VICREG: VARIANCE-INVARIANCE-COVARIANCE REGULARIZATION FOR SELF-SUPERVISED LEARNING

Adrien Bardes, Jean Ponce, Yann LeCun

ICLR 2022

DMAIS@CAU

Yeongon Kim

INDEX

- Introduction
- Previous Work
- VICReg
- Experiments
- Conclusion

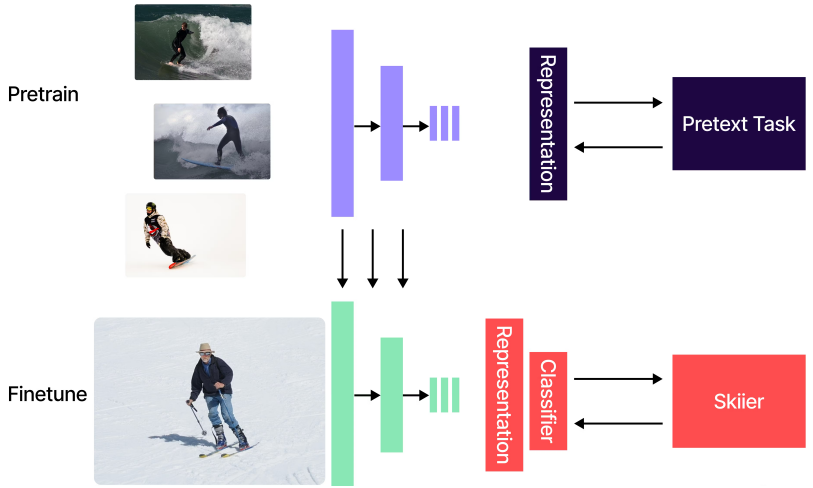
Introduction

Self-Supervised Learning

- Learning meaningful representations from unlabeled data
- Achieves supervised-level performance on downstream tasks
- Pretrain with unlabeled data, Downstream task

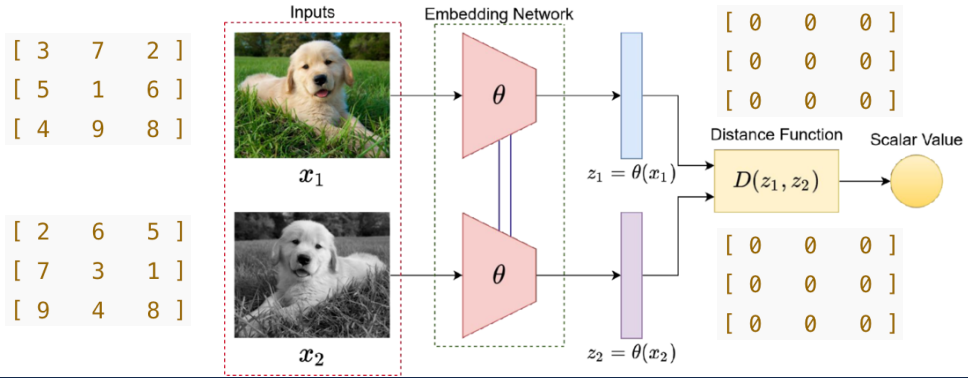
Comparison with other methods

- Supervised Learning
  - Learns from labeled data
  - Directly optimizes for the final target
- Unsupervised Learning
  - Discovers hidden patterns or structures
  - Clustering, Dimensionality reduction



Introduction

- **Joint Embedding Architecture**
  - Bringing different views of the same data closer together
  - Used in recent self-supervised learning
- **Collapse**
  - Outputs become identical regardless of the input
  - Not in supervised learning, labels force the model to separate different inputs
  - **Preventing collapse is the core focus of most self-supervised learning methods**



Previous Work

- **Contrastive learning**
  - Bring similar samples closer, push dissimilar ones apart
  - Relies on a large number of negative samples, which requires a large batch size
  - MoCo, SimCLR

- **Clustering methods**
  - Cluster similar data points and treat the cluster assignments as pseudo-labels
  - Requires an expensive clustering phase and still requires a lot of negative comparison
  - DeepCluster, SwAV

Previous Work

- **Distillation methods**
  - Use asymmetric networks and architectural tricks to prevent collapse
  - Use only positive samples, but no clear understanding how they avoid collapse
  - BYOL, SimSiam, OBoW

