Object Recognition & Computer Vision

Topic D - Self-Supervised Learning for Visual Representations

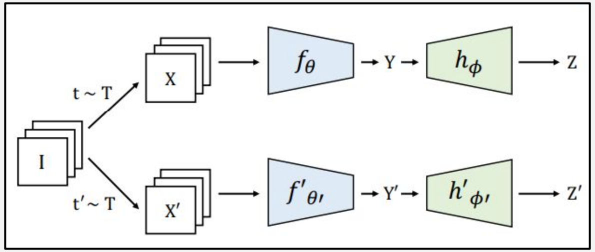


Figure from Bardes et al, ICLR 2022

Théo Di Piazza (theo.dipiazza@gmail.com)

January 2023

1. Introduction, context

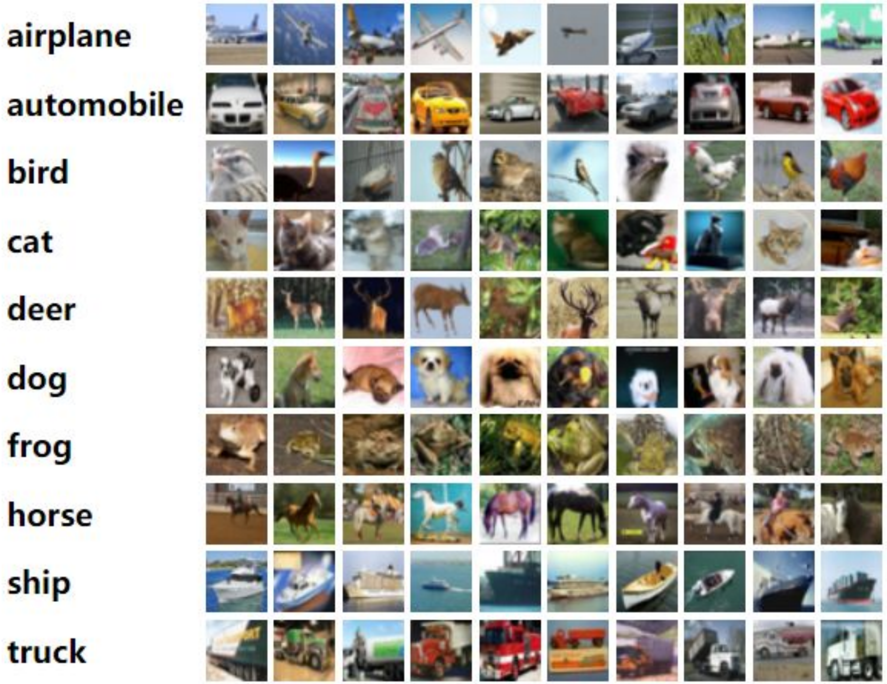


Figure 1.1 : CIFAR-10 datasets with labels

Labelling data :

- Costly
- Time consuming
- Biased towards the responsible

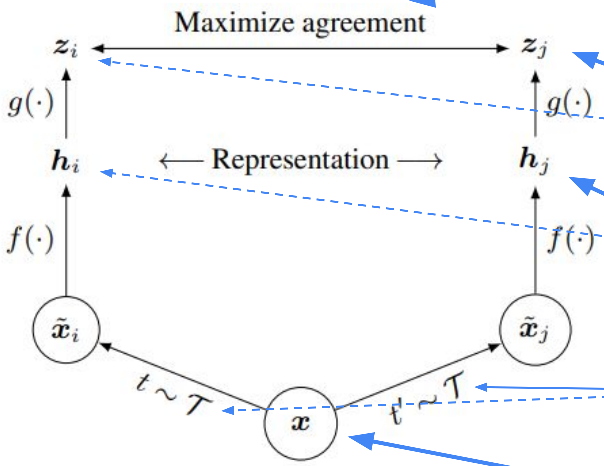
⇒ How to learn visual representations without labels?

