Proceedings of the 2nd International Conference on Software Engineering and Machine Learning
DOI: 10.54254/2755-2721/76/20240571

Interpretability analysis in transformers based on attention visualization

Yuxi Guo

SWUFE-UD Institute of Data Science, Southwestern University of Finance and Economics, Chengdu, 611130, China

m18302211926@163.com

Abstract. Self-attention is the core idea of the transformer, a kind of special structure for models to understand sentences and texts. Transformer is growing fast, but the model's internal unknowns are still out of control. In this work, the research visualizes self-attention and observes those self-attentions in some transformers. Through observation, there are five types of self-attention connections. The research classifies them as Parallel self-attention head, Radioactive self-attention head, Homogeneous self-attention head, X-type self-attention head, and Compound self-attention head. The Parallel self-attention head is the most important. The combination of different types will affect the performance of the transformer. Visualizations can indicate the location of different types. The results show that some homogeneous heads should be more varied in that case the model will perform better. A new training method is called local head training method, and the local training method may be useful during training transformer. The purpose of this study is to lay the foundation for model biology, to take other perspectives to understand transformers, and to fine-tune training methods.

Keywords: Transformer, self-attention, attention visualization, interpretability analysis.

1. Introduction

After the interpretability of transformers becomes more and more popular, the development of attention visualization tools has flourished. In 2019, BertViz appeared and provided attention visualizations for transformers [1]. After attention visualization, the arrangement shape of the self-attention lines inside the head varies greatly. The BertViz also provides visualization of query and key vectors of self-attention in a head [1]. Besides, computer vision, Attention-Viz also handles attention visualization in computer vision transformers effectively [2]. In the field of interpretability analysis, there are several visual tools available for researchers to use. However, unlike the studies of these visualization tools, this work considers the configuration characteristics of attention rays globally. In addition, this work is innovative in that it also presents a quantitative indicator of relative importance and compares them.

Good transformers evolve rapidly and differ in many ways from classical models [3]. In the course of this work, the study visualized the self-attentional shape of good models and old models and came to some interesting findings. The research contents mainly include: analyzing the arrangement characteristics of attention rays, then studying the attention visualization of the old and new models, and exploring the importance of attention heads under the condition of the same input.

DOI: 10.54254/2755-2721/76/20240571

The research approach is like doing experiments in biology because now large models and biological responses have some things in common. Based on the dominant visual level analysis, quantification is used to assist the research in reaching conclusions. For example, they're abstract black boxes that require a detailed analysis of an input and an output, but it is difficult to know what happens in the middle. So the research can explore how intelligent models react to the input and get some experimental conclusions. As biology tries to explore abstract black boxes.

The research significance of this paper is to provide a perspective for the understanding of good models. This work explores a new interpretability analysis, model biology, as a general approach to biological experiments to explore transformer attention mechanisms. It provides a new idea for the training method of the model.

2. Research design

The model name of the relatively new transformer the research chooses is called "xtremedistil-l12-h384-uncased" in Huggingface published in 2021 [4]. Given the same input, the study compares its self-attention visualization with those of the other three models which are respectively BERT [5], GPT2 [6], and RoBERTa [7]. In hugging face, their names are respectively "bert-base-uncased", "gpt2" and "Roberta-base". The number of layers and the number of headers inside their models are the same, 12 layers and 12 heads. The input text is "Mathematics is difficult because it requires not only logic but also imagination." Given the same input, the attention visualization can show how the transformers react to the input. In this work, it is obvious to see the attention differences, and helps us think about what happens in the old models.

