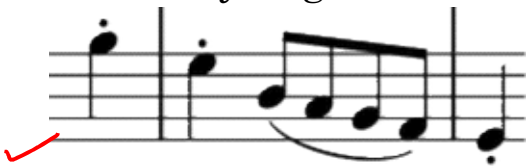
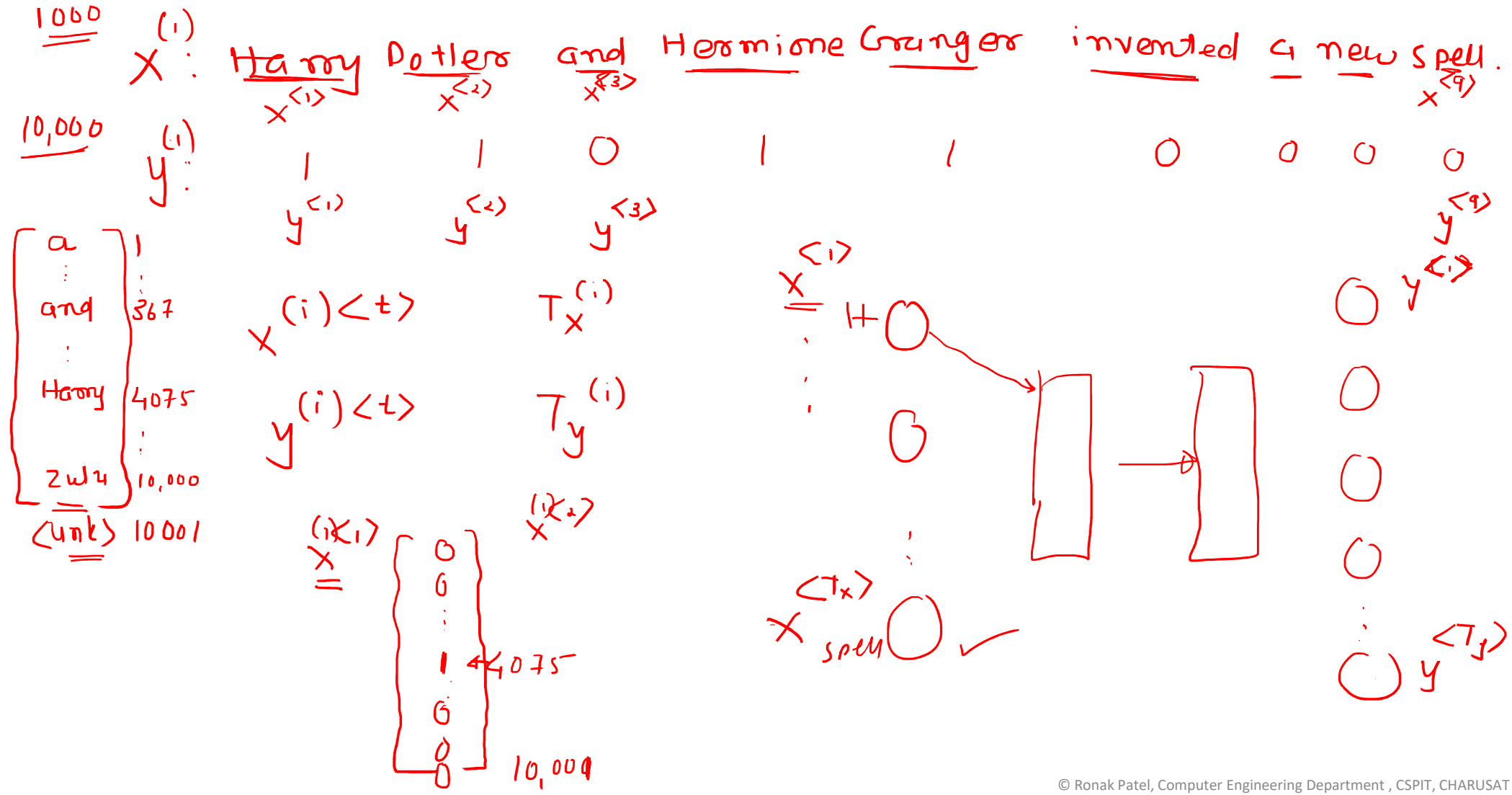
Recurrent Neural Network

