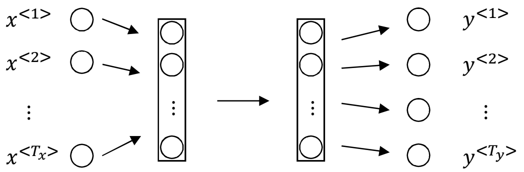
Motivation

Not all problems can be converted into one with fixed-length inputs and outputs

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.

1

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.

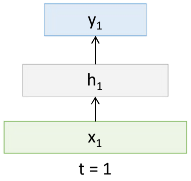
2

Recurrent Neural Networks (RNNs)

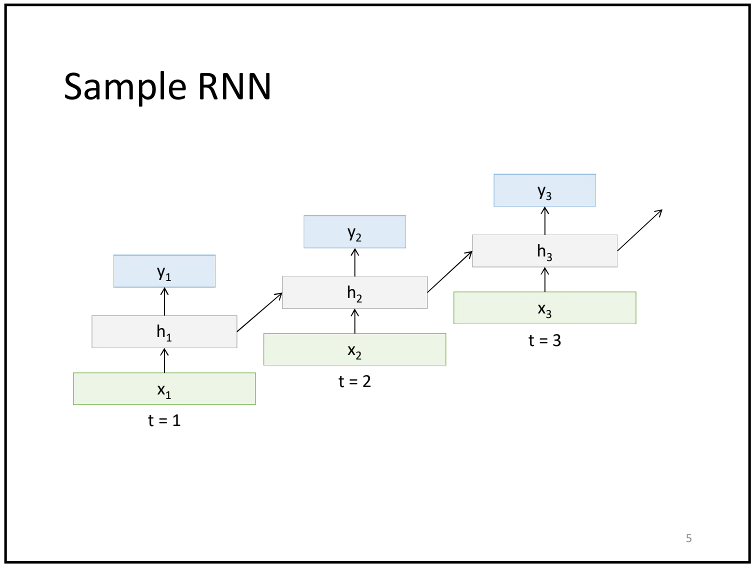
- Recurrent Neural Networks take the previous output or hidden states as inputs.
- The composite input at time t has some historical information about the happenings at time T < t
- RNNs are useful as their intermediate values (state) can store information about past inputs for a time that is not fixed a priori

