An Exploration of Local Neural NLP Explainability Methods

Abhishek Jain,
aj26@illinois.edu

January 4, 2022

Abstract

Deep learning has revolutionized the Natural Language Processing (NLP) area and is able to solve challenging NLP tasks with state-of-the-art results. But at the same time due to black box nature of deep learning models explainability of this models has been challenging and is an active area of research. It is very important to build explainable systems since it drives building trust, fairness and fidelity in the systems. This literature review explores methods and techniques which are centered towards explainability of deep neural NLP models with an special emphasis on local methods.

1 Introduction

In recent years Natural Language Processing (NLP) has seen a major development and it can be credited to the advances in deep learning techniques. Neural NLP architectures have seen a subtle paradigm shift from RNNs, LSTM/GRUs, Attention/Self-Attention and recently to state-of-the-art Transformer based implementations. Due to the black box nature of these deep learning architecture, explainability of these neural NLP models has always been challenging. A culture of user adoption, fairness and trust can be built in these models only if these are made transparent and explainable. This literature review aim to discuss in detail about different neural NLP model explainability techniques with an emphasis on local explainability which enables highlighting important features contributing to the attribution towards the prediction.

This literature review discusses 4 different research papers published in the area of neural NLP explainability. Each paper explores and discusses a unique technique to solve one of different neural NLP architecture style. This is purposely done to review different styles of attribution techniques. Individual research paper is reviewed and details are provided for a) motivation behind research paper selection b) brief summary of the reviewed paper and observations c) assessment of the research and d) contribution of the paper towards neural NLP explainability.

2 Motivation

- Why **Explainability**? NLP models explainability is important from both engineering and business perspective. During model engineering process explainability can help data scientists and machine learning engineers with root cause analysis over incorrect predictions. Also, similar improvements are possible during the user feedback analysis process. This can help render new possibilities for model improvement. This is even more important from a business perspective as explainability develops trust in the tools which further drives user adoption. Beyond this unexplained model can lead in building vulnerable systems which has biases and fairness challenges. So, a transparent, interpretable and explainable NLP systems can not only improve the performance outcomes but will also be more trusted and adopted for real world problems.

- Why **Neural NLP** Explainability? Neural NLP architectures leveraged by modern systems faces challenges with respect to explainability and it is an extensive area of research today. Research scientists have come up with different methods and techniques to open the black box deep learning models and make it more transparent. There are different types of attribution methods like attention, gradient and perturbation based which can be employed to different neural NLP architecture styles.

- Why **Local** Neural NLP Explainability? Global techniques a more classical styled method where expectation is the whole model is made interpretable and decision trees like representation can be drawn. Global approach is too restrictive and challenging. Currently ML community has gravitated towards the point wise notion of local techniques.

3 Background

Model explainability is broad concept of understanding and analyzing input and output to determine which input features and their interactions are important in determining the prediction output. At high level these can be categorized into two different types a) based on whether the process generates explanation for a given sample or holistic model representation b) based on processing whether there is a need for addition processing unit.

- **Global vs Local**: Local explanation (Danilevsky et al., 2020) techniques provides details of prediction rationals for an individual instance. These techniques help us understand why a model has predicted a result and what factors contributes to prediction at what degree. While Global techniques (Danilevsky et al., 2020) deals with internals of the model and try to come up with more generic results representing model in a more holistic way.

- **Self-explaining vs Post-hoc**: Self-explaining (Danilevsky et al., 2020) directly explains the model and does not require any additional step (model) for explaining the reasons. While Post-hoc techniques (Danilevsky et al., 2020) requires an additional processing step, another model, for explanations.

Neural NLP architecture based on deep learning employs different architecture styles like RNN, LSTMs, Attentions and Transformers. Explainability of neural NLP architecture is challenging and requires different types of attribution method for explainability. Below are different methods (Danilevsky et al., 2021):

- **Attention based techniques** (Bahdanau et al., 2014): Attention principle has revolutionized the NLP problem solving techniques and is one of the extensively used style in different NLP architectures. Attention helps models highlight the important input parameters which is leading to a certain outcome. In this type of methods attention weights are explored in combination with other model features.

