

Attention is
interpretable
explanation
not explanation
not not explanation
maybe explanation??

Ádám Kovács

TU Wien

April 13, 2021

Outline

Interpretability

Attention

Self-attention

Attention as explanation

Is Attention Interpretable?

Attention is not explanation

Attention is not not explanation

References

Interpretability, Explainability

From Benedikt's slides:

- ▶ *Interpretability is the degree to which a human can understand the cause of a decision (Tim Miller)*
- ▶ **Faithfulness:** *faithful interpretation is one that accurately represents the reasoning process behind the model's prediction.*
- ▶ LIME, ELI5, SHAP, etc..
- ▶ Traditional ML algorithm can be interpretable, but we still have struggles with black-box DL models

LIME [Ribeiro et al., 2016]

Prediction probabilities



atheism

christian

Posting
0.15
Host
0.14
NNTP
0.11
edu
0.04
have
0.01
There
0.01

Text with highlighted words

From: johnchad@triton.unm.edu (jchadwic)
Subject: Another request for Darwin Fish
Organization: University of New Mexico, Albuquerque
Lines: 11
NNTP-Posting-Host: triton.unm.edu

Hello Gang,

There have been some notes recently asking where to obtain the DARWIN fish.
This is the same question I have and I have not seen an answer on the net. If anyone has a contact please post on the net or email me.

What is attention?

- ▶ In broad terms attention pays greater focus to certain parts of the data
- ▶ Attention can be classified into two classes
- ▶ General attention
 - ▶ between input and output elements
 - ▶ general seq2seq architectures
- ▶ Self-attention
 - ▶ within the input elements
 - ▶ used in Transformer architectures [Vaswani et al., 2017] (BERT, RoBERTa, ALBERT, etc..)

What is attention?

- ▶ In broad terms attention pays greater focus to certain parts of the data
- ▶ Attention can be classified into two classes
- ▶ General attention
 - ▶ between input and output elements
 - ▶ general seq2seq architectures
- ▶ Self-attention
 - ▶ within the input elements
 - ▶ used in Transformer architectures [Vaswani et al., 2017] (BERT, RoBERTa, ALBERT, etc..)

What is attention?

- ▶ In broad terms attention pays greater focus to certain parts of the data
- ▶ Attention can be classified into two classes
 - ▶ General attention
 - ▶ between input and output elements
 - ▶ general seq2seq architectures
 - ▶ Self-attention
 - ▶ within the input elements
 - ▶ used in Transformer architectures [Vaswani et al., 2017] (BERT, RoBERTa, ALBERT, etc..)

What is attention?

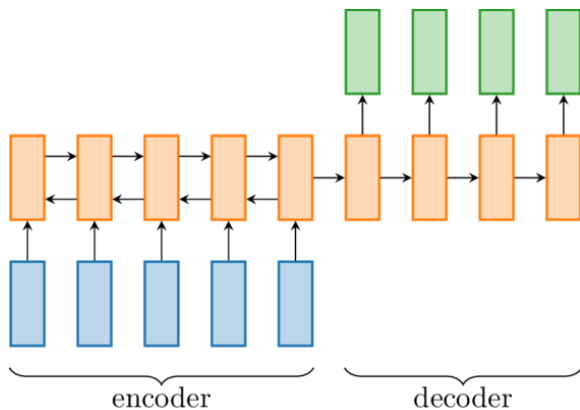
- ▶ In broad terms attention pays greater focus to certain parts of the data
- ▶ Attention can be classified into two classes
- ▶ General attention
 - ▶ between input and output elements
 - ▶ general seq2seq architectures
- ▶ Self-attention
 - ▶ within the input elements
 - ▶ used in Transformer architectures [Vaswani et al., 2017] (BERT, RoBERTa, ALBERT, etc..)

What is attention?

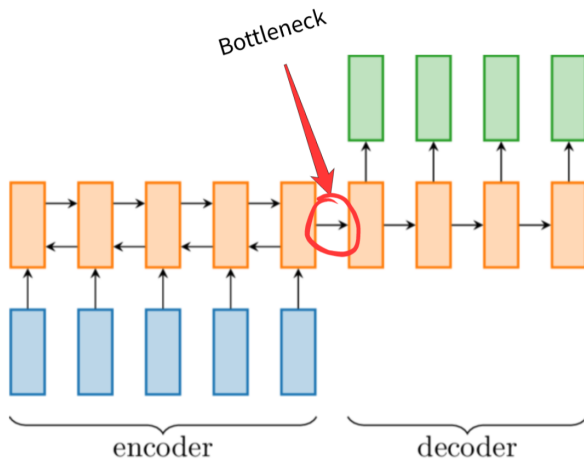
- ▶ In broad terms attention pays greater focus to certain parts of the data
- ▶ Attention can be classified into two classes
- ▶ General attention
 - ▶ between input and output elements
 - ▶ general seq2seq architectures
- ▶ Self-attention
 - ▶ within the input elements
 - ▶ used in Transformer architectures [Vaswani et al., 2017] (BERT, RoBERTa, ALBERT, etc..)

General seq2seq

- ▶ Encoder-Decoder model, popularized in Machine Translation
- ▶ Both the Encoder and the Decoder part are based on RNN structures

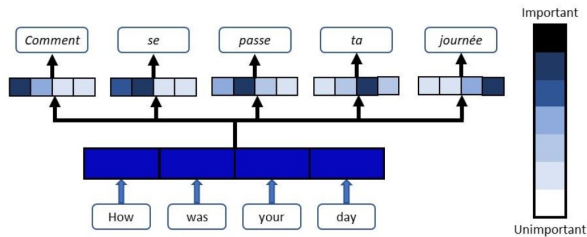


General seq2seq problems



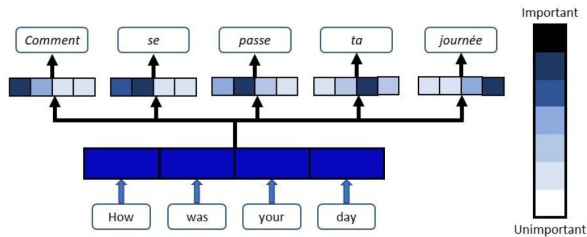
Seq2seq with Attention

- ▶ Attention is an additional layer on top of the encoder RNN structure
- ▶ It will work as a "Query" for the decoder
- ▶ It will assign higher weights to important words
- ▶ These weights assign a score directly to each input



