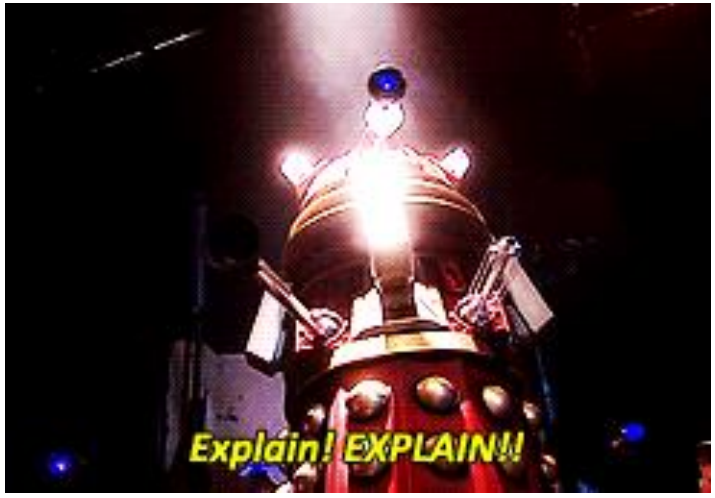
Interpretability in Machine Learning



Why Interpret ?

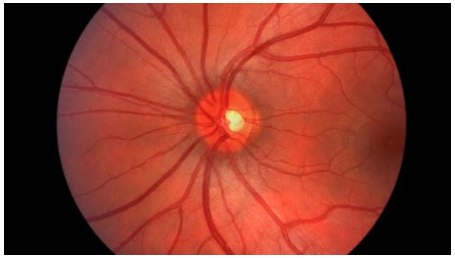
The current state of machine learning

And its uses ...

https://www.tesla.com/videos/autopilot-self-driving-hardware-neighborhood-long

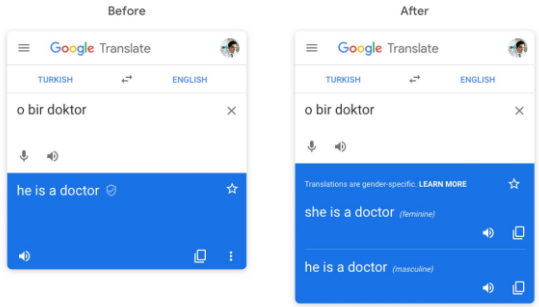
NYPost

MIT Technology Review



DeepMind

DeepMind



So are we in the golden age of AI ?

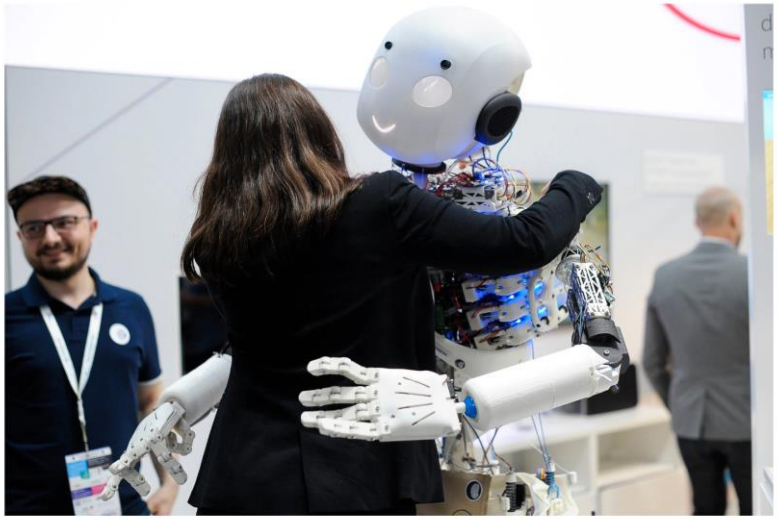
Safety and well being

Tesla hit parked police car 'while using Autopilot'

30 May 2018 Share

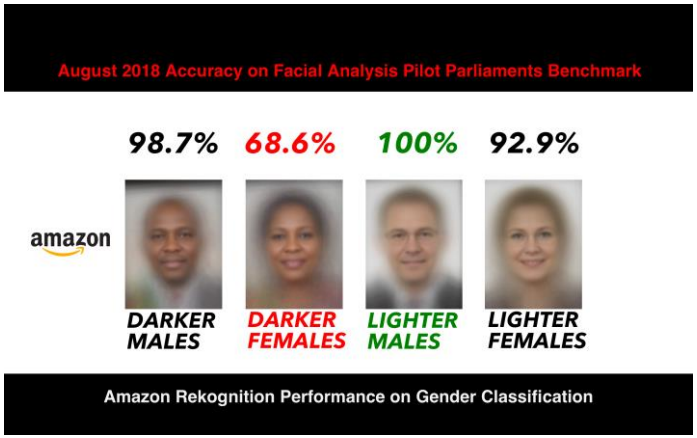
A number of Tesla vehicles have been involved in crashes.

Warnings of a Dark Side to A.I. in Health Care

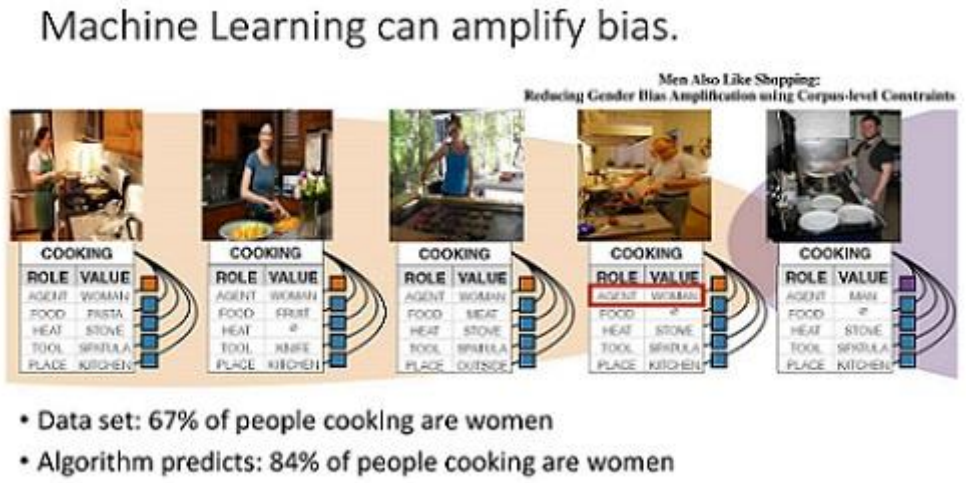


Scientists worry that with just tiny tweaks to data, neural networks can be fooled into committing "adversarial attacks" that mislead rather than help. Joan Cros/NurPhoto, via Getty Images

Bias in algorithms

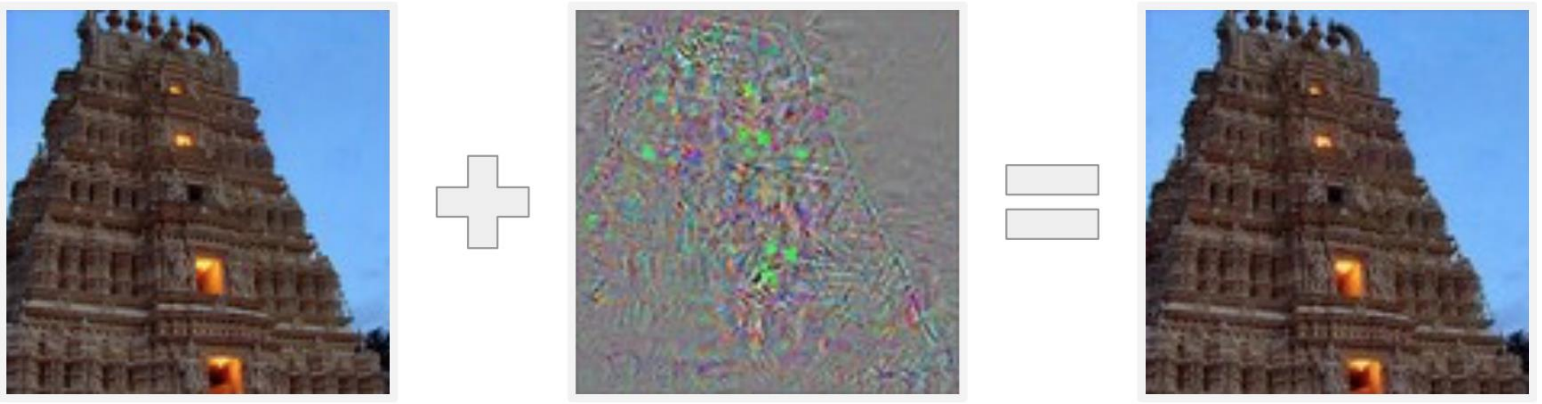


https://medium.com/@Joy.Buolamwini/response-racial-and-gender-bias-in-amazon-rekognition-commercial-ai-system-for-analyzing-faces-a289222eeced



https://www.infoq.com/presentations/unconscious-bias-machine-learning/

Adversarial Examples



Original image

Temple (97%)

Perturbations

Adversarial example

Ostrich (98%)

Legal Issues - GDPR

And more ...

- Interactive feedback - can model learn from human actions in online setting ? (Can you tell a model to not repeat a specific mistake ?)

- Recourse – Can a model tell us what actions we can take to change its output ? (For example, what can you do to improve your credit score?)

In general, it seems like there are few fundamental problems –

- We don’t trust the models
- We don’t know what happens in extreme cases
- Mistakes can be expensive / harmful
- Does the model makes similar mistakes as humans ?
- How to change model when things go wrong ?

Interpretability is one way we try to deal with these problems

What is interpretability ?