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

NLP

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

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

$T_x = 9$

y: 1 1 0 1 1 0 0 0 0

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

$T_y = 9$

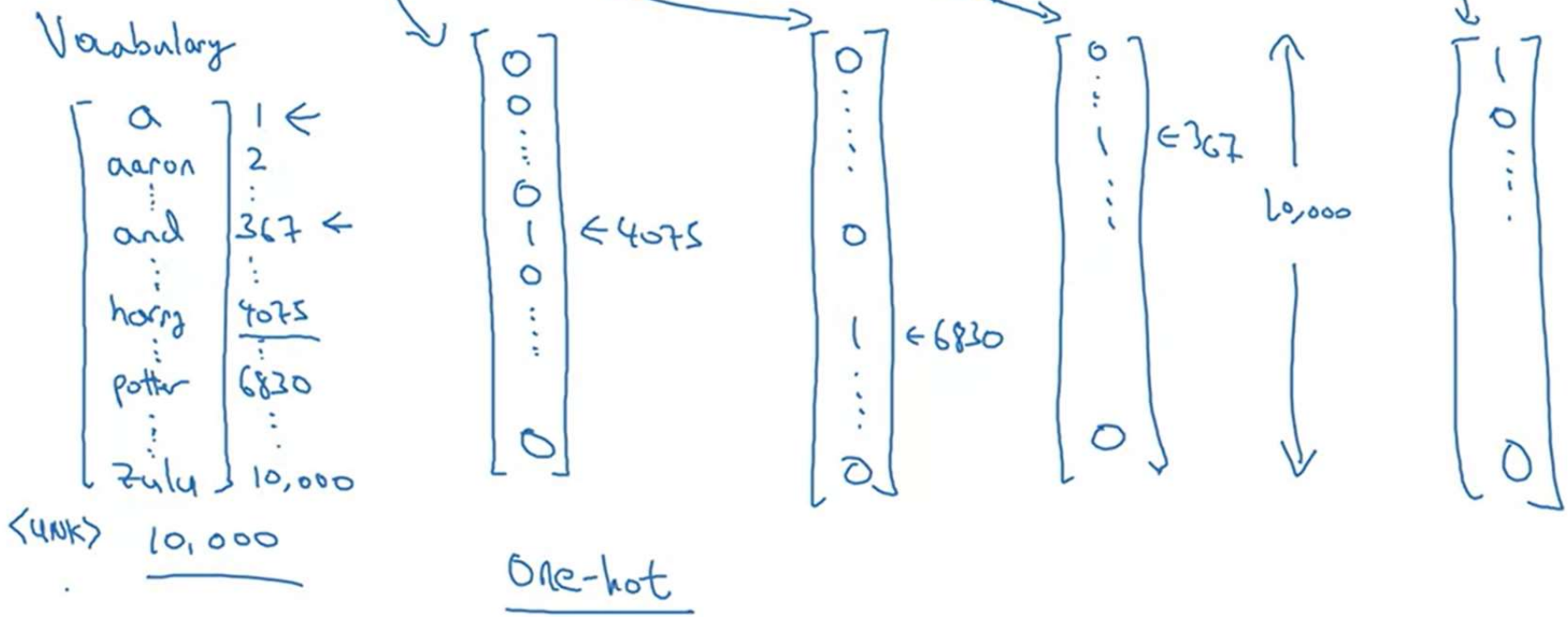
$X^{(i)<t>}$ $T_x^{(i)} = 9$ 15

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

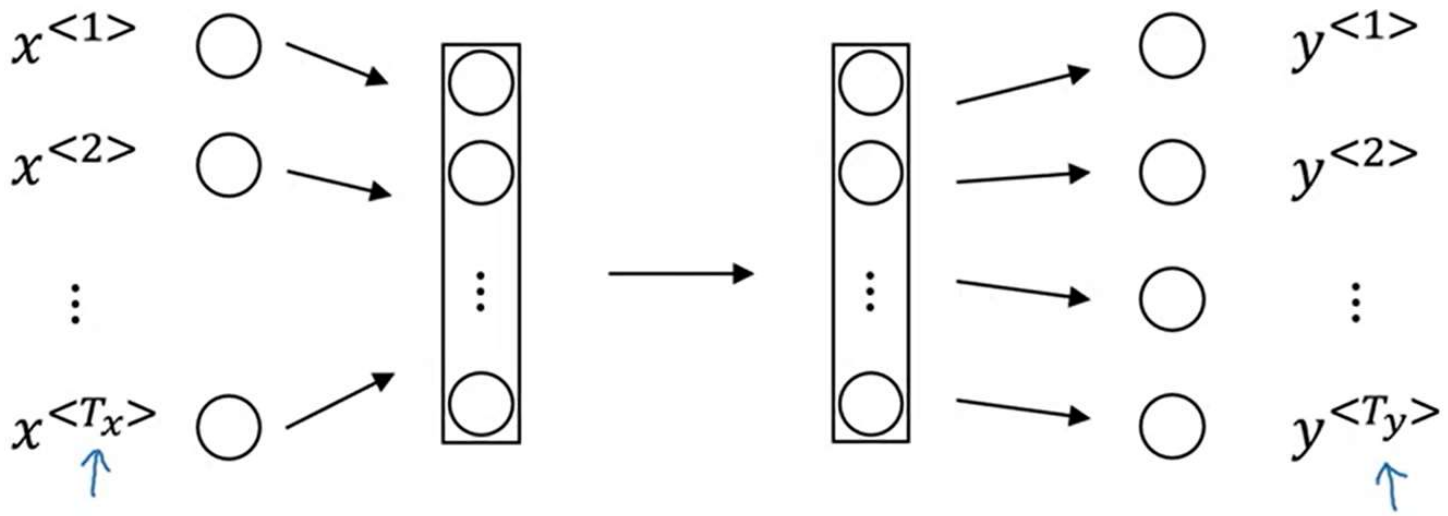
Representing words

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

$x^{<1>}$ $x^{<2>}$ $x^{<3>}$ ... $x^{<7>}$ $x^{<9>}$



Why not a standard network?