3

Sample Feed-forward Network

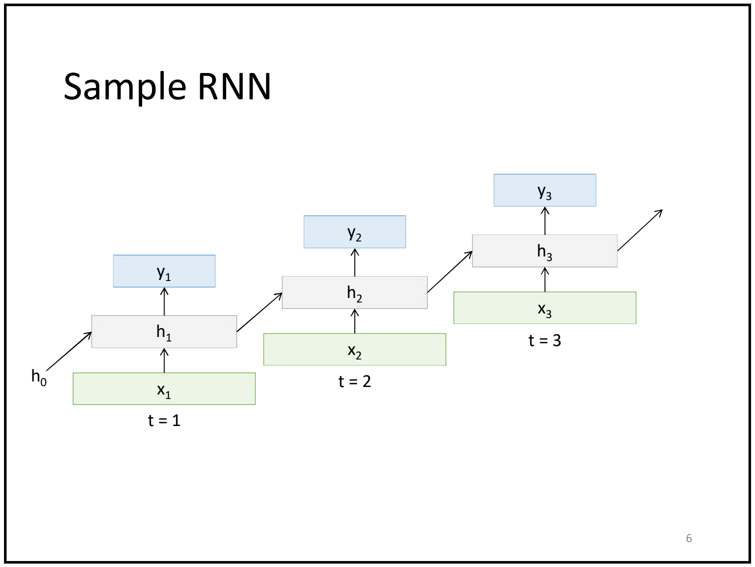


4



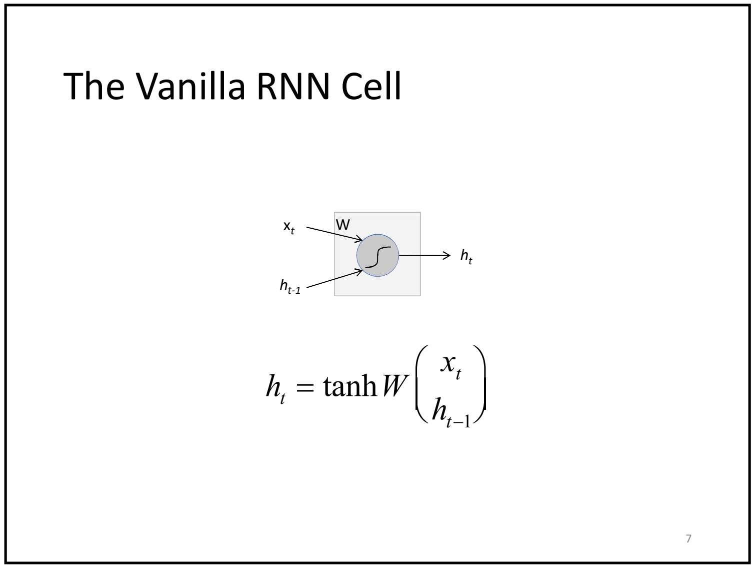
Sample RNN

5



Sample RNN

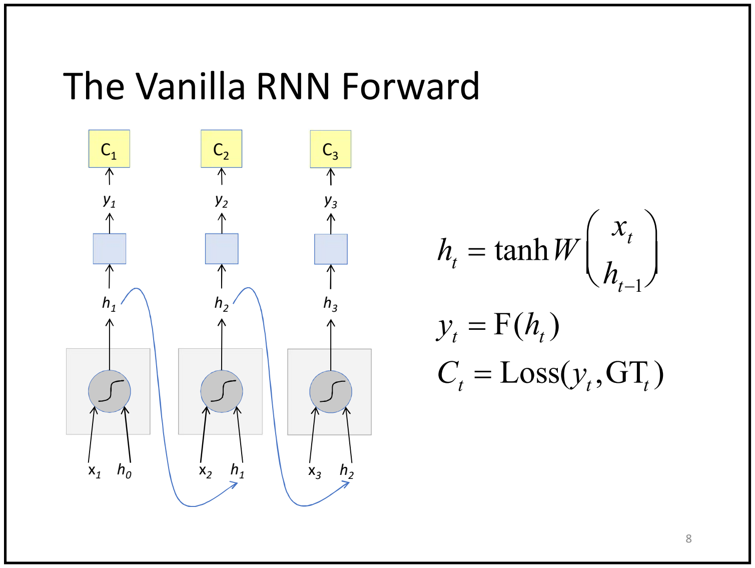
6



The Vanilla RNN Cell

$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

7



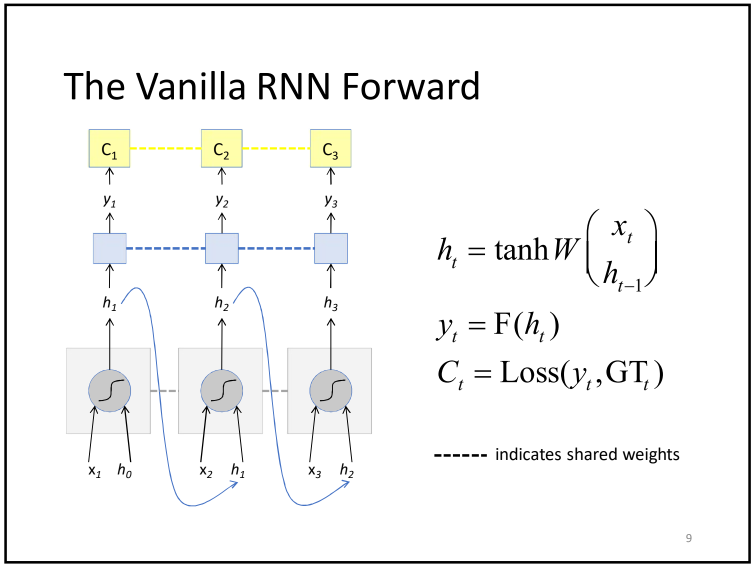
The Vanilla RNN Forward

$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

8



The Vanilla RNN Forward

$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

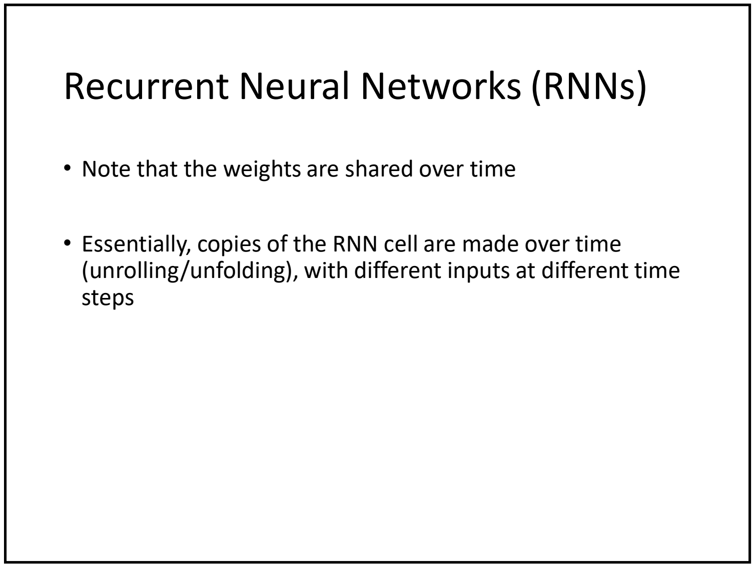
$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

