Medical Imaging Intelligent Reality Lab

ASAN Medical Center

DINO : Emerging Properties in Self-Supervised Vision Transformers

성지민

01 Introduction

- NLP 분야에서는 Bert의 masked language model이나 GPT의 language modeling은 self-supervised learning을 적용하여 성공.
- Self-supervised learning된 ViT에 대한 선행연구를 보면 대체로 좋은 성능을 보임.
- ViT의 self-supervised learning에 대한 선행연구들은 아키텍처는 비슷하지만, Collapse를 막기 위해 다양한 방식을 시도함. - predictor, advanced normalization, contrastive loss 등등

01 Introduction

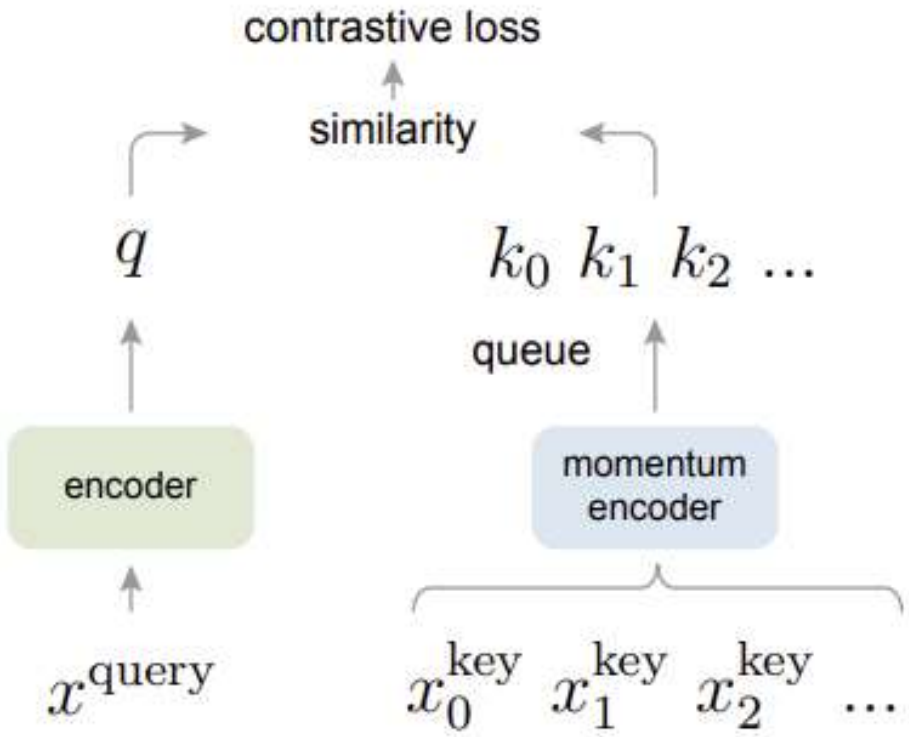
MI2RL Medical Imaging Intelligent Reality Lab

DINO

- momentum encoder을 활용하여 teacher network를 학습.
- Collapse를 방지하기 위해 teacher output을 centering & sharpening.
- 작은 패치를 사용한 ViT-Base 모델에서 ImageNet 선형 분류 벤치마크에서 80.1%의 top-1 정확도를 달성
- ViT를 사용하여 DINO를 학습하는 데에는 8개의 GPU 서버 두 대에서 3일이 소요됨. 이는 ImageNet 벤치마크에서 76.1%의 성능을 달성하여, 유사한 크기의 컨볼루션 신경망 기반 자기지도 학습 시스템보다 계산 요구량을 크게 줄이면서도 더 나은 성능

