CS21si: AI for Social Good

Lecture 5: Recurrent Neural Networks

Plan for Today

- Fake news
- Natural language processing with deep learning
- Language models
- Recurrent neural networks

Fake news

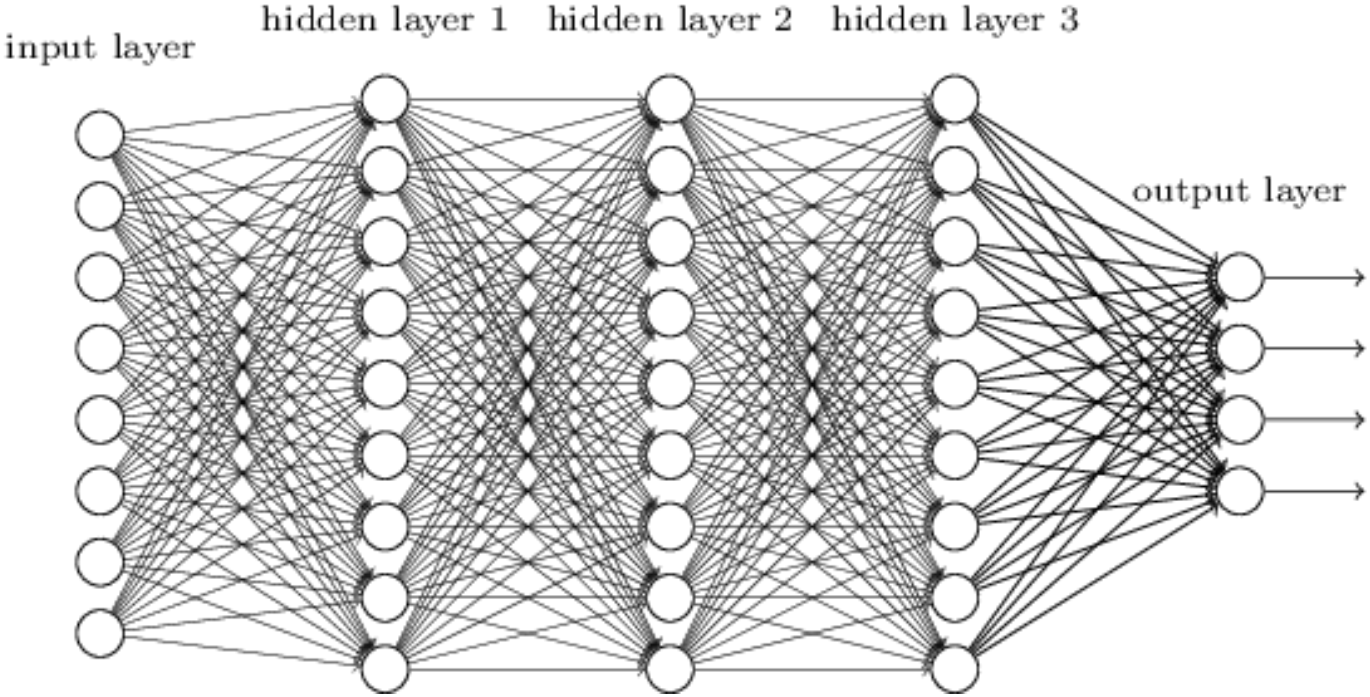
WEEKLY WORLD
NEWS
Coast-to-coast manhunt for cave boy captured just weeks ago!
NEWS
BAT CHILD FOUND IN CAVE!
BAT CHILD ESCAPES!
He's 2-feet-tall and VERY DANGEROUS, warn police

WEEKLY WORLD
NEWS
Space creature survived UFO crash in Arkansas!
HILLARY CLINTON ADOPTS ALIEN BABY
OFFICIAL PHOTO!
Secret Service building special nursery in the White House!

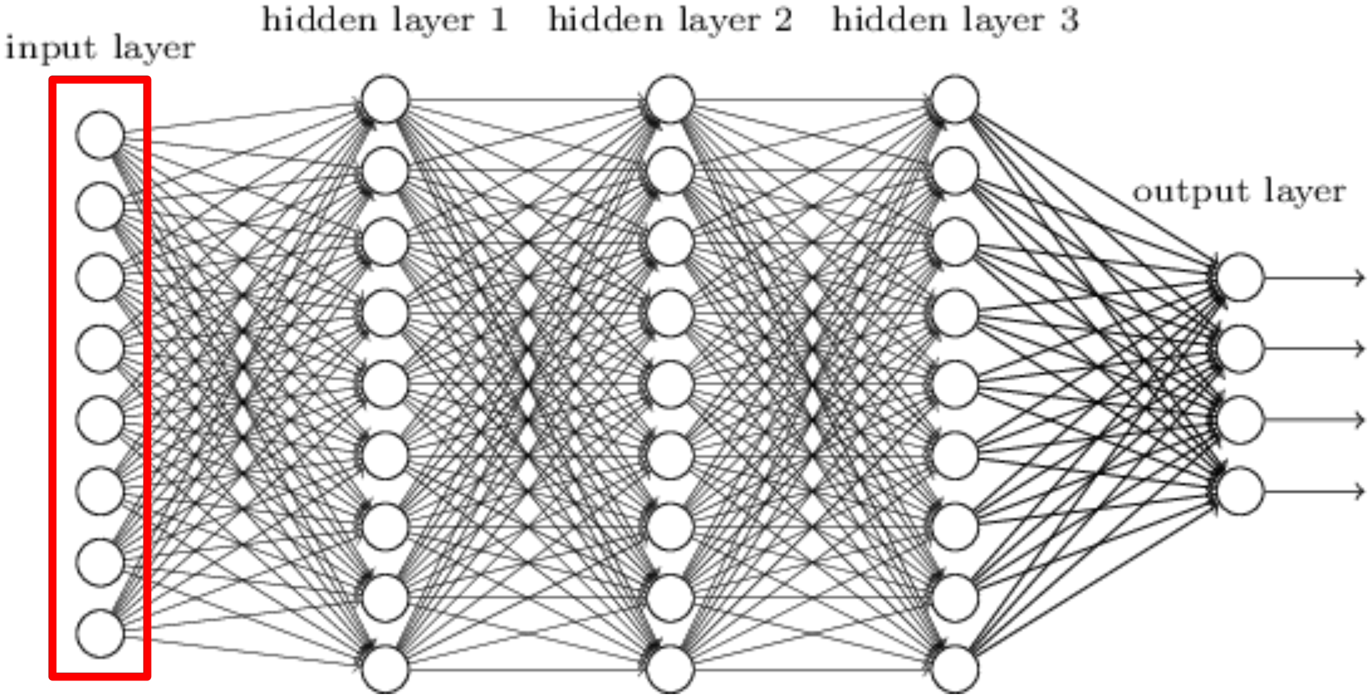
Our Dataset

How do we deal with text data?