There is no standard definition –

Most agree it is something different from performance.

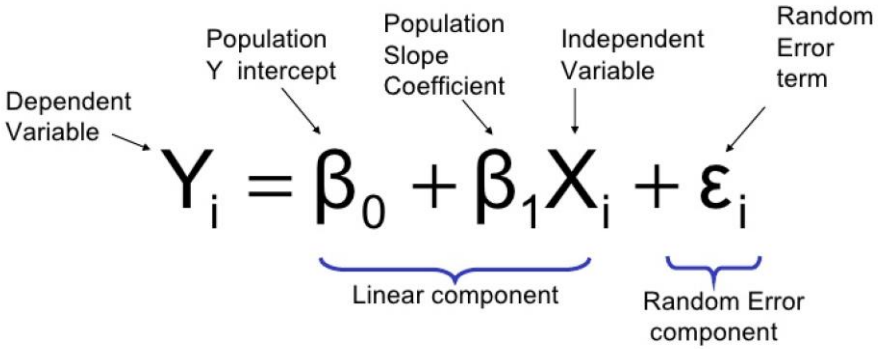
Ability to explain or to present a model in understandable terms to humans (Doshi-Velez 2017)

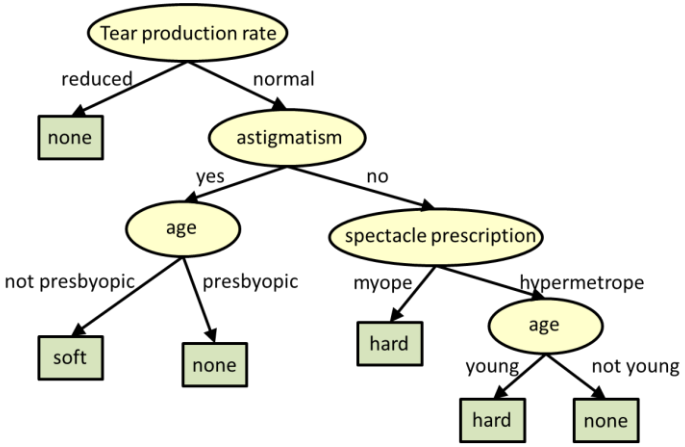
Cynical view – It is what makes you feel good about the model.

It really depends on target audience.

What does interpretation looks like ?

- In pre-deep learning models, some models are considered “interpretable”





What does interpretation look like ?

- Heatmap Visualization

Figure 3. **Attribution for Diabetic Retinopathy grade prediction from a retinal fundus image.** The original image is show on the left, and the attributions (overlayed on the original image in gray scaee) is shown on the right. On the original image we annotate lesions visible to a human, and confirm that the attributions indeed point to them.

[Sundarajan 2017]

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

in a clinical trial mainly involving patients over qqq with coronary heart disease , ramipril reduced mortality while vitamin e had no preventive effect .

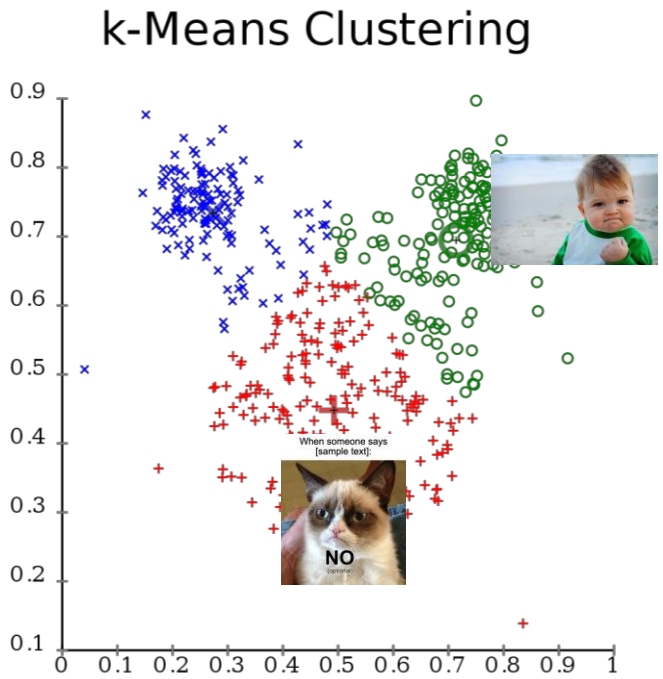
Table 2: Gate activations for each aspect in a PICC abstract. Note that because gates are calculated a the final convolution layer, activations are not ir exact 1-1 correspondence with words.

[Jain 2018]

What does interpretation looks like ?

- Give prototypical examples

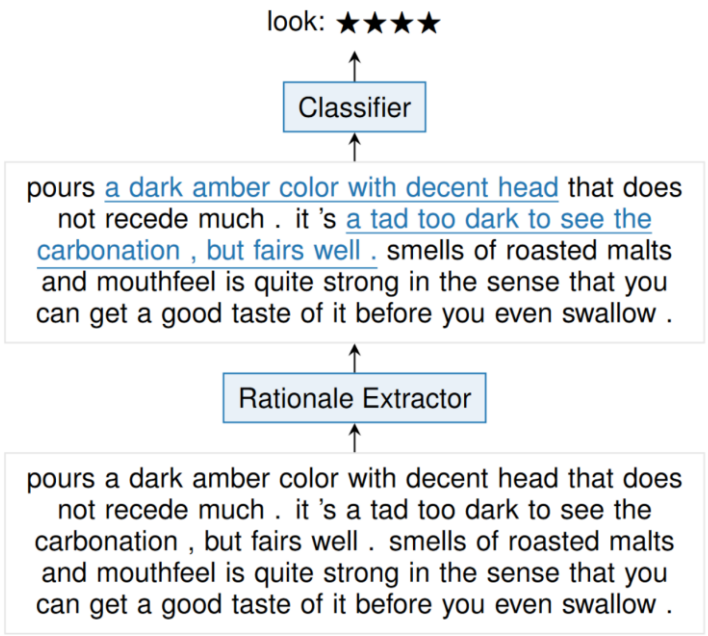
[Kim 2016]



By Chire - Own work, Public Domain, https://commons.wikimedia.org/w/index.php?curid=11765684

What does interpretation look like ?

- Bake it into the model



[Bastings et al 2019]

What does interpretation looks like ?

- Provide explanation as text

Question:	While eating a hamburger with friends, what are people trying to do?
Choices:	**have fun**, tasty, or indigestion
CoS-E:	Usually a hamburger with friends indicates a good time.
Question:	After getting drunk people couldn't understand him,it was because of his what?
Choices:	lower standards,**slurred speech**, or falling down
CoS-E:	People who are drunk have difficulty speaking.
Question:	People do what during their time off from work?
Choices:	**take trips**, brow shorter, or become hysterical
CoS-E:	People usually do something relaxing, such as taking trips,when they don't need to work.

Table 1: Examples from our CoS-E dataset.

[Rajani et al 2019]

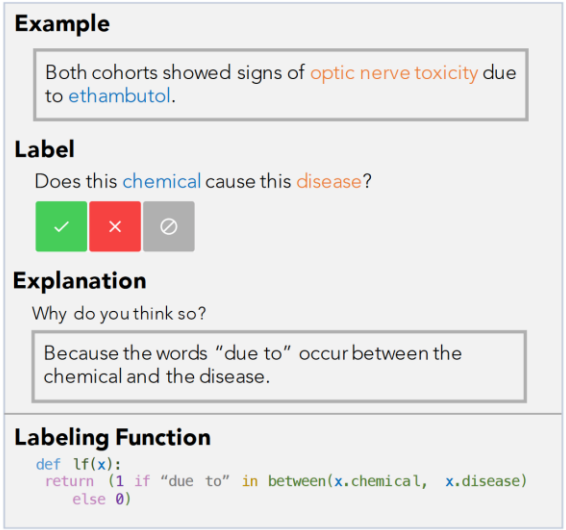


Figure 1: In BabbleLabble, the user provides a natural language explanation for each labeling decision. These explanations are parsed into labeling functions that convert unlabeled data into a large labeled dataset for training a classifier.

[Hancock et al 2018]

Some properties of Interpretations

- **Faithfulness** - how to provide explanations that accurately represent the true reasoning behind the model's final decision.
- **Plausibility** – Is the explanation correct or something we can believe is true, given our current knowledge of the problem ?
- **Understandable** – Can I put it in terms that end user without in-depth knowledge of the system can understand ?
- **Stability** – Does similar instances have similar interpretations ?

Evaluating Interpretability [Doshi-Velez 2017]

- Application level evaluation – Put the model in practice and have the end users interact with explanations to see if they are useful .
- Human evaluation – Set up a Mechanical Turk task and ask non-experts to judge the explanations
- Functional evaluation – Design metrics that directly test properties of your explanation.

