Transformer-specific Interpretability

Hosein Mohebbi[1] **Jaap Jumelet**[2] **Michael Hanna**[2] **Afra Alishahi**[1] **Willem Zuidema**[2]
[1] Tilburg University [2] University of Amsterdam
{h.mohebbi, a.alishahi}@tilburguniversity.edu
{j.w.d.jumelet, m.w.hanna, w.h.zuidema}@uva.nl

Abstract

Transformers have emerged as dominant players in various scientific fields, especially NLP. However, their inner workings, like many other neural networks, remain opaque. In spite of the widespread use of model-agnostic interpretability techniques, including gradient-based and occlusion-based, their shortcomings are becoming increasingly apparent for Transformer interpretation, making the field of interpretability more demanding today. In this tutorial, we will present Transformer-specific interpretability methods, a new trending approach, that make use of specific features of the Transformer architecture and are deemed more promising for understanding Transformer-based models. We start by discussing the potential pitfalls and misleading results model-agnostic approaches may produce when interpreting Transformers. Next, we discuss Transformer-specific methods, including those designed to quantify context-mixing interactions among all input pairs (as the fundamental property of the Transformer architecture) and those that combine causal methods with low-level Transformer analysis to identify particular subnetworks within a model that are responsible for specific tasks. By the end of the tutorial, we hope participants will understand the advantages (as well as current limitations) of Transformer-specific interpretability methods, along with how these can be applied to their own research.

1 Tutorial Description

With Transformers (Vaswani et al., 2017) demonstrating exceptional performance across every domain they venture into such as language, speech, vision, and music, the necessity to understand their underlying mechanisms has become more crucial than ever before. Many model-agnostic interpretability techniques that were commonly used for earlier generations of deep learning architectures, such as probing, occlusion-based, and feature attribution methods, were swiftly adapted for use with the Transformer architecture. However, these approaches demonstrate notable disagreement with each other and a lack of stability when moving from one domain to another (Neely et al., 2022; Pruthi et al., 2020; Krishna et al., 2022). Their effectiveness in drawing reliable conclusions has therefore been an ongoing matter of debate (Bibal et al., 2022).

Recently, a game-changing trend has emerged: the development of analysis methods that are precisely tailored to the model architecture of Transformers, built upon their underlying mathematical foundations. These methods make use of specific features of Transformers, including their layered structure (layers, heads, tokens), the division of labor between the attention mechanism, feed-forward layers, and residual streams. These techniques span from those aimed at measuring token-to-token interactions (known as *context mixing*, Brunner et al., 2020; Kobayashi et al., 2020, 2021; Ferrando et al., 2022b; Mohebbi et al., 2023b,a), to others striving to reverse engineer the model decision and decompose it into understandable pieces (known as *mechanistic interpretability*, Wang et al., 2023; Elhage et al., 2021).

This tutorial focuses on Transformer-specific interpretability methods. We will first briefly review the internal structure of the Transformer architecture to establish our notations. Next, we will explain why it is necessary to design methods tailored to the model architecture, exposing the limitations of model-agnostic approaches when applied to Transformer analysis using practical examples. Subsequently, we will introduce Transformer-specific techniques, delving into their mathematics, and categorizing them according to their purposes, using experimental results across a number of domains, such as text, speech, and music, as well as across several languages. Our tutorial will conclude with a discussion on current limitations in interpretability and promising future directions.

Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics
Tutorial Abstracts, pages 21–26

2 Tutorial Type

The tutorial will be cutting-edge, covering the latest research advancements in the interpretability of Transformers, which serve as the backbone architecture of modern NLP systems.

The only ACL tutorials similar to ours are "Interpretability and Analysis in Neural NLP" (Belinkov et al., 2020) and "Fine-grained Interpretation and Causation Analysis in Deep NLP Models" (Sajjad et al., 2021), held at ACL 2020 and NAACL 2021, respectively. Both focused on general model-agnostic interpretability techniques. Our tutorial, however, will question the effectiveness of those general-purpose analysis methods and mark the next chapter: a transition from model-agnostic approaches to Transformer-specific methods.

3 Target Audience

Given the widespread use of Transformers across various applications in both text and speech, we expect that our audience will be not only folks engaged in interpretability but also those from various tracks within the Computational Linguistics community who have not kept up with the recent advancements within interpretability research. In fact, we have been frequently asked at *ACL conferences and our industry meetings, particularly by individuals outside of the interpretability track, seeking guidance on the most effective interpretability techniques to employ in their projects for non-interpretability purposes, such as training monitoring, model compression, or model tuning.

In terms of expected prerequisite background, we expect audience members to be familiar with the basic concepts of Transformer models. For the Jupyter notebooks that will be covered, we expect experience with PyTorch and the `Transformers` library.

4 Outline of Tutorial Structure

The tutorial will consist of 30 minute slots of lectures and interactive seminars for which we will provide Jupyter notebooks. A small part of the tutorial will be focused on interpretability techniques from the organisers (e.g. Abnar and Zuidema, 2020 and Mohebbi et al., 2023b), but the majority of the work discussed will be work from other labs to provide an honest and broad overview of the current state of interpretability research in NLP.