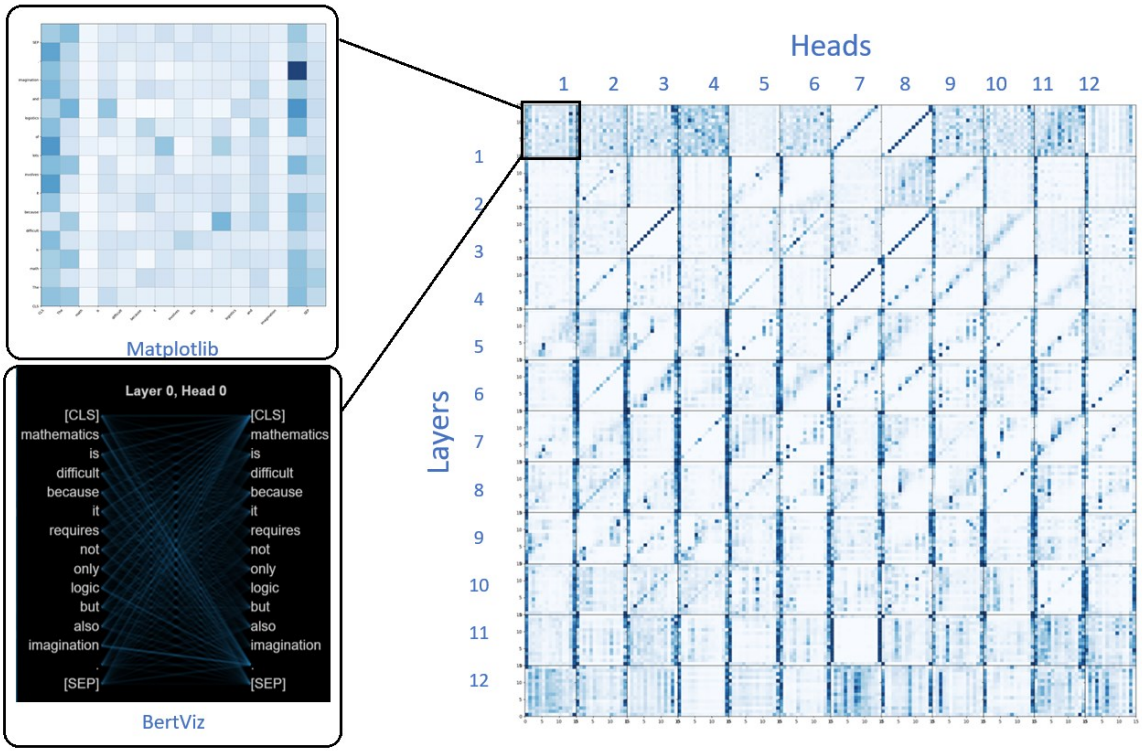

Figure 1. Layer 0 Head 0 self-attention visualization in "xtremedistil-l12-h384-uncased".

In this work, visualization is the first task. Here, a heatmap in Matplotlib is used to show the self-attention in one sentence. After visualization, the vertical axis is the main body to analyze. Figure 1 is a visualization of self-attention in the first head of the first layer in the new model. The above heatmap

DOI: 10.54254/2755-2721/76/20240571

tells us that Layer 0 Head 0 pays more attention to the beginning and end of the sentence. The visualization of BertViz on lefthand can show the self-attention shape.

3. Results

The research results include the following aspects, the first is the classification of attention head. This is followed by a definition of the importance of attention heads, and then a definition of relative importance. The discovery of the direction of attention, and finally the analysis of the species inherent in our heads.

3.1. Different shapes of self-attention

Based on the arrangement of attention rays, it is easy to find that there are many common features in a transformer model. If the arrangement of attention rays is used as a classification standard, it can be divided into the following five categories: Parallel self-attention head, Radioactive self-attention head, Homogeneous self-attention head, X-type self-attention head, and Compound self-attention head.

3.1.1. Parallel self-attention head

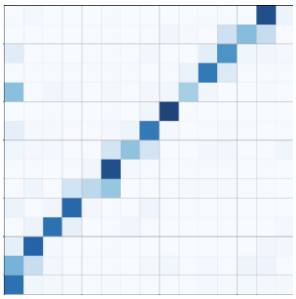

Figure 2. Parallel head heatmap example (Heatmap of self-attention values in one sentence).

The attention rays in one head mainly consist of the parallels so that it is obvious to notice them in the transformer global visualization.