Plan

1. Introduction, context
2. Theory: Contrastive learning & VICReg
3. Handling the method
4. Pretrain backbone
5. Linear evaluation
6. Qualitative results
7. Conclusion

2. Theory: Contrastive Learning & VICReg

Contrastive Learning of Visual Representations



Contrastive Loss

Projection head

Extract representation vectors with backbones

Sample 2 data augmentation operators

From a given image

Figure 2.1 : Scheme of SimCLR
Source: Chen et al, ICML 2020

https://arxiv.org/pdf/2002.05709.pdf

2. Theory: Contrastive Learning & VICReg

Contrastive Learning of Visual Representations

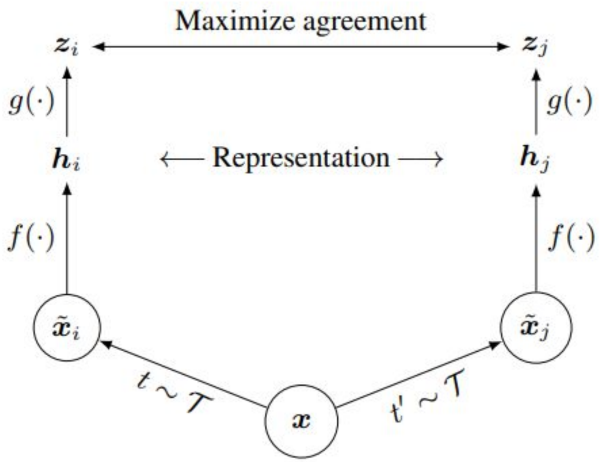


Figure 2.2 : Scheme of SimCLR
Source: Chen et al, ICML 2020

→ Data Augmentation plays a critical role for effective predictive tasks.

→ Adding learnable non-linear transformation between the representation and the loss improve quality of representations.

→ Benefits from larger batch sizes and more training steps.

→ Collapse: Encoders produce constant or non-informative vectors.

https://arxiv.org/pdf/2002.05709.pdf

2. Theory: Contrastive Learning & VICReg

VICReg

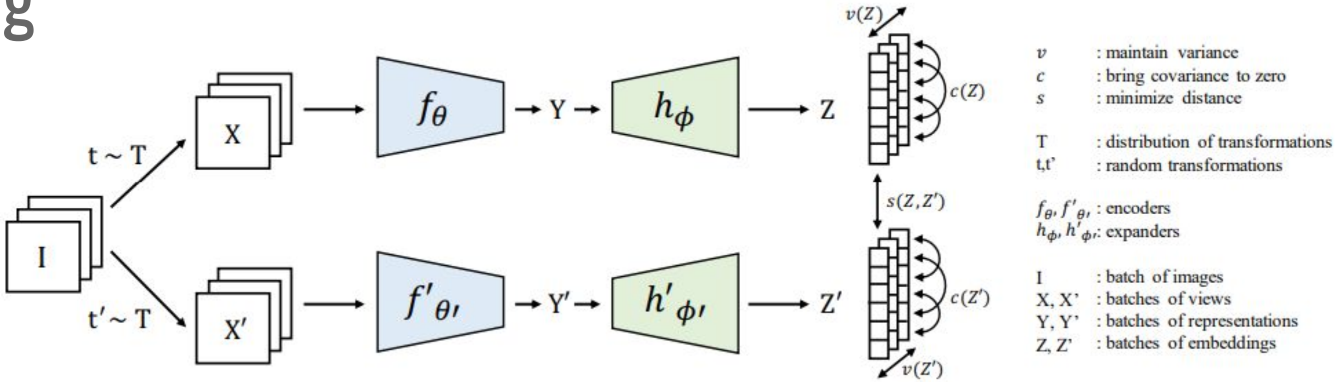


Figure 2.3 : Scheme of VICReg

Source: Bardes et al, ICLR 2022

Details about loss function

$$l(Z,Z') = \lambda s(Z,Z') + \mu[\underbrace{v(Z)+v(Z')}_{\text{Variance}}] + \nu[\underbrace{c(Z)+c(Z')}_{\text{Covariance}}]$$

$$c(z) = \sum_{i \neq j} [C(Z)]_{i,j}^2 \;\; ; \;\; s(Z,Z') = \frac{1}{n} ||z_i - z_i'||_2^2$$