1. 30 minute lecture on **model-agnostic interpretability**:
   - Introduction
   - Model-agnostic approaches: probing, feature attributions, behavioral studies
   - How are model-agnostic approaches adapted to Transformers? What are their limitations?

2. 30 minute lecture on interpretation of **attention** and **context mixing**:
   - Attention analysis (Clark et al., 2019) as a straightforward starting point for measuring context mixing.
   - Limitations of interpreting raw attention scores (Bibal et al., 2022; Hassid et al., 2022)
   - Effective attention scores: rollout (Abnar and Zuidema, 2020), HTA (Brunner et al., 2020), LRP-based attention (Chefer et al., 2020).
   - Expanding the scope of context mixing analysis by incorporating other model components: Attention-Norm (Kobayashi et al., 2020, 2021, 2023), GlobEnc (Modarressi et al., 2022), ALTI (Ferrando et al., 2022b,a), Value Zeroing (Mohebbi et al., 2023b), DecompX (Modarressi et al., 2023).

3. 30 minute interactive tutorial on interpreting context mixing: Jupyter notebooks will be provided (via Google Colab) and can be run interactively while the presenters go through it.

4. Coffee break

5. 30 minute lecture on **mechanistic** and **causality-based** interpretability:
   - Basics of mechanistic interpretability: the residual stream and computational graph views of models, and the circuits framework (Olah et al., 2020; Elhage et al., 2021; Hanna et al., 2023).
   - Finding circuit structure using causal interventions (Vig et al., 2020; Geiger et al., 2021; Wang et al., 2023; Goldowsky-Dill et al., 2023; Conmy et al., 2023; Nanda, 2023; Syed et al., 2023).

- Assigning semantics to circuit components: the logit lens (Nostalgebrist, 2020; Geva et al., 2021), concept erasure (Belrose et al., 2023), and (potentially) polysemanticity and superposition (Elhage et al., 2022).

6. 30 minute interactive tutorial mechanistic interpretability in NLP, notebooks will again be provided.

7. 30 minute slot for discussion, reflection and future outlook: what are open questions in interpretability, what's next, and what's lacking?

5 Reading List

In addition to the key papers mentioned in Section 4, we would recommend attendees that are interested in gaining a broader understanding of general interpretability techniques to explore the following survey papers: (Belinkov and Glass, 2019; Madsen et al., 2021; Raukur et al., 2022; Lyu et al., 2022)

6 Special Requirements

There are no special technical requirements, other than standard conference equipment (computer, screen, and projector). If participants wish to participate in the interactive parts, they should bring their laptops.

7 Diversity

Our tutorial focuses on Transformer-specific interpretability across several domains, including text, speech, music, (and vision, to some extent). As Transformers have gained widespread adoption within the CL community, we anticipate engaging a diverse and extensive audience. To ensure diversity, we have both professors and PhD students on our instructor team.

8 Tutorial Instructors

Hosein Mohebbi is a PhD candidate at Tilburg University. He is part of the InDeep consortium project, doing research on the interpretability of deep neural models for both text and speech. During his Master's, his research revolved around the interpretation of pre-trained language models and the utilization of interpretability techniques to accelerate their inference time. His research has been published in leading NLP venues such as ACL, EACL, EMNLP, and BlackboxNLP, where he also regularly serves as a reviewer. He is also one of the organizers of BlackboxNLP 2023-2024, a workshop focusing on analyzing and interpreting neural networks for NLP.

Jaap Jumelet is a PhD candidate at the Institute for Logic, Language and Computation at the University of Amsterdam. His research focuses on gaining an understanding of how neural models are able to build up hierarchical representations of their input, by leveraging hypotheses from (psycho-)linguistics. His research has been published at leading NLP venues, including TACL, ACL, and CoNLL. He is a co-organiser for BlackboxNLP in 2023-2024. He has been involved in numerous courses in the AI Master of the University of Amsterdam, all with a focus on NLP and interpretability.

Michael Hanna is a PhD candidate at the University of Amsterdam, as part of the Institute for Logic, Language and Computation. His research focuses on understanding the abilities of pre-trained language models, and linking these behaviors to low-level mechanisms using causal methods. His work has been published in leading interpretability and NLP venues such as NeurIPS, EMNLP, and EACL. He previously designed and led a workshop on mechanistic interpretability as part of the University of Amsterdam's artificial intelligence masters program.

Afra Alishahi is an Associate Professor at the Department of Cognitive Science and Artificial Intelligence at Tilburg University, Netherlands. Her main research interests are developing computational models of human language, studying the emergence of linguistic structure in grounded models of language learning, and developing tools and techniques for analyzing linguistic representations in neural models of language. She has served as program chair for CoNLL and as AC and SAC for many recent CL conferences, and is one of the founders of the BlackboxNLP workshops. She has acted as ACL tutorial co-chair and taught tutorials at ACL and ESSLII; most recently she offered a tutorial on *Interpretability of linguistic knowledge in neural language models* as part of Lectures on Computational Linguistics in Pisa, Italy.

