# Self-Supervised Learning Methods for Information Maximization

**2023. 03. 03**

**발표자: 허종국**

# 발표자 소개

❖ 이름 : 허종국 (Jong Kook, Heo)

- Data Mining & Quality Analytics Lab
- Ph.D. Student (2021.03~)
- 지도 교수 : 김성범 교수님

❖ 관심 연구 분야

- Deep Reinforcement Learning
- Graph Neural Networks
- Self-Supervised Learning

❖ 연락망


- E-mail : hjkso1406@korea.ac.kr

# 목 차

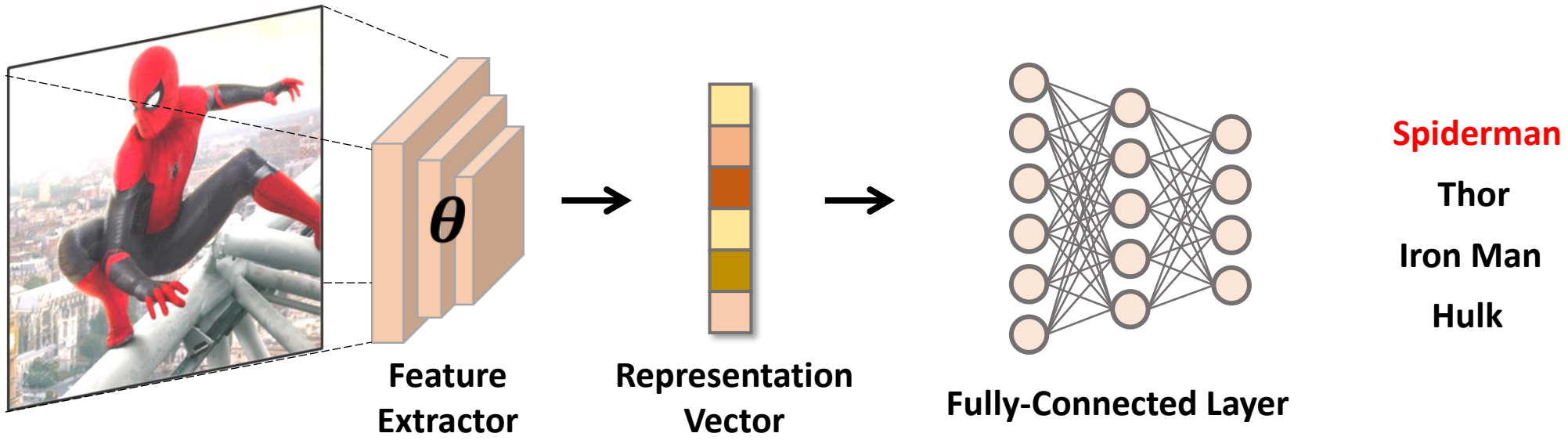


# Introduction

## What is Self-Supervised Learning?

❖ Supervised Learning Framework

- Feature Extractor : 입력 데이터로부터 중요한 정보를 요약, 추출
- Fully-Connected Layer : 요약정보로부터 목적에 알맞는 타겟값을 예측

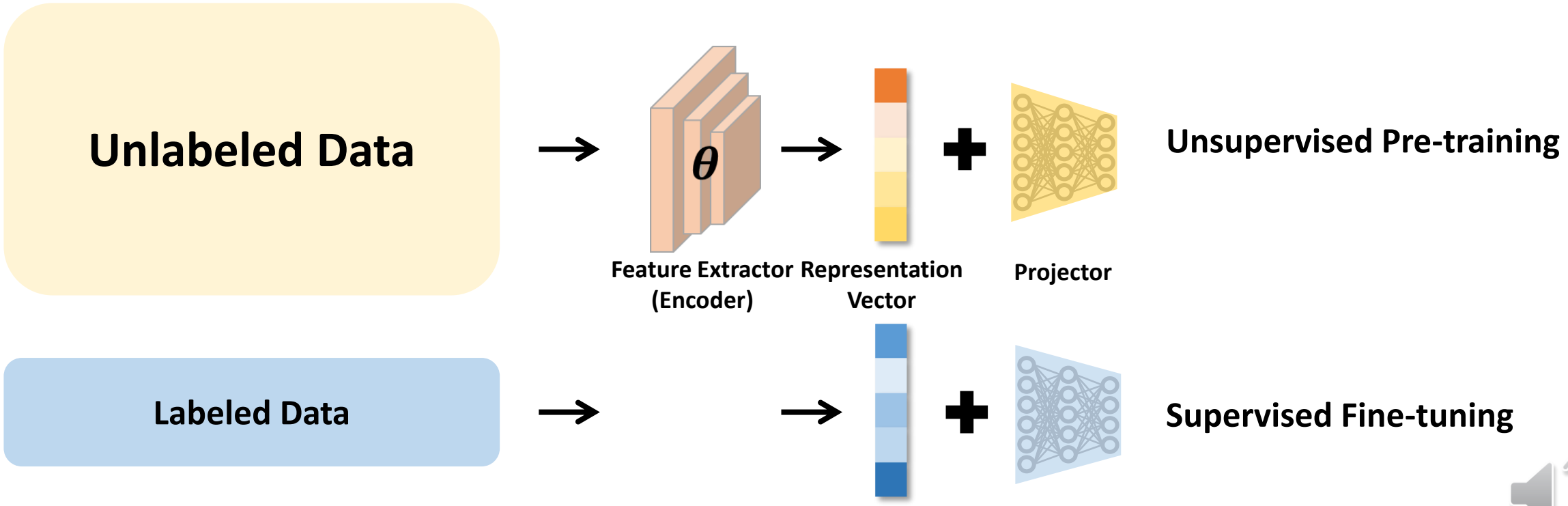


# Introduction

## What is Self-Supervised Learning?

❖ Self-Supervised Learning Framework

- Supervised Learning 을 위한 Labeled Data 를 구축하는 것은 시간/비용적 소모가 큼
- 대량의 Unlabeled Data 를 활용하여 Feature Extractor 를 학습할 수 없을까?

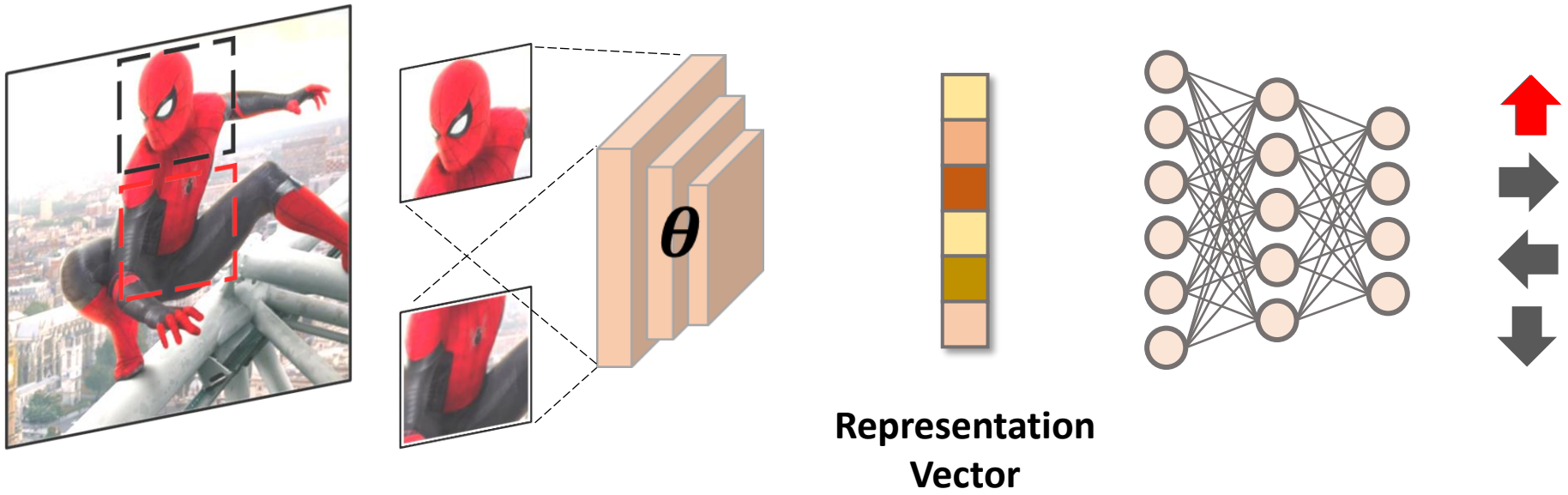


# Previous SSL Methods

## Pretext Task Methods

❖ Pretext Task Method : Context Prediction

- 사용자가 Task-agnostic 한 Label을 만들어 사전 학습을 수행
- Label : 특정 패치에 대한 다른 패치의 상대적인 위치




Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision (pp. 1422-1430).

# Previous SSL Methods

## Pretext Task Methods

- ❖ Others
  - Self-Supervised Representation Learning : http://dmqa.korea.ac.kr/activity/seminar/284
    - ✓ Exemplar, Jigsaw Puzzle, Image Colorization, Count, Rotation etc.

**1. Pretext task 방식은 휴리스틱(heuristic)에 기반하기 때문에 Representation Vector의 일반성이 제한됨**

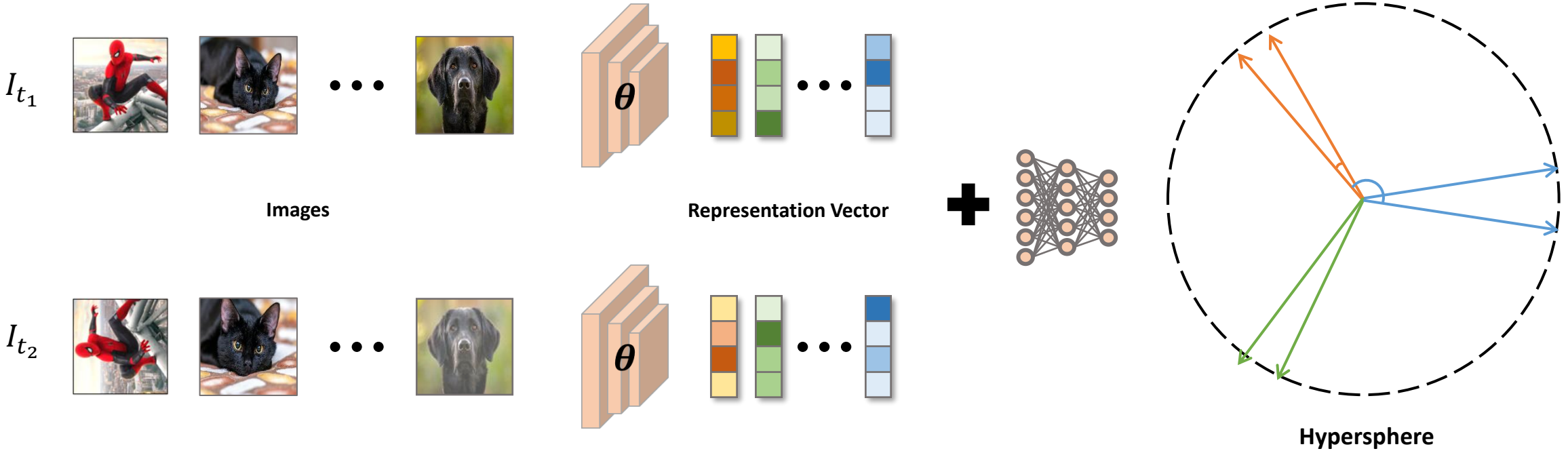
**→ Heuristic에 상관없이 Instance의 표현을 잘 뽑을 수 없을까?**


Doersch, C., Gupta, A., & Efros, A. A. (2015). Unsupervised visual representation learning by context prediction. In Proceedings of the IEEE international conference on computer vision (pp. 1422-1430).

# Previous SSL Methods

## Contrastive Learning Methods

- Contrastive Learning Method : SimCLR
  - Positive Pair : 같은 이미지에서 서로 다른 데이터 증강 기법을 적용한 샘플
  - Negative Pair : 서로 다른 이미지로부터 데이터 증강 기법을 적용한 샘플
  - Positive Pair 는 가깝게, Negative Pair는 멀어지도록 학습(Cosine Similarity)




Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

# Previous SSL Methods

1. Contrastive Learning은 대량의 Negative Sample과 비교할 때 뛰어난 성능을 보임

→ Negative Sample 없이 뛰어난 Representation을 학습할 수는 없을까?

