P8 - Explainable Multi-modal classification

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Salvatore Greco, Eliana Pastor

Project. This research aims to propose new methods or evaluations of XAI methods for multimodal classification models.

Overview.

Multimodal classification research has been gaining popularity, showing the benefits of combining data from multiple sources compared to traditional unimodal data. In recent years, many novel multimodal architectures have been proposed. However, these models are even more complex than the unimodal ones. Thus, their opacity is an even more challenging problem that poses a barrier to their applicability in real-world applications.

Recent advancements in XAI have introduced methods aimed at enhancing interpretability for multimodal models [4, 1, 3, 2]. Still, many gaps exist in this field. This project aims to develop XAI techniques tailored for multimodal models or propose new evaluation methods for XAI multimodal techniques. The focus is to increase the interpretability of the decision-making processes of these classifiers involving multiple modalities.

Goal.

Firstly, the project aims to review existing explanation methods or XAI evaluation for multimodal classifiers. Secondly, the project must identify existing research gaps in the XAI literature for multimodal classifiers. The research gap can be for both XAI techniques or the evaluation of XAI techniques. Examples of proposals can be the investigation of techniques for generating explanations of XAI classifiers (e.g., feature-based, gradient-based, counterfactual, plain-text explanations, etc.) suitable for classifiers involving multiple modalities. Another example can be the implementation of XAI techniques for individual modalities (e.g., LIME, SHAP) for models involving multiple modalities. A possible output of this project can also be a library to democratize the use of XAI methods for multimodal classifiers, or new evaluation techniques for XAI-multimodal methods.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing XAI methods or evaluation methods for explaining the predictions of multimodal classifiers.
- **Identification of Research Gaps.** Identify key research gaps for improving existing XAI techniques or evaluation of XAI techniques for multimodal classifiers.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to address the identified research gap. This may involve 1) proposing a novel XAI approach for multimodal classifiers, 2) improving an existing approach, 3) adapting an existing unimodal technique (e.g., SHAP) to the multimodal domain, 4) implementing a simple interface to apply existing XAI multimodal techniques with well-known libraries such as HuggingFace, or 5) propose XAI evaluations suitable for the multimodal domain.
- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach.

References

[1] Charles A. Ellis et al. "A Gradient-based Approach for Explaining Multimodal Deep Learning Classifiers". In: *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)*. 2021, pp. 1–6. DOI: `10.1109/BIBE52308.2021.9635460`.

[2] Pengbo Hu, Xingyu Li, and Yi Zhou. *SHAPE: An Unified Approach to Evaluate the Contribution and Cooperation of Individual Modalities.* 2022. arXiv: `2205.00302 [cs.LG]`.

[3] Gargi Joshi, Rahee Walambe, and Ketan Kotecha. "A Review on Explainability in Multimodal Deep Neural Nets". In: *IEEE Access* 9 (2021), pp. 59800–59821. DOI: `10.1109/ACCESS.2021.3070212`.

[4] Letitia Parcalabescu and Anette Frank. "MM-SHAP: A Performance-agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2023. DOI: `10.18653/v1/2023.acl-long.223`. URL: `http://dx.doi.org/10.18653/v1/2023.acl-long.223`.