02 Related works

- MoCo : Momentum Contrast for Unsupervised Visual Representation Learning



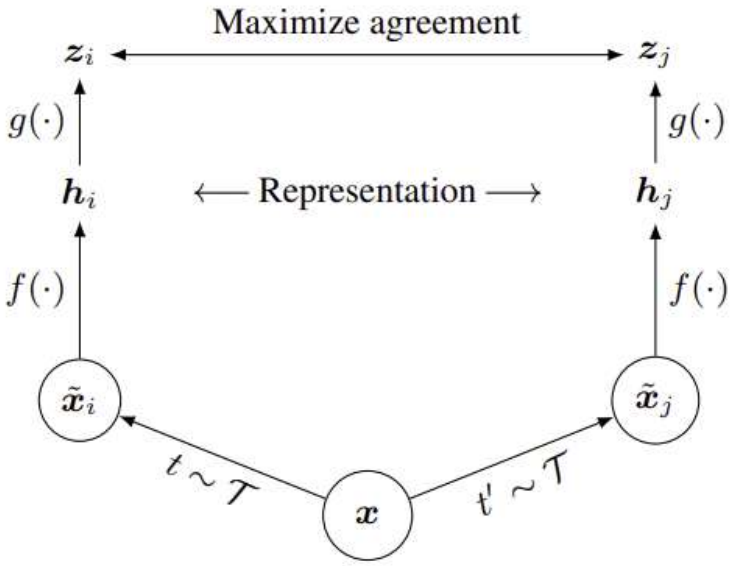
- Query : embedded input image
- Key
  - Positive key : 동일 input image에서 생성된 key
  - Negative keys : 다른 input image에서 생성된 keys
- Contrastive loss
  - Query – Positive key : 거리를 가깝게 함
  - Query – Negative keys : 거리를 멀게 함
- Momentum update

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}.$$

- Dictionary as a queue

02 Related works

- SIMCLR

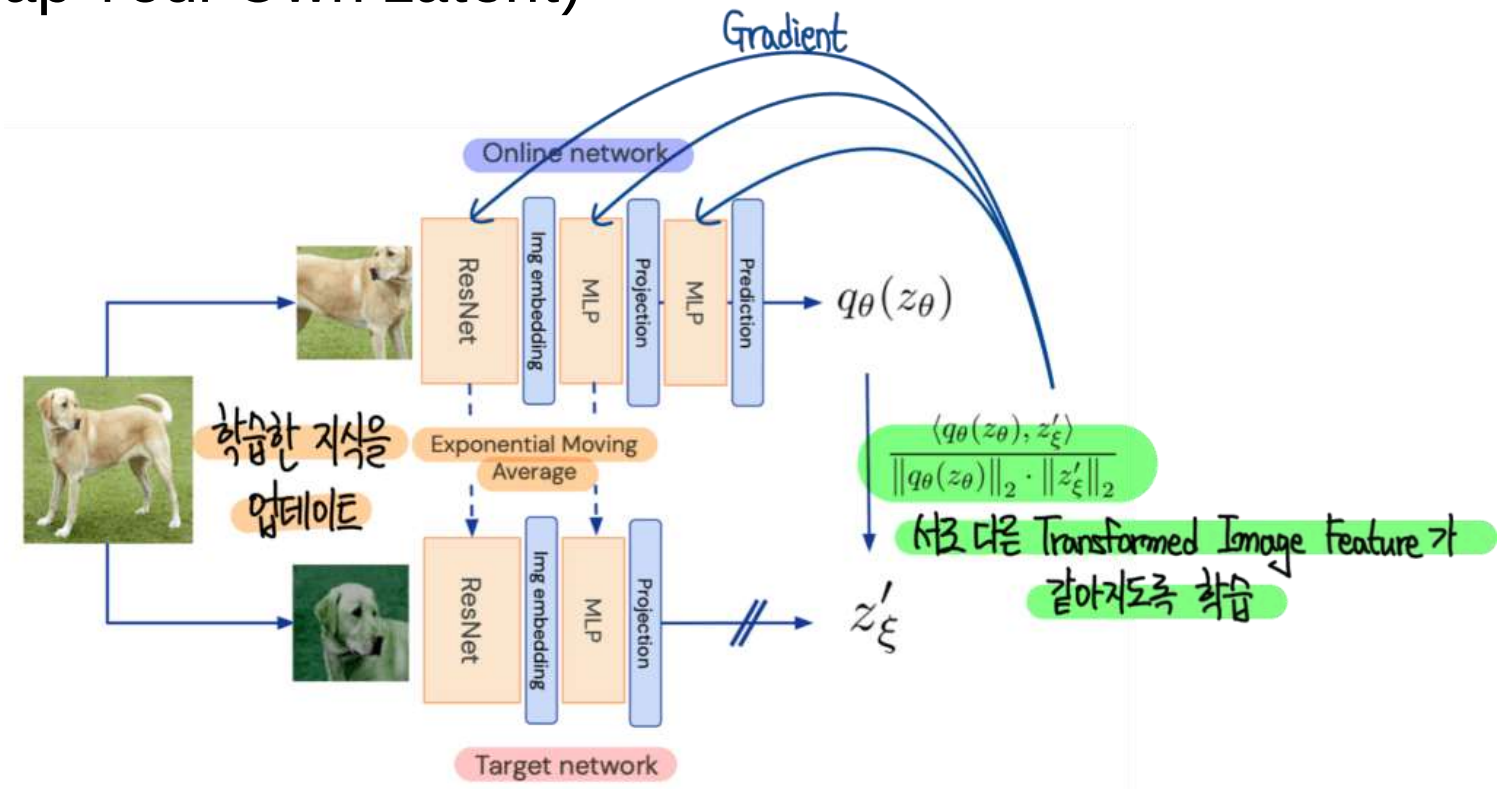


$$l(\,,\,) = -\log\left(\frac{e^{\text{similarity}(\,,\,)}}{e^{\text{similarity}(\,,\,)} + e^{\text{similarity}(\,,\,)} + e^{\text{similarity}(\,,\,)}}\right)$$