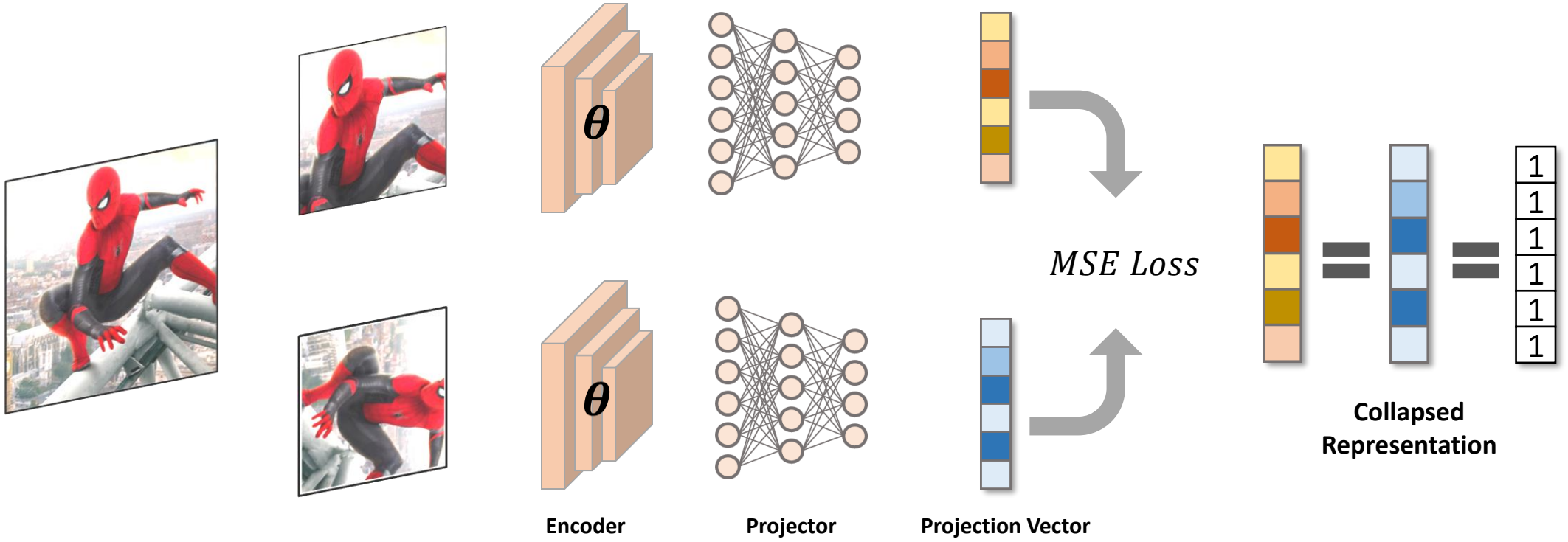
2. 같은 범주/의미론적 유사성을 가지더라도 다른 객체에 대해 Negative로 선정하기 때문에 False Negative 문제가 발생

→ 의미론적 유사성을 고려하기 위해 Clustering 을 활용할 수 있을까?

# Previous SSL Methods

## Distillation Methods

- SSL exploiting only positives
  - Positive Sample만을 활용하여 학습할 경우 이미지 특징과 연관 없는 Constant Vector로 수렴할 수 있음(Collapse)
  - Distillation 기반 방법론은 Negative Sample 없이 Collapsed Representation 을 해결하고자 함



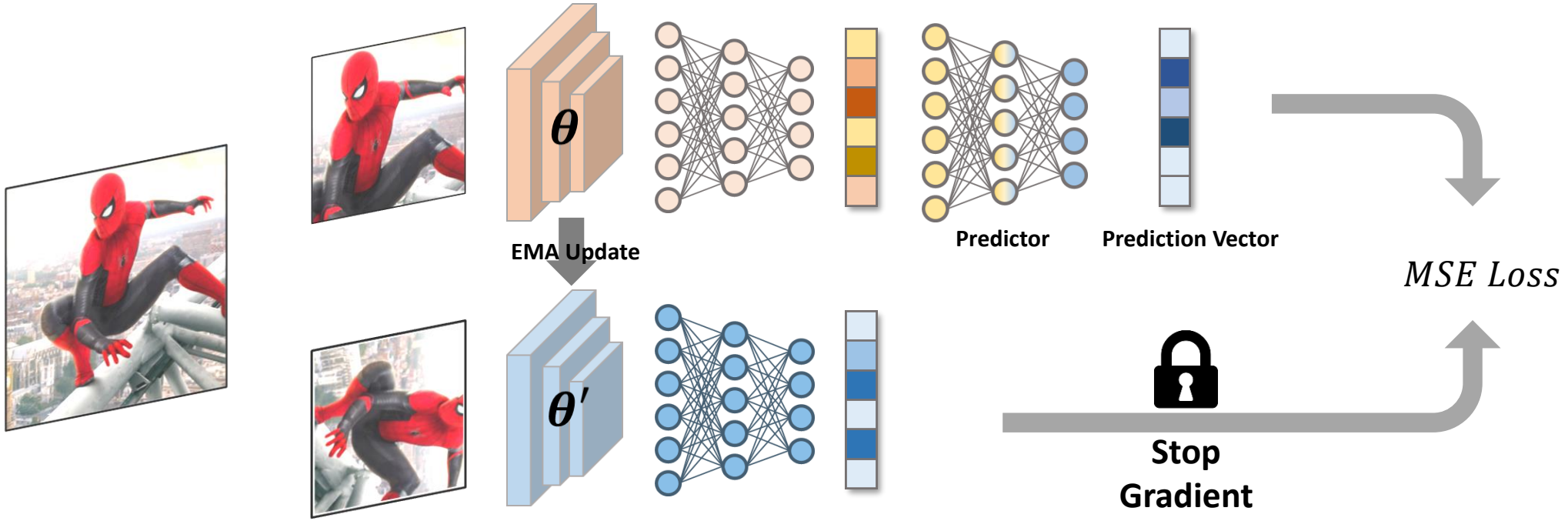

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.

# Previous SSL Methods

## Distillation Methods

❖ Distillation Method : BYOL

- Online Network 와 Target Network에서 Projection Vector 를 추출
- Online Network 의 Embedding Vector 로부터 Target Network 의 Representation Vector 를 예측
- Asymmetric Architecture/Stop Gradient 등을 활용하여 Collapsed Representation 해결



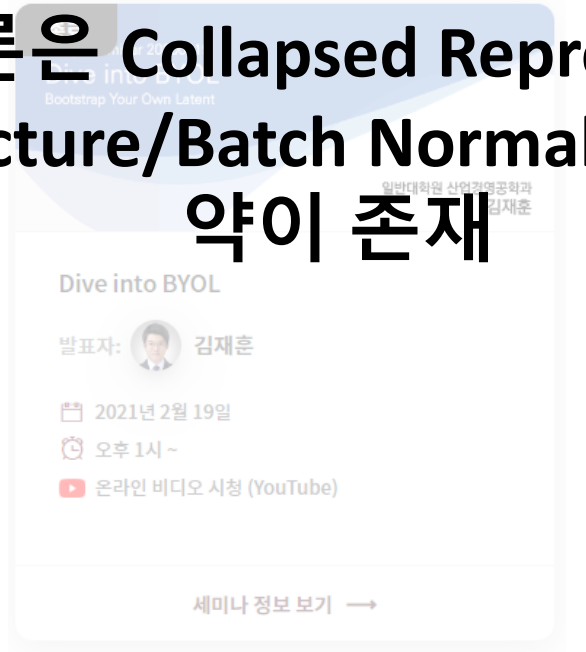

Grill, J. B., Strub, F., Altché, F., Tallec, C., Richemond, P., Buchatskaya, E., ... & Valko, M. (2020). Bootstrap your own latent-a new approach to self-supervised learning. Advances in neural information processing systems, 33, 21271-21284.

# Previous SSL Methods

Distillation Methods

- ❖ Details of BYOL
  - Dive into BYOL : http://dmqa.korea.ac.kr/activity/seminar/310
    - ✓ Motivation/Details/Experiment Results of BYOL

1. **Distillation 기반 방법론은 Collapsed Representation을 해결하기 위해 Asymmetric Architecture/Batch Normalization 등 모델 구조적 제약이 존재**



# Previous SSL Methods

1. Contrastive Learning은 대량의 Negative Sample과 비교할 때 뛰어난 성능을 보임

→ Negative Sample 없이 뛰어난 Representation을 학습할 수는 없을까?

2. 같은 범주/의미론적 유사성을 가지더라도 다른 객체에 대해 Negative로 선정하기 때문에 False Negative 문제가 발생

→ 의미론적 유사성을 고려하기 위해 Clustering 을 활용할 수 있을까?

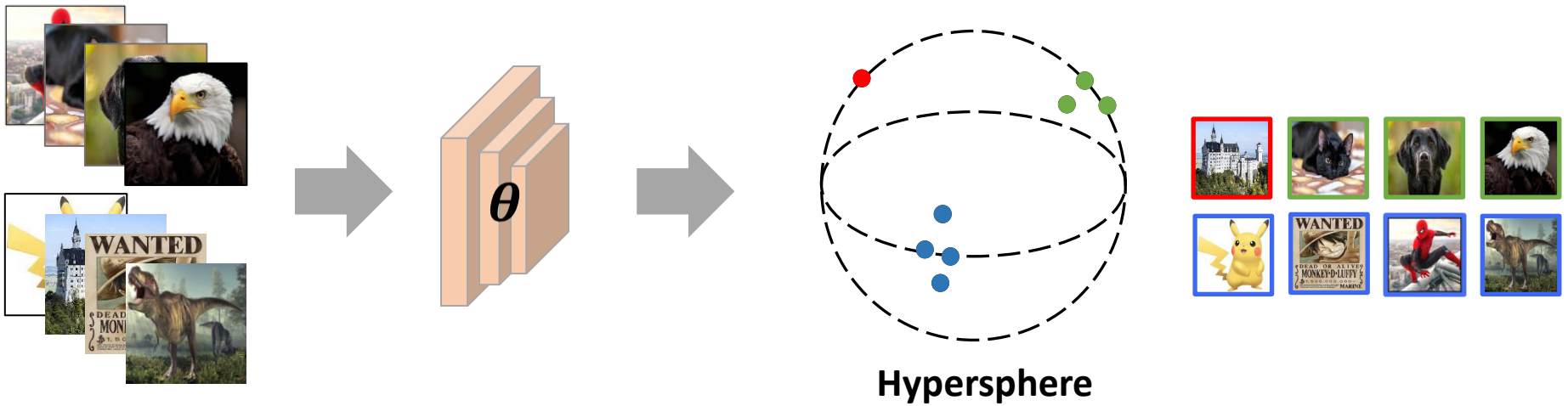
Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020, November). A simple framework for contrastive learning of visual representations. In International conference on machine learning (pp. 1597-1607). PMLR.

# Previous SSL Methods

## Clustering Methods

❖ Clustering Method : Deep Cluster

- Step1 : 추출된 Feature Vector에 대해 k-means Clustering 수행
- Step2 : 할당된 Cluster 를 pseudo-label로 지정하여 Classification 수행
- Step1 & Step2 반복



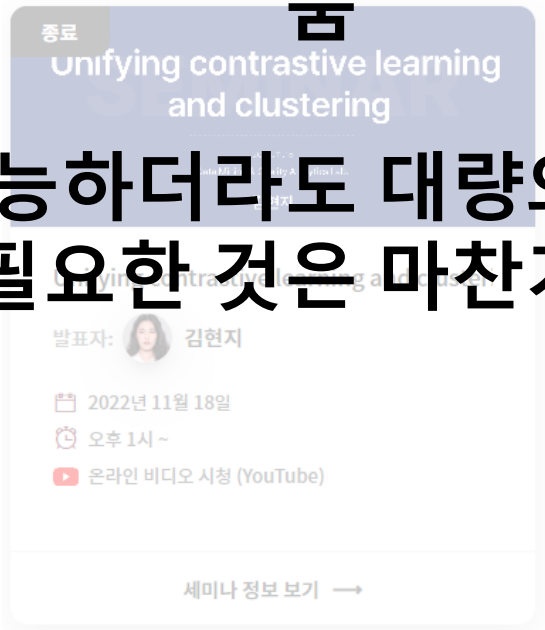

Caron, M., Bojanowski, P., Joulin, A., & Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In Proceedings of the European conference on computer vision (ECCV) (pp. 132-149).

