P1 — Benchmarking C-XAI post-hoc methods

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Gabriele Ciravegna, Eleonora Poeta

Project. The project consists in comparing and benchmarking post-hoc concept-based methods.

Overview.

Post-hoc concept-based methods have been proposed in the literature to provide explanations in terms of higher-level terms (i.e., concepts) without modifying the training paradigm of a model[3]. In particular, these methods, given a trained model, analyze how a set of concepts is represented in its latent space. The goal is to either compare these representations with those of the output classes or to understand which patterns have been learned by the hidden neurons. Several approaches have been proposed, either supervised [2] or unsupervised [1], and with different goals. However, there is a need for a comprehensive benchmark to assess the performance and reliability of these methods across different criteria.

Goal.

The primary goal is to establish a robust benchmarking framework for evaluating post-hoc concept-based XAI methods. By systematically comparing their performance under various conditions and metrics, we aim to identify strengths, weaknesses, and relative performance, providing valuable insights for researchers and practitioners.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic analysis and review of post-hoc concept-based methods. 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 methods will be implemented or adapted within a unified software framework, prioritizing code quality, documentation, and ease of use. The same also holds for the identified metrics.
- **Evaluation**. The selected methods should be tested using the identified metrics 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] Amirata Ghorbani et al. "Towards automatic concept-based explanations". In: *Advances in neural information processing systems* 32 (2019).

[2] Been Kim et al. *Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)*. 2018. arXiv: 1711.11279 [stat.ML].

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