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).

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.

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.

P4 - Explanation evaluation in text classification

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Salvatore Greco

Project. Evaluating the quality of explanations generated by explainability methods is critical for ensuring the reliability of explanations in real-world applications. This project focuses on enhancing the evaluation of explanation quality in the context of text classification.

Overview.

Evaluating the quality of explanations of the prediction of machine learning models is crucial for ensuring the reliability and trust of explanations. Various evaluation methods focused on assessing different aspects of explanation quality such as faithfulness, plausibility, robustness, and compactness [1, 3, 5, 6]. Considering this relevance, different libraries and tools have been proposed for evaluating explanations [2, 4]. Despite the efforts, many explainability libraries only cover a subset of these methods.

Goal.

The task of the project is first to systematically review existing evaluation methods to assess the quality of the explanation. Then, the project aims to improve the evaluation capabilities of a specific package, ferret [4], which focuses on explainability methods tailored for transformers. The package ferret includes only a few faithfulness measures and plausibility measures. The project involves studying other possible metrics suitable for the task of text classification. Once identified, the task of the project is to implement them and assess them.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing evaluation methods for explanation quality assessment in text classification.
- **Identification of Metrics.** Identify a set of metrics suitable for evaluating explanations for text classifiers not yet included in the *ferret* package.

- **Implementation.** Select and implement 2-3 promising metrics within the *ferret* package.
- **Evaluation.** Assess the effectiveness and applicability of the newly implemented metrics by comparing them across different attribution-based explanation methods.

References

[1] Julius Adebayo et al. "Sanity checks for saliency maps". In: vol. 31. 2018.

[2] Chirag Agarwal et al. "OpenXAI: Towards a Transparent Evaluation of Model Explanations". In: *Advances in Neural Information Processing Systems*. Ed. by S. Koyejo et al. Vol. 35. Curran Associates, Inc., 2022, pp. 15784–15799. URL: `https://proceedings.neurips.cc/paper_files/paper/2022/file/65398a0eba88c9b4a1c38ae405b125ef-Paper-Datasets_and_Benchmarks.pdf`.

[3] Pepa Atanasova et al. "A Diagnostic Study of Explainability Techniques for Text Classification". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 3256–3274. DOI: `10.18653/v1/2020.emnlp-main.263`. URL: `https://aclanthology.org/2020.emnlp-main.263`.

[4] Giuseppe Attanasio et al. "ferret: a Framework for Benchmarking Explainers on Transformers". In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*. Ed. by Danilo Croce and Luca Soldaini. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 256–266. DOI: `10.18653/v1/2023.eacl-demo.29`. URL: `https://aclanthology.org/2023.eacl-demo.29`.

[5] Jay DeYoung et al. "ERASER: A Benchmark to Evaluate Rationalized NLP Models". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458. DOI: `10.18653/v1/2020.acl-main.408`. URL: `https://aclanthology.org/2020.acl-main.408`.

[6] Yang Liu et al. "Synthetic Benchmarks for Scientific Research in Explainable Machine Learning". In: *Advances in Neural Information Processing Systems - Datasets and Benchmarks*. 2021.

P5 - Attribution-based explainability methods for speech classification models

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

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

Project. Given the increasing adoption of speech classification models, recent works in explainable AI address the issue of their lack of interpretability. This project focuses on identifying research gaps in feature attribution methods for speech classification models and proposes novel solutions to improve speech model interpretability.

Overview.

Speech classification models are widely adopted for various tasks, such as intent classification, keyword spotting, and emotion recognition. Speech models are increasingly adopted end-to-end, i.e., directly processing the audio signal without transcribing and operating on the text data. However, these models often lack interpretability, not revealing the reason behind individual prediction. To address this limitation, recent works address this issue by proposing feature attribution explainability methods. Some works operate on spectrograms [2, 1] and highlight parts of the spectrogram impacting predictions close to saliency map visualization in image classification tasks. Other works operate on the audio signal itself [3, 4, 6, 5], providing which audio segments impact the prediction, including equal-width audio [3] or phoneme-level [5] or word-level audio segments [4]. Moreover, recent works also focus on the paralinguistic aspects [4].

Goal.

The task of the project is first to review existing explanation methods for speech-based classification systematically. Then, the project aims to identify existing research gaps in the explainable AI literature for speech classification, focusing on feature-attribution explainability methods. Examples of research gaps include the combination of word-level explanations with paralinguistic aspects and assessing the quality of explanations.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing explanation methods for explaining the behavior of speech classification models.
- **Identification of Research Gaps.** Identify key research gaps to address for the context of explainability methods for speech classification.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to address the identified research gap. This may involve proposing a novel approach or adapting existing explainability methods to suit the context of speech classification.
- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach.

References

[1] Sören Becker et al. *AudioMNIST: Exploring Explainable Artificial Intelligence for Audio Analysis on a Simple Benchmark.* 2023. arXiv: `1807.03418 [cs.SD]`.

[2] Marco Colussi and Stavros Ntalampiras. *Interpreting deep urban sound classification using Layer-wise Relevance Propagation.* 2021. arXiv: `2111.10235 [cs.SD]`.

[3] Saumitra Mishra, Bob L Sturm, and Simon Dixon. "Local interpretable model-agnostic explanations for music content analysis." In: *ISMIR.* Vol. 53. 2017, pp. 537–543.

[4] Eliana Pastor et al. "Explaining Speech Classification Models via Word-Level Audio Segments and Paralinguistic Features". In: *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers).* Ed. by Yvette Graham and Matthew Purver. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 2221–2238. URL: `https://aclanthology.org/2024.eacl-long.136`.