# Previous SSL Methods

Clustering Methods

- ❖ Others
  - Unifying Contrastive Learning and Clustering : http://dmqa.korea.ac.kr/activity/seminar/386
    - ✓ DeepCluster, PCL, SwAV

1. **Clustering Phase가 비동기적으로 수행되기 때문에 Scale Up이 어려움**




2. **Online Clustering이 가능하더라도 대량의 Negative Comparison이 필요한 것은 마찬가지**

# Previous SSL Methods

Summary

**Pretext Task**

Unlabeled Data에서 Task Agnostic한 pseudo-label을 생성하여 학습

**Contrastive Learning**

동일 객체(instance) 여부에 따라 positive/negative를 정의하여 Cosine Similarity로 학습

**Distillation**

Negative Sample 없이 Positive Sample 만을 활용하여 학습

**Clustering**

클러스터링을 활용하여 Semantic Similarity를 포착, Instance Discrimination 단점 극복

# Information Maximization Methods



## ❖ Information Maximization Methods

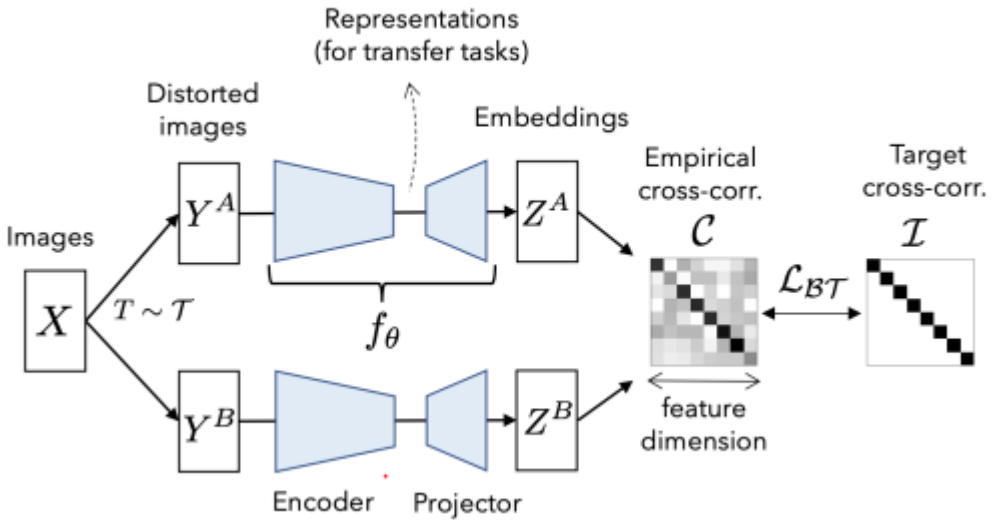
- 동일한 이미지에서 나온 상호 정보를 최대화하도록 학습하는 방법론
  - ✓ **Barlow Twins** : Cross-Correlation 을 최소화하여 차원 간정보의 중복(Redundancy Reduction)을 최소화
  - ✓ **W-MSE** : Whitening Transform 을 통해 Negative Sample 없이 Collapse를 방지
  - ✓ **VICReg** : 3가지 규제화(Regularization) 손실 함수를 통해 임베딩 벡터의 정보 함축 능력을 증가, Collapse 방지

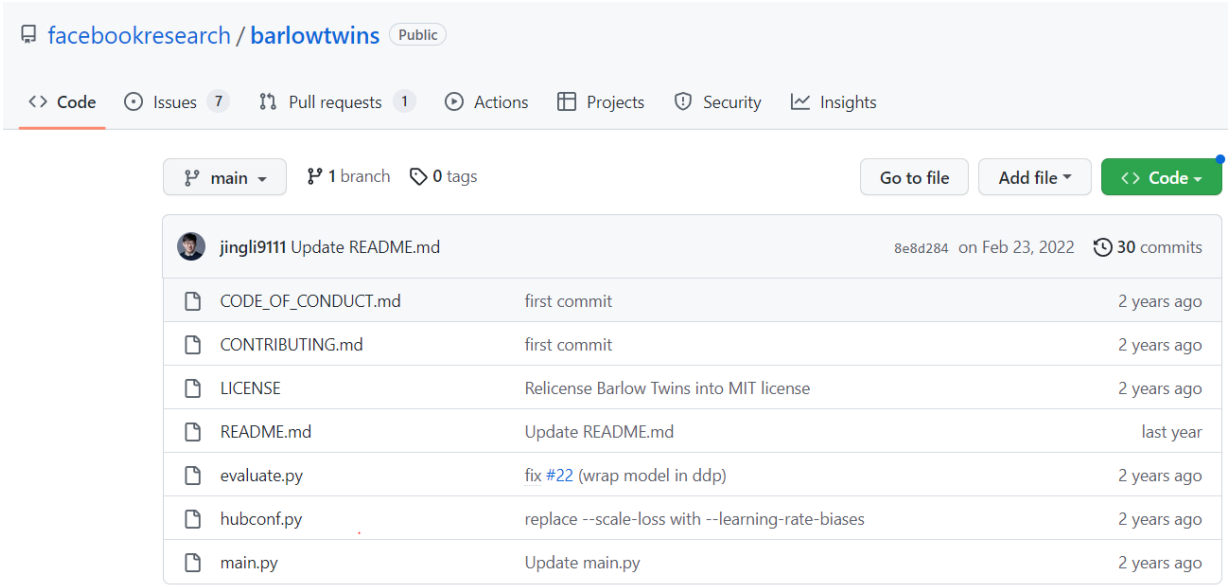
- 기존 방법론들 대비 아래와 같은 장점을 가짐
  - ✓ **vs Pretext Task** : 특정 태스크에 치우치지 않은 일반화된 표현 학습 가능
  - ✓ **vs Contrastive Learning** : Negative Sample로 인한 메모리 제약이 없음
  - ✓ **vs Clustering** : Off-line Clustering, Negative Sample 필요 없음
  - ✓ **vs Distillation** : Assymetric Architecutre, Batch Norm 등 구조적 제약 없음

# Information Maximization Methods

## Barlow Twins

- Barlow Twins : Self-Supervised Learning via Redundancy Reduction(2021, ICML)
  - 2023년 2월 기준 947회 인용
  - Cross-Correlation Matrix를 활용하여 Representation Vector의 정보 중복을 감소(Redundancy Reduction)
  - https://github.com/facebookresearch/barlowtwins



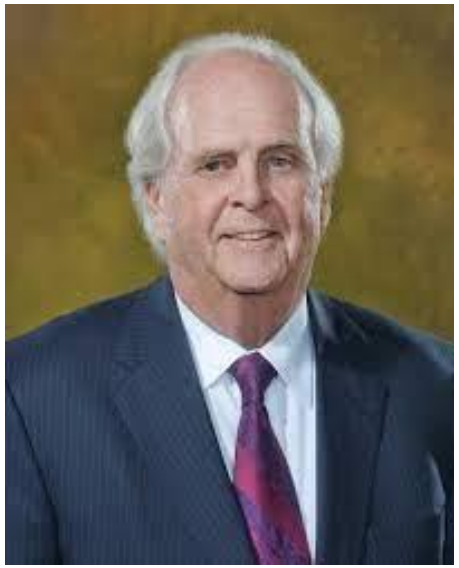




Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

# Information Maximization Methods

## Barlow Twins

- Why does it called "Barlow Twins??"
  - **David H. Barlow** : "신호 처리의 목표는 굉장히 중복된 정보를 독립된 요소의 코드로 인코딩하는 것이다"
  - **Redundancy Reduction** : 제한된 크기의 Representation Vector에서 정보의 중복을 감소 시키는 것



**David H. Barlow**

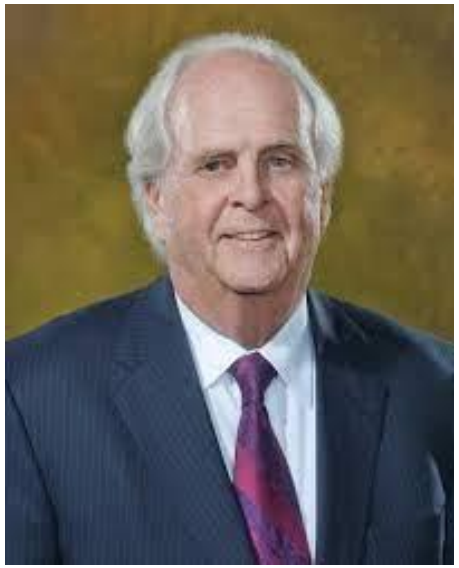
"The goal of sensory processing is to recode highly redundant sensory inputs into a factorial code, a code with statistically independent components"

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

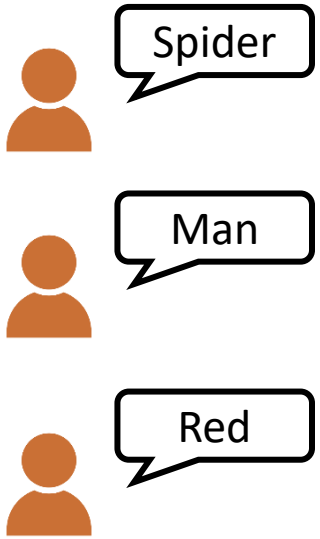
# Information Maximization Methods

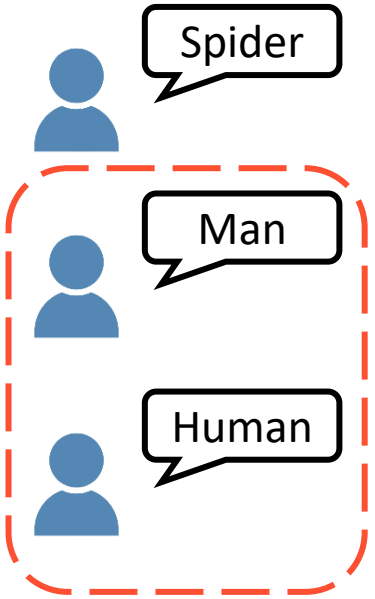
## Barlow Twins

- Why does it called "Barlow Twins??"
  - **David H. Barlow** : "신호 처리의 목표는 굉장히 중복된 정보를 독립된 요소의 코드로 인코딩하는 것이다"
  - **Redundancy Reduction** : 제한된 크기의 Representation Vector에서 정보의 중복을 감소 시키는 것



**David H. Barlow**



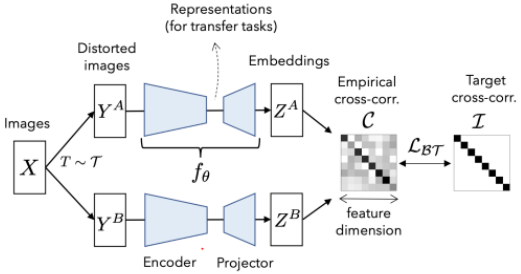




**Highly Correlated**

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

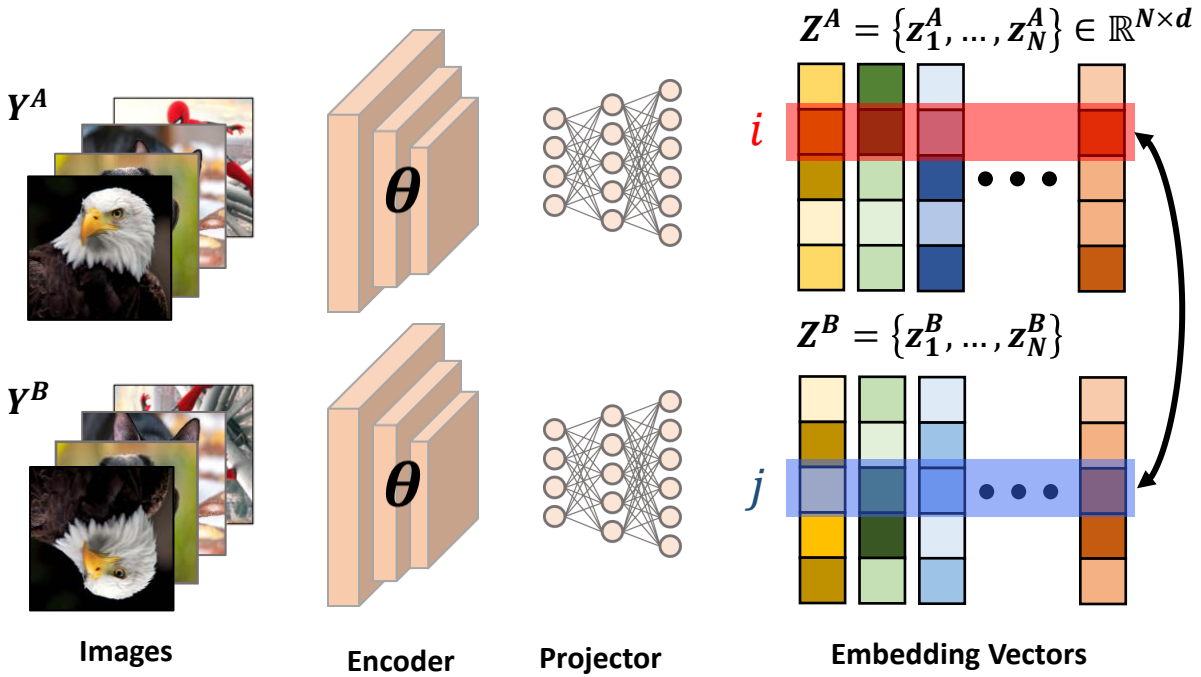
# Information Maximization Methods

## Barlow Twins

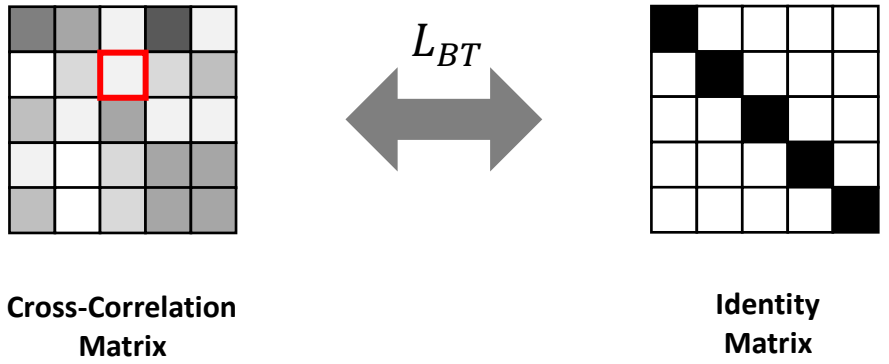



### ❖ Method

- Cross-Correlation Matrix가 Identity Matrix에 근접하도록 학습
  - ✓ **Diagonal Component** : 이미지 증강 기법에 상관 없이 동일한 정보를 인코딩(Invariance)
  - ✓ **Off-Diagonal Component** : 임베딩 벡터의 서로 다른 요소는 독립적인 정보를 인코딩(Redundancy Reduction)




$$C_{ij} \triangleq \frac{\sum_b z_{b,i}^A z_{b,j}^B}{\sqrt{\sum_b (z_{b,i}^A)^2} \sqrt{\sum_b (z_{b,j}^B)^2}}$$