$$L = \frac{\overbrace{[l(\,,\,) + l(\,,\,)]}^{\text{Pair 1 Loss (k=1)}} + \overbrace{[l(\,,\,) + l(\,,\,)]}^{\text{Pair 2 Loss (k=2)}}}{2 * 2}$$

02 Related works

BYOL(Boostrap Your Own Latent)

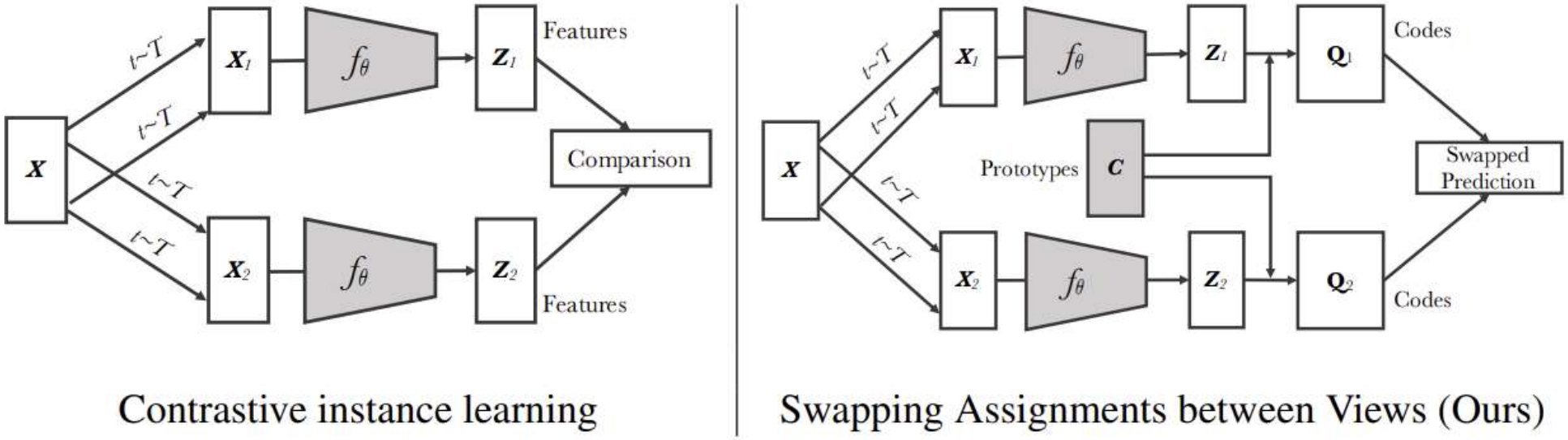


EMA

$$\theta_{\text{target}} \leftarrow \tau\theta_{\text{target}} + (1-\tau)\theta_{\text{online}}$$

