Logic Explanations of Neural Networks

Pietro Barbiero[1], **Gabriele Ciravegna**[2], **Francesco Giannini**[3], **Marco Gori**[2,3], **Pietro Liò**[1], **Marco Maggini**[3], **Stefano Melacci**[3]
[1] University of Cambridge, Cambridge, UK
[2] Université Côte d'Azur, Nice, FR
[3] Universiy of Siena, Siena, IT
pb737@cam.ac.uk, gabriele.ciravegna@unifi.it, francesco.giannini@unisi.it, marco.gori@unisi.it, pl219@cam.ac.uk, marco.maggini@unisi.it, mela@diism.unisi.it

Abstract

Neural networks achieve state-of-the-art performances in several tasks, but their applicability in safety-critical domains has been slowed down by the lack of human-understandable motivations for their decisions. To provide high-performance neural models whose predictions could be used in decision support, we introduced Logic Explained Networks (LEN). LENs allow the extraction of First-Order Logic explanations, while still achieving results close to black-box competitors. The framework allows both the definition of interpretable-by-design models and the explanation of existing classifiers. We experimentally evaluated LENs on several tasks, from clinical data to computer vision. A Python module has been released implementing some state-of-the-art LENs.

1 Research Team

The University of Siena is a CINI unit in the Laboratory of Artificial Intelligence and Intelligent Systems (AIIS). In particular, the Siena Artificial Intelligence Laboratory (SAILAB)[1] consists in a considerable group of professors, researchers and PhD students working on different fields related to machine learning and its applications, such as computer vision, natural language processing, neural-symbolic approaches, and so forth. One of the main topics, subject of many investigations over time, is the integration of learning and reasoning techniques in order to either inject prior knowledge into a neural model or to learn relational knowledge directly from data. The study of neural-symbolic approaches combining both neural architectures and probabilistic logic reasoning is a fundamental problem in AI, and there has been several joint work over the years between SAILAB members and external partners, such as other universities or companies. In the following, we summarize the most relevant people in SAILAB and from other academic institutions that are currently involved in the study of neural-symbolic approaches.

[1] https://sailab.diism.unisi.it/

University of Siena

Marco Gori: Full Professor of Computer Engineering at the University of Siena and 3IA chair at Université Côte d'Azur.
Marco Maggini: Full Professor in Computer Engineering at the University of Siena.
Stefano Melacci: Associate Professor of Computer Engineering at the University of Siena.
Michelangelo Diligenti: Assistant Professor in Computer Engineering at the University of Siena and Google Inc. collaborator.
Francesco Giannini: Post-Doc Researcher at the Consorzio Interuniversitario Nazionale per l'Informatica (CINI).

Other Academic Institutions

Pietro Liò: Full Professor at the University of Cambridge.
Giuseppe Marra: Post-Doc Researcher at KU Leuven.
Gabriele Ciravegna: Post-Doc Researcher at Université Côte d'Azur.
Pietro Barbiero: PhD Student at the University of Cambridge.

2 Research Interests

Neural-symbolic models aim at combining the benefits of both sub-symbolic approaches, like deep neural networks, and symbolic approaches, like probabilistic logic reasoners, in a unified framework [De Raedt *et al.*, 2020; Diligenti *et al.*, 2021]. Neural architectures are very good at learning complex representations of data allowing to make accurate predictions, but in general struggle in capturing relational knowledge about a certain task and are mostly considered as black-boxes. On the other hand, symbolic approaches provide more human-understandability and can easily express and exploit relational knowledge, e.g. by performing probabilistic logic reasoning. However, in general these approaches can be difficultly applied to real-world datasets, involving large amounts of numerical data. In this regards, SAILAB researchers have been focusing on the definition of neural-symbolic models that might be used both to inject relational knowledge into high-performance neural models [Marra *et al.*, 2019; Marra *et al.*, 2020; Marra *et al.*, 2021] and to improve the understanding of the decision process implemented by black-boxes [Ciravegna *et al.*, 2020a; Ciravegna *et al.*, 2020b].

