Is Attention Explanation in Transformers?

Andriy Drozdyuk
Carleton University

Parsa Vafaie
University of Ottawa
School of Engineering

Nkechinyere Ogbuagu
University of Ottawa
School of Engineering

Amirhossein Hajianpour
University of Ottawa
School of Engineering

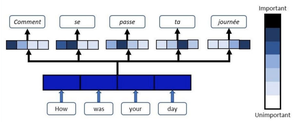
Abstract

Explanation in many Natural Language Processing (NLP) models have been attributed to attention mechanisms. The basis of the claim is due to the fact that attention provides an importance distribution over the input data, identifying the input unit(s) responsible for the output. To challenge this claim, adversarially generated attention weights, different from the model computed attention weights by a threshold, were applied to a variety of NLP tasks. These experiments proved that different weights can produce the same output and therefore should not be considered meaningful explanation. However, experiments were only carried out on sequential models. The transformer model is a non sequential model which at its core consists of multiple self attention heads. We experiment with two approaches to perturbing the attention weights at each layer to test the assumption of transparency. We find that perturbation in different heads at various layers does not produce much effect on the prediction.

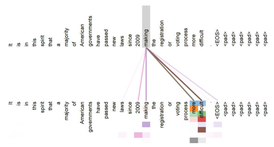
Attention

Attention is a mechanism for measuring interdependence. Attention induces a conditional distribution over input units to compose a weighted vector of importance.

General Attention : Between Input & Output



Self Attention : Among Input Elements



Is Attention Explanation?

Answering this question constitutes an expectation that the following property holds: alternative attention weight distributions yield corresponding changes in prediction. This property assumes that attention weights are *identifiable* i.e. unique.

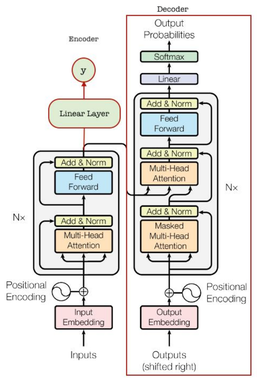
There are two approaches to validating the property above:

1. Attention Permutation
2. Adversarial Attention

Generally, the identifiability of a set of parameters rests on the absence of collinearity in the input data. However, inputs to NLP tasks are encoded in word embeddings with collinear relations.

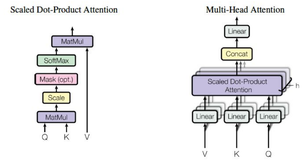
Output Interpretability is dependent on Attention Weights Identifiability which in turn is dependent onInput Non-Collinearity

Transformers



$$\text{Attention}(Q,K,V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$$

$$\text{MultiHead}(Q,K,V) = \text{Concat}(\text{head}_1,...,\text{head}_h)W^O$$
$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$



Self Attention produces contextual word embeddings, aggregating contextual meaning information into input word embeddings.

Multihead Attention allows the transformer to simultaneously attend to information from different representations at different positions.

Is Attention Explanation in Transformers?

In transformers specifically, attention weights for an attention head are identifiable if they can be uniquely identified from the head's output.

$$\text{Attention}(Q,K,V)H = AEW^VH = AT$$

$$\text{rank}(T) \leq \min\left(\text{rank}(E), \text{rank}(W^V), \text{rank}(H)\right) \leq \min(d_s, d, d, d_v, d_v, d) = \min(d_s, d_v).$$

$$\text{null}(T) = \{\tilde{x}^T \in \mathbb{R}^{1\times d_s} | \tilde{x}^T T = 0\}$$

$$\text{for } \tilde{A} = [\tilde{x}_1, \tilde{x}_2, ..., \tilde{x}_{d_s}]^T \text{ where } \tilde{x}_i^T \text{ is any vector in this null space,}$$
$$(A+\tilde{A})T = AT.$$

Due to the Rank Nullity theorem, the dimension of the null space of T is

$$\dim(\text{null}(T)) = d_s - \text{rank}(T) \geq d_s - \min(d_s, d_v) = \begin{cases} d_s - d_v, & \text{if } d_s > d_v \\ 0, & \text{otherwise} \end{cases}$$

Self Attention is only identifiable when the dimension of the null space of T is zero, which occurs when the sentence length is less than or equal to attention head dimension.

