Paper summary

The paper I chose to summarize is : How to train Vision Transformer on Small-scale Datasets ?.

In recent years, Vision Transformers (ViT) have become a great alternative to CNN architectures. But, despite its great performance, it usually needs pre-training on millions of images to be performant on downstream tasks. So it is not given to anybody to pre-train a ViT from scratch as one needs the dataset and also the time to train it. The problem stays the same if you want to train on a small-scale dataset.

To alleviate this issue, this paper came up with a strategy that is simple and yet allows to improve state-of-the-art performances on small-scale datasets.

There are two main steps to this strategy:

- The first step is to initialize the weights on the small-scale dataset using a self-supervised method. Indeed, we train a teacher model and a student model together. In addition to that, they add a concept of **Low-resolutionView Prediction.** It allows the teacher model to be trained on **global** view images representing at least a proportion of 50% of the input image at a low resolution. The student is trained on a **local** view image with a random crop of 20-50% of the input image and with a ratio of 1:4 compared to the teacher view. In both cases, they apply standard augmentation to each local and global view. The self-supervised method optimizes the following : The teacher generates target features for the global view, then the student is trained on local and global views to generate predicted features. The student parameters are updated by an optimization function. Finally, the teacher weights are updated by an Exponential Moving Average.
- After the initialization of the teacher weights in the first step, the teacher weights are used to fine-tune the model on the same dataset. The MLP projection head during the self-supervised training step is replaced by a randomly initialized MLP head that serves for training on the classification objective on the dataset. The loss used is a categorical-crossentropy loss and no other architecture changes are made.

In conclusion, this strategy allows to train a ViT model on a small-scale dataset without having to pre-train the model on millions of images. So it enables to gain time and also have decent performance on a small-scale dataset.

When I read the publication, I was looking for ideas and strategies to implement for the technical project you assigned me. I felt it was particularly interesting because I wanted to use ViT as they were proven to be quite performant for classification tasks in recent years and also it could be applied to the small-scale dataset that was given to me for the task.

Of course, I could have done some strategies to augment the size of the dataset. For example, I could have scrapped images online or even applied offline data augmentation on the images as well.

But I chose to not do it for the following reasons:

- Scrap images online : This is a common idea when one wants to have more images in their dataset but it would have been very time-consuming to clean the dataset afterward.
- Data augmentation : Since the dataset is small, it would have worked to some extent but I believe that the diversity and variety of data would have become a problem and made the model overfitting.

For those reasons, I found this paper really interesting because it could solve the problem I was facing using a model architecture that has proven its performance in addition to using a simple strategy to overcome it.