Deep Neural Networks



Deep Neural Networks



I’ll meet you at the airport in an …

I'll meet you at the airport in an …

$\mathbf{x}^{(1)}$ $\mathbf{x}^{(2)}$ $\mathbf{x}^{(3)}$ $\mathbf{x}^{(4)}$ $\mathbf{x}^{(5)}$ $\mathbf{x}^{(6)}$ $\mathbf{x}^{(7)}$ $\mathbf{x}^{(8)}$

I’ll meet you at the airport in an …

$$x^{(1)} \quad x^{(2)} \quad x^{(3)} \quad x^{(4)} \quad x^{(5)} \quad x^{(6)} \quad x^{(7)} \quad x^{(8)}$$

$$\Big\downarrow \text{Word2Vec}$$

$$e^{(1)} \quad e^{(2)} \quad e^{(3)} \quad e^{(4)} \quad e^{(5)} \quad e^{(6)} \quad e^{(7)} \quad e^{(8)}$$

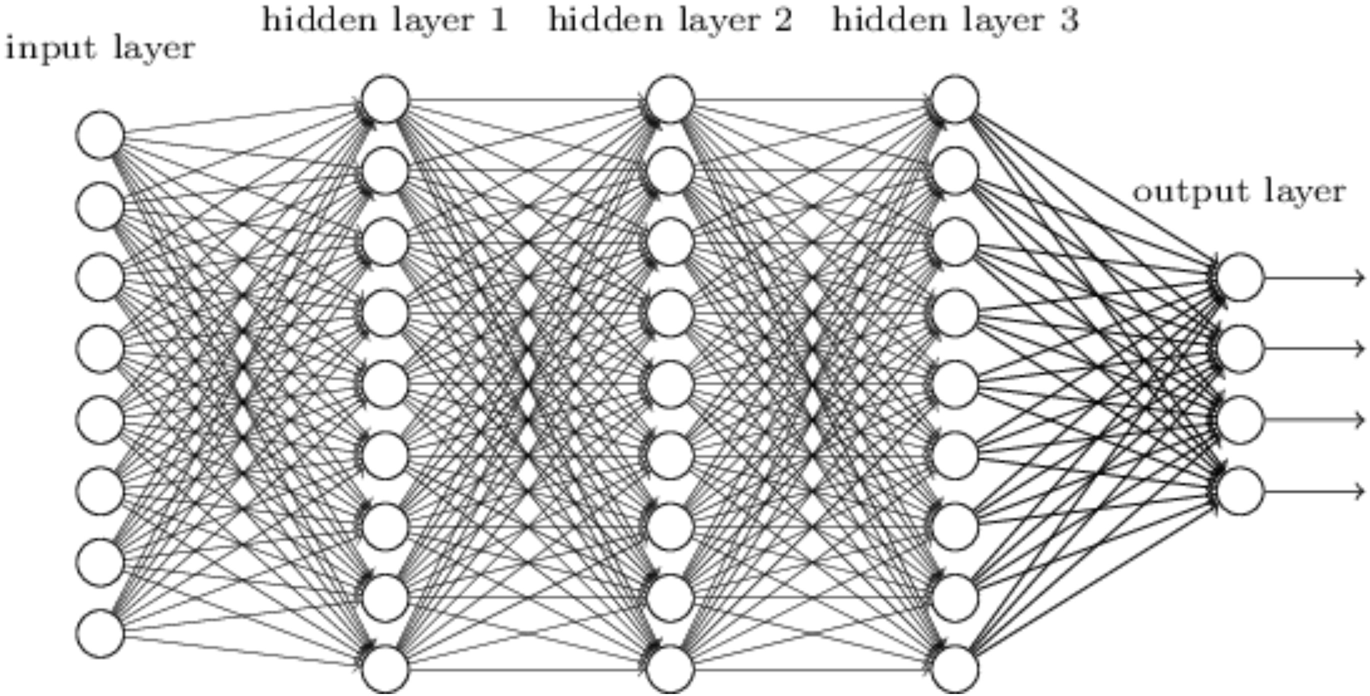
I’ll meet you at the airport in an …

$$\mathbf{x}^{(1)} \quad \mathbf{x}^{(2)} \quad \mathbf{x}^{(3)} \quad \mathbf{x}^{(4)} \quad \mathbf{x}^{(5)} \quad \mathbf{x}^{(6)} \quad \mathbf{x}^{(7)} \quad \mathbf{x}^{(8)}$$

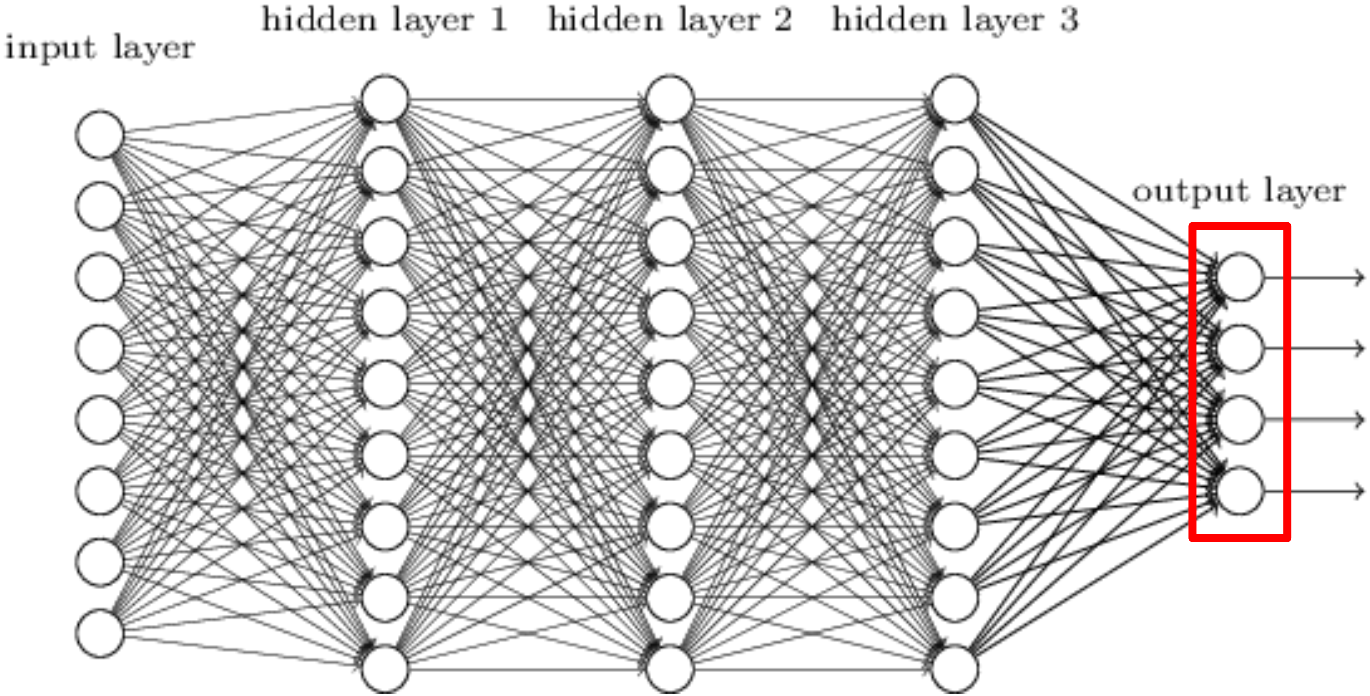
$$\Big\downarrow \ \text{Word2Vec}$$

$$\begin{bmatrix} e^{(1)} & e^{(2)} & e^{(3)} & e^{(4)} & e^{(5)} & e^{(6)} & e^{(7)} & e^{(8)} \end{bmatrix}$$

$$\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ e^{(3)} \\ e^{(4)} \\ e^{(5)} \\ e^{(6)} \\ e^{(7)} \\ e^{(8)} \end{bmatrix}$$



$$\begin{bmatrix} e^{(1)} \\ e^{(2)} \\ e^{(3)} \\ e^{(4)} \\ e^{(5)} \\ e^{(6)} \\ e^{(7)} \\ e^{(8)} \end{bmatrix}$$



Let’s predict the next word!

