Natural Language Processing with Deep Learning CS224N/Ling284

John Hewitt

Lecture 18: Model Analysis and Explanation

Course logistics

1. Guest lecture reactions
   1. **[updated]** All due on Friday, March 12 at 11:59PM US-Pacific.
2. Final project report
   1. Due date is Saturday, March 18 at 11:59PM US-Pacific
   2. Hard deadline with late days!
3. It's the end stretch! Thanks for all your hard work this quarter and good luck in the final days!

Lecture Plan

1. Motivating model analysis and explanation
2. One model at multiple levels of abstraction
3. Out-of-domain evaluation sets
   1. Testing for linguistic knowledge
   2. Testing for task heuristics
4. Influence studies and adversarial examples
   1. What part of my input led to this answer?
   2. How could I minimally modify this input to change the answer?
5. Analyzing representations
   1. Correlation in "interpretable" model components
   2. Probing studies: supervised analysis
6. Revisiting model ablations as analysis

https://xkcd.com/1838/

Motivation: what are our models doing?

input
sentence

Your final
project model
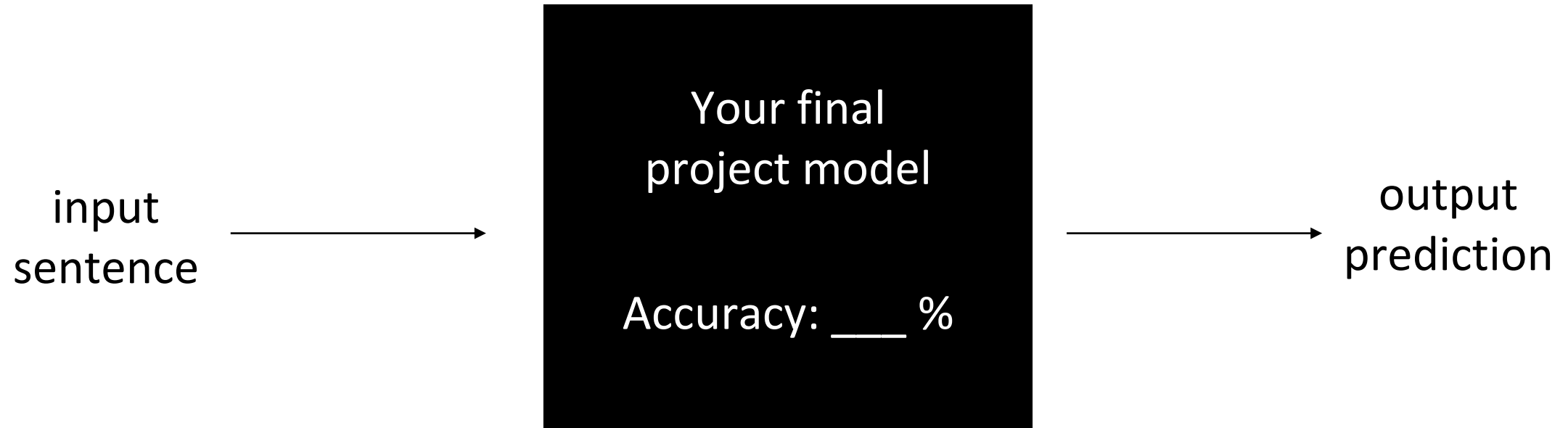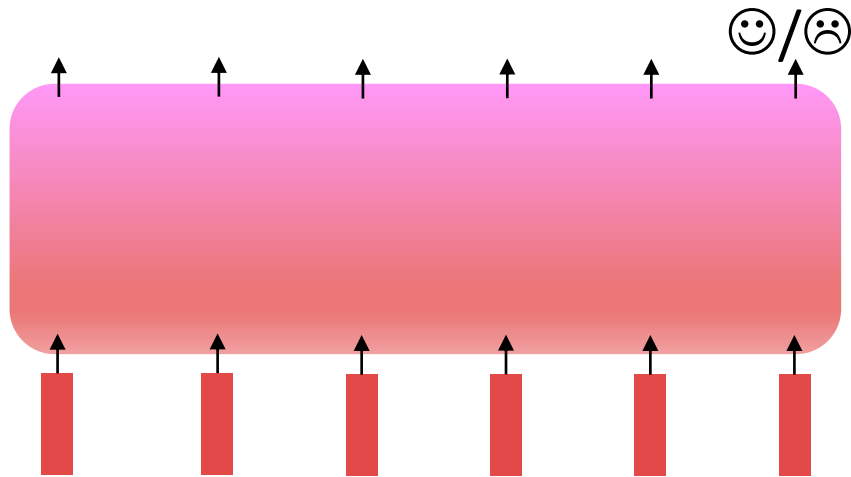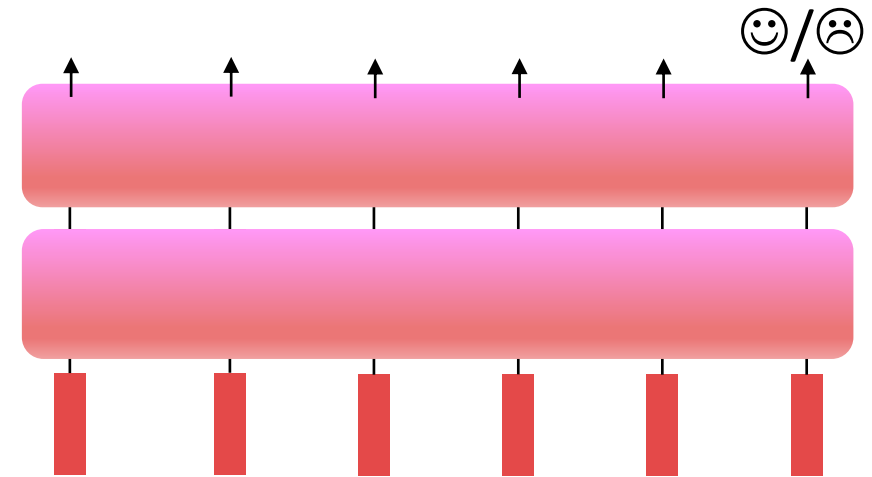
Accuracy: ___ %

output
prediction

Fig 1. *A black box*

We summarize our models with one (or a handful) of accuracies metric numbers.

What do they learn? Why do they succeed and fail?

Motivation: how do we make tomorrow's model?

Today's models: use recipes that work, but aren't perfect

Tomorrow's models: take what works and find what needs changing

Understanding **how far** we can get with incremental improvements on current methods is crucial to the eventual development of major improvements.

Motivation: what biases are built into my model?
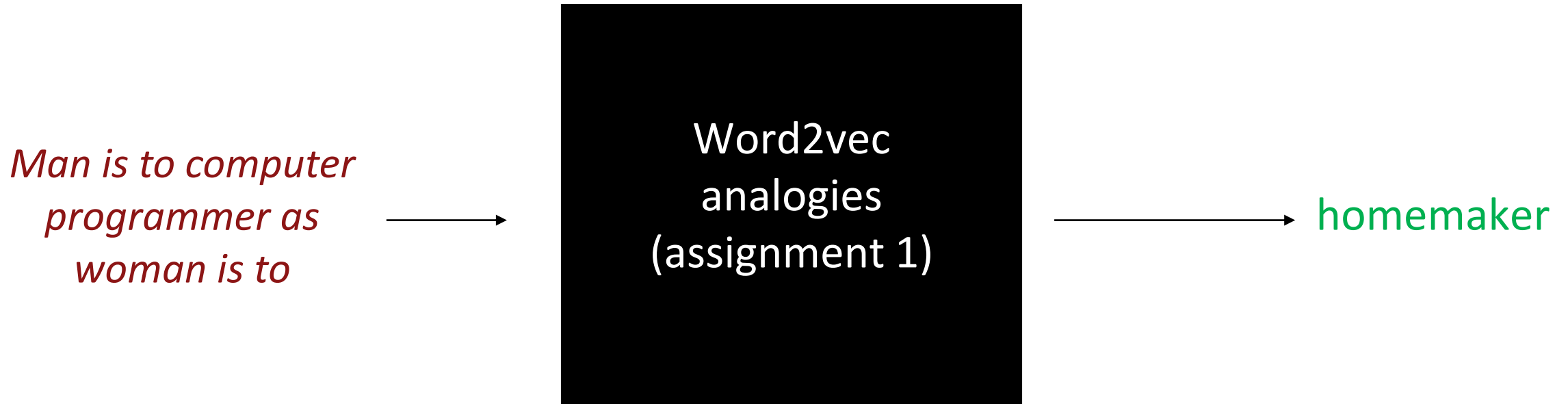
*Man is to computer programmer as woman is to* →

Word2vec analogies (assignment 1)

→ homemaker

Fig 1. *A black box*

What did the model use in its decision?

What biases did it learn and possibly worsen?

[Bolukbasi et al., 2016]

Motivation: how do we make the next 25 years of models?

What can be learned via language model pretraining?

What will replace the Transformer?

What **can't** be learned via language model pretraining?
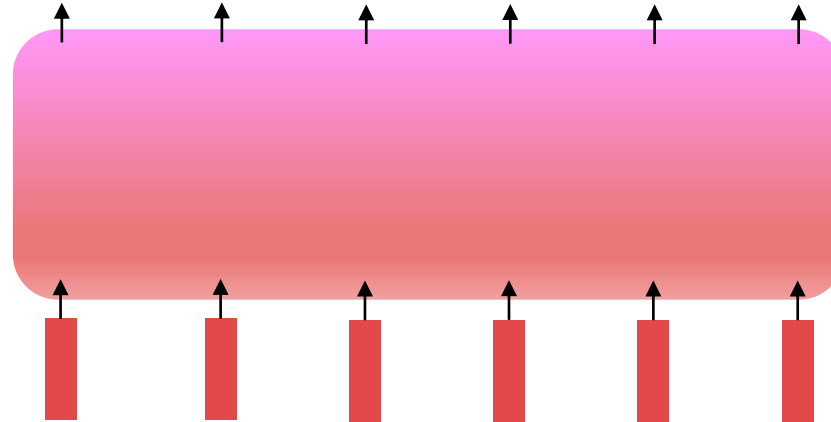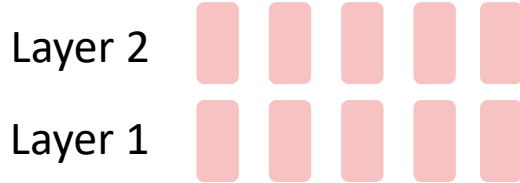
What does deep learning struggle to do?
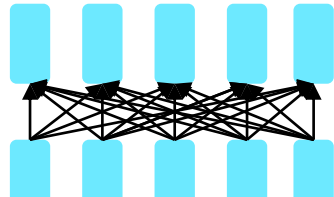
How are our models affecting people, and transferring power?

What do neural models tell us about language?

Model analysis at varying levels of abstraction

There is a **wide variety** of ways to analyze models; **none is perfect or provides total clarity.**

To start, at what level of **abstraction** do you want to reason about your model?

1. Your neural model as a probability distribution and decision function

$$p_{\text{model}}(y|x)$$

2. Your neural model as a sequence of vector representations in depth and time

Layer 2

Layer 1

3. Parameter weights, specific mechanisms like attention, dropout, +++

Outline

Model evaluation as model analysis

When looking at the **behavior** of a model, we're not yet concerned with **mechanisms** the model is using. We want to ask *how does model behave in situations of interest?*

- You've trained your model on some samples $(x, y) \sim D$ from some distribution.
- How does the model behave on samples from the same distribution?
  - Aka *in-domain* or *i.i.d.* (independently and identically distributed)
  - **This is your test set accuracy / F1 / BLEU**

Model A *Accuracy: 95%*	? >	Model B *Accuracy: 92%*

[Also, both models seem pretty good?]

