Improving self-supervised vision transformers with attention

Environment: QARMA (machine learning) Team at Laboratoire d'Informatique et Systèmes (LIS)
Location: Ecole Centrale Marseille (ECM), Technopôle de Château-Gombert, Marseille
Supervisors: R. Sicre (LIS – ECM), S. Ayache (LIS - Polytech)
Salary: legal minimum
Keywords: computer vision, deep learning, interpretability
Contact: ronan.sicre@lis-lab.fr

Computer vision and deep learning received a lot of attention lately due to the great improvements brought by Deep Neural Networks. Over the last decade, these networks are addressing more complex tasks and are reducing their requirement for large amounts of annotated data. Specifically self-supervised method learn good transfereable visual representations without requiring any labels.

Several works have been proposed to learn self supervised representation, based on rotations and deep clustering. Then methods proposed the use of contrastive loss combined with several augmentations, meaning that several augmentations of each image of the dataset are considered as positive while the ramining images of the dataset are negatives. The contrastive loss pulls together positives representation and push away negative ones. This methods was later combined with teacher-student distillation, which shows better generalisation capabilities of the representations [1]. Finally with vision transformers architecture appearing [2], random token masking has been added as another method for self supervision. IBOT [3] combines these three latest methods to perform self supervised learning. The work of [4] then proposed to replace random masking, by attention based masking.

The candidate will first need to setup the experimental protocol for self supervised visual representation following [4]. Representations are learned on a subset of ImageNet. Then the evaluation is performed by using K-nn or linear probing on the representations. Then we will study this masking method further and use attention to define new losses allowing self-supervised learning.

[1] Grill, Jean-Bastien, et al. "Bootstrap your own latent-a new approach to self-supervised learning." NeurIPS 2020
[2] Dosovitskly, Beyer, et al. : "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" ICLR 2021
[3] Zhou, Jinghao, et al. "ibot: Image bert pre-training with online tokenizer." ICLR 2022
[4] Kakogeorgiou, Gidaris et al. : "What to Hide from Your Students: Attention-Guided Masked Image Modeling". ECCV 2022