Is "Attention = Explanation"? Past, Present & Future

Sarthak Jain & Sarah Wiegreffe
Big Picture Workshop, Dec. 2023

aws
Allen Institute for AI
UNIVERSITY of WASHINGTON

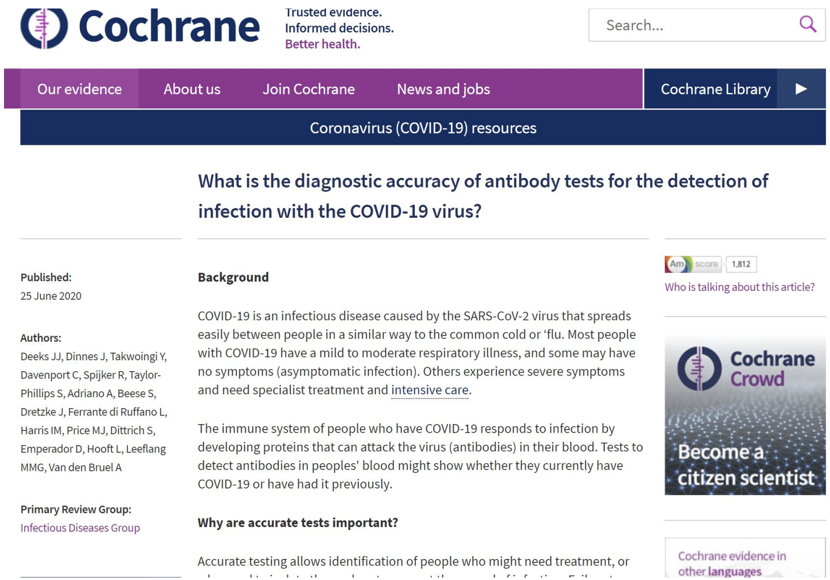
Talk Outline

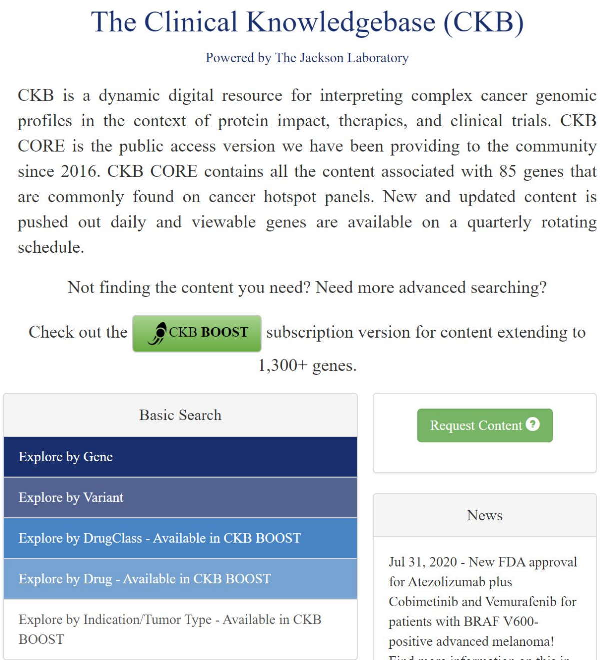
1. Introduction & why we studied this problem
2. Attention is not Explanation
3. Attention is not not Explanation
4. Current & Future Relevance (let’s talk about transformers)

Part 1: Introduction & why we studied this problem

Why was this question interesting to Sarthak?

Help Curators find relevant parts of document





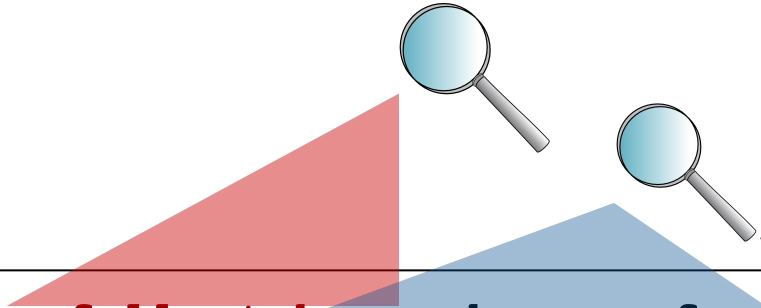
http://evidence-inference.ebm-nlp.com/

aka.ms/hanover

Why was this question interesting to Sarah?

E849.0: Home accidents

801.26: ...subdural, and extradural hemorrhage...



...who sustained a fall at home she was found to have a large acute on chronic subdural hematoma with extensive midline shift...

Mullenbach, Wiegreffe, Duke, Sun, and Eisenstein. [NAACL 2018.]

A Generic Classification Setup

In group A, lower peak (median) plasma levels of procalcitonin (0.2 versus 1.4, p < 0.001), IL 8 (5.6 versus 94.8, p < 0.001), IL 10 (47.2 versus 209.7, p = 0.001), endothelial leukocyte adhesion molecule-1 (88.5 versus 130.6, p = 0.033), intercellular adhesion molecule-1 (806.7 versus 1,375.7, P = 0.001) and troponin-I (0.22 versus 0.66, p = 0.018) were found. There was no significant difference in IL 6, IL-6r and C-reactive protein values between groups. Higher figures of the cardiac index (p = 0.010) along with reduced systemic vascular resistance (p = 0.005) were noted in group A.

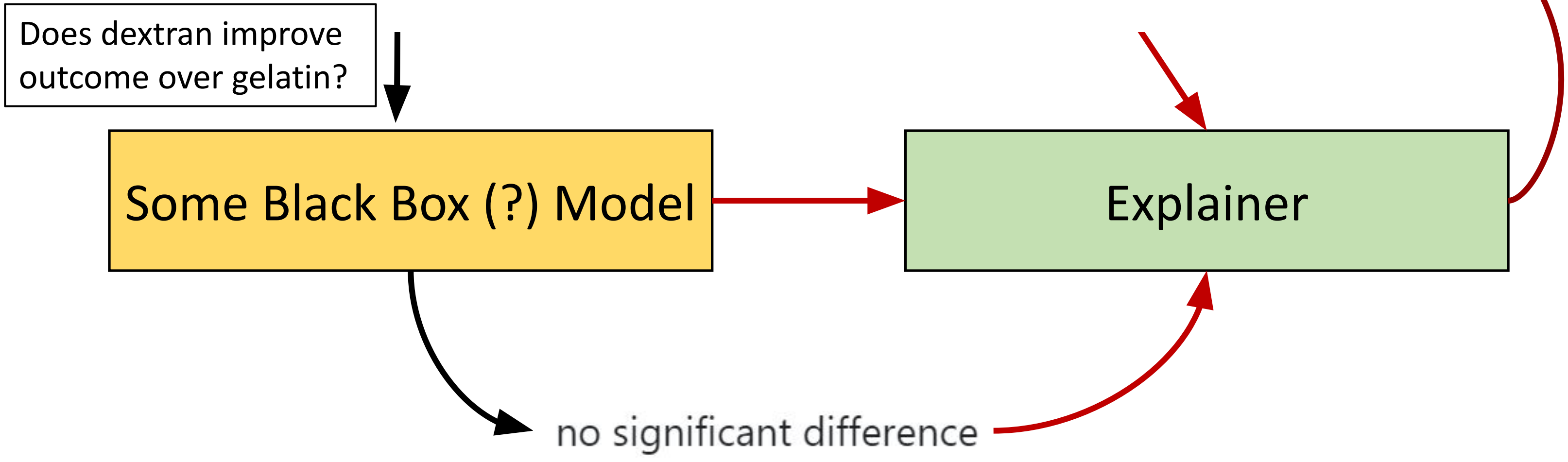
Does dextran improve outcome over gelatin?

Some Black Box (?) Model

no significant difference

A Generic Classification Setup
(with Heatmap based Explanation)

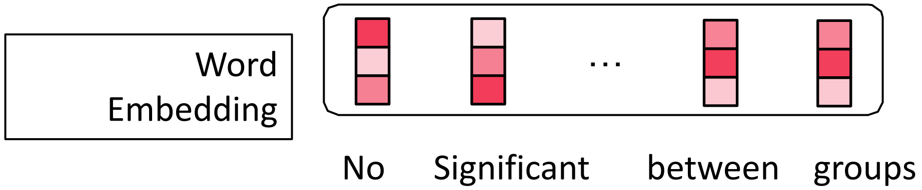
In group A, lower peak (median) plasma levels of procalcitonin (0.2 versus 1.4, $p < 0.001$), IL 8 (5.6 versus 94.8, $p < 0.001$), IL 10 (47.2 versus 209.7, $p = 0.001$), endothelial leukocyte adhesion molecule-1 (88.5 versus 130.6, $p = 0.033$), intercellular adhesion molecule-1 (806.7 versus 1,375.7, $P = 0.001$) and troponin-I (0.22 versus 0.66, $p = 0.018$) were found. There was no significant difference in IL 6, IL-6r and C-reactive protein values between groups. Higher figures of the cardiac index ($p = 0.010$) along with reduced systemic vascular resistance ($p = 0.005$) were noted in group A.