Following the question "Is Attention Explanation", we apply the two approaches for validating output interpretability using attention weights to transformers.

1. Attention Permutation : If we shuffle the weights at each layer randomly, would we get the same results?

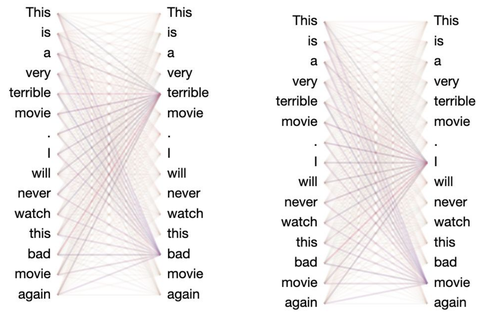
2. Adversarial Attention : Can we find adversarial weights such that we get [almost] identical outputs for different weight distributions distinct by at most an epsilon?

Results

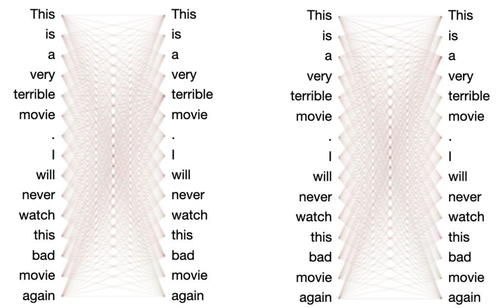
Permutation

Permuting k drops the accuracy of the model by only 6 percent.

Shuffling the weights of layers other than first layer does not change the accuracy.



Both pictures are from the first layer. We can see that the left picture provides explanation, while the right picture provides meaningless attention. The results of the left attention is negative, while the right one is positive.



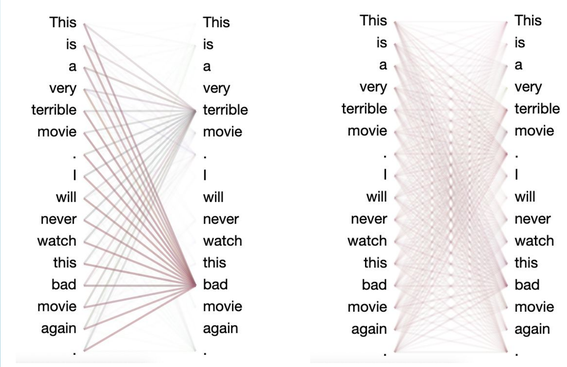
Changing the weights of other layers does not have any effect on the final results. We can see that in other layers, there is no meaningful explanation with the attention weights, and thus, changing them will not change this fact.

Adversarial

We train a new set of Keys which produce the same output, but are different from the original set of Keys.
Our adversarial model is a simple MLP with a relu activation with a hidden size of 256. It takes in a Keys of shape: (batch, num_layers, seq_len, d_model) and produces a new tensor of the same size. The loss it is trained on is given by:

$$\mathcal{L} = \alpha\mathbf{BCE}(\mathbf{y}, \hat{\mathbf{y}}) - (1-\alpha)||k - k_{adv}||_2$$

The first term is the classification loss when using the adversarial keys in the transformer and the second term is the euclidean distance between the original and adversarial keys, which is being maximized.



The results demonstrate that we can have a different set of keys for each input but still get the same classification. The adversarial training seems to remove noise from the keys.

Conclusion and Future

Randomly shuffling the Key weights can provide the same output most of the time
By using adversarial training, we can get the same output, but with a different set of Keys. The adversarial keys seem to be similar to the original but without the noise.
It seems that the adversarial training is able to focus in on the important words, and it would be interesting to see in the future if we can provide explanation by using adversarial training.
Another thing we observed is that the original loss can actually be decreased by adversarial training, so we plan to investigate whether adversarial training can improve transformer performance.

References

1. Jain S., Wallace B. C.(2019). Attention is not Explanation. NAACL 2019
2. Vaswani A. et al. C.(2017). Attention Is All You Need. NeurIPS 2017
3. Brunner G. et al. C.(2019). On Identifiability In Transformers.
4. Bellman R., Astrom K.J., C.(1970). On Structural Identifiability. Mathematical Biosciences Volume 7, Issues 3–4, Pages 329-339