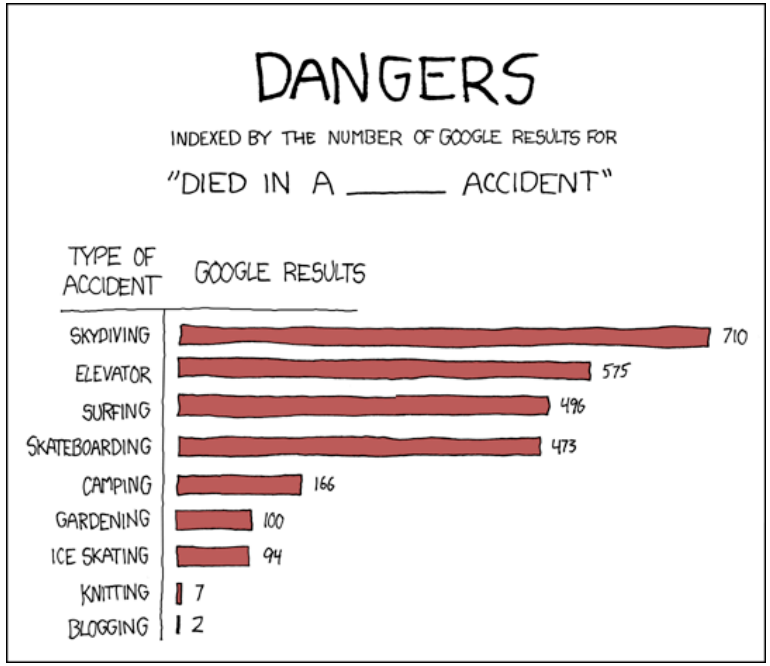
COMP 4446 / 5046
Lecture 3: Models – Non-linear

Jonathan K. Kummerfeld

Semester 1, 2025

Dangers



[Zero results: 'snake charming' and 'haberdashery'. (Things like 'car' and 'boating' and such are of course the highest, by a huge margin.)]

Source: https://xkcd.com/369/

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Deadlines

Assignment 1 - Due TOMORROW 11:59pm

Assignment 2 - Available on Thursday

Slip-day policy reminder:
- Max 2 per assignment
- 5 over the semester

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Quiz

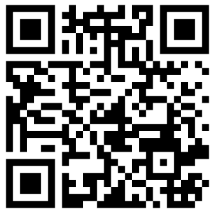
10 minutes long at the start of the lecture

On paper - bring a dark pen

No devices permitted

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

Neural Networks

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

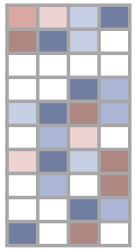
Core idea: Add a non-linear function in the model

Model

Linear model

$$\text{scores(doc)} = \text{features(doc)} \cdot \text{weights}$$

$$= \quad \cdot$$

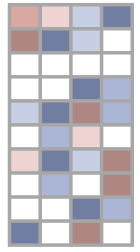


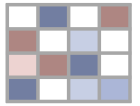
$$f(x) = x \cdot W$$

Non-Linear model

$$\text{scores(doc)} = g(\text{features(doc)} \cdot \text{weights}_1) \cdot \text{weights}_2$$

$$= g(\quad \cdot \quad) \cdot$$





Sigmoid, etc.

$$f(x) = g(x \cdot W_1) \cdot W_2 \quad + \text{bias}$$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

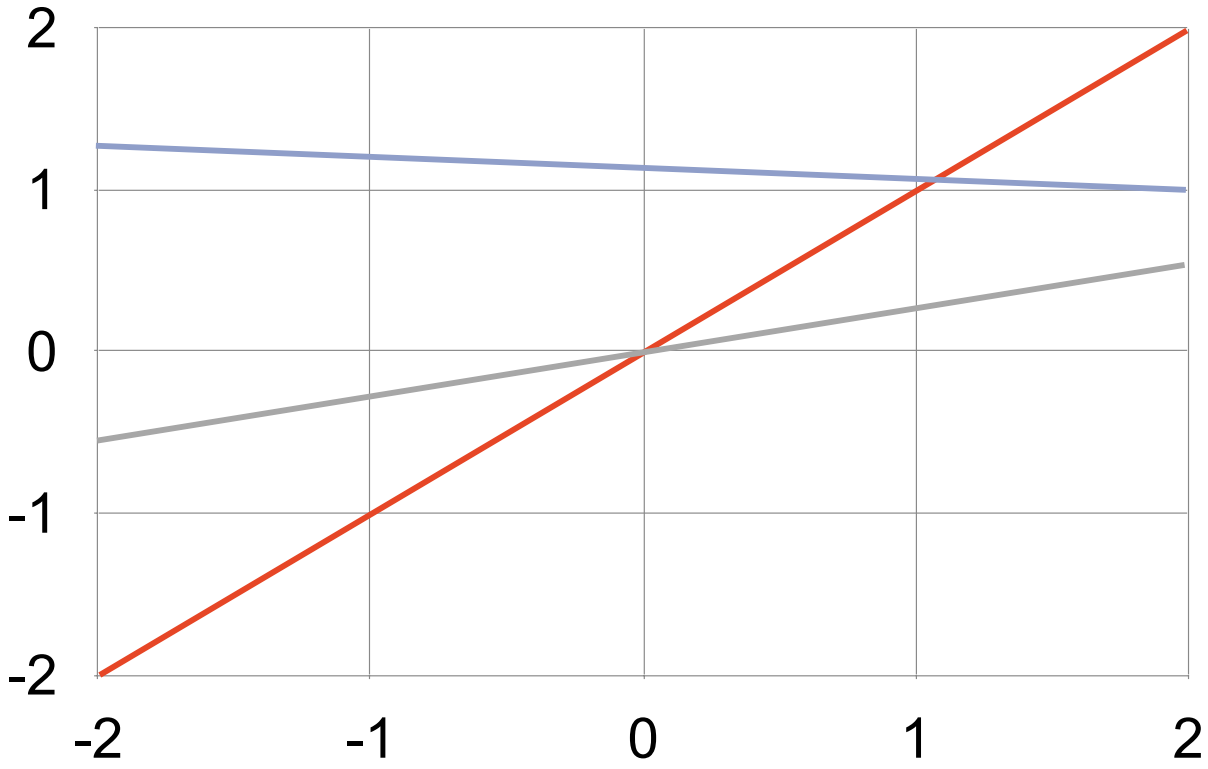
menti.com 1750 7815

What is a non-linear function?

Model

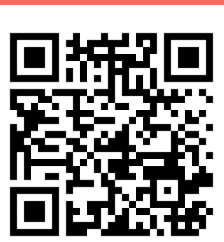
Linear functions

$$f(x) = ax + b$$



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

What is a non-linear function?

Model

Sigmoid

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$

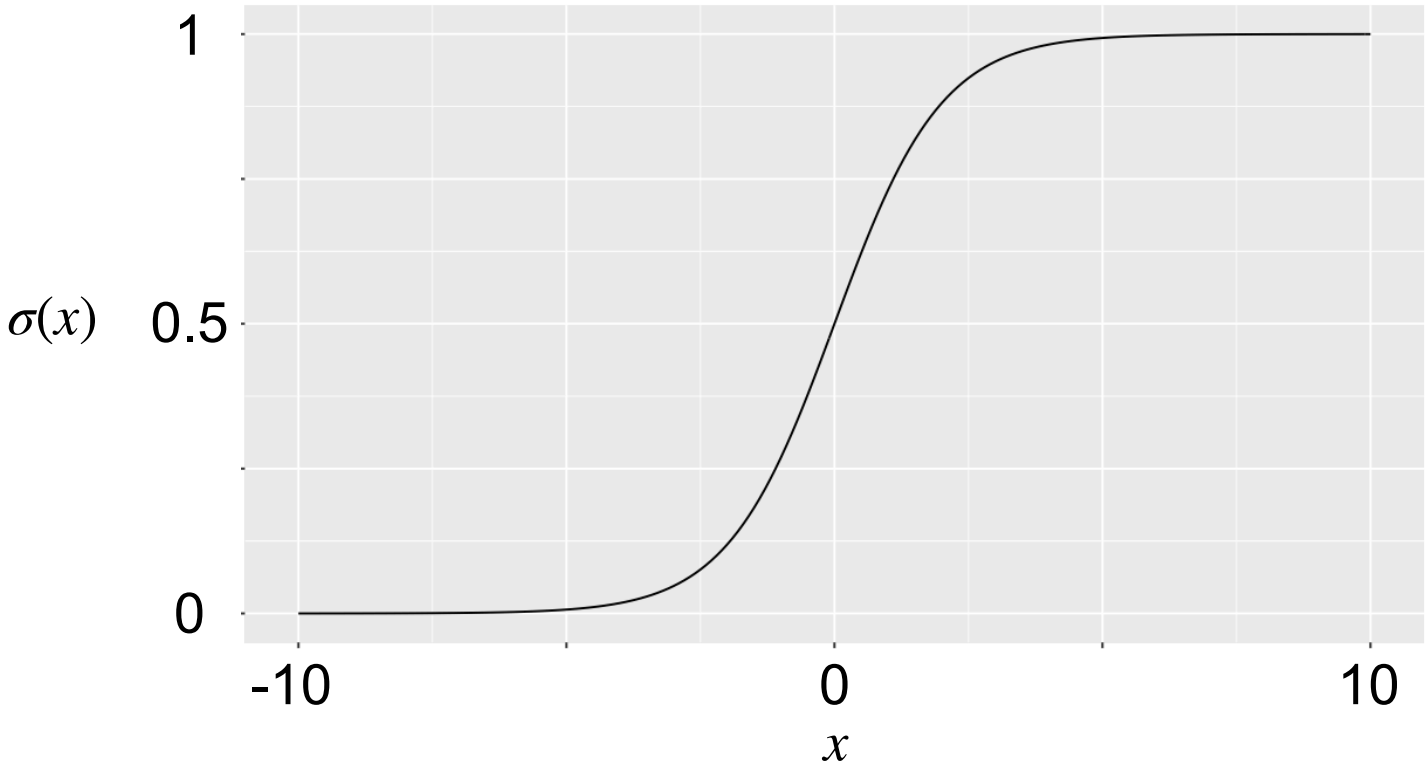
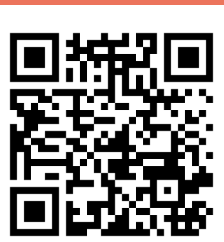


Figure from David Bamman's Lecture Slides for Info 159/259 at UC Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

What is a non-linear function?

Model

Hyperbolic Tangent

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

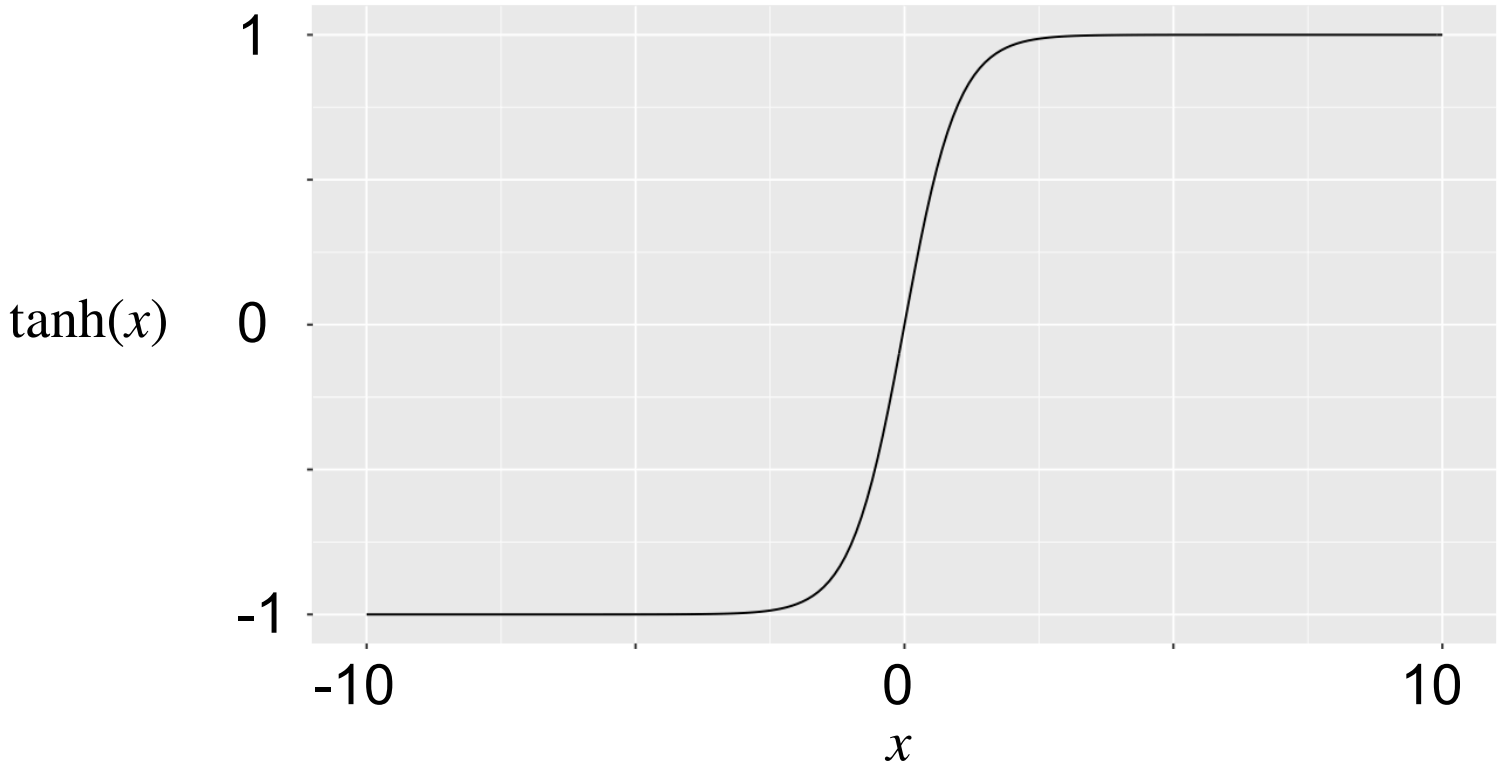


Figure from David Bamman's Lecture Slides for Info 159/259 at UC Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

What is a non-linear function?

Rectified Linear Unit (ReLU)

Model

$$\text{ReLU}(x) = \max(0,x)$$

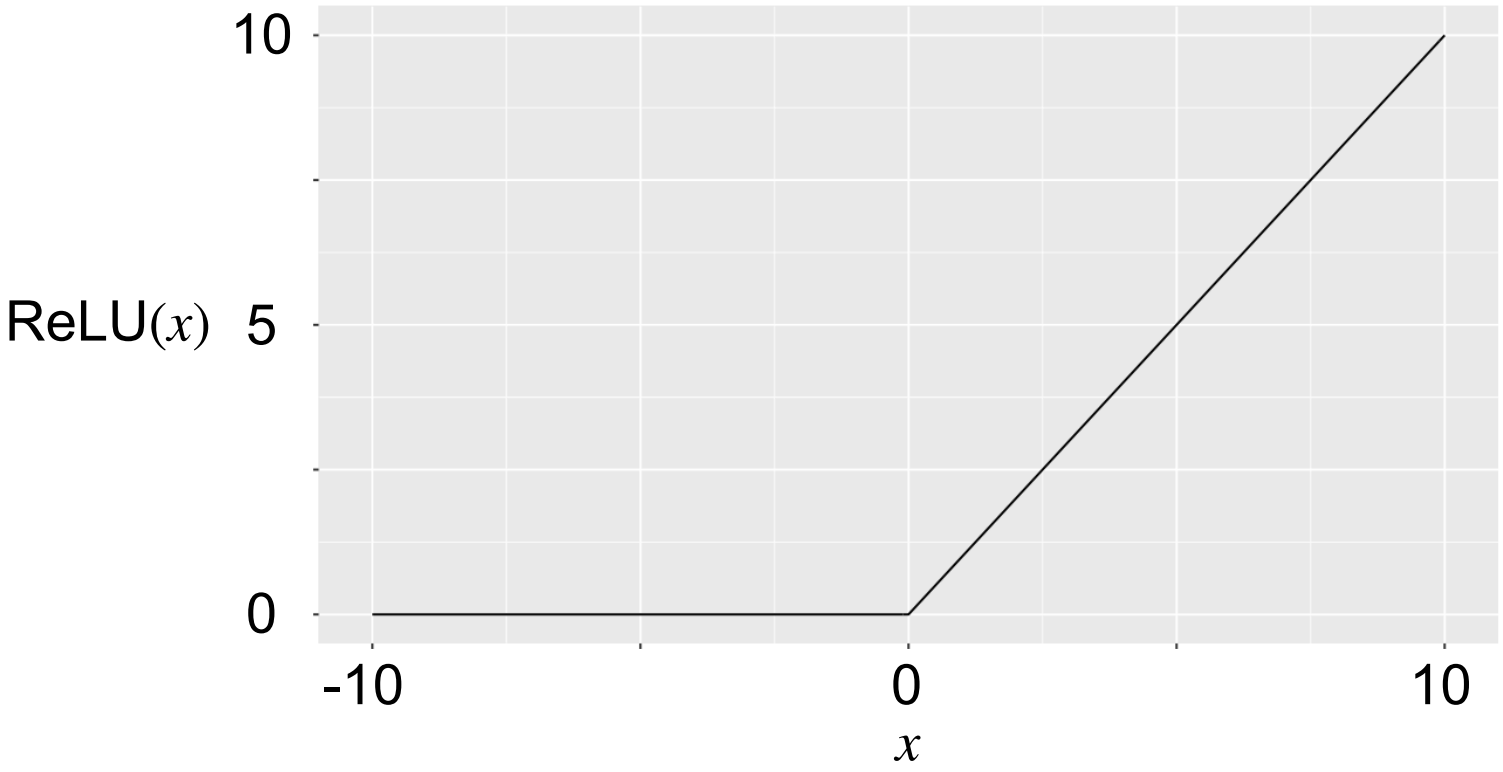


Figure from David Bamman's Lecture Slides for Info 159/259 at UC Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

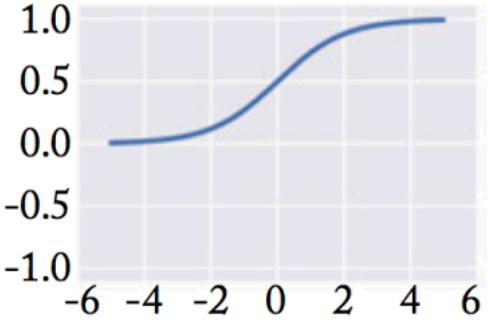
menti.com 1750 7815

What is a non-linear function?

Model

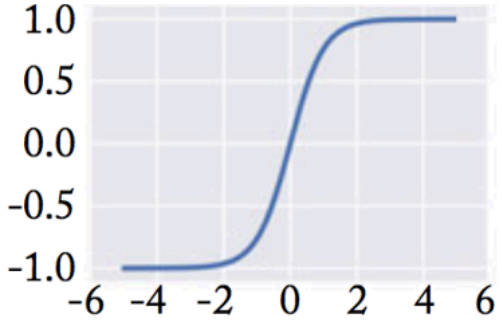
1

$$\sigma(x) = \frac{1}{1 + e^{-x}}$$



2

$$\tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$



3

$$\text{ReLU}(x) = \max(0,x)$$

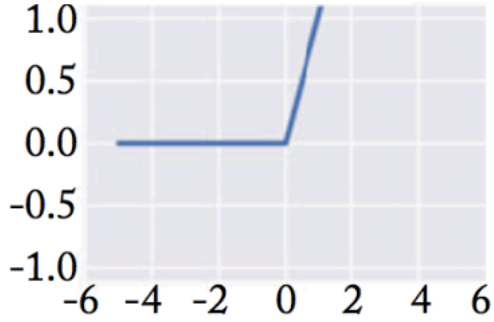


Figure from Goldberg, “Neural Network Methods for Natural Language Processing“

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

What is a non-linear function?

Model

$$\sigma(x) = \frac{1}{1 + e^{-x}} \qquad \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}} \qquad \text{ReLU}(x) = \max(0,x)$$

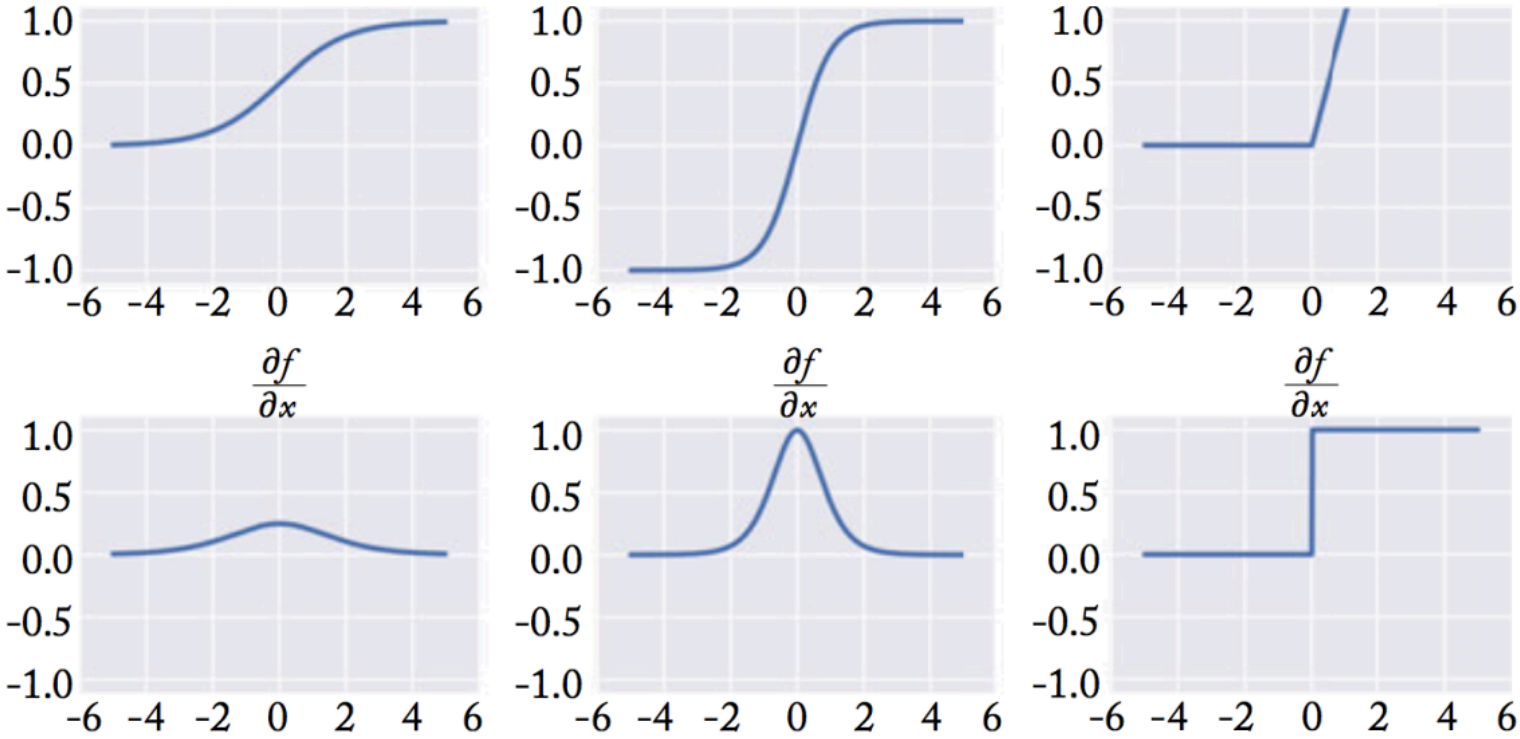


Figure from Goldberg, “Neural Network Methods for Natural Language Processing“

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

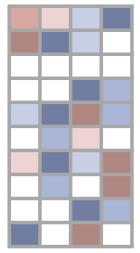
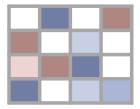
menti.com 1750 7815

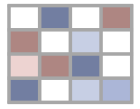
Model

What is a non-linear function?

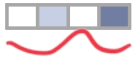
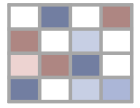
each layer has its own weight

$$\text{scores}(\text{doc}) = g(\text{features}(\text{doc}) \cdot \text{weights}_1) \cdot \text{weights}_2$$

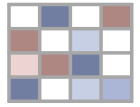
$= g($  $\cdot$  $) \cdot$ 

$= g($  $) \cdot$ 

If $g(x) = \text{ReLU}(x) = \max(0,x)$

$=$  $\cdot$ 

If $g(x) = \tanh(x) = \dfrac{e^x - e^{-x}}{e^x + e^{-x}}$

$=$  $\cdot$ 

连2个 non linear 结果不同

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

What is a non-linear function?

Model

$$f(x) = g(x \cdot W_1) \cdot W_2$$ ← 1 layers

$$f(x) = h(g(x \cdot W_1) \cdot W_2) \cdot W_3$$ ← 2 layers

Each stage of non-linearity + weight multiplication is a layer

We typically also have a bias (fixed value that is added on):

$$f(x) = h(g(x \cdot W_1 + b_1) \cdot W_2 + b_2) \cdot W_3 + b_3$$

Can also think of this as an extra input that is always 1

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

What is a non-linear function?