02 Related works

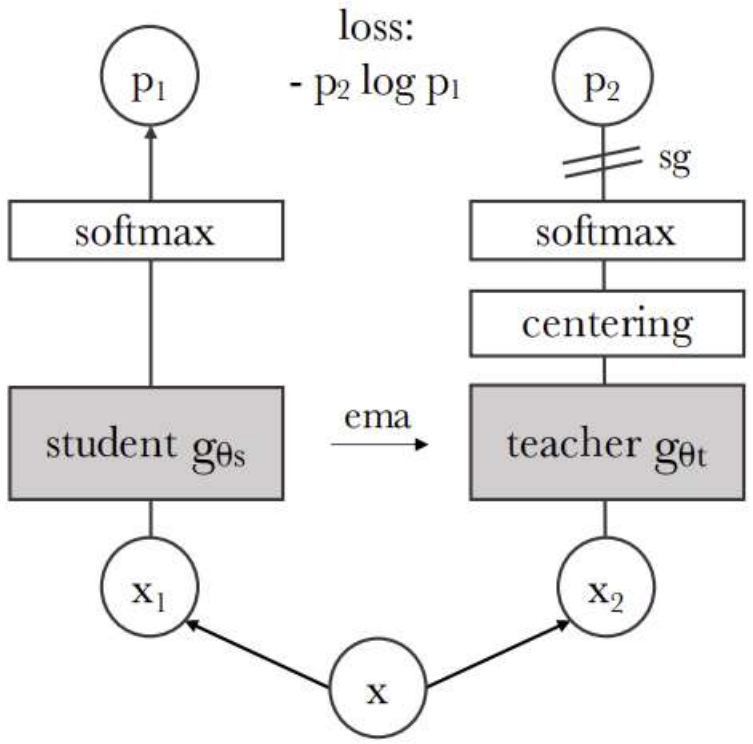
SWAV



$$L(z_1, z_2) = l(z_1, q_2) + l(z_2, q_1)$$

03 Architecture

01 | Self-supervised learning with knowledge distilation



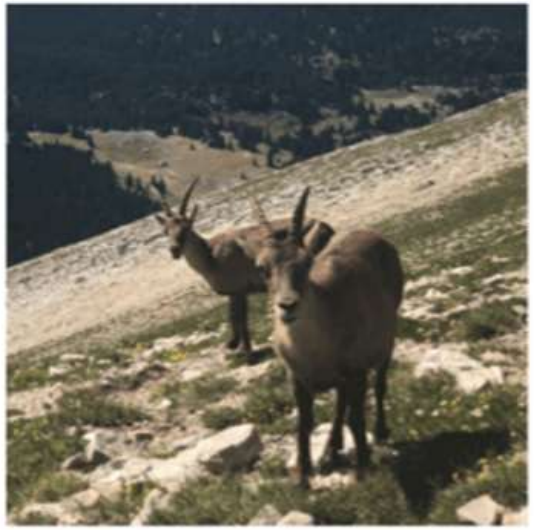
03 Architecture

Medical Imaging Intelligent Reality Lab

01 | Self-supervised learning with knowledge distilation

03 Architecture

*Multi-cropping

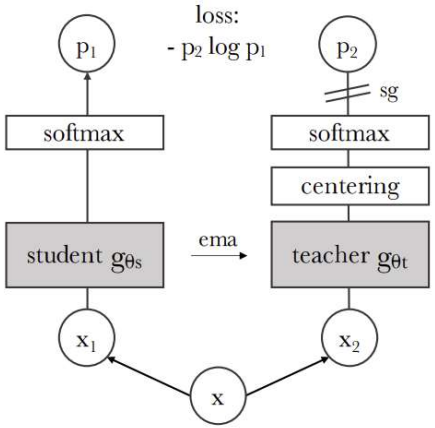


Local view
(96 * 96)

Global view
(224 * 224)

03 Architecture

03 | Network Architecture



Projection head

Vision backbone

$$g = h \circ f$$

Projection head	3-layer MLP(GELU), L2 norm, weight normalized fully connected layer with K dimensions
Vision backbone	ViT, Resnet

03 Architecture

01 | Self-supervised learning with knowledge distilation

$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K}\exp(g_{\theta_s}(x)^{(k)}/\tau_s)}, \quad P_t(x)^{(i)} = \frac{\exp(g_{\theta_t}(x)^{(i)}/\tau_t)}{\sum_{k=1}^{K}\exp(g_{\theta_t}(x)^{(k)}/\tau_t)}$$

$$\min_{\theta_s} H(P_t(x), P_s(x)) \qquad H(P_t, P_s) = h(P_t) + D_{KL}(P_t|P_s).$$

$$\min_{\theta_s} \sum_{x\in\{x_1^g, x_2^g\}} \sum_{x'\in V, x'\neq x} H(P_t(x), P_s(x'))$$

