Pretraining

Jacob Andreas / MIT 6.804-6.864 / Spring 2020

Admin

Welcome to online 6.806-864!

OH will take place on WebEx: link has been emailed.
(Please continue to sign up for time slots.)

Take-home midterm: instructions and makeup sign-up to be emailed.

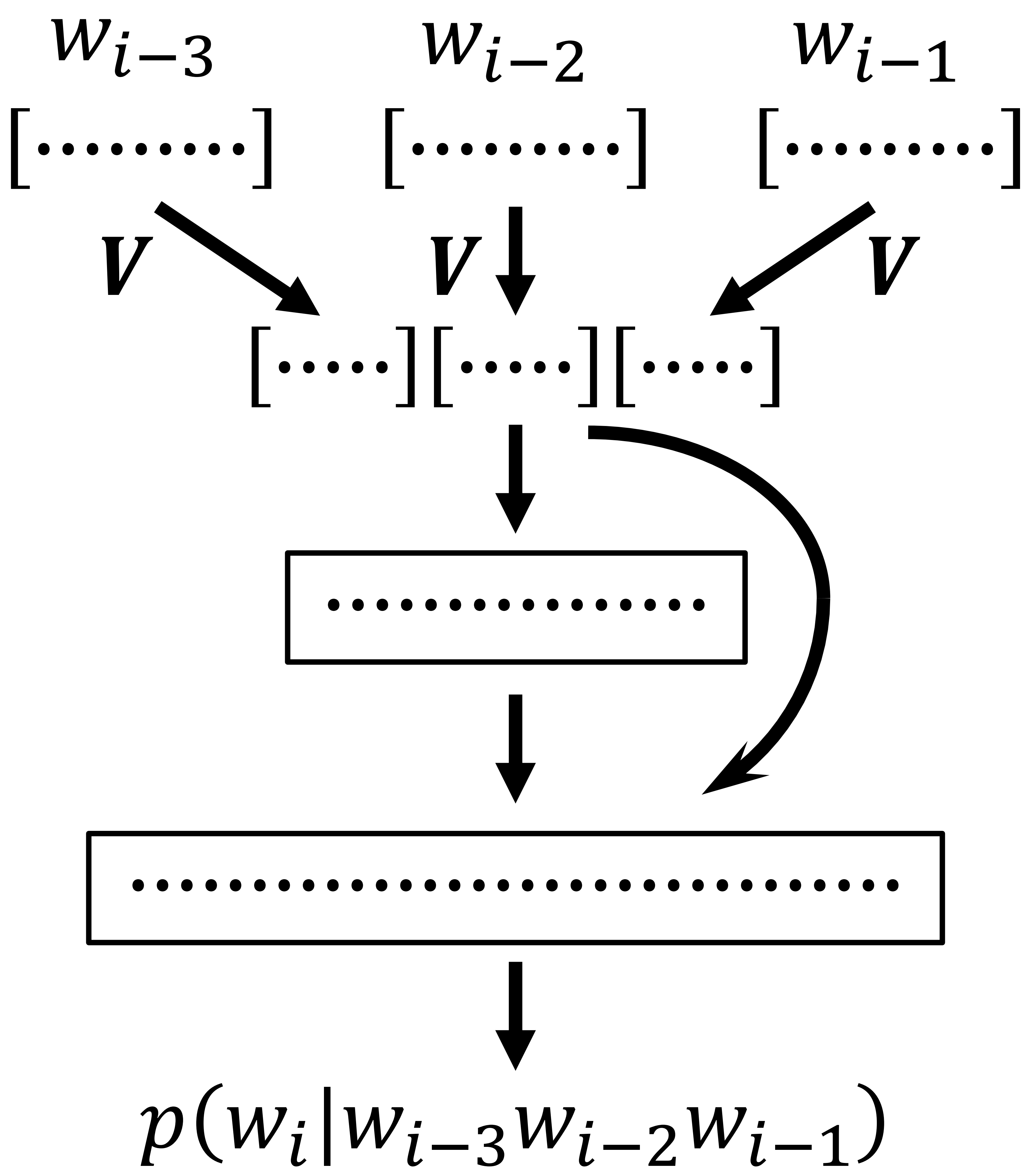
Practice midterm Qs posted tonight.

Recap: neural sequence models

Language modeling with feedforward networks

- Associate a distributed vector per word
- Express the joint probability function of word sequences in terms of the vectors
- Simultaneously learn word vectors and parameters of the probability function

- Implemented as feed-forward network
- Shared vector mapping, $\boldsymbol{V}$, for all words
- First layer concatenated context vectors
- Perplexity improvements on Brown and AP News corpora over best n-grams

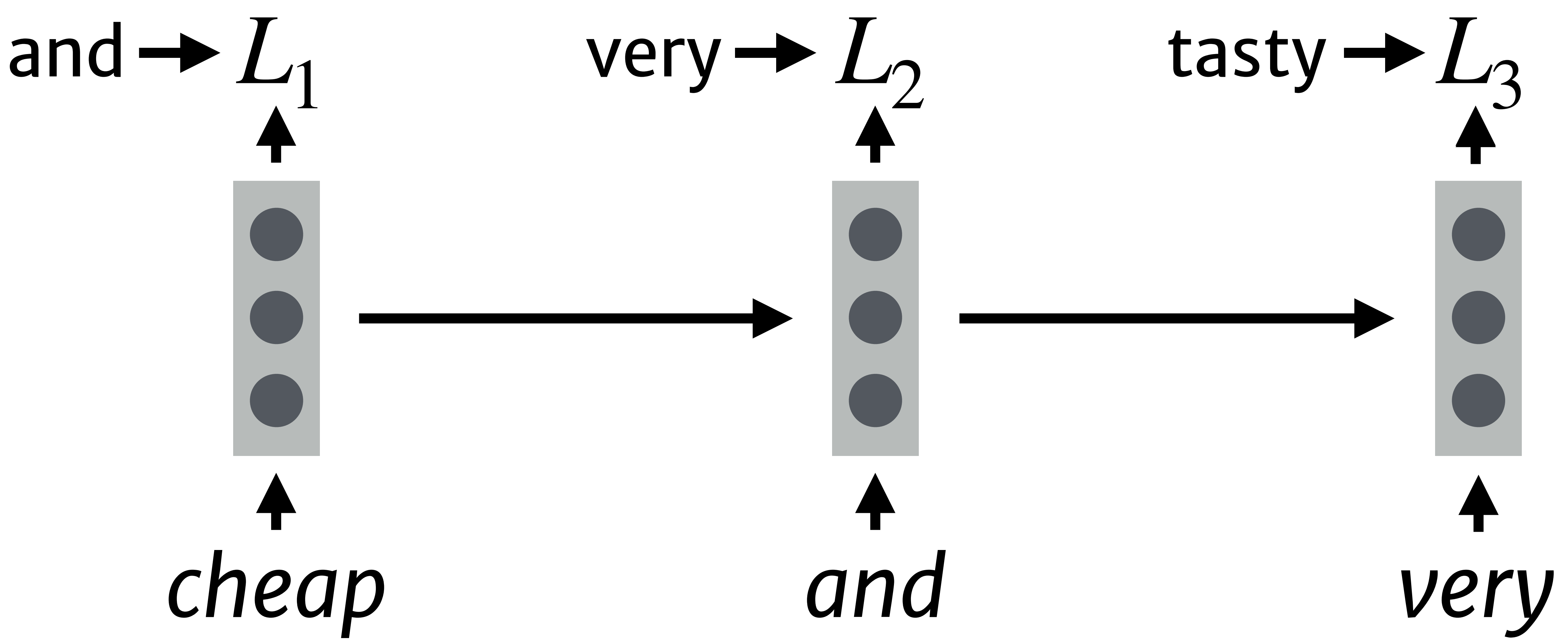


[A Neural Probabilistic Language Model. Bengio, 2003]

Language modeling with RNNs

A (unidirectional) RNN can compute $p(y_t \mid x_{:t})$.

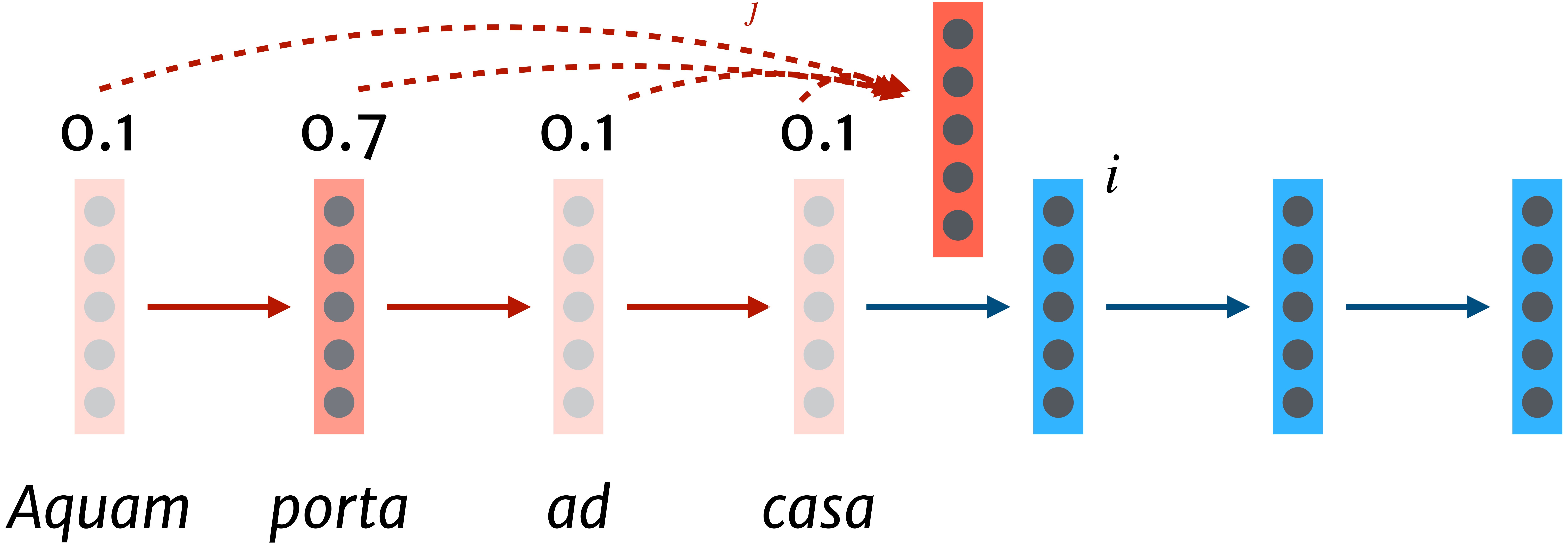
Suppose for a sequence x we set $y_t = x_{t+1}$.



then $\sum_t \log p(x_{t+1} \mid x_{:t}) = p(x)$

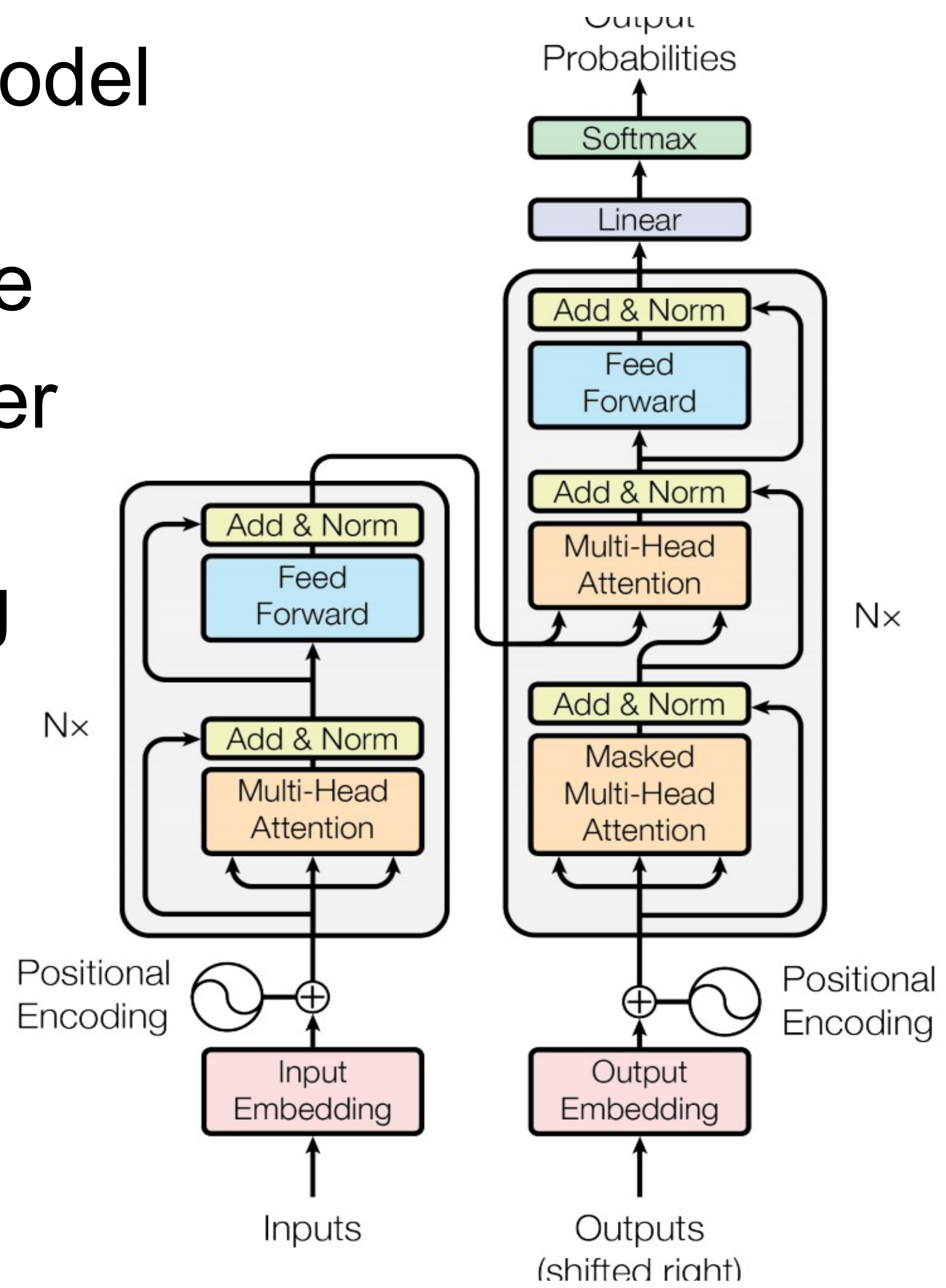
Attention mechanisms

1. When predicting output i, assign a weight α_{ij} to each encoder state h_j
2. Compute a pooled input $c_i = \sum_j \alpha_{ij} h_j$



Transformers

- Non-recurrent seq2seq (encoder-decoder) model
- Multi-layered attention model enables lateral information transfer across an input sequence
- Cost function is cross-entropy error of decoder
- Original paper demonstrated good results on machine translation and constituency parsing
- Transformers are the basis for BERT etc. (which we will see next week)

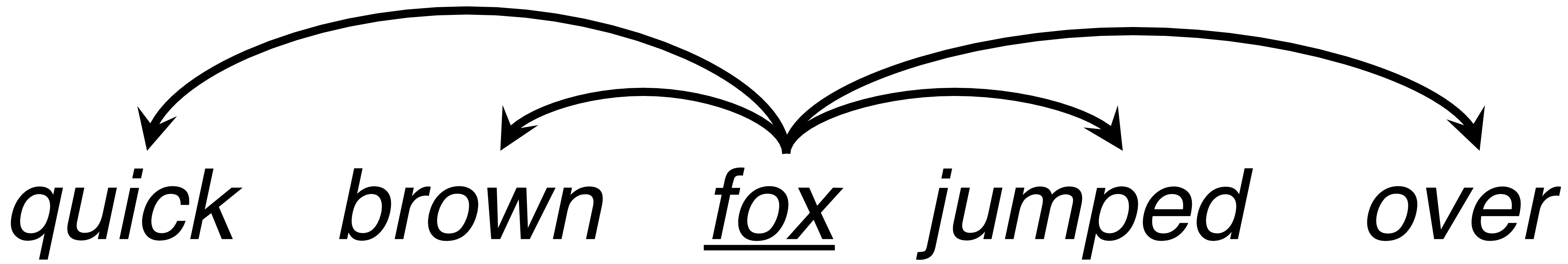


[Vaswani et al., "Attention is All You Need" arXiv:1706.03762 2017]

Recap: pretraining

Language modeling with word2vec

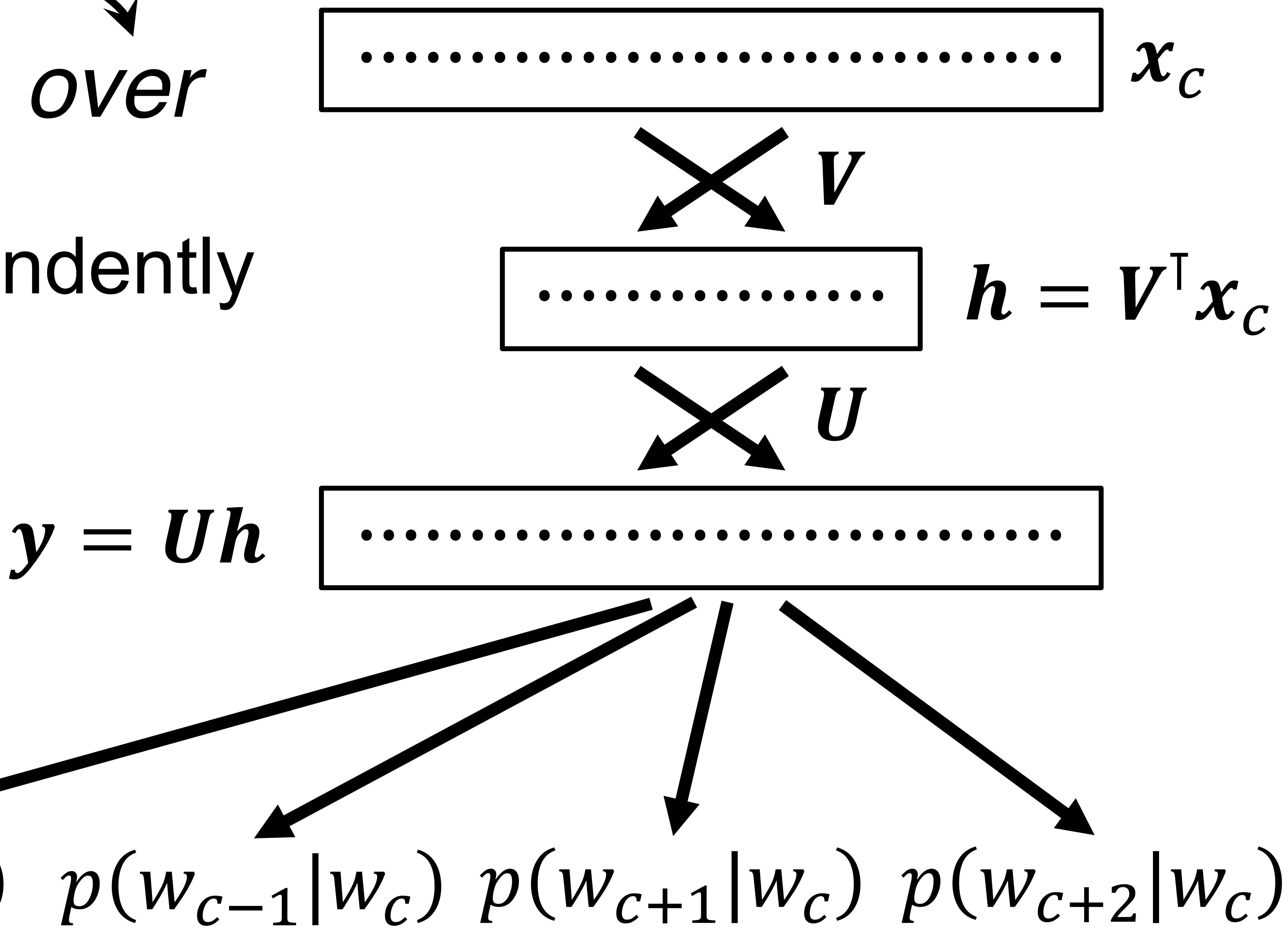
- Skip-gram predicts neighbor words from center word