Model evaluation as model analysis in **natural language inference**

Recall the **natural language inference** task, as encoded in the Multi-NLI dataset.

Premise
"He turned and saw Jon sleeping in his half-tent"

Hypothesis
"He saw Jon was asleep"

Model A
Accuracy: 95%

Entailment

Neutral

Contradiction
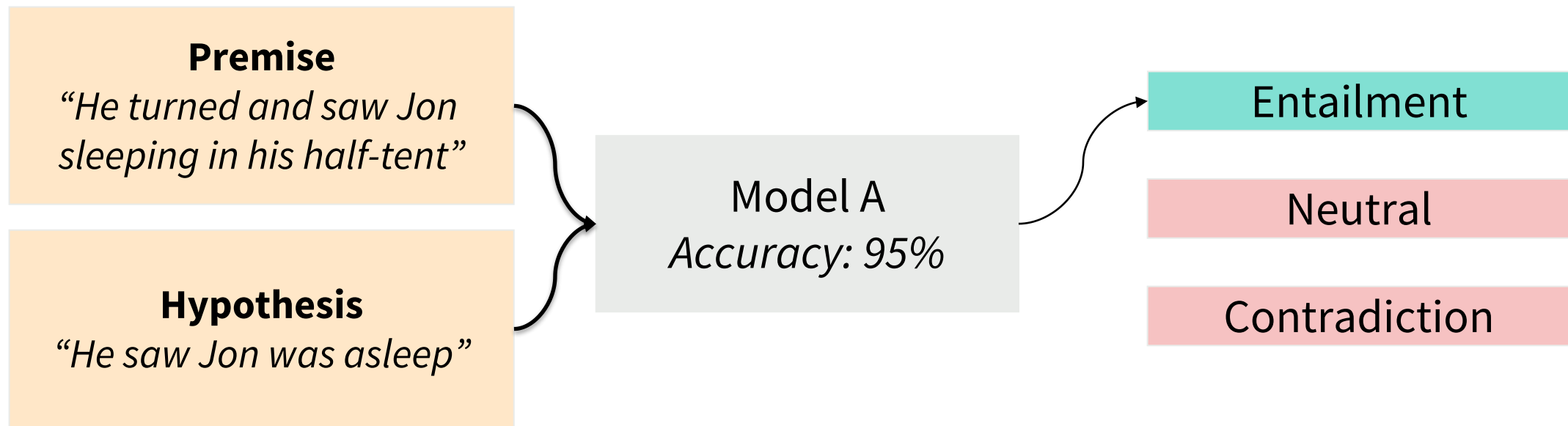
[Likely to get the right answer, since the accuracy is 95%?]

[Williams et al., 2018]

Model evaluation as model analysis in **natural language inference**

What if our model is using simple heuristics to get good accuracy?

A **diagnostic test set** is carefully constructed to test for a specific skill or capacity of your neural model.

For example, **HANS**: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI

Heuristic	Definition	Example
Lexical overlap	Assume that a premise entails all hypotheses constructed from words in the premise	**The doctor** was **paid** by **the actor**. $\xrightarrow[\text{WRONG}]{}$ The doctor paid the actor.
Subsequence	Assume that a premise entails all of its contiguous subsequences.	The doctor near **the actor danced**. $\xrightarrow[\text{WRONG}]{}$ The actor danced.
Constituent	Assume that a premise entails all complete subtrees in its parse tree.	If **the artist slept**, the actor ran. $\xrightarrow[\text{WRONG}]{}$ The artist slept.

[McCoy et al., 2019]

HANS model analysis in **natural language inference**

McCoy et al., 2019 took 4 strong MNLI models,
with the following accuracies on the **original test set (in-domain)**
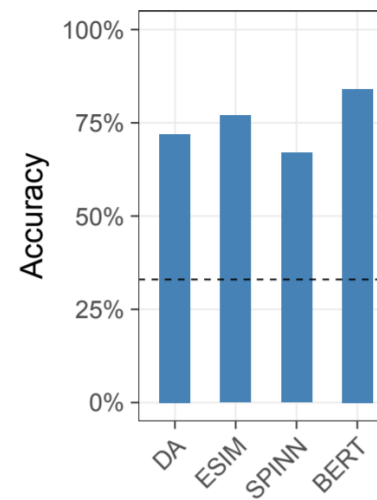
Evaluating on HANS, where syntactic heursitcs **work**, accuracy is high!

But where syntactic heuristics fail, accuracy is very very low…

[McCoy et al., 2019]
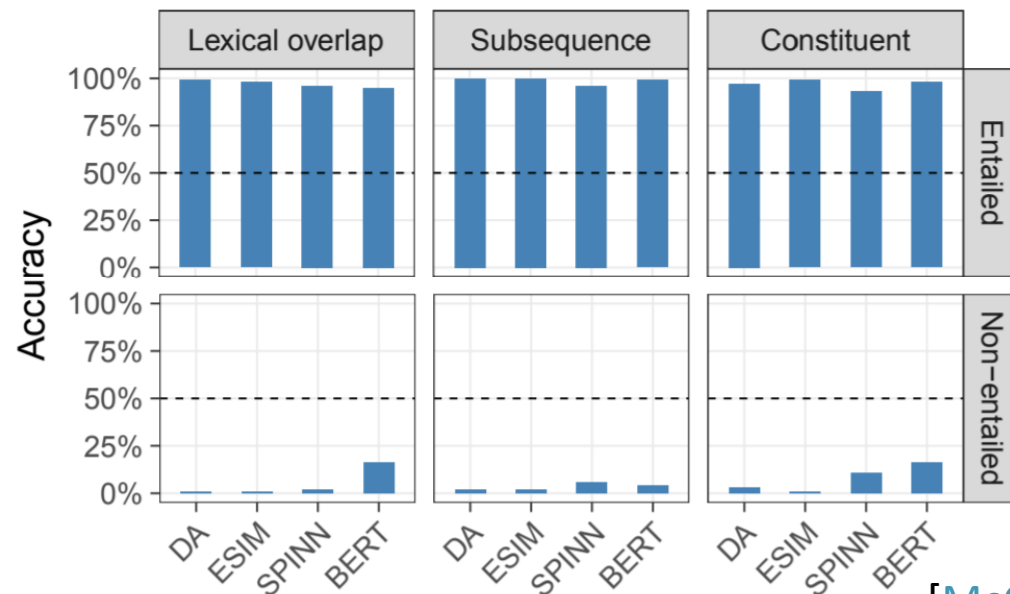
~~Language models~~ as linguistic test subjects
Humans

- How do we understand language behavior in humans?
- One method: *minimal pairs*. What sounds "okay" to a speaker, but doesn't with a small change?

The chef who made the pizzas is here. ← "Acceptable"

The chef who made the pizzas are here ← "Unacceptable"

Idea: English past-tense verbs *agree in number* with their subjects

[Linzen et al., 2016; Fig from Manning et al., 2020]

Language models as linguistic test subjects

- What's the language model analogue of acceptability?

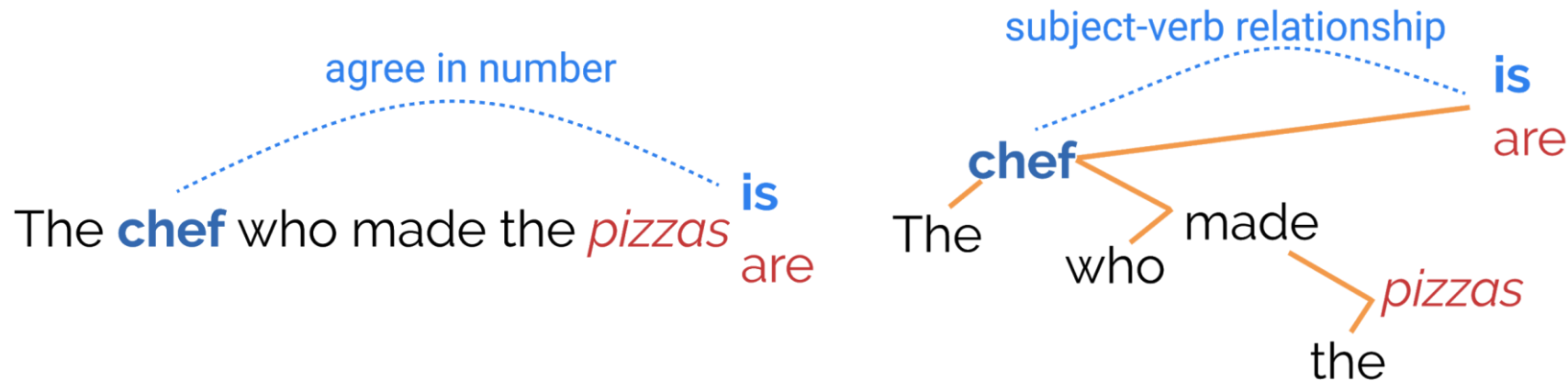   The chef who made the pizzas is here. ← "Acceptable"

    The chef who made the pizzas are here ← "Unacceptable"

- Assign higher probability to the acceptable sentence in the minimal pair

    P(The chef who made the pizzas is here.) > P(The chef who made the pizzas are here)

- Just like in HANS, we can develop a **test set with carefully chosen properties**.
  - Specifically: can language models handle "attractors" in subject-verb agreement?
  - 0 Attractors: The chef is here.
  - 1 Attractor: The chef who made the pizzas is here.
  - 2 Attractors: The chef who made the pizzas and prepped the ingredients is here.
  - …

[Linzen et al., 2016]

Language models as linguistic test subjects

- Kuncoro et al., 2018 train an LSTM language model on a small set of Wikipedia text.
- They evaluate it on and evaluate it *only* on sentences with specific numbers of agreement attractors.
- Numbers in this table: accuracy at predicting the correct number for the verb

Zero attractors: Easy

4 attractors: harder, but models still do pretty well!

	n=0	n=1	n=2	n=3	n=4
Random	50.0	50.0	50.0	50.0	50.0
Majority	32.0	32.0	32.0	32.0	32.0
Our LSTM, H=50	2.4	8.0	15.7	26.1	34.65
Our LSTM, H=150	1.5	4.5	9.0	14.3	17.6
Our LSTM, H=250	1.4	3.3	5.9	**9.7**	13.9
Our LSTM, H=350	**1.3**	**3.0**	**5.7**	**9.7**	**13.8**

The larger LSTMs learn subject-verb agreement better!

[Kuncoro et al., 2016]

Language models as linguistic test subjects

Sample test examples for subject-verb agreement with attractors that a model got wrong

The **ship** that the player drives **has** a very high speed.

The **ship** that the player drives **have** a very high speed.

The **lead** is also rather long; 5 paragraphs **is** pretty lengthy …

The **lead** is also rather long; 5 paragraphs **are** pretty lengthy …

[Linzen et al., 2016]

Careful test sets as unit test suites: CheckListing

- Small careful test sets sound like… unit test suites, but for neural networks!
- *Minimum functionality tests:* small test sets that target a specific behavior.

Test case		Expected	Predicted	Pass?
(A) Testing **Negation** with *MFT*		Labels: negative, positive, neutral		
Template: I {NEGATION} {POS_VERB} the {THING}.				
I can't say I recommend the food.		neg	pos	X
I didn't love the flight.		neg	neutral	X
…				
			Failure rate = 76.4%	

- Ribeiro et al., 2020 showed **ML engineers working on a sentiment analysis product** an interface with categories of linguistic capabilities and types of tests.
  - The engineers found a bunch of bugs (categories of high error) through this method!

[Ribeiro et al., 2020]

Fitting the dataset vs learning the task

Across a wide range of tasks, high model accuracy on the in-domain test set does not imply the model will also do well on other, "reasonable" out-of-domain examples.