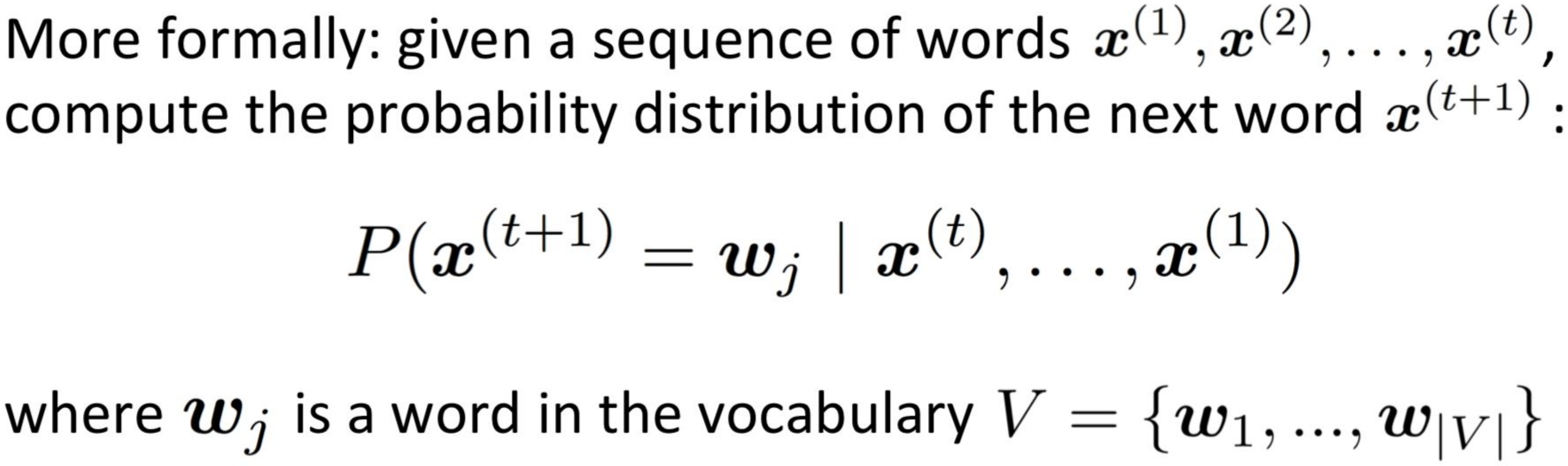
(a.k.a. multi-class classification with |V| classes)

I’ll meet you at the airport in an …

I’ll meet you at the airport in an …

hour?
minute?
automobile?

Language models



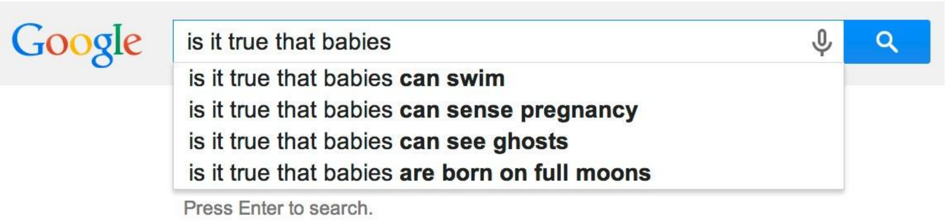
More formally: given a sequence of words $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}$, compute the probability distribution of the next word $\boldsymbol{x}^{(t+1)}$:

$$P(\boldsymbol{x}^{(t+1)} = \boldsymbol{w}_j \mid \boldsymbol{x}^{(t)}, \ldots, \boldsymbol{x}^{(1)})$$

where $\boldsymbol{w}_j$ is a word in the vocabulary $V = \{\boldsymbol{w}_1, ..., \boldsymbol{w}_{|V|}\}$

Questions?

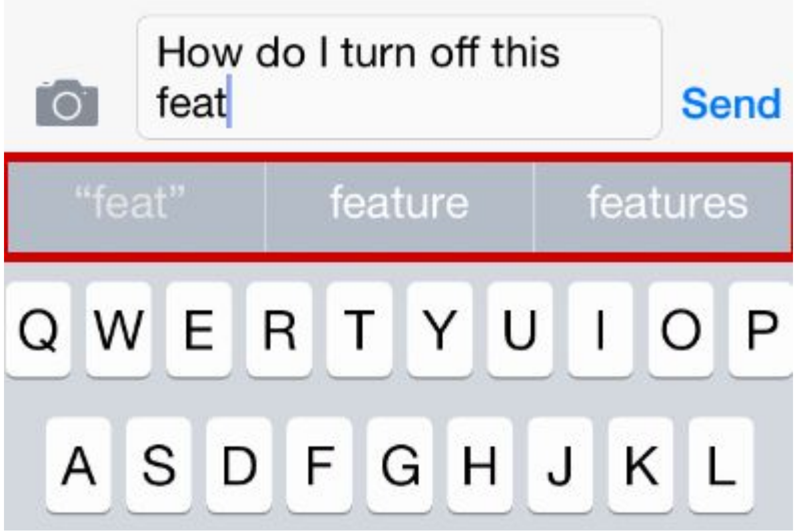
You use language models every day!



is it true that babies

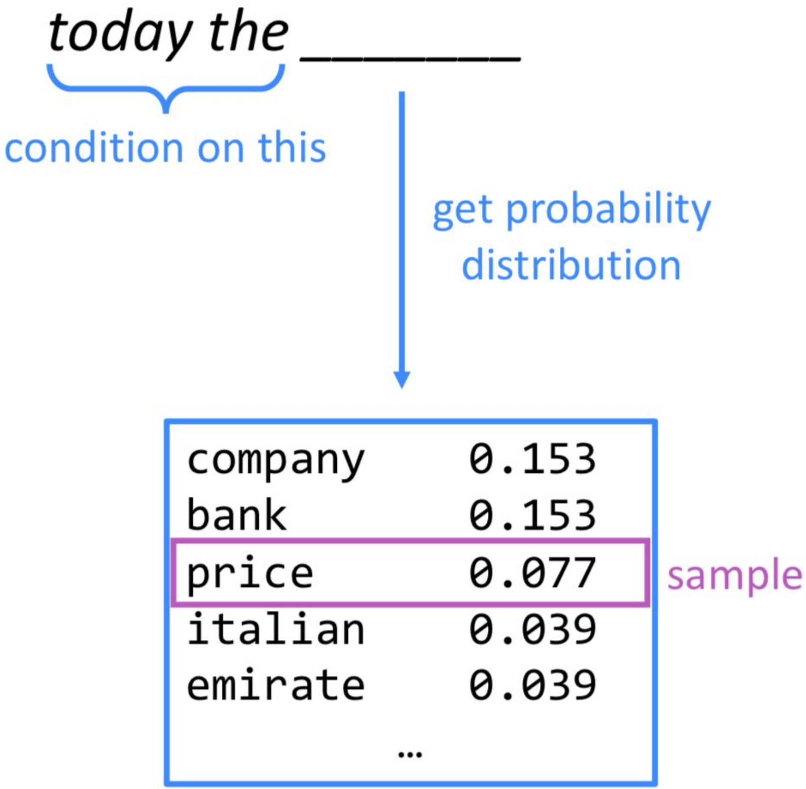
- is it true that babies **can swim**
- is it true that babies **can sense pregnancy**
- is it true that babies **can see ghosts**
- is it true that babies **are born on full moons**

Press Enter to search.



