CS21si: AI for Social Good

Lecture 5: NLP for Social Good

Plan for Today

- Fake news
- Natural language processing
- Language models
- Recurrent neural networks

Fake news

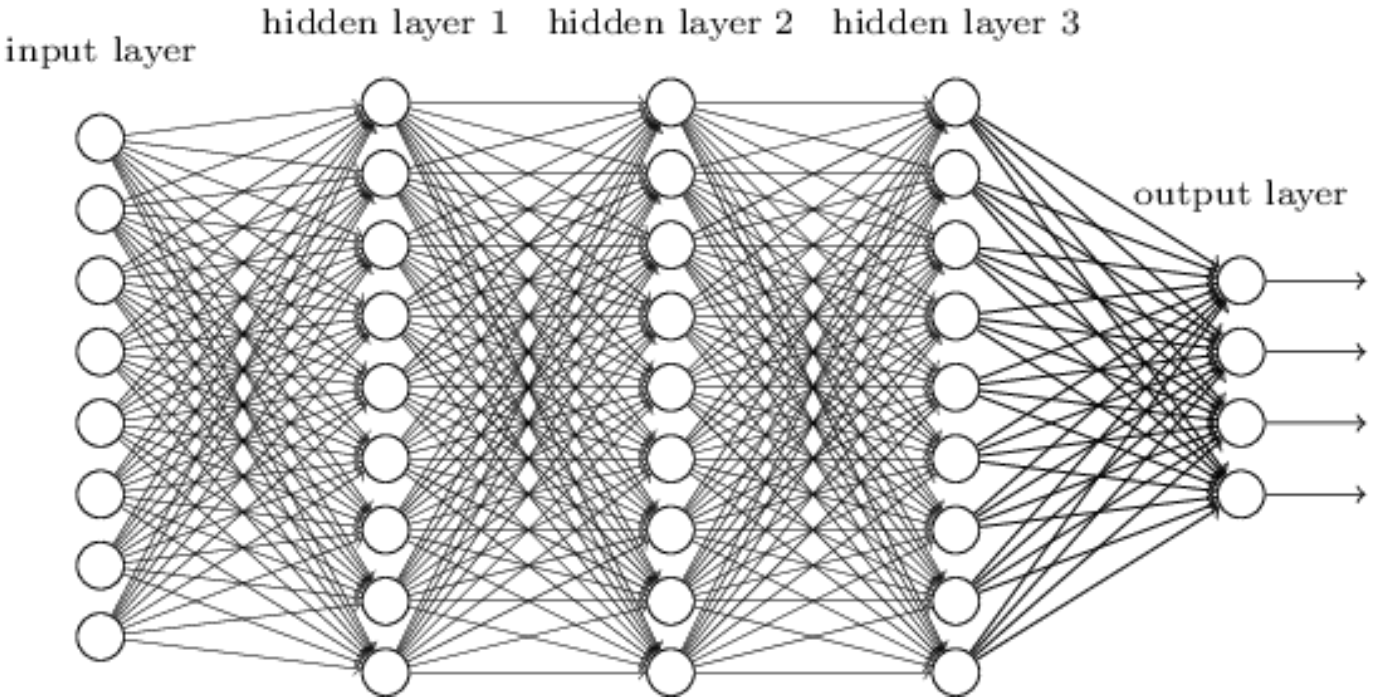

WEEKLY WORLD
NEWS
Coast-to-coast manhunt for cave boy captured just weeks ago!
BAT CHILD ESCAPES!
He's 2-feet-tall and VERY DANGEROUS, warn police


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

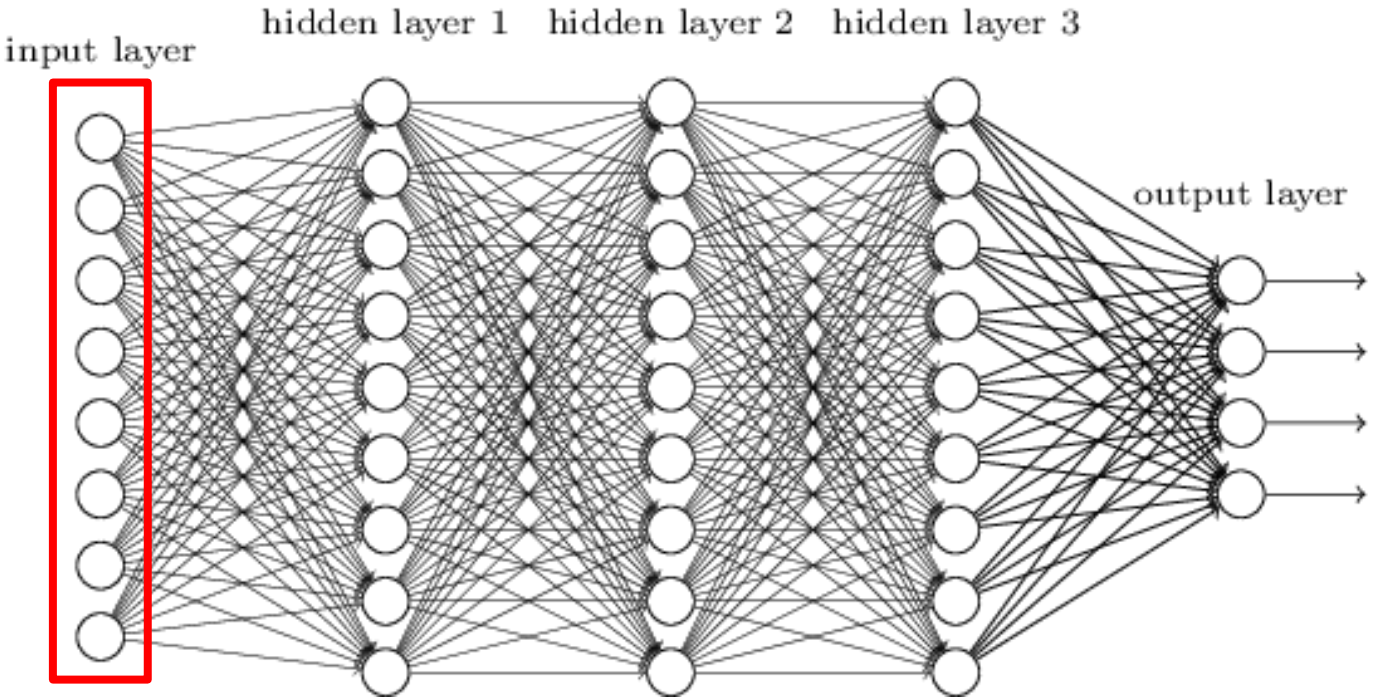
Our Dataset

How do we deal with text data?