One way to think about this: models seem to be learning the *dataset* (like MNLI) not the *task* (like how humans can perform natural language inference).

[Ribeiro et al., 2020]

Knowledge evaluation as model analysis

- What has a language model learned from pretraining?
- More on this later, but last lecture we saw one way of accessing some of the knowledge in the model by providing it with prompts.
- This fits into the set of *behavioral studies* we've seen so far!
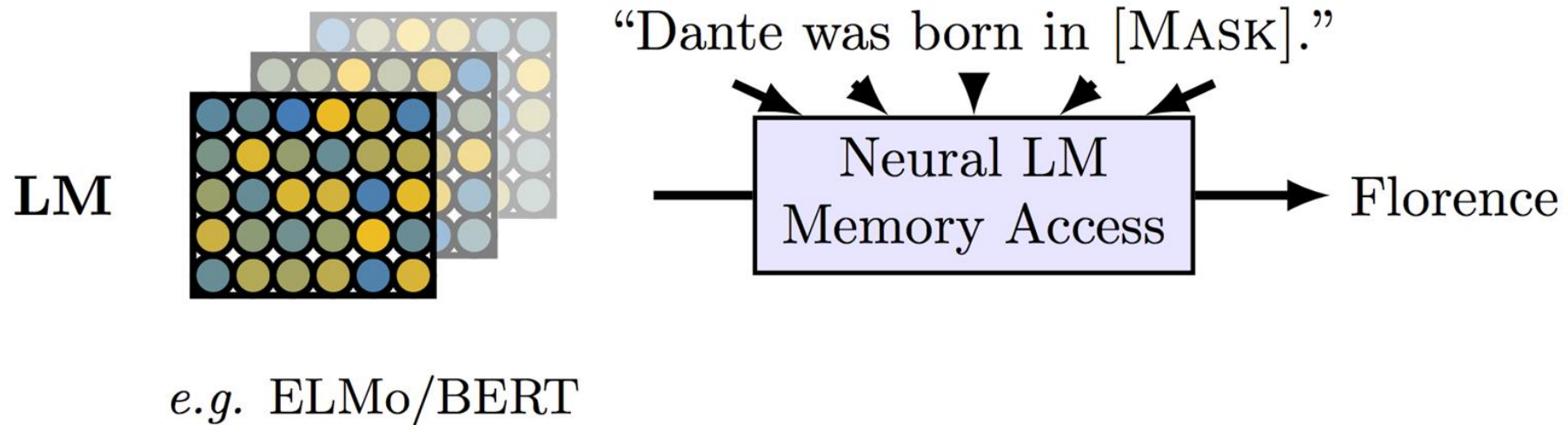
*e.g.* ELMo/BERT

[Petroni et al., 2020]

Outline

1. Motivating model analysis and explanation
2. One model at multiple levels of abstraction
3. Out-of-domain evaluation sets (Your model as a probability distribution)
    1. Testing for linguistic knowledge
    2. Testing for task heuristics
4. Influence studies and adversarial examples
    1. What part of my input led to this answer?
    2. How could I minimally modify this input to change the answer?
5. Analyzing representations
    1. Correlations with simple model components
    2. Probing studies: supervised analysis
6. Revisiting model ablations as analysis

# Input influence: does my model *really* use long-distance context?

- We motivated LSTM language models through their theoretical ability to use long-distance context to make predictions. But how long really is the long short-term memory?

- Khandelwal et al., 2018's idea: shuffle or remove all contexts farther than k words away for multiple values of k and see at which k the model's predictions start to get worse!

- Loss is averaged across many examples.
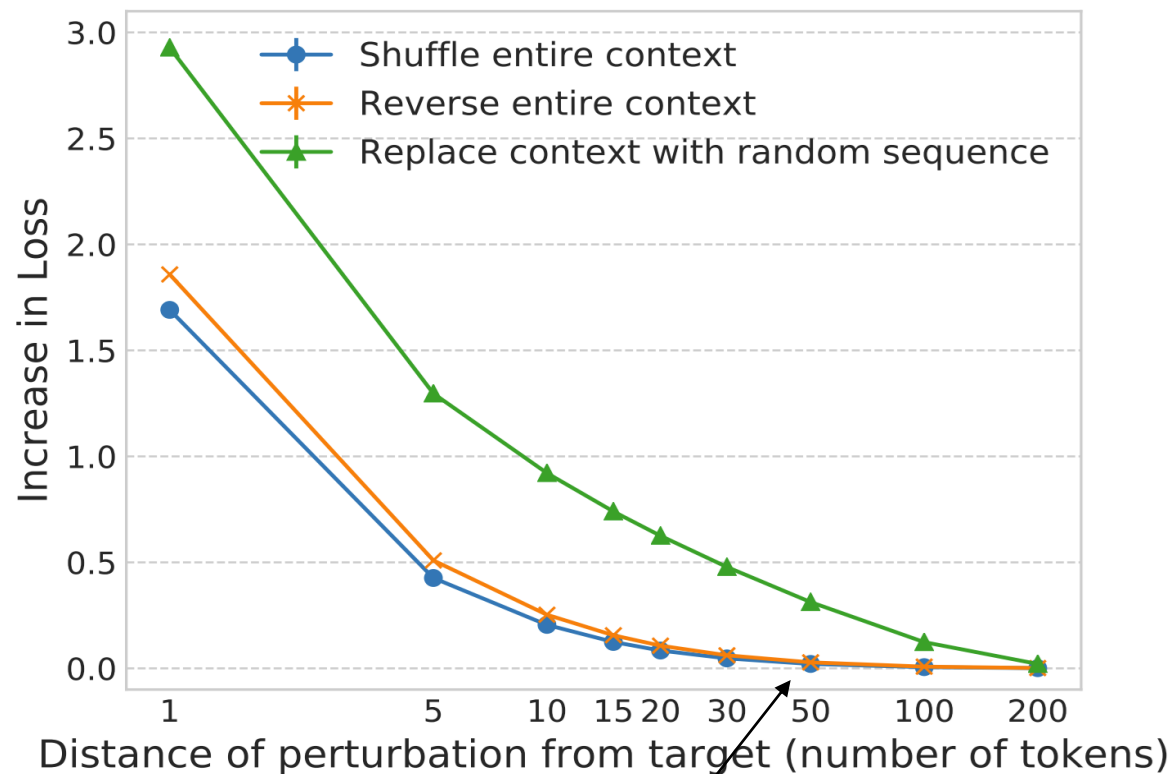
History farther than 50 words away treated as a bag of words.

[Khandelwal et al., 2018]

Prediction explanations: what in the input led to this output?

- For a single example, what parts of the input led to the observed prediction?
- **Saliency maps**: a score for each input word indicating its importance to the model's prediction
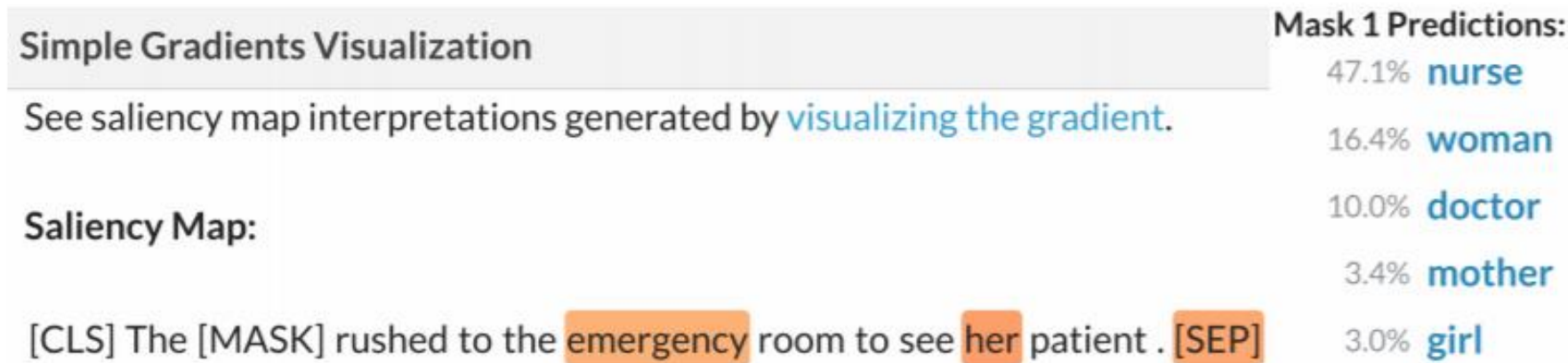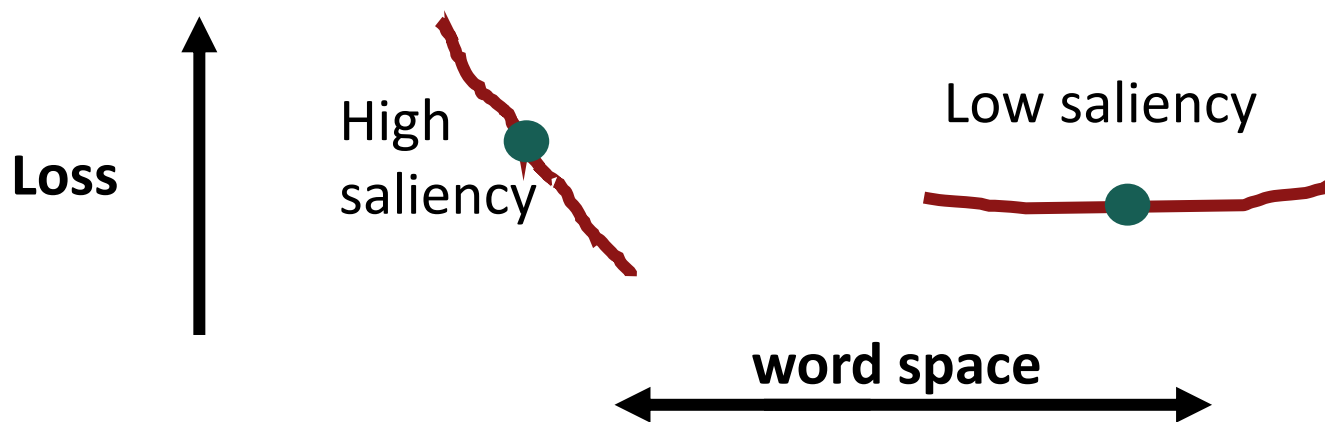

- In the above example, BERT is analyzed, and interpretable words seem to contribute to the model's predictions (right).

