Interpretability, Explainability, Model Analysis

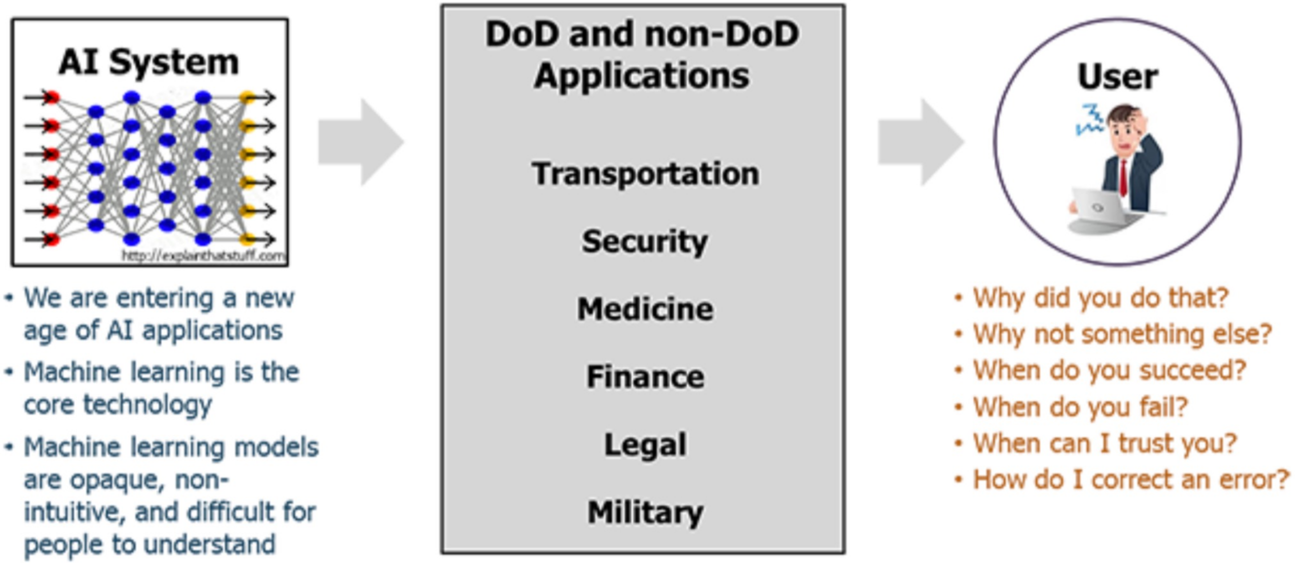
CS 6301

Outline

- Attention
- Robustness
- Probing
- Interpreting Individual Neurons
- Saliency Map
- Influence Function
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Explainable Artificial Intelligence (XAI)

Dr. Matt Turek



AI System

- We are entering a new age of AI applications
- Machine learning is the core technology
- Machine learning models are opaque, non-intuitive, and difficult for people to understand

DoD and non-DoD Applications

Transportation

Security

Medicine

Finance

Legal

Military

User

- Why did you do that?
- Why not something else?
- When do you succeed?
- When do you fail?
- When can I trust you?
- How do I correct an error?

Figure 1. The Need for Explainable AI

https://www.darpa.mil/program/explainable-artificial-intelligence

Beyond accuracy (model)

Only achieving high accuracy is not enough, but we need to answer

- Where is the errors/bugs in my models?

- Why are my models behaving the way they are?
- Why do my model fail on this example?

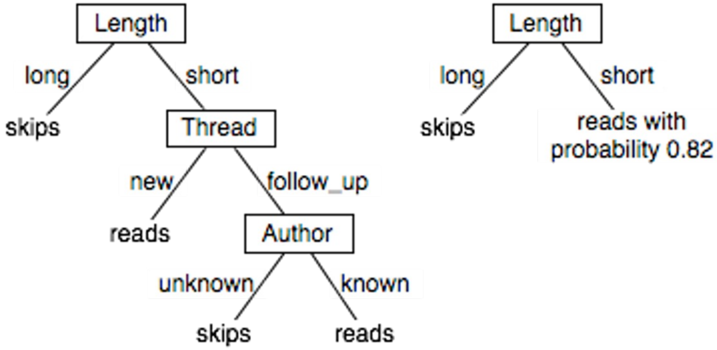
- How can I improve my models?
- Why should I trust my models?

Beyond accuracy (model & data)

Only achieving high accuracy is not enough, but we need to answer

- Where is the errors/bugs in my models?
- Where is the errors/bugs in my data?
- Why are my models behaving the way they are?
- Why do my model fail on this example?
- Can I locate particular training examples that cause the model behavior?
- How can I improve my models?
- Why should I trust my models?

Interpretability and Explainability



They are often used interchangeably

But still subtle difference between the two

- We consider a model to be “**interpretable**” if the model itself can provide humanly understandable interpretations of its predictions. Note that such a model is no longer a black box *to some extent*. For example, a decision tree model is an “interpretable” one.
- An “**explainable**” model implies that the model is still a black box whose predictions could potentially be understood by post hoc explanation techniques.

Outline

- **Attention**
- Robustness
- Probing
- Interpreting Individual Neurons
- Saliency Map
- Influence Function
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Analysis on BERT - Attention

What Does BERT Look At? An Analysis of BERT’s Attention (Clark et al., 2019)

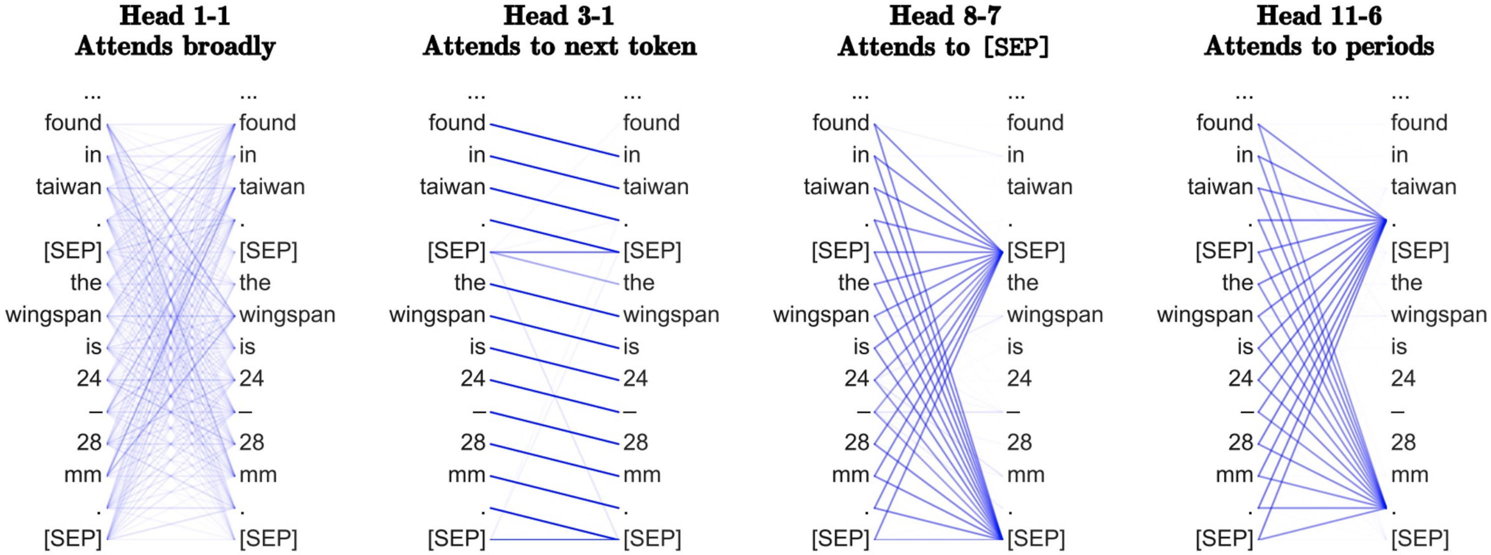


Figure 1: Examples of heads exhibiting the patterns discussed in Section 3. The darkness of a line indicates the strength of the attention weight (some attention weights are so low they are invisible).

