Emerging Properties in Self-Supervised Vision Transformers

School of Industrial and Management Engineering, Korea University

Jong Kook, Heo

KOREA UNIVERSITY

DMQA

Contents

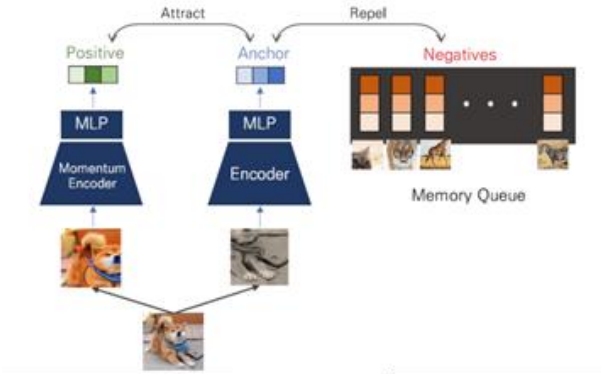
Background

Self-Supervsied Learning

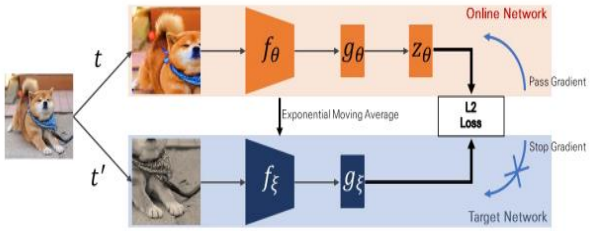
❖ What is Self-Supervised Learning?

- Pre-training method which exploits abundant unlabeled data with pseudo-labels defined by users, to learn good representations
  - ✓ **Pretext task :** 문제를 직접 정의하는 방식 ex) Rotation, Jigsaw Puzzle
  - ✓ **Contrastive Learning :** Noise Contrastive Estimation 방식 ex) MoCo, SimCLR, PIRL
  - ✓ **Non-Constrastive Learning :** Negative Sample 없이 학습하는 방식 ex) BYOL

Jigsaw



MoCo



BYOL

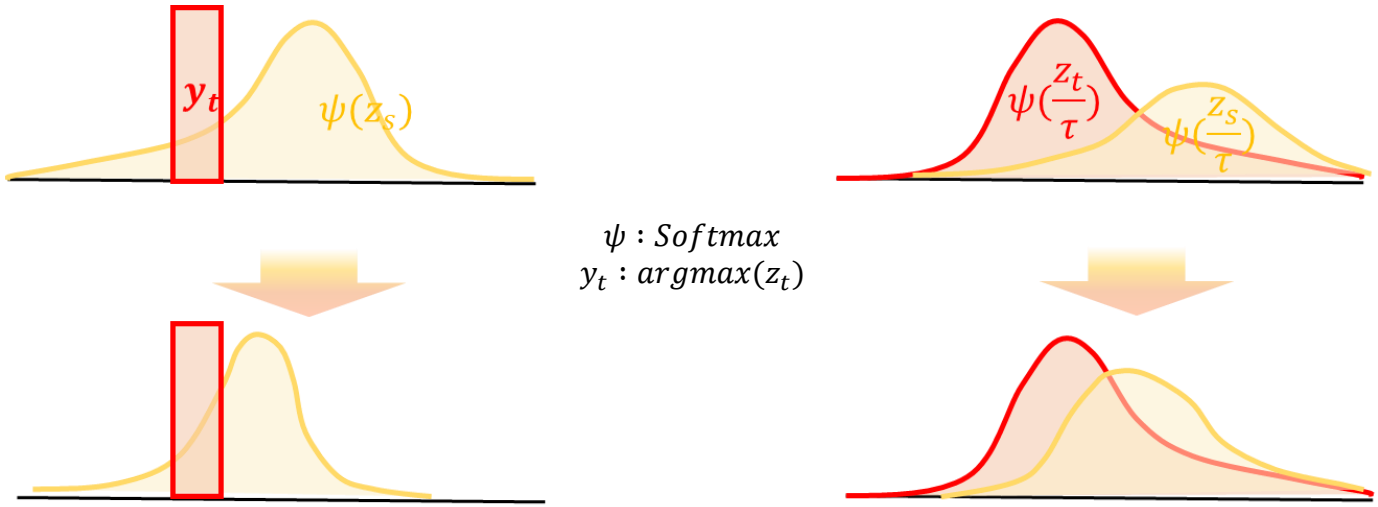
DMQA

Background

Knowledge Distillation

❖ What is Knowledge Distillation??