------ indicates shared weights

9

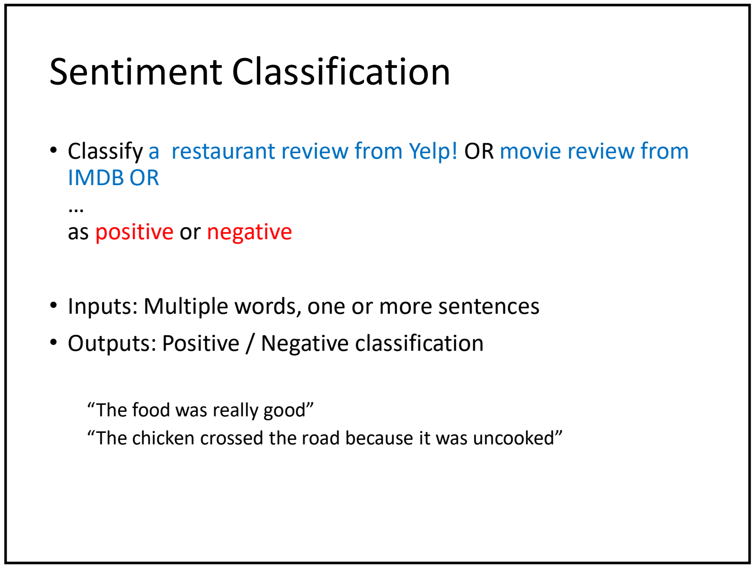
9



Recurrent Neural Networks (RNNs)

- Note that the weights are shared over time
- Essentially, copies of the RNN cell are made over time (unrolling/unfolding), with different inputs at different time steps

10

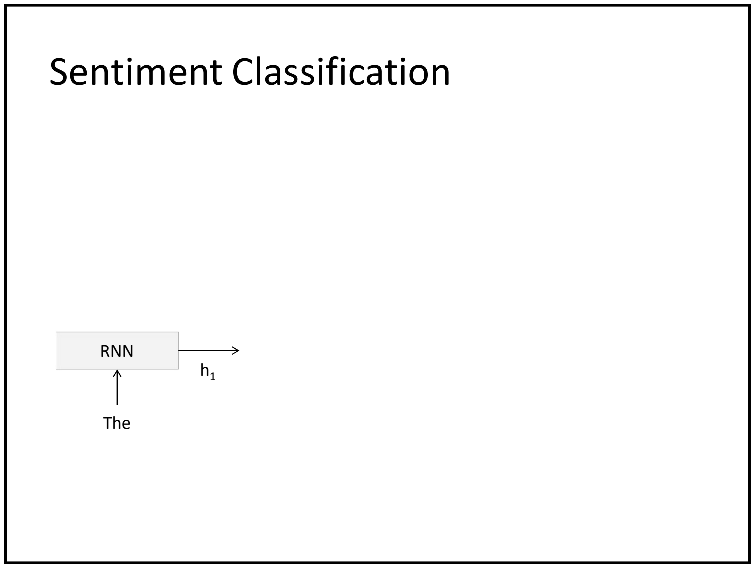


Sentiment Classification

- Classify a restaurant review from Yelp! OR movie review from IMDB OR
  …
  as positive or negative
- Inputs: Multiple words, one or more sentences
- Outputs: Positive / Negative classification

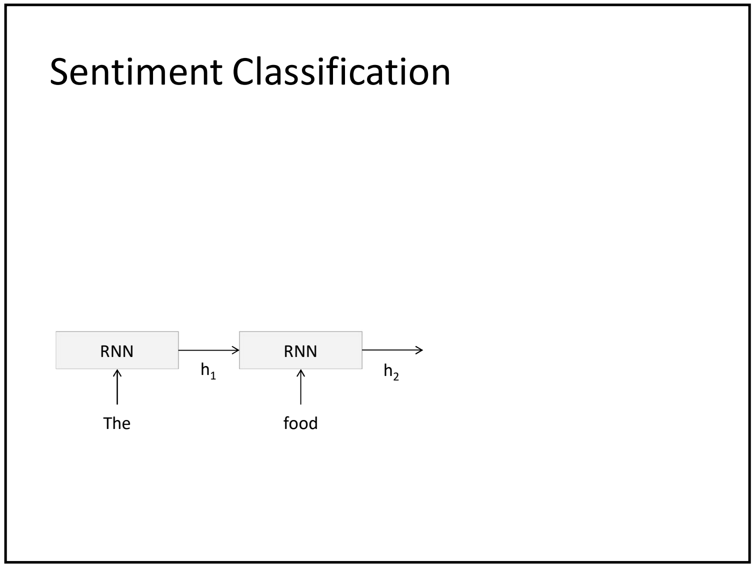
"The food was really good"
"The chicken crossed the road because it was uncooked"