Logic Explained Networks. The research activities on Logic Explained Networks (LEN) [Ciravegna *et al.*, 2021], concerning Explainable Artificial Intelligence (XAI) techniques to logically explain neural networks, are very relevant for the topics of the Ital-IA workshop *AI Responsabile ed Affidabile*. For instance, the adoption of interpretable-by-design neural models, achieving competing results with respect to state-of-the-art black-boxes, would increase the trustworthiness in the AI model, while still preserving high performances in solving a certain task [Barbiero *et al.*, 2021].

Most techniques explaining black boxes focus on finding or ranking the most relevant features used by the model to make predictions [Simonyan *et al.*, 2013; Ribeiro *et al.*, 2016; Selvaraju *et al.*, 2017]. Such feature-scoring methods are very efficient and widely used, but they cannot explain how neural networks compose such features to make predictions [Kindermans *et al.*, 2019; Kim *et al.*, 2018]. By providing human-readable logic explanations of (deep) neural networks as a set of First-Order Logic (FOL) formulas, LENs are also suitable in case of decision making tasks. Indeed, by means of FOL expressions that humans can easily understand, high-level relational knowledge about a set of categories can be naturally expressed.

2.1 Experimental Evaluation and Software

The introduced models have been experimentally evaluated on several tasks and different datasets with promising, and often state-of-the-art, results. Furthermore, the continual collaboration with some companies like Expert.ai and QuestIT allowed the experimental evaluation of the models also on real-world datasets.

Neural-symbolic models like Relational Reasoning Networks [Marra *et al.*, 2021] have been recently applied to knowledge graph embedding benchmarks, like the Countries dataset [Bouchard *et al.*, 2015] and the Nations, Kinship and UMLS datasets [Kok e Domingos, 2007].

Logic Explained Networks have been applied to classification problems ranging from computer vision to medicine, such as (MIMIC-II) [Saeed *et al.*, 2011], (V-Dem) [Coppedge *et al.*, 2021] and (CUB) [Wah *et al.*, 2011], with the aim of solving the classification task, while also providing FOL explanations of the underlying decision process. A visual sketch of some classification problems and a selection of the logic formulas found by the proposed approach is reported in Fig. 1. In [Barbiero *et al.*, 2021] 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/.

2.2 Projects

The described research activities have been further motivated and developed thanks to the participation of SAILAB members in different current and previous research projects regarding the integration of learning and reasoning systems and trustworthy AI.

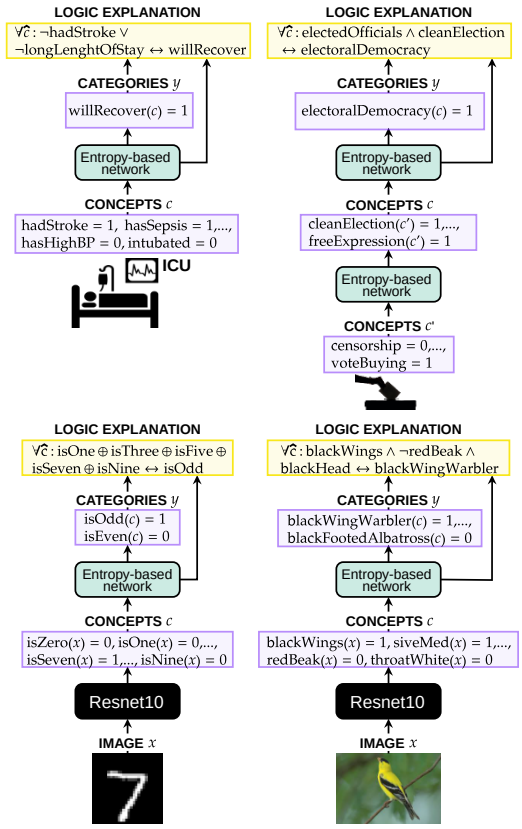

Figure 1: The four case studies show how the proposed Entropy-based networks (green) provide concise logic explanations (yellow) of their own decision process in different real-world contexts. When input features are non-interpretable, as pixel intensities, a "concept decoder" (ResNet10) maps images into concepts. Entropy-based networks then map concepts into target classes.