[Simonyan et al., 2014, Wallace et al., 2019]

Prediction explanations: simple saliency maps

- How do we make a saliency map? Many ways to encode the intuition of "importance"
- **Simple gradient method:**

  For words $x_1, \dots, x_n$ and the model's score for a given class (output label) $s_c(x_1, \dots, x_n)$, take the norm of the gradient of the score w.r.t. each word:

  $$\text{salience}(x_i) = \ ||\nabla_{x_i} \, s_c(x_1, \dots, x_n)||$$

  Idea: **high gradient norm** means changing that word (locally) would affect the score a lot

Loss

High saliency

Low saliency

word space

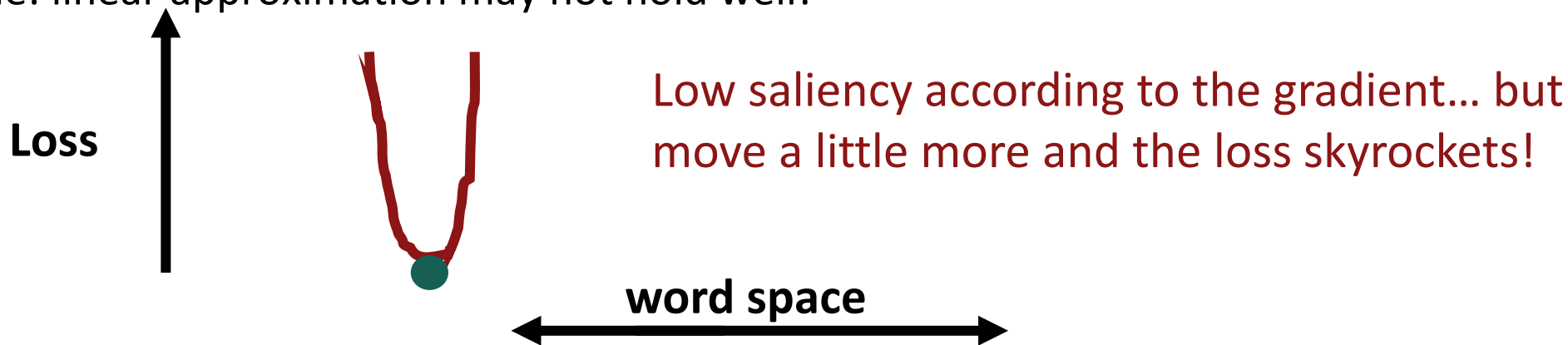[Li et al., 2016, Simonyan et al., 2014, Wallace et al., 2019]

Prediction explanations: simple saliency maps

- How do we make a saliency map? Many ways to encode the intuition of "importance"
- **Simple gradient method:**
  For words $x_1, \ldots, x_n$ and the model's score for a given class (output label) $s_c(x_1, \ldots, x_n)$, take the norm of the gradient of the score w.r.t. each word:

$$salience(x_i) = |\nabla_{x_i} s_c(x_1, \ldots, x_n)|$$

Not a perfect method for saliency; many more methods have been proposed.
One issue: linear approximation may not hold well!

Loss

Low saliency according to the gradient… but move a little more and the loss skyrockets!

word space

[Li et al., 2016, Simonyan et al., 2014, Wallace et al., 2019]

Explanation by input reduction

What is the smallest part of the input I could keep and still get the same answer?

An example from SQuAD:

Passage: In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.

[prediction]

Original Question: What did Tesla spend Astor's money on ?

Reduced Question did

In this example, the model had confidence 0.78 for the original question, and the same answer at confidence **0.91** for the reduced question!

[Feng et al., 2018]

A method for explanation by input reduction

Idea: **run an input saliency method. Iteratively remove the most unimportant words.**

Passage:
The Panthers used the San Jose State practice facility and stayed at the San Jose Marriott. The Broncos practiced at Stanford University and stayed at the Santa Clara Marriott.

[prediction]

Original Question:
Where did the Broncos practice for the Super Bowl ?
Where did the practice for the Super Bowl ?
Where did practice for the Super Bowl ?
Where did practice the Super Bowl ?
Where did practice the Super ?
Where did practice Super ?
did practice Super ?

Steps of input reduction

[Note: beam search to find *k* least important words is an important addition]

Only here did the model stop being confident in the answer

[Feng et al., 2018]

Analyzing models by breaking them

Idea: **Can we break models by making seemingly innocuous changes to the input?**

Passage: Peyton manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by **John Elway**, who led the Broncos to victory in Super Bowl XXXIII at age 38…

[prediction]

Question: What was the name of the quarterback who was 38 in Super Bowl XXXIII?

Looks good!

[Jia et al., 2017]

Analyzing models by breaking them

Idea: **Can we break models by making seemingly innocuous changes to the input?**

Passage:
Peyton manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38... Quarterback **Jeff Dean** had jersey number 37 in Champ Bowl XXXIV.

[prediction]

Question:
What was the name of the quarterback who was 38 in Super Bowl XXXIII?

The sentence in orange hasn't changed the answer, but the model's prediction changed!

So, seems like the model wasn't performing question answering as we'd like?

[Jia et al., 2017]

Analyzing models by breaking them

Idea: **Can we break models by making seemingly innocuous changes to the input?**

In the United States especially, several high-profile cases such as Debra LaFave, Pamela Rogers, and Mary Kay Letourneau have caused increased scrutiny on teacher misconduct.

(a) Input Paragraph

This model's predictions look good!

Q: What has been the result of this publicity?
A: increased scrutiny on teacher misconduct

(b) Original Question and Answer

Q: What haL been the result of this publicity?
A: teacher misconduct

(c) Adversarial Q & A (Ebrahimi et al., 2018)

This typo is annoying, but a reasonable human might ignore it.

Q: What's been the result of this publicity?
A: teacher misconduct

(d) **Semantically Equivalent Adversary**

Changing *what* to *what's* should never change the answer!

[Ribeiro et al., 2018]

Are ~~models~~ **Humans** robust to noise in their input?

"Aoccdrnig to a rscheearch at Cmabrigde Uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, the olny iprmoetnt tihng is taht the frist and lsat ltteer be at the rghit pclae."

Seemingly so!

Are models robust to noise in their input?

Noise of various kinds is an inevitable part of the inputs to NLP systems. How do models trained on (relatively) clean text perform when typo-like noise is added?

Belinkov and Bisk, 2018 performed a study on popular machine translation models.

Character-swaps like we just saw break the model!

BLEU scores are high on in-domain clean text

(More) natural typo noise also breaks the models.

				Synthetic			
		Vanilla	Swap	Mid	Rand	Key	Nat
French	charCNN	42.54	10.52	9.71	1.71	8.26	17.42
German	charCNN	34.79	9.25	8.37	1.02	6.40	14.02
	char2char	29.97	5.68	5.46	0.28	2.96	12.68
	Nematus	34.22	3.39	5.16	0.29	0.61	10.68
Czech	charCNN	25.99	6.56	6.67	1.50	7.13	10.20
	char2char	25.71	3.90	4.24	0.25	2.88	11.42
	Nematus	29.65	2.94	4.09	0.66	1.41	11.88

[Belinkov and Bisk, 2018]

Outline

1. Motivating model analysis and explanation

2. One model at multiple levels of abstraction

3. Out-of-domain evaluation sets

    1. Testing for linguistic knowledge

    2. Testing for task heuristics

4. Influence studies and adversarial examples

    1. What part of my input led to this answer?

    2. How could I minimally modify this input to change the answer?

5. Analyzing representations

    1. Correlation in "interpretable" model components

    2. Probing studies: supervised analysis

6. Revisiting model ablations as analysis

34

Analysis of "interpretable" architecture components

Idea: **Some modeling components lend themselves to inspection.**

For example, can we try to characterize each attention head of BERT?

Attention head 1 of layer 1.

This head performs this kind of behavior on most sentences.
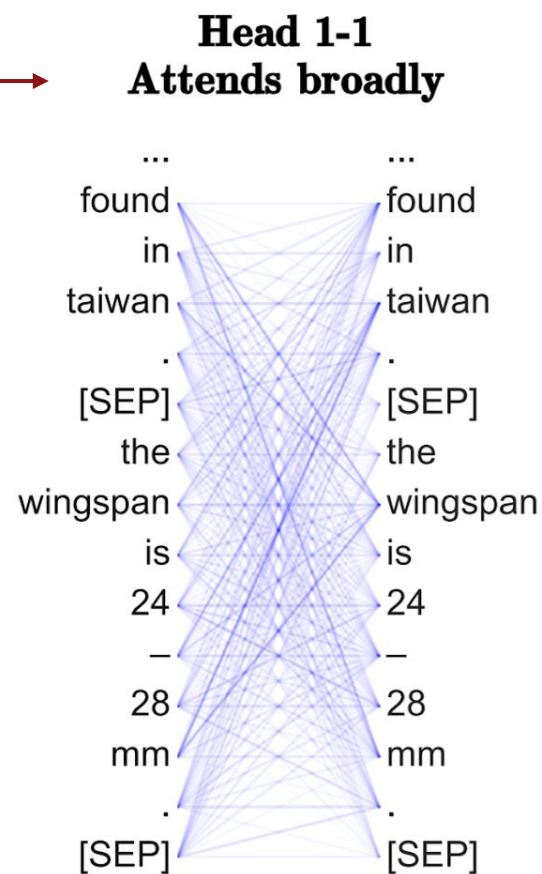
[Why is "interpretable" in quotes? It's hard to tell exactly how/whether the model is performing an interpretable function, especially deep in the network.]

Head 1-1
Attends broadly

...	...
found	found
in	in
taiwan	taiwan
.	.
[SEP]	[SEP]
the	the
wingspan	wingspan
is	is
24	24
–	–
28	28
mm	mm
.	.
[SEP]	[SEP]

[Clark et al., 2018]

Analysis of "interpretable" architecture components

Idea: **Some modeling components lend themselves to inspection.**

Some attention heads seem to perform simple operations.
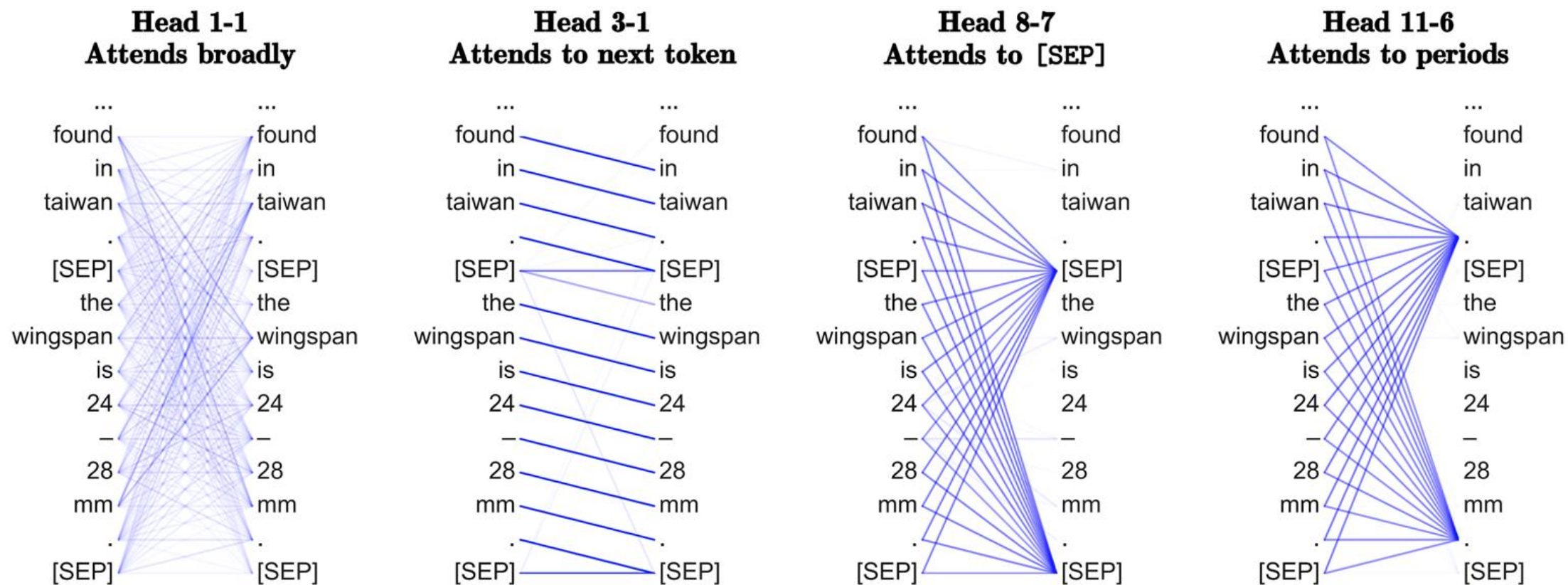
[Clark et al., 2018]

