Insights from Attention: New Approaches to Visualizing Transformer Model Output

Raphael Kunert, Adrian Jobst, Andreas Fricke, Willy Scheibel[0000−0002−7885−9857], and Jürgen Döllner

Hasso Plattner Institute, Digital Engineering Faculty,
University of Potsdam, Germany
raphael.kunert@student.hpi.uni-potsdam.de
{adrian.jobst, andreas.fricke, willy.scheibel,
juergen.doellner}@hpi.uni-potsdam.de

Abstract. Recent advancements in language models, particularly those based on the Transformer architecture, have led to remarkable achievements in natural language processing. However, the increasing complexity and size of these models pose significant challenges for understanding their behavior and decision-making processes. In this work, we propose a set of new attention visualization techniques that address these challenges by improving model explainability and interpretability. The key improvements include new layouts that better handle the large number of tokens present in prompt and answer scenarios, making long distance attention relationships more comprehensible. Our techniques have the potential to enable researchers and practitioners to better understand the decision-making processes of large language models and identify potential sources of bias or errors. While detailed user studies and evaluations are outside the scope of this work, we discuss potential use cases for our visualization techniques and present directions for future research.

Keywords: Visualization · Attention Mechanism · Transformer · Explainable AI

1 Introduction

Significant progress has been made in the field of artificial intelligence in recent years, particularly in language models and image processing, whose machine-learning models may soon find wide application [4]. Retrospectively, a milestone in the area of language models was the introduction of the Transformer architecture [21]. Models based on it, such as GPT [15] and BERT [7], performed well in established benchmarks but were still inferior to human performance [26, 25]. Subsequent work constantly improved the performance of such models [16, 27, 5] and just recently, OpenAI released GPT-4, which can now write convincing texts and can perform at the human level in various tasks [14]. Despite the good this technology could do for fields like healthcare and science, there are also major concerns about potential harms, for instance, its usage for influence

operations [10] or fraud [9]. Even intended usage might affect a vast amount of people. A recent study suggests that around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of modern language models [8]. Apart from economic inequalities a large-scale usage of such technology might introduce, there are also others, for instance, machine learning models might misrepresent people by stereotyping, underrepresent or even erase minorities, or also overrepresent certain groups [4]. Therefore general model understandability and explainability are of utter importance to judging their implications. In the field of explainable AI, visualization is considered a key technique for improving explainability [1] and was already used to show misrepresentations, e.g. gender bias [22].

In this work, we propose a set of new attention visualization techniques that better adapt to the current far larger size of generative transformer models. The key improvement is a few layouts that better handle a large number of tokens in a prompt-and-answer scenario. These aim to make long-distance attention relationships more comprehensible, as well as explore the effects of the input vocabulary on the intermediate attention weights. The evaluation of the visualizations and detailed user studies are out of the scope of this paper, but potential use cases will be discussed. The remainder of this paper is structured as follows. Section 2 presents previous work on attention visualization. Section 3 is explaining our visualization approaches. We conclude this work in section 4 and present directions for future work.

2 Related Work

The current state-of-the-art for attention visualization by Vaswani et al. [21] is shown in fig. 1. It consists of two columns of words, where the left column represents the input of the transformer and the right column represents its output. This design allows for the complete input-output visualization of the original transformer architecture, which was built with machine translation in mind [21]. Similar layouts were used in other works, where the underlying representation was described as a bipartite graph [12, 13, 19]. These layouts work great for smaller machine translation models but were not designed for larger GPT (Generative Pre-trained Transformer) style architectures. A visualization like the one in fig. 1 can not be properly extended to longer sentences, as it is strictly limited by screen size. Also, retracing long-range attention relationships would be quite tedious. A similar approach like in fig. 1 is taken by Vig where a similar layout is used to visualize attention heads which revealed the previously mentioned gender bias misrepresentation [23, 22]. Similarly, Vig and Belinkov used these visualizations to analyze the interaction between attention and syntax [24]. Both also used heatmap views to support their analysis. Heatmap layouts have also been used in previous work to visualize attention matrices [18, 17, 3] and have also often been combined with a layout similar to the one presented by Vaswani [12, 13]. DeRose and Berger embedded matrices in a circular layout in order to visualize attention within and across layers [6]. Other approaches didn't focus solely

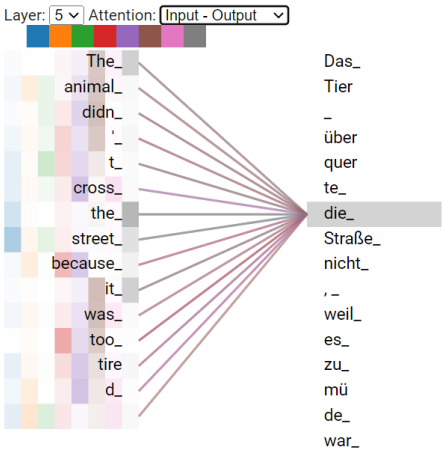

Fig. 1: Interactive visualization included in the `tensor2tensor` [20] python module

Fig. 2: Our text layout

on the attention mechanism, like the work of Aken et al. [2] where a scatter plot is used to show semantic representations, and the work of Lal et al. [11] where multiple layouts where used to track and visualize token embeddings through layers.