Deep Neural Networks



Deep Neural Networks

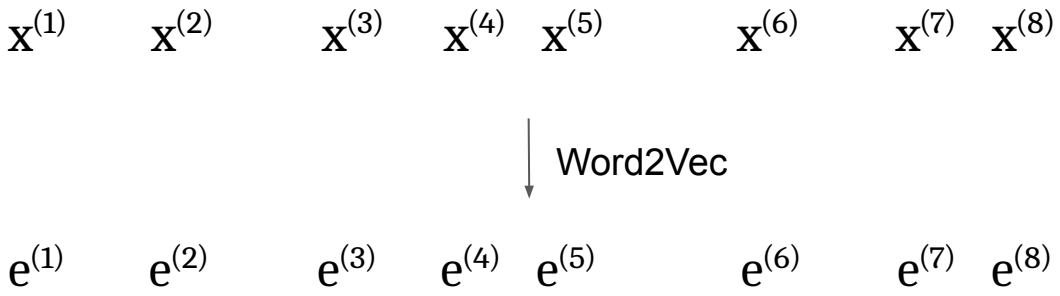


I'll meet you at the airport in an ...

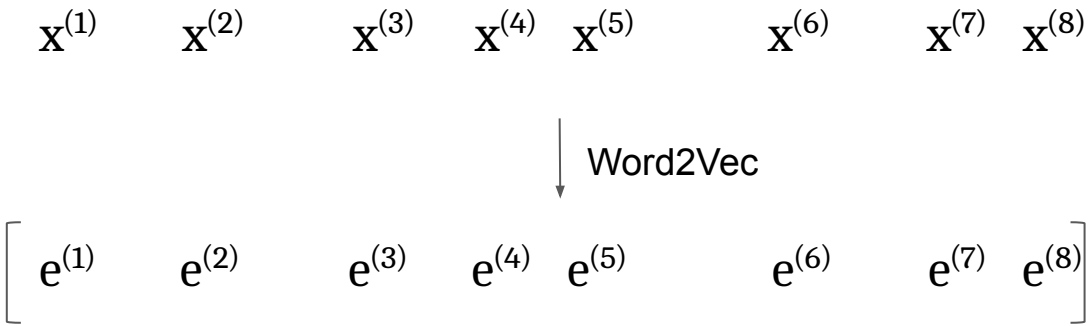
I'll meet you at the airport in an ...

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

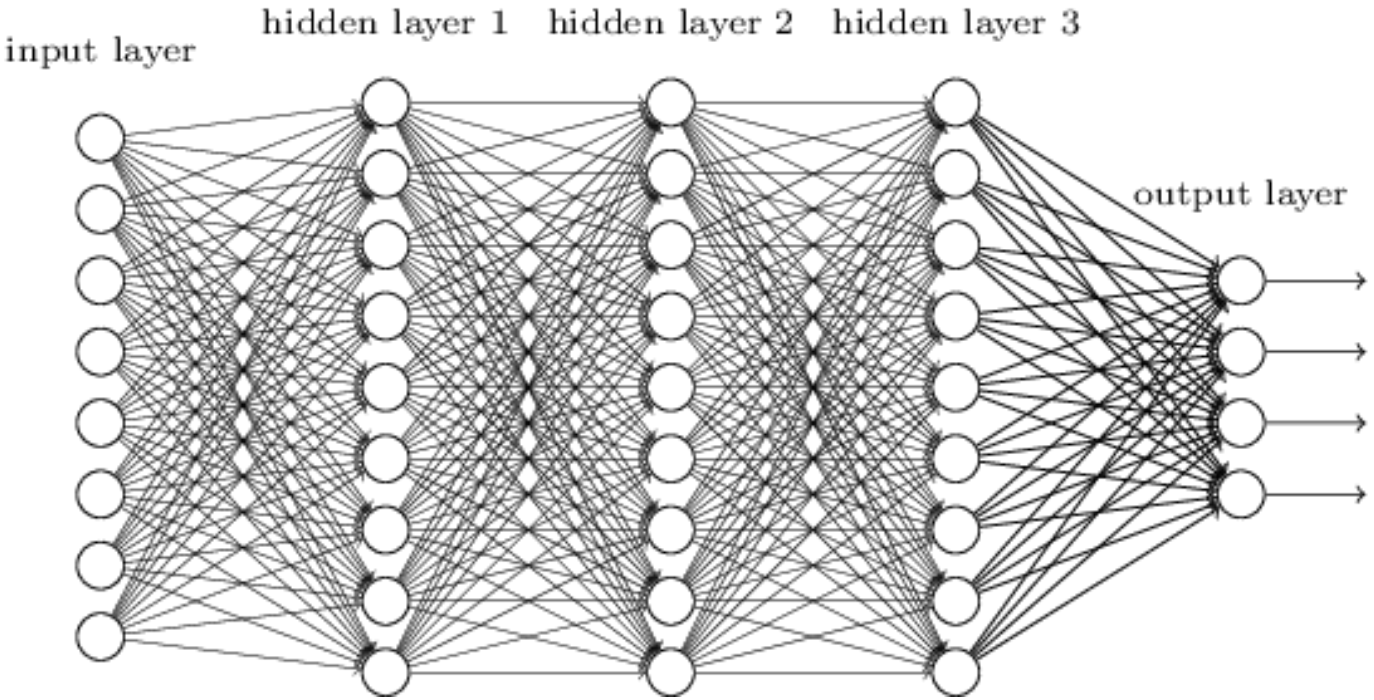
I'll meet you at the airport in an ...



