P2 — Benchmarking Concept-Based Explainable-by-Design Models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

Project. The project aims to compare and benchmark concept-based explainable-by-design models for assessing the performance and interpretability of these models.

Overview.

Concept-based explainable-by-design models offer insights into AI decision-making processes by explicitly representing concepts or abstractions within the model architecture.[6] This is generally done to enhance the interpretability of the model at the cost, sometimes, of its predictive capacity. Several models with different characteristics have been provided, employing both supervised[4, 8, 2], unsupervised[1, 3], and generative approaches[7, 5]. However, a comprehensive benchmarking effort is needed to evaluate and compare these models across various dimensions.

Goal.

The primary objective is to establish a benchmarking framework for assessing the effectiveness and reliability of concept-based explainable-by-design models. By systematically comparing their performance under different conditions and metrics, the project aims to identify strengths, weaknesses, and areas for improvement.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic analysis and review of concept-based explainable-by-design models. Particularly, provide a detailed review of the subset of selected methods.
- **Methodology.** Propose a benchmark to study the characteristics of the selected methods under well-defined conditions. Propose a set of metrics that are meaningful to compare the methods under different aspects. This

framework must include standardized datasets and experimental protocols to ensure consistency and reproducibility.

- **Implementation.** Selected models will be implemented or adapted within a unified software framework, emphasizing code quality, documentation, and usability. The same also holds for the defined metrics.

- **Evaluation**. The implemented models should be tested using the evaluation metrics at over at least one dataset to assess performance across interpretability. The results must be well-documented and discussed, reporting both quantitative and qualitative results to identify trends, strengths, and areas for improvement.

References

[1] David Alvarez Melis and Tommi Jaakkola. "Towards robust interpretability with self-explaining neural networks". In: *Advances in neural information processing systems*. Vol. 31. 2018.

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

[3] Chaofan Chen et al. "This looks like that: deep learning for interpretable image recognition". In: *Advances in neural information processing systems* 32 (2019).

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

[5] Tuomas Oikarinen et al. "Label-free concept bottleneck models". In: *arXiv preprint arXiv:2304.06129* (2023).

[6] Eleonora Poeta et al. "Concept-based explainable artificial intelligence: A survey". In: *arXiv preprint arXiv:2312.12936* (2023).

[7] Yue Yang et al. "Language in a bottle: Language model guided concept bottlenecks for interpretable image classification". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 19187–19197.

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