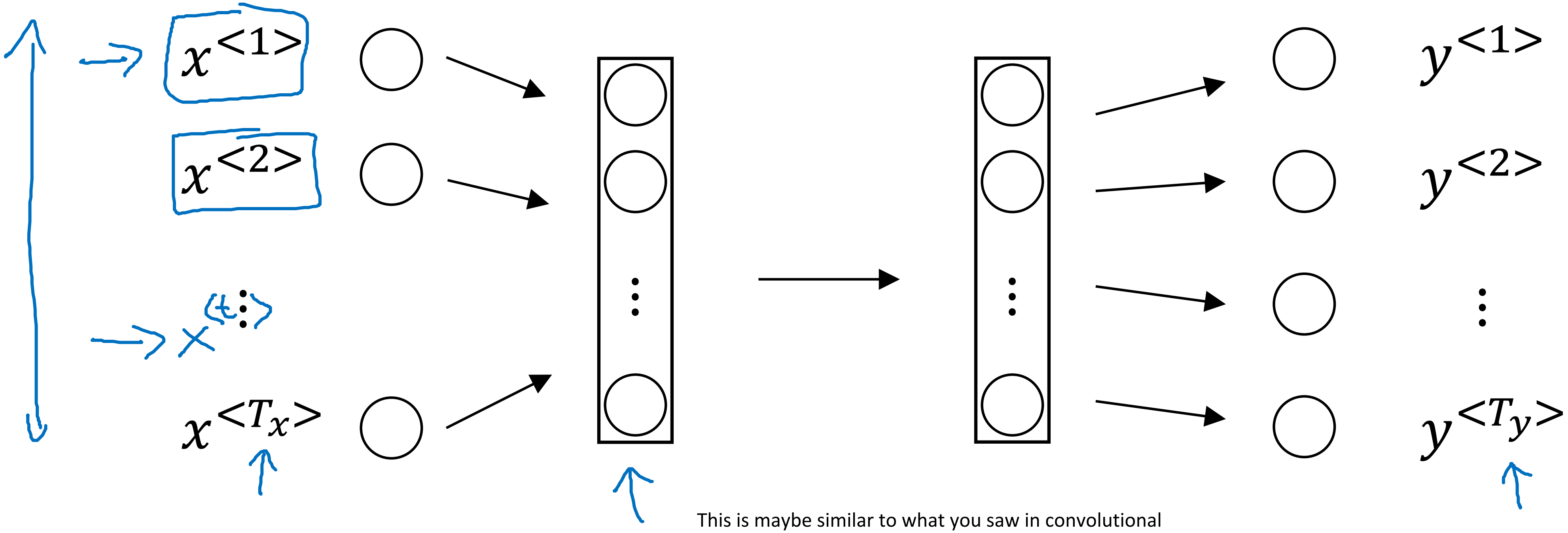
Recurrent Neural Networks

deeplearning.ai

Recurrent Neural Network Model

Why not a standard network?



This is maybe similar to what you saw in convolutional neural network where you want things learned for one part of the image to generalize quickly to the other parts of the image, and we'd like similar effect for sequence data as well. And similar to what you saw with convnets, using a better representation will also let you reduce the number of parameters in your model.

Problems:

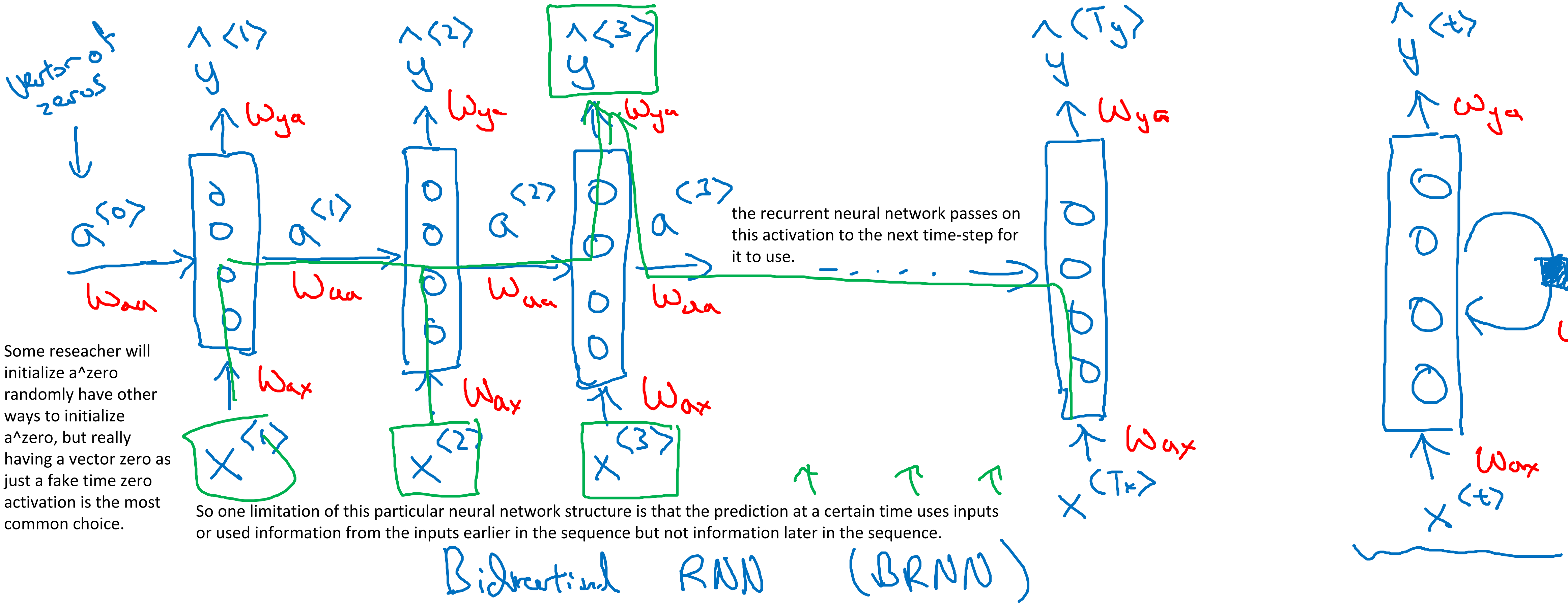
pad / zero pad

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

Andrew Ng

Recurrent Neural Networks

The architecture will change a little bit if T_x and T_y are not identical.



the recurrent neural network passes on this activation to the next time-step for it to use.

Some reseacher will initialize a^zero randomly have other ways to initialize a^zero, but really having a vector zero as just a fake time zero activation is the most common choice.

So one limitation of this particular neural network structure is that the prediction at a certain time uses inputs or used information from the inputs earlier in the sequence but not information later in the sequence.