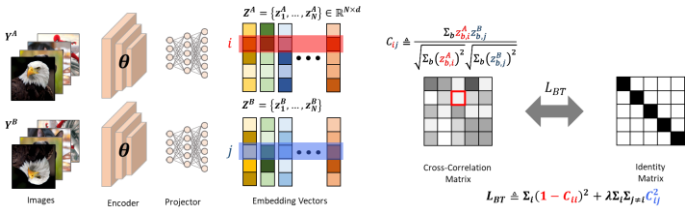



$$L_{BT} \triangleq \Sigma_i (1 - C_{ii})^2 + \lambda \Sigma_i \Sigma_{j \neq i} C_{ij}^2$$

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

# Information Maximization Methods

## Barlow Twins




❖ Main Results

- 기존 SSL 방법론들이 제안했던 트릭 없이 단순한 구조로 높은 성능을 나타냄
  - ✓ **Multi-Crop Augmentation** : SwAV
  - ✓ **Negative Samples** : MoCo, PIRL, SimCLR
  - ✓ **Momentum Encoder** : MoCo, BYOL
  - ✓ **Asymmetric Architecture** : BYOL, SimSiam

*Table 1.* **Top-1 and top-5 accuracies (in %) under linear evaluation on ImageNet**. All models use a ResNet-50 encoder. Top-3 best self-supervised methods are underlined.

| Method | Top-1 | Top-5 |
|---|---|---|
| Supervised | 76.5 | |
| MoCo | 60.6 | |
| PIRL | 63.6 | - |
| SimCLR | 69.3 | 89.0 |
| MoCo v2 | 71.1 | 90.1 |
| SimSiam | 71.3 | - |
| SwAV (w/o multi-crop) | 71.8 | - |
| BYOL | 74.3 | 91.6 |
| SwAV | 75.3 | - |
| Barlow Twins (ours) | 73.2 | 91.0 |

*Table 2.* **Semi-supervised learning on ImageNet** using 1% and 10% training examples. Results for the supervised method are from (Zhai et al., 2019). Best results are in **bold**.

| Method | Top-1 | | Top-5 | |
|---|---|---|---|---|
| | 1% | 10% | 1% | 10% |
| Supervised | 25.4 | 56.4 | 48.4 | 80.4 |
| PIRL | - | - | 57.2 | 83.8 |
| SimCLR | 48.3 | 65.6 | 75.5 | 87.8 |
| BYOL | 53.2 | 68.8 | 78.4 | 89.0 |
| SwAV | 53.9 | **70.2** | 78.5 | **89.9** |
| Barlow Twins (ours) | **55.0** | 69.7 | **79.2** | 89.3 |

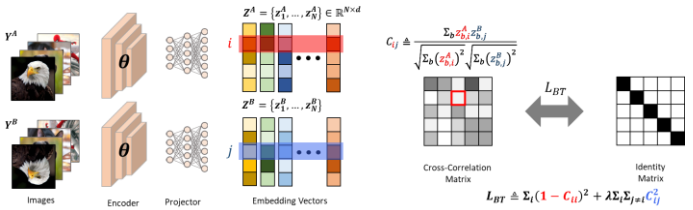
*Table 3.* **Transfer learning: image classification.** We benchmark learned representations on the image classification task by training linear classifiers on fixed features. We report top-1 accuracy on Places-205 and iNat18 datasets, and classification mAP on VOC07. Top-3 best self-supervised methods are underlined.

| Method | Places-205 | VOC07 | iNat18 |
|---|---|---|---|
| Supervised | 53.2 | 87.5 | 46.7 |
| SimCLR | 52.5 | 85.5 | 37.2 |
| MoCo-v2 | 51.8 | 86.4 | 38.6 |
| SwAV (w/o multi-crop) | 52.8 | 86.4 | 39.5 |
| SwAV | 56.7 | 88.9 | 48.6 |
| BYOL | 54.0 | 86.6 | 47.6 |
| Barlow Twins (ours) | 54.1 | 86.2 | 46.5 |

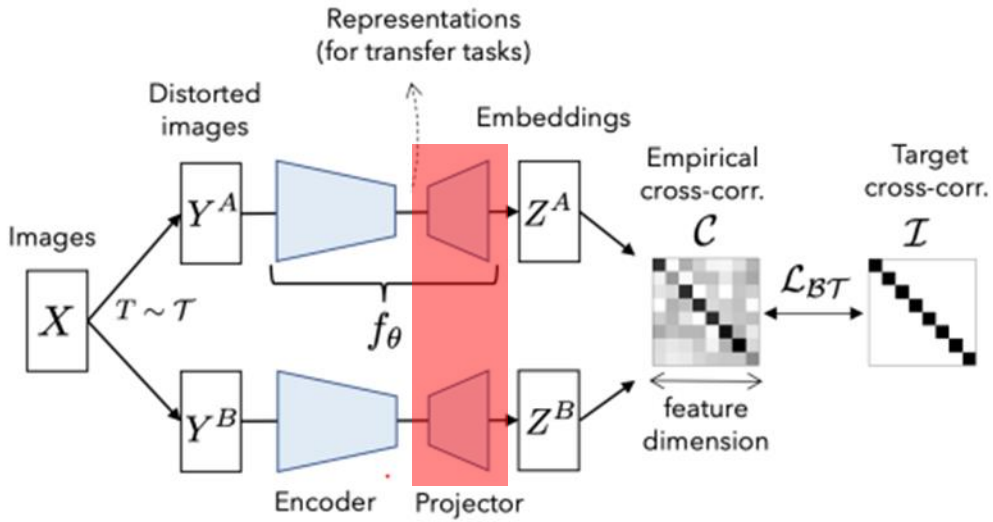
Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

# Information Maximization Methods

## Barlow Twins



- Dimensionality of Projector Network
  - Projector Network의 차원이 클 수록 성능이 향상 됨
  - 다른 SSL 방법론과 달리 Representation Vector의 차원보다 Projection Vector의 차원이 더욱 큼



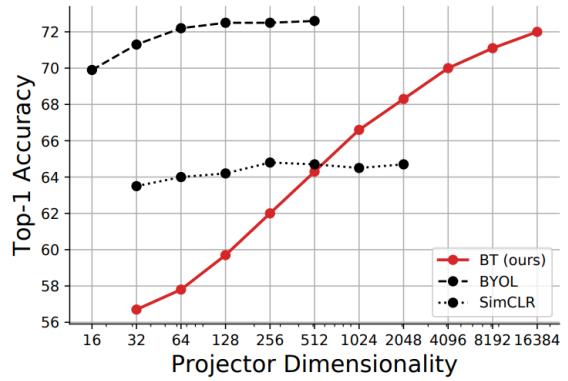



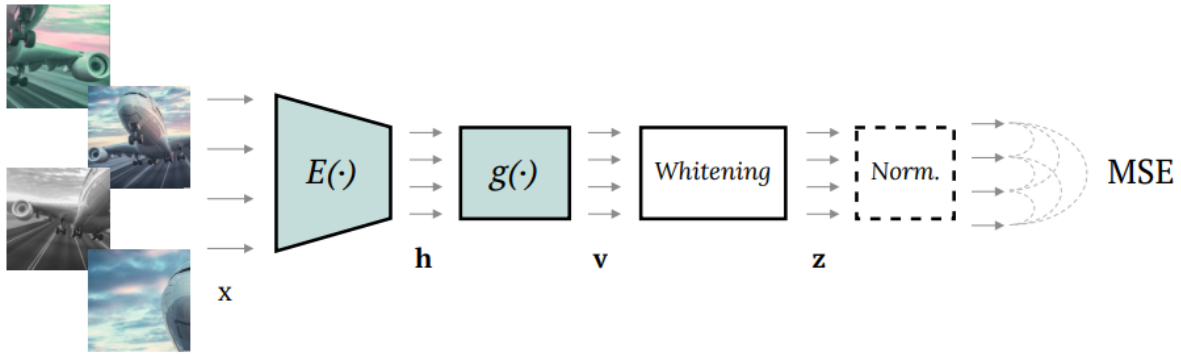

*Figure 4.* Effect of the dimensionality of the last layer of the projector network on performance. The parameter $\lambda$ is kept fix for all dimensionalities tested. Data for SIMCLR is from (Chen et al., 2020a) fig 8; Data for BYOL is from (Grill et al., 2020) Table 14b.

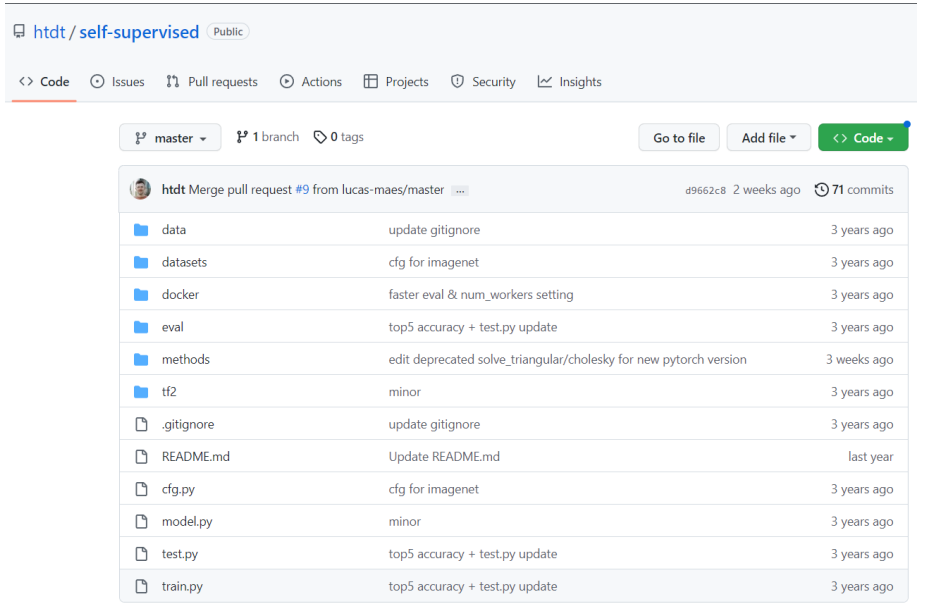

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021, July). Barlow twins: Self-supervised learning via redundancy reduction. In International Conference on Machine Learning (pp. 12310-12320). PMLR.

# Information Maximization Methods

## W-MSE

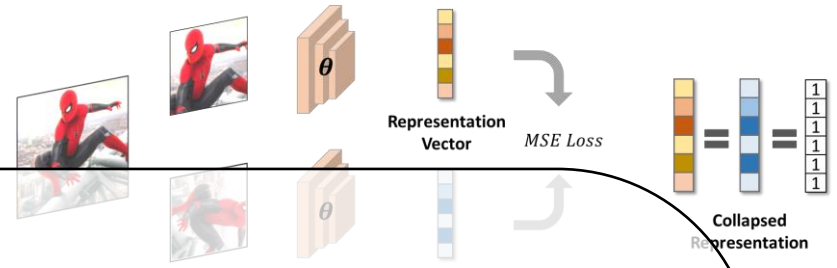
- Whitening for Self-Supervised Representation Learning(2021, ICML)
  - 2023년 2월 기준 142회 인용
  - Whitening Transformation을 통해 Negative Sample 없이 Collapse를 방지하여 데이터를 매핑
  - https://github.com/htdt/self-supervised




Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

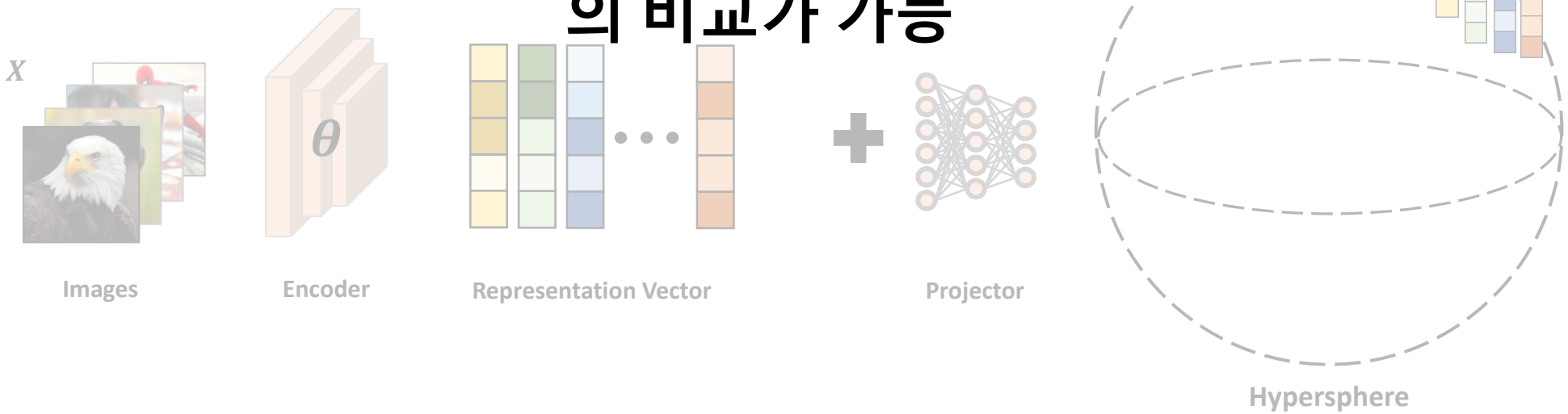
# Information Maximization Methods




## W-MSE

- What is "Whitening"??
  - Representation Vector를 Spherical Distribution으로 매핑하는 작업(Scattering)
    - ✓ 단순히 MSE만 쓸 경우 모든 Representation Vector가 데이터에 상관없이 Constant Vector로 매핑됨
    - ✓ Whitening은 임베딩 벡터를 흩뿌림으로써 기존의 Positive-Negative Instance Contrast를 대체하고자함