Model

So far: all feed-forward neural networks

Also called multi-layer perceptrons (MLP)

We'll see other types in the weeks ahead

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How do we update the models? First consider linear

Learning Method

$$\text{score}(\text{doc}) = \text{features}(\text{doc}) \cdot \text{weights}$$

$$= \quad \cdot$$

$$f(x) = x \cdot W$$ ← linear model

Week 2: Perceptron

$$W_i = \begin{cases} W_i, \text{ if correct} \\ W_i - \text{sign}(x_i), \text{ if incorrect} \end{cases}$$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

通过 Loss Function iteration

How do we update the models? First consider linear

Learning Method

$$\text{score}(\text{doc}) = \text{features}(\text{doc}) \cdot \text{weights}$$

$$= \quad \cdot$$

$$f(x) = x \cdot W$$

Week 3: Move based on the derivative of loss / cost

- Loss is a mathematical function measuring how wrong we are
- Derivative is the direction we should move to increase the function
- Move in the negative derivative to get better!

COMP 4446 / 5046
Lecture 3, 2025

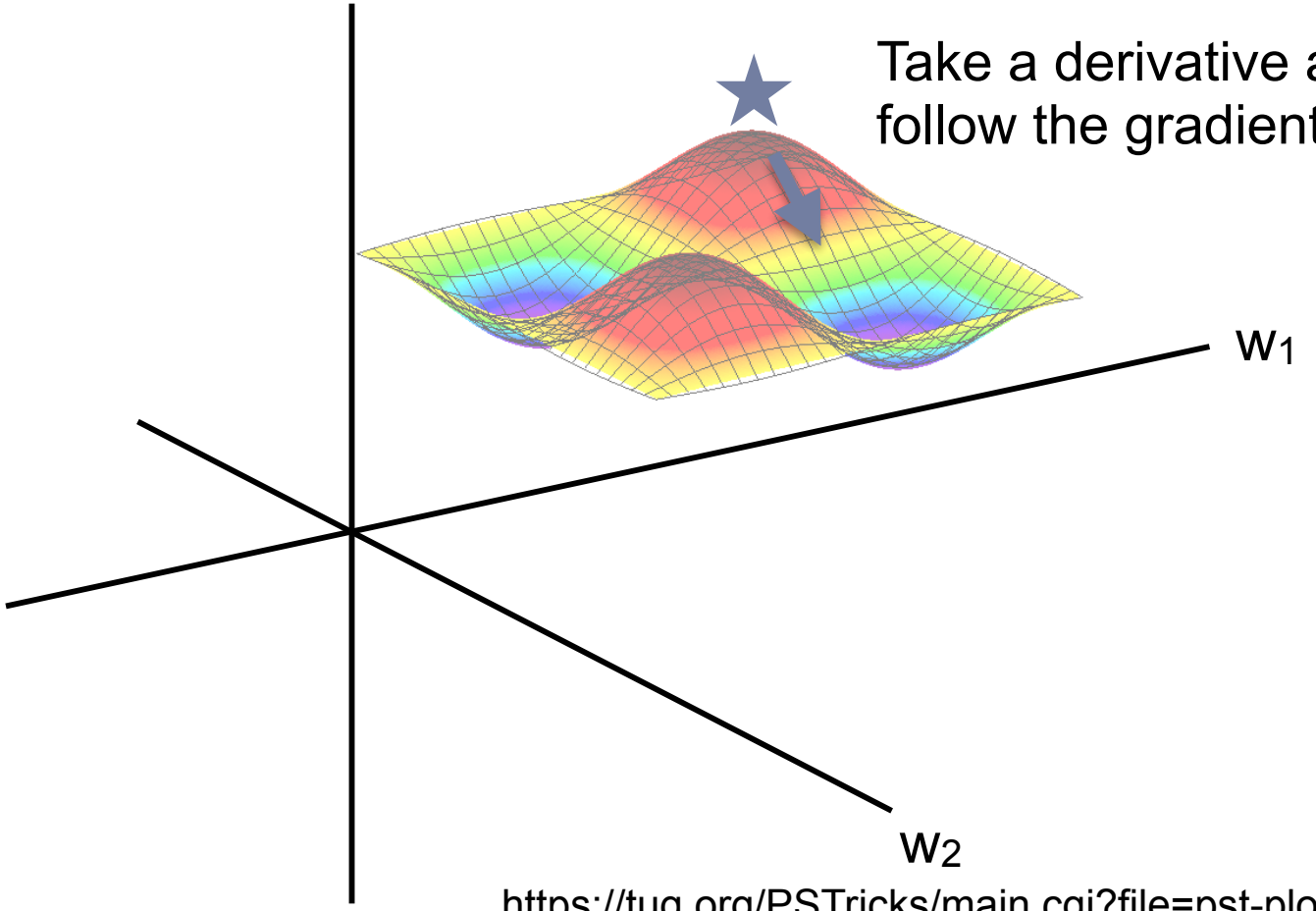
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How do we update the models? First consider linear

Learning Method

y = loss / cost, ie., how many errors we are making



https://tug.org/PSTricks/main.cgi?file=pst-plot/3D/contour

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How do we update the models? First consider linear

Learning Method

$$\text{score}(\text{doc}) = \text{features}(\text{doc}) \cdot \text{weights}$$

$$= \quad \cdot$$

$$f(x) = x \cdot W$$

Week 3: Move based on the derivative of loss

$$\boxed{loss} = -\log(p(y \mid x))$$

$$= -[y \log \hat{y} + (1-y) \log(1-\hat{y})]$$

sigmoid function

$$= -[y \log \sigma(x \cdot W) + (1-y) \log(1-\sigma(x \cdot W))]$$

$$= \begin{cases} -\log \sigma(x \cdot W), \text{if } y = 1 \\ \log(1-\sigma(x \cdot W)), \text{if } y = 0 \end{cases}$$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How do we update the models? First consider linear

Learning Method

$$\text{score}(\text{doc}) = \text{features}(\text{doc}) \cdot \text{weights}$$

$$= \quad \cdot$$

$$f(x) = x \cdot W$$

This is the sigmoid function from earlier, which conveniently maps values to between 0 and 1

Week 3: Move b ss

$$loss = -log(p$$

$$= -[y\ log$$

$$= -[y\ log\ \sigma(x \cdot W) + (1-y)\ log(1-\sigma(x \cdot W))]$$

$$\frac{d\ loss}{dW_j} = (\sigma(x \cdot W) - y)x_j$$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

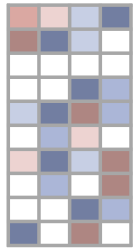
back propagation

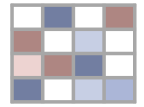
How do we update the models? Now non-linear

Learning Method

$$\text{scores(doc)} = g(\text{features(doc)} \cdot \text{weights}_1) \cdot \text{weights}_2$$

$$= g(\quad \cdot \quad) \cdot$$





$$f(x) = g(x \cdot W_1) \cdot W_2$$

Just another function!

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

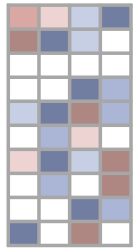
menti.com 1750 7815

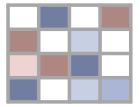
How can we avoid doing so much differentiation?

Learning Method

$$\text{scores}(\text{doc}) = g(\text{features}(\text{doc}) \cdot \text{weights}_1) \cdot \text{weights}_2$$

$$= g(\quad \cdot \quad) \cdot$$





$$f(x) = g(x \cdot W_1) \cdot W_2$$

Need derivatives for W_1 and W_2 and more as we add layers

$$f(x) = h(g(x \cdot W_1) \cdot W_2) \cdot W_3$$

Chain rule!

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

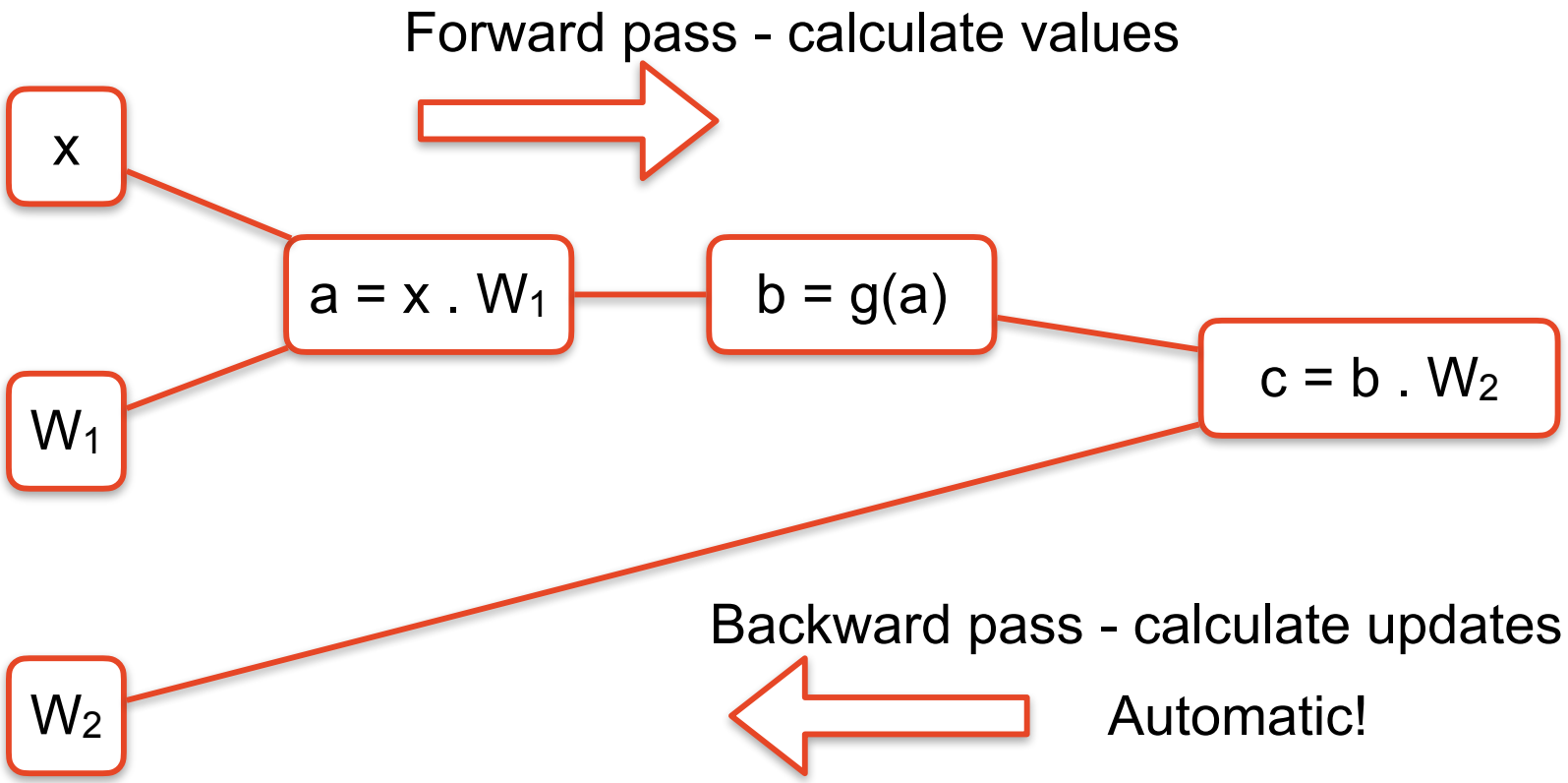
menti.com 1750 7815

Computation graphs and backpropagation

Learning Method

$$f(x) = g(x \cdot W_1) \cdot W_2$$

The equation above can be represented as a graph



This is one loop

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Computation graphs and backpropagation

Learning Method

Chain rule

$$f(x) = h(g(x)) \qquad \frac{df}{dx} = \frac{dh}{dg} \cdot \frac{dg}{dx}$$

$$f(x) = g_3(g_2(g_1(x))) \qquad \frac{df}{dx} = \frac{dg_3}{dg_2} \cdot \frac{dg_2}{dg_1} \cdot \frac{dg_1}{dx}$$

$$\frac{d}{dx}\left[f(g(x))\right] = f'(g(x))\,g'(x)\,x'$$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Computation graphs and backpropagation

Learning Method

There is some randomness to it because it depends on the order we see data in

We go downwards in loss space

Stochastic gradient descent

We differentiate to get the gradient

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Computation graphs and backpropagation

Learning Method

Libraries do the calculations for us:

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

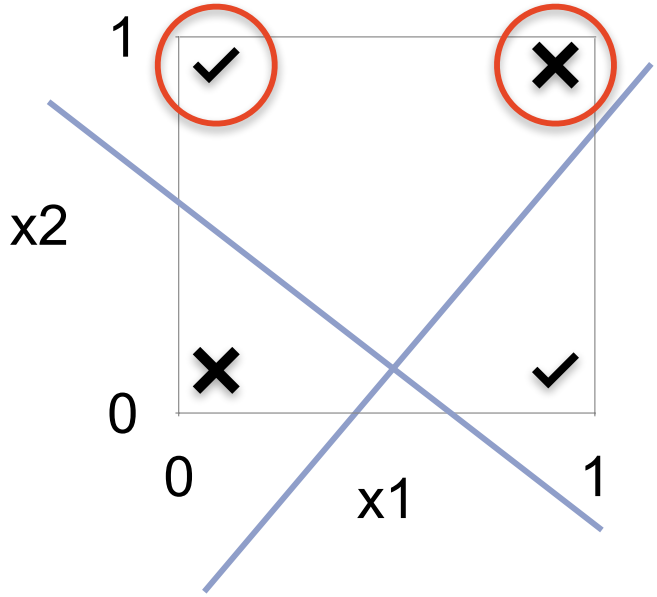
menti.com 1750 7815

What can we learn with this?

Model

XOR: Two inputs, x1 and x2

x2 \ x1	0	1
0	✗	✓
1	✓	✗



cannot use linear to represent it

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

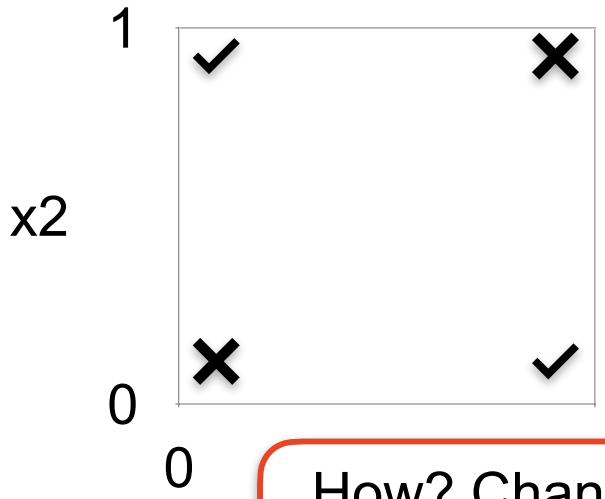
menti.com 1750 7815

What can we learn with this?

Model

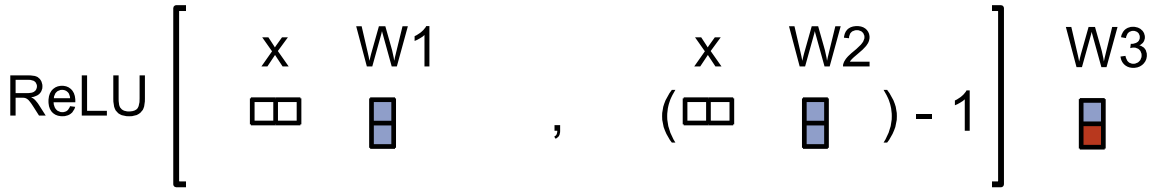
XOR: Two inputs, x1 and x2

x2 \ x1	0	1
0	✗	✓
1	✓	✗



How? Change representation / space

ReLU(x1+ x2) - 2 (ReLU(x1 + x2 - 1)



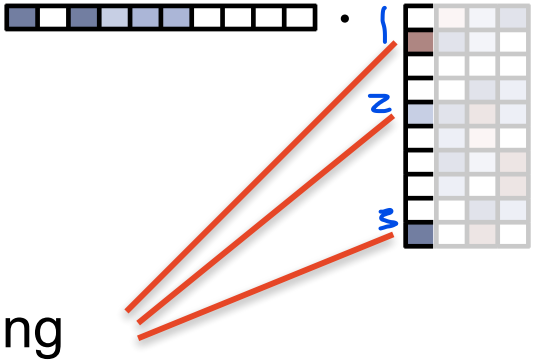
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

What can we learn with this?

Model



Combining
these 3 inputs

e.g., are 'very', and 'good' present, but not 'bad'?

positive

prediction

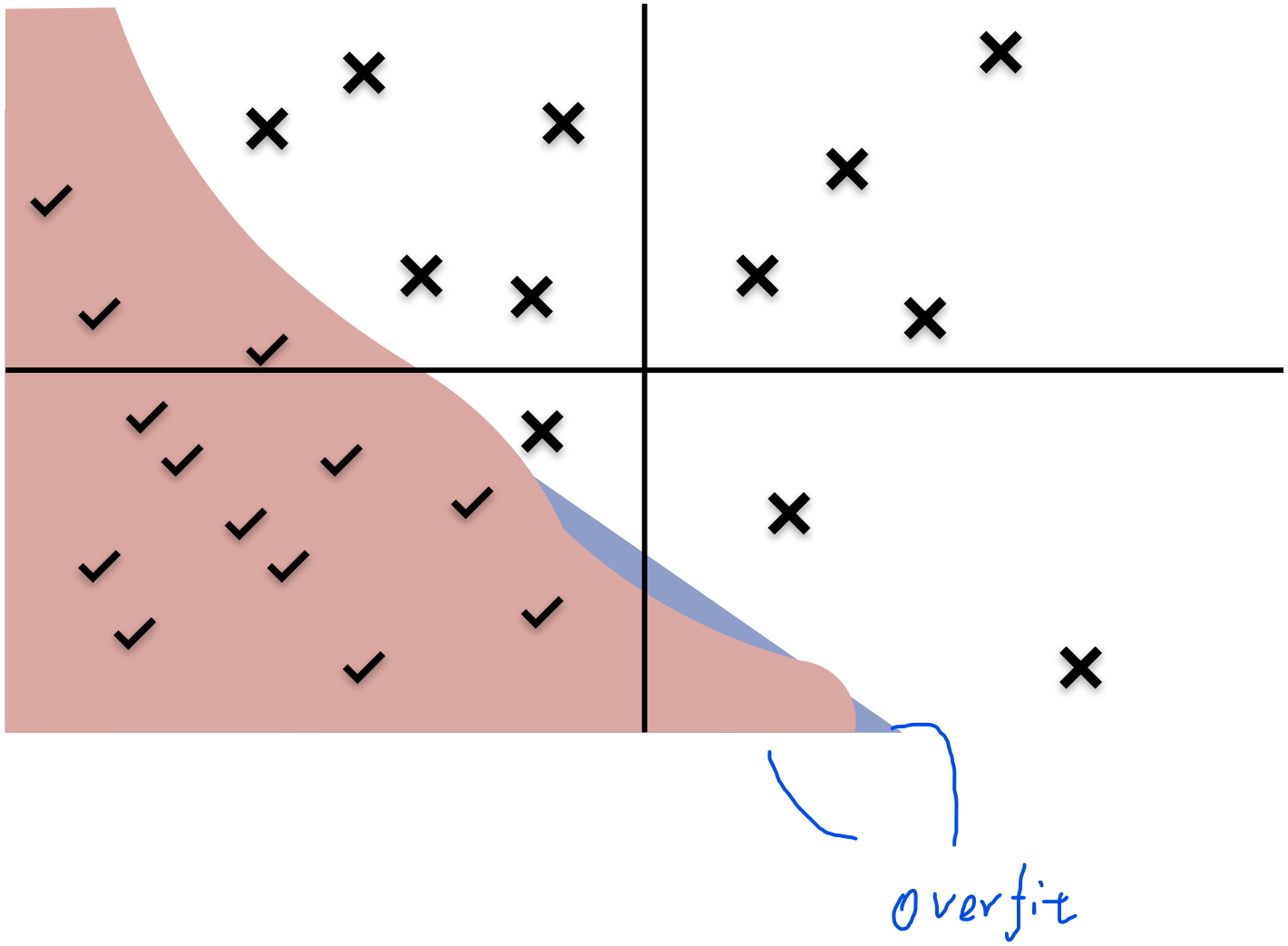
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Overfitting risk

Model



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Overfitting risk

Model

Original loss

$$-log(p(y|x))$$

Loss with regularization

$$-log(p(y|x)) + regularizer(W)$$

An extra cost that depends on the weights

Effect - push learning to favour smaller weights or more zero weights

越少0的 weight
更容易被 penalize

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Methods of regularization

Learning Method

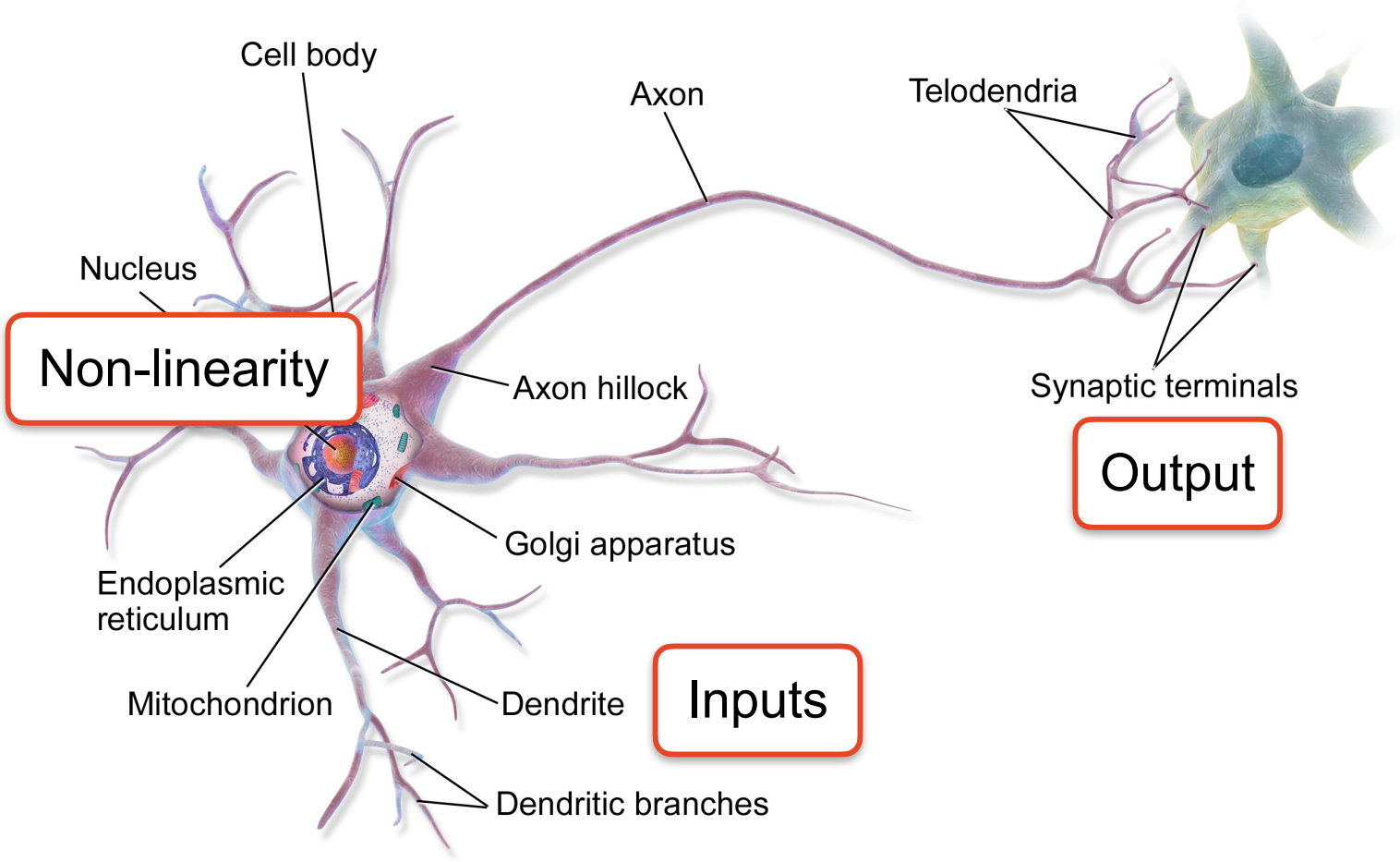
(1) Modify loss with cost, e.g, l1, l2
(2) Dropout, during training, randomly set some weights to zero
(3) Early stopping, don't backpropagate error further once it gets small