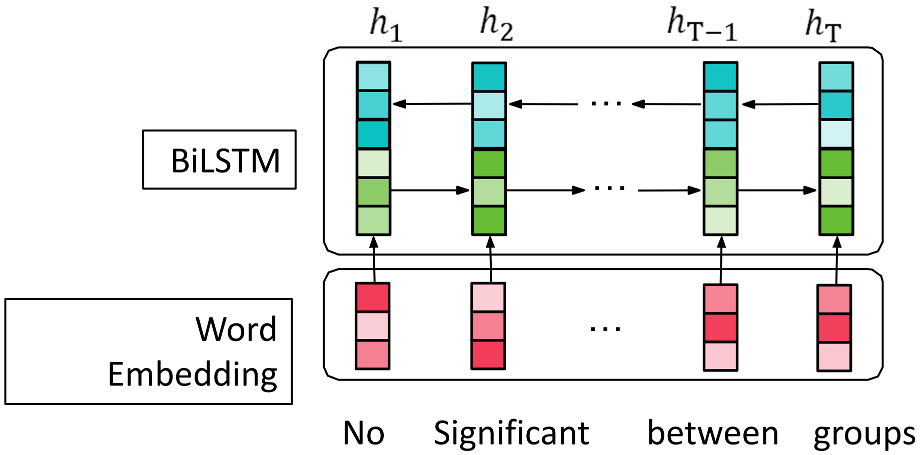


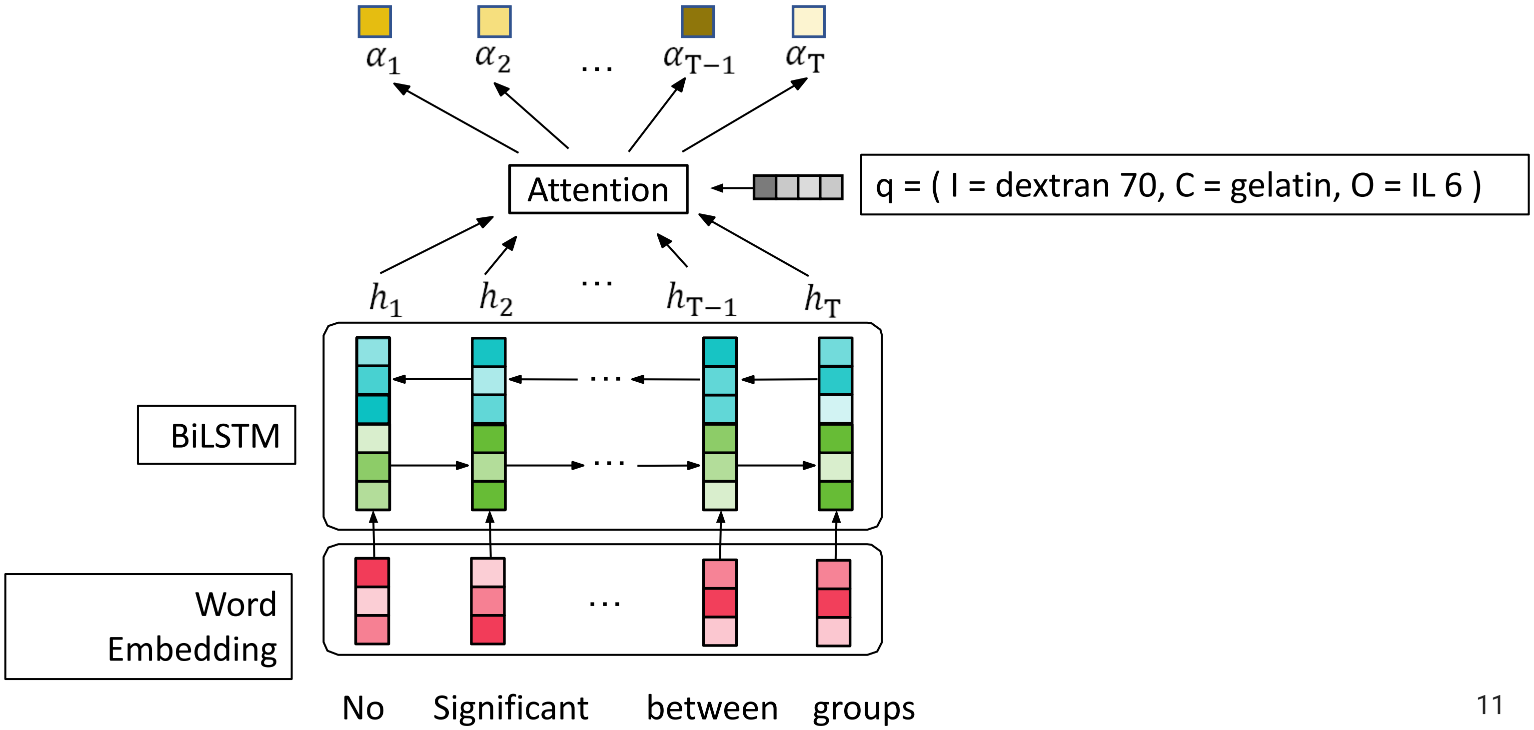
Neural Attention

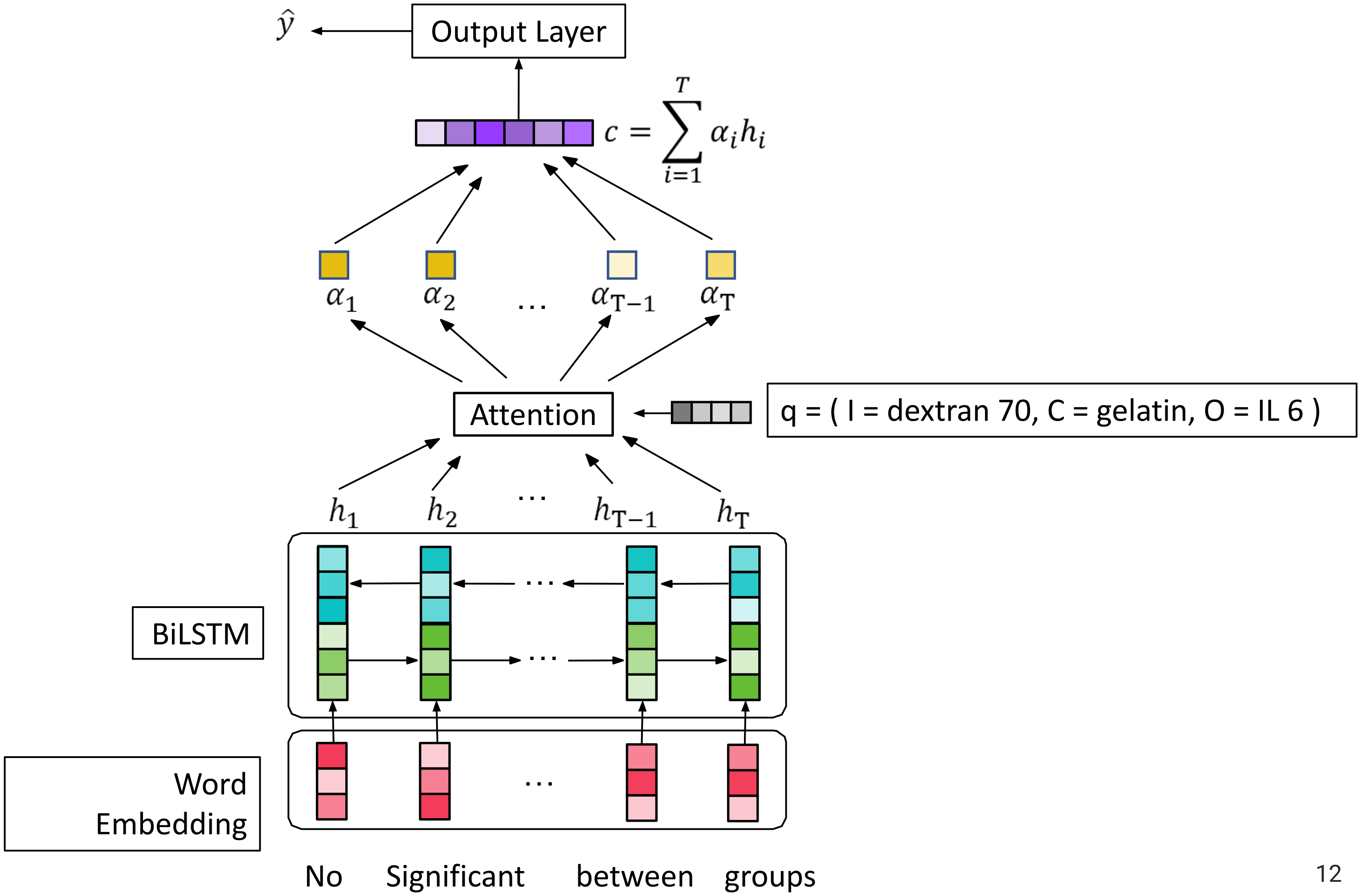
A stop sign is on a road with a mountain in the background.

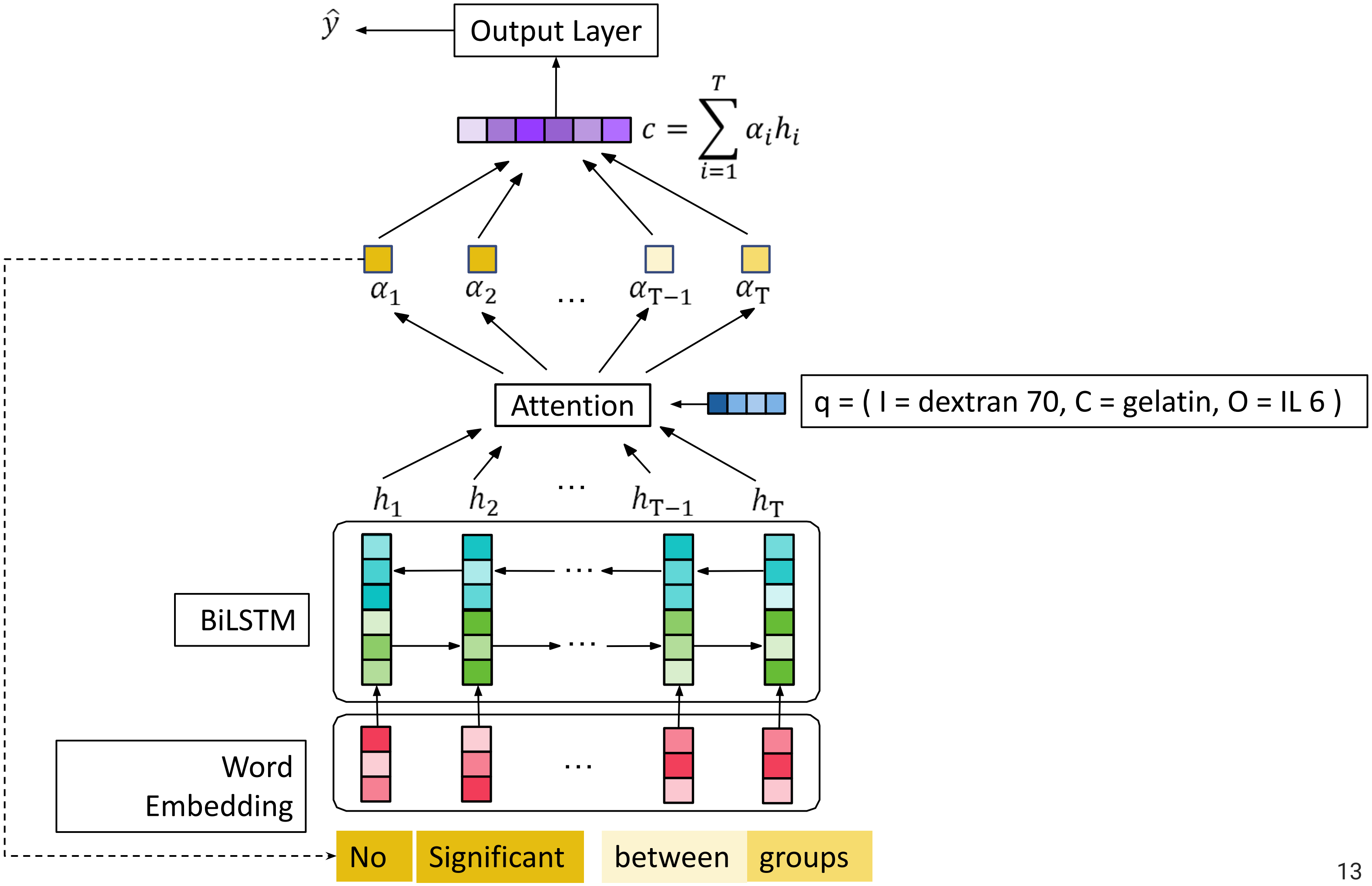
In group A, lower peak (median) plasma levels of procalcitonin (0.2 versus 1.4, p < 0.001), IL 8 (5.6 versus 94.8, p < 0.001), IL 10 (47.2 versus 209.7, p = 0.001), endothelial leukocyte adhesion molecule-1 (88.5 versus 130.6, p = 0.033), intercellular adhesion molecule-1 (806.7 versus 1,375.7, P = 0.001) and troponin-I (0.22 versus 0.66, p = 0.018) were found. There was no significant difference in IL 6, IL-6r and C-reactive protein values between groups. Higher figures of the cardiac index (p = 0.010) along with reduced systemic vascular resistance (p = 0.005) were noted in group A.


