Is Attention Points Out the Only ”Path”? An Analysis on Transformers

Zijiao Yang
Computer Science Department
University of Colorado
Boulder, CO 80309, USA
zijiao.yang@colorado.edu

Abstract

Attention mechanism is popular throughout the Natural Language Processing (NLP) research community nowadays. yet little has been understood about attention. It remains largely a outcome of ”whatever works the best”, a empirical result. Nonetheless, attention is effective in terms of improving various of scores of measurement. Naturally, this leads to arise of analysis work centering at attention mechanism. This project extends from previous work of (Jain and Wallace, 2019) and (Wiegreffe and Pinter, 2019) that argues about if attention could serve as faithful explanations by adding similar analysis for self-attention mechanism of transformer, and for sequence tagging task. Specifically, I selected natural language chunking task(Sang and Buchholz, 2000) and transformer structure.(Vaswani et al., 2017) Trained a global adversarial model that is similar to (Wiegreffe and Pinter, 2019). But acts on sequence tagging task rather than just classification task, which makes more sense since attention mechanism got its attention from sequence generation tasks.

1 Introduction

Attention mechanism, gets intuition from human visual attention, an oversimplified description is that human perception process a scene by attend to different areas of the scene, obtain the information needed. Then those knowledge accumulated over time and formulate a scene representation. On top of that, further understanding and reasoning could happen. In natural language processing, the attention mechanism originated from (Mnih et al., 2014), (Bahdanau et al., 2014), initially is developed for solving the information bottleneck issue of Encoder-Decoder structure for machine translation task. Before, the interface that connects Encoder and Decoder are just a fixed length vector (context vector). Intuitively, that interface serve as a tunnel for decoder to access information from encoder at predication stage, also a interface to pass information back from decoder to encoder during training stage. And the interface is usually the last hidden states of encoder. Thought it is quite interesting approach, its capacity is limited by the fact that all information has to be encoded into a fixed-length vector, and that the vector does not change during after passed to decoder. Attention mechanism, instead of passing hard-formulated information segments, it allow the decoder part to actively search for the context it is needed at that moment of processing with a real time fashion. Namely the attention mechanism gives decoder access to previously computed hidden states of encoder, also the ability of changing its preference towards those hidden states actively.

Attention has since become a standard technique for a lot of NLP tasks. Also different forms of attention have evolved. A big impact for NLP community that centers around attention mechanism is the development of Transformer structure

On the other hand, since the successes of attention mechanism, researchers have started to make claims that attention trained afford transparency of the model’s inner mechanism, since for a given output, one could trace back to how much preference each input element are given for the specific output. In addition, visualizing attention map is often times adopted as a illustration of how the model works, and a effective analysis tool for debugging. Though attention has been ubiquitously applied. It is not validated if attention really provide a faithful explanation that relates inputs, and outputs or just merely a effective method to ease the training.

(Jain and Wallace, 2019) proposed that in order for attention to be faithful explanation, it needs to be consistent with other feature-importance measure, they also argue that as faithful explanation, the attention distribution should be exclusive respective to certain prediction result, namely, if one could find distinct attention distribution that leads to a similar prediction with respect to the original model prediction, then the attention distribution of original model is said to be not exclusive and cannot be use as a faithful explanation. (Wiegreffe and Pinter, 2019) mainly dispute with (Jain and Wallace, 2019) of their claim of attention distribution needs to be exclusive by saying that their experiment setting left a large amount of freedom that the result from the experiment setting is hard to reason about. They made claim that to find out if attention distribution is exclusive, the attention module should not be detached from other parts of the model, as attention mechanism is not a standalone, and other parts directly involved in the computation as well as the optimization of attention module. Also the per-instance adversary attentions (Jain and Wallace, 2019) is naturally easy to find under their problem setting. (Wiegreffe and Pinter, 2019) countered the claims and results from (Jain and Wallace, 2019) and give more perspectives on attention analysis. And I will elaborate the experiments they did and discuss the relationship with this project in next section.

