Machine Learning CS 460 - Project

VARIANCE INVARIANCE COVARIANCE REGULARIZATION

by Aaditya Vicram Saraf, under guidance of Dr. Subhankar Mishra
School of Computer Sciences, NISER Bhubaneswar

aadityavicram.saraf@niser.ac.in

Introduction

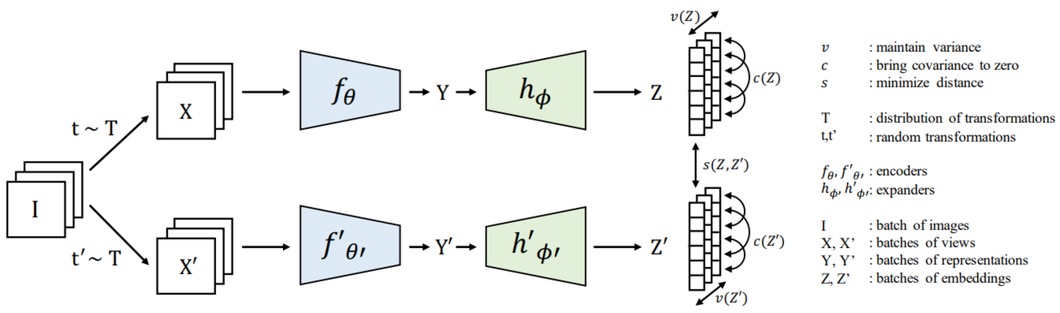
The problem to be discussed at hand is to train a model which takes image inputs and transforms them into embedding vectors, in such a way that the vectors obtained for different views of the same image have minimal distance, while dissimilar images are assigned distant vectors. This is useful as it helps to identify important features of an image and hence aid in downstream tasks.

In a self-supervised learning approach, to transform input data into meaningful embedding, a neural network architecture called encoders are used. Encoders bring along *collapse problem* in which the encoders produce a constant or non-informative embedding, leading to loss of useful information. VIGReg is a method for self-supervised learning that prevents collapse by applying variance and covariance regularization to the embedding vectors, to aid in downstream tasks. It is more robust than Barlow Twins and SimCLR, and does not require weight sharing, batch normalization, etc.

How it works?

The model

Given a batch of images I, two batches of different views X and X′ are produced and are then encoded into representations Y and Y′. The representations are fed to an expander producing the embeddings Z and Z′. The distance between two embeddings from the same image is minimized, the variance of each embedding variable over a batch is maintained above a threshold, and the covariance between pairs of embedding variables over a batch are attracted to zero, decorrelating the variables from each other. Although the two branches do not require identical architectures nor share weights, in most of our experiments, they are Siamese with shared weights: the encoders are ResNet-50 backbones with output dimension 2048. The expanders have 3 fully-connected layers of size 8192.[2]



VICReg Joint embedding with variance, invariance and covariance regularization

Image augmentation includes changing an input image by applying one of the effects shown below. The aim of self-supervised learning is to embed these images into representations that are highly similar provided the images are augmented versions of a parent image.

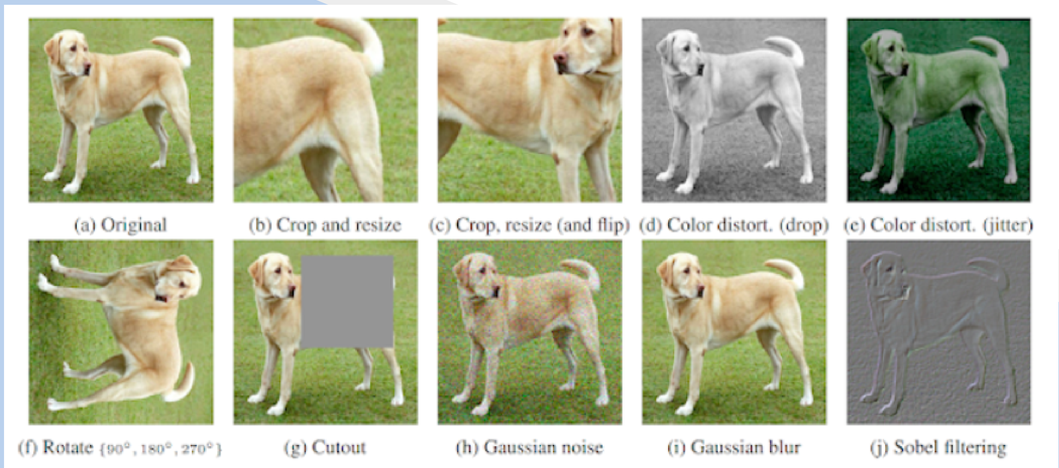


Figure 1: Augmented versions of the original image of a dog. Source : SimCLR post.[1]

