Asha Aher, Chinmay Vadagma, Ronak Mehta, Manmeet Singh

In our project, we will focus on using contrastive semi-supervised learning to classify the disease or anomaly type in an X-Ray or MRI scan. We will use the SimCLRv2 model underneath in order to perform semi-supervised learning. The model will be trained using small segments of various datasets in order to broad spectrum classifier. A key requirement is for it to be able to reliably classify for a set of various inputs that may be from different MRI/X-Ray machines (ie. Different datasets for scans taken generally the same purpose).

We will fine tune SimCLRv2 for our multi-class classification problem and demonstrate how training on additional datasets (for more disease types, etc.) can be used as a form of additive learning

We will build a webapp around this and allow classification on custom user scans. This could be integrated into a radiologist centric application which is easily extensible to new disease types and requires low amount of data to "learn". small datasets required = smaller model size? Means we can train a SimCLR model on multiple datasets simultaneously

We have identified some initial datasets to start with:

https://www.kaggle.com/c/1056lab-brain-cancer-classification
https://www.kaggle.com/paultimothymooney/predicting-pathologies-in-x-ray-images

End project deliverables:

- Web application hosting the model capable of multi-class classification on Chest/brain images.
- Colab with the model code

Inputs: MRI or X-Ray images containing multiple types of anomalies in a particular type of scan

Output: Disease/anomaly Classification based on training with relatively small amounts of data (20-30% of total).

Key evaluation metrics: Precision, Recall, F1 score

The overall objective of the project is to fine tune SimCLRv2 to another prevalent problem, namely that of disease classification. Additionally, the goal is to explore creating a full end-to-end ML lifecycle pipeline which can ultimately be used to expand the learning of this semi-supervised learning model with very few input instances of a novel anomaly type.

This project is dissimilar from our Master's project where we're currently focusing on building a generative Alzheimer's disease classifier based on GANs, and the overall using GAN based approaches for classification.