How do I turn off this feat

You can use language models to generate new text!



today the ________

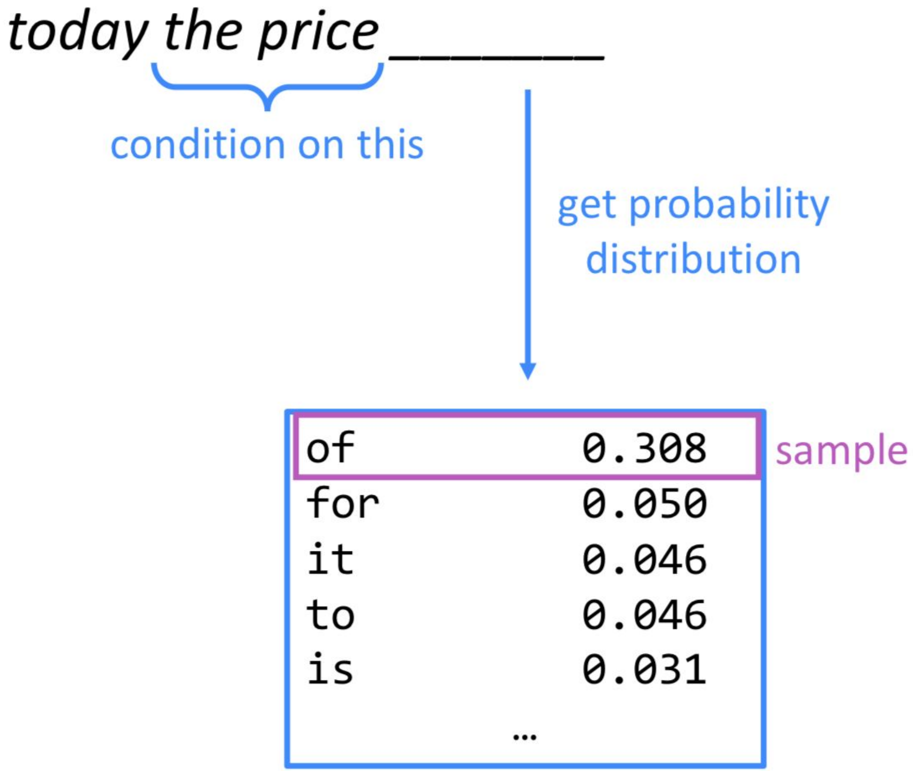
condition on this

get probability distribution

company	0.153
bank	0.153
price	0.077
italian	0.039
emirate	0.039
…	

sample

You can use language models to generate new text!



today the price ________

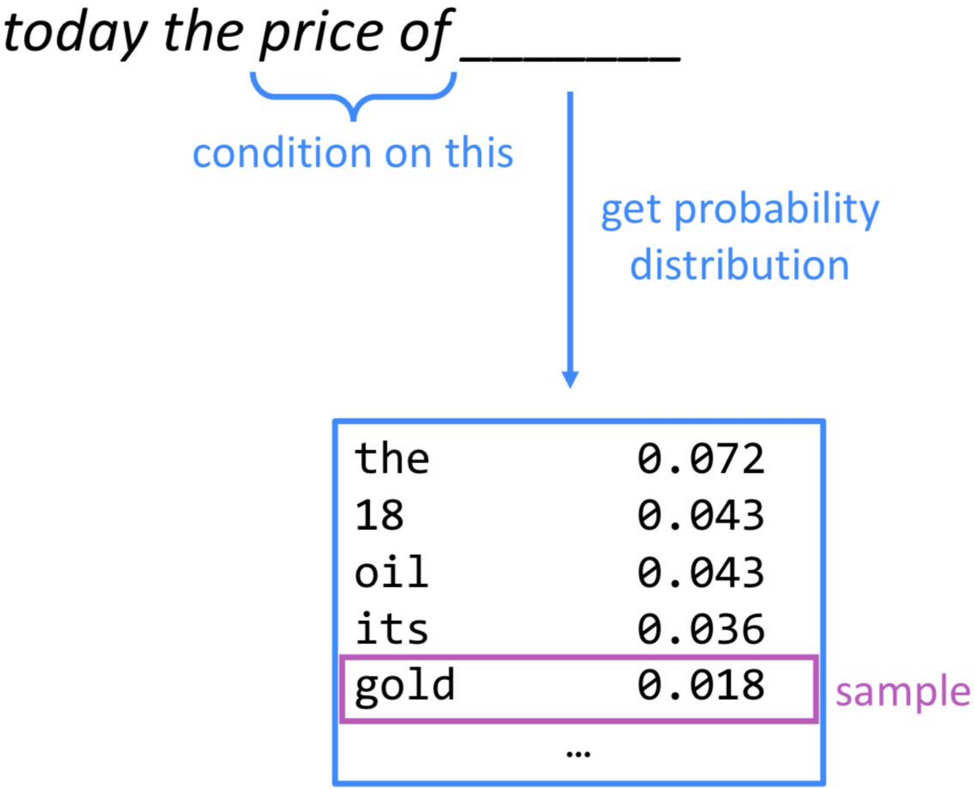
condition on this

get probability distribution

of	0.308
for	0.050
it	0.046
to	0.046
is	0.031
…	

sample

You can use language models to generate new text!