Willem Zuidema is Associate Professor of NLP, Explainable AI and Cognitive Modelling at the University of Amsterdam. He has published widely in NLP, AI and Cognitive Science venues, including TACL, JAIR, ACL, EMNLP and NeurIPS. Since 2016, many of his publications have focused on interpretability in AI. He has taught many undergraduate and graduate courses (including Interpretability and Explainability in AI in Amsterdams's MSc AI, 2022, 2023), and two courses at graduate summerschools (ESSLLI 2008, 2015). He leads a project on interpretability that involves 5 universities ('InDeep', 2021-2026). He has served on many program committees, including ACL, NAACL, EMNLP, BlackboxNLP, and helped organize workshops and conferences; in 2016, he was tutorial co-chair for ACL.

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Yonatan Belinkov, Sebastian Gehrmann, and Ellie Pavlick. 2020. Interpretability and analysis in neural NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 1–5, Online. Association for Computational Linguistics.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. 2023. Leace: Perfect linear concept erasure in closed form.

Adrien Bibal, Rémi Cardon, David Alfter, Rodrigo Wilkens, Xiaoou Wang, Thomas François, and Patrick Watrin. 2022. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3889–3900, Dublin, Ireland. Association for Computational Linguistics.

Gino Brunner, Yang Liu, Damian Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. 2020. On identifiability in transformers. In *International Conference on Learning Representations*.

Hila Chefer, Shir Gur, and Lior Wolf. 2020. Transformer interpretability beyond attention visualization. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

Javier Ferrando, Gerard I. Gállego, Belen Alastruey, Carlos Escolano, and Marta R. Costa-jussà. 2022a. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8756–8769, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-jussà. 2022b. Measuring the mixing of contextual information in the transformer. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8698–8714, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9574–9586.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching.

Michael Hanna, Ollie Liu, and Alexandre Variengien. 2023. How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith, and Roy Schwartz. 2022. How much does attention actually attend? questioning the importance of attention in pretrained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1403–1416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2020. Attention is not only a weight: Analyzing transformers with vector norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7057–7075, Online. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2021. Incorporating Residual and Normalization Layers into Analysis of Masked Language Models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4547–4568, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Goro Kobayashi, Tatsuki Kuribayashi, Sho Yokoi, and Kentaro Inui. 2023. Feed-forward blocks control contextualization in masked language models. *ArXiv*, abs/2302.00456.

Satyapriya Krishna, Tessa Han, Alex Gu, Javin Pombra, Shahin Jabbari, Steven Wu, and Himabindu Lakkaraju. 2022. The disagreement problem in explainable machine learning: A practitioner's perspective. *ArXiv*, abs/2202.01602.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2022. Towards faithful model explanation in nlp: A survey. *ArXiv*, abs/2209.11326.

Andreas Madsen, Siva Reddy, and A. P. Sarath Chandar. 2021. Post-hoc interpretability for neural nlp: A survey. *ACM Computing Surveys*, 55:1 – 42.

Ali Modarressi, Mohsen Fayyaz, Ehsan Aghazadeh, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2023. DecompX: Explaining transformers decisions by propagating token decomposition. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2649–2664, Toronto, Canada. Association for Computational Linguistics.

Ali Modarressi, Mohsen Fayyaz, Yadollah Yaghoobzadeh, and Mohammad Taher Pilehvar. 2022. GlobEnc: Quantifying global token attribution by incorporating the whole encoder layer in transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 258–271, Seattle, United States. Association for Computational Linguistics.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023a. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Hosein Mohebbi, Willem Zuidema, Grzegorz Chrupała, and Afra Alishahi. 2023b. Quantifying context mixing in transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3378–3400, Dubrovnik, Croatia. Association for Computational Linguistics.

Neel Nanda. 2023. Attribution patching: Activation patching at industrial scale.

Michael Neely, Stefan F. Schouten, Maurits J. R. Bleeker, and Ana Lucic. 2022. A song of (dis)agreement: Evaluating the evaluation of explainable artificial intelligence in natural language processing. In *HHAI*.

Nostalgebrist. 2020. interpreting GPT: the logit lens.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Danish Pruthi, Bhuwan Dhingra, Livio Baldini Soares, Michael Collins, Zachary Chase Lipton, Graham Neubig, and William W. Cohen. 2020. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375.

Tilman Raukur, An Chang Ho, Stephen Casper, and Dylan Hadfield-Menell. 2022. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks. *2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 464–483.

Hassan Sajjad, Narine Kokhlikyan, Fahim Dalvi, and Nadir Durrani. 2021. Fine-grained interpretation and causation analysis in deep NLP models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 5–10, Online. Association for Computational Linguistics.

Aaquib Syed, Can Rager, and Arthur Conmy. 2023. Attribution patching outperforms automated circuit discovery.

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

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart M. Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*.