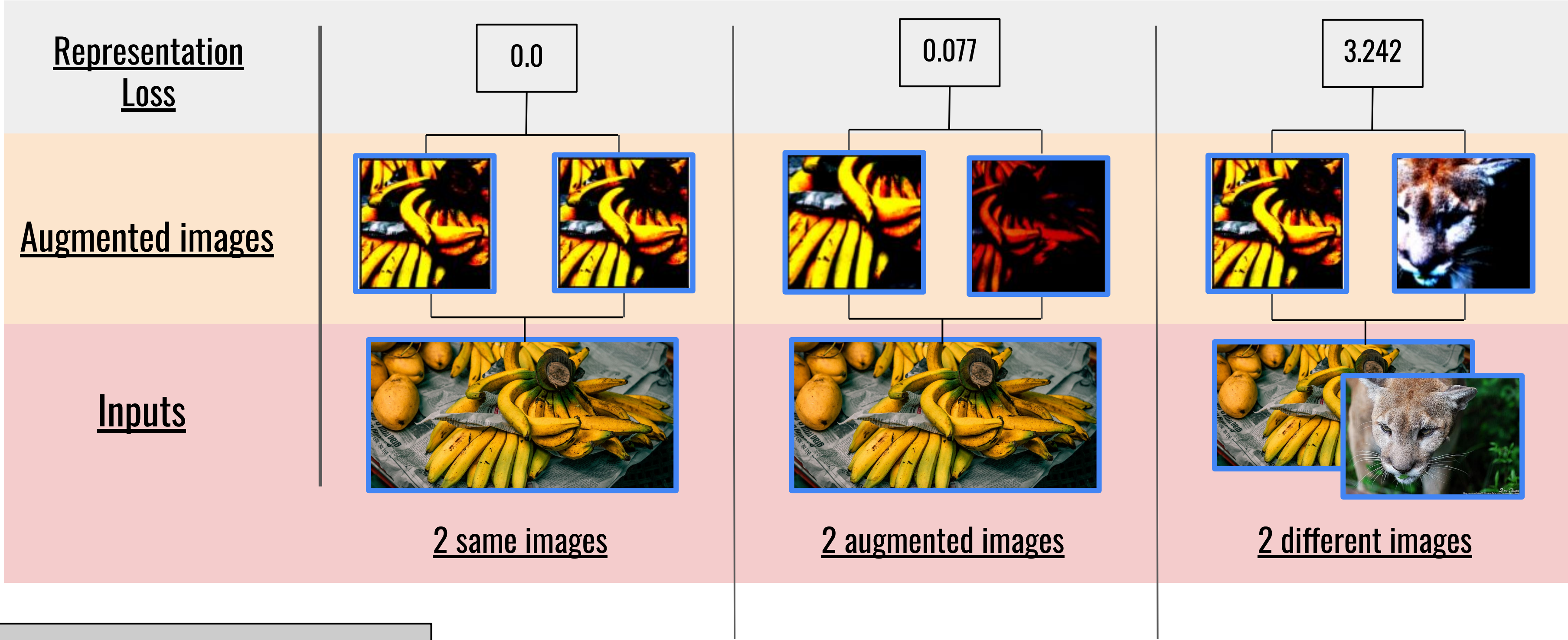
$$C(Z) = \frac{1}{n-1} \sum_{i=1}^{n} (z_i - \hat{z})(z_i - \hat{z})^T$$

$$v(Z) = \frac{1}{d} \sum_{j=1}^{d} max(0, \gamma - \sqrt{Var(z_j) + \varepsilon})$$