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

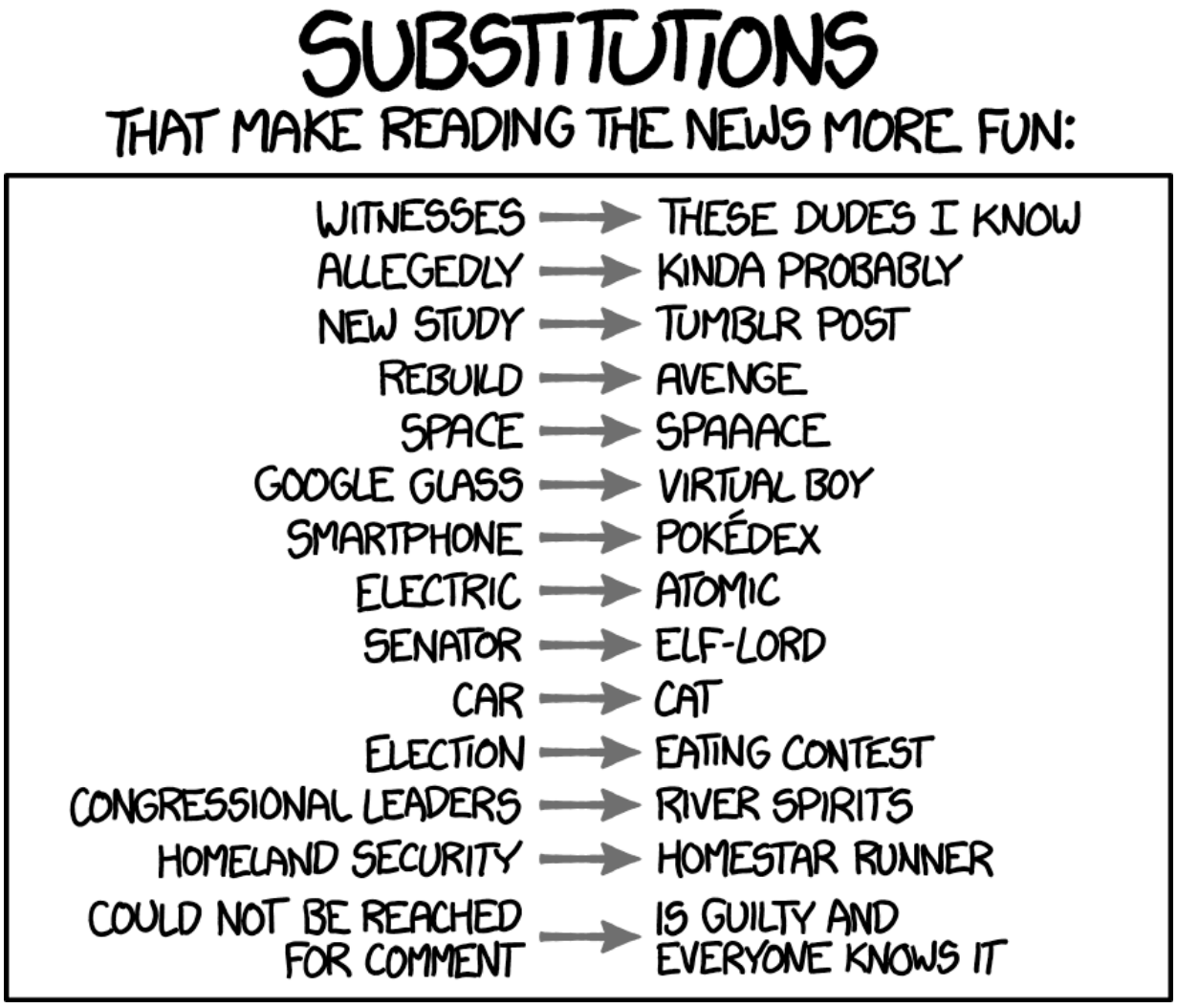
Why are they called ‘neural networks’?



https://commons.wikimedia.org/w/index.php?curid=28761830

3 minute Break - stretch and visit Menti

menti.com
1750 7815



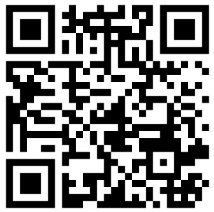
Substitutions

[INSIDE ELON MUSK'S NEW ATOMIC CAT]

Source: https://xkcd.com/1288/

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

Recurrent Models

Structure and concepts drawn from CS288 at Berkeley, which cites UT Austin and Stanford courses as sources

COMP 4446 / 5046
Lecture 3, 2025

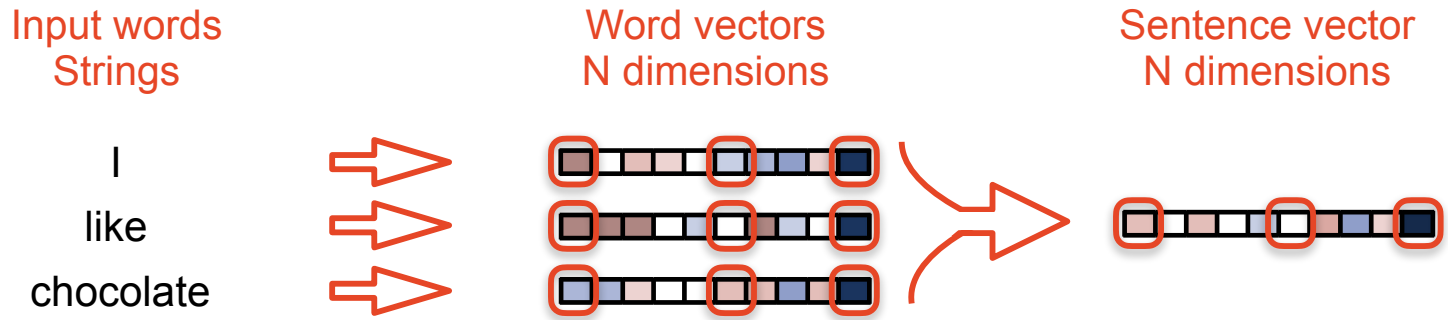
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How to handle variable length input

Option 1

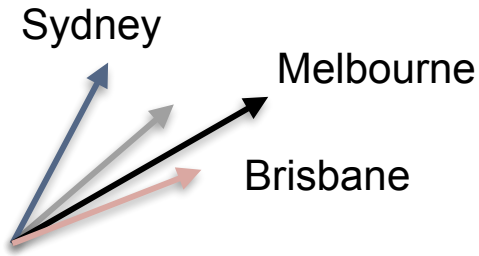
- Look up word vectors
- Take the mean



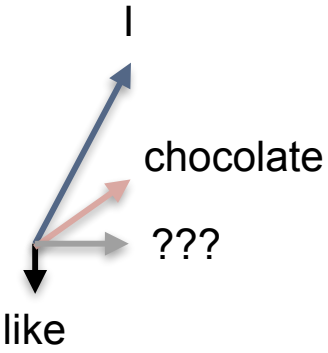
Problem: word properties are washed out / diluted

More broadly - why would this work?

Averaging city vectors will give something city-like in the middle of them

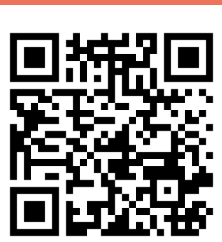


What does averaging all different words do?



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

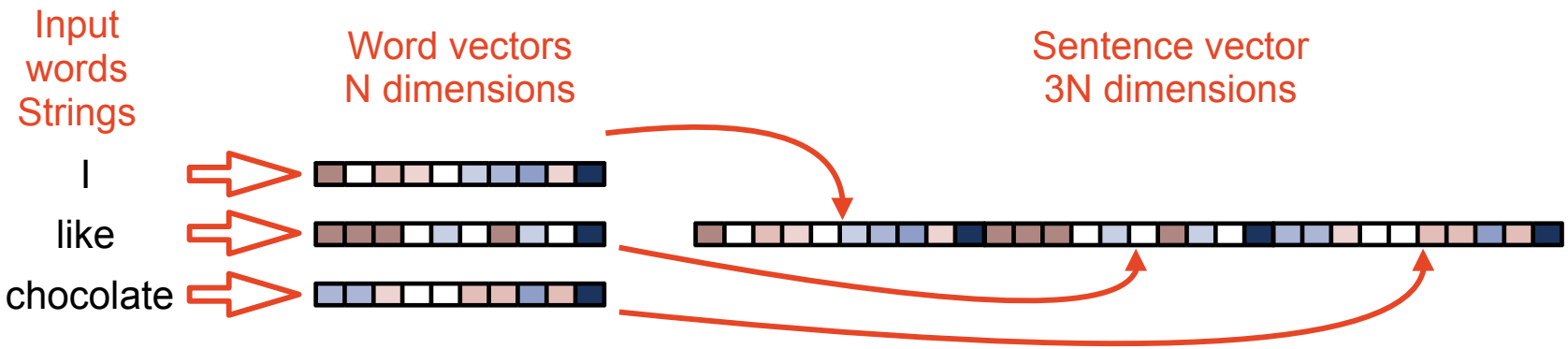


menti.com 1750 7815

How to handle variable length input

Option 2

- Look up word vectors
- Concatenate
- Pad and/or Truncate



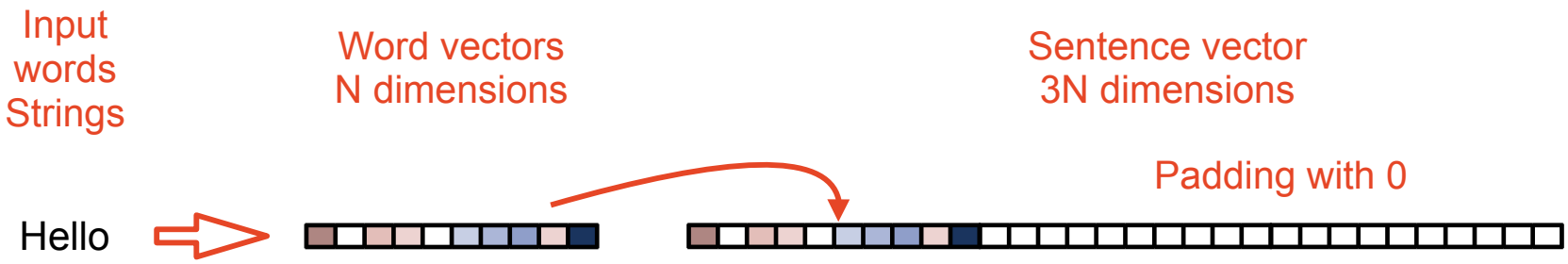
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How to handle variable length input

Option 2
- Look up word vectors
- Concatenate
- **Pad** and/or Truncate



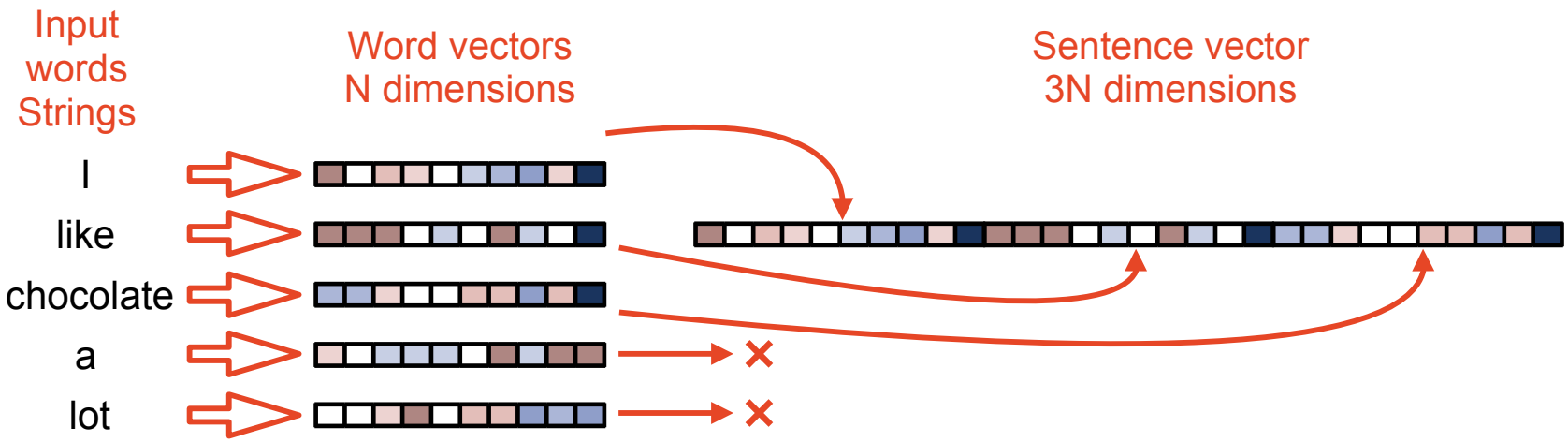
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How to handle variable length input

Option 2
- Look up word vectors
- Concatenate
- Pad and/or **Truncate**



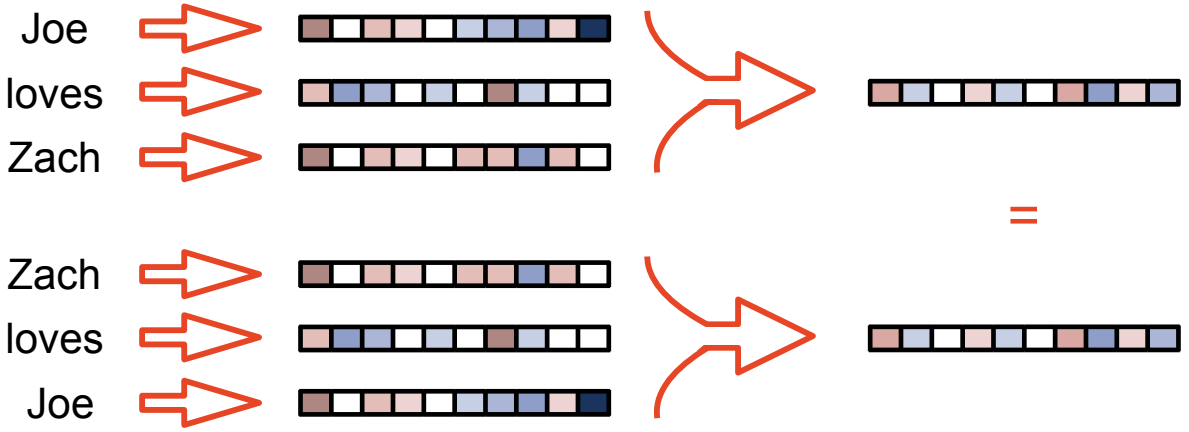
COMP 4446 / 5046
Lecture 3, 2025

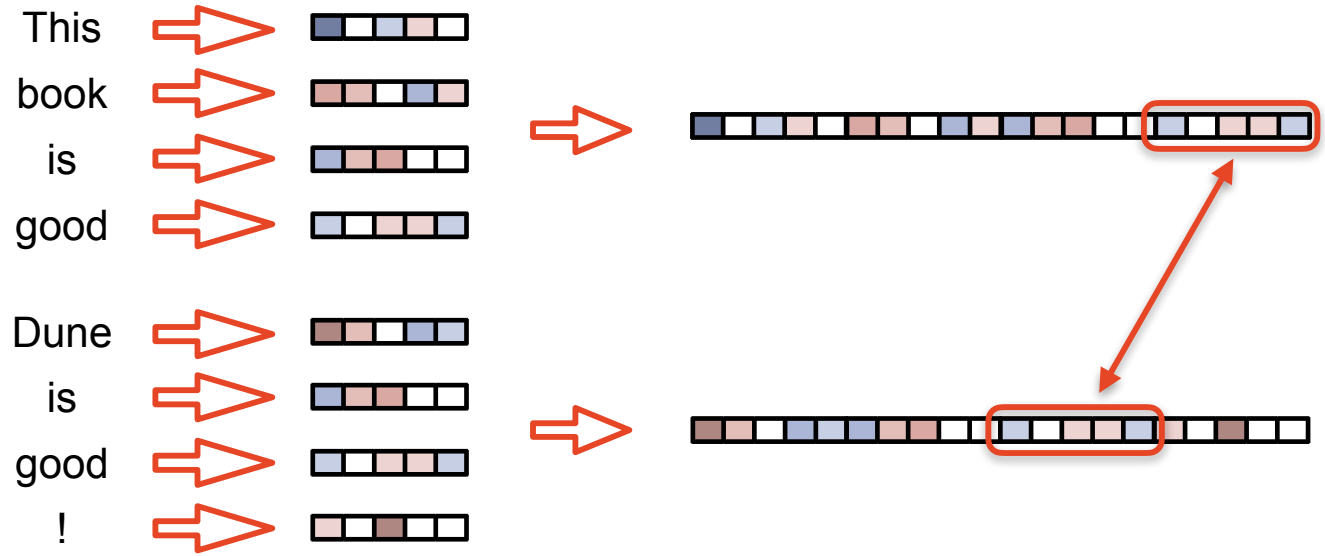
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Issues with these two approaches

→ RNN 可以解决



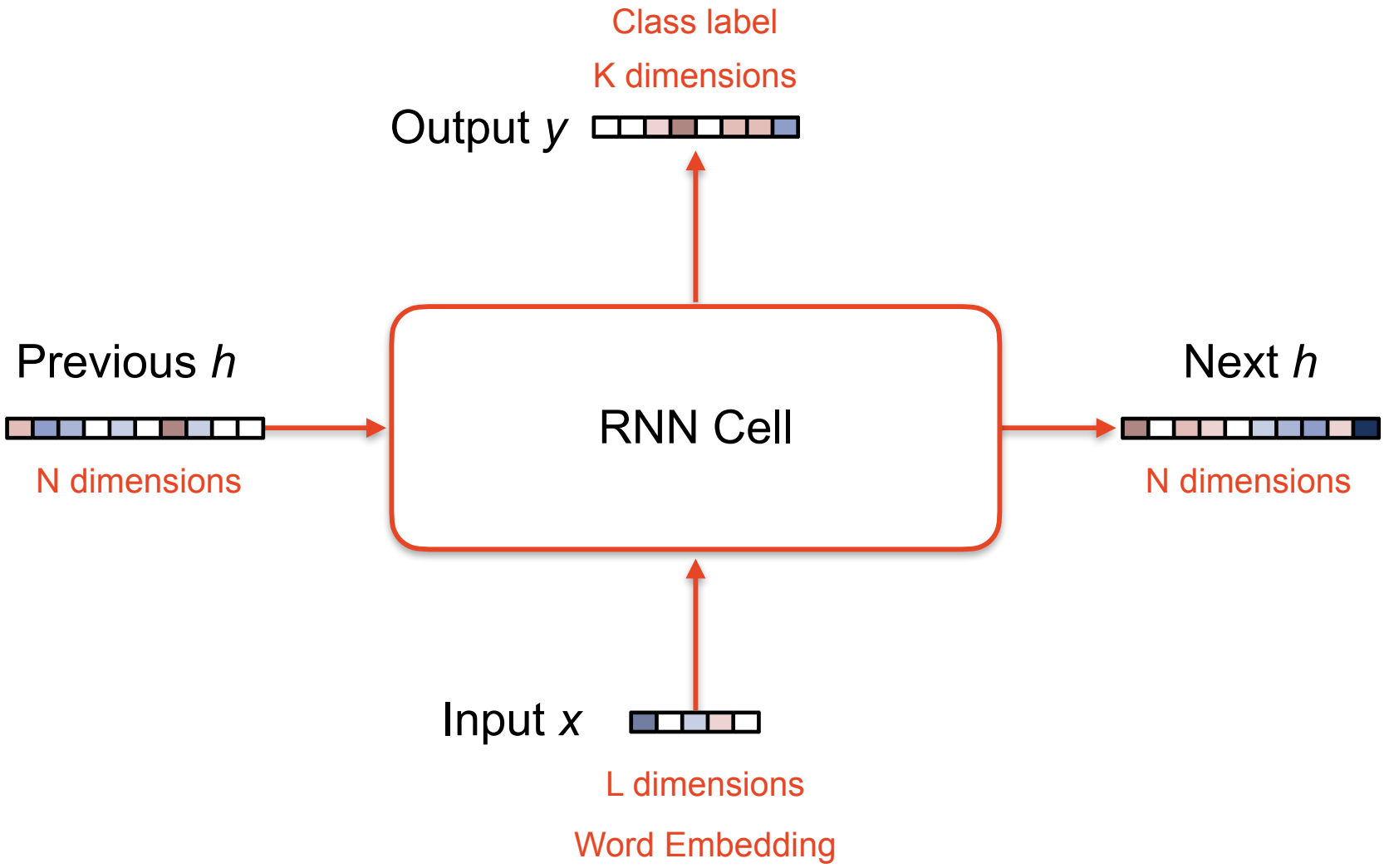


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Recurrent Neural Networks (RNNs) address this issue with context-dependent representations



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Simple RNN

Recurrent Neural Networks (RNNs) address this issue with context-dependent representations

Class label

K dimensions

Output y_t

h_{t-1}

N dimensions

Basic, Elman (1990) Cell:

$$\mathbf{h}_t = \tanh(W\,\mathbf{x}_t + V\,\mathbf{h}_{t-1} + \mathbf{b}_\mathbf{h})$$

$$\mathbf{y}_t = \tanh(U\,\mathbf{h}_t + \mathbf{b}_\mathbf{y})$$

h_t

N dimensions

Input x_t

L dimensions

Word Embedd

t for 'time' because the 1990 paper framed it that way. Note, this was building on Jordan (1986) which used a ReLU

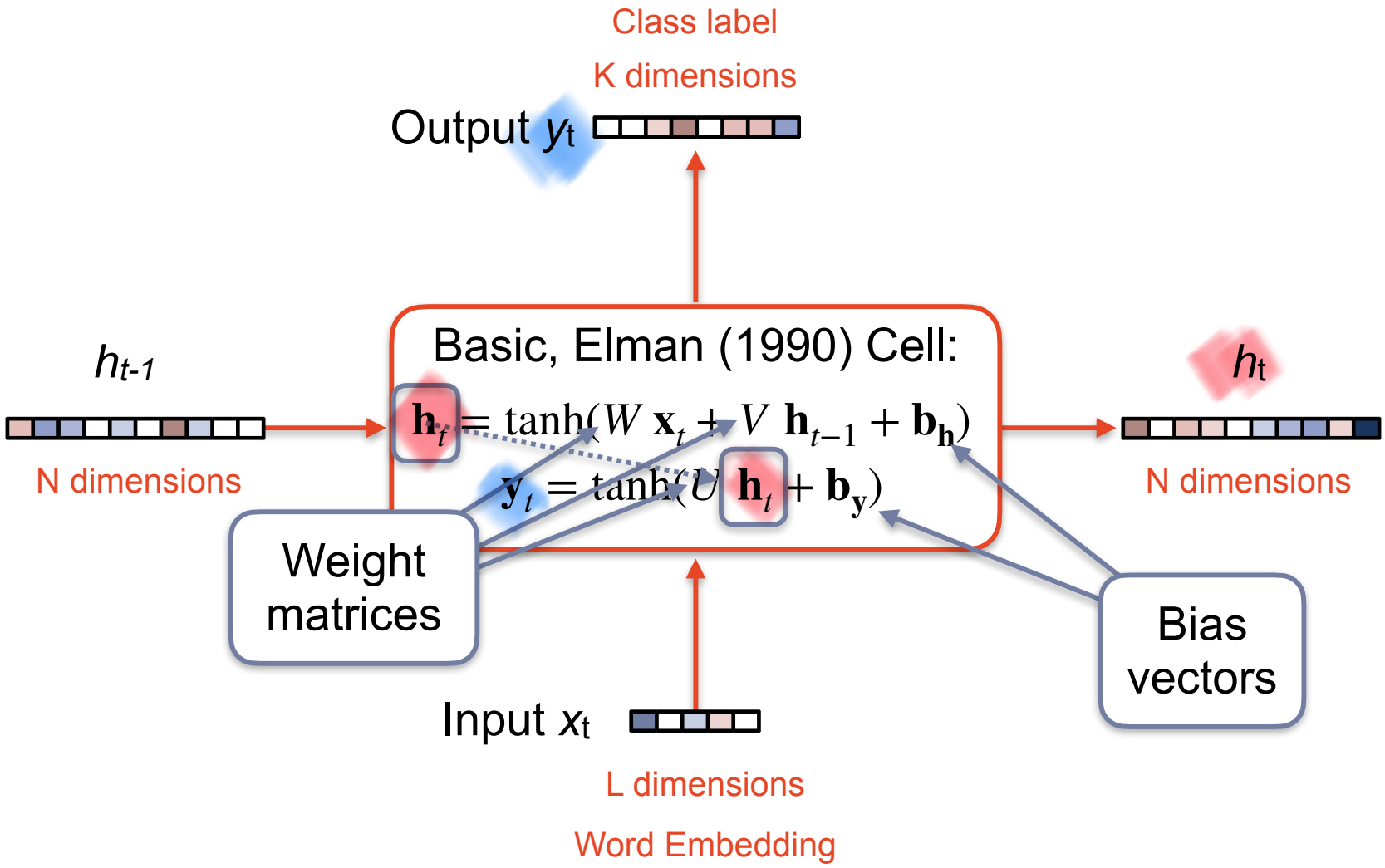
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Simple RNN

Recurrent Neural Networks (RNNs) address this issue with context-dependent representations



