GlanceNets: Interpretable, Leak-proof Concept-based Models

Emanuele Marconato[1,2,*], Andrea Passerini[1] and Stefano Teso[3]

[1]DISI, University of Trento, Italy
[2]DII, University of Pisa, Italy
[3]CIMeC, University of Trento, Italy

Abstract

In this extended abstract, we briefly outline GlanceNets [1], a new class of deep learning classifiers that acquire high-level *concepts* from data and use them for both computing predictions and generating *ante-hoc* explanations of those predictions. In contrast with other concept-based networks, GlanceNets ensure the learned concepts, and the explanations built on them, are human interpretable, even in out-of-distribution scenarios. The core ideas at the heart of GlanceNets extend naturally to other Neuro-Symbolic architectures involving reasoning during inference.

Keywords

Concept-based Models, Concept Learning, Representation Learning, Explainable AI

There is growing interest in *trustworthy* deep learning models that attain high-performance and explainability by acquiring and reasoning with a vocabulary of high-level symbolic *concepts*. A key requirement is that the concepts can be *understood by human stakeholders*. Existing neural architectures tackle this desideratum using a variety of heuristics based on unclear notions of interpretability, failing to acquire concepts *with the intended semantics*. We address this by providing a clear definition of interpretability in terms of *alignment* between the model's representation and an underlying data generation process, and introduce GlanceNets, a new architecture that leverages ideas from deep generative modeling, causal representation learning and open-set recognition to *achieve alignment*, thus improving the interpretability of the learned concepts. In a longer version of the paper [1], we show that GlanceNets achieve better alignment than state-of-the-art approaches while preventing spurious concepts from unintentionally affecting its predictions in out-of-distribution scenarios. The code is available at: https://github.com/ema-marconato/glancenet.

References

[1] E. Marconato, A. Passerini, S. Teso, Glancenets: Interpretable, leak-proof concept-based models, in: Advances in Neural Information Processing Systems, 2022.

The 17th International Workshop on Neural-Symbolic Learning and Reasoning, July 3–5, 2023, Siena, Italy
*Corresponding author.
emanuele.marconato@unitn.it (E. Marconato); andrea.passerini@unitn.it (A. Passerini); stefano.teso@unitn.it (S. Teso)