eval-rationales: An End-to-End Toolkit to Explain and Evaluate Transformers-Based Models

Khalil Maachou, Jesús Lovón-Melgarejo, Jose G Moreno, Lynda Tamine

eval-rationales: An End-to-End Toolkit to Explain and Evaluate Transformers-Based Models

Khalil Maachou, Jesús Lovón-Melgarejo(✉), Jose G. Moreno, and Lynda Tamine

Université Paul Sabatier, IRIT, UMR 5505 CNRS, Toulouse, France
khalil.maachou@univ-tlse3.fr, {jesus.lovon,jose.moreno,tamine}@irit.fr

Abstract. State-of-the-art (SOTA) transformer-based models in the domains of Natural Language Processing (NLP) and Information Retrieval (IR) are often characterized by their opacity in terms of decision-making processes. This limitation has given rise to various techniques for enhancing model interpretability and the emergence of evaluation benchmarks aimed at designing more transparent models. These techniques are primarily focused on developing interpretable models with the explicit aim of shedding light on the rationales behind their predictions. Concurrently, evaluation benchmarks seek to assess the quality of these rationales provided by the models. Despite the availability of numerous resources for using these techniques and benchmarks independently, their seamless integration remains a non-trivial task. In response to this challenge, this work introduces an end-to-end toolkit that integrates the most common techniques and evaluation approaches for interpretability. Our toolkit offers user-friendly resources facilitating fast and robust evaluations.

Keywords: Interpretable · Evaluation · Transformers

1 Introduction

The rapid evolution of neural models in Natural Language Processing (NLP) and Information Retrieval (IR) has yielded exceptional performance on various benchmarks [11,18]. However, the high performance achieved often obfuscates the inner workings of these models, reducing their interpretability [12]. Consequently, the field of eXplainable Artificial Intelligence (XAI) has garnered increased attention within the AI community, aiming to enhance approaches for elucidating the decision-making processes of these models [3,5].

Particularly, in the context of transformers-based models for NLP [10,14] and IR [2,8], different efforts have emerged towards explaining model predictions at different granularity levels. Most common approaches focus on *local* explanations, which seek to clarify the reasons behind a single input's prediction, while less common approaches for *global* explanations aim to explain predictions for general input. Moreover, while there are different approaches to understanding

N. Goharian et al. (Eds.): ECIR 2024, LNCS 14612, pp. 212–217, 2024.
https://doi.org/10.1007/978-3-031-56069-9_20

what makes a model interpretable, we specifically focus on the use of *local rationales*, following recent work [4,6,9]. These *rationales* are subsets of the input elements (words, phrases, or sentences) that explain a prediction [17].

The rationale-based XAI techniques are fundamentally focused on elucidating the "why" behind a model's specific output. A step forward in this research line seeks to evaluate the "quality" of the explanations found by these methods in order to compare them. However, as a novel field, there is little agreement on how this evaluation should be performed; consequently, comprehensive benchmarks such as ERASER have emerged [4].

Various programming frameworks and libraries have been developed to facilitate the implementation of these XAI techniques, including well-known tools such as LIME [13], Captum[1], ELI5[2], Shap [15], and others. Similarly, evaluation resources, like the ERASER benchmark, are readily accessible online. However, integrating the evaluation of XAI techniques with predefined benchmarks is a challenging task. For a regular user, the integration of both frameworks can be a limiting factor and a reason to avoid adopting benchmark evaluation as a good practice to evaluate their models. Moreover, as the number of freely available pre-trained models and datasets increases, using hubs such as HuggingFace has become an important source to explore and enhance new approaches. However, the available tools are limited to a fixed list of models and datasets, and integrating HuggingFace resources is not straightforward.

To address these challenges, we present *eval-rationales*, an end-to-end toolkit that integrates *local* XAI techniques from different libraries with the ERASER evaluation benchmark. Our toolkit can also integrate transformers-based models and datasets from the HuggingFace hub, thus facilitating the integration of state-of-the-art models for evaluation and exploration. Our toolkit integrates the main functions of the Captum, LIME, and ERASER libraries, and we abstract all the integration processes of the mentioned frameworks to compute the rationales prediction and evaluation. The user is only required to input a Transformers model, a Dataset, and an XAI technique. As a final output, the user obtains detailed metrics for their inputs, reflecting the quality of their model. Furthermore, we have empowered this toolkit with classical functions such as highlighted visualization for each input.

2 The *eval-rationales* Toolkit

The presented end-to-end toolkit provides a robust set of rationale-based metrics, enabling researchers and practitioners to evaluate the quality of explanations generated by transformers-based models on evaluation datasets. Specifically, our toolkit consists of two interconnected modules: i) the prediction rationales module, responsible for predicting rationales from model inputs using various XAI techniques, and ii) the metrics module, which computes the quality metrics based on these predicted rationales (see Fig. 1).