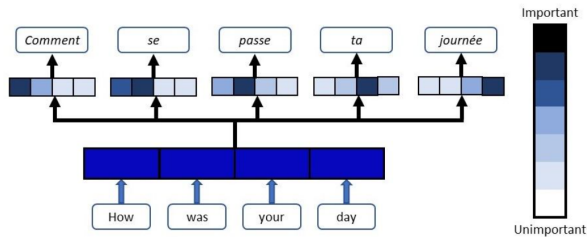
Seq2seq with Attention

- ▶ Attention is an additional layer on top of the encoder RNN structure
- ▶ It will work as a "Query" for the decoder
- ▶ It will assign higher weights to important words
- ▶ These weights assign a score directly to each input



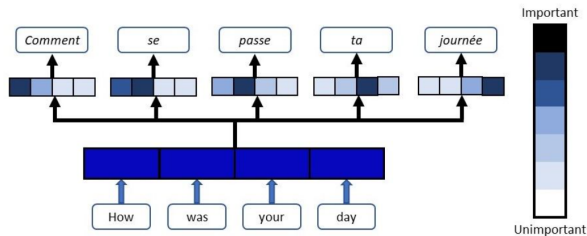
Seq2seq with Attention

- ▶ Attention is an additional layer on top of the encoder RNN structure
- ▶ It will work as a "Query" for the decoder
- ▶ It will assign higher weights to important words
- ▶ These weights assign a score directly to each input



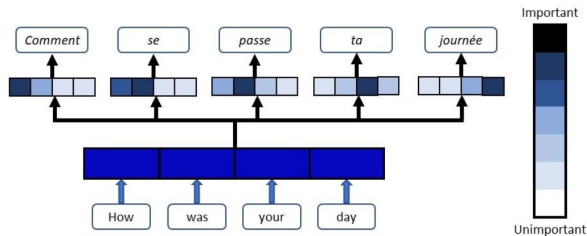
Seq2seq with Attention

- ▶ Attention is an additional layer on top of the encoder RNN structure
- ▶ It will work as a "Query" for the decoder
- ▶ It will assign higher weights to important words
- ▶ These weights assign a score directly to each input

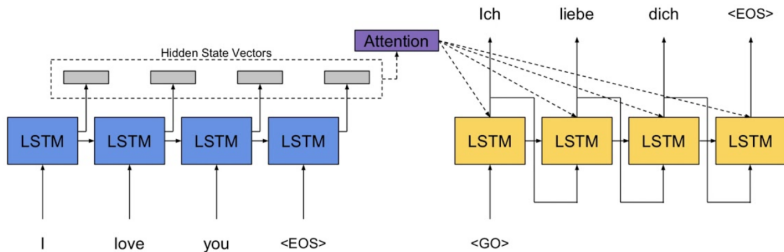


Seq2seq with Attention

- ▶ Attention is an additional layer on top of the encoder RNN structure
- ▶ It will work as a "Query" for the decoder
- ▶ It will assign higher weights to important words
- ▶ These weights assign a score directly to each input

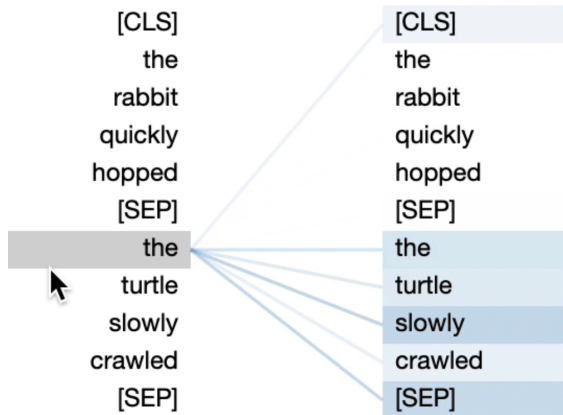


Seq2seq with Attention



Self attention

- ▶ Self attention assigns weights to each input word
- ▶ For each word we query the most important words in context
- ▶ Used mostly for classification and language modeling tasks



Types of attention

- ▶ We have many types of attention
- ▶ 2 different major types of Attention
- ▶ Bahdanau attention (additive attention)
[Bahdanau et al., 2015]
- ▶ Luong attention (multiplicative attention) [Luong et al., 2015]

Bahdanau attention

The process is the following¹:

- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Bahdanau attention

The process is the following¹:

- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Bahdanau attention

The process is the following¹:

- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Bahdanau attention

The process is the following¹:

- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Bahdanau attention

The process is the following¹:

- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Bahdanau attention

The process is the following¹:

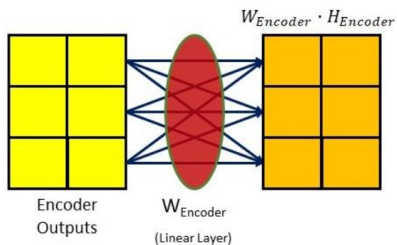
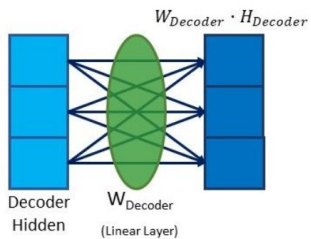
- ▶ Producing the hidden states from the encoder
- ▶ Calculating alignment scores
- ▶ Softmaxing the alignment scores
- ▶ Calculating the context vector
- ▶ Decoding the output

¹The images are from this great blog

Calculating alignment scores

$$score_{alignment} = W_{combined} \cdot \tanh(W_{decoder} \cdot H_{decoder} + W_{encoder} \cdot H_{encoder})$$

Calculating alignment scores



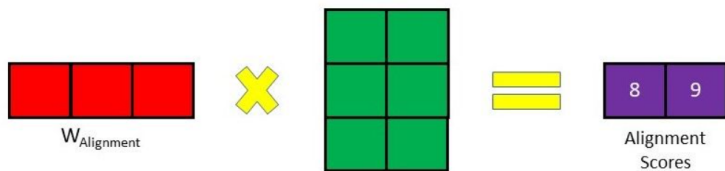
Calculating alignment scores

$$\tanh \left(\begin{matrix} W_{Decoder} \cdot H_{Decoder} & W_{Encoder} \cdot H_{Encoder} \end{matrix} \right) = \begin{matrix} \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \\ \blacksquare & \blacksquare \end{matrix}$$

The diagram illustrates the calculation of alignment scores. It shows the \tanh function applied to the sum of two matrices. The first matrix is a 3x1 blue matrix representing $W_{Decoder} \cdot H_{Decoder}$. The second matrix is a 3x2 orange matrix representing $W_{Encoder} \cdot H_{Encoder}$. The sum of these two matrices is then passed through the \tanh function, resulting in a 3x2 green matrix.