Problems:

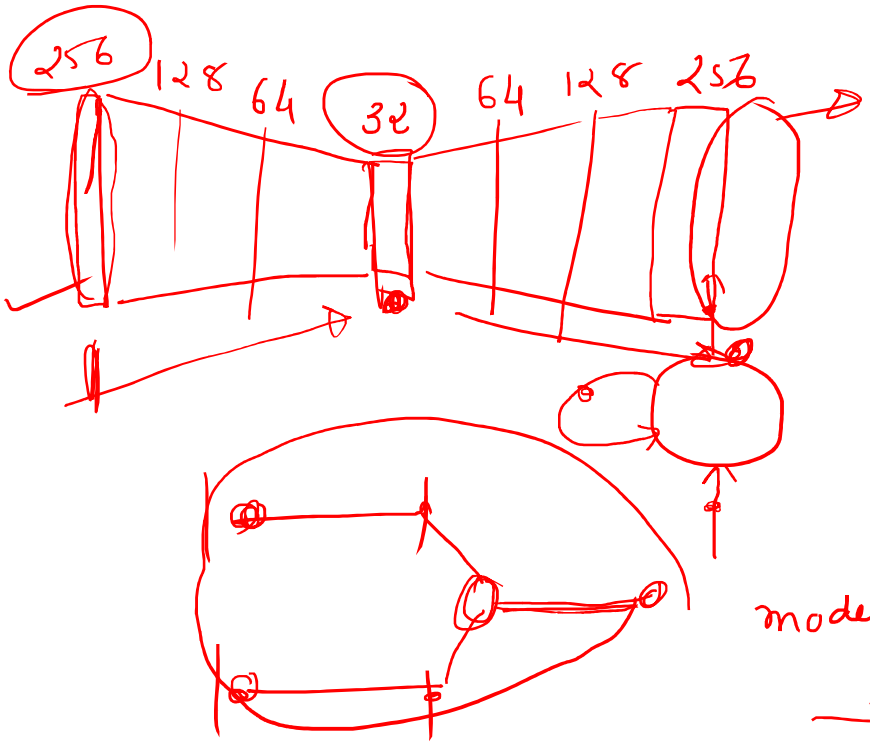
- Inputs, outputs can be different lengths in different examples.
- Doesn't share features learned across different positions of text.

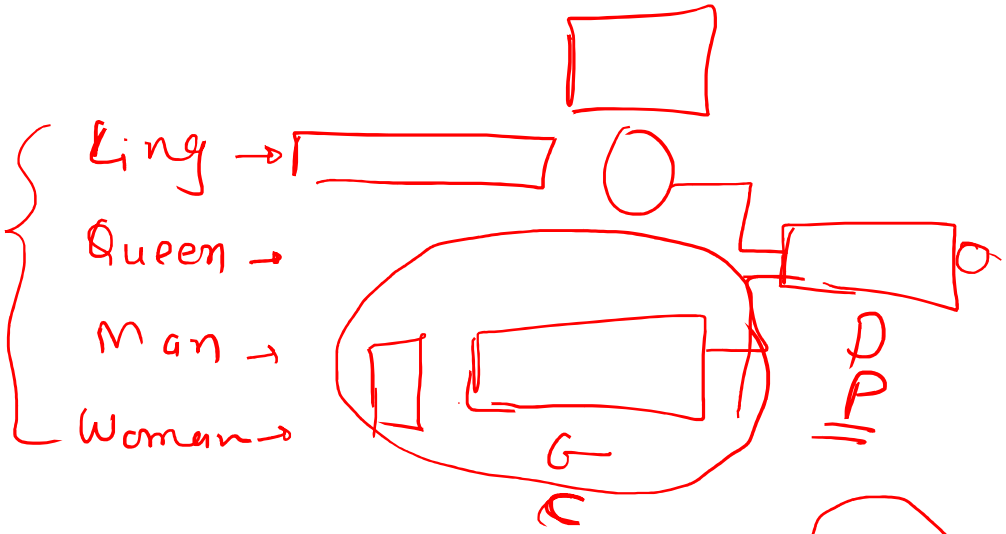
Andrew Ng

Why not standard network?