https://arxiv.org/pdf/2105.04906.pdf

3. Handling the method

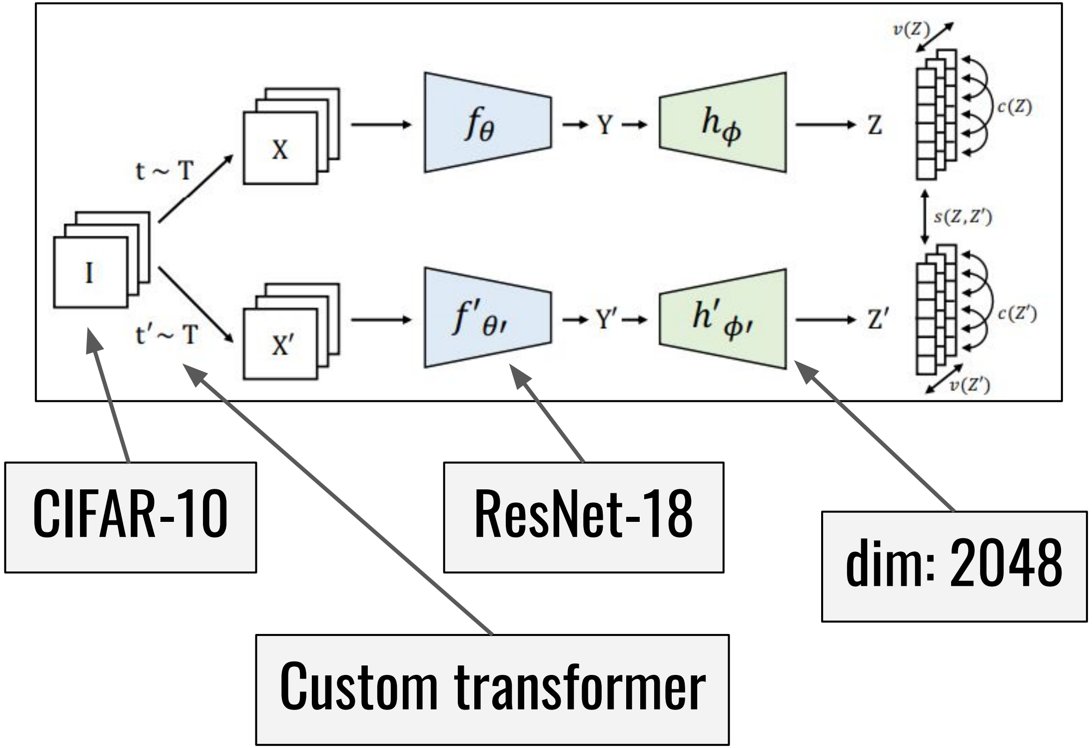
VICReg forward propagation



ImageNet https://www.image-net.org/

4. Pretrain backbone, linear evaluation

Figure 4.1 : Proposed architecture



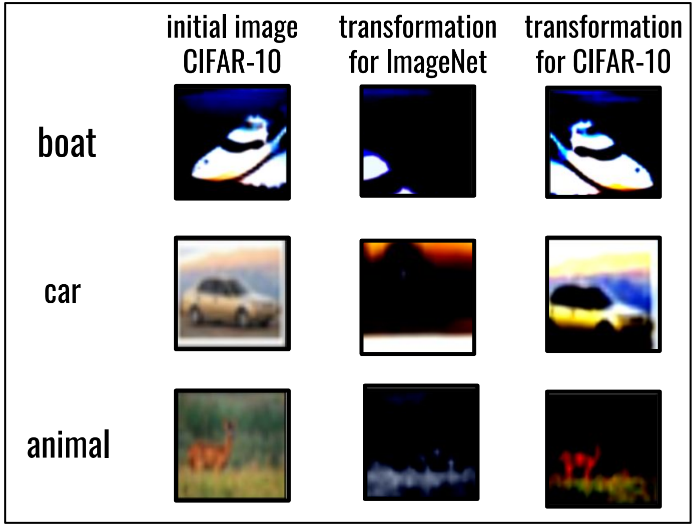
3 samples from CIFAR-10

4. Pretrain blackbone

transformations - Details

from the paper	Horizontal flip	RandomResizedCrop	Brightness 0.4	Contrast 0.4	Saturation 0.2	hue 0.1
for CIFAR10	Horizontal flip	RandomResizedCrop Scale to 0.4 to 1	Brightness 0.2	Contrast 0.2	Saturation 0.1	hue 0.025

\+ Normalize images : Mean: (0.4914, 0.4822, 0.4465) - Std: (0.247, 0.243, 0.261)



Source (CIFAR)
https://www.cs.toronto.edu/~kriz/cifar.html

4. Pretrain backbone, linear evaluation

<u>Training details</u>

Adam	Adam	Adam	Adam	LARS
lr = 1e-2	lr = 1e-3	lr = 1e-4	lr = 1e-5	lr : 0.2

For each optimizer
  Pretrain backbone on 20 epochs
  Perform Linear Evaluation every 4 epochs
⇒ Keep the optimizer with the best results

