

The in the interpretability room

Why use attention as explanation when we have saliency methods?



Jasmijn Bastings & Katja Filippova

Introduction

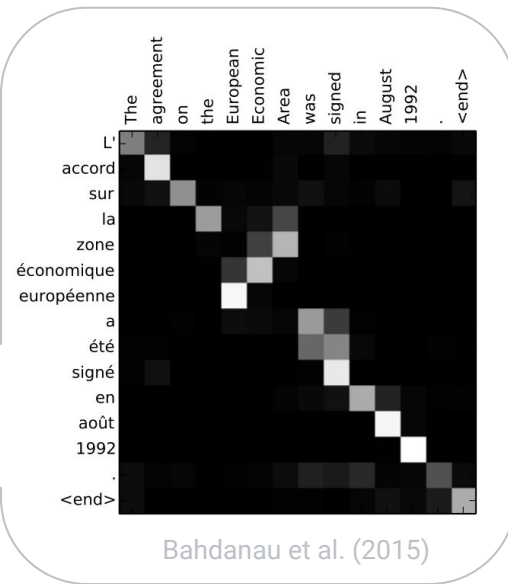
Attention mechanisms have allowed for huge performance gains in NLP.

They offer a **window** into how a model is operating.

Example: word alignment in MT.

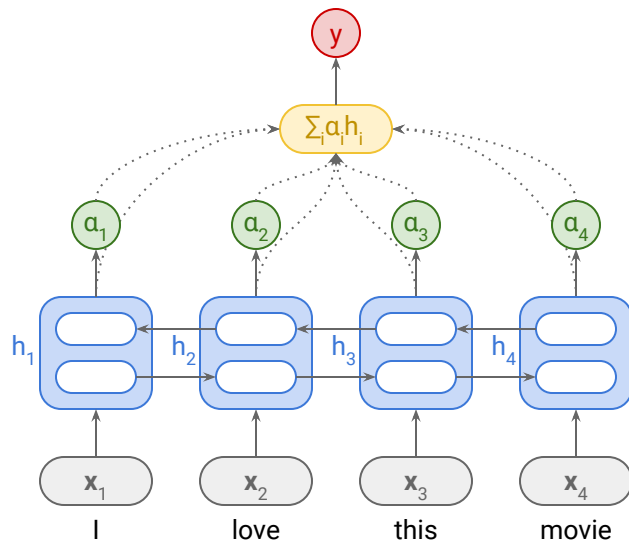
Does that window amount to **explanation**?

Who is the **user**? What is the **goal**?



The Attention Debate

A typical model in the debate



Task: text classification

Model: BiLSTM with MLP-based attention mechanism

Attention is not explanation

Jain & Wallace (2019) show that it is possible to find completely **different attention weights** that result in the **same prediction** on an example.

Attention is often **uncorrelated** with gradient-based feature importance measures.

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

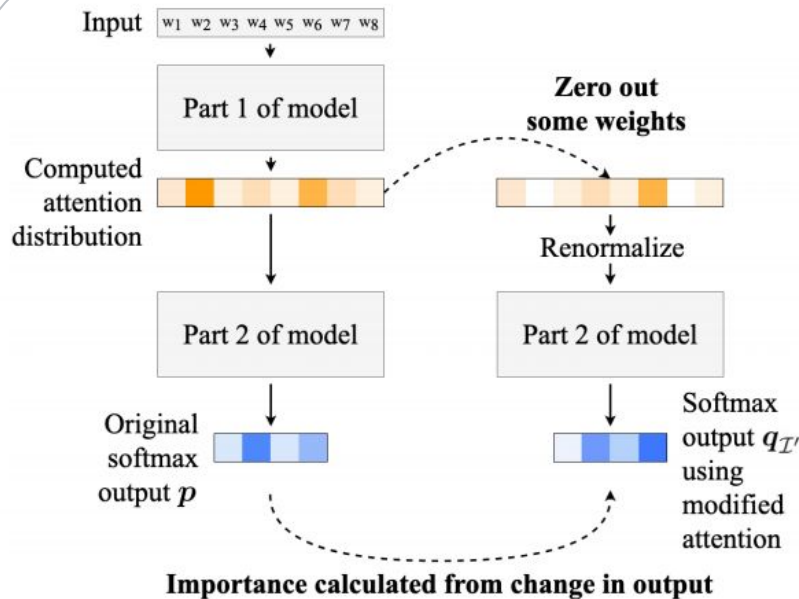
after 15 minutes watching the movie i was asking myself what to do leave the theater sleep or try to keep watching the movie to see if there was anything worth i finally watched the movie what a waste of time maybe i am not a 5 years old kid anymore

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

Is attention interpretable?

Serrano & Smith (2019) **erase** (some) attention weights, and compute importance from the change in output.

*"While we observe some ways in which higher attention weights correlate with greater impact on model predictions, **we also find many ways** in which this does not hold, i.e., where gradient-based rankings of attention weights better predict their effects than their magnitudes."*



Attention is not not explanation

Wiegreffe & Pinter (2019) propose **4 tests** to determine when/whether attention can be used as explanation.

1. a simple uniform-weights baseline
2. a variance calibration based on multiple random seed runs
3. a diagnostic framework using frozen weights from pretrained models
4. an end-to-end adversarial attention training protocol.

“we have confirmed that adversarial distributions can be found for LSTM models in some classification tasks”

Can attention be manipulated?

Pruthi et al. (2020) propose a method that **reduces** how much weight is assigned to a set of 'impermissible' tokens... even when the models demonstratively **rely** on those tokens for their predictions.