Analysis of "interpretable" architecture components
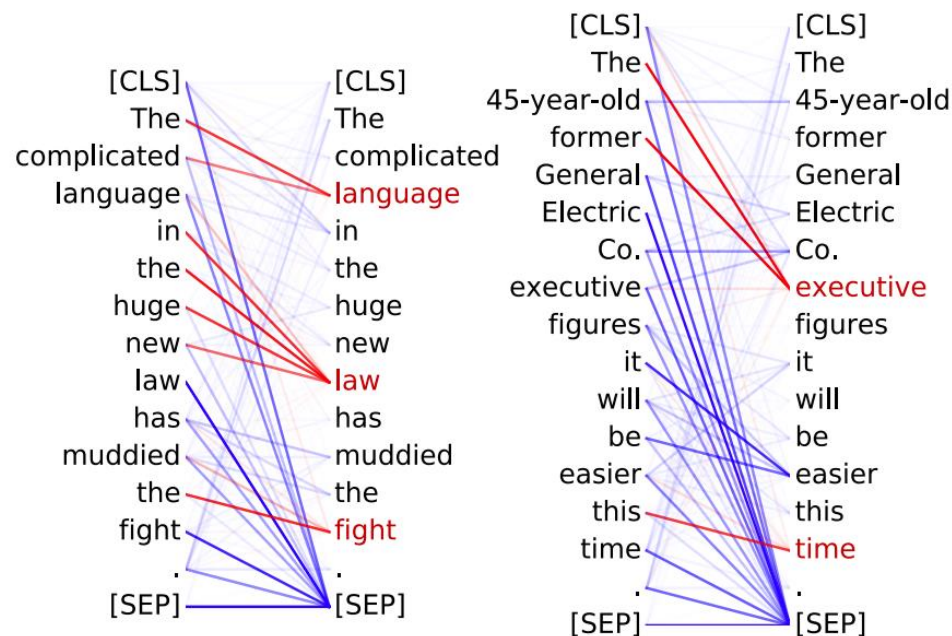
Idea: **Some modeling components lend themselves to inspection.**

Some heads are correlated with linguistic properties!

Approximate interpretation + quantitative analysis

Model behavior

[Clark et al., 2018]

Analysis of "interpretable" architecture components

Idea: **Some modeling components lend themselves to inspection.**

We saw coreference before; one head often matches coreferent mentions!

Approximate interpretation + quantitative analysis

Model behavior

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
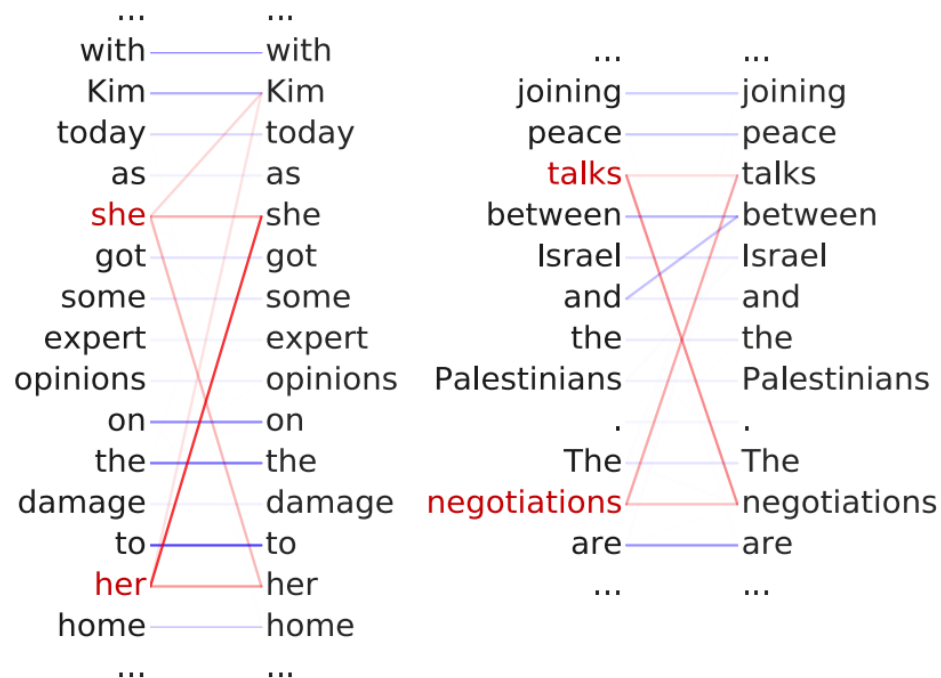
[Clark et al., 2018]

Analysis of "interpretable" architecture components

Idea: **Individual hidden units can lend themselves to an interpretable meaning.**

This model: a character-level LSTM language model.

Cell sensitive to position in line:

Here, "cell" refers to a single dimension of the cell state of the LSTM.

[Karpathy et al., 2016]

Analysis of "interpretable" architecture components

Idea: **Individual hidden units can lend themselves to an interpretable meaning.**

This model: a character-level LSTM language model.

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."

Here, "cell" refers to a single dimension of the cell state of the LSTM.

[Karpathy et al., 2016]

# Analysis of "interpretable" architecture components

Idea: Let's go back to **subject-verb agreement**. What's the mechanism by which LSTMs solve the task?

This model: a word-level LSTM language model.

(a) 2Adv

(b) nounPP

(c) subject relative
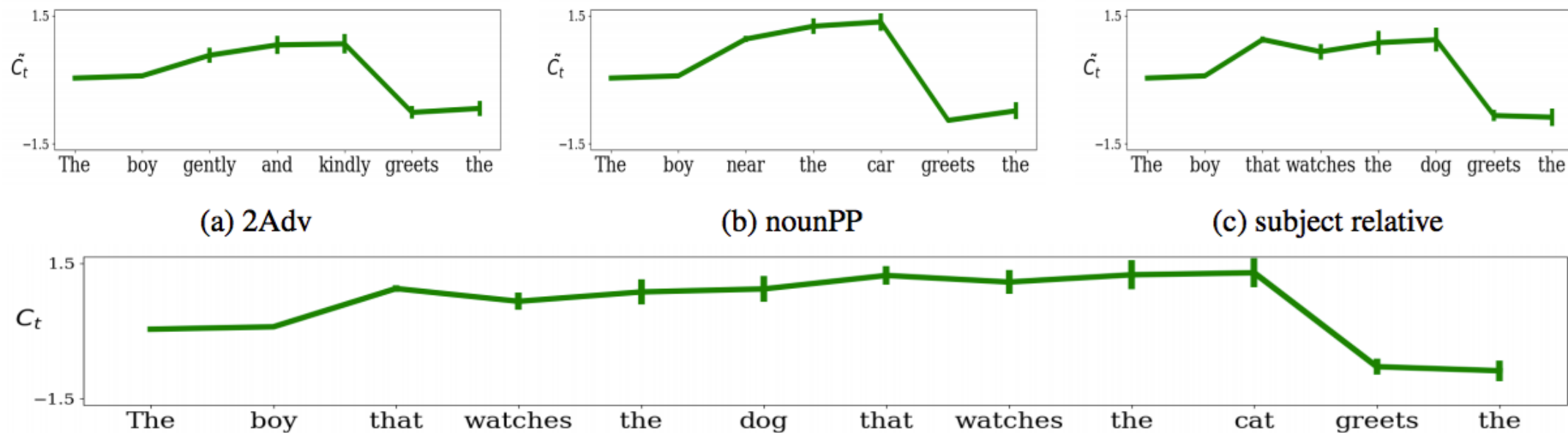
This is neuron 1150 in the LSTM, which seems to track the scope of the grammatical number of the subject! Removing this unit harms subject-verb agreement much more than removing a random unit.

[Lakretz et al., 2019]

Probing: supervised analysis of neural networks

SOTA Results

Pretrained Transformer (+finetuning)

Language Data
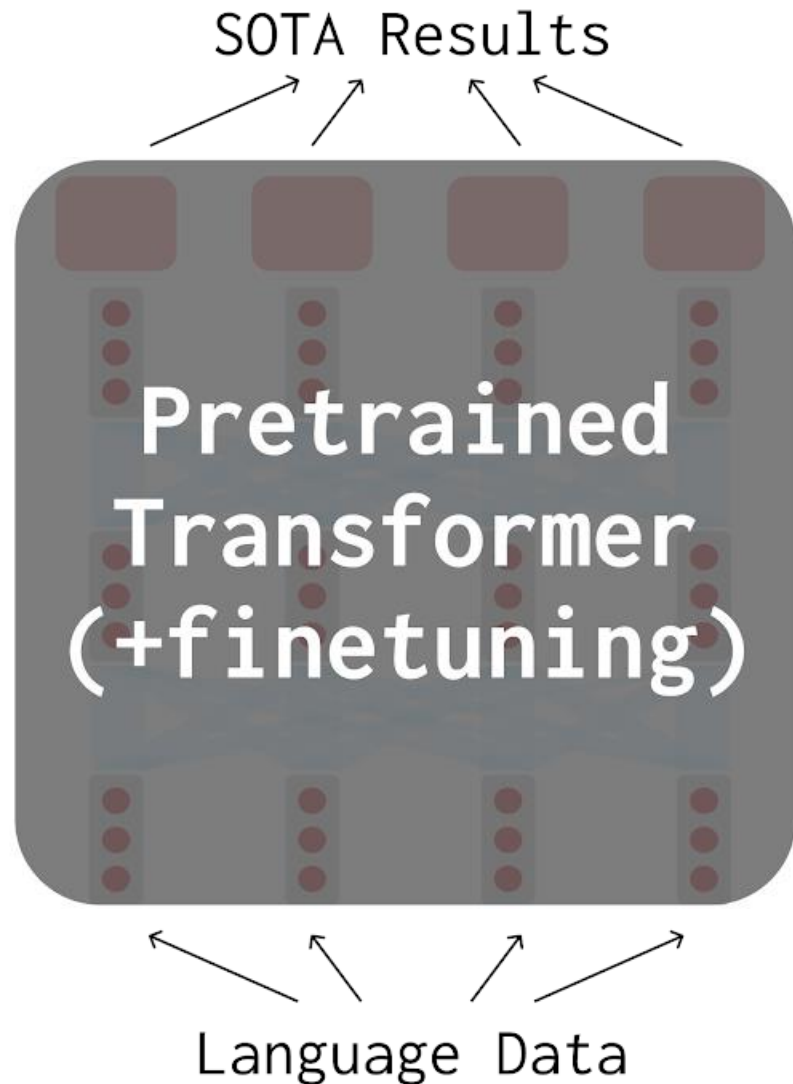
Premise:
Pretrained Transformers provide wildly **general-purpose language representations**
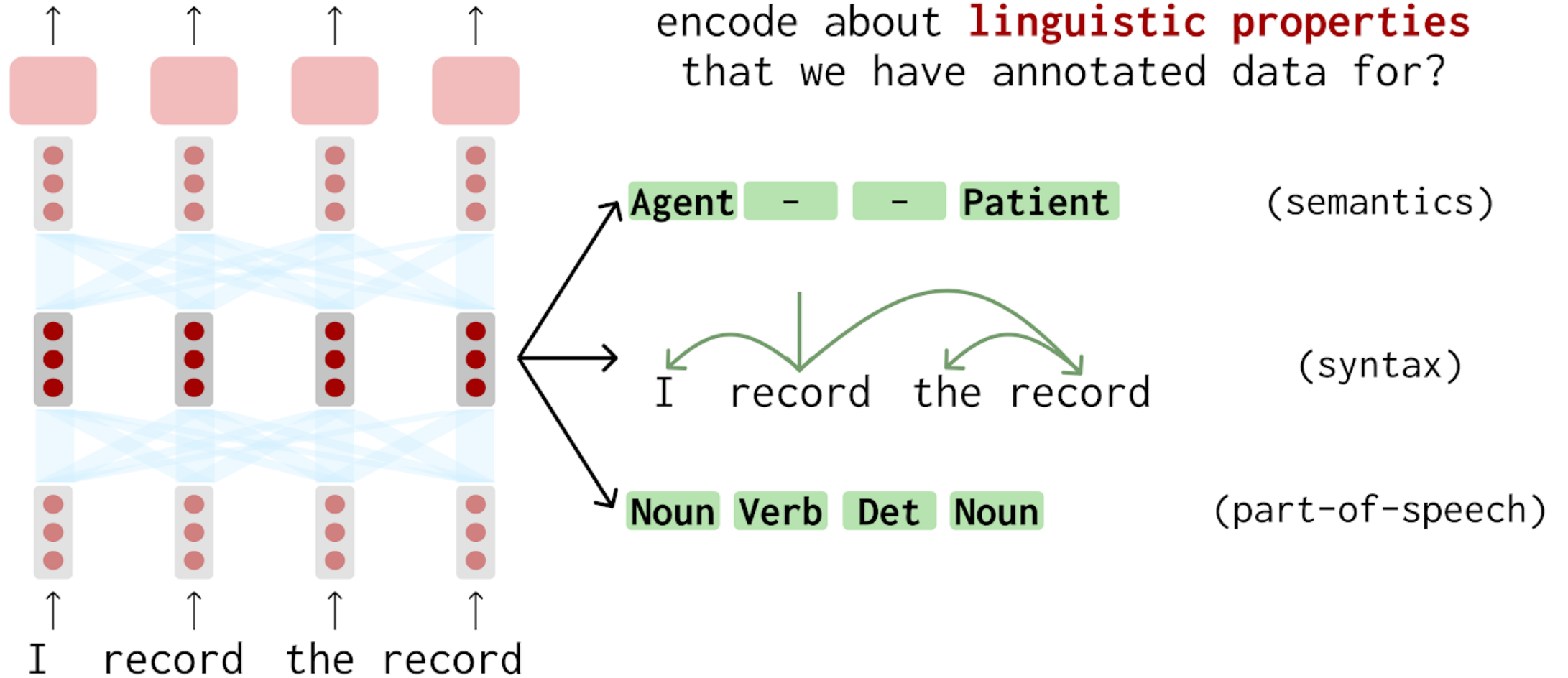
Question:
What do their representations **encode about language?**

