Interpretable Neural-Symbolic Concept Reasoning*

Pietro Barbiero[1,2,*], Gabriele Ciravegna[3], Francesco Giannini[4], Mateo Espinosa Zarlenga[1], Lucie Charlotte Magister[1], Alberto Tonda[5,6], Pietro Lió[1], Frederic Precioso[3], Mateja Jamnik[1] and Giuseppe Marra[7]

[1]Department of Computer Science and Technology, University of Cambridge (UK)
[2]Faculty of Informatics, Università della Svizzera Italiana (Switzerland)
[3]Maasai, Inria, I3S, CNRS, Université Côte d'Azur, (France)
[4]Consorzio Interuniversitario Nazionale per l'Informatica, CINI, (Italy)
[5]UMR 518 MIA-PS, INRAE, Université Paris-Saclay, 91120, Palaiseau (France)
[6]Institut des Systèmes Complexes de Paris Île-de-France (ISC-PIF) - UAR 3611 CNRS, Paris (France)
[7]Department of Computer Science, KU Leuven, (Belgium)

Deep learning methods are highly accurate, yet their opaque decision process prevents them from earning full human trust [1]. Concept-based models [2, 3] aim to address this issue by learning tasks based on a set of human-understandable concepts. However, state-of-the-art concept-based models rely on high-dimensional concept embedding representations which lack a clear semantic meaning, thus questioning the interpretability of their decision process [4, 5]. To overcome this limitation, we propose the *Deep Concept Reasoner* (DCR) [6], the first interpretable concept-based model that builds upon concept embeddings. In DCR framework, the utilization of neural networks does not involve direct task predictions. Instead, neural networks are employed to construct syntactic rule structures through the utilization of concept embeddings. These concept embeddings serve as a foundation for the subsequent rule evaluation, enabling the DCR to derive its final prediction. Notably, the evaluation of rules occurs based on the truth values associated with the concepts themselves rather than their embeddings. Consequently, the DCR framework upholds clear semantic principles, facilitating the provision of fully interpretable decision outcomes. The overarching process is illustrated in the accompanying Figure 1-left. One of the key advantages of DCR lies in its differentiability, which enables its effective training as an independent module on concept databases.

Experimental Results. The carried out experimental analysis aims at addressing two main research questions, i.e. the generalization and interpretability of DCR in different settings. In particular, our experiments show that DCR: (i) improves up to +25% w.r.t. state-of-the-art interpretable concept-based models on challenging benchmarks (Figure 1-right) (ii) discovers meaningful logic rules matching known ground truths even in the absence of concept supervision

NeSy 2023, 17th International Workshop on Neural-Symbolic Learning and Reasoning, Certosa di Pontignano, Siena, Italy
*Corresponding author.
pb737@cam.ac.uk (P. Barbiero)
https://www.pietrobarbiero.eu/ (P. Barbiero)
0000-0003-3155-2564 (P. Barbiero)

during training, and (iii), facilitates the generation of counterfactual examples providing the learnt rules as guidance.

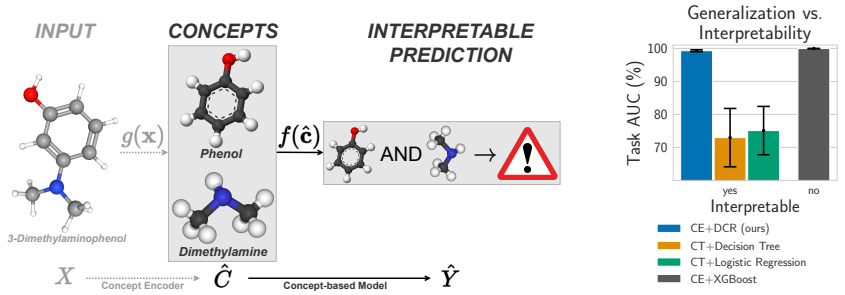

Figure 1: (left) An interpretable concept-based model f maps concepts $\hat{C}$ to tasks $\hat{Y}$ generating an interpretable rule. When input features are not semantically meaningful, a concept encoder g can map raw features to concepts. (right) DCR outperforms interpretable concept-based models. *CE* stands for concept embeddings and *CT* for concept truth values.

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] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature machine intelligence 1 (2019) 206–215.

[2] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, P. Liang, Concept bottleneck models, in: International Conference on Machine Learning, PMLR, 2020, pp. 5338–5348.

[3] Z. Chen, Y. Bei, C. Rudin, Concept whitening for interpretable image recognition, Nature Machine Intelligence 2 (2020) 772–782.

[4] D. Kazhdan, B. Dimanov, M. Jamnik, P. Liò, A. Weller, Now you see me (cme): concept-based model extraction, arXiv preprint arXiv:2010.13233 (2020).

[5] M. E. Zarlenga, P. Barbiero, G. Ciravegna, G. Marra, F. Giannini, M. Diligenti, Z. Shams, F. Precioso, S. Melacci, A. Weller, et al., Concept embedding models, arXiv preprint arXiv:2209.09056 (2022).

[6] P. Barbiero, G. Ciravegna, F. Giannini, M. E. Zarlenga, L. C. Magister, A. Tonda, P. Lio, F. Precioso, M. Jamnik, G. Marra, Interpretable neural-symbolic concept reasoning, arXiv preprint arXiv:2304.14068 (2023).