The main purpose of this project is to explore: when self-attention mechanism is applied for sequence tagging task, what properties do the obtained self-attention distribution possess? Is it more exclusive compared to when applying attention to classification task. Does it possess similar model-independent information as in (Wiegreffe and Pinter, 2019)? Could it be more stable in terms of training with different initialization.

The main experimental steps of this project are:

1. Arguments on why we need to extend the attention analysis to other task, and why the natural language chunking task and Transformer structure is chosen

2. Train a global adversarial model of attention for natural language chunking task

3. Performed probing classifier with obtained adversarial attention

I will start by describe the essence of experiment design and insight from (Jain and Wallace, 2019) and (Wiegreffe and Pinter, 2019) in section 2 and discuss the their relationship with my extension. Then in section 3, I will describe the task and structure of my choice, and the reason why I make such a choice. The next section 4.3 will state the experiment design, which consist of the description for global adversarial model training, and probing classifier training. Finally the result and analysis is in section 2 and followed by a discussion more generally about attention and model transparency.

2 Attention as Explanation

(Jain and Wallace, 2019) did two experiments, one that try to find correlation between attention distribution and existing feature importance measures (*Gradient based measure*, *Leave one out measure*. The second one tries to find, for similar model prediction result, distinct attention distribution (**Adversarial attention**). And the attention part is detached from other parts of the model during equation solving. For the second experiment, the objective is to find a attention distribution as distinct as possible compared to original attention distribution. To account for the divergence between two distribution, **Jensen-Shannon Divergence**:

$$\mathrm{JSD}(\alpha_1, \alpha_2) = \frac{1}{2}\mathrm{KL}[\alpha_1||\bar{\alpha}] + \frac{1}{2}\mathrm{KL}[\alpha_2||\bar{\alpha}]$$
$$\bar{\alpha} = \frac{\alpha_1 + \alpha_2}{2}$$

The KL stands for **KL-divergence**, α_1 and α_2 are different attention distributions. To measure the distance between the predictions, Total Variation Distance (TVD) is considered:

$$\mathrm{TVD}(\hat{y_1}, \hat{y_2}) = \frac{1}{2}\sum_{i=1}^{|Y|} |\hat{y_{1i}} - \hat{y_{2i}}|$$

y_1, y_2 are the outputs of two models.

If the defined adversarial attention could be obtained, it means there could be different path from input to similar output, and the path is represented by attention distributions. And indeed their experiments successfully found adversarial attention distributions that fits the definition, for each instance locally. So they argue the attention could not be faithful explanation because they are not exclusive.

(Wiegreffe and Pinter, 2019) mainly criticized the second experiment's assumptions: First, the attention does not need to be exclusive for serving as explanation of the model, and the per-instance adversarial attentions tend to be easy to find for the classification tasks. Also, the attention should not be isolated from rest of the model. They design four experiments: First, they want to eliminate the tasks that are too simple by comparing the performance of model for a specific task with and without attention mechanism being activated. As it makes more sense to perform attention analysis if attention is actually needed. They freeze the attention with uniform weights over hidden states, and find two data set: AgNews and 20 NewsGroups did not utilize attention. The next experiment they try to provide the baseline of the variance of attention under different parameter initialization. So as to have a valid result to be compared with. If the variance of attention is already large enough, then we might not attribute the reason of the divergence between adversarial attention and original attention to "attention is not exclusive". And it is indeed testified for in their results of Diabets data set. And the third experiment tries to use attention trained to guide prediction for MLP model, and showed that attention could provide beneficial signal with a model-independent fashion. The final experiment redefine the adversarial attention with the concept of "global adversarial training": To train an adversarial attention model by optimizing the whole model towards a distinct attention distribution compared to original attention distribution and similar predictions. The result of fourth experiment does not invalidate the claim of (Wiegreffe and Pinter, 2019) but rather provide a framework for attention analysis.

