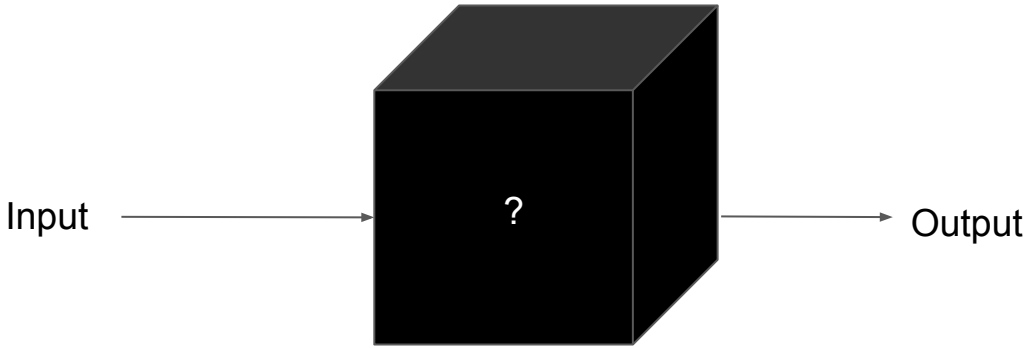
Attention is Not Explanation

(or is it?)

Stefan F. Schouten, Michael J. Neely

Neural models are normally black boxes...

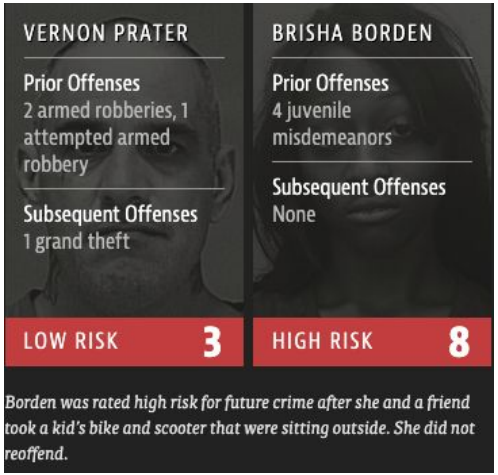


And this can be problematic...

“The data subject should have the right **not to be subject to a decision**, which may include a measure, evaluating personal aspects relating to him or her which is **based solely on automated processing** … such processing should be subject to suitable safeguards, which should include specific information to the data subject and **the right to obtain human intervention**, to express his or her point of view, t**o obtain an explanation of the decision** reached after such assessment and to challenge the decision” - GDPR Recital 71[1]

Hence the inclusion of this course

- Fairness
- Accountability
- Confidentiality
- Transparency (Explanation)



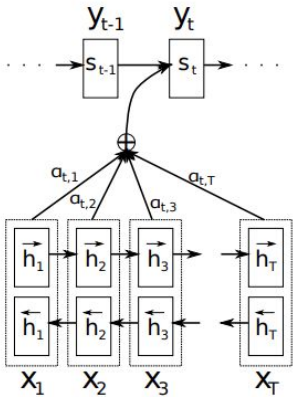
Racial bias in a recidivism algorithm used in the United States[2]

Methods to ‘explain’ model decisions

- Attention [Bahdanau et al. 2015]
- Shapely sampling values [Štrumbelj and Kononenko 2013]
- LIME [Ribeiro et al. 2016]
- SHAP [Lundberg and Lee 2017]
- Contextual Decomposition [Jumelet, Zuidema, and Hupkes 2019] (UvA)

Attention

Keep intermediate encoder hidden states to build a *context vector* so the decoder can (soft-)search for the sequence tokens relevant for prediction.