Above outputs combined and \tanh applied

Calculating alignment scores

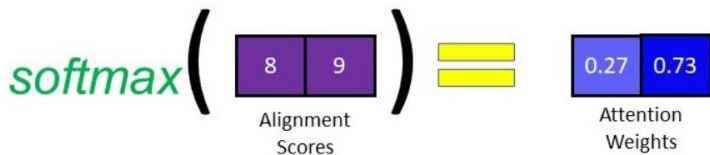


Softmaxing the alignment scores

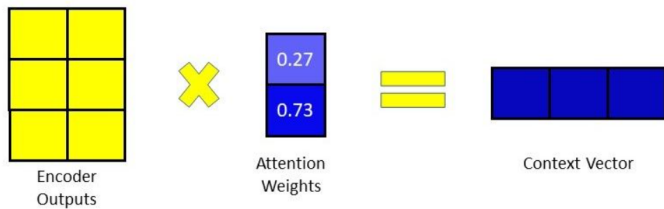
$$\text{softmax} \left(\begin{array}{|c|c|} \hline 8 & 9 \\ \hline \end{array} \right) = \begin{array}{|c|c|} \hline 0.27 & 0.73 \\ \hline \end{array}$$

Alignment Scores

Attention Weights



Calculating context vector



Attention is all you need [Vaswani et al., 2017]

- ▶ Recurrent neural networks are hard to parallelize and train
- ▶ Transformer-based architectures replace RNN-s with self-attention and Linear layers
- ▶ State-of-the art methods in most of the NLP tasks
- ▶ BERT [Devlin et al., 2019], ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], etc..

Attention is all you need [Vaswani et al., 2017]

- ▶ Recurrent neural networks are hard to parallelize and train
- ▶ Transformer-based architectures replace RNN-s with self-attention and Linear layers
- ▶ State-of-the art methods in most of the NLP tasks
- ▶ BERT [Devlin et al., 2019], ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], etc..

Attention is all you need [Vaswani et al., 2017]

- ▶ Recurrent neural networks are hard to parallelize and train
- ▶ Transformer-based architectures replace RNN-s with self-attention and Linear layers
- ▶ State-of-the art methods in most of the NLP tasks
- ▶ BERT [Devlin et al., 2019], ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], etc..

Attention is all you need [Vaswani et al., 2017]

- ▶ Recurrent neural networks are hard to parallelize and train
- ▶ Transformer-based architectures replace RNN-s with self-attention and Linear layers
- ▶ State-of-the art methods in most of the NLP tasks
- ▶ BERT [Devlin et al., 2019], ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], etc..

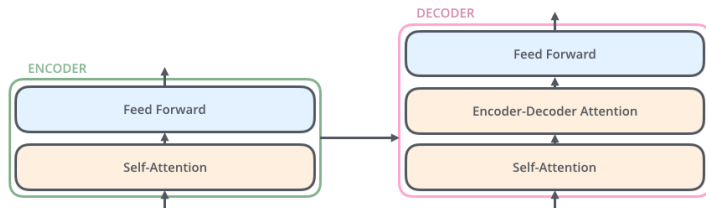
Attention is all you need [Vaswani et al., 2017]

- ▶ Recurrent neural networks are hard to parallelize and train
- ▶ Transformer-based architectures replace RNN-s with self-attention and Linear layers
- ▶ State-of-the art methods in most of the NLP tasks
- ▶ BERT [Devlin et al., 2019], ALBERT [Lan et al., 2019], RoBERTa [Liu et al., 2019], etc..

The architecture

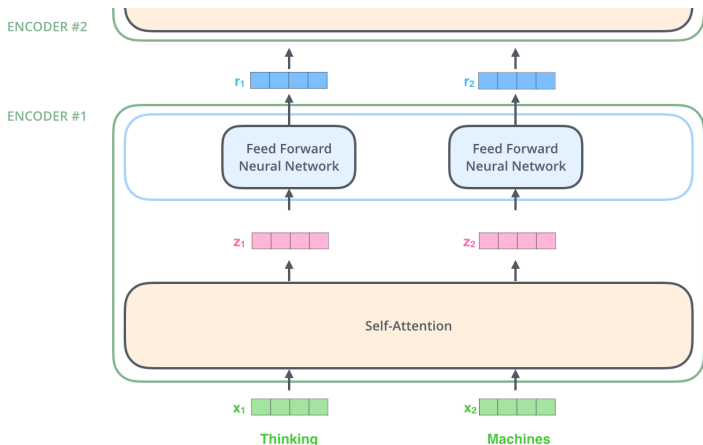
pictures are taken from this post

- ▶ Self-attention in both the encoder and decoder
- ▶ Encoder-Decoder attention can be still present



