Interactively Providing Explanations for Transformer Language Models

Felix Friedrich [a,b], Patrick Schramowski [a,b], Christopher Tauchmann [a,b] and Kristian Kersting [a,b]

[a] *Computer Science Department, TU Darmstadt, Germany*
[b] *Hessian Center for AI (hessian.AI)*
{lastname}@cs.tu-darmstadt.de

Transformer language models (LMs) are state of the art in a multitude of NLP tasks. Despite these successes, their opaqueness remains problematic, especially as the training data might be unfiltered and contain biases. As a result, ethical concerns about these models arise, which can have a substantial negative impact on society as they get increasingly integrated into our lives [1]. Therefore, it is not surprising that a growing body of work aims to provide interpretability and explainability to black-box LMs [2]: Recent evaluations of saliency or attribution methods [3,4] find that, while intriguing, different methods assign importance to different inputs for the same outputs, thus encouraging misinterpretation and reporting bias [5,6]. Moreover, these methods primarily focus on post-hoc explanations of (sometimes spurious) input-output correlations. Instead, we emphasize using (interactive) prototype networks directly incorporated into the model architecture and hence explain the reasoning behind the network's decisions.

Interactive Prototype Learning. In order to address the black-box character of current LMs, we here focus on providing case-based reasoning explanations [7] during the inference process (*cf.* Fig. 1). We enhance the basic transformer architecture with a prototype layer and propose *Prototypical-Transformer Explanation* (Proto-Trex) Networks. Proto-Trex provides an explanation as a prototypical example for a specific model prediction, which is similar to (training-)samples of the same label. This approach not only increases interpretability [8] but is ideally suited for user interaction.

We also propose an interactive learning setting, iProto-Trex, enhancing Proto-Trex with user-revision. In addition to simply revealing the network's reasoning by providing prototype explanations, our approach further enables users to revise the network's explanations according to their preferences. In this way, we use human capabilities to incorporate knowledge outside of the rigid range of purely data-driven approaches. To this end,

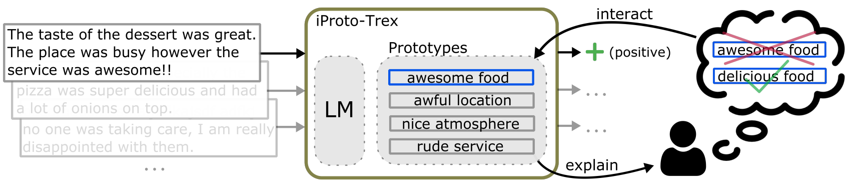

Figure 1. Interactive prototype learning: iProto-Trex classifies the input and gives the user an explanation based on a prototype. The user can directly, e.g., replace a given explanation with a self-chosen sequence.

Type of interaction	Acc.	Prototype/ Explanation
no interaction	93.64	Horrible customer service and service does not care about safety features. That's all I'm going to say. Oh they also don't care about their customers
soft replace (< 1)	93.79	I really don't recommend this place. The food is not good, service is bad. The entertainment is so cheesy. Not good
soft replace (1)	93.79	They offer a bad service.

Table 1. Showcasing interactive prototype learning: a user manipulates a model iteratively, i.e. softly replaces a prototype, with varying certainty. Thereby, he adapts it to his preferences without performance loss.

we integrate eXplanatory Interactive Learning (XIL) into prototype networks, which, in contrast to previous XIL methods [9], avoids tracing gradients and allows "Plug & Play", i.e. directly interacts on prototypes (*cf.* Fig 1). This combination is exciting and arguably necessary as explanation quality is normative, and no direct optimization is available. In order to address *suboptimal* explanations, user revision promotes *good* explanation quality wrt. individual notions of human subjects. To this end, we provide users with several methods, including the incorporation of strong- and weak-knowledge, as well as user certainty, to interact on the explanation. Users can directly replace *weak* prototypes, i.e. explanations, or steer the model to provide better ones, regarding their viewpoint. Interaction via explanations can be valuable already in the model building and understanding phase, avoiding Clever-Hans moments early on, increasing the explanation quality of the model and, in turn, user trust [13,14,15].

Language Model	Yelp	Movie
SBERT [10]	$94.92_{\pm 0.01}$	$84.56_{\pm 0.91}$
SBERT (Proto-Trex)	$93.59_{\pm 0.16}$	$80.05_{\pm 0.26}$
CLIP [11]	$93.78_{\pm 0.00}$	$75.49_{\pm 0.21}$
CLIP (Proto-Trex)	$87.16_{\pm 1.56}$	$63.52_{\pm 0.66}$
GPT-2 [12]	$93.78_{\pm 0.41}$	$87.05_{\pm 0.31}$
GPT-2 (Proto-Trex)	$95.32_{\pm 0.06}$	$84.57_{\pm 0.31}$
SBERT (iProto-Trex)	$93.81_{\pm 0.03}$	$80.24_{\pm 0.31}$
GPT-2 (iProto-Trex)	$95.25_{\pm 0.11}$	$84.80_{\pm 0.17}$

Table 2. Average accuracy of (i)Proto-Trex with different LMs compared to their baselines.