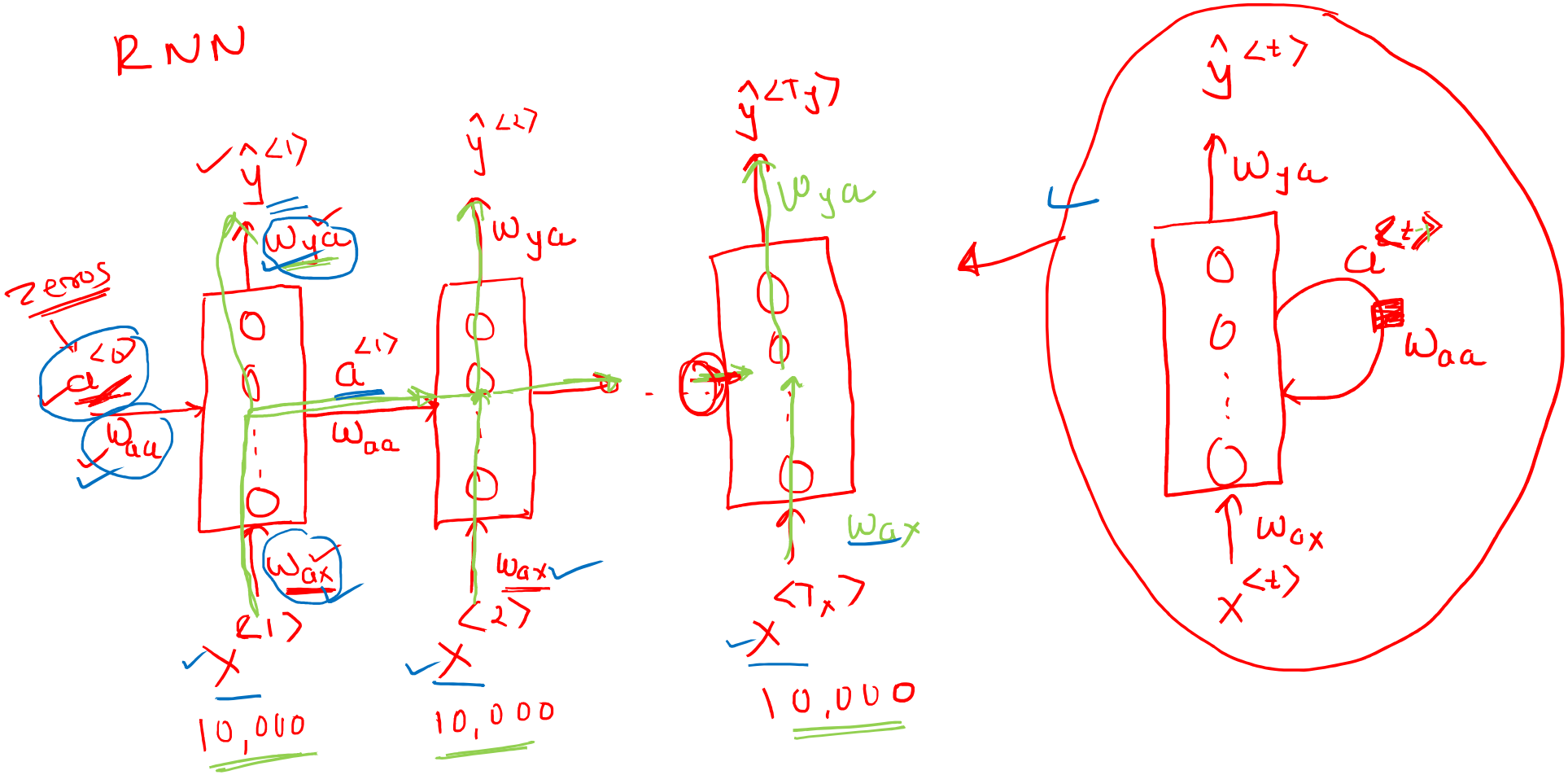
Image generator

- Inputs, outputs can be different length in different example.
- Doesn't share features learned across different position of text.

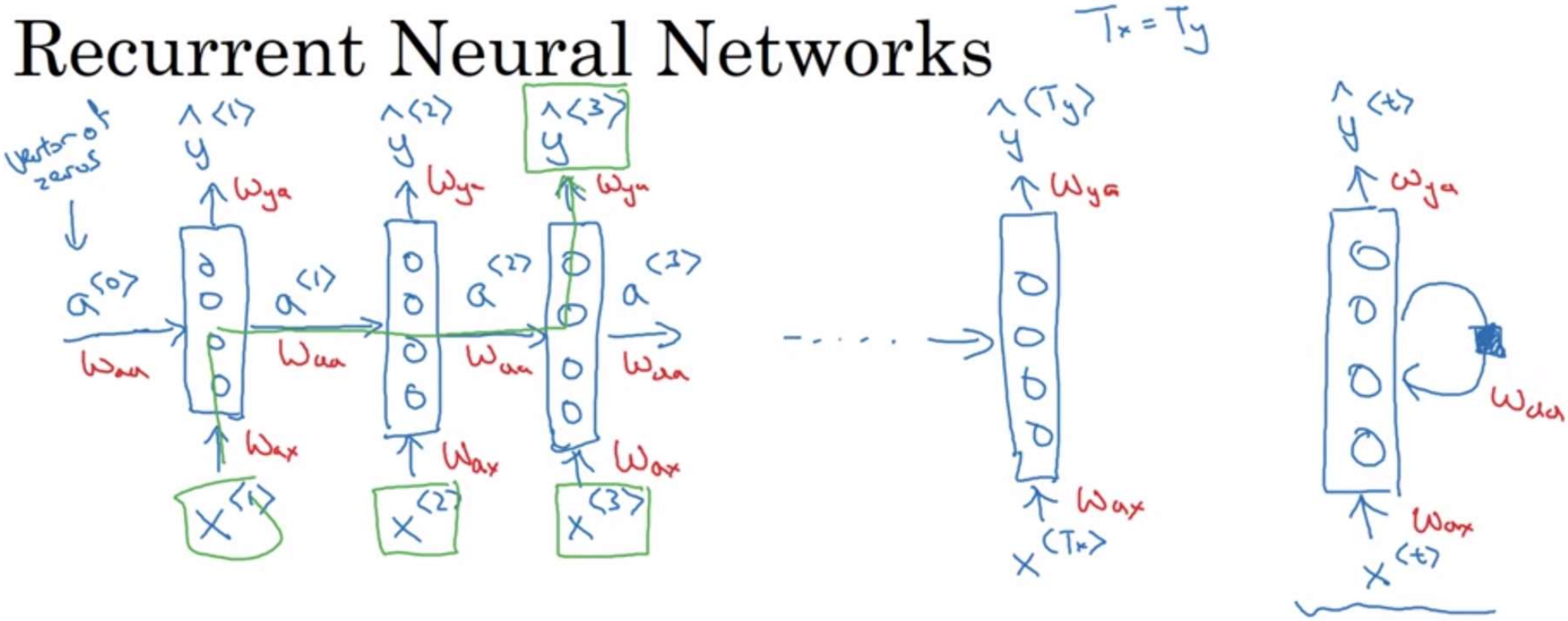




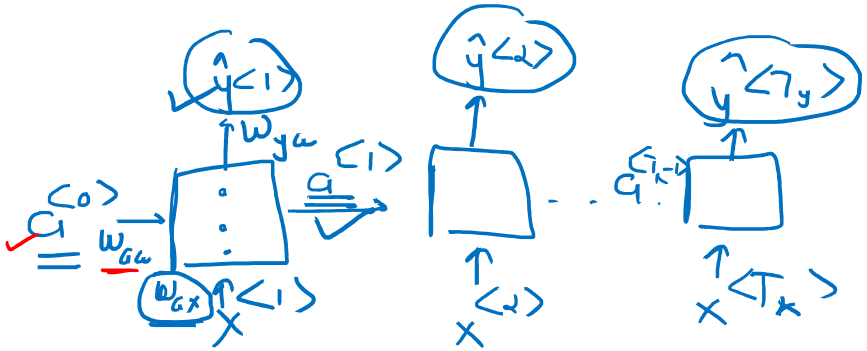
model fit (X, Y, - - - King - Man + Woman = Queen)