today the price of ________

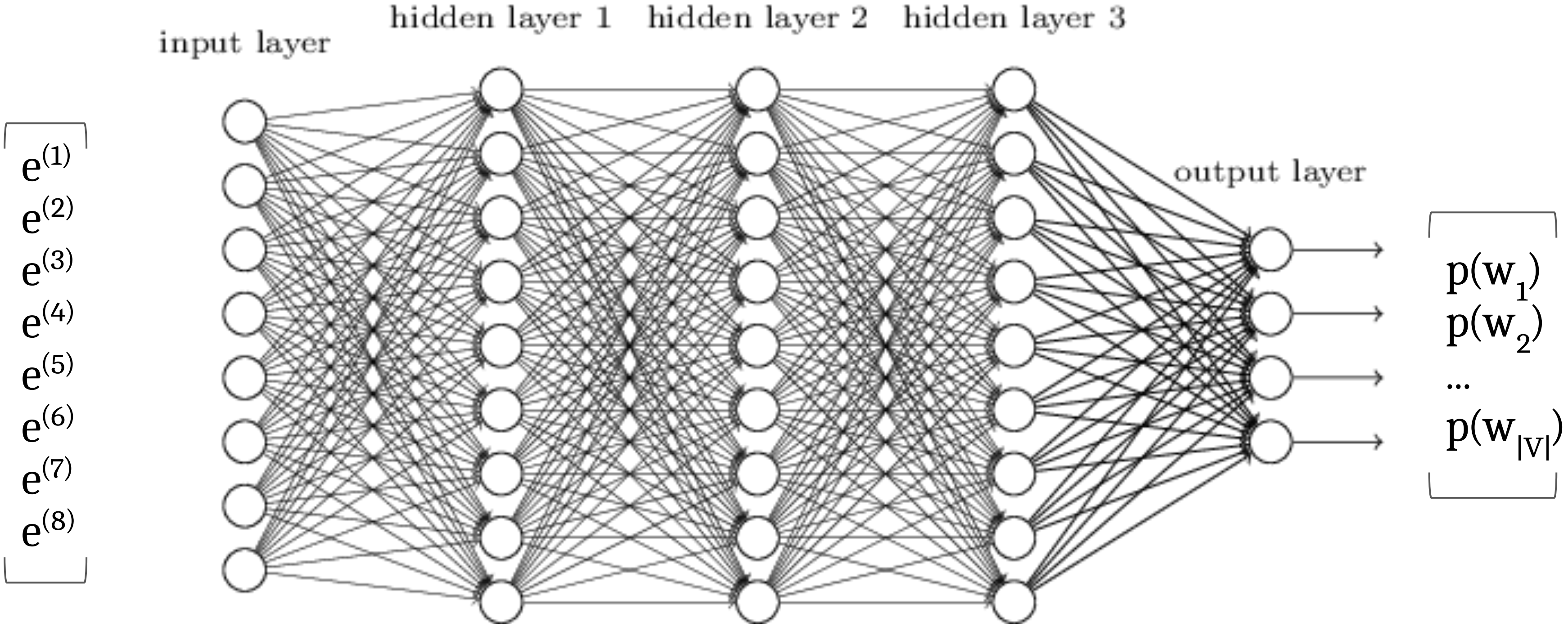
condition on this

get probability distribution

the	0.072
18	0.043
oil	0.043
its	0.036
gold	0.018
…	

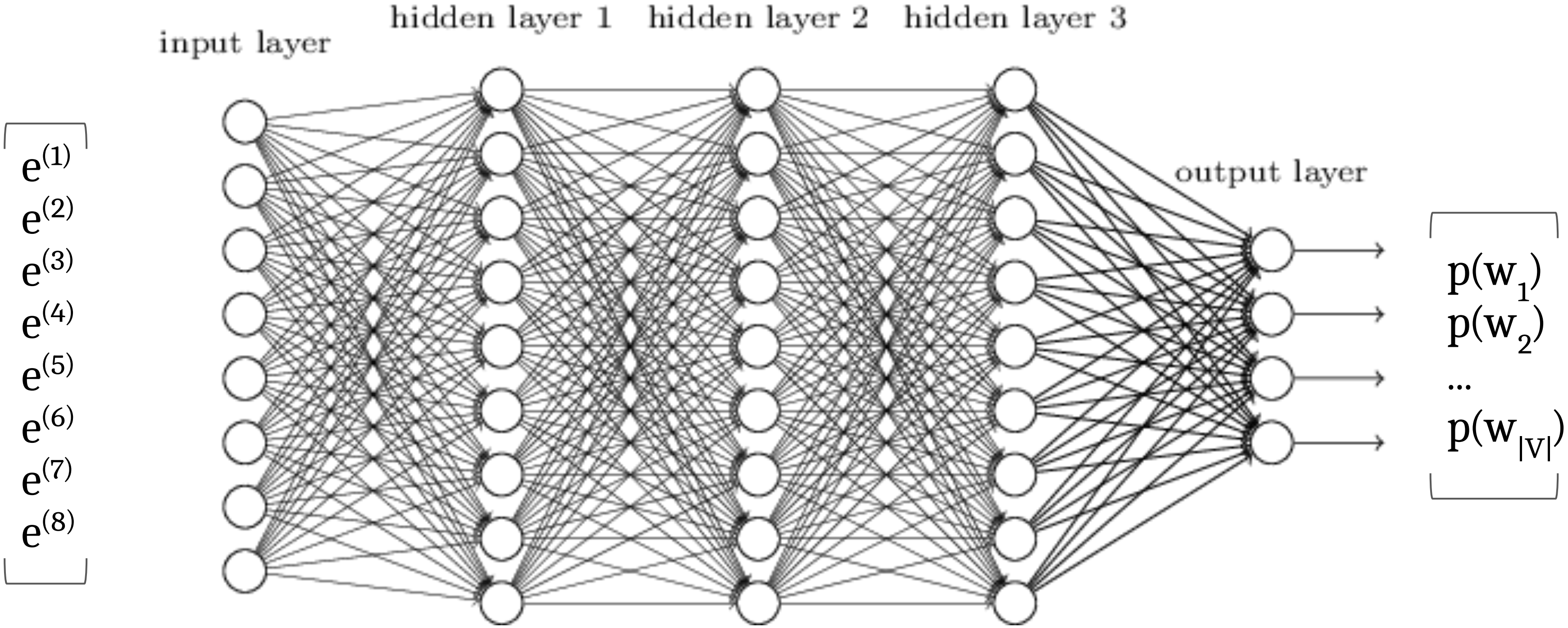
sample

Deep Learning + NLP: Attempt #1



Class Exercises Part 1:
Neural NLP Warmup

Deep Learning + NLP: First Attempt



What’s wrong with our model?

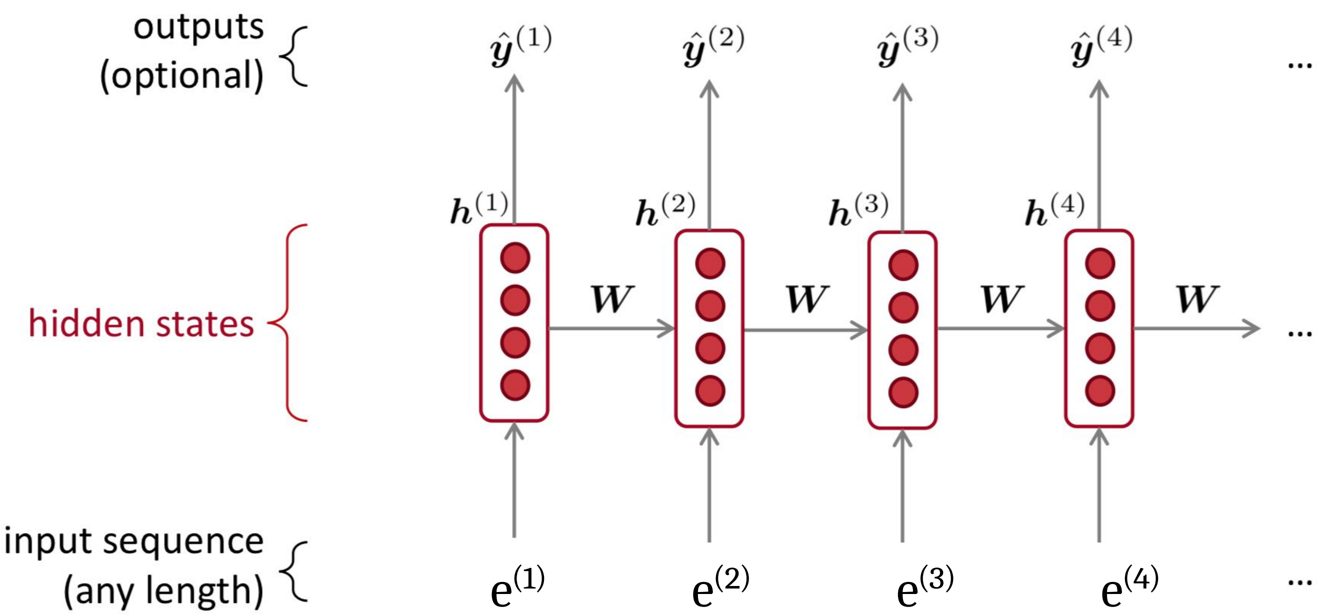
What’s wrong with our model?