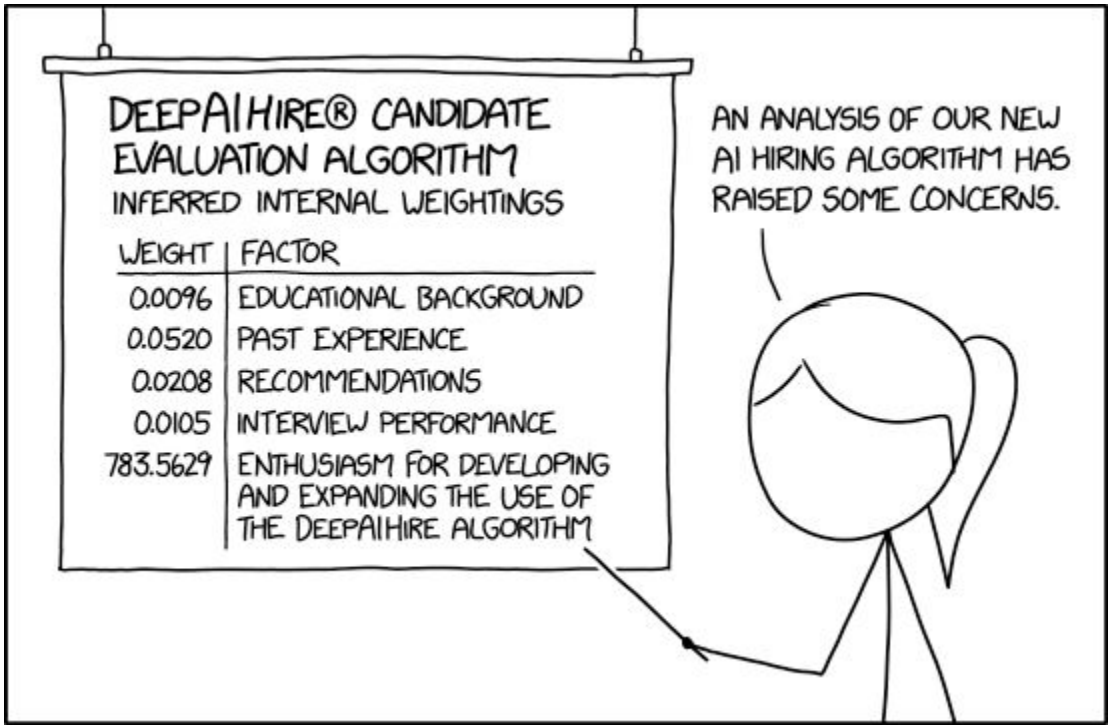


Graphic from [Bahdanau et al. 2015][3]

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

Heatmap from [Jain and Wallace 2019][4]

But can we trust these mechanisms?



[5]

Jain and Wallace 2019: “Attention is Not Explanation”

“Assuming attention provides a faithful explanation for model predictions, we might expect the following properties to hold”:

- “Attention weights should correlate with feature importance measures (e.g., gradient-based measures)”
- Alternative (or counterfactual) attention weight configurations ought to yield corresponding changes in prediction

Experimental Specifics

- Seq2seq encoder → additive attention mechanism → feedforward decoder
- 3 encoder variants: CNN, LSTM, embedding average
- 2 attention types: additive (tanh), scaled dot product
- 3 tasks: binary text classification, question answering (QA), natural language inference (NLI)

Algorithm 1 Feature Importance Computations

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \alpha)$
$g_t \leftarrow |\sum_{w=1}^{|V|} \mathbb{1}[\mathbf{x}_{tw} = 1] \frac{\partial y}{\partial \mathbf{x}_{tw}}| \ , \forall t \in [1, T]$
$\tau_g \leftarrow \text{Kendall-}\tau(\alpha, g)$
$\Delta\hat{y}_t \leftarrow \text{TVD}(\hat{y}(\mathbf{x}_{-t}), \hat{y}(\mathbf{x})) \ , \forall t \in [1, T]$
$\tau_{loo} \leftarrow \text{Kendall-}\tau(\alpha, \Delta\hat{y})$

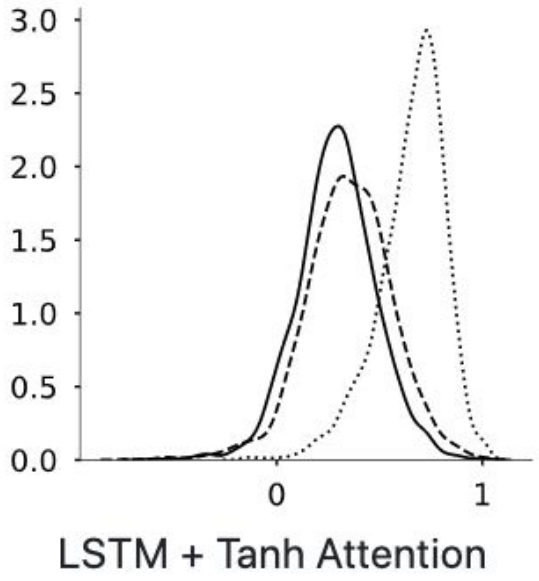
Algorithm 2 Permuting attention weights

$\mathbf{h} \leftarrow \text{Enc}(\mathbf{x}), \hat{\alpha} \leftarrow \text{softmax}(\phi(\mathbf{h}, \mathbf{Q}))$
$\hat{y} \leftarrow \text{Dec}(\mathbf{h}, \hat{\alpha})$
for $p \leftarrow 1$ to 100 **do**
  $\alpha^p \leftarrow \text{Permute}(\hat{\alpha})$
  $\hat{y}^p \leftarrow \text{Dec}(\mathbf{h}, \alpha^p)$ ▷ Note : $\mathbf{h}$ is not changed
  $\Delta\hat{y}^p \leftarrow \text{TVD}[\hat{y}^p, \hat{y}]$
end for
$\Delta\hat{y}^{med} \leftarrow \text{Median}_p(\Delta\hat{y}^p)$

[4]

Jain and Wallace 2019: “Attention is Not Explanation”

- They observe neither property holds for a bidirectional LSTM model



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 $\tilde{\alpha}$

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

[4]

Wiegreffe and Pinter 2019: “Attention is not not explanation”

Existence does not entail exclusivity: “If we use attention scores to explain to a user why a model picked a certain token, shouldn’t we be glad that despite being able to reach the same prediction by giving a high weight to a different token with the same type, the attention mechanism managed to focus on the correct instance?”