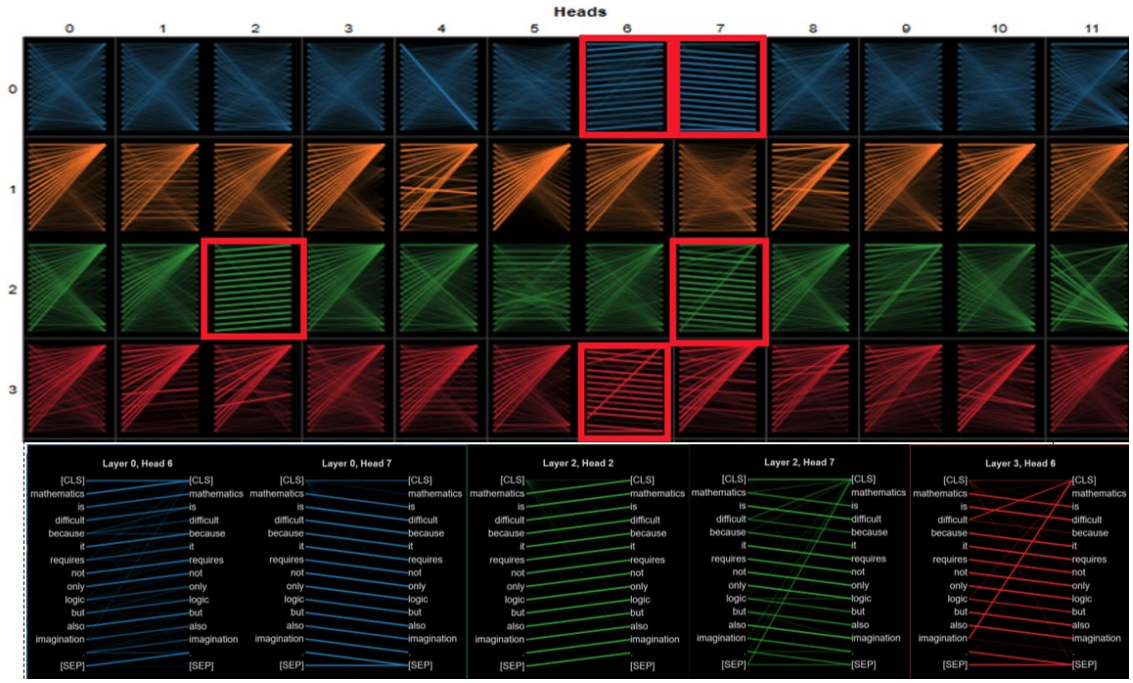

Figure 3. Parallel head location examples.

DOI: 10.54254/2755-2721/76/20240571

This kind of head has common features which are parallel shapes overall. The parallels mean that tokens point clearly to those tokens to which they pay more attention. The main distribution of parallel self-attention heads is mainly in the first four layers. Attention rays in one head mainly consist of the parallels so that it is obvious to notice them in the transformer global visualization. This kind of head in this transformer is rare. The total number of heads is 144, but the parallel self-attention head only takes up five of them.

3.1.2. Radioactive self-attention head

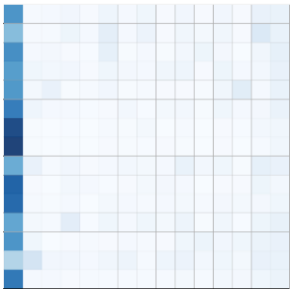

Figure 4. Radioactive head heatmap example (Heatmap of self-attention values in one sentence).

The attention rays point to the start or end tag intensively, and in global visualization, it is like a right triangle. This type of attention head tells us that the transformer's eyes move to one end of the sentence, in this case focusing on the first tag.

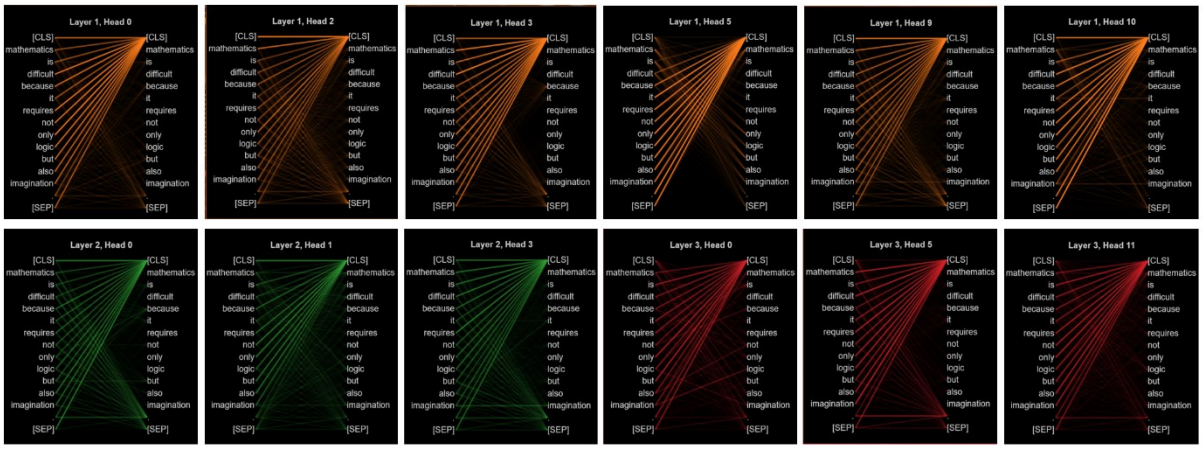

Figure 5. Radioactive head location examples.

This kind of head mainly consists of the attention rays which focus on the start or end tag, [CLS] label, or [SEP] label. For this new model, they prominently appear in the first four layers.

DOI: 10.54254/2755-2721/76/20240571

3.1.3. Homogeneous self-attention head

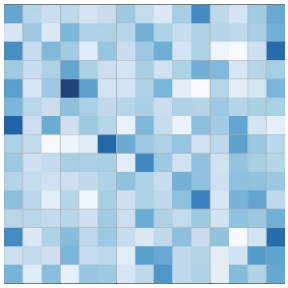

Figure 6. Homogeneous head heatmap example (Heatmap of self-attention values in one sentence).

Attention flows evenly to each token, in this process, meaning that a token has similar attention to all tokens, and tokens without significant attention appear to spread the attention rays evenly over the whole sentence, and the shape appears to be fully connected.