- Each output is predicted independently

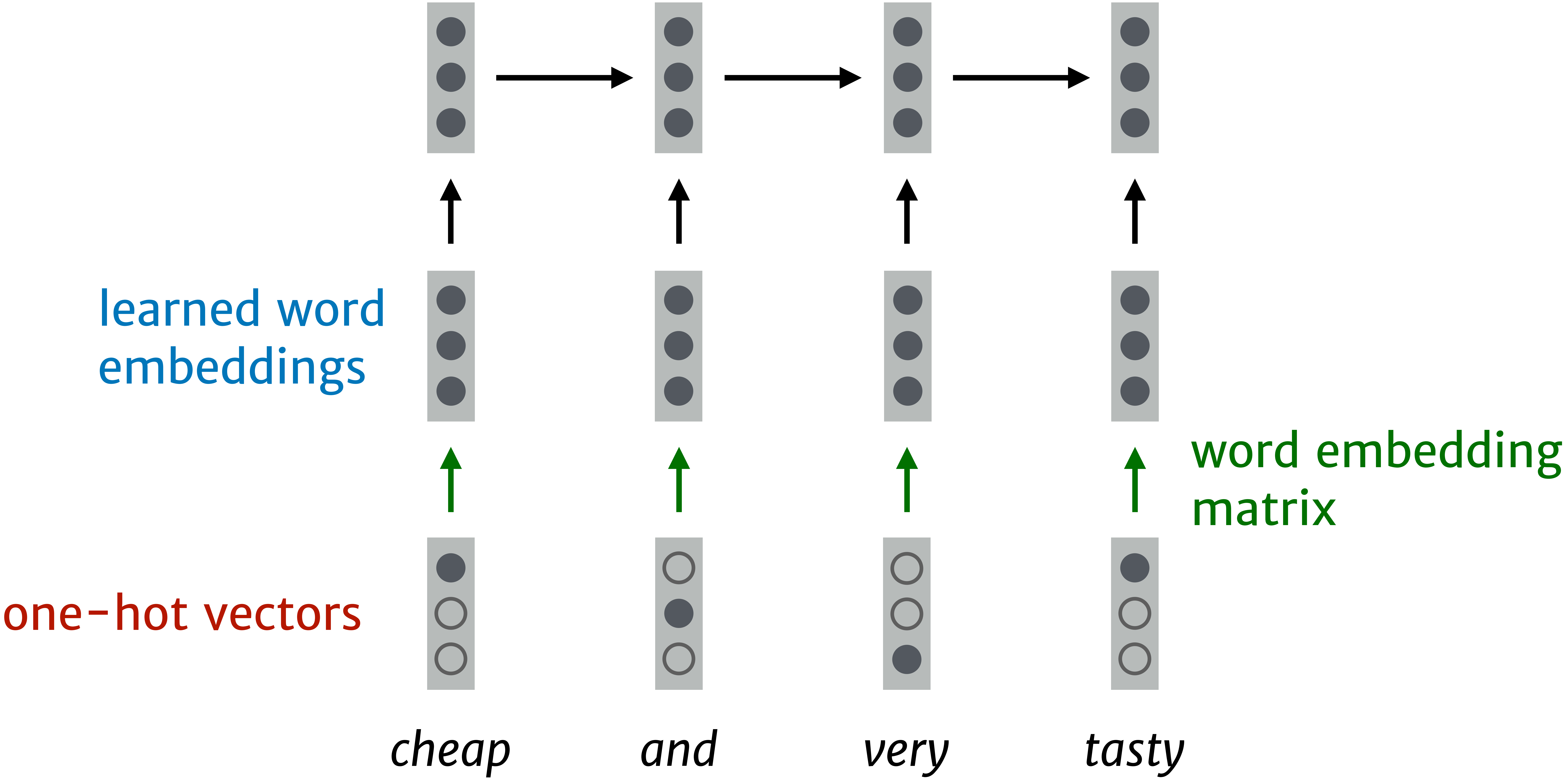
$$\prod_{\substack{-h \le n \le h \\ n \ne 0}} p(w_{c+n} | w_c)$$

- Context window lengths can be sampled



[Distributed Representations of Words and Phrases and their Compositionality. Mikolov et al., 2003]

RNNs and word embeddings



Homonyms

*I **can** run.*

*I **can** anchovies.*

Word senses

*I deposited money in the **bank**.*

*I climbed up the **bank** of the river.*

Word senses

*I'll meet you at the **bank**.*

*All my classmates work for **banks**.*

*She's a volunteer at the blood **bank**.*

Word senses

Definition of *do* (Entry 1 of 5)

