이미지 자기지도학습
I-JEPA
표현 공간에서의 예측 학습

I-JEPA의 구조와 성능

학부생 전범기

26.09.18

목 차

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

“Self-Supervised Learning From Images With a Joint-Embedding Predictive Architecture”

저자

Mahmoud Assran[123] · Quentin Duval[1] · Ishan Misra[1] · Piotr Bojanowski[1]
Pascal Vincent[1] · Michael Rabbat[13] · Yann LeCun[14] · Nicolas Ballas[1]

[1] Meta AI (FAIR) [2] McGill University
[3] Mila, Quebec AI Institute [4] New York University

학회

IEEE/CVF Conference on Computer Vision and Pattern Recognition [June 2023]

핵심

이미지의 픽셀 자체가 아니라 의미 있는 표현 자체를 학습하는 자기지도학습 방식

백그라운드

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

자기지도 학습이란?

사람이 정답(label)을 직접 붙이지 않아도, 데이터 자체에서 학습 문제를 만들어 표현을 배우는 방법

불변성 기반 학습

Invariance-based / Joint-Embedding

원본 이미지 → 서로 다른 View → 임베딩을 같게

장점 높은 의미(semantic) 수준의 표현 → 분류·전이학습에 강함

단점 수작업이 강한 편향을 만들 수 있고,
입력 데이터의 종류가 바뀌면 설계를 다시 고민해야 함

생성형·복원 기반 학습

Generative / Masked Reconstruction

원본 이미지 → 마스킹 → 픽셀·토큰 복원

장점 증강 설계 의존이 적고, 마스킹 원리는 여러 모달리티로 확장 용이

단점 픽셀 수준 세부정보까지 복원하느라 의미 수준이 낮아질 수 있고,
좋은 다운스트림 성능을 위해 full fine-tuning이 필요한 경우가 많음

자기지도 학습의 분류

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

Overview : I - JEPA

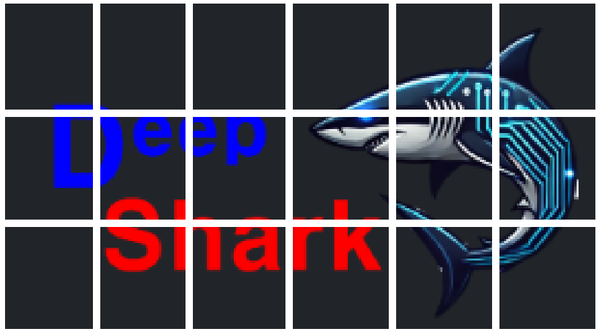
픽셀을 직접 복원하지 않고,
이미지의 의미 있는 표현을 예측한다.

I-JEPA의 구조

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

Context/Target Sampling



Input

사용자가 넣은 이미지

Context

이미지에서 인코더가 직접보고
정보를 추출하는 큰 블록
(전체 크기의 0.85~1.0 배)

Target

모델이 예측해야 하는
이미지 내의 작은 블록
(0.15~0.2 배)

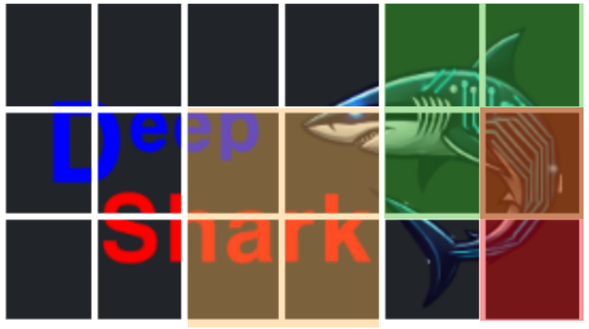
I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

Elimination (Masked Image Modeling방식, Masked AutoEncoder)

Context Block 마스크 내에서 Target Block 마스크와 겹치는 모든 영역을 제거

Context



Target

Context

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

Context Encoder

Context를 입력받아 패치 단위 표현 S_x 생성

Target Encoder

원본 이미지 전체를 입력받아 각 패치의 정답 표현 S_y 생성

Target Encoder는 Context Encoder 가중치를 지수이동평균 방식으로 복사 업데이트
→ 학습 안정성, 표현 붕괴 방지

Predictor

Context Encoder 출력값과 예측하고자 하는 위치 정보를 받아 해당 위치의 타겟 표현 예측

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

세 블록 모두 Vision Transformer를 활용합니다.

