Deep Learning Project Proposal

Léo Tronchon
Ens Paris Saclay
leo.tronchon@ens-paris-saclay.fr

Arthur Zucker
Ens Paris Saclay
arthur.zucker@ens-paris-saclay.fr

Abstract

In recent years, deep learning with self-supervision for images has gained a lot of traction in the vision community. As self-supervised models' performances keep getting closer to their supervised counterparts, 2 recent papers have stood out as breakthroughs in the field. First, DINO [1] has set a new SOTA on Imagenet [2] and shown how Vision Transfomers learn to pay attention to important elements in Self-Supervised settings. Second, although not beating the SOTA, Barlow Twins [5] proved that Self-Supervised models can naturally avoid collapse by using a cross correlation matrix as the loss function, and enforce its convergence towards the identity matrix.
Our primary goal is to combine ideas from DINO and Barlow Twins to design a new self-supervised architecture featuring both a cross entropy loss and a loss based on a cross correlation matrix. As a secondary task, we will attempt to leverage the stability induced by the Barlow Twins' loss to discard some of the hyperparameters used in the DINO architecture. Finally, depending on how well our model performs, we will investigate either the attention maps obtained by the new architecture, or the ones obtained with a ViT-based [3] Barlow Twins.

1. Project proposal

1.1. Motivation

It is our understanding that, while the cross correlation loss in Barlow Twins creates rich embeddings for images, its objective is fundamentally different from cross entropy. The cross entropy used in the context of DINO learns implicit classes from self supervision, via the distribution created by the last layers of the network. On the other hand, the cross correlation loss pushes the model to describe the image with an uncorrelated feature representation. Moreover, the DINO architecture relies on multiple hyperparameters which have to be optimized and curated for a given dataset. We believe that having a combination of the two losses as a double objective could be of great benefit to a self-supervised learning model. It could potentially increase its explanatory power while regularizing the learning process.

1.2. Proposed architectures

Figure 1 and 2 feature the two main architectures we wish to explore during this project.

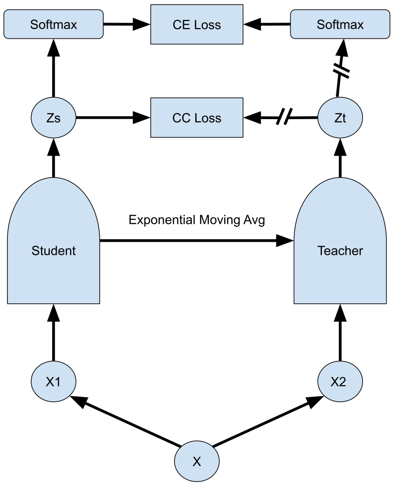

Figure 1. First architecture

Those designs are obviously subject to changes depending on the results obtained after implementation and analysis. Self-supervised models are often subject to *mode collapse*, therefore it is difficult to anticipate the performances of a specific architecture.

- We will use a similar image preprocessing to that of DINO, with student fed local and global crops while teacher is fed global crops only. The recent BYOL [4] augmentation method will also be re-used.
- We will use CIFAR-10, CIFAR-100, and Imagenet-32 datasets initially as they are the only datasets we can afford to use with the limited GPU space at our disposition. We will use datasets with larger images if the model performs well.

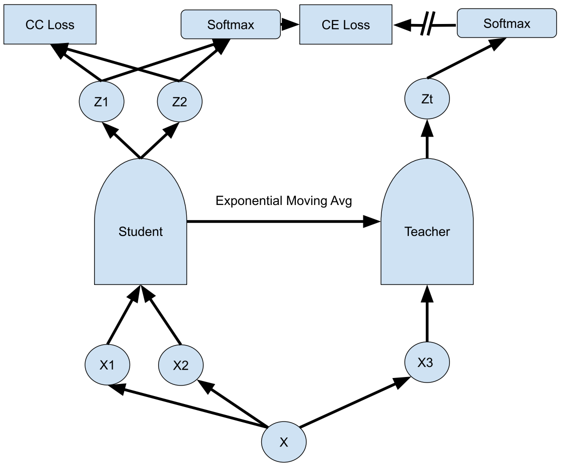

Figure 2. Second architecture

2. Alternatives

In the event that our architectures cannot avoid collapse, we will simply reproduce the Barlow Twins model with a ViT backbone as well as the DINO architecture. We will then investigate whether the attention maps obtained from the 2 architectures differ significantly on our datasets of 32x32 images. Additionally, using the attention map learned by the multi head attentions presented in [1], we want to explore possible transfer learning and benchmark the usage of self-supervision in semantic segmentation tasks. This can be tested on ImageNet, which also includes semantic segmentation masks.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*, 2021. 1, 2

[2] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1

[3] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 1

[4] Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning, 2020. 1

[5] Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. *arXiv preprint arXiv:2103.03230*, 2021. 1