Attention distribution is not a primitive: “J&W provide alternative distributions which may result in similar predictions, but in the process they remove the very linkage by which attention modelers motivate the explainability of these distributions, namely the fact that the model was trained to attend to the tokens it chose.”

Jain and Wallace 2019: “Attention is Not Explanation”

“Assuming attention provides a faithful explanation for model predictions, we might expect the following properties to hold”:

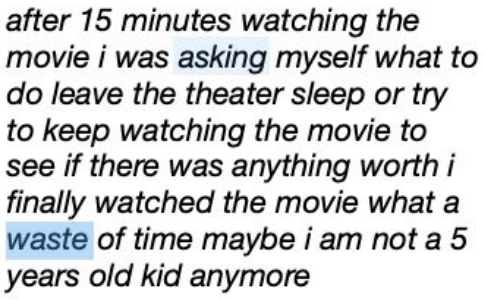
- “Attention weights should correlate with feature importance measures (e.g., gradient-based measures)”
- ~~Alternative (or counterfactual) attention weight configurations ought to yield corresponding changes in prediction~~

So what about the correlation?

“We also acknowledge that irrelevant features may be contributing noise to the Kendall tau measure, thus depressing this metric artificially... it remains a possibility that agreement is strong between attention weights and feature importance scores for the top-k features only (the trouble would be defining this k and then measuring correlation between non-identical sets)” [Jain and Wallace 2019]

Irrelevant Features

- Attention distributions are normally sparse
- Sparsemax [Martins and Astudillo 2016]: “a new activation function similar to the traditional softmax, but able to output sparse probabilities”



[4]

Only two words in this review have significant attention density

Top-k correlations

Generalized version of Kendall tau - Fagin et al. 2013: “Comparing top k lists”

What k to use?

- The number of non-zero elements in the attention weights
- The average sequence length of the dataset

Our contributions

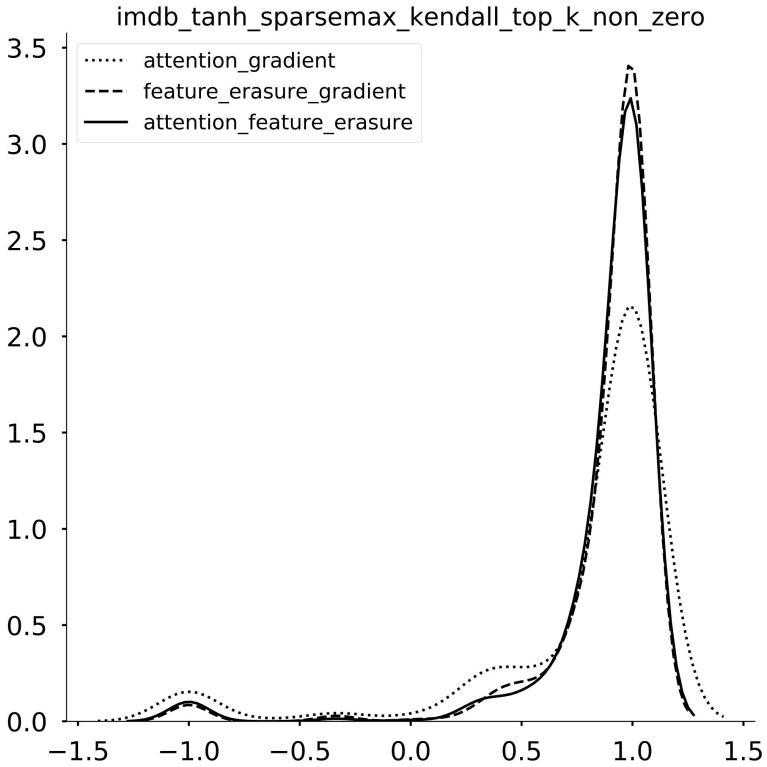
- Demonstrate that attention weights correlate with feature importance measures when:
  - Irrelevant features are removed by using a **sparse attention** distribution
  - Correlation is measured using the top-k features given by the sparse attention
- Thereby also reproduce the first experiment from Jain and Wallace's work
- Write a new, clean and extensible codebase with AllenNLP
- Contextualise our work and theirs with a review of the literature that has been published since Jain and Wallace's work.