Context Encoder

Context를 입력받아 패치 단위 표현 S_x 생성

Target Encoder

원본 이미지 전체를 입력받아 각 패치의 정답 표현 S_y 생성

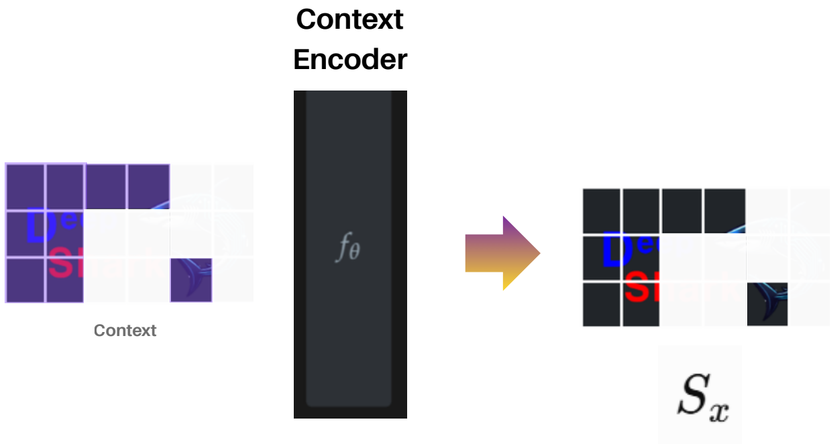
Target Encoder는 Context Encoder 가중치를 지수이동평균 방식으로 복사 업데이트
→ 학습 안정성, 표현 붕괴 방지

