Assignment 4: Interpretability

The deadline for this assignment is ***June 21, 2022, 22:00***

You must submit your solution electronically via the Absalon home page.

Important rules and notes:

- *All assignments in the course are individual:*
- *You are not allowed to collaborate with anyone on the assignment and you are not allowed to communicate your solutions to other students.*
- *You must not ask for help from anyone except the teachers and TAs on the course.*
- *On the other hand, we encourage you to use the exercise classes and the Absalon forum to get help. The exercise sessions exist to help you with the assignments, and you are welcome to ask any questions related to the teaching material and the assignments on the forum.*
- *If your solution contains material from other sources than the assignment text, you must cite the source of the material and any changes you have made. This also applies to material from textbooks, Absalon, etc.*
- *If your solution uses methods or notation which are not used in the course material, you must specify where you have found the method or notation.*
- *If you are in doubt about plagiarism or citation rules, please ask the teachers or TAs.*

Please be very observant of these rules. We do not want any plagiarism cases, both for your and our sake.

This assignment has two parts.

Part A. Fill out the Google Form multiple-choice test at:

https://docs.google.com/forms/d/e/1FAIpQLSeJiKIpfE28yxc3b-KMP0xpLacitfYZsybUEX0lG_ijASXbyw/viewform?usp=sf_link

Part B. Write a **one-page research report** following the instructions in Appendix.

Appendix. Train a classifier on the MNIST dataset. This can be a binary classifier, say, distinguishing 0s from 1s. It can be any type of neural network. Now

a) produce pixel-wise heat maps using four different interpretability methods for two pictures in your validation split. You can use existing implementations, including Pytorch-compatible libraries.
b) compute the Spearman correlation across the pixels for each pair of interpretability methods, averaged across *all* the data points in your validation split, and present these coefficients in a 4x4 table.
c) Try to explain any differences in the coefficients.