[1] https://captum.ai/.
[2] https://eli5.readthedocs.io/en/latest/.

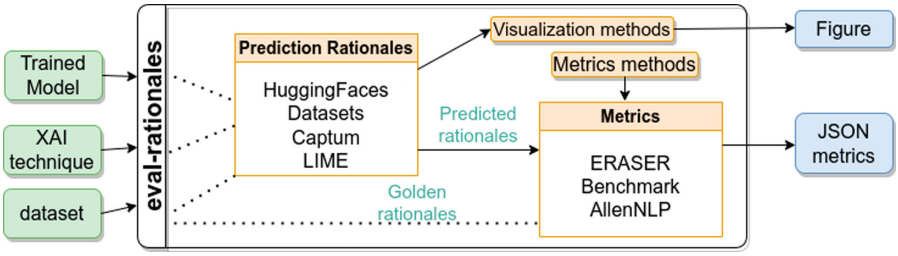


Fig. 1. The eval-explanation toolkit consists of two modules: prediction rationales, and metrics. The output is a file json file containing the computed metrics.

The prediction rationales module of our toolkit takes as input three elements: a model, an evaluation dataset, and an XAI technique. Our contribution is tailored to assess models primarily oriented toward sequence classification tasks, specifically focusing on binary classification models and datasets. To foster integration with the latest developments in NLP and IR, we have adapted our toolkit to support models based on the HuggingFace library, which can be seamlessly imported from the online repository or utilized from local storage. Regarding the XAI techniques, our toolkit includes attention-based methods [1,16], LIME [13], and gradient-based [7] approaches, mainly based on the Captum library. Additionally, we have included a Random method, which serves as a baseline, following the methodology proposed in previous work [4].

The metrics module leverages rationale-based metrics following the ERASER benchmark. Specifically, we compute two essential metrics: comprehensiveness and sufficiency. Comprehensiveness evaluates whether the predicted rationales include all the features necessary to make a prediction. Sufficiency assesses whether the extracted rationales contain enough signal to make an informed decision. Additionally, we provide the Area Over the Perturbation Curve (AOPC), originally proposed by ERASER, to further gauge the quality of explanations.

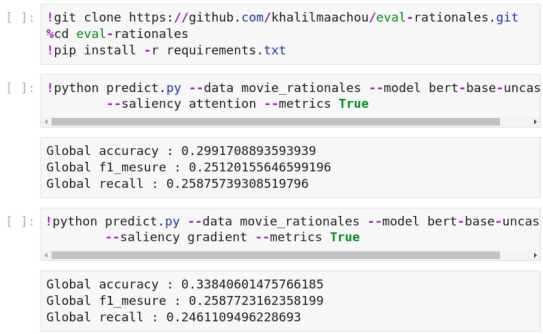
Furthermore, when the evaluation dataset includes golden rationales, we compute metrics at a token-level for accuracy, recall, and F1 score. These metrics are stored in a JSON file for easy access and manipulation by users.

In the following section, we demonstrate the utility and versatility of our evaluation toolkit through a case study involving two different datasets.

3 Case Study

Let us consider a hypothetical scenario where a user implements and trains a transformer-based model for enhanced interpretability using the Hugging Face library. This user's main objective is to assess the quality of the model's rationales predictions evaluated against the ERASER benchmark with Movie Rationales[3] and MIMIC IV[4] datasets. This evaluation process involves implementing

[3] https://huggingface.co/datasets/movie_rationales.
[4] https://physionet.org/content/mimiciv/0.4/.



```
[ ]: !git clone https://github.com/khalilmaachou/eval-rationales.git
     %cd eval-rationales
     !pip install -r requirements.txt

[ ]: !python predict.py --data movie_rationales --model bert-base-uncas
            --saliency attention --metrics True

     Global accuracy : 0.29917088935939
     Global f1_mesure : 0.25120155646599196
     Global recall : 0.25875739308519796

[ ]: !python predict.py --data movie_rationales --model bert-base-uncas
            --saliency gradient --metrics True

     Global accuracy : 0.33840601475766185
     Global f1_mesure : 0.2587723162358199
     Global recall : 0.2461109496228693
```

(a)

Model	XAI Technique	Comp.	Suff.
Dataset: *Movie Rationales*			
BERT + LSTM	Gradient	0.29	0.10
BERT + LSTM	Attention	0.26	0.13
BERT + LSTM	LIME	0.40	0.04
Dataset: *MIMIC IV*			
Bio-Clinical BERT	Gradient	0.26	0.10
Bio-Clinical BERT	Attention	0.15	0.01
Bio-Clinical BERT	LIME	0.14	-0.03

(b)

Fig. 2. (a) eval-rationales running on a fresh JupyterLab noteebok. This code includes installation and running two XAI techniques for a given model. (b) Sample results for Movie Rationales and MIMIC IV datasets.

