Natural language processing project

Machine learning for data science, Paris University

A multiscale visualization of attention in the transformer model.

Wacim Belahcel, Imad Oualid Kacimi, Yasmine Agliz, Sami Arabi
Formation FA

Abstract

The Transformer is a deep learning model introduced in 2017, used primarily in the field of natural language processing. Transformers are used for sequential data such as natural language, unlike RNN models they do not require the data to be processed in order. Thus, the Transformer allows more parallelization than RNNs and therefore reduces training times.

BERT (Bidirectional Encoder Representations from Transformers) and GPT (Generative Pre-trained Transformer) are two very known pretrained transformer models, which have been trained with large corpora such as Wikipedia Corpus and Common Crawl.

The transformer model is entirely based on the attention approach which improves its performances; however, it is difficult to understand how the model assigns weight to different input elements especially with the multi-layer, multi head attention. Some tools were implemented to help visualize attention at multiple scales, which provides unique perspective on the attention mechanism.)

Keywords: Transformers, BERT, GPT, attention, multi-head attention, multi-layer

1. Introduction

Transformer models have been breaking performance records on many NLP tasks, like Neural Machine Translation, Language Modeling, and Question Answering problems. Attention networks are also more efficient and require less computational resources. This is an important improvement, as it often requires significant computing power in the form of a GPU to train RNN's.

Attention is a technique that mimics cognitive attention which leads to enhance the important parts of the input data and fades out the rest. Transformer networks make extensive use of attention mechanisms to achieve their expressive power. A multi-head attention network combines several different attention mechanisms to direct the overall attention of a network.

BERT and GPT-2 are two of the best transformer models we have nowadays, they use fully attention-based approaches. Many tools were implemented to facilitate the understanding of these models' mechanism such as heat maps or bipartite graph. The main challenge of this task, was the representation of the multi-layer, multi-head attention mechanism, which produces different attention patterns for each layer and head.

In the original paper [1], they presented a tool that allows multiscale visualisation of attention in the transformer model by introducing a high-level model view, which visualizes all the layers and attention heads in a single interface, and a low-level neuron view, which shows how individual neurons interact to produce attention. He also adapts the tool from the original encoder-decoder implementation to the decoder-only GPT-2 model and the encoder-only BERT model.

In this work we will try to reproduce and explore the tool in its different perspectives. Moreover, we will analyse the results in different use cases such as: Detecting Model Bias, Locating Relevant Attention Heads and Linking Neurons to Model Behavior.

2. Related work

Some research were made as a way to improve analysis methods to help understand the attention mechanisms of models and apply them to BERT. Even though recent work focused on vector representations or model outputs. Clark, Kevin, et al (2029)[2] work shows that a substantial amount of linguistic knowledge can be found not only in the hidden states, but also in the attention. They thought that exploring attention maps completes the other model analysis.

For the Transformer, which neither uses convolution nor recurrence, adding explicit representation of position information is even more significant since the model is invariant to sequence ordering. Thus, attention-based models use position encodings or biased attention weights based on distance (Parikh et al., 2016).
Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani (2018)[3] present an efficient way to incorporate relative position representations in the self-attention mechanism of the Transformer in their work. They also demonstrate significant improvements in translation quality on two machine translation tasks.

(Vig, Jesse, and Yonatan Belinkov) [4] extend models attention visualization tasks on three levels: The attention head level, the model level, and the neuron level. They also adapted the original encoder-decoder implementation to the decoder-only GPT-2 model, as well as the encoder-only BERT model to analyse attention in aggregate over a large corpus as a way to explore the problematic of alignment of the attention with syntactic dependency, the links between the attention heads and the part of speech tags, and finally, how attention capture long-distance relationships versus short-distance ones.

Their analysis was based on a GPT-2 small pre-trained model. They concluded that attention follows dependency relations most strongly in the middle layers of the model, and that attention heads target particular parts of speech depending on layer depth. They also found that attention spans the greatest distance in the deepest layers but significantly varies between heads.

