Attention is not not Explanation

Sarah Wiegreffe* and Yuval Pinter*

@sarahwiegreffe

@yuvalpi

http://github.com/sarahwie/attention

Background

- Can attention weights serve as a form of explanation?
  - Jain & Wallace 2019, Serrano & Smith 2019

Background

- Can attention weights serve as a form of explanation?
  - Jain & Wallace 2019, Serrano & Smith 2019

Background

- Can attention weights serve as a form of explanation?
  - Jain & Wallace 2019, Serrano & Smith 2019

***Plausible* Explainability**

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

Background

- Can attention weights serve as a form of explanation?
  - Jain & Wallace 2019, Serrano & Smith 2019

***Plausible* Explainability**

- Rationale generation
  (Ehsan et al. 2019, Riedl 2019)

***Faithful* Explainability**

- Understanding correlation between inputs and output (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

Background

- Can attention weights serve as a form of explanation?
  - Jain & Wallace 2019, Serrano & Smith 2019

brilliant and moving performances by tom and peter finch

***Plausible* Explainability**

- Rationale generation (Ehsan et al. 2019, Riedl 2019)

***Faithful* Explainability**

- Understanding correlation between inputs and output (Lipton 2016, Rudin 2018)
- Models' explanations are exclusive

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

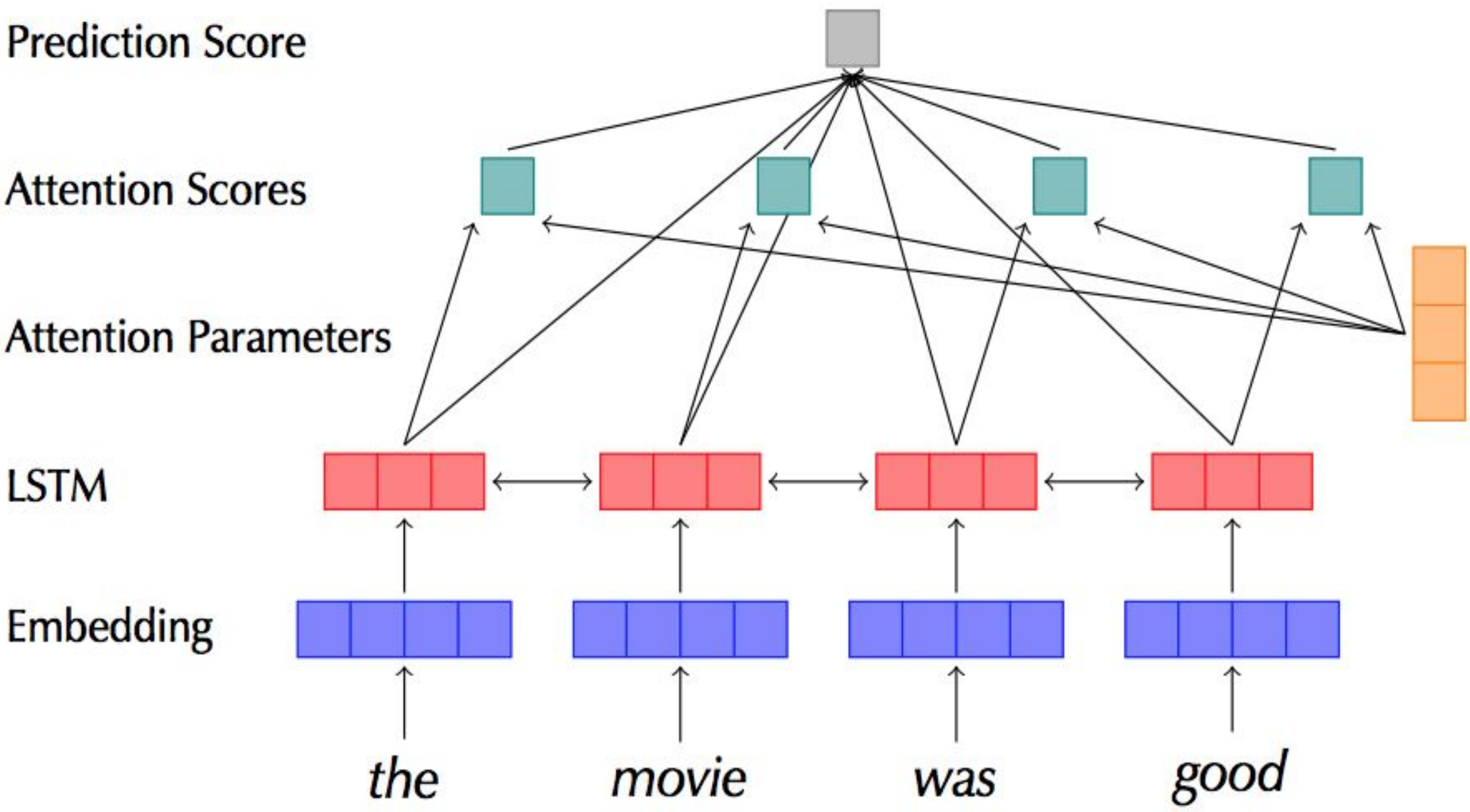
Selecting Meaningful Tasks

Necessary

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

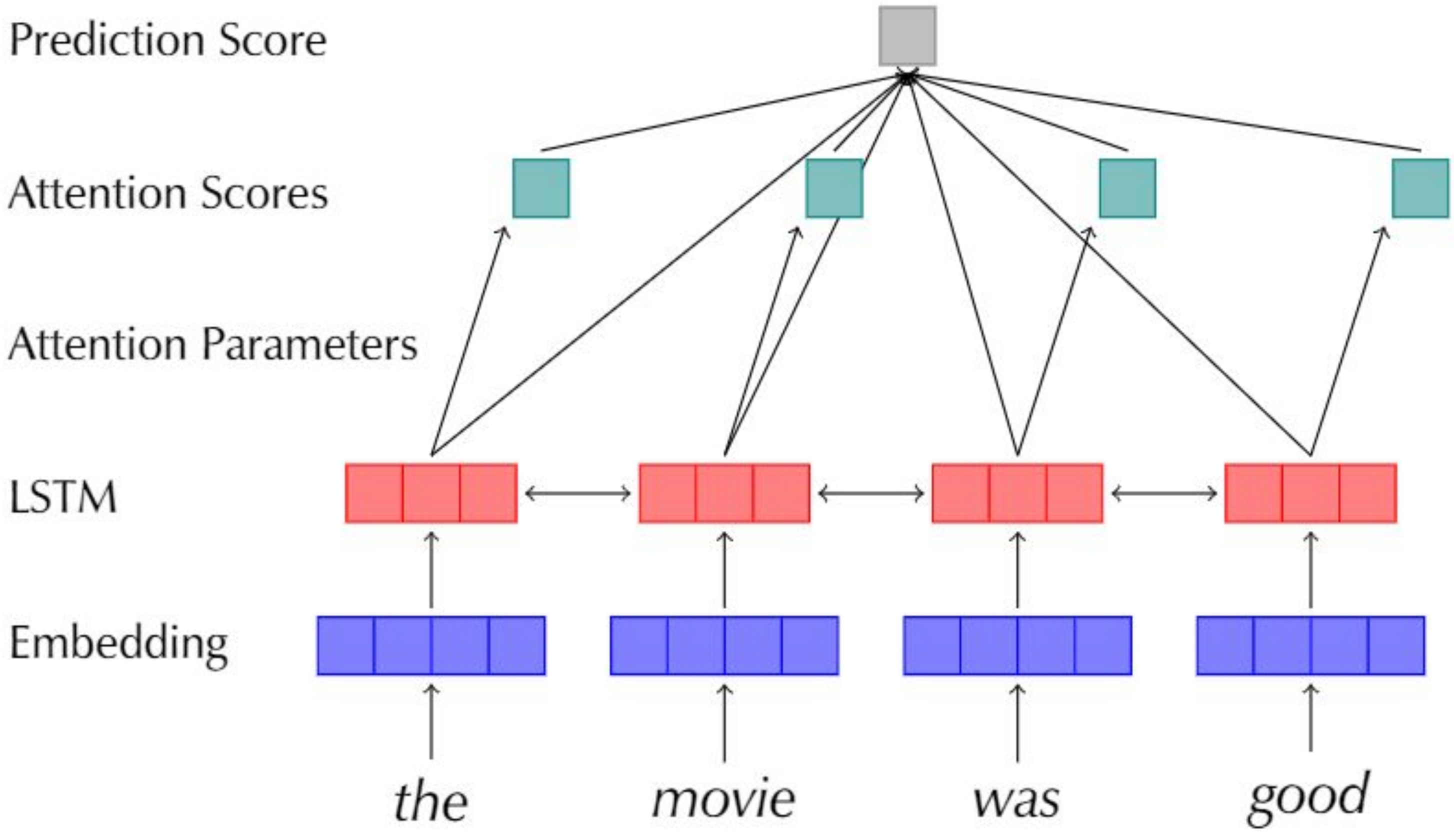
Selecting Meaningful Tasks

Necessary



Selecting Meaningful Tasks

Necessary



Set to Uniform

Selecting Meaningful Tasks

Necessary

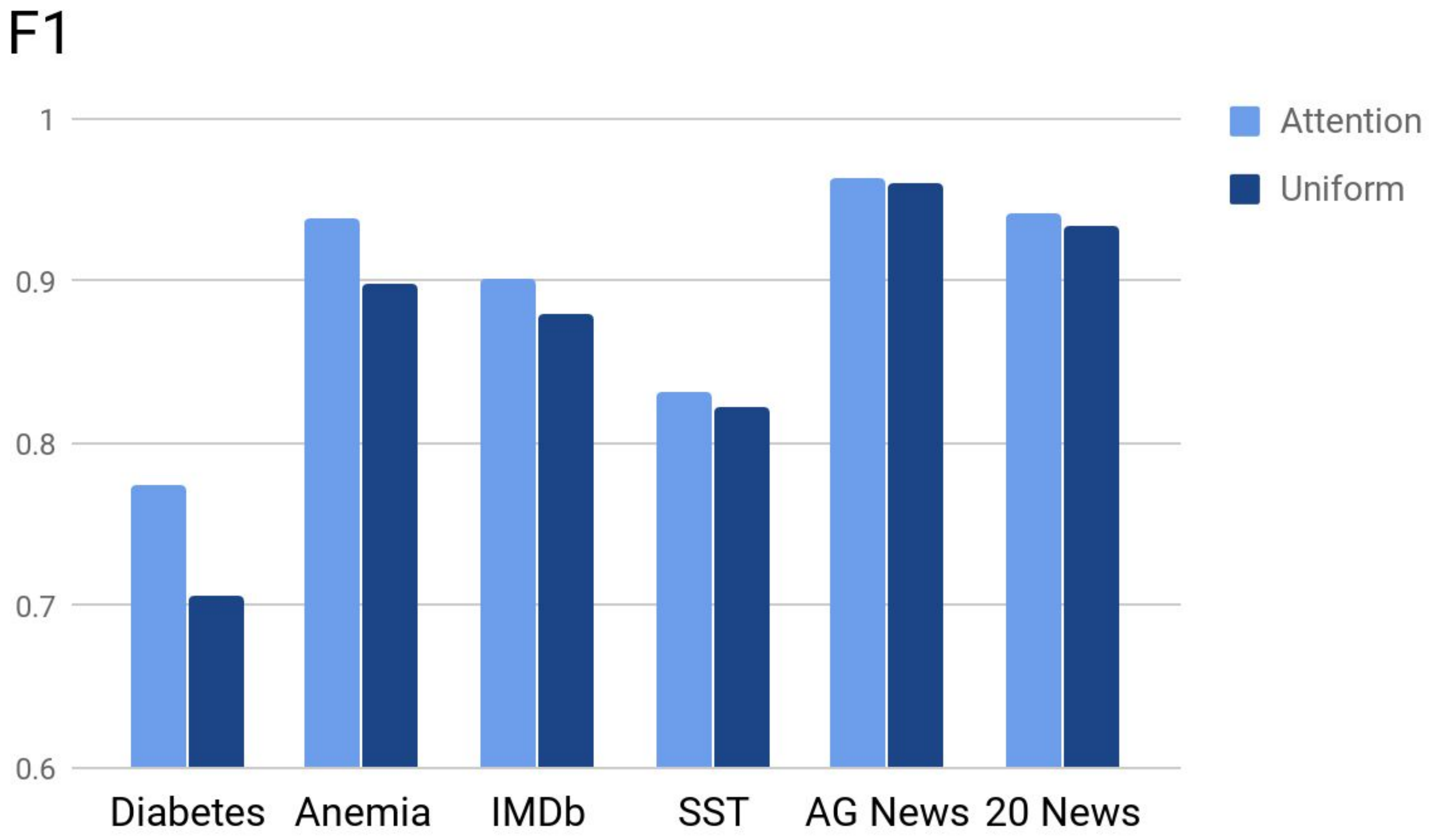
- Diabetes (MIMIC-III)
- Anemia (MIMIC-III)