Word
Embedding
No
Significant
...
between
groups


h_1
h_2
h_{T-1}
h_T
BiLSTM
Word Embedding
No
Significant
between
groups


α_1 α_2 … α_{T-1} α_T
Attention
q = (I = dextran 70, C = gelatin, O = IL 6)
h_1 h_2 … h_{T-1} h_T
BiLSTM
Word Embedding
No Significant between groups


$\hat{y}$
Output Layer
$c = \sum_{i=1}^{T} \alpha_i h_i$
α_1
α_2
...
α_{T-1}
α_T
Attention
q = (I = dextran 70, C = gelatin, O = IL 6)
h_1
h_2
...
h_{T-1}
h_T
BiLSTM
Word Embedding
No
Significant
between
groups



$\hat{y}$
Output Layer
$c = \sum_{i=1}^{T} \alpha_i h_i$
α_1
α_2
...
α_{T-1}
α_T
Attention
q = (I = dextran 70, C = gelatin, O = IL 6)
h_1
h_2
...
h_{T-1}
h_T
BiLSTM
Word Embedding
No
Significant
between
groups

Unclear Questions

What does Attention heatmap tell us – How "important" a word is?

Is there really a 1:1 mapping between Attention and input tokens?

Does Attention tell us how a model reached its prediction?

Part 2: Attention is not Explanation

Jain, S., & Wallace, B.C. (2019). **Attention is not Explanation**. *NAACL-HLT*.

Empirical Questions

1. Do Attention weights correlate with existing feature importance measures (gradients and leave-one-out) ?

2. **Uniqueness:** Had we attended to different inputs, would the prediction have been different ?

Tasks and Datasets

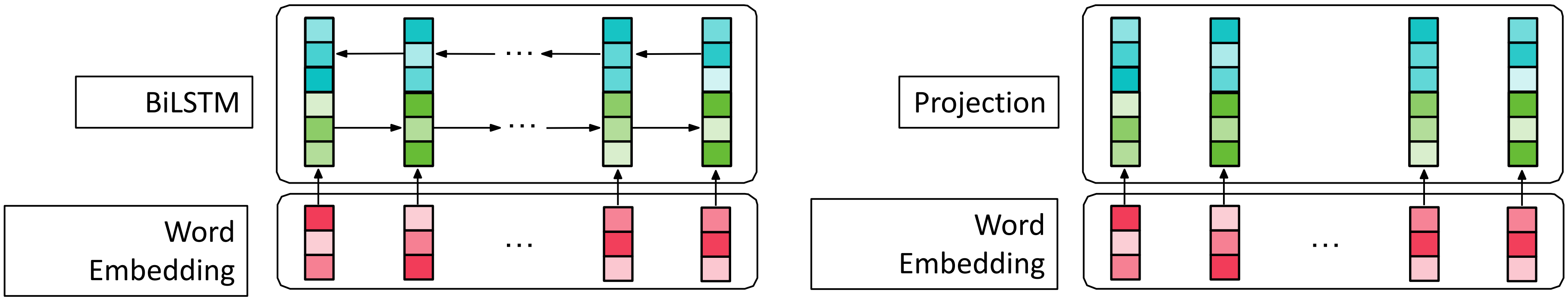
- **Binary Classification**
  - Sentiment Classification – Stanford Sentiment Treebank, IMDB
  - Topic Classification – 20NewsGroup, AGNews
  - Diagnosis (MIMIC-III) – **Diabetes**, Anemia
  - Twitter – Adverse Drug Reaction

- **Multiple Choice Question Answering**
  - CNN News, bAbI

- **Entailment**
  - SNLI

Encoder Models

- We aim to evaluate whether Attention weights provide transparency, under different encoders consistently



Empirical Questions

1. Do Attention weights correlate with existing feature importance measures (gradients and leave-one-out) ?

2. **Uniqueness:** Had we attended to different inputs, would the prediction have been different ?

Feature Importance – Experiments

- Rank Correlation (Kendall-Tau) between Attention Scores and Feature Importance Measures (gradients and leave-one-out)

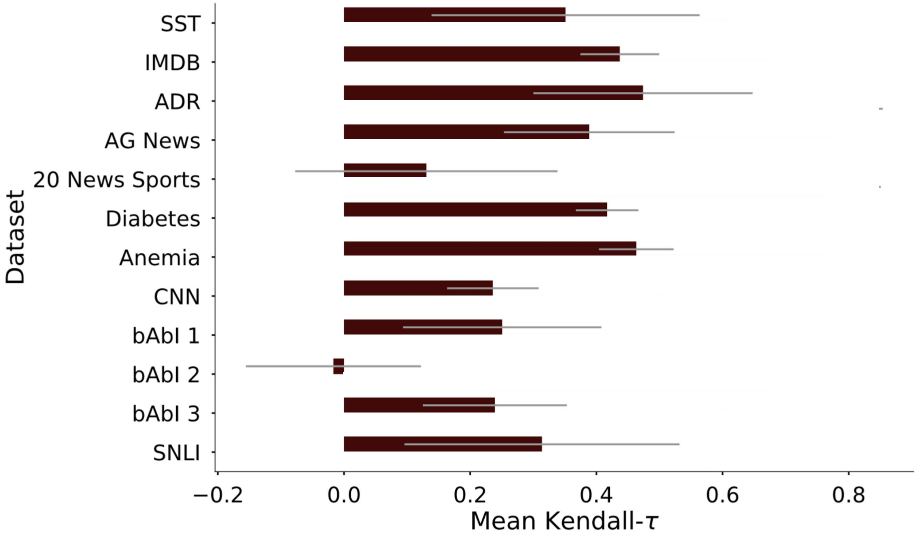
- 0 = no correlation, 1 = perfect correlation

- Total Variation Distance: for comparing class predictions between 2 models

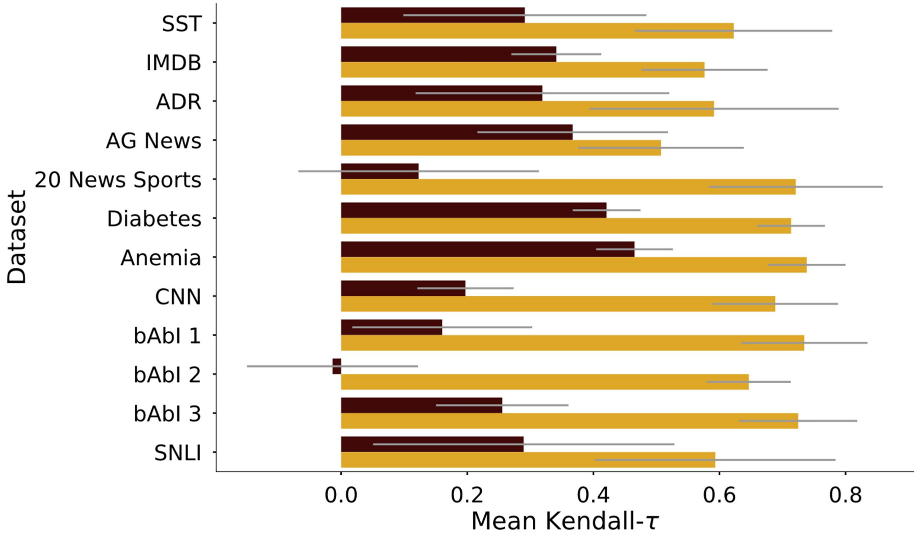
$$\mathrm{TVD}(\hat{y}_1, \hat{y}_2) = \frac{1}{2} \sum_{i=1}^{|\mathcal{Y}|} |\hat{y}_{1i} - \hat{y}_{2i}|$$

Feature Importance – Results

GRADIENTS VS ATTENTION



LEAVE-ONE-OUT VS ATTENTION



BiLSTM Projection

Empirical Questions

1. Do Attention weights correlate with existing feature importance measures (gradients and leave-one-out) ?

2. **Uniqueness:** Had we attended to different inputs, would the prediction have been different ?

Counterfactual Experiments

Empirical questions to measure Uniqueness:

How much on average does an output change if we randomly permute Attention scores?

Can we find maximally different Attention that doesn't change the output by more than some epsilon?