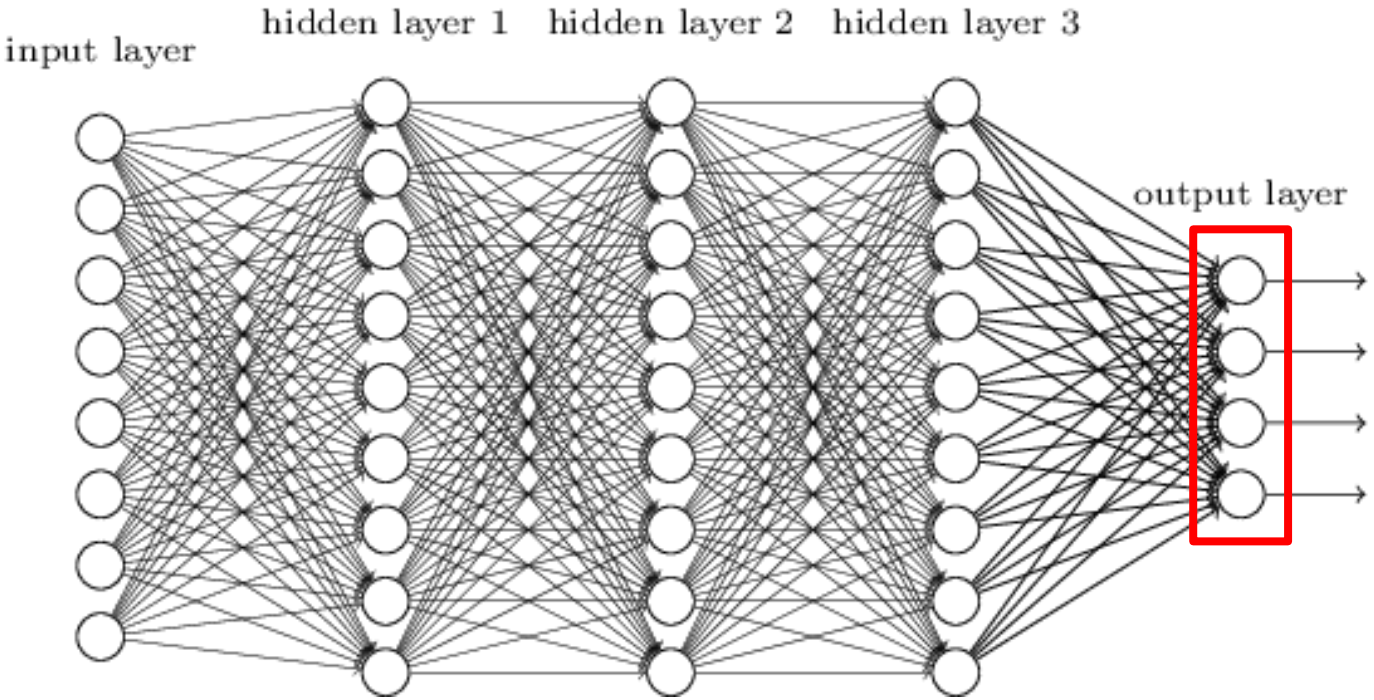
I'll meet you at the airport in an ...



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



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



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

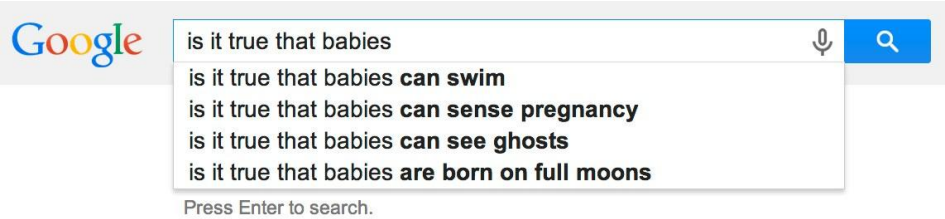
More formally: given a sequence of words $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots, \boldsymbol{x}^{(t)}$, compute the probability distribution of the next word $\boldsymbol{x}^{(t+1)}$:

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

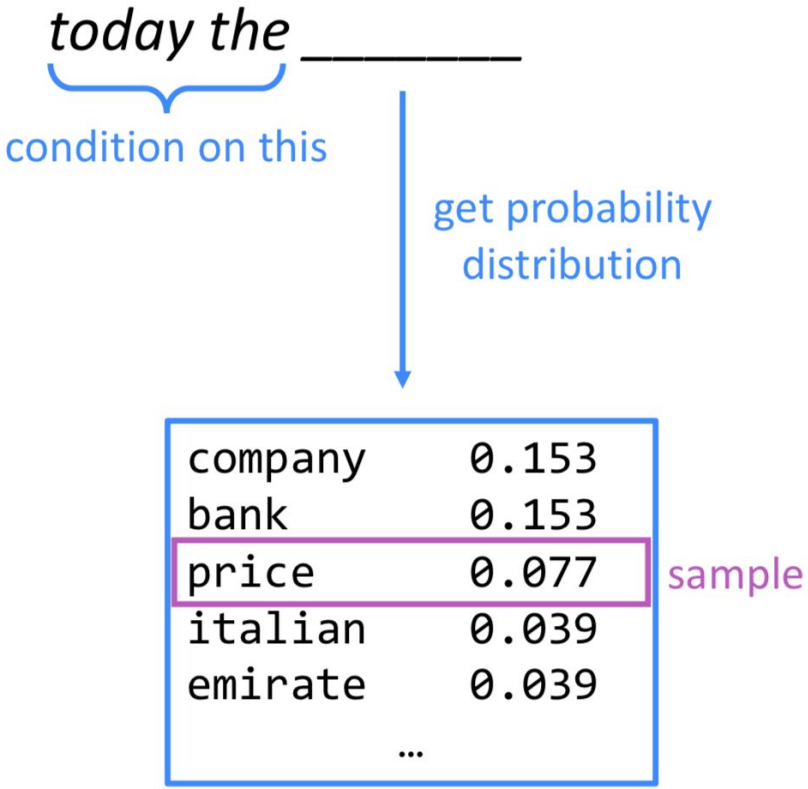
where $\boldsymbol{w}_j$ is a word in the vocabulary $V = \{\boldsymbol{w}_1, \ldots, \boldsymbol{w}_{|V|}\}$

Questions?

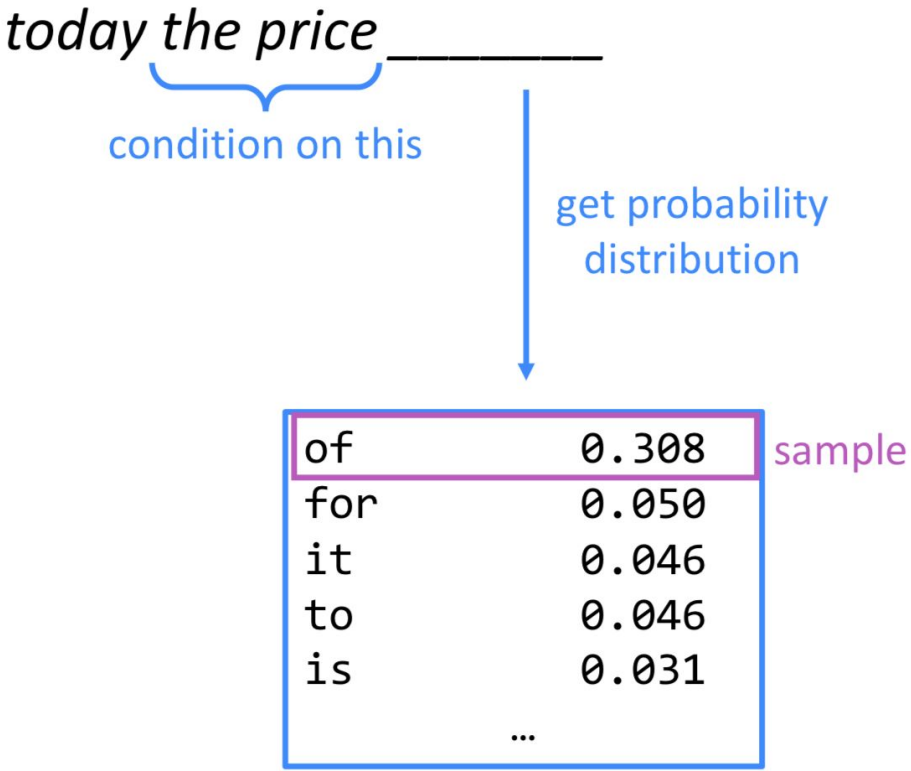
You use language models every day!