- IMDb Movie Reviews
- Stanford Sentiment Treebank (SST)

- AG News
- 20 Newsgroups

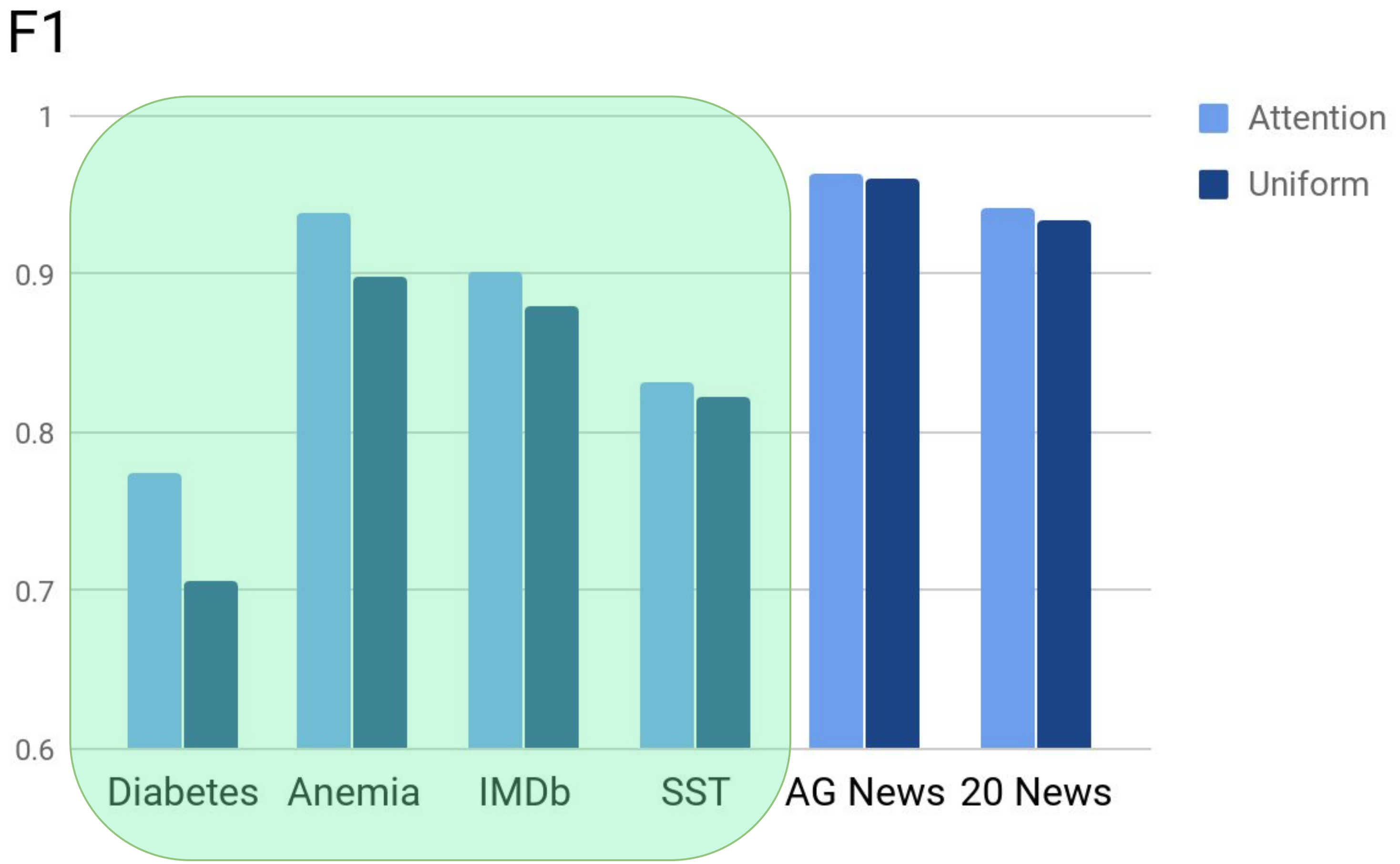
Selecting Meaningful Tasks

Necessary



Selecting Meaningful Tasks

Necessary



Searching for Adversarial Models

Hard to manipulate

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

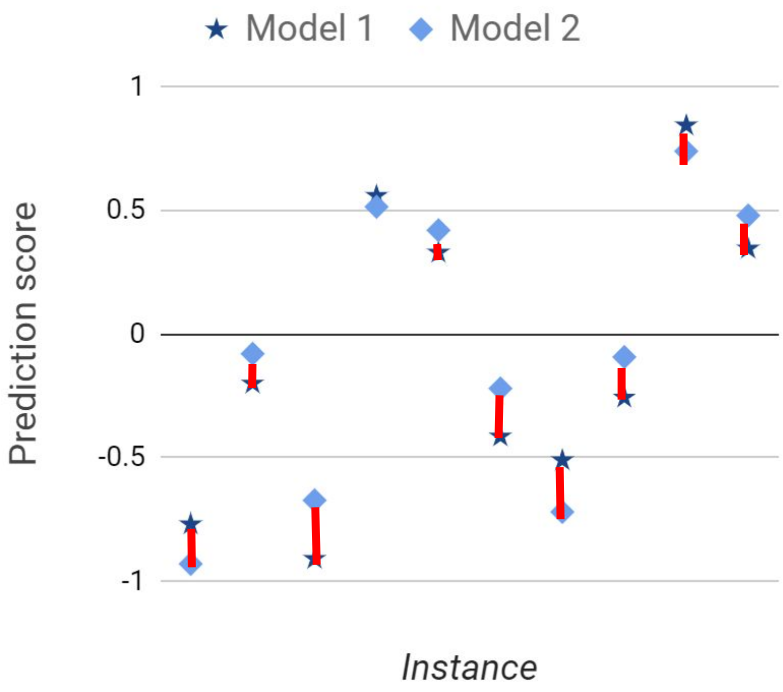
2. If **trained models** can vary in attention distributions while giving similar predictions, they might be bad for explanation

Measures

Hard to manipulate

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

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



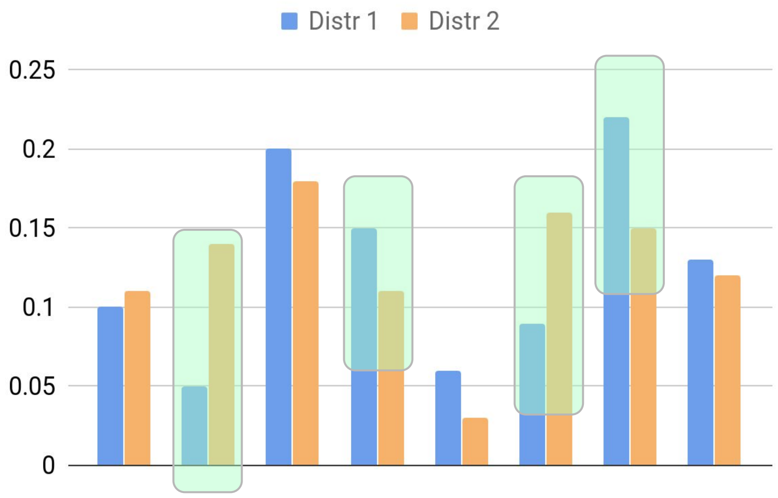
Measures

Hard to manipulate

- Jensen–Shannon Divergence: for comparing 2 distributions

$$\mathrm{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}\mathrm{KL}[\alpha_1 \parallel \bar{\alpha}] + \frac{1}{2}\mathrm{KL}[\alpha_2 \parallel \bar{\alpha}],$$

where $\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}$.