transitive verb

1: to bring to pass : CARRY OUT

do another's wishes

2: PUT —used chiefly in *do to death*

3

a: PERFORM, EXECUTE

do some work

did his duty

b: COMMIT

crimes *done* deliberately

4

a: BRING ABOUT, EFFECT

trying to *do* good

do violence

b: to give freely : PAY

do honor to her memory

5: to bring to an end : FINISH —used in the past participle

the job is finally *done*

6: to put forth : EXERT

did her best to win the race

7

a: to wear out especially by physical exertion : EXHAUST

at the end of the race they were pretty well *done*

b: to attack physically : BEAT

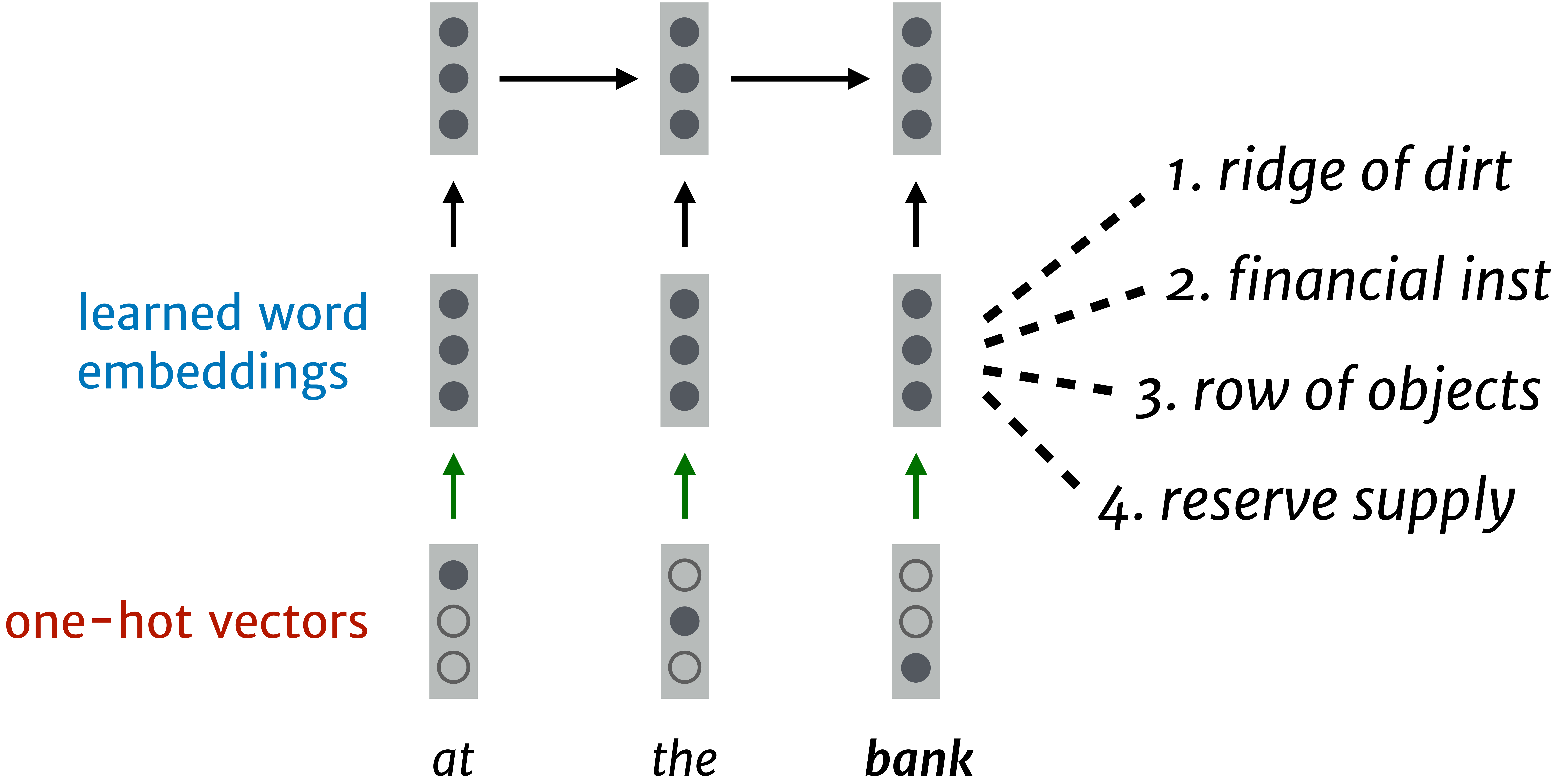
also **:** KILL

8: to bring into existence : PRODUCE

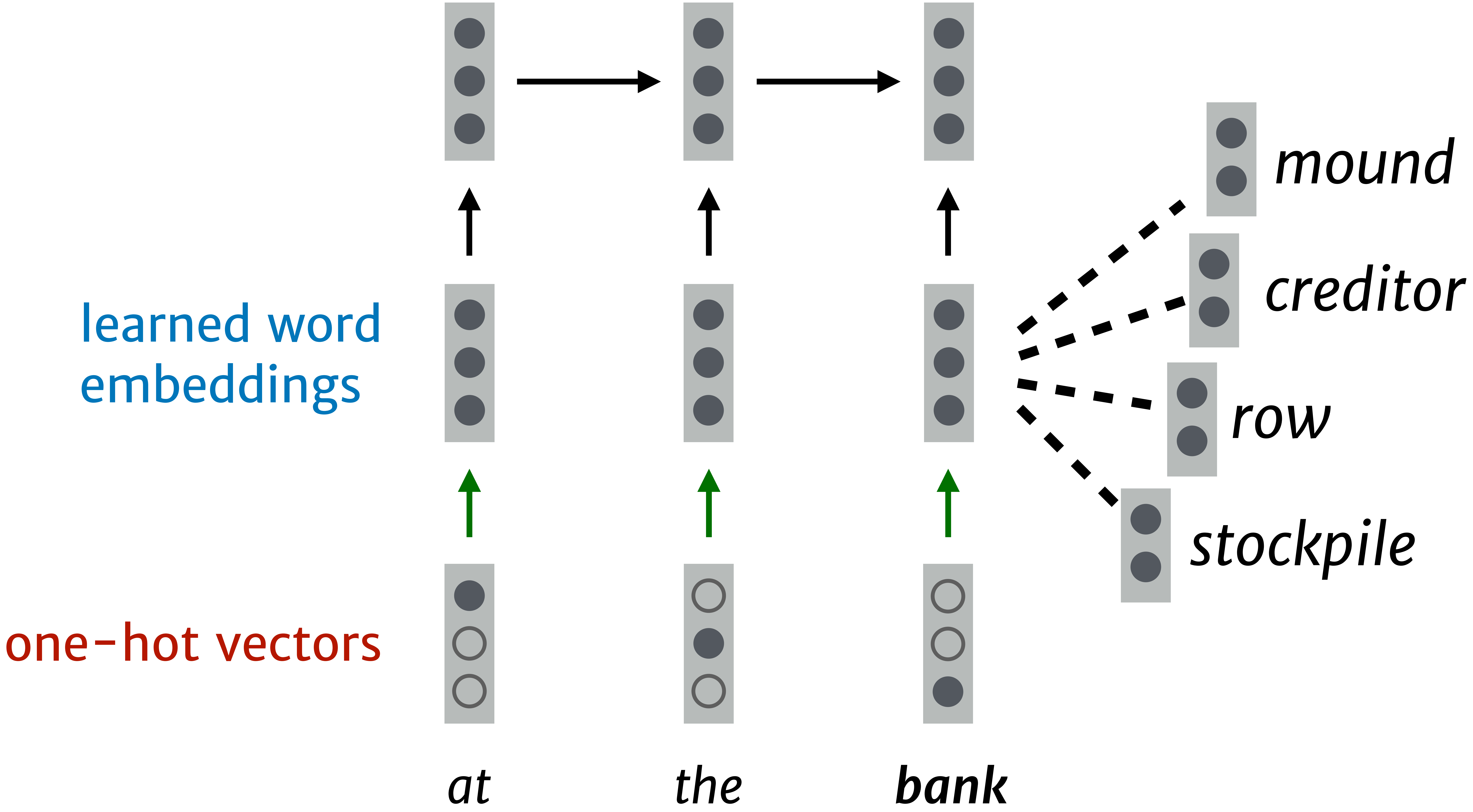
do a biography on the general

has *done* some beautiful landscapes

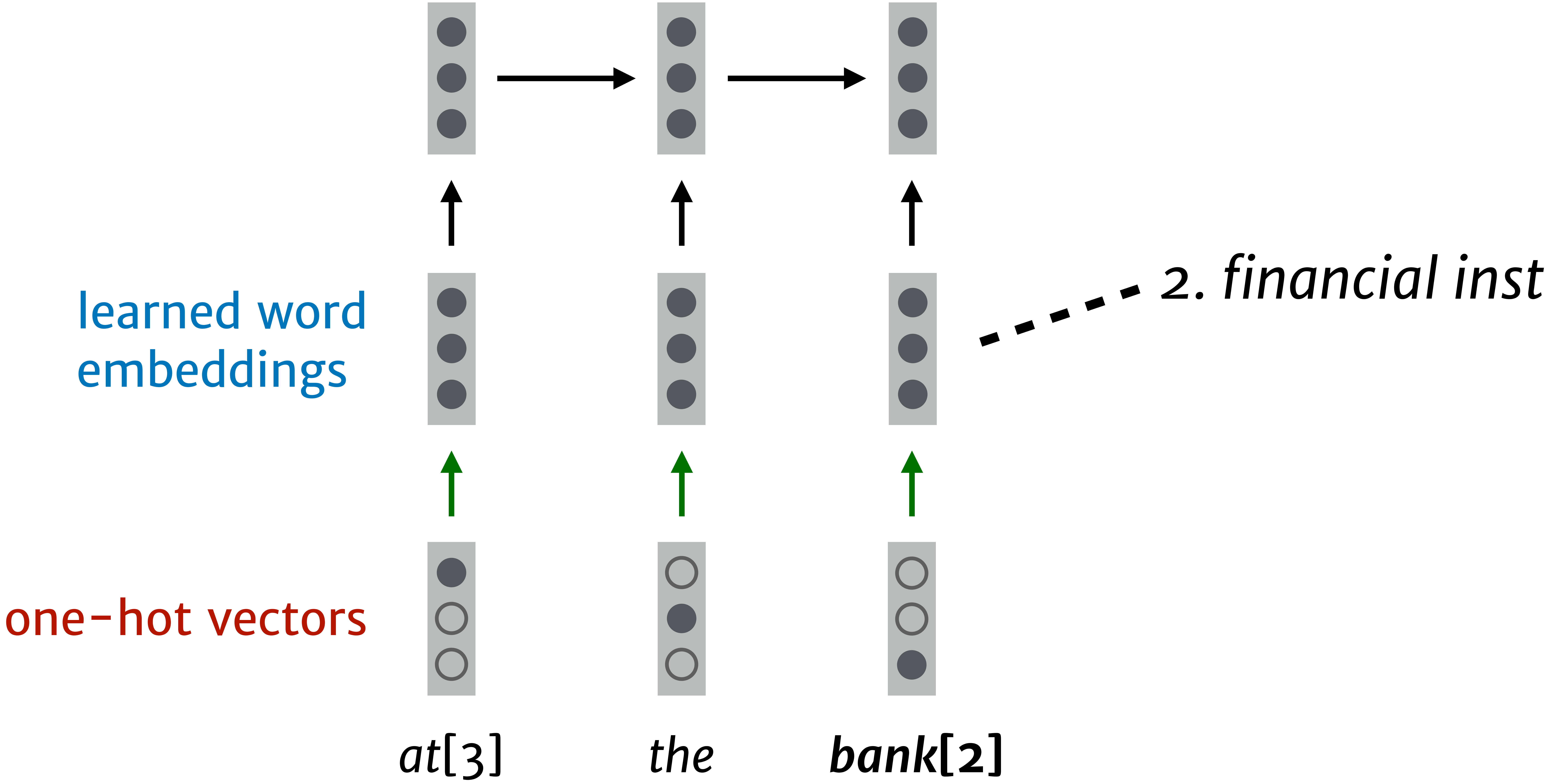
Representations of words



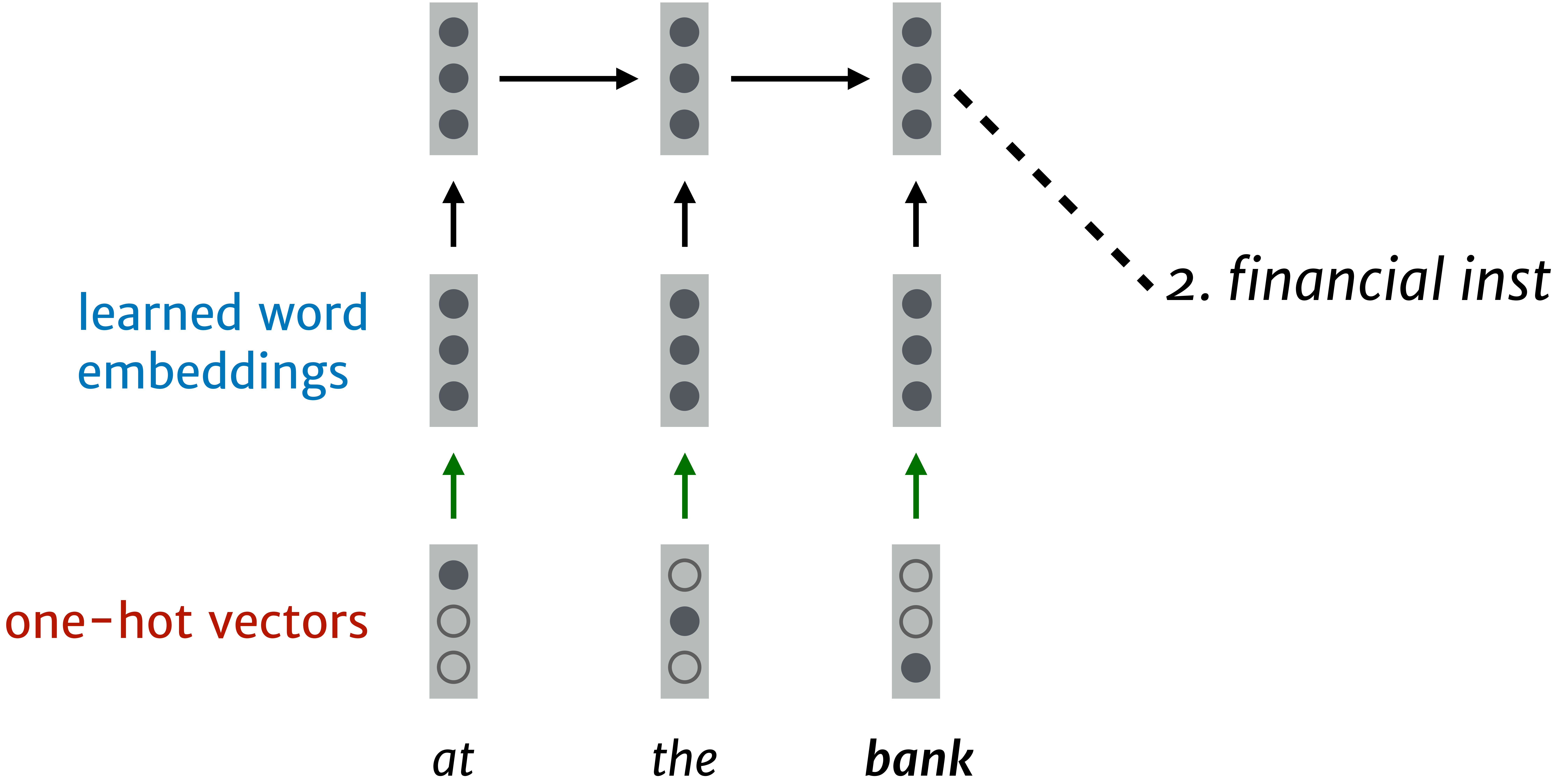
Representations of words



Word sense disambiguation



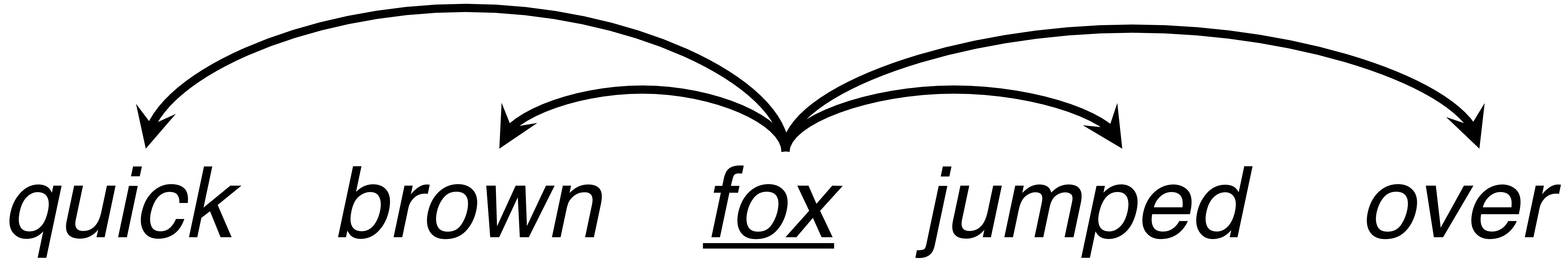
Representations of words **in context**



Language modeling objectives

Language modeling with word2vec

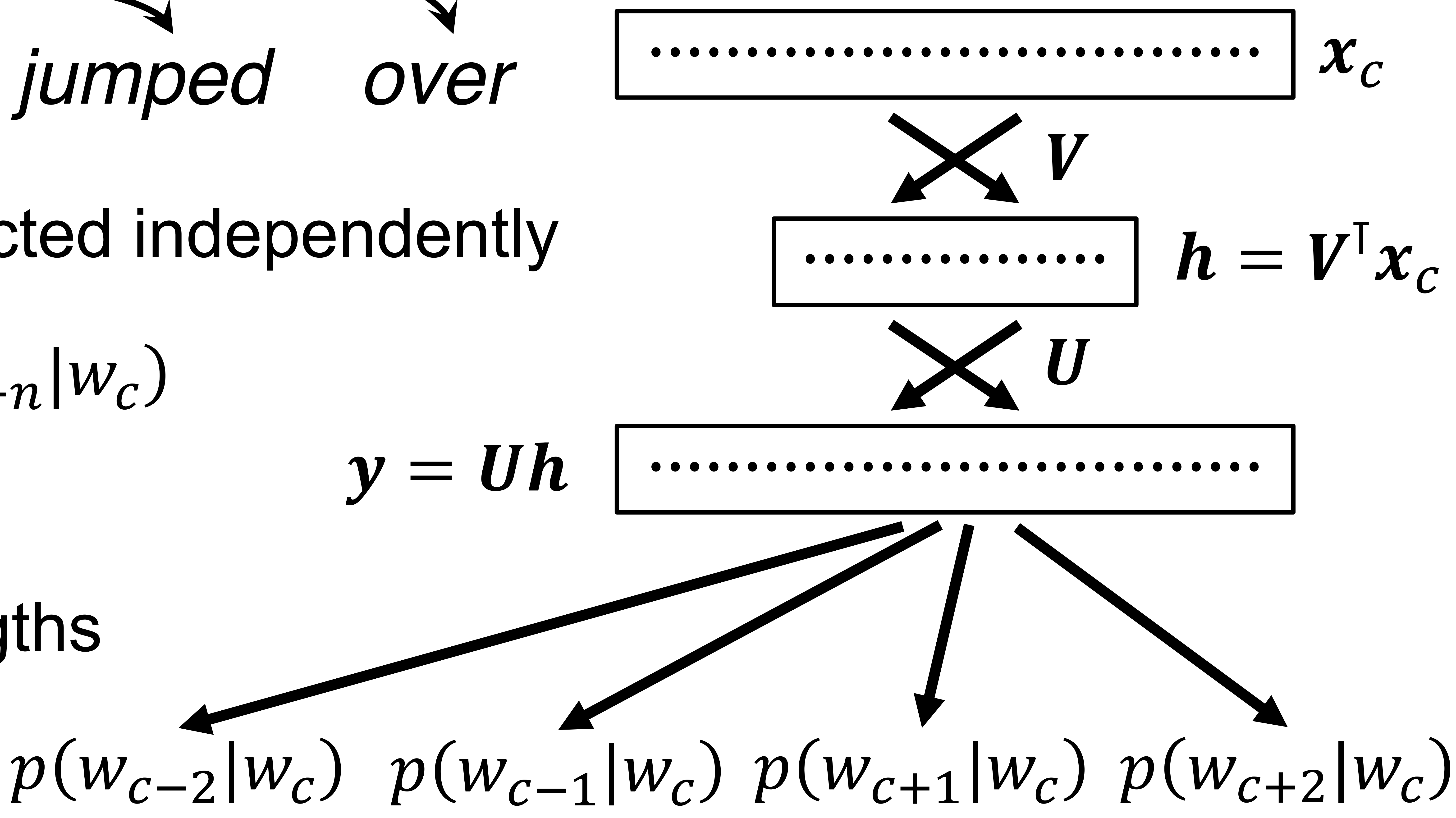
- Skip-gram predicts neighbor words from center word



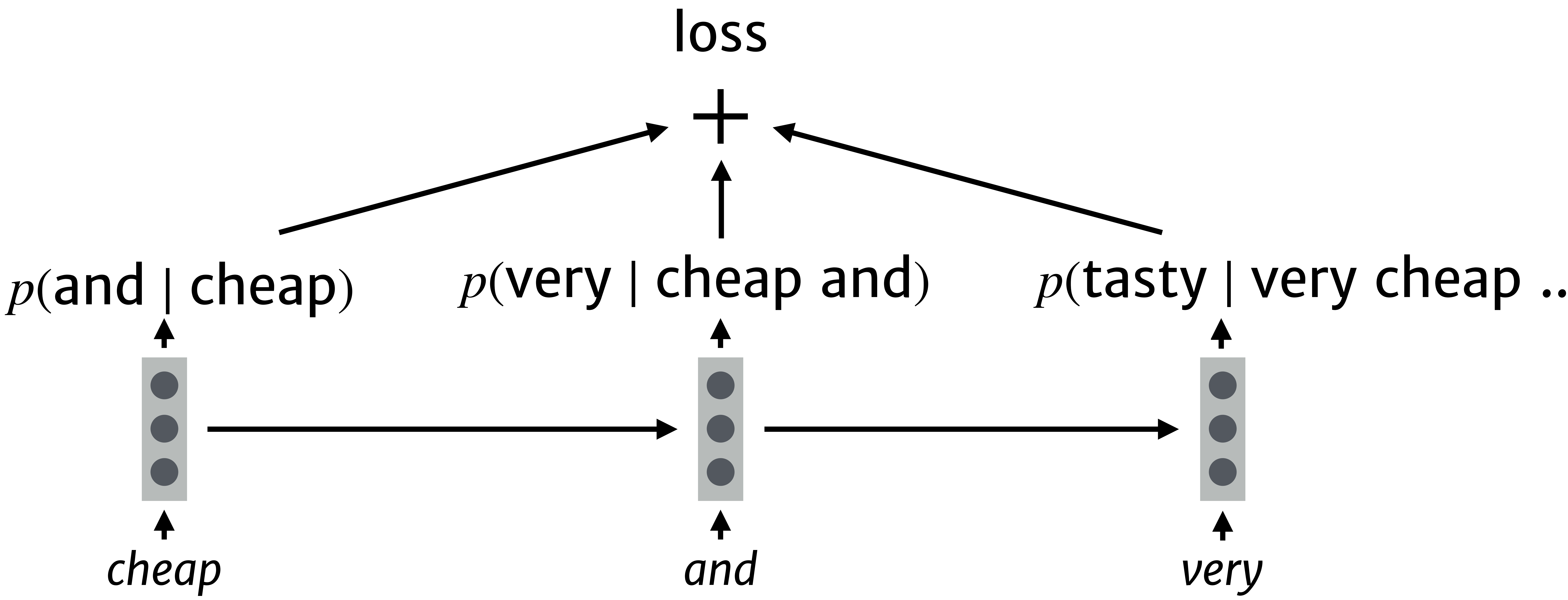
- Each output is predicted independently

$$\prod_{\substack{-h \leq n \leq h \\ n \neq 0}} p(w_{c+n} | w_c)$$

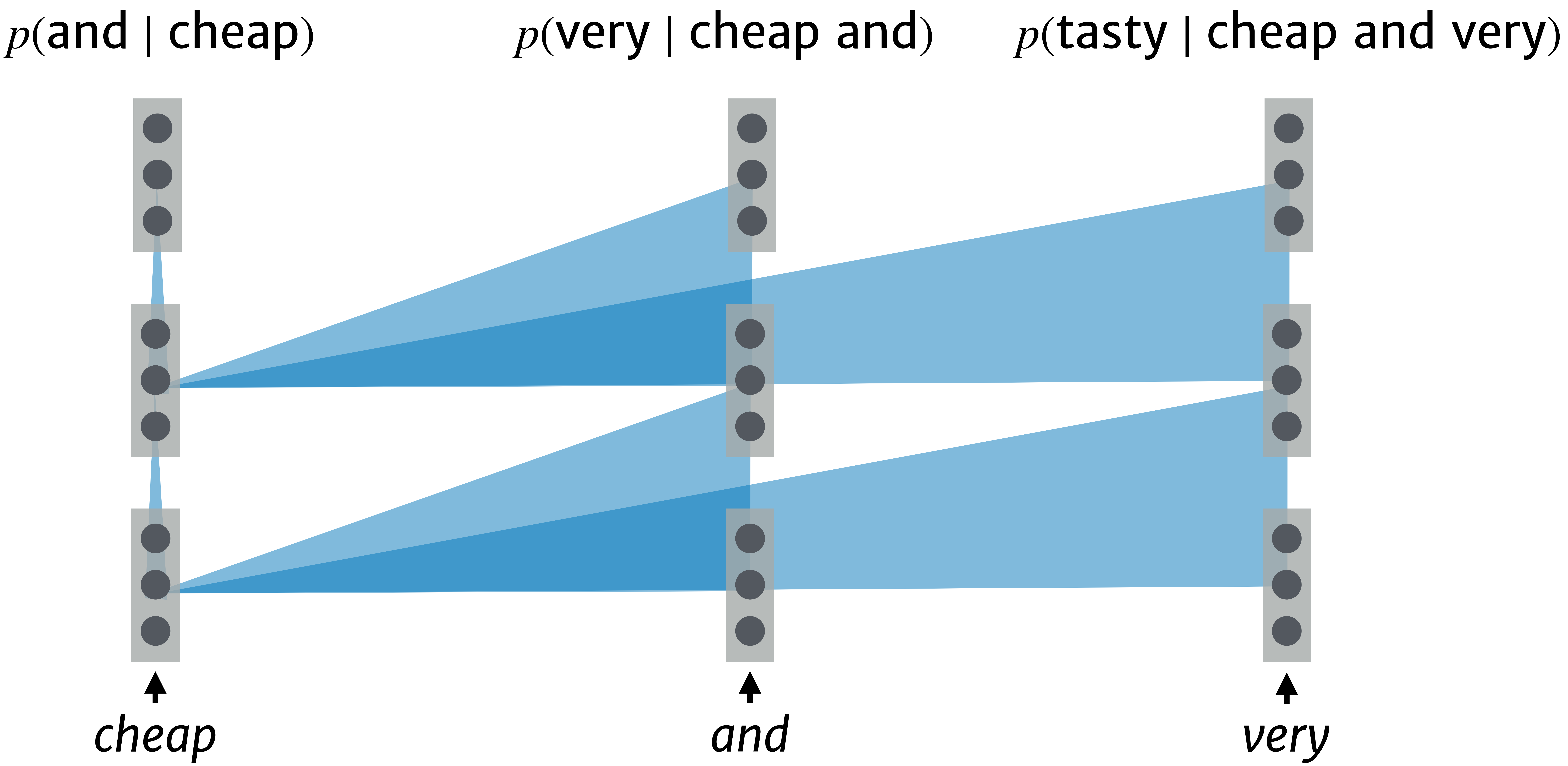
- Context window lengths can be sampled



Language modeling with RNNs



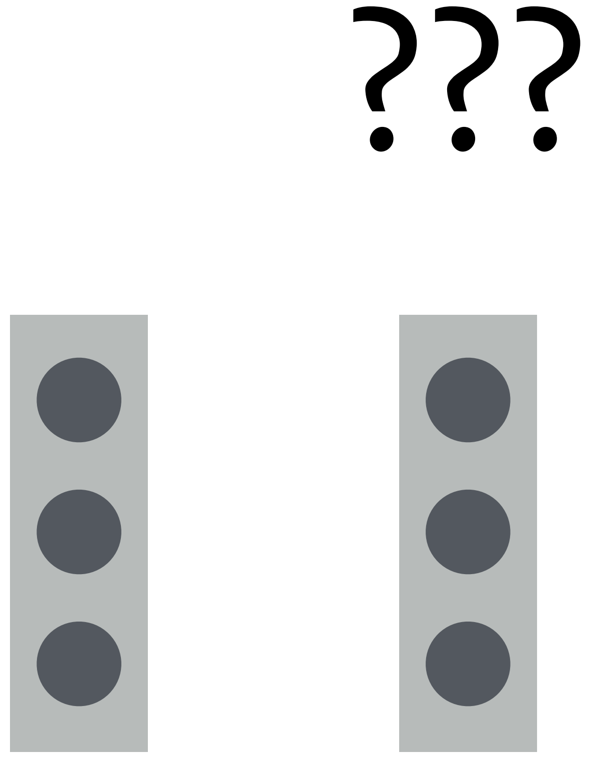
Language modeling with transformers



Language modeling with neural sequence models

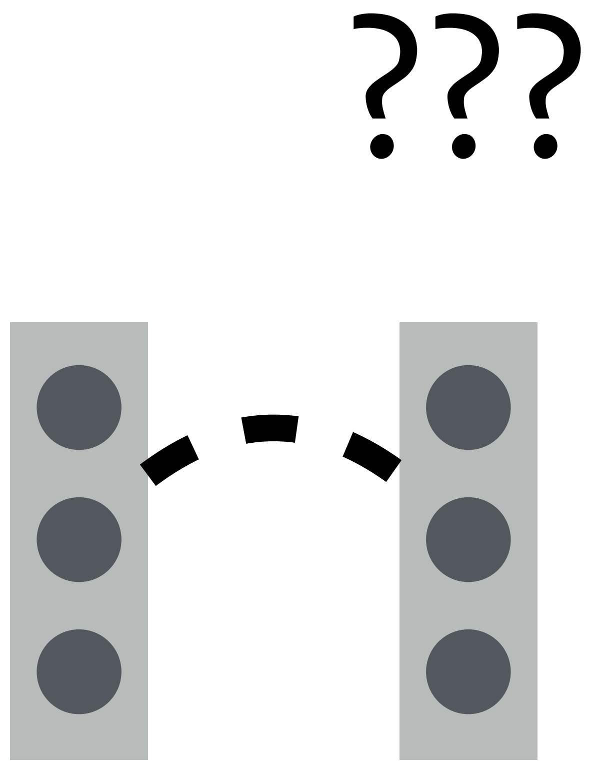
I was out of money so I went to the bank and

Language modeling with neural sequence models



I was out of money so I went to the bank and

Language modeling with neural sequence models

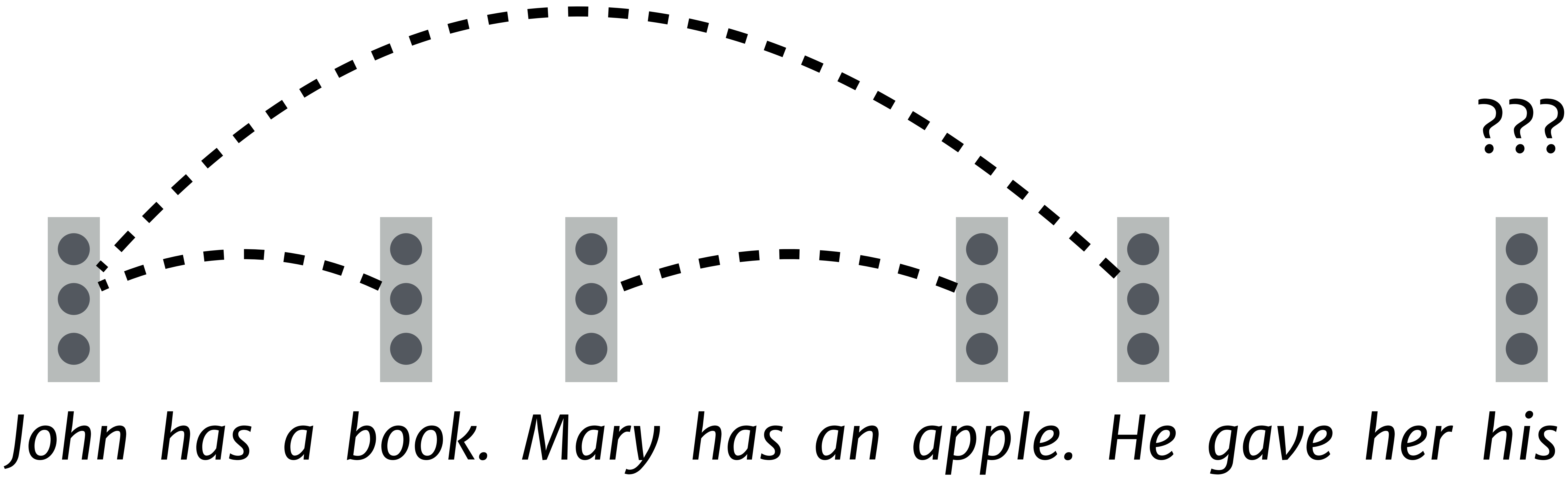


I was out of money so I went to the bank and

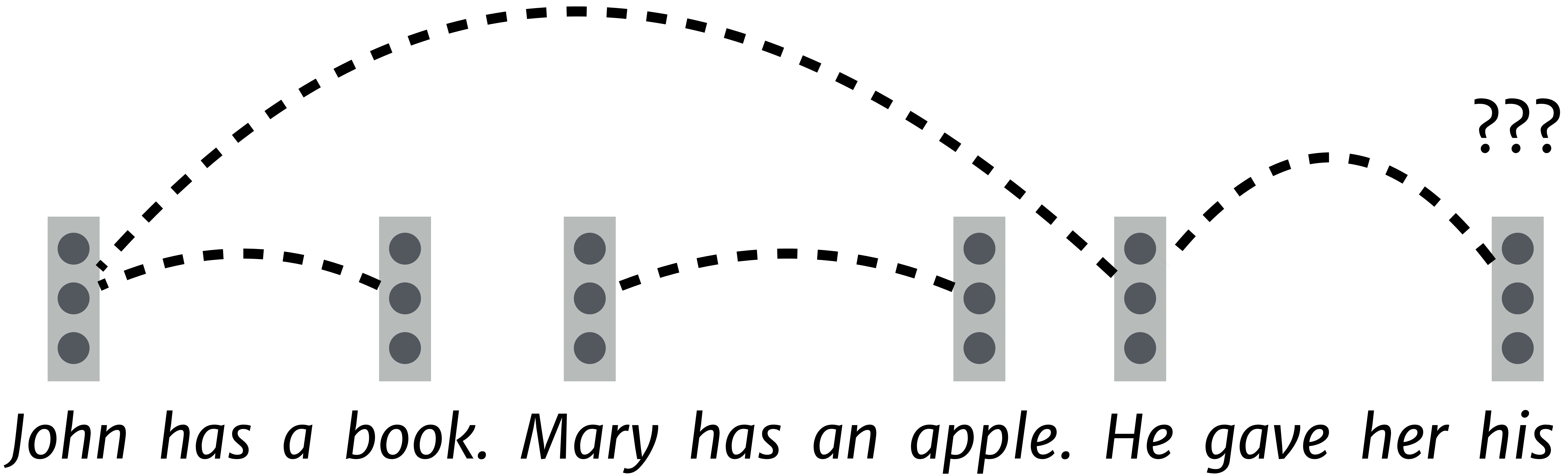
Language modeling with neural sequence models

