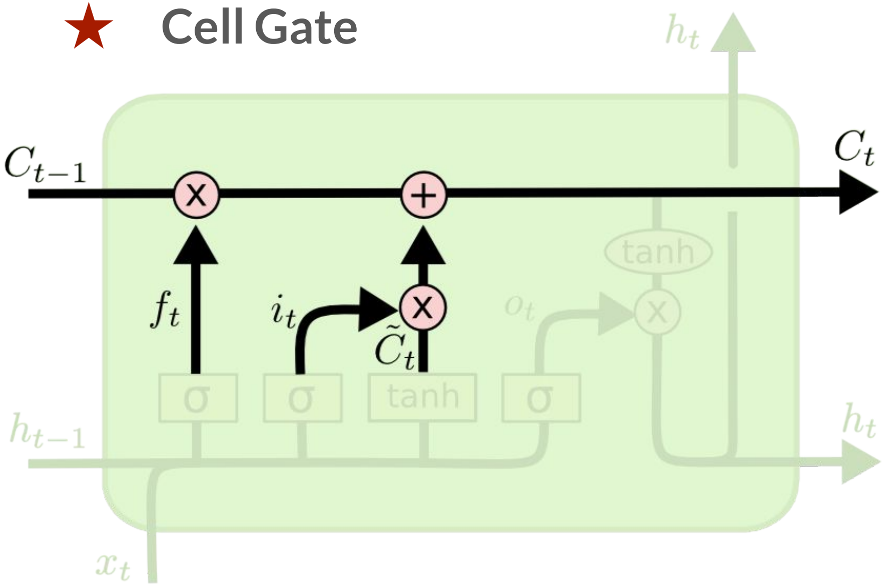
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] \ + \ b_C)$$



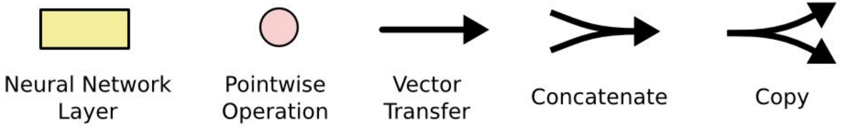
Source: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

LSTM

★ **Cell Gate**



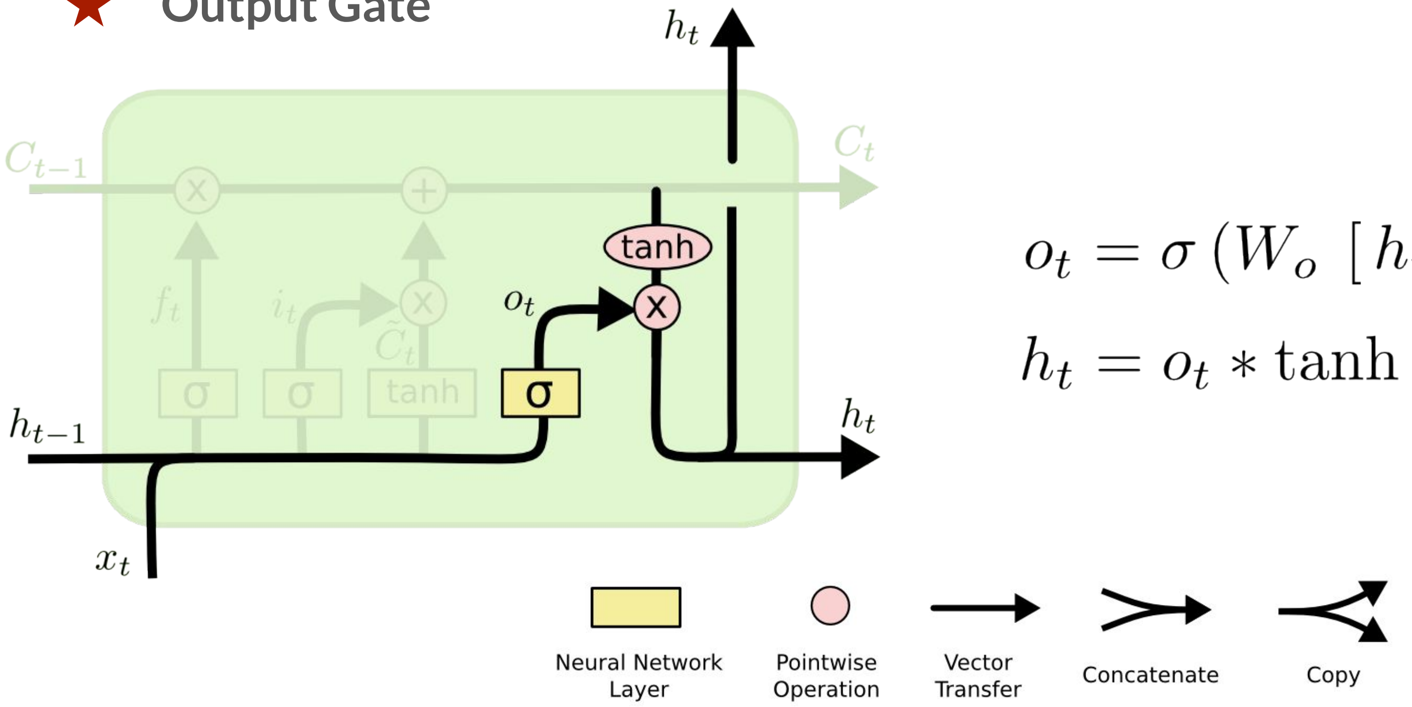
$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t$$



