

Unifying contrastive learning and clustering

2022.11.18

Data Mining & Quality Analytics Lab.

김현지

발표자 소개



이름

김현지

학력 사항

고려대학교 산업경영공학과 석사과정 (2022.03 ~ 현재)
Data Mining & Quality Analytics 연구실 (지도교수: 김성범 교수님)

관심 연구 분야

Deep Learning Algorithm for Multivariate Signal Analysis
Representation Learning for Time-Series Data

E-Mail

99ktxx@korea.ac.kr

목차

01 Introduction

Instance-wise contrastive learning: Instance discrimination
Clustering: a classical school of unsupervised learning

02 Clustering for deep representation learning - DeepCluster

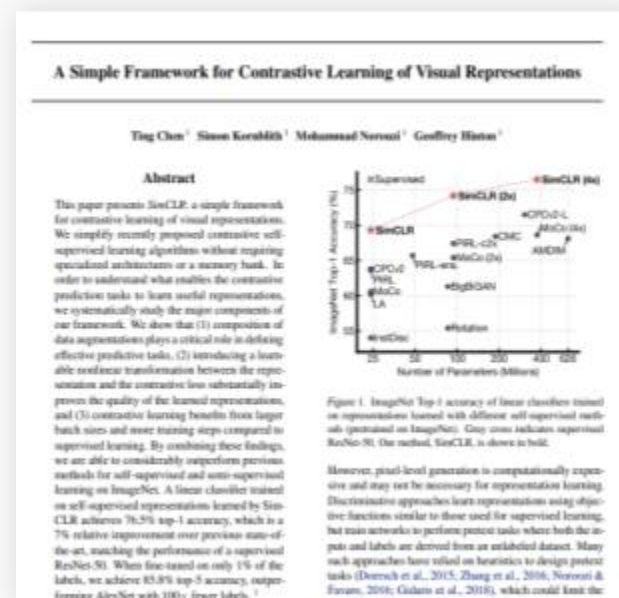
03 Unifying contrastive learning and clustering

- PCL
- SwAV

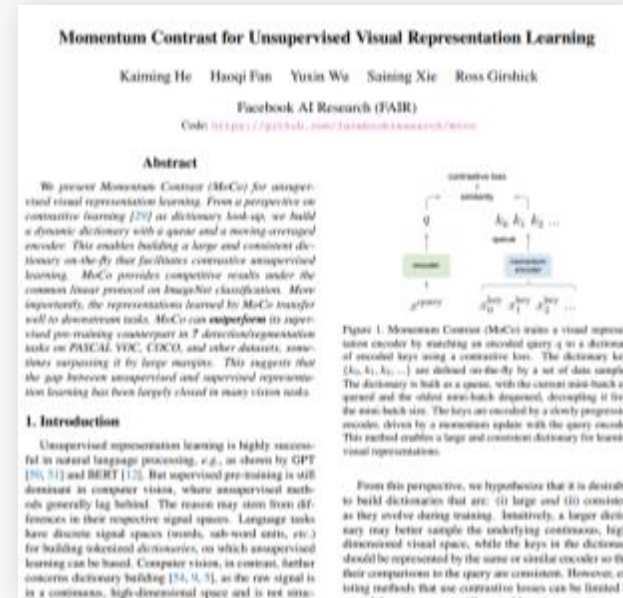
Instance-wise contrastive learning

Background

- 최근 Unsupervised visual representation learning의 성능은 supervised method의 성능과 유사한 수준까지 도달함
- 이러한 발전은 주로 instance discrimination 작업을 기반으로 한 instance-wise contrastive learning을 중심으로 이루어지고 있음



SimCLR



MoCo

Instance-wise contrastive learning

Background

- 최근 Unsupervised visual representation learning의 성능은 supervised method의 성능과 유사한 수준까지 도달함
- 이러한 발전은 주로 instance discrimination 작업을 기반으로 한 instance-wise contrastive learning을 중심으로 이루어지고 있음


관련 DMQA OPEN SEMINAR

종료

Self-Supervised Representation Learning

Seokho Moon
May 1, 2020

Self-Supervised Representation Learning

발표자:  문석호

📅 2020년 5월 1일
🕒 오후 1시 ~
📍 화상 프로그램 이용(Zoom)

세미나 정보 보기 →

종료

Towards Contrastive Learning

발표자:  관민구

📅 2021년 1월 29일
🕒 오후 1시 ~
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →


종료

Deal with Contrastive Learning

고은성

Korea University
Data Mining & Quality Analytics Lab.

Deal with Contrastive Learning

발표자:  고은성

📅 2021년 9월 10일
🕒 오전 1시 ~
📺 온라인 비디오 시청 (YouTube)

세미나 정보 보기 →

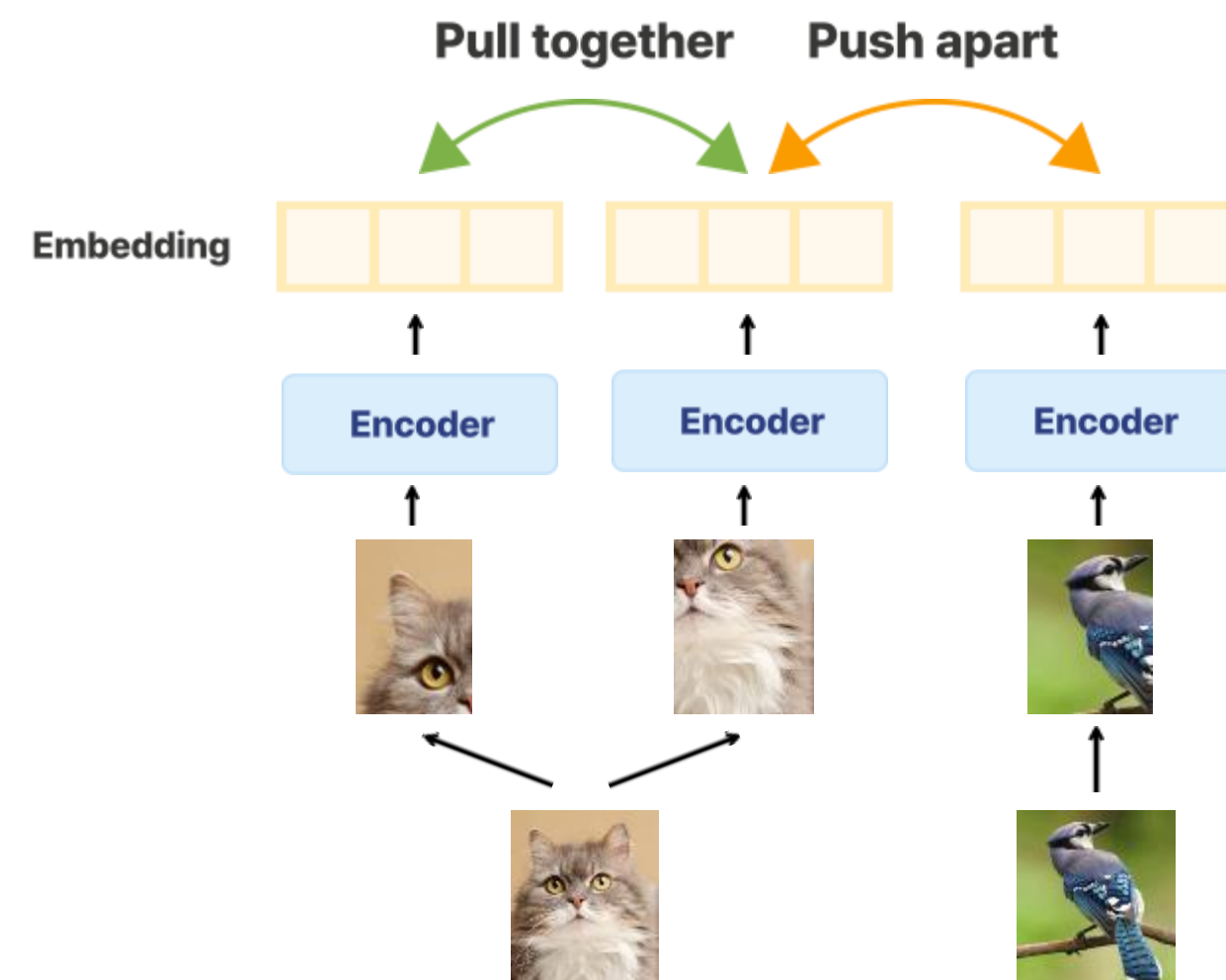
Instance-wise contrastive learning

Background

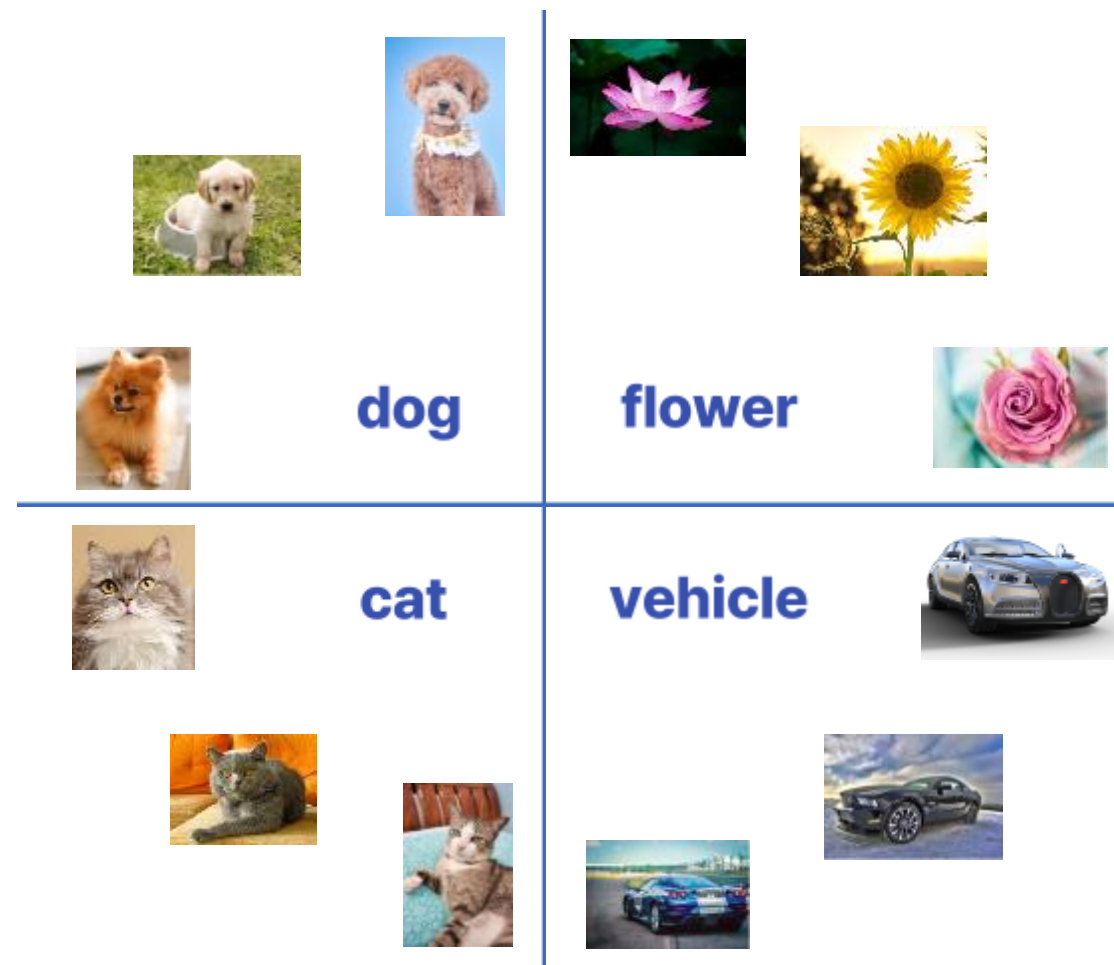
Instance discrimination

두 이미지 crop이 동일한 source 이미지에서 왔는지에 대한 여부를 분류하도록 인코더를 학습

- 인코더는 동일한 source의 임베딩은 서로 더 가까이 끌어당기면서 다른 source의 임베딩은 멀어지도록 함

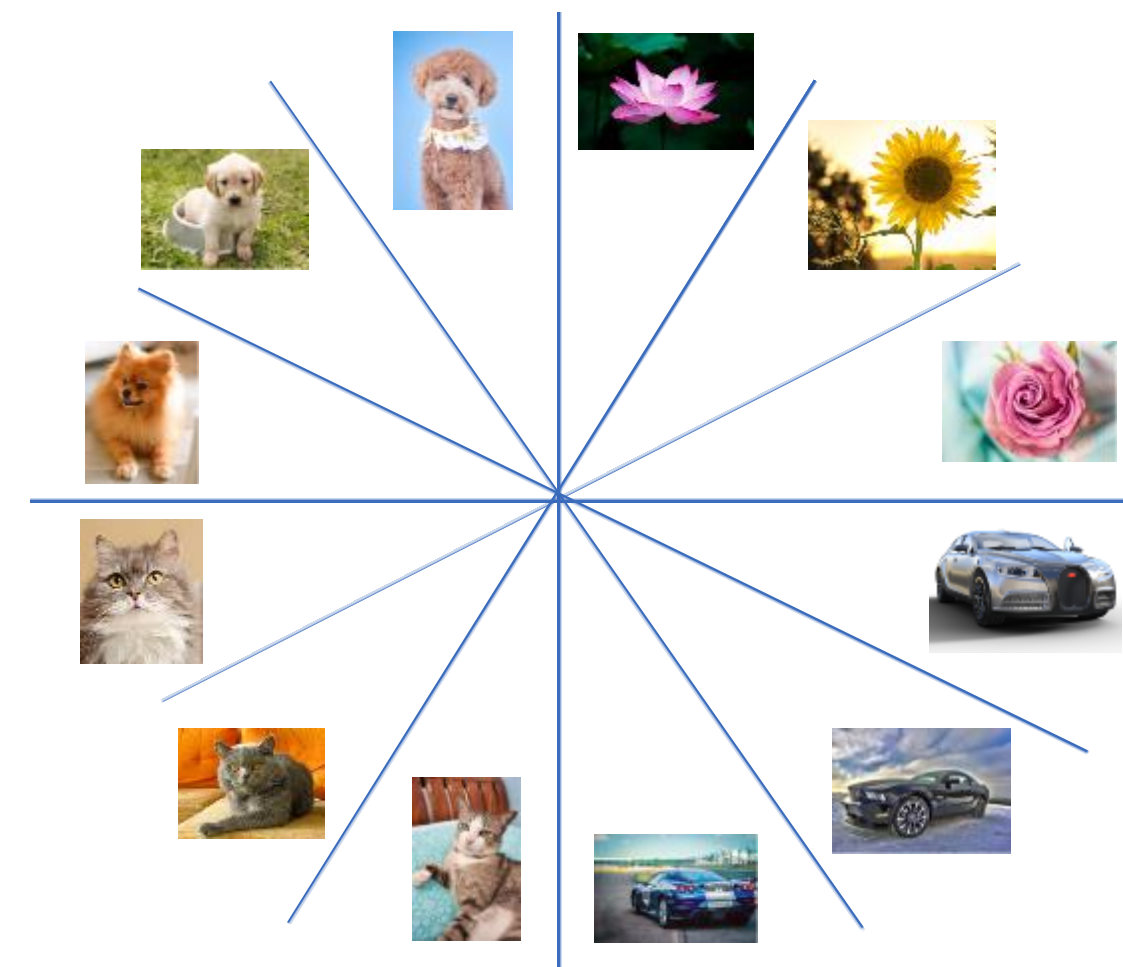


Instance discrimination



Classification

VS



Instance discrimination

Instance discrimination

Instance discrimination은 unsupervised representation learning에서 큰 성능 향상을 보여주었음
그러나 해당 방법론에는 주요한 한계점 두 가지가 존재함