- Propagating the knowledge of "teacher model" to lightweight "student model" with reducing performance degradation
  - ✓ Hard Distillation : Teacher Model 의 예측 값을 따르도록 학습
  - ✓ Soft Distillation : Teacher Model 의 예측 분포에 유사하도록 학습



DMQA

Research Purpose

❖ DINO : Distillation with no-labels

- ViT 아키텍처에 Self-supervised Learning 을 접목
- 2021 CVPR 에서 발표, 2021.07.07 기준 인용횟수 18회

Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron[1,2] Hugo Touvron[1,3] Ishan Misra[1] Hervé Jegou[1]
Julien Mairal[2] Piotr Bojanowski[1] Armand Joulin[1]

[1] Facebook AI Research [2] Inria* [3] Sorbonne University

Figure 1: **Self-attention from a Vision Transformer with 8×8 patches trained with no supervision.** We look at the self-attention of the [CLS] token on the heads of the last layer. This token is not attached to any label nor supervision. These maps show that the model automatically learns class-specific features leading to unsupervised object segmentations.

DMQA

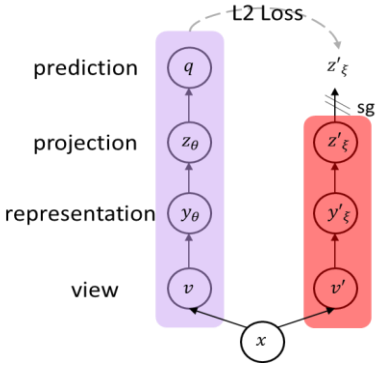
Research Purpose

DINO(Self-Distillation with No-Labels)

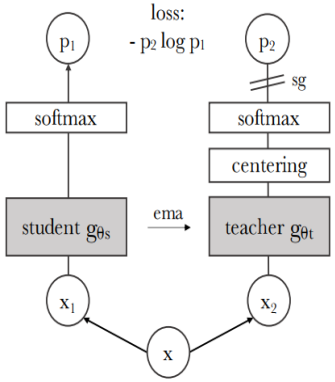
❖ Motivation

- ViT 는 기존 CNN 모델에 비해 명확한 이점을 가지지 못하였음(연산량과 데이터도 더 많이 요구되었음)
- 저자는 Transformer 가 NLP 에서 큰 성공을 거둔 주된 요소는 self-supervised pretraining 이라고 주장함
  - ✓ BERT : Masked Language Model, Next Sentence Prediction
  - ✓ GPT : Autoregressive Language Modeling

- Teacher Network 의 output 을 직접 예측하는 self-supervised Learning 방법론 제시
  - ✓ **저자는 이를 Pretrained teacher model 과 labeled data 가 필요하지 않는 knowledge distillation, 즉 DINO(Self-Distillation with No labels) 라고 명명**
  - ✓ 기존의 momentum encoder(mean teacher) 방식을 쓰며 negative sample 이 필요하지 않아서 BYOL 과 매우 유사



BYOL



DINO

DINO 는 BYOL과 거의 유사하나 augmentation, loss function, collapse avoiding strategy 에서 차이를 보인다.

DMQA

Overview : DINO

Interesting properties of DINO

❖ Self-supervised ViT features are…

- 객체의 경계나 배경을 잘 탐지한다
  - ✓ [CLS] 토큰에 대한 last layer self-attention map 이 segmentation mask 와 매우 유사.
  - ✓ 각 self-attention head 가 서로 다른 물체나 경계를 집중적으로 잘 포착함

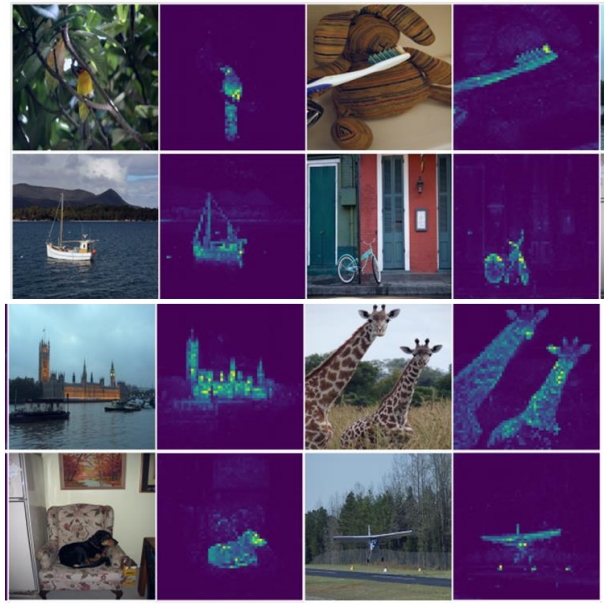


Fig 1 : DINO 를 통해 레이블없이 모델이 자동적으로 class-specific feature 를 학습할 수 있으며, 이를 통해 unsupervised object segmentation 이 가능하다고 주장.