Predictor

Context Encoder 출력값과 예측하고자 하는 위치 정보를 받아 해당 위치의 타겟 표현 예측

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

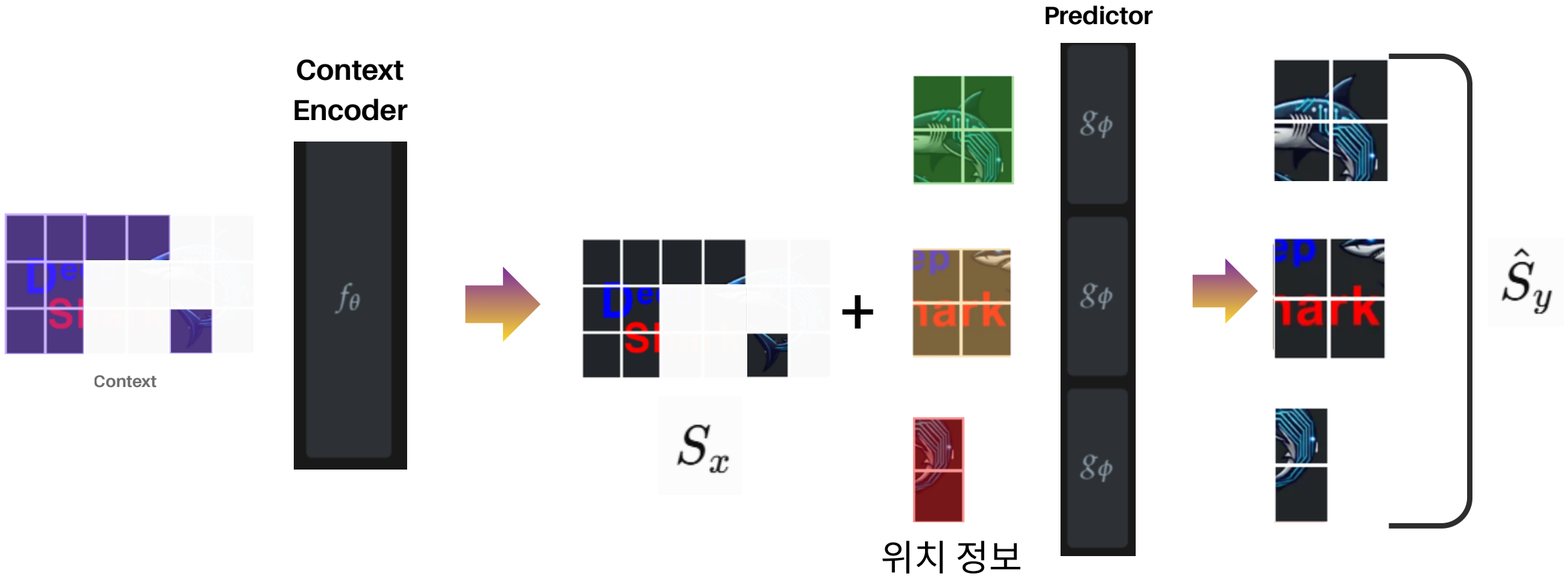


Context Encoder

Context를 입력받아 패치 단위 표현 S_x 생성

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문



Context Encoder

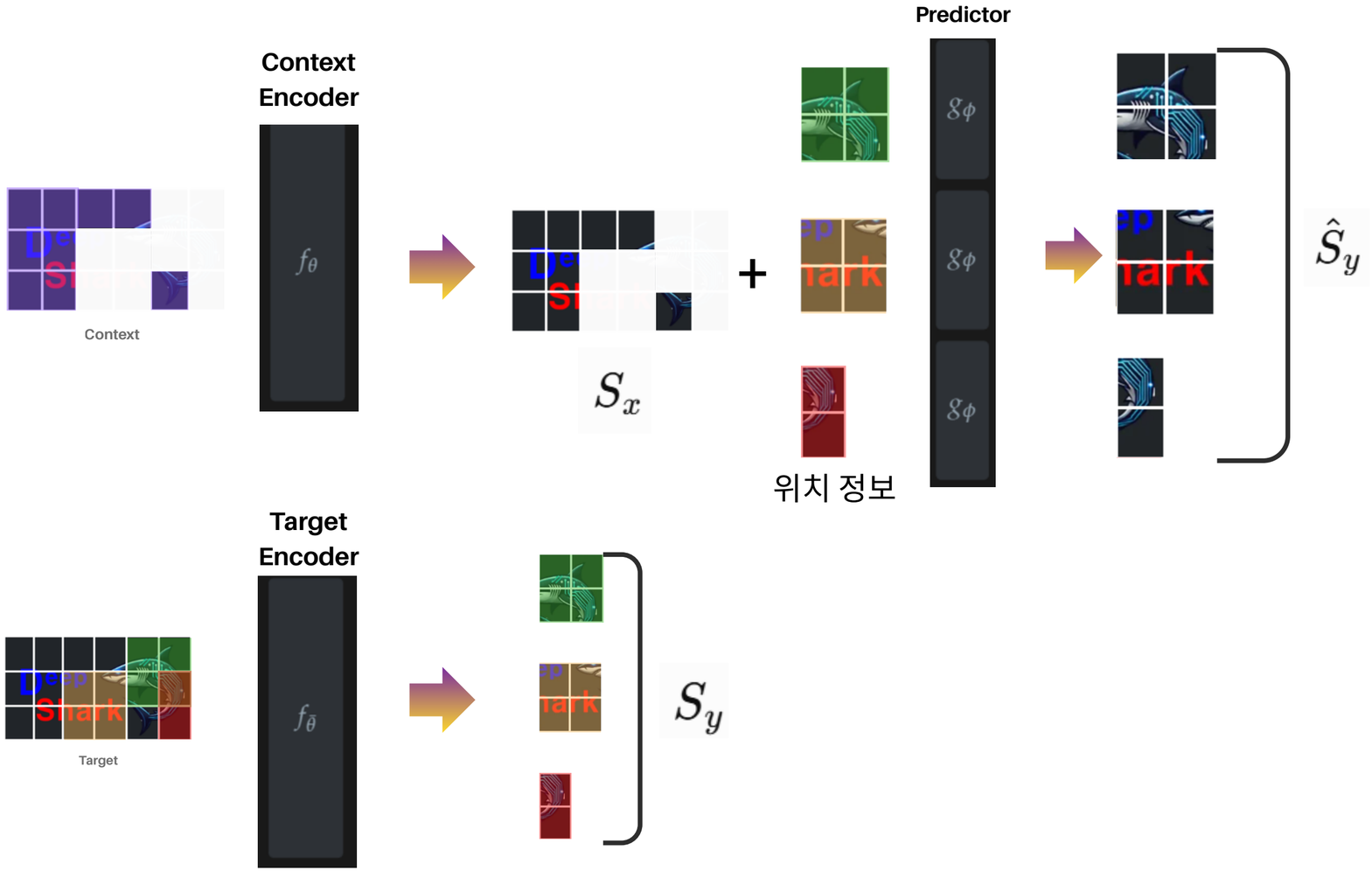