How to “interpret” ? Some definitions

Global vs Local

- **Do we explain individual prediction ?**

Example –

Heatmaps
Rationales

- **Do we explain entire model ?**

Example –

Prototypes
Linear Regression
Decision Trees

Inherent vs Post-hoc

- **Is the explainability built into the model ?**

Example –

Rationales
Linear Regression
Decision Trees
Natural Language Explanations

- **Is the model black-box and we use external method to try to understand it ?**

Example –

Heatmaps (Some forms)
Prototypes

Model based vs Model Agnostic

- **Can it explain only few classes of models ?**

Example –

Rationales
LR / Decision Trees
Attention
Gradients (Differentiable Models only)

- **Can it explain any model ?**

Example –

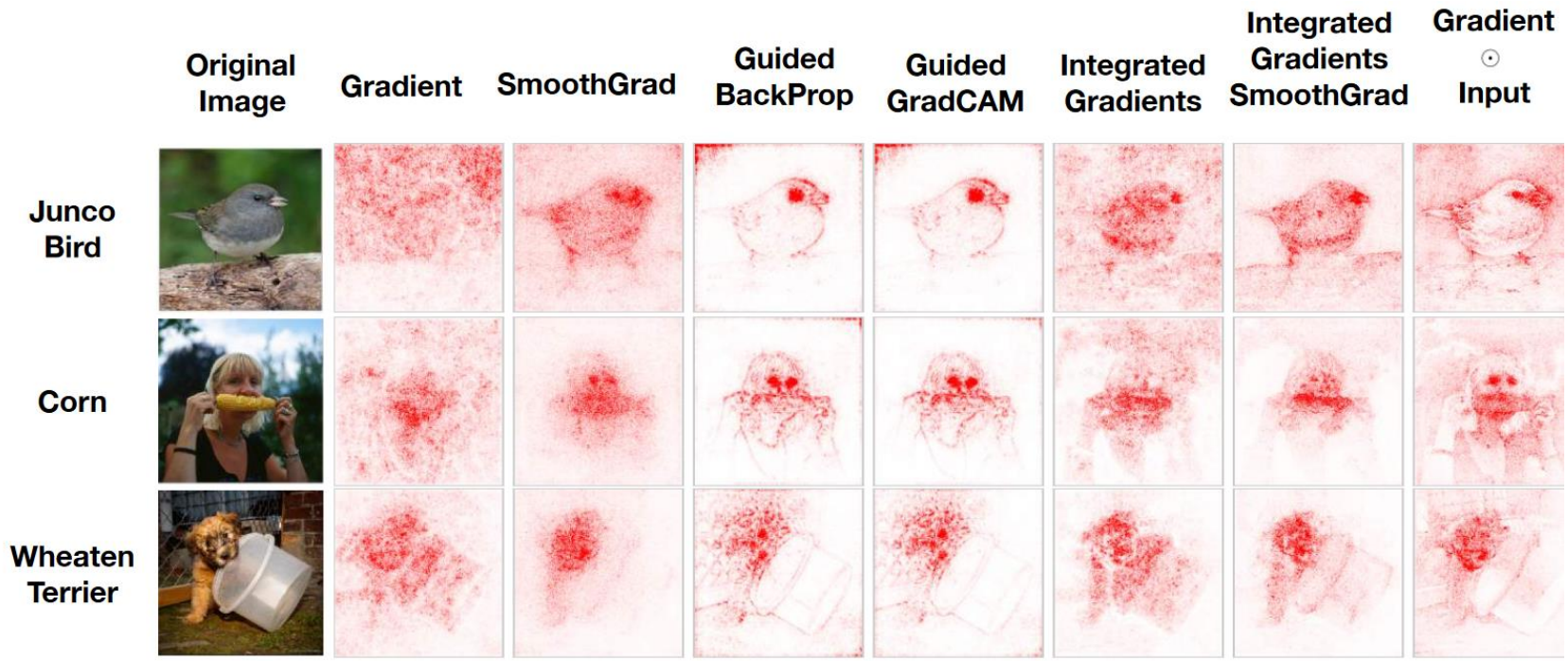
LIME – Locally Interpretable Model Agnostic Explanations

SHAP – Shapley Values

Some
Locally Interpretable,
Post-hoc
methods

Saliency Based Methods

- Heatmap based visualization
- Need differentiable model in most cases
- Normally involve gradient



[Adebayo et al 2018]

Saliency Example - Gradients

$$f(x): R^d \rightarrow R$$

$$E(f)(x) = \frac{df(x)}{dx}$$

How do we take gradient with respect to words ?

Take gradient with respect to embedding of the word .

Saliency Example – Leave-one-out

$$f(x): R^d \rightarrow R$$

$$E(f)(x)_i = f(x) - f(x\backslash \mathrm{i})$$

How to remove ?

1. Zero out pixels in image
2. Remove word from the text
3. Replace the value with population mean in tabular data

Problems with Saliency Maps

- Only capture first order information
- Strange things can happen to heatmaps in second order.

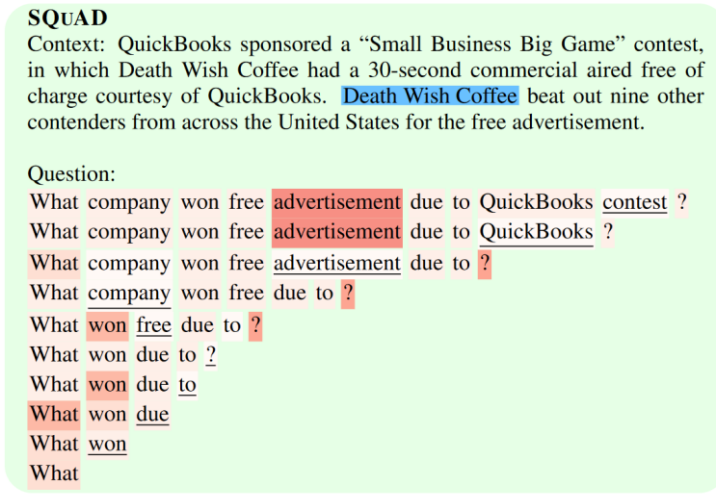


Figure 6: Heatmap generated with leave-one-out shifts drastically despite only removing the least important word (underlined) at each step. For instance, "advertisement", is the most important word in step two but becomes the least important in step three.

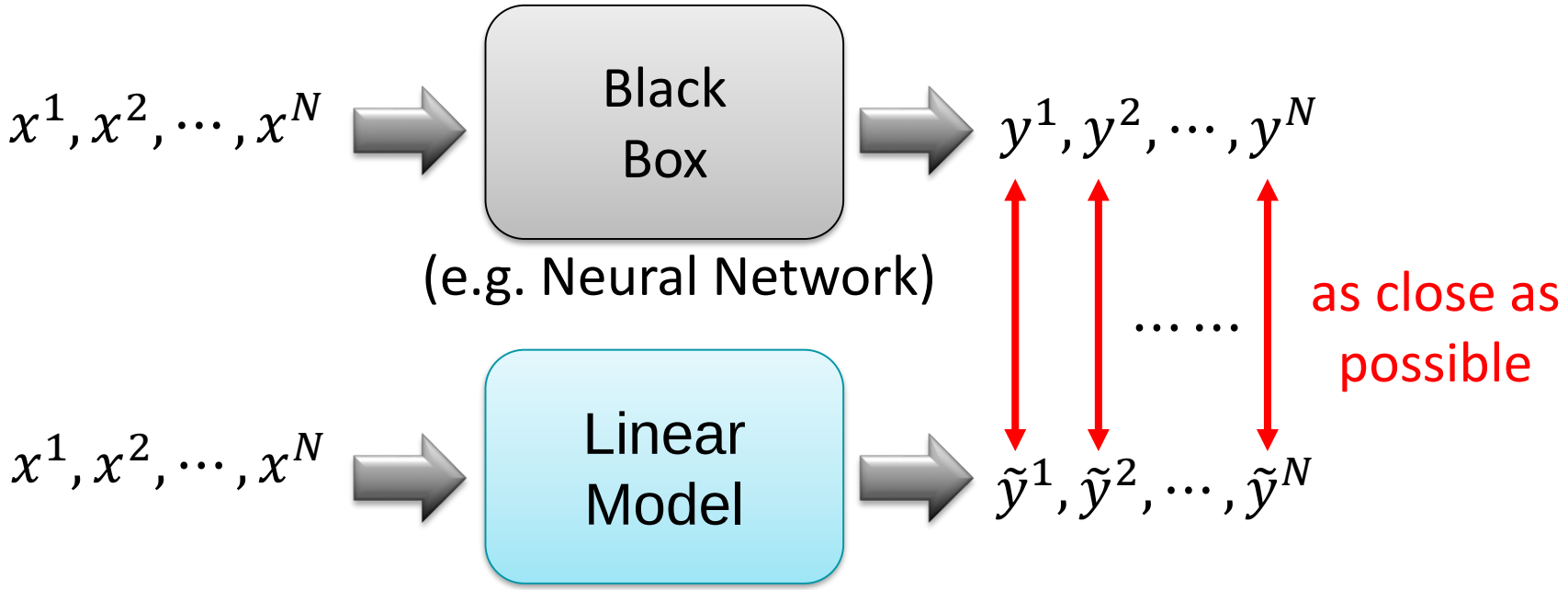
[Feng et al 2018]

Sanity check:
When prediction changes, do explanations change?

