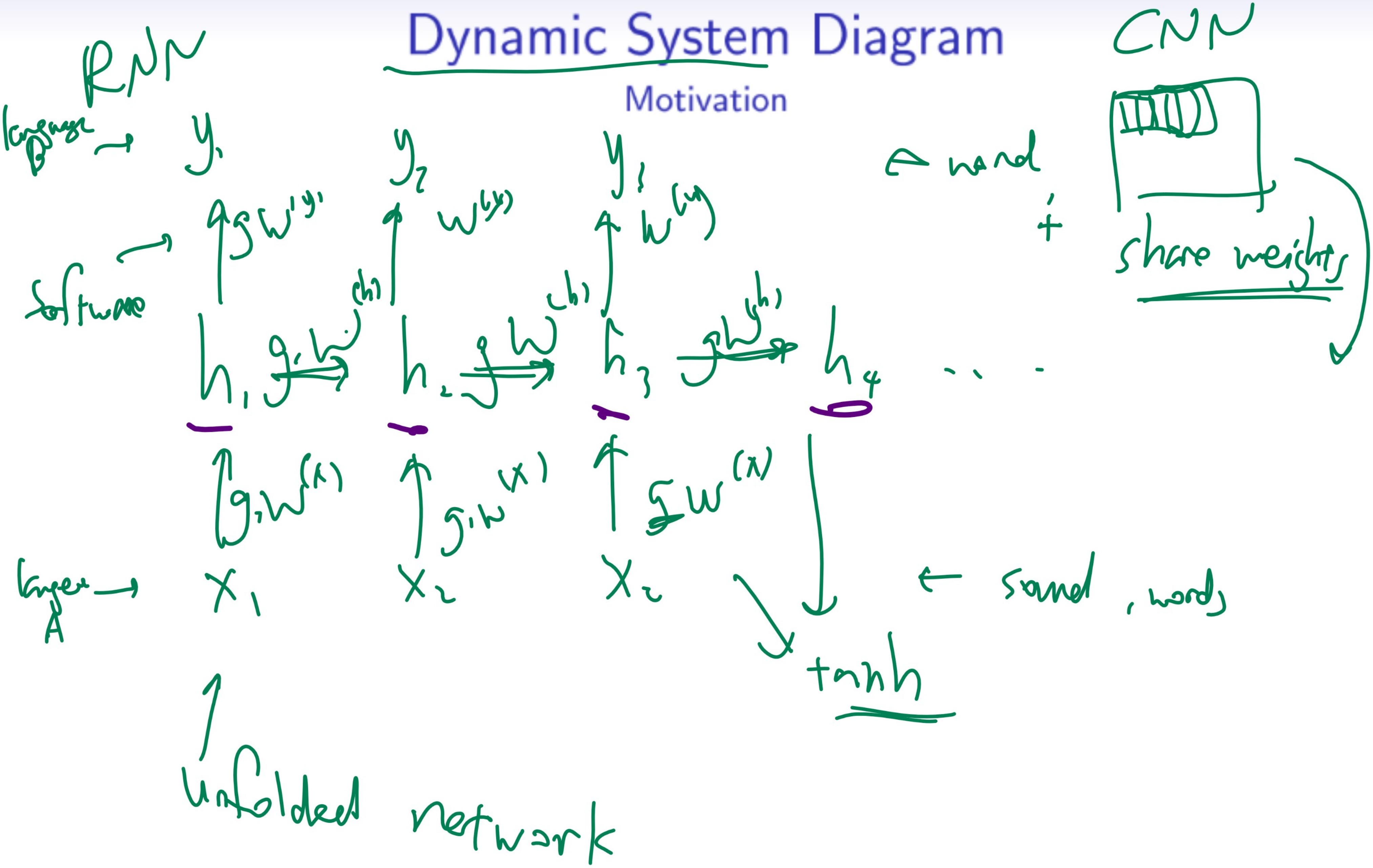
CS540 Introduction to Artificial Intelligence
Lecture 24