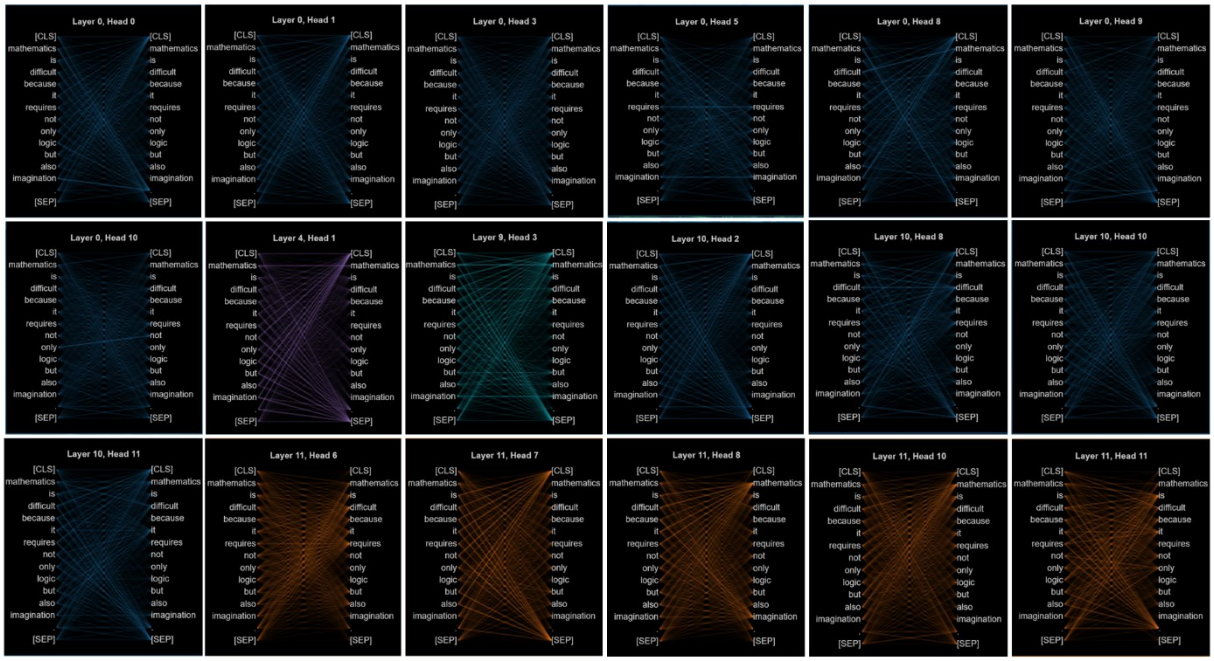

Figure 7. Homogeneous head location examples.

This kind of self-attention head is so homogeneous that it is not easy to know the relationship between the two tokens.

3.1.4. X-type self-attention head

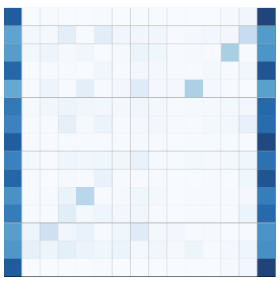

Figure 8. X-type head heatmap example (Heatmap of self-attention values in one sentence).

DOI: 10.54254/2755-2721/76/20240571

As the focus of attention flows towards the beginning label and the end label, the shape of this attention ray takes on the shape of an X letter.