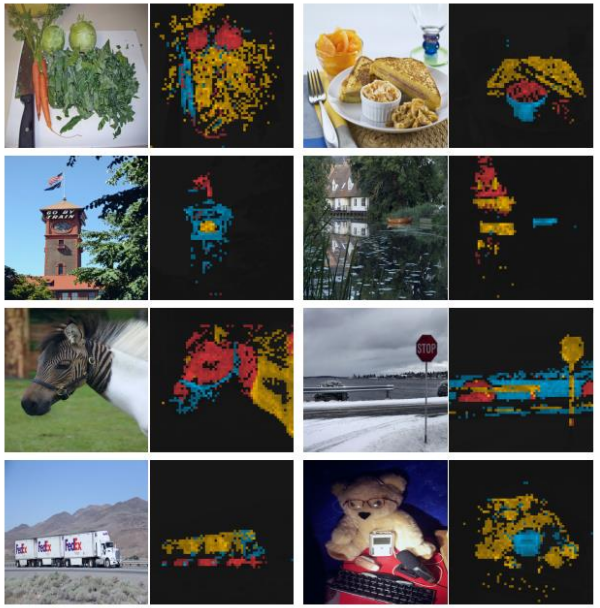


Fig 3 : attention head 마다 다른 색깔로 [CLS] Token 에 대한 score map 계산. 이를 통해 각 head 가 서로 다른 물체나 파츠를 보고 있는 것을 알 수 있음

Overview : DINO

Interesting properties of DINO

❖ Self-supervised ViT features are…

- Fine-tuning 이나 hyper-parameter tuning 없이 k NN Classifier 로도 높은 성능을 보여줌
  - ✓ Multi-crop augmention 과 momentum encoder 가 있어야 k NN 성능에 크게 기여
  - ✓ 패치 사이즈가 작아질수록 throughtput 은 작아지지만 더 좋은 질의 feature 를 추출함
  - ✓ 기존 Convnet 에서도 해당 방법론을 사용할 수는 있지만 ViT 계열과 궁합이 더 잘 맞음

	Method	Mom.	SK	MC	Loss	Pred.	k-NN	Lin.
1	DINO	✓	✗	✓	CE	✗	72.8	76.1
2		✗	✗	✓	CE	✗	0.1	0.1
3		✓	✓	✓	CE	✗	72.2	76.0
4		✓	✗	✗	CE	✗	67.9	72.5
5		✓	✗	✓	MSE	✗	52.6	62.4
6		✓	✗	✓	CE	✓	71.8	75.6
7	BYOL	✓	✗	✗	MSE	✓	66.6	71.4
8	MoCov2	✓	✗	✗	INCE	✗	62.0	71.6
9	SwAV	✗	✓	✓	CE	✗	64.7	71.8

SK: Sinkhorn-Knopp, MC: Multi-Crop, Pred.: Predictor
CE: Cross-Entropy, MSE: Mean Square Error, INCE: InfoNCE

Table 7. Momentum encoder, Loss, Predictor, Multi-Crop 등 방법론에 대한 ablation study

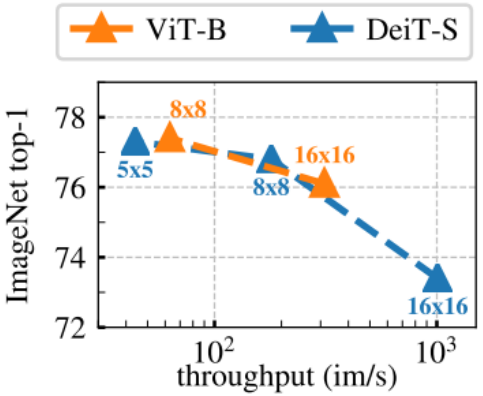


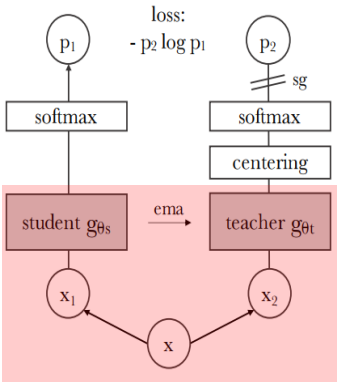
Fig 5. patch size 가 작아지면 throughput 은 작아지지만 성능은 좋아짐