Context를 입력받아 패치 단위 표현 S_x 생성

Predictor

Context Encoder 출력값과 예측하고자 하는 위치 정보를 받아 해당 위치의 타겟 표현 예측

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문



Context Encoder

Context를 입력받아 패치 단위 표현 S_x 생성

Predictor

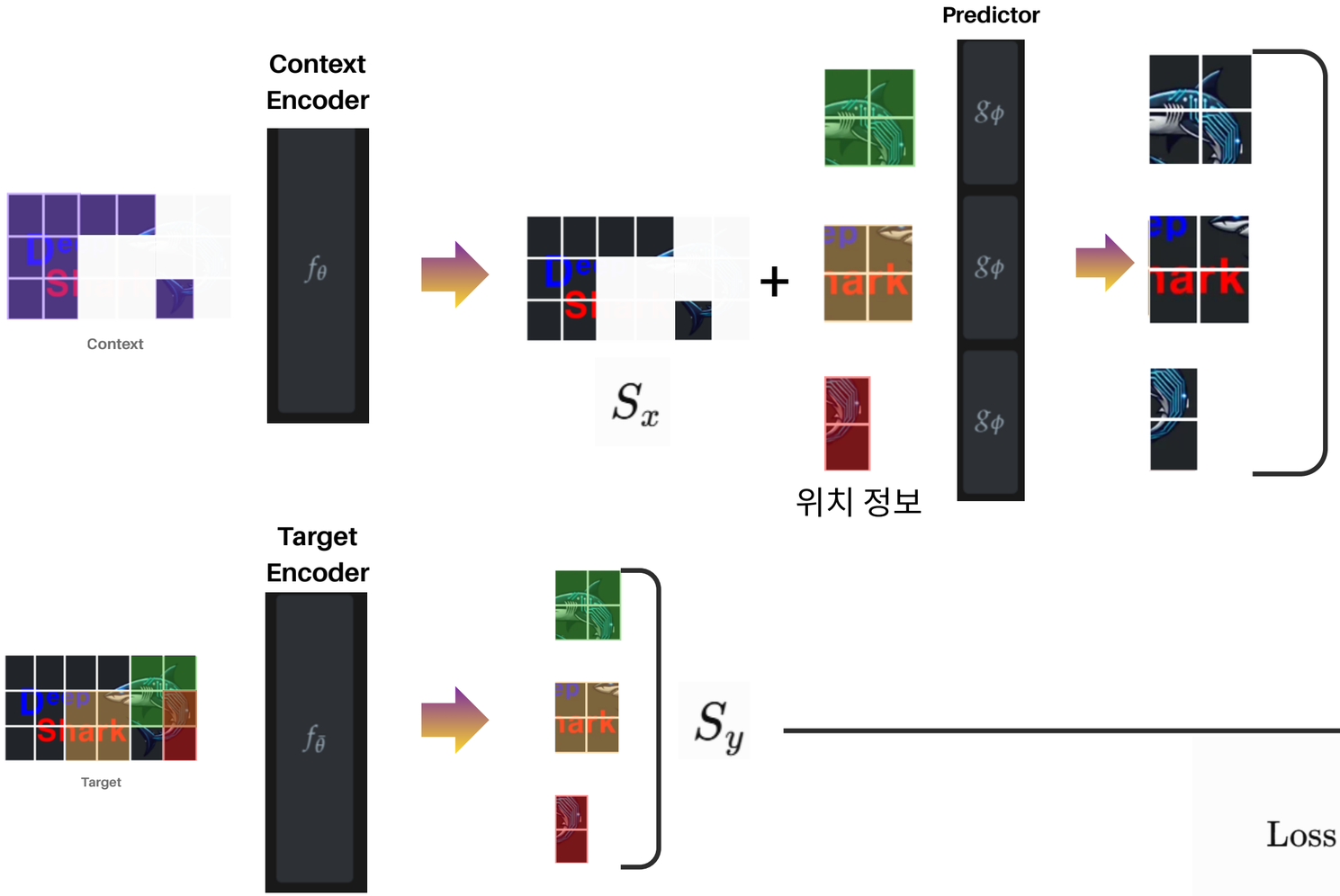
Context Encoder 출력값과 예측하고자 하는 위치 정보를 받아 해당 위치의 타겟 표현 예측

