Attention is not not Explanation

Sarah Wiegreffe* (presenting) and Yuval Pinter*

http://github.com/sarahwie/attention

Follow me on Twitter: @sarahwiegreffe

Motivation:

- Can attention weights serve as a form of explanation?

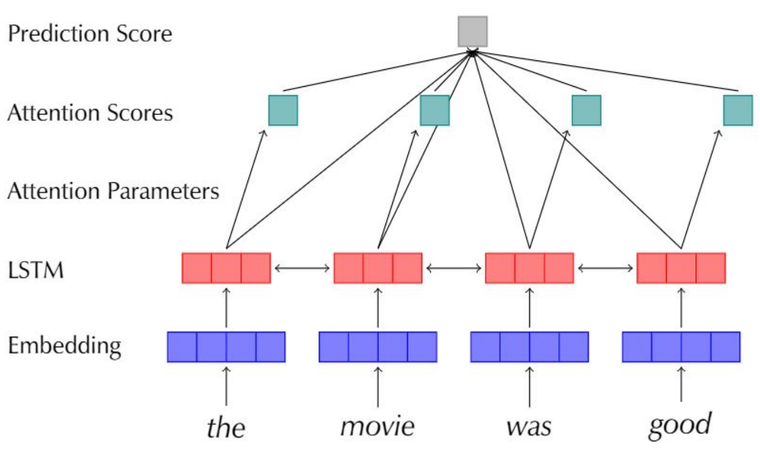
brilliant and moving performances by tom and peter finch

- Faithful Explainability (Jain & Wallace 2019, Serrano & Smith 2019)
  - Understanding correlation between inputs and output
  - Models' explanations are exclusive

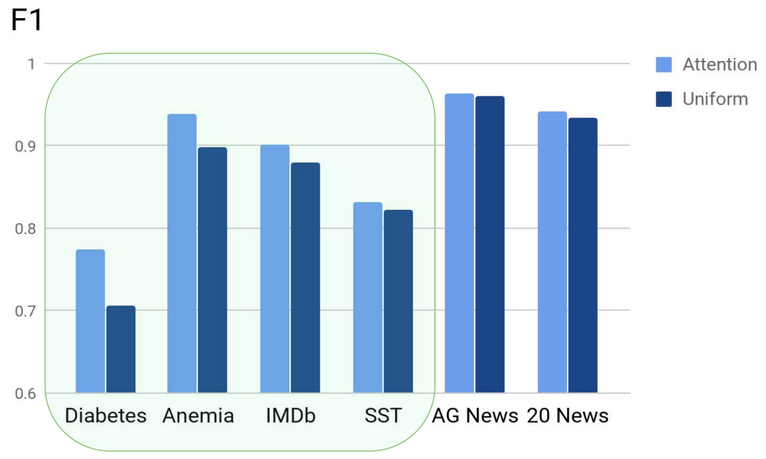
Thesis: ***If Attention is (Faithful) Explanation, then***

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

Experiment 1: Selecting Meaningful Tasks



Set to Uniform



Experiment 2: Searching for Adversarial Models

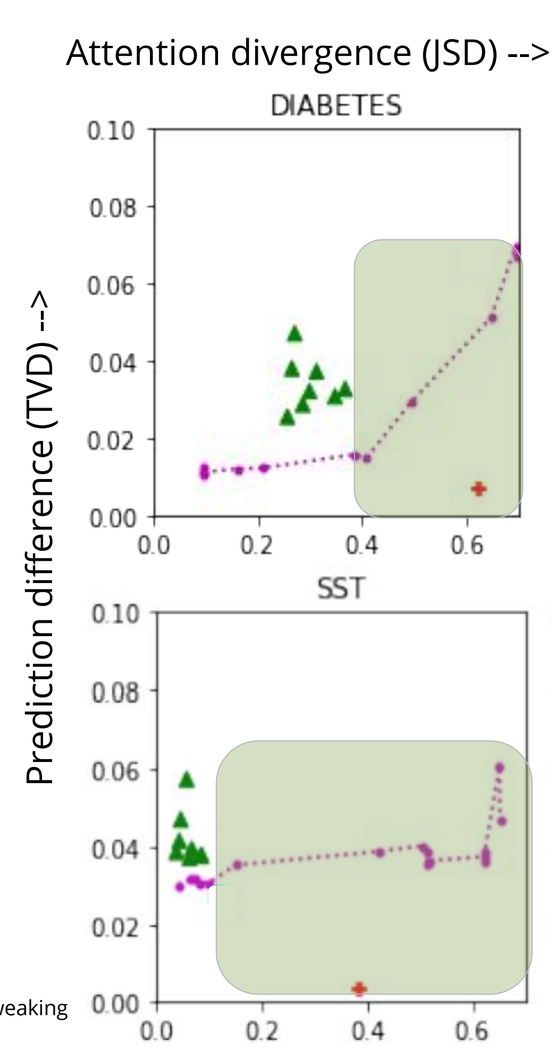
1. Train a base model (M_b)
2. Train an adversary (M_a) that minimizes change in prediction scores while maximizing changes in the learned attention distributions.