- Window size is fixed
- Window size can never be big enough
- Weights are not shared between timesteps

Questions?

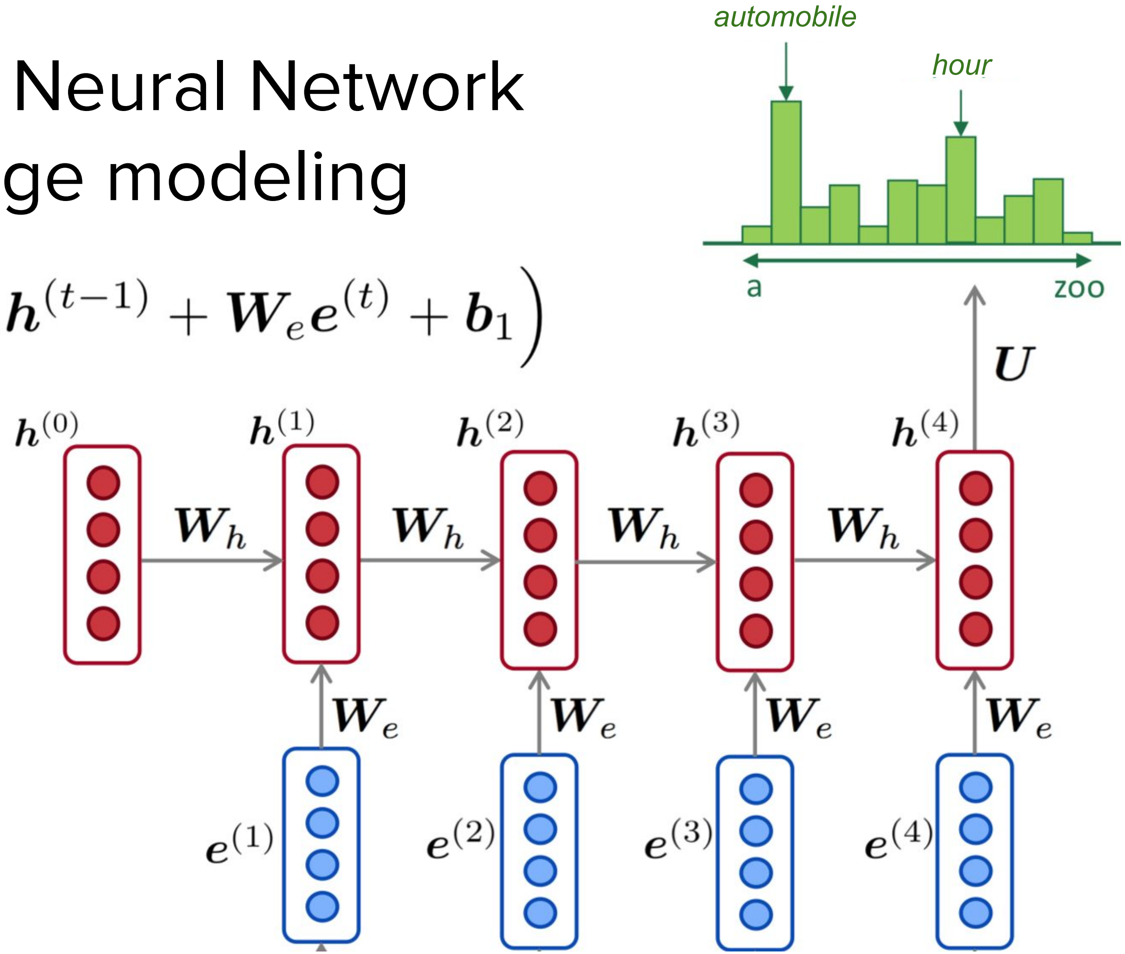
What if we share weights across timesteps?

Deep Learning + NLP: Attempt #2
Recurrent Neural Network



Recurrent Neural Network for language modeling

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h\boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e\boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$$



What’s wrong with our model?

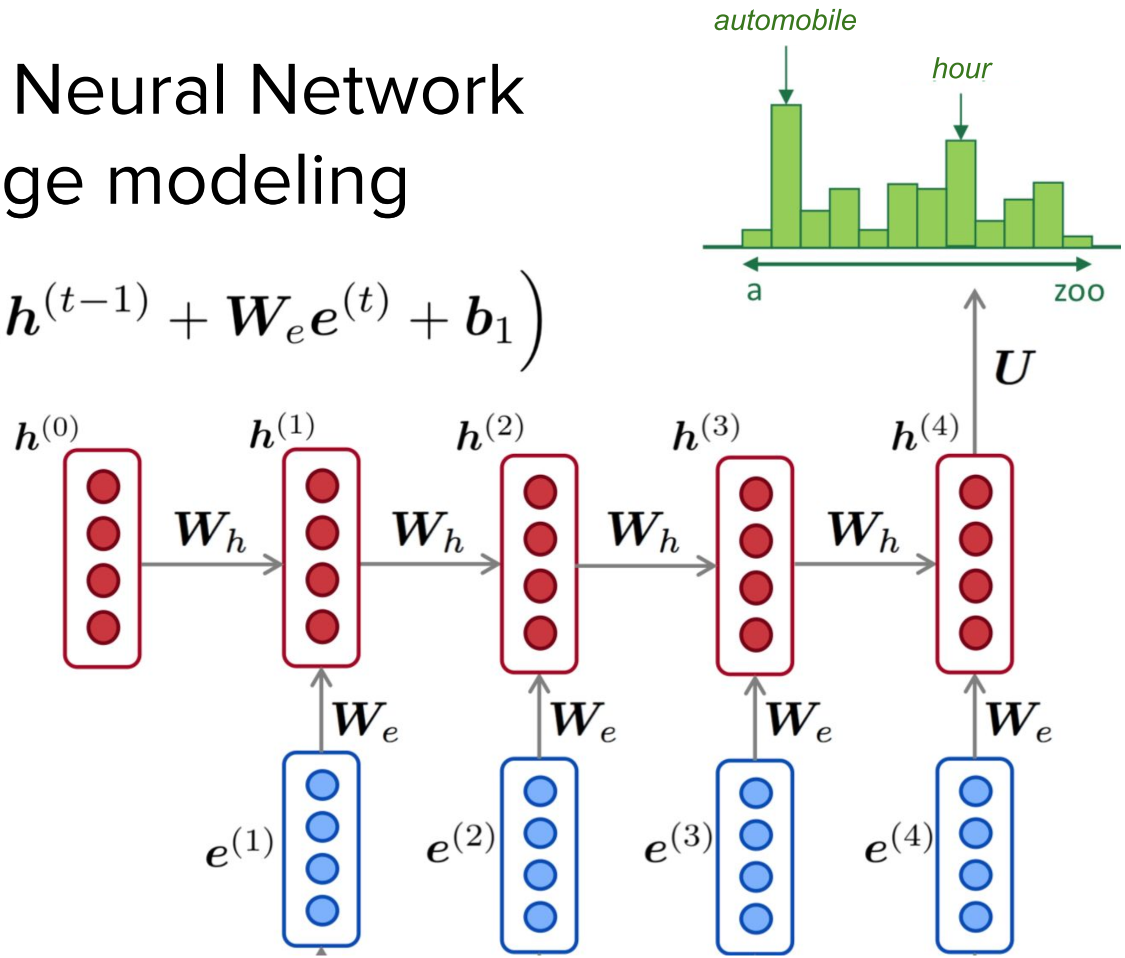
- ~~Window size is fixed~~
- ~~Window size can never be big enough~~
- ~~Weights are not shared between timesteps~~