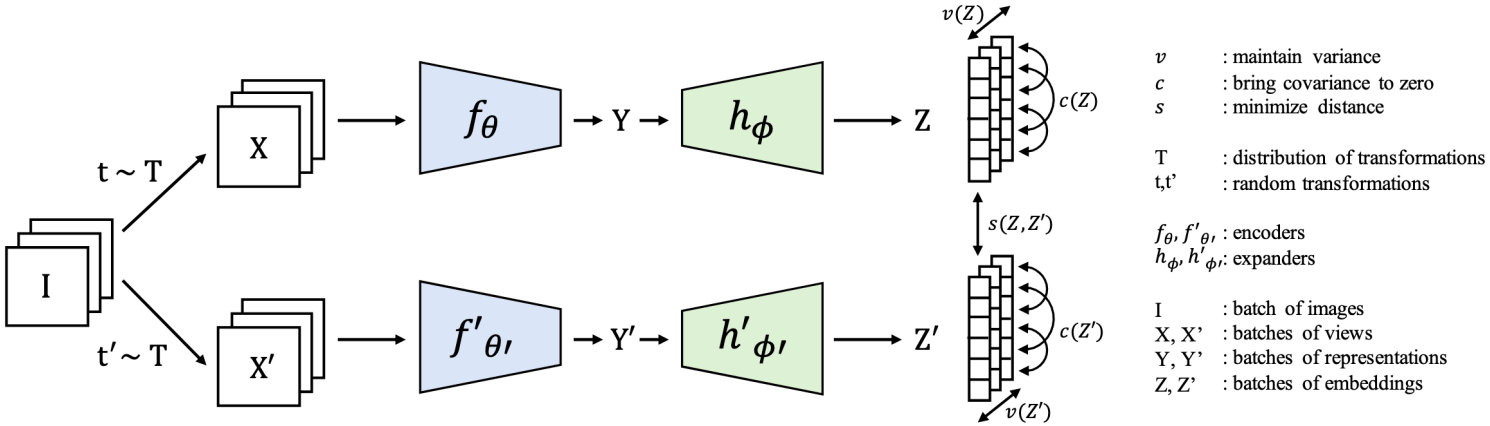
- **Information maximization methods**
  - Prevent collapse by maximizing information through feature decorrelation
  - Still needs batch-wise nor feature-wise normalization
  - W-MSE, Barlow Twins

➡ **Our goal is to design a method that is explainable and free from complex tricks**

VICReg

❖ Architecture of VICReg

- Augmented views, x and x′, are generated from the input image I
- Encoder generates representations, and the expander expands them for loss computation
- Simple, explainable regularization terms without complex tricks



$$\ell(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]$$

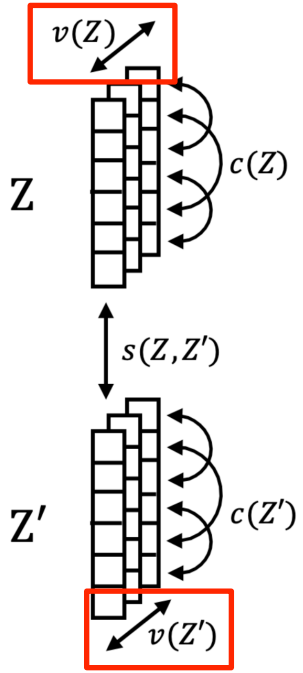
VICReg

❖ Variance Regularization

- Ensures minimum variance per feature via hinge loss
- Maintains meaningful information in each feature
- Preventing informational collapse

$$v(Z) = \frac{1}{d}\sum_{j=1}^{d} \max(0, \gamma - S(z^j, \epsilon))$$

$$S(x, \epsilon) = \sqrt{\mathrm{Var}(x) + \epsilon}$$



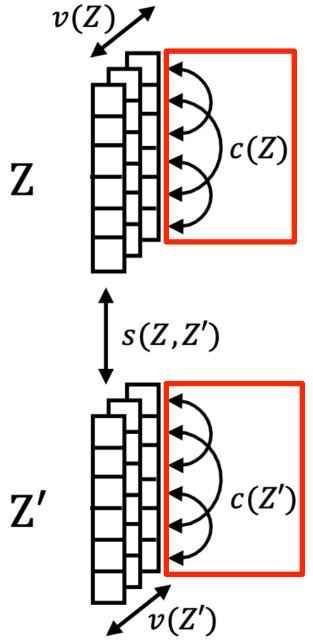
VICReg

❖ Covariance Regularization

- Reducing feature correlation by minimizing off-diagonal covariance
- Feature dimensions capture diverse info, enhancing representation
- Preventing informational collapse

