Attention is not not Explanation

By Sarah Wiegreffe Yuval Pinter

Presenter: Zijie Pan

https://qdata.github.io/deep2Read/

02/21/2020

Motivation

Attention is not Explanation is overstating

Key points in previous paper:

- Attention weights should correlate with feature importance measures.
- Counterfactually, attention weight configurations ought to yield corresponding changes in prediction (and if they do not then are equally plausible as explanations)

Methods

- Feature erase method
- Generate alternative weight distribution

- **Attention Distribution is not a Primitive**

  “ From a modeling perspective, detaching the attention scores obtained by parts of the model degrades the model itself ”

- **Existence does not Entail Exclusivity**

Uniform as the Adversary

- Examine whether attention is necessary in every dataset (ex. very simple task)

- Uniform Model Variant (attention weight distribution is frozen to uniform weights while training)

- Expectation: Large drop of performance if attention is a necessary component

Uniform as the Adversary Results

Dataset	Attention (Base)		Uniform
	Reported	Reproduced	
Diabetes	0.79	0.775	0.706
Anemia	0.92	0.938	0.899
IMDb	0.88	0.902	0.879
SST	0.81	0.831	0.822
AgNews	0.96	0.964	0.960
20News	0.94	0.942	0.934

Table 2: Classification F1 scores (1-class) on attention models, both as reported by Jain and Wallace and in our reproduction, and on models forced to use uniform attention over hidden states.

Variance with a Model

- To test whether the variances (J&W) between trained attention scores and adversarially-obtained ones are unusual.

- Different initialization random seeds -> variance of attention distribution (baseline variance)

Variance with a Model Results

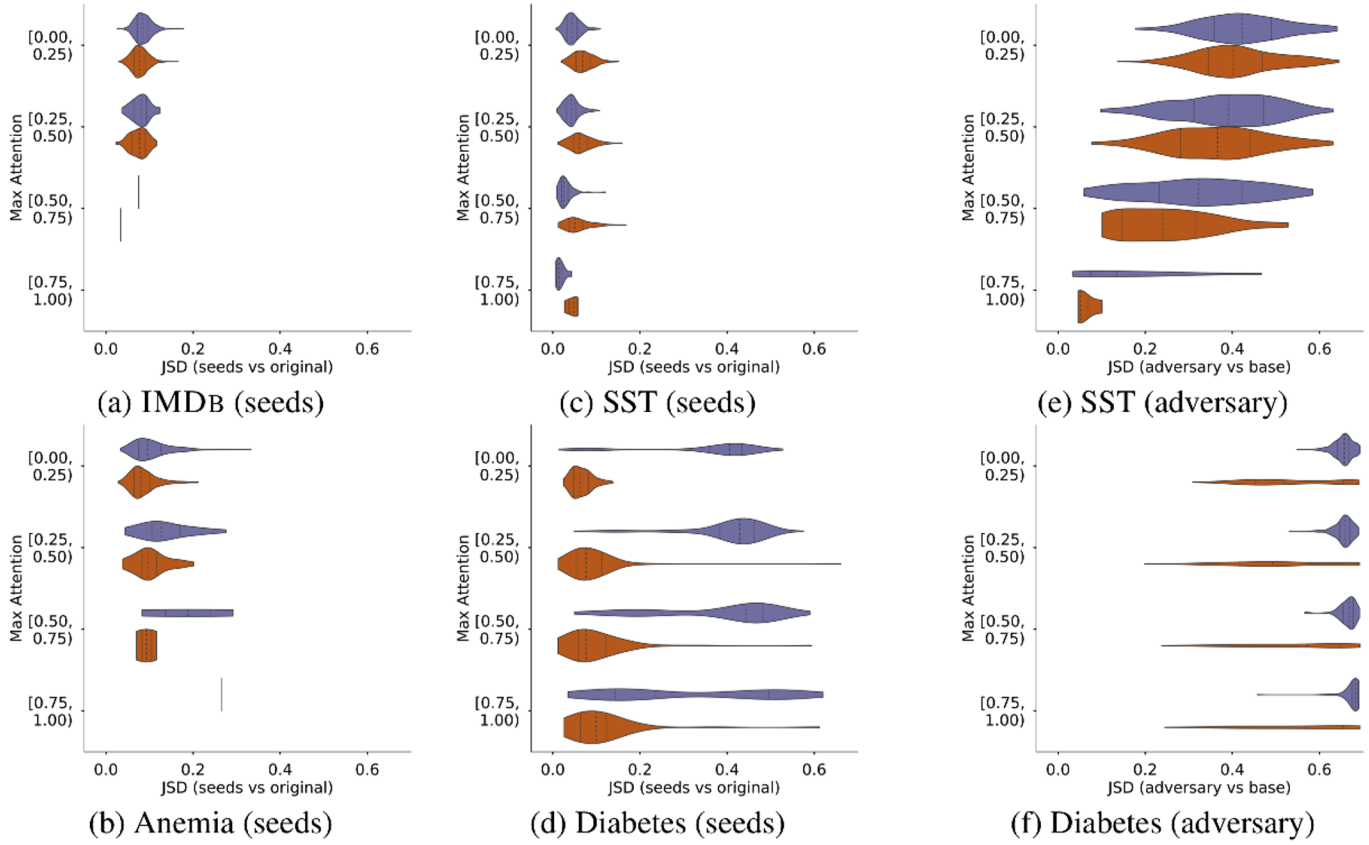
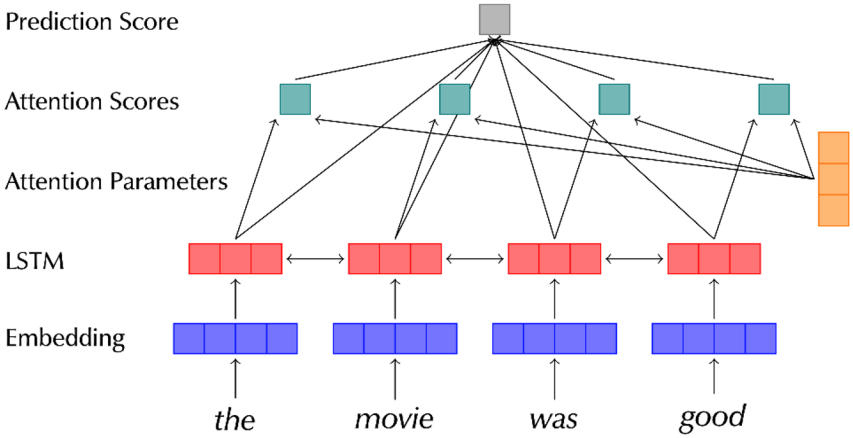