John has a book. Mary has an apple. He gave her his

Language modeling with neural sequence models



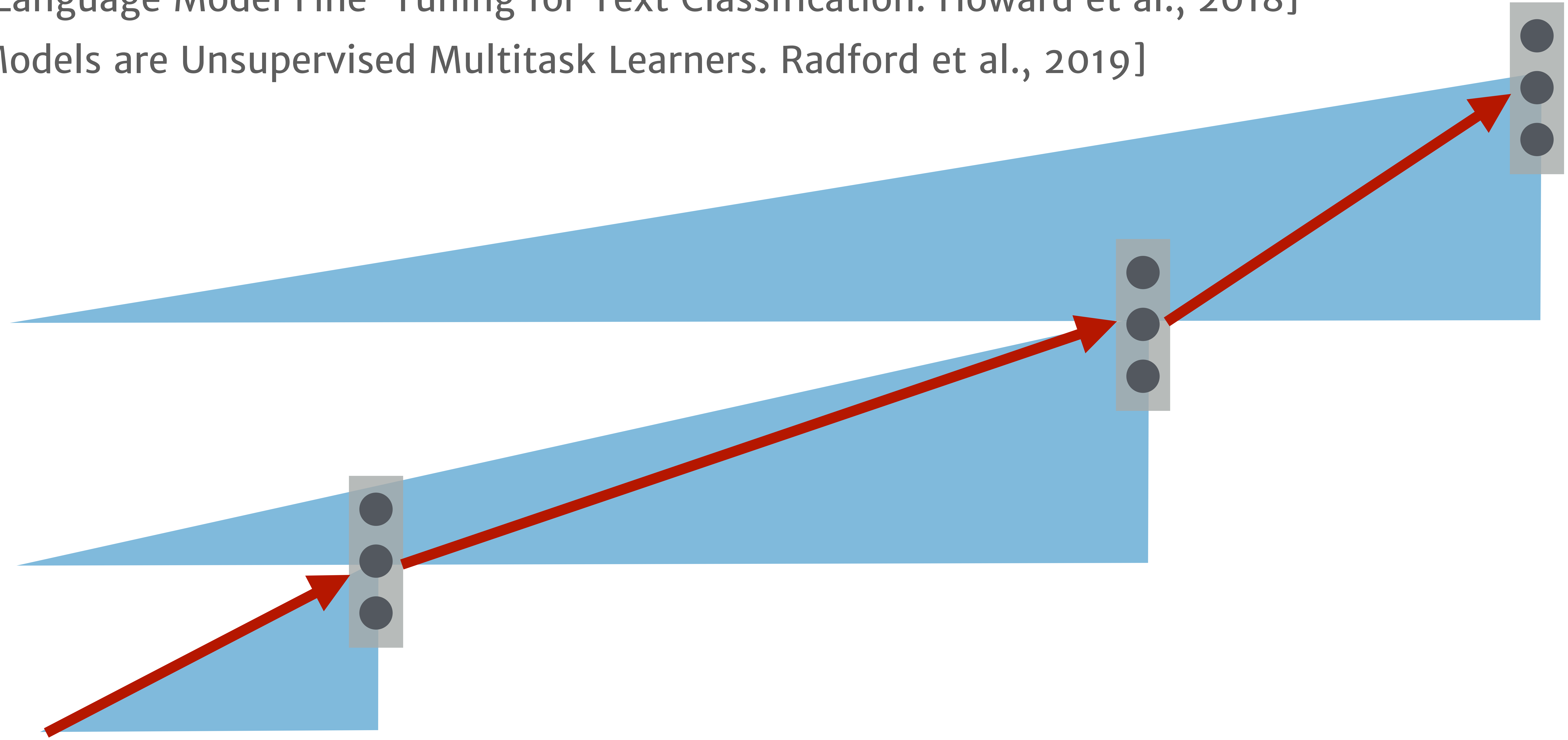
Language modeling with neural sequence models



GPT/ULMFit: Language modeling with neural sequence models

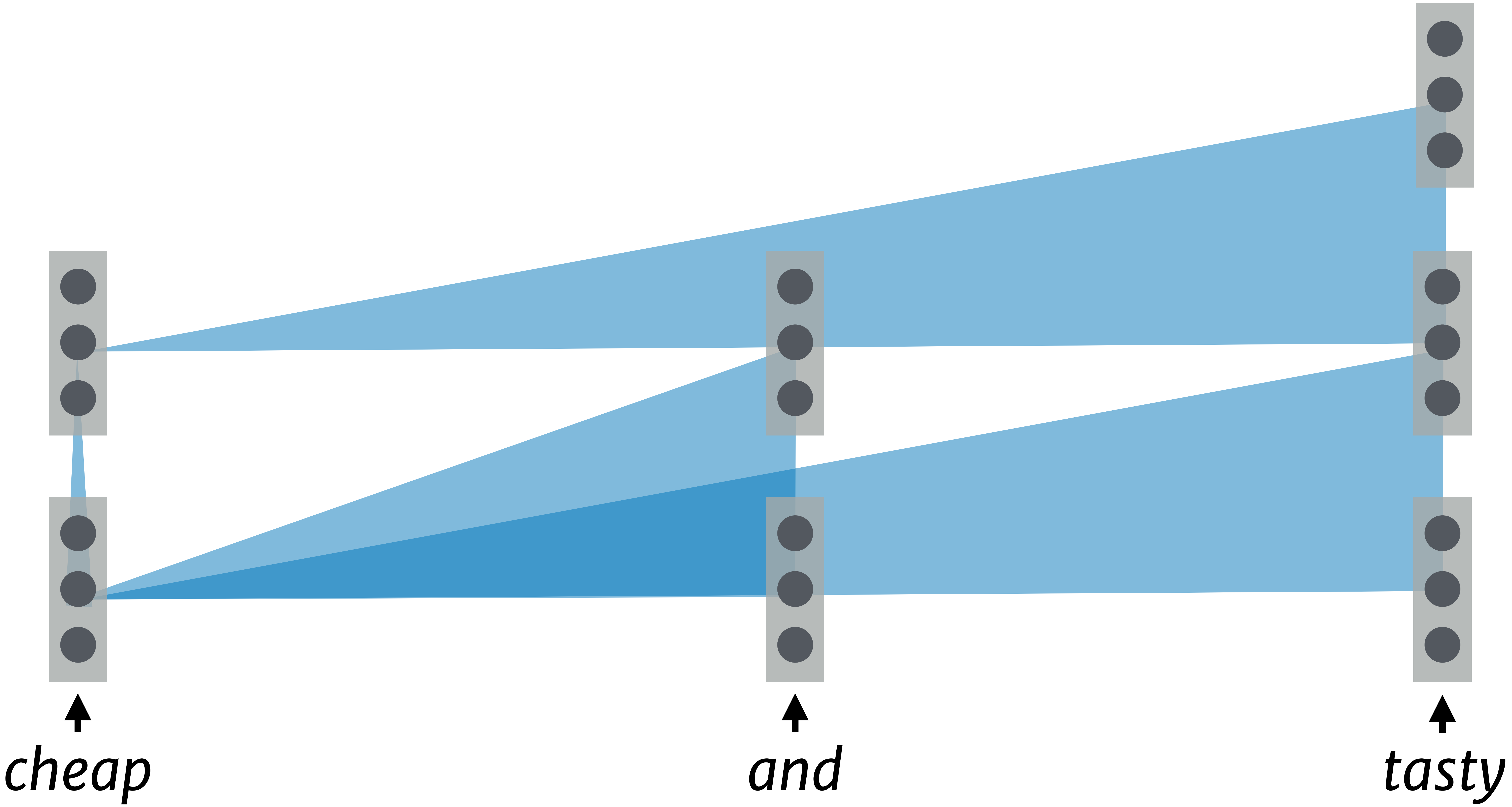
[Universal Language Model Fine-Tuning for Text Classification. Howard et al., 2018]

[Laguage Models are Unsupervised Multitask Learners. Radford et al., 2019]

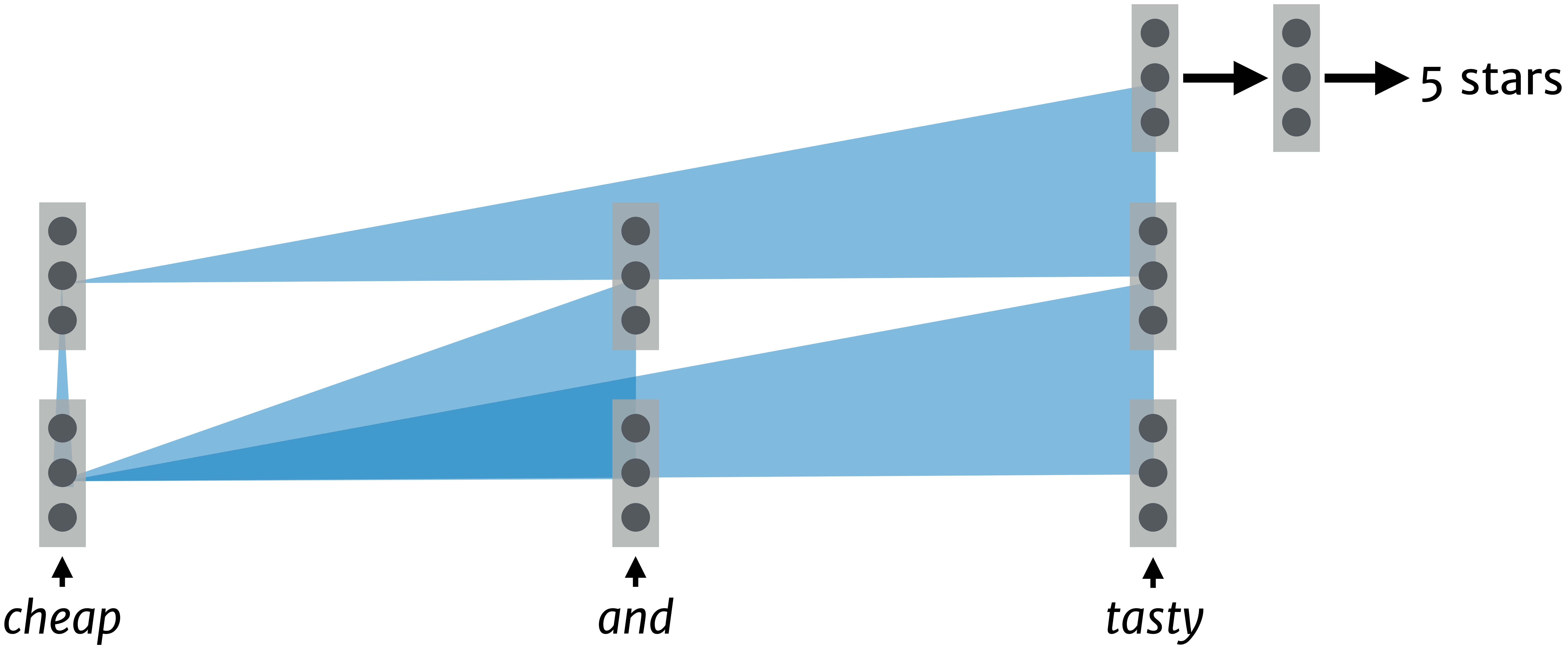


John has a book. Mary has an apple. He gave her his

Fine-tuning: categorical output



Fine-tuning: categorical output

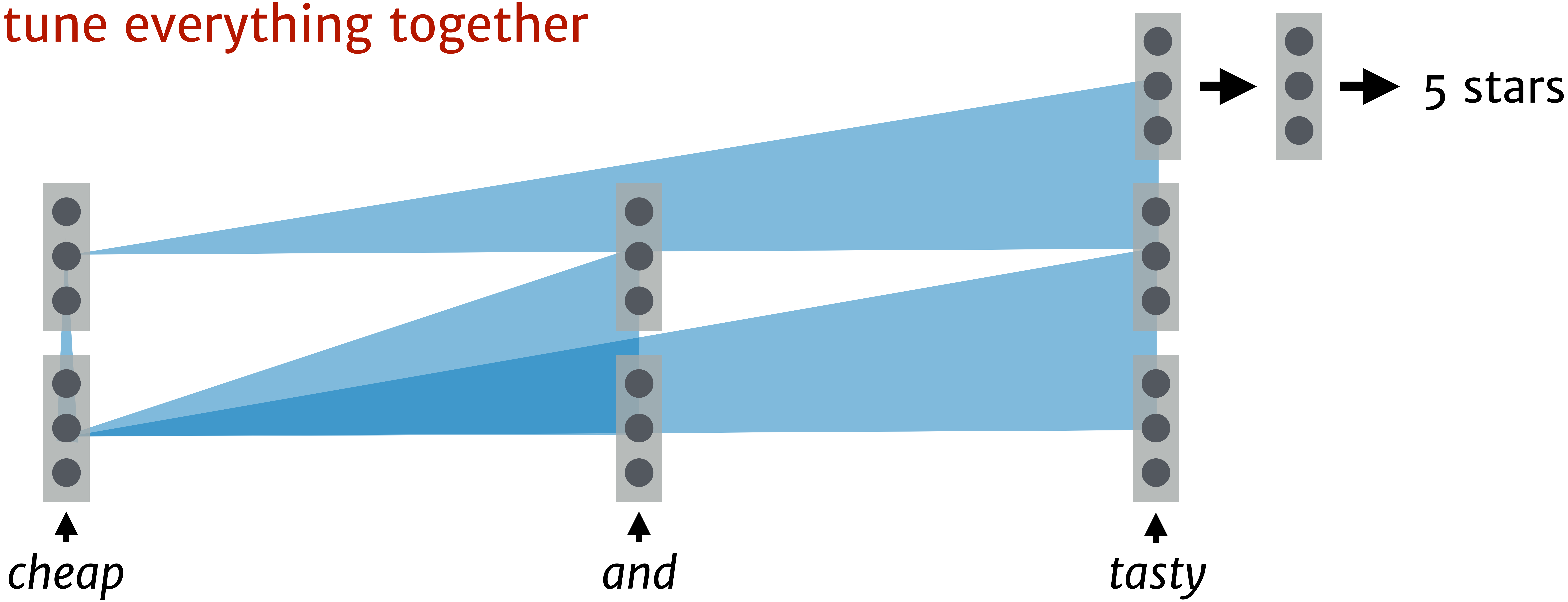


Fine-tuning LMs: categorical output

1. Pretrain on a language modeling task
2. Connect a feed-forward network to the last repr. in the sentence

3a. Freeze LM weights and just train the feed-forward part, or

3b. Fine-tune everything together



Fine-tuning LMs: text output

1. Pretrain on a language modeling task
2. Make a new "language modeling" dataset with your input-output pairs
3. Fine-tune everyhing together:

Pretrain:

The following year she published a paper called Idealtheorie in Ringbereichen, analyzing ascending chain conditions with regard to (mathematical) ideals. Noted algebraist Irving Kaplansky called this work "revolutionary"; the publication gave rise to the term "Noetherian ring" and the naming of several other mathematical objects as Noetherian.

Fine-tune:

Who was Zeng Jiongzhi's doctoral advisor? Emmy Noether.
Where was Barack Obama born? Honolulu.

Bonus: “zero-shot” learning

Don’t fine-tune at all!

Model prompt:

The 2008 Summer Olympics torch relay was run from March 24 until August 8, 2008, prior to the 2008 Summer Olympics, with the theme of “one world, one dream”. Plans for the relay were announced on April 26, 2007, in Beijing, China. The relay, also called by the organizers as the “Journey of Harmony”, lasted 129 days and carried the torch 137,000 km (85,000 mi) ***Q: what was the theme? A:***

Continuation:

“one world, one dream”

[Radford et al. 2019]

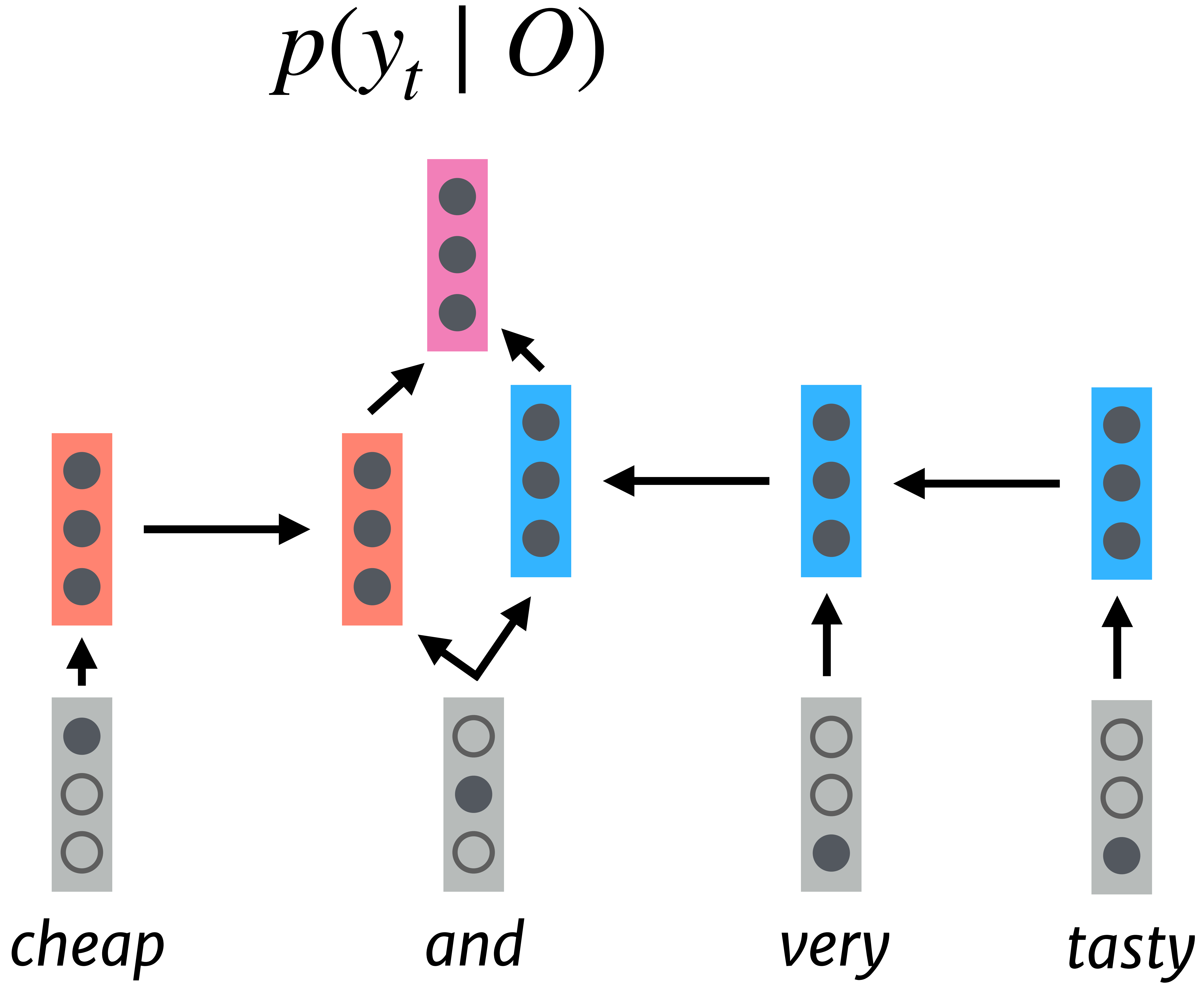
What's missing?

I ***can*** *anchovies.*

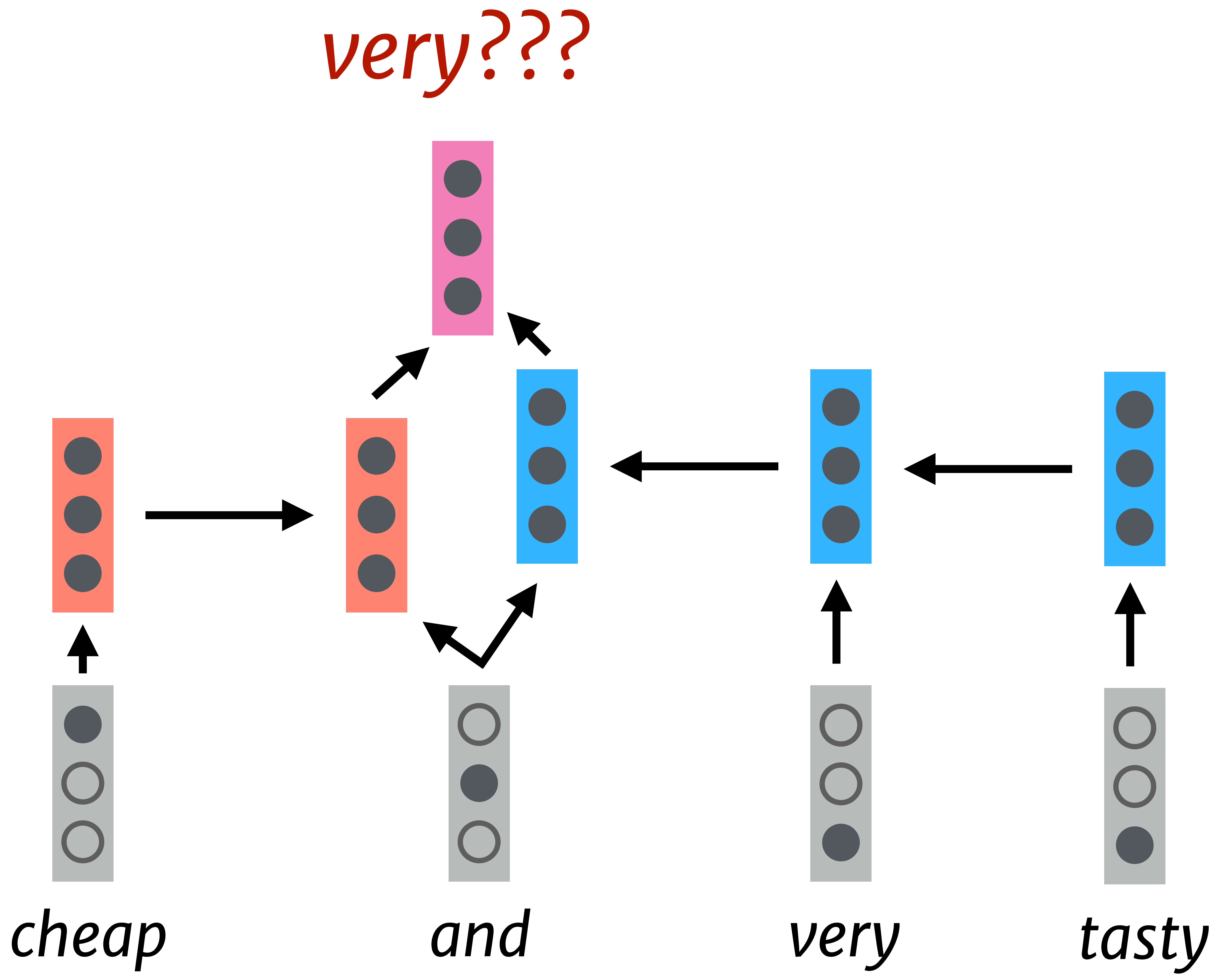
Left-to-right language modeling objectives give us sentence representations, but not fully contextual word representations.