**Positive-Negative Contrast가 필요 없기 때문에 Multi-Positive Sample간의 비교가 가능**




Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

❖ Whitening MSE Loss

- Formulation
  - ✓ $\mathbf{v}_i, \mathbf{v}_j$ : Embedding Vectors of positive image pairs $(x_i, x_j)$

$$\min_{\theta} E\left[dist\left(\mathbf{v}_i, \mathbf{v}_j\right)\right]$$
$$s.t.\, cov(\mathbf{v}_i, \mathbf{v}_i) = cov\left(\mathbf{v}_j, \mathbf{v}_j\right) = \mathbf{I}$$

$$dist\left(\mathbf{v}_i, \mathbf{v}_j\right) = \left\| \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} - \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|_2} \right\|_2^2$$

Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

❖ Whitening MSE Loss

- Formulation
  - ✓ $d$ : $\#\ of\ positive\ examples(augmentation)$
  - ✓ $N$ : $\#\ of\ Original\ Images$
  - ✓ $V$ : $Embedding\ Vectors\ in\ a\ batch\ \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K\}, K = Nd$

$$L_{MSE}(V) = \frac{2}{Nd(d-1)} \sum_{pos(i,j)=True} dist(\mathbf{v}_i, \mathbf{v}_j)$$

$$dist(\mathbf{v}_i, \mathbf{v}_j) = \left\| \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} - \frac{\mathbf{v}_j}{\|\mathbf{v}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_i\|_2 - \|\mathbf{v}_j\|_2}$$

Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

❖ Whitening MSE Loss

- Formulation
  - ✓ $d$ : # of positive examples(augmentation)
  - ✓ $N$ : # of Original Images
  - ✓ $V$ : Embedding Vectors in a batch $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_K\}, K = Nd$

$$L_{MSE}(V) = \frac{2}{Nd(d-1)} \sum_{pos(i,j)=True} dist(\mathbf{z}_i, \mathbf{z}_j)$$

$$\mathbf{z} = Whitening(\mathbf{v}) = W_V(\mathbf{v} - \mu_V)$$

$$\mu_V = \frac{1}{K}\sum_k \mathbf{v}_k$$

$$\Sigma_V = \frac{1}{K-1}\sum_k (\mathbf{v}_k - \mu_V)(\mathbf{v}_k - \mu_V)^T$$

$$W_V^T W_V = \Sigma_V^{-1}$$

$$dist(\mathbf{z}_i, \mathbf{z}_j) = \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\|_2 - \|\mathbf{z}_j\|_2}$$

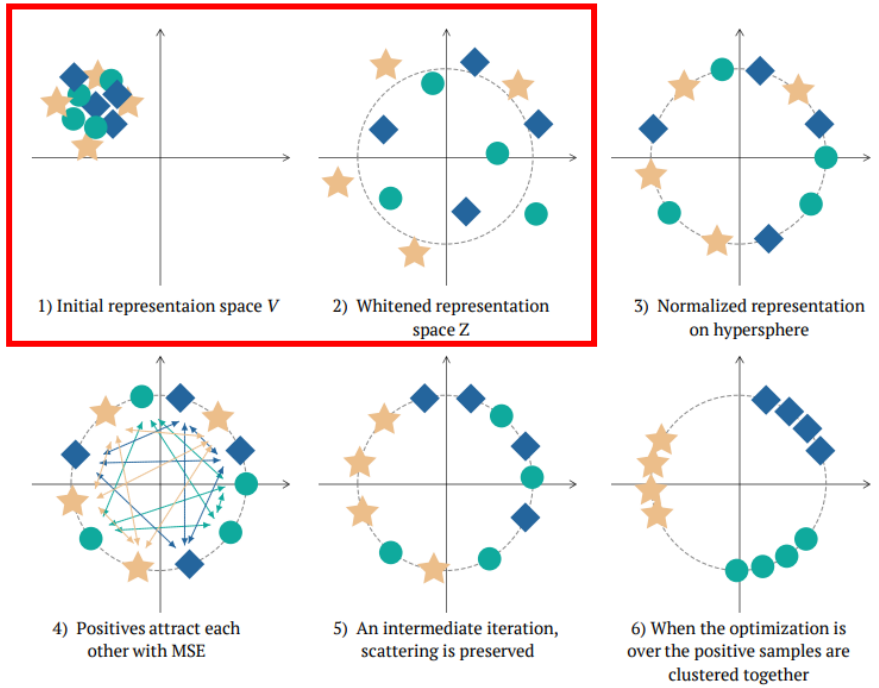
Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

### ❖ Whitening MSE Loss

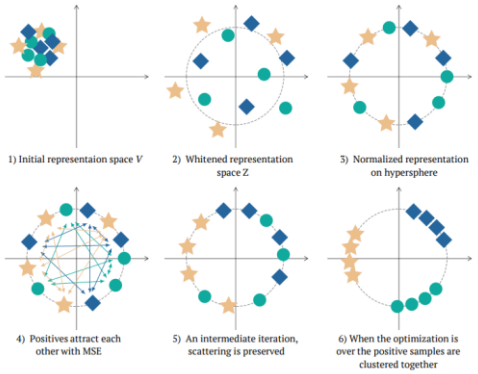
- What is role of "**Whitening**"??
  - ✓ 임베딩 벡터를 평균이 0이고 공분산 행렬이 단위 행렬인 분포에 매핑시키는 역할
  - ✓ **Redundancy Reduction + Collapse Prevention**




Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE




❖ Overall Process

- **Whitening Transform** : Redundancy Reduction + Preventing Collapse
- **MSE Loss** : Representation Invariance

$$L_{MSE}(V) = \frac{2}{Nd(d-1)} \sum_{pos(i,j)=True} dist(\mathbf{v}_i, \mathbf{v}_j) \qquad \frac{2}{Nd(d-1)} \sum_{pos(i,j)=True} dist(\mathbf{z}_i, \mathbf{z}_j)$$

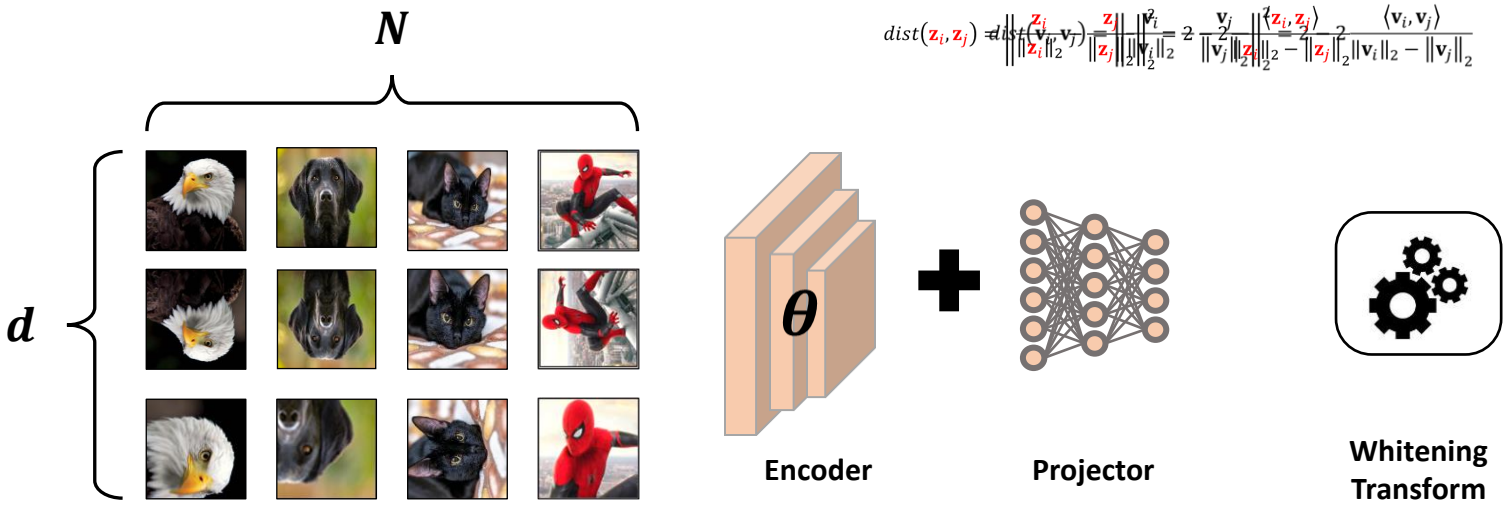
$$\mathbf{z} = Whitening(\mathbf{v}) = W_V(\mathbf{v} - \mu_V)$$

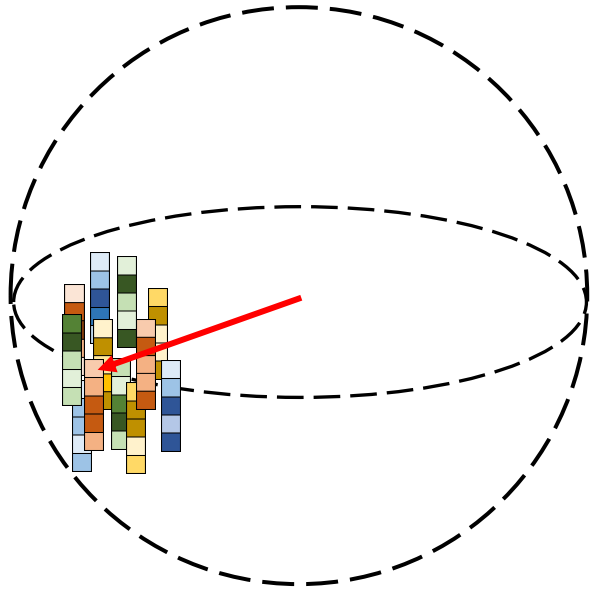
$$\mu_V = \frac{1}{K}\sum_k \mathbf{v}_k$$

$$\Sigma_V = \frac{1}{K-1}\sum_k (\mathbf{v}_k - \mu_V)(\mathbf{v}_k - \mu_V)^T$$

$$W_V^T W_V = \Sigma_V^{-1}$$

$$dist(\mathbf{z}_i, \mathbf{z}_j) = \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2} = 2 - 2\frac{\langle \mathbf{v}_i, \mathbf{v}_j \rangle}{\|\mathbf{v}_i\|_2 \|\mathbf{v}_j\|_2}$$




Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

$$L_{MSE}(V) = \frac{2}{Nd(d-1)} \sum_{pos(i,j)=True} dist(\mathbf{z}_i, \mathbf{z}_j)$$

$$\mathbf{z} = Whitening(\mathbf{v}) = W_V(\mathbf{v} - \mu_V)$$

$$\mu_V = \frac{1}{K}\sum_k \mathbf{v}_k$$

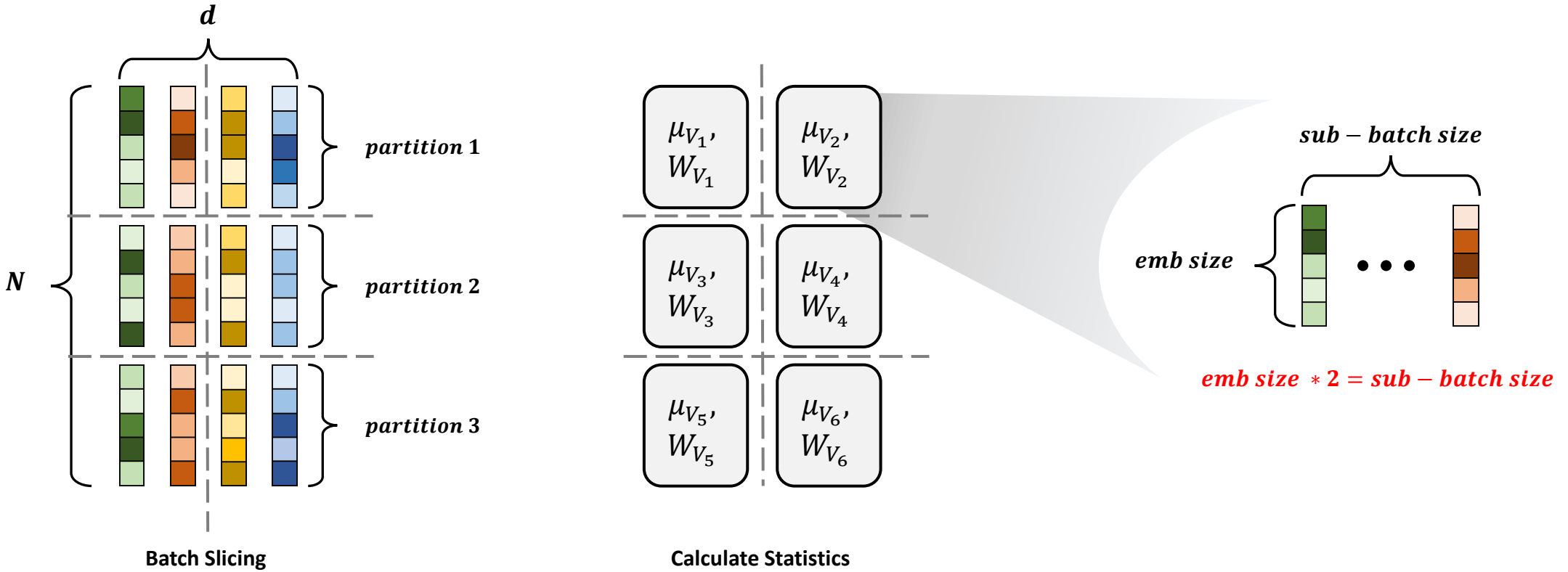
$$\Sigma_V = \frac{1}{K-1}\sum_k (\mathbf{v}_k - \mu_V)(\mathbf{v}_k - \mu_V)^T$$

$$W_V^T W_V = \Sigma_V^{-1}$$

$$dist(\mathbf{z}_i, \mathbf{z}_j) = \left\| \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|_2} - \frac{\mathbf{z}_j}{\|\mathbf{z}_j\|_2} \right\|_2^2 = 2 - 2\frac{\langle \mathbf{z}_i, \mathbf{z}_j \rangle}{\|\mathbf{z}_i\|_2 - \|\mathbf{z}_j\|_2}$$