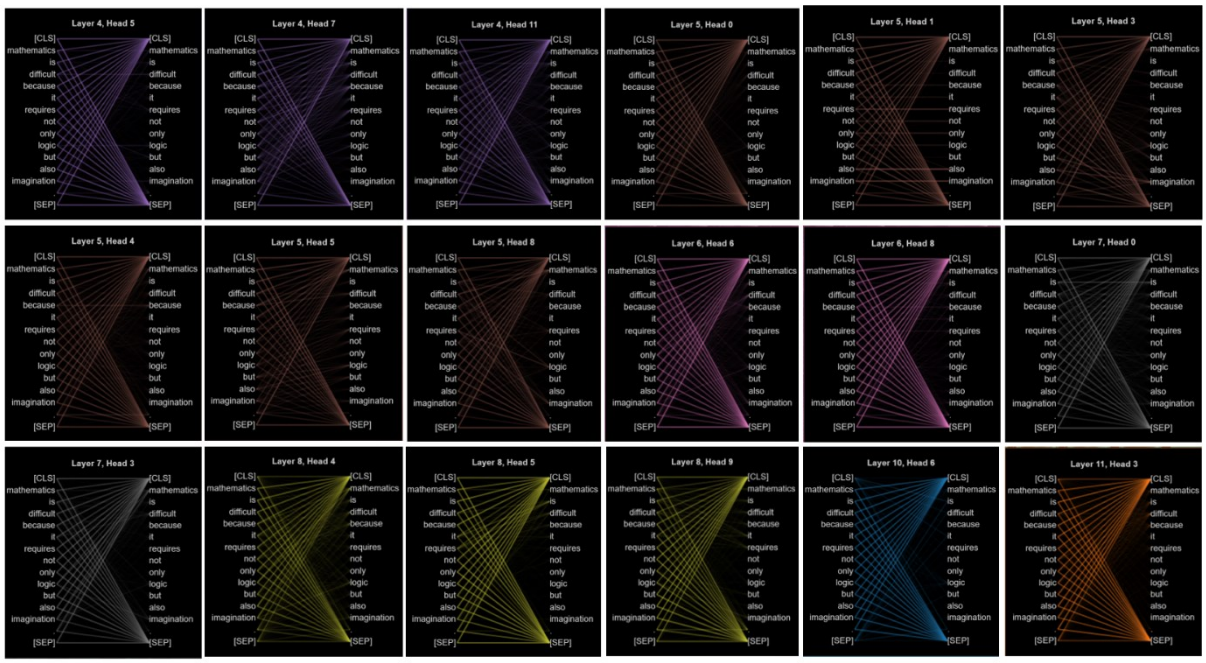

Figure 9. X type head location examples.

This type of self-attentional head pays equal attention to the beginning and end labels of the sentence, making it impossible for researchers to know which end of the sentence it is leaning towards, resulting in a lack of explainability of the entire attention flow, and this mystery will be mentioned again in the topic of where the attention flows.

3.1.5. Compound Self-attention head

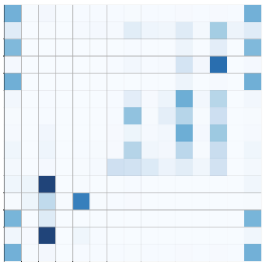

Figure 10. Compound head heatmap example (Heatmap of self-attention values in one sentence).

This kind of head whose attention rays are messy and have no definite shape, but they may play some important roles inside the transformer is called by “Compound Self-attention head”.

This type of self-attention has a strong directivity in the process of attention flow of the model and can focus on several individual tokens, which has a mysterious effect on the model but cannot be ignored.

3.2. The importance of different shapes

This research uses the same standard importance score for each self-attention head to characterize their importance, and combined with our above classification of attention heads, we obtained the following findings.

DOI: 10.54254/2755-2721/76/20240571

3.2.1. Definition of importance. For a transformer, the Q, K, and V matrixes are used to calculate the attention values as follows [9]:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

The importance score of a head is defined by the maximum of the attention matrix.[10]

$$\text{Im}_{i,j} = \max\left(\text{Attention}_{i,j}\right) \tag{2}$$

Where $\text{Attention}_{i,j}$ is the self-attention matrix at the i th layer, j th head in a transformer.

3.2.2. Visualization of importance

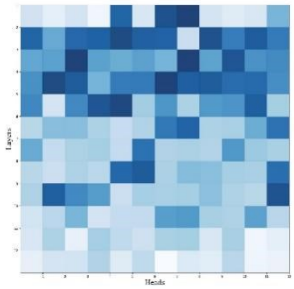

Figure 11. Importance scores visualization of "xtremedistil-l12-h384-uncased".

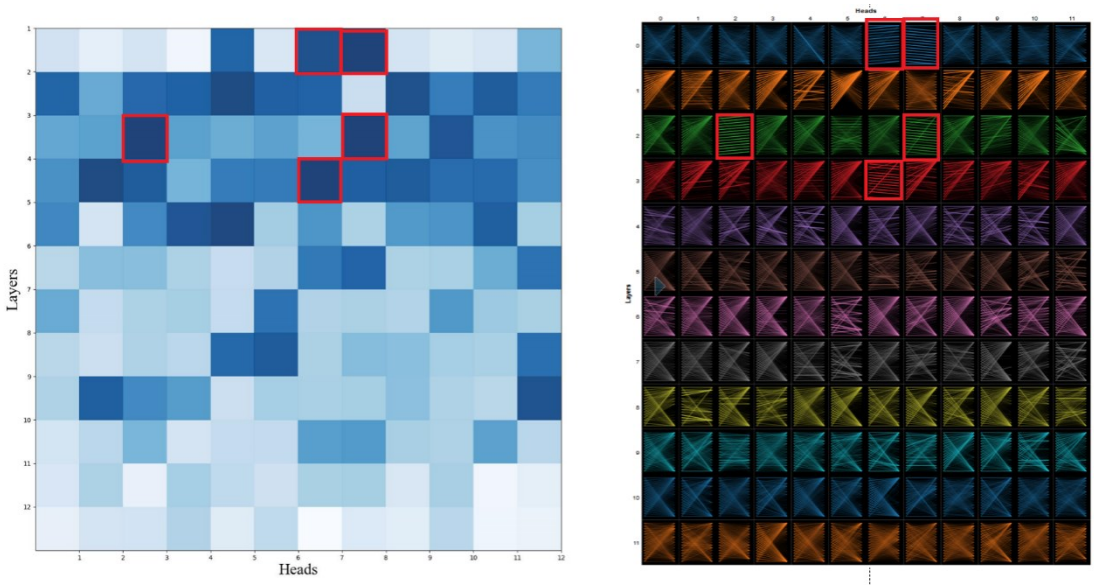

Figure 12. The importance scores of Parallel head visualization.

