P6 - Example-Based Explanations for Audio Classification

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Eleonora Poeta (+ Alkis Koudounas)

Project. This research proposes to investigate prototype-based/example-based explanation methods for audio classification to provide interpretable insights into the model decisions via listenable example-based explanations.

Overview.

Audio classification is the task of categorizing audio data into predefined classes. This task finds multiple applications, such as music genre classification and environmental sound monitoring. Recent solutions proposed prototype-based solutions to address the need for more interpretability of these models [7, 8, 9]. This project proposes to investigate prototype-based explanation methods for audio classification to provide interpretable insights into the model decisions via listenable example-based explanations.

Goal.

The task of the project is first to review existing explanation methods for audio classification, analyzing both by design and post-processing explainability. Then, the project aims to identify existing research gaps in the explainable AI literature for audio classification, focusing on prototype-based (or instance-based) explainability methods. Examples of research gaps include proposed example-based post-hoc methods and associating prototypes with specific class labels or groups of samples. Examples of analysis can be the investigation of techniques for generating prototypes that capture the distinct characteristics of each audio class. This process may involve clustering algorithms, such as K-means, to identify representative instances for each class (e.g., as [6] for time series). Other analyses can investigate the adoption of prototype networks proposed in the context of image classification [2, 3, 4, 5] in the context of audio classification for an interpretable-by design solution (e.g., [7, 8, 9, 1].

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing explanation methods for audio classification, both by design and post-hoc. Focus the analysis on prototype-based/explanation-based explanation approaches.
- **Identification of Research Gaps.** Identify key research gaps in the context of explainability methods for audio classification.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to address the identified research gap. This process may involve proposing a novel approach or adapting existing explainability methods to suit the context of audio classification.
- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach.

References

[1] Pablo Alonso-Jiménez et al. "Leveraging Pre-Trained Autoencoders for Interpretable Prototype Learning of Music Audio". In: *arXiv preprint arXiv:2402.09318* (2024).

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

[3] Chaofan Chen et al. "This looks like that: deep learning for interpretable image recognition". In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates Inc., 2019.

[4] Peter Hase et al. "Interpretable image recognition with hierarchical prototypes". In: *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*. Vol. 7. 2019, pp. 32–40.

[5] Oscar Li et al. "Deep learning for case-based reasoning through prototypes: a neural network that explains its predictions". In: AAAI'18/ IAAI'18/ EAAI'18. New Orleans, Louisiana, USA: AAAI Press, 2018. ISBN: 978-1-57735-800-8.

[6] Christoph Obermair et al. "Example or Prototype? Learning Concept-Based Explanations in Time-Series". In: *Proceedings of The 14th Asian Conference on Machine Learning*. Ed. by Emtiyaz Khan and Mehmet Gonen. Vol. 189. Proceedings of Machine Learning Research. PMLR, 2023, pp. 816–831. URL: `https://proceedings.mlr.press/v189/obermair23a.html`.

[7] Zhao Ren, Thanh Tam Nguyen, and Wolfgang Nejdl. "Prototype learning for interpretable respiratory sound analysis". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2022, pp. 9087–9091.

[8] Pablo Zinemanas et al. "An interpretable deep learning model for automatic sound classification". In: *Electronics* 10.7 (2021), p. 850.

[9] Pablo Zinemanas et al. "Toward interpretable polyphonic sound event detection with attention maps based on local prototypes". In: *Proceedings of the 6th Workshop on Detection and Classification of Acoustic Scenes and Events (DCASE 2021)*. Universitat Pompeu Fabra. Music Technology Group. 2021.