Both the work in (Jain and Wallace, 2019) and (Wiegreffe and Pinter, 2019) have experiments centers around if attention is exclusive in terms of specific data instance. (Wiegreffe and Pinter, 2019) counters the (Jain and Wallace, 2019) by saying that it does not have to be exclusive. In deed, for classification task the freedom in the model is large. Quoted from (Wiegreffe and Pinter, 2019) when describing the task setting: "This mathematically flexible production capacity is particularly evident in binary classifiers". My question is what if the task itself requires more exclusiveness, could attention help achieve that, could attention under that setting to be more exclusive, thus has potential to be provided as a rather faithful explanation tool: There are many ways to says no to people for various reasons, but how many ways do you have for describing the process of cooking a omelette. And I will describe my task selection and experiment design in the coming sections.

3 Task and Structure choice justification

I choose Transformer structure (Vaswani et al., 2017) as my target model structure. The main reason is the Transformer based model structure has brought about the 'image-net' moment for NLP research community by breaking a lot of benchmark tasks by a large margin (Devlin et al., 2018). And it is mainly based on the mechanism of self-attention, and multi-head attention (which is just the combination of many self attention). Besides people have been investigating the properties of the attention that the transformer is built on. (Clark et al., 2019) has looked at the attention head of BERT (Transformer based model) to analyse their specific functions, and showed attention heads could capture some aspects of syntactic features. And (Voita et al., 2019) analysed the redundancy of multi-head attention and found out some of those heads could be pruned out without affect the model performance much.

Also, alignment attention used for sequence to sequence generation, is computed next attention distribution based on previously computed results (Follow the property of recurrent neural network structure, one context vector is computed, and that is feed into generation of next decoder hidden state, and that hidden state is then used to computed another attention distribution to generate yet another context vector), therefore, it is relatively hard to define the adversarial attention training procedure. As opposite of that, the computation of self-attention does not rely on previous computation, therefore, it is more convenient for analysis in terms of simplicity (at least on surface level). In summary, self-attention based model, Transformer, possess a lot of interesting properties, and is mainly based on attention mechanism, and it fulfills my needs for performing sequence generation tasks. So I chose it.

In order to focus on analysis on self-attention mechanism as a initial start, I choose sequence tagging task, this type of task requires a sequence of output and a sequence of input, but the input and output length are the same, the former properties ensures that the task is more complex than classification tasks that the task possess innate restriction to the freedom of the model parameters more strictly than for classification tasks. And the second property avoid the need of decoding structure, as we only need encoder part for tagging task, this restrict the model under attention to be consist of only self-attention rather than both self-attention and alignment-attention (Bahdanau et al., 2014) as we typically see in (Devlin et al., 2018) which utilizes encoder-decoder structure. Also, it makes the computation of the divergence of two prediction sequence easy.

Because of the former described needs, I chose the natural language chunking task, and the data set is CONLL2000, in which I used 8042 training instance, and 2012 test instance. (Sang and Buchholz, 2000), chunking is a process to extract phrase from text data, for example, verb phrase, none phrase, etc. the following listing shows an example of chunking.

```
{
    ’fail’:’B-VP’,
    ’to’:’I-VP’,
    ’show’: ’I-VP’,
    ’a’: ’B-NP’,
    ’substantial’:’I-NP’,
    ’improvement’:’I-NP’,
    ’from’:’B-PP’,
    ’July’: ’B-NP’,
    ’and’:’I-NP’,
    Augustt’:’I-NP’,
    "’s": ’B-NP’,
    ’near-record’:’I-NP’,
    ’deficits’:’I-NP’
 }
```

Listing 1: Example of generated json formatted data: for each verb contained in the sentence, a description of predicate argument structure is given, a sequence of semantic role tags in BIO formatting is given, and the tokenized words for the sentence are given

4 Experiment Design

In this section, I will give the description of my model, along with detailed description of how adversarial model in trained, and the procedure currently used for probing classifier.