- **Perturbation methods**: Perturbation is used for local explanation for a particular instance and is often combined with surrogate model. In this methodology where the model function F is not explainable another surrogate explainable model function G is trained using perturbed instances. These perturbed instances are generated using sampling of training data set or using data generation functions.

- **Gradient methods**: In this technique attributions are calculated using the gradients and limiting it to the first order derivatives. As per this for a non linear model function f(x) the attribution of xi feature is calculated by multiplying xi with partial derivative of f(xi). A special type of gradient method, *Saliency* (Ding and Koehn, 2021) is used to interpret a specific prediction made by a neural network model by assigning a distribution of importance over the input feature set.

- **Guided back propagation**: Here attributions are derived by tuning the back propagation in network. Negative values of derivatives are filtered out when passing through relu activation.

4 Neural NLP Explainability Methods

4.1 Attention Saliency

Interpreting Recurrent and Attention-Based Neural Models: a Case Study on Natural Language Inference (Ghaeini et al., 2018)

- **Motivation**: This research work focuses on use of attention based techniques to interpret complex NLI (Natural Language Inference) tasks. Beyond attention, it examines and proposes attention saliency method for improved inferences. Given attention is one of the widely used NLP explanability method, this research work discusses additional improvements over attentions which can be used.

- **Summary**: In this paper authors have considered attention based modeling for explainability and further introduced new technique which can be augmented with attention for neural NLP models. They have proposed a) *attention with saliency* and b) *LSTM gating signals*. Both these modelling techniques are studied to understand the importance of input words towards the final decision. For the experimentation, authors has used a complex Natural Language Inference (NLI) task and compared the performance with respect to an attention based baseline model.

  In the paper saliency was experimented over the attentions and observation were recorded. It is quite evident the proposed method, attention with saliency, clearly stands out in inferring the importance of input with respect to the prediction result.

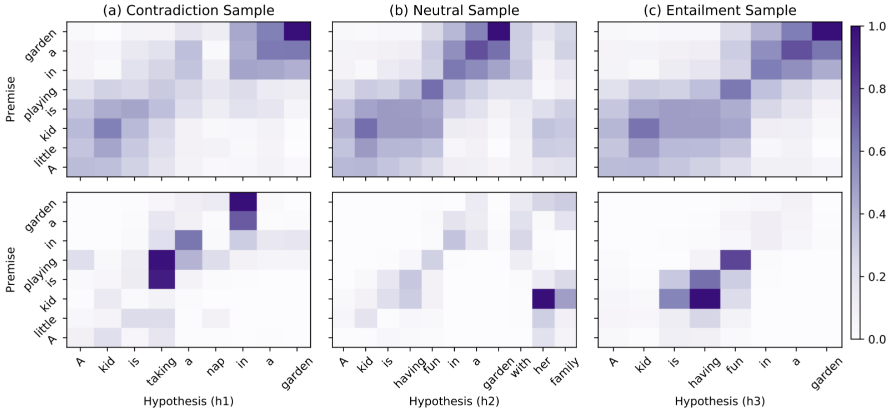


Figure 1: Attention vs Attention Saliency Outcomes (Ghaeini et al., 2018)

Figure 1 represents the comparison of attention based baseline vs attention saliency based visualization. First row in the fig represents attention based visualization and second row represents attention saliency based results. It can be clearly observed that attention model has similar impact for all cases and is independent of outcome but attention with saliency learns and highlight different features which are inline with the outcomes.