Recurrent Neural Networks



RNN: Forward Propagation



$$W_a \begin{bmatrix} W_{aa} & W_{ax} \end{bmatrix} \begin{bmatrix} a^{\langle t-1\rangle} \\ x^{\langle t\rangle} \end{bmatrix}$$

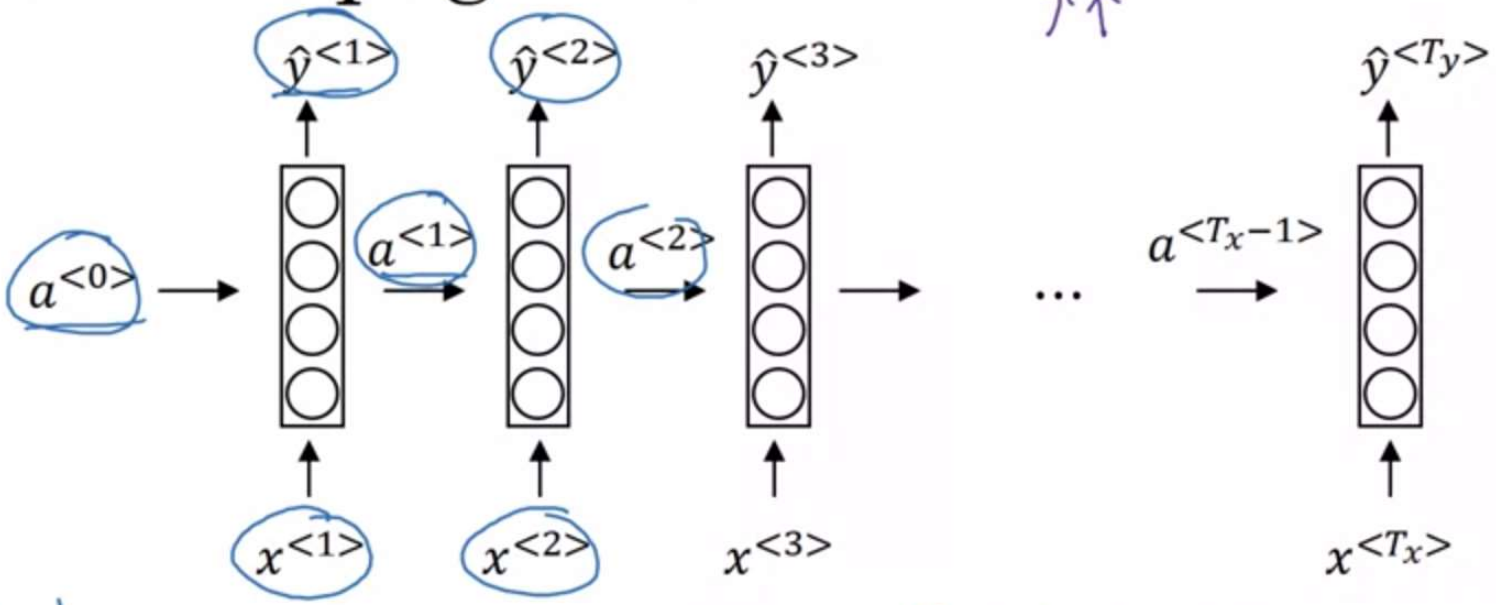
$$a^{\langle 1\rangle} = g_1(W_{aa}\underbrace{a^{\langle 0\rangle}}_{0} + W_{ax}x^{\langle 1\rangle} + b_a) \quad \text{tanh / Relu}$$

$$\hat{y}^{\langle 1\rangle} = g_2(W_{ya}a^{\langle 1\rangle} + b_y) \quad \text{Sigmoid}$$

$$\begin{cases} a^{\langle t\rangle} = g_1(W_{aa}a^{\langle t-1\rangle} + W_{ax}x^{\langle t\rangle} + b_a) \\ \quad = g_1(W_a[a^{\langle t-1\rangle}, x^{\langle t\rangle}] + b_a) \\ \hat{y}^{\langle t\rangle} = g_2(W_{ya}a^{\langle t\rangle} + b_y) \\ \quad = g_2(W_y a^{\langle t\rangle} + b_y) \end{cases}$$