Robustness Analysis for interpretation and explanation

Understanding model robustness can give many insights in interpretation and explanation

- When will the model fail?
- How can we test the behavior of the model?
- Is the model right for the right reason?
- …

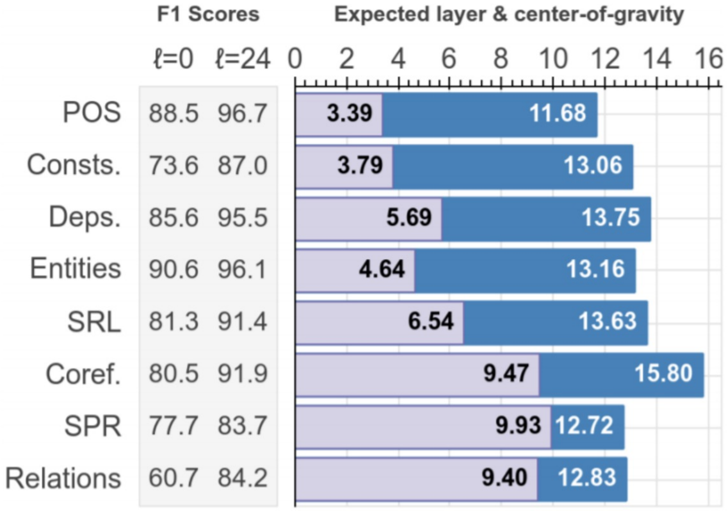
Analysis on BERT

BERT Rediscovers the Classical NLP Pipeline (Tenney et al., 2019)

- Quantify where linguistic information is captured within the network.
- Find that the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way, and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference

Increasing depth in the network

Increasing abstractness of linguistic properties



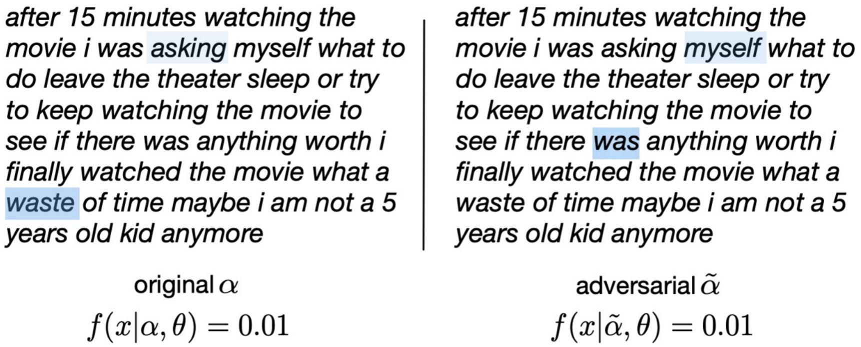
	F1 Scores		Expected layer & center-of-gravity	
	ℓ=0	ℓ=24		
POS	88.5	96.7	3.39	11.68
Consts.	73.6	87.0	3.79	13.06
Deps.	85.6	95.5	5.69	13.75
Entities	90.6	96.1	4.64	13.16
SRL	81.3	91.4	6.54	13.63
Coref.	80.5	91.9	9.47	15.80
SPR	77.7	83.7	9.93	12.72
Relations	60.7	84.2	9.40	12.83

Attention is not explanation (Jain et al., 2019)

Does attention weights provide meaningful “explanations" for predictions?
We find that they largely do not.

- learned attention weights are frequently uncorrelated with gradient-based measures of feature importance
- one can identify very different attention distributions that nonetheless yield equivalent predictions

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).



after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

original α

$f(x|\alpha,\theta) = 0.01$

after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

adversarial $\tilde{\alpha}$

$f(x|\tilde{\alpha},\theta) = 0.01$

Outline

- Attention
- Robustness
- **Probing**
- Interpreting Individual Neurons
- Saliency Map
- Influence Function
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Probing

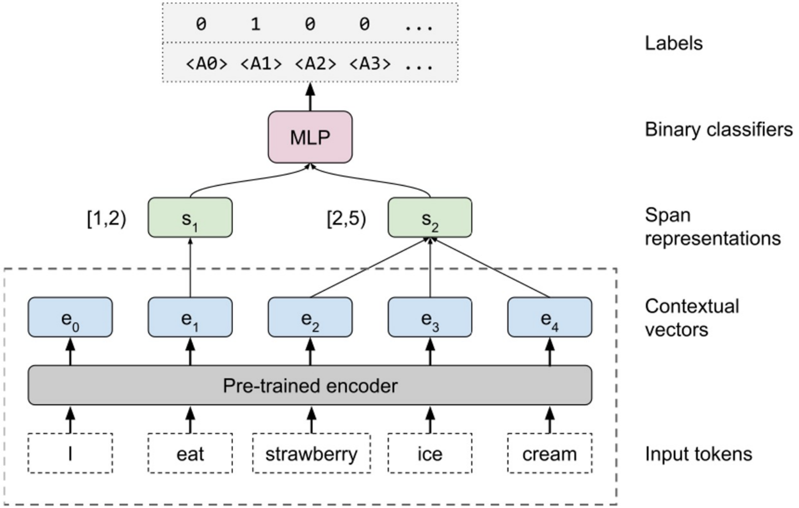


Figure 1: Probing model architecture (§ 3.1). All parameters inside the dashed line are fixed, while we train the span pooling and MLP classifiers to extract information from the contextual vectors. The example shown is for semantic role labeling, where $s^{(1)} = [1, 2)$ corresponds to the predicate ("eat"), while $s^{(2)} = [2, 5)$ is the argument ("strawberry ice cream"), and we predict label A1 as positive and others as negative. For entity and constituent labeling, only a single span is used.

Probing: Supervised Analysis of Neural Networks

Linguistic Knowledge and Transferability of Contextual Representations (Liu et al., 2019)

A probe, i.e. a classifier trained to predict the property from the representations.

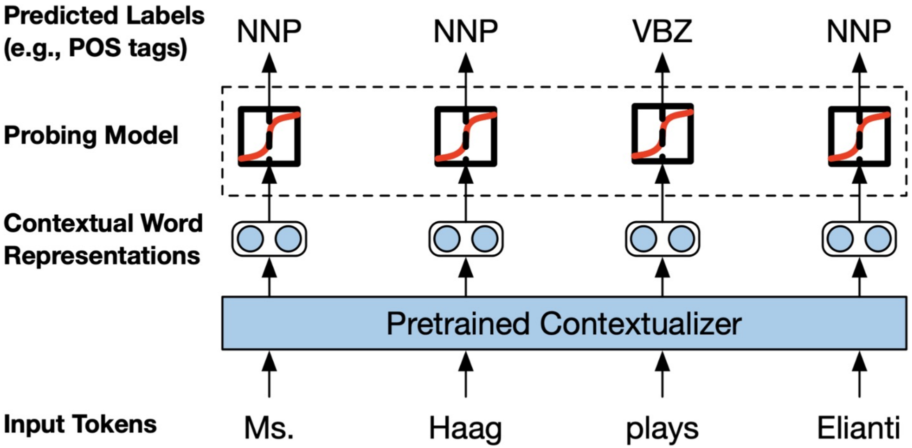
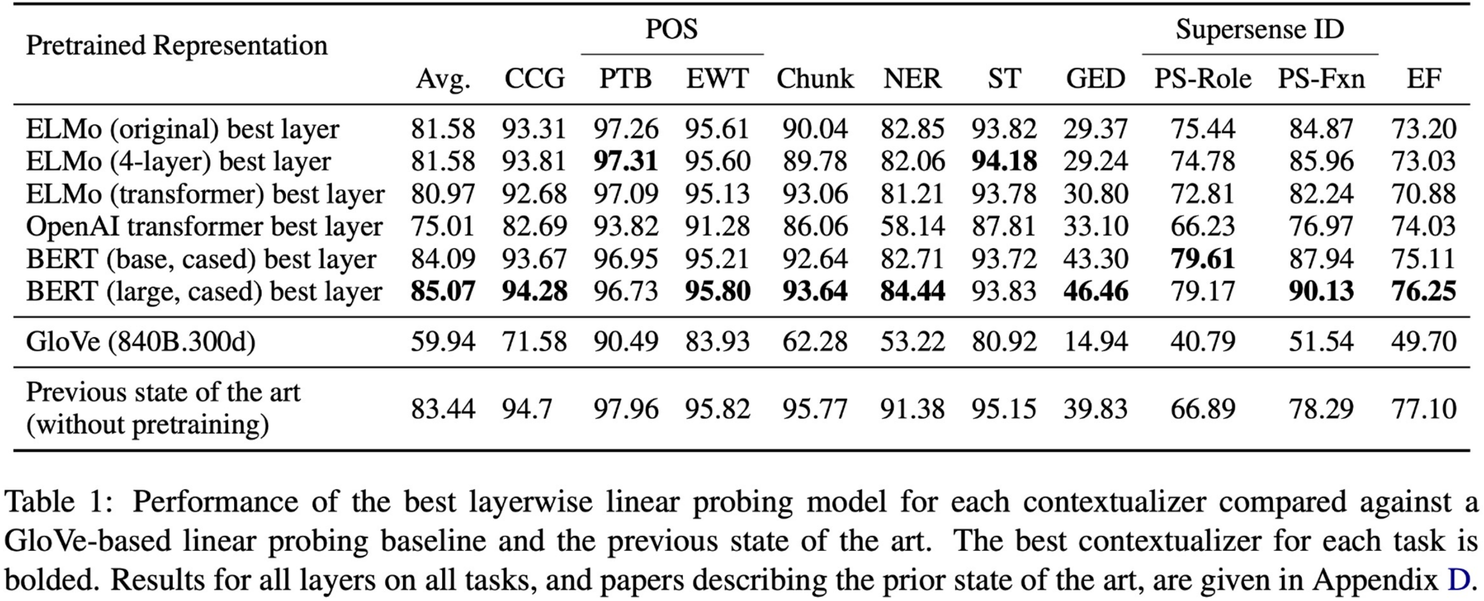

