Interactively Providing Explanations for Transformer Language Models

Felix Friedrich, Patrick Schramowski, Christopher Tauchmann and Kristian Kersting

Computer Science Department, TU Darmstadt, Germany
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.

To enhance Proto-Trex, we propose an interactive learning setting, iProto-Trex. 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, we inte-

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.

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].

Results. As Tab. 2 shows and expected from the literature, 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. 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