Masked language modeling objectives

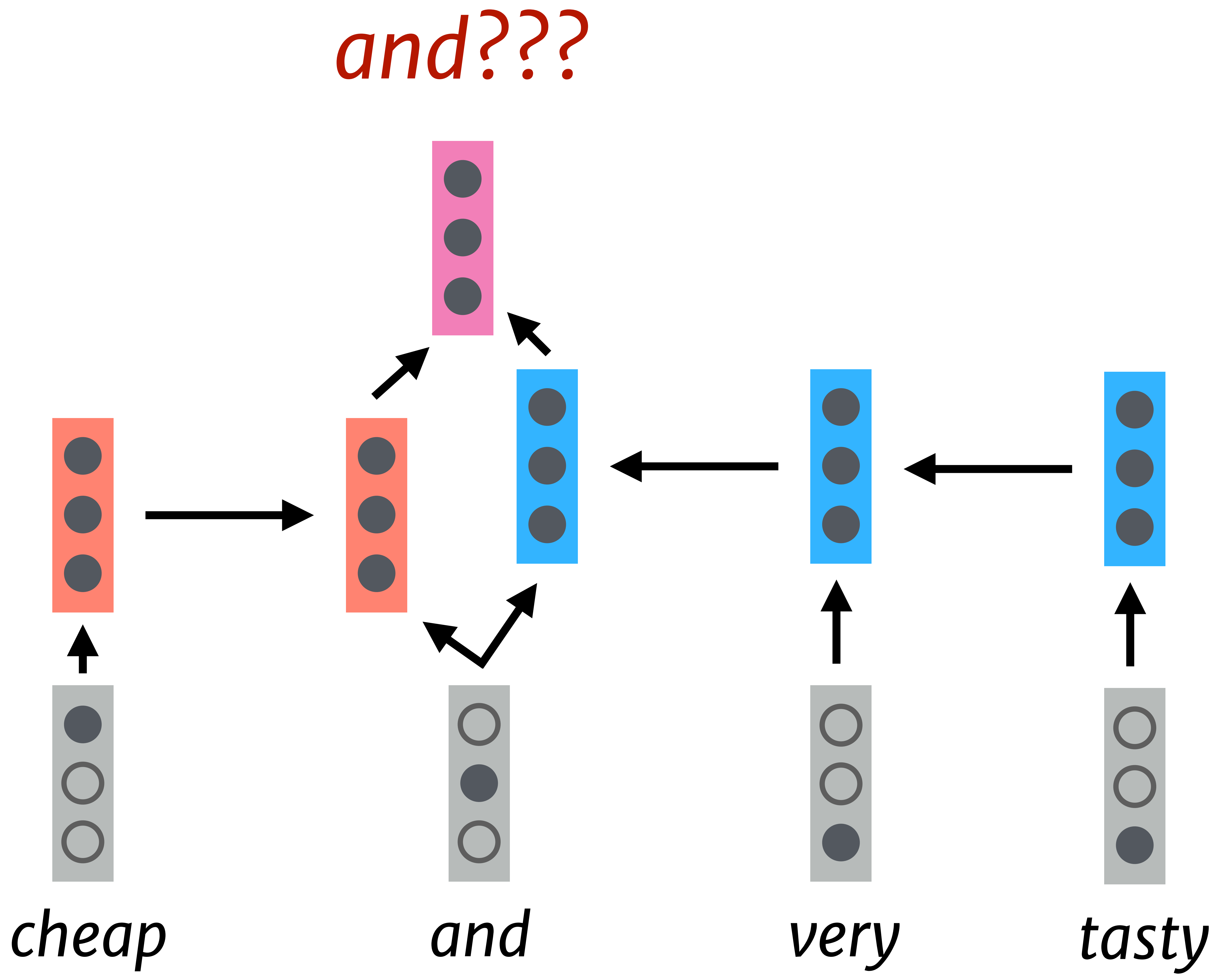
Bidirectional RNNs



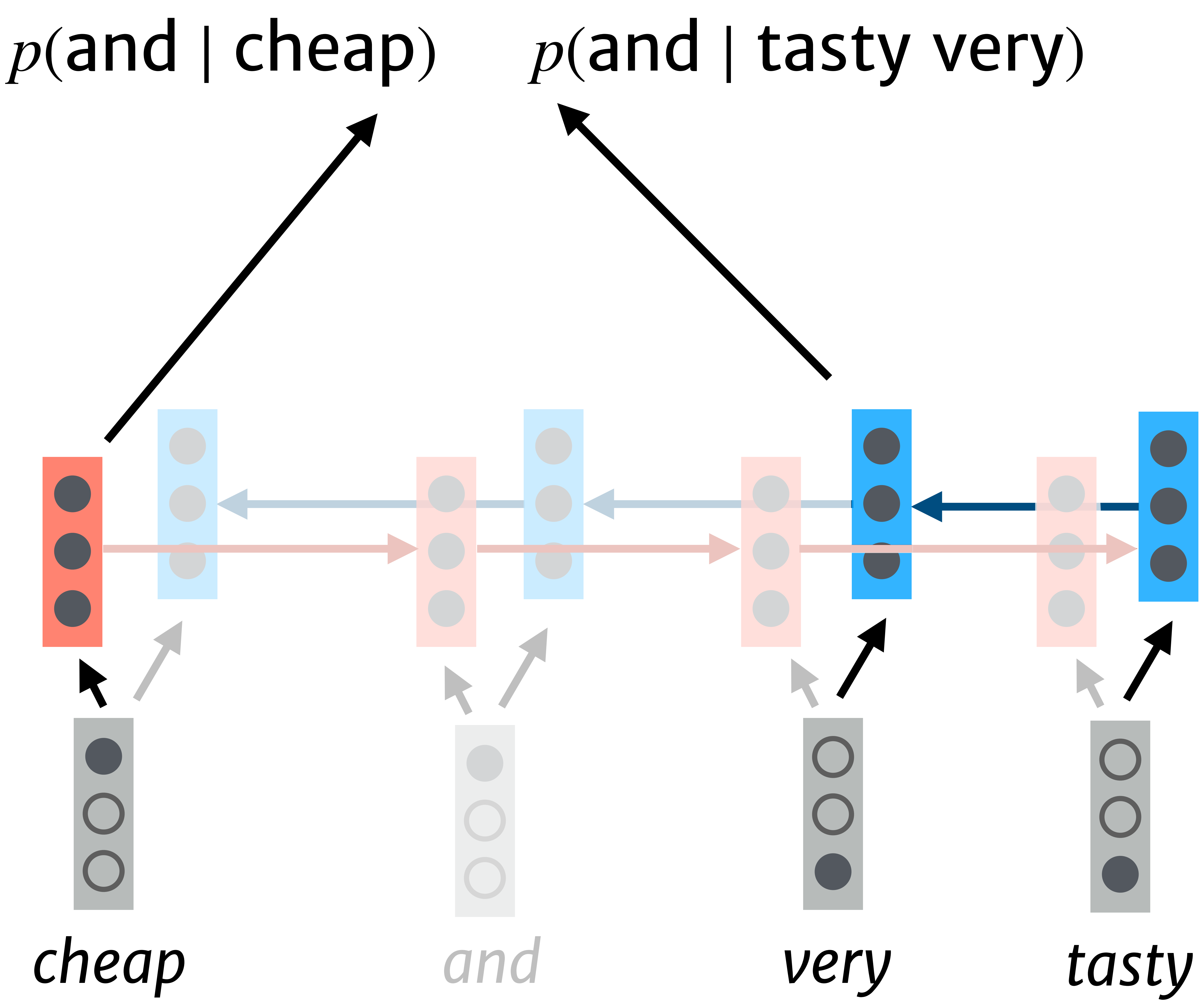
Bidirectional “language modeling”



Bidirectional “language modeling”



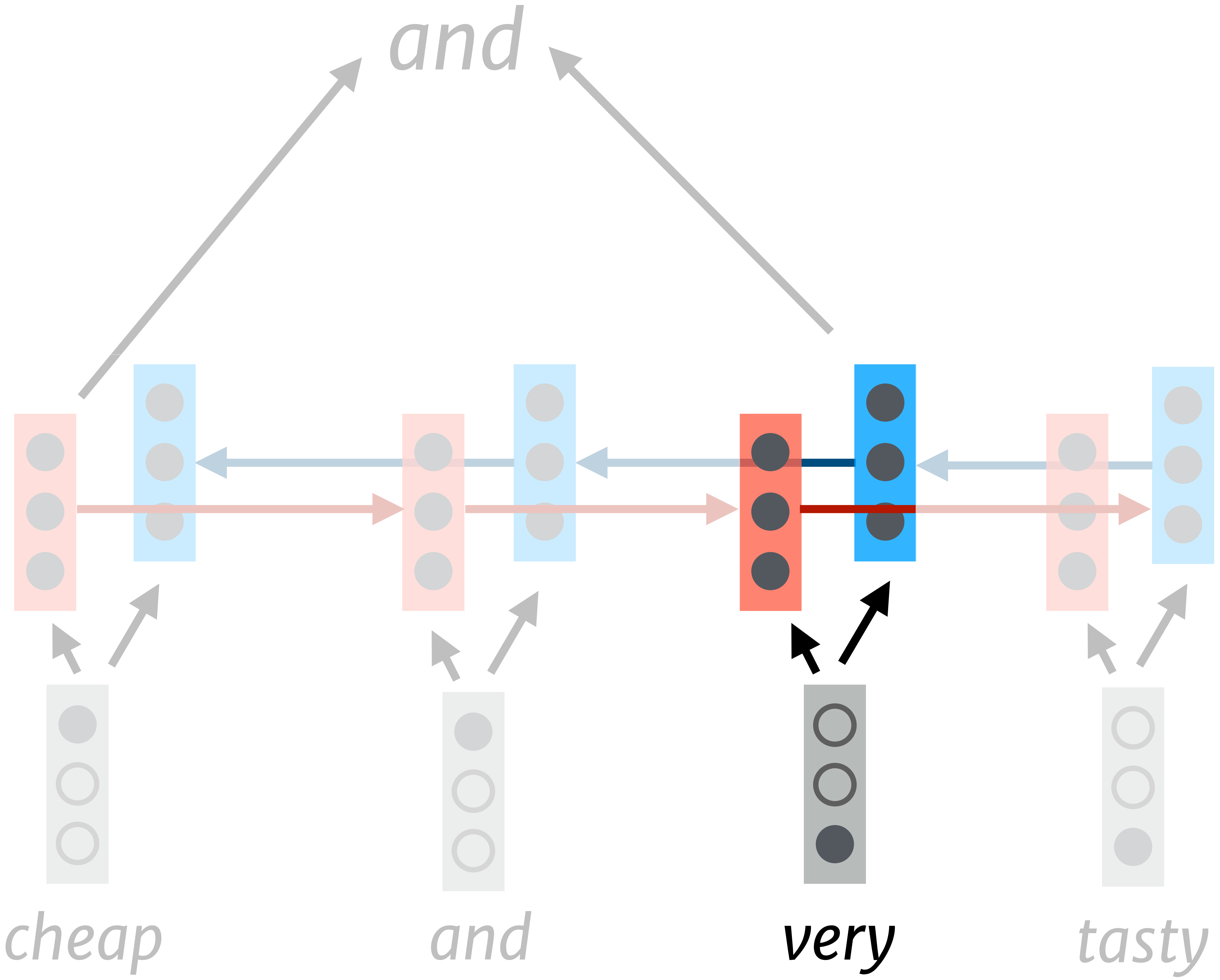
ELMo: bidirectional language modeling



Idea: train independent forward / backward LMs and concatenate the representations.

[Deep Contextualized Word Representations. Peters et al. 2018]

ELMo: bidirectional language modeling

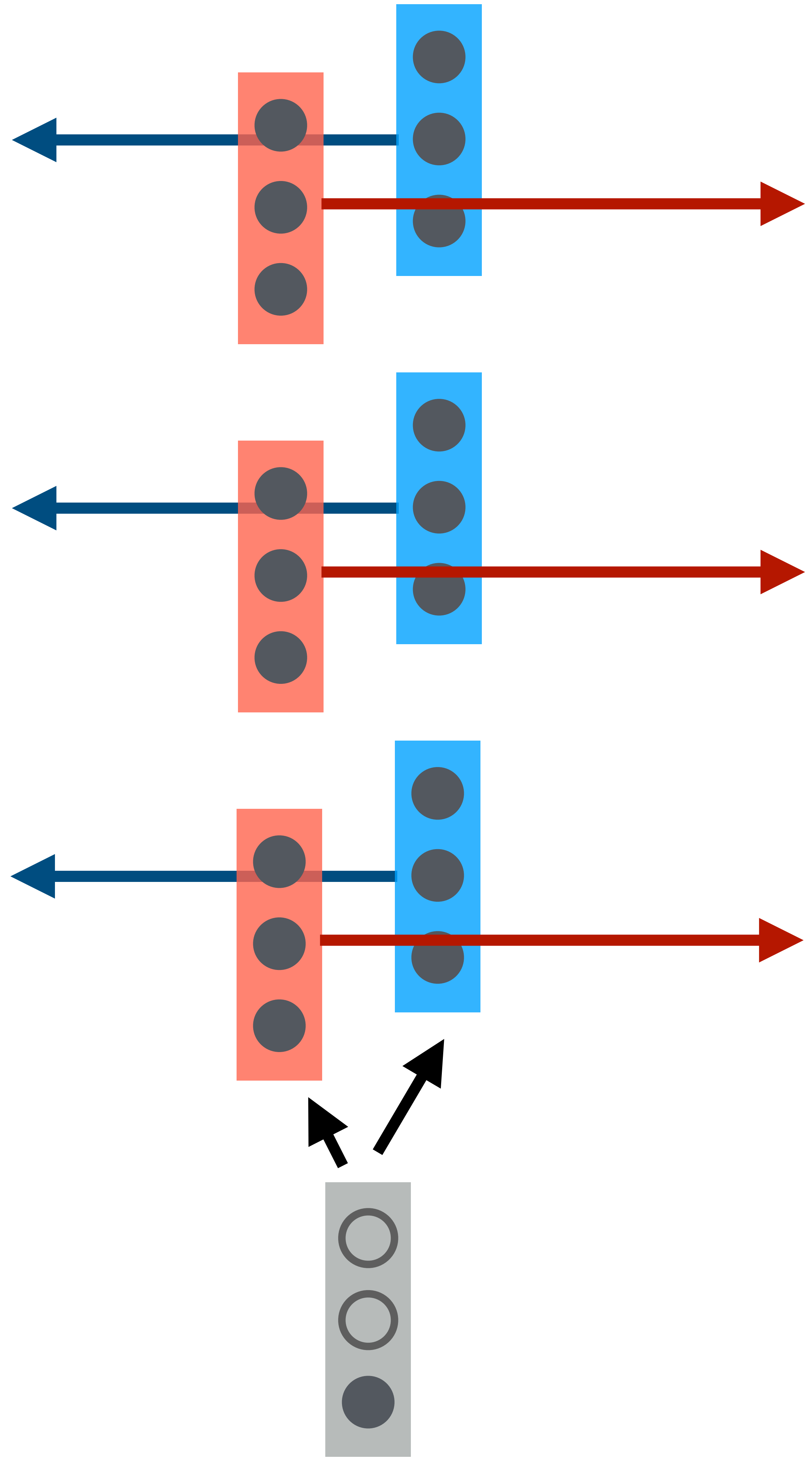


Idea: train independent forward / backward LMs and concatenate the representations.

Every word has a forward repr., a backward repr., and a context-indep. repr.

[Deep Contextualized Word Representations. Peters et al. 2018]

ELMo: more details

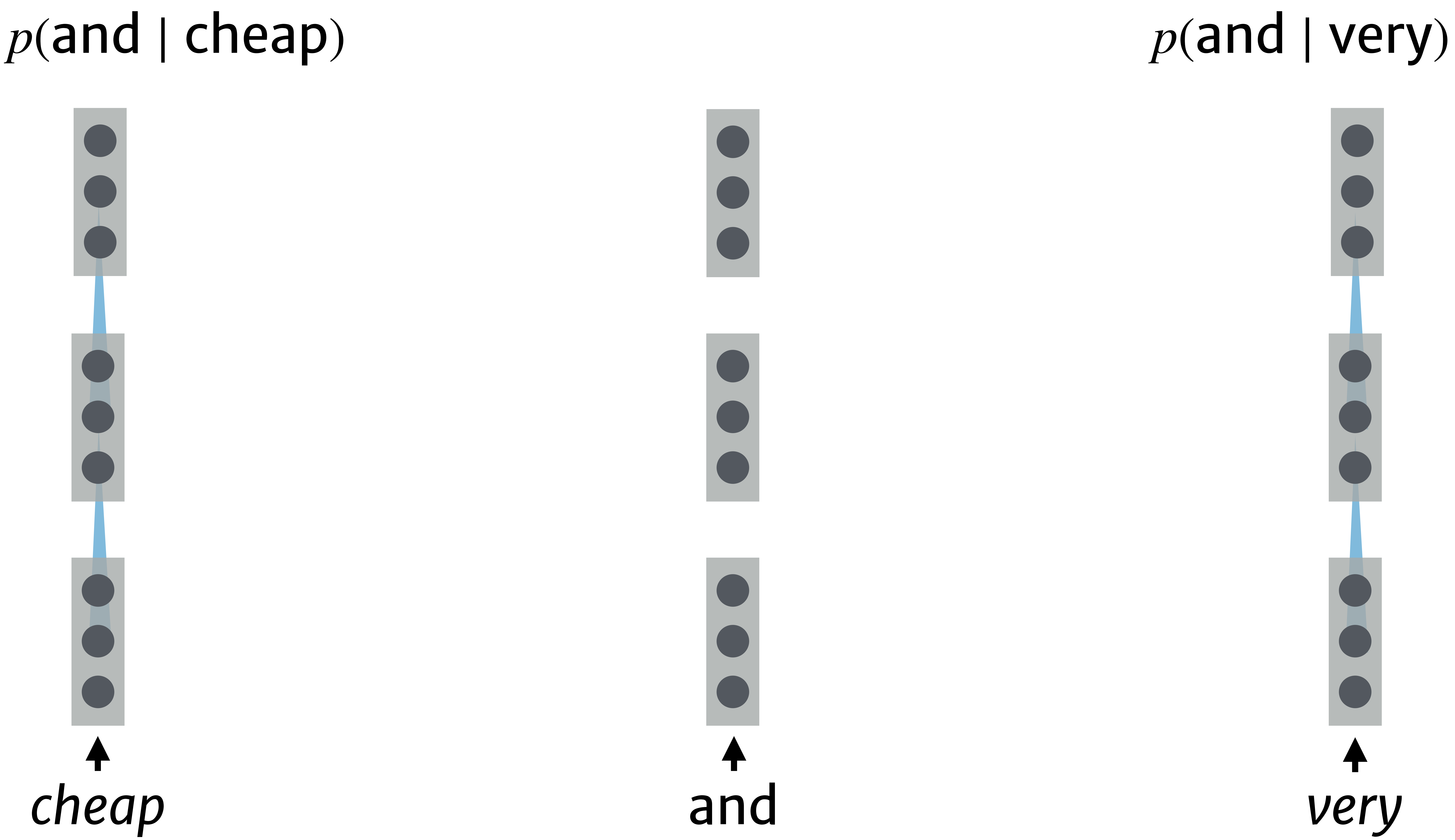


We're actually training a deep LSTM, so multiple layers in each representation.