Young Wu
Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 22, 2020

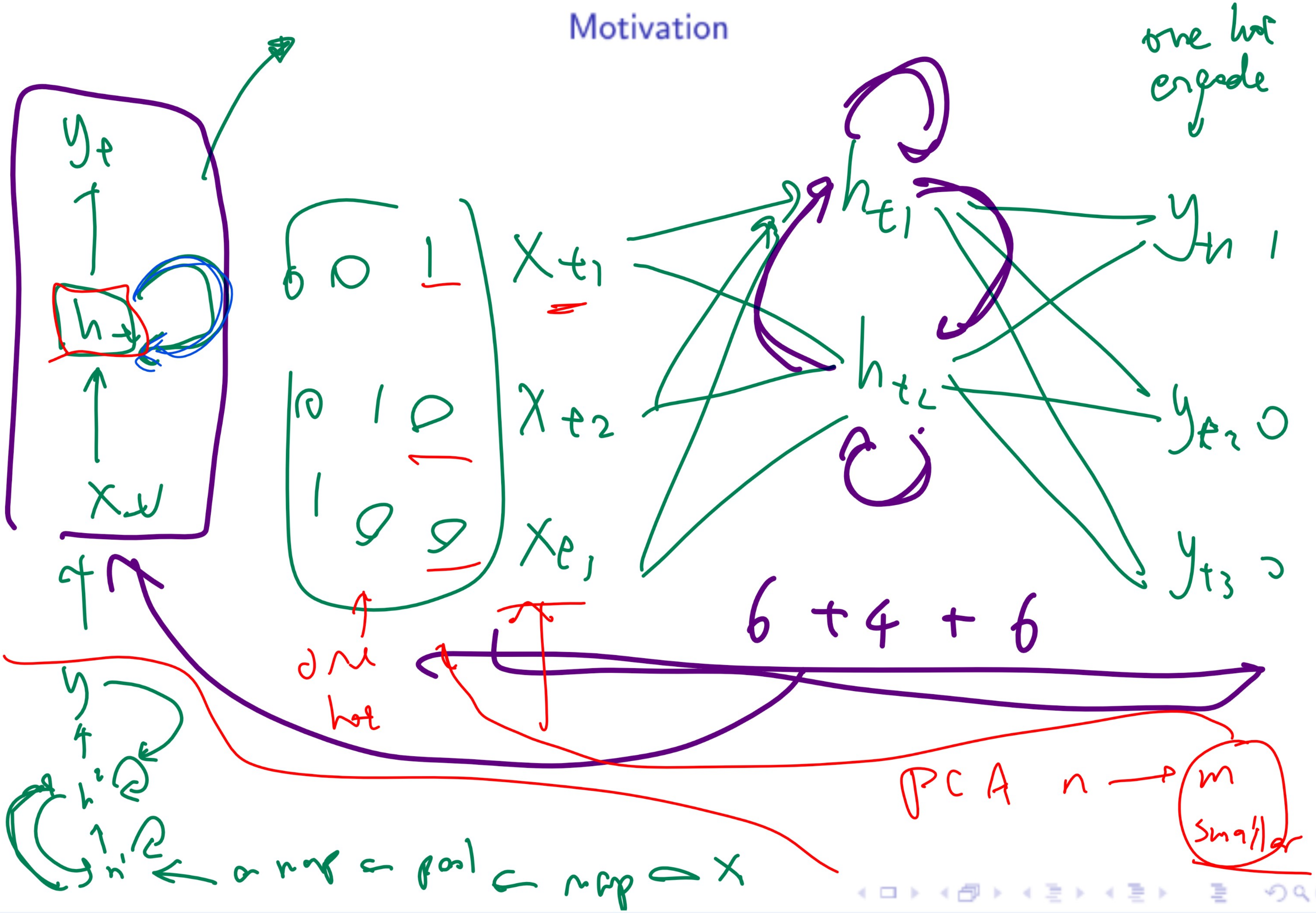
Dynamic System Diagram

Motivation



Recurrent Neural Network Structure Diagram

Motivation



Activation Functions

Definition

- The hidden layer activation function can be the tanh activation, and the output layer activation function can be the softmax function.

$$
\begin{aligned}
z_t^{(x)} &= \boxed{W^{(x)} x_t} + W^{(h)} \overset{h_{t-1}}{a_{t-1}^{(x)}} + b^{(x)} \\
h_t \quad a_t^{(x)} &= g\left(z_t^{(x)}\right), g\left(\boxed{\cdot}\right) = \tanh\left(\boxed{\cdot}\right) \\
z_t^{(y)} &= W^{(y)} a^{(1,t)} + b^{(y)} \\
a_t^{(y)} &= g\left(z_t^{(y)}\right), g\left(\boxed{\cdot}\right) = \text{softmax}\left(\boxed{\cdot}\right)
\end{aligned}
$$

$W^{(x)} \cancel{x} x_t$

Cost Functions

Definition

- Cross entropy loss is used with softmax activation as usual.