11



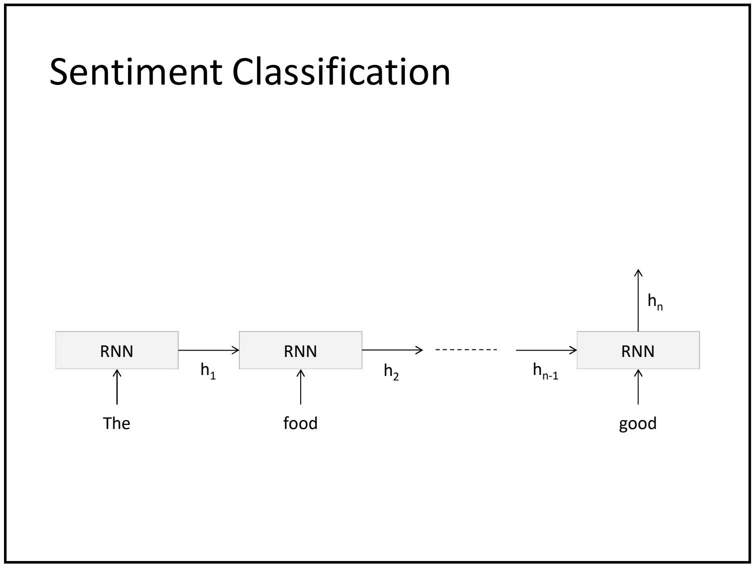
Sentiment Classification

12



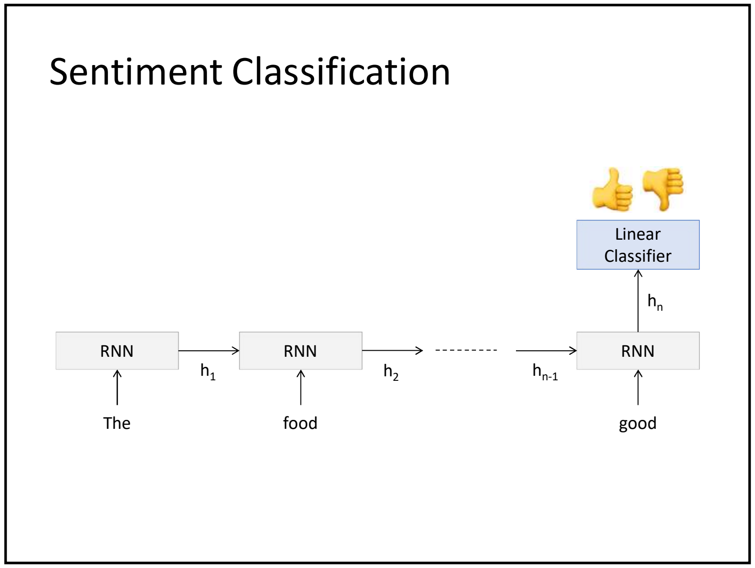
Sentiment Classification

13



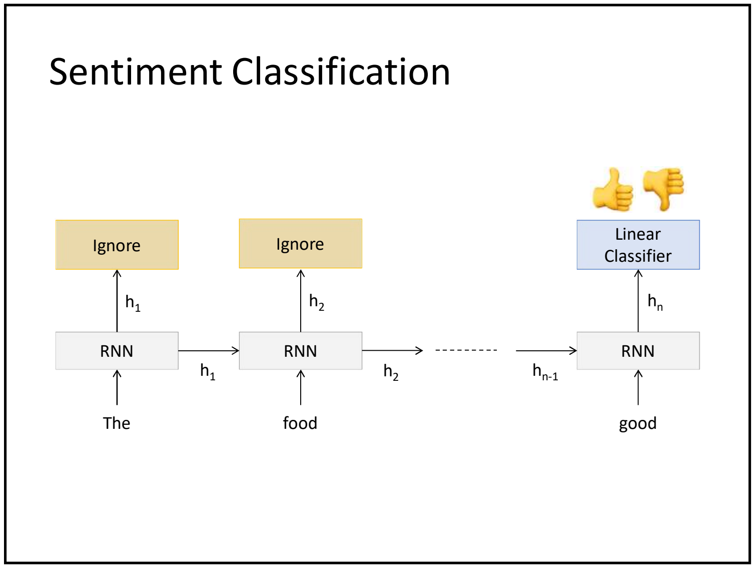
Sentiment Classification

14



Sentiment Classification

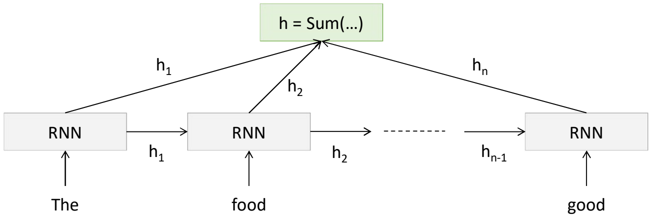
15



Sentiment Classification

16

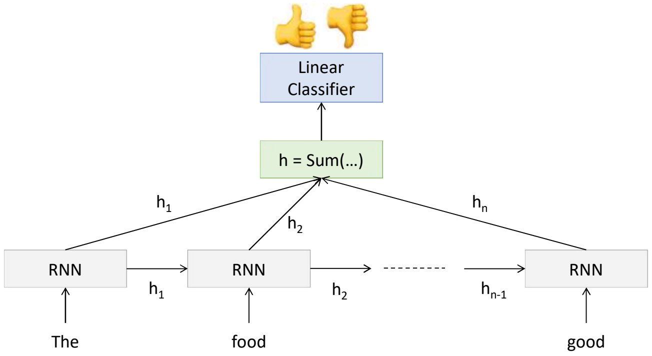
Sentiment Classification



http://deeplearning.net/tutorial/lstm.html

17

Sentiment Classification



http://deeplearning.net/tutorial/lstm.html

18

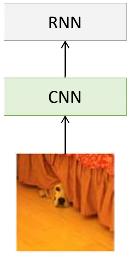
Image Captioning