- **Interpretations**: Below are some interpretations derived based on this research paper review:
  - Though attentions are important feature that model learn and it helps in predictions but these cannot be used for explainability needs.
  - When attention gradients, Saliency, are observed to improve the explainability outcomes.]
  - Since this method is based totally on attention it cannot be applied to traditional deep learning based on RNNs, LSTMs and GRUs.
  - This style can be used for architecture which uses attention mechanism.
- **Contribution**: This research work leverages two attribution techniques a) attention and b) gradient and suggest a better explainability technique for attention based neural NLP models.

4.2 Self-Attention Attribution

Self-Attention Attribution: Interpreting Information Interactions Inside Transformer Hao et al. (2021)

- **Motivation**: In recent days all state-of-the-art implementations of NLP architectures are based on transformers and its explainability is very critical for adoption of these models in real world applications. It is challenging to interpret the self attention distribution learned in transformer models. This paper claims to have solved the problem of transformers explanability and hence considered in this literature review.
- **Summary**: This paper proposes a self-attention attribution framework to interpret interactions inside the transformer model. When transformer models are trained the higher attention score

learned does not mean the respective input element pair are important in final prediction. Rather the attribution of self-attention score to final score is important. Further transformer models like BERT which has multi-heads analysing and interpreting these multi-heads is far more challenging. This paper further proposes a pruning technique to prune heads based on attribution score.

$$\mathrm{A\tilde{t}tr}_h(A) = \frac{A_h}{m} \odot \sum_{k=1}^{m} \frac{\partial \mathrm{F}(\frac{k}{m}A)}{\partial A_h}$$

Figure 2: Attribution Score (Hao et al., 2021) mathematical representation. For the h-th attention head, attribution score matrix is calculated using above formula where h-th head's attention weight matrix and differential computes the gradient of model F.

Experiments were conducted over BERT model and below results were observed:

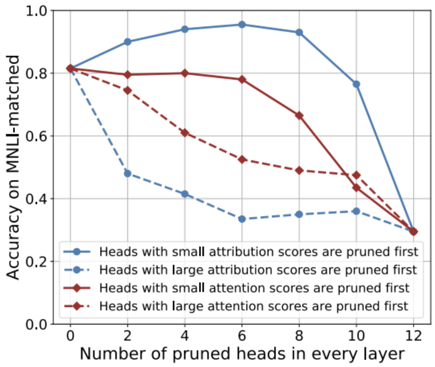


Figure 3: Self-Attention attribution vs attention BERT results: (Hao et al., 2021) Attribution score based results showed significant improvement over attention score. Also, attention head pruning based on score values improves the performance further.

- **Interpretation**:
  - Self-attention attribution method best fit for self-attention and transformer based neural NLP architecture explainability.
  - In multi-headed self attention architecture not all heads are important for explainability. Some heads can be pruned based on self-attention attribution scores.
- **Contribution**: This research paper introduces a new methodology specifically for state-of-the-art self-attention based architectures which are extensively used today for different NLP tasks.

4.3 Integrated Gradients

Axiomatic Attribution for Deep Networks (Sundararajan et al., 2017)

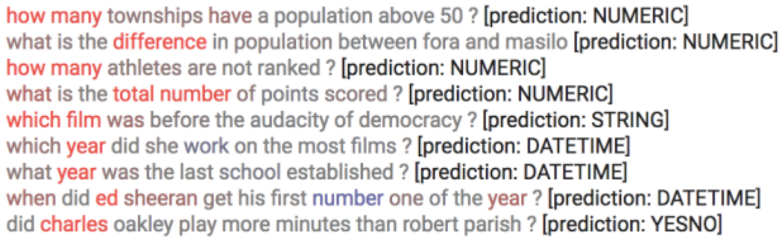
- **Motivation**: Traditional gradient based attribution techniques seems to fail with highly dimensional non linear deep learning neural NLP models. This paper proposes an improved way to address the challenge faced with gradient based approach and is called as *Integrated Gradients*. Integrated Gradients can be applied to wide variety of neural NLP technique and hence is worth considering for review.