Classification

Instance discrimination

Limitations of instance-wise contrastive learning

Limitation 1

데이터의 의미론적 구조(semantic structure)를 인코딩하지 못함

Limitation 2

좋은 특징을 추출하기 위해 큰 batch size가 필요함

Limitations of instance-wise contrastive learning

Limitation 1

데이터의 의미론적 구조(semantic structure)를 인코딩하지 못함

- Instance discrimination 작업은 low-level 신호를 이용하여 해결할 수 있으므로 학습된 임베딩이 반드시 high-level semantic을 포착할 필요가 없음
 - ✓ Instance classification의 정확도는 종종 높은 수준으로 빠르게 수렴함
 - ✓ Instance discrimination에서는 높은 성능을 보이지만 downstream task에서는 좋지 않은 성능을 보일 수 있음 (Tschannen et al., 2020)

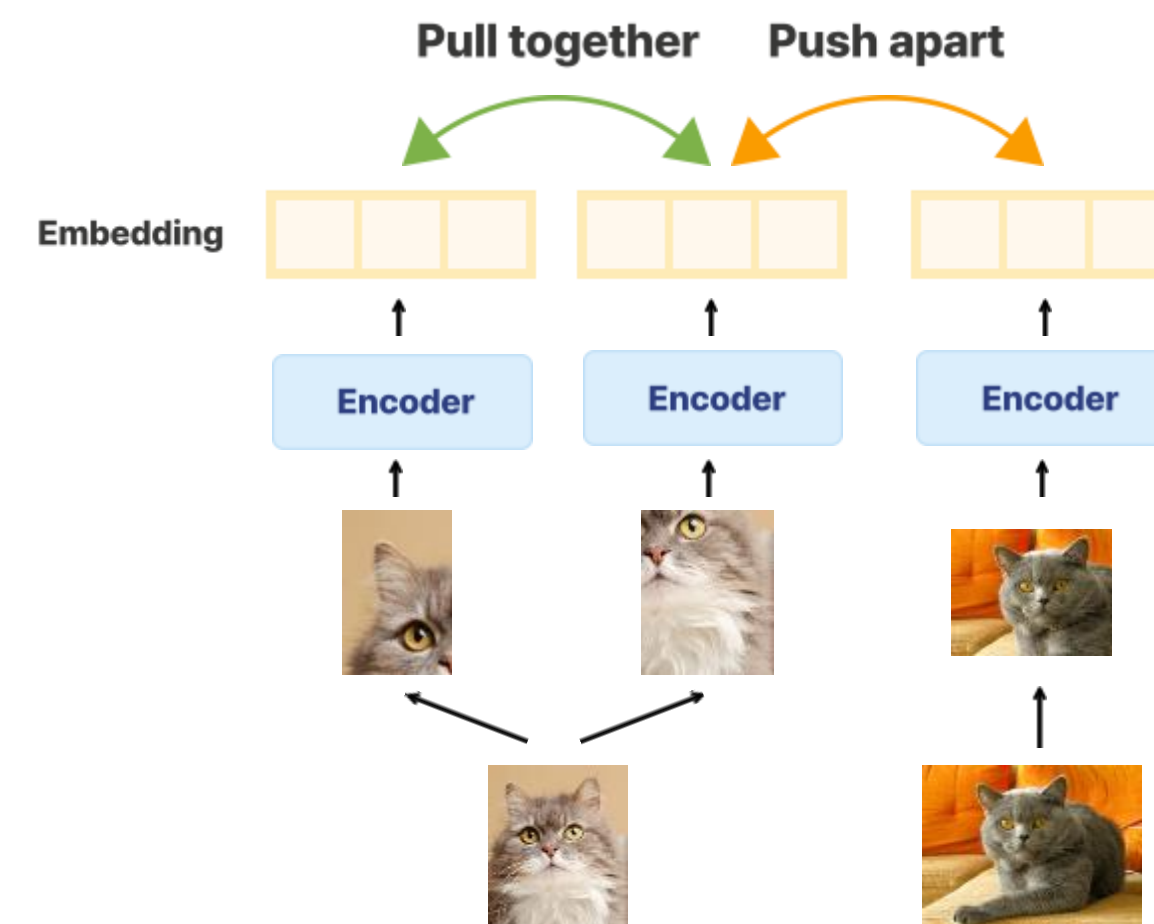
Limitations of instance-wise contrastive learning

Limitation 1

데이터의 의미론적 구조(semantic structure)를 인코딩하지 못함

➤ Semantic similarity에 상관없이 서로 다른 instance는 negative pair로 취급됨

- ✓ 이러한 문제는 클래스 충돌(class collision)로 정의되며 representation 학습에 부정적인 영향을 미침 (Saunshi et al., 2019)



Limitations of instance-wise contrastive learning

Limitation 2

좋은 특징을 추출하기 위해 큰 batch size가 필요함

- 학습이 진행되면서 target이 변하는 online 형태로 학습이 진행되기 때문에, 많은 positive, negative 샘플 간 비교를 위해 큰 batch size가 필요함
 - ✓ 큰 memory bank나 특별한 momentum network가 필요함
 - ✓ 큰 batch size가 필요한 것은 실용적이지 않을 수 있음

Limitations of instance-wise contrastive learning

Limitation 1

데이터의 의미론적 구조(semantic structure)를 인코딩하지 못함

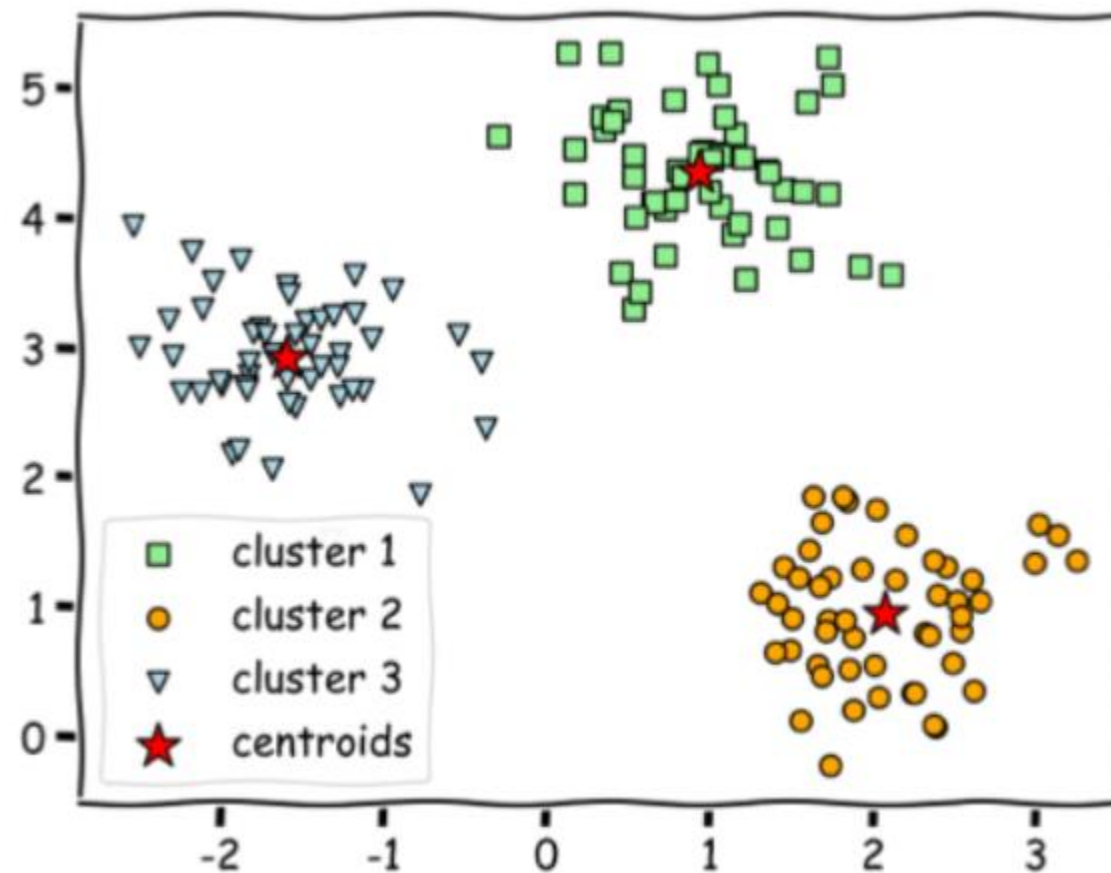
Limitation 2

좋은 특징을 추출하기 위해 큰 batch size가 필요함



Clustering 기법을 통해 극복!

Clustering: a classical school of unsupervised learning



데이터 셋의 여러 그룹으로 분할하는 과정으로, 유사한 데이터 포인트가 서로 그룹화되고, 서로 다른 데이터 포인트들은 다른 그룹에 속하도록 함

2

Clustering for deep representation learning

Deep clustering for unsupervised learning of visual features

(ECCV 2018)

Deep Clustering for Unsupervised Learning of Visual Features

Mathilde Caron, Piotr Bojanowski, Armand Joulin, and Matthijs Douze

Facebook AI Research

Abstract. Clustering is a class of unsupervised learning methods that has been extensively applied and studied in computer vision. Little work has been done to adapt it to the end-to-end training of visual features on large scale datasets. In this work, we present DeepCluster, a clustering method that jointly learns the parameters of a neural network and the cluster assignments of the resulting features. DeepCluster iteratively groups the features with a standard clustering algorithm, k -means, and uses the subsequent assignments as supervision to update the weights of the network. We apply DeepCluster to the unsupervised training of convolutional neural networks on large datasets like ImageNet and YFCC100M. The resulting model outperforms the current state of the art by a significant margin on all the standard benchmarks.

Keywords: unsupervised learning, clustering

Motivation

Random AlexNet으로 1000개의 클래스를 분류한다면?



Motivation

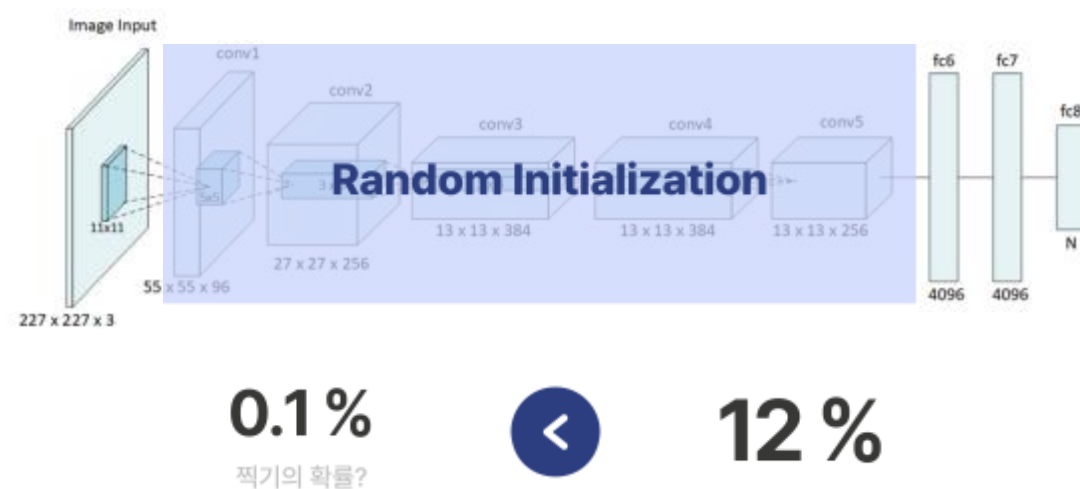
Random AlexNet으로 1000개의 클래스를 분류한다면?**0.1 %**

찍기의 확률?

**12 %**

Motivation

Feature extractor가 주는 정보를 최대한 활용한 pseudo label



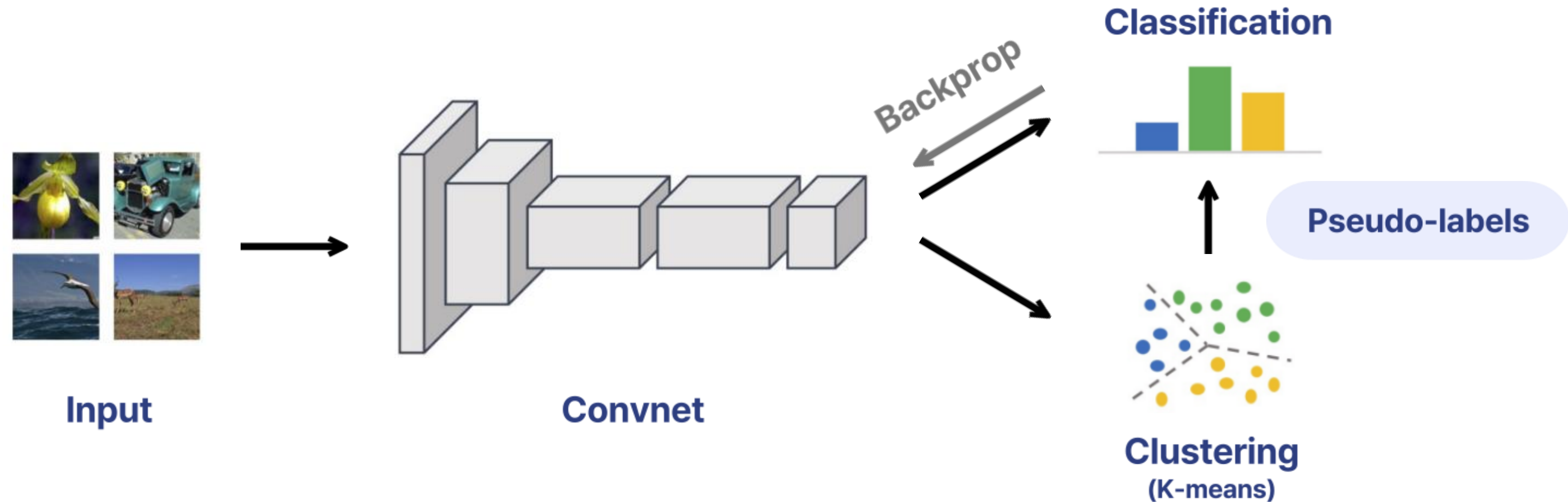
- Random AlexNet의 성능이 무작위로 예측했을 때보다 훨씬 우수한 이유
: Input signal에 strong prior를 주는 convolutional neural networks의 구조적인 이점 때문
- ✓ 초기 convnet의 output에도 어느 정도 유의미한 정보가 있음
- ✓ 초기의 weak signal부터 convnet의 이점을 최대화하는 것을 시도

Main idea

Convnet의 output을 clustering한 후 cluster assignments를 pseudo label 로 사용하자

Method

Framework of DeepCluster



" Feature vector들에 대한 clustering과 cluster assingment를 예측하여 convnet의 가중치를 업데이트 하는 것을 반복 "

Method

Framework of DeepCluster

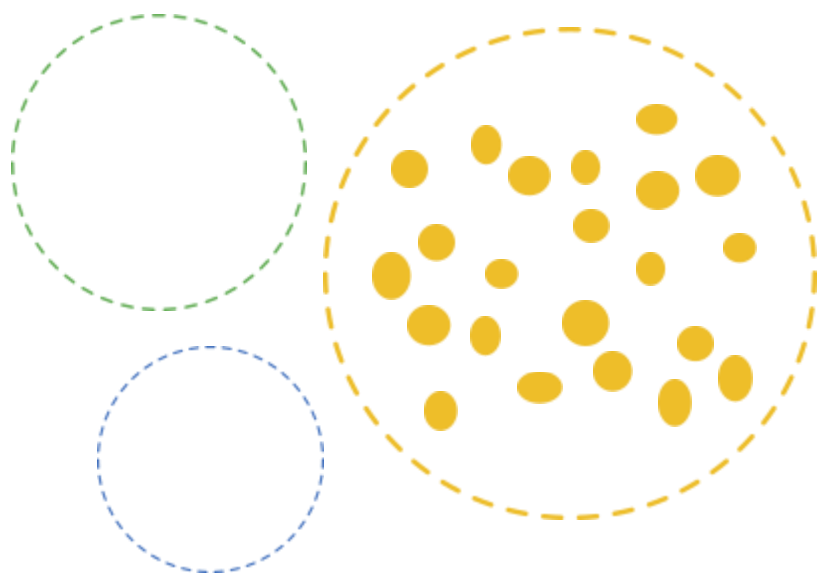


" Feature vector들에 대한 clustering과 cluster assignment를 예측하여 convnet의 가중치를 업데이트 하는 것을 반복 "