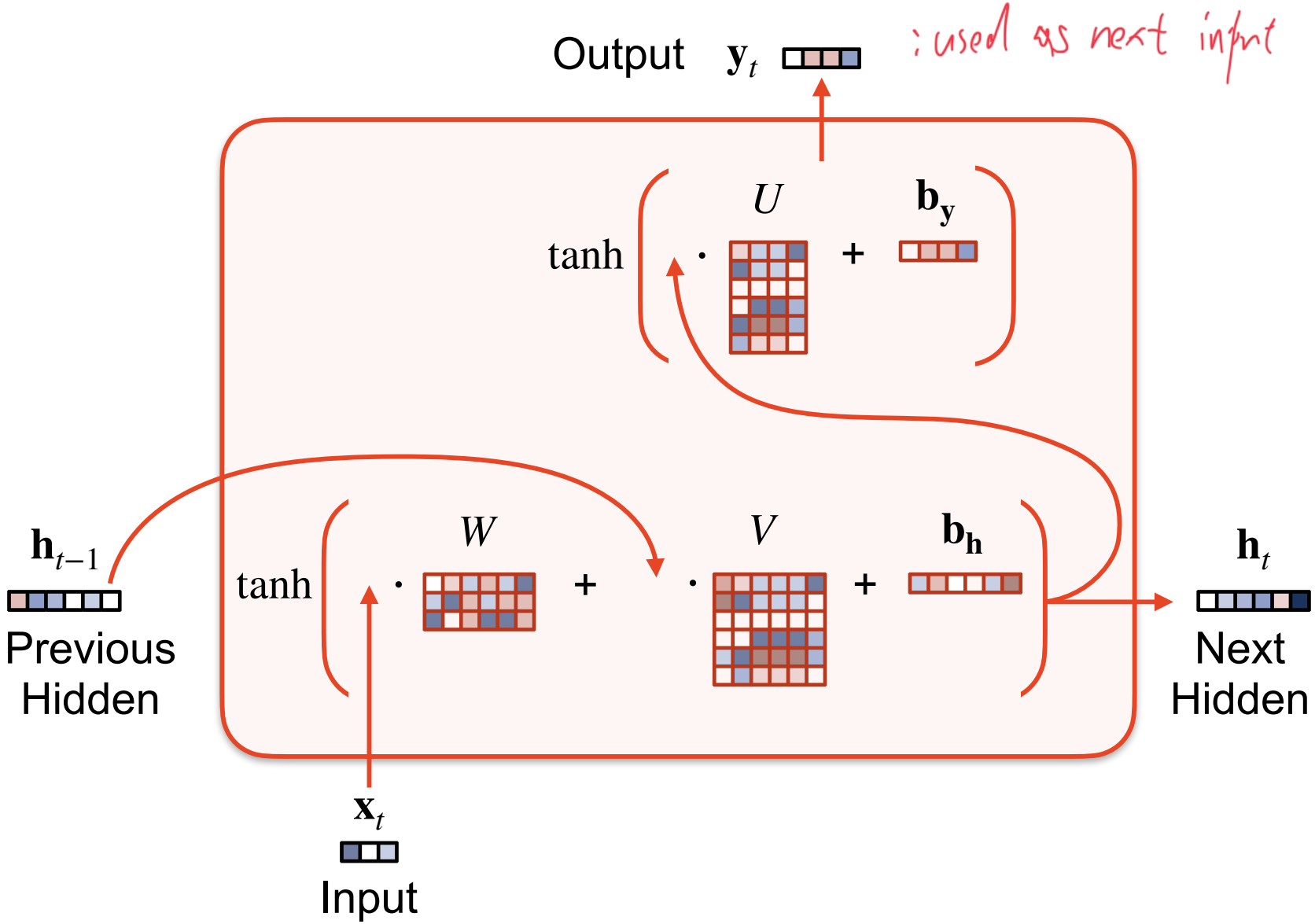
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Simple RNN

We can visualise these cells as a series of calculations



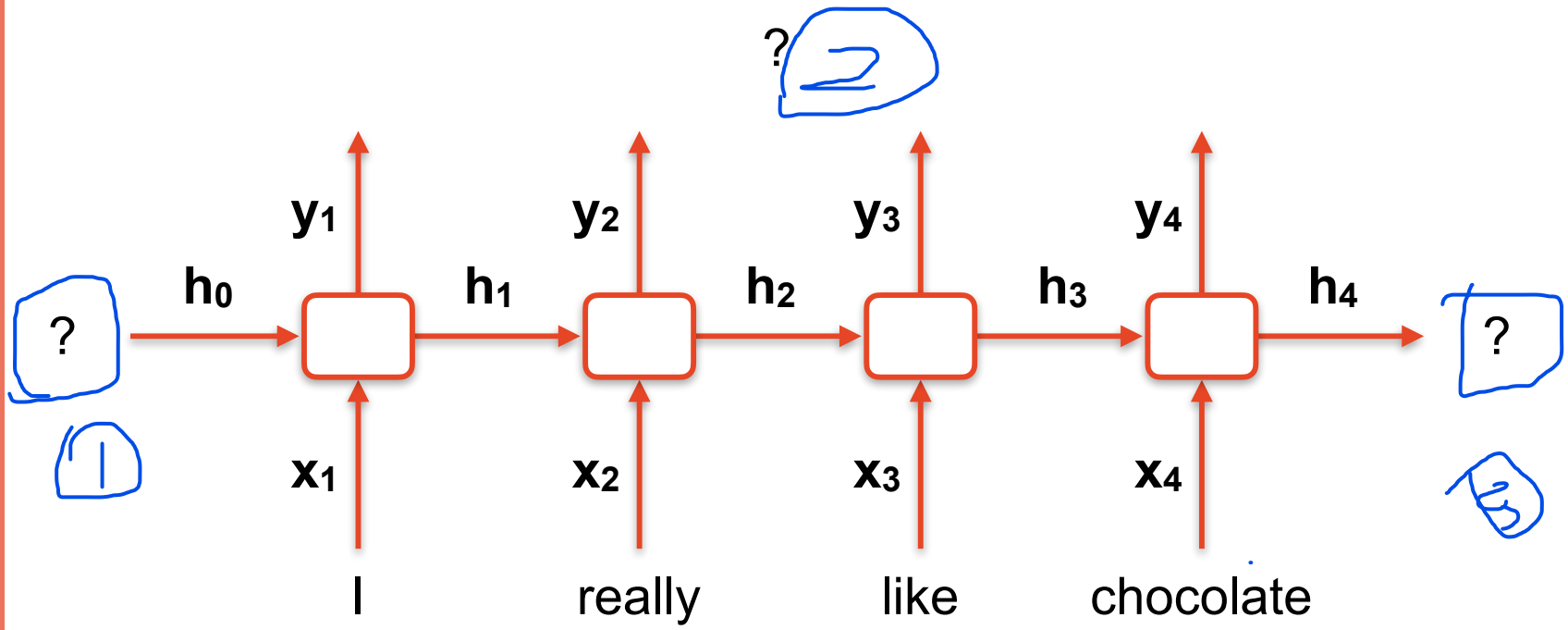
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

By putting a sequence of cells together, we process a multi-word input

一个个车前入词 再 Predict



COMP 4446 / 5046
Lecture 3, 2025

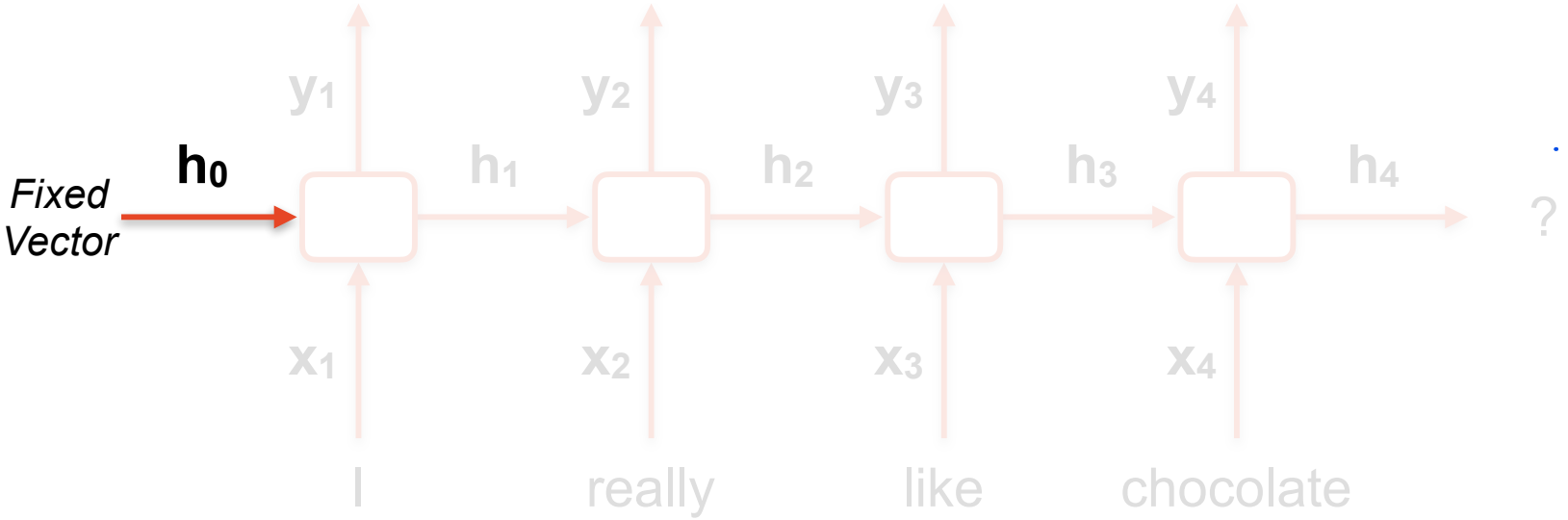
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

To initialise the sequence, use a fixed vector

Option 1: All zero

Option 2: Learn a set of weights

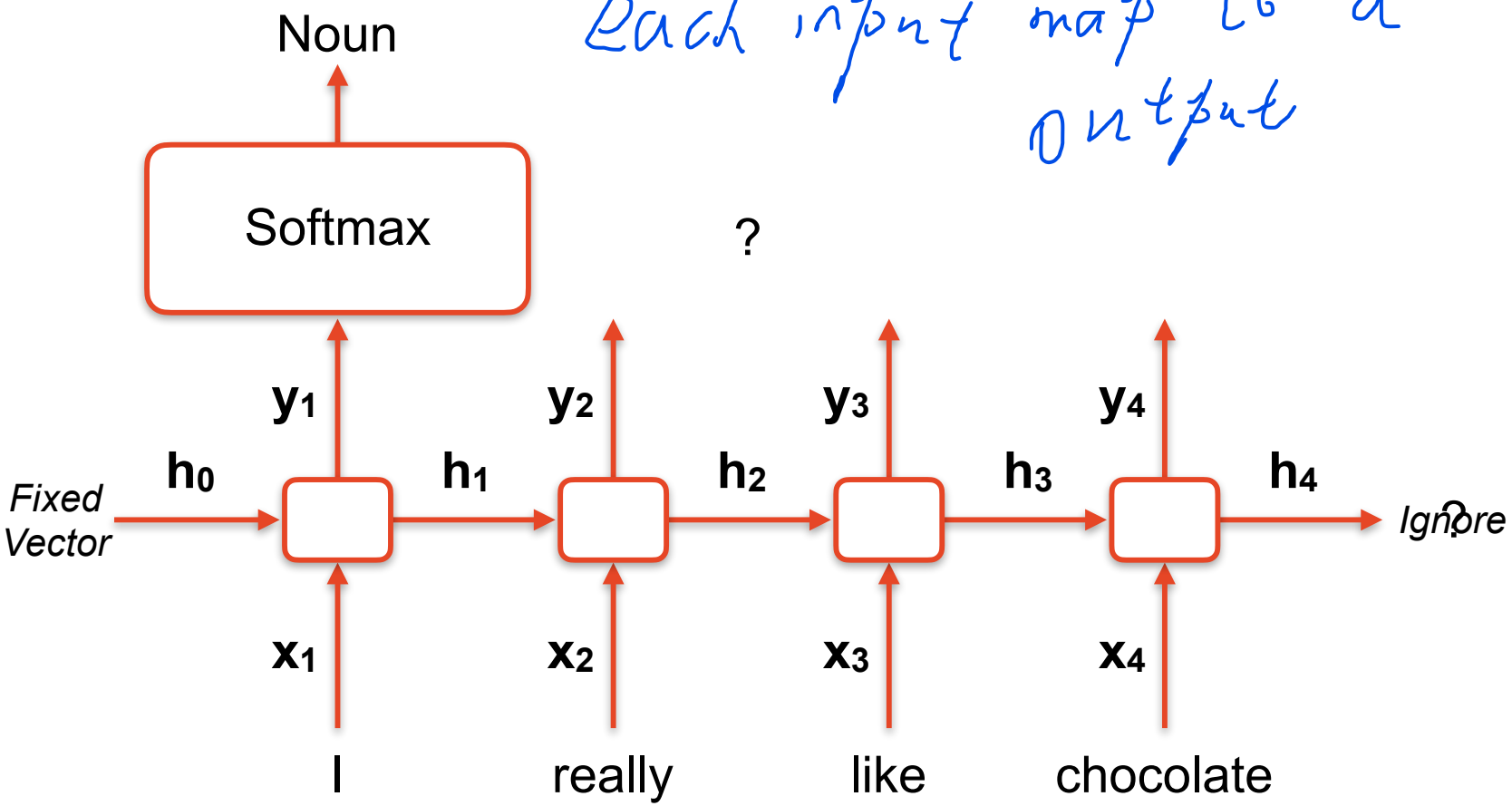


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

One way to use them is as a **transducer**

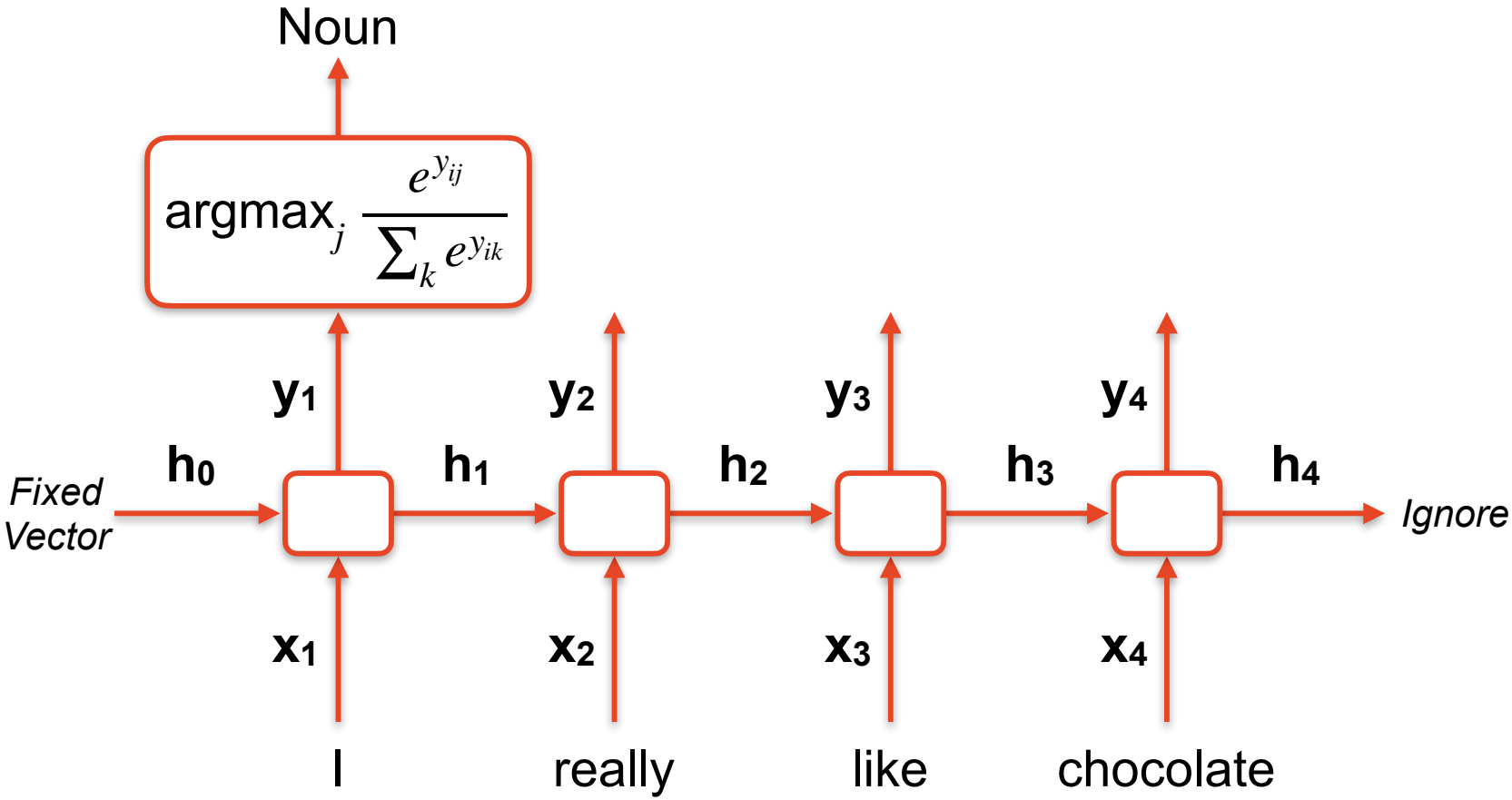


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

One way to use them is as a **transducer**

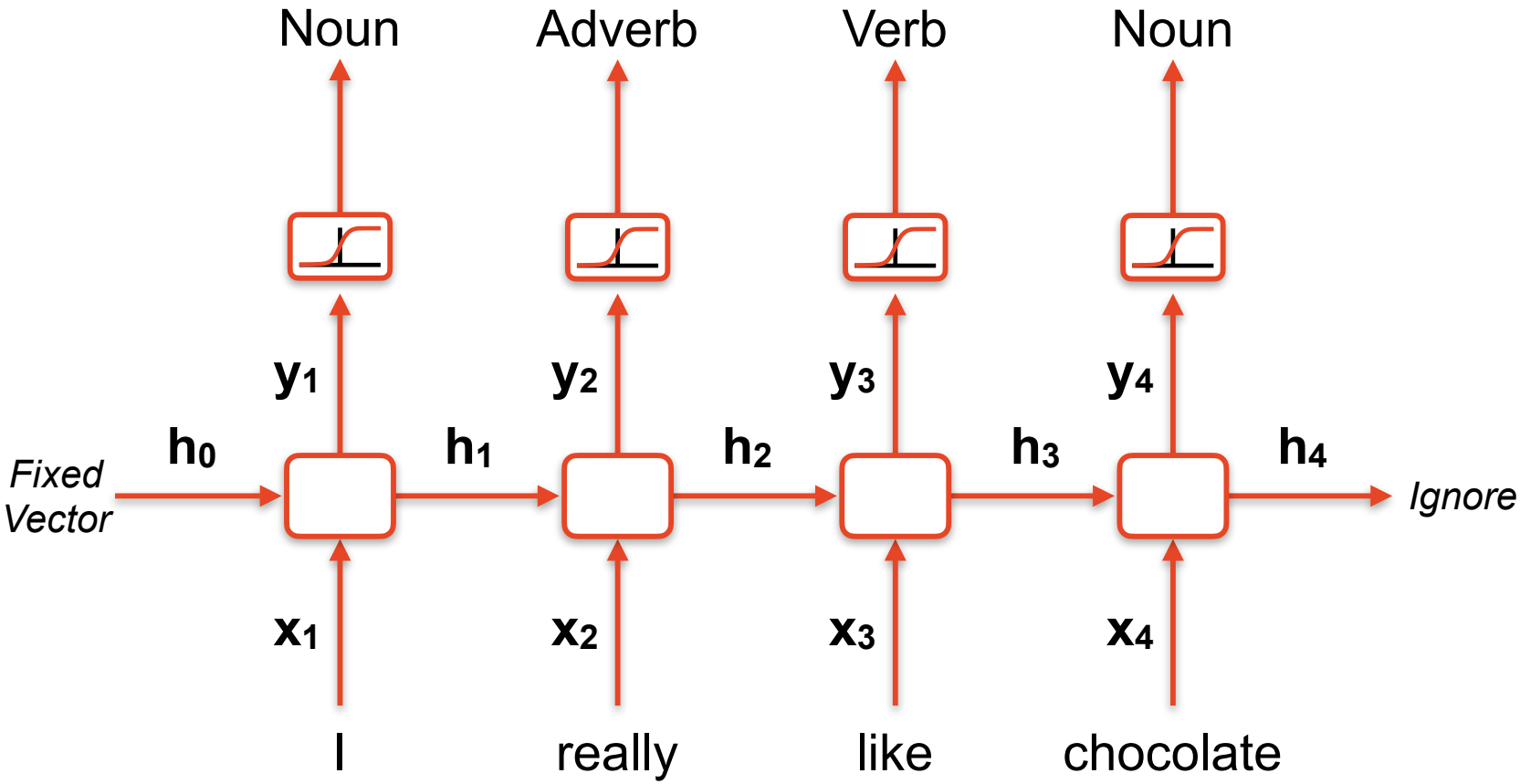


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

One way to use them is as a **transducer**

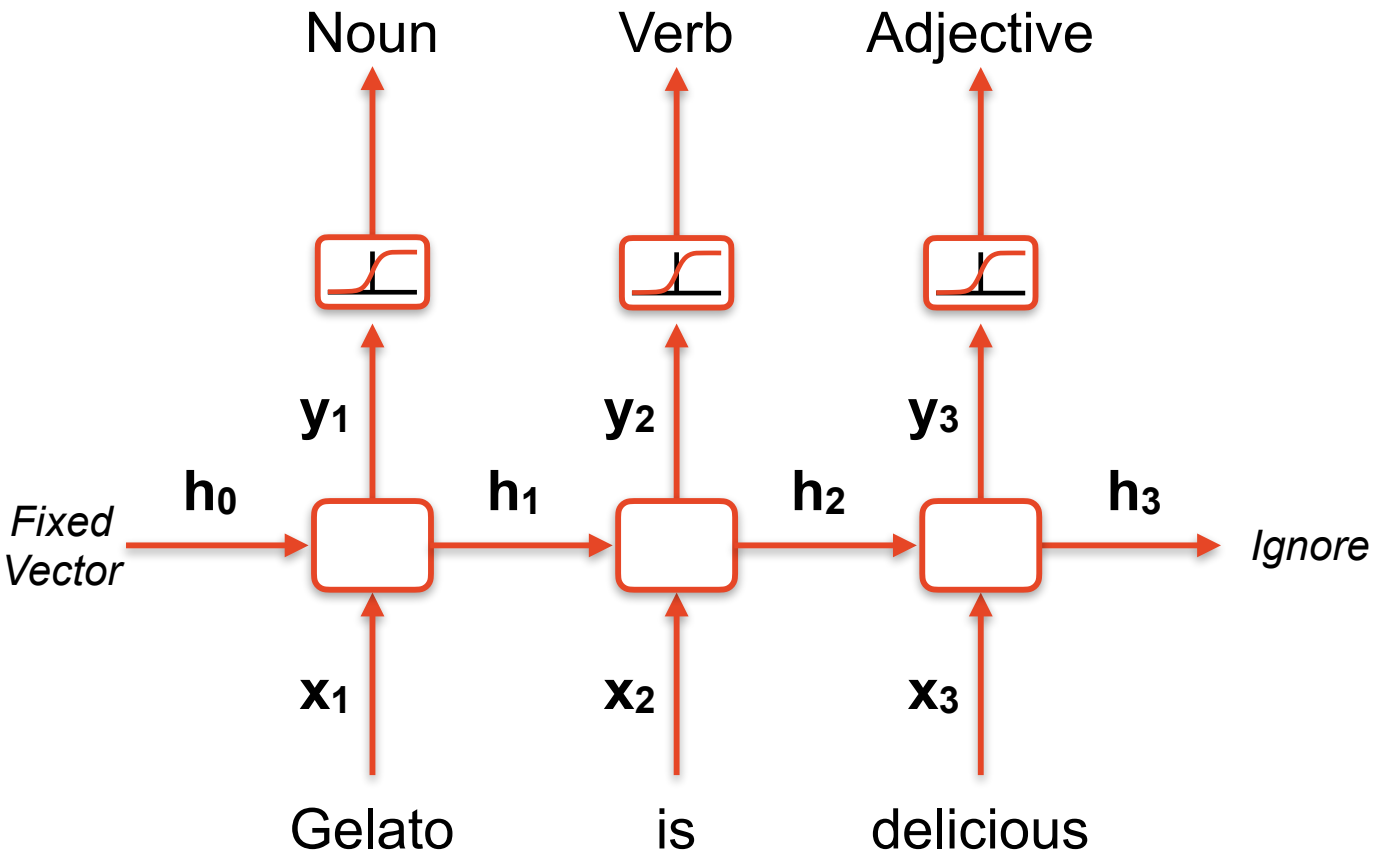


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

One way to use them is as a **transducer**



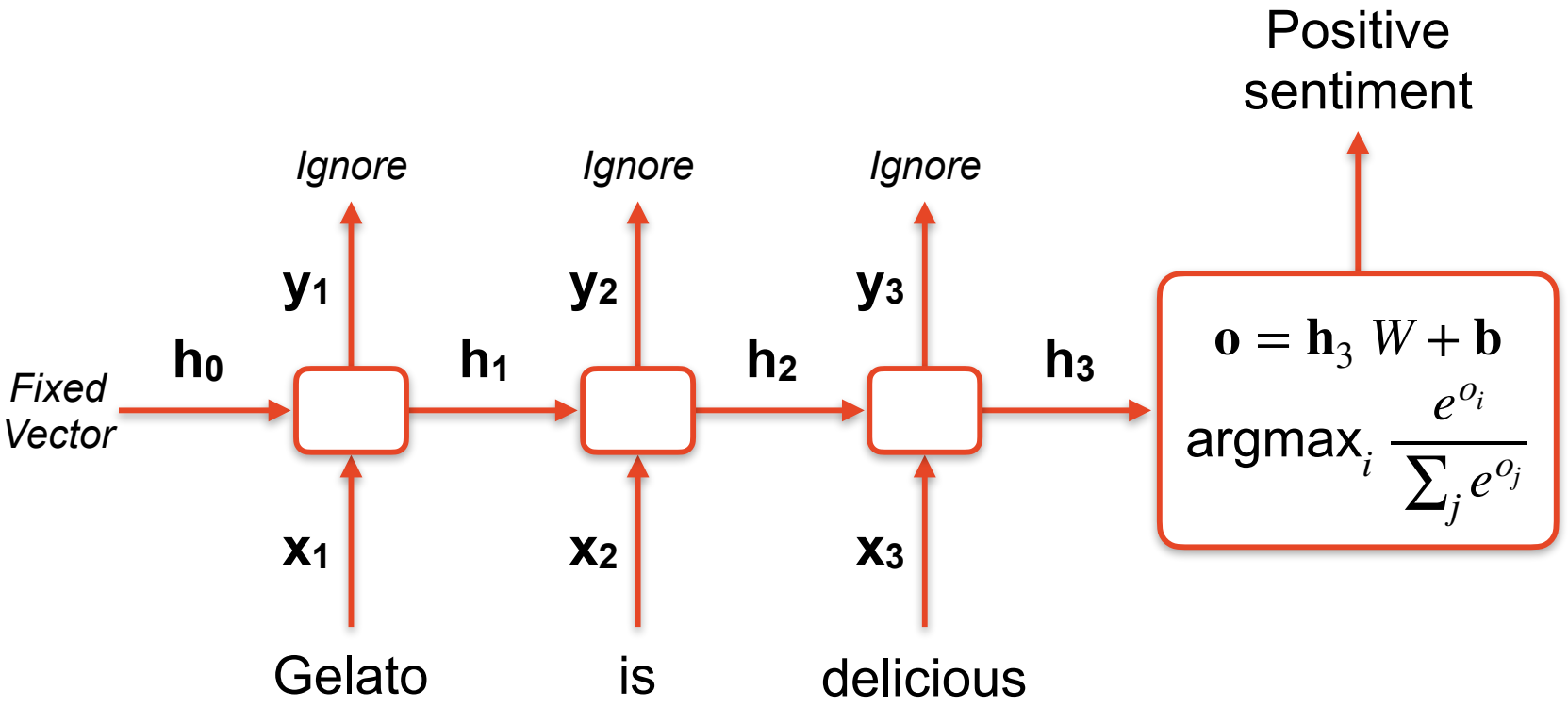
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Single output