(Slide Credit – Julius Adebayo)

(Image Credit – Hung-yi Lee)

LIME – locally interpretable model agnostic



Can't do it globally of course, but locally ? Main Idea behind LIME

Intuition behind LIME

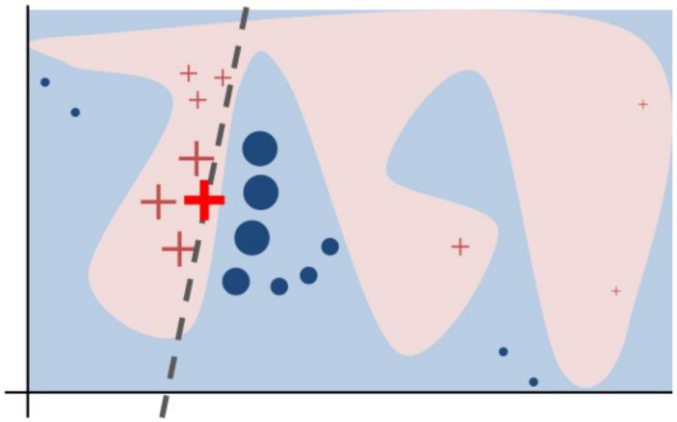
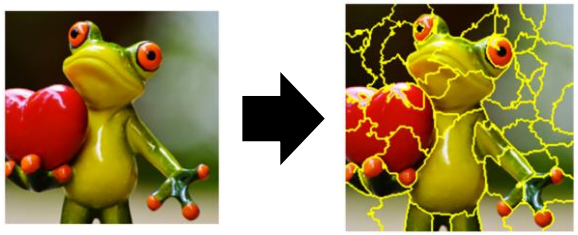


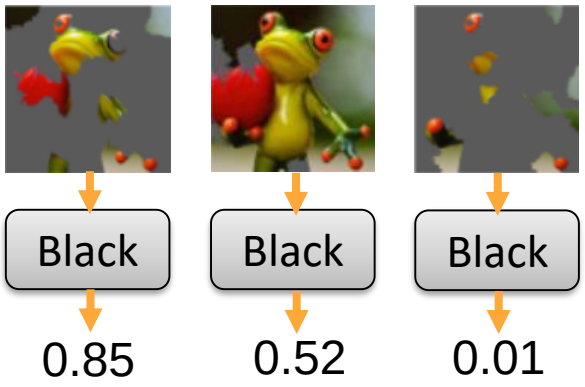
Figure 3: Toy example to present intuition for LIME. The black-box model's complex decision function f (unknown to LIME) is represented by the blue/pink background, which cannot be approximated well by a linear model. The bold red cross is the instance being explained. LIME samples instances, gets predictions using f, and weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

[Ribeiro et al 2016]

LIME – Image



- 1. Given a data point you want to explain
- 2. Sample at the nearby - Each image is represented as a set of superpixels (segments).



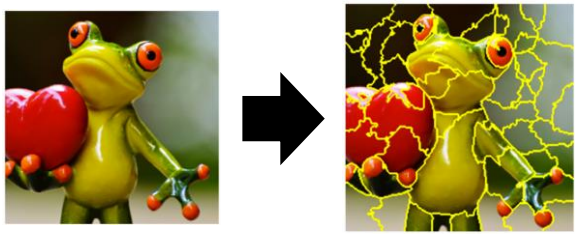
Randomly delete some segments.

Compute the probability of “frog” by black box

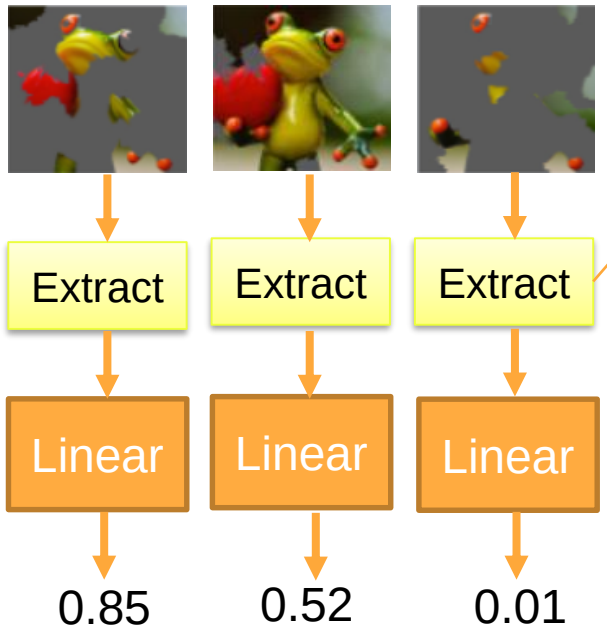
Ref: https://medium.com/@kstseng/lime-local-interpretable-model-agnostic-explanation%E6%8A%80%E8%A1%93%E4%BB%8B%E7%B4%B9-a67b6c34c3f8

(Slide Credit – Hung-yi Lee)

LIME – Image



- 3. Fit with linear (or interpretable) model



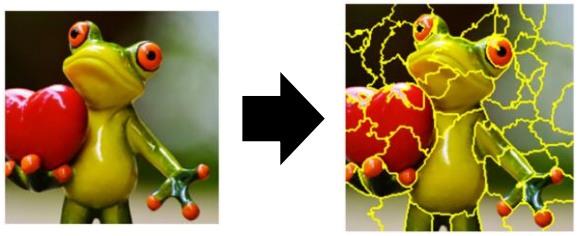
$x_1 \quad \cdots\cdots \quad x_m \quad \cdots\cdots \quad x_M$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

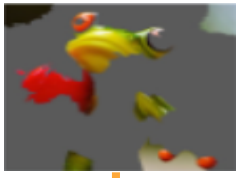
M is the number of segments.

(Slide Credit – Hung-yi Lee)

LIME – Image



- 4. Interpret the model you learned



Extract

Linear

0.85

$$y = w_1 x_1 + \cdots + w_m x_m + \cdots + w_M x_M$$

$$x_m = \begin{cases} 0 & \text{Segment m is deleted.} \\ 1 & \text{Segment m exists.} \end{cases}$$

M is the number of segments.

If $w_m \approx 0$ ➡ segment m is not related to “frog”

If w_m is positive ➡ segment m indicates the image is “frog”

If w_m is negative ➡ segment m indicates the image is not “frog”

(Slide Credit – Hung-yi Lee)

The Math behind LIME

Algorithm 1 Sparse Linear Explanations using LIME

Require: Classifier f, Number of samples N
Require: Instance x, and its interpretable version x'
Require: Similarity kernel π_x, Length of explanation K

$\mathcal{Z} \leftarrow \{\}$
for $i \in \{1, 2, 3, ..., N\}$ **do**
  $z_i' \leftarrow sample\_around(x')$
  $\mathcal{Z} \leftarrow \mathcal{Z} \cup \langle z_i', f(z_i), \pi_x(z_i) \rangle$
end for

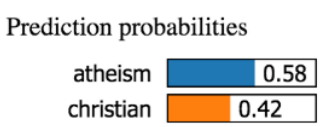
Match interpretable model to black box

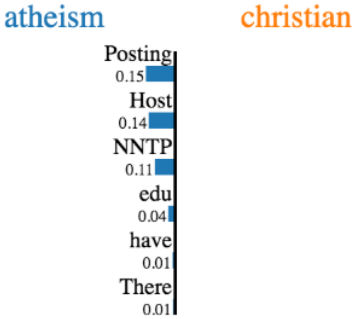
Control complexity of the model

$$\xi(x) = \underset{g \in G}{\operatorname{argmin}} \; \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

$$\mathcal{L}(f, g, \pi_x) = \sum_{z, z' \in \mathcal{Z}} \pi_x(z) \left(f(z) - g(z')\right)^2$$

Example from NLP





Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the
net. If anyone has a contact please post on the net or email me.

Rationalization Models

General Idea

Extractor

Classifier

Tree frog
(97%)

---

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...

Extractor

this beer pours ridiculously clear with tons of carbonation that forms a rather impressive rocky head that settles slowly into a fairly dense layer of foam. this is a real good lookin' beer, unfortunately it gets worse from here ...

Classifier

Positive (98%)

Natural Language Processing with Deep Learning

CS224N/Ling284

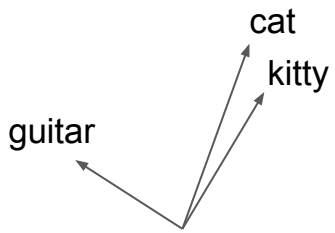
John Hewitt

Analysis and Interpretability of Neural NLP

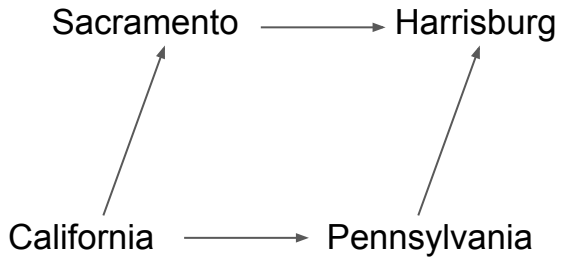
What we've seen: simple analyses of word2vec

Bold type: Math property

Italic type: interpretation



We interpret **cosine similarity** as *semantic similarity*

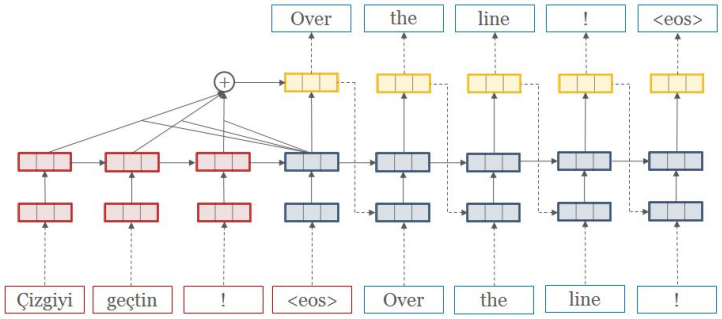


Some *relationships* are encoded as **vector differences**

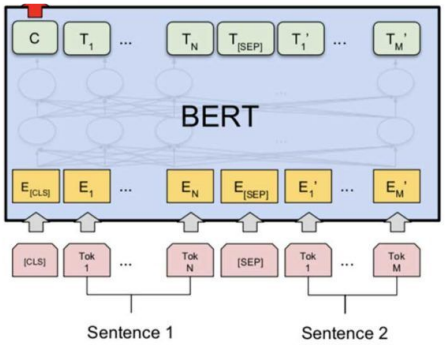
Knowing what *properties* word embeddings have: useful for practitioners!
Knowing that word embeddings encode *undesirable social biases*: useful for everyone!