Questions?

Class Exercises Part 2: RNN Warmup

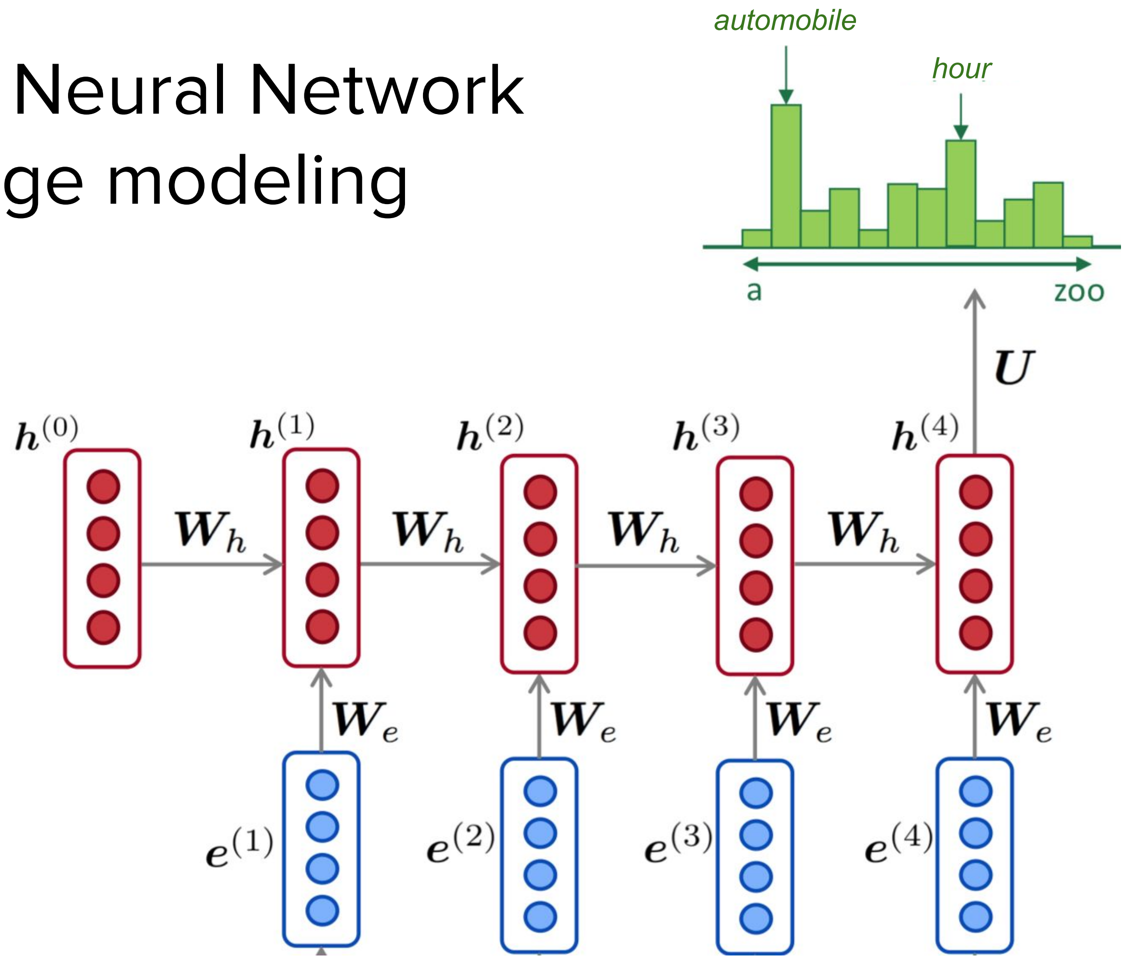
Recurrent Neural Network for language modeling

$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$$



How do we train these weights?

Recurrent Neural Network for language modeling



What’s wrong with our model?

- In practice, it’s difficult for the model to “remember” what it has seen many timesteps ago
    - “Vanishing gradients”

Questions?

RNN Variants!

Solution: use different hidden “cells”!

- Vanilla RNN: $\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$
- Gated Recurrent Unit (GRU)
- Long Short-Term Memory (LSTM)

Solution: use different hidden “cells”!

- Vanilla RNN: $\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h\boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e\boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$
- Gated Recurrent Unit (GRU)
- **Long Short-Term Memory (LSTM)**