[5] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. "Can we trust explainable ai methods on asr? an evaluation on phoneme recognition". In: *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE. 2024, pp. 10296–10300.

[6] Xiaoliang Wu, Peter Bell, and Ajitha Rajan. "Explanations for Automatic Speech Recognition". In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* 2023, pp. 1–5. DOI: `10.1109/ICASSP49357.2023.10094635`.

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.

P7 - Problematic subgroup identification on text

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Salvatore Greco

Project. This research aims to propose new methods for the identification and mitigation of disadvantaged subgroups in text classifiers.

Overview.

The identification of subgroups in which a model performs differently than overall behavior enables model understanding at the subgroup level and investigates their fairness and robustness. Existing solutions focus on tabular [5, 3, 1, 4] and speech data [2]. For speech data, they focus on the interpretable representation of utterances (e.g., the gender of the speaker, the speaking rate, and the level of noise). Few works focus on textual data, and most of them usually rely on template-generated evaluation data [6].

Goal.

The project involves identifying interpretable representations of textual data. A possible direction is to use LLM-prompts to derive categories (close to concept-based approaches). A definition of a hierarchy of categories could further enable the understanding. Once identified, we can leverage existing approaches for problematic subgroup identification.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing NLP fairness metrics, and techniques for the identification and evaluation of subgroups in textual data.
- **Identification of Research Gaps.** Identify key research gaps for identifying and evaluating subgroups performance, and optionally for mitigating performance disparities.
- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to identify and evaluate subgroups' performance. This may involve 1) using LLM-prompting in zero or few-shot

learning for annotating metadata in texts or 2) deriving categories (e.g., concept-based). Then, propose or leverage existing techniques of subgroups' performance disparities mitigation.

- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach for at least two datasets.

References

[1] Yeounoh Chung et al. "Slice finder: Automated data slicing for model validation". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. 2019, pp. 1550–1553.

[2] Alkis Koudounas et al. "Exploring subgroup performance in end-to-end speech models". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[3] Eliana Pastor, Luca De Alfaro, and Elena Baralis. "Looking for trouble: Analyzing classifier behavior via pattern divergence". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 1400–1412.

[4] Svetlana Sagadeeva and Matthias Boehm. "Sliceline: Fast, linear-algebra-based slice finding for ml model debugging". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 2290–2299.

[5] Nima Shahbazi et al. "Representation bias in data: a survey on identification and resolution techniques". In: *ACM Computing Surveys* 55.13s (2023), pp. 1–39.

[6] Tony Sun et al. "Mitigating gender bias in natural language processing: Literature review". In: *arXiv preprint arXiv:1906.08976* (2019).

P7 - Problematic subgroup identification on text

Explainable and Trustworthy AI Course

Politecnico di Torino - 2023/2024

Reference teachers: Eliana Pastor, Salvatore Greco

Project. This research aims to propose new methods for the identification and mitigation of disadvantaged subgroups in text classifiers.

Overview.

The identification of subgroups in which a model performs differently than overall behavior enables model understanding at the subgroup level and investigates their fairness and robustness. Existing solutions focus on tabular [5, 3, 1, 4] and speech data [2]. For speech data, they focus on the interpretable representation of utterances (e.g., the gender of the speaker, the speaking rate, and the level of noise). Few works focus on textual data, and most of them usually rely on template-generated evaluation data [6].

Goal.

The project involves identifying interpretable representations of textual data. A possible direction is to use LLM-prompts to derive categories (close to concept-based approaches). A definition of a hierarchy of categories could further enable the understanding. Once identified, we can leverage existing approaches for problematic subgroup identification.

Required analysis, implementation, and evaluation.

- **Literature Review.** Conduct a systematic review of existing NLP fairness metrics, and techniques for the identification and evaluation of subgroups in textual data.

- **Identification of Research Gaps.** Identify key research gaps for identifying and evaluating subgroups performance, and optionally for mitigating performance disparities.

- **Implementation.** Select a specific research gap to address. Propose and implement a methodology to identify and evaluate subgroups' performance. This may involve 1) using LLM-prompting in zero or few-shot learning for annotating metadata in texts or 2) deriving categories (e.g., concept-based). Then, propose or leverage existing techniques of subgroups' performance disparities mitigation.

- **Evaluation**. Assess the effectiveness and applicability of the newly implemented approach for at least two datasets.

References

[1] Yeounoh Chung et al. "Slice finder: Automated data slicing for model validation". In: *2019 IEEE 35th International Conference on Data Engineering (ICDE)*. IEEE. 2019, pp. 1550–1553.

[2] Alkis Koudounas et al. "Exploring subgroup performance in end-to-end speech models". In: *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE. 2023, pp. 1–5.

[3] Eliana Pastor, Luca De Alfaro, and Elena Baralis. "Looking for trouble: Analyzing classifier behavior via pattern divergence". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 1400–1412.

[4] Svetlana Sagadeeva and Matthias Boehm. "Sliceline: Fast, linear-algebra-based slice finding for ml model debugging". In: *Proceedings of the 2021 International Conference on Management of Data*. 2021, pp. 2290–2299.

[5] Nima Shahbazi et al. "Representation bias in data: a survey on identification and resolution techniques". In: *ACM Computing Surveys* 55.13s (2023), pp. 1–39.

[6] Tony Sun et al. "Mitigating gender bias in natural language processing: Literature review". In: *arXiv preprint arXiv:1906.08976* (2019).

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.