❖ Additional Trick – Batch Slicing

- Iterative Batch 별 $\mu_V, W_V$의 Variance가 매우 큼
- 학습의 안정성을 위해 Sub-Batch 별 $\mu_V, W_V$ 을 구한 후 별도로 Whitening





Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

# Information Maximization Methods

## W-MSE

❖ Main Results

- SimCLR, BYOL 과 비교
- positive sample 개수에 따른 성능 비교

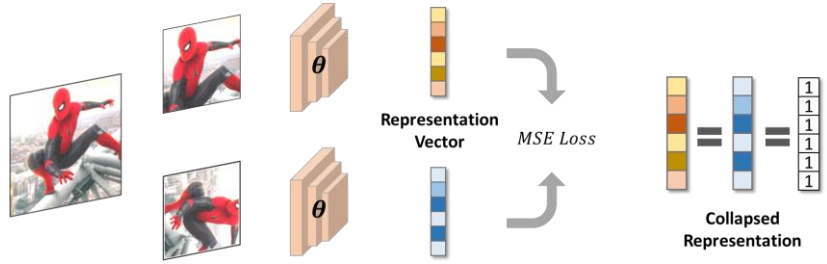
*Table 1.* Classification accuracy (top 1) of a linear classifier and a 5-nearest neighbors classifier for different loss functions and datasets with a ResNet-18 encoder.

| Method | CIFAR-10 | | CIFAR-100 | | STL-10 | | Tiny ImageNet | |
|---|---|---|---|---|---|---|---|---|
| | linear | 5-nn | linear | 5-nn | linear | 5-nn | linear | 5-nn |
| SimCLR (Chen et al., 2020a) (our repro.) | 91.80 | 88.42 | 66.83 | 56.56 | 90.51 | 85.68 | 48.84 | 32.86 |
| BYOL (Grill et al., 2020) (our repro.) | 91.73 | 89.45 | 66.60 | **56.82** | **91.99** | **88.64** | **51.00** | **36.24** |
| W-MSE 2 (ours) | 91.55 | 89.69 | 66.10 | 56.69 | 90.36 | 87.10 | 48.20 | 34.16 |
| W-MSE 4 (ours) | **91.99** | **89.87** | **67.64** | 56.45 | 91.75 | 88.59 | 49.22 | 35.44 |

Ermolov, A., Siarohin, A., Sangineto, E., & Sebe, N. (2021, July). Whitening for self-supervised representation learning. In International Conference on Machine Learning (pp. 3015-3024). PMLR.

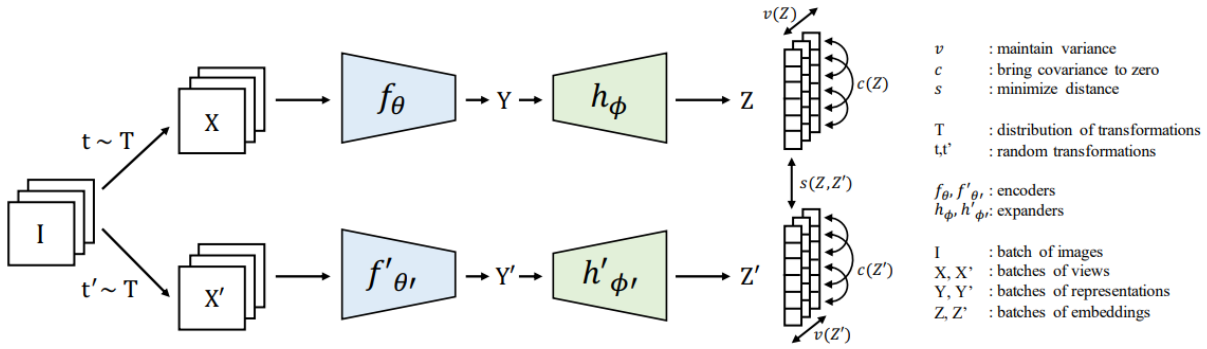
# Information Maximization Methods

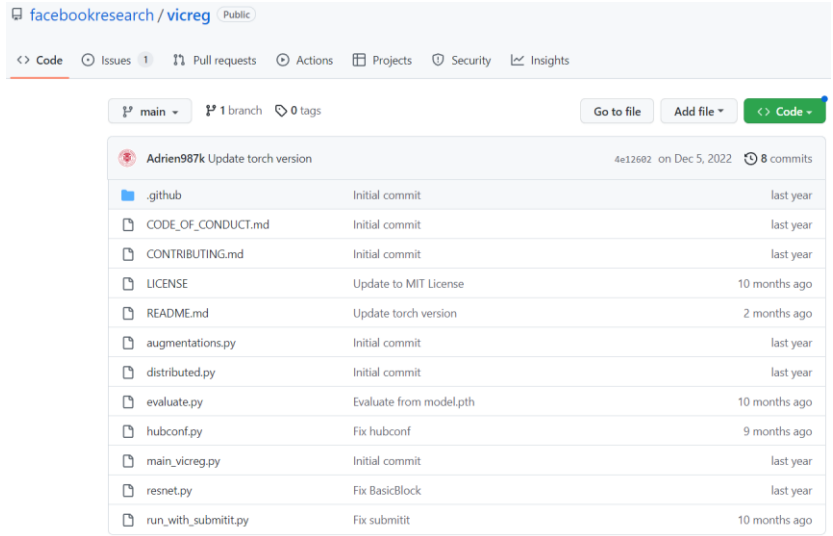
## VICReg




❖ VICREG: Variance-Invariance-Covariance Regularization for Self-Supervised Learning(2022, ICLR)

- 2023년 2월 기준 311회 인용
- 단순한 MSE Loss(Invariance Term)로는 Collapse가 발생해서 non-informative constant vector가 생성됨
  - ✓ 2가지 규제(Regularization) 함수를 추가하여 Collapse를 방지하였음
  - ✓ Weight Sharing, Batch Norm, Feature-wise Norm, Stop Gradient, Memory Bank 등 기존 SSL에서 학습의 안정성을 위해 추가한 트릭들이 필요없는게 장점
- https://github.com/facebookresearch/vicreg

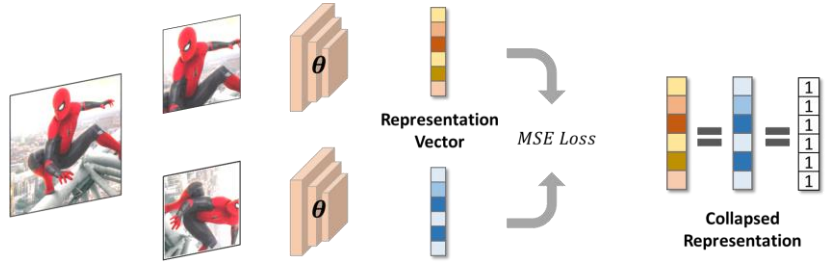



Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg




❖ Why does it called VICReg??

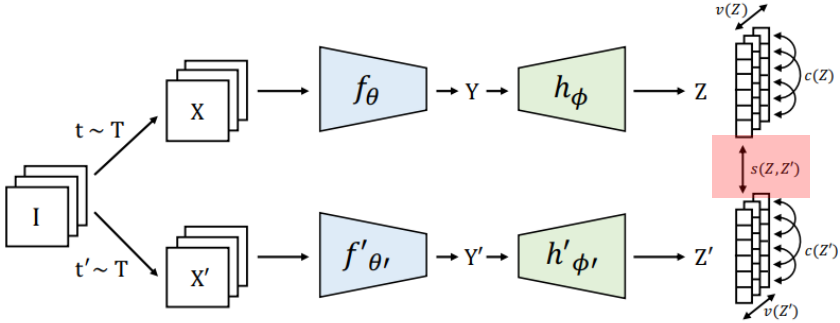
- **Variance Regularization** : 각 임베딩 차원의 분산이 0보다 크도록 하자
- **Invariance Regularization** : 같은 객체로부터 나온 임베딩이 의미론적으로 같도록 하자
- **Covariance Regularization** : 임베딩 차원 간의 상관성이 작아지도록 하자

- **Invariance**: the mean square distance between the embedding vectors.
- **Variance**: a hinge loss to maintain the standard deviation (over a batch) of each variable of the embedding above a given threshold. This term forces the embedding vectors of samples within a batch to be different.
- **Covariance**: a term that attracts the covariances (over a batch) between every pair of (centered) embedding variables towards zero. This term decorrelates the variables of each embedding and prevents an *informational collapse* in which the variables would vary together or be highly correlated.

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

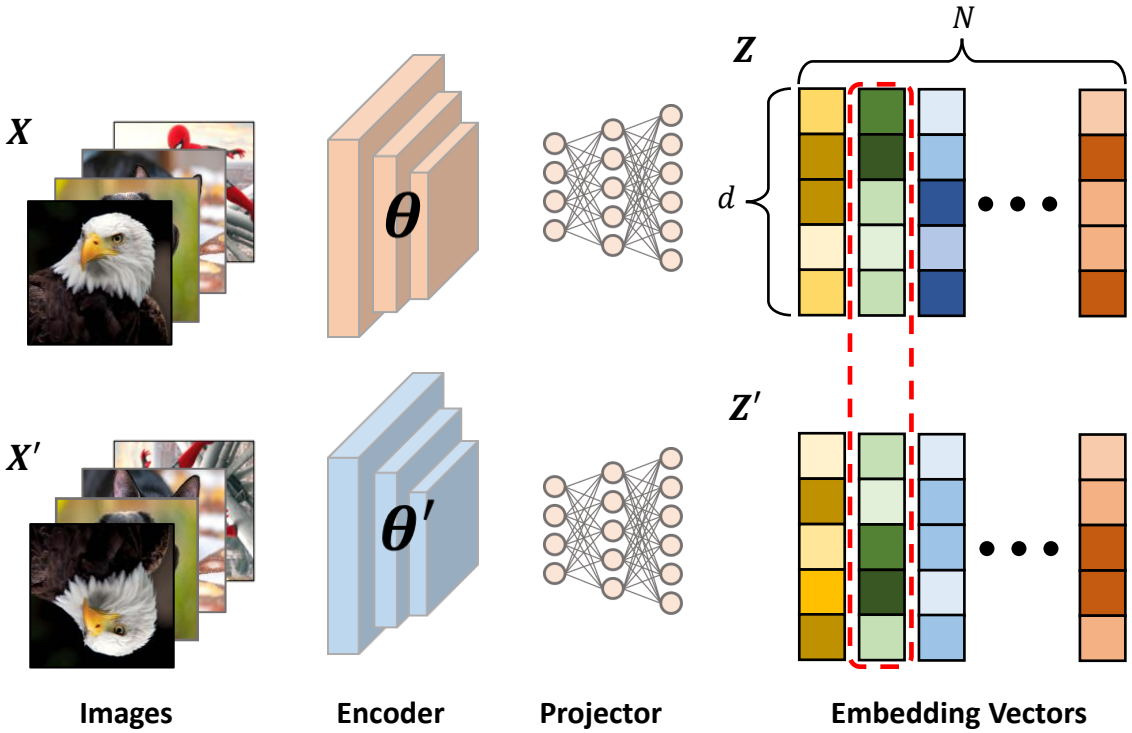
# Information Maximization Methods

## VICReg




❖ Regularization Terms

- **Invariance Regularization – Distance between the embedding vectors towards zero**
- Data Augmentation 에 상관없이 공통된 특징을 추출, 해당 Loss만 사용할 경우 Collapse 발생