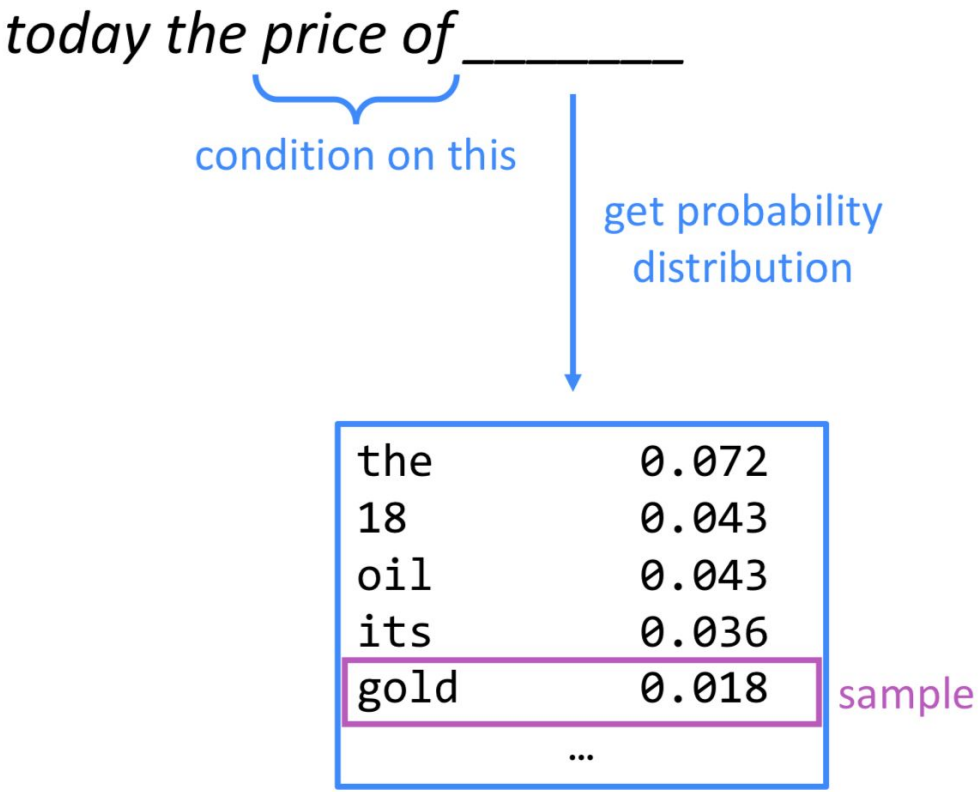
You can use language models to generate new text!



You can use language models to generate new text!



You can use language models to generate new text!



Better Language Models and Their Implications (OpenAI)

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

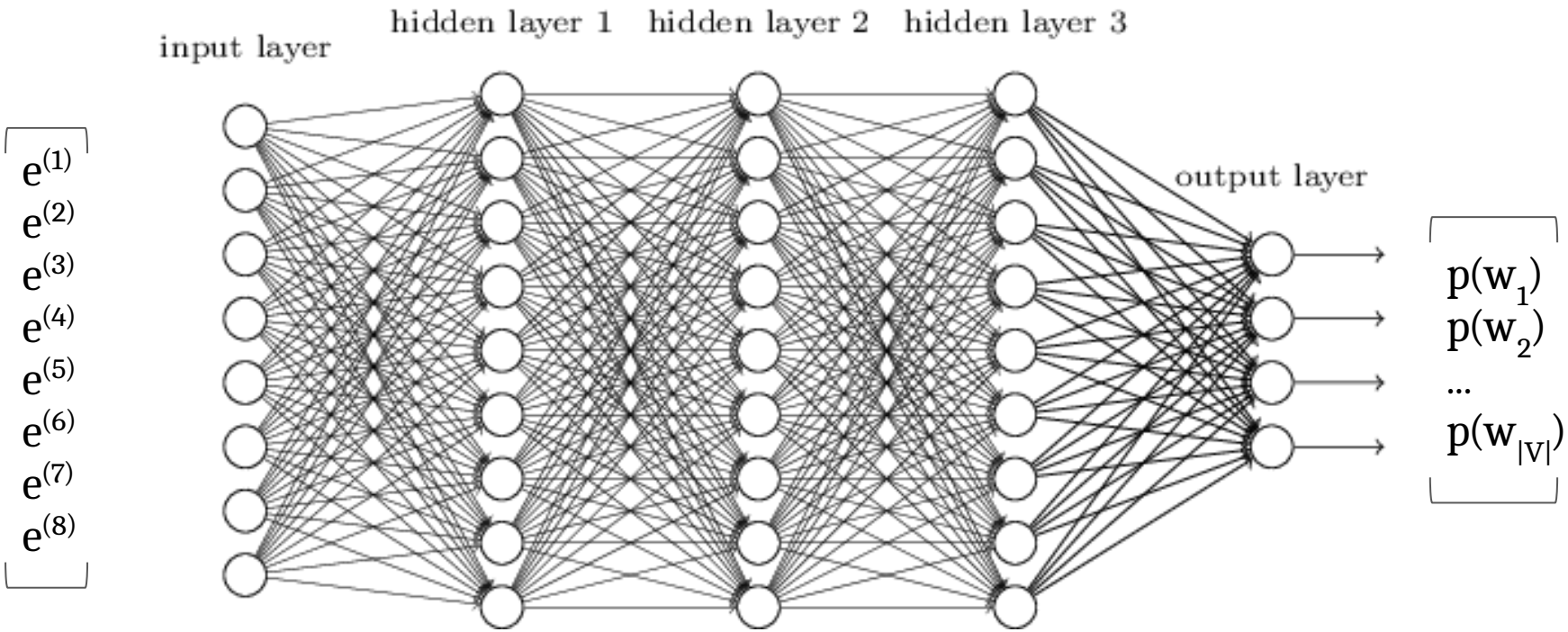
The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