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

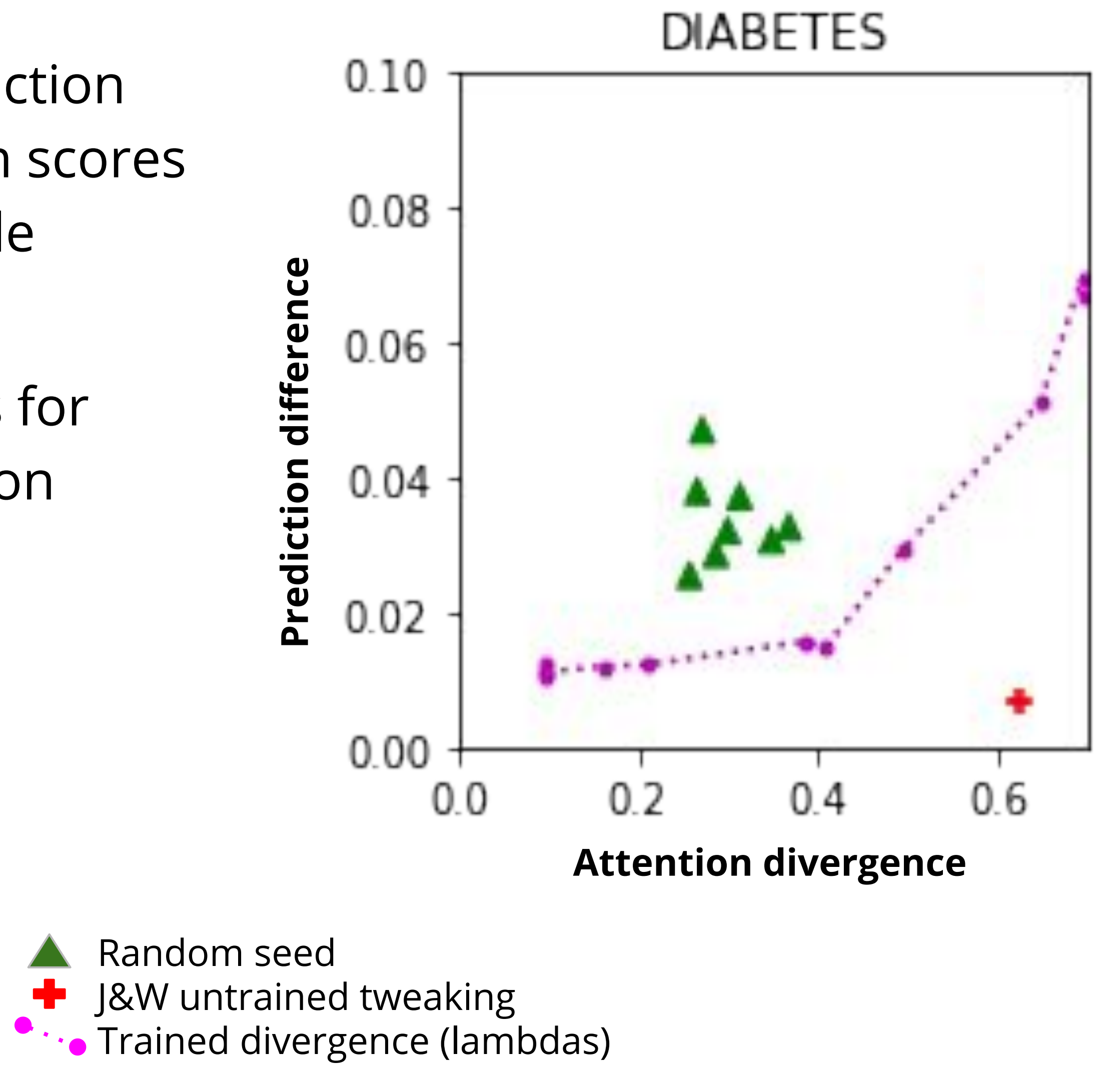
Hard to manipulate

- Random seed variance
  - Re-running the **base setup** with multiple random seeds to calibrate what we expect for variance in attention weights
- Jain & Wallace (2019)
  - Finding adversarial attention maps by post-hoc tweaking
  - **No model trained**

Adversarial Results

Hard to manipulate

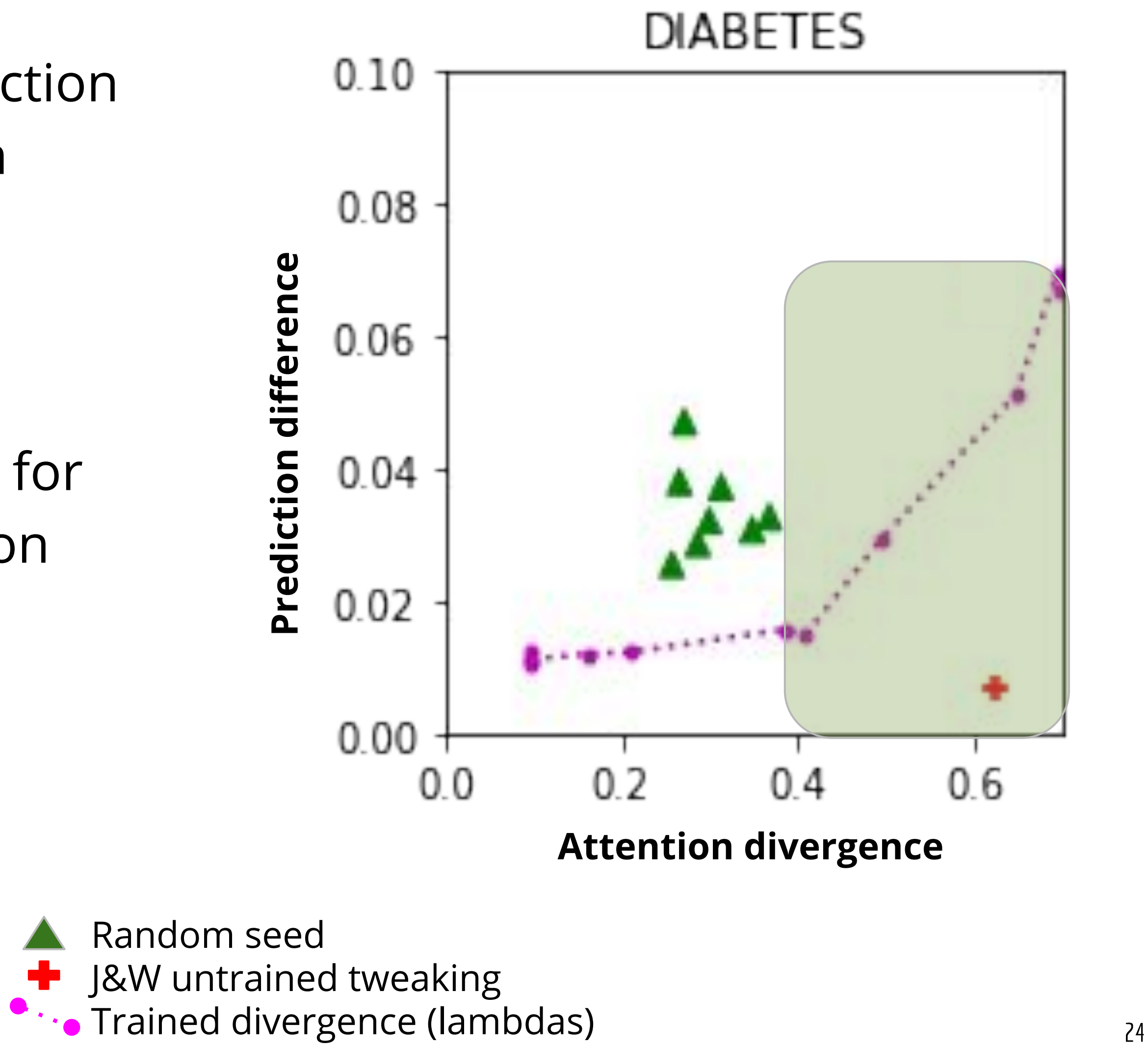
- Fast increase in prediction difference = attention scores not easily manipulable
    - Supports use of attention weights for faithful explanation