- Given an image, produce a sentence describing its contents
- Inputs: Image feature (from a CNN)
- Outputs: Multiple words (let's consider one sentence)

: The dog is hiding

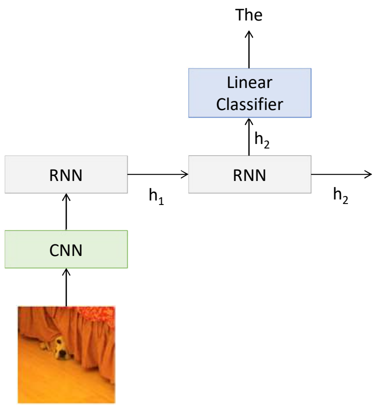
19

Image Captioning



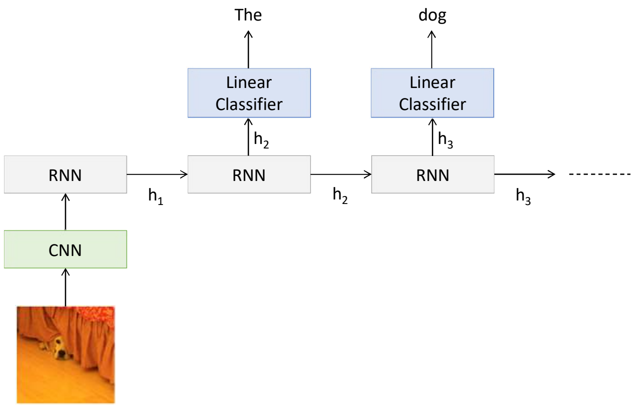
20

Image Captioning



21

Image Captioning



22

RNN Outputs: Image Captions

A person riding a motorcycle on a dirt road.

Two dogs play in the grass.

A herd of elephants walking across a dry grass field.

A group of young people playing a game of frisbee.

Two hockey players are fighting over the puck.

A close up of a cat laying on a couch.