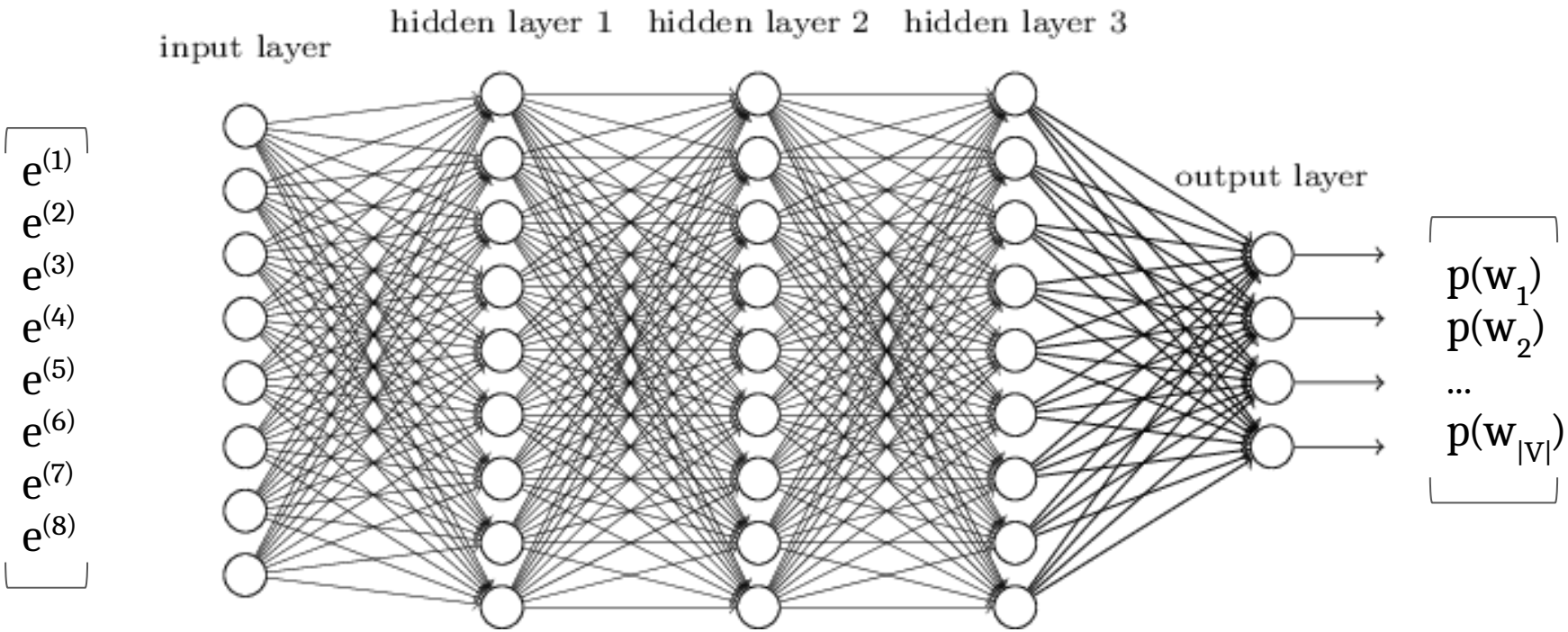
https://openai.com/blog/better-language-models/

Deep Learning + NLP: Attempt #1



Class Exercises Part 1: Neural NLP Warmup

Deep Learning + NLP: First Attempt



What's wrong with our model?

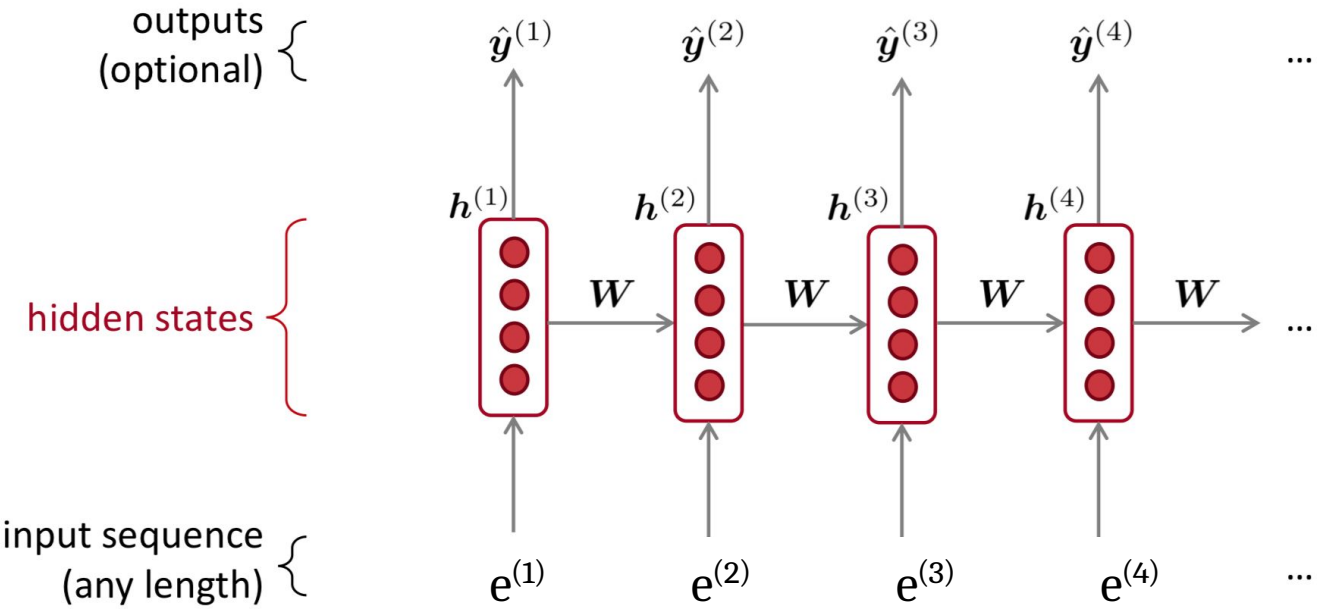
What's wrong with our model?

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

Questions?

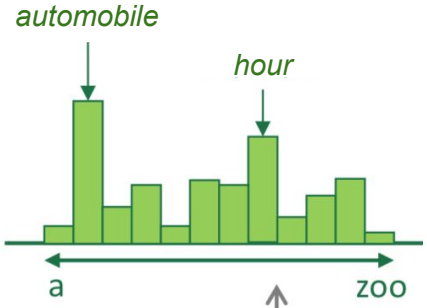
What if we share weights across timesteps?