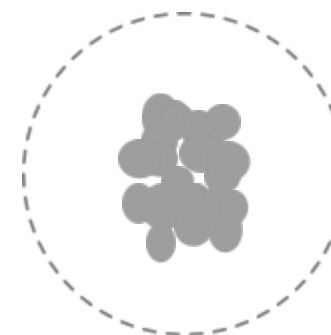
Method

Trivial solution

유형 1

**모든 feature를 한 cluster에 할당, 빈 cluster 발생**

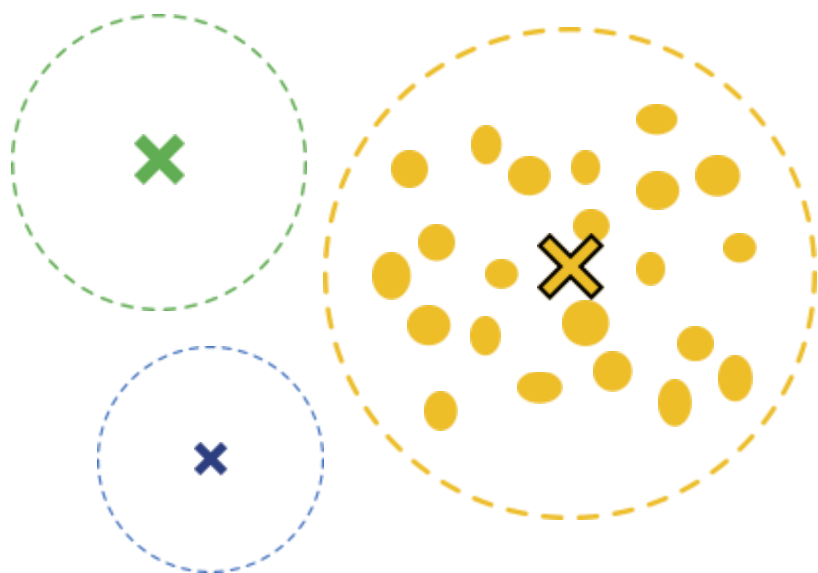
유형 2

**Major cluster에만 할당되도록 feature가 추출됨**

Method

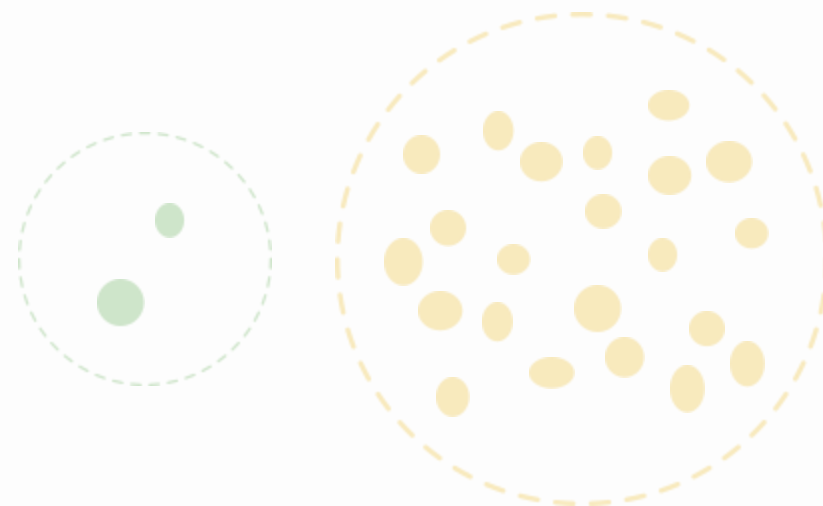
Avoiding trivial solution

유형 1



빈 cluster의 중심을 주변 데이터가 있는 곳으로
옮긴 후 다시 cluster를 할당

유형 2

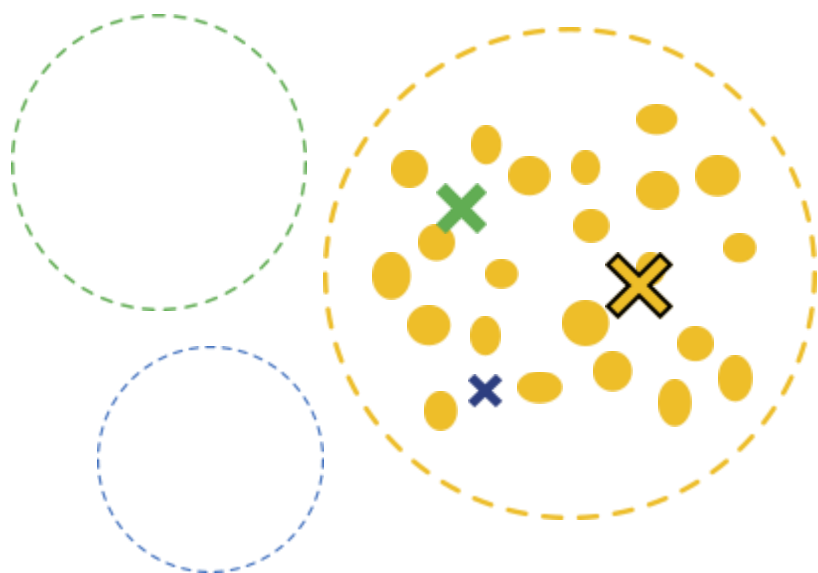


데이터 불균형으로 인해 발생하는 문제로, 각 cluster에
할당된 데이터 개수의 역수를 loss에 곱해줌

Method

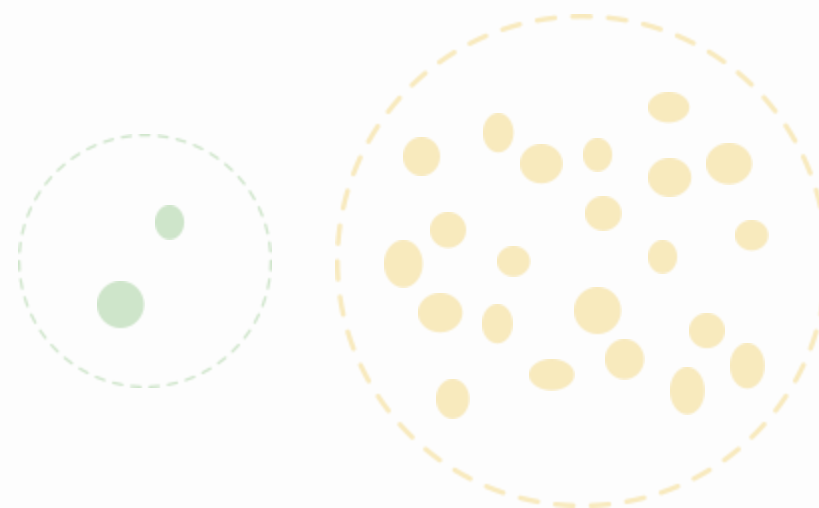
Avoiding trivial solution

유형 1



빈 cluster의 중심을 주변 데이터가 있는 곳으로
옮긴 후 다시 cluster를 할당

유형 2

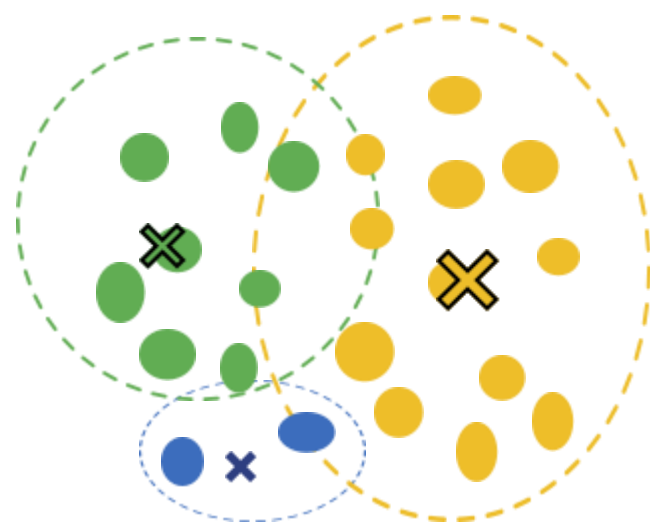


데이터 불균형으로 인해 발생하는 문제로, 각 cluster에
할당된 데이터 개수의 역수를 loss에 곱해줌

Method

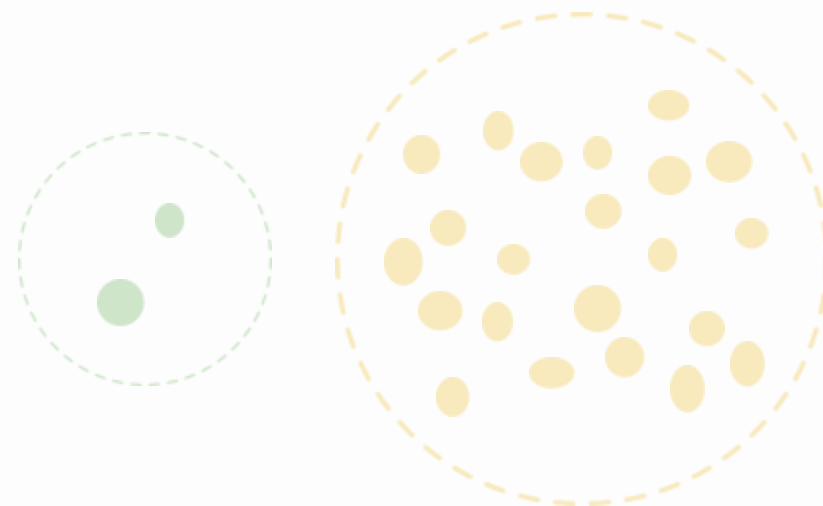
Avoiding trivial solution

유형 1



빈 cluster의 중심을 주변 데이터가 있는 곳으로
옮긴 후 다시 cluster를 할당

유형 2



데이터 불균형으로 인해 발생하는 문제로, 각 cluster에
할당된 데이터 개수의 역수를 loss에 곱해줌

Method

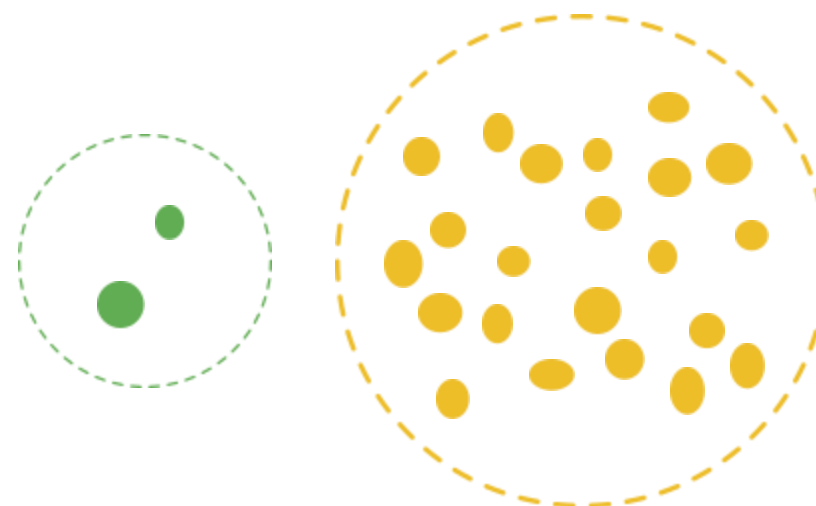
Avoiding trivial solution

유형 1



빈 cluster의 중심을 주변 데이터가 있는 곳으로
옮긴 후 다시 cluster를 할당

유형 2



데이터 불균형으로 인해 발생하는 문제로, 각 cluster에
할당된 데이터 개수의 역수를 loss에 곱해줌

Experiment

Performance in a downstream task

Method	Classification		Detection		Segmentation	
	FC6-8	ALL	FC6-8	ALL	FC6-8	ALL
ImageNet labels	78.9	79.9	–	56.8	–	48.0
Random-rgb	33.2	57.0	22.2	44.5	15.2	30.1
Random-sobel	29.0	61.9	18.9	47.9	13.0	32.0
Pathak <i>et al.</i> [38]	34.6	56.5	–	44.5	–	29.7
Donahue <i>et al.</i> [20]*	52.3	60.1	–	46.9	–	35.2
Pathak <i>et al.</i> [27]	–	61.0	–	52.2	–	–
Owens <i>et al.</i> [44]*	52.3	61.3	–	–	–	–
Wang and Gupta [29]*	55.6	63.1	32.8 [†]	47.2	26.0 [†]	35.4 [†]
Doersch <i>et al.</i> [25]*	55.1	65.3	–	51.1	–	–
Bojanowski and Joulin [19]*	56.7	65.3	33.7 [†]	49.4	26.7 [†]	37.1 [†]
Zhang <i>et al.</i> [28]*	61.5	65.9	43.4 [†]	46.9	35.8 [†]	35.6
Zhang <i>et al.</i> [43]*	63.0	67.1	–	46.7	–	36.0
Noroozi and Favaro [26]	–	67.6	–	53.2	–	37.6
Noroozi <i>et al.</i> [45]	–	67.7	–	51.4	–	36.6
DeepCluster	70.4	73.7	51.4	55.4	43.2	45.1

대규모 데이터 셋에서 세 가지 downstream task에 대해 성능 평가

→ 모든 task에서 DeepCluster(제안 방법론)가 기존의 unsupervised 방법론의 성능을 능가함

Summary

DeepCluster

Clustering 알고리즘을 사용한 unsupervised representation learning 기법인 DeepCluster를 제안

- Feature vector 들에 대한 clustering과 cluster assignment를 예측하여 convnet의 가중치를 업데이트하는 것을 반복
 - ✓ Convnet의 구조적 강점을 극대화하여 유의미한 feature 추출
 - ✓ Trivial solution을 해결하기 위해 두 가지 방법 적용
- 대규모 데이터셋에서의 transfer task에서 기존 unsupervised 방법론보다 우수한 성능을 보임

Key point

Representation learning에 있어 clustering을 통해 convnet의 구조적인 이점을 최대한 활용할 수 있음

3-1

Unifying

contrastive learning and clustering

- to encode semantic structures

PROTOTYPICAL CONTRASTIVE LEARNING OF UNSUPERVISED REPRESENTATIONS

(ICLR 2021)

Published as a conference paper at ICLR 2021

PROTOTYPICAL CONTRASTIVE LEARNING OF UNSUPERVISED REPRESENTATIONS

Junnan Li, Pan Zhou, Caiming Xiong, Steven C.H. Hoi
Salesforce Research
{junnan.li, pzhou, cxiong, shoi}@salesforce.com

ABSTRACT

This paper presents Prototypical Contrastive Learning (PCL), an unsupervised representation learning method that bridges contrastive learning with clustering. PCL not only learns low-level features for the task of instance discrimination, but more importantly, it encodes semantic structures discovered by clustering into the learned embedding space. Specifically, we introduce prototypes as latent variables to help find the maximum-likelihood estimation of the network parameters in an Expectation-Maximization framework. We iteratively perform E-step as finding the distribution of prototypes via clustering and M-step as optimizing the network via contrastive learning. We propose ProtoNCE loss, a generalized version of the InfoNCE loss for contrastive learning, which encourages representations to be closer to their assigned prototypes. PCL outperforms state-of-the-art instance-wise contrastive learning methods on multiple benchmarks with substantial improvement in low-resource transfer learning. Code and pretrained models are available at <https://github.com/salesforce/PCL>.