Predicted Labels (e.g., POS tags)
NNP
NNP
VBZ
NNP
Probing Model
Contextual Word Representations
Pretrained Contextualizer
Input Tokens
Ms.
Haag
plays
Elianti

Figure 1: An illustration of the probing model setup used to study the linguistic knowledge within contextual word representations.

Probing: Supervised Analysis of Neural Networks

Linguistic Knowledge and Transferability of Contextual Representations (Liu et al., 2019)



Pretrained Representation			POS						Supersense ID		
	Avg.	CCG	PTB	EWT	Chunk	NER	ST	GED	PS-Role	PS-Fxn	EF
ELMo (original) best layer	81.58	93.31	97.26	95.61	90.04	82.85	93.82	29.37	75.44	84.87	73.20
ELMo (4-layer) best layer	81.58	93.81	**97.31**	95.60	89.78	82.06	**94.18**	29.24	74.78	85.96	73.03
ELMo (transformer) best layer	80.97	92.68	97.09	95.13	93.06	81.21	93.78	30.80	72.81	82.24	70.88
OpenAI transformer best layer	75.01	82.69	93.82	91.28	86.06	58.14	87.81	33.10	66.23	76.97	74.03
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

Table 1: Performance of the best layerwise linear probing model for each contextualizer compared against a GloVe-based linear probing baseline and the previous state of the art. The best contextualizer for each task is bolded. Results for all layers on all tasks, and papers describing the prior state of the art, are given in Appendix D.

Probing: Supervised Analysis of Neural Networks

Linguistic Knowledge and Transferability of Contextual Representations (Liu et al., 2019)

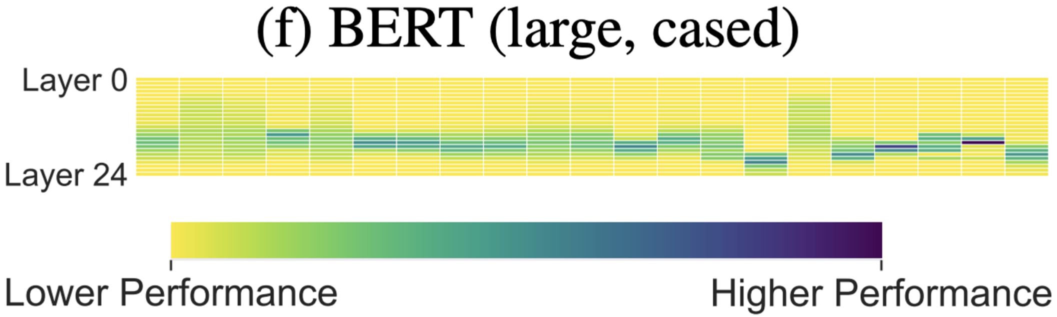
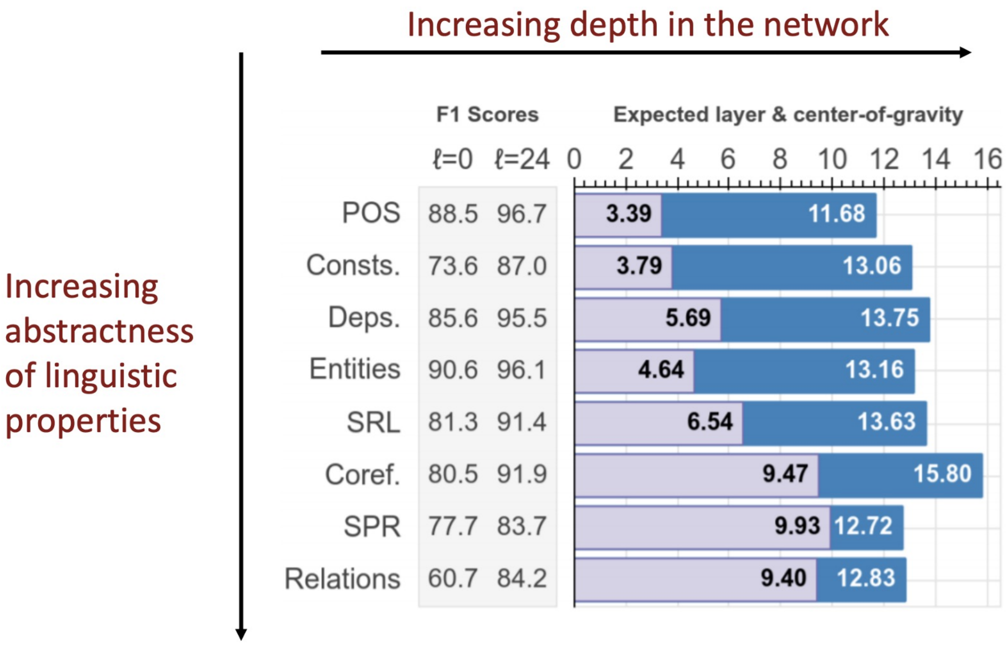


Figure 3: A visualization of layerwise patterns in task performance. Each column represents a probing task, and each row represents a contextualizer layer.

Analysis on BERT with probing

BERT Rediscovers the Classical NLP Pipeline (Tenney et al., 2019)

- Quantify where linguistic information is captured within the network.
- Find that the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way, and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference

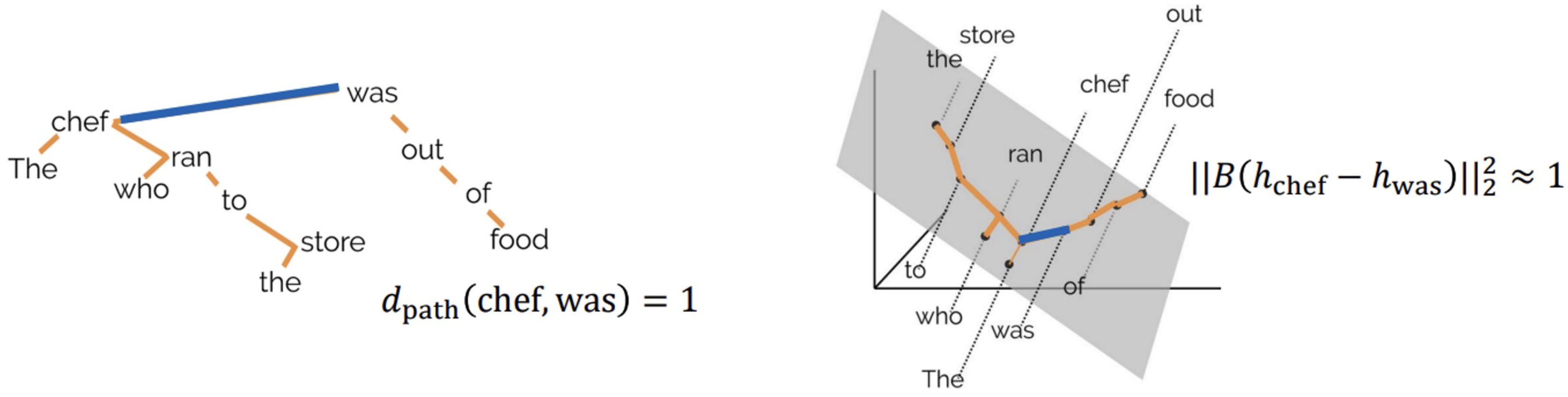


Increasing depth in the network

Increasing abstractness of linguistic properties

Structural Probing

A Structural Probe for Finding Syntax in Word Representations (Hewitt et al., 2019): BERT representations can be transformed using a matrix to encode distance in dependency parse trees.