Show and Tell: A Neural Image Caption Generator, CVPR 15

23

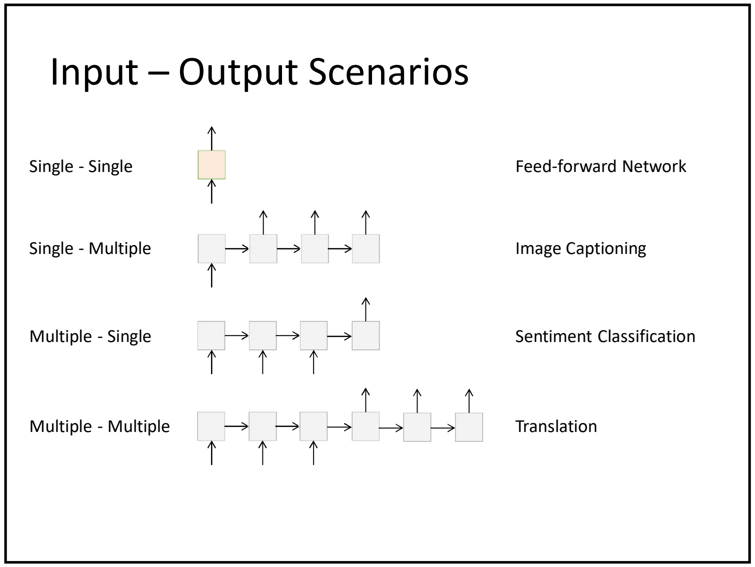
RNN Outputs: Language Modeling

VIOLA:
Why, Salisbury must find his flesh and thought
That which I am not aps, not a man and in fire,
To show the reining of the raven and the wars
To grace my hand reproach within, and not a fair are hand,
That Caesar and my goodly father's world;
When I was heaven of presence and our fleets,
We spare with hours, but cut thy council I am great,
Murdered and by thy master's ready there
My power to give thee but so much as hell:
Some service in the noble bondman here,
Would show him to her wine.

KING LEAR:
O, if you were a feeble sight, the courtesy of your law,
Your sight and several breath, will wear the gods
With his heads, and my hands are wonder'd at the deeds,
So drop upon your lordship's head, and your opinion
Shall be against your honour.

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

24



Input – Output Scenarios

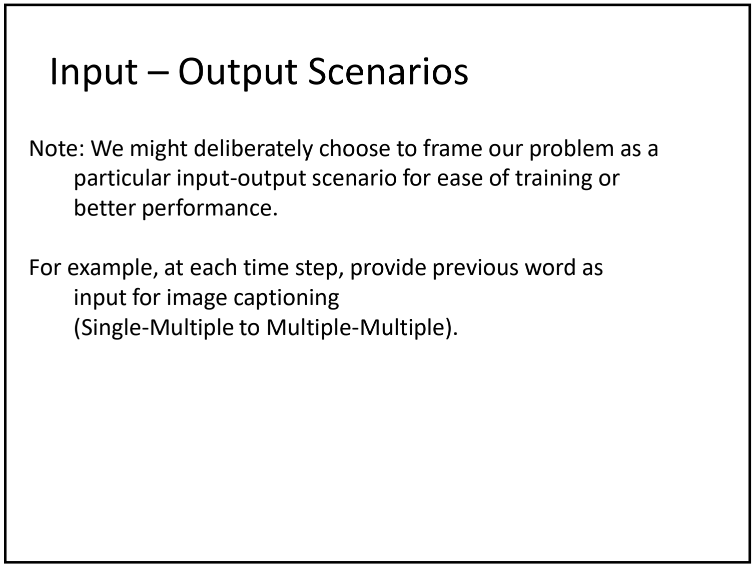
Feed-forward Network

Image Captioning

Sentiment Classification

Translation

25

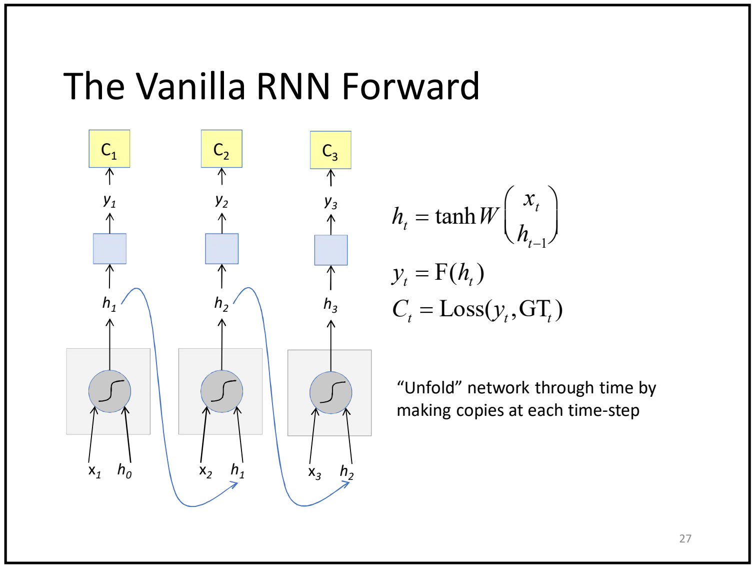


Input – Output Scenarios

Note: We might deliberately choose to frame our problem as a particular input-output scenario for ease of training or better performance.