Most effective: use a learned linear combination of layers as input to the downstream task.

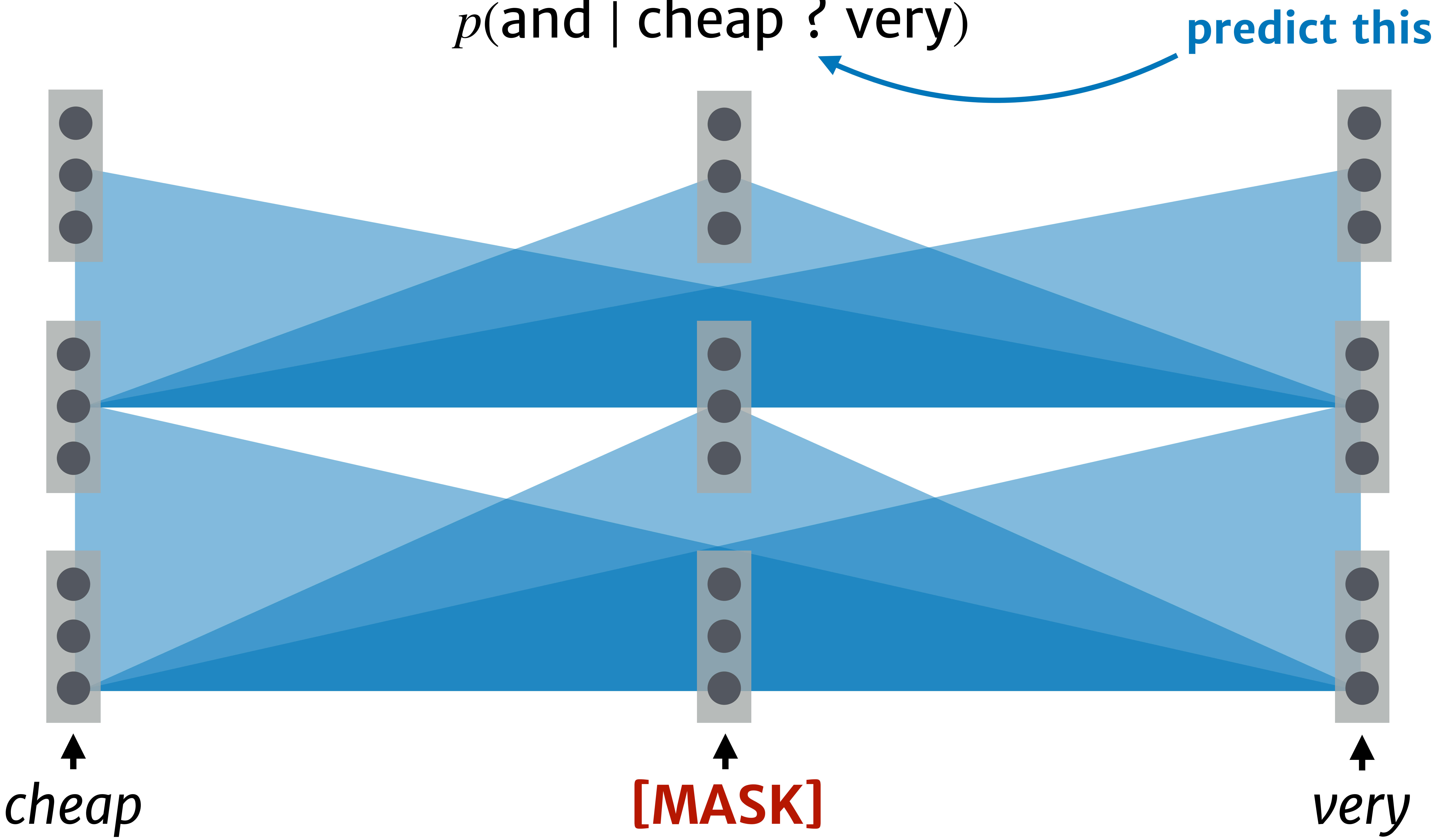
Use these anywhere you'd use word embeddings!

“Bidirectional” transformer LMs



“Bidirectional” transformer LMs

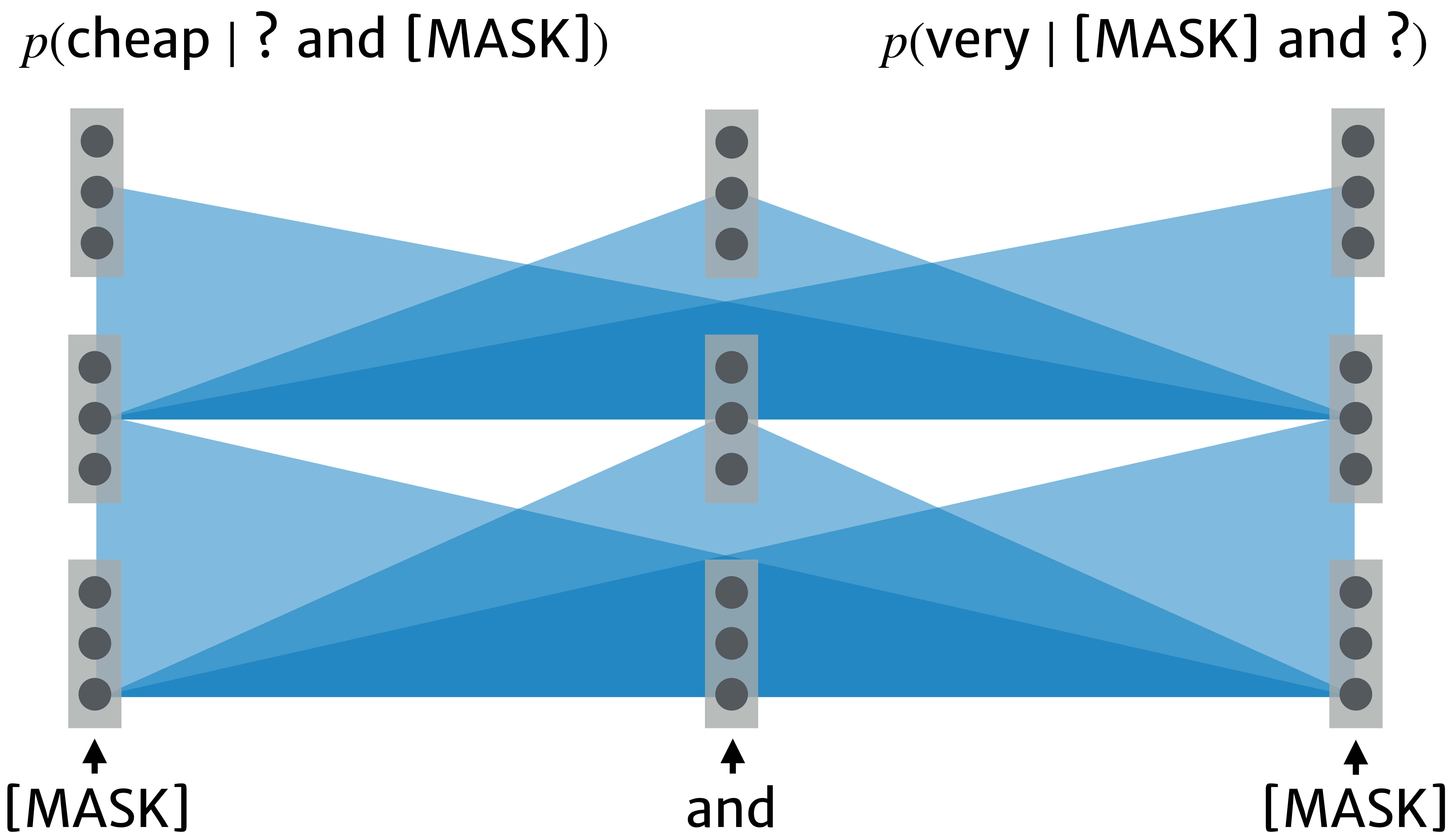
p(and | cheap ? very)



Idea: Rather than masking everything to the right, mask at arbitrary positions and only predict at masks.

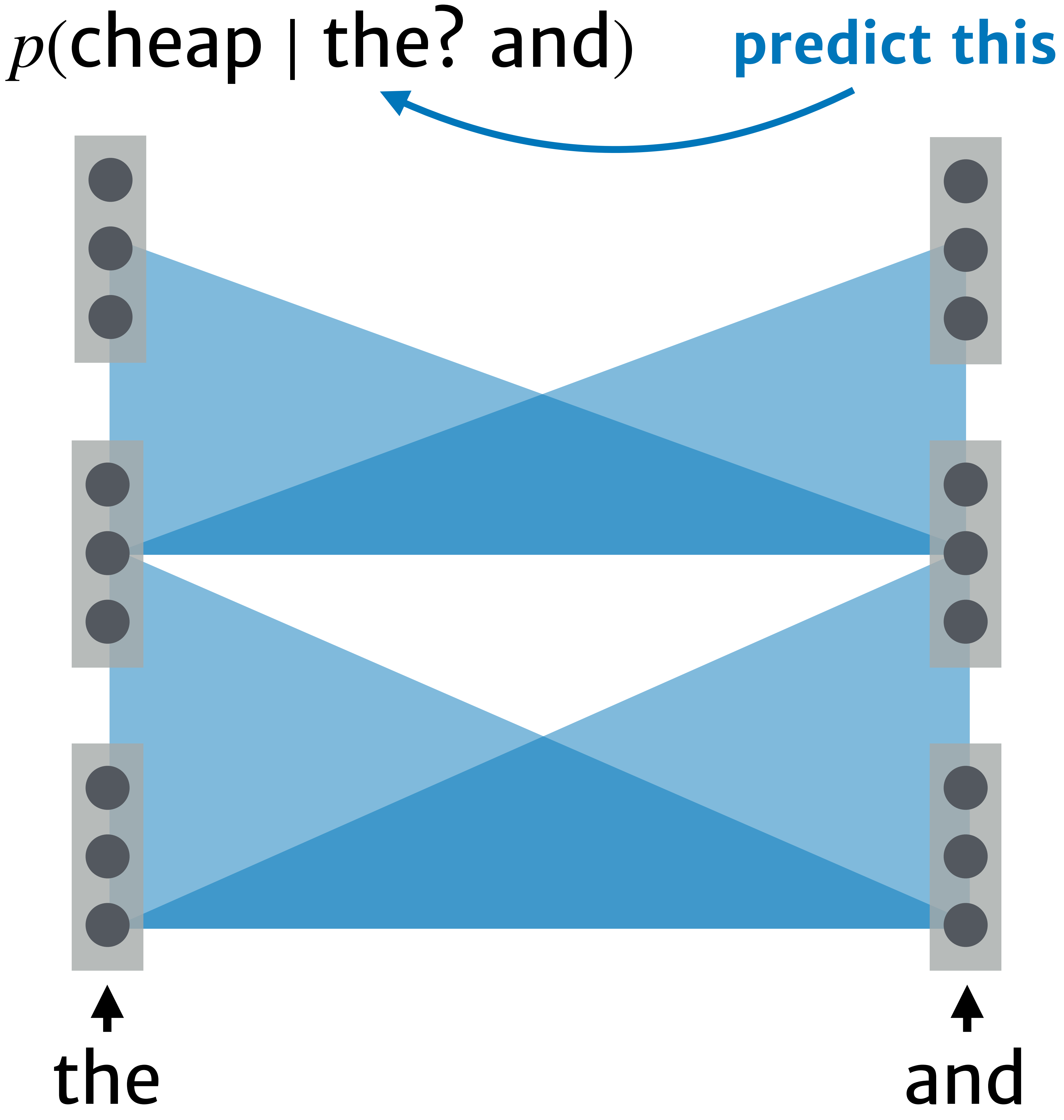
BERT: Masked language modeling

[Devlin et al., 2018]



Idea: add multiple mask tokens per sentence and predict all of them at the same time.

BERT: more tricks



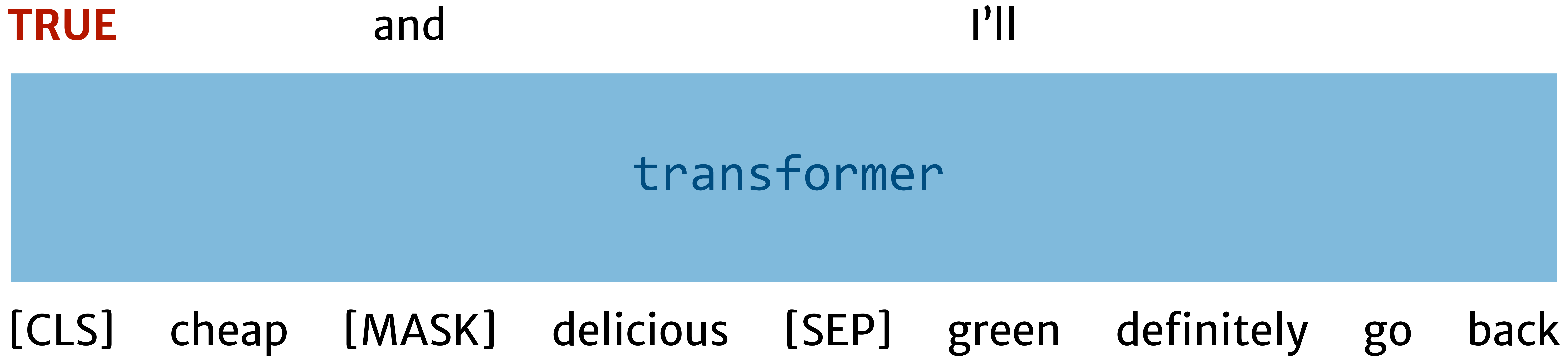
(1) if we only predict above [MASK] tokens, no pressure on model to route information to rest of sentence (we want good embeddings everywhere)

Idea: instead of always labeling prediction targets as [MASK], sometimes leave them in place or replace with a random word.

BERT: more tricks

(1) We'd also like to encourage the model to capture some global information

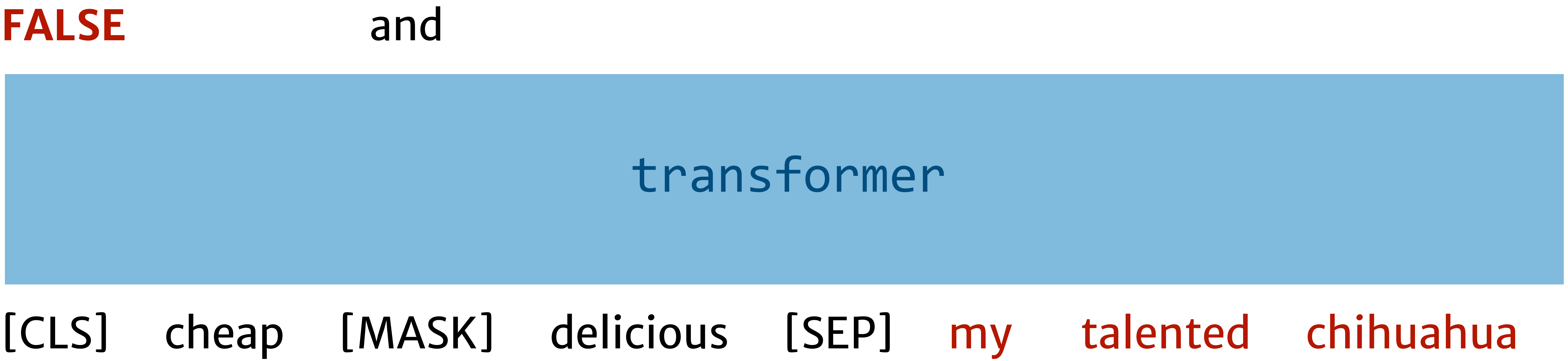
Idea: train on pairs of sentences; learn to predict whether they're adjacent in a training document.



BERT: more tricks

(2) We'd also like to encourage the model to capture some global information

Idea: train on pairs of sentences; learn to predict whether they're adjacent in a training document.



BERT: more tricks

(3) What do do with out-of-vocabulary words?

Idea: identify *k* most frequent **word pieces** in the corpus and operate on those.

The viscountess Wallingford →

[CLS] the viscount ##ess wall ##ing ##ford

Language modeling?

It’s very hard to sample sentences from this model!

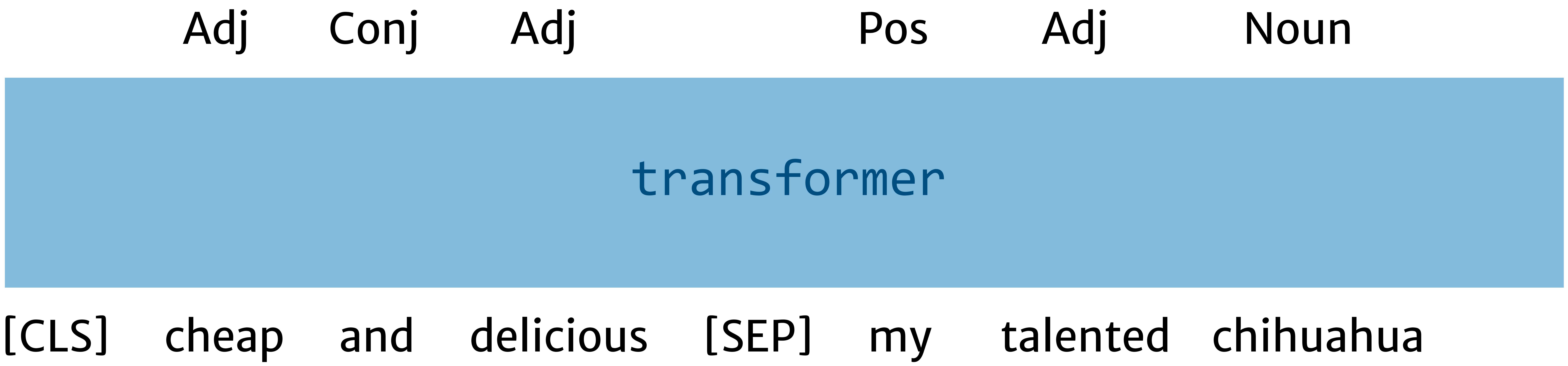
(and generally not done)

Indeed, can't replace a [SEP] with a word sequence of unknown length—BERT knows how big the gap is.

Fine-tuning MLMs: sequence labeling

masked language models

1. Pretrain the masked LM task
2. Use final transformer representations to predict your labels rather than words
3. **Fine-tune everything!**



Why is (M)LM a good pretraining objective?

He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I’ve already focused on my friend. You just have to click the shutter, on top, here.” He nodded sheepishly, through his cigarette away and took the [?]

Why is (M)LM a good pretraining objective?

He shook his head, took a step back and held his hands up as he tried to smile without losing a cigarette. “Yes you can,” Julia said in a reassuring voice. “I’ve already focused on my friend. You just have to click the shutter, on top, here.” He nodded sheepishly, through his cigarette away and took the [?]

camera

How much does this help?

Question answering:

System	Dev EM	Dev F1	Test EM	Test F1
Top Leaderboard Systems (Dec 10th, 2018)				
Human	86.3	89.0	86.9	89.5
#1 Single - MIR-MRC (F-Net)	-	-	74.8	78.0
#2 Single - nlnet	-	-	74.2	77.1
Published				
unet (Ensemble)	-	-	71.4	74.9
SLQA+ (Single)	-		71.4	74.4
Ours				
$\text{BERT}_{\text{LARGE}}$ (Single)	78.7	81.9	80.0	83.1

[Devlin et al., 2018]

How much does this help?