Results. As Tab. 2 shows and expected from literature [7], interpretability comes along with a trade-off in accuracy. Still, our first experimental results demonstrate that Proto-Trex networks perform on par with non-interpretable baseline LMs. More importantly, we showcase that users can interact with ease by simply manipulating the interpretable layer, i.e. a prototype (*cf.* Fig. 1). In Tab. 1, a user manipulates a prototypical explanation successively. While the accuracy remains unchanged, the user applies interactions with different certainty levels to give feedback and manipulate the model regarding his preferences. A certainty of < 1 yields a prototype close to the user preference, whereas 1 means the user's prototype is adopted. Interactive learning (Tab. 1) enables a loop between humans and AI, adapting the network according to user preferences of *good* prototypical explanations along with high accuracy. This loop can be repeated multiple times with different feedback methods, including different knowledge and certainty levels.

Conclusion. We introduce prototype networks for transformer LMs (Proto-Trex) to provide explanations. Importantly, to improve prototype explanations, we provide a novel interactive prototype learning setting (iProto-Trex) accounting for user knowledge and certainty[1]. An exciting future avenue is to equip prototype networks with a more flexible interaction policy, i.e. components beyond user certainty, to promote a greater human-AI communication towards what might be called cooperative AI [16].

Acknowledgment. This work benefited from the Hessian Ministry of Science and the Arts (HMWK) projects "The Third Wave of Artificial Intelligence - 3AI" and hessian.AI.

[1] full paper available at arxiv.org/abs/2110.02058 and code at github.com/felifri/XAITransformer

References

[1] Bender EM, Gebru T, McMillan-Major A, Shmitchell S. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In: Elish MC, Isaac W, Zemel RS, editors. Conference on Fairness, Accountability, and Transparency (FAccT); 2021. p. 610-23.

[2] Sokol K, Flach P. Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; 2020. p. 56–67.

[3] Ding S, Koehn P. Evaluating Saliency Methods for Neural Language Models. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL); 2021. p. 5034-52.

[4] Pezeshkpour P, Jain S, Wallace BC, Singh S. An Empirical Comparison of Instance Attribution Methods for NLP. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL); 2021. p. 967-75.

[5] Stammer W, Schramowski P, Kersting K. Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations. In: Proceedings of the 2021 Conference on Computer Vision and Pattern Recognition (CVPR); 2021. p. 3618-28.

[6] Gordon J, Van Durme B. Reporting Bias and Knowledge Acquisition. In: Proceedings of the 2013 Workshop on Automated Knowledge Base Construction (AKBC); 2013. p. 25-30.

[7] Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This Looks Like That: Deep Learning for Interpretable Image Recognition. In: Proceedings of the 2019 Conference on Advances in Neural Information Processing Systems (NeurIPS); 2019. p. 1-12.

[8] Ming Y, Xu P, Qu H, Ren L. Interpretable and Steerable Sequence Learning via Prototypes. In: Proceedings of the 25th International Conference on Knowledge Discovery and Data Mining (KDD); 2019. p. 903-13.

[9] Friedrich F, Stammer W, Schramowski P, Kersting K. A Typology to Explore and Guide Explanatory Interactive Machine Learning. arXiv preprint arXiv:220303668. 2022.

[10] Reimers N, Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP); 2019. p. 3982-92.

[11] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al. Learning Transferable Visual Models From Natural Language Supervision. In: Proceedings of the 38th International Conference on Machine Learning (ICML); 2021. p. 8748-63.

[12] Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language Models are Unsupervised Multitask Learners. arXiv preprint arXiv:201209699. 2019.

[13] Schramowski P, Stammer W, Teso S, Brugger A, Shao X, Luigs HG, et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence. 2020:476-86.

[14] Teso S, Kersting K. Explanatory interactive machine learning. In: Proceedings of the 2019 Conference on AI, Ethics, and Society (AIES); 2019. p. 239-45.

[15] Ribeiro MT, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In: Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining (KDD); 2016. p. 1135-44.

[16] Dafoe A, Bachrach Y, Hadfield G, Horvitz E, Larson K, Graepel T. Cooperative AI: machines must learn to find common ground. Nature. 2021:33-6.