→

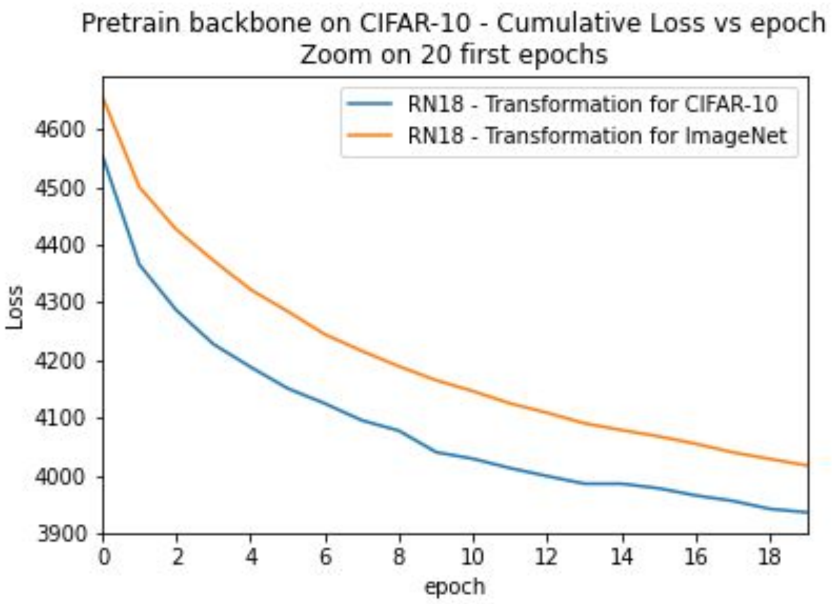
Adam
lr = 1e-4

↓

epochs : 50
batch size : 128
Adam optimizer :
  lr : 1e-4
    1e-5 after 30 epochs
train set : 40k samples

4. Pretrain blackbone

Training details - Loss function



It is not because the <u>loss decreases</u> that the model learns "good" visual representations.

epochs : 50
batch size : 128
Adam optimizer :
lr : 1e-4
1e-5 after 30 epochs
train set : 40k samples