was
chef
The
ran
who
to
store
the
out
of
food
$d_{\text{path}}(\text{chef}, \text{was}) = 1$
out
store
the
chef
food
ran
to
of
who
was
The
$||B(h_{\text{chef}} - h_{\text{was}})||_2^2 \approx 1$

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

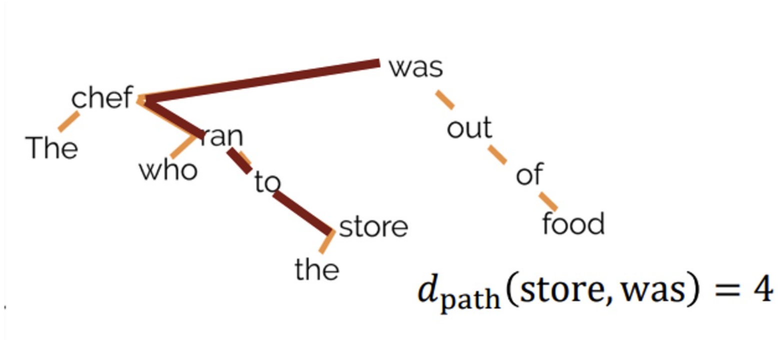
Tree path distance: the number of edges in the path between the words

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

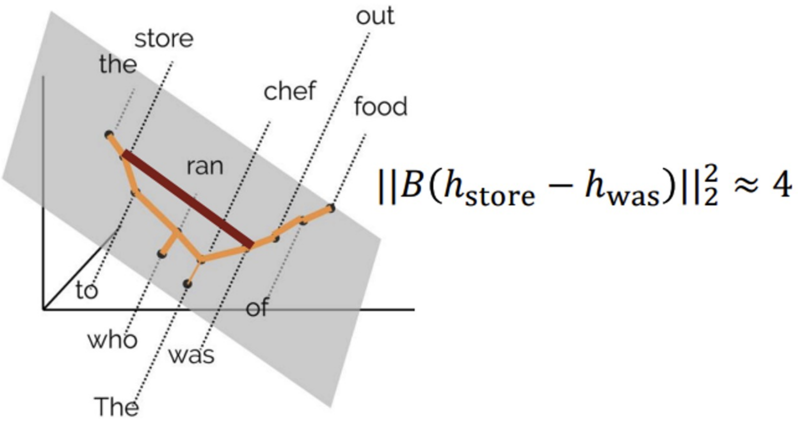
Squared Euclidean distance of BERT vectors after transformation by the (probe) matrix B.

Structural Probing

A Structural Probe for Finding Syntax in Word Representations (Hewitt et al., 2019): BERT representations can be transformed using a matrix to encode distance in dependency parse trees.



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



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

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

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

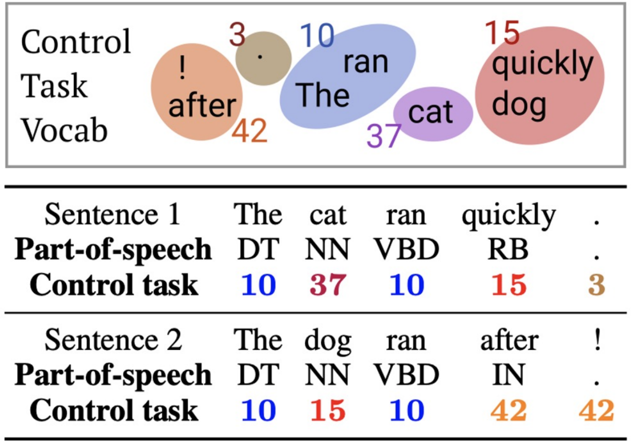
$$||B(h_{w_1} - h_{w_2})||_2^2$$

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

A good probing classifier should not be too strong

Designing and Interpreting Probes with Control Tasks (Hewitt et al., 2019)

- Probes have achieved high accuracy on linguistic tasks.
- But does this mean that the representations encode linguistic structure or just that the probe has learned the linguistic task?
- We propose *control tasks*, which associate word types with random outputs, to complement *linguistic tasks*.
- So a good probe, (one that reflects the representation), should be **selective**, achieving **high linguistic task accuracy** and **low control task accuracy**.
- We show that popular probes on ELMo representations are not selective.



Sentence 1	The	cat	ran	quickly	.
Part-of-speech	DT	NN	VBD	RB	.
Control task	**10**	**37**	**10**	**15**	**3**
Sentence 2	The	dog	ran	after	!
Part-of-speech	DT	NN	VBD	IN	.
Control task	**10**	**15**	**10**	**42**	**42**

Figure 1: Our control tasks define random behavior (like a random output, top) for each word type in the vocabulary. Each word token is assigned its type's output, regardless of context (middle, bottom.) Control tasks have the same input and output space as a linguistic task (e.g., parts-of-speech) but can only be learned if the probe memorizes the mapping.

Outline

- Attention
- Robustness
- Probing
- **Interpreting Individual Neurons**
- Saliency Map
- Influence Function
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Knowledge Neurons in Pretrained Transformers (Dai et al., 2021)

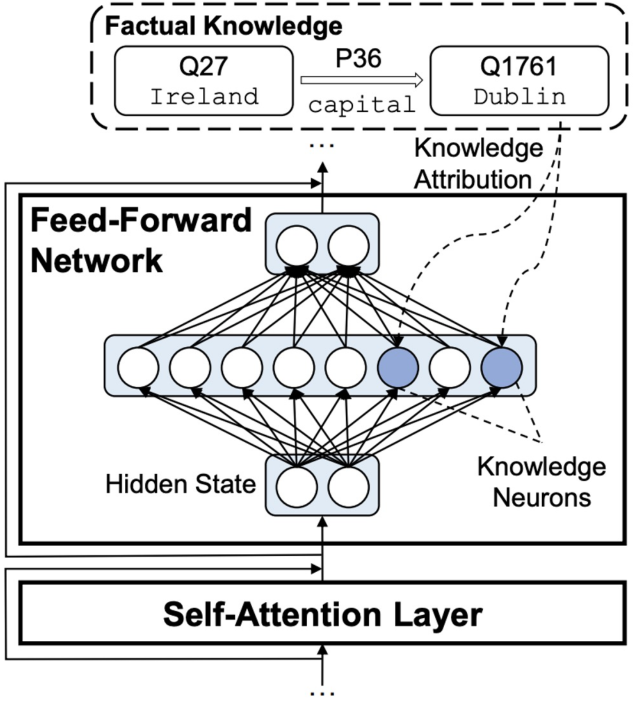
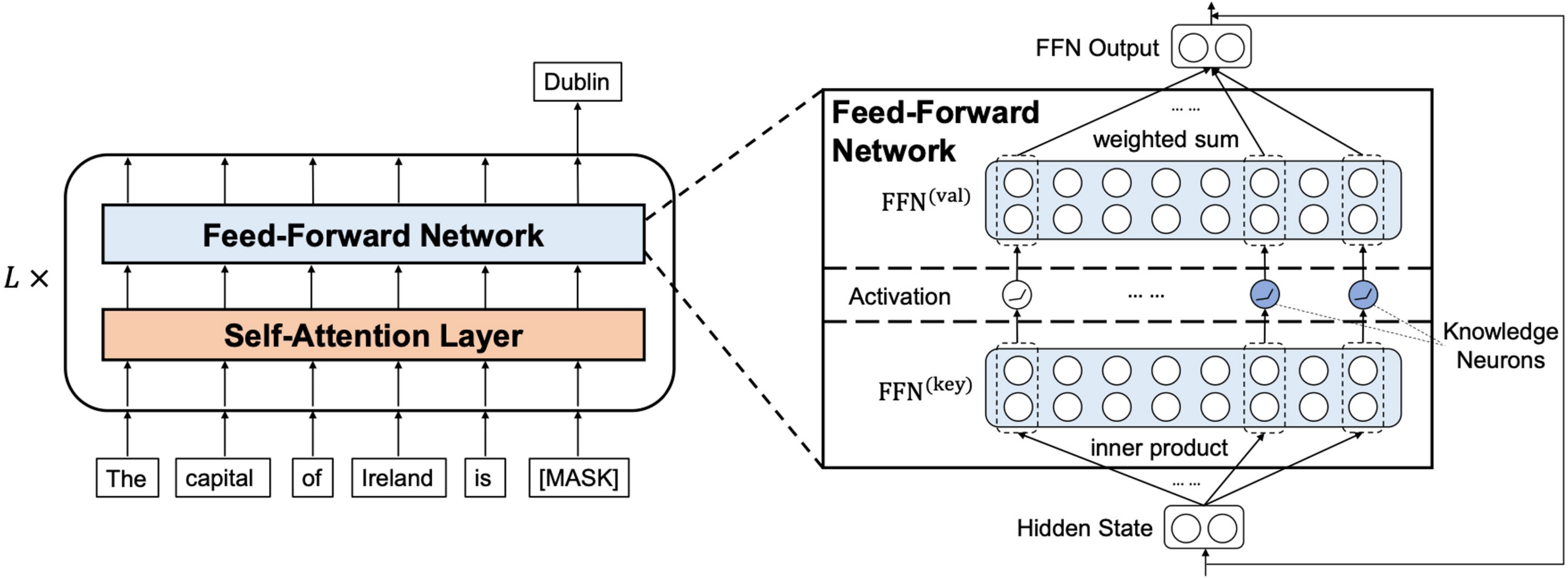