Method	Arch.	Param.	im/s	Linear	k-NN
Supervised	RN50	23	1237	79.3	79.3
SCLR [12]	RN50	23	1237	69.1	60.7
MoCov2 [14]	RN50	23	1237	71.1	61.9
InfoMin [64]	RN50	23	1237	73.0	65.3
BarlowT [78]	RN50	23	1237	73.2	66.0
OBoW [25]	RN50	23	1237	73.8	61.9
BYOL [28]	RN50	23	1237	74.4	64.8
DCv2 [10]	RN50	23	1237	75.2	67.1
SwAV [10]	RN50	23	1237	**75.3**	65.7
DINO	RN50	23	1237	**75.3**	**67.5**
Supervised	DeiT-S	21	1007	79.8	79.8
BYOL* [28]	DeiT-S	21	1007	71.4	66.6
MoCov2* [14]	DeiT-S	21	1007	72.7	64.4
SwAV* [10]	DeiT-S	21	1007	73.5	66.3
DINO	DeiT-S	21	1007	**77.0**	**74.5**
Comparison across architectures					
SCLR [12]	RN50w4	375	117	76.8	69.3
SwAV [10]	RN50w2	93	384	77.3	67.3
BYOL [28]	RN50w2	93	384	77.4	–
DINO	ViT-B/16	85	312	78.2	76.1
SwAV [10]	RN50w5	586	76	78.5	67.1
BYOL [28]	RN50w4	375	117	78.6	–
BYOL [28]	RN200w2	250	123	79.6	73.9
DINO	DeiT-S/8	21	180	79.7	**78.3**
SCLRv2 [13]	RN152w3+SK	794	46	79.8	73.1
DINO	ViT-B/8	85	63	**80.1**	77.4

Table 2. backbone 에 상관없이 DINO 가 SSL 방법론보다 우수함을 보여줌. Backbone 은 ConvNet 보다 ViT 계열이 더 궁합이 좋음