5. Linear evaluation

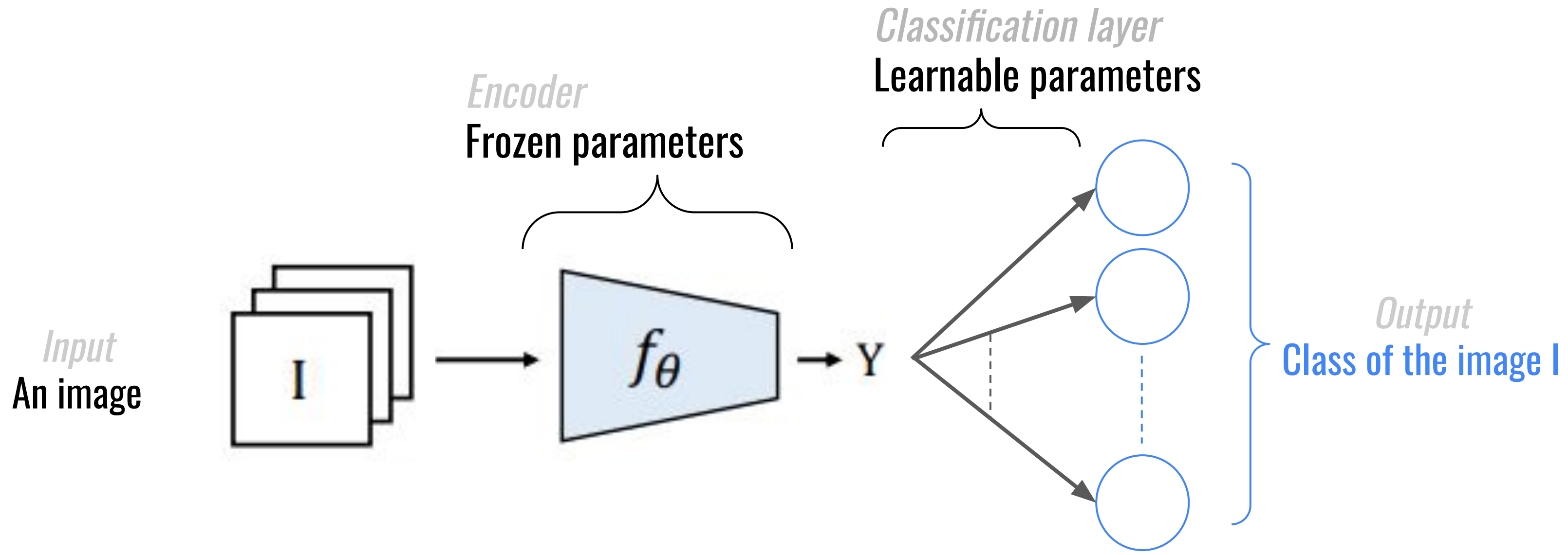


Figure 5.1 : Architecture for linear evaluation

Deng et al. (2009) linear evaluation protocol

5. Linear evaluation - CIFAR-10

Train 15k samples	Test 5k samples

Train/Test split

3 samples from CIFAR-10

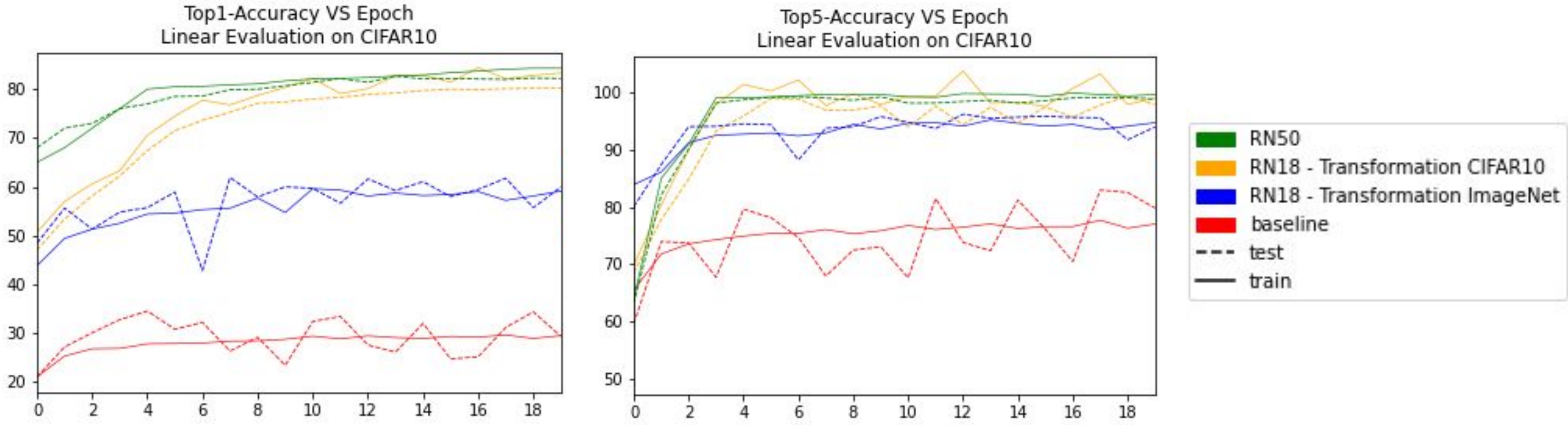
epochs : 20
batch size : 64
SVM optimizer :
- lr : 0.3
- weight decay : 1e-6
- momentum : 0.9

train : 15k samples
test : 5k samples

Source (CIFAR) : https://www.cs.toronto.edu/~kriz/cifar.html

5. Linear evaluation on CIFAR-10

Training details



Results on test

Metrics on CIFAR-10		
Model	Top-1 accuracy	Top-5 accuracy
ResNet-18 ∅	29.3%	79.8%
ResNet-18 - IN	59.9% ↑ +30.6%	93.9% ↑ +14.1%
ResNet-18 - C10	80.2% ↑ +52.0%	**99.1%** ↑ +23.0%
ResNet-50	**82.2%** ↑ +52.9%	98.8% ↑ +19.0%