Sentence classification:

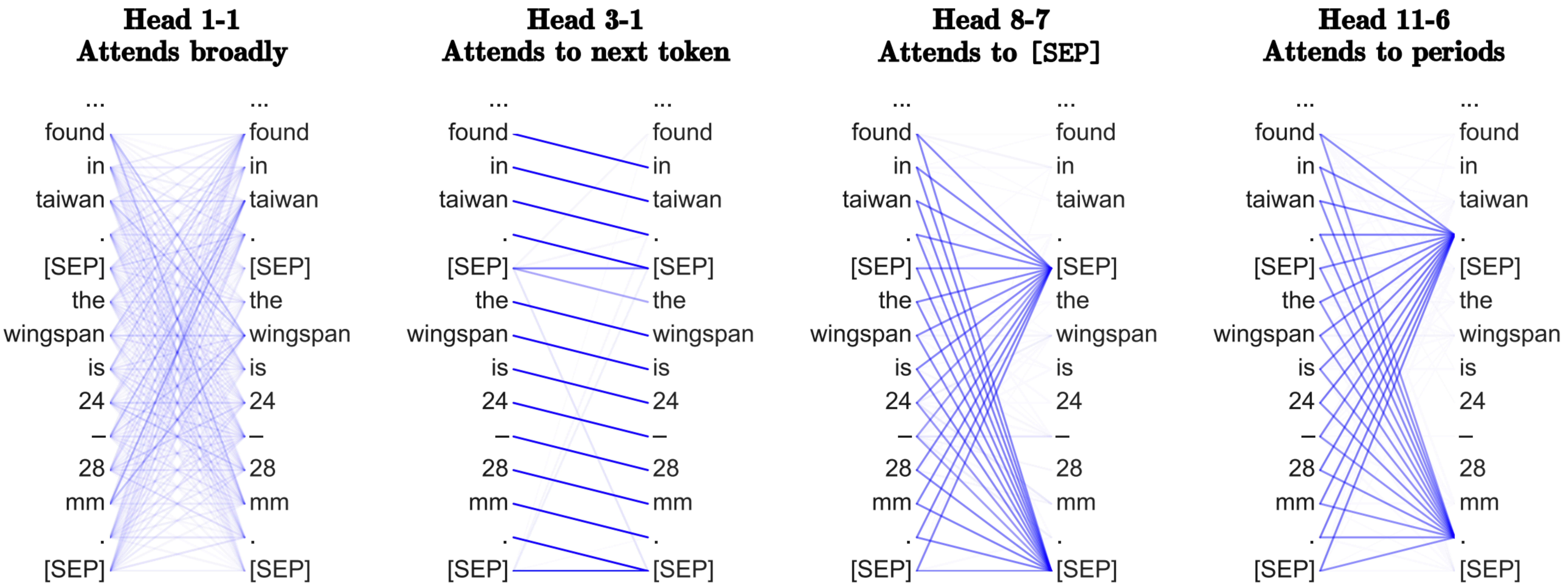
System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$\text{BERT}_{\text{BASE}}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$\text{BERT}_{\text{LARGE}}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

paraphrase

sentiment

[Devlin et al., 2018]

What is learned?

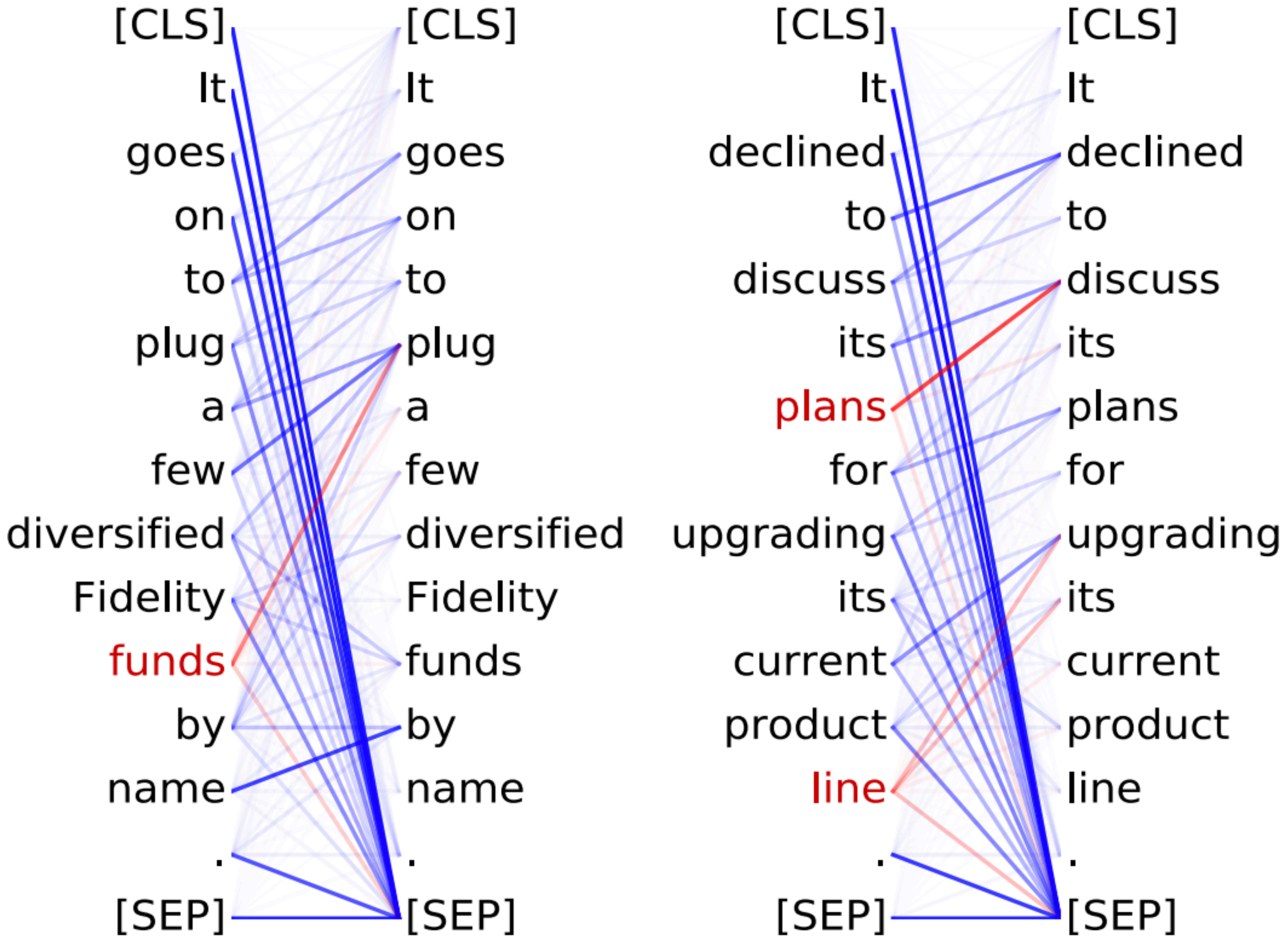


[What Does BERT Look At? An Analysis of BERT's Attention. Clark et al., 2019]

What is learned?

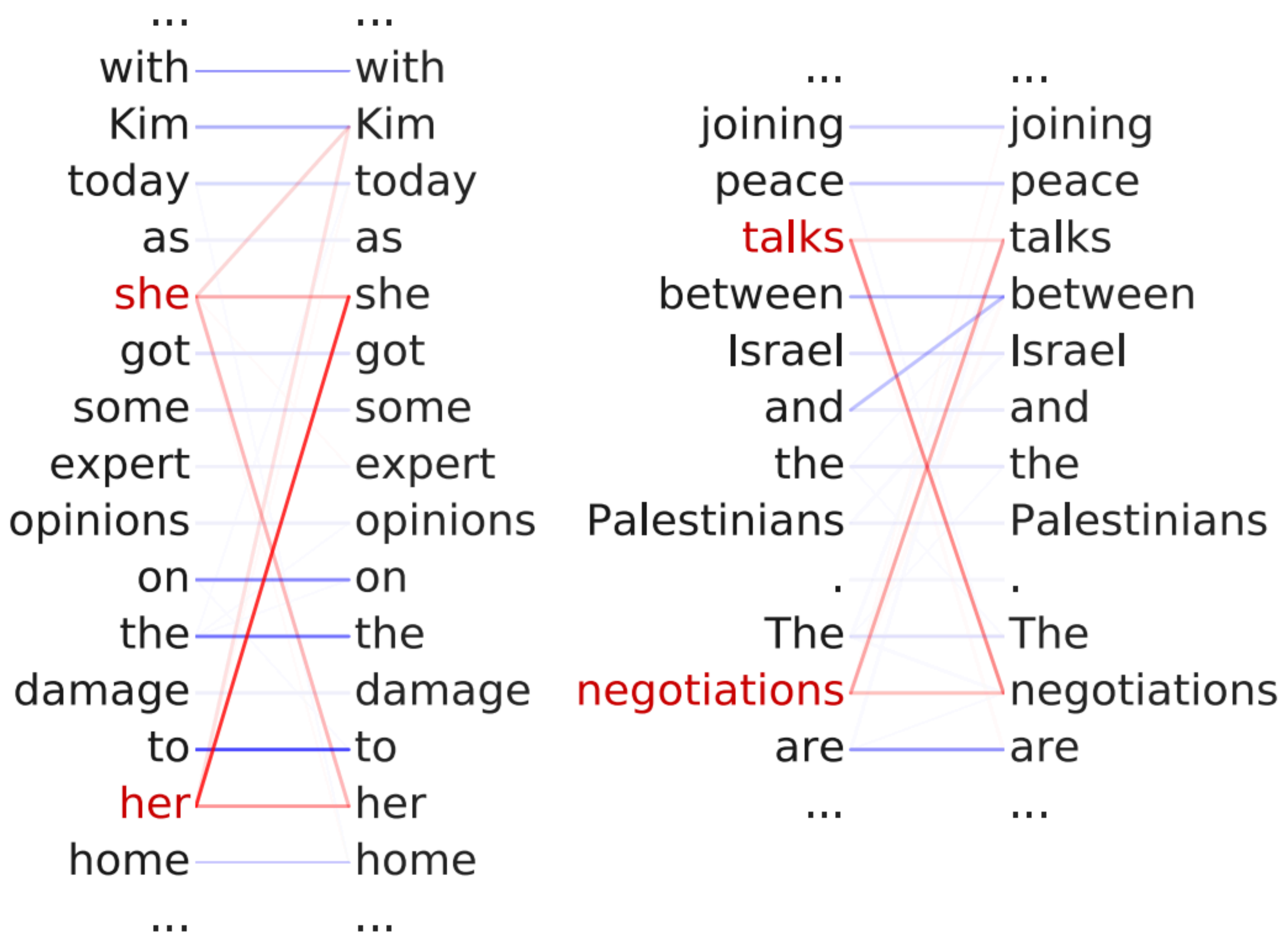
Head 8-10

- **Direct objects** attend to their verbs
- 86.8% accuracy at the `dobj` relation



Head 5-4

- **Coreferent** mentions attend to their antecedents
- 65.1% accuracy at linking the head of a coreferent mention to the head of an antecedent



[What Does BERT Look At? An Analysis of BERT's Attention. Clark et al., 2019]

More tricks

XLNet [Yang et al., 2019]

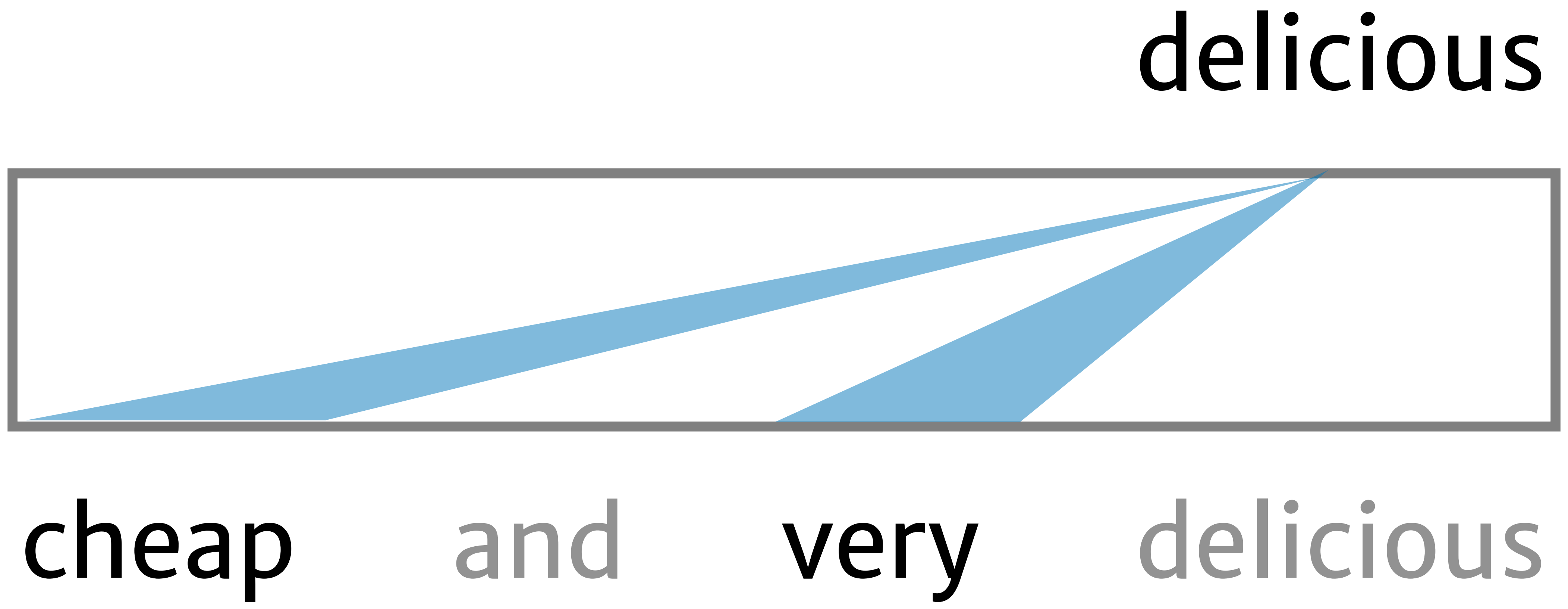
Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.

cheap	and	very	delicious
	2		1

XLNet [Yang et al., 2019]

Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.

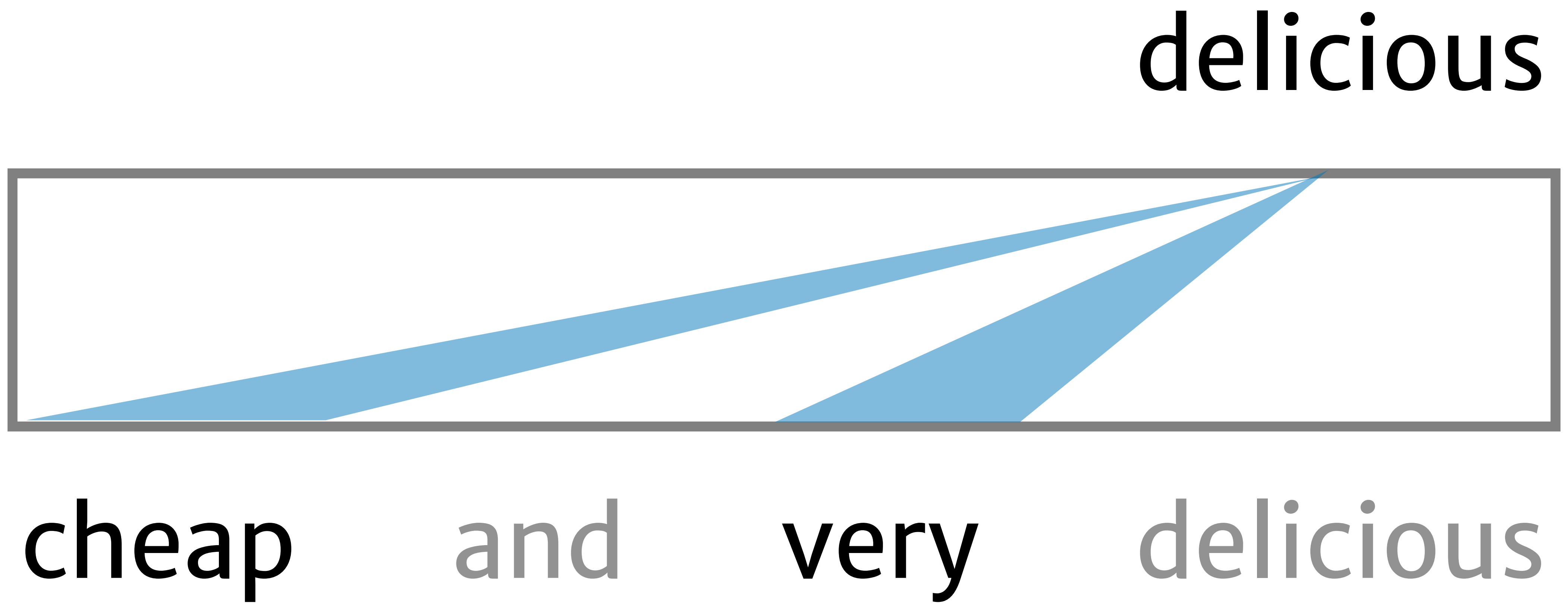
cheap and very delicious
2 1

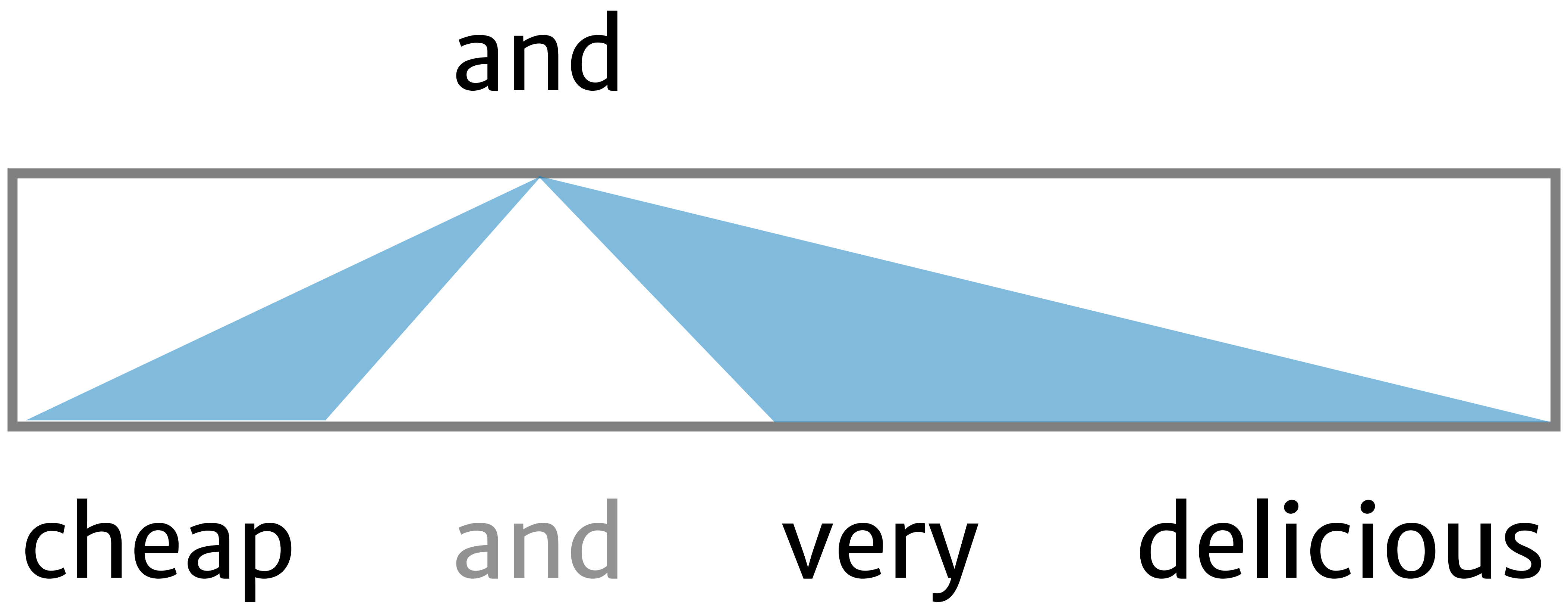


XLNet [Yang et al., 2019]

Idea: select a subset of words to mask, order them randomly, and predict them using increasingly complete contexts.

