P3 - Interpretable CEM

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

Project. The project consists of developing a variant of Deep Concept Reasoning (DCR)[1], an explainable-by-design concept-based model providing interpretable predictions over concept embeddings. The variant should maintain the generalization capability of a Concept Embedding Model (CEM)[3] while providing interpretable predictions.

Overview.

To overcome the limited representation capability of the Concept Bottleneck Model (CBM)[2], the Concept Embedding Model (CEM)[3] has been proposed in the literature, representing concepts as embeddings instead of single neurons. This allows for retaining the same generalization capability of a black-box end-to-end model without losing the capability to interact with the model. However, when employing a white-box task classifier on top of the concept representation, CBM is a globally interpretable model. On the contrary, even when using a white box classifier, CEM is non-interpretable since the single dimensions of the concept embeddings are not interpretable. To overcome this limit, the DCR[1] model has been proposed. It is an end-to-end trainable model providing interpretable predictions in terms of logic rules. More specifically, for each sample, DCR predicts a rule that holds for a given sample over the concept embeddings. This rule is then symbolically executed over the concept scores.

Goal.

The task of the project is to create a variant of DCR[1]. The variant should either improve one of the characteristics of standard DCR (e.g., locality of the interpretation, robustness, etc.), or it should change the interpretability paradigm, providing a different type of interpretable prediction (e.g., by means of a linear equation).

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic analysis and review of the existing supervised concept-based model, jointly training the concept and

task classifiers with a specific focus on CEM[3] and DCR[1].

- **Methodology.** Suggest either a variant of the DCR model (e.g., having global validity) or a different model providing a different type of interpretable prediction (e.g., by means of a linear equation).
- **Implementation.** The method should be implemented in a clear and documented way. The proposed model should be tested on a couple of datasets (both toy and real ones) together with DCR
- **Evaluation**. The proposed model should be compared with DCR under different aspects, e.g., its generalization performance but also its interpretability. To achieve this task, a number of metrics have to be identified and evaluated.

References

[1] Pietro Barbiero et al. "Interpretable neural-symbolic concept reasoning". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 1801–1825.

[2] Pang Wei Koh et al. "Concept bottleneck models". In: *International conference on machine learning*. PMLR. 2020, pp. 5338–5348.

[3] Mateo Espinosa Zarlenga et al. "Concept embedding models". In: *NeurIPS 2022-36th Conference on Neural Information Processing Systems*. 2022.