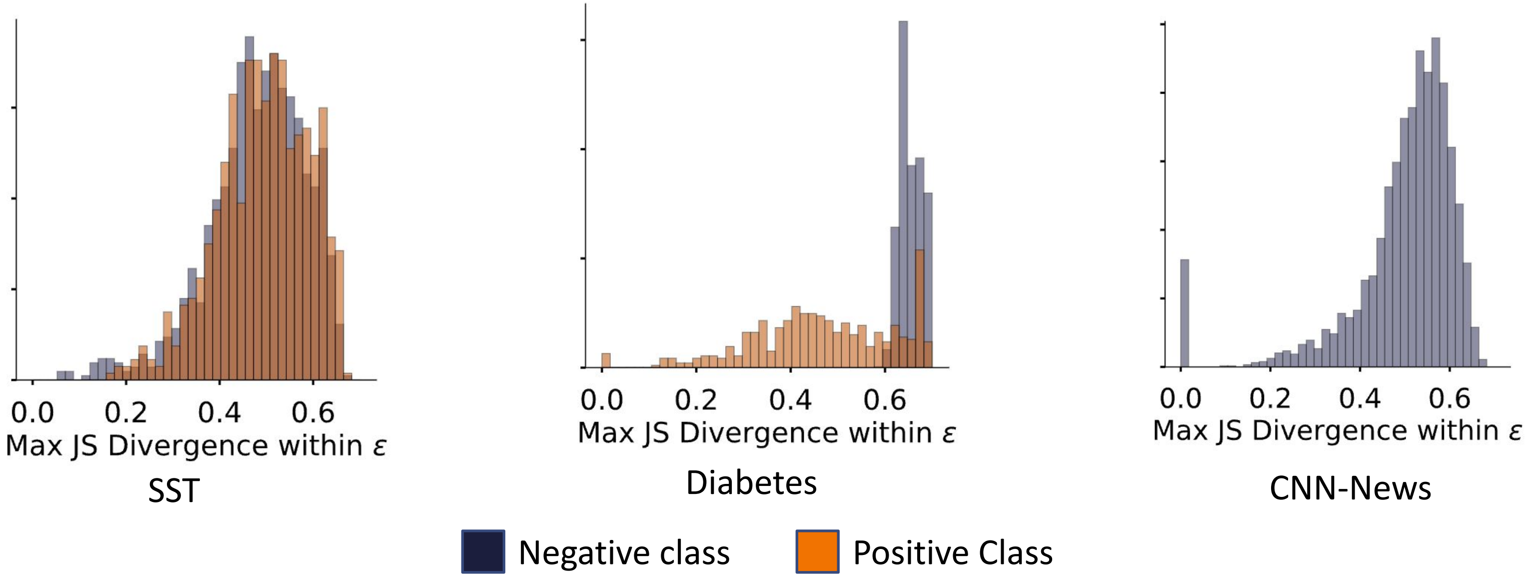
Adversarial Attention – Experiment

For each example -

Find k Attention distributions that are → $\text{argmax over } \{\alpha^{(1)}, \ldots, \alpha^{(k)}\}$

1. Maximally different from observed Attention $\hat{\alpha}$, using JS Divergence. → $$\sum_{i=1}^{k} JSD(\alpha^{(i)}, \hat{\alpha}) +$$
2. Maximally different from each other → $$\frac{1}{k(k-1)} \sum_{j>i} JSD(\alpha^{(i)}, \alpha^{(j)})$$
3. Doesn't change output by more than epsilon (= 0.001). → $$s.t.\ TVD(y^{(i)}, \hat{y}) < \epsilon, \forall i \in \{1, \ldots, k\}$$

Adversarial Attention – Results (BiLSTM)



Original: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film

Adversarial: reggio falls victim to relying on the very digital technology that he fervently scorns creating a meandering inarticulate and ultimately disappointing film $\Delta\hat{y}$: *0.005*

Conclusions

Correlation between Attention and Feature Importance scores are often low

Attention distributions do not uniquely characterize *why* a model made a given prediction; alternative heatmaps would have yielded the same output

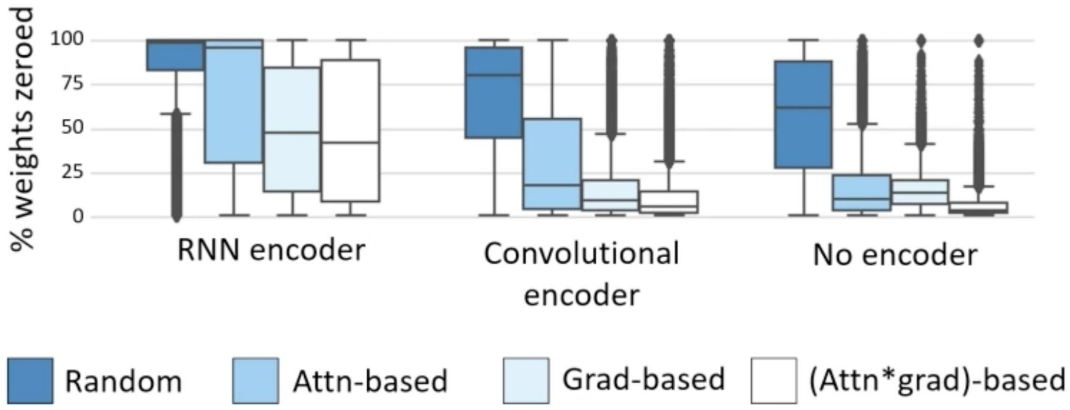
Takeaway

- Attention do not provide clear and consistent interpretation of why a model made a prediction.

- We should question what the author is trying to convey with the heatmap.

Concurrent Relevant Work: Serrano & Smith

- Focused on whether Attention provides relative importance of hidden states themselves
- How quickly does Attention flip when zeroing out attention scores according to their rank?




PUBLISHED
PUBLISHED
PUBLISHED

Part 3: Attention is not *not* Explanation

Wiegreffe, S.*, & Pinter, Y.* (2019). Attention is not not Explanation. *EMNLP*.

Blogpost #1

Attention is not not Explanation

Yuval Pinter · Follow
8 min read · Apr 21, 2019

303 1

[**Update, August 13 — December 6, 2019**: Sarah Wiegreffe and I performed experiments to follow up on the points here, as well as constructive setups for detecting and claiming faithful explanation, presented at EMNLP 2019. The paper is available here.

Byron Wallace responded to the paper here.]

[This post is intended for an NLP practitioner audience, and assumes its readers know what attention modules are and how they are being used. **All feedback is welcome, either here or to *uvp@gatech.edu* or to *@yuvalpi* on Twitter.**]

An upcoming NAACL paper was uploaded to arXiv earlier this month, and has been making the rounds on social media. The title chosen for it was Attention is not Explanation; the authors are Sarthak Jain and Byron C. Wallace (from here on I will refer to them, and the paper, as J&W). Such a title sets high expectations for a rigorous, convincing proof of the claim. In this post I argue that it does not deliver on them.

Briefly, my main points are:

link

Main Arguments

1. Explanation can be **many things**
2. Rank Correlation is not always appropriate + missing baselines
3. Counterfactual Distributions are **not Counterfactual Weights**
    a. Attention distribution is **not a primitive**

link

Explanation can be many things

- Explainability = **both** post-hoc rationalizations and faithful "interpretability".
- Human explanation is post-hoc
    - invent a story that plausibly justifies our actions, even if it not an entirely accurate reconstruction

link

Counterfactual Distributions are *not* Counterfactual Weights

- Detaching attention scores from the attention mechanism degrades the model itself.
  - Attention scores are not assigned arbitrarily by the model.
  - Jain & Wallace removed the linkage that motivates the original claim of attention distribution explainability.
- Adversarial search was *per-instance*
- Too high degree of freedom

link

Blogpost #2: Response to Sarah/Yuval

“Attention is not Explanation” - Assumption or Conclusion?

Strengthening the Feature Importance Correlation Experiments

If Attention distribution is not a primitive, what do heatmaps tell us?

link

Blogpost #2

“Attention is not Explanation” - Assumption or Conclusion?

- Why expect attention to have any identification with input tokens, given contextualization layer?

- We assumed faithfulness as necessary component of any explanation method, but didn’t clarify it enough.

link

Blogpost #2

Strengthening the Feature Importance Correlation Experiments

- Does gradient and Leave-one-out correlate with each other?
- Rank Correlation metrics do not take account magnitudes and long tail can artificially depress the correlation scores.

link

Blogpost #2

If Attention distribution is not a primitive, what do heatmaps tell us?

- Attention model rather than Attention heatmap is the valid primitive - **we agree**. But then why show heat-maps over a handful of examples?

- Multiple valid causes can exist - **we agree**. But does attention tell us which one model used?

link

Attention is not not Explanation

1. Explanation can be **many things**
2. Rank Correlation is not always appropriate + **missing baselines**
3. Counterfactual Distributions are **not Counterfactual Weights**

Attention is not not Explanation

1. Explanation can be **many things**
2. ~~Rank Correlation is not always appropriate +~~ **missing baselines**
3. Counterfactual Distributions are **not Counterfactual Weights**
4. **Random seed variance** as a baseline for adversaries

What is explanation?

Plausible **Explainability**

- **Goal:** increasing user trust, satisfaction, or understanding
- Rationale generation (Ehsan et al. 2019, Riedl 2019)
- **Evaluation:** users

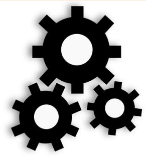
What is explanation?

Plausible Explainability

- **Goal:** increasing user trust, satisfaction, or understanding
- Rationale generation (Ehsan et al. 2019, Riedl 2019)
- **Evaluation:** users

Faithful Explainability

- **Goal:** understanding how models make predictions (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive
- **Evaluation:** not exclusively users



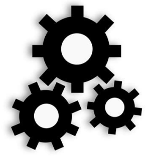
What is explanation?