Based on the above visualization and previous studies [10], it is expected that the first four layers of heads are more important. Combining the previous knowledge of different types of heads, we can see that parallel heads are more important than other types of heads by comparing their importance.

3.2.3. Relative importance

The relative importance of the transformer can be regarded as the differences in important scores of heads. The standard deviation can describe this relative importance, that is, the magnitude of the

DOI: 10.54254/2755-2721/76/20240571

difference in the importance of heads from head to head. If the standard deviation is large, the importance score fluctuates greatly, indicating that the importance of heads is varied.

$$RI = \sqrt{\frac{1}{n}\sum_{1\le i\le m,1\le j\le k}\left(Im_{i,j} - \overline{Im}\right)^2} \tag{3}$$

The following are the statistics of the importance scores obtained by different models under the same input text.

Table 1. Importance score statistical results.

	xtremedistil	bert	gpt2	roberta
Number of heads	144	144	144	144
Mean	0.591469	0.839072	1.0	0.827780
RI	0.224299	0.156224	0.0	0.167467
Min	0.185897	0.219898	1.0	0.231707
25%	0.430913	0.766107	1.0	0.742653
50%	0.498125	0.896529	1.0	0.885998
75%	0.799879	0.965649	1.0	0.955006
Max	0.999994	1.000000	1.0	1.000000

The important scores of gpt2 are identical, which means that it is not proper to prune any head from it or it is proper to give up all heads in it. Based on the classification of self-attentional heads, it is easy to find that the same type of self-attentional heads may be consistent with the relative importance of zero here. It is impossible to tell which is more important because it is the same type.

For a normal model, in contrast to the layers that are in the back, the layers that are in the front are more important, which previous studies[10] have shown. The sparse parallel heads happen to be mainly present in the front. This is also known based on the importance score of the parallel head, which is the more important head.

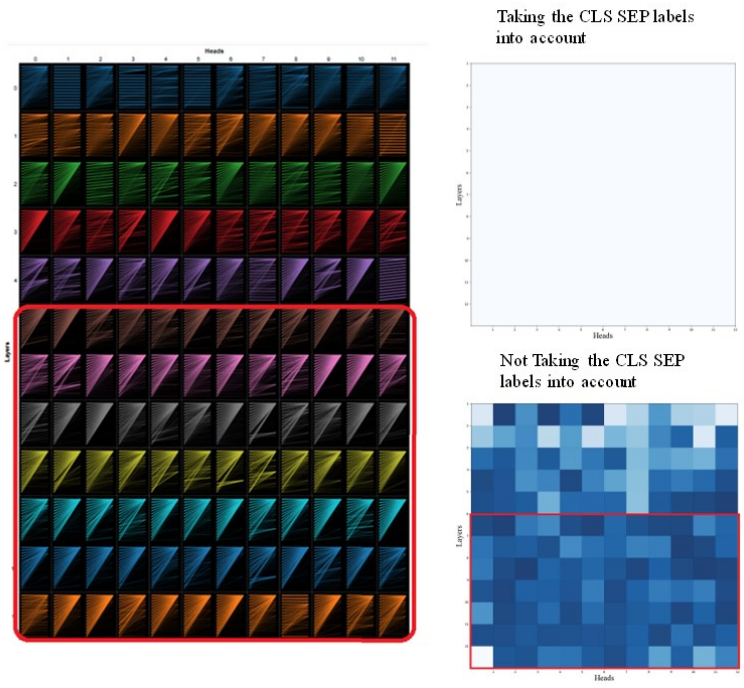

Figure 13. Global visualization of self-attention in “gpt2”.

DOI: 10.54254/2755-2721/76/20240571

Figure 13 is the visualization of gpt2. On the right-hand side is a heat map of its importance scores. Taking the [CLS] and [SEP] tags into account, each head plays the same importance in gpt2.

The effect of this type of attention on its importance is particularly strong in the latter layer, where the importance scores of both radioactive heads are very close to each other, while the parallel heads remain the most important.

Therefore, generally based on importance scores, different types of heads may have different importance. The difference in importance of the same head is not obvious. As a result, the transformer has different attention heads with their roles to carry out information transmission, just like an organism.