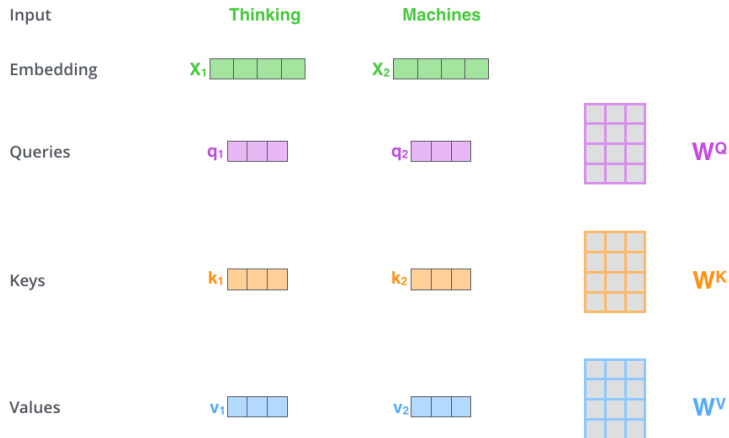
The architecture

- ▶ Attention score is calculated for each word against the other words



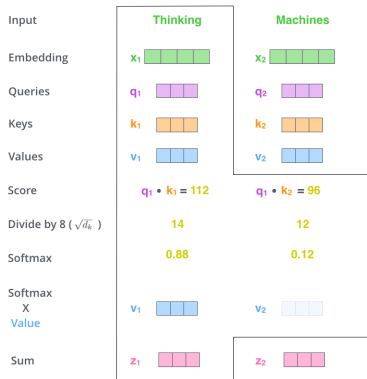
The architecture

- ▶ We have Query, Key, Value trainable matrices



The architecture

1. Dot product of the query vector with the key vector of the respective word we're scoring
2. Softmax and multiply with the value vector
3. Sum the weighted value vectors



The architecture

$$\text{softmax} \left(\frac{\begin{matrix} \mathbf{Q} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix} \times \begin{matrix} \mathbf{K}^T \\ \begin{matrix} \square & \square \\ \square & \square \\ \square & \square \end{matrix} \end{matrix}}{\sqrt{d_k}} \right) \begin{matrix} \mathbf{V} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix} \\ = \begin{matrix} \mathbf{Z} \\ \begin{matrix} \square & \square & \square \\ \square & \square & \square \end{matrix} \end{matrix}$$

Attention as explanation

- ▶ Various attention mechanisms exist
- ▶ Each has the same high-level goal
 - ▶ calculate nonnegative weights for the input components
 - ▶ it should sum to 1
 - ▶ multiply the weights with the representations
 - ▶ sum the resulting vectors into a single representation
- ▶ Attention calculates a distribution over the inputs
- ▶ It has been used as an interpretation of the model
[Wang et al., 2016, Lee et al., 2017, Lin et al., 2017, Ghaeini et al., 2018]

Attention as explanation

- ▶ Various attention mechanisms exist
- ▶ Each has the same high-level goal
 - ▶ calculate nonnegative weights for the input components
 - ▶ it should sum to 1
 - ▶ multiply the weights with the representations
 - ▶ sum the resulting vectors into a single representation
- ▶ Attention calculates a distribution over the inputs
- ▶ It has been used as an interpretation of the model [Wang et al., 2016, Lee et al., 2017, Lin et al., 2017, Ghaeini et al., 2018]

Attention as explanation

- ▶ Various attention mechanisms exist
- ▶ Each has the same high-level goal
 - ▶ calculate nonnegative weights for the input components
 - ▶ it should sum to 1
 - ▶ multiply the weights with the representations
 - ▶ sum the resulting vectors into a single representation
- ▶ Attention calculates a distribution over the inputs
- ▶ It has been used as an interpretation of the model
[Wang et al., 2016, Lee et al., 2017, Lin et al., 2017, Ghaeini et al., 2018]

Attention as explanation

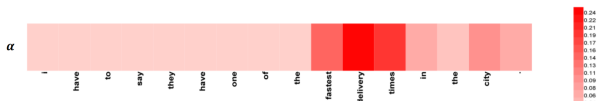
- ▶ Various attention mechanisms exist
- ▶ Each has the same high-level goal
 - ▶ calculate nonnegative weights for the input components
 - ▶ it should sum to 1
 - ▶ multiply the weights with the representations
 - ▶ sum the resulting vectors into a single representation
- ▶ Attention calculates a distribution over the inputs
- ▶ It has been used as an interpretation of the model
[Wang et al., 2016, Lee et al., 2017, Lin et al., 2017, Ghaeini et al., 2018]

Attention as explanation

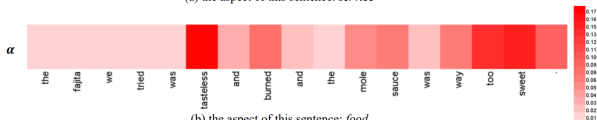
- ▶ Various attention mechanisms exist
- ▶ Each has the same high-level goal
 - ▶ calculate nonnegative weights for the input components
 - ▶ it should sum to 1
 - ▶ multiply the weights with the representations
 - ▶ sum the resulting vectors into a single representation
- ▶ Attention calculates a distribution over the inputs
- ▶ It has been used as an interpretation of the model
[Wang et al., 2016, Lee et al., 2017, Lin et al., 2017, Ghaeini et al., 2018]

Attention as explanation

- ▶ We can look at the local weights for each prediction
- ▶ The weights can serve as an explanation for that specific decision



(a) the aspect of this sentence: *service*



(b) the aspect of this sentence: *food*

Attention as explanation?

Multiple works have appeared that try to understand what these attention weights actually communicate:

- ▶ Is Attention Interpretable? [Serrano and Smith, 2019]
- ▶ Attention is not explanation [Jain and Wallace, 2019]
- ▶ Attention is not not explanation [Wiegreffe and Pinter, 2019]

Is Attention Interpretable?

- ▶ Paper accepted to: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
- ▶ Often assumed that attention identifies information that models found important
- ▶ The paper tests this hypothesis on text classification datasets
- ▶ If a model is interpretable, it must suggest an explanation and ensure that the explanation represents the true reason for the decision

Is Attention Interpretable?

- ▶ Paper accepted to: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
- ▶ Often assumed that attention identifies information that models found important
- ▶ The paper tests this hypothesis on text classification datasets
- ▶ If a model is interpretable, it must suggest an explanation and ensure that the explanation represents the true reason for the decision

Is Attention Interpretable?

- ▶ Paper accepted to: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
- ▶ Often assumed that attention identifies information that models found important
- ▶ The paper tests this hypothesis on text classification datasets
- ▶ If a model is interpretable, it must suggest an explanation and ensure that the explanation represents the true reason for the decision

Is Attention Interpretable?

- ▶ Paper accepted to: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
- ▶ Often assumed that attention identifies information that models found important
- ▶ The paper tests this hypothesis on text classification datasets
- ▶ If a model is interpretable, it must suggest an explanation and ensure that the explanation represents the true reason for the decision

Is Attention Interpretable?