Another use is as an **accepter** or **encoder**

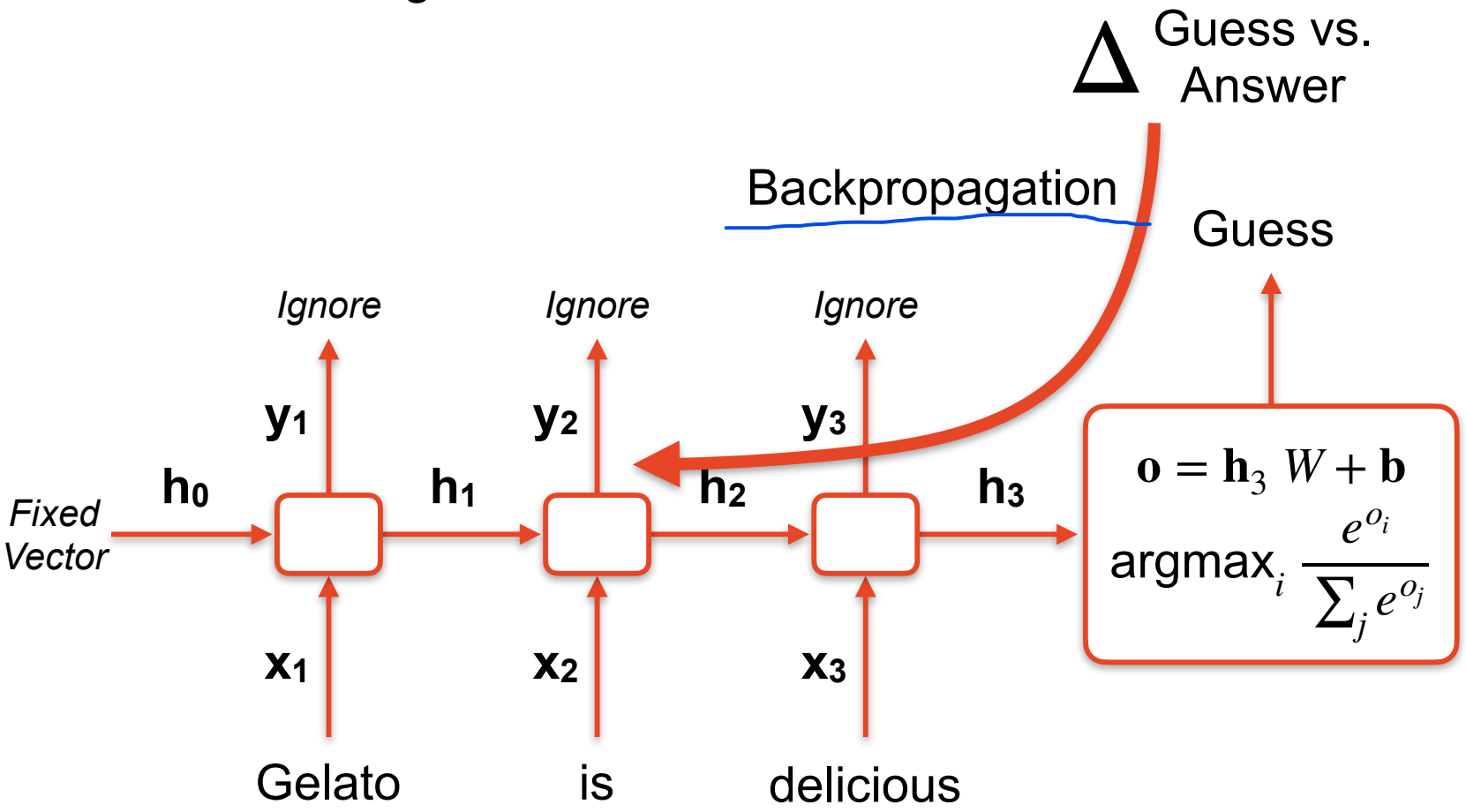


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How does learning work?

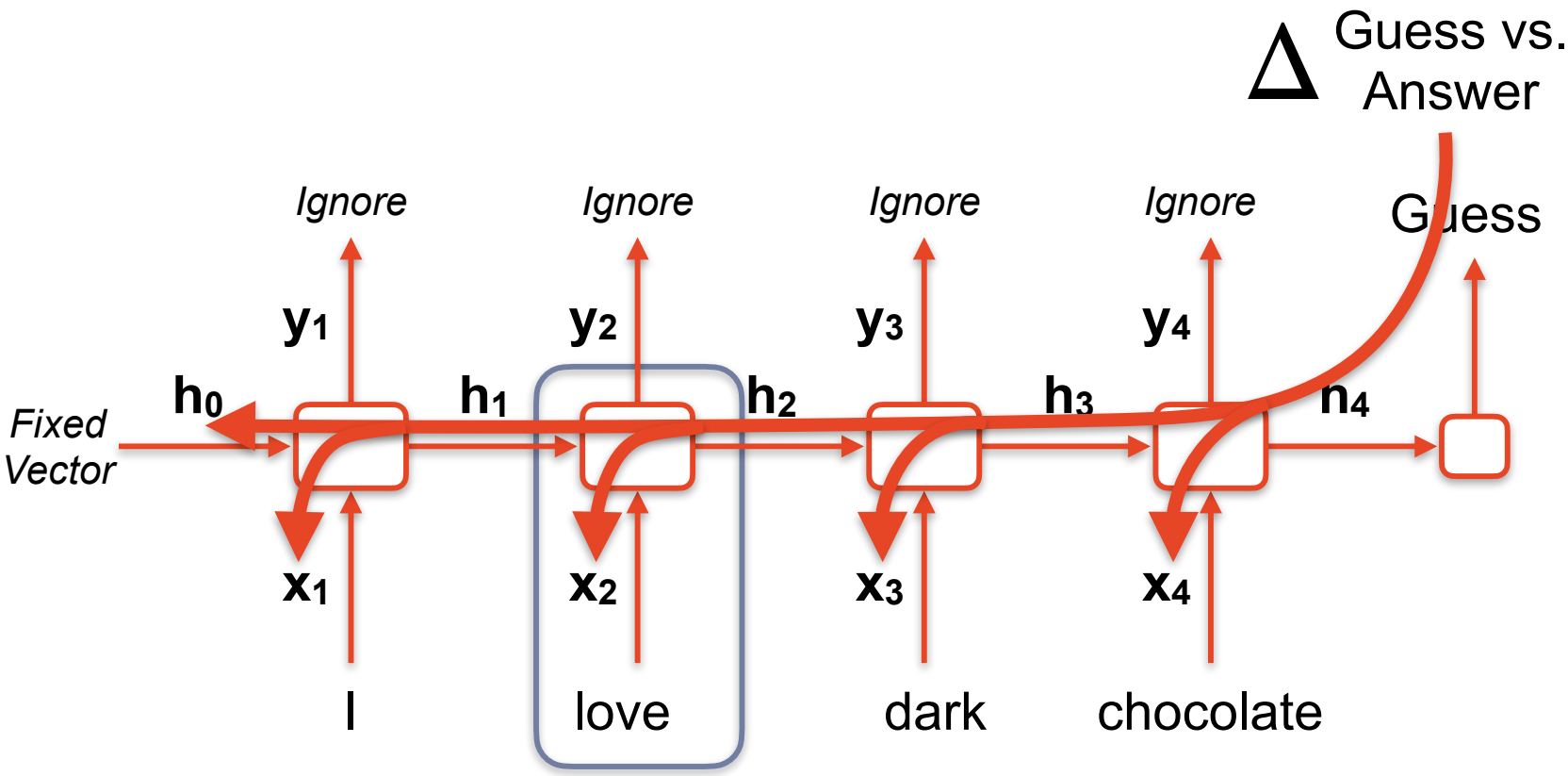


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How does learning work?



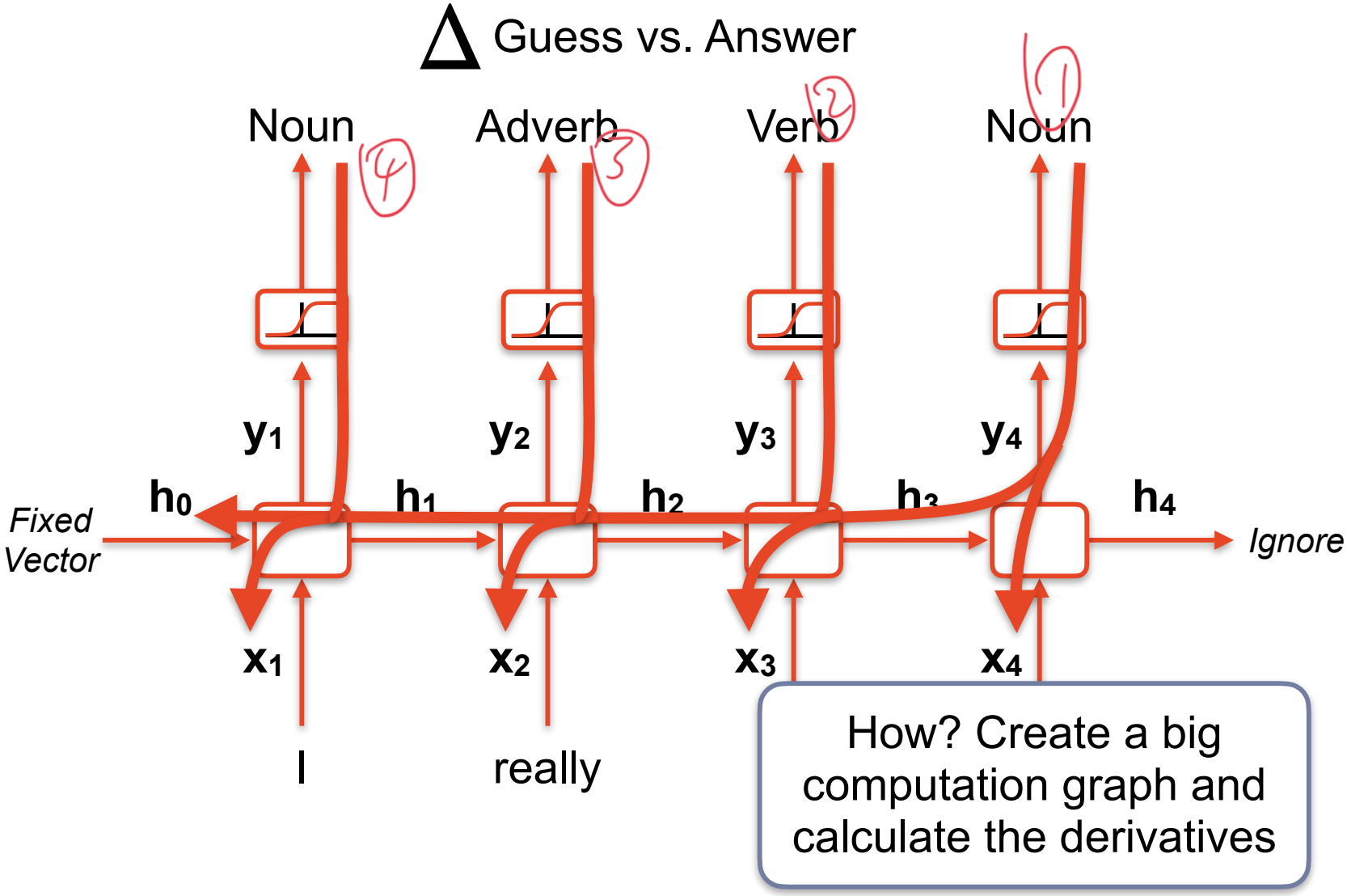
To fix the error, the derivative needs to update all the way back to to $\mathbf{x}_2$

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How does learning work?



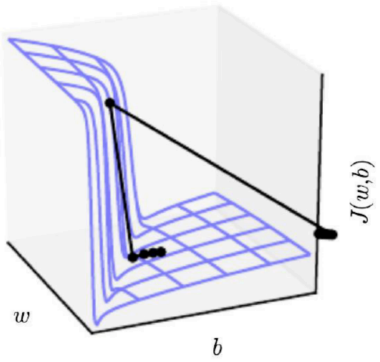
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Two key challenges in learning: (1) Exploding gradients

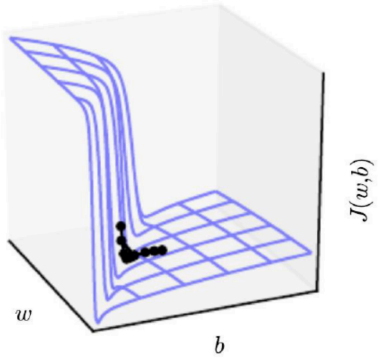
What happens if we have a very steep gradient?



- Might jump far away to a bad weight
- Might overflow the floating point number storage

Soln

Clipping the gradient, i.e., using min(gradient, k), solves this:



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

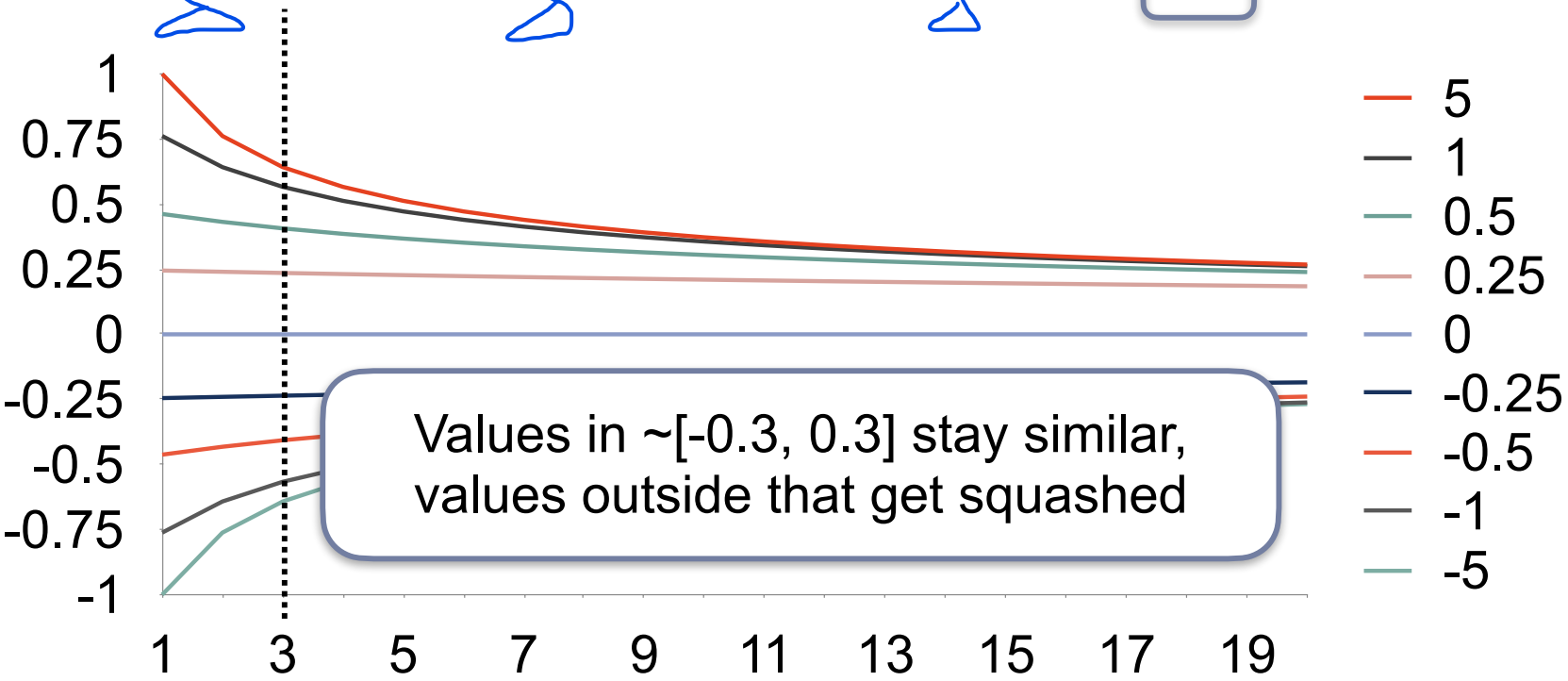
Two key challenges in learning: (2) Vanishing gradients

$$\mathbf{h}_t = \tanh(\mathbf{hb} + W\,\mathbf{x}_t + V\,\mathbf{h}_{t-1})$$

$$= \tanh(\mathbf{hb} + W\,\mathbf{x}_t + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-1} + V\,\mathbf{h}_{t-2})$$

$$= \tanh(\mathbf{hb} + W\,\mathbf{x}_t + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-1} + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-2} + V\,\mathbf{h}_{t-3})))$$

The values in this input go through tanh three times



Values in ~[-0.3, 0.3] stay similar, values outside that get squashed

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Two key challenges in learning: (2) Vanishing gradients

The bias is added every time

The values in this input is multiplied by W three times

$$= \tanh(\mathbf{hb} + W\,\mathbf{x}_t + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-1} + V\,\mathbf{h}_{t-2}))$$

$$= \tanh(\mathbf{hb} + W\,\mathbf{x}_t + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-1} + V\,\tanh(\mathbf{hb} + W\,\mathbf{x}_{t-2} + V\,\mathbf{h}_{t-3})))$$

New inputs are being added

Result? model forgets quickly / mainly captures local information

Small effect => Small gradient => Learning is hard

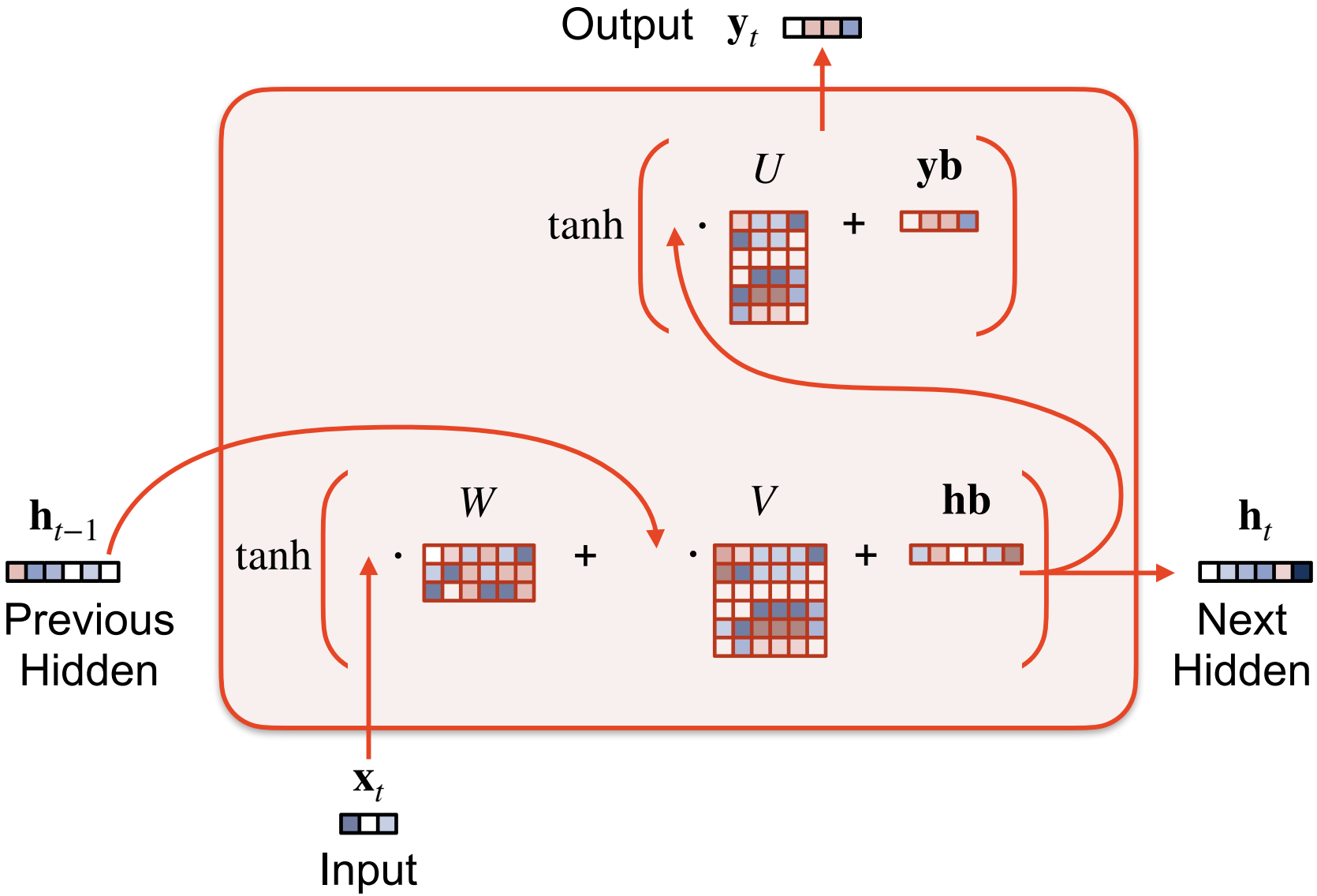
Vanishing Gradient description

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Key idea: maintain state in a non-decaying way

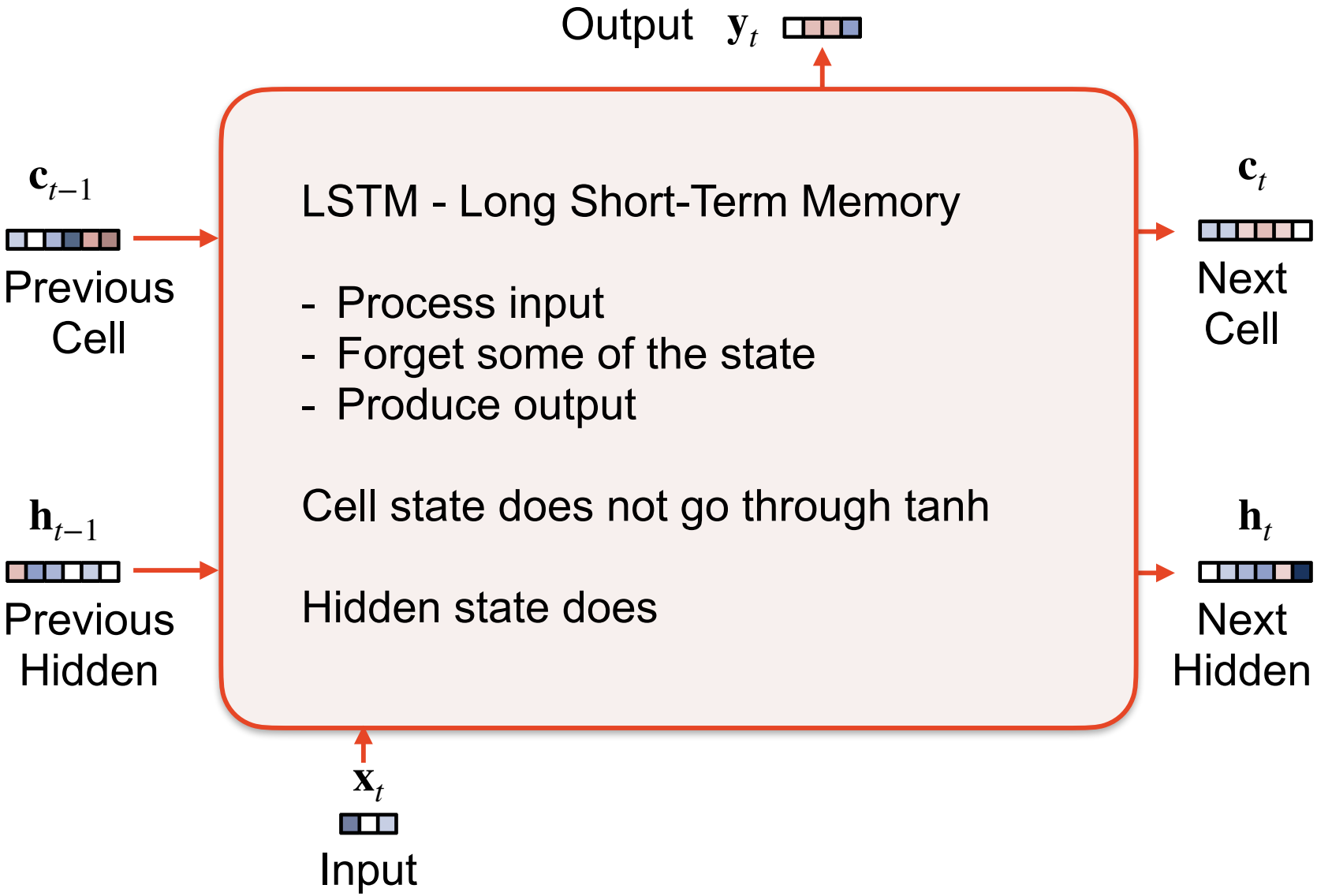


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Key idea: maintain state in a non-decaying way