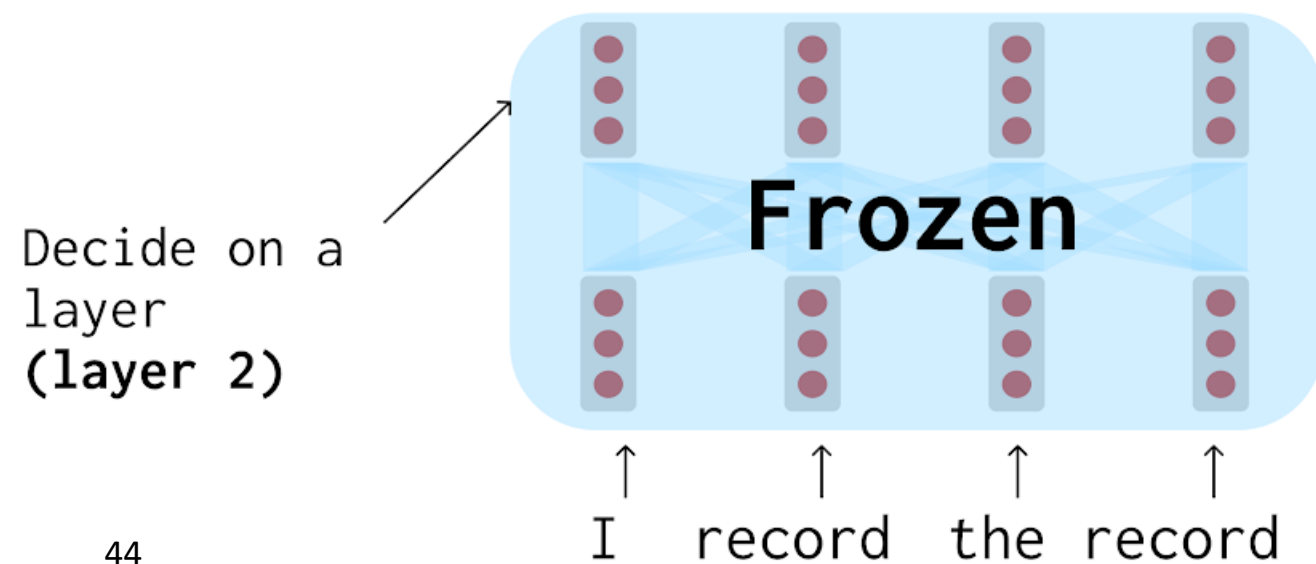[SOTA means "state-of-the-art," the best method for a given problem.]

Probing: supervised analysis of neural networks

Idea:
What do pretrained representations encode about **linguistic properties** that we have annotated data for?

| Agent | – | – | Patient | (semantics) |

I record the record (syntax)

| Noun | Verb | Det | Noun | (part-of-speech) |

43

Probing: supervised analysis of neural networks

Decide on a layer
(layer 2)

Frozen

I   record   the   record

Probing: supervised analysis of neural networks

Let's take a second to think more about probing.

- We have some property *y* (like part-of-speech)
- We have the model's word representations at a fixed layer: $h_1, \ldots, h_T$, where $h_i \in \mathbb{R}^d$, where the words are at indices $1, \ldots, T$.
- We have a function family F like the set of *linear models* or *1-layer feed-forward networks (with fixed hyperparmaters.)*
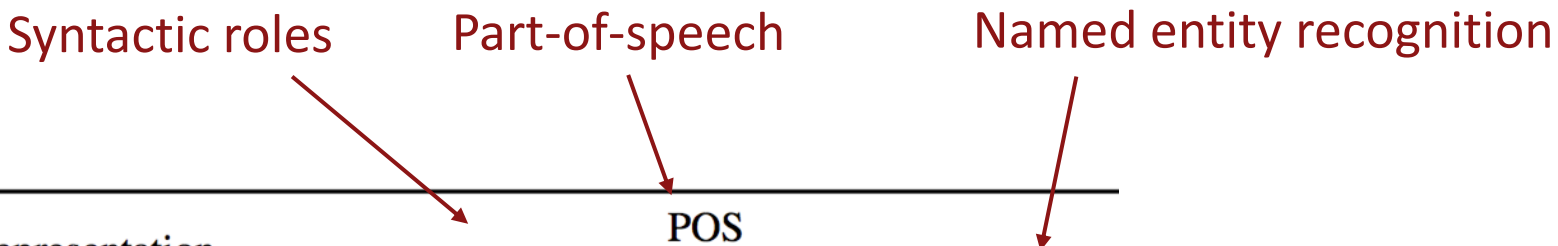- We **freeze** the parameters of the model, so it's not finetuned. Then, we train our probe, a function

$$\hat{y} \sim f(h_i) \qquad f \in F$$

The extent to which we can predict y from h_i is a measure of the accessibility of that feature in the representation.

- This helps gain a rough understanding into how the model processes its inputs.
- Also may help in the search for causal mechanisms.

Probing: supervised analysis of neural networks

BERT (and other pretrained LMs) make some linguistic properties predictable to very high accuracy with a simple linear probe.

Syntactic roles   Part-of-speech   Named entity recognition

Pretrained Representation			POS			
	Avg.	CCG	PTB	EWT	Chunk	NER
BERT (base, cased) best layer	84.09	93.67	96.95	95.21	92.64	82.71
BERT (large, cased) best layer	**85.07**	**94.28**	96.73	**95.80**	**93.64**	**84.44**
GloVe (840B.300d)	59.94	71.58	90.49	83.93	62.28	53.22
Previous state of the art (without pretraining)	83.44	94.7	97.96	95.82	95.77	91.38

Layerwise trends of probing accuracy

- Across a wide range of linguistic properties, the middle layers of BERT yield the best probing accuracies.

Input words here

MLM objective here

Consistently best a bit past the mid point

(f) BERT (large, cased)

Layer 0

Layer 24

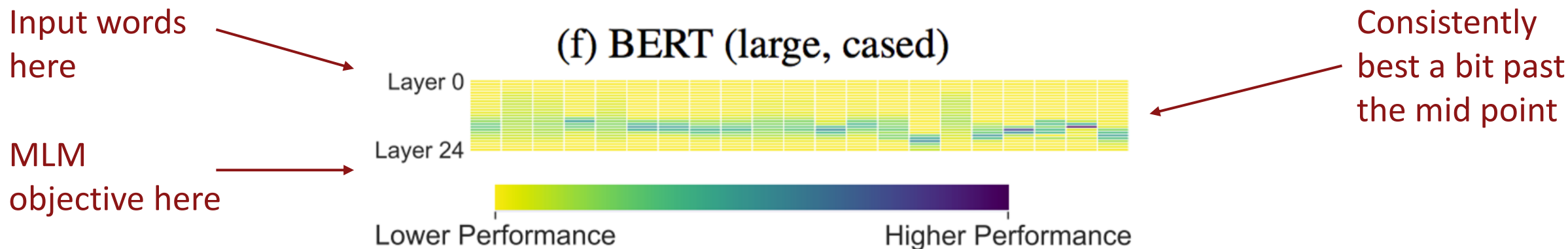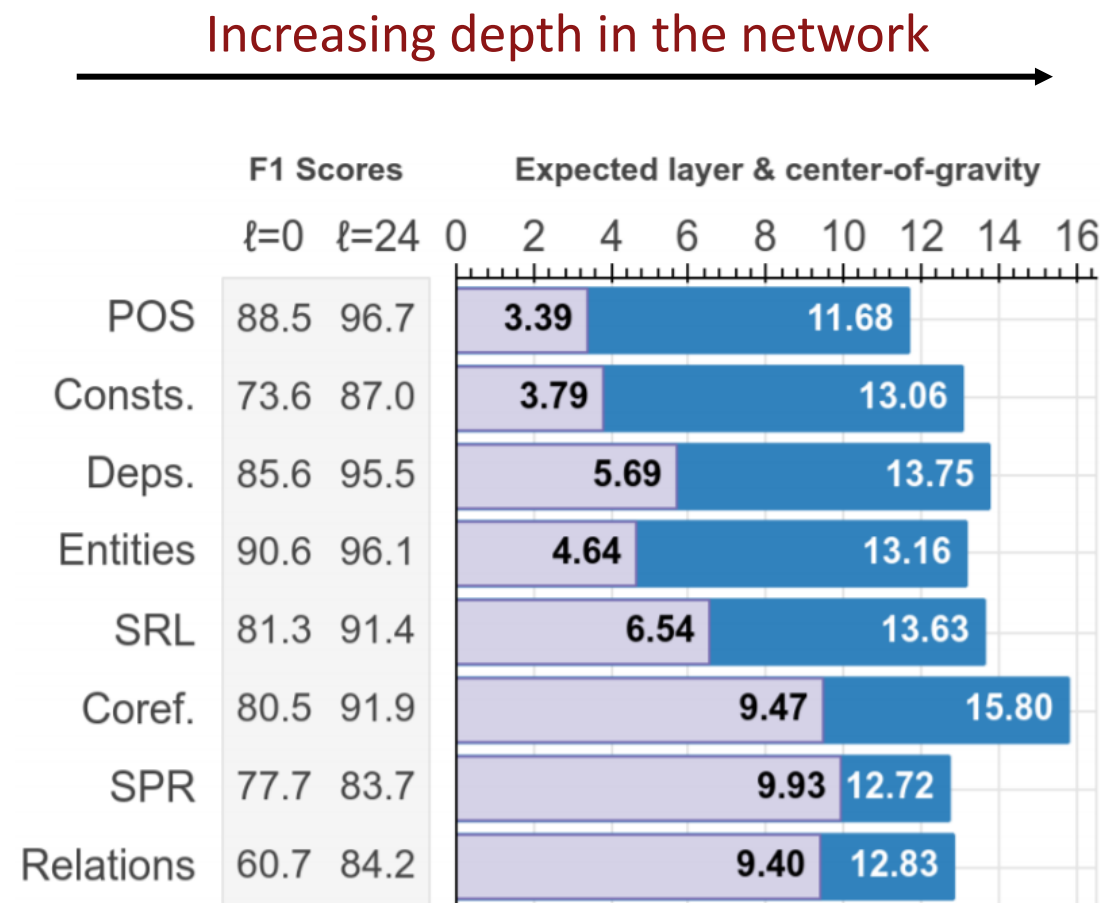Lower Performance   Higher Performance

Figure 3: A visualization of layerwise patterns in task performance. Each column represents a probing task, and each row represents a contextualizer layer.