Adversarial Results

Hard to manipulate

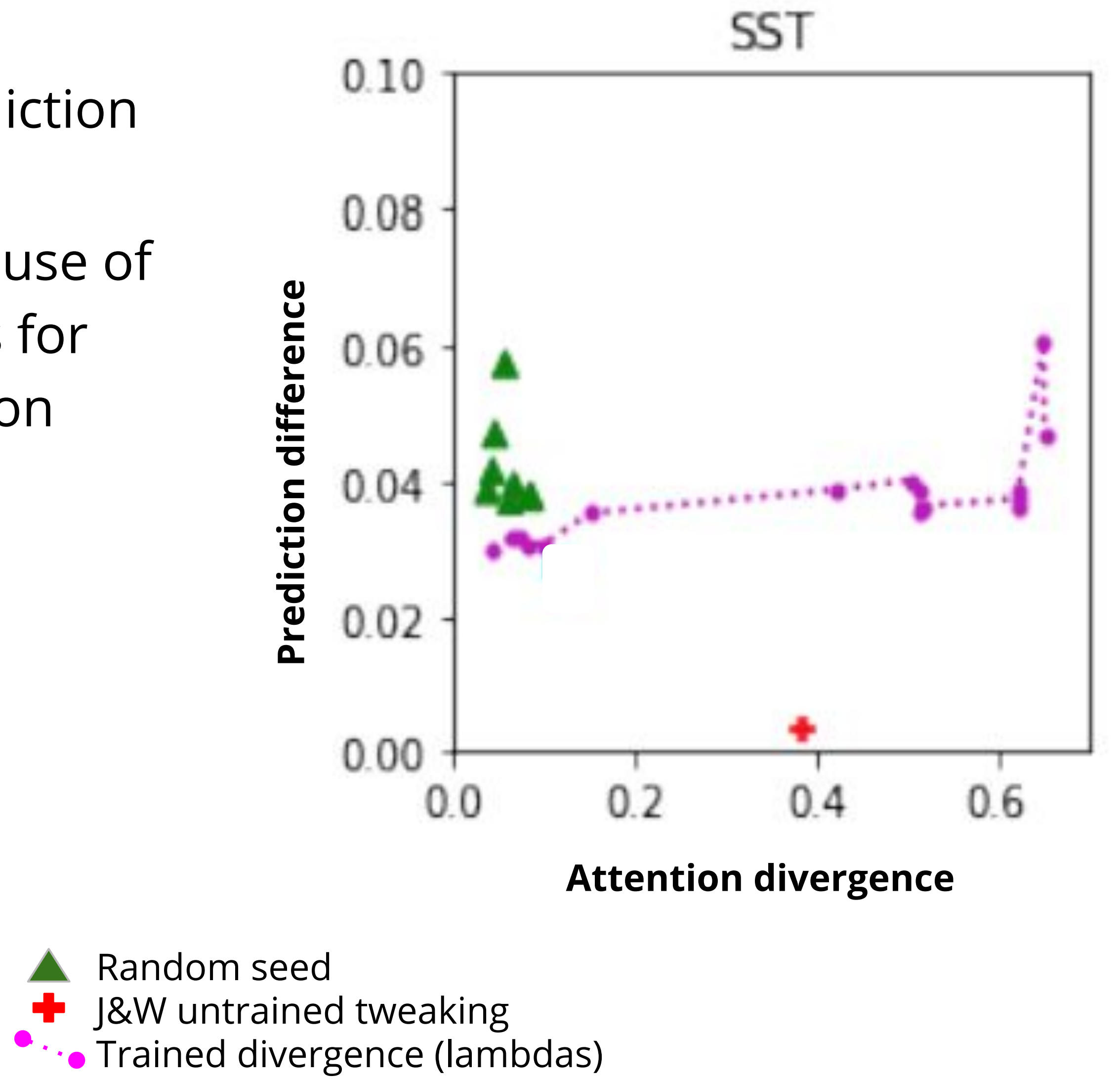
- Fast increase in prediction difference = attention scores not easily manipulable
  - Supports use of attention weights for faithful explanation