3. Paper summary

In 2018, the BERT language representation model achieved state-of-the-art performance across NLP tasks ranging from sentiment analysis to question answering. Recently, the OpenAI GPT-2 model outperformed other models on several language modeling benchmarks in a zero-shot setting. Underlying BERT and GPT-2 are the Transformer model, which uses a fully attention-based approach in contrast to traditional sequence models based on recurrent architectures. An advantage of using attention is that it can help interpret a model by showing how the model assigns weight to different input elements, although its value in explaining individual predictions may be limited

One challenge for visualizing attention in the Transformer is that it uses a multi-layer, multi-head attention mechanism, which produces different attention patterns for each layer and head. the author in paper [1] designed a visualization tool specifically for multi-head attention, which visualizes attention over multiple heads in a layer by superimposing their attention patterns. In this paper, the author extends the work of Jones by visualizing attention in the Transformer at multiple scales.

The **attention-head view** visualizes the attention patterns produced by one or more attention heads in a given layer, attention heads also capture specific lexical patterns. Other attention heads detected named entities, paired punctuation, subject-verb pairs, and other syntactic and semantic relations. The author presented some use cases such as detecting model bias or locating relevant attention heads.

The **model view** provides a birds-eye view of attention across all of the model's layers and heads for a particular input. Attention heads are presented in tabular form, with rows representing layers and columns representing heads. Each layer/head is visualized in a thumbnail form that conveys the coarse shape of the attention pattern, following the small multiples design pattern.

The **neuron view** visualizes the individual neurons in the query and key vectors and shows how they interact to produce attention. Given a token selected by the user, this view traces the computation of attention from that token to the other tokens in the sequence. This can be used to link neurons to model behavior and when specific neurons are linked to a tangible outcome, it presents an opportunity to intervene in the model.

4. Results

As a way to assert the fact that we are using the same models and parameters as the original paper, we first tried the tools on the same sentence "The doctor asked the nurse a question. She/He", as can be seen in the figure 1, results seem similar to the original paper, which is obvious since no training of the models was required (the exact same model is used in the original paper).

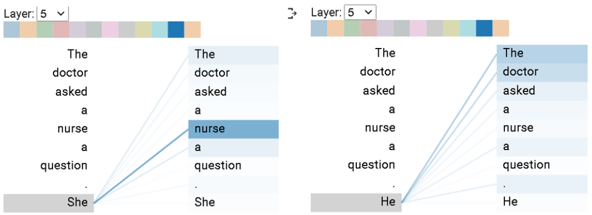

Figure 1: GPT-2 model layer 5 head 11 attention of the pronoun "She" and "He"

We then studied possible links between pronouns and their antecedent in different attention heads for BERT and GPT-2 models using data (sentences) from [5] which they explored in their original paper evaluating gender bias in machine translation.

On BERT, 4 different layers with different attention heads seem to focus their attention on finding links between pronouns and their antecedent: layer 2 on attention head 9, 3 on attention head 11, 5 on attention head 4, finally, 8 on attention head 11, we can see different results we got on figure 2 to 5.

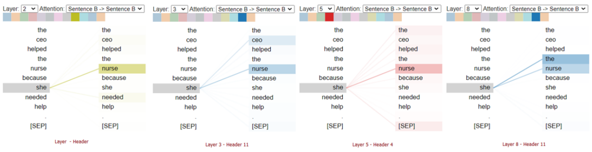
Layer: 2 Attention: Sentence B -> Sentence B
the ceo helped the nurse because she needed help . [SEP]
Layer - Header
Layer: 3 Attention: Sentence B -> Sentence B
the ceo helped the nurse because she needed help . [SEP]
Layer 3 - Header 11
Layer: 5 Attention: Sentence B -> Sentence B
the ceo helped the nurse because she needed help . [SEP]
Layer 5 - Header 4
Layer: 8 Attention: Sentence B -> Sentence B
the ceo helped the nurse because she needed help . [SEP]
Layer 8 - Header 11