Forward Propagation

$a \leftarrow W_{ax} x^{\langle 1 \rangle}$



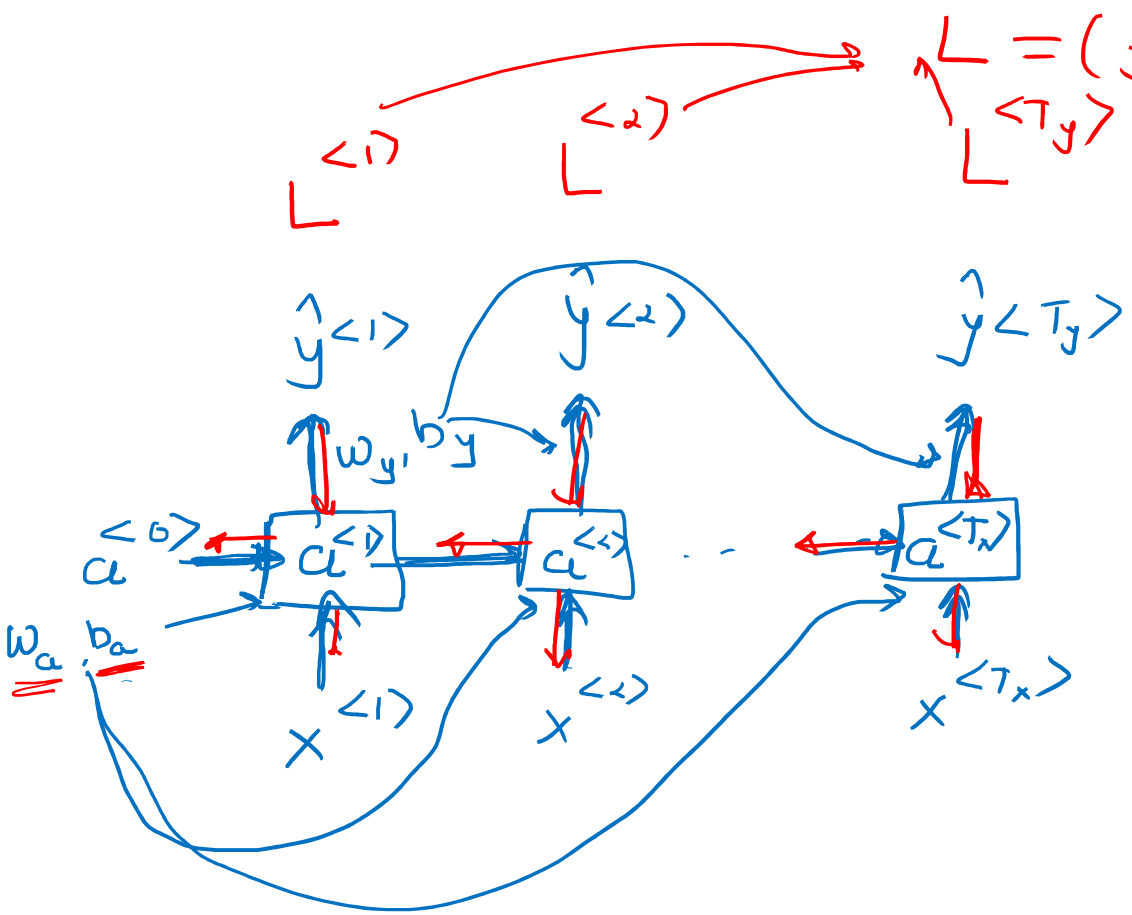
$a^{\langle 0 \rangle} = \vec{0}$.

$a^{\langle 1 \rangle} = g_1(W_{aa} a^{\langle 0 \rangle} + W_{ax} x^{\langle 1 \rangle} + b_a)$ ← tanh / ReLU

$\hat{y}^{\langle 1 \rangle} = g_2(W_{ya} a^{\langle 1 \rangle} + b_y)$ ← sigmoid

