Project description for master’s thesis – Adrian Duric

Temporary title

The project is titled “Explainable Artificial Intelligence for improved machine learning performance”.

Temporary thesis statement

The overall goal of the project is to develop methods that use Explainable Artificial Intelligence (XAI) to potentially improve Machine Learning (ML) prediction performance.

Methodology

The project will utilize the HyperKvasir dataset (link), consisting of image data. On this data, the project will examine various saliency map methods to see which features of an image contribute the most to image classification. The thesis will primarily utilize well-established and widely used methods, though it could become relevant to explore newer ones as well. One of the most standard approaches to creating saliency maps is known as Grad-CAM (link). Other potential approaches to examine include newer variants of Grad-CAM, such as Grad-CAM++ (link), Smooth Grad-CAM++ (link) and Eigen-CAM (link).

Convolutional Neural Networks (CNNs) are often paired with saliency maps in scientific literature and will also be utilized for the project. As with saliency map algorithms, the thesis will mainly use well-known methods but might also explore novel approaches. Potential algorithms to examine in this project includes one of the most standard CNN architectures of recent years in ResNet (link), as well as its more recent variants, such as EfficientNet (link).

Once images have been classified by the CNN and saliency maps have been produced, the main focus of the thesis comes into play. The thesis will explore how the information present in the saliency maps can be used to improve the performance of the CNN used in classification. For example, this could be examined by retraining the ML model to emphasize the parts of the input image that were important for prediction, for instance through adding noise to the less important parts of the image.

Progress plan

The second semester (Spring 2024) will be used to examine relevant literature and write an essay as per the requirements for a long thesis (60 credits, the essay writing corresponds to 10 credits).

The remaining 50 credits are made up of working on the thesis in the following two semesters (Autumn 2024 and Spring 2025). Some of the work to be done in this period includes:

- Program a CNN to perform image classification on data from HyperKvasir. This will function as a baseline performance to compare later results against.

- Implement a saliency map algorithm to create saliency maps using the classified pictures.

- Based on the values from the saliency map, create an algorithm to retrain the CNN in a way that emphasizes the important parts of the input from the initial prediction.

- Measure performance before and after retraining and compare results.

- Potentially examine effects of using different CNNs or different saliency map methods.

- Examine related literature.

- Write thesis covering this methodology and its results.

Relevant curriculum

Relevant curriculum centers around topics related to image classification (CNNs) and XAI.

To cover image classification, I plan to take **IN4310 – Deep Learning for Image Analysis**, which directly covers this topic.

As for XAI, I plan to take **STK4900 – Statistical Methods and Applications**. This is due to there being close connections between XAI and statistics, meaning I can gain a deeper understanding of XAI by furthering my knowledge of statistical methods.