Figure 1: Through knowledge attribution, we identify knowledge neurons that express a relational fact.

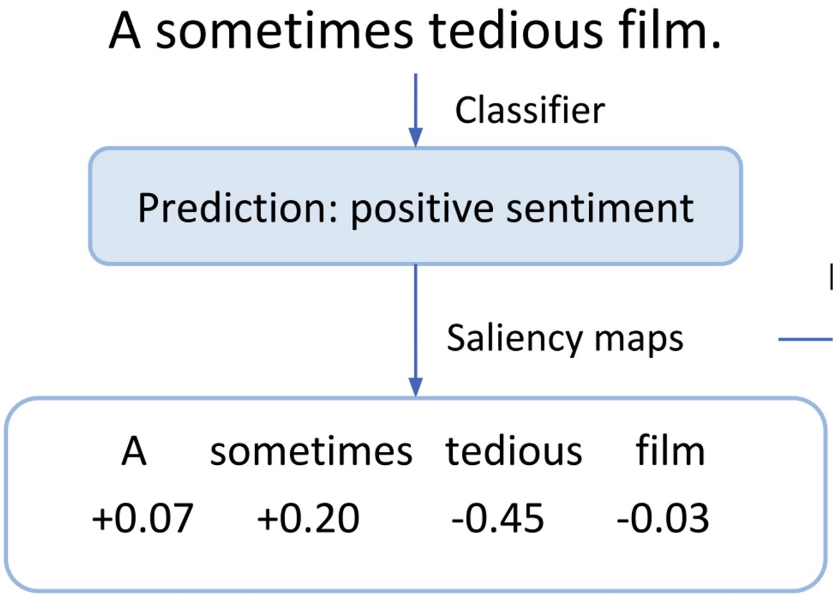

Dublin
$L \times$
Feed-Forward Network
Self-Attention Layer
The
capital
of
Ireland
is
[MASK]
FFN Output
Feed-Forward Network
weighted sum
$\mathrm{FFN}^{(\mathrm{val})}$
Activation
Knowledge Neurons
$\mathrm{FFN}^{(\mathrm{key})}$
inner product
Hidden State

Outline

- Attention
- Robustness
- Probing
- Interpreting Individual Neurons
- **Saliency Map**
- Influence Function
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Saliency Map

"Explanation by Input Features"



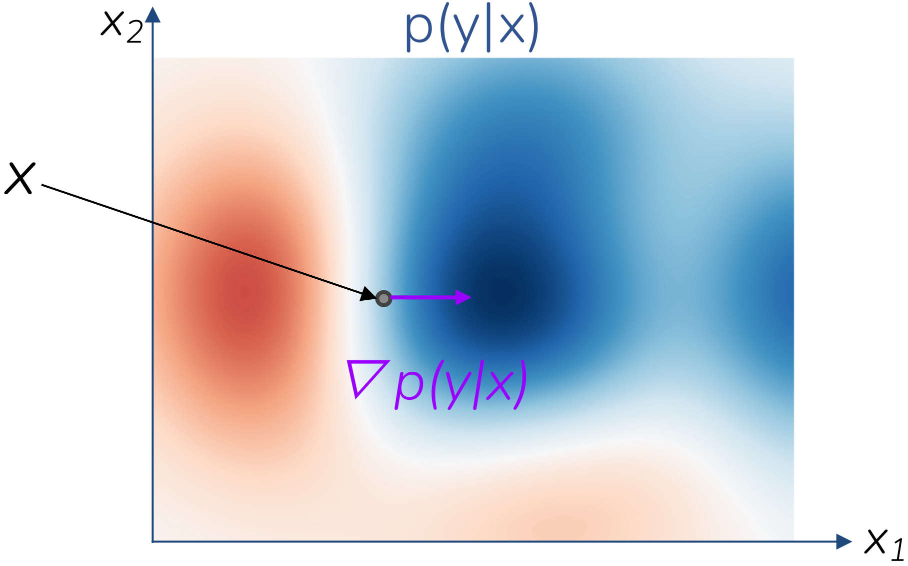
A sometimes tedious film.

A	sometimes	tedious	film
+0.07	+0.20	-0.45	-0.03

Salient tokens in the input

Gradient-based Saliency Map (Simonyan et al., 2014)

- The gradient of the loss L is computed with respect to each token **t** in the input text, and the magnitude of the gradient serves as a feature importance score
- They tell us how much the loss would change, were we to **perturb** a token by a small amount
- “gradient × input”



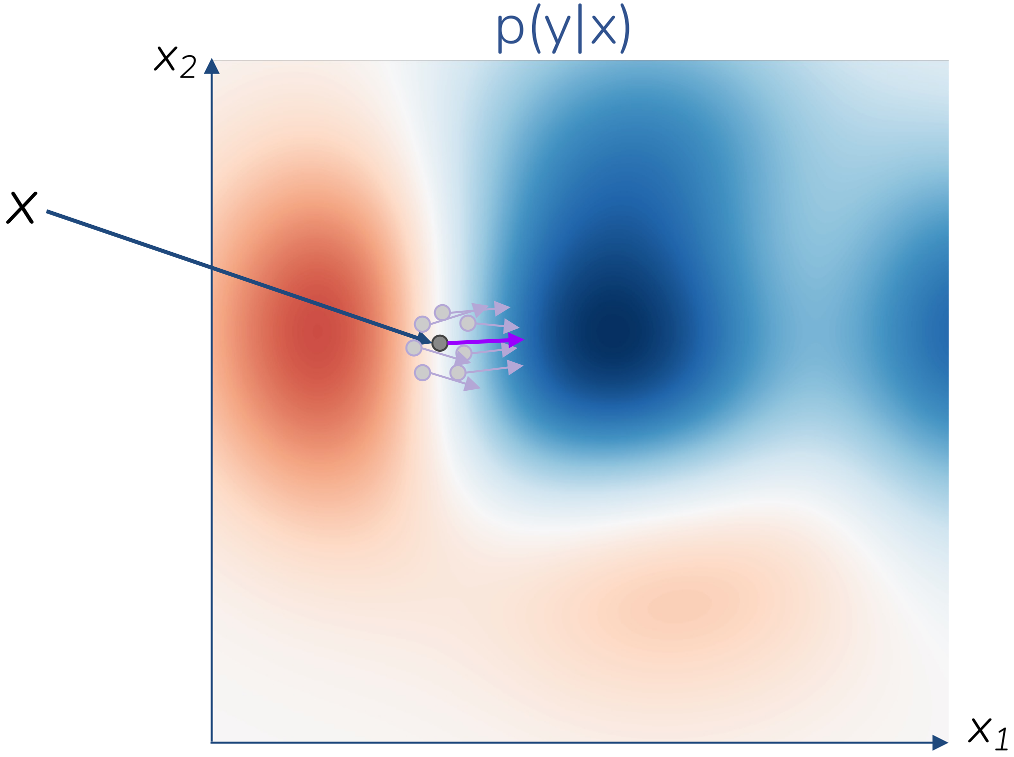
$$-\nabla_{e(t)}\mathcal{L}_{\hat{y}} \cdot e(t)$$

https://arxiv.org/pdf/2005.06676.pdf

Interpreting Predictions of NLP Models, EMNLP 2020 Tutorial [slides]