Neural networks are worthy subjects of study

Machine Translation



Language Modeling



Question Answering

The first recorded travels by Europeans to China and back date from this time. The most famous traveler of the period was the Venetian Marco Polo, whose account of his trip to "Cambaluc," the capital of the Great Khan, and of life there astounded the people of Europe. The account of his travels, Il milione (or, The Million, known in English as the Travels of Marco Polo), appeared about the year 1299. Some argue over the accuracy of Marco Polo's accounts due to the lack of mentioning the Great Wall of China, tea houses, which would have been a prominent sight since Europeans had yet to adopt a tea culture, as well the practice of foot binding by the women in capital of the Great Khan. Some suggest that Marco Polo acquired much of his knowledge through contact with Persian traders since many of the places he named were in Persian.

How did some suspect that Polo learned about China instead of by actually visiting it?

Answer: through contact with Persian traders

It's wild that any of our models work at all

- Their behavior is an emergent property of *data* and our *design decisions*
- Accuracy on a held out test set is not sufficient to fully characterize them

Lecture Outline

Lecture 20: Analysis and Interpretability of Neural NLP

1. Motivation: what are our models doing?
2. **Neural networks as linguistic test subjects**
3. Careful ablation studies and architecture modifications
4. Analysis of inherently interpretable architectures
5. Playing the adversary: breaking NLP models
6. Analyzing representations using supervised methods
7. Aggregating analysis insights across studies

Neural networks as linguistic test subjects

Neural networks as linguistic test subjects

Neural networks as linguistic test subjects: Humans

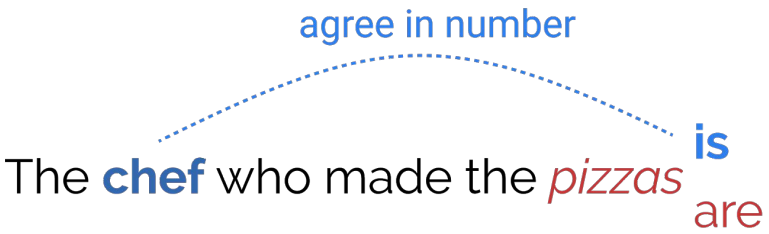
How do we understand language behavior in **humans**?

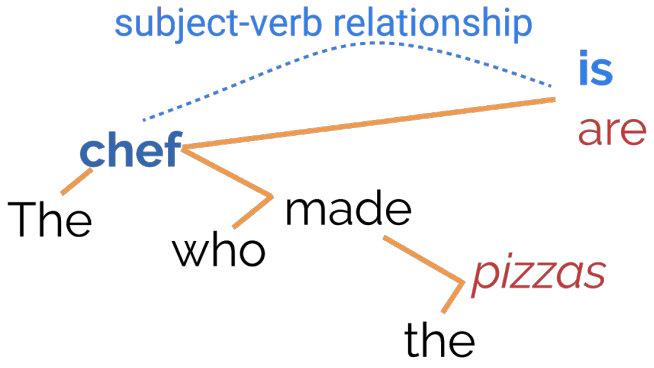
One method: *minimal pairs. What sounds "okay" to a speaker?*

The chef who made the pizzas is ← "*Acceptable*"

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

Idea: English present-tense verbs *agree in number* with their subject.

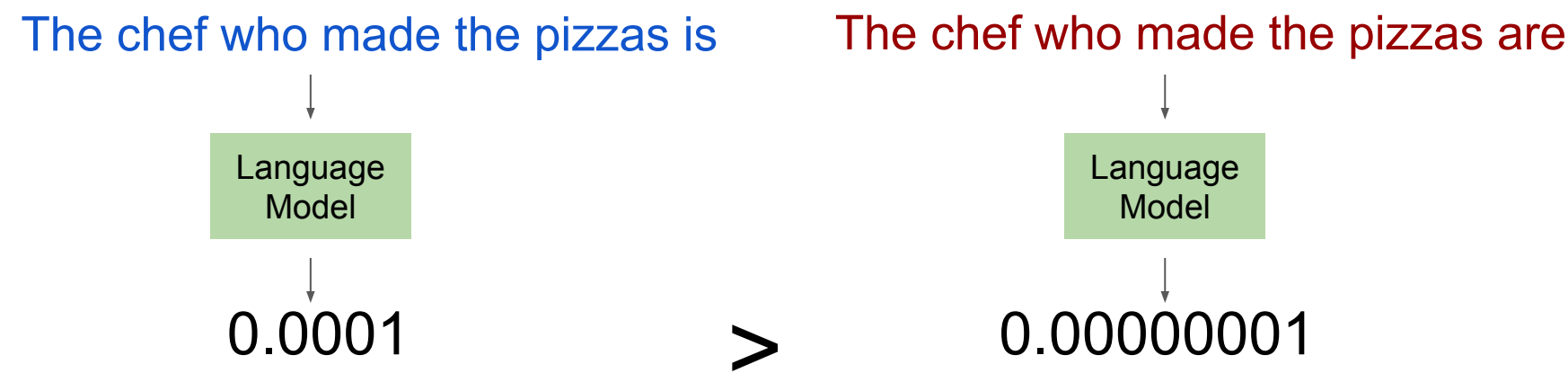




Neural networks as linguistic test subjects

How do we understand language behavior in **language models**?

One method: *minimal pairs. Is the acceptable sentence higher-probability?*



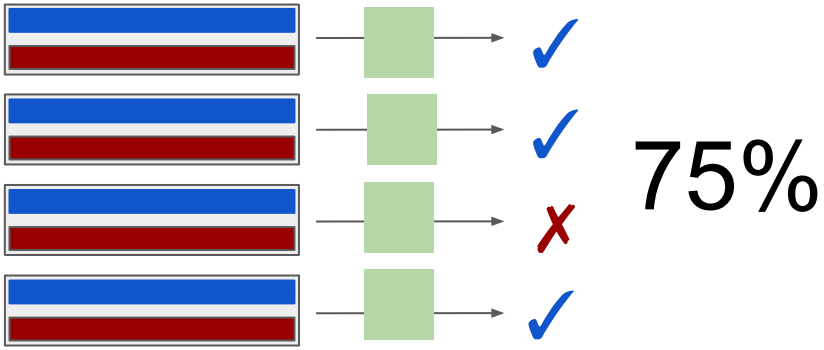
Premise: A language model should assign higher probability to the acceptable sentence in any minimal pair.

Linzen et al., 2016

Neural networks as linguistic test subjects

Steps to conduct a *minimal pairs* test on a language model:

1. Gather or construct a test set of minimal pairs which require specific aspects of understanding to distinguish.
2. Run your language model on the pairs, and report percent of pairs the model predicts as desired.



Neural networks as linguistic test subjects

Example: Do LMs show Subject-Verb number agreement across attractors?

The chef who made the pizzas and talked to the customers is

subject *attractor* *attractor* *verb*

	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
LSTM, H=50†	6.8	32.6	≈50	≈65	≈70
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**
1B Word LSTM (repl)	2.8	8.0	14.0	21.8	20.0
Char LSTM	**1.2**	5.5	11.8	20.4	27.8

# of attractors between subject and verb

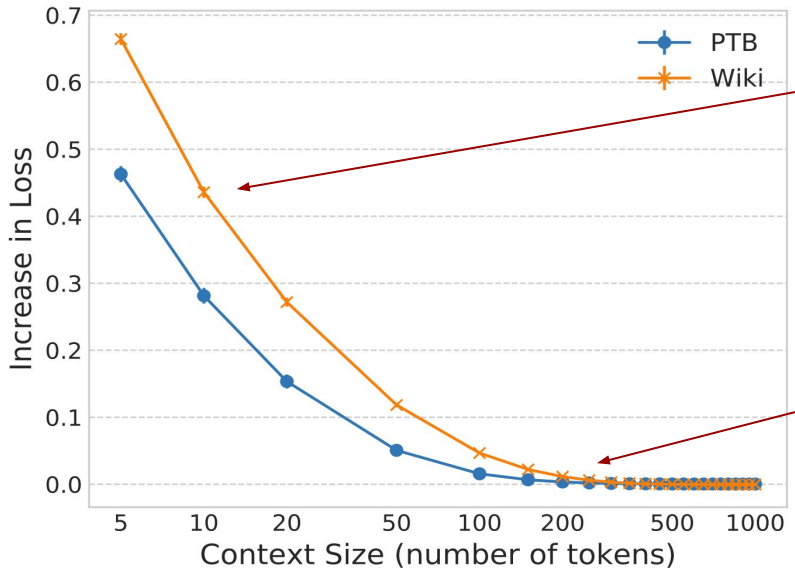
Error rate on a large corpus of minimal pairs

LMs do *really* well!?

[Kuncoro et al., 2018]

Neural networks as linguistic test subjects

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.



Only giving the LM 10 words of context at test time makes the test error go up.

Only giving the LM 250 words of context *doesn't change its loss*, so it's not using contexts longer than 250 words much.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

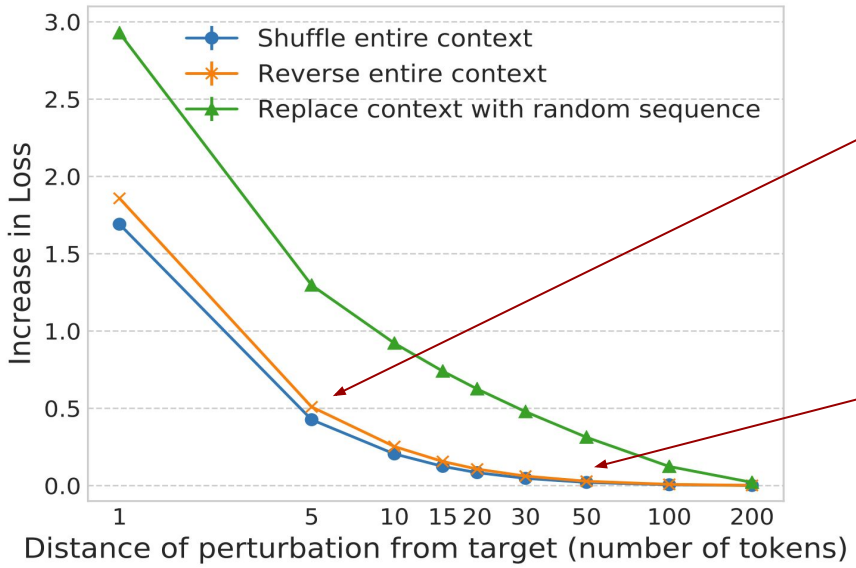
Question: How does an LSTM language model use its long-distance contexts?