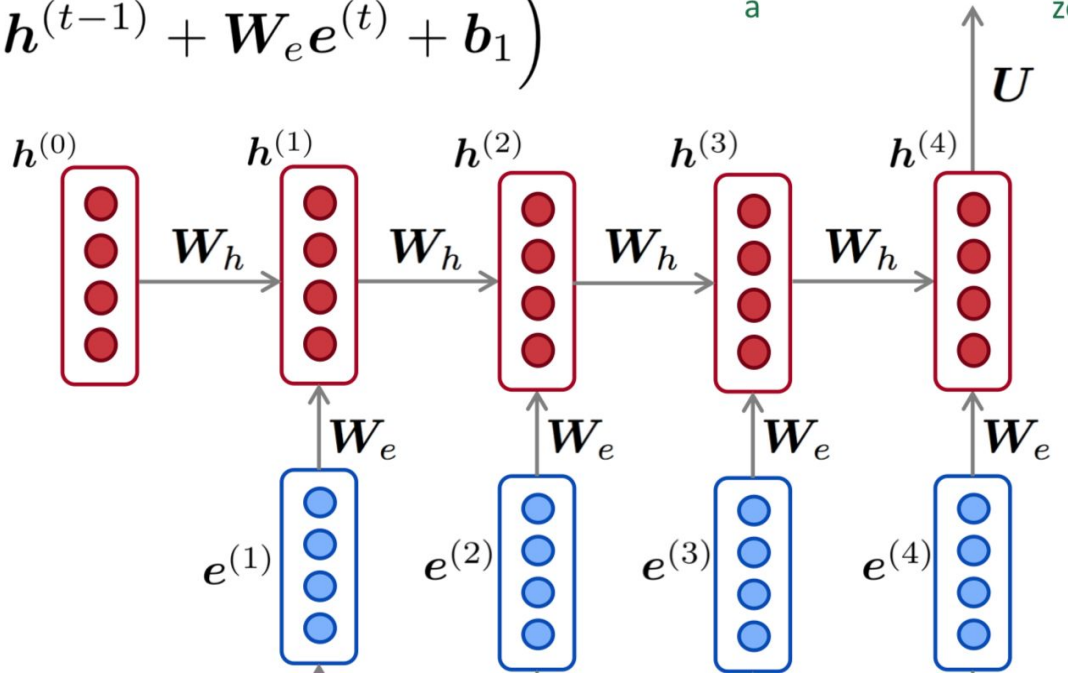
Deep Learning + NLP: Attempt #2
Recurrent Neural Network



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?

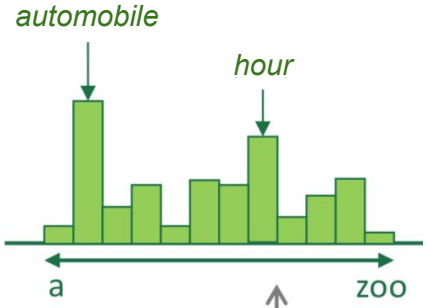
- ~~Window size is fixed~~
- ~~Window size can never be big enough~~
- ~~Weights are not shared between timesteps~~

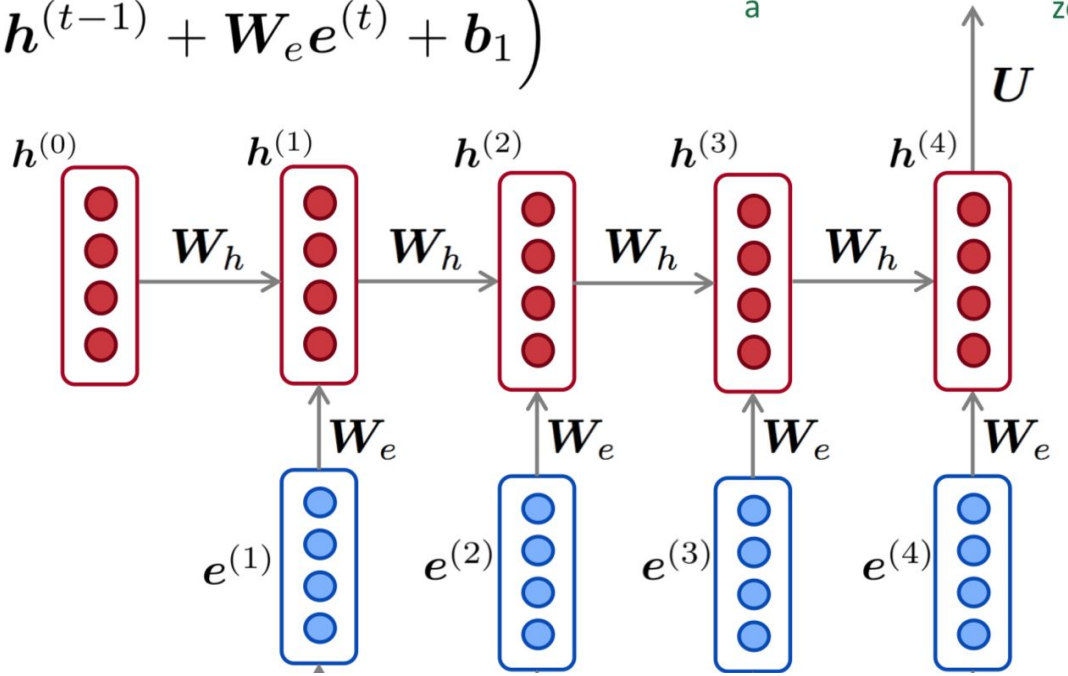
Questions?