Captum library code for rationale prediction and adapting the ERASER evaluation code, typically requiring significant coding effort.

In contrast, our toolkit provides a concise one-line command that seamlessly integrates the user's trained model into the end-to-end toolkit. By reducing the coding workload associated with rationale generation and evaluation, our toolkit empowers researchers to focus on refining the core model. It is developed on Python and tailored to the prevailing framework, Hugging Face, preferred by NLP and IR practitioners. Once installed, one can simply run the evaluation with a single-line command or use it as a library in a given script. Figure 2 presents an example of installation and usage of the toolkit, and a sample of our experiment results for this case study. The *eval-rationales* toolkit is freely available on GitHub[5], where detailed instructions are given and a video demo[6].

4 Conclusion and Future Work

This paper introduces an end-to-end toolkit to simplify the evaluation of predicted rationales generated by interpretable models. Our toolkit offers seamless integration with HuggingFace models, Datasets, and the ERASER benchmark, mitigating the challenges of integrating these distinct frameworks. This tool can benefit researchers and practitioners, significantly simplifying the model evaluation process. We plan to broaden the toolkit's adaptability with more XAI techniques tailored to textual-based transformer models. We also aim to integrate innovations and resources from the interpretability domain into ERASER benchmark and evaluation metrics.

[5] https://github.com/khalilmaachou/eval-rationales.
[6] https://youtu.be/3M1MJPhmMQE.

Acknowledgments. The authors would thank the support of the In-Utero project funded by HDH (France) and FRQS (Canada).

References

1. Bahdanau, D., Cho, K.H., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015 (2015)
2. Bhattarai, B., Granmo, O.C., Jiao, L.: An interpretable knowledge representation framework for natural language processing with cross-domain application. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13980, pp. 167–181. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_11
3. Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B., Sen, P.: A survey of the state of explainable AI for natural language processing. In: Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pp. 447–459 (2020)
4. DeYoung, J., et al.: ERASER: a benchmark to evaluate rationalized NLP models. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 4443–4458. Association for Computational Linguistics (2020). https://doi.org/10.18653/v1/2020.acl-main.408
5. Dwivedi, R., et al.: Explainable AI (XAI): core ideas, techniques, and solutions. ACM Comput. Surv. **55**(9), 1–33 (2023)
6. Hayati, S.A., Kang, D., Ungar, L.: Does BERT learn as humans perceive? Understanding linguistic styles through lexica. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 6323–6331. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (2021). https://doi.org/10.18653/v1/2021.emnlp-main.510
7. Karlekar, S., Niu, T., Bansal, M.: Detecting linguistic characteristics of Alzheimer' dementia by interpreting neural models. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers), pp. 701–707 (2018)
8. Lyu, L., Anand, A.: Listwise explanations for ranking models using multiple explainers. In: Kamps, J., et al. (eds.) ECIR 2023. LNCS, vol. 13980, pp. 653–668. Springer, Cham (2023). https://doi.org/10.1007/978-3-031-28244-7_41
9. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: Hatexplain: a benchmark dataset for explainable hate speech detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35, no. 17, pp. 14867–14875 (2021). https://doi.org/10.1609/aaai.v35i17.17745
10. Narang, S., Raffel, C., Lee, K., Roberts, A., Fiedel, N., Malkan, K.: Wt5?! training text-to-text models to explain their predictions. arXiv preprint arXiv:2004.14546 (2020)
11. Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., Huang, X.: Pre-trained models for natural language processing: a survey. SCIENCE CHINA Technol. Sci. **63**(10), 1872–1897 (2020)
12. Ras, G., Xie, N., Van Gerven, M., Doran, D.: Explainable deep learning: a field guide for the uninitiated. J. Artif. Intell. Res. **73**, 329–396 (2022)

13. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why should I trust you?": explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016, pp. 1135–1144 (2016)
14. Ross, A., Marasović, A., Peters, M.E.: Explaining NLP models via minimal contrastive editing (MICE). In: Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 3840–3852 (2021)
15. Sundararajan, M., Najmi, A.: The many shapley values for model explanation. In: Proceedings of the 37th International Conference on Machine Learning, ICML 2020. JMLR.org (2020)
16. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
17. Wiegreffe, S., Marasovic, A.: Teach me to explain: a review of datasets for explainable natural language processing. In: Thirty-Fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1) (2021)
18. Yates, A., Nogueira, R., Lin, J.: Pretrained transformers for text ranking: BERT and beyond. In: Kondrak, G., Bontcheva, K., Gillick, D. (eds.) Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials, pp. 1–4. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-tutorials.1