TU Wien, LV Assignment 01, LV 185.A83 Machine Learning for Health Informatics Class of 2021

Tutor: Marcus BLOICE, HCAI Lab

Semi-supervised and self-supervised learning in computer vision and image analysis

Semi-supervised learning is a sub-field of machine learning that has been popular in the field of natural language processing (NLP) for several years and is the basis for some of the most powerful language models currently in use, such as BERT [1].

It is used in situations where both **labelled** and **unlabelled** data are available. It is especially used in cases where the number of unlabelled samples far exceeds the number of labelled samples. Very recently, the technique is also beginning to see adoption in deep learning for **computer vision and image analysis**.

Interestingly, we are now approaching the point where unsupervised and semi-supervised image classification algorithms are **outperforming purely supervised methods on the same datasets**, using **only a fraction of the number of labelled samples**.

In this Mini Project, you will apply semi and self-supervised learning by implementing the SimCLR algorithm described in [2] See also this video, which describes the paper and algorithm in detail, by Yannic KILCHER, ETH Zürich: https://www.youtube.com/watch?v=2lkUNDZld-4

This paper actually is a good example, as the authors applied unsupervised, supervised, and self-supervised approaches in a single workflow, and you can become familiar with all concepts.

Your tasks in this project are:

- Understand what semi-supervised and self-supervised algorithms are, how they work, and when they are applied.
- Implement a version of SimCLR from the paper above (a GitHub repository is available with source code, and it should not require large amounts of programming to get started: https://github.com/google-research/simclr)
- Choose a dataset, and compare a purely supervised approach to your semi-supervised approach. Change the ratio of labelled versus unlabelled samples and experiment with how this affects accuracies. Report on your results.

Before selecting a dataset, consult with your tutor about whether this dataset is suitable. You need to choose a labelled dataset, and discard a proportion of the labels to simulate the unsupervised aspect of the SimCLR workflow. By choosing a labelled dataset, we can compare the semi-supervised approach to a purely supervised approach.
Contact directly: marcus.bloice AT medunigraz.at

[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. https://arxiv.org/abs/1810.04805
[2] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi & Geoffrey Hinton (2020). Big self-supervised models are strong semi-supervised learners. https://arxiv.org/abs/2006.10029