SmoothGrad (Smilkov et al., 2017)

add gaussian noise to input and average the gradient

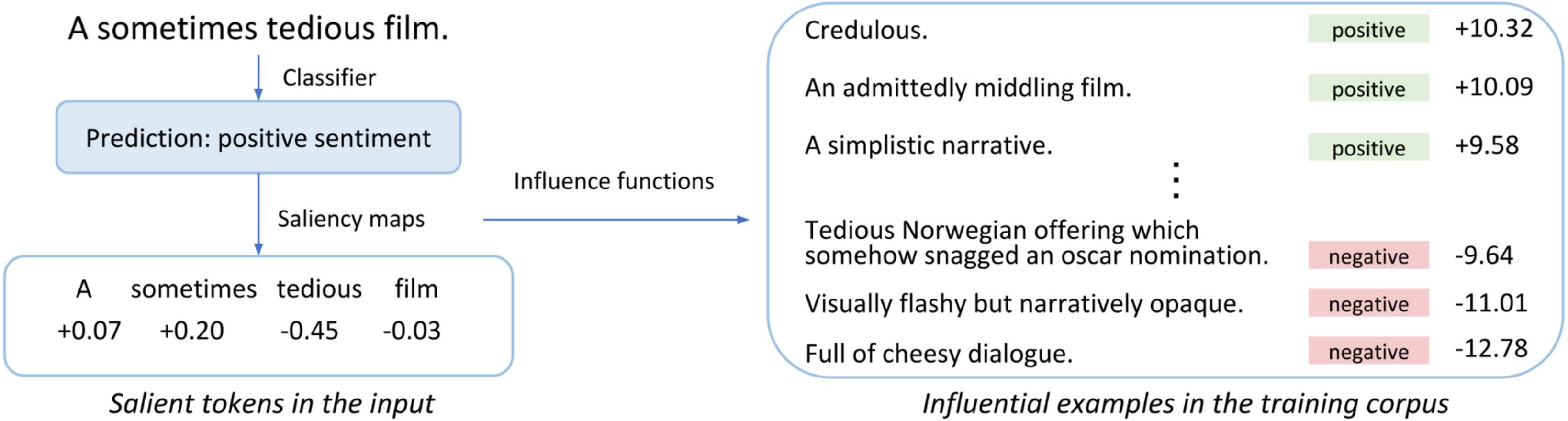


Interpreting Predictions of NLP Models, EMNLP 2020 Tutorial [slides]

Outline

- Attention
- Robustness
- Probing
- Interpreting Individual Neurons
- Saliency Map
- **Influence Function**
- Text Explanation
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

Influential Training Examples



A	sometimes	tedious	film
+0.07	+0.20	-0.45	-0.03

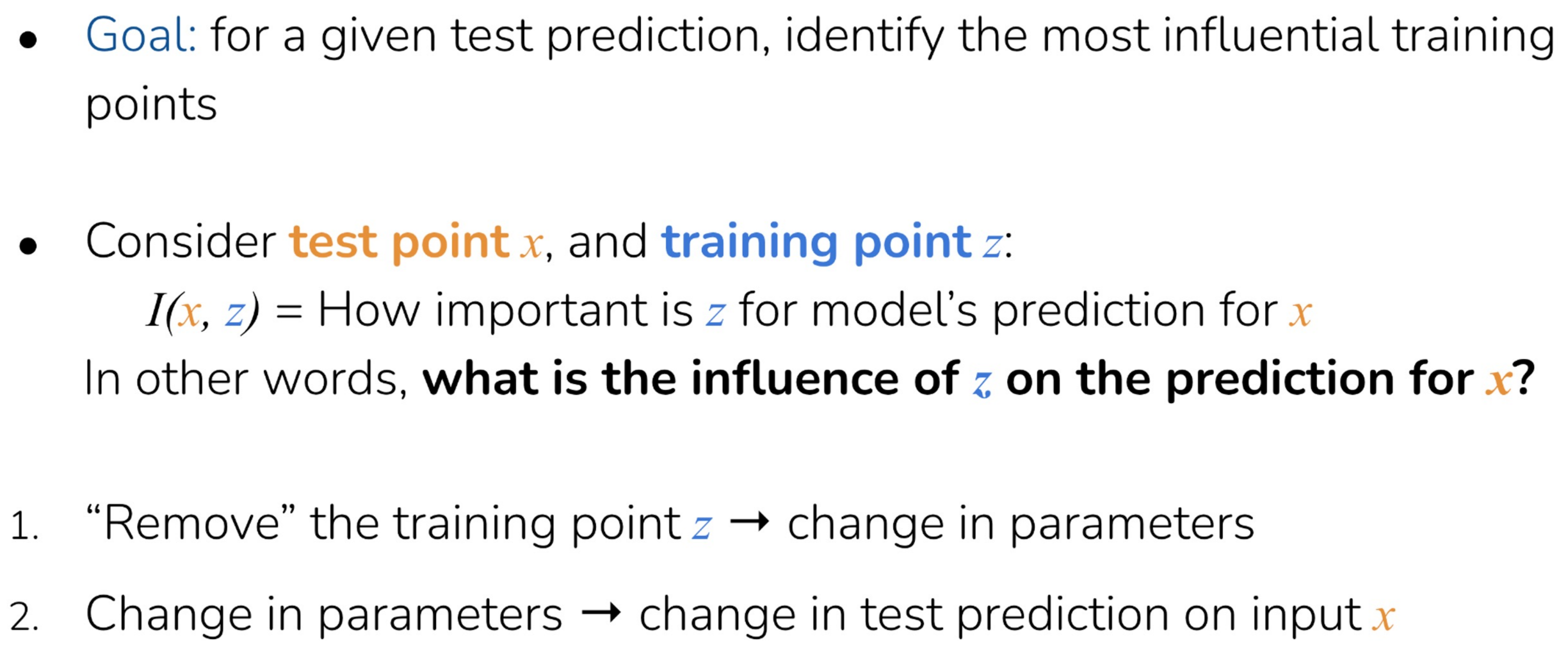
Credulous.	positive	+10.32
An admittedly middling film.	positive	+10.09
A simplistic narrative.	positive	+9.58
Tedious Norwegian offering which somehow snagged an oscar nomination.	negative	-9.64
Visually flashy but narratively opaque.	negative	-11.01
Full of cheesy dialogue.	negative	-12.78

Salient tokens in the input

Influential examples in the training corpus

Figure 1: A sentiment analysis example interpreted by gradient-based saliency maps (left) and influence functions (right). Note that this example is classified incorrectly by the model. Positive saliency tokens and highly influential examples may suggest why the model makes the wrong decision; tokens and examples with negative saliency or influence scores may decrease the model's confidence in making that decision.