3 Attention Visualization

The main premise in developing a new visualization for attention weights was to update it to work with GPT-style transformer models and to make it more scalable. The following subsections describe the different approaches and outline the benefits and possible improvements. In all of the following visualizations, tokens are represented as circular nodes with the corresponding word in the center. In contrast to the state-of-the-art approach, only a single set of tokens is used, since the GPT architecture combines input and output into one set. All nodes are color-coded so that all the same tokens have the same color. This is to help identify repeating patterns in the output and analyze word frequency. The nodes are connected with lines, where opacity and stroke weight are proportional to the attention weights measured at training or inference.

3.1 Data

The data presented in this section were generated using a simple Transformer model trained on "The Adventures of Sherlock Holmes" by Arthur Conan Doyle. It consists of an embedding block, that handles both token embedding and positional embedding, a single modified encoder layer that allows for masked attention, and a language model head that converts the attended and forwarded

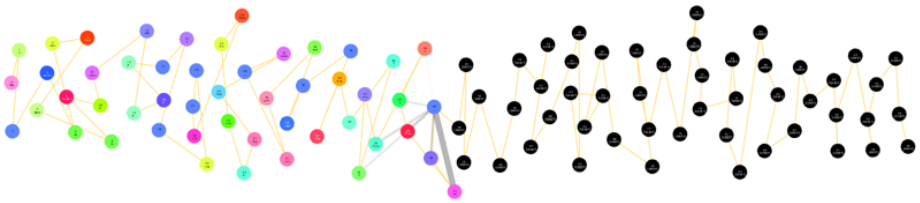

Fig. 3: Dispersed line layout

embedding vectors into vocabulary-sized probability distributions. To extract data from the model, the attention layer code was modified to store attention weights in global storage before storing them in a file. Thus, attention can be stored both during training and inference. In addition to the attention weights, we also store the entire vocabulary as well as the set of tokens used for training or generated during inference.

3.2 Text Layout

In this layout, all tokens are displayed in several lines below each other. To facilitate the analysis of the sentence structure, the text is separated by punctuation marks. The result can be seen in fig. 2. Sentence-by-sentence viewing allows for a more readable view than one line and allows repetitive patterns to be identified. However, the text visualization approach has drawbacks because nodes are usually close together and attention lines in between sentences are difficult to distinguish.

3.3 Dispersed Line Layout

In this layout nodes are placed next to each other and are moved up or down by a random amount. This then allows a more space-saving arrangement by moving all nodes closer together. In the resulting layout, however, overlapping nodes may occur; these are separated by iteratively moving all nodes apart from each other. This approach of iterative distancing is also followed in the other layouts described later. The dispersed line layout can be seen in fig. 3. The final visualization maps all tokens to a 2D space in which the attention lines are largely free of overlap. However, overlaps may still occur with certain constellations of nodes. This layout solves both problems mentioned with the existing state-of-the-art approaches. It is both compact and intended for GPT-style models, as it works with just one set of tokens. However, this layout has a tradeoff between saving space and readability. To mitigate readability issues due to dispersed nodes, we connect all nodes in text order with a line called a *storyline*.

3.4 Data Dispersed Line Visualization

Instead of using random offsets to move nodes, this extra dimension can also be used to display additional information. In fig. 4, this dimension encodes the