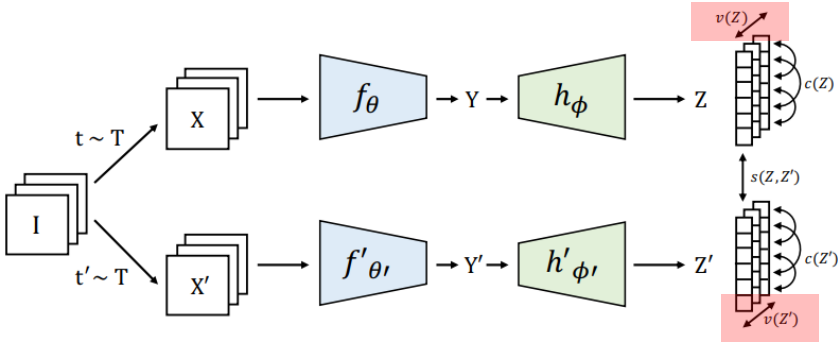



$$s(Z, Z') = \frac{1}{n}\Sigma_i^n \|z_i - z'_i\|_2^2$$

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

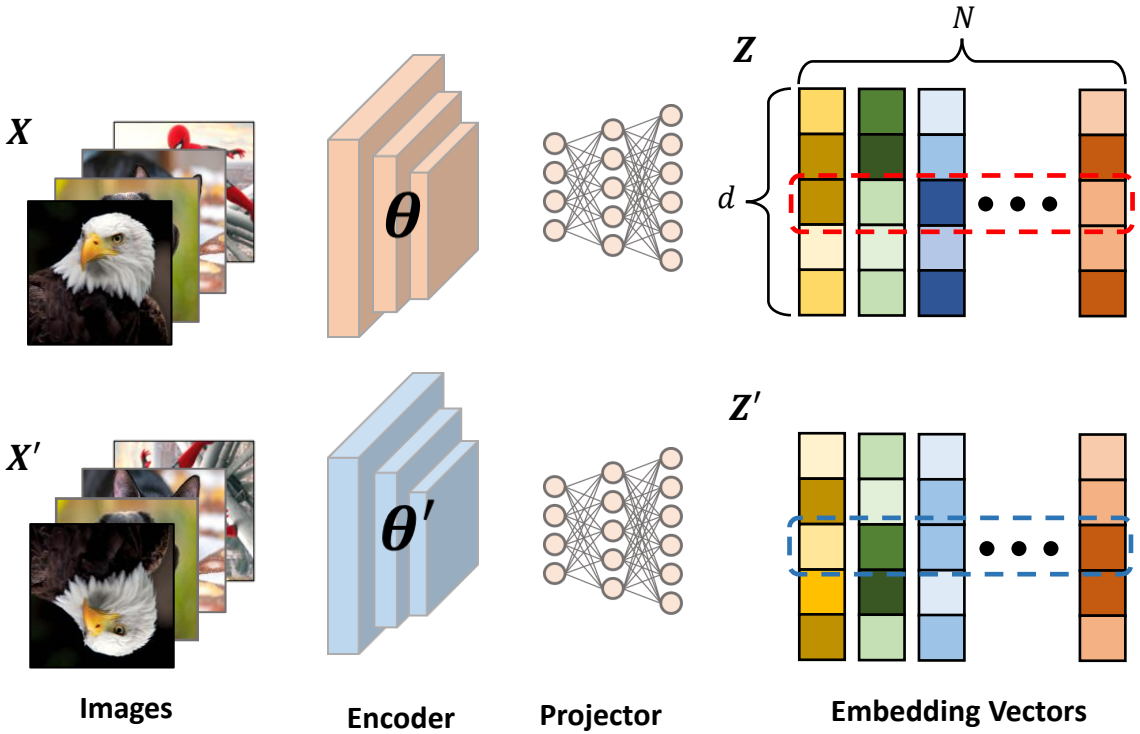
# Information Maximization Methods

## VICReg




❖ Regularization Terms

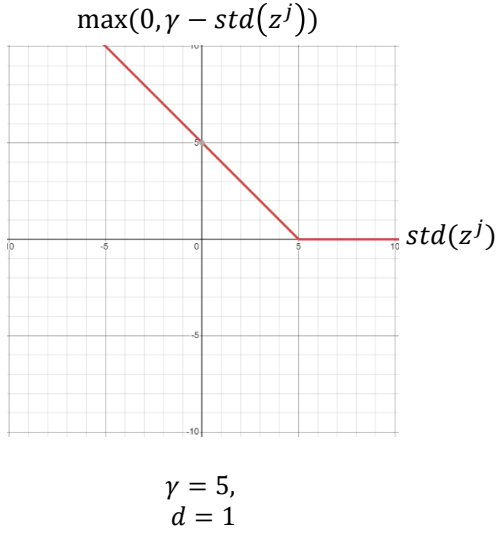
- **Variance Regularization – Standard deviation of each variable(dimension) of the embedding above a given threshold**
- 임베딩 벡터들이 동일한 값을 가지는 벡터로 매핑되는 Collapse를 방지




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

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

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

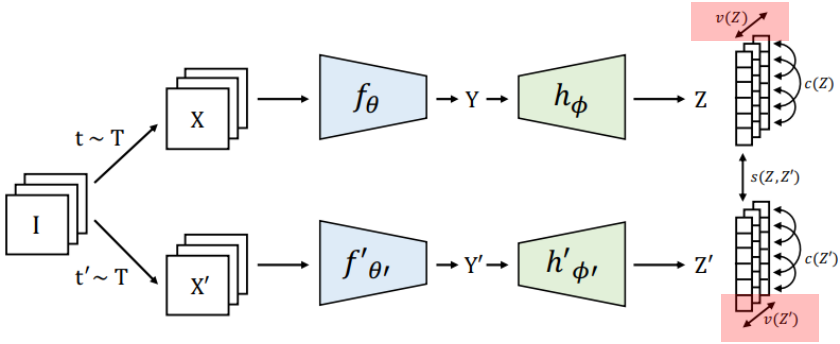




Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

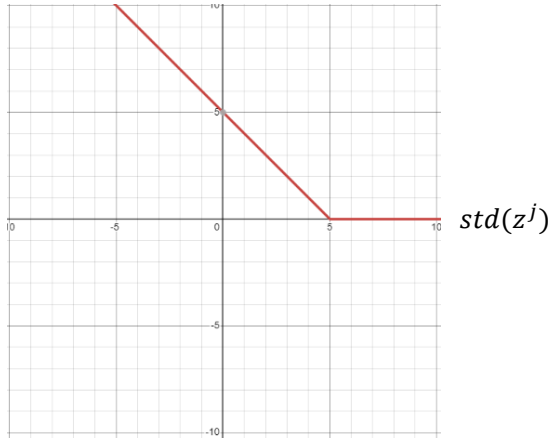
## VICReg




❖ Regularization Terms

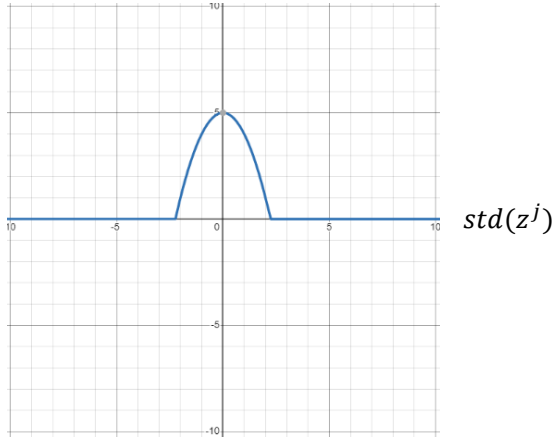
- **Variance Regularization – Standard deviation of each variable(dimension) of the embedding above a given threshold**
- Why Standard Deviation instead of Variance??

$\max(0, \gamma - std(z^j))$




$\gamma = 5,$
$d = 1$

$\max(0, \gamma - Var(z^j))$

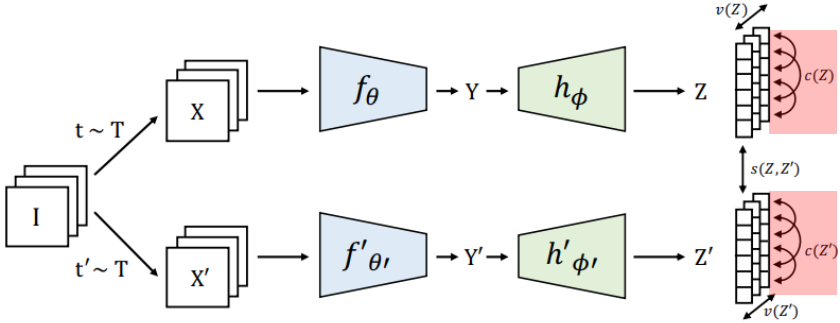



$\gamma = 5,$
$d = 1$

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

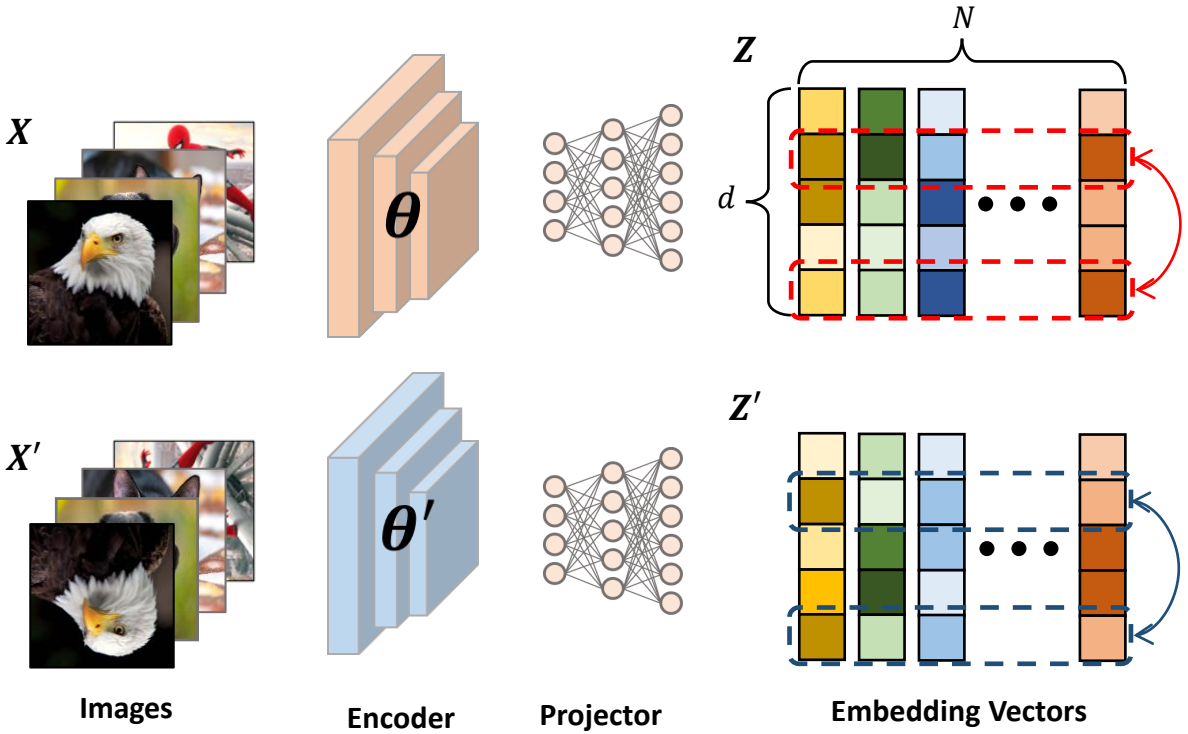
# Information Maximization Methods

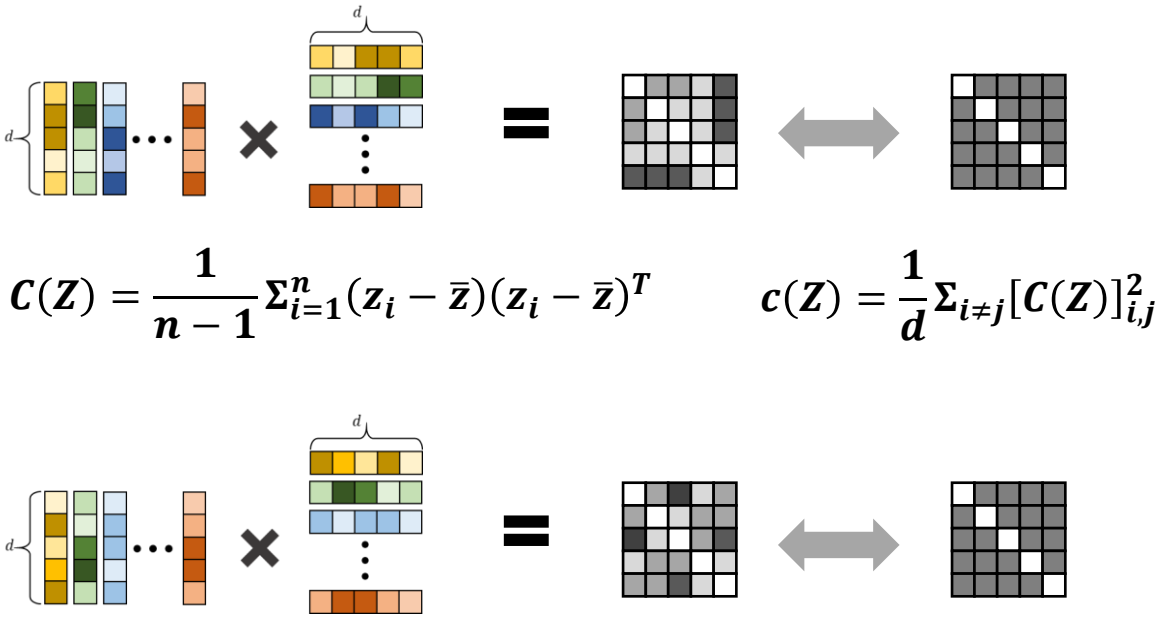
## VICReg




❖ Regularization Terms

- **Covariance Regularization – Covariances between every pair of centered embedding variables towards zero**
- 각 임베딩 차원이 동일한 정보를 인코딩하는 것을 방지(Redundancy Reduction)