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

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

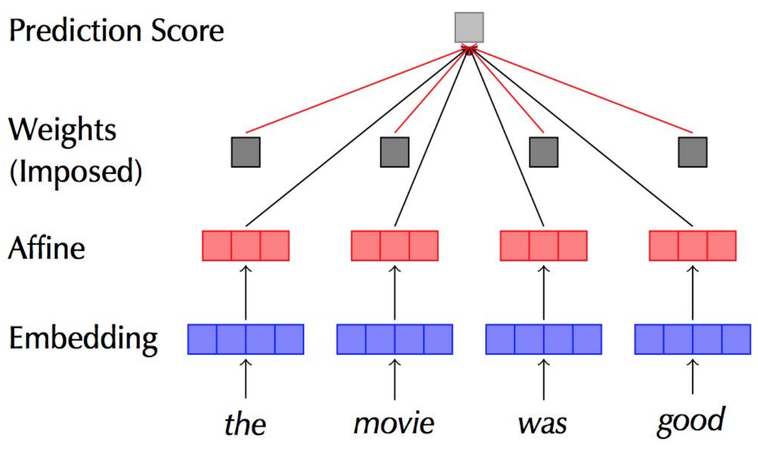
- Metrics:
  - Total Variation Distance: for comparing class predictions
  - Jensen-Shannon Divergence (JSD): for comparing 2 distributions

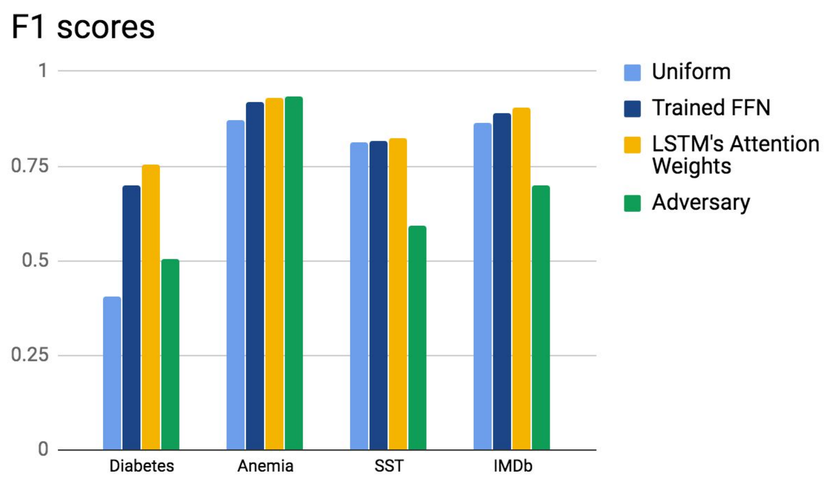
- Looking for fast vs. slow increase in prediction difference
  - Attention scores easily manipulable? (fast=no, slow=yes)
  - Supports use of attention weights for faithful explanation? (fast=yes, slow=no)



Experiment 3: Using Attention as a Guide

- Non-contextualized model
- High performance → attention scores capture relationship between inputs and output





Takeaways:

- Performance highly task-specific
- Use guides to judge token-output correlation
- Use adversarial models to investigate exclusivity
- Calibrate your notion of variance
- Investigate models & tasks where attention is necessary