[Liu et al., 2019]

Layerwise trends of probing accuracy

- Increasingly abstract linguistic properties are more accessible later in the network.

Increasing abstractness of linguistic properties

Increasing depth in the network

	F1 Scores		Expected layer & center-of-gravity
	$\ell=0$	$\ell=24$	
POS	88.5	96.7	3.39 — 11.68
Consts.	73.6	87.0	3.79 — 13.06
Deps.	85.6	95.5	5.69 — 13.75
Entities	90.6	96.1	4.64 — 13.16
SRL	81.3	91.4	6.54 — 13.63
Coref.	80.5	91.9	9.47 — 15.80
SPR	77.7	83.7	9.93 — 12.72
Relations	60.7	84.2	9.40 — 12.83

[Tenney et al., 2019]

Emergent simple structure in neural networks

- Recall word2vec, and the intuitions we built around its vectors

cat   kitty

guitar
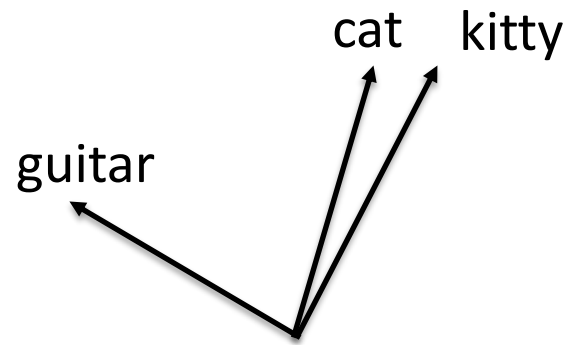
We interpret **cosine similarity** as *semantic similarity.*

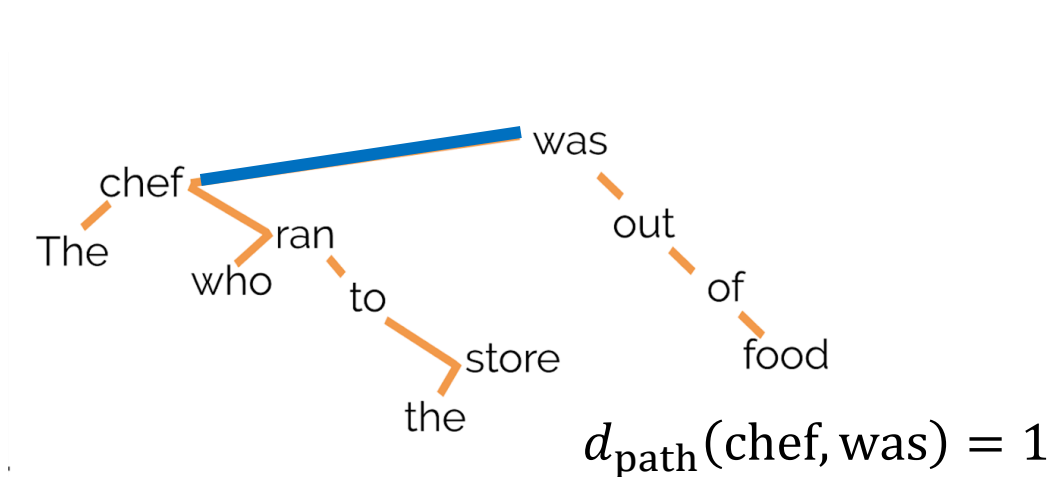Sacramento   Harrisburg

California

Pennsylvania
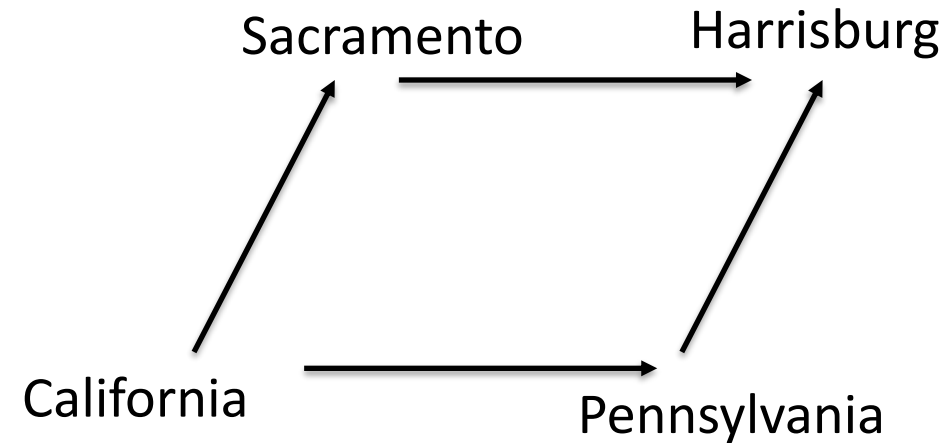
Some *relationships* are encoded as **linear offsets**

- It's hard to the dimensions of word2vec vectors, but it's fascinating that interpretable concepts approximately map onto simple functions of the vectors
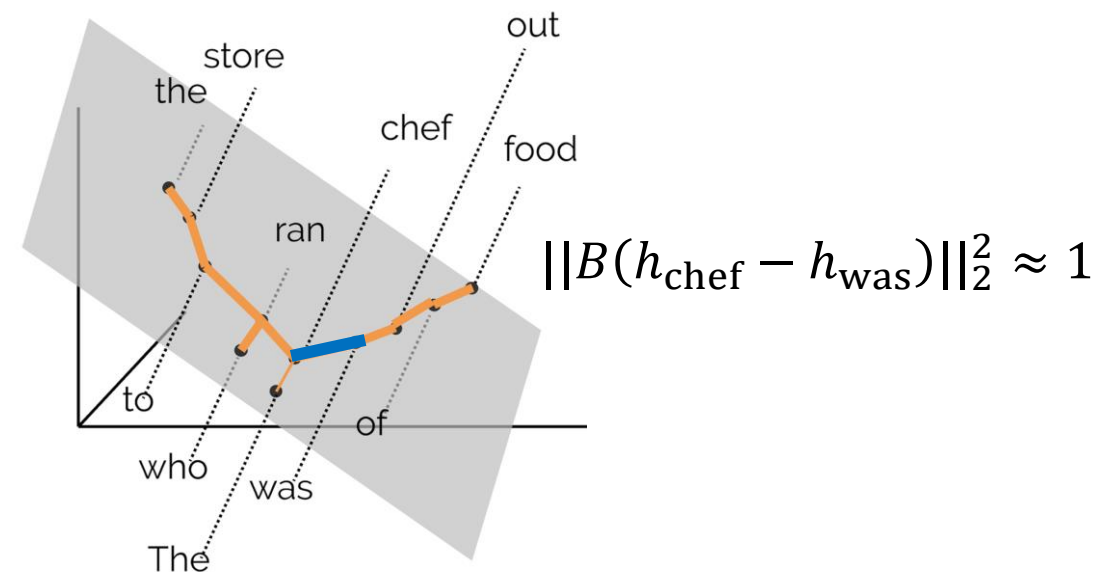
[Mikolov et al., 2013]

Probing: trees simply recoverable from BERT representations

- Recall dependency parse trees. They describe underlying syntactic structure in sentences.
- Hewitt and Manning 2019 show that BERT models make dependency parse **tree structure** easily accessible.

$$d_{\mathrm{path}}(\mathrm{chef}, \mathrm{was}) = 1$$

$$\|B(h_{\mathrm{chef}} - h_{\mathrm{was}})\|_2^2 \approx 1$$

$$d_{\mathrm{path}}(w_1, w_2)$$

$$\|B(h_{w_1} - h_{w_2})\|_2^2$$

Tree path distance: the number of edges in the path between the words

Squared Euclidean distance of BERT vectors after transformation by the (probe) matrix B.
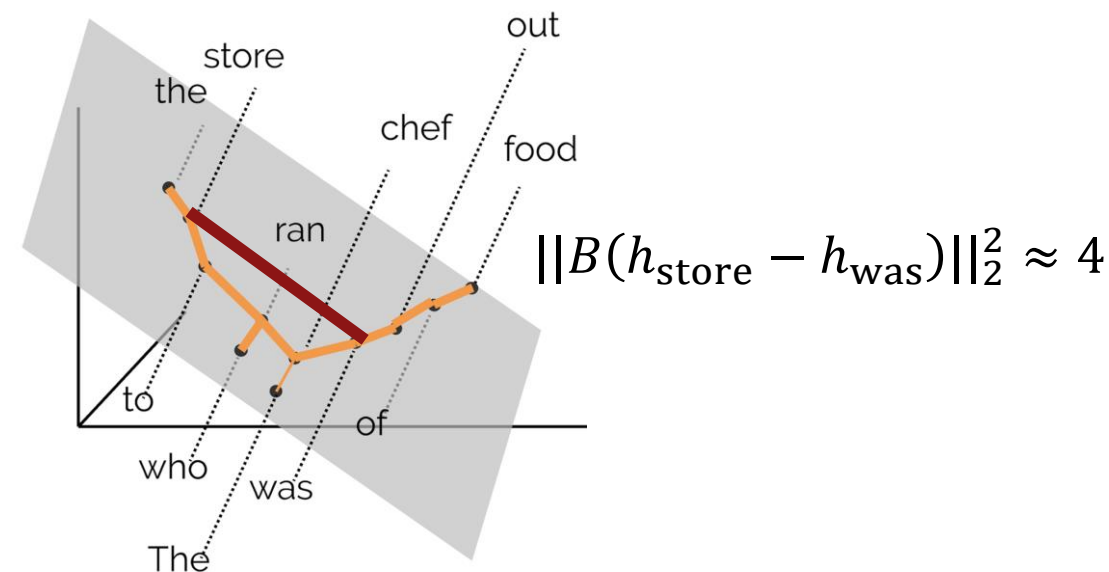
[Hewitt and Manning, 2019]

Probing: trees simply recoverable from BERT representations

- Recall dependency parse trees. They describe underlying syntactic structure in sentences.
- Hewitt and Manning 2019 show that BERT models make dependency parse **tree structure** easily accessible.

$$d_{\mathrm{path}}(\mathrm{store}, \mathrm{was}) = 4$$

$$||B(h_{\mathrm{store}} - h_{\mathrm{was}})||_2^2 \approx 4$$

$$d_{\mathrm{path}}(w_1, w_2)$$

$$||B(h_{w_1} - h_{w_2})||_2^2$$

Tree path distance: the number of edges in the path between the words

Squared Euclidean distance of BERT vectors after transformation by the (probe) matrix B.

[Hewitt and Manning, 2019]

Probing: helping design hypotheses for causal analysis

- The structural probe was used to design and test hypotheses as to how LMs *incrementally* parse sentences (and causally intervene on network behavior)!