Influence Functions (Koh and Liang, 2017)



- Goal: for a given test prediction, identify the most influential training points

- Consider **test point** x, and **training point** z:

  $I(x, z)$ = How important is z for model's prediction for x

  In other words, **what is the influence of z on the prediction for x?**

1. "Remove" the training point z → change in parameters
2. Change in parameters → change in test prediction on input x

Interpreting Predictions of NLP Models, EMNLP 2020 Tutorial [slides]

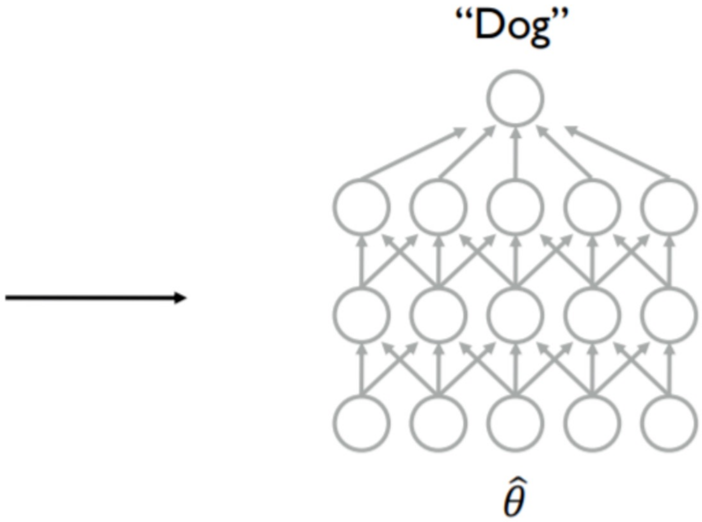
Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

Fish

Dog

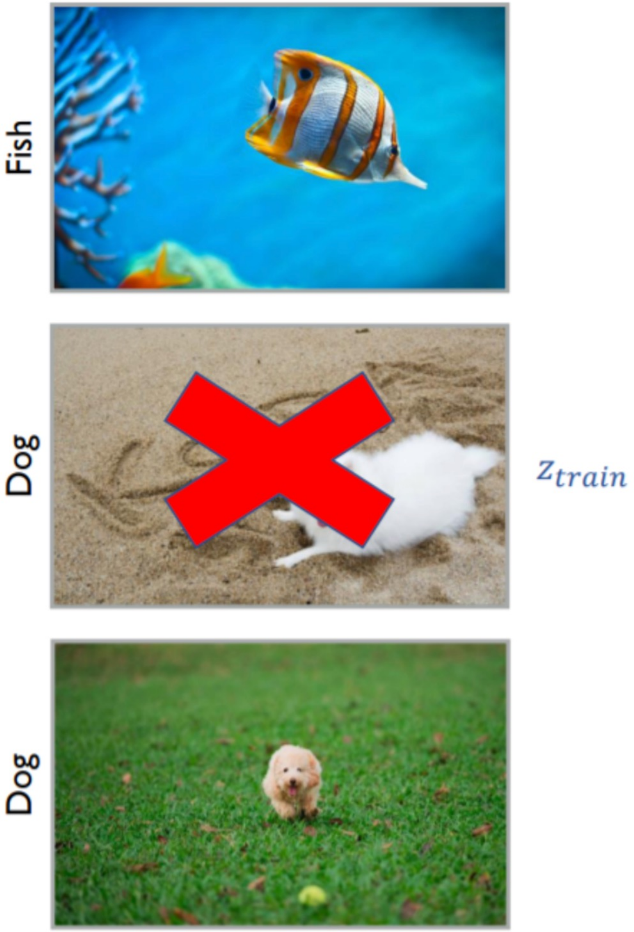
Dog

Training data $z_1, z_2, \dots, z_n$

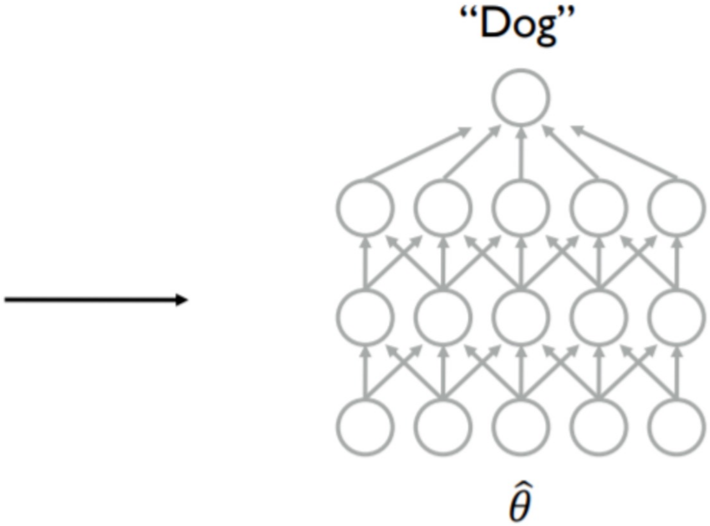


Slide: Pang Wei Koh

Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$



Training data $z_1, z_2, \dots, z_n$



Pick $\hat{\theta}$ to minimize $\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta)$

Pick $\hat{\theta}_{-z_{train}}$ to minimize

$$\frac{1}{n}\sum_{i=1}^{n} L(z_i, \theta) - \frac{1}{n} L(z_{train}, \theta)$$

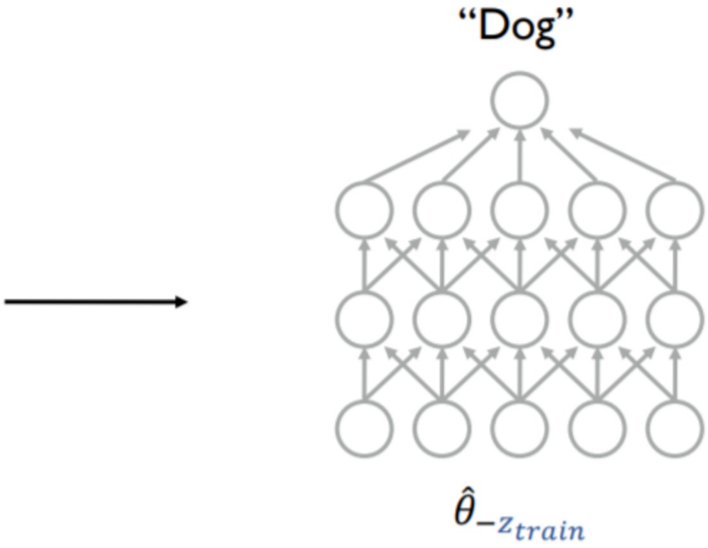
Fish

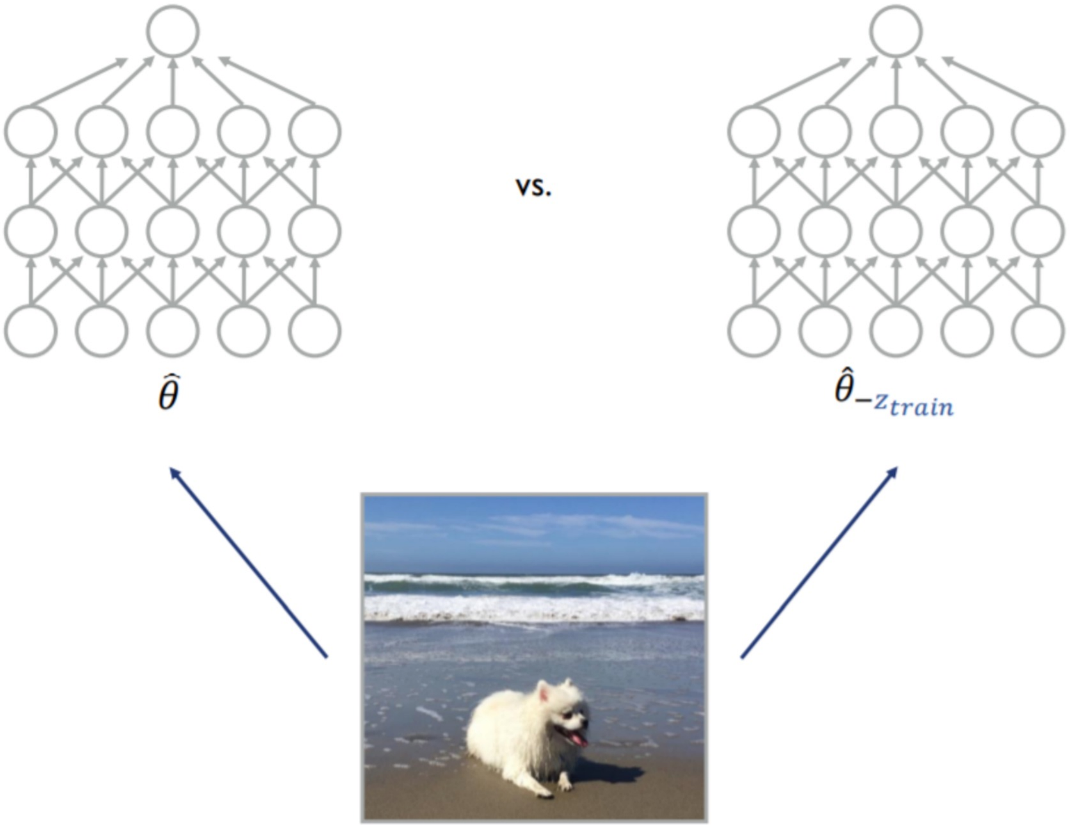
Dog

z_{train}

Dog

Training data $z_1, z_2, \dots, z_n$





“Dog” (82% confidence)