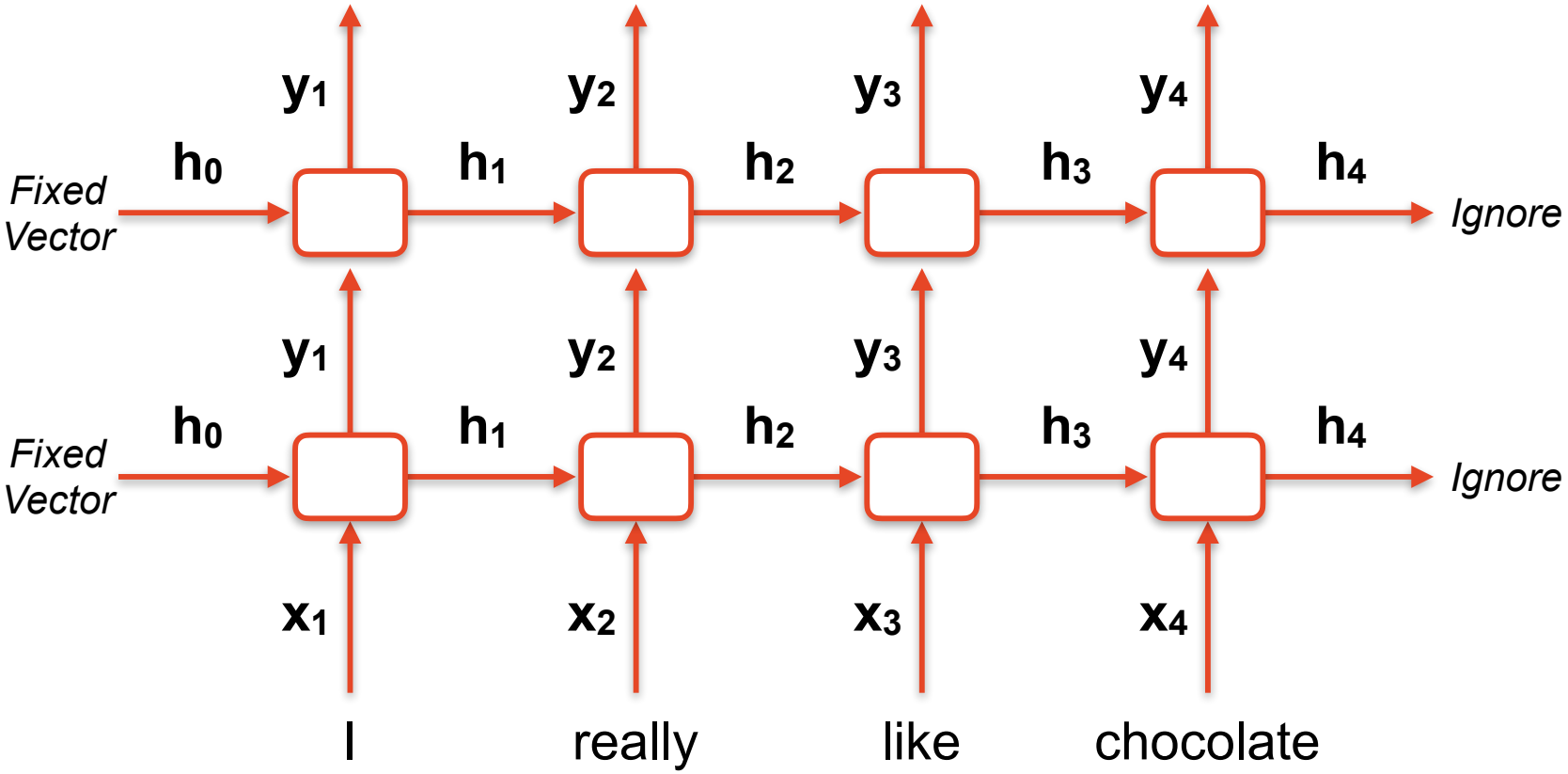
COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can use the output of one as the input to another

Stack RNN

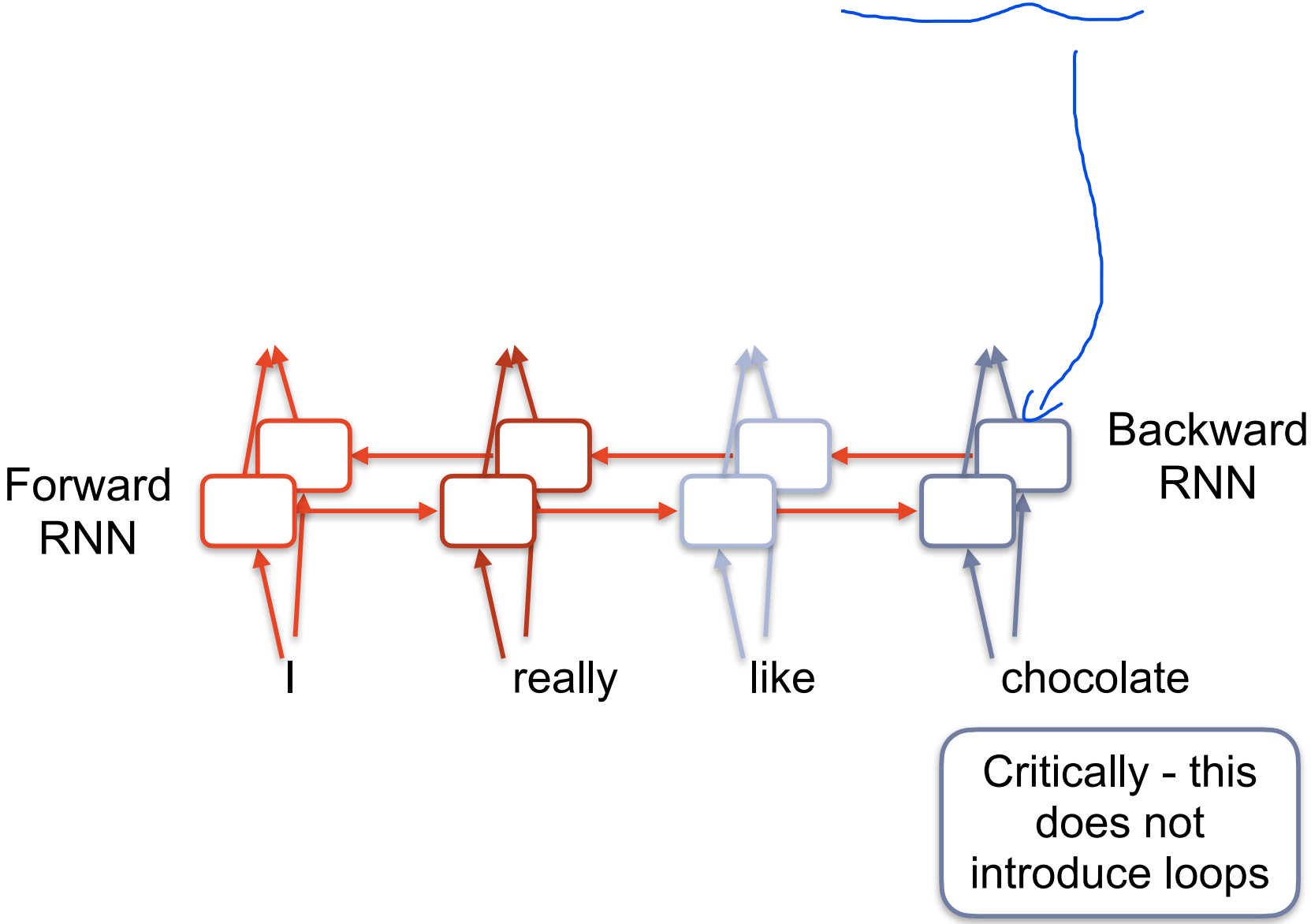


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can have two processing the input in reversed directions



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Part of Speech tagging

Noun	Adverb	Verb	Noun
I	really	like	chocolate

Categories of words, defined by:

Morphological context - combine with similar affixes (ie., prefixes and suffixes)

Verbs

	-s	-ed	-ing
walk	walks	walked	walking
slice	slices	sliced	slicing
believe	believes	believed	believing

David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Part of Speech tagging

Noun	Adverb	Verb	Noun
I	really	like	chocolate

Categories of words, defined by:

Morphological context - combine with similar affixes (ie., prefixes and suffixes)

		-s	-ed	-ing
Verbs	walk	walks	walked	walking
	slice	slices	sliced	slicing
	believe	believes	believed	
Not Verbs	of	*ofs	*ofed	
	red	*reds	*redded	

Exceptions:
Sleep
Eat
Give

Problem

David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

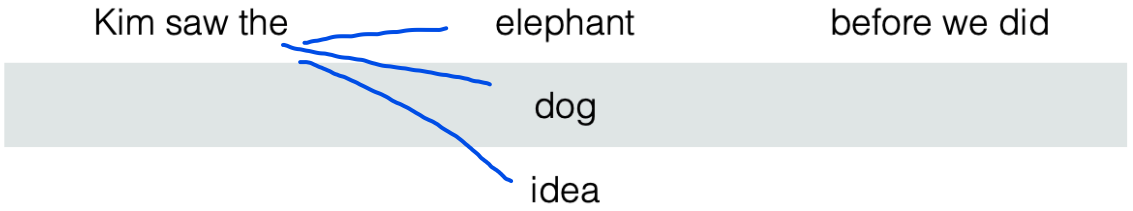
menti.com 1750 7815

Part of Speech tagging

Noun	Adverb	Verb	Noun
I	really	like	chocolate

Categories of words, defined by:

Syntactic distribution - can be replaced with another word and remain grammatical



David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Part of Speech tagging

Noun	Adverb	Verb	Noun
I	really	like	chocolate

Categories of words, defined by:

Syntactic distribution - can be replaced with another word and remain grammatical

Kim saw the	elephant	before we did
	dog	
	idea	
	*of	
	*goes	

May be too strict, e.g., "Jonathan" cannot go here

David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Part of Speech tagging can help distinguish meanings



NN	NN	VB	DT	NN
Time	flies	like	an	arrow


NN	VBZ	IN	DT	NN
Time	flies	like	an	arrow

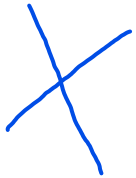
David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Part of Speech tagging can help indicate pronunciation



Noun	Verb
My conduct is great	I conduct myself well
She won the contest	I contest the ticket
He is my escort	He escorted me
That is an insult	Don't insult me
Rebel without a cause	He likes to rebel
He is a suspect	I suspect him

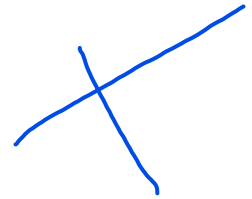
David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

How well do current models do?



English News

- Rule based (guess most frequent label): 92%
- Machine learning: 97%

But!

- There are ~20 words per sentence…
  sentence accuracy is 55%
- Drops on other text types (out of domain)

David Bamman's Info 159/259 lectures at Berkeley

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Where can I find data?

Universal Dependencies

Abaza	1	<1K		Northwest Caucasian
Afrikaans	1	49K		IE, Germanic
Akkadian	2	25K		Afro-Asiatic, Semitic
Akuntsu	1	1K		Tupian, Tupari
Albanian	1	<1K	W	IE, Albanian
Amharic	1	10K		Afro-Asiatic, Semitic
Ancient Greek	3	456K		IE, Greek
Ancient Hebrew	1	39K		Afro-Asiatic, Semitic
Apurina	1	<1K		Arawakan
Arabic	3	1,042K	W	Afro-Asiatic, Semitic
Armenian	2	94K	W	IE, Armenian
Assyrian	1	<1K		Afro-Asiatic, Semitic
Bambara	1	13K		Mande
Basque	1	121K		Basque
Beja	1	1K		Afro-Asiatic, Cushitic
Belarusian	1	305K	W	IE, Slavic
Bengali	1	<1K		IE, Indic
Bhojpuri	1	6K		IE, Indic
Bororo	1	1K		Bororoan
Breton	1	10K	W	IE, Celtic
Bulgarian	1	156K		IE, Slavic
Buryat	1	10K		Mongolic
Cantonese	1	13K		Sino-Tibetan
Catalan	1	553K		IE, Romance
Cebuano	1	1K		Austronesian, Central Philippine
Chinese	7	309K	W	Sino-Tibetan
Chukchi	1	6K		Chukotko-Kamchatkan
Classical Armenian	1	13K		IE, Armenian
Classical Chinese	1	433K		Sino-Tibetan
Coptic	1	57K		Afro-Asiatic, Egyptian
Croatian	1	199K	W	IE, Slavic
Czech	6	2,253K	W	IE, Slavic
Danish	1	100K		IE, Germanic
Dutch	2	306K	W	IE, Germanic
English	10	726K	W	IE, Germanic
Erzya	1	20K		Uralic, Mordvin

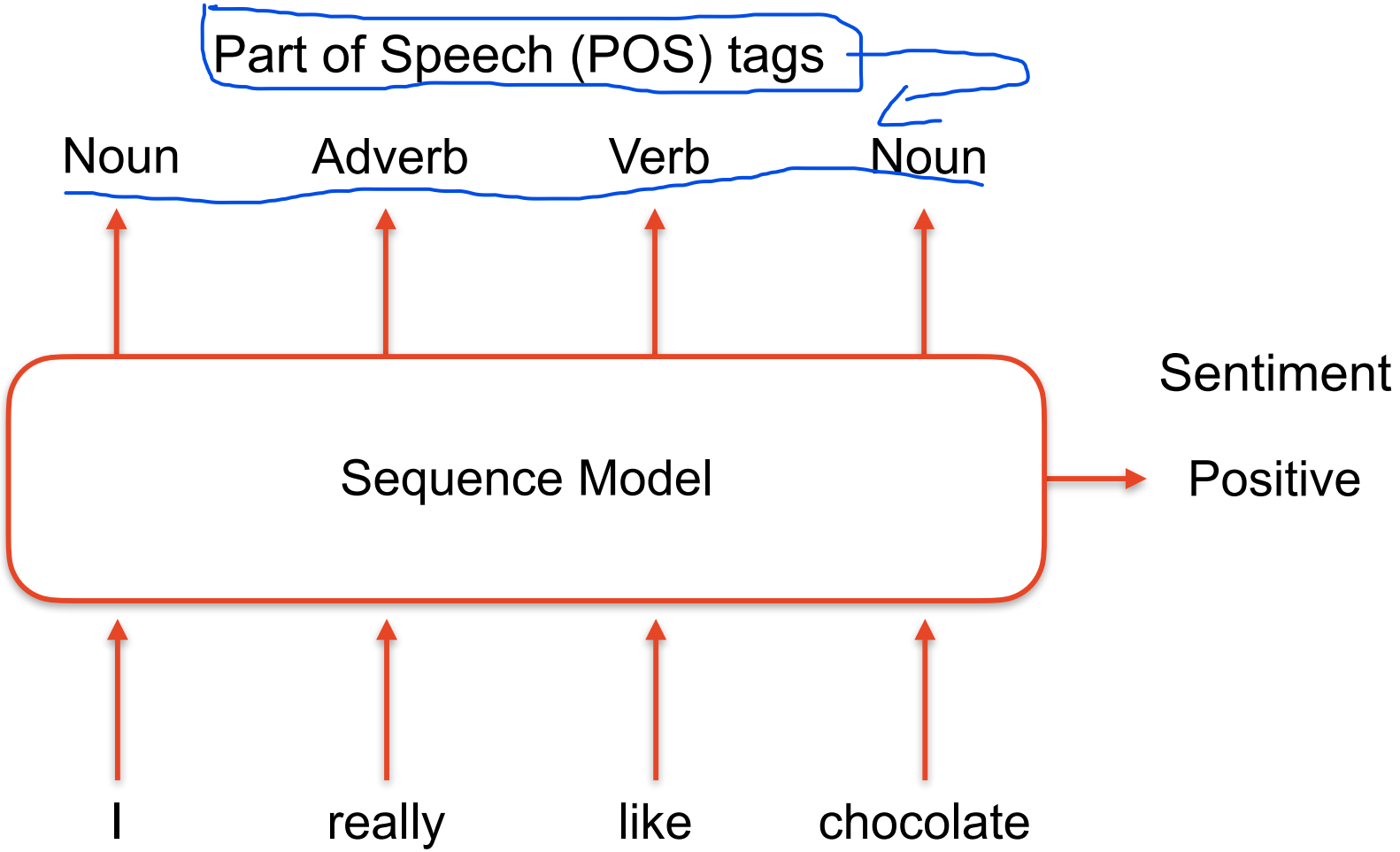
https://universaldependencies.org/

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can do multiple tasks simultaneously with a single model

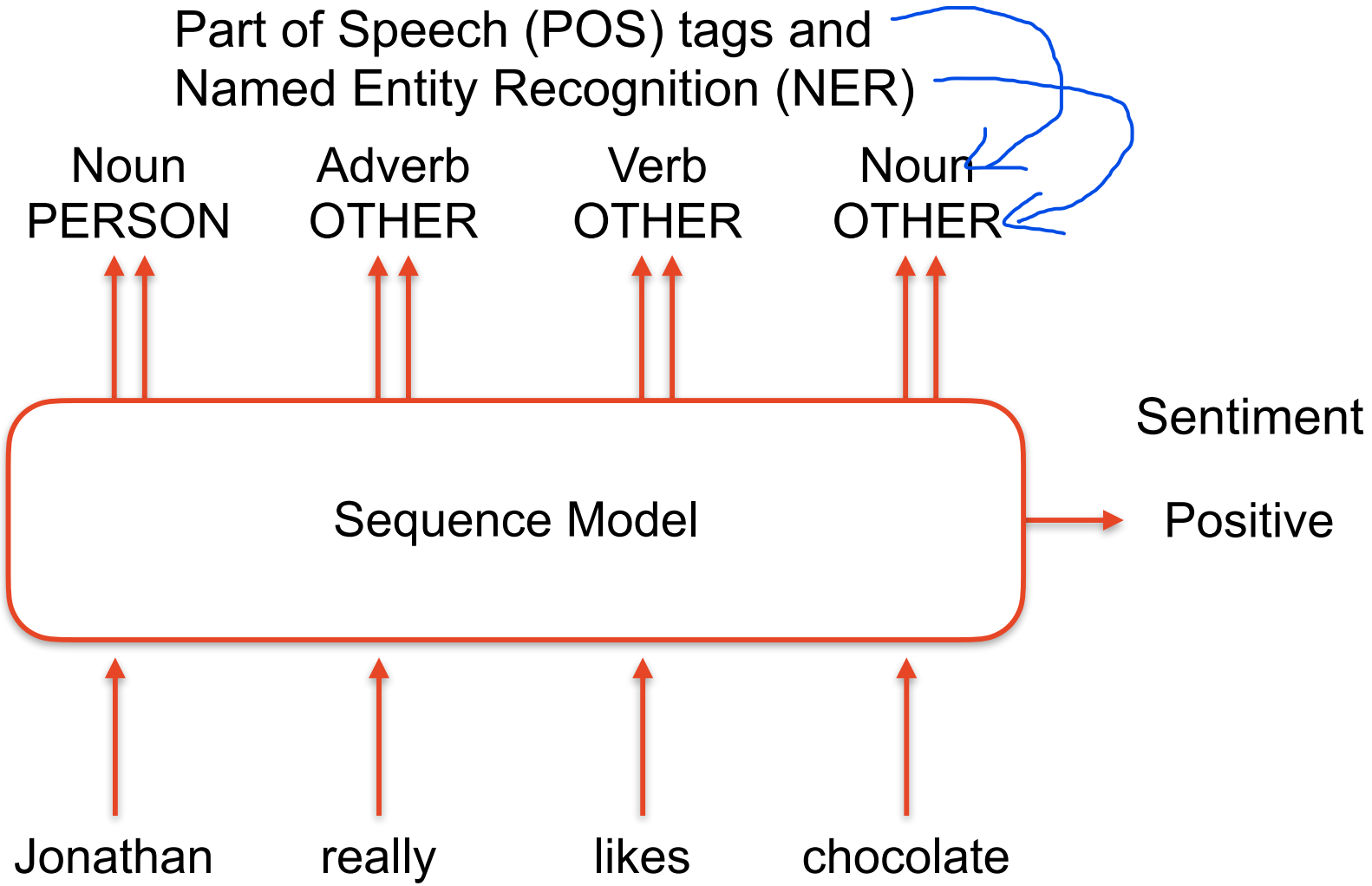


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can train on multiple related tasks

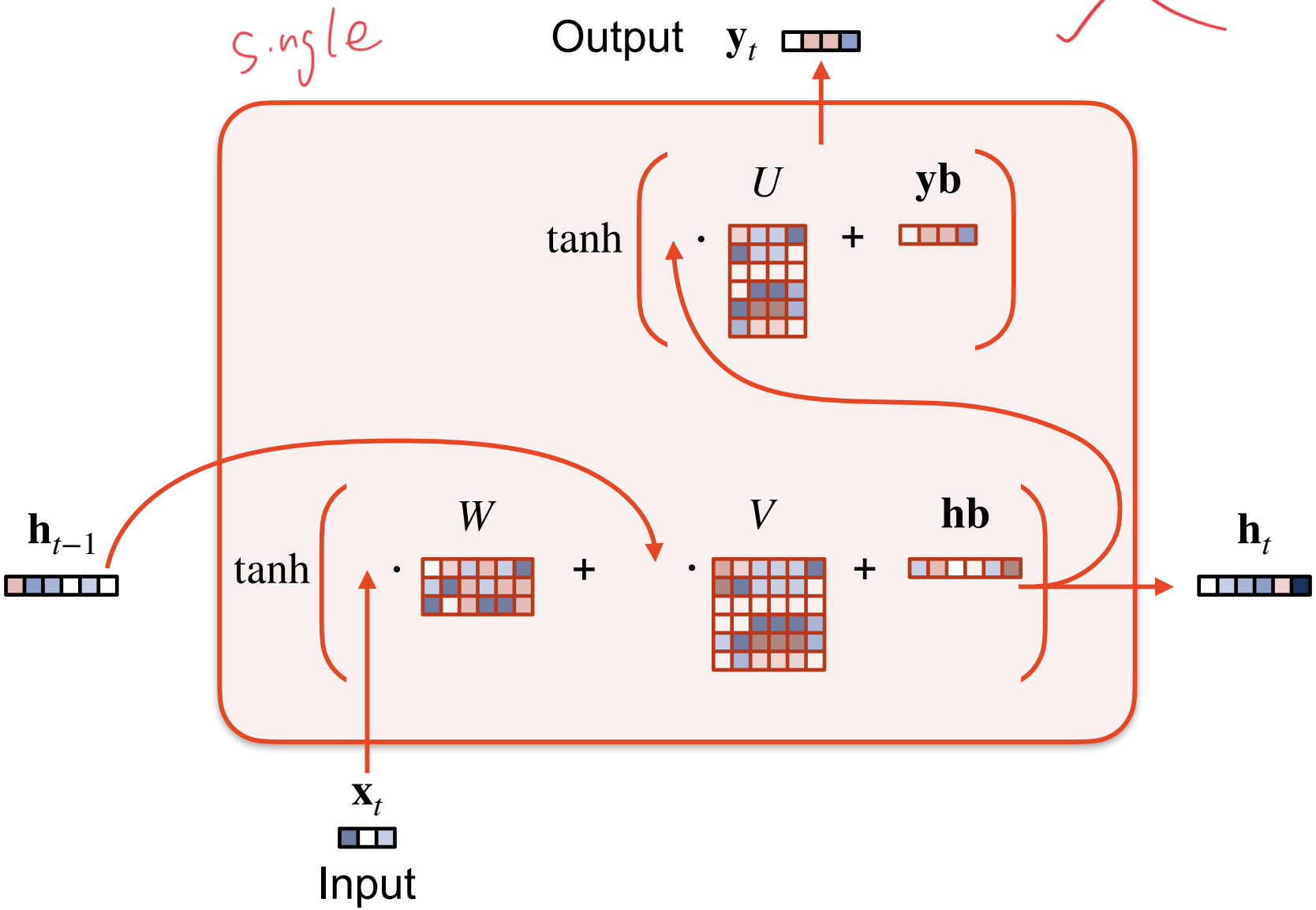


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can train on multiple related tasks