Conclusion: we cannot rely on attention as explanation. It can be easily manipulated.

Attention	Biography	Label
Original	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician
Ours	Ms. X practices medicine in Memphis, TN and is affiliated ... Ms. X speaks English and Spanish.	Physician

Is a causal definition assumed?

Grimsley et al. (2020) go as far as saying that **attention is not explanation by definition**, if a **causal** definition of explanation is assumed.

Causal explanations presuppose that a **surgical intervention** is possible which is not the case with deep neural networks.

You **cannot intervene** on attention while keeping all the other variables **invariant**.

Can attention be improved?

Mohankumar et al. (2020) observe **high similarity** among LSTM states and propose a diversity-driven training objective that makes them more diverse.

The resulting attention weights result in decision flips more easily as compared to vanilla attention.

Tutek and Snajder (2020) use a word-level objective to achieve a **stronger connection** between hidden states and the words they represent, which in turn affects the attention weights.

Saliency Methods

Gradient-based methods

1. Li et al. (2016) use the **gradient** of the output (of class c) w.r.t. each input embedding \mathbf{x}_i :

$$\nabla_{\mathbf{x}_i} f_c(\mathbf{x}_{1:n})$$

2. Denil et al. (2015) use **gradient x input**:

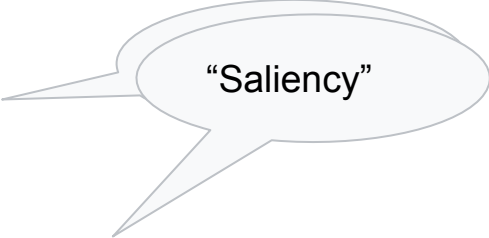
$$\nabla_{\mathbf{x}_i} f_c(\mathbf{x}_{1:n}) \cdot \mathbf{x}_i$$

3. Sundararajan et al. (2017) propose **integrated gradients** (IG):

In m steps, **interpolate** the inputs $\mathbf{x}_{1:n}$ with a baseline $\mathbf{b}_{1:n}$ (e.g., all zeros), **average** those gradients, and multiply with the input (minus baseline).



“Sensitivity”



“Saliency”

Propagation-based methods

Layer-wise Relevance Propagation (LRP) starts with a **forward pass** to obtain the output, which is the top level **relevance**.

They then use a **special backward pass** that, at each layer, redistributes the incoming relevance among the inputs of that layer.

Each kind of layer has its own propagation rules.

For example, there are different rules for feed-forward layers (Bach et al., 2015) and LSTM layers (Arras et al., 2017)

Occlusion-based methods

Occlusion-based methods (Zeiler and Fergus, 2014; Li et al., 2016b) compute input saliency by **occluding** (or **erasing**) input features and measuring how that affects the model.

$$f_c(\mathbf{x}_{1:n}) - f_c(\mathbf{x}_{1:n} \mid \mathbf{x}_i = 0)$$

That is, the saliency of each input is the **difference in output** with and without that input erased.

We can also use this for **evaluating** other methods.

Attention vs. Saliency

Attention vs. Saliency (1/2)

In many of the cited papers, the **goal** of the explanation is to reveal which input words are the most important ones for the prediction.

The intended **user** for the explanation is often not stated, but typically that user is a model developer, and not a non-expert end user.

For model developers, **faithfulness**, the need for an explanation to accurately represent the reasoning of the model, is a key concern.

Attention vs. Saliency (2/2)

Input saliency methods are addressing the goal **head-on!**

They show how relevant each input word was to one particular prediction.

Saliency methods take the **entire computation path** into account, all the way from input to output.

Attention weights do not: they reflect, at one point in the computation, how much the model attends to each input representation, but those representations might *already have mixed in information from other inputs*.

Attention is often evaluated against a gradient-based method, so why not use that method?

While extracting attention is **easy** and only requires a **forward pass**, most saliency methods only require an extra **backward pass**, and are easily implemented.

Final notes & Conclusions

Attention is not not interesting

We criticized the use of attention to assess input saliency for the benefit of the model developer.

Yet, understanding the **role** of the attention mechanism is a perfectly justified research goal!

E.g. finding the role of attention heads in transformers, adding linguistic biases, etc.

For **different goals and users**, attention might become useful as explanation (e.g., Strout et al., 2019).

Is saliency the ultimate answer?

While we have argued that saliency methods are a good fit for our goal, there are other goals for which different methods can be a better fit.

Some examples:

- Counterfactual analysis with visualization tools
- DiffMask (learn with differentiable gates what inputs can be masked out and where information is stored) (De Cao et al., 2020)
- Rationalizing neural predictions (Lei et al., 2016; Bastings et al., 2019)

Limitations of Saliency

- **A known problem with erasure:** changes in output may be due to the fact that **the corrupted input falls off the training manifold** (Hooker et al., 2019).
- Saliency methods are not always reliable and can produce **unintuitive** results (Kindermans et al., 2017).
- Per-token saliency weights can be called an explanation only in a **very narrow sense!**
- Input feature **interactions** are largely unexplored.
- It is hardly possible to fully understand why a deep non-linear model produced a certain prediction by only looking at the input tokens

Conclusions

We summarized the debate on whether attention is explanation, and observed that the **goal** for explanation is often to determine what inputs are the most relevant to the prediction.

The **user** for that explanation is typically assumed to be a **model developer**.

With this goal and user clearly stated, **we argued that input saliency are better suited** than attention.

We hope, at least for the goal and user that we identified, that the focus shifts from attention to input saliency methods, and perhaps to entirely different methods, goals, and users.