- ▶ Paper accepted to: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics
- ▶ Often assumed that attention identifies information that models found important
- ▶ The paper tests this hypothesis on text classification datasets
- ▶ If a model is interpretable, it must suggest an explanation and ensure that the explanation represents the true reason for the decision

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

- ▶ Take the attention weights as ranking: τ
- ▶ If $i \in \tau$ is higher than $j \in \tau$ then i is more important to the output
- ▶ **Question:** Does τ faithfully describe the output?
[Ghorbani et al., 2018]
- ▶ **Method:** Select $\tau' \subset \tau$
- ▶ Run the model without modification and with modification of the attention weights
- ▶ Modifications: Zero out weights and re-normalize the distribution

The setting

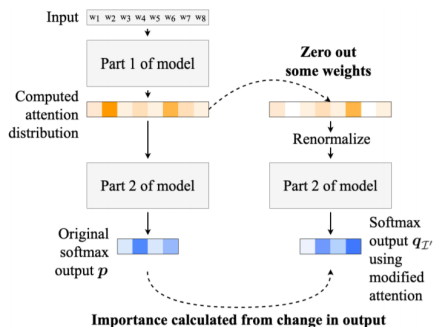


Figure 1: Our method for calculating the importance of representations corresponding to zeroed-out attention weights, in a hypothetical setting with four output classes .

The setting

They trained HAN (Hierarchical Attention Network) based neural networks on classification datasets.

Dataset	Av. # Words	(s.d.)	Av. # Sents.	(s.d.)	# Train. + Dev.	# Test	# Classes
Yahoo Answers	104	(114)	6.2	(5.9)	1,400,000	50,000	10
IMDB	395	(259)	16.2	(10.7)	122,121	13,548	10
Amazon	73	(48)	4.3	(2.6)	3,000,000	650,000	5
Yelp	125	(109)	7.0	(5.6)	650,000	50,000	5

Table 1: Dataset statistics.

Goals

The paper has two main goals:

- ▶ How \mathbf{p} and \mathbf{q} (label distributions) correlate - Jensen-Shannon (JS) divergence between output distributions
- ▶ How the argmaxes of \mathbf{p} and \mathbf{q} differ, indicating a decision flip

Attention weight importance

- ▶ $i^* \in \tau$ is the component with the highest attention, α_{i^*} is its attention
- ▶ Compare the output with removing i^* and with removing r , a randomly drawn variable
- ▶ Use JS divergence on the output distributions
- ▶ Plot this quantity against $\alpha_{i^*} - \alpha_r$

$$\Delta\text{JS} = \text{JS}(\mathbf{p}, \mathbf{q}_{\{i^*\}}) - \text{JS}(\mathbf{p}, \mathbf{q}_{\{r\}})$$

Attention weight importance

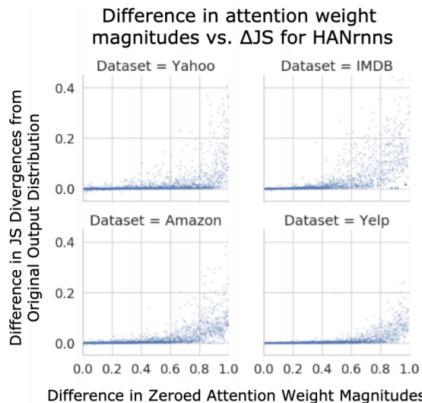


Figure 3: Difference in attention weight magnitudes versus ΔJS for HANrnns, comparable to results for the other architectures; for their plots, see Appendix A.2.

Decision flips

- ▶ The second experiment of the paper looks at the argmaxes of the decisions, indicating a decision flip
- ▶ Zero out α_{j^*} and see if there was a decision flip
- ▶ Zero out α_r and see if there was a decision flip
- ▶ **Result:** in the majority of the cases, zeroing out α_{j^*} does not result in a decision flip

Decision flips

- ▶ The second experiment of the paper looks at the argmaxes of the decisions, indicating a decision flip
- ▶ Zero out α_{j^*} and see if there was a decision flip
- ▶ Zero out α_r and see if there was a decision flip
- ▶ **Result:** in the majority of the cases, zeroing out α_{j^*} does not result in a decision flip

Decision flips

- ▶ The second experiment of the paper looks at the argmaxes of the decisions, indicating a decision flip
- ▶ Zero out α_{j^*} and see if there was a decision flip
- ▶ Zero out α_r and see if there was a decision flip
- ▶ **Result:** in the majority of the cases, zeroing out α_{j^*} does not result in a decision flip

Decision flips

- ▶ The second experiment of the paper looks at the argmaxes of the decisions, indicating a decision flip
- ▶ Zero out α_{j^*} and see if there was a decision flip
- ▶ Zero out α_r and see if there was a decision flip
- ▶ **Result:** in the majority of the cases, zeroing out α_{j^*} does not result in a decision flip

Decision flips

- ▶ The second experiment of the paper looks at the argmaxes of the decisions, indicating a decision flip
- ▶ Zero out α_{j^*} and see if there was a decision flip
- ▶ Zero out α_r and see if there was a decision flip
- ▶ **Result:** in the majority of the cases, zeroing out α_{j^*} does not result in a decision flip

Decision flips

		Remove random: Decision flip?				
		Yahoo		IMDB		
Remove i^* : Decision flip?	Yes	0.5	8.7	Yes	2.2	12.2
	No	1.3	89.6	No	1.4	84.2
			Amazon		Yelp	
	Yes	2.7	7.6	Yes	1.5	8.9
No	2.7	87.1	No	1.9	87.7	

Table 2: Percent of test instances in each decision-flip indicator variable category for each HANrnn.