Plausible Explainability

- **Goal:** increasing user trust, satisfaction, or understanding
- Rationale generation (Ehsan et al. 2019, Riedl 2019)
- **Evaluation:** users

Faithful Explainability

- **Goal:** understanding how models make predictions (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive
- **Evaluation:** not exclusively users



If Attention is (Faithful) Explanation:

1. Attention should be a **necessary component** for good performance

Necessary

If Attention is (Faithful) Explanation:

1. Attention should be a **necessary component** for good performance
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation

Necessary

Hard to manipulate

If Attention is (Faithful) Explanation:

1. Attention should be a **necessary component** for good performance
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation
3. Attention weights should work well in **uncontextualized settings**

Necessary

Hard to manipulate

Work out of context

Selecting Meaningful Tasks

Necessary

1. Attention should be a **necessary component** for good performance

Searching for Adversarial Models

Hard to manipulate

1. Attention should be a **necessary component** for good performance
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation

Adversarial Training

Hard to manipulate

1. Train a base model (*M_b*)
2. Train an adversary (*M_a*) that **minimizes change in prediction scores** from the base model, while *maximizing changes in the learned attention distributions.*

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \mathrm{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \lambda\ \mathrm{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$

Adversarial Training

Hard to manipulate

1. Train a base model (M_b)
2. Train an adversary (M_a) that **minimizes change in prediction scores** from the base model, while *maximizing changes in the learned attention distributions.*

$$\mathcal{L}(\mathcal{M}_a, \mathcal{M}_b)^{(i)} = \mathrm{TVD}(\hat{y}_a^{(i)}, \hat{y}_b^{(i)}) - \boxed{\lambda}\,\mathrm{KL}(\boldsymbol{\alpha}_a^{(\mathbf{i})} \parallel \boldsymbol{\alpha}_b^{(\mathbf{i})})$$

Comparisons

Hard to manipulate

1. Random seed variance
   a. Re-running the **base setup** with multiple random seeds to calibrate what we expect for variance in attention weights
2. Jain & Wallace (2019)
   a. Instance-specific adversarial attention weights
   b. No consistency requirement
   c. No model trained

Result Sample (IMDb)

Hard to manipulate

Base model

Result Sample (IMDb)

Hard to manipulate

Base model

Unconstrained adversary (“not”)

Result Sample (IMDb)

Hard to manipulate

Base model

brilliant and moving performances by tom and peter finch

Unconstrained adversary (“not”)

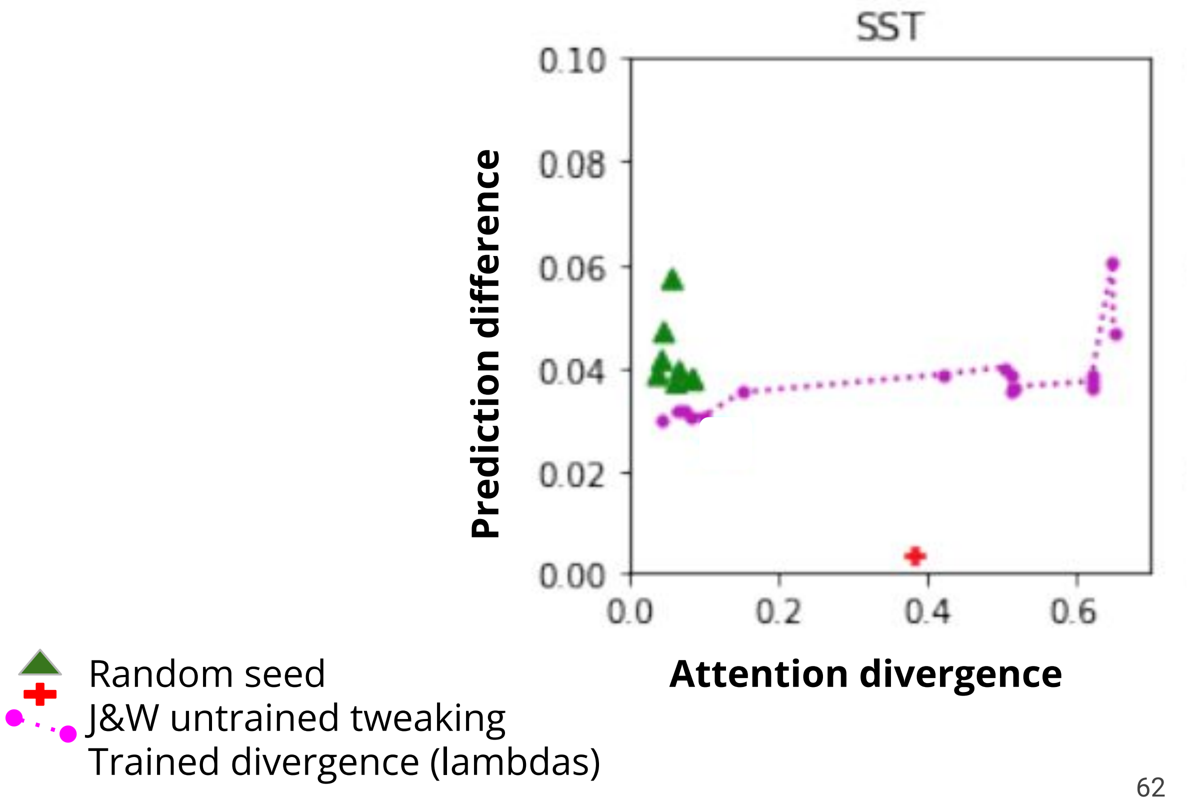
brilliant and moving performances by tom and peter finch

Trained adversary (“not not”)

brilliant and moving performances by tom and peter finch

Adversarial Results

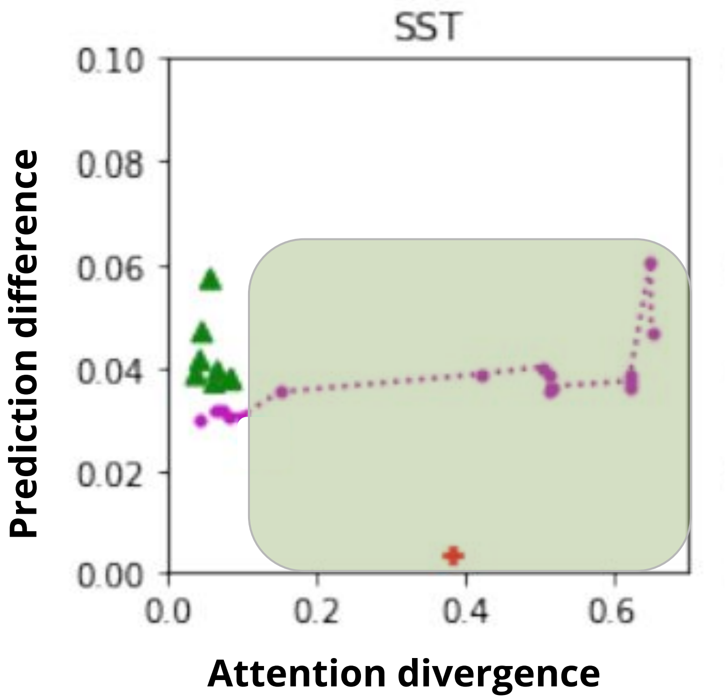
Hard to manipulate



Adversarial Results

Hard to manipulate

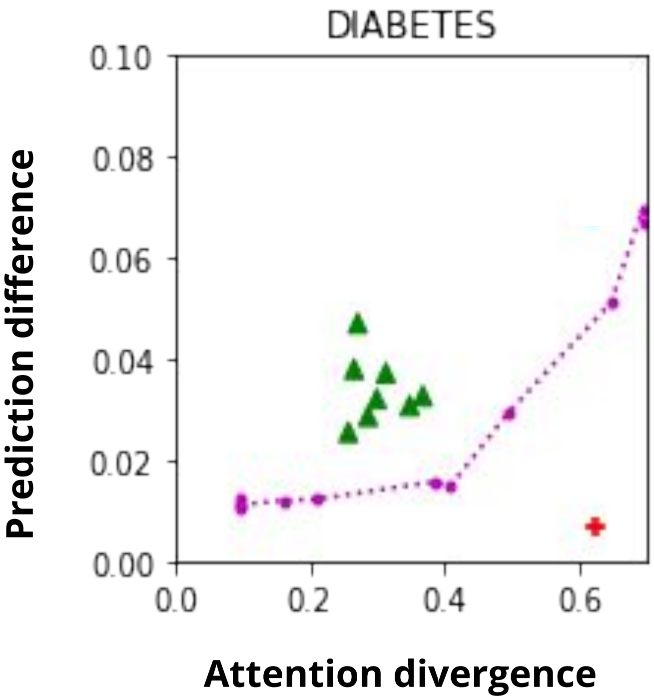
- Slow increase in prediction difference
  - *Does not* support use of attention weights for faithful explanation



Random seed
J&W untrained tweaking
Trained divergence (lambdas)

Adversarial Results

Hard to manipulate

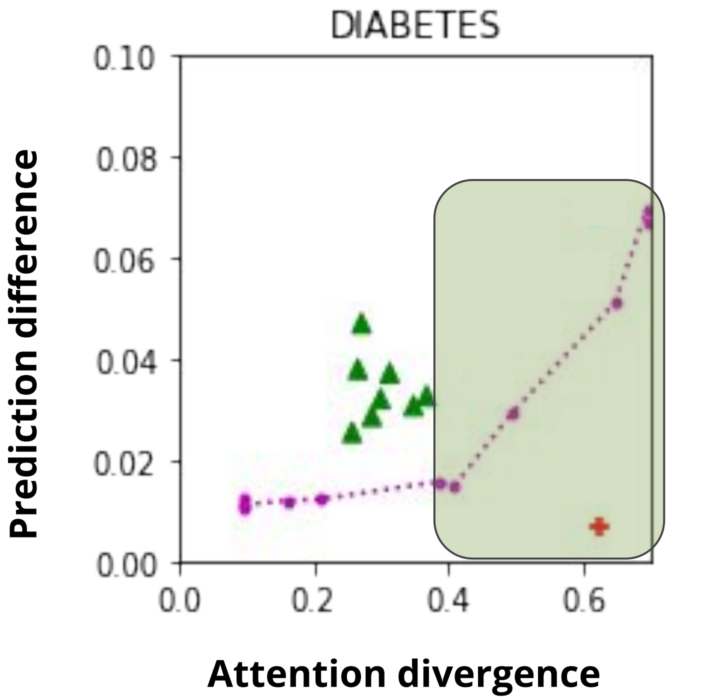


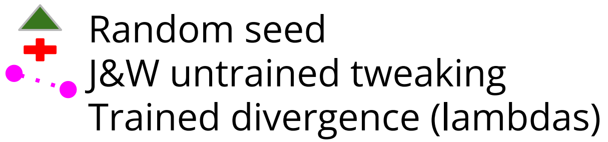
Random seed
J&W untrained tweaking
Trained divergence (lambdas)

Adversarial Results

Hard to manipulate

- Fast increase in prediction difference = attention scores not easily manipulable
    - Supports use of attention weights for faithful explanation