Motivation

Clustering을 통한 의미론적 구조 인코딩

Limitation 1

데이터의 의미론적 구조(semantic structure)를 인코딩 하지 못함

Limitation 2

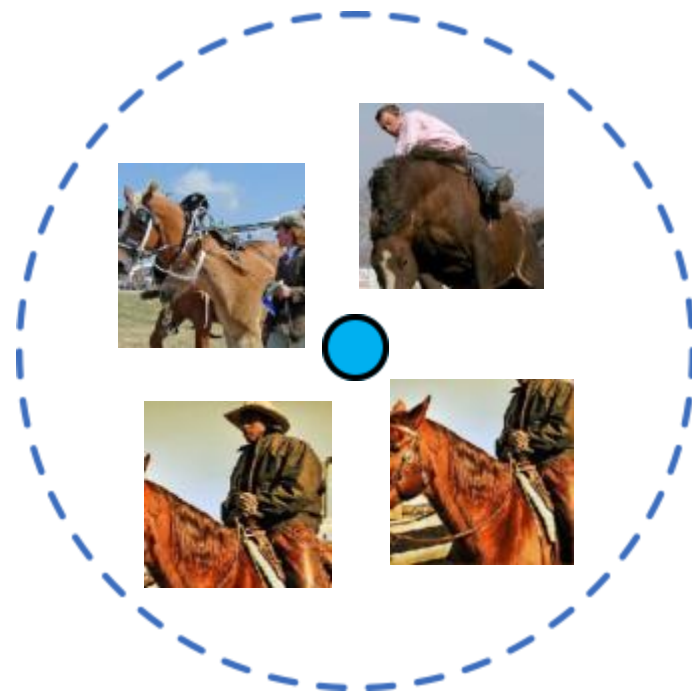
좋은 특징을 추출하기 위해 큰 batch size가 필요함

Solution

Instance discrimination 작업 뿐 아니라 학습된 임베딩 공간에서의 clustering을 통해 발견된 의미론적 구조를 인코딩함

Method

Prototypical Contrastive Learning

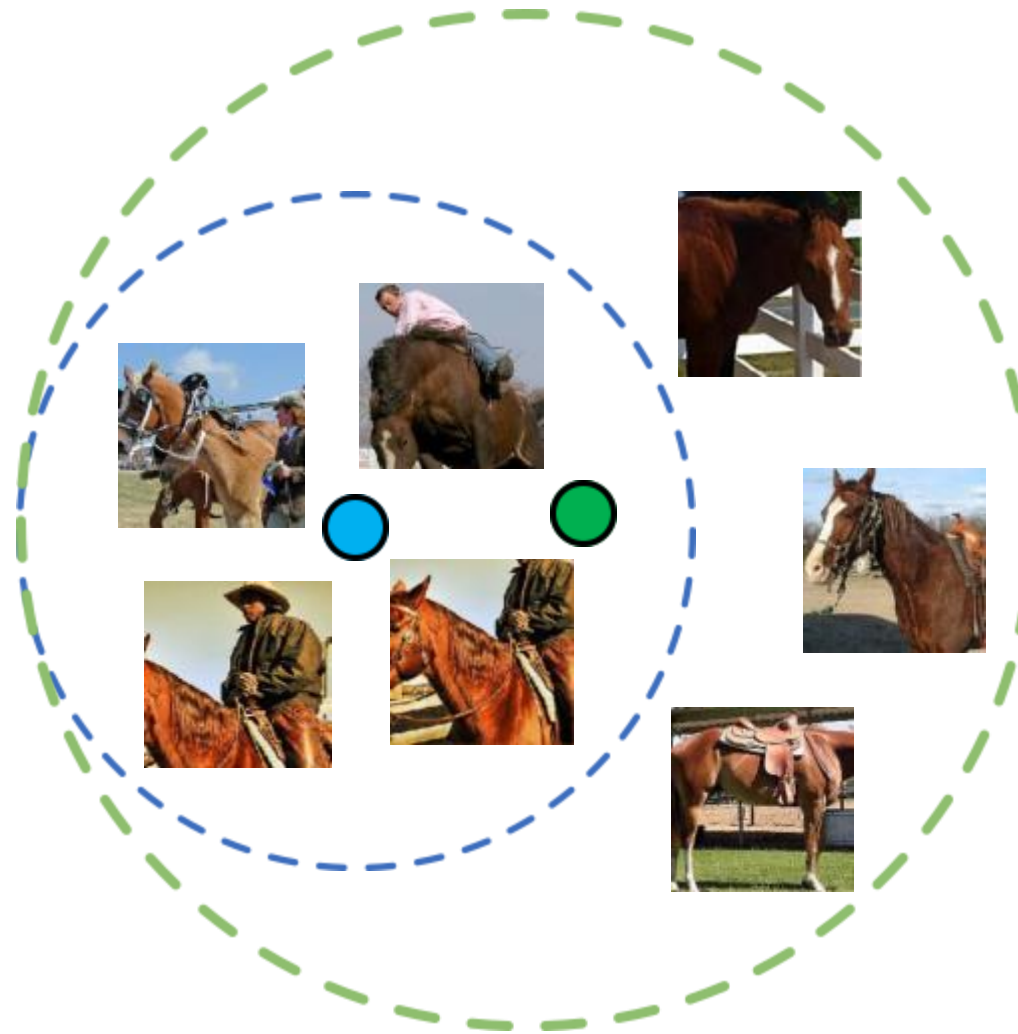


" Prototype "

의미적으로 유사한 인스턴스 그룹들의 대표적인 임베딩
= 유사한 이미지로 형성된 클러스터의 중심

Method

Prototypical Contrastive Learning

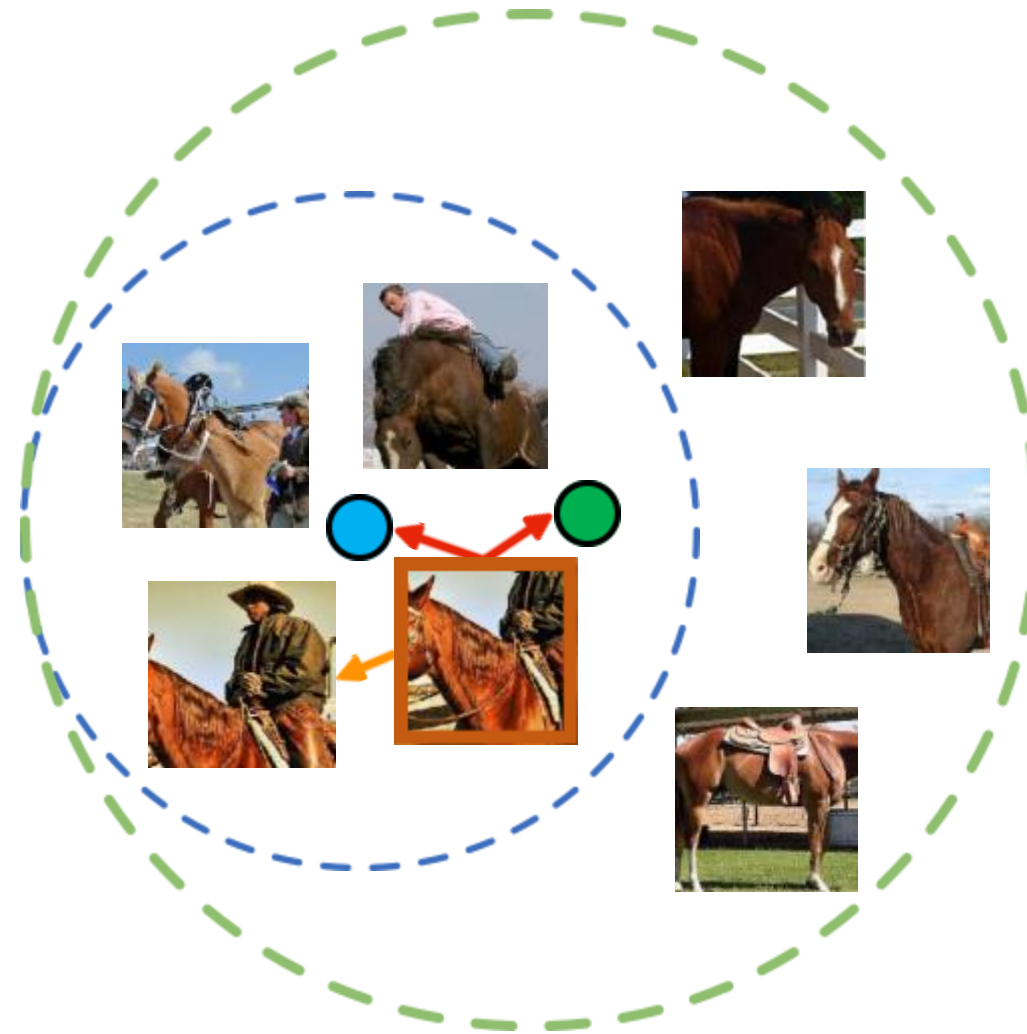


“ 각 인스턴스는 서로 다른 세분성(**granularity**)의 여러 **prototype**에 할당 될 수 있음 ”

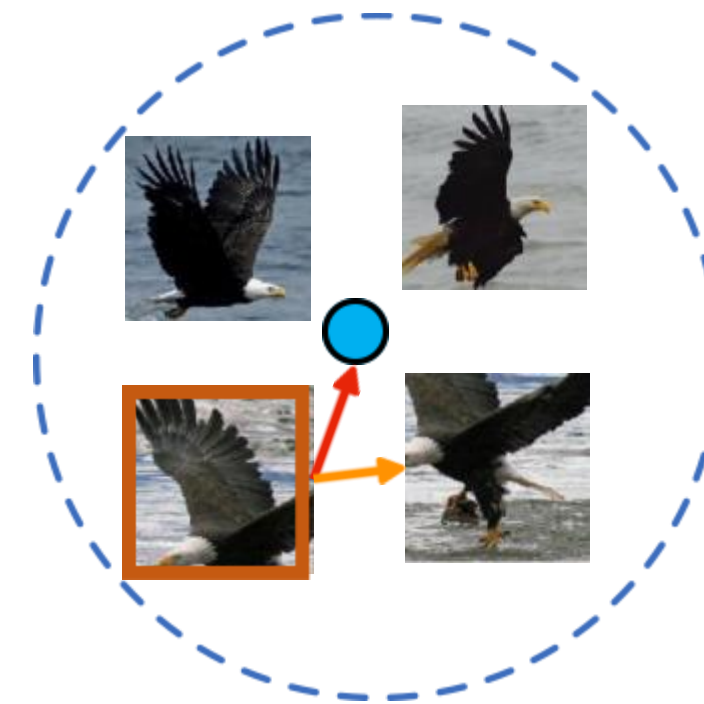
- Fine-grained prototypes (e.g. horse with man)
- Coarse-grained prototypes (e.g. horse)

Method

Prototypical Contrastive Learning



- Fine-grained prototypes (e.g. horse with man)
- Coarse-grained prototypes (e.g. horse)
- Instance-wise Contrastive Learning
- Prototypical Contrastive Learning

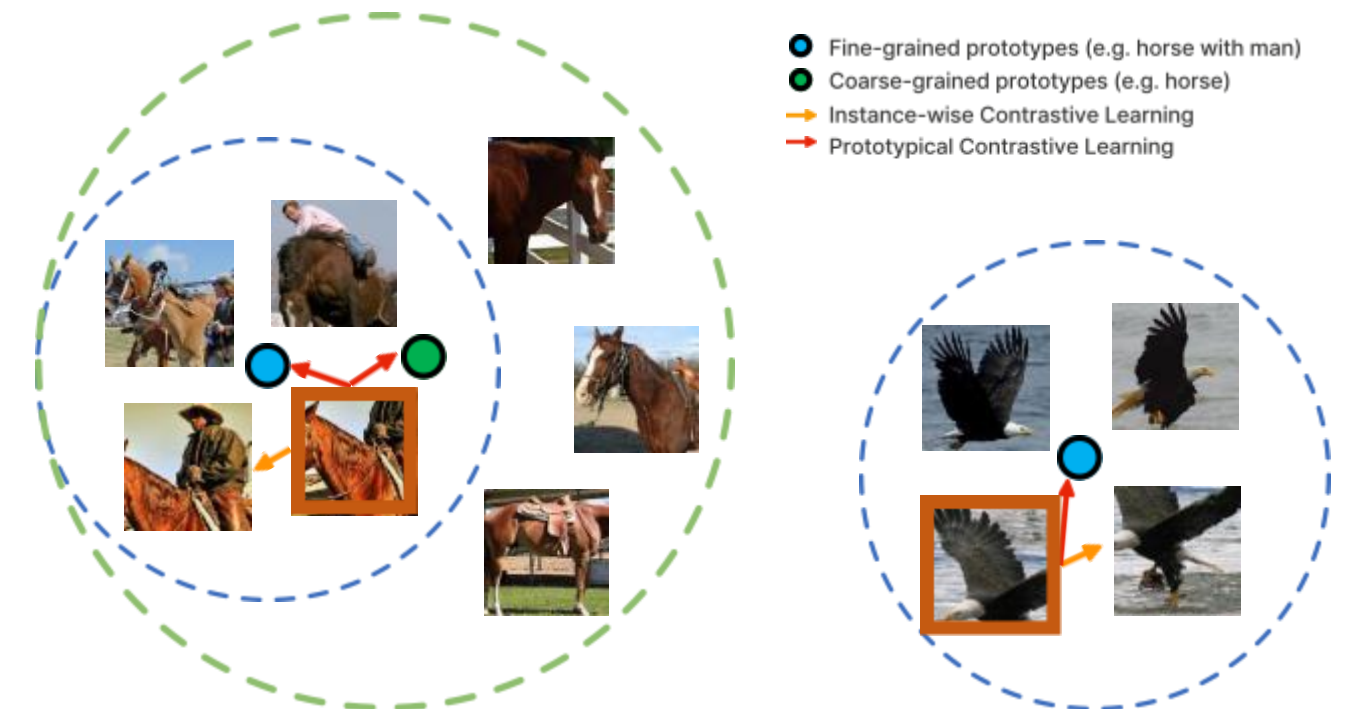


Method

Prototypical Contrastive Learning

Instance-wise contrastive learning과 prototypical contrastive learning 함께 수행

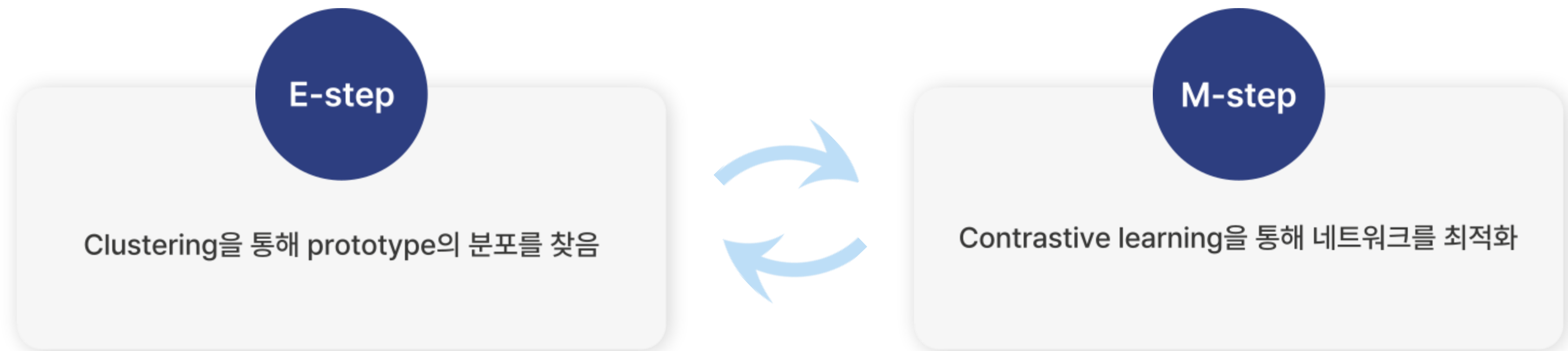
- Instance-wise contrastive learning: Instance끼리 비교하여 밀고 당기며 학습
 - ✓ Instance discrimination 작업을 통해 low-level feature 학습
- Prototypical contrastive learning: Instance, prototype 간 비교하여 밀고 당기며 학습
 - ✓ Clustering을 통해 발견된 의미론적 구조(semantic structures)를 인코딩