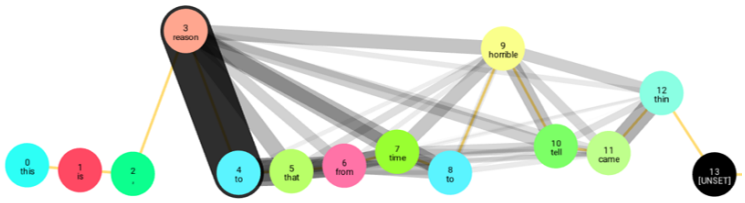

Fig. 4: Data dispersed line layout without focus

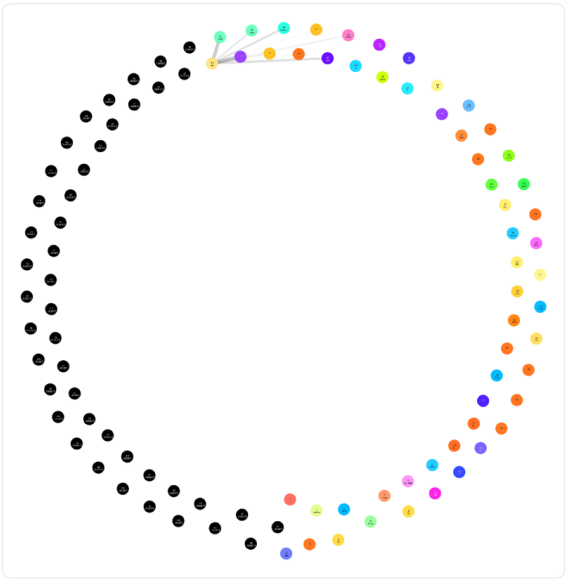

Fig. 5: Dispersed circle layout

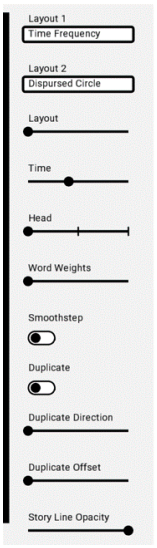

Fig. 6: Side bar

word frequencies computed from the training data, where rare words are moved toward the top. For instance, one can see that simple words like the prompt "this" and "is" are at the bottom, while less occuring words like, "reason" or "horrible" are at the top.

3.5 Dispersed Circle Layout

The final layout arranges all tokens in a circle, as seen in fig. 5. The chosen arrangement allows a high number of tokens to be displayed, and attention lines are also limited by the circumference of the circle. This makes it easy to trace even distant relationships. Another advantage is that the curvature of the circle does not allow for completely overlapping lines, so the viewer can distinguish all lines.

3.6 Sidebar UI

The sidebar in the visualization framework was built from scratch in Processing, as the platform doesn't offer built-in UI elements, which can be seen in fig. 6. It contains various interactive elements, such as two layout dropdowns, a slider for interpolation, a time slider for advancing the epoch or inference step, a slider for controlling the strength of the word weight visualization, and a selection slider for the attention head. An additional smooth step operation performed on the displayed attention lines can be toggled. This reduces the opacity and size of small lines and increases the strength of larger lines. A toggle for a second set of nodes is needed for the state-of-the-art visualization Additionally, the sidebar includes two sliders for offset length and direction and a slider for setting the opacity of the yellow storyline.

3.7 Use Cases

The visualizer created in Processing accepts data in JSON format, making it adaptable to different sources and web APIs with minor adjustments. It has various use cases, such as displaying tokens in a natural language format for finding sentence and word patterns using the layout from section 3.2, analyzing long attention spans using the dispersed circle layout in section 3.5, finding highly attended nodes by adjusting the word weight slider as well as exploring the impact of the vocabulary and its word frequencies on attention with the dispersed line layout in section 3.3.

4 Conclusions

In this paper, we discussed our visualization framework written in Processing that allows users to visualize multi-head attention data from larger GPT-style models interactively and we also described potential use cases. We finally want to suggest a few improvements to the techniques described in this paper. Firstly, the proposed techniques could be developed as a Python module to make the framework more accessible to a wider range of users. Additionally, exploring alternative color schemes, such as perceptually uniform color maps, could improve the visualization's accessibility. Furthermore, incorporating additional dimensions on the second axis of the line layouts, such as grammatical word categories or cumulative attention, could provide more insights into the attention patterns of language models. Addressing these future work items would enable users to effectively interpret and utilize the data provided by the attention visualization framework.

References

1. Adadi, A., Berrada, M.: Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). IEEE Access **6**, 52138–52160 (2018). https://doi.org/10.1109/ACCESS.2018.2870052

