PhD Course: Rise of the Transformer

Hackathon Task Description;

(1st of February 2023)

Your task for the Hackathon is to implement in Google Colab a CLIP model which embeds a set of images and corresponding captions (text descriptions of the images) into the same embedding space:

- See this blog post: https://openainode src/server.js.com/blog/clip/
- And this paper: https://arxiv.org/pdf/2103.00020.pdf

The CLIP model was discussed during the lecture of Prof Boracchi. Once the CLIP model has been learnt, **the aim will be to use the model to perform image retrieval**, i.e. to find the most related image in the collection to the text query (or vice versa, find the caption in the dataset which is closest to a query image).

NOTE: You are expected to implement CLIP (possibly in TensorFlow 2 / Keras) and not to copy/paste it from the OpenAI PyTorch repository.

The dataset of images and text that we will use for this task is a simplified version of the ROCO (Radiology Objects in COntext) dataset:

- https://link.springer.com/chapter/10.1007/978-3-030-01364-6_20

Note that this dataset is available only for research purposes, so **use it only for the hackathon task** and **please delete it afterwards**.

The dataset contains **thousands of medical images**, **each with a corresponding caption** (text describing the image). In order to be possible to train the model with limited resources in Google Colab, the images have been reduced in size (by lowering the resolution) from those in the original dataset.

- The reduced images can be found in the file: **Train/resized_train.zip**
- The corresponding text captions are found in: **Train/caption_prediction_train.csv**

Each image is also associated with a set of **concept labels** from a set of over 8 thousand classes present in the file “Train/concept_detection_train.csv”. You do NOT NEED to make use of these labels, but you may decide to do so. For example, you could pre-train the image encoder as a multi-label classifier, or make use of the string defining each label (see file “concepts.txt”) as additional text to augment the captions associated with the images.

In the Hackathon, once you have trained CLIP and tested the text/image retrieval, you are free to address more sophisticated problems such as zero/few shot learning. This was mentioned during prof Boracchi' lecture and you can find how to use CLIP for this purpose in the CLIP paper. Considering that exam for MSc students has not to be finalised today, we leave this as an opportunity for further work in their projects.