Method

PCL as Expectation-Maximization

- Log-likelihood function을 반복적으로 근사하고 최대화하여 데이터 분포를 가장 잘 설명하는 심층 신경망의 매개변수를 찾는 것이 목표인 Expectation-Maximization (EM) 알고리즘으로 prototypical contrastive learning을 구현함



Method

PCL as Expectation-Maximization (1) Problem Definition

- EM 알고리즘의 목표: log-likelihood function을 반복적으로 근사하고 최대화하여 데이터 분포를 가장 잘 설명하는 심층 신경망의 매개변수를 찾는 것

Maximum Likelihood Estimation(최대우도법)

- ✓ 다음과 같이 5개의 데이터가 주어졌을 때, 데이터는 정규 분포에서 추출되었다고 가정
- ✓ 목표: 데이터에 대해 정규분포를 가정했을 때, 주어진 데이터를 가장 잘 설명하는 분포의 파라미터(eg. 평균, 분산)를 찾는 것



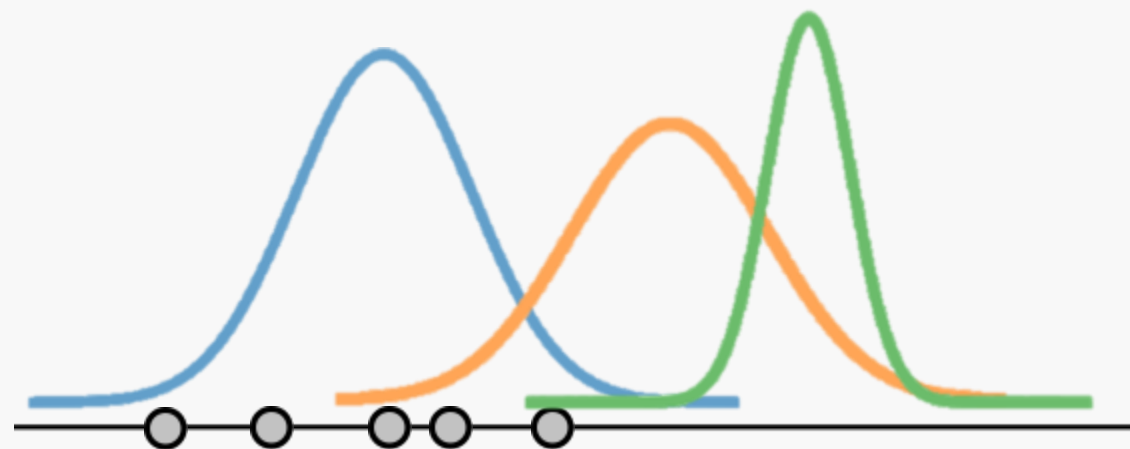
Method

PCL as Expectation-Maximization (1) Problem Definition

- EM 알고리즘의 목표: log-likelihood function을 반복적으로 근사하고 최대화하여 데이터 분포를 가장 잘 설명하는 심층 신경망의 매개변수를 찾는 것

Maximum Likelihood Estimation(최대우도법)

- ✓ 다음과 같이 5개의 데이터가 주어졌을 때, 데이터는 정규 분포에서 추출되었다고 가정
- ✓ 목표: 데이터에 대해 정규분포를 가정했을 때, 주어진 데이터를 가장 잘 설명하는 분포의 파라미터(eg. 평균, 분산)를 찾는 것



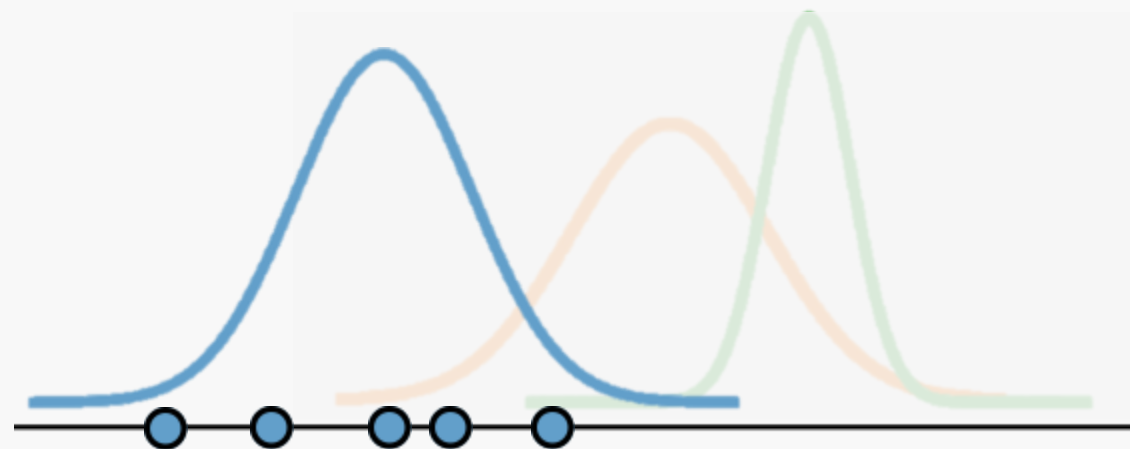
Method

PCL as Expectation-Maximization (1) Problem Definition

- EM 알고리즘의 목표: log-likelihood function을 반복적으로 근사하고 최대화하여 데이터 분포를 가장 잘 설명하는 심층 신경망의 매개변수를 찾는 것

Maximum Likelihood Estimation(최대우도법)

- ✓ 다음과 같이 5개의 데이터가 주어졌을 때, 데이터는 정규 분포에서 추출되었다고 가정
- ✓ 목표: 데이터에 대해 정규분포를 가정했을 때, 주어진 데이터를 가장 잘 설명하는 분포의 파라미터(eg. 평균, 분산)를 찾는 것



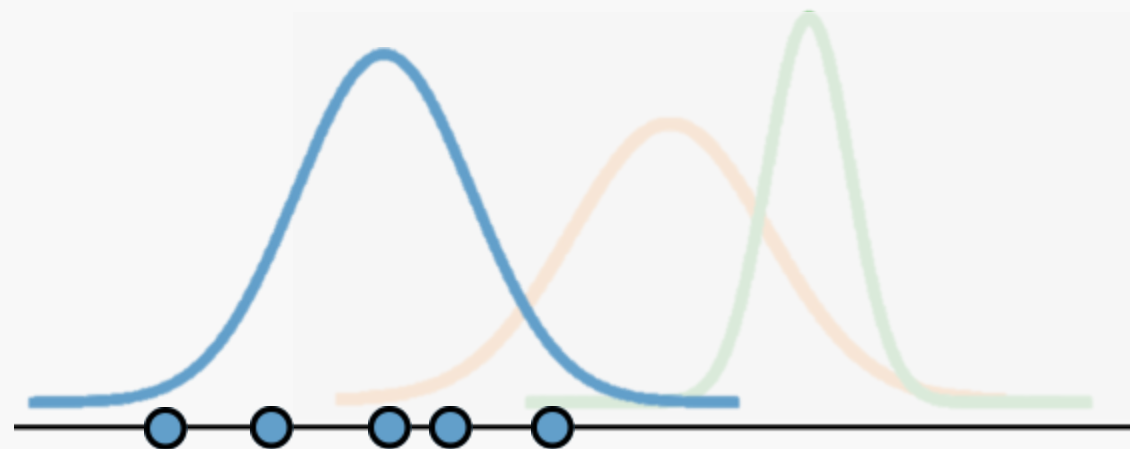
Method

PCL as Expectation-Maximization (1) Problem Definition

- EM 알고리즘의 목표: log-likelihood function을 반복적으로 근사하고 최대화하여 데이터 분포를 가장 잘 설명하는 심층 신경망의 매개변수를 찾는 것

Maximum Likelihood Estimation(최대우도법)

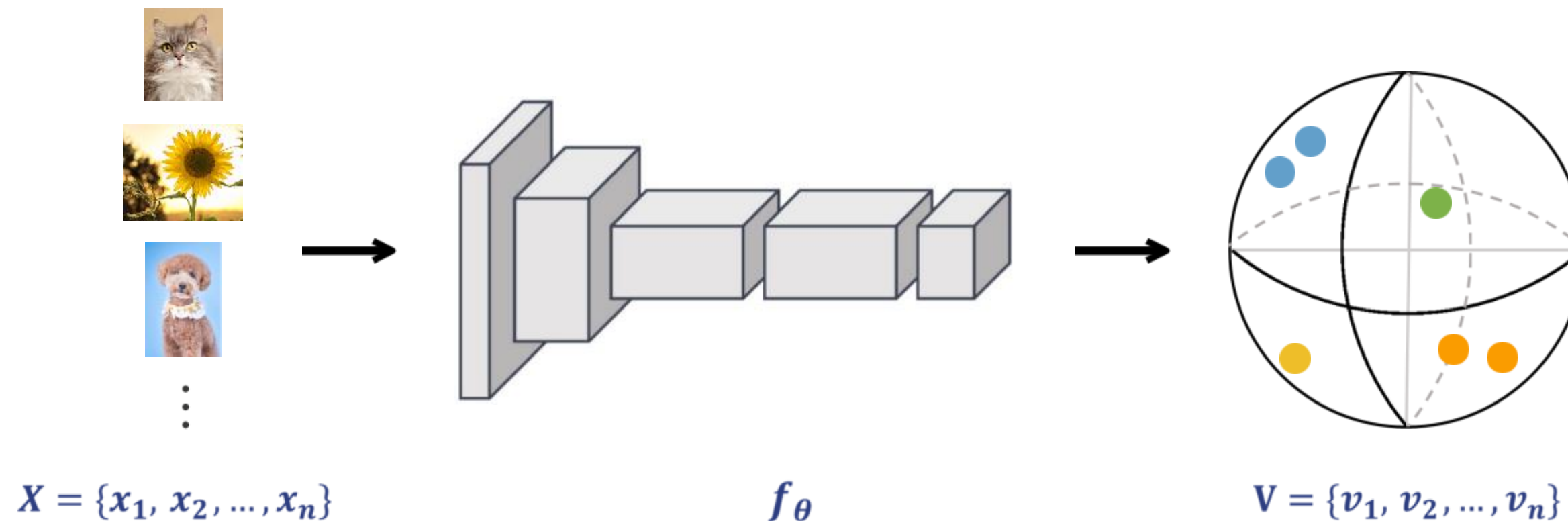
- ✓ 목표: 데이터에 대해 정규분포를 가정했을 때, 주어진 데이터를 가장 잘 설명하는 분포의 파라미터(eg. 평균, 분산)를 찾는 것
- ✓ Likelihood function이 최대가 되게하는 파라미터를 찾으면 됨!



Method

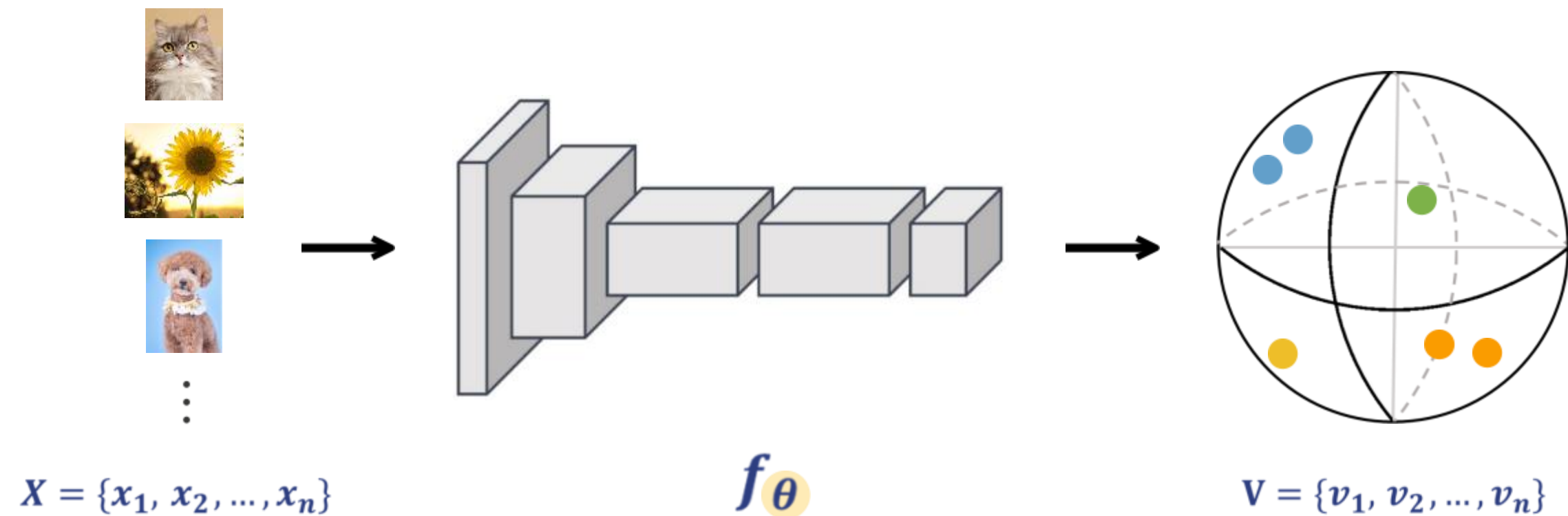
PCL as Expectation-Maximization (1) Problem Definition

- Unsupervised visual representation learning에서는 n 개의 이미지 $X = \{x_1, x_2, \dots, x_n\}$ 가 주어졌을 때, x_i 를 $v_i = f_\theta(x_i)$ 로 $V = \{v_1, v_2, \dots, v_n\}$ 에 매핑하는 임베딩 함수 f_θ 를 학습하는 것을 목표로 하며 v_i 는 x_i 를 가장 잘 설명해야 함



Method

PCL as Expectation-Maximization (1) Problem Definition



Goal

데이터 x 가 주어졌을 때 데이터의 분포를 가장 잘 설명하는 심층 신경망의 매개 변수를 찾는 것
 = 주어진 n 개의 데이터의 log likelihood function을 최대화하는 심층 신경망의 매개 변수 θ 를 찾는 것

Method

PCL as Expectation-Maximization (1) Problem Definition

Goal

주어진 n 개의 데이터의 log likelihood function을 최대화하는 심층 신경망의 매개 변수 θ 를 찾는 것

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i; \theta)$$

$p(x_i; \theta)$ 를 직접 계산하는 것은 매우 까다로움

" 마음대로 **latent variable**을 정의해서 더 풀기 쉬운 문제로 만들자! "

Method

PCL as Expectation-Maximization (1) Problem Definition

Latent variable $C = \{c_i\}_{i=1}^k$ 도입

데이터의 prototype들 (k는 하이퍼파라미터)

" 주어진 데이터들은 latent variable과 관련이 있다고 가정 "

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log \sum_{c_i \in C} p(x_i, c_i; \theta)$$

Marginal distribution $p(x_i; \theta)$ 보다 계산이 훨씬 쉬운 join distribution $p(x_i, c_i; \theta)$ 으로 만듦!