DMQA

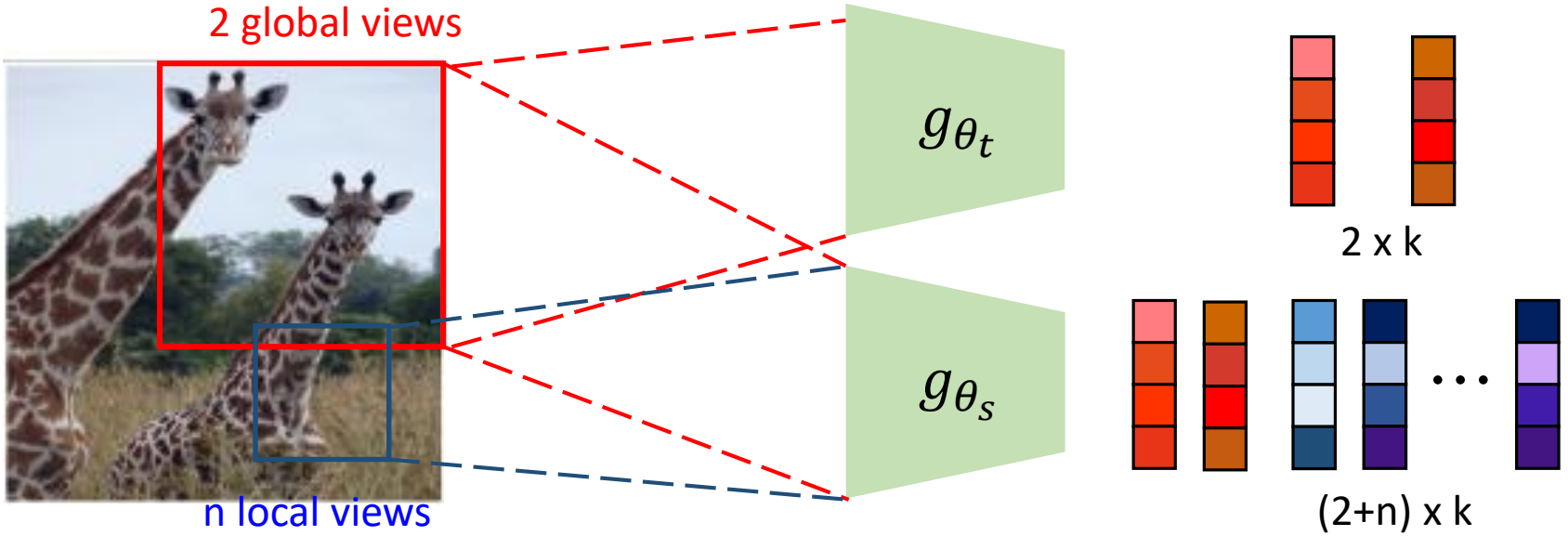
Overview : DINO



Process

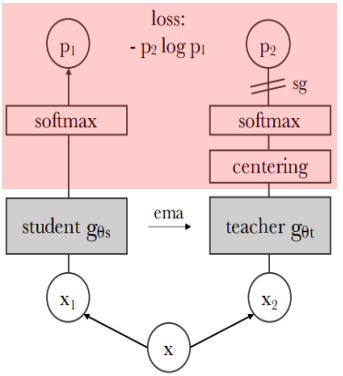
❖ Augmentation & Forward

- SWaV(Caron et al, 2020) 의 Multi-crop 에서 차용
- 원본 이미지에서 2개의 Global View(원본 크기의 50% ↑), n 개의 Local View(원본 크기의 50% 미만↓) 생성
- 각 view 는 독립적으로 crop 후 augmentation 진행(ex : Global view : 224 x 224, Local view : 96 x 96)
  - ✓ Teacher network 는 2개의 Global View 를 입력으로 받음
  - ✓ **Student network 는 2개의 Global View 와 n 개의 Local View 모두 입력으로 받음**
  - ✓ 각 네트워크는 이후 3 층의 layer 로 구성된 MLP 와 Layer norm 을 거쳐 K dimensional vector 를 뱉음



DMQA

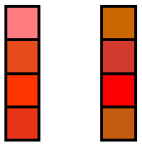
Overview : DINO



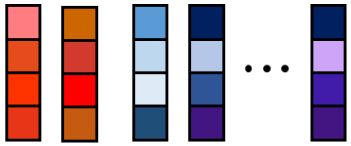
Process

❖ Softmax with Centering

- Teacher Representation $g_{\theta_t}(x)$: Softmax + sharpening+ centering
- Student Representation $g_{\theta_s}(x)$: Softmax + sharpening
- Centering 과 sharpening 이 model Collapse 를 방지할 수 있다고 주장



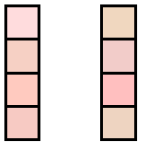
2 x k

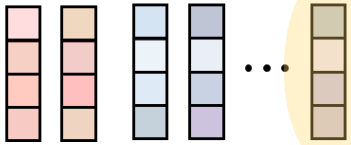


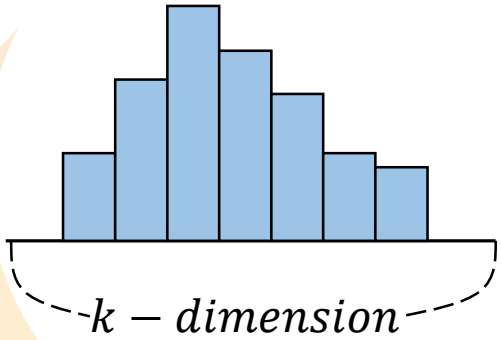
(2+n) x k

$$P_t(x)^i = Softmax\left(\frac{g_{\theta_t}(x) - C}{\tau_t}\right)_i$$

$$P_s(x)^i = Softmax\left(\frac{g_{\theta_s}(x)}{\tau_s}\right)_i$$





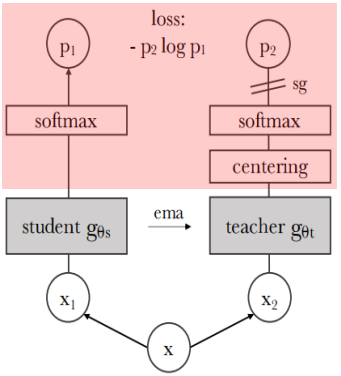


Softmax over k dimension per each view

τ_t: teacher model temperature
τ_s: student model temperature
C : centering factor