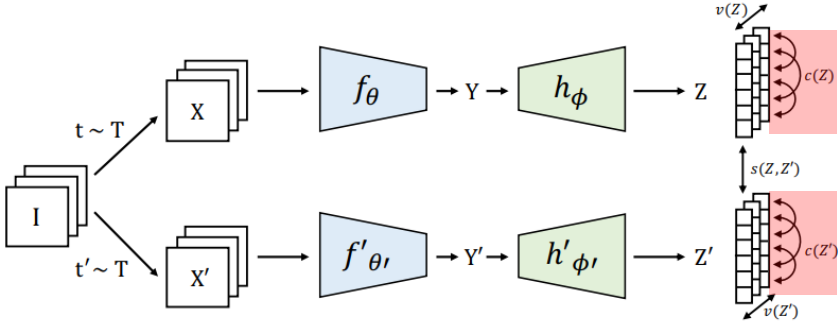

$$C(Z) = \frac{1}{n-1}\Sigma_{i=1}^{n}(z_i - \bar{z})(z_i - \bar{z})^T \qquad c(Z) = \frac{1}{d}\Sigma_{i \neq j}[C(Z)]_{i,j}^2$$

$$C(Z') = \frac{1}{n-1}\Sigma_{i=1}^{n}(z'_i - \bar{z}')(z'_i - \bar{z}')^T \qquad c(Z') = \frac{1}{d}\Sigma_{i \neq j}[C(Z')]_{i,j}^2$$


Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

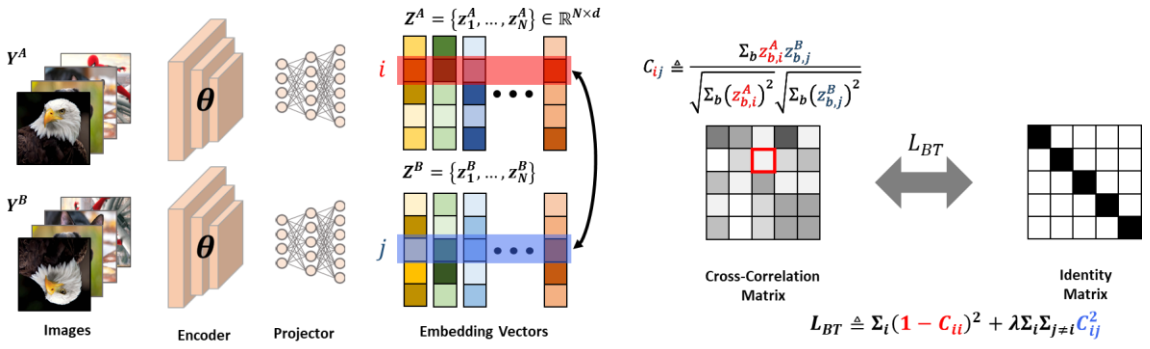
# Information Maximization Methods

## VICReg

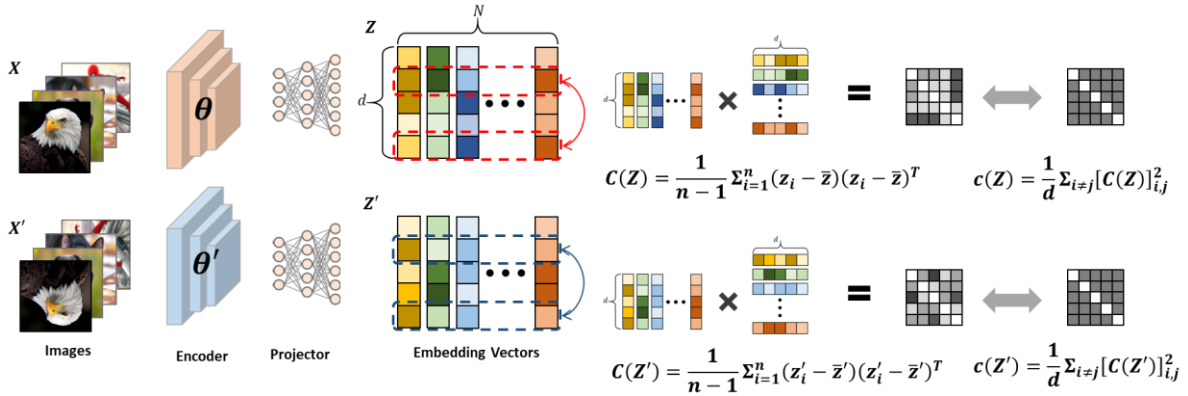



❖ Regularization Terms

- **Covariance Regularization – Covariances between every pair of centered embedding variables towards zero**
- Difference between Barlow Twins and VICReg
  - ✓ **Barlow Twins** : 서로 다른 인코더에서 나온 Embedding Variable 간의 상관관계 계산(Cross-Correlation)
  - ✓ **VICReg** : 각각의 인코더에서 나온 Embedding Variable 간의 상관관계 계산(Correlation)




**Barlow Twins**

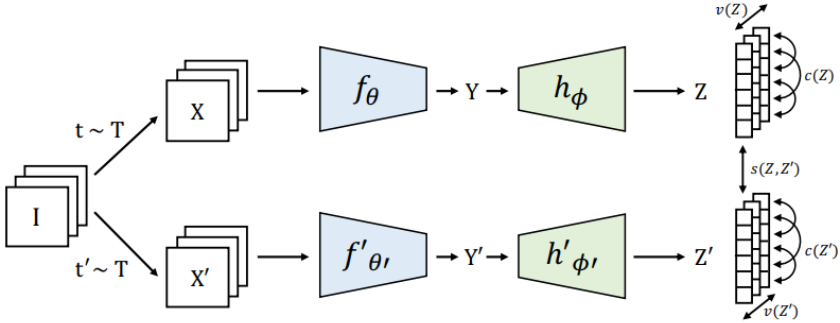



**VICReg**


Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg



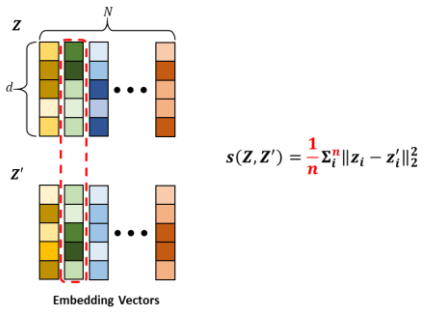

❖ Total Loss Function

- Hyper parameters(Grid Search)
  - ✓ $\nu = 1$
  - ✓ $\lambda = \mu > 1$

$$\boldsymbol{l(Z, Z') = \lambda s(Z, Z') + \mu[v(Z) + v(Z')] + \nu[c(Z) + c(Z')]}$$

### Invariance Term

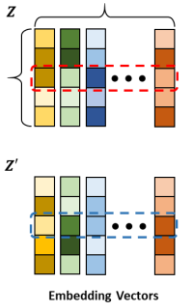
- **Encoding Representation**




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

### Variance Term

- **Collapse Prevention**



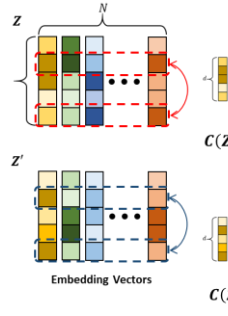

$$S(x, \epsilon) = \sqrt{Var(x) + \epsilon}$$

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

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

### Covariance Term

- **Redundancy Reduction**




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

$$C(Z') = \frac{1}{n-1}\Sigma_{i=1}^n (z'_i - \bar{z}')(z'_i - \bar{z}')^T$$


Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg

### ❖ Main Results

- Linear Classification/Semi-Supevised Classification from ImageNet
- Linear Classification/Object Detection from Places, VOC, COCO

Table 1: **Evaluation on ImageNet.** Evaluation of the representations obtained with a ResNet-50 backbone pretrained with VICReg on: (1) linear classification on top of the frozen representations from ImageNet; (2) semi-supervised classification on top of the fine-tuned representations from 1% and 10% of ImageNet samples. We report Top-1 and Top-5 accuracies (in %). Top-3 best self-supervised methods are underlined.

| 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 |

Table 2: **Transfer learning on downstream tasks.** Evaluation of the representations from a ResNet-50 backbone pretrained with VICReg on: (1) linear classification tasks on top of frozen representations, we report Top-1 accuracy (in %) for Places205 Zhou et al. (2014) and iNat18 Horn et al. (2018), and mAP for VOC07 Everingham et al. (2010); (2) object detection with fine-tunning, we report $AP_{50}$ for VOC07+12 using Faster R-CNN with C4 backbone Ren et al. (2015); (3) object detection and instance segmentation, we report AP for COCO Lin et al. (2014) using Mask R-CNN with FPN backbone He et al. (2017). We use † to denote the experiments run by us. Top-3 best self-supervised methods are underlined.

| 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 |

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg

❖ Multi-Modal Availability

- MS-COCO Dataset에서 이미지-캡션(텍스트)로 사전 학습 후 Content Retrieval 성능 비교
  - ✓ VICReg는 각각의 인코더를 별개로 규제화하기 때문에 각 인코더의 Modality가 다르거나 혹은 통계량이 달라도 사용 가능
  - ✓ Barlow Twins의 경우 각 인코더의 결과로부터 Cross Correlation을 구하기 때문에 동일한 Modality 혹은 통계량을 기대

Table 3: **Evaluation on MS-COCO 5K retrieval tasks.** Comparison of VICReg with the contrastive loss of VSE++ Faghri et al. (2018), and with Barlow Twins, pretrain on the training set of MS-COCO. In all settings, the encoder for text is a word embedding followed by a GRU layer, the encoder for images is a ResNet-152.

| 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 |

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg

- Options
  - Different Architecture? Sharing Weights?

Table 5: **Impact of sharing weights or not between branches.** Top-1 accuracy on linear classification with 100 pretraining epochs. The encoder and expander of both branches can share the same architecture and share their weights (SW), share the same architecture with different weights (DW), or have different architectures (DA). The encoders can be ResNet-50, ResNet-101 or ViT-S.

| | 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 |

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Information Maximization Methods

## VICReg

- Combining with other Methods
  - BYOL/SimSiam + Variance Regularization

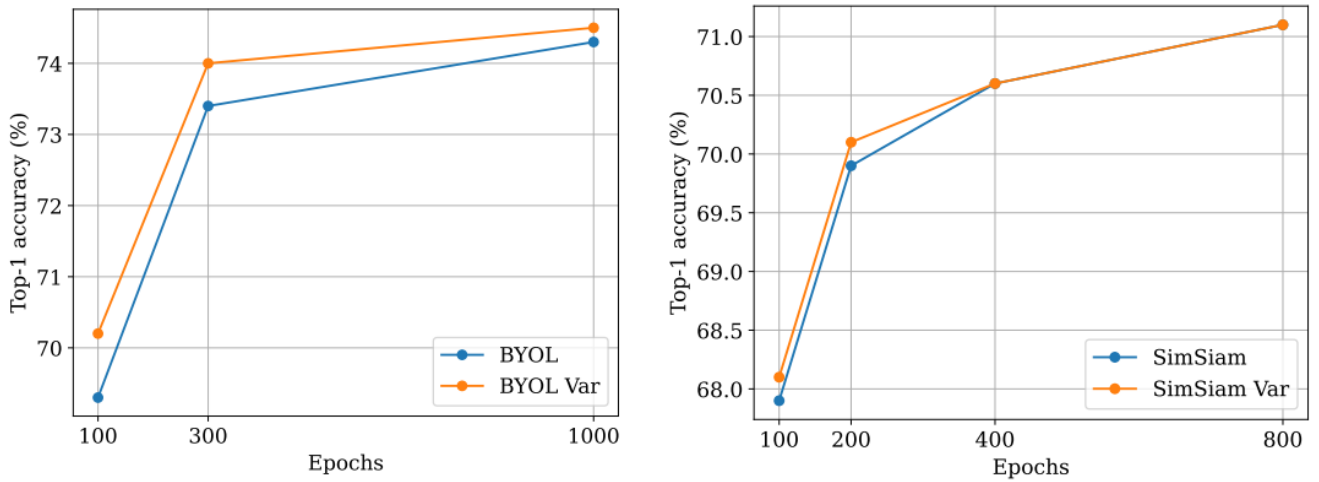



Figure 3: **Incorporating variance regularization in BYOL and SimSiam.** Top-1 accuracy on the linear evaluation protocol for different number of pretraining epochs. For both methods pre-training follows the optimization and data augmentation protocol of their original paper but is based on our implementation. *Var* indicates variance regularization

Bardes, A., Ponce, J., & LeCun, Y. (2021). Vicreg: Variance-invariance-covariance regularization for self-supervised learning. arXiv preprint arXiv:2105.04906.

# Conclusion

## Summary

❖ Limitation of Previous SSL Methods

- 기존 SSL 방법론은 의미있는 정보를 추출하였으나 Large Batch Size, Batch Normalization, Assymetric Architecture 를 요구
  - ✓ 메모리 비효율성 혹은 구조적 제약을 필요로함

❖ Information Maximiaztion Methods

- Redundancy Reduction, Invariance 를 해결하면서 Collapse를 방지하고자함
  - ✓ 모델 구조적 제약, 메모리 부담이 적음
- BarlowTwins : Cross-Correlation 을 감소시켜 Redundancy Reduction 해결
- W-MSE : Whitening Transform을 통해 데이터를 흩뿌림으로써 Negative Sample 없이 Collapse를 방지
- VICReg : Variance/Invariance/Covariance Regularization Term을 활용해 Collapse 방지와 Redundancy Reduction 수행