Method

PCL as Expectation-Maximization (1) Problem Definition

수식 정리

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log \sum_{c_i \in C} p(x_i, c_i; \theta)$$

$$\sum_{i=1}^n \log \sum_{c_i \in C} p(x_i, c_i; \theta) = \sum_{i=1}^n \log \sum_{c_i \in C} Q(c_i) \frac{p(x_i, c_i; \theta)}{Q(c_i)} \geq \sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log \frac{p(x_i, c_i; \theta)}{Q(c_i)},$$

$$Q(c_i) = \frac{p(x_i, c_i; \theta)}{\sum_{c_i \in C} p(x_i, c_i; \theta)} = \frac{p(x_i, c_i; \theta)}{p(x_i; \theta)} = p(c_i; x_i, \theta)$$

$$\sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i; \theta)$$

Jensen's inequality에 따른 하한 정의

Jensen's inequality에서 등식을 만족할 경우 prototype의 분포

x와 관련 없는 constant 부분 제외

Method

PCL as Expectation-Maximization (1) Problem Definition

수식 정리

$$\theta^* = \arg \max_{\theta} \sum_{i=1}^n \log p(x_i; \theta) = \arg \max_{\theta} \sum_{i=1}^n \log \sum_{c_i \in C} p(x_i, c_i; \theta)$$

$$\sum_{i=1}^n \log \sum_{c_i \in C} p(x_i, c_i; \theta) = \sum_{i=1}^n \log \sum_{c_i \in C} Q(c_i) \frac{p(x_i, c_i; \theta)}{Q(c_i)} \geq \sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log \frac{p(x_i, c_i; \theta)}{Q(c_i)},$$

$$Q(c_i) = \frac{p(x_i, c_i; \theta)}{\sum_{c_i \in C} p(x_i, c_i; \theta)} = \frac{p(x_i, c_i; \theta)}{p(x_i; \theta)} = p(c_i; x_i, \theta)$$

$$\sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i; \theta)$$

Jensen's inequality에 따른 하한 정의

Jensen's inequality에서 등식을 만족할 경우 prototype의 분포

“ 최종적으로 최대화할 함수! ”

Method

PCL as Expectation-Maximization (1) Problem Definition

$$\sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i; \theta)$$

θ 와 $Q(c_i)$ 를 joint optimize하기는 어려우므로 한 variable을 고정하고 나머지를 update 한 후,
나머지 variable을 같은 방식으로 update하는 alternating method를 사용!



Expectation-Maximization(EM) 알고리즘!

Method

PCL as Expectation-Maximization (1) Problem Definition

$$\sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i; \theta)$$

Expectation-Maximization(EM) 알고리즘!

E-step

θ 를 고정하고 $Q(c_i)$ 추정



M-step

$Q(c_i)$ 를 고정하고, lower-bound를 최대화
시키는 θ 추정

Method

PCL as Expectation-Maximization (2) E-step

Goal θ 를 고정하고 $Q(c_i) = p(c_i; x_i, \theta)$ 추정

- Momentum encoder로 추출된 feature vector에 대해 k-means clustering을 수행
- i 번째 cluster의 centroid를 c_i 라고 할 때, 각 데이터 x_i 가 c_i 가 속한 cluster에 속하는지 여부로 추정

$$p(c_i; x_i, \theta) = \mathbb{I}(x_i \in c_i) \quad \mathbb{I}(x_i \in c_i): x_i \text{가 } c_i \text{가 속한 cluster에 속한다면 1 아니면 0}$$

고정된 파라미터를 가진 momentum encoder로 추출된 feature를 clustering (k-means)

Method

PCL as Expectation-Maximization (3) M-step

Goal $Q(c_i)$ 를 고정하고, lower-bound를 최대화 시키는 θ 추정

E-step에서 추정한 prototype의 분포를 고정

$$\sum_{i=1}^n \sum_{c_i \in C} Q(c_i) \log p(x_i, c_i; \theta) = \sum_{i=1}^n \sum_{c_i \in C} p(c_i; x_i, \theta) \log p(x_i, c_i; \theta) = \sum_{i=1}^n \sum_{c_i \in C} \mathbb{1}(x_i \in c_i) \log p(x_i, c_i; \theta)$$

Cluster의 centroid가 uniform prior를 따른다는 가정과 prototype 주변 분포가 isotropic Gaussian을 따른다는 가정에 MLE 추정치

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^n -\log \frac{\exp(v_i \cdot c_s / \phi_s)}{\sum_{j=1}^k \exp(v_i \cdot c_j / \phi_j)},$$

Method

PCL as Expectation-Maximization (3) M-step

Goal $Q(c_i)$ 를 고정하고, lower-bound를 최대화 시키는 θ 추정

- 샘플들을 M번 서로 다른 수 $K = \{k_m\}_{m=1}^M$ 로 clustering → 계층적 구조를 인코딩할 수 있고, prototype의 확률적 추정을 robust하게 할 수 있음
- Local smoothness를 유지하도록 InfoNCE loss 추가

➔ **최종 목적함수: ProtoNCE**

$$\theta^* = \arg \min_{\theta} \sum_{i=1}^n -\log \frac{\exp(v_i \cdot c_s / \phi_s)}{\sum_{j=1}^k \exp(v_i \cdot c_j / \phi_j)}, \quad \Rightarrow \quad \mathcal{L}_{\text{ProtoNCE}} = \sum_{i=1}^n -\left(\log \frac{\exp(v_i \cdot v'_i / \tau)}{\sum_{j=0}^r \exp(v_i \cdot v'_j / \tau)} + \frac{1}{M} \sum_{m=1}^M \log \frac{\exp(v_i \cdot c_s^m / \phi_s^m)}{\sum_{j=0}^r \exp(v_i \cdot c_j^m / \phi_j^m)} \right).$$

Estimated log-likelihood를 최대화하는 것과 ProtoNCE를 최소화하는 것은 같음 → ProtoNCE를 최소화하도록 네트워크 파라미터 학습

Method

PCL as Expectation-Maximization (3) M-step**Prototypical contrastive loss**

- Instance contrastive loss: Instance끼리 비교하여 학습
- Prototype contrastive loss: Instance, prototype 간 비교하여 학습

$$\mathcal{L}_{\text{ProtoNCE}} = \sum_{i=1}^n - \left(\underbrace{\log \frac{\exp(v_i \cdot v'_i / \tau)}{\sum_{j=0}^r \exp(v_i \cdot v'_j / \tau)}}_{\text{Instance contrastive loss}} + \underbrace{\frac{1}{M} \sum_{m=1}^M \log \frac{\exp(v_i \cdot c_s^m / \phi_s^m)}{\sum_{j=0}^r \exp(v_i \cdot c_j^m / \phi_j^m)}}_{\text{Prototype contrastive loss}} \right).$$

Method

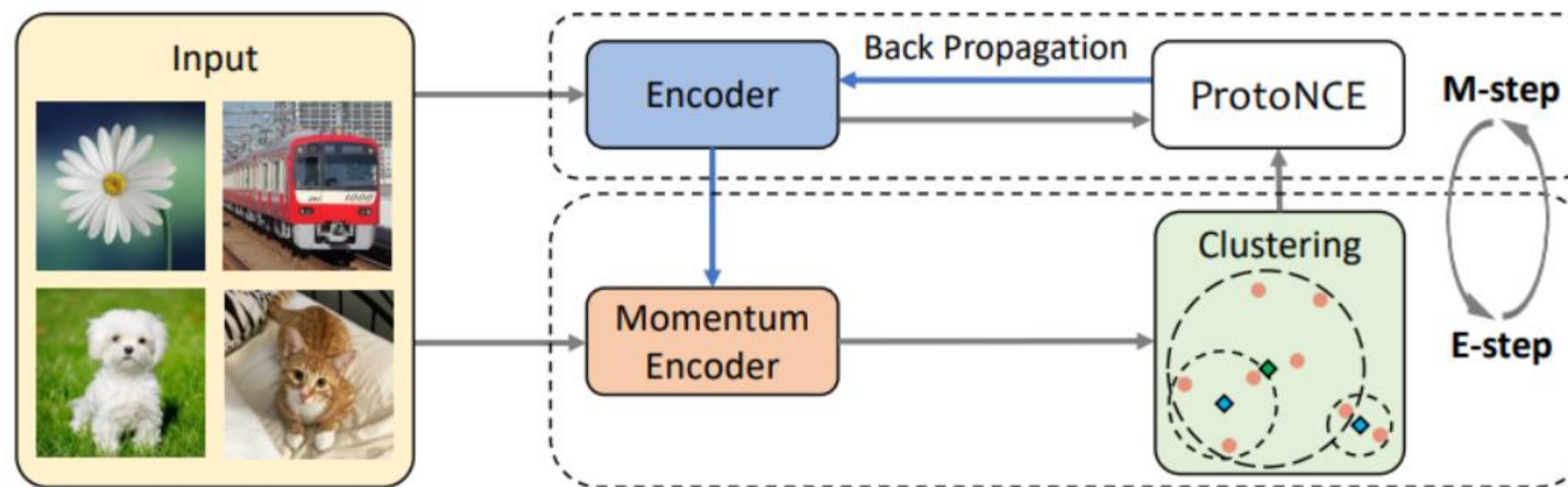
PCL as Expectation-Maximization (3) M-step

- Concentration estimation: Prototype 주변 feature 분포의 concentration level (cluster 내 밀집도)
 - ✓ InfoNCE의 temperature parameter와 유사한 역할
 - ✓ 모든 feature가 하나의 centroid에 모이는 trivial solution을 피하도록 함
 - ✓ Cluster 내 밀집도에 따라 scaling이 되므로 결과적으로는 비슷한 밀집도를 가지는 balanced cluster들이 생기게 됨

$$\mathcal{L}_{\text{ProtoNCE}} = \sum_{i=1}^n - \left(\underbrace{\log \frac{\exp(v_i \cdot v'_i / \tau)}{\sum_{j=0}^r \exp(v_i \cdot v'_j / \tau)}}_{\text{Instance contrastive loss}} + \underbrace{\frac{1}{M} \sum_{m=1}^M \log \frac{\exp(v_i \cdot c_s^m / \phi_s^m)}{\sum_{j=0}^r \exp(v_i \cdot c_j^m / \phi_j^m)}}_{\text{Prototype contrastive loss}} \right).$$

Method

The framework of Prototypical Contrastive Learning (PCL)



“ 미니 배치 내에서 ”

ProtoNCE loss를 최소화 하도록 모델을 학습하여
estimated log-likelihood 최대화

“ 모든 학습 데이터에 대해 ”

K-means clustering을 수행하여 prototype의
확률 추정

Experiment

Low-shot transfer learning

- Unlabeled ImageNet 데이터 셋으로 사전 훈련된 ResNet모델을 VOC07 데이터 셋의 object recognition과 Places205 데이터 셋의 scene classification의 새로운 task으로 trasfer함
- 각 downstream task에서는 매우 적은 수(k)의 labeled sample만 제공됨
- 제안 방법론인 PCL이 가장 좋은 성능을 보임

Method	architecture	VOC07					Places205				
		k=1	k=2	k=4	k=8	k=16	k=1	k=2	k=4	k=8	k=16
Random Supervised	ResNet-50	8.0	8.2	8.2	8.2	8.5	0.7	0.7	0.7	0.7	0.7
Jigsaw	ResNet-50	54.3	67.8	73.9	79.6	82.3	14.9	21.0	26.9	32.1	36.0
MoCo		26.5	31.1	40.0	46.7	51.8	4.6	6.4	9.4	12.9	17.4
PCL (ours)		31.4	42.0	49.5	60.0	65.9	8.8	13.2	18.2	23.2	28.0
SimCLR	ResNet-50-MLP	46.9	56.4	62.8	70.2	74.3	11.3	15.7	19.5	24.1	28.4
MoCo v2		32.7	43.1	52.5	61.0	67.1	9.4	14.2	19.3	23.7	28.3
PCL v2 (ours)		46.3	58.3	64.9	72.5	76.1	10.9	16.3	20.8	26.0	30.1
		47.9	59.6	66.2	74.5	78.3	12.5	17.5	23.2	28.1	32.3

Experiment

Semi-supervised learning

- 레이블이 없는 ImageNet 데이터들로 ResNet을 사전 학습하고, 레이블이 있는 ImageNet 데이터를 사용하여 classifier를 fine-tuning함
- 제안 방법론인 PCL이 다른 방법론의 성능을 큰 차이로 능가함

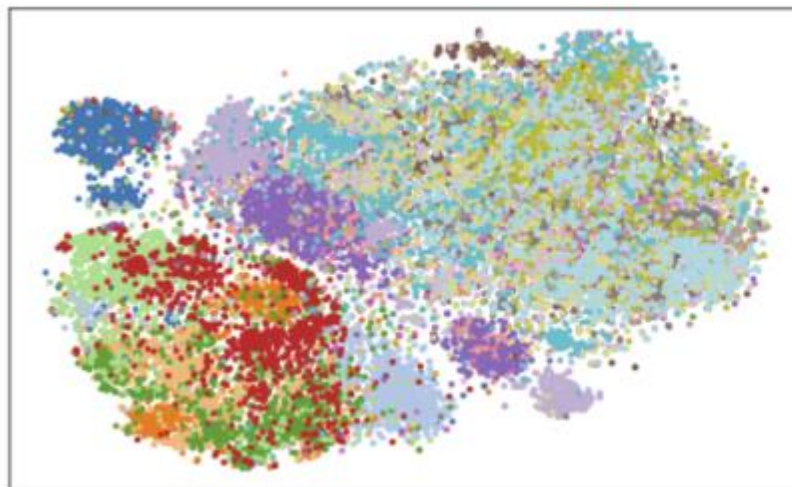
Method	architecture	#pretrain epochs	Top-5 Accuracy	
			1%	10%
Random (Wu et al., 2018)	ResNet-50	-	22.0	59.0
Supervised baseline (Zhai et al., 2019)	ResNet-50	-	48.4	80.4
<i>Semi-supervised learning methods:</i>				
Pseudolabels (Zhai et al., 2019)	ResNet-50v2	-	51.6	82.4
VAT + Entropy Min. (Miyato et al., 2019)	ResNet-50v2	-	47.0	83.4
S ⁴ L Rotation (Zhai et al., 2019)	ResNet-50v2	-	53.4	83.8
<i>Self-supervised learning methods:</i>				
Instance Discrimination (Wu et al., 2018)	ResNet-50	200	39.2	77.4
Jigsaw (Noroozi & Favaro, 2016)	ResNet-50	90	45.3	79.3
SimCLR (Chen et al., 2020a)	ResNet-50-MLP	200	56.5	82.7
MoCo (He et al., 2020)	ResNet-50	200	56.9	83.0
MoCo v2 (Chen et al., 2020b)	ResNet-50-MLP	200	66.3	84.4
PCL v2 (ours)	ResNet-50-MLP	200	73.9	85.0
PCL (ours)	ResNet-50	200	75.3	85.6
PIRL (Misra & van der Maaten, 2020)	ResNet-50	800	57.2	83.8
SimCLR Chen et al. (2020a)	ResNet-50-MLP	1000	75.5 [†]	87.8 [†]
BYOL (Grill et al., 2020)	ResNet-50-MLP _{big}	1000	78.4 [†]	89.0 [†]
SwAV (Caron et al., 2020)	ResNet-50-MLP	800	78.5 [‡]	89.9 [‡]