- **Summary**: This technique requires a baseline to define the attribution problem. In a neural NLP case this baseline can be input sentence of zero vectors. Further for the problem setup it conditions two axioms: a) *Sensitivity*: for every input and baseline which differ in one feature and has different prediction should be given a non zero attribution. b) *Implementation In-variance*: two networks are functionally equivalent independent of their implementation if they have the same outputs for the given inputs. Further , integrated gradients are defined as the path integral of the gradients along the straight-line path from the baseline x' to the input x represented by below:

$$\mathsf{IntegratedGrads}_i(x) ::= (x_i - x'_i) \times \int_{\alpha=0}^{1} \tfrac{\partial F(x' + \alpha \times (x - x'))}{\partial x_i}\, d\alpha$$

Figure 4: Integrated Gradient (Sundararajan et al., 2017) along the ith dimension along the input x and baseline x'

Below results discussed in paper clearly showcase the effectiveness of integrated gradients in attribution of question classification problem:



how many townships have a population above 50 ? [prediction: NUMERIC]
what is the difference in population between fora and masilo [prediction: NUMERIC]
how many athletes are not ranked ? [prediction: NUMERIC]
what is the total number of points scored ? [prediction: NUMERIC]
which film was before the audacity of democracy ? [prediction: STRING]
which year did she work on the most films ? [prediction: DATETIME]
what year was the last school established ? [prediction: DATETIME]
when did ed sheeran get his first number one of the year ? [prediction: DATETIME]
did charles oakley play more minutes than robert parish ? [prediction: YESNO]

Figure 5: integrated gradient results for question classification problem. Highlighted words has attributions toward the final prediction represented in the figure (Sundararajan et al., 2017)

- **Interpretation**:
  - Gradient based techniques seems to fail since neural models are trained to saturation. So when the gradients are calculated the model function is always flat in the vicinity of the input and could not help in input importance.
  - With integrated gradients the intuition is we are moving from a baseline to see at what point prediction changes and hence help in determining for which input factor impacts the final decision.
  - Integrated Gradient methodology is only applicable to models which are differentiable and provides access to the gradients of model.

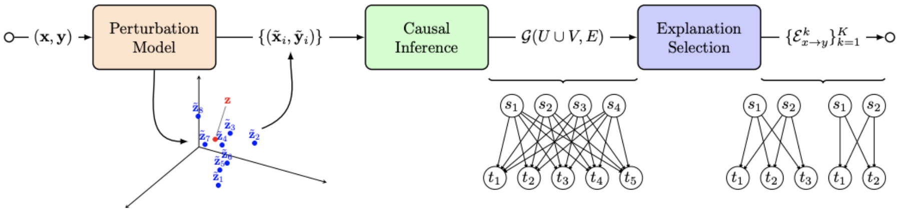
- **Contribution**: This class of explainability method provides a wider solution multiple types of deep networks. Also this has a strong theoretical justification which is widely accepted in the community.

4.4 Perturbation Method

A causal framework for explaining the predictions of black-box sequence-to-sequence models (Alvarez-Melis and Jaakkola, 2017)

- **Motivation**: This paper introduces to a different class of explanability technique which is perturbation. Rather than a sampling based approach which is one of the most common perturbation technique, authors has approached it by generating vector samples.

- **Summary**: This technique can explain predictions of any black box structured input output model around a specific instance. Framework approaches solving explanability by generating causally related input and output data by passing it through the black box model. Specifically for a NLP models auto encoders are used to generate these input and output data. These generated input vectors are minor perturbations to the given input instance. The dependencies between the input and output is inferred by querying the black box model with perturbed inputs. Below is the framework proposed by the authors:



$(\mathbf{x}, \mathbf{y})$
Perturbation Model
$\{(\tilde{\mathbf{x}}_i, \tilde{\mathbf{y}}_i)\}$
Causal Inference
$\mathcal{G}(U \cup V, E)$
Explanation Selection
$\{\mathcal{E}^k_{x \to y}\}^K_{k=1}$
$\mathbf{z}$
$\tilde{\mathbf{z}}_1$ $\tilde{\mathbf{z}}_2$ $\tilde{\mathbf{z}}_3$ $\tilde{\mathbf{z}}_4$ $\tilde{\mathbf{z}}_5$ $\tilde{\mathbf{z}}_6$ $\tilde{\mathbf{z}}_7$ $\tilde{\mathbf{z}}_8$
s_1 s_2 s_3 s_4
t_1 t_2 t_3 t_4 t_5
s_1 s_2 s_1 s_2
t_1 t_2 t_3 t_1 t_2

Figure 6: Structured Output Causal Rationalizer (SoCaRat) Framework Alvarez-Melis and Jaakkola (2017)

Structured Output Causal Rationalizer framework (SOCaRat) has three components a) *Perturbation Model* which generates similar input output data vectors. Variational Autoencoders are used for perturbation generation b) *Causal Inference* component which computes the causal relationship between input output c) *Explanation Selection* component which uses graph partitioning based algorithm to select most relevant sets of association.

- **Interpretations**:
  - This method is model agnostic and can be applied to any complex model not just limited to deep learning based model.
  - The results demonstrated in the paper proves the results are reasonable and coherent.

- **Contribution**: This paper provides a different dimension to NLP explainability methods with perturbation based approach. Rather than simple sampling based technique this leverages auto encoder based techniques for sample generation which is useful for deep networks.

5 Conclusion

As discussed there no one method which can be used for neural NLP explainability and several methods have been devised to cater different NLP architectures. Gradient based approaches needs model function to be differentiable and as discussed there are methods to lift the performance of these kind of methods. Recent state-of-the-art models uses attention and self-attention based methods. Attention weights can be used for explainability needs but it needs to be augmented with saliency for better interpretations. Perturbation based techniques can be used for any kind of model. It depends on sampling technique or sample generation techniques for its implementation. Below table (Captum) summarizes and highlight key difference between different explainability methods discussed in the previous section:

Method	Type	Additional Req.	Application	Complexity
Attention Saliency	Attention Gradient	Self-sufficient	Any attention based model.	O(examples * features)
Self-Attention Attribution	Attention Gradient	Self-sufficient	Transformers and Self-Attention models	O(examples * features)
Integrated Gradient	Gradient	Self-sufficient	Any model that can be represented as a differentiable function.	O(steps * examples * features)
Causal Perturbation	Perturbation	Post-hoc	Any traditional or neural network model.	O(examples * features * perturbations)

Beyond exploring methods for explainability, it is also important to evaluate the correctness of these explainability methods. This area of explainability evaluation is another area of research. Different techniques consider in this areas leverages informal examination, comparison to ground truth and human evaluation method.

References

David Alvarez-Melis and Tommi S. Jaakkola. 2017. A causal framework for explaining the predictions of black-box sequence-to-sequence models. *CoRR*, abs/1707.01943.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. Cite arxiv:1409.0473Comment: Accepted at ICLR 2015 as oral presentation.

Captum. Algorithm comparison matrix.

Marina Danilevsky, Shipi Dhanorkar, Yunyao Li, Lucian Popa, Kun Qian, and Anbang Xu. 2021. Explainability for natural language processing. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 4033–4034.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *A Survey of the State of Explainable AI for Natural Language Processing*, volume abs/2010.00711.

Shuoyang Ding and Philipp Koehn. 2021. Evaluating saliency methods for neural language models. *CoRR*, abs/2104.05824.

Reza Ghaeini, Xiaoli Z. Fern, and Prasad Tadepalli. 2018. Interpreting recurrent and attention-based neural models: a case study on natural language inference. *CoRR*, abs/1808.03894.

Yaru Hao, Li Dong, Furu Wei, and Ke Xu. 2021. Self-attention attribution: Interpreting information interactions inside transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12963–12971.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. *CoRR*, abs/1703.01365.