3.2.4. The comparison of RI

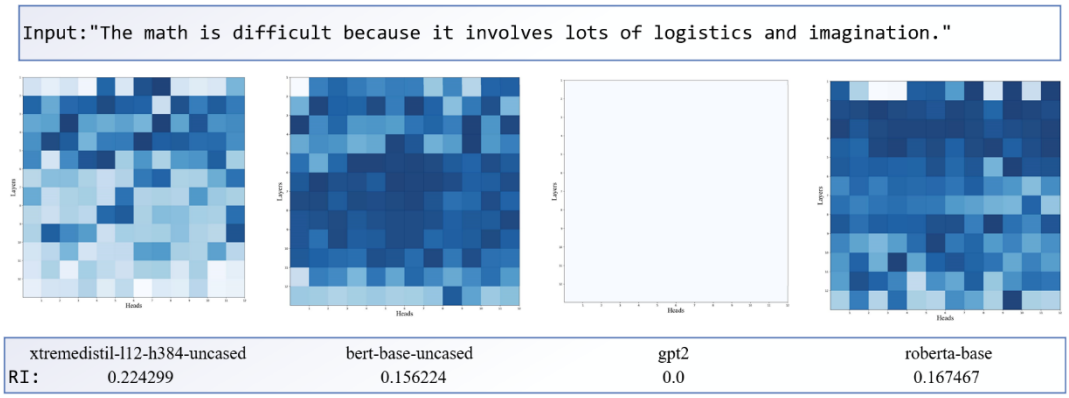

Figure 14. The comparison of RI.

The relative importance RI of the head inside the new model is greater, meaning that the importance of the head varies greatly from head to head. While the old models with poor relative performance, their importance scores are not different, and the RI index is low, which means that the importance of the head to the model is similar, and this point will be echoed in the diversity topic of the head later. If the head is equally important to the model, this may be the result of inadequate training that does not differentiate the head from its function.

3.3. The direction of flow attention

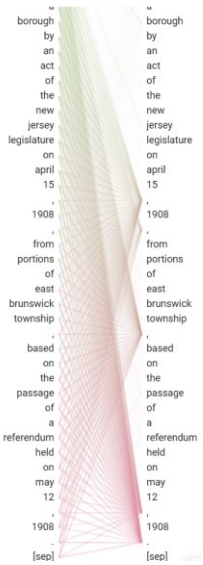

Figure 15. Attention flow of Bert on the Attention-Viz [2].

DOI: 10.54254/2755-2721/76/20240571

Attention tends to flow to CLS and SEP tags, as has been shown in previous studies[8]. In the work here, it is often reflected in the place of the last head on the last layer. Through some observation, attention often tends to go to the middle of the sentence or the end of the sentence. This means that the cls and sep tags need to be discussed separately.

3.4. The Comparison of Reactions

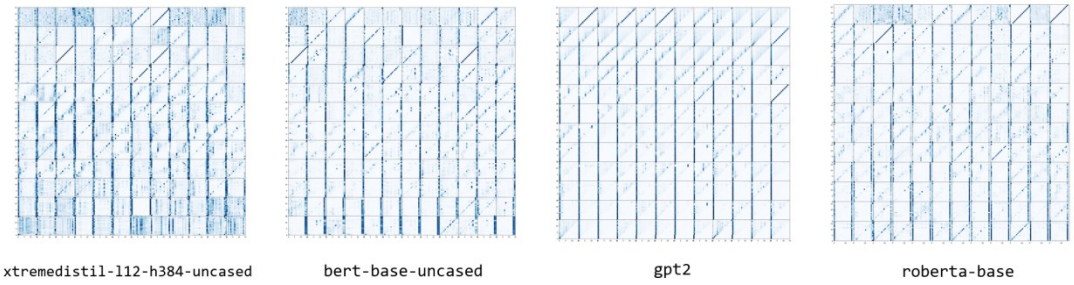

Figure 16. The comparison of attention reaction heatmaps.

It can be seen that the new model, will react more strongly to the same input, while the other three old models have relatively more white space in the heat map, especially in the comparison of RI, it can be seen that its head's response to the input is quite homogenized, making the flow of attention in gpt2 very monotonous. This may be the result of inadequate training and utilization of the attention matrix.

