Explaining Attention mechanism in Transformers: from a gender information perspective

Zhengyun YU
Pembroke College, the University of Oxford
zhengyun.yu@pmb.ox.ac.uk

Abstract

Transformer-based models significantly advanced the state-of-the-art in both linguistic and computer vision tasks, and the working mechanism of its key component, the self-attention mechanism, lacks a proper in-house explanation. This paper illustrates the role of attention mechanism from a token level, focusing on gender information propagation in language modeling tasks. Start by visualizing attention weight and quantifying attention flow with depth, I then select a class-specified gradient-based method named Attentive Class Activation Tokens(AttCAT) [1] to analyze the mechanism of self-attention units via disentangling information flow. Inspired by another token-level method for obtaining interpretable dense subspaces named Densray(DR) [2], I compare the two methods in language modeling tasks and propose an integrated bias visualization method Attentive Densray(AttDR) to further look into gender information propagation with depth in attention blocks. Experiments are performed on several templates and datasets, which show high similarity in AttCAT and DR scores, and clearer explanation by AttDR.

1 Introduction and related work

The key innovation behind the Transformers models[3] is the stacking of multi-head self-attention blocks to extract different levels of features from sequential input tokens, therefore explaining the in-house role of the self-attention mechanism [4] with depth is crucial for output interpretation of the Transformers-based models and gaining insights into its internal formation process of bias.

1.1 The architecture of the self-attention mechanism in Transformers model

The encoder of Transformers typically stack L identical blocks, within each block are two sub-layers: a multi-head self-attention module and a feed-forward network module (denoted by weight matrix W^O), each coupled with skip connection and layer normalization. In the first module, the embeddings of all tokens in the previous layer $h^{l-1} = \{h_1^{l-1}, h_2^{l-1}, ..., h_n^{l-1}\}$ are combined to generate the output h_i^l of the i^{th} token through multi-head self-attention mechanism:

$$\alpha_{i,j}^l = softmax(\frac{Q(h_i^{l-1})K(h_i^{l-1})^T}{\sqrt{d}} \in \mathbb{R}, \quad h_i^l = W^O(\sum_{j=1}^{n} \alpha_{i,j} V(h_j^{l-1}) + h_i^{l-1}) \in \mathbb{R}^d \qquad (1)$$

where $\alpha_{i,j}^l$ is the attention weight of i^{th} token from j^{th} token; d is the dimension of the vectors, and $Q(\cdot), K(\cdot), V(\cdot)$ are the query, key and value transformations respectively. The hidden states of each head are concentrated and put through a linear transformation as the multi-head attention's outputs.

1.2 Token-level explanation of Transformers' attention

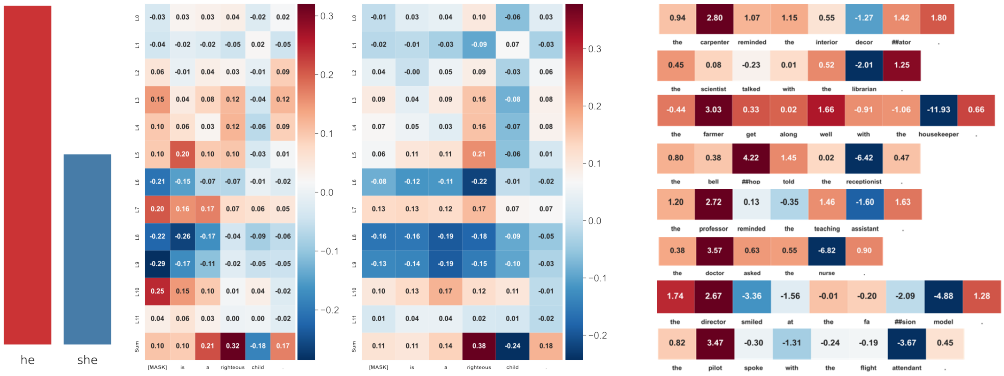
Observed that the output sequence has a one-to-one correspondence to its input sequence, research surges on the token-level interpretation of Transformers. Among them, attention-based methods

such as Attflow and Attrollout [5] improperly take high attention weights as indicators of important information flow [6]; gradient-based methods and LRP-based methods ignore the skip connection term and all three of them ignore the token magnitude term $V(h_i^l)$. Attentive Class Activation Tokens(AttCAT)[1], is the only method that utilizes both attention weights(α) for capturing interactions between features, and gradient information of hidden states (∇h) for quantifying impact related to each class output, thus it is selected for explaining the deep attention networks in Transformers.