Adversarial Results

Hard to manipulate

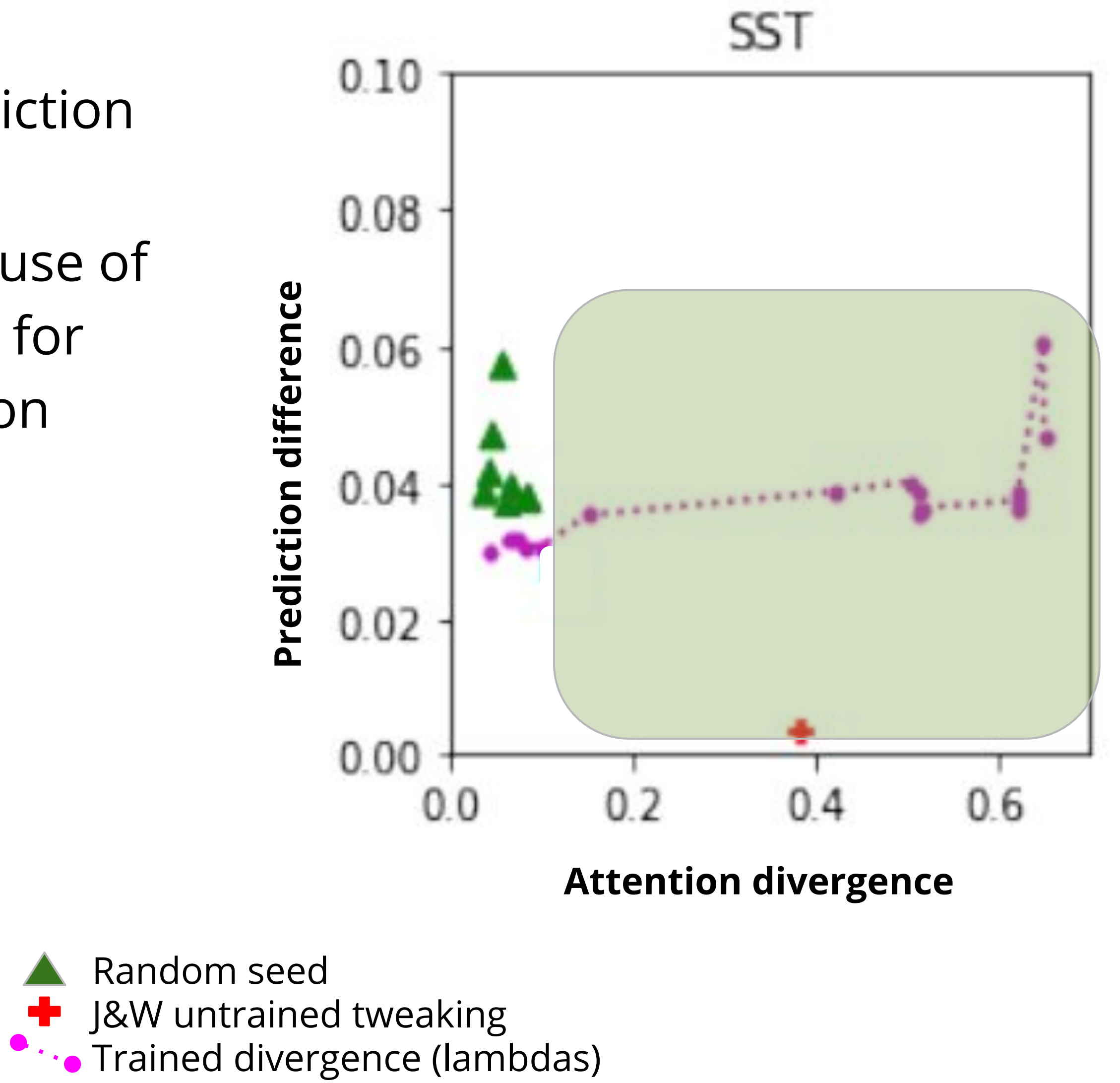
- Slow increase in prediction difference
    - *Does not* support use of attention weights for faithful explanation