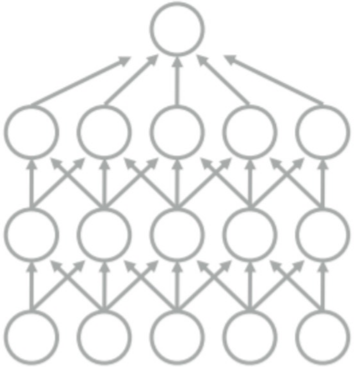
“Dog” (79% confidence)

vs.

$\hat{\theta}$

$\hat{\theta}_{-z_{train}}$

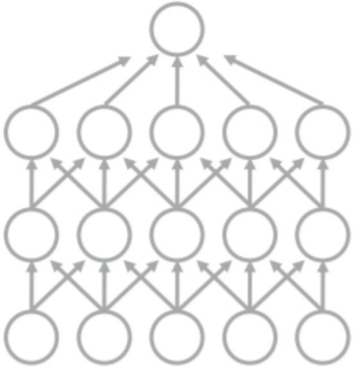
Test input z_{test}



“Dog” (82% confidence)

$\hat{\theta}$

vs.



“Dog” (79% confidence)

$\hat{\theta}_{-z_{train}}$

What is $L(z_{test}, \hat{\theta}_{-z_{train}}) - L(z_{test}, \hat{\theta})$?

Influence Functions in Deep Learning Are Fragile (Basu et al., 2021)

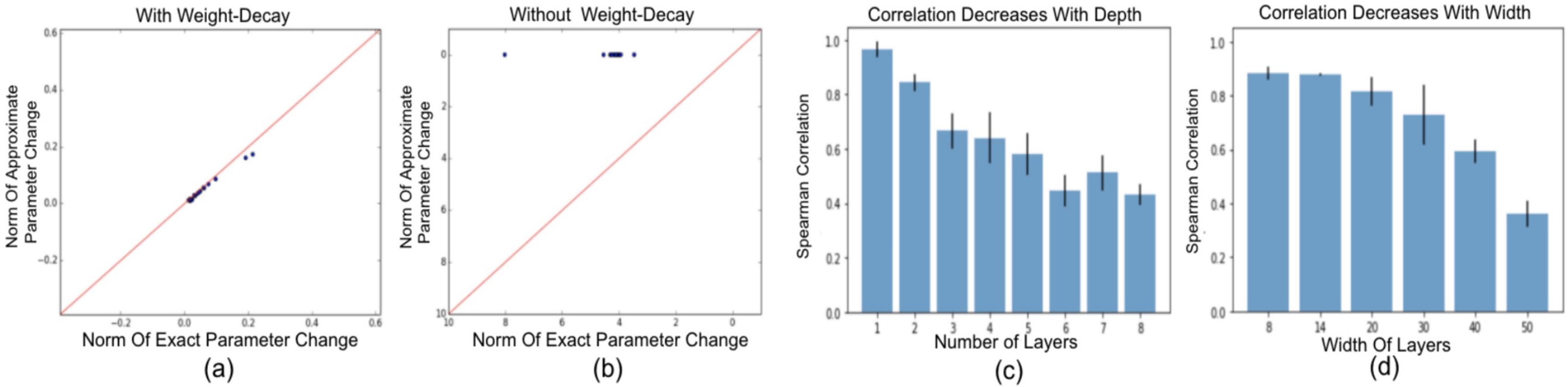
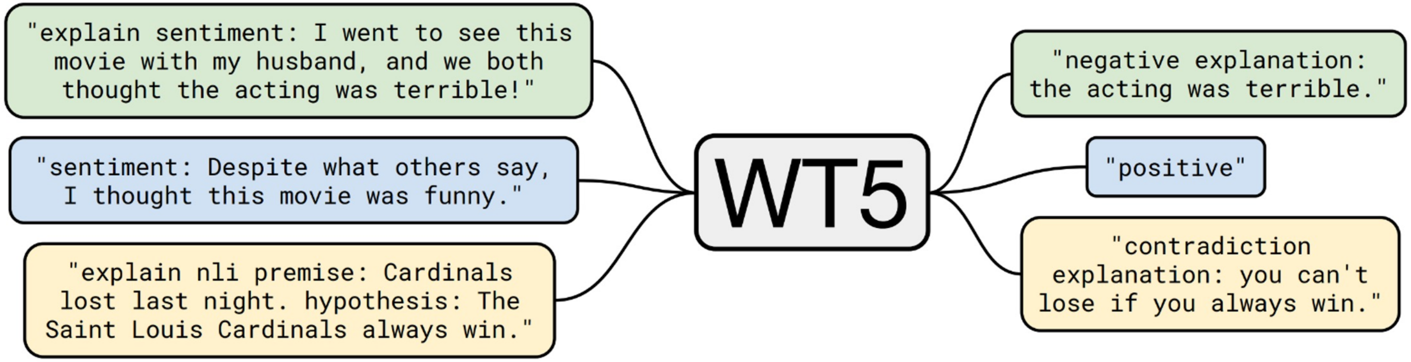


Figure 1: Iris dataset experimental results - (a,b) Comparison of norm of parameter changes computed with influence function vs re-training; (a) trained with weight-decay; (b) trained without weight-decay. (c) Spearman correlation vs. network depth. (d) Spearman correlation vs. network width.

Outline

- Attention
- Robustness
- Probing
- Interpreting Individual Neurons
- Saliency Map
- Influence Function
- **Text Explanation**
- Causal Inference for Interpretation
- Mathematical Frameworks for Transformers

WT5?! Training Text-to-Text Models to Explain their Predictions (Narang et al., 2020)



"explain sentiment: I went to see this movie with my husband, and we both thought the acting was terrible!"

"sentiment: Despite what others say, I thought this movie was funny."

"explain nli premise: Cardinals lost last night. hypothesis: The Saint Louis Cardinals always win."

"negative explanation: the acting was terrible."

"positive"

"contradiction explanation: you can't lose if you always win."

Figure 2: Diagram of our method for training a text-to-text model to explain its predictions. We train the model to generate an explanation when the text "explain" is prepended to the input. The model can still be trained for classification (without an explanation) simply by omitting the "explain" keyword. This approach is readily applicable to sentiment analysis, natural language inference (NLI), and other text tasks.

WT5?! Training Text-to-Text Models to Explain their Predictions (Narang et al., 2020)

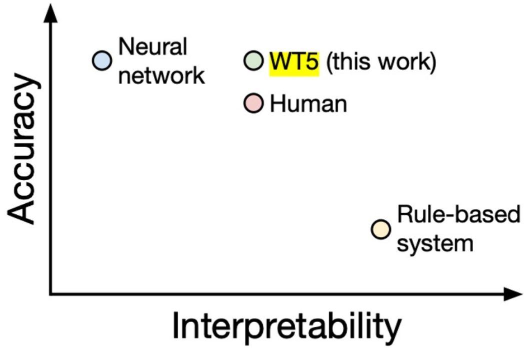


Figure 1: Illustration of our perspective on the accuracy and interpretability of different models. Neural networks (blue) can attain superhuman performance, but are notoriously hard to interpret. A rule-based system (yellow) is easy to interpret but rarely performs well on difficult tasks. Humans (red) are reasonably accurate and provide some degree of interpretability by being able to verbally explain their predictions. In this work, our model (green) is trained both to be highly accurate (in some cases, more accurate than a human) and provide explanations for its predictions as humans do.