L193 - Explainable Artificial Intelligence
Practical #1

Mateja Jamnik, Zohreh Shams, Mateo Espinosa Zarlenga

February 14th, 2025

1 Introduction

In this practical, you will explore some of the e**X**plainable **A**rtificial **I**ntelligence (XAI) techniques we learnt about in our lectures by applying them to a real-world task. Specifically, we will show you how to use feature attribution methods, both model-agnostic and model-specific algorithms, to try and debug a misbehaving Deep Neural Network (DNN) that will prove to be very good at hiding its flaws. For this, we will learn how to use Python's standard library for SHapley Additive exPlanations (SHAP) and then move to explore DNN-specific saliency methods, spending a good amount of time gaining intuition about how these methods work and how they can be improved. Logistically, this practical will be almost entirely run as a Google Colab notebook where all instructions and your written solutions will live. Therefore, the purpose of this document is to help you set up the necessary machinery to get a shinning Colab GPU running for you to use for the duration of the practical.

2 Practical Setup

Throughout this practical, we will be dealing with DNNs and some of their known issues. Because these models tend to be computationally demanding, we will piggyback on top of Google's Colab free GPU allocation and let all the computational hard work be done on their servers. For you to be able to do this, you need to follow the following steps:

1. First and foremost, have your login details for a Gmail-accessible account you can use at your disposal (your Cam account could work!). If you don't have one, you can quickly set one up here.

2. Familiarise yourself with Google Colab by skimming over this page. In short, Google Colab allows you to run iPython notebooks using their hardware accelerators (GPUs and TPUs). For the purposes of this practical, we will focus on using GPUs only.

3. Open the following Colab Notebook and copy it to your personal cloud drive. You may do this by clicking on `File > Save a copy in Drive`. For the sake of making sure things work as best as they can, we recommend that students open this Colab notebook using Chrome.

4. Follow the instructions in the notebook to complete the different exercises of this practical.

3 Submitting Your Answers

We are aware that the number of exercises in the notebook we provided is considerable given the amount of time allotted for this practical. Therefore, we do not expect all exercises to be completed by the end of the practical (although it would be amazing if they are!). The point of this session is for you to get familiar with at least the first two parts of the practical and to ask any questions you may have to the instructors if you get stuck.

To get **assessed for this practical**, please submit your answers, in notebook format, via **Moodle by the 24$^{\text{th}}$ of February at 14:00**. We ask you to please **clean up the notebook before submission** (e.g., remove unnecessary cells and outputs) and make sure all questions are answered clearly in their respective sections. **We will grade almost entirely based on analysis and understanding of the material rather than on your ability to code**. We expect students to complete about 1/2 to 2/3 of the exercises in the practical class and complete the rest as homework. Each practical will be worth 10%.