DMQA

Overview : DINO



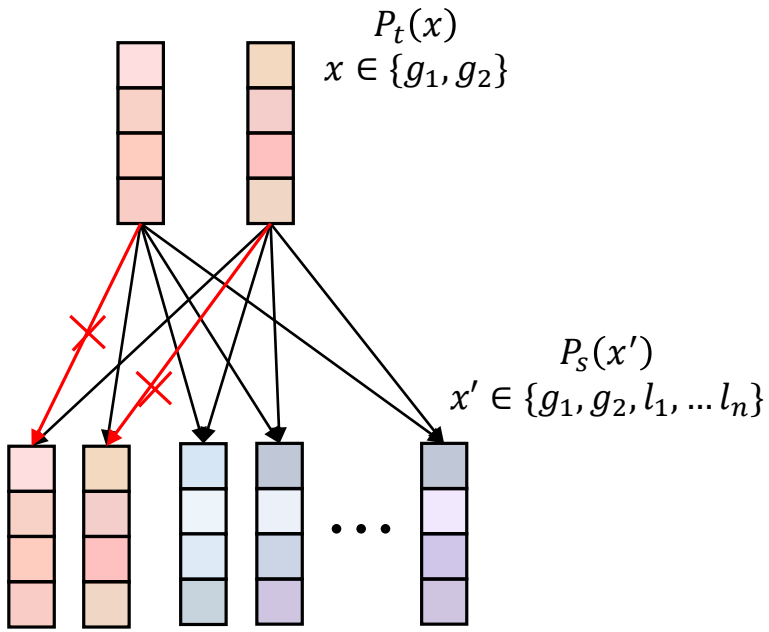
Process

❖ Calcualte Loss and Gradient

- Cross Entropy Loss 를 계산(Target 값에 대해서는 stop gradient!!)
- 같은 view 에 대해서는 loss 를 계산하지 않음

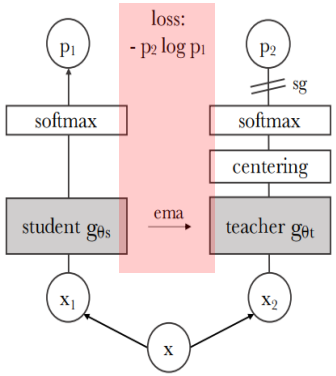
$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in all\ views \\ x' \neq x}} -P_t(x) log P_s(x')$$

Don't Calculate on same view pairs!!



DMQA

Overview : DINO



Process

❖ Update

- Student Model : Teacher Model Representation 과의 KL Divergence 로 업데이트
- Teacher Model : student 모델의 가중치를 통해 모멘텀 업데이트(exponential moving avg like BYOL)
- Centering Factor : teacher output 의 batch center 를 통해 업데이트(exponential moving avg)

Student Model

$$\min_{\theta_s} \sum_{x \in \{x_1^g, x_2^g\}} \sum_{\substack{x' \in all\ views \\ x' \neq x}} -P_t(x) log P_s(x')$$

Teacher Model

$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s,$$
$$\lambda = 0.996 \rightarrow 1 (cosine\ schedule)$$

Centering factor

$$c \leftarrow mc + (1-m)c\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i)$$

τ_t: teacher model temperature
τ_s: student model temperature
C : centering factor
m: rate parameter

Overview : DINO

Details : Centering and Sharpening

❖ How to avoid Collapse in DINO

- 기존의 SSL 방법론들 Contrastive Loss, Clustering Constraints, Predictor, BatchNorm 등을 통해 Collapse 를 방지
- DINO 에서는 momentum teacher output 에 **Centering 과 Sharpening**만 적용해도 Collapse 를 방지할 수 있다고 주장!!
  - ✓ Centering : K 차원 중 특정 차원이 dominate 하는 것을 방지하지만, uniform distribution 으로 collapse 시킬 수 있음
  - ✓ Centering 은 teacher output 에 bias term 을 붙이는 것으로 해석 될 수 있음
  - ✓ 이 bias term 은 teacher output 의 1st order batch statistic 에 영향을 받음
  - ✓ Sharpening(Temperature scaling **where $0 < \tau < 1$**): Centering 과 반대의 효과

$$P_t(x)^i = Softmax\left(\frac{g_{\theta_t}(x) - C}{\tau_t}\right)_i$$

Teacher output with centering and sharpening

$$c \leftarrow mc + (1-m)c\frac{1}{B}\sum_{i=1}^{B} g_{\theta_t}(x_i)$$

C 는 teacher output 의 batch mean 으로 모멘텀 업데이트

Overview : DINO

Details : Centering and Sharpening

❖ How to avoid Collapse in DINO

- 만약 Centering 이나 Sharpening 둘 중 하나가 없다면 오른 쪽 그림과 같이 teacher 와 student 사이의 KL divergence 가 0으로 수렴(Collapse)
- 왼쪽 그림은 **Centering 이 없을 때 teacher model entropy 가 0으로**, **Sharpening 이 없을 때는 log K 로** 가는 것을 나타내며 두 operation 이 각각 다른 Collapse 를 유발하는 것을 나타낸다.
- 오른쪽 그림에서 **둘 다 사용했을 경우, KL divergence 가 적절하게 감소**하는 것을 알 수 있다.