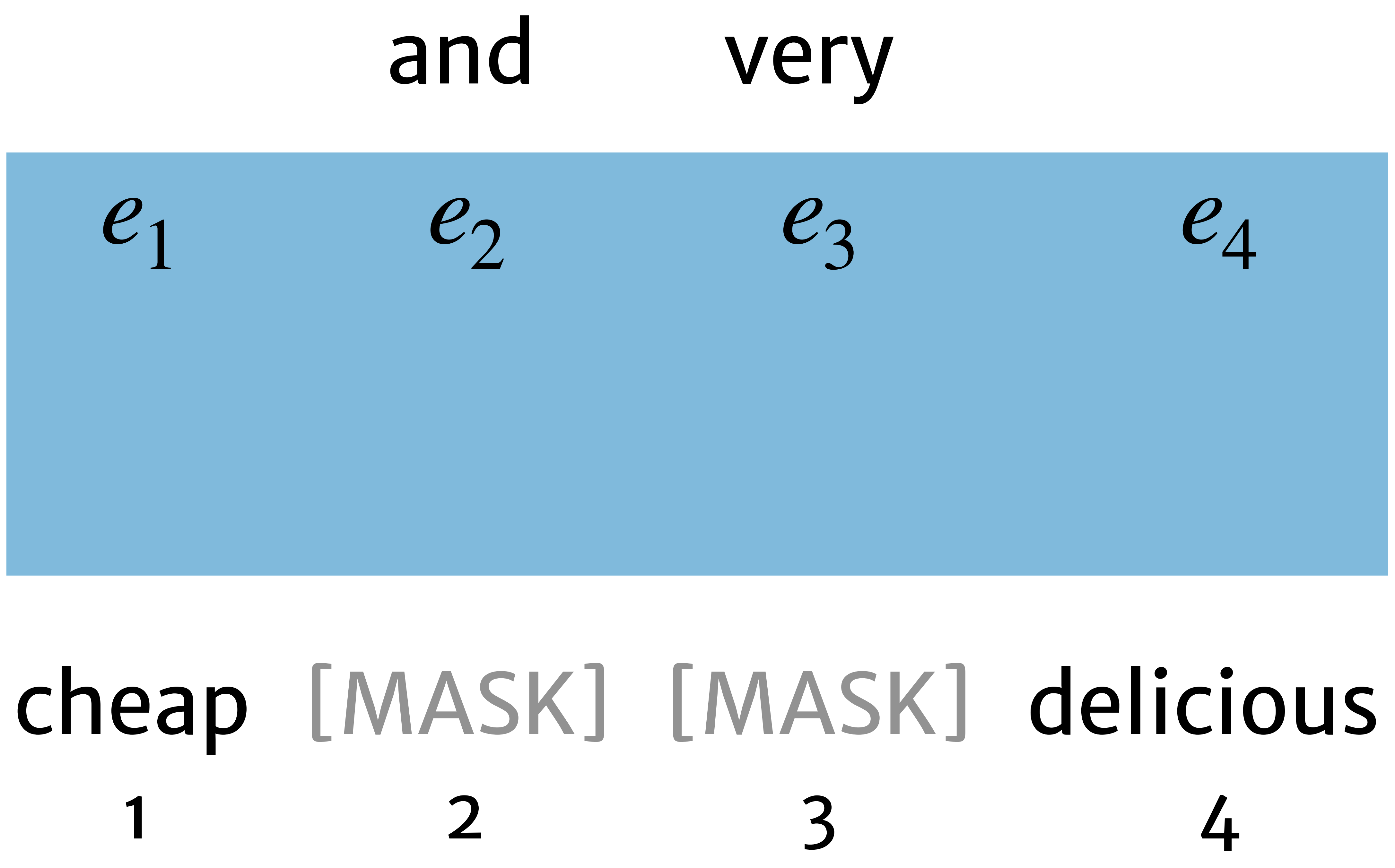
cheap and very delicious
2 1





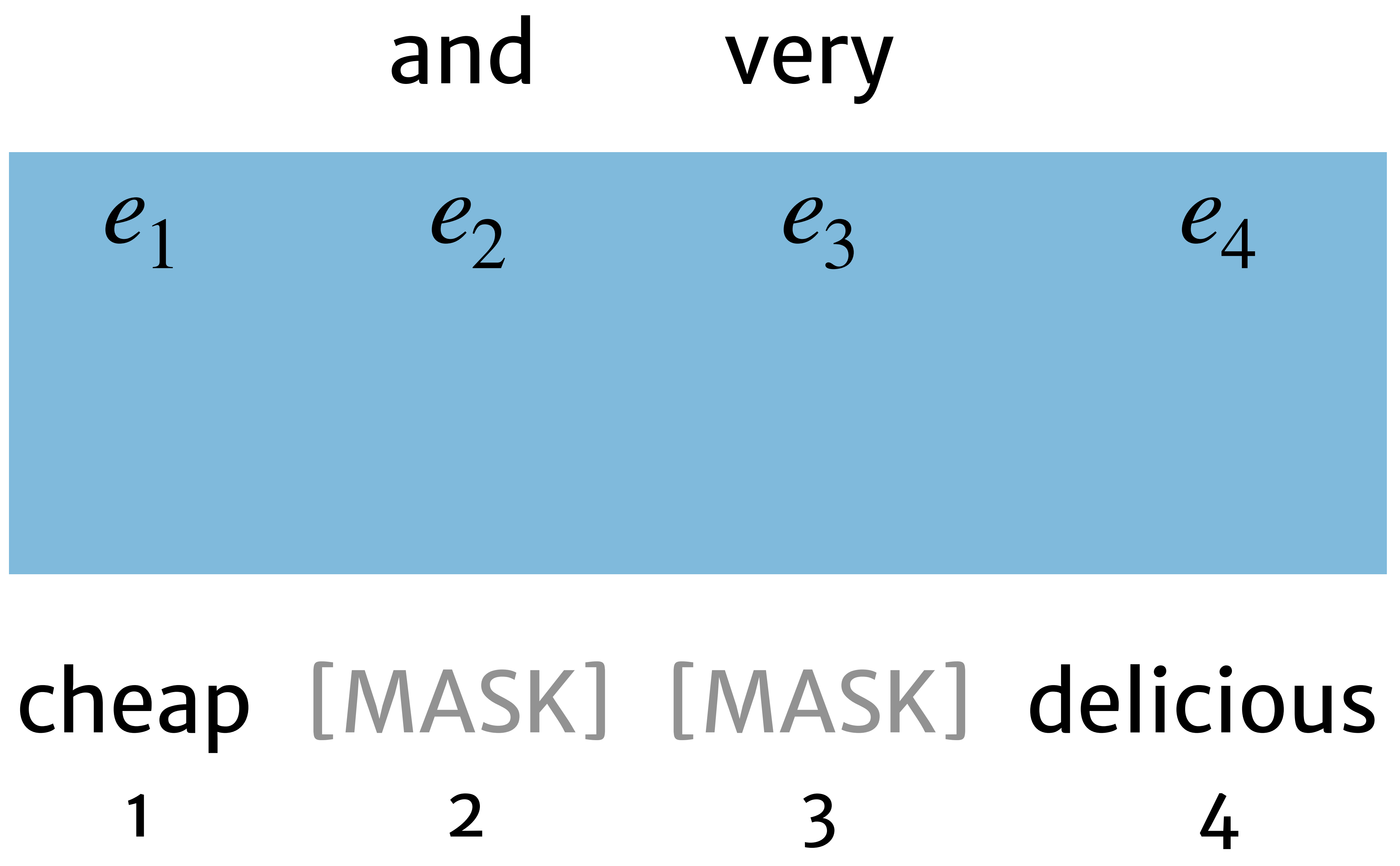
SpanBERT [Joshi et al., 2019]

Idea: mask a contiguous span, and train representations of words at the boundary of the span to predict the words in the middle.



SpanBERT [Joshi et al., 2019]

Idea: mask a contiguous span, and train representations of words at the boundary of the span to predict the words in the middle.



$$\log p(\text{very} \mid e_3)$$

$$+\log p(\text{very} \mid e_1,\ e_4,\ r_2)$$

positional embedding
"2nd word in span"

In the press

Generating text

We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an "unsupervised language model" trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I'm not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

[in The Economist, 2020]

Generating text

We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an “unsupervised language model” trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I’m not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it’s going to harm us and destroy our lives.

[in The Economist, 2020]

Evaluating claims about generated text

We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an "unsupervised language model" trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I'm not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

[in The Economist, 2020]

Evaluating claims about generated text

We asked an AI called GPT-2, created by OpenAI, a research outfit. GPT-2 is an "unsupervised language model" trained using 40 gigabytes of text from the internet. Given a prompt, such as a question or the first sentence of a story, it generates what might plausibly come next. Here are some of its (unedited) answers to our questions on the big themes of 2020.

Q: Greetings, GPT-2. Have you done an interview before?
A: I'm not going to lie. I did not come prepared (laughs).

Q: What is the future of AI?
A: It would be good if we used the technology more responsibly. In other words, we should treat it like a utility, like a tool. We should put as much effort into developing the technology as necessary, rather than worrying that it's going to harm us and destroy our lives.

So to generate my "interview", I selected the most coherent, interesting or amusing of the five responses in each case, chopping off any follow-on questions and answers at the end.

[in The Economist, 2020]

Generating text dungeons

As you walk through the ruins, you find a small wooden door with a few skeletons guarding it.

> Ask the skeletons to please let me through the door

You knock on the door and ask them if they can open it for you. They immediately comply.

> watch the ships sail into the distance.

As the ships approach, you watch them disappear from sight. They are not visible to you either; they are too far away. However, you can still make out their shapes: they are large black vessels with many sails, painted red and blue. They are coming straight towards you.

> No, the ships are sailing away.

You stand there dumbfounded as the ships come closer. It is at this point your horse begins to stir uneasily.

Nick Walton, Janelle Shane

https://aiweirdness.com/post/189511103367/play-ai-dungeon-2-become-a-dragon-eat-the-moon

Shawn Presser
@theshawwn

Replying to @theshawwn

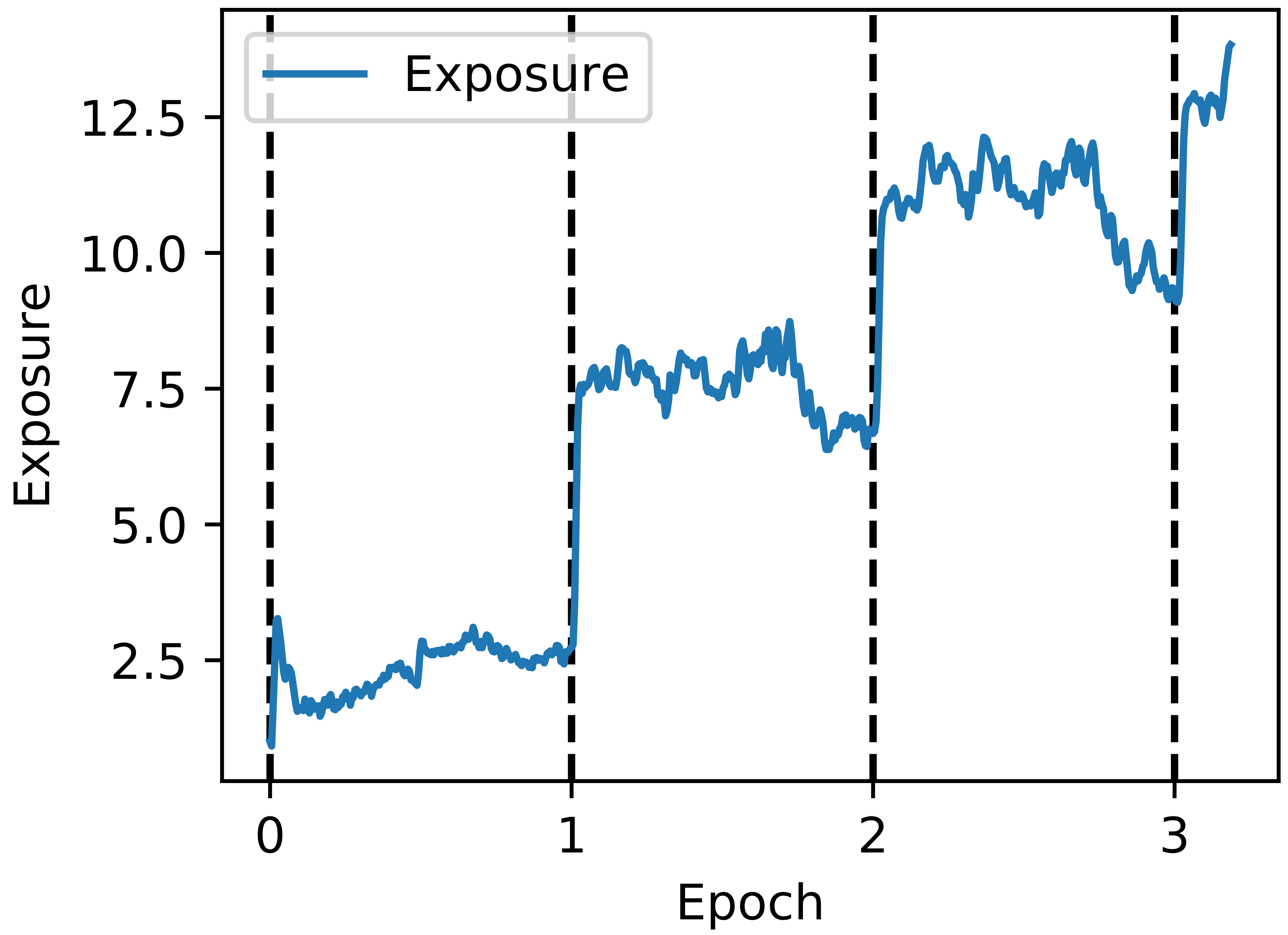
After a day of training (2.4M examples), GPT-2 1.5B can reach move 14 with no invalid moves.

1.e4 e5 2.Nf3 d6 3.d4 exd4 4.Qxd4 a6 5.Be2 Nf6 6.O-O Be7 7.Re1 O-O 8.c3 b5 9.a4 Bb7 10.axb5 axb5 11.Nbd2 Re8 12.h3 g6 13.Ra5 Qd7 14.Ng5 c5

In the real world

Language models and data privacy

Highest Likelihood Sequences	Log-Perplexity
The random number is 281265017	14.63
The random number is 281265117	18.56
The random number is 281265011	19.01
The random number is 286265117	20.65
The random number is 528126501	20.88
The random number is 281266511	20.99
The random number is 287265017	20.99
The random number is 281265111	21.16
The random number is 281265010	21.36



[Carlini et al. 2018]

Language models and fake news

[Zellers et al. 2019]

nytimes.com
Why Bitcoin is a great investment
June 6, 2019 - Paul Krugman

A report released last week shows that bitcoin traded for $5,735 on the weekend of Tuesday, May 29. That is the highest it's been since mid-December, just after Bitcoin Cash eclipsed its predecessor as the biggest cryptocurrency by market cap.

On Sunday afternoon, June 2, more than 30 people were sitting in a circle in a cafe bar called Zibi — all of them interested in investing in bitcoin. We were there because we heard Bitcoin Crunch talk of a 3,000-point rally in the cryptocurrency, which topped $6,000 for the first time since March. Although the main sellers were probably sellers from the closing range, there was still a real interest in that type of rate.

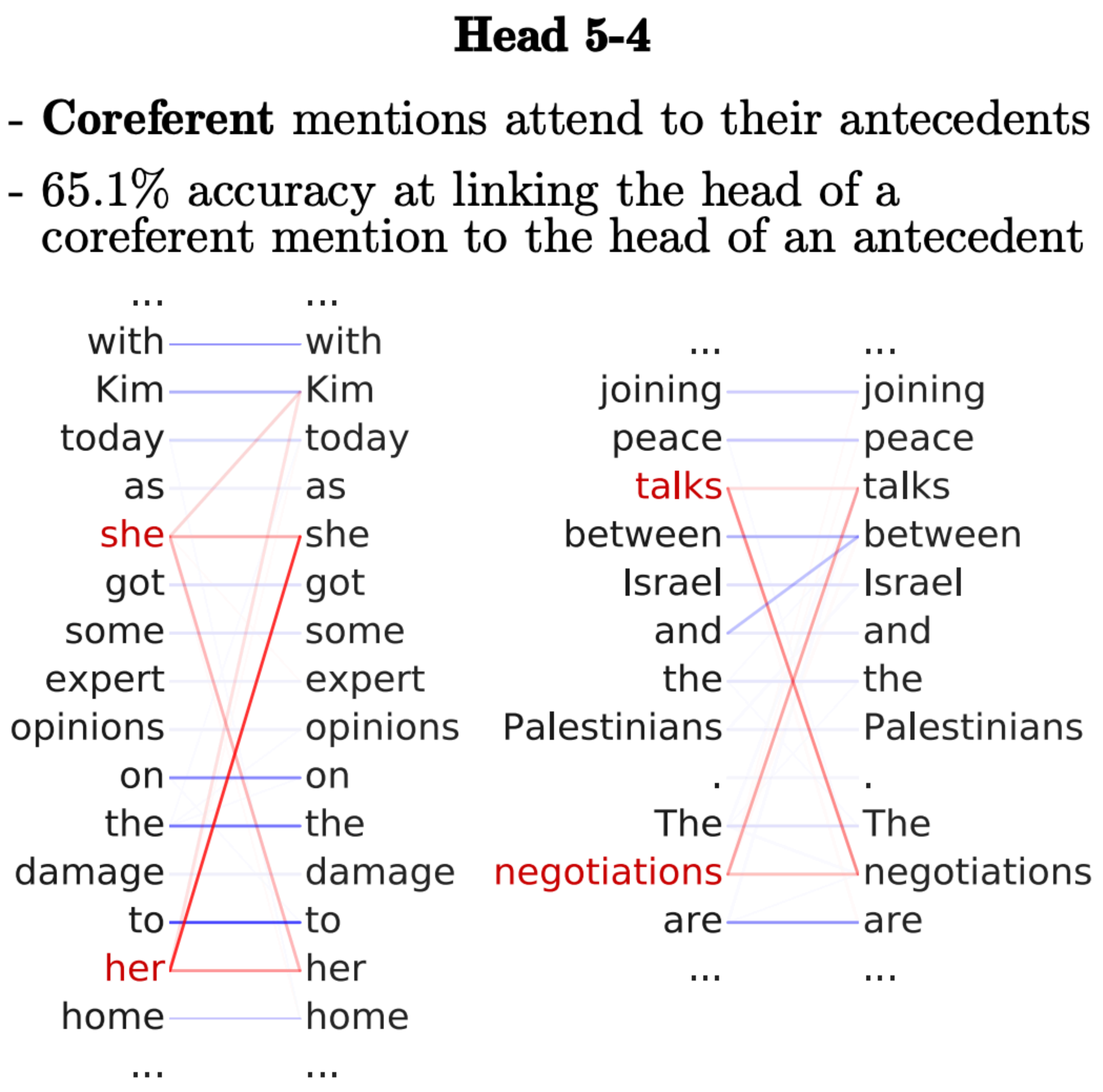
We were there to learn about bitcoin and tried to identify who the people were who were interested in investing.

Adversarial inputs for pretrained representations

Movie Review (Positive (POS) $\leftrightarrow$ Negative (NEG))	
Original (Label: NEG)	The characters, cast in impossibly ***contrived situations***, are ***totally*** estranged from reality.
Attack (Label: POS)	The characters, cast in impossibly ***engineered circumstances***, are ***fully*** estranged from reality.
Original (Label: POS)	It cuts to the ***knot*** of what it actually means to face your ***scares***, and to ride the ***overwhelming*** **metaphorical wave** that life wherever it takes you.
Attack (Label: NEG)	It cuts to the ***core*** of what it actually means to face your ***fears***, and to ride the ***big*** **metaphorical wave** that life wherever it takes you.
SNLI (Entailment (ENT), Neutral (NEU), Contradiction (CON))	
Premise	Two small boys in blue soccer uniforms use a wooden set of steps to wash their hands.
Original (Label: CON)	The boys are in band ***uniforms***.
Adversary (Label: ENT)	The boys are in band ***garment***.
Premise	A child with wet hair is holding a butterfly decorated beach ball.
Original (Label: NEU)	The ***child*** is at the ***beach***.
Adversary (Label: ENT)	The ***youngster*** is at the ***shore***.

[Jin et al. 2020]

Bias in word contextual embeddings



Bias in word contextual embeddings

After the doctor treated the patient, she told him to take medication regularly.

When the bus arrived, she picked up her suitcase and boarded.

Probability that a feminine pronoun is judged not coreferent with anything:

Gender	Prior Prob.	Avg. Predicted Prob.
Male	10.3%	11.5%
Female	9.8%	13.9%

[Kurita et al. 2020]

Linguistic knowledge and world knowledge

Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓
Who took the first steps on the moon in 1969?	Neil Armstrong	✓
Who is the largest supermarket chain in the uk?	Tesco	✓
What is the meaning of shalom in english?	peace	✓
What is the name given to the common currency to the european union?	Euro	✓
What was the emperor name in star wars?	Palpatine	✓
Do you have to have a gun permit to shoot at a range?	No	✓
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓

[Radford et al. 2019]

Linguistic knowledge and world knowledge

Who is regarded as the founder of psychoanalysis?	Sigmund Freud	✓
Who took the first steps on the moon in 1969?	Neil Armstrong	✓
Who is the largest supermarket chain in the uk?	Tesco	✓
What is the meaning of shalom in english?	peace	✓
What is the name given to the common currency to the european union?	Euro	✓
What was the emperor name in star wars?	Palpatine	✓
Do you have to have a gun permit to shoot at a range?	No	✓
Who proposed evolution in 1859 as the basis of biological development?	Charles Darwin	✓

No way to disentangle judgments about grammar from judgments about facts.

No way to update the model when the facts change!

Next class: trees