Experiment

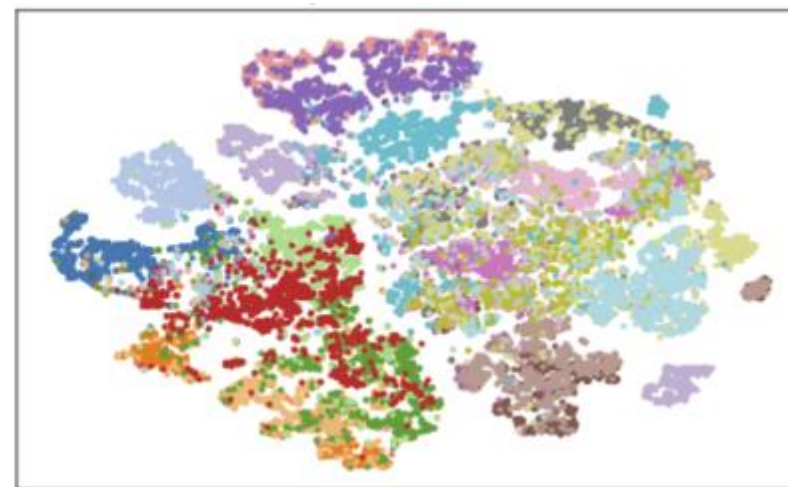
Visualization of learned representation

- MoCo를 통해 학습된 representation과 비교하여, 제안 방법론인 PCL이 학습한 representation은 더 많은 분리된 cluster를 형성하며, 동일한 클래스의 이미지가 함께 clustering 된 것을 볼 수 있음

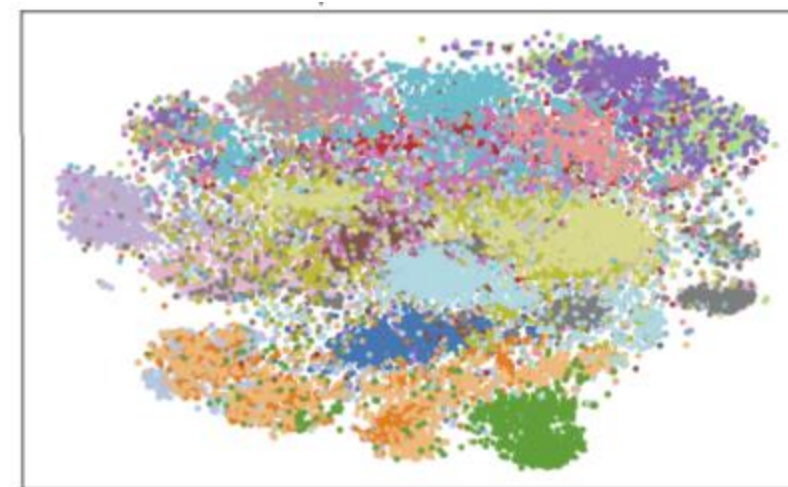
MoCo Representation: Class 1-20



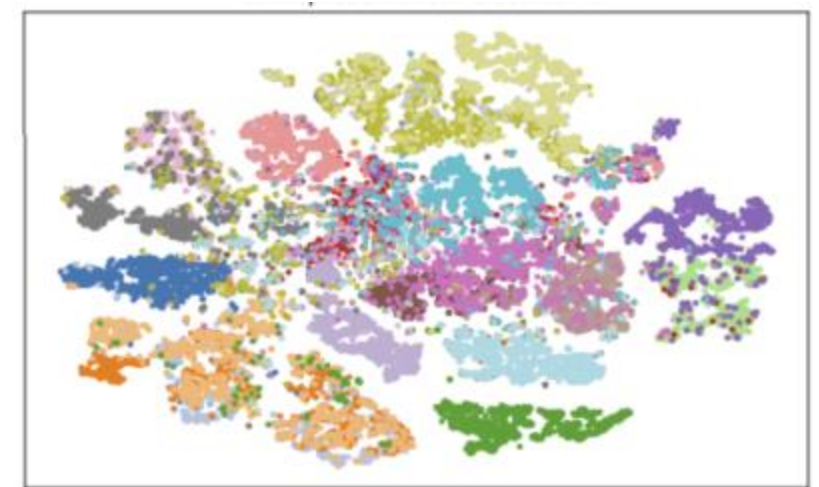
PCL Representation: Class 1-20



MoCo Representation: Class 21-40



PCL Representation: Class 21-40



Summary

PCL

Low-level feature 뿐 아니라 clustering을 통해 의미론적 구조를 인코딩하는 PCL 제안

- ProtoNCE loss
 - ✓ Instance 간 contrastive learning을 통해 low-level feature 학습
 - ✓ Instance, prototype 간 contrastive learning을 통해 high-level feature 학습
- EM 알고리즘을 통해 prototypical contrastive learning 구현

3-1

Unifying

contrastive learning and clustering

– to overcome the need for a large batch size

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

(NeurIPS 2020)

Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

Mathilde Caron^{1,2}

Ishan Misra²

Julien Mairal¹

Priya Goyal²

Piotr Bojanowski²

Armand Joulin²

¹ Inria*

² Facebook AI Research

Abstract

Unsupervised image representations have significantly reduced the gap with supervised pretraining, notably with the recent achievements of contrastive learning methods. These contrastive methods typically work online and rely on a large number of explicit pairwise feature comparisons, which is computationally challenging. In this paper, we propose an online algorithm, SwAV, that takes advantage of contrastive methods without requiring to compute pairwise comparisons. Specifically, our method simultaneously clusters the data while enforcing consistency between cluster assignments produced for different augmentations (or “views”) of the same image, instead of comparing features directly as in contrastive learning. Simply put, we use a “swapped” prediction mechanism where we predict the code of a view from the representation of another view. Our method can be trained with large and small batches and can scale to unlimited amounts of data. Compared to previous contrastive methods, our method is more memory efficient since it does not require a large memory bank or a special momentum network. In addition, we also propose a new data augmentation strategy, multi-crop, that uses a mix of views with different resolutions in place of two full-resolution views, without increasing the memory or compute requirements. We validate our findings by achieving 75.3% top-1 accuracy on ImageNet with ResNet-50, as well as surpassing supervised pretraining on all the considered transfer tasks.

Motivation

Feature간 직접 비교가 아닌 cluster 활용**Limitation 1**

데이터의 의미론적 구조(semantic structure)를 인코딩 하지 못함

Limitation 2

좋은 특징을 추출하기 위해 큰 batch size가 필요함

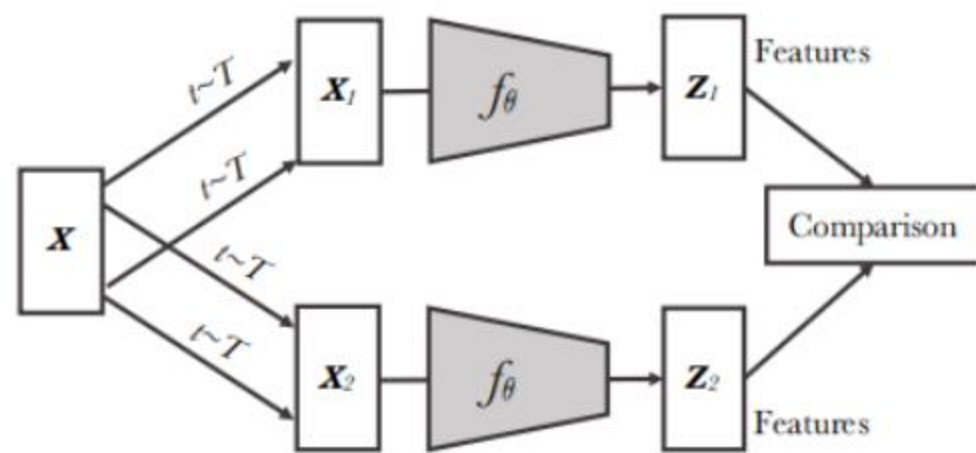
**Solution**

데이터를 clustering하는 동시에 동일한 이미지의 다른 augmentation에 대한 cluster assignment 간의 일관성 강화 (pairwise comparison이 필요 없어짐)

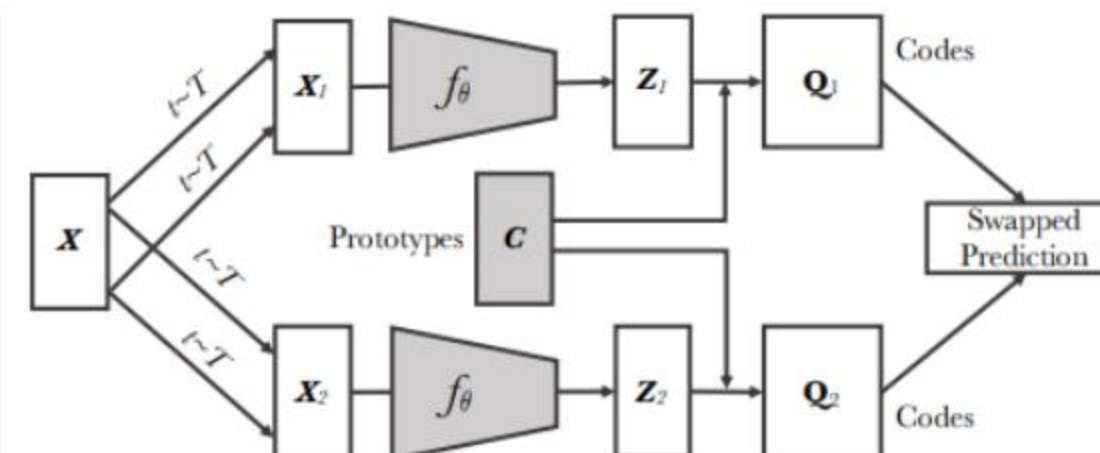
Motivation

Pair 간 feature 비교를 피하는 방법

- Contrastive instance learning에서는 동일한 이미지에 대한 서로 다른 augmentation의 feature들을 직접 비교함 → 매우 계산 집약적인 작업
- 본 논문에서 제안하는 SwAV에서는 prototype vector에 feature를 할당하여 code를 얻고, 한 데이터 augmented view에서 얻은 code를 다른 view를 사용해 예측하는 'swapped' 예측 문제를 해결함 → 이미지 feature들을 직접 비교하지 않음



Contrastive instance learning



Swapping Assignments between Views

Motivation

Architecture

Step 1

같은 이미지에서 서로 다른 augmentation의 feature vector z_1, z_2 추출

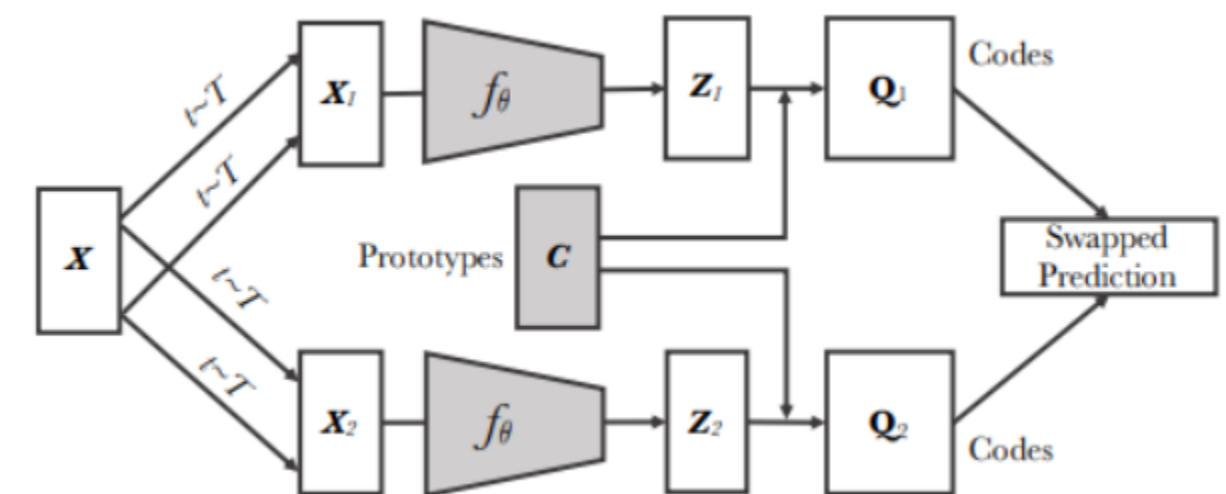
Step 2

Feature vector들을 K 개의 prototypes $\{c_1, \dots, c_k\}$ 와 매칭하여 그들의 code q_1, q_2 계산
cluster의 대표 representation

Step 3

Swapped 예측 문제 해결

: 이미지의 한 augmented 버전으로부터 code를 계산하고, 같은 이미지의 다른 augmented 버전으로부터 이 code를 예측함



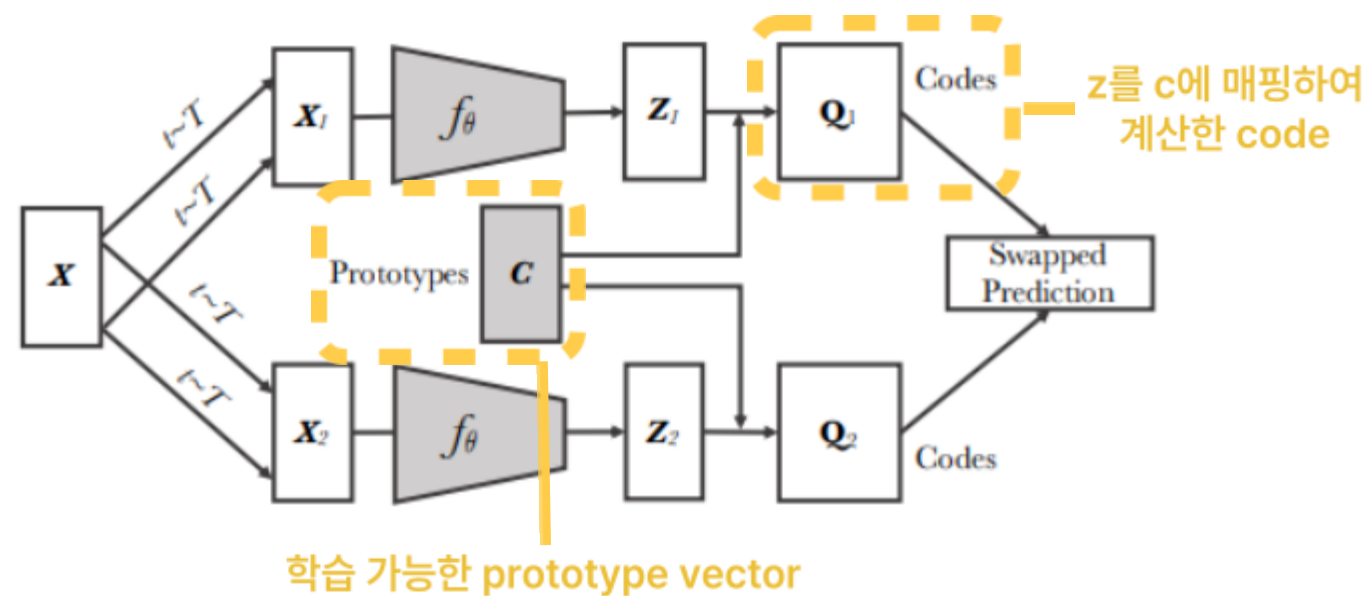
SwAV

Method

Online clustering

- 일반적인 clustering 기반 방법들은 cluster 할당과 training step을 번갈아 가면서 진행하는 **offline 방식**
- 이러한 방법은 모든 데이터에 대해 feature를 매번 새로 추출해야 하기 때문에 **target이 계속 변하는 online 학습에는 실용적이지 않음**

➡ **Online clustering**-based self-supervised method 제안!



어떻게 online 방식으로 q 를 계산하고, $\{c_1, \dots, c_K\}$ 를 업데이트 할 수 있을까?