4.1 Model Description

My model consist of a 1 layer transformer that contains 4 attention head, the attention mask is disable and the concatenation of context computed from multi-head attention is passed through a linear layer followed by a softmax layer to make final predictions. Follows the nature of self-attention, each token in the sequence has a corresponding context vector computed, and the final prediction is made based on them, following is the formulation of self attention as well as muli-head attention:

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

$$\textbf{MultiHead}(q, k, b) = Concat(head_1, ..., head_h)W^O$$
$$\textbf{where } head_i = \textbf{Attention}(qW_i^q, kW_i^k, vW_i^v)$$

In the equation above, the q, k, v are identical here, and is computed by passing the input word embeddings to a layer transformation followed by adding time signals. Q, K, V are query, key, value vectors, which is computed by passing q, k, v through a linear transformation respectively. Indeed, the encoder part setup is identical to (Vaswani et al., 2017). And the loss used for training is the **Negative log likelihood loss** and the code as well as hyperparameter setting can be found here: Transformer-Attention

4.2 Global Adversarial Training

For training global adversarial model, I follow largely from (Wiegreffe and Pinter, 2019). the attention computed following the above equation for each head and for each training instance as well as testing instance is saved for later use. And also the prediction results of original model are saved.

During adversarial training, a loss account for the distance between original predictions and adversarial model prediction, as well as the divergence between original model attention distribution and adversarial attention distributions. To account for distance between original predictions and adversarial model predictions, the original model loss is used, namely the **Negative Log likelihood Loss**. To account for the divergence between attention distributions, **KL-divergence** is used. One thing to be note here is, the divergence for attention distribution computed here is with respect to the attention (self attention context vector) of each token of each attention head. Overall, during global adversarial training, we want the adversarial model prediction to be as close to original prediction as possible, and we want the divergence between original and adversarial attention distributions to be as large as possible. in **??**, y_1, y_2 stands for model predictions, α_1, α_2 stands for different attention distributions.

$$\textbf{Loss} = \textsc{NLLLoss}(y_1, y_2) - \lambda * \text{KL}(\alpha_1, \alpha_2)$$

4.3 Probing Classifier

In order to test if the attention obtained possess model-independent information that could aid the training of a different model, also to compare the guiding performance difference between original attention and the adversarial attention. The attention computation step in the multi-head attention is substitute by pre-trained attention distributions, yet the linear transformations that compute the query, key, value from original word embedding is kept, as well as the later linear transformation after the concatenation of context vectors are kept into optimization.

5 Results and Analysis

In this section, the results from adversarial training are reported, as well as the result of probing classifier are reported. The results are reported using F1 score on the test split of the data set. The F1 score of original model and adversarial model for $\lambda = 0.2$ and $\lambda = 2.5$ are reported in 1. And we could see as divergence between attentions grows, the performance drops.

Table 1: F1 score of original and adversarial models

ExpName	F1-score
original	0.90
adv-0.2	0.90
adv-3	0.75

Table 2: F1 score of probing classifiers

ExpName	F1-score
prob-ori	0.90
prob-adv02	0.90
prob-adv3	0.89

The above table reported the results from probing classifiers. Again, the probing classifier, as described in section 4.3 is basically the original with pre-trained attention substituting the attention computation step for each training instance. Here, ”ori” stands for probing classifier using the attention obtained from original model, ”adv02” stands for using adversarial attention that is obtained when $\lambda = 0.2$, ”adv3” stands for when $= 3$. As we could see, the F1-scores are close, I think it is likely due to my current probing classifier kept large part of attention module in the overall training process, therefore, there might be more information left to the probing classifier other than just the pre-trained attention distributions.

The results are experiments of adversarial training of ranging from $0.2 - 3$. The smaller the λ, the model train prefer a closer distance between prediction rather than a large divergence between attentions, so as $lambda$ grows, the divergence encouraged grows, and the gap between the predictions grows. 1 shows the relationship between the divergence of attentions (x-asix) and predictions (y-axis). We could see that as the divergence between original and adversarial grows, the gap between predictions grows as we expected (indicated by increasing NLLLoss). But the NLLLoss / KL div was initially grows slowly which indicates change of attention won’t affect much of model performance, and that indicates the model has tolerance for a attention distributions within a certain variance. But after around KL divergence is 0.33, the loss of prediction grows substantially compared to little increment of attention divergence, which coudl result from the attention distribution has pass a certain ”safe point”. Overall, this result might indicates that self-attention mechanism utilized in transformer structure tends to have tolerance within a range of variance, but after a certain point, the tolerance is exhausted, and model broken down suddenly. That could means some level of exclusiveness exists for this structure for this specific type of task. But of course we need further investigation. For example, we could compare between the adversarial attention distribution obtained before the ”safe point” and the adversarial attention obtained after of their attention weights ranking. The JSD vs TVD results from (Wiegreffe and Pinter, 2019) is here **??** (it is a direct screenshot from their paper).