3.5. Inherent properties of the self-attentional head

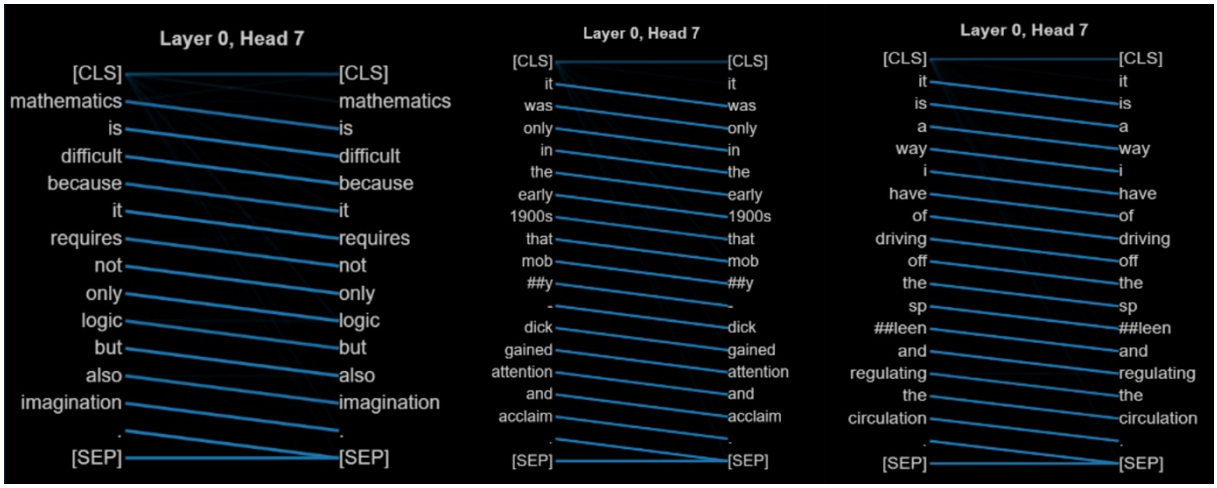

Figure 17. The attention rays of different inputs in Layer 0 Head 7.

The above three sets of attention rays come from the same head of the new model, and it can be found that they are all shaped like parallel heads. There are also a lot of groups of heads that are not convenient to show in the paper.

However, the other head types also remained largely unchanged, perhaps due to the internal parameters of the trained model. So the types of heads become, in the course of training, inherent properties of the new models they form.

4. Conclusion

Through the above visual observation and quantification to a certain extent, it is easy to find that some heads have obvious common characteristics, and they can be classified as Parallel self-attention heads,

DOI: 10.54254/2755-2721/76/20240571

Radioactive self-attention heads, Homogeneous self-attention heads, X-type self-attention heads, and Compound self-attention heads. Parallel heads are the rarest and usually have the highest importance scores. The parallel heads usually appear in the previous layers, and the later layers are less important. Model performance may not be as good when the importance and variety of heads are more homogeneous. In excellent new models, CLS and SEP have little influence on the relative importance of their heads, while attention tends to move to these two labels, and the relative importance difference is significant. The relative attention of the old model will be easily drawn to CLS and SEP, and the relative importance difference is not as significant as that of the new model. This paper hopes to provide some new perspectives in the analysis of interpretability, make use of the characteristics of the black box, do more experiments to observe the model's attention response, and give some new training ideas, such as training individual heads separately to differentiate the heads inside the model. The disadvantage of the research is that the support of the underlying mathematics is thin and the number of examples of the model is limited. In the future, we can try to study the current open problems such as model biology and the direction of attention.

References

[1] Vig, J. (2019). A multiscale visualization of attention in the transformer model. arXiv (Cornell University). http://export.arxiv.org/pdf/1906.05714

[2] Yeh, C. V., Chen, Y., Wu, A., Chen, C., Viégas, F. B., & Wattenberg, M. (2023). AttentionViZ: A Global View of Transformer Attention. IEEE Transactions on Visualization and Computer Graphics, 1–11. https://doi.org/10.1109/tvcg.2023.3327163

[3] Lin, T., Wang, Y., Li, X., & Qiu, X. (2022). A survey of transformers. AI Open, 3, 111–132. https://doi.org/10.1016/j.aiopen.2022.10.001

[4] Mukherjee, S., Awadallah, A. H., & Gao, J. (2021). XtremeDistilTransformers: Task transfer for task-agnostic distillation. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2106.04563

[5] Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv (Cornell University). https://arxiv.org/pdf/1810.04805v2

[6] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. Language Models are Unsupervised Multitask Learners (openai.com)

[7] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). ROBERTA: A robustly optimized BERT pretraining approach. arXiv (Cornell University). https://doi.org/10.48550/arxiv.1907.11692

[8] Hao, Y., Dong, L., Wei, F., & Xu, K. (2021). Self-Attention Attribution: Interpreting information interactions inside the transformer. Proceedings of the . . . AAAI Conference on Artificial Intelligence, 35(14), 12963–12971. https://doi.org/10.1609/aaai.v35i14.17533

[9] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All Need. arXiv (Cornell University), 30, 5998–6008. https://arxiv.org/pdf/1706.03762v5

[10] Wang, Z., Turko, R., & Chau, D. H. (2021). Dodrio: Exploring Transformer Models with Interactive Visualization. arXiv (Cornell University). http://arxiv.org/pdf/2103.14625.pdf