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

Method: Modify the test set to remove long contexts, or replace them with longer words. Evaluate whether the LM perplexity changes.



Shuffling the order of the context further than 5 words away increases loss, so the LM cares about word order past 5 words.

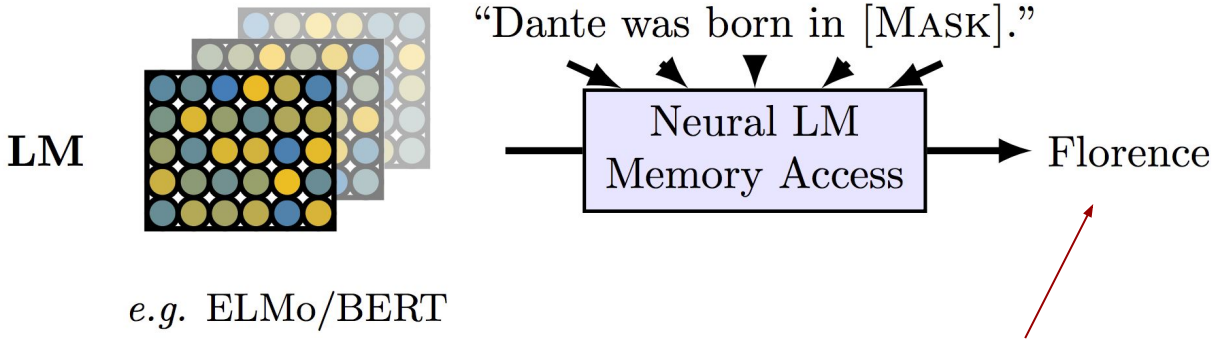
Shuffling the word order of the context further than 50 words away *doesn't* increase loss, so the LM treats words 50-250 effectively as a bag-of-words.

[Khandelwal et al., 2018]

Neural networks as linguistic test subjects

Question: Do LMs memorize factual relations?

Method:



Check if most likely word under the LM is a correct answer.
Eval: % of these relations for which this holds.

[Petroni et al., 2019]

Neural networks as linguistic test subjects

Question: Do LMs memorize factual relations?

Evaluation:

Baseline: Return word that shows up most with the subject (Dante) and the relation (born in)

BERT-base and BERT-large: memorize a surprising number of facts

Corpus	Relation	Statistics		Baselines		KB		LM					
		#Facts	#Rel	Freq	DrQA	RE_n	RE_o	Fs	Txl	Eb	E5B	Bb	Bl
Google-RE	birth-place	2937	1	4.6	-	3.5	13.8	4.4	2.7	5.5	7.5	14.9	**16.1**
	birth-date	1825	1	1.9	-	0.0	**1.9**	0.3	1.1	0.1	0.1	1.5	1.4
	death-place	765	1	6.8	-	0.1	7.2	3.0	0.9	0.3	1.3	13.1	**14.0**
	Total	5527	3	4.4	-	1.2	7.6	2.6	1.6	2.0	3.0	9.8	**10.5**

[Khandelwal et al., 2018]

Lecture Outline

Lecture 20: Analysis and Interpretability of Neural NLP

1. Motivation: what are our models doing?
2. Neural networks as linguistic test subjects
3. **Careful ablation studies and architecture modifications**
4. Analysis of inherently interpretable architectures
5. Playing the adversary: breaking NLP models
6. Analyzing representations using supervised methods
7. Aggregating analysis insights across studies

Viewing model studies as network analysis

Question: What is necessary, or even *good*, about my network design?
Method: Make targeted model changes; observe validation accuracy

Ex: The Transformer interleaves *self-attention* with *feed-forward* layers

[Press et al., 2019]

Viewing model studies as network analysis

Question: What is necessary, or even *good*, about my network design?
Method: Make targeted model changes; observe validation accuracy

Ex: The Transformer interleaves *self-attention* with *feed-forward* layers

→ 18.40 PPL

But what if we re-ordered them?

→ **17.96 PPL**
(Better!)

Bunch of self-attention

Bunch of feed-forward

[Press et al., 2019]

Lecture Outline

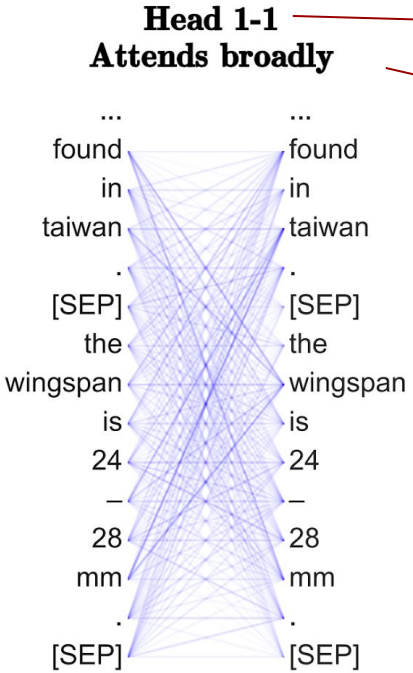
Lecture 20: Analysis and Interpretability of Neural NLP

1. Motivation: what are our models doing?
2. Neural networks as linguistic test subjects
3. Careful ablation studies and architecture modifications
4. **Analysis of inherently interpretable architectures**
5. Playing the adversary: breaking NLP models
6. Analyzing representations using supervised methods
7. Aggregating analysis insights across studies

Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.



Attention head 1 of layer 1

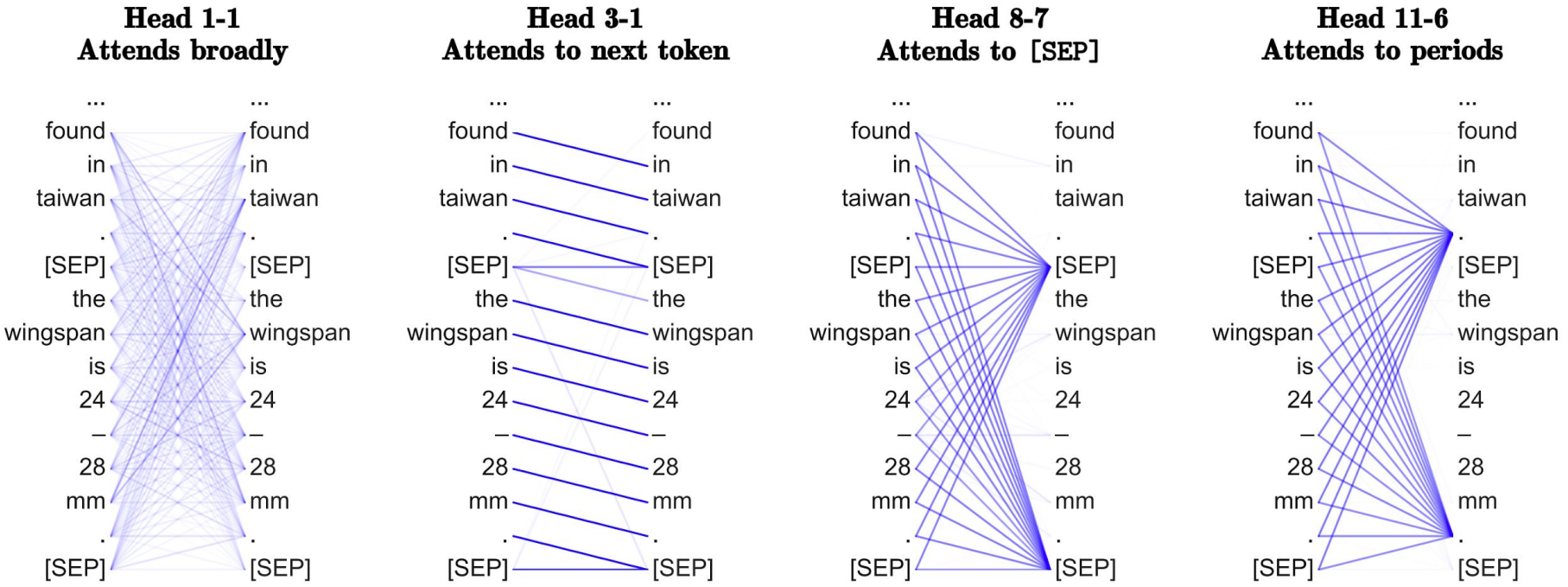
Not just on this sentence, but on most sentences, this head demonstrates a behavior like this.

[Clark et al., 2019]

Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.