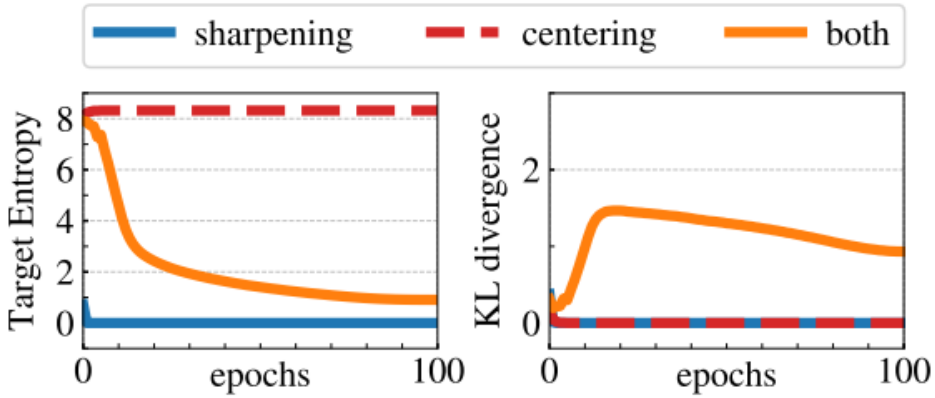


Figure 7: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

Overview : DINO

Total Review

❖ Reminder : Pseudo Code

- 해당 pseudo code 는 multi-crop 이 포함되지 않은 코드(multi-crop augmentation 이 들어간 pseudo code 는 논문에 기재 X)

Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# gs, gt: student and teacher networks
# C: center (K)
# tps, tpt: student and teacher temperatures
# l, m: network and center momentum rates
gt.params = gs.params
for x in loader: # load a minibatch x with n samples
    x1, x2 = augment(x), augment(x) # random views

    s1, s2 = gs(x1), gs(x2) # student output n-by-K
    t1, t2 = gt(x1), gt(x2) # teacher output n-by-K

    loss = H(t1, s2)/2 + H(t2, s1)/2
    loss.backward() # back-propagate

    # student, teacher and center updates
    update(gs) # SGD
    gt.params = l*gt.params + (1-l)*gs.params
    C = m*C + (1-m)*cat([t1, t2]).mean(dim=0)