2. Aken, B.v., Winter, B., Löser, A., Gers, F.A.: VisBERT: Hidden-state visualizations for transformers. In: Proc. Web Conference. pp. 207–211. WWW '20, ACM (2020). https://doi.org/10.1145/3366424.3383542
3. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proc. 3rd International Conference on Learning Representations. ICLR '15, arXiv (2015). https://doi.org/10.48550/arXiv.1409.0473
4. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., et al.: On the opportunities and risks of foundation models. CoRR **cs.LG**(arXiv:2108.07258) (2022). https://doi.org/10.48550/arXiv.2108.07258, preprint
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, NIPS '20, vol. 33, pp. 1877–1901. Curran Associates, Inc. (2020)
6. DeRose, J.F., Wang, J., Berger, M.: Attention flows: Analyzing and comparing attention mechanisms in language models. IEEE TVCG **27**(2), 1160–1170 (2020). https://doi.org/10.1109/TVCG.2020.3028976
7. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proc. Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186. ACL (2019). https://doi.org/10.18653/v1/N19-1423
8. Eloundou, T., Manning, S., Mishkin, P., Rock, D.: GPTs are GPTs: An early look at the labor market impact potential of large language models. CoRR **econ.GN**(arXiv:2303.10130) (2023). https://doi.org/10.48550/arXiv.2303.10130, preprint
9. Europol: ChatGPT - the impact of large language models on law enforcement. Tech. rep., Tech Watch Flash Report from the Europol Innovation Lab, Publications Office of the European Union, Luxembourg (2023). https://doi.org/10.2813/255453
10. Goldstein, J.A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., Sedova, K.: Generative language models and automated influence operations: Emerging threats and potential mitigations. CoRR **cs.CY**(arXiv:2301.04246) (2023). https://doi.org/10.48550/arXiv.2301.04246, preprint
11. Lal, V., Ma, A., Aflalo, E., Howard, P., Simoes, A., Korat, D., Pereg, O., Singer, G., Wasserblat, M.: InterpreT: An interactive visualization tool for interpreting transformers. In: Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations. pp. 135–142. EACL '21 (2021). https://doi.org/10.18653/v1/2021.eacl-demos.17
12. Lee, J., Shin, J.H., Kim, J.S.: Interactive visualization and manipulation of attention-based neural machine translation. In: Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 121–126. EMNLP '17, ACL (2017). https://doi.org/10.18653/v1/D17-2021
13. Liu, S., Li, T., Li, Z., Srikumar, V., Pascucci, V., Bremer, P.T.: Visual interrogation of attention-based models for natural language inference and machine comprehension. In: Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 36–41. EMNLP '18 (2018). https://doi.org/10.18653/v1/D18-2007

14. OpenAI: GPT-4 technical report. CoRR **cs.CL**(arXiv:2303.08774) (2023). https://doi.org/10.48550/arXiv.2303.08774
15. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I.: Improving language understanding with unsupervised learning. Tech. rep., OpenAI (2018), https://openai.com/research/language-unsupervised
16. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. Tech. rep., OpenAI (2019)
17. Rocktäschel, T., Grefenstette, E., Hermann, K.M., Kociský, T., Blunsom, P.: Reasoning about entailment with neural attention. In: Proc. 4th International Conference on Learning Representations. ICLR '16, arXiv (2016). https://doi.org/10.48550/arXiv.1509.06664
18. Rush, A.M., Chopra, S., Weston, J.: A neural attention model for abstractive sentence summarization. In: Proc. Conference on Empirical Methods in Natural Language Processing. pp. 379–389. EMNLP '15, ACL (2015). https://doi.org/10.18653/v1/D15-1044
19. Strobelt, H., Gehrmann, S., Behrisch, M., Perer, A., Pfister, H., Rush, A.M.: Seq2seq-Vis: A visual debugging tool for sequence-to-sequence models. IEEE TVCG **25**(1), 353–363 (2018). https://doi.org/10.1109/TVCG.2018.2865044
20. Vaswani, A., Bengio, S., Brevdo, E., Chollet, F., Gomez, A.N., Gouws, S., Jones, L., Kaiser, L., Kalchbrenner, N., Parmar, N., Sepassi, R., Shazeer, N., Uszkoreit, J.: Tensor2Tensor for neural machine translation. In: Proc. Conference of the Association for Machine Translation in the Americas. AMTA '18, AMTA (2018)
21. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems. NIPS '17, vol. 30. Curran Associates, Inc. (2017)
22. Vig, J.: A multiscale visualization of attention in the transformer model. In: Proc. 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations. pp. 37–42. ACL '19, ACL (2019). https://doi.org/10.18653/v1/P19-3007
23. Vig, J.: Visualizing attention in transformer-based language representation models. CoRR **cs.HC**(arXiv:1904.02679) (2019). https://doi.org/10.48550/arXiv.1904.02679
24. Vig, J., Belinkov, Y.: Analyzing the structure of attention in a transformer language model. In: Proc. Analyzing and Interpreting Neural Networks for NLP. pp. 63–76. BlackboxNLP '19, ACL (2019). https://doi.org/10.18653/v1/W19-4808
25. Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: Superglue: A stickier benchmark for general-purpose language understanding systems. In: Advances in Neural Information Processing Systems. NIPS '32, vol. 32. Curran Associates, Inc. (2019)
26. Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.: GLUE: A multitask benchmark and analysis platform for natural language understanding. In: Proc. EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. pp. 353–355. ACL (2018). https://doi.org/10.18653/v1/W18-5446
27. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., Rush, A.: Transformers: State-of-the-art natural language processing. In: Proc. Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 38–45. ACL (2020). https://doi.org/10.18653/v1/2020.emnlp-demos.6