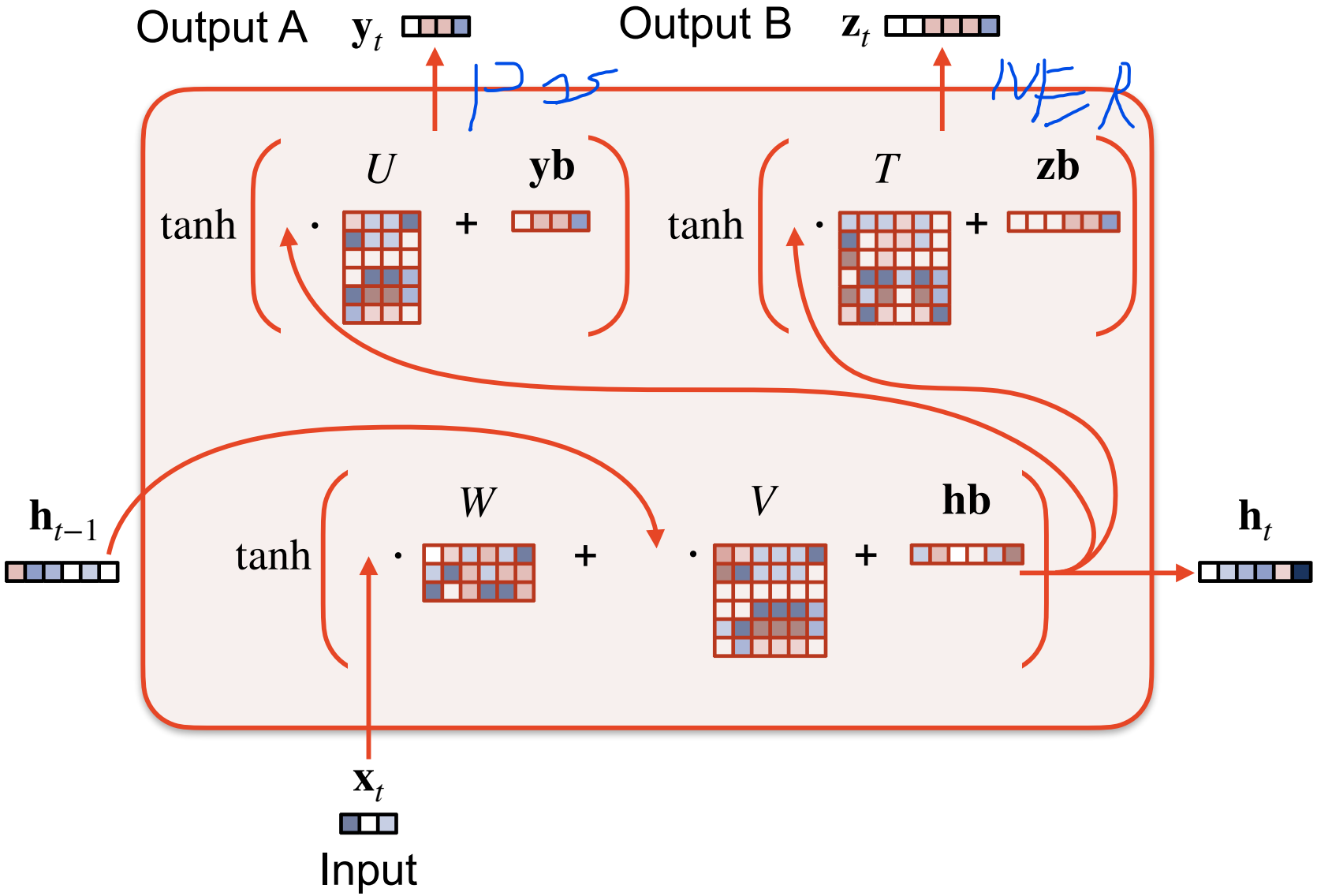


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

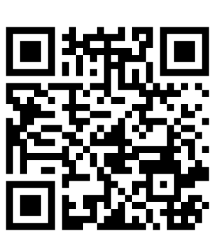
menti.com 1750 7815

We can train on multiple related tasks



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

Named Entity Recognition

NER

Person	Other	Other	Other
Jonathan	really	likes	chocolate

Identifying proper nouns:

Coarse-grained:

- Locations
- Geo-Political Entities
- People
- Organisations

Example

- Blackwattle Bay
- Sydney
- Jonathan
- University of Sydney

Fine-grained:

- 112 categories in Ling and Weld (2012)!

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

Named Entity Recognition

Person Other Other Other
Jonathan really likes chocolate

Identifying proper nouns:

Words	IO Label	BIO Label	BIOES Label
Jane	I-PER	B-PER	B-PER
Villanueva	I-PER	I-PER	E-PER
of	O	O	O
United	I-ORG	B-ORG	B-ORG
Airlines	I-ORG	I-ORG	I-ORG
Holding	I-ORG	I-ORG	E-ORG
discussed	O	O	O
the	O	O	O
Chicago	I-LOC	B-LOC	S-LOC
route	O	O	O
.	O	O	O

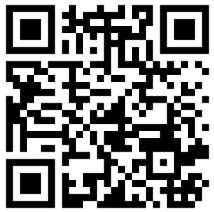
Why does it matter?

I gave Joe Smith a book
I gave Joe Zach's book

Jurafsky and Martin, chapter 8

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

Analysis

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

In lecture 2, we saw we could form a confusion matrix

	Answer: Spam	Answer: Not Spam
Guess: Spam	● ● ● ●	△ △ △
Guess: Not Spam	○ ○ ○ ○ ○ ○	▲ ▲ ▲ ▲ ▲ ▲ ▲

● Spam and predicted Spam

▲ Not spam and predicted Not spam

○ Spam, predicted Not spam

△ Not spam, predicted Spam

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

In lecture 2, we saw we could form a confusion matrix

		True Answer				
		A	B	C	…	None
Guess	A	TP				FP
	B		TP			FP
	C			TP		FP
	…				TP	FP
	None	FN	FN	FN	FN	TN

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

In lecture 2, we saw we could form a confusion matrix

		True Answer				
		Person	Location	Org	…	Other
Guess	Person	TP				FP
	Location		TP			FP
	Org			TP		FP
	…				TP	FP
	Other	FN	FN	FN	FN	TN

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

For NER, we consider typed spans / chunks

Convert tags into labeled spans:

	Words	BIO Label
0	Jane	B-PER
1	Villanueva	I-PER
2	of	O
3	United	B-ORG
4	Airlines	I-ORG
5	Holding	I-ORG
6	discussed	O
7	the	O
8	Chicago	B-LOC
9	route	O
10	.	O

[0, 1, PER]
[3, 5, ORG]
[8, 8, LOC]

Compare true answer and guess:

Match? TP
Item in guess, but not answer? FP
Item in answer, but not guess? FN

Note: Possible outputs = | Labels | * Length * Length
TN will be huge!

Jurafsky and Martin, chapter 8

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

NER is a good example of where micro-P, micro-R, and micro-F are used

$$\text{Precision}_{\text{micro}} = \frac{\sum_{l \in \text{labels}} \text{TP}_l}{\sum_{l \in \text{labels}} \text{TP}_l + \text{FP}_l}$$

$$\text{Recall}_{\text{micro}} = \frac{\sum_{l \in \text{labels}} \text{TP}_l}{\sum_{l \in \text{labels}} \text{TP}_l + \text{FN}_l}$$

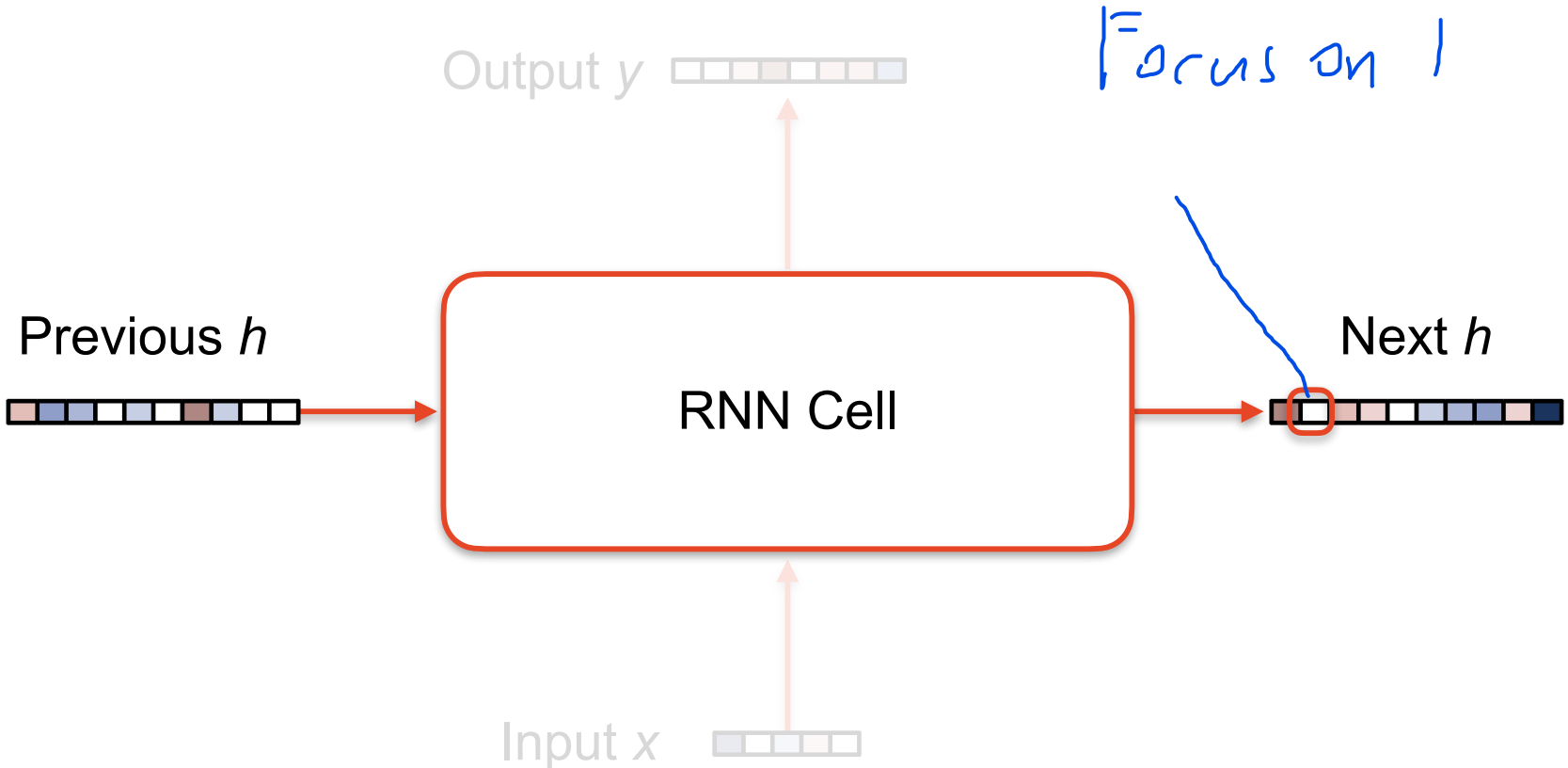
Why? Some labels are less common and we care about them in proportion to their frequency

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state



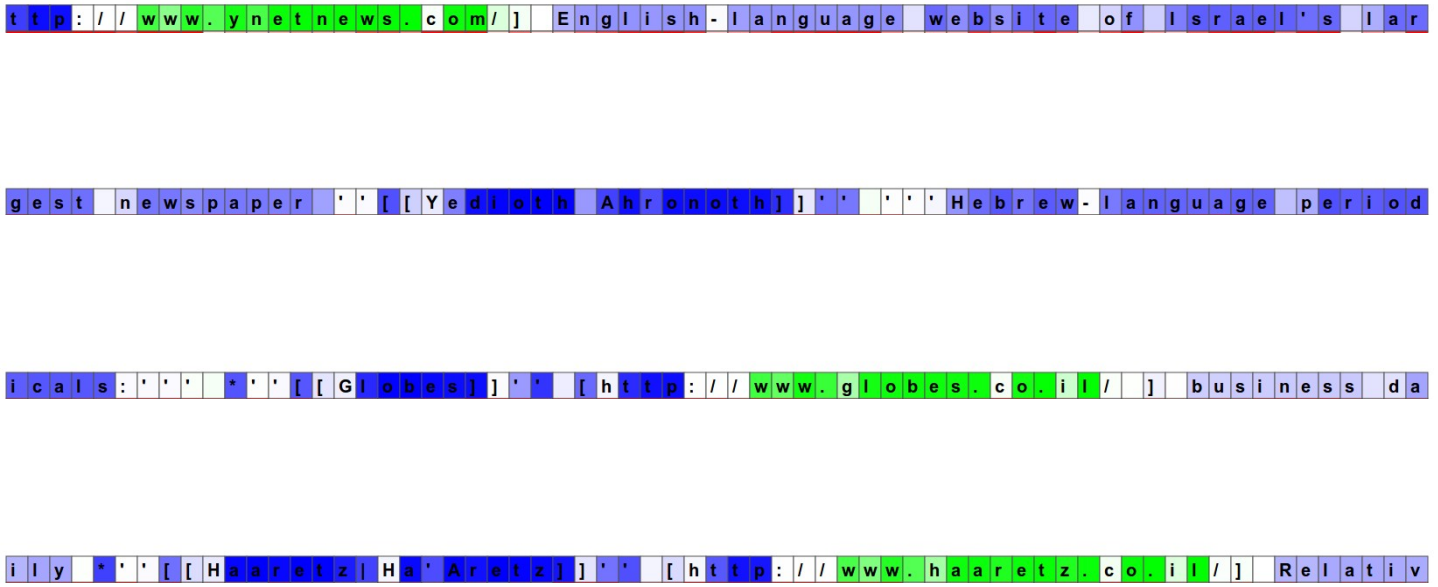
"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state



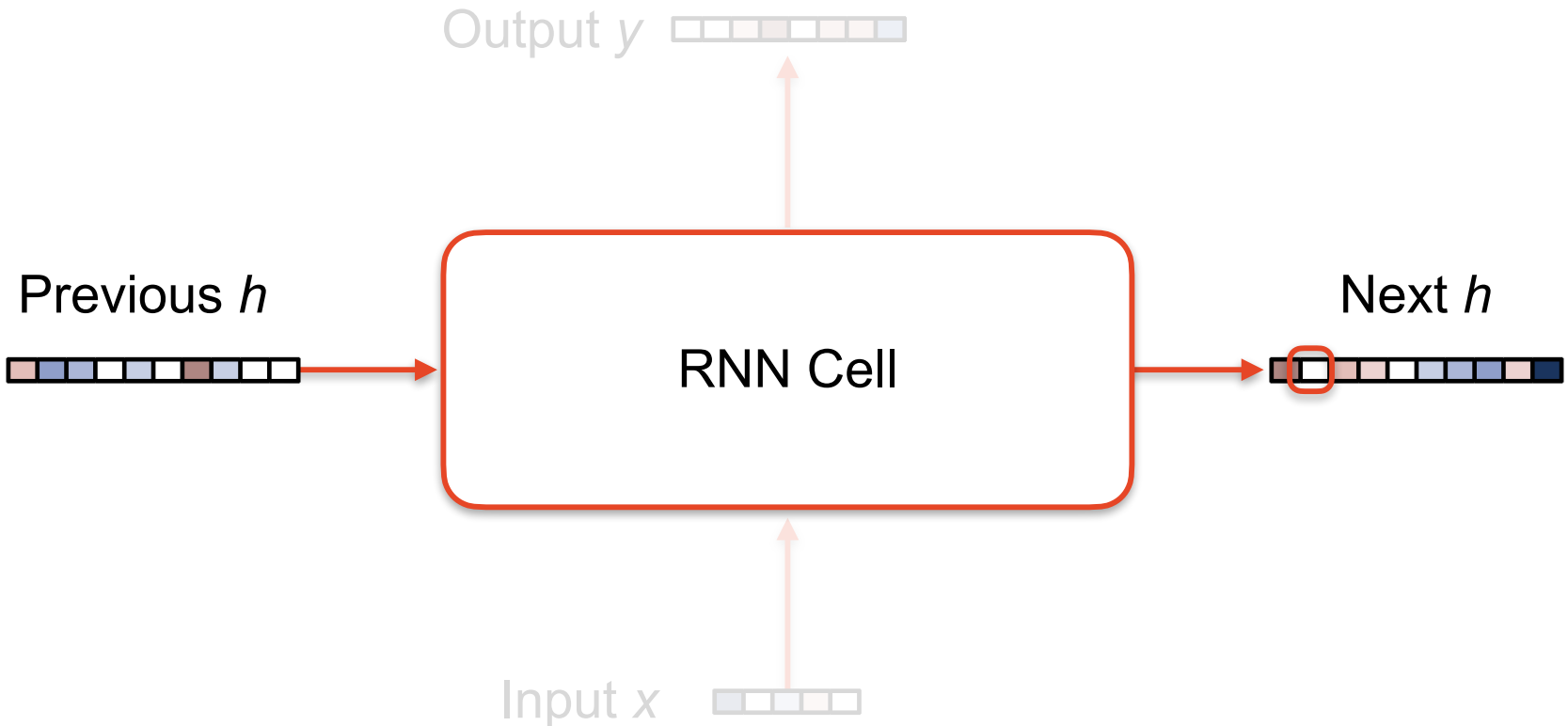
"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state



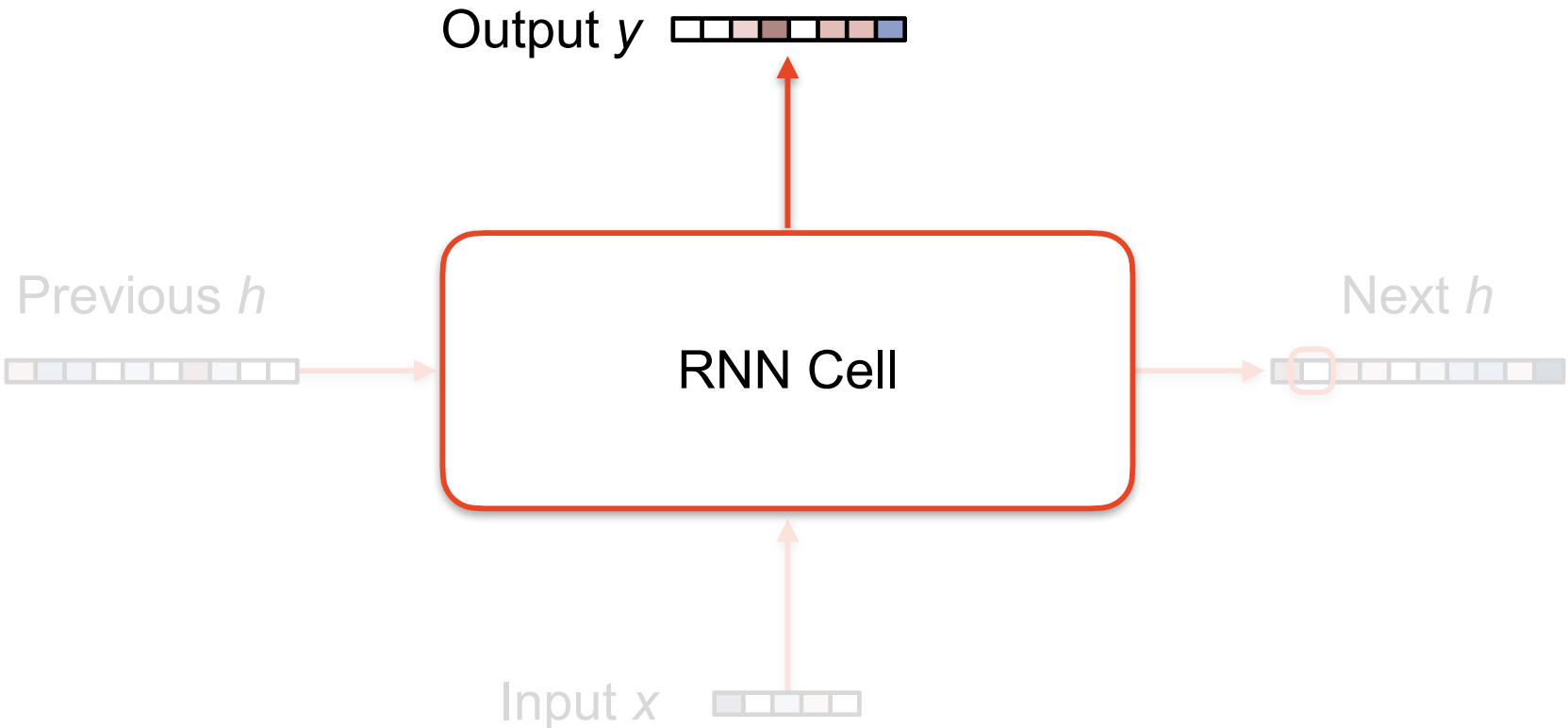
"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state



"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

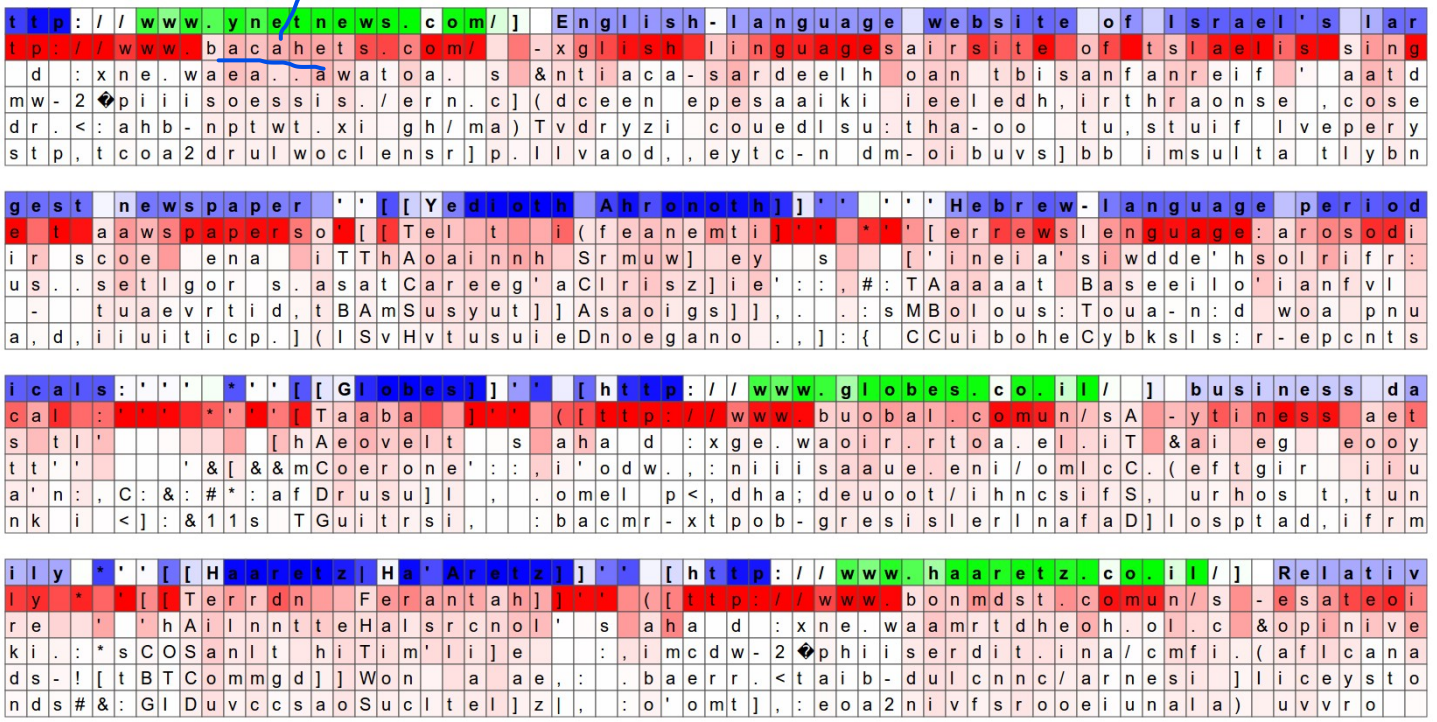
menti.com 1750 7815

We can visualise (1) values in the hidden state, and (2) the output distribution

confident predict — not confident

Input

Guesses



"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

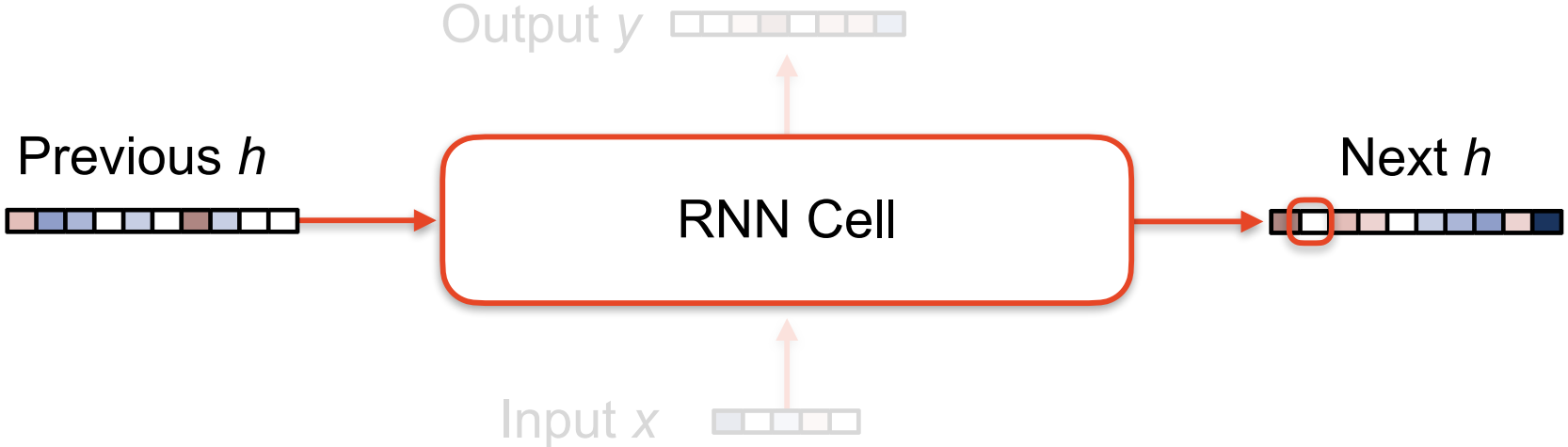
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state, and (2) the output distribution

Cell sensitive to position in line:

The sole importance of the crossing of the Berezina lies in the fact that it plainly and indubitably proved the fallacy of all the plans for cutting off the enemy's retreat and the soundness of the only possible line of action--the one Kutuzov and the general mass of the army demanded--namely, simply to follow the enemy up. The French crowd fled at a continually increasing speed and all its energy was directed to reaching its goal. It fled like a wounded animal and it was impossible to block its path. This was shown not so much by the arrangements it made for crossing as by what took place at the bridges. When the bridges broke down, unarmed soldiers, people from Moscow and women with children who were with the French transport, all--carried on by vis inertiae--pressed forward into boats and into the ice-covered water and did not surrender.