$$C_t = \underline{H}\left(y_t, a_t^{(y)}\right)$$

$$\underline{C} = \underline{\sum_t C_t}$$

BackPropogation Through Time

Definition

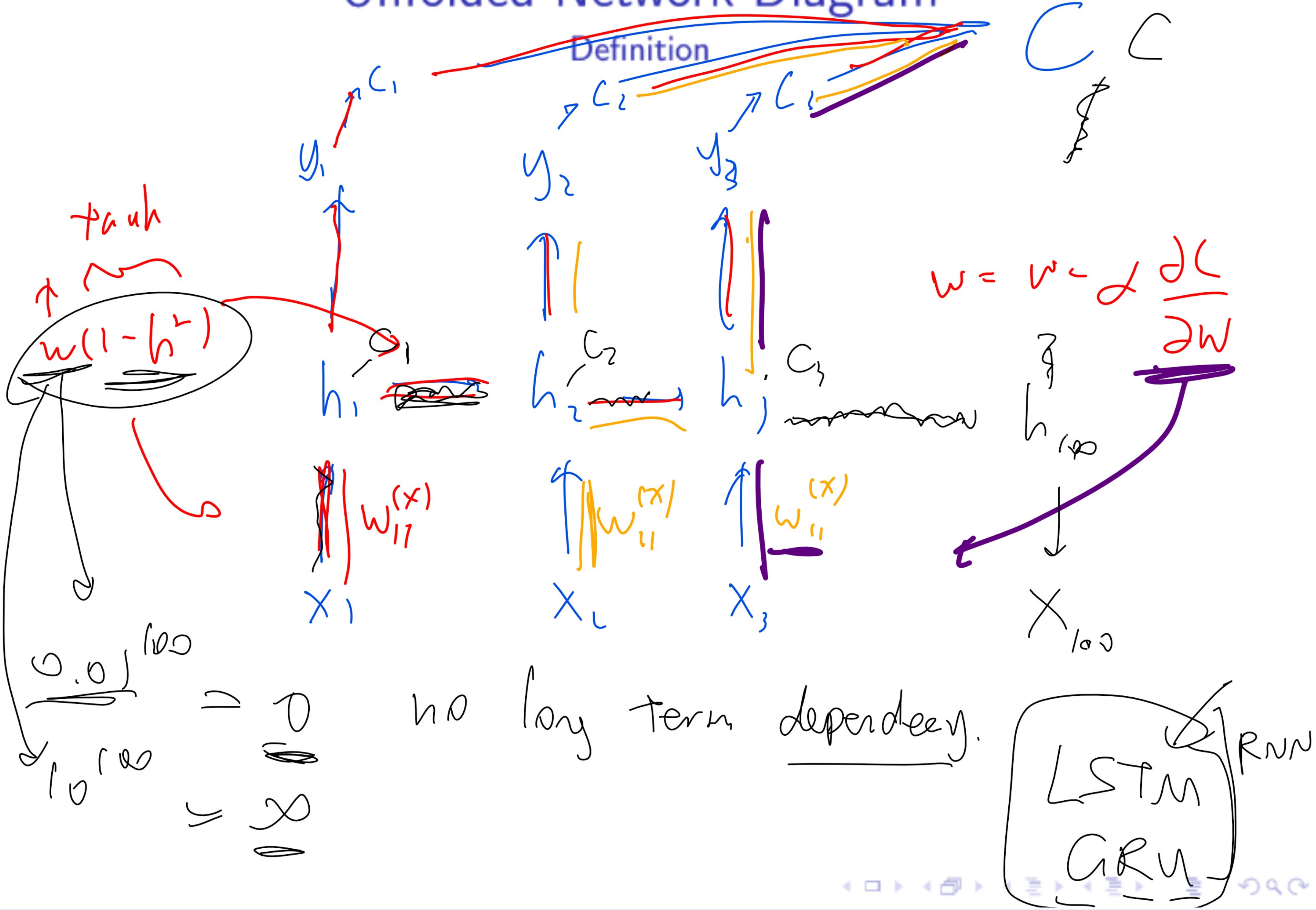
- The gradient descent algorithm for recurrent neural networks is called BackPropogation Through Time (BPTT). The update procedure is the same as standard neural networks using the chain rule.

$$w = w - \alpha \frac{\partial C}{\partial w}$$

$$b = b - \alpha \frac{\partial C}{\partial b}$$

Unfolded Network Diagram

Definition



$$w = w - \alpha \frac{\partial C}{\partial w}$$

$$0.01^{100} = 0$$

$$10^{100} = \infty$$

no long term dependency.

Backpropagation Diagram 1

Definition

Backpropagation Diagram 2

Definition

Vanishing and Exploding Gradient

Discussion

- If the weights are small, the gradient through many layers will shrink exponentially. This is called the vanishing gradient problem.
- If the weights are large, the gradient through many layers will grow exponentially. This is called the exploding gradient problem.
- Fully connected and convolutional neural networks only have a few hidden layers, so vanishing and exploding gradient is not a problem in training those networks.
- In a recurrent neural network, if the sequences are long, the gradients can easily vanish or explode.

Long Short Term Memory

Discussion

- Long Short Term Memory (LSTM) network adds more connected hidden units for memories controlled by gates. The activation functions used for these gates are usually logistic functions.
- An LSTM unit usually contains an input gate, an output gate, and a forget gate, to keep track of the dependencies in the input sequence.

Long Short Term Memory Diagram

Discussion

Gated Recurrent Unit

Discussion

- Gated Recurrent Unit (GRU) does something similar to an LSTM unit.
- A GRU contains input and forget gates, and does not contain an output gate.

Gated Recurrent Unit Diagram

Discussion