Target Encoder

원본 이미지 전체를 입력받아 각 패치의 정답 표현 S_y 생성

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문



Context Encoder

Context를 입력받아 패치 단위 표현 S_x 생성

Predictor

Context Encoder 출력값과 예측하고자 하는 위치 정보를 받아 해당 위치의 타겟 표현 예측

Target Encoder

원본 이미지 전체를 입력받아
각 패치의 정답 표현 S_y 생성

Loss

L2 거리를 최소화 하는 방향으로 학습

$$\text{Loss} = \frac{1}{M} \sum_{i=1}^{M} D\left(\hat{s}_{y(i)}, s_{y(i)}\right)$$

L_2

M=타깃 블럭의 갯수

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

구조적 특징

1. 표현 공간에서의 예측으로 의미있는 특징 학습

픽셀을 복원하려면 색상·질감·배경처럼 이미지의 의미를 이해하는데
덜 중요한 세부 정보까지 예측해야 합니다.
표현 공간에서는 이러한 정보를 추상화할 수 있어,
의미 있는 특징에 집중하도록 학습할 수 있습니다.

2. MIM 마스킹으로 표현의 품질 향상

이미지 일부를 가리면 모델은 보이는 정보를 그대로 복사하기 어렵습니다.
가려진 영역을 예측하려면 주변 영역의 특징과 관계를 활용해야 하므로,
이미지의 구조를 이해하도록 학습을 유도할 수 있습니다.
동시에 충분한 context를 제공하고 여러 target을 예측하게 하여,
다양한 영역 사이의 관계를 학습합니다.

Experiments

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

PRETRAIN

Pretraining

1K ImageNet-1K

pretraining dataset

Backbone

ViT-B/16	Base
ViT-L/16	Large
ViT-H/14	Huge

평가 절차

Linear Evaluation

사전학습된 인코더를 업데이트 하지 않고
선형 분류기만 학습한 뒤,
인코더만의 표현의 품질을 평가합니다.

Pretrained
Encoder

→

Linear
Classifier

가중치 고정

학습

Key idea
복잡한 분류기를 붙이면 분류기 자체의 능력으로 성능을 높일 수 있으므로, 선형 분류기를 학습 시켜 활용한다.