Class Exercises Part 2: RNN Warmup

Recurrent Neural Network for language modeling

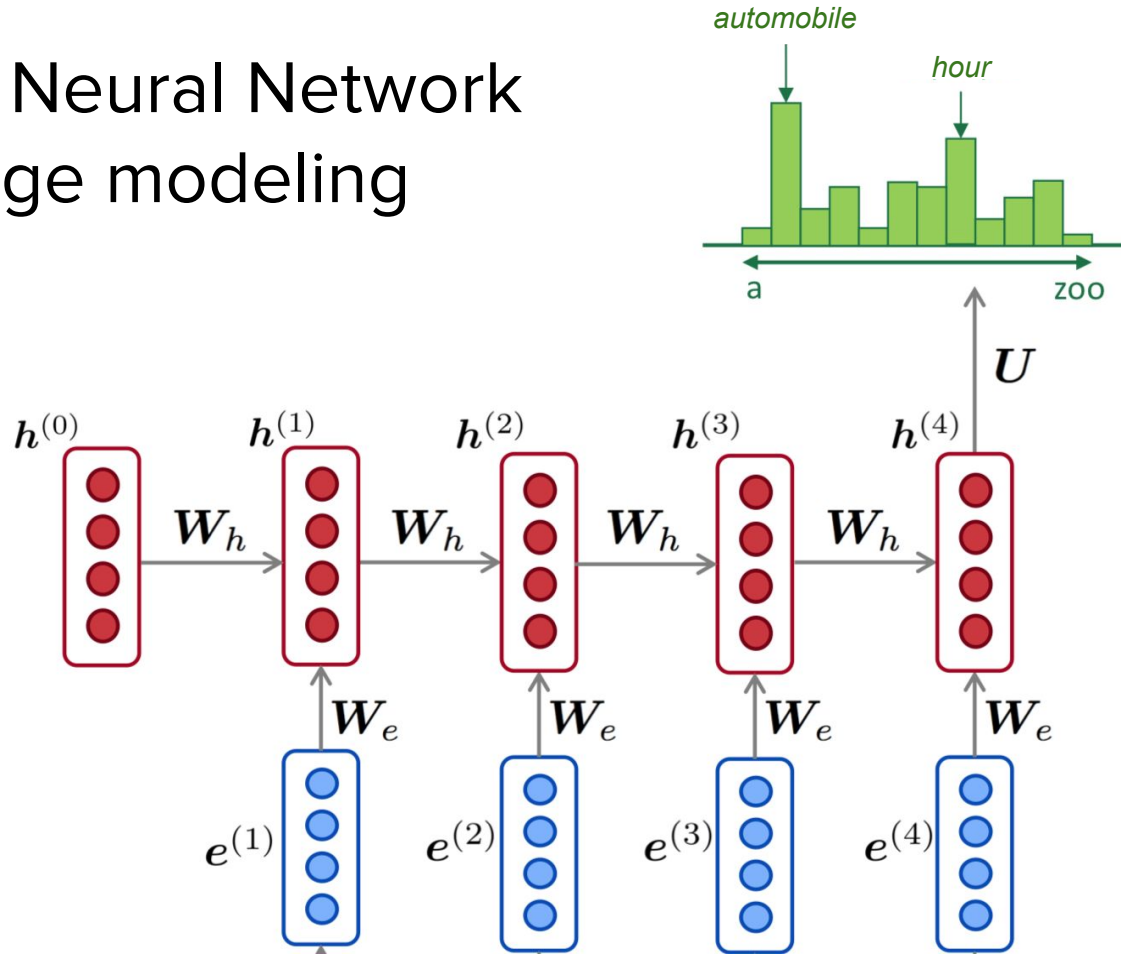
$$\boldsymbol{h}^{(t)} = \sigma\left(\boldsymbol{W}_h \boldsymbol{h}^{(t-1)} + \boldsymbol{W}_e \boldsymbol{e}^{(t)} + \boldsymbol{b}_1\right)$$





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

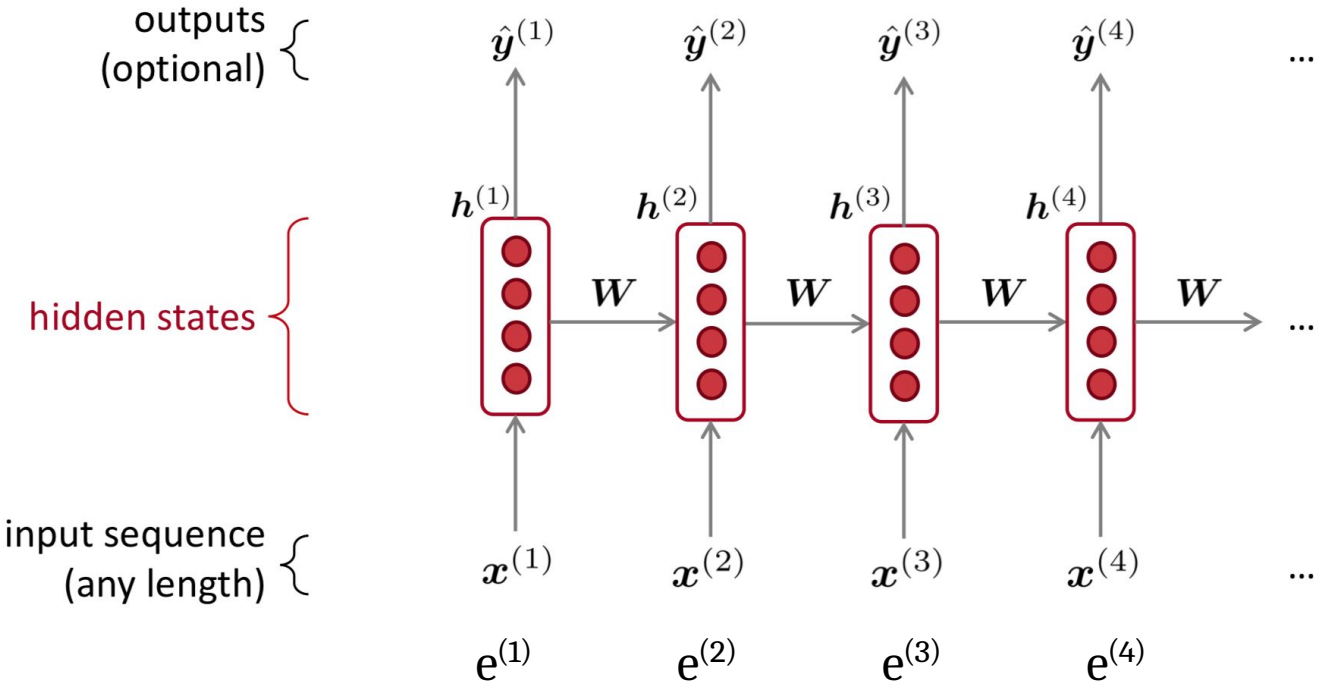
Final memory: $$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$

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~~

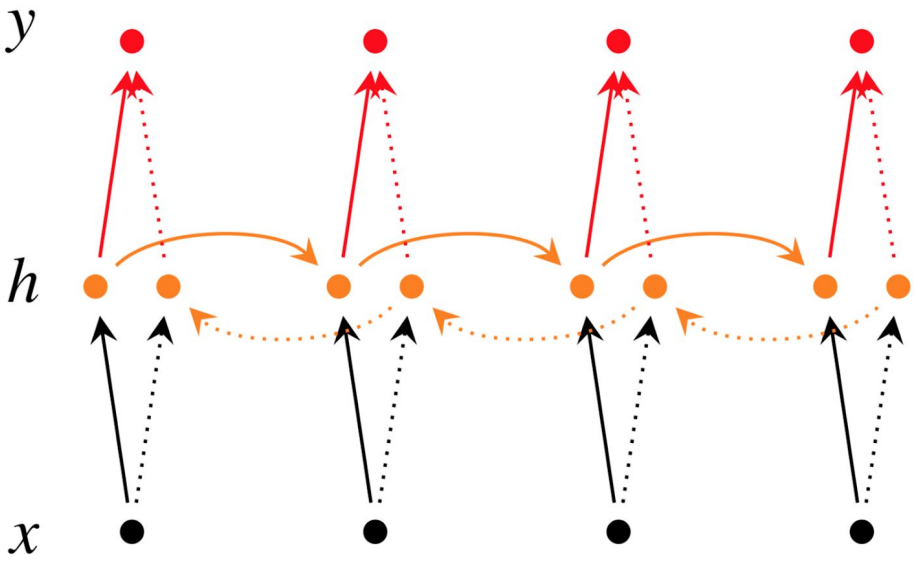
Outputs can be at every step!