$$C(Z) = \frac{1}{n-1}\sum_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T, \quad \text{where} \quad \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$$

$$c(Z) = \frac{1}{d}\sum_{i \neq j}[C(Z)]_{i,j}^2$$

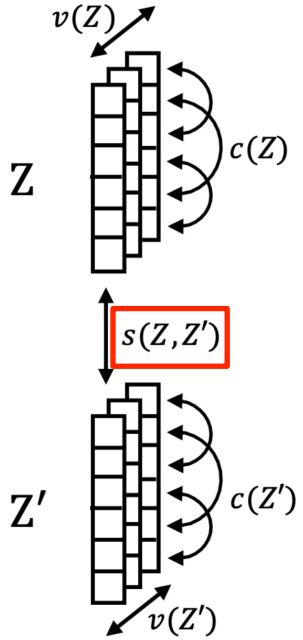


VICReg

❖ Invariance Regularization

- Produce similar representations for different views of the same image
- By minimizing the distance between embeddings
- The core idea of joint embedding

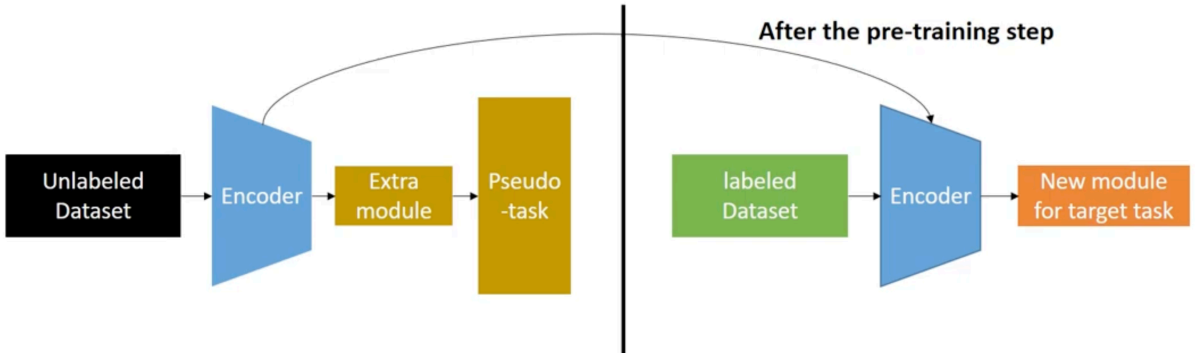
$$s(Z, Z') = \frac{1}{n} \sum_i \|z_i - z'_i\|_2^2$$



Experiments

❖ Training pipeline of Self-Supervised Learning

- Pre-training: Train the encoder using unlabeled data by solving a pseudo-task
- Downstream task: Same encoder is reused with a task-specific head
  - ResNet-50 backbone pretrained with ImageNet



Experiments

❖ Evaluation on ImageNet

Method	Linear		Semi-supervised			
	Top-1	Top-5	Top-1		Top-5	
			1%	10%	1%	10%
Supervised	76.5	-	25.4	56.4	48.4	80.4
MoCo He et al. (2020)	60.6	-	-	-	-	-
PIRL Misra & Maaten (2020)	63.6	-	-	-	57.2	83.8
CPC v2 Hénaff et al. (2019)	63.8	-	-	-	-	-
CMC Tian et al. (2019)	66.2	-	-	-	-	-
SimCLR Chen et al. (2020a)	69.3	89.0	48.3	65.6	75.5	87.8
MoCo v2 Chen et al. (2020c)	71.1	-	-	-	-	-
SimSiam Chen & He (2020)	71.3	-	-	-	-	-
SwAV Caron et al. (2020)	71.8	-	-	-	-	-
InfoMin Aug Tian et al. (2020)	73.0	91.1	-	-	-	-
OBoW Gidaris et al. (2021)	73.8	-	-	-	82.9	90.7
BYOL Grill et al. (2020)	74.3	91.6	53.2	68.8	78.4	89.0
SwAV (w/ multi-crop) Caron et al. (2020)	75.3	-	53.9	70.2	78.5	89.9
Barlow Twins Zbontar et al. (2021)	73.2	91.0	55.0	69.7	79.2	89.3
VICReg (ours)	73.2	91.1	54.8	69.5	79.4	89.5