def H(t, s):
    t = t.detach() # stop gradient
    s = softmax(s / tps, dim=1)
    t = softmax((t - C) / tpt, dim=1) # center + sharpen
    return - (t * log(s)).sum(dim=1).mean()
```

Experimental Results

Other experiments

❖ 기타 실험 장표

- Image Retrieval and Copy Detection

Table 3: **Image retrieval.** We compare the performance in retrieval of off-the-shelf features pretrained with supervision or with DINO on ImageNet and Google Landmarks v2 (GLDv2) dataset. We report mAP on revisited Oxford and Paris. Pretraining with DINO on a landmark dataset performs particularly well. For reference, we also report the best retrieval method with off-the-shelf features [55].

			$\mathcal{R}$Ox		$\mathcal{R}$Par	
Pretrain	Arch.	Pretrain	M	H	M	H
Sup. [55]	RN101+R-MAC	ImNet	49.8	18.5	74.0	**52.1**
Sup.	DeiT-S/16	ImNet	33.5	8.9	63.0	37.2
DINO	ResNet-50	ImNet	35.4	11.1	55.9	27.5
DINO	DeiT-S/16	ImNet	41.8	13.7	63.1	34.4
DINO	DeiT-S/16	GLDv2	**51.5**	**24.3**	**75.3**	51.6

Table 4: **Copy detection.** We report the mAP performance in copy detection on Copydays "strong" subset [20]. For reference, we also report the performance of the multigrain model [5], trained specifically for particular object retrieval.

Method	Arch.	Dim.	Resolution	mAP
Multigrain [5]	ResNet-50	2048	224^2	75.1
Multigrain [5]	ResNet-50	2048	largest side 800	82.5
Supervised [66]	ViT-B/16	1536	224^2	76.4
DINO	ViT-B/16	1536	224^2	81.7
DINO	ViT-B/8	1536	320^2	**85.5**

Experimental Results

Other experiments

❖ 기타 실험 장표

- Fine tuning

Table 6: **Transfer learning by finetuning pretrained models on different datasets.** We report top-1 accuracy. Self-supervised pretraining with DINO transfers better than supervised pretraining.

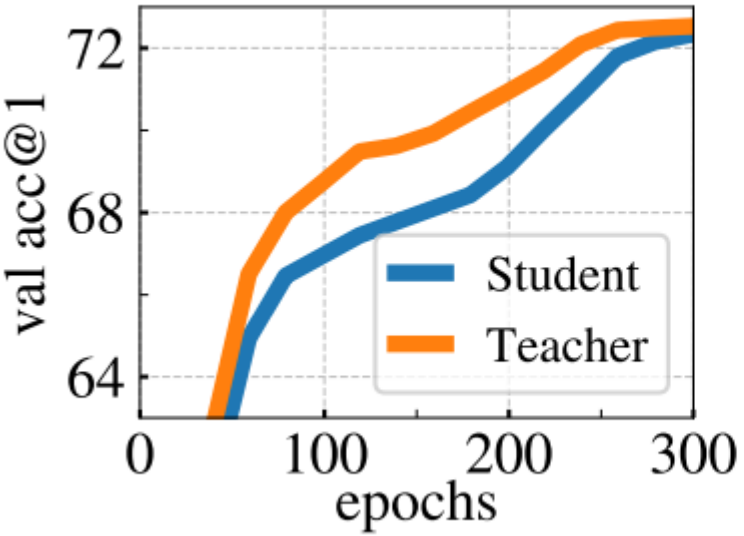
	Cifar_{10}	Cifar_{100}	INat_{18}	INat_{19}	Flwrs	Cars	INet
DeiT-S/16							
Sup. [66]	**99.0**	89.5	70.7	76.6	98.2	92.1	79.9
DINO	**99.0**	**90.5**	**72.0**	**78.2**	**98.5**	**93.0**	**81.5**
ViT-B/16							
Sup. [66]	99.0	90.8	**73.2**	77.7	98.4	92.1	81.8
DINO	**99.1**	**91.7**	72.6	**78.6**	**98.8**	**93.0**	**82.8**

Experimental Results

Other experiments

❖ 기타 실험 장표

- Update Options



Teacher	Top-1
Student copy	0.1
Previous iter	0.1
Previous epoch	66.6
Momentum	72.8

Conclusion

- ❖ 단순히 ViT 에 SSL 을 적용한 것에 의의를 둔것이 해당 방법론을 적용 했을 때, Supervised ViT 나 ConvNet 에서 나타나지 않았던 현상을 발견한 것이 큰 의미가 있음
  - ✓ DINO 를 적용한 ViT 는 객체의 경계나 배경을 매우 잘 탐지하며, 이러한 특징들은 linear layer fine-tuning 없이 k NN classifier 로도 높은 성능을 나타냄
  - ✓ Segmentation 등의 다른 task를 주지 않고 positive sample 로 representation learning 만을 했을 뿐인데 결과가 기존 segmentation 방법론에 근접
  - ✓ 논문에서도 언급하듯이, Computer vision 에서도 GPT 나 BERT 같이 다른 분야의 태스크도 수행가능한 pre-trained model 이 더욱 발전되리라 기대됨

- ❖ 엄청나게 많은 실험을 진행
  - ✓ 방법론의 타당성을 위해 Momentum Encoder, Loss 등 옵션에 대해 모두 ablation study 를 진행
  - ✓ Predictor 나 Negative Sample 없이도 단순 연산만으로 Collapse 를 방지하였는 것은 굉장히 큰 의미가 있다고 봄
  - ✓ 특히 Sharpening 과 Centering 가 서로 상충된 효과를 내지만 둘 중 하나가 없으면 서로 다른 방향으로 collapse 가 진행된다는 것을 보여준 것이 재밌었음

Reference

- Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., & Joulin, A. (2021). Emerging properties in self-supervised vision transformers. *arXiv preprint arXiv:2104.14294*.

- Noroozi, M., & Favaro, P. (2016, October). Unsupervised learning of visual representations by solving jigsaw puzzles. In *European conference on computer vision* (pp. 69-84). Springer, Cham.

- http://dmqa.korea.ac.kr/activity/seminar/310
  - ✓ Dive into BYOL