For example, at each time step, provide previous word as input for image captioning
(Single-Multiple to Multiple-Multiple).

26



The Vanilla RNN Forward

$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

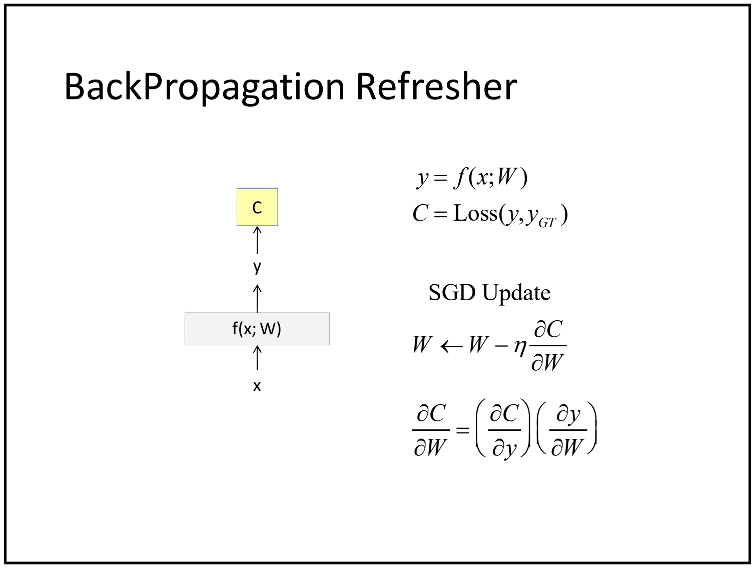
$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

“Unfold” network through time by making copies at each time-step

27

27



BackPropagation Refresher

$$y = f(x;W)$$

$$C = \mathrm{Loss}(y, y_{GT})$$

SGD Update

$$W \leftarrow W - \eta \frac{\partial C}{\partial W}$$

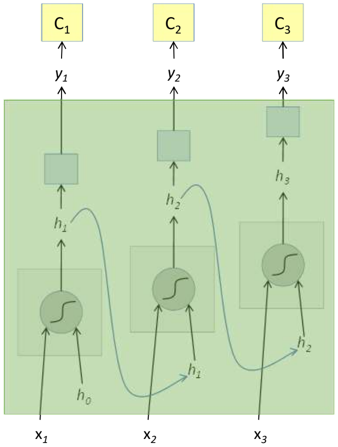
$$\frac{\partial C}{\partial W} = \left(\frac{\partial C}{\partial y}\right)\left(\frac{\partial y}{\partial W}\right)$$

28

BackPropagation Through Time (BPTT)

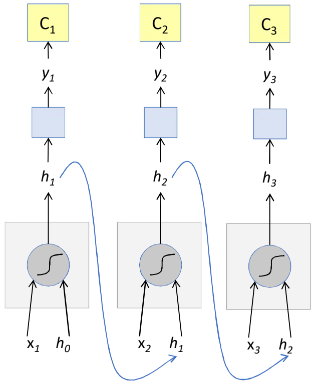
- One of the methods used to train RNNs
- The unfolded network (used during forward pass) is treated as one big feed-forward network
- This unfolded network accepts the whole time series as input
- The weight updates are computed for each copy in the unfolded network, then summed (or averaged) and then applied to the RNN weights

The Unfolded Vanilla RNN

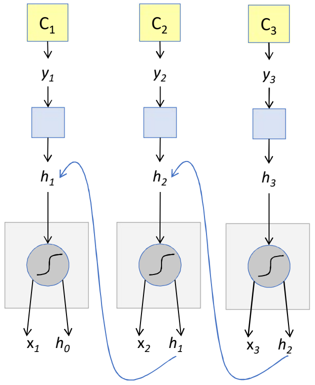


- Treat the unfolded network as one big feed-forward network!
- This big network takes in entire sequence as an input
- Compute gradients through the usual backpropagation
- Update shared weights