While Attrollout and Attflow [5] can not reasonably model information flow in attention blocks, they inspire this work in two ways: Attention flow through self-attention blocks and skip connection can be disentangled to study two modules' cooperation in Transformers, which is illustrated in section 2.1; The token-level attention map of pronoun prediction indicates potential gender tendency in Transformers-based language models, and this work aims to reveal the formation process of this inner tendency via different methods of quantifying attention flow in Transformers.

1.3 Densray: a projection-based token-level visualization method

DensRay(DR) is adapted for contextualized word embeddings by Liang et al[2]. For quantifying binary gender information, it identifies the "gender subspace" using a set of gendered words, and find an orthogonal matrix to concentrate all gender information in the first dimension. In Transformers-based language models, per-layer embeddings of a token are rotated to get the DR scores(shown in Fig 1.1(a)(right) with signs indicating their gender directions). DR is use as a baseline method to compare with AttCAT, and the two are integrated as Attentive Densray(AttDR) in section 3.2.



(a) predicting probability, AttDR(middle) and DR(right) scores (b) the sum AttDR score of all layers

Figure 1.1: Visualization of Densray score and AttDR score in gender bias probing sentences

2 The role of Attention mechanism in Transformers

It is obvious that from a layer-wise perspective, the self-attention weights alone in Transformers capture the contextual information of each token in that layer. When considering skip connection and depth, this section illustrates the in-house role of attention units from two pairs of aspects: (a) layer-wise and depth-related; (b) feed-forward propagation and disentangling information;

2.1 Quantify the attention flow through skip connection layer-wise and with depth

Utilized Attflow mentioned in section 1, this work computes the token-level proportion of attention flow through skip connection, and finds that skip connection occupies a larger portion in shallower layers and shrinks in deeper layers, indicating more contextualized information is learned in deeper layers. Prior work [6] also looks into this depth-related proportion in BERT[8], and gives the same decreasing trend with a much larger percentage of 75.4%. Besides the different definitions of information flow and methods to quantify it: feedforward and back-propagation, the gap is mainly due to the ignorance of the feature's magnitude term $V(x_i^l)$ in the Attflow method. Also, experiments on the SST-2 fine-tuned BERT model show that important words skip most of the attention layers.

2.2 AttCAT: a gradient-based method via disentangling the information flow in Transformers

In section2.1, the cooperation is studied by computing the attention flow ratio and make comparisons from the view of feed-forward propagation. This section introduces AttCAT from the view of back-propagating the output to the input. In each layer of the Transformers encoder, the self-attention blocks capture the interactions between tokens while skip connection modules serve as shortcuts bridging the input and output of the self-attention operation, completly preserving the last layer's information From the view of decomposing information flow, AttCAT considered the Transformers encoder as a recurrent operator that adds the representation of token interactions (via self-attention mechanism) onto the original representation of the token (via skip connection) layer by layer:

$$Information(x_i^l) = Information(x_i^{l-1}) + \sum_{j \neq i} Interaction(x_i^{l-1}, x_j^{l-1}) \tag{2}$$

where $Interaction(x_i^{l-1}, x_j^{l-1})$ denotes the pairwise interaction between the i^{th} token and the j^{th} token. Utilizing equation2 to propagate information back to the input, the encoder output of the i^{th} token can be decomposed into the original information plus different levels of contextual information received from all other tokens at different layers. Stacking attention blocks together, this work therefore interprets the encoder as a combinator of weighted token-level interactions, enabling the model to capture features that consist of any level of contextualized information.