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



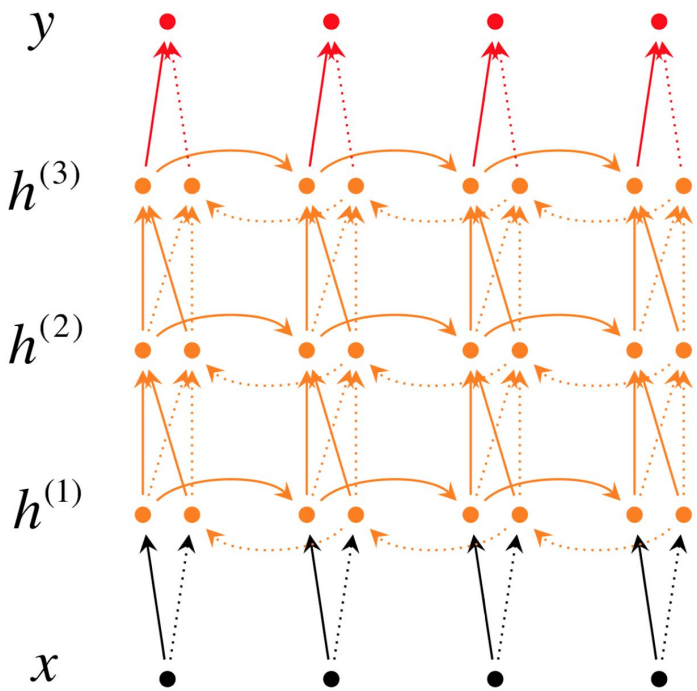
$$\vec{h}_t = f(\vec{W}x_t + \vec{V}\vec{h}_{t-1} + \vec{b})$$

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

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

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

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?

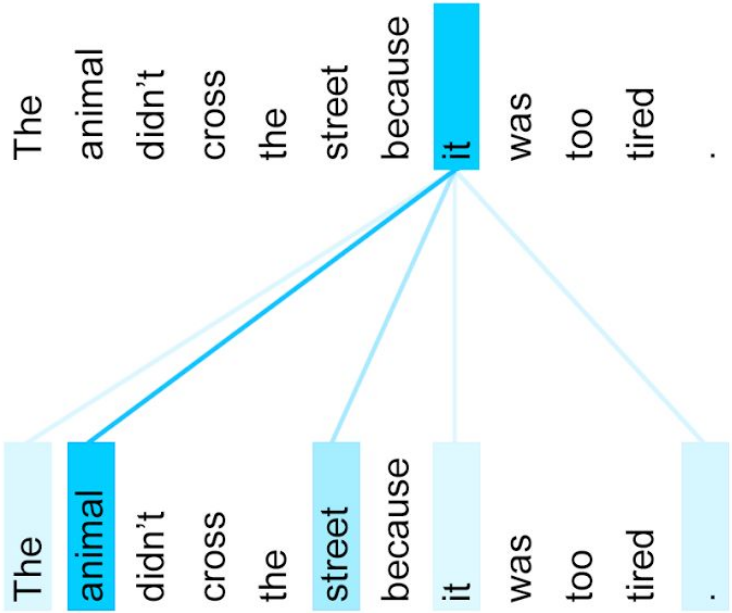
Practical RNN Tips

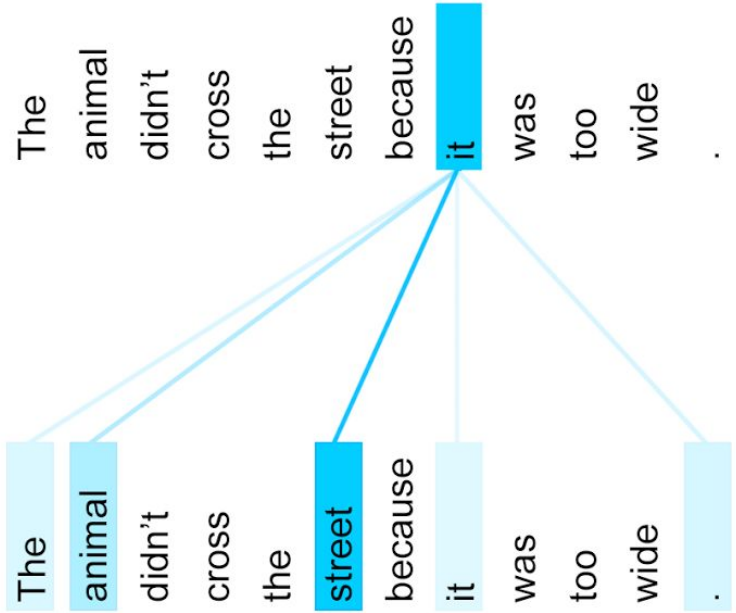
- Don’t use a “vanilla” RNN
- LSTMs generally work well for most tasks
- Use bidirectional whenever it makes sense
- Don’t stack too many layers (too computationally expensive)

What's wrong with our model?

- ~~In practice, it's difficult for the model to "remember" what it has seen many timesteps ago~~ Still a problem!
- Not parallelizable!

Transformers





Better Language Models and Their Implications (OpenAI)

SYSTEM PROMPT (HUMAN-WRITTEN)

In a shocking finding, scientist discovered a herd of unicorns living in a remote, previously unexplored valley, in the Andes Mountains. Even more surprising to the researchers was the fact that the unicorns spoke perfect English.

MODEL COMPLETION (MACHINE-WRITTEN, 10 TRIES)

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science.

Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved.

Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. Pérez noticed that the valley had what appeared to be a natural fountain, surrounded by two peaks of rock and silver snow.

https://openai.com/blog/better-language-models/

Class Exercises Part 3: Generating Fake News

Homework:
Fake News Evaluation

Summary of Today

- Introduction to NLP
- Language modeling of fake news
- Recurrent neural networks and variants
- Transformers

Questions?