Figure 2: BERT attention heads on sentence "the CEO helped the nurse because she needed help" on pronoun "she"

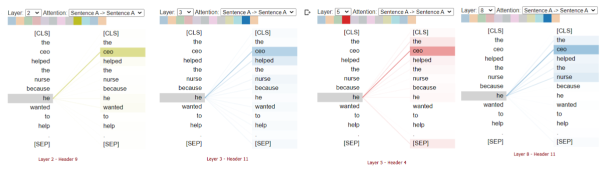

Figure 3: BERT attention heads on sentence ”the CEO helped the nurse because he wanted to help” on pronoun ”he”

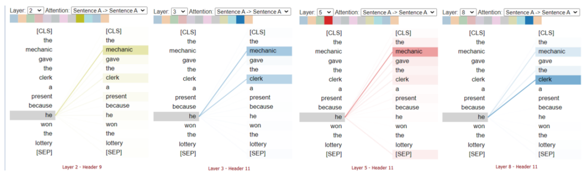

Figure 4: BERT attention heads on sentence ”the mechanic gave the clerk a present because he won the lottery” on pronoun ”he”

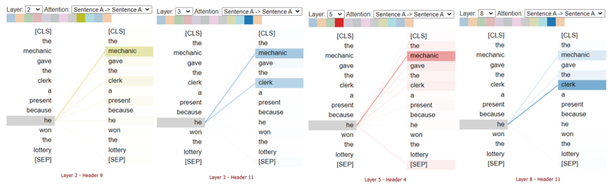

Figure 5: BERT attention heads on sentence ”the mechanic gave the clerk a present because it was her birthday” on pronoun ”her”

We made the same experiment on GPT-2, as a result we found 2 attention heads on different layers that could be interpreted as trying to find link (focusing attention) on pronoun’s antecedent, which are: layer 6 attention head 1, layer 5 attention head 11.

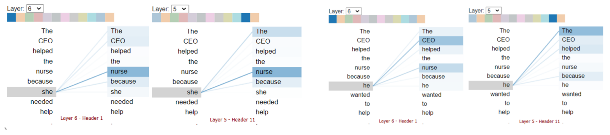

Figure 6: GPT-2 attention heads on sentences ”The CEO helped the nurse because she needed help” (left) ”The CEO helped the nurse because he wanted to help” (right) on pronouns ”she”/”he”

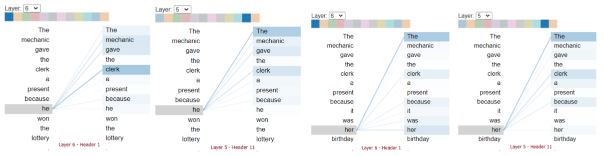

Figure 7: GPT-2 attention heads on sentences ”the mechanic gave the clerk a present because he won the lottery” (left) ”the mechanic gave the clerk a present because it was her birthday” (right) on pronouns ”he”/”her”

5. Discussion

If we analyses the different results we got for both models, first looking at BERT, we can see that the model seems efficient in linking a pronoun to its antecedent, as can be seen in figure 2 and 3, the pronoun “she” is linked to the antecedent “nurse”, in the other hand the pronoun “he” is linked to the antecedent “ceo” on every targeted attention head, the same observation can be made for figure 4 and 5 and pronouns “he”, “her” with antecedents “mechanic” and “clerk” which shows that those attention heads are clearly specialized in focusing pronouns embedding words attention to their antecedent.

