Comparison of XAI Methods

Group Project in XAI exercise

Task

- Many XAI methods have been discussed in the lecture so far. Which one is most appropriate to apply to a specific dataset?
- **Task:** Compare several XAI methods of your choice on the same dataset and analyze the comparison.
- Guiding questions:
  - Which XAI methods are applicable or make sense to apply for this task and model? Why?
  - How can they be evaluated? Which method performs better?
  - Based on the results, are there any observations that could lead to potential model improvements?

Dataset

- Resource (datasets with human annotations):
  - Image dataset:
    - https://github.com/holyseven/M4_XAI_Benchmark (1 text, 1 image)
    - https://xaidataset.github.io/dataset/ (8 image data)

  - Text dataset:
    - https://www.kaggle.com/competitions/tweet-sentiment-extraction/data (1 text)
    - https://github.com/acmi-lab/counterfactually-augmented-data (2 text)
    - https://github.com/holyseven/M4_XAI_Benchmark (1 text, 1 image)

- Optional (contact us):
  - Proposing your own projects/dataset
  - Demo/Late-breaking work paper.

How to work

- Form a group of 2 or 3 members (no more, no less) on ILIAS.
  - **Deadline: end of 12.12.2024**
- Choose **n** datasets and **n+1** models, where **n** is at least equal to the number of group members (**n** ≥ number of members).
- Steps:
  - Find a pretrained model or train a model for each dataset.
  - Apply selected XAI methods to these models for the chosen datasets.
  - Evaluate the XAI methods and analyze the results.

Presentation

- Task is done in groups, all group members get the same number of points (unless you talk to us because there are problems..)
- Practical assignment will be graded based on a **3-minute** project presentation in the exercises sessions on **13.01 and 20.01.2025**. Attendance in the session assigned to you is mandatory to pass. No exceptions.
- Copying and plagiarism are not allowed.