" via **Swapped Prediction Problem** "

Method

Online clustering (1) Swapped prediction problem

- Feature인 z_t, z_s 를 직접 비교하지 않고, code인 q_t, q_s 를 통해 비교
 - 같은 이미지로부터 나온 z_t, z_s 가 비슷한 정보를 가지고 있다면 z_t 로 q_s 를 예측하는 것이 가능할 것이라는 가정
- ➡ 같은 이미지에서 파생된 augmented 이미지 code(cluster)들의 일관성을 강화하는 방향으로 학습

Loss function

$$L(\mathbf{z}_t, \mathbf{z}_s) = \underbrace{\ell(\mathbf{z}_t, \mathbf{q}_s)}_{\text{Feature } z_t \text{로부터 code } q_s \text{ 예측}} + \underbrace{\ell(\mathbf{z}_s, \mathbf{q}_t)}_{\text{Feature } z_s \text{로부터 code } q_t \text{ 예측}}$$

(각 term은 CE loss) $\ell(\mathbf{z}_t, \mathbf{q}_s) = - \sum_k \mathbf{q}_s^{(k)} \log \mathbf{p}_t^{(k)}, \quad \text{where} \quad \mathbf{p}_t^{(k)} = \frac{\exp(\frac{1}{\tau} \mathbf{z}_t^\top \mathbf{c}_k)}{\sum_{k'} \exp(\frac{1}{\tau} \mathbf{z}_t^\top \mathbf{c}_{k'})}$

Method

Online clustering (1) Swapped prediction problem

- Feature인 z_t, z_s 를 직접 비교하지 않고, code인 q_t, q_s 를 통해 비교
 - 같은 이미지로부터 나온 z_t, z_s 가 비슷한 정보를 가지고 있다면 z_t 로 q_s 를 예측하는 것이 가능할 것이라는 가정
- ➡ 같은 이미지에서 파생된 augmented 이미지 code(cluster)들의 일관성을 강화하는 방향으로 학습

Total loss function

(over all image & pair of data augmentation)

$$-\frac{1}{N} \sum_{n=1}^N \sum_{s,t \sim \mathcal{T}} \left[\frac{1}{\tau} \mathbf{z}_{nt}^\top \mathbf{C} \mathbf{q}_{ns} + \frac{1}{\tau} \mathbf{z}_{ns}^\top \mathbf{C} \mathbf{q}_{nt} - \log \sum_{k=1}^K \exp \left(\frac{\mathbf{z}_{nt}^\top \mathbf{c}_k}{\tau} \right) - \log \sum_{k=1}^K \exp \left(\frac{\mathbf{z}_{ns}^\top \mathbf{c}_k}{\tau} \right) \right]$$

" Feature extractor의 파라미터와 prototype vector C에 의해 최적화됨 "

Method

Online clustering (2) Computing codes online

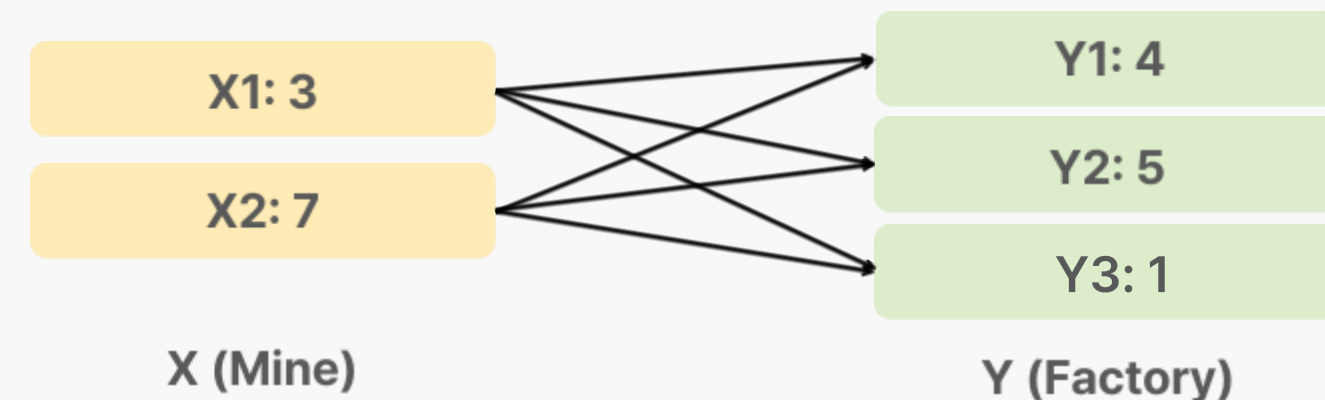
- Prototype C는 서로 다른 batch에서 공유됨
- Code는 한 배치 내에서 계산되며, 모든 sample이 같은 code로 mapping되는 trivial solution을 방지하기 위해 batch 안의 서로 다른 sample들이 prototype C에 의해 균등하게 분배되도록 함

Transportation problem

- ✓ Cost를 고려하여 X에서 Y로 가는 최적의 transportation plan를 찾는 문제
- ✓ 목표: cost를 최소화하는 transportation plans를 찾는 것!

→ Optimal transportation problem

→ Sinkhorn Algorithm



Method

Online clustering (2) Computing codes online

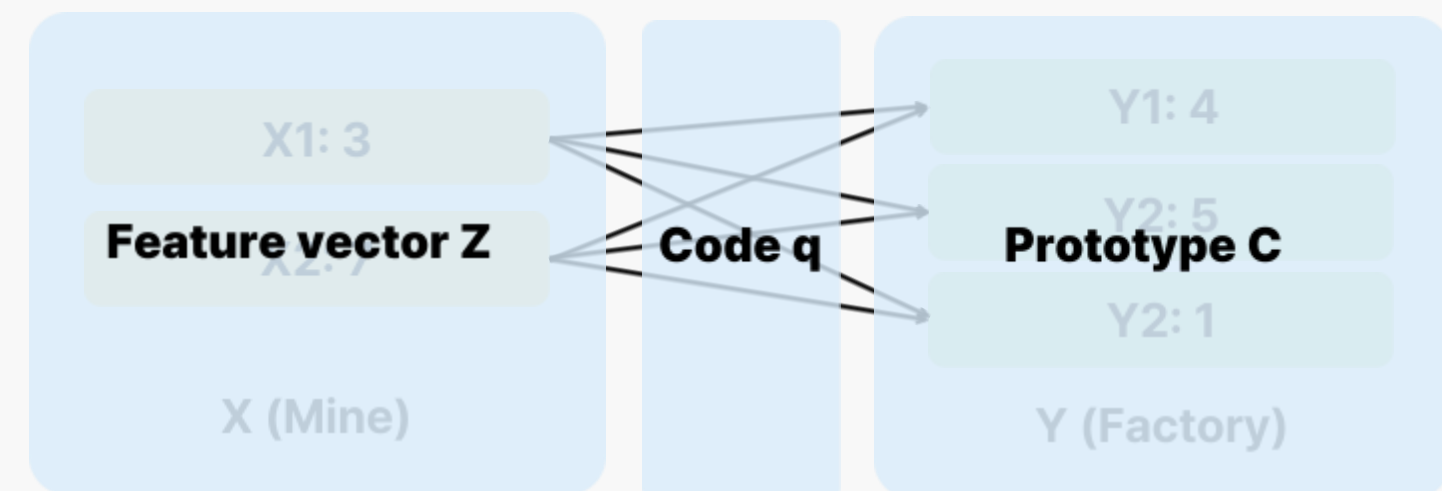
- Prototype C는 서로 다른 batch에서 공유됨
- Code는 한 배치 내에서 계산되며, 모든 sample이 같은 code로 mapping되는 trivial solution을 방지하기 위해 batch 안의 서로 다른 sample들이 prototype C에 의해 균등하게 분배되도록 함

Transportation problem

- ✓ Cost를 고려하여 X에서 Y로 가는 최적의 transportation plan를 찾는 문제
- ✓ 목표: cost를 최소화하는 transportation plans를 찾는 것!

→ Optimal transportation problem

→ Sinkhorn Algorithm



Method

Online clustering (2) Computing codes online

- Feature vectors: $\mathbf{Z} = [z_1, \dots, z_B]$
- Codes: $\mathbf{Q} = [q_1, \dots, q_B]$
- Prototype vectors: $\mathbf{C} = [c_1, \dots, c_K]$
- \mathbf{Q} : \mathbf{Z} 와 \mathbf{C} 를 연결하는 code matrix (transportation plan)
- $\mathbf{C}^T \mathbf{Z}$: (negative) cost matrix (cost)
- H : Entropy function
- ε : Mapping의 smoothness를 조절하는 파라미터
 - 값이 크면 모든 sample들이 unique representation으로 모이는 trivial solution이 발생할 수 있음

$$\max_{\mathbf{Q} \in \mathcal{Q}} \text{Tr}(\mathbf{Q}^T \mathbf{C}^T \mathbf{Z}) + \varepsilon H(\mathbf{Q})$$

$$\mathcal{Q} = \left\{ \mathbf{Q} \in \mathbb{R}_+^{K \times B} \mid \mathbf{Q} \mathbf{1}_B = \frac{1}{K} \mathbf{1}_K, \mathbf{Q}^T \mathbf{1}_K = \frac{1}{B} \mathbf{1}_B \right\}$$

$$\mathbf{Q}^* = \text{Diag}(\mathbf{u}) \exp\left(\frac{\mathbf{C}^T \mathbf{Z}}{\varepsilon}\right) \text{Diag}(\mathbf{v})$$

최적화 식

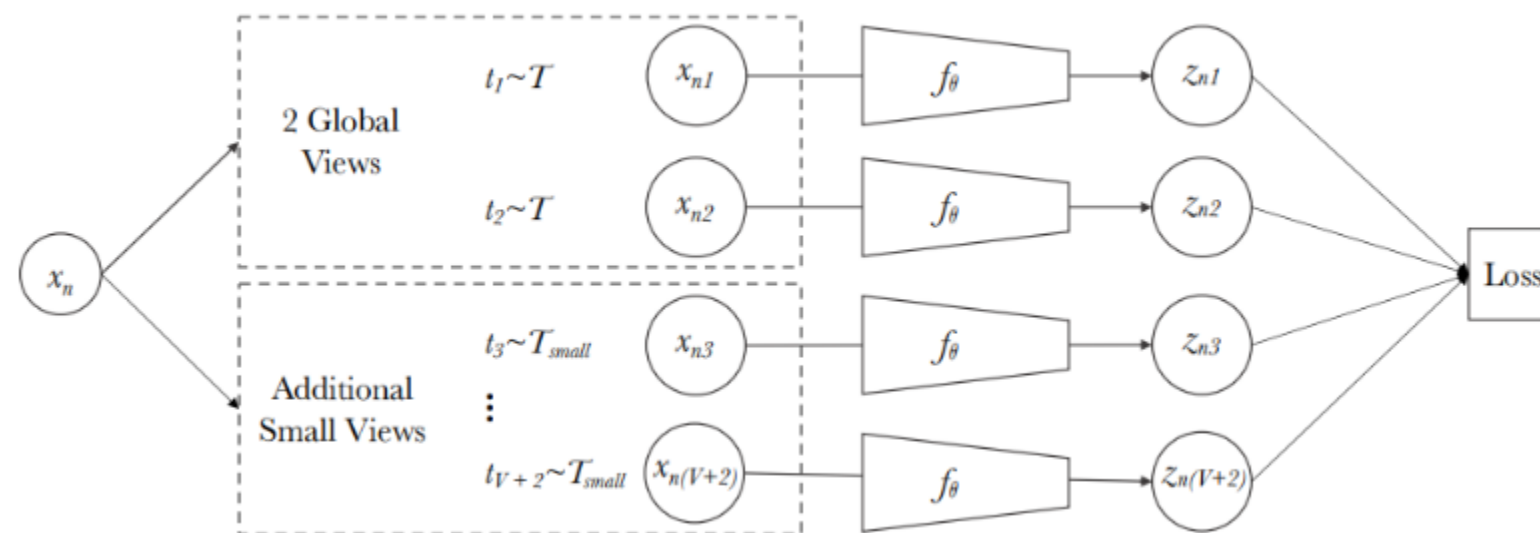
Optimal transport를 이용하기 위한 제약조건

Sinkhorn-Knopp 알고리즘을 통해 얻은 최적값 \mathbf{Q}^*

Method

Multi-crop: Augmenting views with smaller images

- 일반적으로 사용되는 augmentation 기법인 random crop은 views를 늘릴수록 메모리 및 컴퓨팅 부담이 증가함
- 이러한 문제를 해결하기 위해 2개의 일반적인 random crop을 수행한 후 V 개의 해상도가 낮고 이미지의 매우 작은 부분만 포함하는 추가적인 crop을 생성
- $V+2$ 개의 crop에 대해 모두 code를 계산하는 것이 아니라 일반적인 2개의 random crop에 대해서만 code를 계산하고 V 개의 저해상도 crop은 code 예측을 위한 feature로만 사용
- SimCLR와 DeepCluster등에 다른 방법론에서도 multi-crop을 적용한 결과 성능 향상을 보임



$$L(\mathbf{z}_{t_1}, \mathbf{z}_{t_2}, \dots, \mathbf{z}_{t_{V+2}}) = \sum_{i \in \{1, 2\}} \sum_{v=1}^{V+2} \mathbf{1}_{v \neq i} \ell(\mathbf{z}_{t_v}, \mathbf{q}_{t_i})$$

Experiment

Transfer learning on downstream tasks

- Unlabeled ImageNet 데이터 셋으로 사전 훈련된 모델을 다른 데이터 셋의 linear classification과 objection detection task으로 trasfer함
- 두 task 모두에서 supervised learning의 성능을 뛰어넘음

	Linear Classification			Object Detection		
	Places205	VOC07	iNat18	VOC07+12	COCO	COCO
				(Faster R-CNN R50-C4)	(Mask R-CNN R50-FPN)	(DETR)
Supervised	53.2	87.5	46.7	81.3	39.7	40.8
SwAV	56.7	88.9	48.6	82.6	41.6	42.1

Experiment

Training with small batches

- 256 batch size에서 성능 실험
- Mocov2의 경우 momentum encoder를 위해 65536개의 feature를 저장하고 있어야 하지만 SwAV는 3840개의 feature를 저장
- SimCLR와 MoCov2보다 좋은 성능을 보이며 MoCov2에 비해 4배 적은 epoch 만에 72% 성능에 도달

Method	Mom. Encoder	Stored Features	multi-crop	epoch	batch	Top-1
SimCLR		0	2×224	200	256	61.9
MoCov2	✓	65,536	2×224	200	256	67.5
MoCov2	✓	65,536	2×224	800	256	71.1
SwAV		3,840	$2 \times 160 + 4 \times 96$	200	256	72.0
SwAV		3,840	$2 \times 224 + 6 \times 96$	200	256	72.7
SwAV		3,840	$2 \times 224 + 6 \times 96$	400	256	74.3

Summary

SwAV

직접 feature를 비교하는 대신 cluster assignments 간의 일관성을 강화하는 SwAV 제안

- 다른 view의 representation으로부터 view의 code를 예측하는 'swapped' prediction mechanism 사용
- 같은 이미지에서 파생된 augmented 이미지 code(cluster)들의 일관성을 강화하는 방향으로 학습
- Online 방식으로 prototype을 업데이트하고 code를 계산하는 online clustering-based self-supervised method 제안
- 새로운 augmentation 기법인 multi-crop 제안

Summary

Unifying contrastive learning and clustering

Instance-wise contrastive learning의 한계점을 극복하는 contrastive learning과 clustering을 통합한 방법론 소개

- Instance-wise contrastive learning은 다음과 같은 한계점을 지님
 - ✓ 데이터의 의미론적 구조(semantic structure)를 인코딩하지 못함
 - ✓ 좋은 representation을 추출하기 위해 큰 batch size가 필요함
- Contrastive learning과 clustering을 통합하여 위의 한계점을 극복한 방법론들이 제안됨
 - ✓ PCL: Clustering을 통해 의미론적 구조를 인코딩함
 - ✓ SwAV: Feature 간 직접 비교를 하는 것이 아닌 cluster를 활용한 비교를 통해 작은 batch size에서도 좋은 성능을 냄

감사합니다:)

Reference

- [1] 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).
- [2] Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., & Joulin, A. (2020). Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems*, 33, 9912-9924.
- [3] Li, J., Zhou, P., Xiong, C., & Hoi, S. C. (2020). Prototypical contrastive learning of unsupervised representations. *arXiv preprint arXiv:2005.04966*.