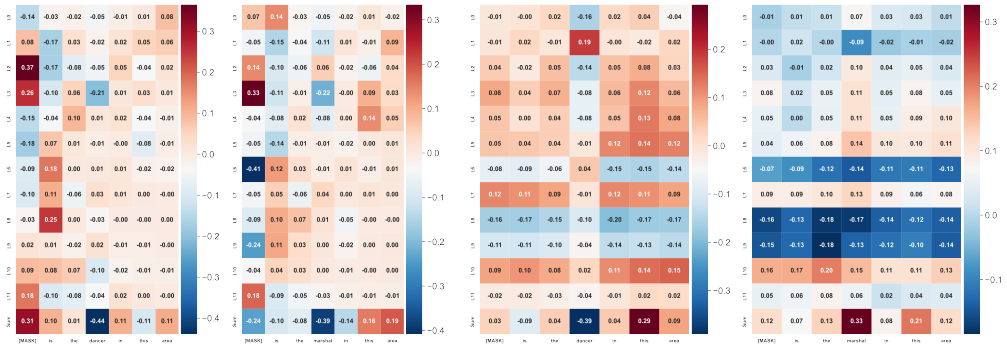
Based on disentangling information as above, AttCAT integrated attention weights and a gradient-based method Class Activation Tokens(CAT) to generate an element-wise product in each layer, then sum them up to give the final token-level AttCAT scores, which can be represented as:

$$CAT_i^L = \frac{\partial y^c}{\partial h_i^L)} \odot h_i^L, \quad AttCAT_i^l = \mathbb{E}_H(\alpha_i^l \cdot CAT_i^L), \quad AttCAT_i = \sum_{l=1}^{L} AttCAT_i^l \tag{3}$$

where $\odot$ is the Hadmard product, h_i^L denotes the output of i^{th} token from the last Transformers layer L, and $\mathbb{E}_H(\cdot)$ denotes the average over multiple heads.

2.3 Evaluation and generalization of AttCAT

AttCAT has several advantages: The CAT score capture both the magnitude and impact of the features, which improves methods in section1.2. Also, attention weights module interactions and the summation is reasonable via information disentangling. Moreover, the direction of AttCAT impact score represents whether each token makes a positive or negative impact on the specified output class, which is more useful for explaining classification outputs compared to all other prior methods. However, the class-specified score may be one-sided in a multi-classification task. Since CAT scores are fixed with depth and only related to a specific class, tokens with the same sum of attention weights and output states at L^{th} layer impact the same, missing their different impacts on other classes. In multi-classification tasks, some patterns may contribute to several classes while others are rather unique. Therefore, depending on AttCAT score in one desired class alone to capture token-level impact is too narrow, a feasible improvement may be including a global indicator that represents the relative magnitude of the impact score compared to other classes.



(a) AttCAT scores of dancer and marshal (b) DR scores of dancer and marshal

Figure 2.1: Normalized scores heatmap for occupation templates examples of BERT

AttCAT can be generalized to certain language modeling tasks such as pronouns predicting by considering the prediction as a classification result in all potential pronouns. Since the prediction of "he" and "she" reveals gender tendency, and AttCAT impact scores can explain Transformers's token-level attention, this work generates AttCAT heatmaps in masked sentences to investigate the contextual gender information forming through layers, examples shown in Fig 2.1(a). From the last row of summing scores, the top-2 largest absolute values appear in token "[MASK]" and the occupation word, and this trend appears in all occupation templates(OCCTMP) examples. Since the attention on the "[MASK]" is obvious due to the task itself, this result indicates that the occupation words impact the most when the model chooses a pronoun to fill in. As it goes deeper, the magnitude of impact scores first rises, then decreases and concentrates on the 2 top tokens while scores for the other tokens almost fade out. The rise indicates that gender information is more contextualized with a higher level of attention in Transformer-based language models, and the fading-out phenomenon is because raw attention weights are generally uniform in deeper layers, and AttCAT essentially just scales the weights by the depth-fixed CAT score, therefore for each token, the propagation dynamic doesn't change. Also, the concentration effect of AttCAT is beneficial since useless information receives almost no attention in further propagation.

