MO435 — Probabilistic Machine Learning
INSTITUTO DE COMPUTAÇÃO — UNICAMP

Self-supervised learning: non-generative representation learning

Lesson No. 07

Leonardo Alves de Melo - RA: 156188

1 Introduction

In Machine Learning, a supervised learning model tries to predict a label of a given data after been exposed to a dataset that matches each data to its respective true label. The problem is that sometimes it is very hard or even impossible to have a label for each data, being necessary to use another type of model. In this scenario, a self-supervised leaning model can be useful, because it can handle this problem even without someone explicitly tells what are the labels.

In this summary it will be discussed self-supervised models, and it will be detailed the state-of-the-art called SimCLR [1].

2 Formal Definition of Self-supervised Learning

A self-supervised model is defined as being a model that automatically creates a supervisory signal of the given dataset to resolve the given task. The way that this model generates this signal depends on the context, and in this summary it will be presented how SimCLR labels images in the training process.

3 SimCLR

SimCLR is the current state-of-the-art of self-supervised models, it is a framework for contrastive learning of visual representations. The main idea is to compare each data and use this comparisons to learn useful representations of the data.

The first step of this framework is the data augmentation. For each image x_i in dataset, some filters were applied to them, creating a new image x_j. Then, a base encoder, which is a convolutional neural network variant based on the ResNet architecture [2] and is denoted as $f(\cdot)$, is applied to x_i and x_j, generating $f(x_i) = h_i$ and $f(x_j) = h_j$. The next step is the projection head, denoted as $g(\cdot)$, that is responsible to calculate the non-linear projection of h, generating $g(h_i) = z_i$ and $g(h_j) = z_j$, using a Multilayer Perceptron (MLP) in a way that maximizes the agreement between z_i and z_j. All this steps can be seen in Figure 1.

3.1 Training

The training process starts with a batch of images of the dataset, then for each image in this batch the following process occur: all the parameters of the encoder and projection head are optimized in a way to minimize the loss function, given by Equation 2, known as NT-Xent, which normalizes the temperature-scaled cross entropy. The main idea is to maximizes the similarity between a projection z_i of an image x_i and the projection z_j of its augmented image x_j, and minimizes the similarity between z_i and the projection of other images of the batch. The similarity function could be computed using for example the cosine similarity, given by Equation 1.

$$s_{i,j} = \frac{x_i \cdot x_j}{\|x_i\| \cdot \|x_j\|} \tag{1}$$

$$\ell_{i,j} = -\log \frac{\exp(s(z_i, z_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k\neq i]} \exp(s(z_i, z_k)/\tau)} \tag{2}$$

where $\mathbb{1}_{[k\neq i]} \in 0,1$ is an indicator function evaluating to one if and only if $k \neq i$ and τ is a temperature parameters.

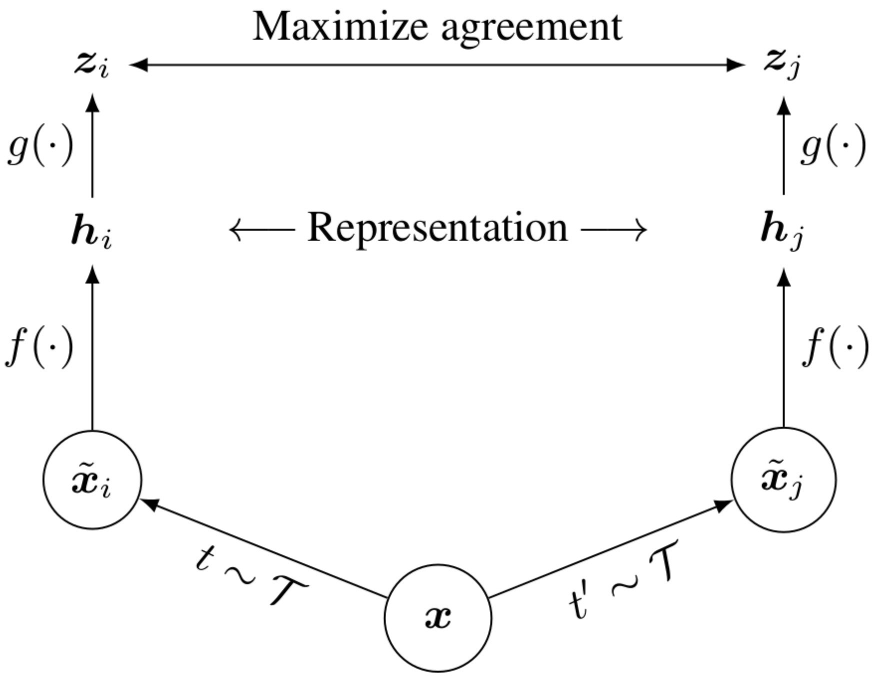


Figure 1. SimCLR Framework.

3.2 Results

SimCLR reached as good results as some supervised prediction models, which can be seen in the graph of Figure 2. This results are obtained using the concept of Linear Evaluation Protocol, which consists of using the result h_i of the encoder of an image x_i and the true label of x_i to train a linear classifier, like Linear Regression, and then calculates the accuracy. An image of the results using different transformations can be seen in Figure 3, where the transformation *crop* with *colors changes* represent the best transformations.