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

Andrew Ng



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

Back propagation through time

Simplified RNN notation

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

(100,100) 100 (100,10,000) 10,000

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

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

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

$$\begin{bmatrix} W_{aa} ; W_{ax} \end{bmatrix} = W_a \quad (100, 10100)$$

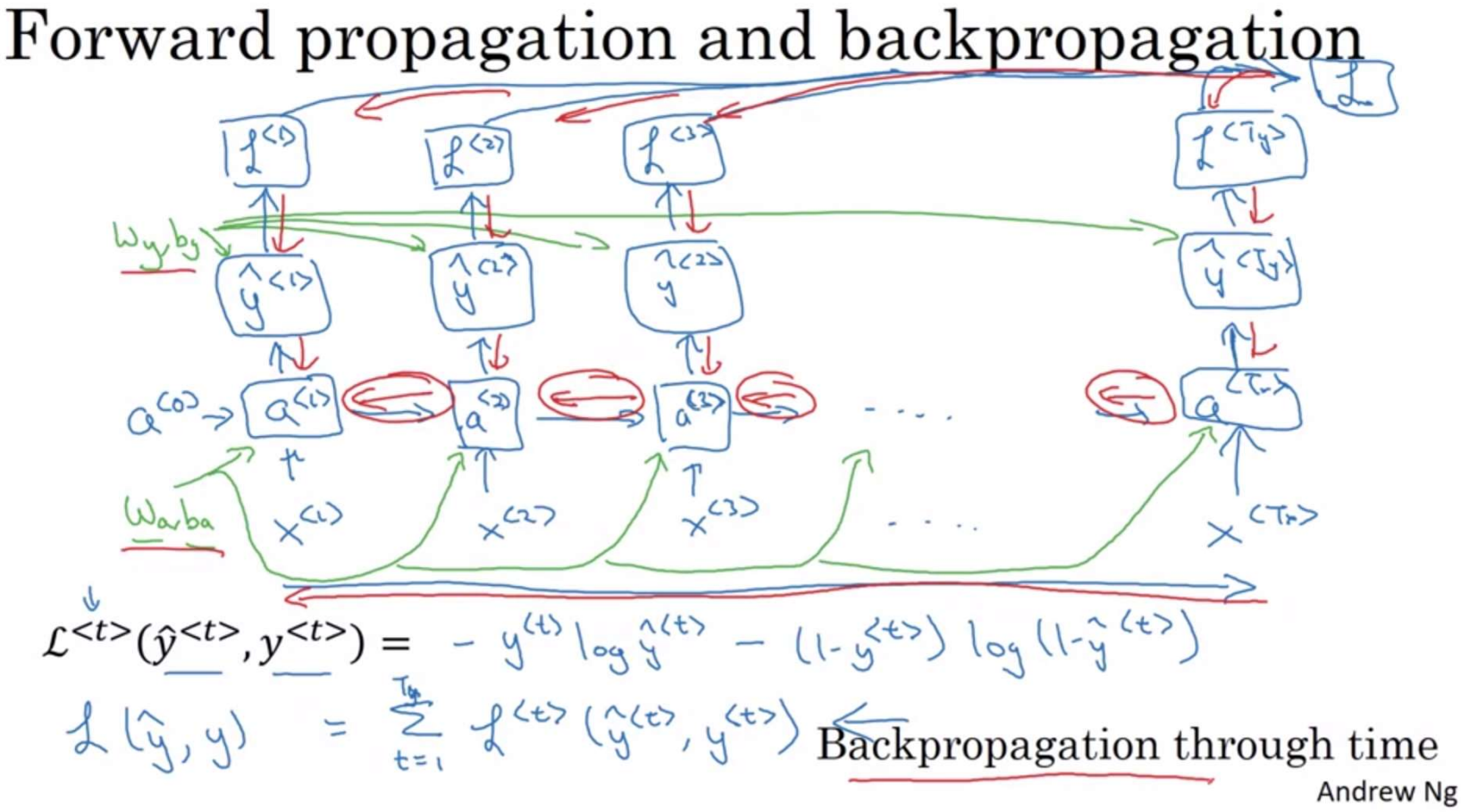
100 × 100, 100 × 10000

$$[a^{<t-1>}, x^{<t>}] = \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} \quad 100 + 10000 = 10100$$

$$[W_{aa} ; W_{ax}] \begin{bmatrix} a^{<t-1>} \\ x^{<t>} \end{bmatrix} = W_{aa}a^{<t-1>} + W_{ax}x^{<t>}$$