Intuition behind VICReg

VICReg introduces Variance Invariance and Covariance Regularization to train joint embedding architectures, based on principles or preserving the information content of the embeddings. The loss function uses 3 terms :

- **Invariance :** This minimizes the mean square distance between embedding vectors produced by different encoders for the same original data.

$$s(Z, Z') = \frac{1}{n}\sum_i \|z_i - z'_i\|_2^2$$

- **Variance :** A hinge loss to maintain standard deviation above a certain threshold within a batch. This forces embedding vectors obtained from data within a batch to be significantly different. $v(Z)$ denotes the variance term, defined as below :

$$\mathrm{Var}(z_{:j}) = \frac{1}{n-1}\sum_{i=1}^{n}(z_{ij} - \bar{z}_j)^2, \text{ where } \bar{z}_j = \frac{1}{n}\sum_{i=1}^{n} z_{ij}$$

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max\left(0, \gamma - \sqrt{\mathrm{Var}(z_{:j}) + \epsilon}\right)$$

- **Covariance :** This term minimizes the covariance (over a batch) between every pair of (centered) embedding variables. This term decorrelates the variables of each embedding and prevents an informational collapse in which the variables would vary together or be highly correlated. $c(Z)$ defines the covariance term, where

$$C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_{i:} - \bar{z}_{i:})^T, \text{ where } \bar{z}_{i:} = \frac{1}{n}\sum_{i=1}^{n} z_{i:}, \text{ then } c(Z) = \frac{1}{d}\sum_{l\neq m} C(Z)_{lm}^2$$

Application

In other recent algorithms like BYOL and Simsiam, without contrastive samples, high quality representations are obtained mainly due to shared weights or symmetrical architecture. Apart from that they rely on batch-wise or feature-wise normalization, which play the role of repulsive term.
An experiment was conducted where correlation matrices for BYOL, Simsiam, VICReg, VICReg without covariance regularization were all computed and it was found that even without covariance regularization, BYOL and Simsiam would itself minimize correlation in vectors. This mostly would rely on dependency of the architectures, and it is difficult to clearly interpret the reason for this. VICReg explicitly defines Variance and Covariance terms making it easy to understand, with encoders in independent branches.

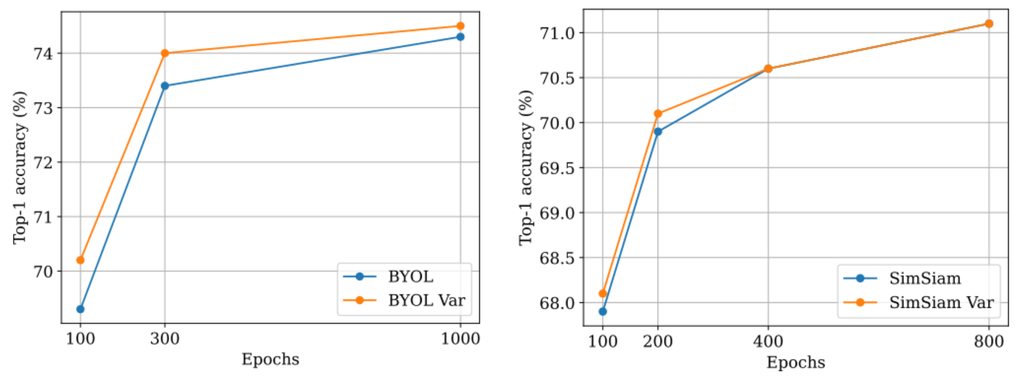


Figure 2: Incorporating variance regularization in BYOL and SimSiam. Top-1 accuracy on the linear evaluation protocol for different number of pretraining epochs. For both methods pre-training follows the optimization and data augmentation protocol of their original paper but is based on our implementation..

References

[1] Vicreg: Tutorial and lightweight pytorch implementation. https://imbue.com/open-source/2022-04-21-vicreg/. Accessed: 2022-04-21.

[2] A. Bardes, J. Ponce, and Y. LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.