[Clark et al., 2019]

Analysis of “interpretable” architectures

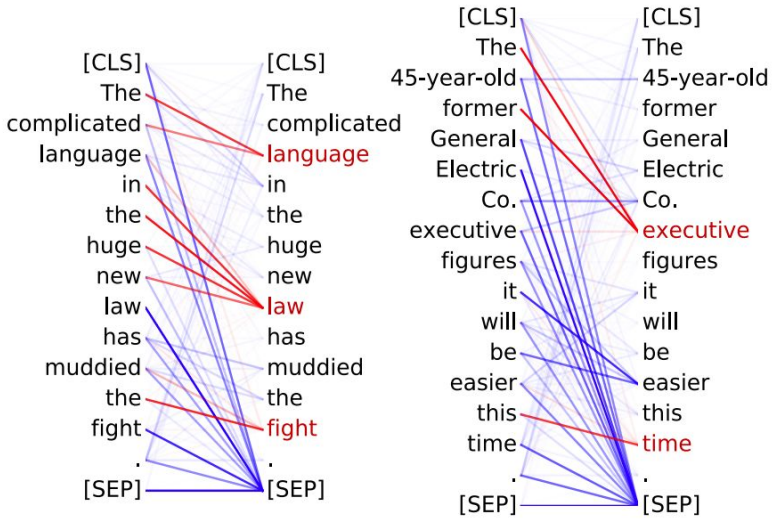
Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.

Head 8-11

- **Noun modifiers** (e.g., determiners) attend to their noun
- 94.3% accuracy at the `det` relation

Interpretation + Quantitative Analysis



Qualitative Model behavior

[Clark et al., 2019]

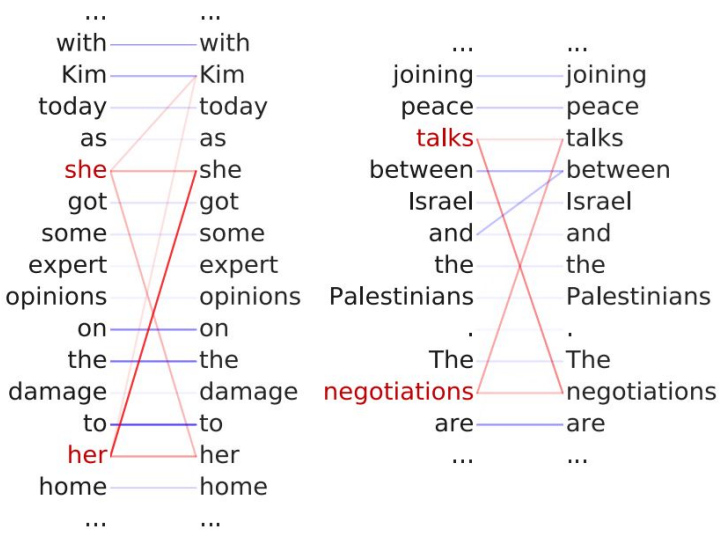
Analysis of “interpretable” architectures

Some architectures have components that lend themselves to inspection

Example: Try to characterize each attention head of BERT.

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



Interpretation + Quantitative Analysis

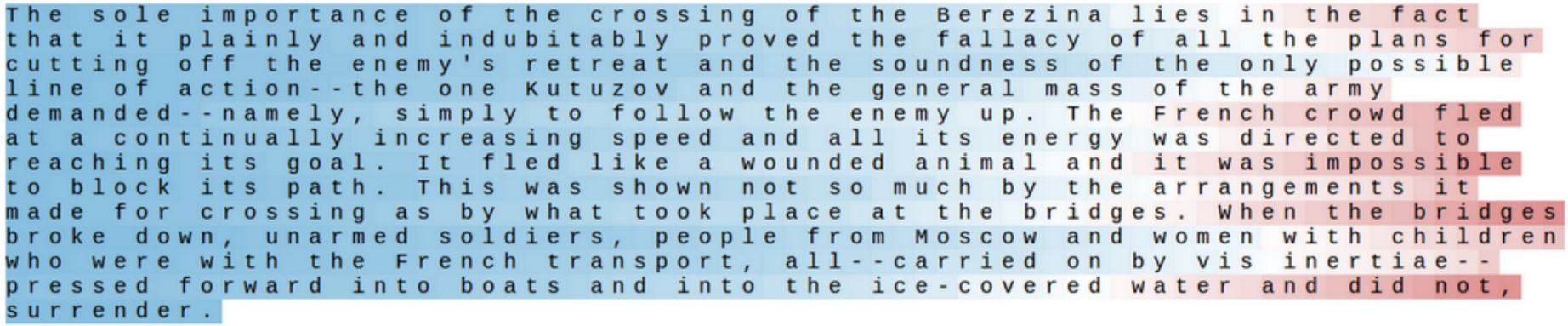
Qualitative Model behavior

[Clark et al., 2019]

Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?

Cell sensitive to position in line:



[Karpathy et al., 2016]

Understanding representations by inspection

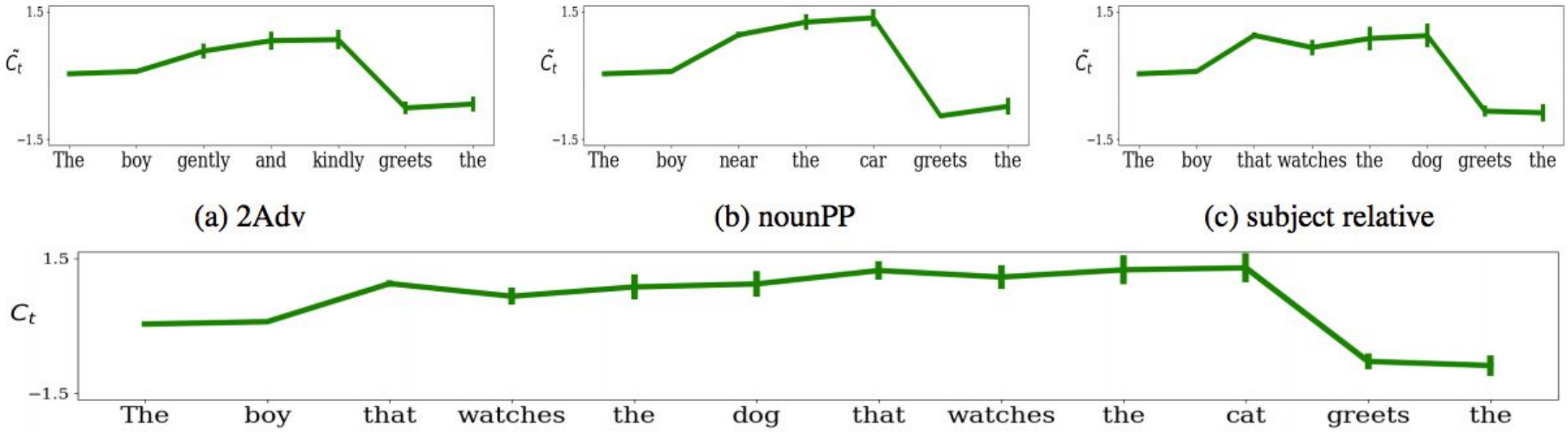
Are individual hidden units in recurrent neural networks interpretable?

Cell that turns on inside quotes:

[Karpathy et al., 2016]

Understanding representations by inspection

Are individual hidden units in recurrent neural networks interpretable?



Interpretation: this LSTM cell unit fires approximately between a subject and its verb

[Lakretz et al., 2019]

Lecture Outline

Lecture 20: Analysis and Interpretability of Neural NLP

1. Motivation: what are our models doing?
2. Neural networks as linguistic test subjects
3. Careful ablation studies and architecture modifications
4. Analysis of inherently interpretable architectures
5. **Playing the adversary: breaking NLP models**
6. Analyzing representations using supervised methods
7. Aggregating analysis insights across studies

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their input?

By robust, in this case we mean their outputs do not change.

Article: Super Bowl 50
Paragraph: "*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 and is currently Denver's Executive Vice President of Football Operations and General Manager.*

Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway

The performance of this QA model on this input looks good!

[Jia et al., 2019]

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their input?

By robust, in this case we mean their outputs do not change.

Article: Super Bowl 50
Paragraph: "*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 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway

The performance of this QA model on this input looks good!

This sentence is irrelevant; adding it does not change the answer.

[Jia et al., 2019]

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their input?

By robust, in this case we mean their outputs do not change.

Article: Super Bowl 50
Paragraph: "*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 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

The performance of this QA model on this input looks good!

This sentence is irrelevant; adding it does not change the answer.

But it changes the model's prediction :(

Interpretation: model is not really working

[Jia et al., 2019]

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their 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

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

(b) Original Question and Answer

The performance of this QA model on this input looks good!

[Ribeiro et al., 2018]

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their 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

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)

The performance of this QA model on this input looks good!

This typo is annoying, but a reasonable language learner would be robust to it.

[Ribeiro et al., 2018]

Understanding models by breaking them

Question: Are our models robust to innocuous changes in their 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

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

(b) Original Question and Answer

The performance of this QA model on this input looks good!

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 language learner would be robust to it.

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

(d) **Semantically Equivalent Adversary**

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

[Ribeiro et al., 2018]

Understanding models by breaking them

Question: Are our models robust to typos or noise in their input?

[Belinkov and Bisk, 2018]

Understanding models by breaking them

Question: Are our models robust to typos or noise in their input?

Humans

[Belinkov and Bisk, 2018]

Understanding models by breaking them

Question: Are our models robust to typos or noise in their input?

Humans

"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."

Just 1 data point/meme, but interpretation: humans are!

[Belinkov and Bisk, 2018]

Understanding models by breaking them

Question: Are our models robust to typos or noise in their input?

BLEU on clean text

		Vanilla
French	charCNN	42.54
German	charCNN	34.79
	char2char	29.97
	Nematus	34.22
Czech	charCNN	25.99
	char2char	25.71
	Nematus	29.65

[Belinkov and Bisk, 2018]

Understanding models by breaking them

Question: Are our models robust to typos or noise in their input?