Importance sets

- ▶ The authors also investigated a set of components in τ
- ▶ How multiple attention weights perform together
- ▶ Setup:
 - ▶ rank attentions by their weights
 - ▶ determine a minimal set that causes a decision flip
 - ▶ the top items are expected to have this characteristic

Importance sets

- ▶ The authors also investigated a set of components in τ
- ▶ How multiple attention weights perform together
- ▶ Setup:
 - ▶ rank attentions by their weights
 - ▶ determine a minimal set that causes a decision flip
 - ▶ the top items are expected to have this characteristic

Importance sets

- ▶ The authors also investigated a set of components in τ
- ▶ How multiple attention weights perform together
- ▶ Setup:
 - ▶ rank attentions by their weights
 - ▶ determine a minimal set that causes a decision flip
 - ▶ the top items are expected to have this characteristic

Importance sets

- ▶ The authors also investigated a set of components in τ
- ▶ How multiple attention weights perform together
- ▶ Setup:
 - ▶ rank attentions by their weights
 - ▶ determine a minimal set that causes a decision flip
 - ▶ the top items are expected to have this characteristic

Importance sets

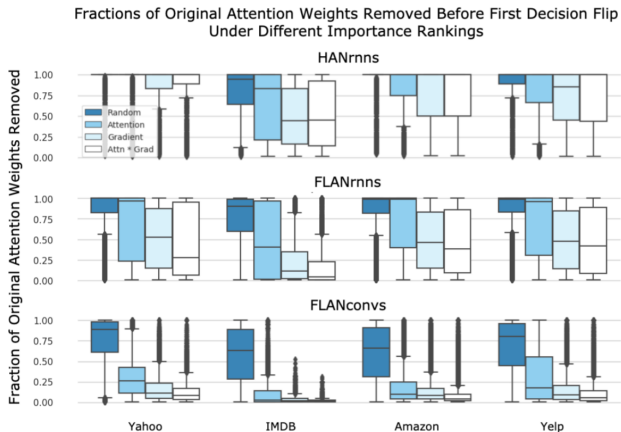


Figure 5: The distribution of fractions of items removed before first decision flips on three model architectures under different ranking schemes. Boxplot whiskers represent the highest/lowest data point within 1.5 IQR of the higher/lower quartile, and dataset names at the bottom apply to their whole column. In several of the plots, the median or lower quartile aren't visible; in these cases, the median/lower quartile is either 1 or very close to 1.

Conclusion

- ▶ the highest attention weights fail to have a large impact
- ▶ in multi-weight tests, we see that attention weights often fail to identify the sets of representations most important to the model's final decision
- ▶ in the papers settings attention is not an optimal method of identifying relevant input elements responsible for the output
- ▶ Attention may yet be interpretable in other ways, but as an importance ranking, it fails to explain model decisions.

Conclusion

- ▶ the highest attention weights fail to have a large impact
- ▶ in multi-weight tests, we see that attention weights often fail to identify the sets of representations most important to the model's final decision
- ▶ in the papers settings attention is not an optimal method of identifying relevant input elements responsible for the output
- ▶ Attention may yet be interpretable in other ways, but as an importance ranking, it fails to explain model decisions.

Conclusion

- ▶ the highest attention weights fail to have a large impact
- ▶ in multi-weight tests, we see that attention weights often fail to identify the sets of representations most important to the model's final decision
- ▶ in the papers settings attention is not an optimal method of identifying relevant input elements responsible for the output
- ▶ Attention may yet be interpretable in other ways, but as an importance ranking, it fails to explain model decisions.

Conclusion

- ▶ the highest attention weights fail to have a large impact
- ▶ in multi-weight tests, we see that attention weights often fail to identify the sets of representations most important to the model's final decision
- ▶ in the papers settings attention is not an optimal method of identifying relevant input elements responsible for the output
- ▶ Attention may yet be interpretable in other ways, but as an importance ranking, it fails to explain model decisions.

Conclusion

- ▶ the highest attention weights fail to have a large impact
- ▶ in multi-weight tests, we see that attention weights often fail to identify the sets of representations most important to the model's final decision
- ▶ in the papers settings attention is not an optimal method of identifying relevant input elements responsible for the output
- ▶ Attention may yet be interpretable in other ways, but as an importance ranking, it fails to explain model decisions.

Attention is not explanation

The main claims of the paper:

- ▶ Correlation between standard feature importance and attention weights are weak
- ▶ Randomly permuting the attention weights doesn't change the output significantly
- ▶ *"These results suggest that while attention modules consistently yield improved performance on NLP tasks, their ability to provide transparency for model predictions is questionable"*

Attention is not explanation

The main claims of the paper:

- ▶ Correlation between standard feature importance and attention weights are weak
- ▶ Randomly permuting the attention weights doesn't change the output significantly
- ▶ *"These results suggest that while attention modules consistently yield improved performance on NLP tasks, their ability to provide transparency for model predictions is questionable"*

Attention is not explanation

The main claims of the paper:

- ▶ Correlation between standard feature importance and attention weights are weak
- ▶ Randomly permuting the attention weights doesn't change the output significantly
- ▶ *"These results suggest that while attention modules consistently yield improved performance on NLP tasks, their ability to provide transparency for model predictions is questionable"*

Attention is not explanation

The main claims of the paper:

- ▶ Correlation between standard feature importance and attention weights are weak
- ▶ Randomly permuting the attention weights doesn't change the output significantly
- ▶ *"These results suggest that while attention modules consistently yield improved performance on NLP tasks, their ability to provide transparency for model predictions is questionable"*

The setting

- ▶ Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.. (text classification tasks using standard encoders with attention mechanism)
- ▶ Empirical analysis between gradient base feature importance and attention
- ▶ Also between 'leave-one-out' (LOO) and attention
- ▶ Generate counterfactual attention distributions that doesn't change the output - \hat{z} attention doesn't provide unique explanation

The setting

- ▶ Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.. (text classification tasks using standard encoders with attention mechanism)
- ▶ Empirical analysis between gradient base feature importance and attention
- ▶ Also between 'leave-one-out' (LOO) and attention
- ▶ Generate counterfactual attention distributions that doesn't change the output - \hat{z} attention doesn't provide unique explanation

The setting

- ▶ Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.. (text classification tasks using standard encoders with attention mechanism)
- ▶ Empirical analysis between gradient base feature importance and attention
- ▶ Also between 'leave-one-out' (LOO) and attention
- ▶ Generate counterfactual attention distributions that doesn't change the output - \hat{z} attention doesn't provide unique explanation

The setting

- ▶ Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.. (text classification tasks using standard encoders with attention mechanism)
- ▶ Empirical analysis between gradient base feature importance and attention
- ▶ Also between 'leave-one-out' (LOO) and attention
- ▶ Generate counterfactual attention distributions that doesn't change the output - \hat{z} attention doesn't provide unique explanation