Adversarial Results

Hard to manipulate

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



Probing Attention

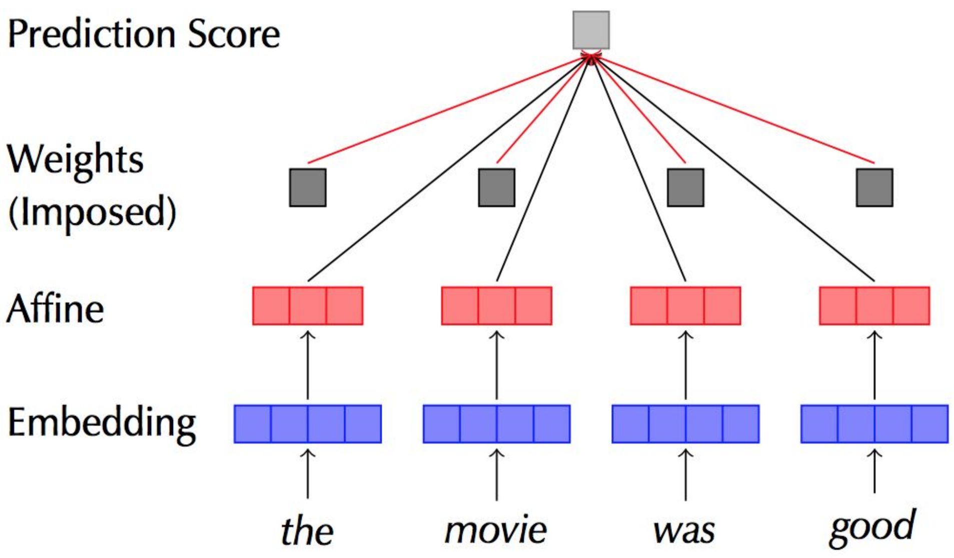
Work out of context

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**

Probing Attention

Work out of context

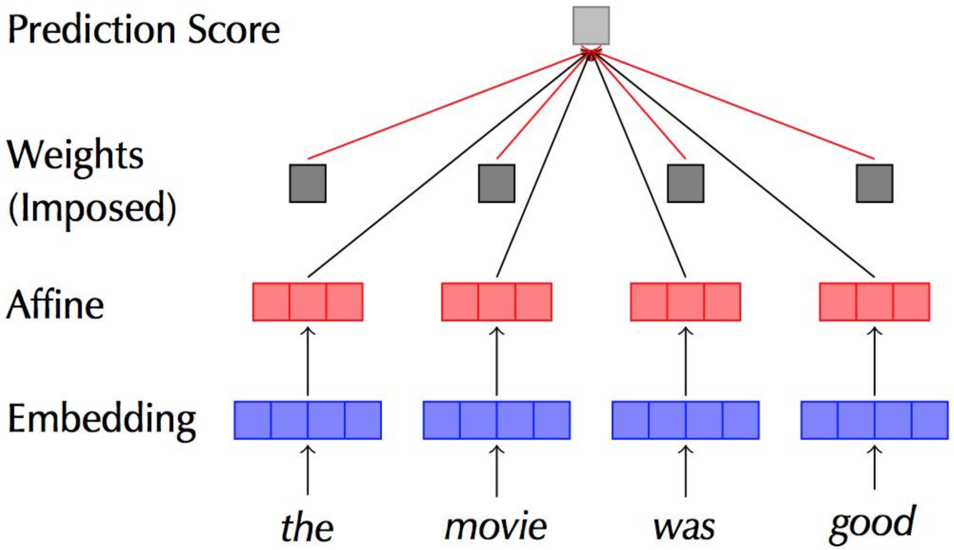
- Treat the learned attention weights as a **guide** in a non-contextualized, bag-of-word-vectors model



Probing Attention

Work out of context

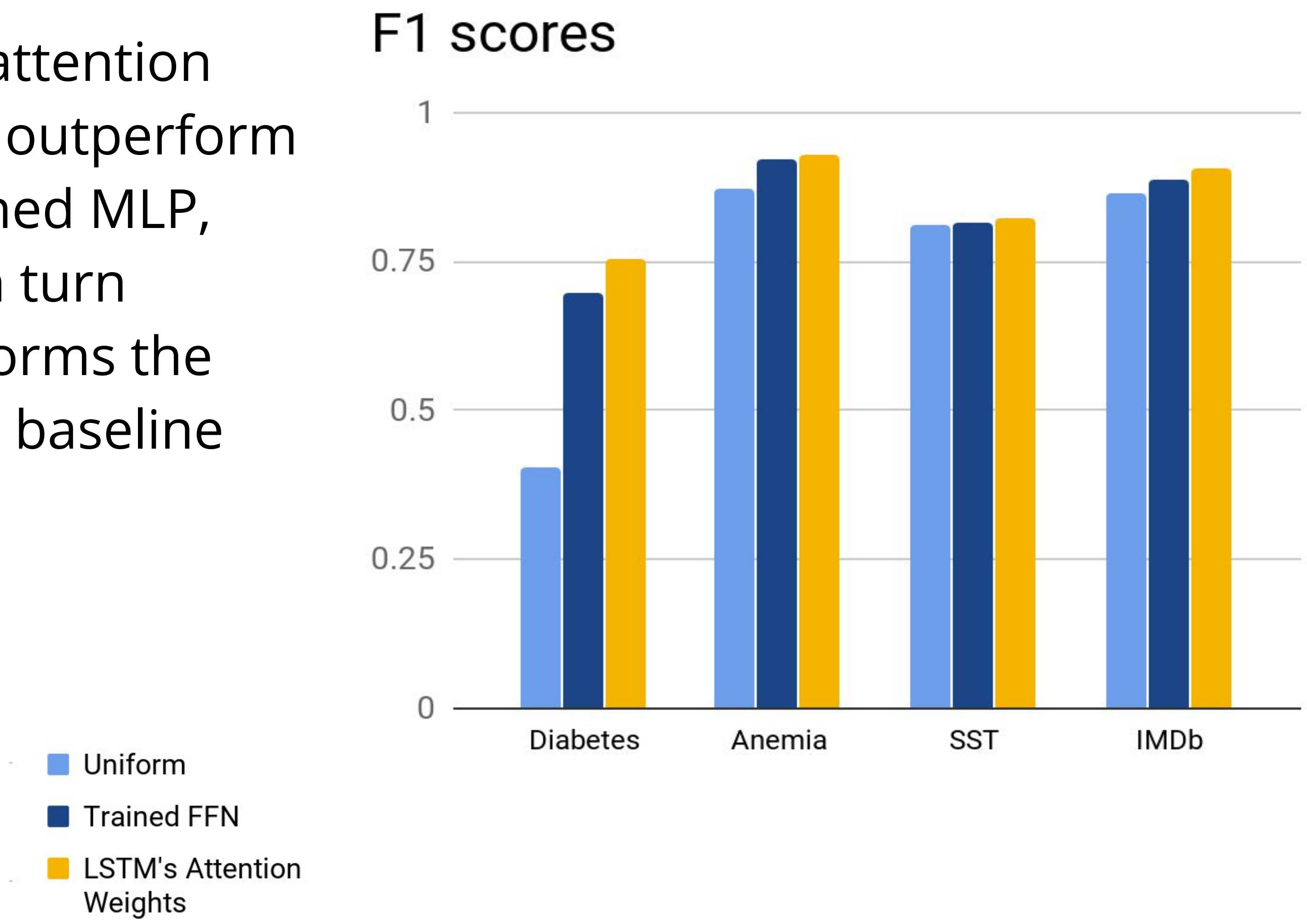
- Treat the learned attention weights as a **guide** in a non-contextualized, bag-of-word-vectors model
- High performance → attention scores capture relationship between inputs and output



Results

Work out of context

- LSTM’s attention weights outperform the trained MLP, which in turn outperforms the uniform baseline



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 Guides

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 Guides

Recommendations

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

Code: http://github.com/sarahwie/attention

Thanks!

- Acknowledgements: Yoav Goldberg, Erik Wijmans, Sarthak Jain, Byron Wallace, and members of the Georgia Tech Computational Linguistics Lab, particularly Jacob Eisenstein and Murali Raghu Babu Balusu.
- Yuval Pinter is supported by a Bloomberg Data Science Fellowship.

Twitter: @sarahwiegreffe @yuvalpi

Code: http://github.com/sarahwie/attention