EVALUATION

Evaluation Suite

01 **ImageNet 선형 평가**
전역적 의미 표현의 품질

02 **ImageNet 1% 레이블 평가**
적은 레이블로도 유용한 표현

03 **전이 학습**
CIFAR100 · Places205 · iNat18

04 **지역 정보 예측 과제**
객체 개수 · 공간 / 깊이 정보

Baselines MAE · data2vec · DINO · iBOT

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

01

ImageNet 선형 평가

전역적 의미 표현의 품질

인코더를 고정하고, ImageNet 학습 이미지와 정답으로 선형 분류기만 학습합니다.
평가 이미지에서 정확도를 측정해 기존 표현만으로 클래스가 잘 구분되는지 확인합니다.

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [8]	ViT-L/16	1600	77.3
MAE [36]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
CAE [22]	ViT-B/16	1600	70.4
	ViT-L/16	1600	78.1
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [21]	RN152 (2×)	800	79.1
DINO [18]	ViT-B/8	300	80.1
iBOT [79]	ViT-L/16	250	**81.0**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

01

ImageNet 선형 평가

전역적 의미 표현의 품질

인코더를 고정하고, ImageNet 학습 이미지와 정답으로 선형 분류기만 학습합니다.
평가 이미지에서 정확도를 측정해 기존 표현만으로 클래스가 잘 구분되는지 확인합니다.

시사점

1. 데이터 비증강 모델 중 분류 성능이 좋음
2. 더 적은 Epoch로 성능을 얻음
3. 모델과 입력 해상도가 커짐에 따라 성능이 올라감
4. 데이터 증강 모델과 비교했을때 어느정도 경쟁력있는지 확인 가능.

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [8]	ViT-L/16	1600	77.3
MAE [36]	ViT-B/16	1600	68.0
	ViT-L/16	1600	76.0
	ViT-H/14	1600	77.2
CAE [22]	ViT-B/16	1600	70.4
	ViT-L/16	1600	78.1
I-JEPA	ViT-B/16	600	72.9
	ViT-L/16	600	77.5
	ViT-H/14	300	79.3
	ViT-H/16_{448}	300	**81.1**
Methods using extra view data augmentations			
SimCLR v2 [21]	RN152 (2×)	800	79.1
DINO [18]	ViT-B/8	300	80.1
iBOT [79]	ViT-L/16	250	**81.0**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

02

ImageNet 1% 레이블 평가

적은 레이블로도 유용한 표현

분류기 등을 학습할 때 전체 정답 레이블의 1%만 사용합니다.
클래스당 약 12~13장 수준으로, 적은 레이블 데이터로도 높은 성능을 얻는지 확인합니다.
사전학습 데이터를 1%로 줄인다는 뜻은 아닙니다.
이 평가는 방법에 따라 선형 평가 또는 fine-tuning을 사용합니다
→ Finetune: 1% 레이블로 인코더를 50 epochs fine-tuning

Finetune Not Finetune

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [8]	ViT-L/16	1600	73.3
MAE [36]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [79]	ViT-B/16	400	69.7
DINO [18]	ViT-B/8	300	70.0
SimCLR v2 [35]	RN151 (2×)	800	70.2
BYOL [35]	RN200 (2×)	800	71.2
MSN [4]	ViT-B/4	300	**75.7**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

02

ImageNet 1% 레이블 평가

적은 레이블로도 유용한 표현

분류기 등을 학습할 때 전체 정답 레이블의 1%만 사용합니다.
클래스당 약 12~13장 수준으로, 적은 레이블 데이터로도 높은 성능을 얻는지 확인합니다.
사전학습 데이터를 1%로 줄인다는 뜻은 아닙니다.
이 평가는 방법에 따라 선형 평가 또는 fine-tuning을 사용합니다
→ Finetune: 1% 레이블로 인코더를 50 epochs fine-tuning