The setting

- ▶ Data: Common NLP benchmarks like IMdB, 20 News Groups, SST, etc.. (text classification tasks using standard encoders with attention mechanism)
- ▶ Empirical analysis between gradient base feature importance and attention
- ▶ Also between 'leave-one-out' (LOO) and attention
- ▶ Generate counterfactual attention distributions that doesn't change the output - \hat{z} attention doesn't provide unique explanation

Experiment 1

- ▶ Correlation between attention and gradient base importance (τ_g) and LOO (τ_{loo})
- ▶ (These methods are also insufficient for interpreting DL methods, but they might provide feature importance)
- ▶ Measures:
 - ▶ Gradient base methods: Total Variation Distance, JS-Divergence
 - ▶ LOO: model confidence before and after leaving a feature out

Experiment 1

- ▶ Correlation between attention and gradient base importance (τ_g) and LOO (τ_{loo})
- ▶ (These methods are also insufficient for interpreting DL methods, but they might provide feature importance)
- ▶ Measures:
 - ▶ Gradient base methods: Total Variation Distance, JS-Divergence
 - ▶ LOO: model confidence before and after leaving a feature out

Experiment 1

- ▶ Correlation between attention and gradient base importance (τ_g) and LOO (τ_{loo})
- ▶ (These methods are also insufficient for interpreting DL methods, but they might provide feature importance)
- ▶ Measures:
 - ▶ Gradient base methods: Total Variation Distance, JS-Divergence
 - ▶ LOO: model confidence before and after leaving a feature out

Experiment 1

- ▶ Correlation between attention and gradient base importance (τ_g) and LOO (τ_{loo})
- ▶ (These methods are also insufficient for interpreting DL methods, but they might provide feature importance)
- ▶ Measures:
 - ▶ Gradient base methods: Total Variation Distance, JS-Divergence
 - ▶ LOO: model confidence before and after leaving a feature out

Experiment 1

Result: Not *really*

Dataset	Class	Gradient (BiLSTM) τ_g		Gradient (Average) τ_g		Leave-One-Out (BiLSTM) τ_{100}	
		Mean \pm Std.	Sig. Frac.	Mean \pm Std.	Sig. Frac.	Mean \pm Std.	Sig. Frac.
SST	0	0.34 \pm 0.21	0.48	0.61 \pm 0.20	0.87	0.27 \pm 0.19	0.33
	1	0.36 \pm 0.21	0.49	0.60 \pm 0.21	0.83	0.32 \pm 0.19	0.40
IMDB	0	0.44 \pm 0.06	1.00	0.67 \pm 0.05	1.00	0.34 \pm 0.07	1.00
	1	0.43 \pm 0.06	1.00	0.68 \pm 0.05	1.00	0.34 \pm 0.07	0.99
ADR Tweets	0	0.47 \pm 0.18	0.76	0.73 \pm 0.13	0.96	0.29 \pm 0.20	0.44
	1	0.49 \pm 0.15	0.85	0.72 \pm 0.12	0.97	0.44 \pm 0.16	0.74
20News	0	0.07 \pm 0.17	0.37	0.79 \pm 0.07	1.00	0.06 \pm 0.15	0.29
	1	0.21 \pm 0.22	0.61	0.75 \pm 0.08	1.00	0.20 \pm 0.20	0.62
AG News	0	0.36 \pm 0.13	0.82	0.78 \pm 0.07	1.00	0.30 \pm 0.13	0.69
	1	0.42 \pm 0.13	0.90	0.76 \pm 0.07	1.00	0.43 \pm 0.14	0.91
Diabetes	0	0.42 \pm 0.05	1.00	0.75 \pm 0.02	1.00	0.41 \pm 0.05	1.00
	1	0.40 \pm 0.05	1.00	0.75 \pm 0.02	1.00	0.45 \pm 0.05	1.00
Anemia	0	0.47 \pm 0.05	1.00	0.77 \pm 0.02	1.00	0.46 \pm 0.05	1.00
	1	0.46 \pm 0.06	1.00	0.77 \pm 0.03	1.00	0.47 \pm 0.06	1.00
CNN	Overall	0.24 \pm 0.07	0.99	0.50 \pm 0.10	1.00	0.20 \pm 0.07	0.98
bAb1 1	Overall	0.25 \pm 0.16	0.55	0.72 \pm 0.12	0.99	0.16 \pm 0.14	0.28
bAb1 2	Overall	-0.02 \pm 0.14	0.27	0.68 \pm 0.06	1.00	-0.01 \pm 0.13	0.27
bAb1 3	Overall	0.24 \pm 0.11	0.87	0.61 \pm 0.13	1.00	0.26 \pm 0.10	0.89
SNLI	0	0.31 \pm 0.23	0.36	0.59 \pm 0.18	0.80	0.16 \pm 0.26	0.20
	1	0.33 \pm 0.21	0.38	0.58 \pm 0.19	0.80	0.36 \pm 0.19	0.44
	2	0.31 \pm 0.21	0.36	0.57 \pm 0.19	0.80	0.34 \pm 0.20	0.40

Table 2: Mean and std. dev. of correlations between gradient/leave-one-out importance measures and attention weights. *Sig. Frac.* columns report the fraction of instances for which this correlation is statistically significant; note that this largely depends on input length, as correlation does tend to exist, just weakly. Encoders are denoted parenthetically. These are representative results; exhaustive results for all encoders are available to browse online.

Experiment 2

- ▶ Scrambling the attention weights
- ▶ Re-assign each value to a randomly sampled input
- ▶ Also, generate an *adversarial attention distribution*
 - ▶ set of attention weights maximally distinct from the original weights
 - ▶ but yields the same prediction

Experiment 2

- ▶ Scrambling the attention weights
- ▶ Re-assign each value to a randomly sampled input
- ▶ Also, generate an *adversarial attention distribution*
 - ▶ set of attention weights maximally distinct from the original weights
 - ▶ but yields the same prediction

Experiment 2

- ▶ Scrambling the attention weights
- ▶ Re-assign each value to a randomly sampled input
- ▶ Also, generate an *adversarial attention distribution*
 - ▶ set of attention weights maximally distinct from the original weights
 - ▶ but yields the same prediction

Experiment 2

- ▶ Scrambling the attention weights
- ▶ Re-assign each value to a randomly sampled input
- ▶ Also, generate an *adversarial attention distribution*
 - ▶ set of attention weights maximally distinct from the original weights
 - ▶ but yields the same prediction

Experiment 2

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 $\tilde{\alpha}$

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

Figure 1: Heatmap of attention weights induced over a negative movie review. We show observed model attention (left) and an adversarially constructed set of attention weights (right). Despite being quite dissimilar, these both yield effectively the same prediction (0.01).