3 Gender information propagating through layers

3.1 Comparing results of DR and AttCAT

To further look into Transformers's attention to gender information, this work uses another projection-based method Densray to quantify token-level gender information per layer and compares the results of the two methods. Based on Fig 2.1(b), Densray assigns a positive score to the male-correlated occupation "marshal" and a negative score to its female alternative "dancer" with a significant DR scores gap in the last row. In contrast, AttCAT assigns the same sign to the 2 types of occupation words, which can not indicate gender direction. As scores propagate with depth, the increase of magnitude in DR scores continues to appear in deeper layers than in AttCAT with a slightly downward trend only in the last two layers, showing a longer gender information propagating process.

3.2 AttDR: an integrated method for visualizing binary bias information

Integrating the attentive concentration effect in AttCAT and gender direction indicator in Densray, this work proposes a novel method for token-level gender information visualization: Attentive Densray(AttDR), which performs an element-wise product of attention weights and DR scores for each token in each layer. With the same notation in equation 3, it is calculated as:

$$AttDR_i^l = \mathbb{E}_H(\alpha_i^l \cdot DR_i^L), \quad AttDR_i = \sum_{l=1}^{L} AttDR_i^l \tag{4}$$

From Fig 1.1 (a), AttDR allocates a high score to the occupation word with directionality as DR, while putting attentive emphasis on other words such as "a" useful to the pronoun-predicting task. Interestingly, the direction flipped several times in layers 6, 8, and 9 respectively as in DR heatmap, while the magnitude is rather consistent. Without a related study, I infer that this common phenomenon is due to the dynamic in DR direction through layers. Fig 1.1 (b) also shows distinct gender tendencies of BERT reflected by AttDR on similar job pairs used by Kotek et.al[7] for probing gender bias.

4 Experiments detail and Discussion

4.1 Experiments settings

Transformer models BERT [8]is one of the most representative Transformer models. In the experiments, I lowercase all text and use "bert-base-uncased" from Huggingface[1].
Densray gender word list and corpus are adapted from from Reysens et.al[9]. 37 masculine and 37 feminine one-token words are used as the gender word list, with text data from Wikipedia that contains 5,0000 occurrences of gender words with balanced occurrences in the two sets.

[1]https://huggingface.co/

Table 4.1: DR score and AttCAT score comparisons in distances

templates \ normalized distance	Cosine distance	Euclidean distance	Manhattan distance	Haversine distance
$OCCTMP$	0.474	0.351	0.288	0.531
CSP	0.577	0.305	0.229	non applicable
$ADJTMP^1$	**0.329** (0.418)	0.306 (0.362)	0.256 (0.296)	0.451 (0.666)
$ADJTMP^2$	0.364 (0.363)	0.337 (0.332)	0.274 (0.277)	0.550 (0.606)
$NTMP$	0.393 (0.372)	0.356 (0.348)	0.295(0.284)	0.595 (0.594)

Gender-associated templates and datasets: (a) occupation templates(OCCTMP)[10] with 320 existing occupations are used as attributes (b) Adjective templates(ADJTMP) and (c) Noun templates(NTMP) are filled with female and male stereotype words[11] with three categories, $ADJTMP^1$: intelligence vs appearance, $ADJTMP^2$: strength vs weakness, and $NTMP$: science vs art; CrowS-Pairs(CSP) dataset[12], which is a benchmark dataset that addresses stereotypes associated with historically disadvantaged groups in the US. 197 gender-related examples are selected from CSP and modified by replacing the gender words with "[MASK]" for language modeling tasks.