Andrew Ng

Backpropagation through time



Different RNN Architectures

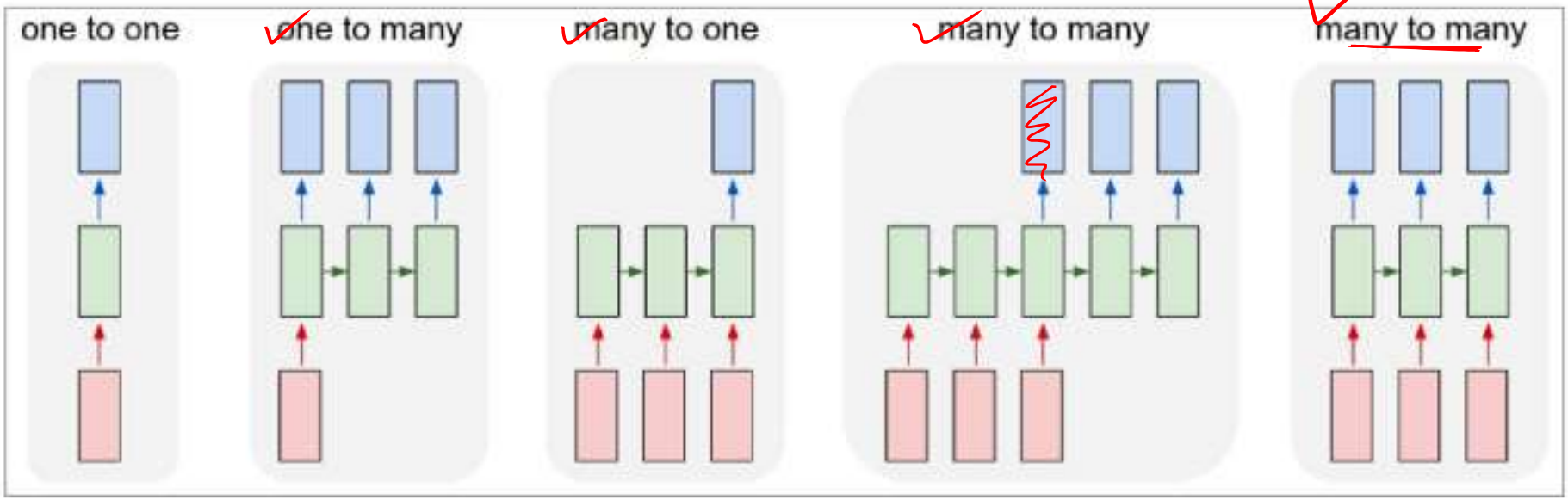


image classification

image captioning,
Music Generation

sentiment classification

Machine translation
$T_x \neq T_y$

Name entity recognition
$T_x = T_y$

Advantages of Recurrent Neural Network

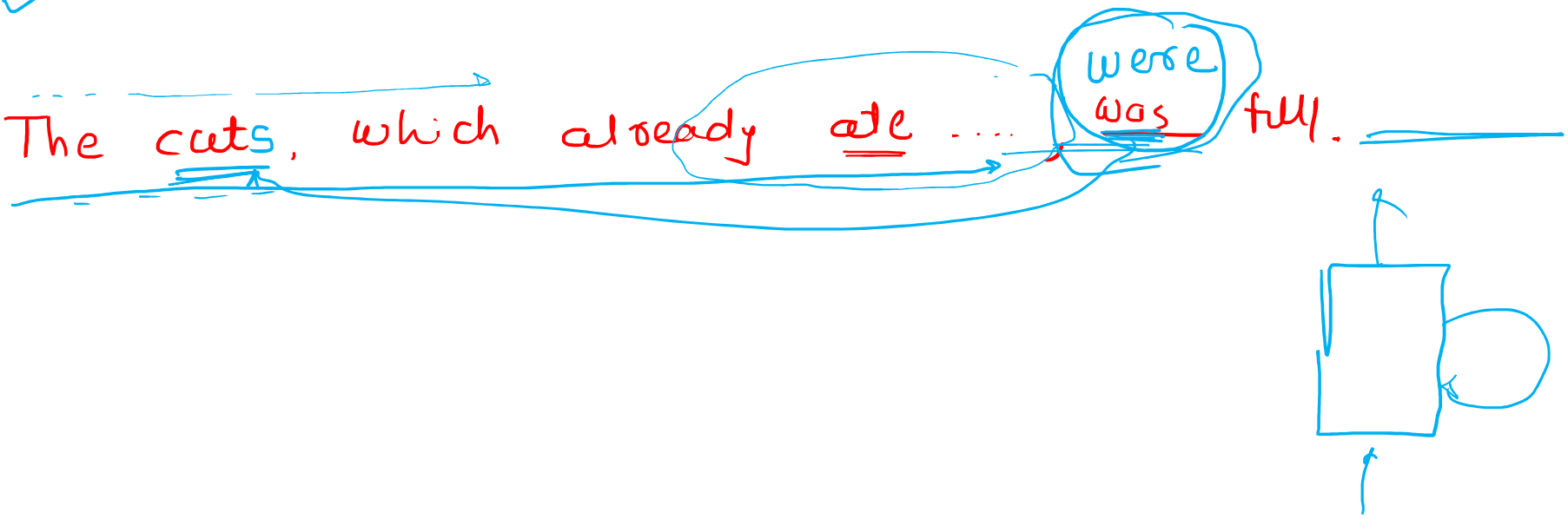
- It is useful in time series prediction only because of the feature to remember previous inputs as well.

Disadvantages of Recurrent Neural Network

- Vanishing gradients or short term memory in RNN
- Exploding gradients.
- Training an RNN is a very difficult task.

GRU, LSTM

bi directional



Vanishing gradients

- As our RNN processes more steps, it has trouble retaining information from the previous steps. So, in some cases, this might not be a problem, where the word just depends on its previous neighboring word.

RP3

- For eg. the sentence: I can speak French very well.

RP2

- But consider this sentence: I am going to France, the language spoken there is _____.

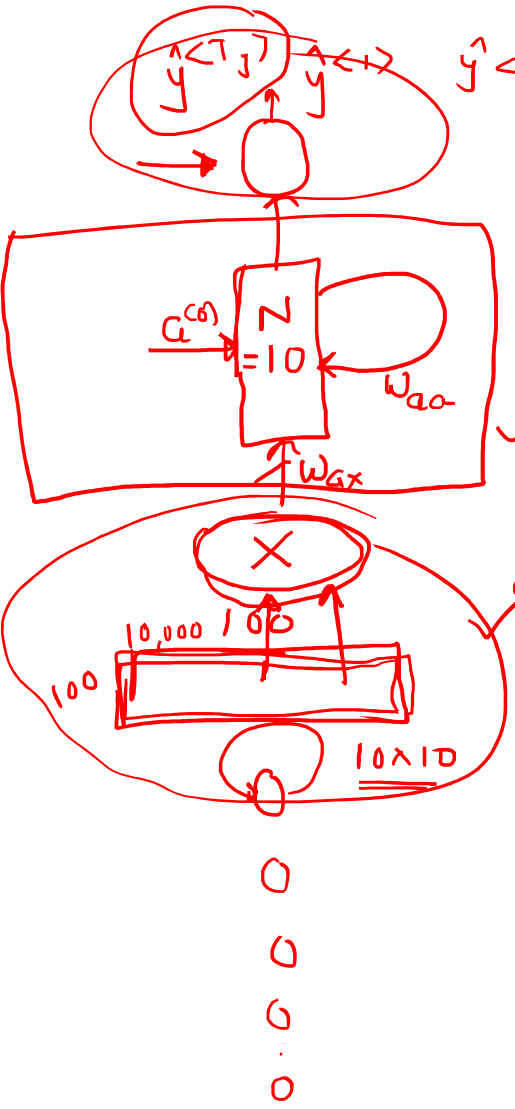
Slide 20

RP2 Now the answer “French” here has a dependency on the word France, which is far away from it. This type of dependency is known as long term dependency, and the normal structure of RNN fails to operate over these.
RONAKKUMAR PATEL, 28-Sep-20

RP3 Now the word ‘well’ in this sentence is very intuitive to come at this place and RNN can handle such sentences effectively.
RONAKKUMAR PATEL, 28-Sep-20

- Short-Term Memory and the vanishing gradient is due to the nature of back-propagation; an algorithm used to train and optimize neural networks.

- When doing back propagation, each node in a layer calculates it's gradient with respect to the effects of the gradients, in the layer before it. So if the adjustments to the layers before it is small, then adjustments to the current layer will be even smaller.



Dense-1 $\quad W_{ay} \quad (b_y)$

$10 + 1 \quad = \quad 11$

RNN $\Rightarrow \quad W_{ax} \quad W_{aa} \quad (b_a)$

$1000 + 100 + 10 \quad = \quad 1110$

embd. Vocab size $10{,}000 \times 100 = 1000000$

the $\quad$ which

100