Experiment 2

- ▶ Attention Permutation: Authors were able to randomly permute attention weights without significantly changing the output
- ▶ Adversarial Attention: Authors also were able to perturb the original attention without significantly affecting the output

Attention is not not explanation

- ▶ One month later a work by Sarah Wiegreffe and Yuval Pinter has appeared [Wiegreffe and Pinter, 2019]
- ▶ Raises issues regarding the experiments of [Jain and Wallace, 2019]

Issues

- ▶ Kendall-tau measures are unfavorable to contextual models - might be the reason why averaged models performed better
- ▶ Attention scores are claimed to provide *an* explanation, not *the* explanation
- ▶ *"my explanation for why it's raining today may involve the ocean streams, atmospheric pressure, cloud formations. An alternative explanation could cite anger from the god of thunder. It yields the same prediction, but I wouldn't call it equally plausible."*

Issues

- ▶ Kendall-tau measures are unfavorable to contextual models - might be the reason why averaged models performed better
- ▶ Attention scores are claimed to provide *an* explanation, not *the* explanation
- ▶ *"my explanation for why it's raining today may involve the ocean streams, atmospheric pressure, cloud formations. An alternative explanation could cite anger from the god of thunder. It yields the same prediction, but I wouldn't call it equally plausible."*

Issues

- ▶ Kendall-tau measures are unfavorable to contextual models - might be the reason why averaged models performed better
- ▶ Attention scores are claimed to provide *an* explanation, not *the* explanation
- ▶ *"my explanation for why it's raining today may involve the ocean streams, atmospheric pressure, cloud formations. An alternative explanation could cite anger from the god of thunder. It yields the same prediction, but I wouldn't call it equally plausible."*

Issues

- ▶ Kendall-tau measures are unfavorable to contextual models - might be the reason why averaged models performed better
- ▶ Attention scores are claimed to provide *an* explanation, not *the* explanation
- ▶ *"my explanation for why it's raining today may involve the ocean streams, atmospheric pressure, cloud formations. An alternative explanation could cite anger from the god of thunder. It yields the same prediction, but I wouldn't call it equally plausible."*

Conclusion

- ▶ Attention can mean a lot of things
- ▶ Attention as a sanity check: the first paper cares about this.
 - ▶ "we, who built the (say) translation model, have an idea which words in the source text "should" map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."
- ▶ Attention as a tool: the second cares about this
 - ▶ "the model [...] tells us through attention which part of the text caused it to make the prediction."

Conclusion

- ▶ **Attention can mean a lot of things**
- ▶ Attention as a sanity check: the first paper cares about this.
 - ▶ "we, who built the (say) translation model, have an idea which words in the source text "should" map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."
- ▶ Attention as a tool: the second cares about this
 - ▶ "the model [...] tells us through attention which part of the text caused it to make the prediction."

Conclusion

- ▶ Attention can mean a lot of things
- ▶ Attention as a sanity check: the first paper cares about this.
 - ▶ "we, who built the (say) translation model, have an idea which words in the source text "should" map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."
- ▶ Attention as a tool: the second cares about this
 - ▶ "the model [...] tells us through attention which part of the text caused it to make the prediction."

Conclusion

- ▶ Attention can mean a lot of things
- ▶ Attention as a sanity check: the first paper cares about this.
 - ▶ "we, who built the (say) translation model, have an idea which words in the source text "should" map to which words in the target text, and it would be a neat demo if a component in the model shows us exactly the patterns we expect."
- ▶ Attention as a tool: the second cares about this
 - ▶ "the model [...] tells us through attention which part of the text caused it to make the prediction."

Thank you for your *attention!*

- ▶ <https://medium.com/@byron.wallace/thoughts-on-attention-is-not-not-explanation-b7799c4c3b24>
- ▶ <https://medium.com/@yuvalpinter/attention-is-not-not-explanation-dbc25b534017>
- ▶ <http://jalammar.github.io/illustrated-transformer/>

References I



Bahdanau, D., Cho, K., and Bengio, Y. (2015).
Neural machine translation by jointly learning to align and translate.
In International Conference on Learning Representations (ICLR 2015).



Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019).
BERT: Pre-training of deep bidirectional transformers for language understanding.
In Proc. of NAACL.



Ghaeini, R., Fern, X., and Tadepalli, P. (2018).
Interpreting recurrent and attention-based neural models: a case study on natural language inference.
In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 4952–4957, Brussels, Belgium. Association for Computational Linguistics.



Ghorbani, A., Abid, A., and Zou, J. (2018).
Interpretation of neural networks is fragile.



Jain, S. and Wallace, B. C. (2019).
Attention is not explanation.



Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019).
ALBERT: A lite BERT for self-supervised learning of language representations.
CoRR, abs/1909.11942.



Lee, J., Shin, J.-H., and Kim, J.-S. (2017).
Interactive visualization and manipulation of attention-based neural machine translation.
In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 121–126, Copenhagen, Denmark. Association for Computational Linguistics.



Lin, Z., Feng, M., dos Santos, C. N., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017).
A structured self-attentive sentence embedding.

References II



Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019).

RoBERTa: A robustly optimized bert pretraining approach.



Luong, T., Pham, H., and Manning, C. D. (2015).

Effective approaches to attention-based neural machine translation.

In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421. Association for Computational Linguistics.



Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).

"why should i trust you?": Explaining the predictions of any classifier.



Serrano, S. and Smith, N. A. (2019).

Is attention interpretable?



Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017).

Attention is all you need.

In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.



Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016).

Attention-based LSTM for aspect-level sentiment classification.

In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615, Austin, Texas. Association for Computational Linguistics.



Wiegrefe, S. and Pinter, Y. (2019).

Attention is not not explanation.