"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

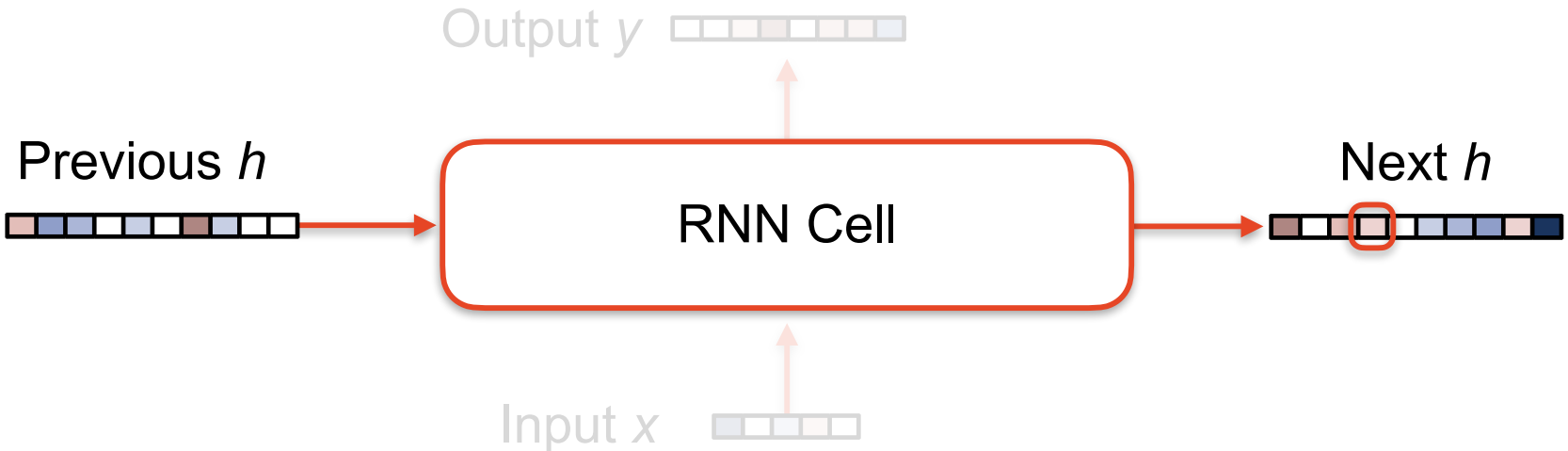
menti.com 1750 7815

We can visualise (1) values in the hidden state, and (2) the output distribution

Cell that turns on inside quotes:

"You mean to imply that I have nothing to eat out of.... On the contrary, I can supply you with everything even if you want to give dinner parties," warmly replied Chichagov, who tried by every word he spoke to prove his own rectitude and therefore imagined Kutuzov to be animated by the same desire.

Kutuzov, shrugging his shoulders, replied with his subtle penetrating smile: "I meant merely to say what I said."



"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

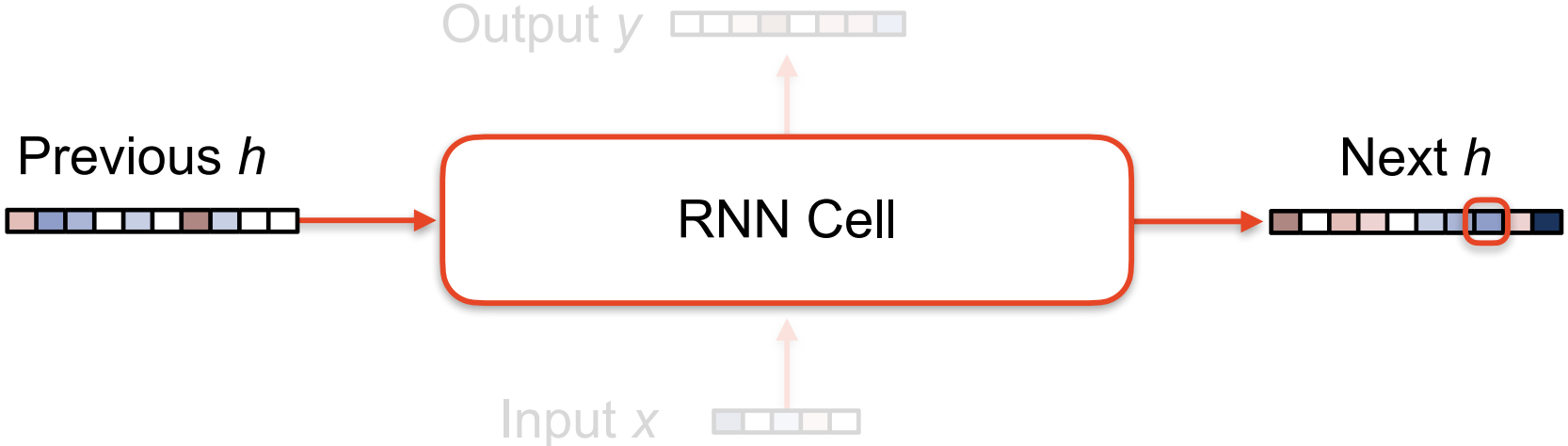
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state, and (2) the output distribution

Cell that robustly activates inside if statements:

```
static int __dequeue_signal(struct sigpending *pending, sigset_t *mask,
        siginfo_t *info)
{
  int sig = next_signal(pending, mask);
  if (sig) {
    if (current->notifier) {
      if (sigismember(current->notifier_mask, sig)) {
        if (!(current->notifier)(current->notifier_data)) {
          clear_thread_flag(TIF_SIGPENDING);
          return 0;
        }
      }
    }
    collect_signal(sig, pending, info);
  }
  return sig;
}
```



"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

COMP 4446 / 5046
Lecture 3, 2025

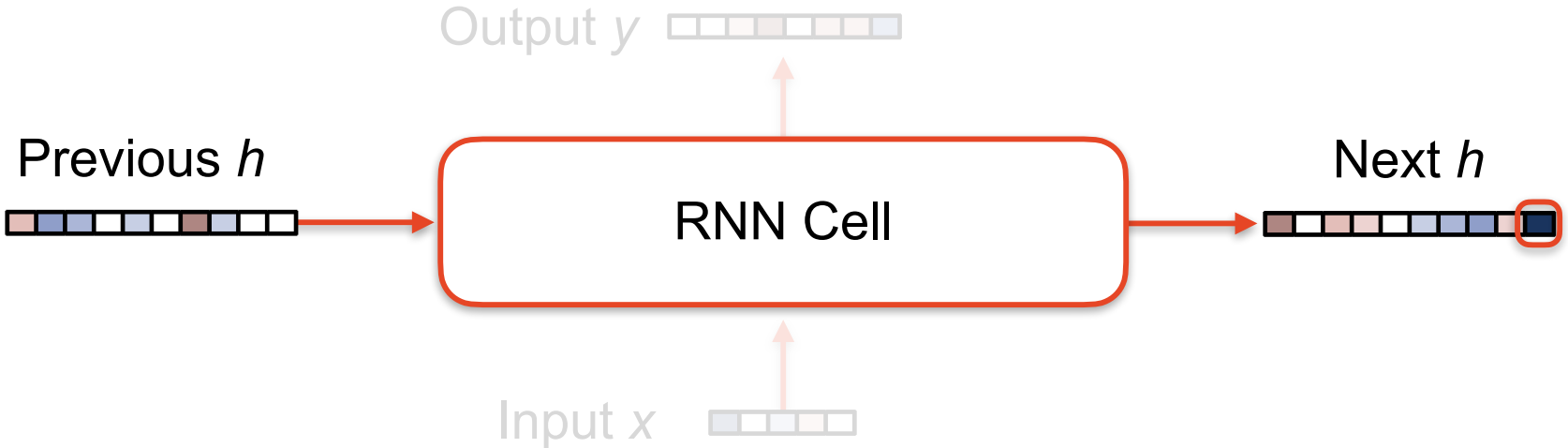
Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 1750 7815

We can visualise (1) values in the hidden state, and (2) the output distribution

A large portion of cells are not easily interpretable. Here is a typical example:

```
/* Unpack a filter field's string representation from user-space
 * buffer. */
char *audit_unpack_string(void **bufp, size_t *remain, size_t len)
{
  char *str;
  if (!*bufp || (len == 0) || (len > *remain))
    return ERR_PTR(-EINVAL);
  /* Of the currently implemented string fields, PATH_MAX
   * defines the longest valid length.
   */
```

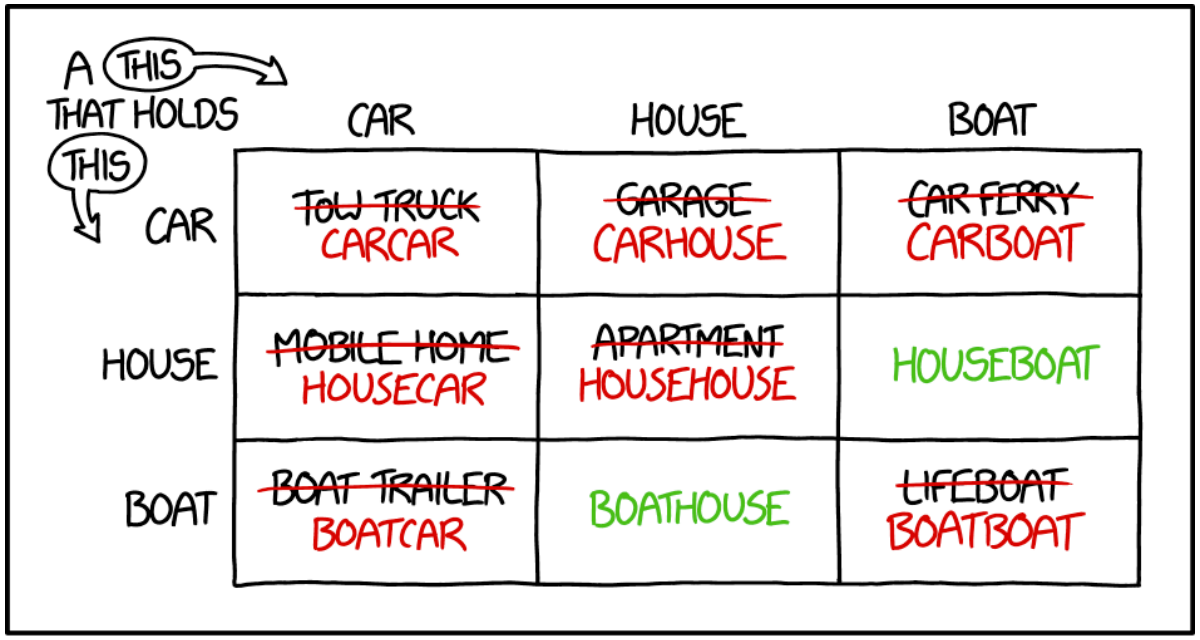


"The Unreasonable Effectiveness of Recurrent Neural Networks", Andrej Karpathy, 2015

3 minute Break - stretch and visit Menti

menti.com
1750 7815

Boathouses and Houseboats

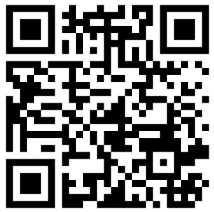


[The <x> that is held by <y> is also a <y><x>, so if you go to a food truck, the stuff you buy is truck food. A phone that's in your car is a carphone, and a car equipped with a phone is a phonecar. When you play a mobile racing game, you're in your phonecar using your carphone to drive a different phonecar. I'm still not sure about bananaphones.]

Source: https://xkcd.com/2043/

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview



menti.com 1750 7815

Workshop Preview

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 7927 6661

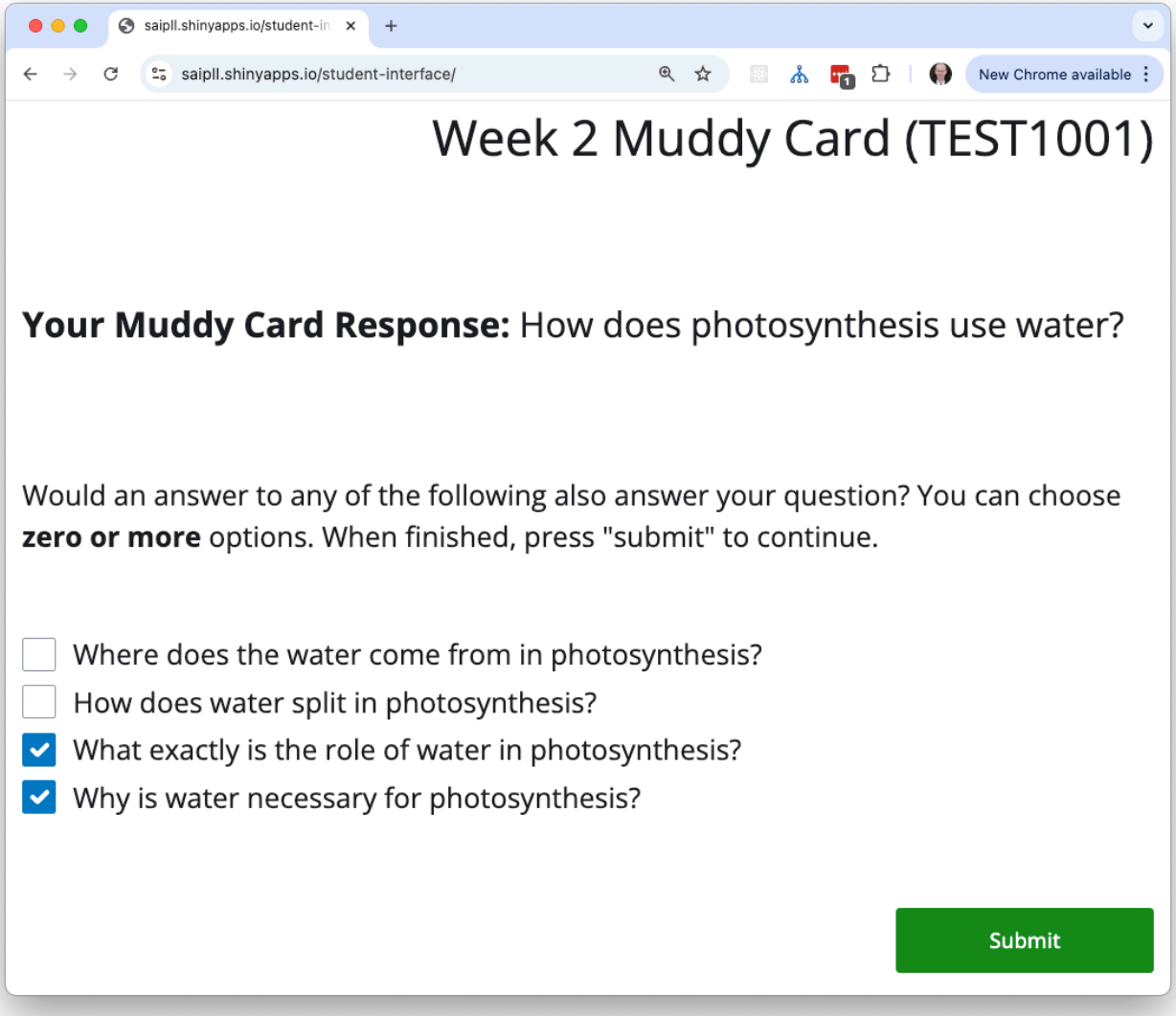
Workshop:

COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 7927 6661

Muddy Card

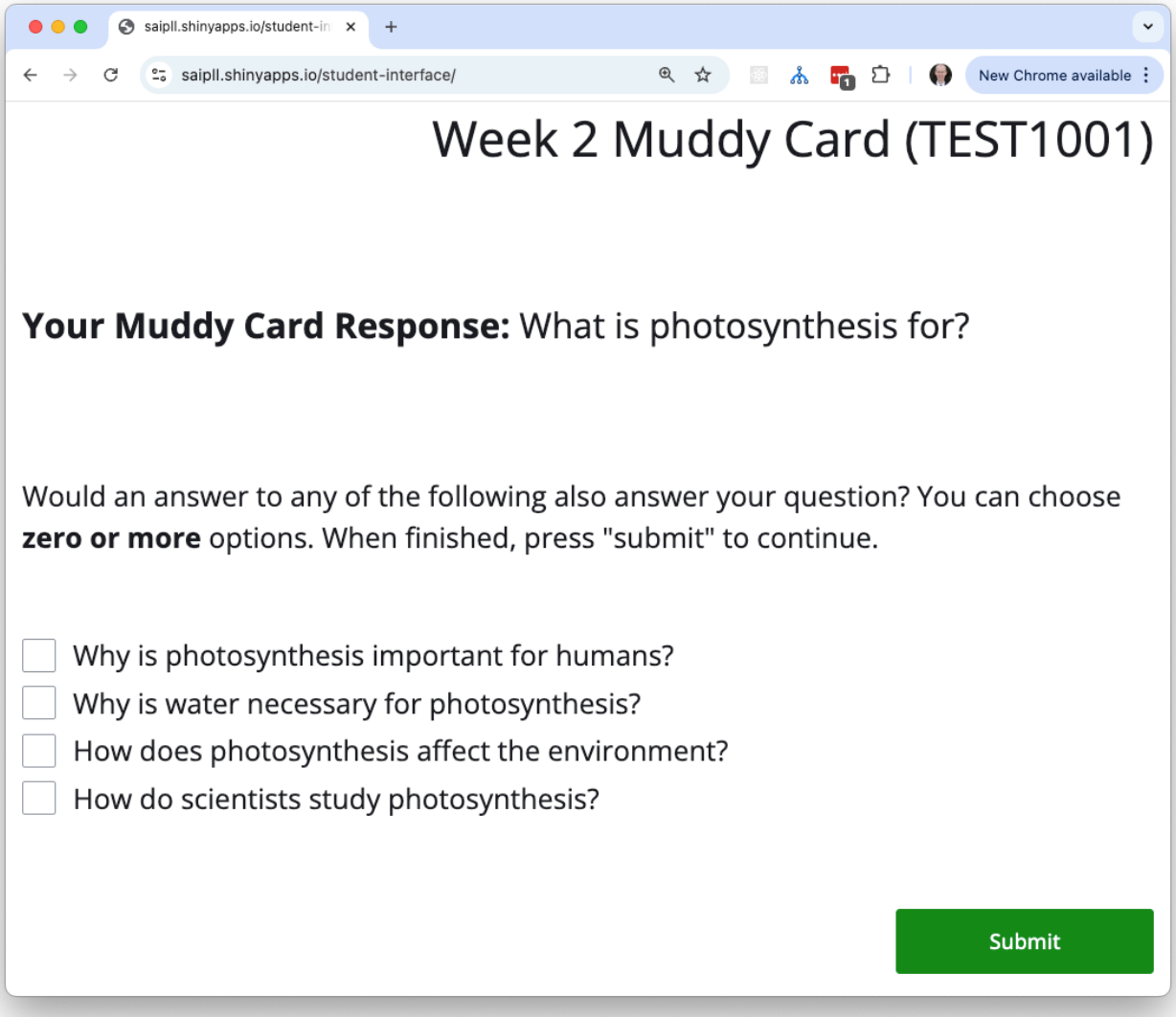


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 7927 6661

Muddy Card

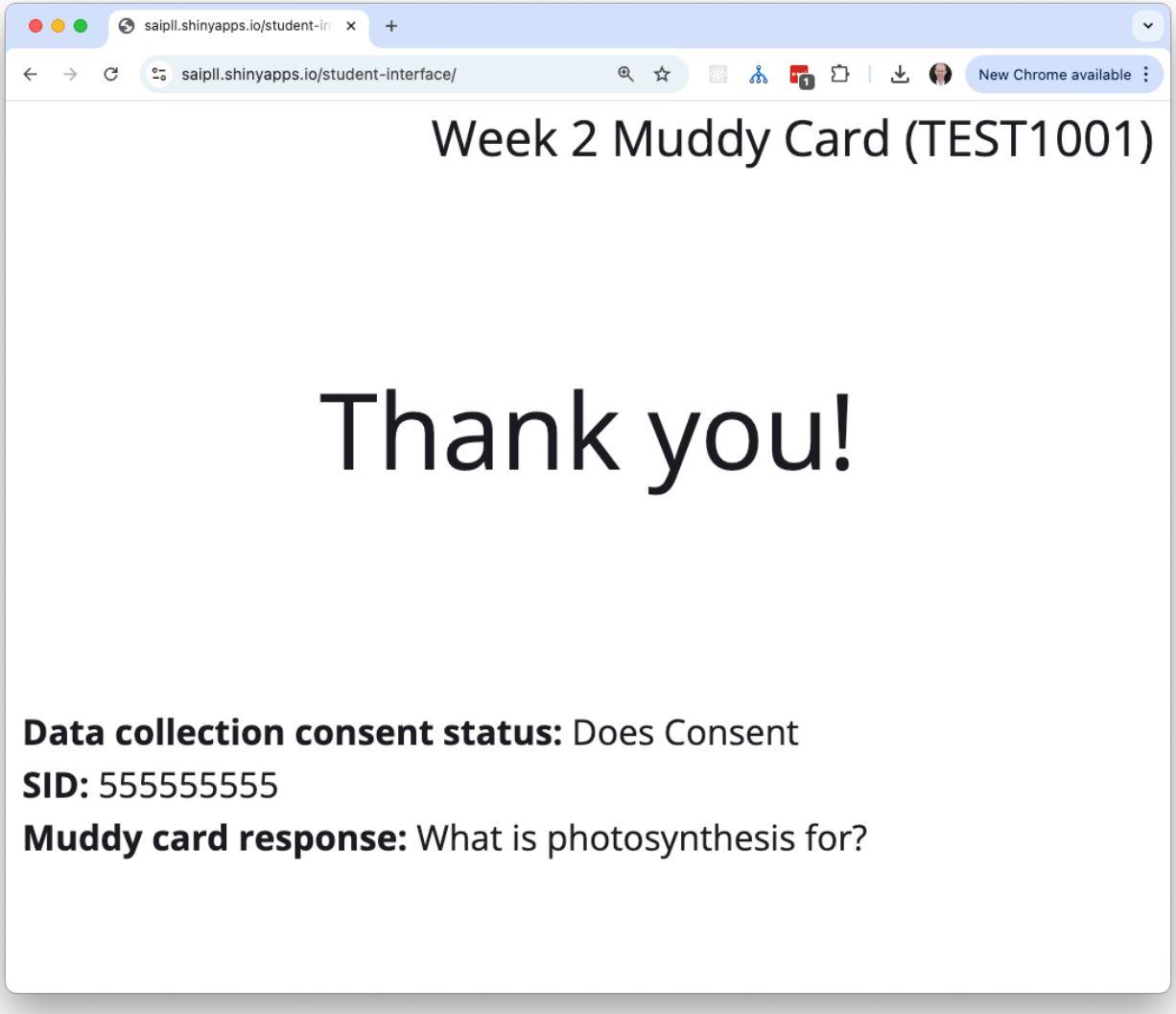


COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 7927 6661

Muddy Card



COMP 4446 / 5046
Lecture 3, 2025

Neural Networks
Recurrent Models
Analysis
Workshop Preview

menti.com 7927 6661

Muddy Card

Open in a moment, closes at 7:05pm

https://saipll.shinyapps.io/student-interface/

If you do not wish to participate in the study, use the Ed form instead

Go to Ed → Lessons → Muddy Card Lecture 3