TAILOR: Foundations of Trustworthy AI - Integrating Reasoning, Learning and Optimization. TAILOR is an EU-funded ICT-48 Network (GA 952215) with the purpose of building the capacity of providing the scientific foundations for Trustworthy AI in Europe by developing a network of research excellence centres leveraging and combining learning, optimization and reasoning. The University of Siena is involved as a member of the CINI partner, and leader of the Task 4.3 on "Learning and Reasoning with Embeddings, Knowledge Graphs & Ontologies".

HumanE-AI-Net: HumanE AI Network. HumanE-AI-Net is an EU-funded ICT-48 Network (GA 761758) with the purpose of facilitating a European brand of trustworthy, ethical AI that enhances Human capabilities and empowers citizens and society to effectively deal with the challenges of an interconnected globalized world. The University of Siena is involved as a third part of the partner University of Pisa.

AI4EU: A European AI On Demand Platform and Ecosystem. AI4EU is an EU-funded ICT-26 Network (GA 825619) with the purpose of building a comprehensive European AI-on-demand platform to facilitate the exchange of

showcasing services, expertise, algorithms, software frameworks, development tools, modules, data, and so forth, for the European AI community. The University of Siena has been partner of this project, by participating especially in Task 7.4 at the development of AI assets for integrative learning.

3 Conclusions and Future Work

The adoption of neural-symbolic models can improve the human-understandability of the decision process underlying the predictions of neural architectures, while also providing a natural platform to integrate learning from raw data and symbolic reasoning. Nowadays, an important limitation of this kind of approaches is that their applicability on large relational domains can still be impractical when working on a fully grounded knowledge base. For this reason, one of the main challenges that we are currently working on is the definition of techniques for the automatic selection of a subset of relevant ground atoms to be considered within the learning process.

Logic Explained Networks allow the use of a neural model either to provide explanations for a black-box or to solve a classification problem in an interpretable manner. The explanations provided by LENs in general consists of FOL rules relating a set of final categories with respect to a set of intermediate human-understandable concepts, like in concept-bottleneck models [Yeh *et al.*, 2019; Koh *et al.*, 2020]. The choice of which and how to represent these concepts may drastically affect the quality of the extracted rules and the learning capabilities of the model. In some contexts, such as computer vision, the use of concept-based approaches may require additional annotations to get a consistent symbolic layer of concepts. Recent works on automatic concept extraction may alleviate the related costs, leading to more effective concept annotations [Ghorbani *et al.*, 2019; Kazhdan *et al.*, 2020]. In particular, we are working on developing techniques for automatic concept extraction from hidden layer nodes, e.g. by defining explainability methods to extract new knowledge from embedded representations.

From the experimental point of view, we are currently applying LENs to different domains, like NLP and graph data structures, with different neural models, like Recurrent Neural Networks and Graph Neural Networks.

To conclude, in our vision this framework would express its full potential by interacting with the external world and especially with humans in a dynamic way. In this case, logic formulas might be rewritten as sentences to allow for a more natural interaction with end users. Moreover, a dynamic interaction may call for an extended expressivity leading to higher-order or temporal logic explanations.

References

[Barbiero *et al.*, 2021] Pietro Barbiero, Gabriele Ciravegna, Francesco Giannini, Pietro Lió, Marco Gori, e Stefano Melacci. Entropy-based logic explanations of neural networks. *arXiv preprint arXiv:2106.06804*, 2021.