LSTM

Input gate: $i_t = \sigma\left(W^{(i)}x_t + U^{(i)}h_{t-1}\right)$

Forget gate: $f_t = \sigma\left(W^{(f)}x_t + U^{(f)}h_{t-1}\right)$

Output gate: $o_t = \sigma\left(W^{(o)}x_t + U^{(o)}h_{t-1}\right)$

New memory: $\tilde{c}_t = \tanh\left(W^{(c)}x_t + U^{(c)}h_{t-1}\right)$

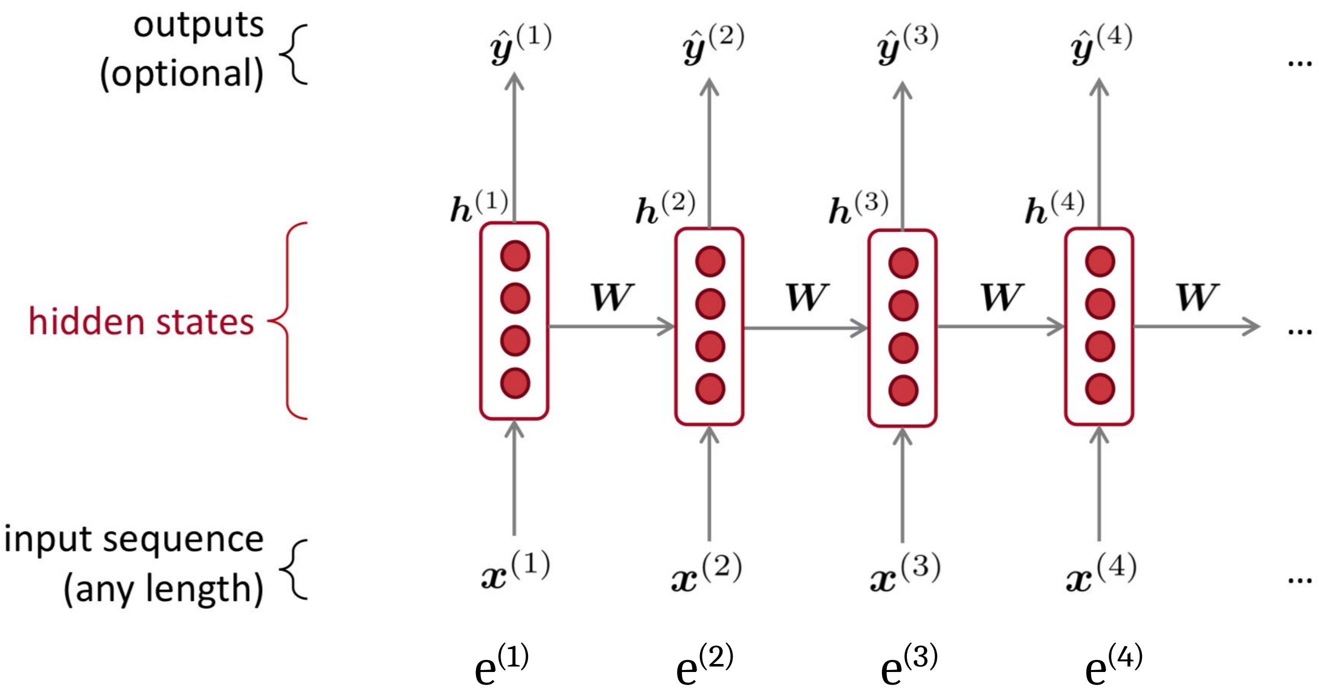
Final memory: $c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$

Final state: $h_t = o_t \circ \tanh(c_t)$

What’s wrong with our model?

- ~~In practice, it’s difficult for the model to “remember” what it has seen many timesteps ago~~

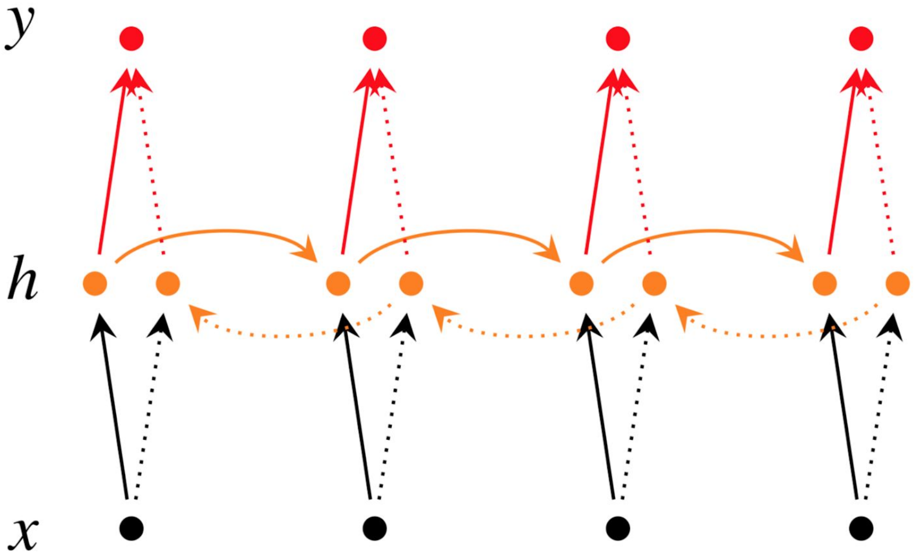
Outputs can be at every step!



What’s wrong with our model?

- ~~In practice, it’s difficult for the model to “remember” what it has seen many timesteps ago~~
- Intermediate steps don’t have access to inputs from future steps

Bidirectional RNN



$$\overrightarrow{h}_t = f(\overrightarrow{W} x_t + \overrightarrow{V}\overrightarrow{h}_{t-1} + \overrightarrow{b})$$

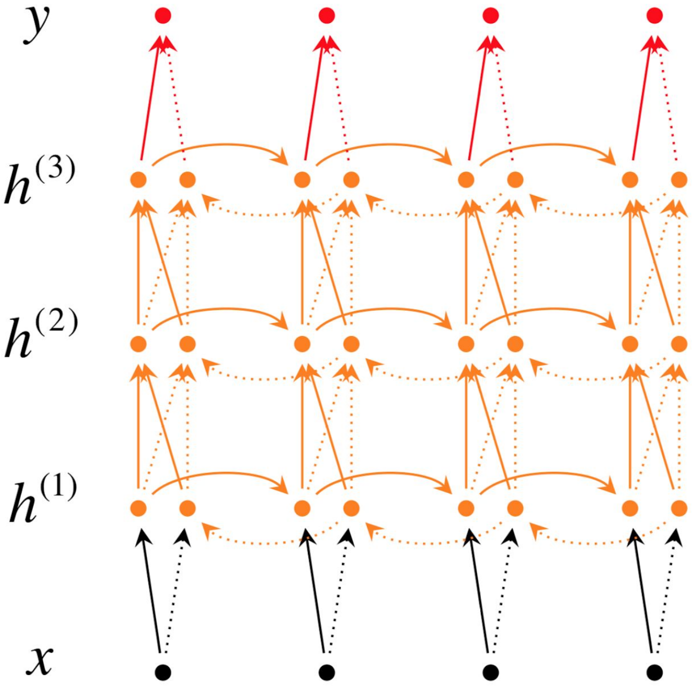
$$\overleftarrow{h}_t = f(\overleftarrow{W} x_t + \overleftarrow{V}\overleftarrow{h}_{t+1} + \overleftarrow{b})$$

$$y_t = g(U[\overrightarrow{h}_t;\overleftarrow{h}_t] + c)$$

$h = [\overrightarrow{h};\overleftarrow{h}]$ now represents (summarizes) the past and future

Questions?

Deep Bidirectional RNN



$$\overrightarrow{h}_t^{(i)} = f(\overrightarrow{W}^{(i)} h_t^{(i-1)} + \overrightarrow{V}^{(i)} \overrightarrow{h}_{t-1}^{(i)} + \overrightarrow{b}^{(i)})$$

$$\overleftarrow{h}_t^{(i)} = f(\overleftarrow{W}^{(i)} h_t^{(i-1)} + \overleftarrow{V}^{(i)} \overleftarrow{h}_{t+1}^{(i)} + \overleftarrow{b}^{(i)})$$

$$y_t = g(U[\overrightarrow{h}_t^{(L)}; \overleftarrow{h}_t^{(L)}] + c)$$

Questions?

Practical Tips

- Don’t use a “vanilla” RNN
- LSTMs generally work well for most tasks
- Use bidirectional whenever it makes sense
- Don’t stack too many layers (too computationally expensive)

Class Exercises Part 3: Generating Fake News

Homework:
Fake News Evaluation

Summary of Today

- Introduction to natural language processing using machine learning
- Language modeling
- Recurrent neural networks
- RNN variants

Questions?