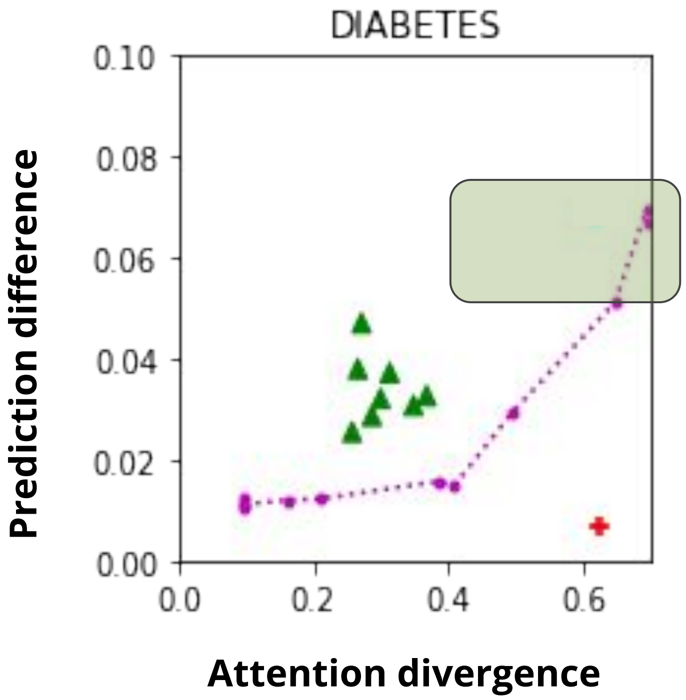




Adversarial Results

Hard to manipulate

- Fast increase in prediction difference = attention scores not easily manipulable
  - Supports use of attention weights for faithful explanation
- **Another interpretation:** y-axis differences are small & random seed variance is high
  - *Does not* support use of attention weights for faithful explanation



Probing Attention

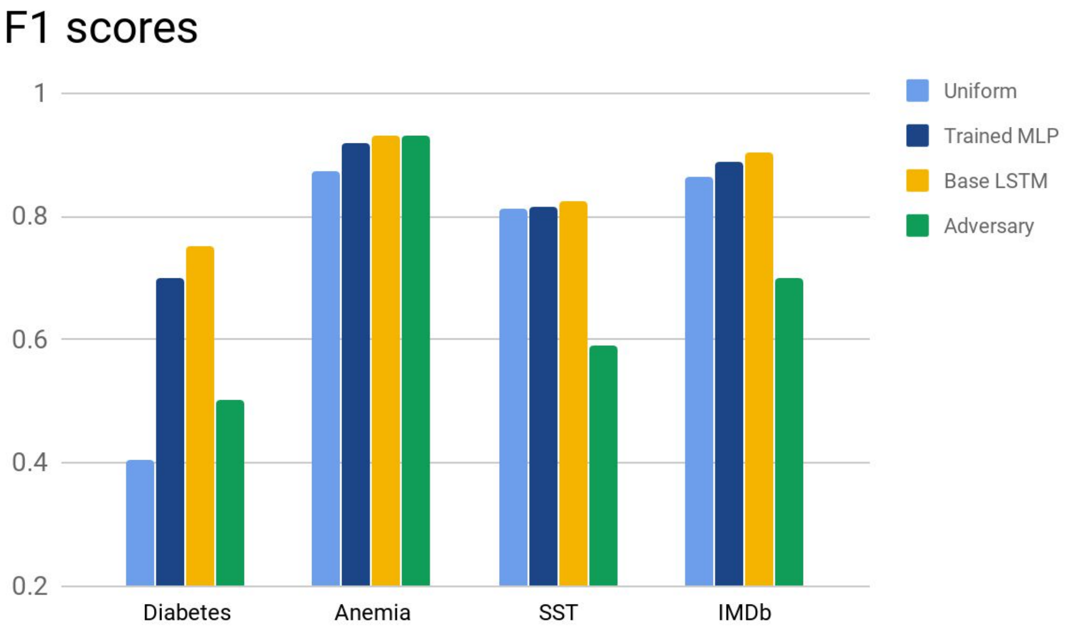
Work out of context

1. Attention should be a **necessary component** for good performance
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation
3. Attention weights should work well in **uncontextualized settings**

Results

Work out of context

- Adversaries' attention scores **don't transfer well.**
- Situation is not nearly as bleak as previously portrayed.



Conclusion

- 3 desiderata of attention for “faithful” explanation

Necessary

Hard to manipulate

Work out of context

Conclusion

- 3 desiderata of attention for “faithful” explanation
- 3 methods to measure the utility of attention distributions for faithful explanation

Necessary	Select Meaningful Tasks
Hard to manipulate	Search for Adversaries
Work out of context	Use Attention as Guide

Conclusion

- 3 desiderata of attention for “faithful” explanation
- 3 methods to measure the utility of attention distributions for faithful explanation
- Results showing performance is **highly task-dependent**

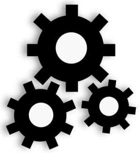
Necessary	Select Meaningful Tasks
Hard to manipulate	Search for Adversaries
Work out of context	Use Attention as Guide

2019 Takeaways

1. Use guides to judge token-output correlation
2. Use adversarial models to investigate exclusivity
3. Calibrate your notion of variance
4. Investigate models & tasks where attention is necessary

We agreed on many things

- We both valued & wanted to investigate ***faithful*** instance-level explanations.



- Both of our search procedures ultimately found adversarial distributions (though with varying levels of success).
- Attention as explanation depends on dataset & model.
- Different (valid) experiments can reach different views on the utility of model internals.

Our Takeaways (now)

1. Faithfulness and plausibility are **different criteria** with distinct merits that **must be evaluated separately.**

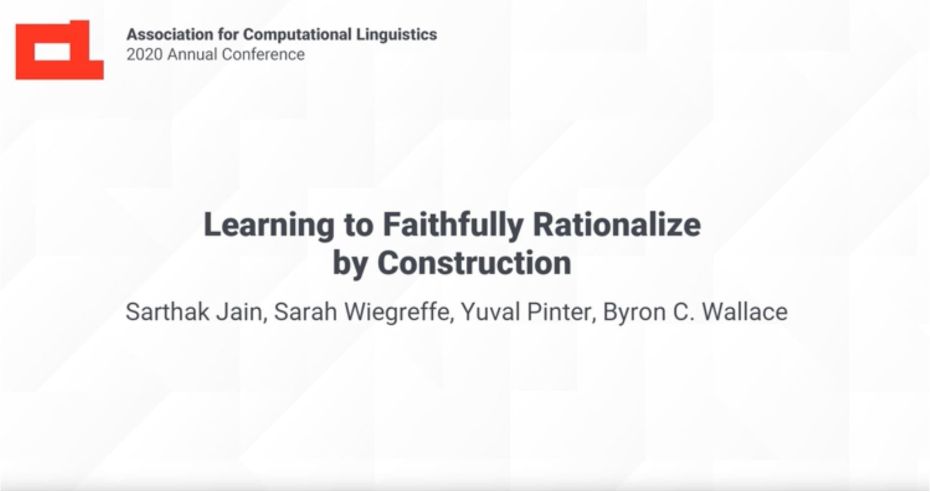
2. Attention mechanisms in LSTM networks can serve as faithful explanation **under certain conditions; there is no one-size-fits-all answer.**

Our Takeaways (now)

3. Faithfulness evaluation is difficult due to **lack of ground-truth.**
    a. Researchers must convince the audience of the meaningfulness of their desiderata.

4. It's important to be careful when drawing analogies between machines and human behavior.
    a. Attention is easy to compute and its qualitative results are cognitively satisfying.

We collaborated on another paper!



- About building faithfulness directly into neural architectures (with BERT)
- Threshold attention to obtain an *explanation* first, then classify.

[link]

Related & Subsequent Work

Checkout survey Is Attention Explanation? An Introduction to the Debate (2022)

Why is Attention not faithful explanation?

(Grimsley et al. 2020, Sun & Lu 2020)

Do our results generalize to other NLP tasks?

(Vashishth et al. 2019, Pruthi et al. 2020)

How to evaluate faithfulness?

(Jacovi & Goldberg 2020)

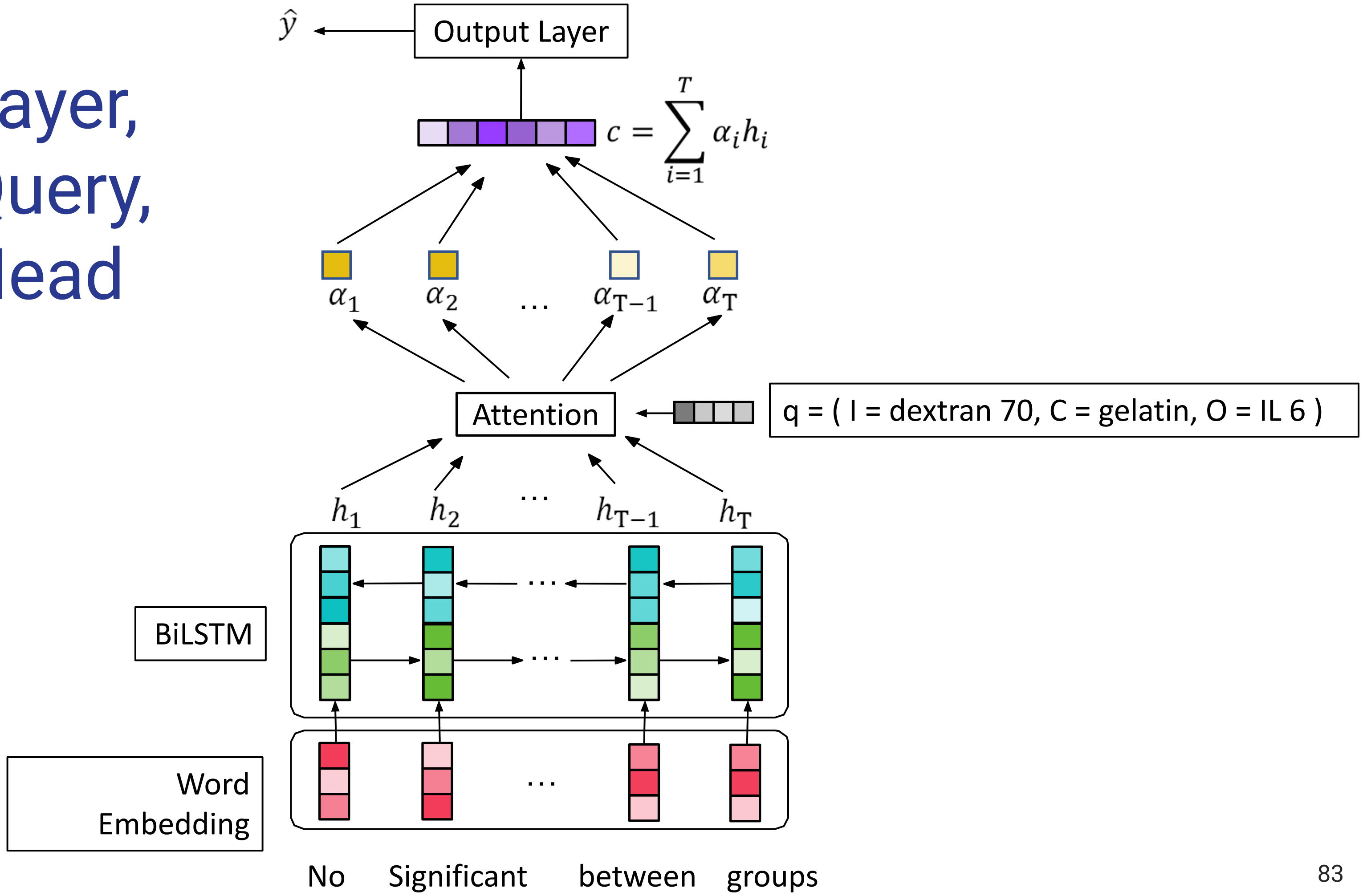
How to improve faithfulness?