5. Linear evaluation - CIFAR-100

Train 45k samples	Test 15k samples

Train/Test split

3 samples from CIFAR-100

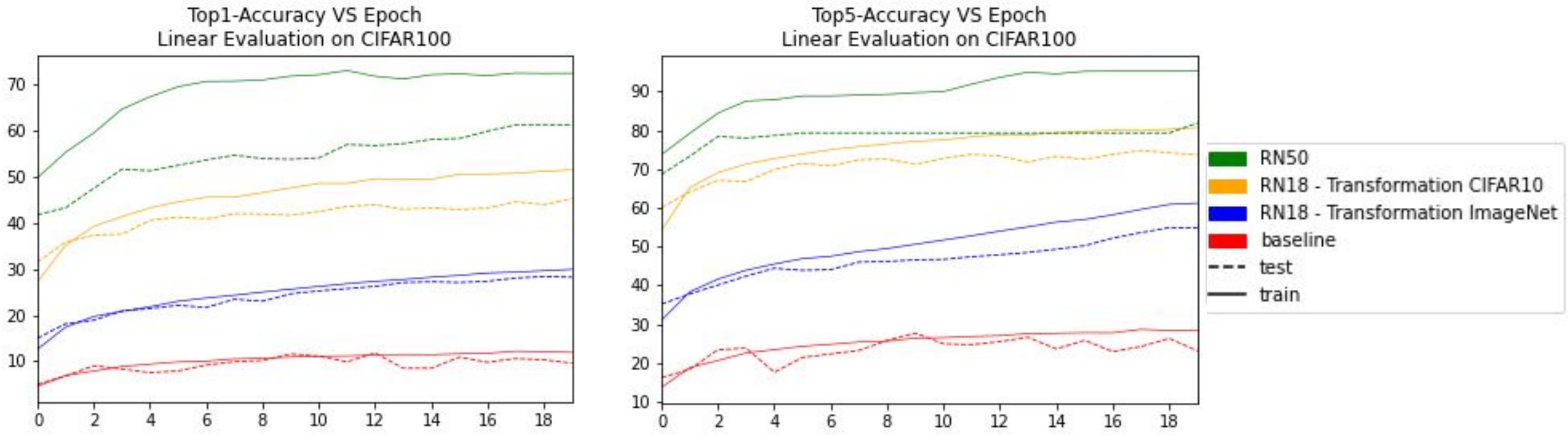
epochs : 20
batch size : 64
SVM optimizer :
- lr : 0.3
- weight decay : 1e-6
- momentum : 0.9

train : 45k samples
test : 15k samples

Source (CIFAR) : https://www.cs.toronto.edu/~kriz/cifar.html

5. Linear evaluation on CIFAR-100

Training details



Results on test

Metrics on CIFAR-100		
Model	Top-1 accuracy	Top-5 accuracy
ResNet-18 ∅	11.7%	25.4%
ResNet-18 - IN	28.2% ↑ +16.5%	54.8% ↑ +29.4%
ResNet-18 - C10	45.2% ↑ +33.5%	74.9% ↑ +42.9%
ResNet-50	**54.6%** ↑ +49.5%	**81.9%** ↑ +56.5%

6. Qualitative results

			VICReg ResNet18	VICReg ResNet50
banana			0.172	0.246
animal			0.004	0.001
fox			0.091	0.120
guacamole			0.039	0.039

4 examples of representative loss obtained with my ResNet18 and ResNet50 from the paper

ImageNet https://www.image-net.org/

6. Qualitative results

			VIC Reg ResNet 18	VIC Reg ResNet 50
banana			0.297	0.201
animal			0.715	1.956
fox			0.714	0.519
guacamole			0.412	0.317

4 examples of representative loss obtained with my ResNet18 and ResNet50 from the paper

ImageNet https://www.image-net.org/

6. Conclusion

- Handling the VICReg method and code
- Application of data augmentation methods
- Pretrain backbone ResNet-18 on Cifar-10
- Linear evaluation of backbones on Cifar-10/100
- Comparison of the results obtained

- Improving the backbone pretrain
  - ⇒ Hyper-parameters
  - ⇒ Data augmentation

- Fine-grain classification

Code base : VICReg Github

https://github.com/facebookresearch/vicreg

Implementation/Experimentation for :

- Data Loader
- Data Augmentation
- Training pipeline
- Qualitative results

Main references

A Simple Framework for Contrastive Learning of Visual Representations, Chen et al, ICML 2020.

VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning, Bardes et al, ICLR 2022