The rising popularity of deep learning has brought to light a fundamental limitation of neural network architectures: they lack the ability to provide interpretable justifications for their decisions, making them unsuitable for contexts where human experts require transparent explanations [1]. This abstract summarizes a newly introduced comprehensive approach to Explainable Artificial Intelligence (XAI), which demonstrates how a deliberate design of neural networks produces a family of interpretable deep learning models known as Logic Explained Networks (LEN) [2]. LENs only necessitate human-understandable predicates as input concepts and offer logic explanations of the output predictions via a set of First-Order Logic (FOL) formulas built on these predicates (see an example in Figure 1). A very interesting feature of this model is its versatility, indeed LENs can be applied in many use cases, including as interpretable classifiers or to explain another black-box model. In case of interpretable classification, some design choices, like learning criterion and parsimony index, allows to achieve state-of-the-art results in the prediction accuracy while gaining transparency on the model’s decision process [3]. Concerning the learning paradigms, LENs can be successfully trained to learn and provide explanations both in supervised and unsupervised learning settings [2, 4].

Experimental Analysis Experimental findings on several datasets and tasks demonstrate that LENs can yield superior classifications compared to established white-box models such as decision trees and Bayesian rule lists[5], while providing more succinct and meaningful explanations. For instance, LENs have been applied to classification problems ranging from computer vision to medicine, such as (MIMIC-II) [6] and (CUB) [7], and recently also to NLP tasks [8], always with the aim of solving the classification task, while also providing FOL explanations of the underlying decision process. In [3] six quantitative metrics are defined and used to compare the proposed approach with other state-of-the-art methods. In addition, in order to make LENs accessible to the whole community, we released the library PyTorch,
Explain! as a Python package on PyPI: https://pypi.org/project/torch-explain/ with an extensive documentation that is available on read at https://pytorch-explain.readthedocs.io/en/latest/

Figure 1: Example of a possible instance of a LEN on the CUB 200-2011 fine-grained classification dataset. Here, a LEN is placed on top of a convolutional neural network $g(\cdot)$ in order to (i) classify the species of the bird in input and (ii) provide an explanation on why it belongs to this class.

Acknowledgments

This work was supported by TAILOR and by HumanE-AI-Net, projects funded by EU Horizon 2020 research and innovation programme under GA No 952215 and No 952026, respectively.

References

[1] A. Chander, R. Srinivasan, S. Chelian, J. Wang, K. Uchino, Working with beliefs: Ai transparency in the enterprise., in: IUI Workshops, volume 1, 2018.

[2] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Lió, M. Maggini, S. Melacci, Logic explained networks, Artificial Intelligence 314 (2023) 103822.

[3] P. Barbiero, G. Ciravegna, F. Giannini, P. Lió, M. Gori, S. Melacci, Entropy-based logic explanations of neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 36, 2022, pp. 6046–6054.

[4] G. Ciravegna, F. Giannini, S. Melacci, M. Maggini, M. Gori, A constraint-based approach to learning and explanation, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 3658–3665.

[5] H. Yang, C. Rudin, M. Seltzer, Scalable bayesian rule lists, in: International conference on machine learning, PMLR, 2017, pp. 3921–3930.

[6] M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, R. G. Mark, Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database, Critical care medicine 39 (2011) 952.

[7] C. Wah, S. Branson, P. Welinder, P. Perona, S. Belongie, The caltech-ucsd birds-200-2011 dataset (2011).

[8] R. Jain, G. Ciravegna, P. Barbiero, F. Giannini, D. Buffelli, P. Lio, Extending logic explained networks to text classification, in: Empirical Methods in Natural Language Processing, 2022.