(Mohankumar et al. 2020, Tutek & Snajder 2020)

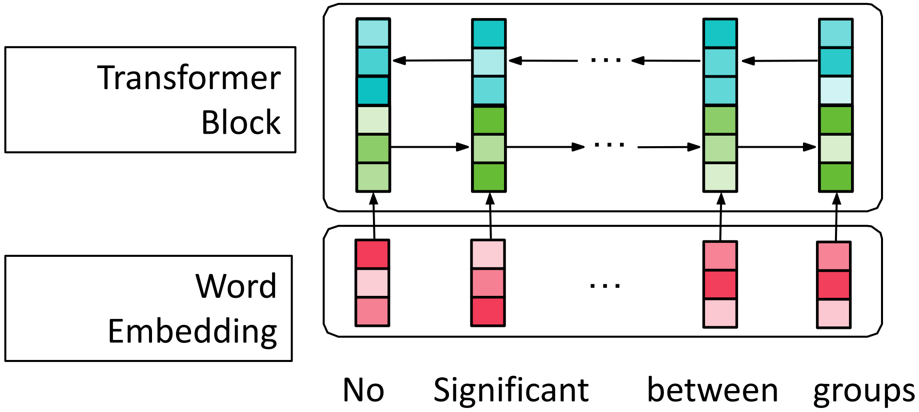
Part 4: Current & Future Relevance

(let’s talk about transformers)