시사점

1. MAE보다 epcoh은 적지만 더 좋은 성능을 보임.
2. 규모·해상도를 확대한 설정에서 표의 최고 성능(실험환경 다름 주의)

Finetune Not Finetune

Method	Arch.	Epochs	Top-1
Methods without view data augmentations			
data2vec [8]	ViT-L/16	1600	73.3
MAE [36]	ViT-L/16	1600	67.1
	ViT-H/14	1600	71.5
I-JEPA	ViT-L/16	600	69.4
	ViT-H/14	300	73.3
	ViT-H/16_{448}	300	**77.3**
Methods using extra view data augmentations			
iBOT [79]	ViT-B/16	400	69.7
DINO [18]	ViT-B/8	300	70.0
SimCLR v2 [35]	RN151 (2×)	800	70.2
BYOL [35]	RN200 (2×)	800	71.2
MSN [4]	ViT-B/4	300	**75.7**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

03

전이 학습

CIFAR100 · Places205 · iNat18

ImageNet에서 사전학습한 인코더를 다른 데이터셋에 가져가 사용합니다.
새로운 데이터셋에서 인코더를 고정하고 새로운 선형 분류기를 학습합니다.
ImageNet에만 유용한 특징을 배웠는지, 다른 분류 문제에도 통하는지 확인합니다.

Method	Arch.	CIFAR100	Places205	iNat18
Methods without view data augmentations				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	**87.5**	**58.4**	**47.6**
Methods using extra view data augmentations				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	**88.3**	**60.4**	**57.3**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

03

전이 학습

CIFAR100 · Places205 · iNat18

ImageNet에서 사전학습한 인코더를 다른 데이터셋에 가져가 사용합니다.
새로운 데이터셋에서 인코더를 고정하고 새로운 선형 분류기를 학습합니다.
ImageNet에만 유용한 특징을 배웠는지, 다른 분류 문제에도 통하는지 확인합니다.

Method	Arch.	CIFAR100	Places205	iNat18
Methods without view data augmentations				
data2vec [8]	ViT-L/16	81.6	54.6	28.1
MAE [36]	ViT-H/14	77.3	55.0	32.9
I-JEPA	ViT-H/14	**87.5**	**58.4**	**47.6**
Methods using extra view data augmentations				
DINO [18]	ViT-B/8	84.9	57.9	55.9
iBOT [79]	ViT-L/16	**88.3**	**60.4**	**57.3**

시사점

1. 다른 비증강 모델보다 높은 표현을 학습함.
2. 인코더를 고정한 상태에서도 다양한 데이터셋에 효과적으로 활용 가능하다.

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

04

지역 정보 예측 과제

객체 개수 · 공간 / 깊이 정보

CLEVR 이미지에서 객체 개수(Count)와 거리·깊이 관련 정보(Dist)를 예측합니다.
인코더는 고정하고 과제용 선형 head를 학습합니다.
'무슨 대상인가'라는 의미 정보뿐 아니라 개수와 공간 정보도 표현에 남아 있는지 확인하는 평가입니다.

예를들어 Clevr/Count는 "이 장면에 빨간 구슬이 몇 개 있나요?" 같은 개수를 묻는 질문, Clevr/Dist는 "이 이미지에서 카메라에 가장 가까운 객체는 어느 깊이 구간에 있나요?"같은 질문을 하여 거리별 클래스를 구간별 6개로 나눠 고르도록 한다.
→ 구간: 8.0, 8.5, 9.0, 9.5, 10.0, 100.0

Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

04

지역 정보 예측 과제

객체 개수 · 공간 / 깊이 정보

CLEVR 이미지에서 객체 개수(Count)와 거리·깊이 관련 정보(Dist)를 예측합니다.
인코더는 고정하고 과제용 선형 head를 학습합니다.
'무슨 대상인가'라는 의미 정보뿐 아니라 개수와 공간 정보도 표현에 남아 있는지 확인하는 평가입니다.