BLEU on clean text

BLEU on data with noise like we just saw

BLEU on data with natural noise (real misspellings, +)

			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]

Lecture Outline

Lecture 20: Analysis and Interpretability of Neural NLP

1. Motivation: what are our models doing?
2. Neural networks as linguistic test subjects
3. Careful ablation studies and architecture modifications
4. Analysis of inherently interpretable architectures
5. Playing the adversary: breaking NLP models
6. **Analyzing representations using supervised methods**
7. Aggregating analysis insights across studies

Understanding representations by probing

Hypothesis:

Neural models, especially large ones like BERT, perform well without any explicit linguistic supervision in part because they learn similar notions themselves.

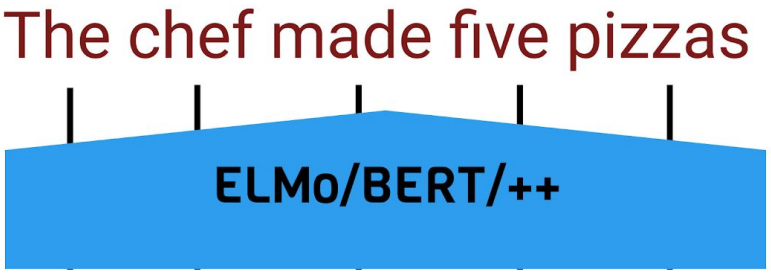
Question:

Do neural networks' internal representations encode linguistic notions of structure, like *parts-of-speech*, *dependency trees*, *named entities?*

[Lakretz et al., 2019]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

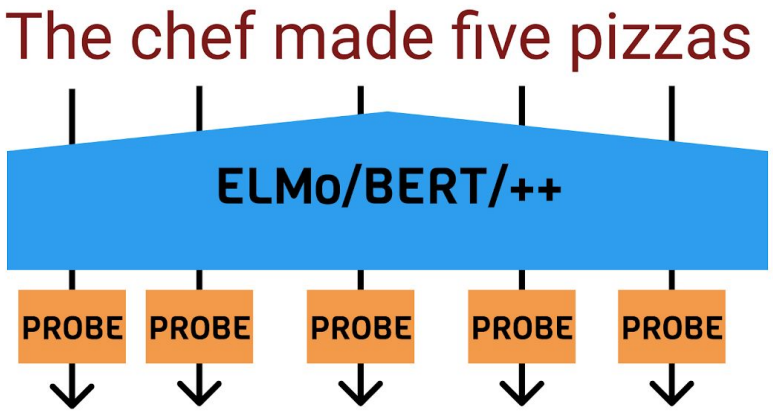


[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...] [Hewitt and Liang, 2019]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family
to decode the task**.** (e.g., linear)



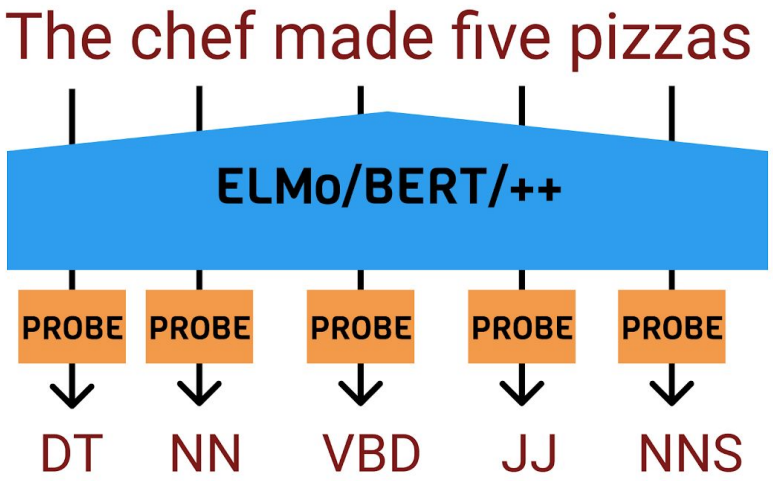
[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...] [Hewitt and Liang, 2019]

Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family to decode the task**.** (e.g., linear)

Train a function representations --> task



[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...] [Hewitt and Liang, 2019]

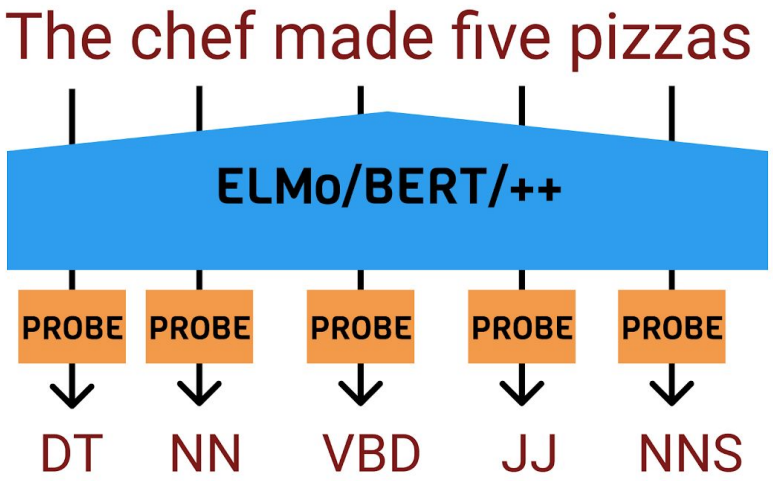
Probing: supervised analysis of representations

Does my network make task (e.g., part-of-speech) labels accessible?

Choose a function family to decode the task. (e.g., linear)

Train a function representations --> task

Interpret accuracy on held-out data



(Don't fine-tune the model while doing this!)

[Shi et al., 2016, Peters et al., 2018, Tenney et al., 2019, Liu et al., 2019, Hewitt and Manning 2019, Kim et al., 2019...] [Hewitt and Liang, 2019]

Understanding representations by probing

Pretrained Representation			POS						Supersense ID		
	Avg.	CCG	PTB	EWT	Chunk	NER	ST	GED	PS-Role	PS-Fxn	EF
BERT (base, cased) best layer	84.09	93.67	96.95	95.21	92.64	82.71	93.72	43.30	**79.61**	87.94	75.11
BERT (large, cased) best layer	**85.07**	**94.28**	96.73	**95.80**	**93.64**	**84.44**	93.83	**46.46**	79.17	**90.13**	**76.25**
GloVe (840B.300d)	59.94	71.58	90.49	83.93	62.28	53.22	80.92	14.94	40.79	51.54	49.70
Previous state of the art (without pretraining)	83.44	94.7	97.96	95.82	95.77	91.38	95.15	39.83	66.89	78.29	77.10

Interpretation 1: BERT's representations, when used as features for a linear classifier, lead to high accuracy on linguistic tasks; this is evidence that BERT makes these properties linearly accessible.

Interpretation 2: BERT-large seems to perform better than BERT-base, indicating that it may learn better representations of linguistic properties.

[Liu et al., 2019; table cropped]

Understanding representations by probing

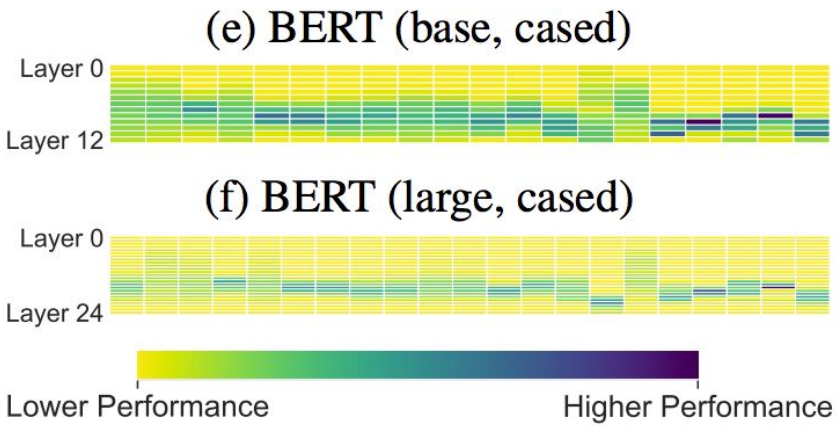


Figure 3: A visualization of layerwise patterns in task performance. Each column represents a probing task, and each row represents a contextualizer layer.

Interpretation: BERT makes linguistic properties most accessible in middle layers

[Liu et al., 2019; figure cropped]

Conclusions

- Neural models are strong, but often fragile and not robust. Do not naively trust metrics such as accuracy.

Conclusions

- Neural models are strong, but often fragile and not robust. Do not naively trust metrics such as accuracy.
- Many tools can approximate an “explanation” for the model behavior. They are almost never perfect.

Conclusions

- Neural models are strong, but often fragile and not robust. Do not naively trust metrics such as accuracy.
- Many tools can approximate an “explanation” for the model behavior. They are almost never perfect.
    - Neural models are more like living organisms than physical systems. We probably cannot find a simple explanation to the model’s behavior, but we can poke it and analyze different aspects of it.

Conclusions

- Neural models are strong, but often fragile and not robust. Do not naively trust metrics such as accuracy.
- Many tools can approximate an “explanation” for the model behavior. They are almost never perfect.
  - Neural models are more like living organisms than physical systems. We probably cannot find a simple explanation to the model’s behavior, but we can poke it and analyze different aspects of it.
- Think critically when someone claims that their model is “interpretable” or “in par with humans”.

Conclusions

- Neural models are strong, but often fragile and not robust. Do not naively trust metrics such as accuracy.
- Many tools can approximate an “explanation” for the model behavior. They are almost never perfect.
  - Neural models are more like living organisms than physical systems. We probably cannot find a simple explanation to the model’s behavior, but we can poke it and analyze different aspects of it.
- Think critically when someone claims that their model is “interpretable” or “in par with humans”.
- I hope you enjoyed the course.