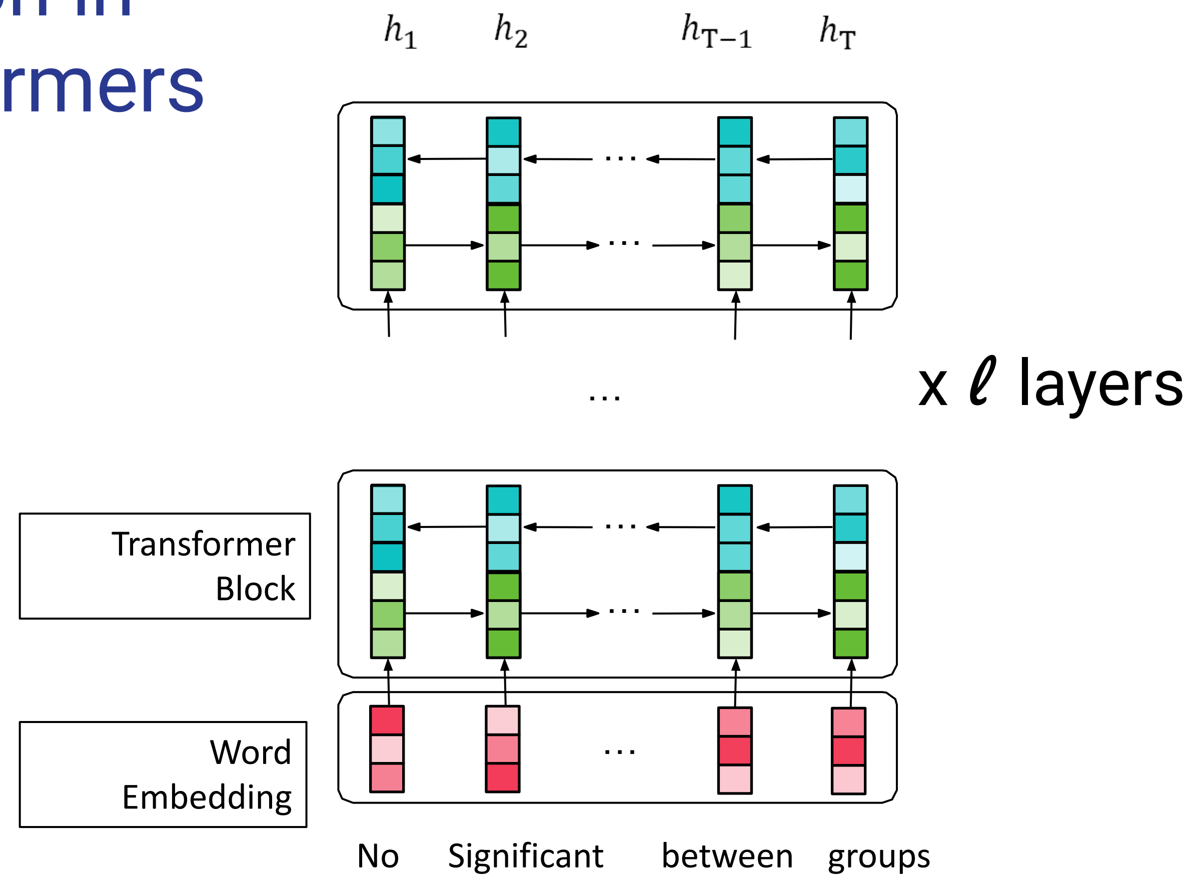
Single Layer, Single Query, Single Head



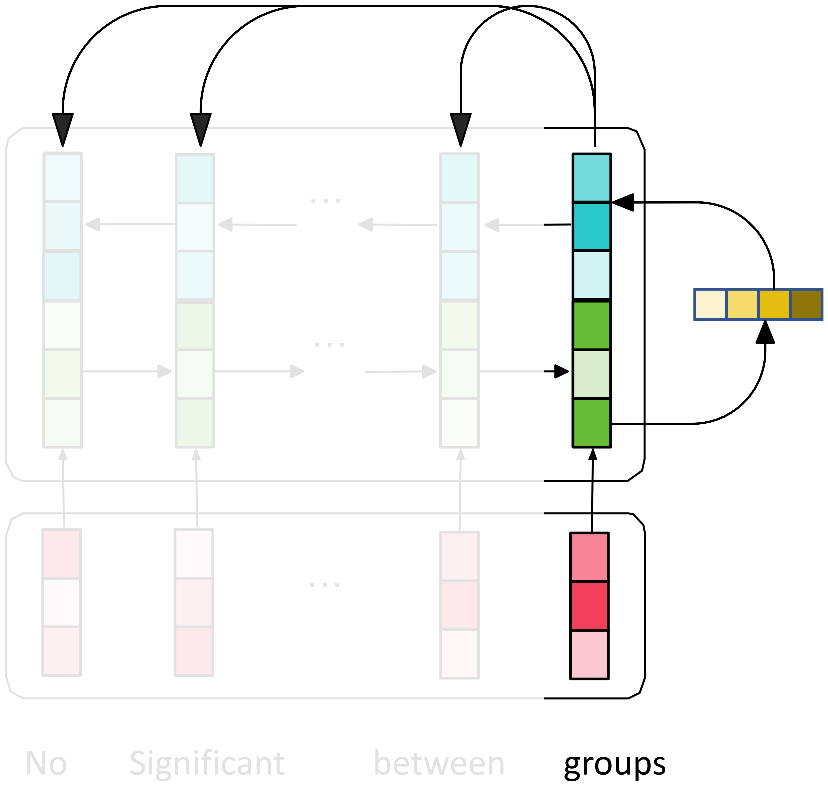
Attention in Transformers



Attention in Transformers



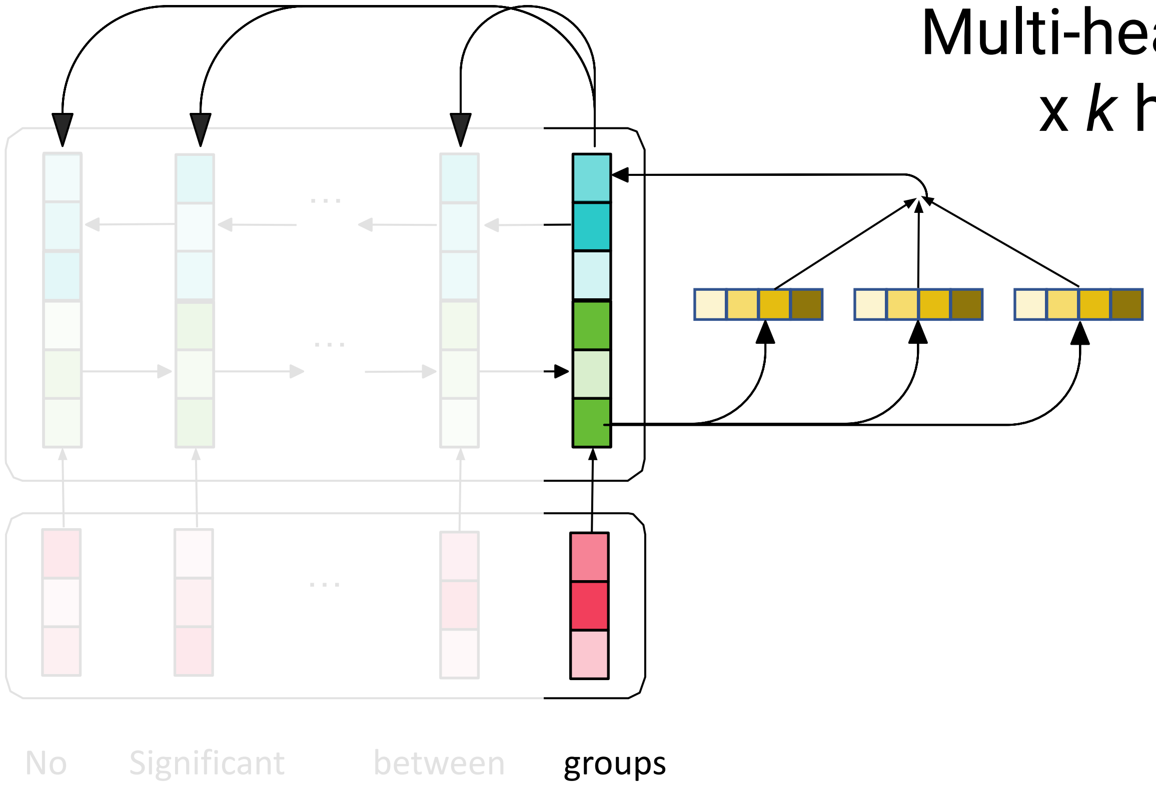
Attention in Transformers



Self-attention:
x *n* input tokens

Attention in Transformers

Multi-headed:
x *k* heads



Attention in Transformers: Challenges

- Sheer ***amount*** of attention
  - e.g., 13B LLaMA model: 40 layers x 100 input tokens x 40 heads
  - ~= 160,000 individual attention patterns which could be studied.

Attention in Transformers: Challenges

- Sheer ***amount*** of attention
  - e.g., 13B LLaMA model: 40 layers x 100 input tokens x 40 heads
  - ~= 160,000 individual attention patterns which could be studied.

- Simplifying approach for BERT:
  - Final-layer attention paid by the [CLS] token to all other tokens (aggregated over heads)

Attention in Transformers: Findings

- Can attention be used in Transformers to provide heatmap based explanation?
  - Token Identifiability, Adversarial Attention Distributions, Effective Attention, Attention Flows
- Do all attention distributions in transformers really matter?
  - Ablation & Pruning
- What can attention tell us about the global mechanisms used by Transformer models?
  - Linguistic Subtasks, Copying Behavior, Factual Knowledge, Token Identifiability

Attention in Transformers: Findings

1. **Token Identifiability**
    - *On Identifiability in Transformers* (Brunner et al. 2020)

2. **Adversarial attention distributions exist for BERT**
    - *Learning to Deceive with Attention-Based Explanations* (Pruthi et al. 2020)

Attention in Transformers: Findings

3. **Modifications to attention scores to improve their interpretability:**

Effective Attention

- *On Identifiability in Transformers* (Brunner et al. 2020)
- *Effective Attention Sheds Light On Interpretability* (Sun & Marasović 2021)

Attention Flows

- *Quantifying Attention Flow in Transformers* (Abnar & Zuidema 2020)
- *Attention Flows are Shapley Value Explanations* (Ethayarajh & Jurafsky 2021)

Attention in Transformers: Findings

4. **Ablation + Pruning of heads: possible**
   - *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned* (Voita et al. 2019)
   - *Are Sixteen Heads Really Better than One?* (Michel et al. 2019)
   - *Revealing the Dark Secrets of BERT* (Kovaleva et al. 2019)
   - *Self-Attention Attribution: Interpreting Information Interactions Inside Transformer* (Hao et al. 2021)

Attention in Transformers: Findings

5. Specialization of attention heads to linguistic subtasks (e.g., syntax/PoS/coreference).

- *Analyzing the Structure of Attention in a Transformer Language Model* (Vig & Belinkov 2019)
- *What Does BERT Look At? An Analysis of BERT's Attention* (Clark et al. 2019)
- *Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned* (Voita et al. 2019)
- *Attention is Not Only a Weight: Analyzing Transformers with Vector Norms* (Kobayashi et al. 2020)

Attention in Transformers: Findings

6. **Attention promotes copying behavior**
    - *A Mathematical Framework for Transformer Circuits* (Elhage et al. 2021)
    - *In-context Learning and Induction Heads* (Olsson et al. 2022)
    - *Locating and editing factual associations in GPT* (Meng et al. 2022)

Attention in Transformers: Findings

7. **Attention on key entities can predict model correctness**
    - *Attention Satisfies: A Constraint-Satisfaction Lens on Factual Errors of Language Models* (Yuksekgonul et al. 2023)

Current & Future Relevance: Community-Level Shifts

1. Types of tasks we care about

2. Generality of behavior we want to explain

Current & Future Relevance: Community-Level Shifts

1. Types of tasks we care about

- Attention is **no longer very useful** for instance-level explanations

In group A, lower peak (median) plasma levels of procalcitonin (0.2 versus 1.4, $p < 0.001$), IL 8 (5.6 versus 94.8, $p < 0.001$), IL 10 (47.2 versus 209.7, $p = 0.001$), endothelial leukocyte adhesion molecule-1 (88.5 versus 130.6, $p = 0.033$), intercellular adhesion molecule-1 (806.7 versus 1,375.7, $P = 0.001$) and troponin-I (0.22 versus 0.66, $p = 0.018$) were found. There was no significant difference in IL 6, IL-6r and C-reactive protein values between groups. Higher figures of the cardiac index ($p = 0.010$) along with reduced systemic vascular resistance ($p = 0.005$) were noted in group A.

no significant difference

Did Aristotle use a laptop?

StrategyQA

The following are multiple choice questions about high school mathematics.

How many numbers are in the list 25, 26, ..., 100?
(A) 75 (B) 76 (C) 22 (D) 23
Answer: B

Compute $i + i^2 + i^3 + \cdots + i^{258} + i^{259}$.
(A) -1 (B) 1 (C) i (D) $-i$
Answer: A

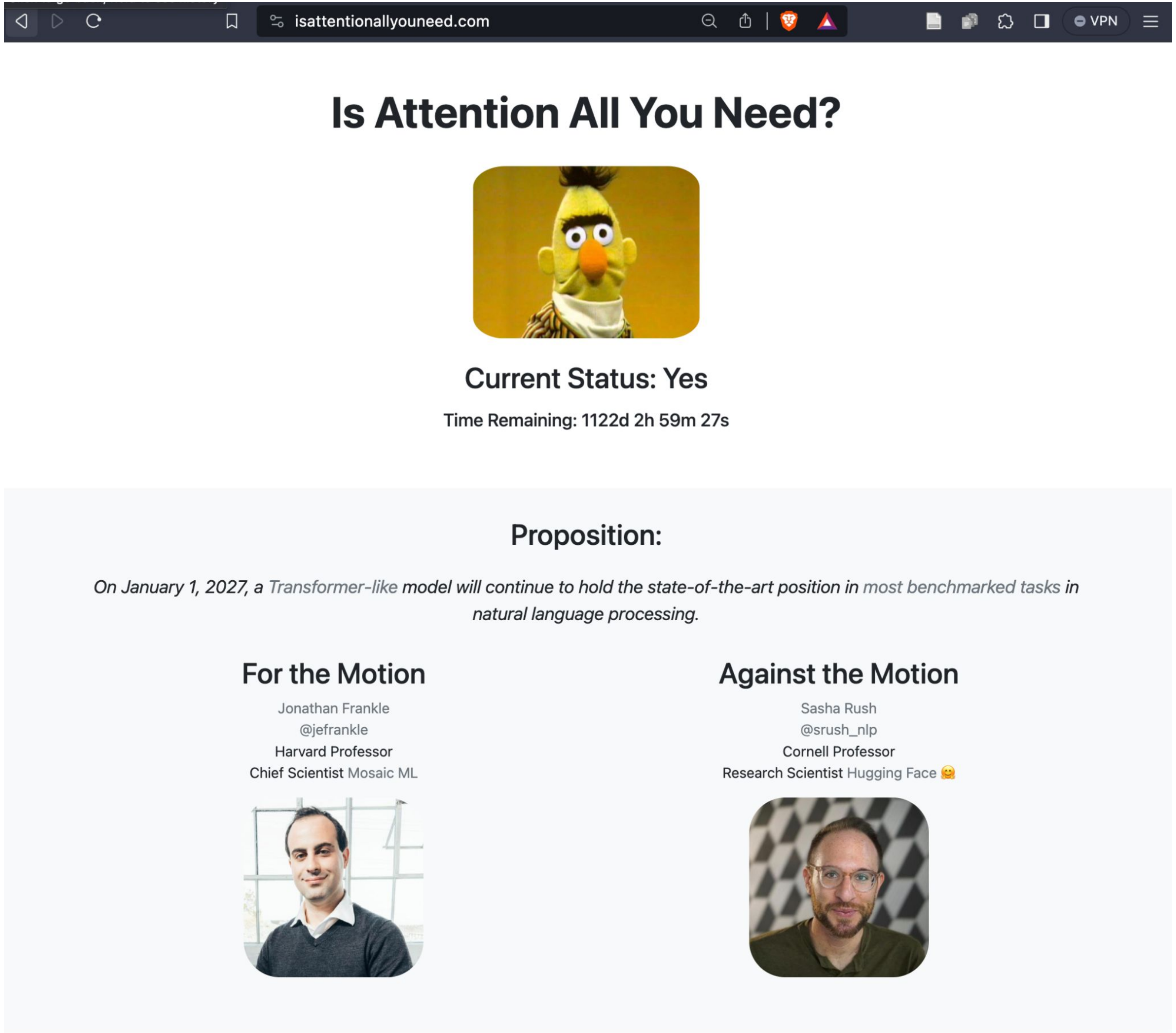
MMLU

Current & Future Relevance: Community-Level Shifts

2. Generality of behavior we want to explain

- Our focus: providing **instance-level explanations** of model behavior
- Current focus: understanding the **mechanisms** underlying general-purpose Transformers
  - Beyond specific **models**, **datasets** and even **architectures**, **tasks**
  - <u>Understanding attention is **still important**</u> ⭐

Attention is still important



isattentionallyouneed.com

Thank you!

@successar_nlp, sarahwiegreffe

{successar, wiegreffesarah}@gmail.com