[Bouchard *et al.*, 2015] Guillaume Bouchard, S. Singh, e Theo Trouillon. On approximate reasoning capabilities of low-rank vector spaces. In *AAAI Spring Symposia*, 2015.

[Ciravegna *et al.*, 2020a] Gabriele Ciravegna, Francesco Giannini, Marco Gori, Marco Maggini, e Stefano Melacci. Human-driven fol explanations of deep learning. In *Twenty-Ninth International Joint Conference on Artificial Intelligence and Seventeenth Pacific Rim International Conference on Artificial Intelligence {IJCAI-PRICAI-20}*, pages 2234–2240, 2020.

[Ciravegna *et al.*, 2020b] Gabriele Ciravegna, Francesco Giannini, Stefano Melacci, Marco Maggini, e Marco Gori. A constraint-based approach to learning and explanation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3658–3665, 2020.

[Ciravegna *et al.*, 2021] Gabriele Ciravegna, Pietro Barbiero, Francesco Giannini, Marco Gori, Pietro Lió, Marco Maggini, e Stefano Melacci. Logic explained networks. *arXiv preprint arXiv:2108.05149*, 2021.

[Coppedge *et al.*, 2021] Michael Coppedge, John Gerring, Carl Henrik Knutsen, Staffan I Lindberg, Jan Teorell, David Altman, Michael Bernhard, Agnes Cornell, M Steven Fish, Lisa Gastaldi, et al. V-dem codebook v11, 2021.

[De Raedt *et al.*, 2020] Luc De Raedt, Sebastijan Dumančić, Robin Manhaeve, e Giuseppe Marra. From statistical relational to neuro-symbolic artificial intelligence. *arXiv preprint arXiv:2003.08316*, 2020.

[Diligenti *et al.*, 2021] Michelangelo Diligenti, Francesco Giannini, Marco Gori, Marco Maggini, e Giuseppe Marra. A constraint-based approach to learning and reasoning. In *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pages 192–213. IOS Press, 2021.

[Ghorbani *et al.*, 2019] Amirata Ghorbani, James Wexler, James Zou, e Been Kim. Towards automatic concept-based explanations. *arXiv preprint arXiv:1902.03129*, 2019.

[Kazhdan *et al.*, 2020] Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, e Adrian Weller. Now you see me (cme): Concept-based model extraction. *arXiv preprint arXiv:2010.13233*, 2020.

[Kim *et al.*, 2018] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International conference on machine learning*, pages 2668–2677. PMLR, 2018.

[Kindermans *et al.*, 2019] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, e Been Kim. The (un) reliability of saliency methods. In *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer, 2019.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, e Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

[Kok e Domingos, 2007] Stanley Kok e Pedro Domingos. Statistical predicate invention. In *Proceedings of the International Conference on Machine Learning*, 2007.

[Marra *et al.*, 2019] Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, e Marco Gori. Integrating learning and reasoning with deep logic models. *arXiv preprint arXiv:1901.04195*, 2019.

[Marra *et al.*, 2020] Giuseppe Marra, Michelangelo Diligenti, Francesco Giannini, Marco Gori, e Marco Maggini. Relational neural machines. In *ECAI 2020*, pages 1340–1347. IOS Press, 2020.

[Marra *et al.*, 2021] Giuseppe Marra, Michelangelo Diligenti, e Francesco Giannini. Learning representations for sub-symbolic reasoning. *arXiv preprint arXiv:2106.00393*, 2021.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, e Carlos Guestrin. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*, 2016.

[Saeed *et al.*, 2011] Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, e Roger G Mark. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952, 2011.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, e Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Simonyan *et al.*, 2013] Karen Simonyan, Andrea Vedaldi, e Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[Wah *et al.*, 2011] C. Wah, S. Branson, P. Welinder, P. Perona, e S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[Yeh *et al.*, 2019] Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, e Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. *arXiv preprint arXiv:1910.07969*, 2019.