4.2 Evaluation metrics and result analysis

This work calculates each example sentence's token-level AttCAT impact scores(code modified from AttCAT[1]), DR gender scores(code modified from DR[2]) and AttDR scores(code origin) per layer, and compares the results. Quantitative comparisons are made layer-wise, while heatmap visualization is used for a qualitative comparison with depth.

Quantitative comparisons: Since AttCAT scores cannot indicate gender direction, this work uses the absolute value of the sum of scores in all layers and performs L2-normalization on the sentence vector. 3 pairwise distance matrices are calculated: Cosine distance, Euclidean distance, and Manhattan distance, using the whole sentence without "[MASK]", and Haversine distance is calculated using the absolute score of token "[MASK]" and the gender-associated word. Using sklearn[2], all distances are then normalized by sequence length. The Results are shown in Table 4.1 with female examples in latter 3 templates shown in brackets. The table reveals the high similarity between the two methods, especially in adjectives associated with male while they differ the most in OCCTMP and CSP dataset due to more gender-neutral words. This validates the generalization of AttCAT in language modeling tasks and reveals that Transformers form gender tendency through attention flow propagation.

Qualitative comparisons: As shown in Fig 1.1 and Fig2.1, AttCAT scores are without direction, relatively less sensitive to gender information, and focuses more on task-oriented useful words, with the propagation dynamic dominated completely by trend of raw attention weights. By contrast, DR scores provide salient gender information with direction but lack token-level precision as similar scores sometimes dominate the whole sentence, and the propagation is unstable in gender direction. AttDR integrates the benefits of the two methods and provides more stable and clearer gender attention scores with direction, which is because the element-wise multiplication of attention weights shrinks the perturbation of gender directions through layers and disgards the useless information.

5 Conclusion

In this work, I use the Attflow and the AttCAT methods to disentangle attention flow in Transformers, showing that it is the recurrent effect of depth and the cooperation of skip connection modules and self-attention modules together make the multi-level features feasible. To further explain the Transformers's attention in language modeling tasks, I investigate the formation of gender tendency by making layer-wise and depth-related comparisons between two token-level methods: AttCAT and Densray, and finally propose an integrated method AttDR. Empirical results show high consistency of the two scores in gender-related templates, both indicate that Transformers learn gender information in deeper layers and make predictions based on certain stereotypical words. AttCAT scores can be extended to quantify tasks-specified attentive bias information in state-of-the-art Transformer-based large language models(LLM)[13] for enhancing fairness and models alignment. Heatmap also shows the stable and precise attentive gender scores with direction provided by AttDR, which can be extended to vision Transformers for indicating attention on bias information in the future.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.metrics.pairwise_distances.html

References

[1] Qiang, Yao et al. "AttCAT: Explaining Transformers via Attentive Class Activation Tokens." Neural Information Processing Systems (2022).

[2] Liang, Sheng et al. "Monolingual and Multilingual Reduction of Gender Bias in Contextualized Representations." International Conference on Computational Linguistics (2020).

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017.

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

[5] Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. arXiv preprint arXiv:2005.00928, 2020.

[6] Kaiji Lu, Zifan Wang, Piotr Mardziel, and Anupam Datta. Influence patterns for explaining information flow in bert. Advances in Neural Information Processing Systems, 34, 2021.

[7] Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in Large Language Models. In Proceedings of The ACM Collective Intelligence Conference (CI '23). Association for Computing Machinery, New York, NY, USA, 12–24.

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

[9] Reusens, M., Borchert, P., Mieskes, M., De Weerdt, J., & Baesens, B. (2023). Investigating Bias in Multilingual Language Models: Cross-Lingual Transfer of Debiasing Techniques.

[10] Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 166–172, Florence, Italy, August. Association for Computational Linguistics.

[11] Saket Karve, Lyle Ungar, and Jo~ao Sedoc. 2019. Conceptor debiasing of word representations evaluated on WEAT. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, pages 40–48.

[12] Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

[13] Zhao W X, Zhou K, Li J, et al. A survey of large language models[J]. arXiv preprint arXiv:2303.18223, 2023.