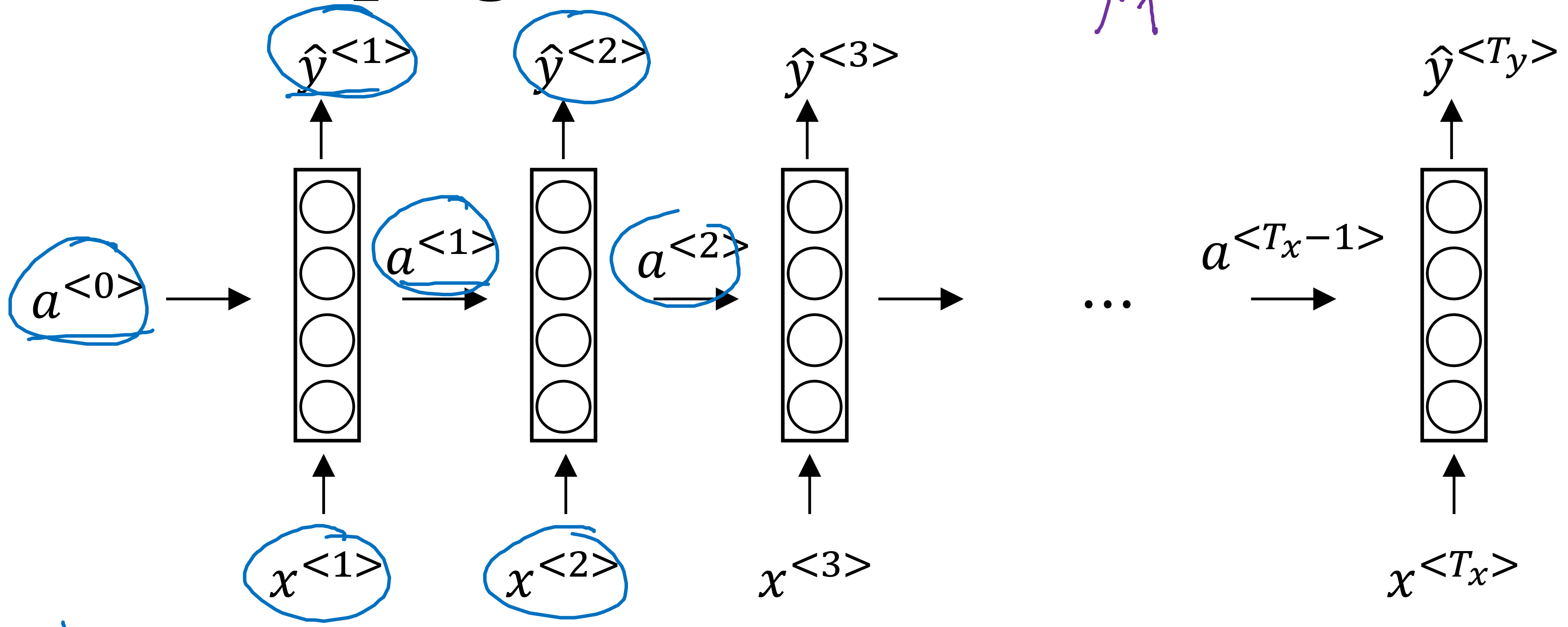
He said, "Teddy Roosevelt was a great President."

He said, "Teddy bears are on sale!"

Andrew Ng

Forward Propagation

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



$a^{<0>} = \vec{0}.$

$a^{<1>} = g_1(W_{aa} a^{<0>} + W_{ax} x^{<1>} + b_a) \leftarrow$ tanh / ReLU

$\hat{y}^{<1>} = g_2(W_{ya} a^{<1>} + b_y) \leftarrow$ Sigmoid

$$\boxed{\begin{aligned} a^{<t>} &= g(W_{aa} a^{<t-1>} + W_{ax} x^{<t>} + b_a) \\ \hat{y}^{<t>} &= g(W_{ya} a^{<t>} + b_y) \end{aligned}}$$

tanh is actually a pretty common choice
we have other ways of preventing the
vanishing gradient problem.

Andrew Ng

Simplified RNN notation

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

100 $\quad$ $10{,}000$

$(100, 100)$ $\quad$ $(100, 10{,}000)$

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

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

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

$$\begin{array}{cc} 100 & \left[\begin{array}{c;{2pt/2pt}c} W_{aa} & W_{ax} \end{array}\right] \\ & \quad \underset{100}{\longleftrightarrow} \qquad \underset{10\,000}{\longleftrightarrow} \end{array} = W_a$$

$(100, 10100)$

$$[a^{\langle t-1 \rangle}, x^{\langle t \rangle}] = \begin{bmatrix} a^{\langle t-1 \rangle} \\ \hline x^{\langle t \rangle} \end{bmatrix} \begin{array}{l} \updownarrow 100 \\ \updownarrow 10000 \end{array} \quad \updownarrow 10100$$

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

Andrew Ng