예를들어 Clevr/Count는 "이 장면에 빨간 구슬이 몇 개 있나요?" 같은 개수를 묻는 질문, Clevr/Dist는 "이 이미지에서 카메라에 가장 가까운 객체는 어느 깊이 구간에 있나요?"같은 질문을 하여 거리별 클래스를 구간별 6개로 나눠 고르도록 한다.
→ 구간: 8.0, 8.5, 9.0, 9.5, 10.0, 100.0

Method	Arch.	Clevr/Count	Clevr/Dist
Methods without view data augmentations			
data2vec [8]	ViT-L/16	85.3	71.3
MAE [36]	ViT-H/14	**90.5**	**72.4**
I-JEPA	ViT-H/14	86.7	**72.4**
Methods using extra data augmentations			
DINO [18]	ViT-B/8	86.6	53.4
iBOT [79]	ViT-L/16	85.7	62.8

시사점

1. 증강 모델과 달리 개수와 깊이 정보를 유지하는 능력이 뛰어나다.
2. Dist에서는 MAE와 같은 성능을 달성한다.

Ablations

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

표현 공간에서의 예측 우수성

Targets	Arch.	Epochs	Top-1
Target-Encoder Output	ViT-L/16	500	**66.9**
Pixels	ViT-L/16	800	40.7

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

마스킹 설정에 따른 큰 성능 변화

Targets		Context	
Scale	Freq.	Scale	**Top-1**
(0.075, 0.2)	4	(0.85, 1.0)	19.2
(0.1, 0.2)	4	(0.85, 1.0)	39.2
(0.125, 0.2)	4	(0.85, 1.0)	42.4
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**
(0.2, 0.25)	4	(0.85, 1.0)	38.9
(0.2, 0.3)	4	(0.85, 1.0)	33.6

Targets		Context	
Scale	Freq.	Scale	**Top-1**
(0.15, 0.2)	4	(0.40, 1.0)	31.2
(0.15, 0.2)	4	(0.65, 1.0)	47.1
(0.15, 0.2)	4	(0.75, 1.0)	49.3
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**

Targets		Context	
Scale	Freq.	Scale	**Top-1**
(0.15, 0.2)	1	(0.85, 1.0)	9.0
(0.15, 0.2)	2	(0.85, 1.0)	22.0
(0.15, 0.2)	3	(0.85, 1.0)	48.5
(0.15, 0.2)	4	(0.85, 1.0)	**54.2**

I-JEPA

CVPR 2023 · Meta AI (FAIR) 논문

5.3배의 적은 시간으로 경쟁령 있는 성능

Method	Arch.	Epochs	Top-1
MAE [36]	ViT-H/14_{448}	1600	**87.8**
I-JEPA	ViT-H/16_{448}	300	87.1

Conclusion

Conclusion

Point 1

표현공간에서의 예측우수성

픽셀을 복원하는 대신 가려진 영역의 표현 벡터를 예측하여 유용한 semantic representation학습

Point 2

마스킹 설계

충분히 큰 target block, 정보가 풍부한 context, 여러 target에 대한 예측이 표현의 품질을 높인다.

Point 3

다양한 downstream task 에 활용

ImageNet 분류와 전이 학습에서 강한 성능을 보이며, 객체 개수와 깊이 관련 정보도 유지한다.

Point 4

data augmentation 의존도를 줄이면서 효율적으로 학습

사람이 설계한 augmentation을 강제하지 않고, 적은 epoch로 경쟁력 있는 성능을 달성한다.

부록

Vision Transformer 모델별 차이

모델	Transformer Blocks	Hidden Dim	MLP Dim	Attention Heads	파라미터
ViT-B	12	768	3,072	12	약 86M
ViT-L	24	1,024	4,096	16	약 307M
ViT-H	32	1,280	5,120	16	약 632M

부록

RSDM Visulization

디퓨전 모델로 다시 시각화 할 수 있습니다.

감사합니다.