Figure 3: Densities of maximum JS divergences (x-axis) as a function of the max attention (y-axis) in each instance between the base distributions and: (a-d) models initialized on different random seeds; (e-f) models from a per-instance adversarial setup (replication of Figure 8a, 8c resp. in Jain and Wallace (2019)). In each max-attention bin, top (blue) is the negative-label instances, bottom (red) positive-label instances.

Diagnosing Attention Distributions by Guiding Simpler Models

- A complementary approach to (J&W)
- Mitigate the contextual influence
- Replace LSTM with MLP
- Four weight settings:
  - Uniform weights
  - not freezing weights layer and train with MLP
  - Base LSTM
  - Weights found adversarially

Experiment Setup



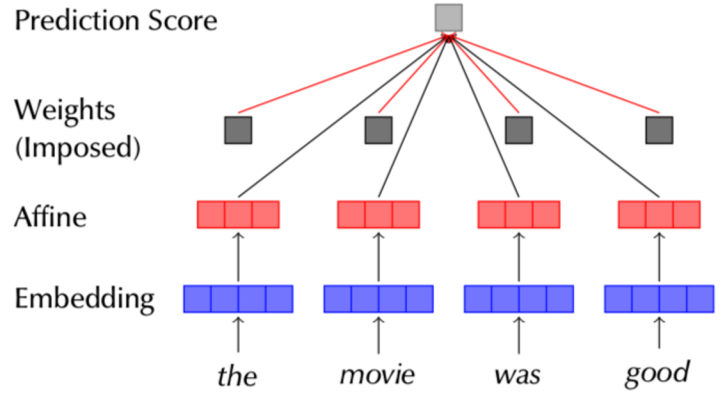


Figure 4: Diagram of the setup in §3.4 (except TRAINED MLP, which learns weight parameters).

Results

Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	**0.753**	0.931	**0.824**	**0.905**
ADVERSARY (4)	0.503	**0.932**	0.592	0.700

Table 3: F1 scores on the positive class for an MLP model trained on various weighting guides. For ADVERSARY, we set $\lambda \leftarrow 0.001$.

Training an Adversary

- Model-consistent training protocol for finding adversarial attention distribution, which can be used in faithful explainability
- To train model a to provide similar prediction scores for each instance as base model b,but distance its attention distribution from that of model b
- Loss function:

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

- TVD and JSD tradeoff

Guide weights	Diab.	Anemia	SST	IMDb
UNIFORM	0.404	0.873	0.812	0.863
TRAINED MLP	0.699	0.920	0.817	0.888
BASE LSTM	**0.753**	0.931	**0.824**	**0.905**
ADVERSARY (4)	0.503	**0.932**	0.592	0.700

Explainability has many Definitions

- Explainable AI: *transparency, explainability, interpretability*
- Plausible or faithful or both
- Adversarial found distributions confirmed and indicate not a faithful interpretation of the model