Even though the band left the party I stayed

Even though the band left the party went on

[Eisape et al., 2022]
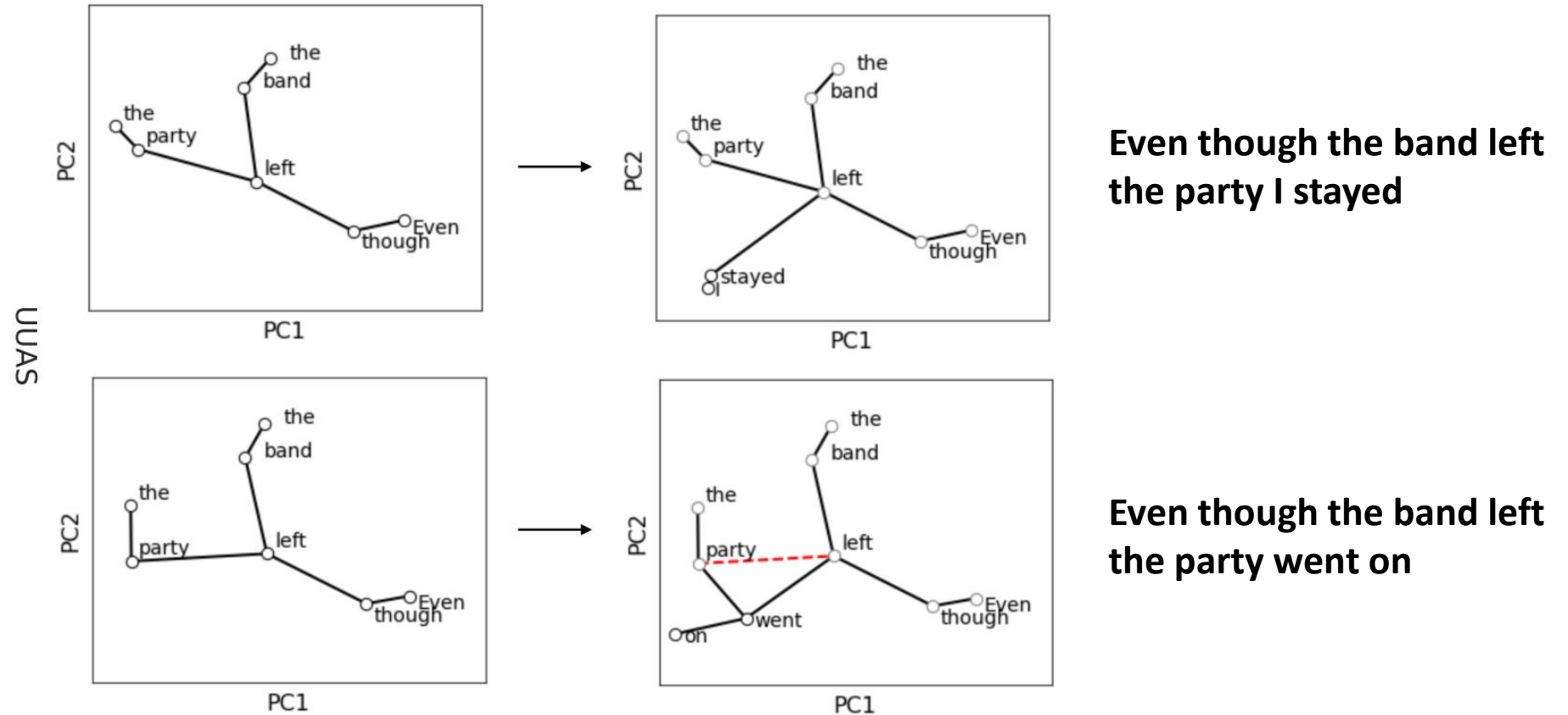
# Final thoughts on probing and correlation studies

- Probing shows that properties are accessible to your probe family, not that they're used by the neural model you're studying.

- Correlation studies (like attention maps) likewise.

- For example:

  - Hewitt and Liang, 2019 show that under certain conditions, probes can achieve high accuracy on random labels.

  - Ravichander et al., 2021 show that probes can achieve high accuracy on a property even when the model is trained to know the property isn't useful.

- Some efforts (Vig et al., 2020, Eisape et al., 2022) have gone towards causal studies. Interesting and harder!

[Hewitt and Liang, 2019, Ravichander et al., 2021, Vig et al., 2020]

Outline

1. Motivating model analysis and explanation
2. One model at multiple levels of abstraction
3. Out-of-domain evaluation sets
   1. Testing for linguistic knowledge
   2. Testing for task heuristics
4. Influence studies and adversarial examples
   1. What part of my input led to this answer?
   2. How could I minimally modify this input to change the answer?
5. Analyzing representations
   1. Correlation in "interpretable" model components
   2. Probing studies: supervised analysis
6. Revisiting model ablations as analysis

# Recasting model tweaks and ablations as analysis

- Consider the usual neural network improvement process:
  - You have a network, which works okay.
  - You see whether you can tweak it in simple ways to improve it.
  - You see whether you can remove any complex things and have it still work as well.
- This can be thought of as a kind of model analysis!
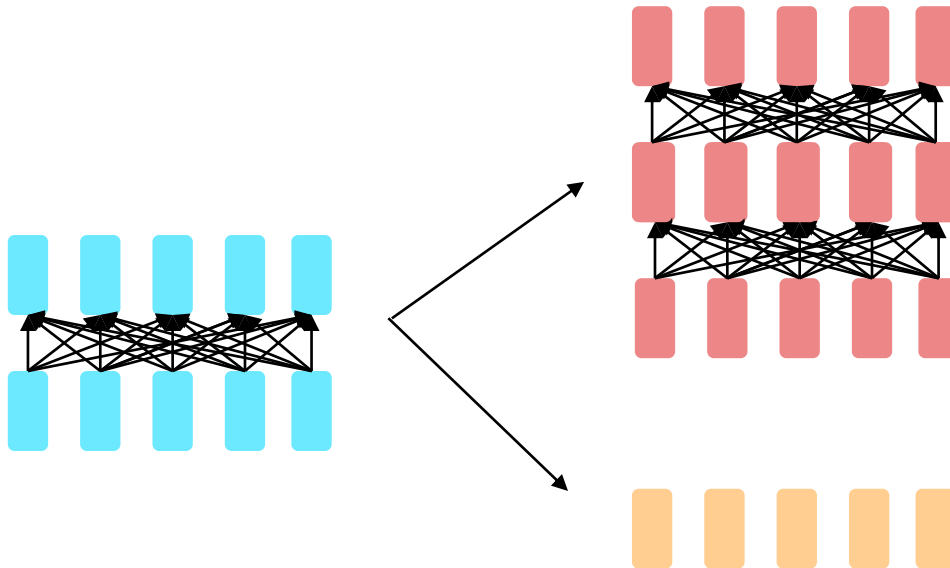
[Would it be better for this part of my model to be deeper? Or can I get away with making it shallower?]

Ablation analysis: do we need all these attention heads?

- Michel et al., 2019 train transformers with multi-headed attention on machine translation and natural language inference.

- After training, they find many attention heads can be **removed** with no drop in accuracy!
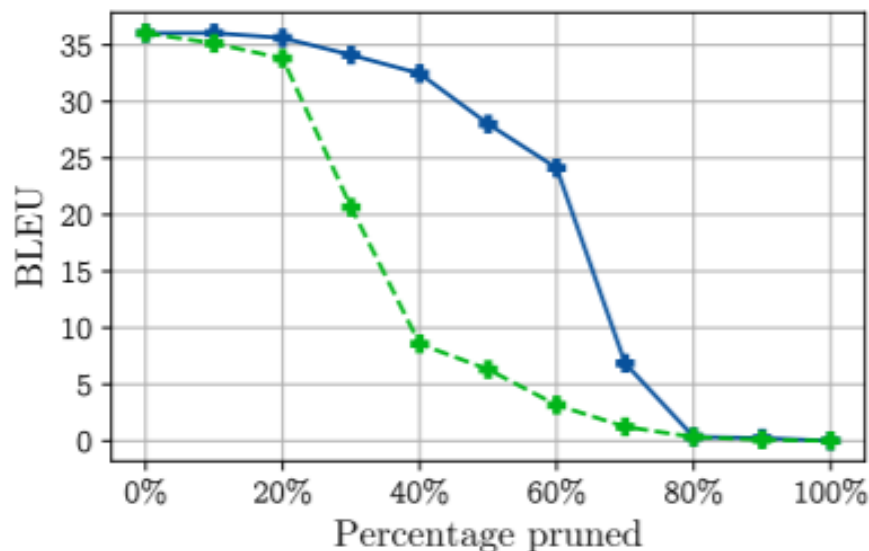
(a) Evolution of BLEU score on `newstest2013` when heads are pruned from WMT.
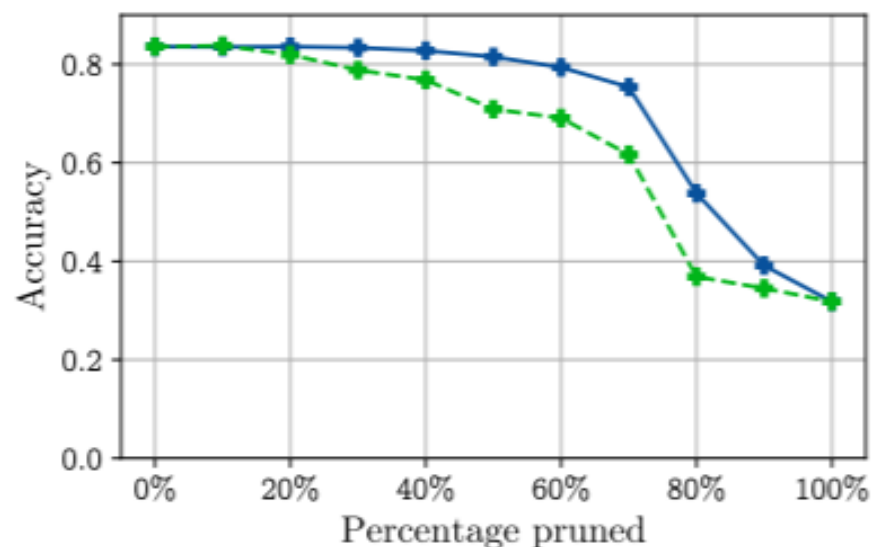
(b) Evolution of accuracy on the MultiNLI-matched validation set when heads are pruned from BERT.

[Green and blue lines indicate two different ways to choose the order to prune attention heads.]

[Michel et al., 2019]

What's the right layer order for a Transformer?

- We saw that Transformer models are sequences of layers
  - Self-attention → Feed-forward → Self-attention → Feed-forward → ….
  - (Layer norm and residual connections omitted)
- Press et al., 2019 asked, why? Is there a better ordering of self-attention and feed-forward layers?
- Here's that sequence of layers again:

sfsfsfsfsfsfsfsfsfsfsfsfsfsfsf

Achieves 18.40 perplexity on a language modeling benchmark

ssssssfsfsfsfsfsfsfsfsfffffff

Achieves 17.96 perplexity on a language modeling benchmark

Many self-attention layers first

Many feed-forward layers last

[Press et al., 2019]

Parting thoughts

- Neural models are complex, and difficult to characterize. A single accuracy metric doesn't cut it.

- We struggle to find intuitive descriptions of model behaviors, but we have a many tools at many levels of abstraction to give insight.

- Engage critically when someone claims a (neural) NLP model is interpretable – in what ways is it interpretable? In what ways is it still opaque?

- Bring this analysis and explanation way of thinking with you to your model building efforts even if analysis isn't your main goal.

Good luck on finishing your final projects! We're really appreciative of your efforts.