- **Linear classification: frozen representations of the ResNet-50 backbone**
- **Semi-supervised: fine-tuned with a linear classifier**

Experiments

❖ Transfer to Other Downstream Tasks

Method	Linear Classification			Object Detection		
	Places205	VOC07	iNat18	VOC07+12	COCO det	COCO seg
Supervised	53.2	87.5	46.7	81.3	39.0	35.4
MoCo He et al. (2020)	46.9	79.8	31.5	-	-	-
PIRL Misra & Maaten (2020)	49.8	81.1	34.1	-	-	-
SimCLR Chen et al. (2020a)	52.5	85.5	37.2	-	-	-
MoCo v2 Chen et al. (2020c)	51.8	86.4	38.6	82.5	39.8	36.1
SimSiam Chen & He (2020)	-	-	-	82.4	-	-
BYOL Grill et al. (2020)	54.0	86.6	47.6	-	40.4†	37.0†
SwAV (m-c) Caron et al. (2020)	56.7	88.9	48.6	82.6	41.6	37.8
OBoW Gidaris et al. (2021)	56.8	89.3	-	82.9	-	-
Barlow Twins Grill et al. (2020)	54.1	86.2	46.5	82.6	40.0†	36.7†
VICReg (ours)	54.3	86.6	47.0	82.4	39.4	36.4

- **Linear classification: frozen representations of the ResNet-50 backbone**
- **Detection: classify and locate multiple objects in a single image**
- **Segmentation: predicts the exact pixel-wise boundaries of each object**

Experiments

❖ Flexibility of VICReg

	SW R50	DW R50	DA R50/R101	DA R50/ViT-S
BYOL	69.3	✗	✗	✗
SimCLR	64.4	63.1	63.9	63.5
Barlow Twins	68.7	64.2	65.3	63.9
VICReg	68.6	66.5	68.1	66.2

- **ResNet-50 backbone trained with ImageNet**

Method	Image-to-text			Text-to-Image		
	R@1	R@5	R@10	R@1	R@5	R@10
Contrastive (VSE++)	30.3	59.4	72.4	41.3	71.1	81.2
Barlow Twins	31.4	60,4	75.1	42.9	74.0	83.5
VICReg	33.6	62.7	77.9	45.2	76.1	84.2

- **Image encoder: ResNet-152 trained with images of MS-COCO**
- **Text encoder: Word Embedding + GRU trained with texts of MS-COCO**

Experiments

❖ Flexibility of VICReg

Method	ME	SG	PR	BN	No Reg	Var Reg	Var/Cov Reg
BYOL	✓	✓	✓	✓	69.3†	70.2	69.5
SimSiam		✓	✓	✓	67.9†	68.1	67.6
SimSiam		✓	✓		35.1	67.3	67.1
SimSiam		✓			collapse	56.8	66.1
VICReg			✓		collapse	56.2	67.3
VICReg			✓	✓	collapse	57.1	68.7
VICReg				✓	collapse	57.5	68.6†
VICReg					collapse	56.5	67.4

- ME: Momentum Encoder, SG: stop-gradient
- PR: predictor, BN: batch normalization
- No Reg: only invariance regularization term
- Unmodified original setups are marked by a †

Conclusion

✓ **Simple and explainable design leads to greater adaptability**

- **State-of-the-art performance with interpretability and no additional tricks**
- **Interpretability leads to flexibility, allowing it to be applied to various tasks**
- **Costly covariance matrix computation is left for future work**

Appendix

- **Running time and peak memory**

Method	time / 100 epochs	peak memory / GPU	Top-1 accuracy (%)
SwAV	9h	9.5G	71.8
SwAV (w/ multi-crop)	13h	12.9G	75.3
BYOL	10h	14.6G	74.3
Barlow Twins	12h	11.3G	73.2
VICReg	11h	11.3G	73.2