4 Conclusion

In this summary it was possible to understand what is a self-supervised model and its difference between a supervised model. It was also explained how it works the state-of-the-art SimCLR, detailing the training process with an image representation and the formulas used, and, in the end, how well this model is compared to others.

References

[1] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, *A simple framework for contrastive learning of visual representations*, 2020. eprint: arXiv:2002.05709.

[2] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV].

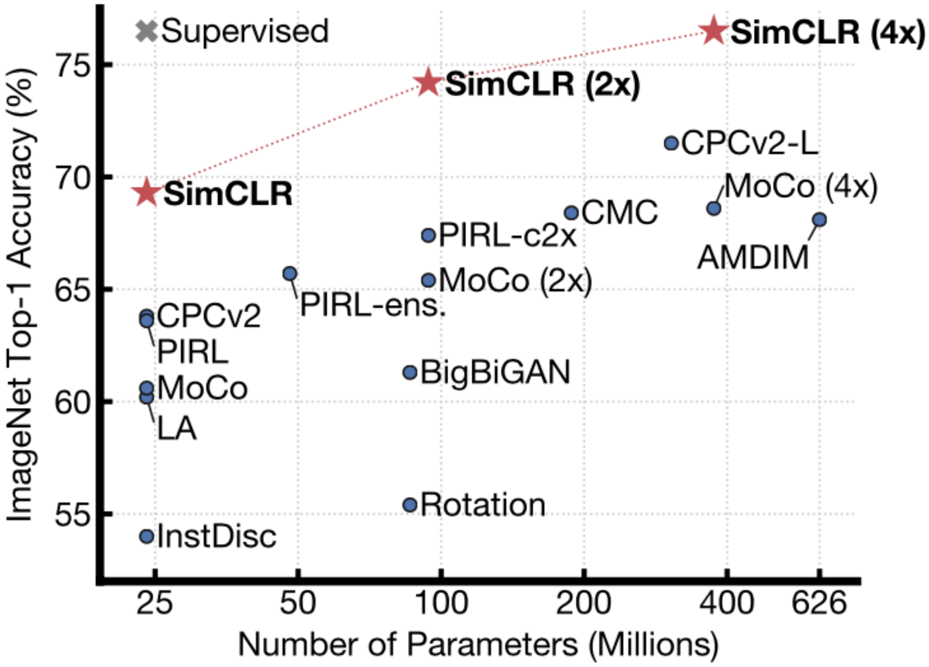
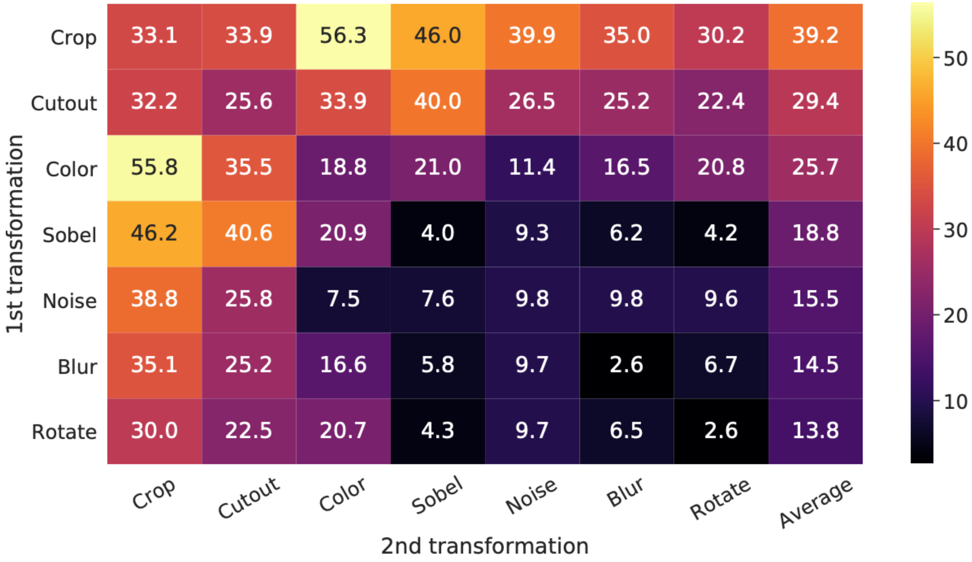


Figure 2. SimCLR results compared to other models.



1st transformation \ 2nd transformation	Crop	Cutout	Color	Sobel	Noise	Blur	Rotate	Average
Crop	33.1	33.9	56.3	46.0	39.9	35.0	30.2	39.2
Cutout	32.2	25.6	33.9	40.0	26.5	25.2	22.4	29.4
Color	55.8	35.5	18.8	21.0	11.4	16.5	20.8	25.7
Sobel	46.2	40.6	20.9	4.0	9.3	6.2	4.2	18.8
Noise	38.8	25.8	7.5	7.6	9.8	9.8	9.6	15.5
Blur	35.1	25.2	16.6	5.8	9.7	2.6	6.7	14.5
Rotate	30.0	22.5	20.7	4.3	9.7	6.5	2.6	13.8

Figure 3. SimCLR linear evaluation over transformations.