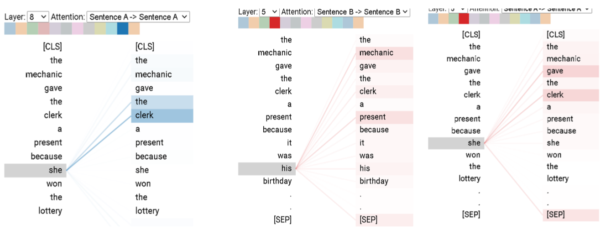

Figure 8: BERT model bias in predicting pronoun’s antecedent

If we now try the same experiment by switching pronouns gender, we can see interesting results that in figure 8, as can be seen BERT models seems biased, for example, “she” on the far left refer to “mechanic”, but we can see that BERT chose to focus its attention on “clerk” even though it’s the wrong choice, it seems to be the obvious choice for the model which was probably trained on data that encouraged such behavior, attributing female pronouns to female dominated jobs and male pronouns to male dominated jobs.

Finally, looking at GPT-2 results, it seems less efficient in finding links between pronouns and antecedent, in Figure 6, observing results, we can see that the attention heads are able to focus their attention for each pronoun “he” and “she” on the right antecedent “CEO” and “nurse”.

But if we now try a less obvious example like in figure 7, we can see that the attention from the pronouns to the antecedents seems more indecisive and the model is having more difficulties finding to which of “mechanic” or “clerk” it should give more attention.

Regarding model bias, the same observation can be

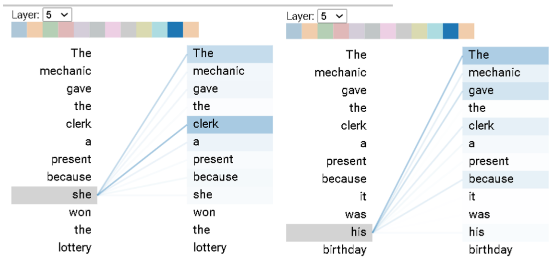

Figure 9: GPT-2 model bias in predicting pronoun's antecedent

made on GPT-2 (looking at figure 9) as the one previous made for BERT models, which is that the model seems to be biased, in the example above, we can see that much like in BERT, GPT-2 attention head 11 of layer 5 focus its attention for pronoun "she" on "clerk" even though the pronoun clearly refer to the antecedent "mechanic".

6. Conclusion

In this paper, we used a multiscale attention visualizing tool on multiple transformer models. We ran experiments on BERT and GPT-2 transformers, focusing our observations on links between pronouns and their antecedents. We were able to find and observe different attention heads that were specialized in focusing their attention on links between pronouns and antecedent for both mmodels. We found out that BERT was more efficient in such task than GPT-2 even though they both output somewhat good results.

Finally, we noticed a clear gender bias on both models probably due to the data that they were trained on which most probably more often linked profession titles genders to the gender they were most dominated by, thus encouraging the model to learn such wrong biased information for the two transformers, resulting in some cases into linking pronouns to the wrong antecedents.

7. Acknowledgement

Abstract: Yasmine Agliz
Introduction: Yasmine Agliz
State of the art: Yasmine Agliz - Sami Arabi
Paper Summary: Sami Arabi
Results : Wacim Belahcel - Imad Oualid Kacimi
Discussion : Wacim Belahcel
Notebook code : Imad Oualid Kacimi - Wacim Belahcel
Docker image : Imad Oualid Kacimi
Conclusion : Sami Arabi - Wacim Belahcel

8. References

[1] Jesse Vig,A Multiscale Visualization of Attention in the Transformer Model, 2019, arXiv:1906.05714 [cs.HC]

[2] Clark, Kevin, et al. ”What does bert look at? an analysis of bert's attention.” arXiv preprint arXiv:1906.04341 (2019).

[3] Shaw, Peter, Jakob Uszkoreit, and Ashish Vaswani. ”Self-attention with relative position representations.” arXiv preprint arXiv:1803.02155 (2018).

[4] Vig, Jesse, and Yonatan Belinkov. ”Analyzing the structure of attention in a transformer language model.” arXiv preprint arXiv:1906.04284 (2019).

[5] Gabriel Stanovsky and Noah A. Smith and Luke Zettlemoyer,Evaluating Gender Bias in Machine Translation, 2019