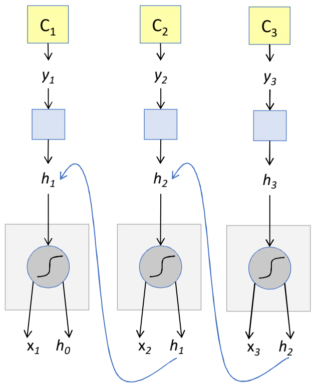
The Unfolded Vanilla RNN Forward



The Unfolded Vanilla RNN Backward



The Vanilla RNN Backward



$$h_t = \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$$

$$y_t = \mathrm{F}(h_t)$$

$$C_t = \mathrm{Loss}(y_t, \mathrm{GT}_t)$$

$$\frac{\partial C_t}{\partial h_1} = \left(\frac{\partial C_t}{\partial y_t}\right)\left(\frac{\partial y_t}{\partial h_1}\right)$$

$$= \left(\frac{\partial C_t}{\partial y_t}\right)\left(\frac{\partial y_t}{\partial h_t}\right)\left(\frac{\partial h_t}{\partial h_{t-1}}\right)\cdots\left(\frac{\partial h_2}{\partial h_1}\right)$$

33

Issues with the Vanilla RNNs

- Information morphing (fundamental) : inability to keep the memory content for more than a few step
- Gradient vanishing (technical) and Exploding

[1] On the difficulty of training recurrent neural networks, Pascanu *et al.*, 2013

34

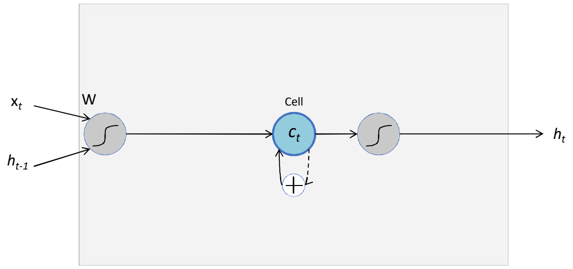
Long Short-Term Memory (LSTM)[1]

- The LSTM uses this idea of "Constant Error Flow" for RNNs to create a "Constant Error Carousel" (CEC) which ensures that gradients don't decay
- The key component is a memory cell that acts like an accumulator (contains the identity relationship) over time
- Instead of computing new state as a matrix product with the old state, it rather computes the difference between them. Expressivity is the same, but gradients are better behaved

[1] Long Short-Term Memory, Hochreiter *et al.*, 1997

35

The LSTM Idea

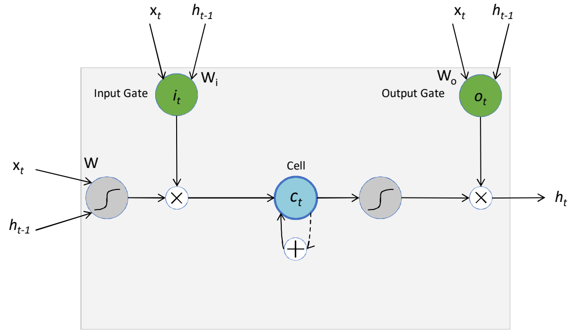


$$c_t = c_{t-1} + \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} \qquad h_t = \tanh c_t$$

* Dashed line indicates time-lag

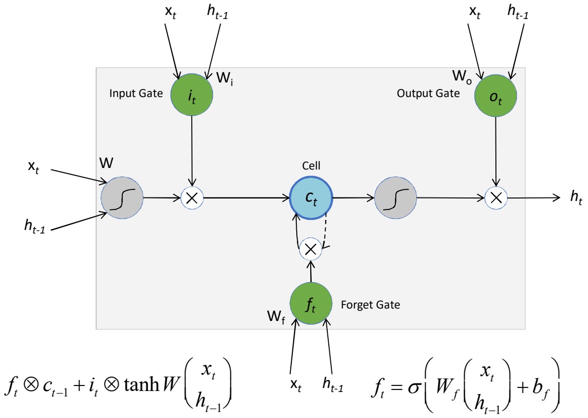
36

The Original LSTM Cell



$c_t = c_{t-1} + i_t \otimes \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$ $h_t = o_t \otimes \tanh c_t$ $i_t = \sigma\left(W_i \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_i \right)$ Similarly for o_t

The Popular LSTM Cell



$c_t = f_t \otimes c_{t-1} + i_t \otimes \tanh W \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix}$ $f_t = \sigma\left(W_f \begin{pmatrix} x_t \\ h_{t-1} \end{pmatrix} + b_f \right)$

- ***Input gate****:* controls the extent to which a new value flows into the cell,
- ***Forget gate****:* controls the extent to which a value remains in the cell and
- ***Output gate****:* controls the extent to which the value in the cell is used to compute the output activation of the LSTM unit.

Summary

- RNNs allow for processing of variable length inputs and outputs by maintaining state information across time steps
- Various Input-Output scenarios are possible (Single/Multiple)
- Vanilla RNNs are improved upon by LSTMs which address the vanishing gradient problem
- Exploding gradients are handled by gradient clipping