6 Conclusion

In this project, I did mainly two investigation, one is training a global adversarial attention model for transformer structure on chunking task. Then a probing classifier tries to understand if adversarial attention could provide information to guide other model in prediction process, as well as under what circumstance, the adversarial attention become unstable in terms of information providing. In last section, I pointed out that based on the

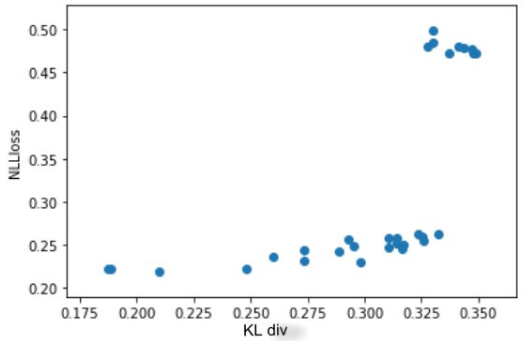

Figure 1: 4 Layer BiLSTM with highway connections, the pink arrows represent highway connections and the red boxes represent *transform gates* r_t that controls the weight of linear and non-linear information flows between layers.

current result we see in 1, it is reasonable to think of self attention as having some tolerance within a certain variance. And after pass the ”safe point”, it could tend to be more exclusive so as to keep the model performance stable. In future work, I think it testify this hypothesis is my priority. Then other forms of probing classifier could be utilized for the analysis, for example: We could pass the sequence through a feed forward layer, then pull out the attention matrix, and extract corresponding weight vector for each token, next, we use each of the weight vector to weight formerly computed hidden states for form context vectors, and make prediction based on those context vectors.

Acknowledgments

I would like to thank Professor Chenhao Tan for his suggestions and comments for this project, also I would like to thank my classmates for useful discussion both in and out of the class. I am grateful to be given a chance to take on this class and this project.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. 2019. What does bert look at? an analysis of bert's attention. *arXiv preprint arXiv:1906.04341*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

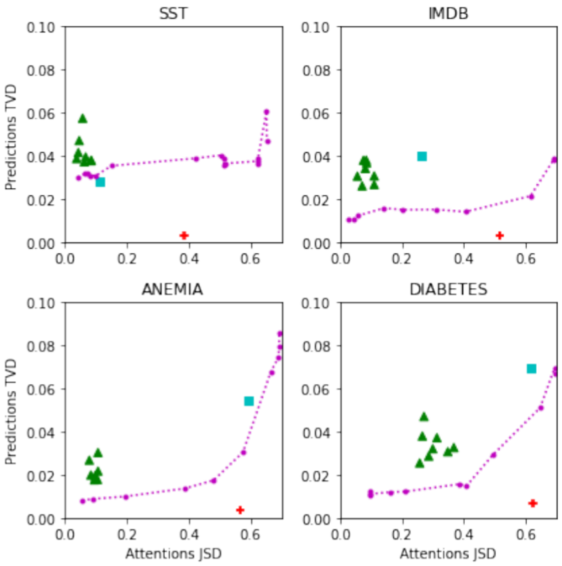

Figure 5: Averaged per-instance test set JSD and TVD from base model for each model variant. JSD is bounded at ∼ 0.693. ▲: random seed; ■: uniform weights; dotted line: our adversarial setup as λ is varied; +: adversarial setup from Jain and Wallace (2019).

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *ArXiv*, abs/1902.10186.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212.

Erik F Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: Chunking. *arXiv preprint cs/0009008*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *arXiv preprint arXiv:1905.09418*.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*.