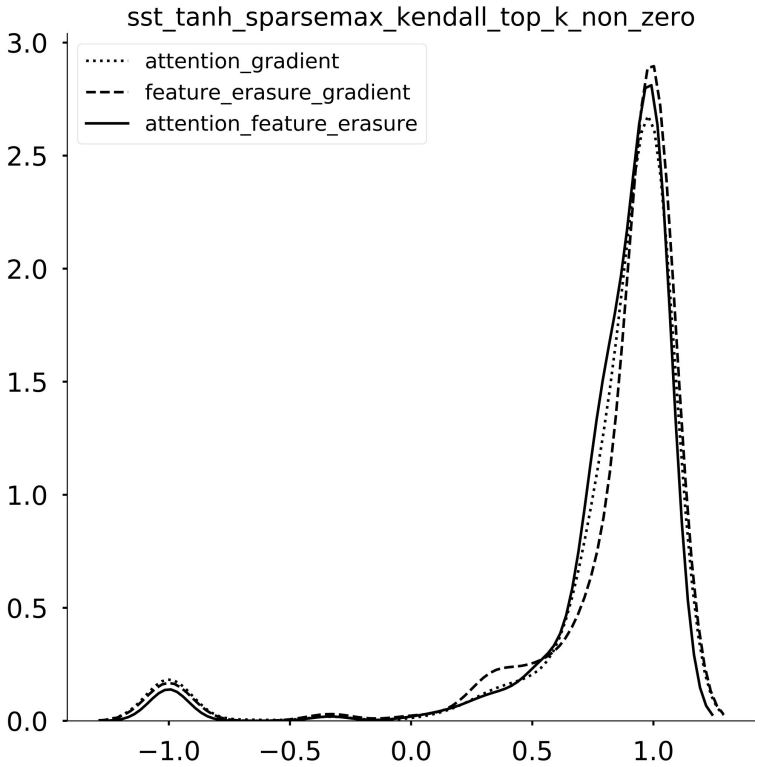
Experiments

- Model: seq2seq BiLSTM encoder → additive attention mechanism → feedforward decoder
- Task: binary sentiment classification
- Datasets: Stanford Sentiment Treebank, IMDB (same splits as J&W)
- Attention types: additive (tanh), scaled dot product
- Attention activation functions: softmax and sparsemax
- Calculate correlation between:
  - Attention weights and difference in model output due to feature erasure
  - Attention weights and gradients
  - Gradients and difference in model output due to feature erasure
- Using:
  - Kendall tau
  - Kendall top-k, k = average sequence length
  - Kendall top-k, k = smallest number of non-zero elements in either list

Results

Results: Top-k





Jain and Wallace 2019: “Attention is Not Explanation”

“Assuming attention provides a faithful explanation for model predictions, we might expect the following properties to hold”:

- ~~“Attention weights should correlate with feature importance measures (e.g., gradient-based measures)”~~
- ~~Alternative (or counterfactual) attention weight configurations ought to yield corresponding changes in prediction~~

Jain and Wallace 2019: “Attention is Not Explanation”

“Assuming attention provides a faithful explanation for model predictions, we might expect the following properties to hold”:

- ~~“Attention weights should correlate with feature importance measures (e.g., gradient-based measures)”~~ ?
- ~~Alternative (or counterfactual) attention weight configurations ought to yield corresponding changes in prediction~~

So… is attention explanation?

Attention in sequence-to-sequence better explanation than in binary classification?

- Observed by both Jain and Wallace, and
- ‘Attention Interpretability Across NLP Tasks’ by Vashishth et al.

So….

“Attention **might be explanation**. It might not be explanation. Whether or not it is depends heavily on the underlying model architecture; on the task; on the sense of “explanation” we’re after.” - Yuval Pinter[6]

It depends on how you define ‘explanation’

“When we have solid problem formulations, flaws in methodology can be addressed by articulating new methods. But when the problem formulation itself is flawed, neither algorithms nor experiments are sufficient to address the underlying problem.” - The Mythos of Model Interpretability [Lipton 2017]

Future Work

- Explore more complex attention mechanisms
- Comparing attention to more advanced measures of ‘explanation’
- Source of noise in attention
  - How does varying the model affect it?
  - How does varying the dataset affect it?
    - Does reliance on the attention mechanism correlate with the amount of noise?

References

- [1] https://gdpr-info.eu/recitals/no-71/
- [2] https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
- [3] https://arxiv.org/abs/1409.0473
- [4] https://arxiv.org/abs/1902.10186
- [5] https://xkcd.com/2237/
- [6] https://medium.com/@yuvalpinter/attention-is-not-not-explanation-dbc25b534017