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

LSTM

★ **Output Gate**

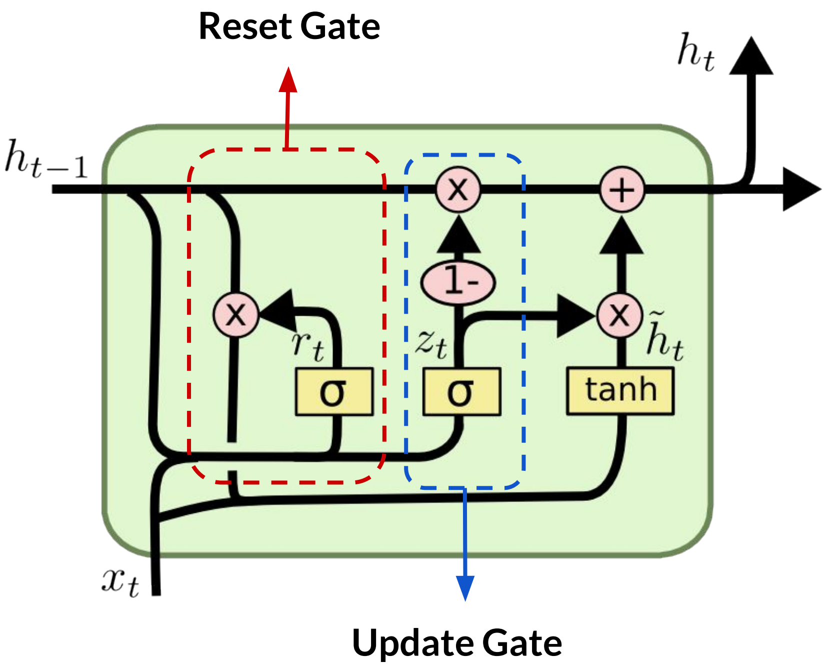


$$o_t = \sigma\left(W_o\ [h_{t-1}, x_t]\ +\ b_o\right)$$

$$h_t = o_t * \tanh\left(C_t\right)$$

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

GRU



$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$

$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$

$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

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

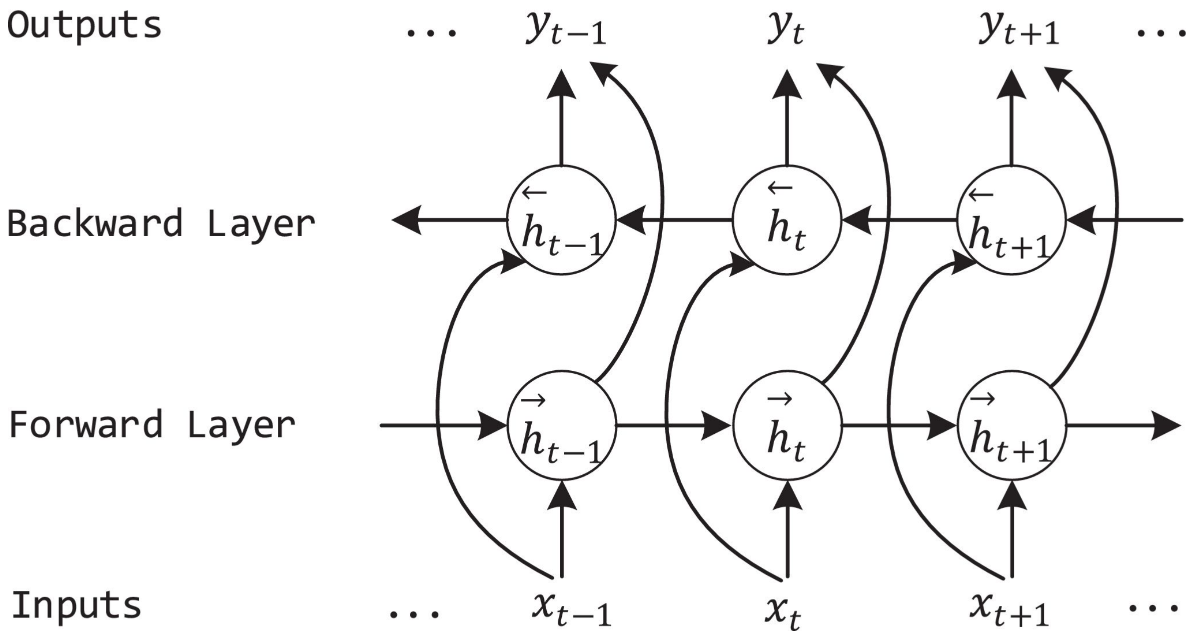
Bidirectional RNN

★ **Problem of unidirectional network:**

- He said, “Teddy Roosevelt was a great president”
- He said, “Teddy bears are on sale!”

Is “Teddy” a PERSON’s name?

Bidirectional RNN



$$h_i = [\overrightarrow{h_i}; \overleftarrow{h_i}]$$