03 Architecture

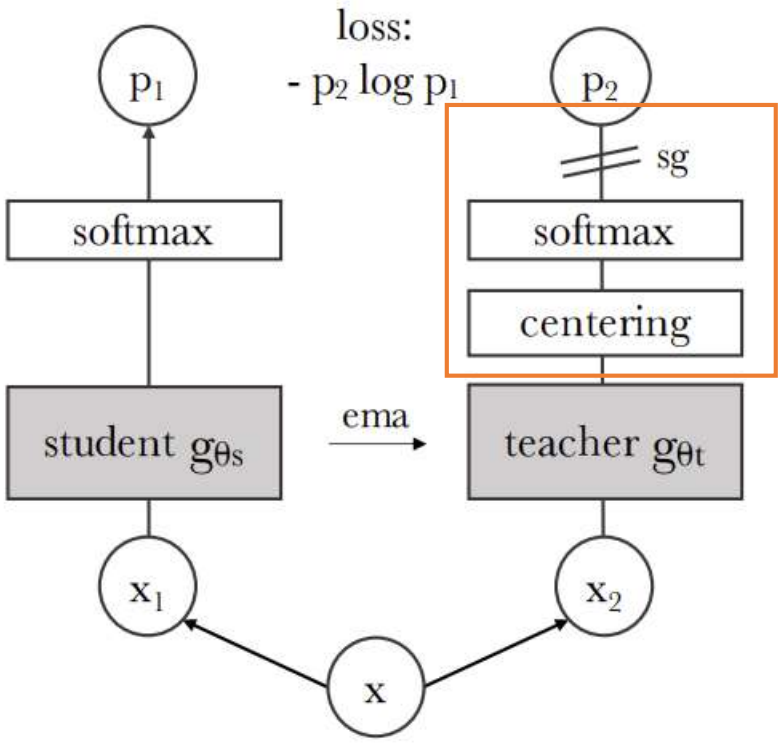
02 | Teacher network.

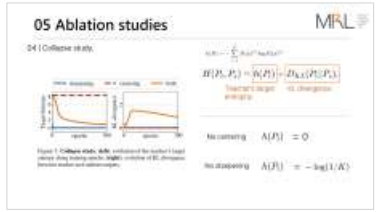
$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

λ = 0.996 ~ 1 (cosine schedule)

03 Architecture

03 | Network Architecture





Centering

$$g_t(x) \leftarrow g_t(x) + c$$

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

Sharpening

$$P_t(x)^{(i)} = \frac{\exp\left(g_{\theta_t}(x)^{(i)} / \tau_t\right)}{\sum_{k=1}^{K} \exp\left(g_{\theta_t}(x)^{(k)} / \tau_t\right)}$$

03 Architecture

*Multi-cropping

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()
```

04 Results

Medical Imaging Intelligent Reality Lab

01 | Comparing with SSL frameworks on ImageNet

Table 2: **Linear and k-NN classification on ImageNet.** We report top-1 accuracy for linear and k-NN evaluations on the validation set of ImageNet for different self-supervised methods. We focus on ResNet-50 and ViT-small architectures, but also report the best results obtained across architectures. * are run by us. We run the k-NN evaluation for models with official released weights. The throughput (im/s) is calculated on a NVIDIA V100 GPU with 128 samples per forward. Parameters (M) are of the feature extractor.

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 [15]	RN50	23	1237	71.1	61.9
InfoMin [67]	RN50	23	1237	73.0	65.3
BarlowT [81]	RN50	23	1237	73.2	66.0
OBoW [27]	RN50	23	1237	73.8	61.9
BYOL [30]	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	ViT-S	21	1007	79.8	79.8
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-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 [30]	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 [30]	RN50w4	375	117	78.6	–
BYOL [30]	RN200w2	250	123	79.6	73.9
DINO	ViT-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

04 Results

02 | Nearest neighbor retrieval with DINO ViT

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 [57].

			$\mathcal{R}$Ox		$\mathcal{R}$Par	
Pretrain	Arch.	Pretrain	M	H	M	H
Sup. [57]	RN101+R-MAC	ImNet	49.8	18.5	74.0	**52.1**
Sup.	ViT-S/16	ImNet	33.5	8.9	63.0	37.2
DINO	ResNet-50	ImNet	35.4	11.1	55.9	27.5
DINO	ViT-S/16	ImNet	41.8	13.7	63.1	34.4
DINO	ViT-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 [21]. 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 [69]	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**

04 Results

03 | Discovering the semantic layout of scenes

Table 5: **DAVIS 2017 Video object segmentation.** We evaluate the quality of frozen features on video instance tracking. We report mean region similarity $\mathcal{J}_m$ and mean contour-based accuracy $\mathcal{F}_m$. We compare with existing self-supervised methods and a supervised ViT-S/8 trained on ImageNet. Image resolution is 480p.

Method	Data	Arch.	$(\mathcal{J}\&\mathcal{F})_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Supervised					
ImageNet	INet	ViT-S/8	66.0	63.9	68.1
STM [48]	I/D/Y	RN50	81.8	79.2	84.3
Self-supervised					
CT [71]	VLOG	RN50	48.7	46.4	50.0
MAST [40]	YT-VOS	RN18	65.5	63.3	67.6
STC [37]	Kinetics	RN18	67.6	64.8	70.2
DINO	INet	ViT-S/16	61.8	60.2	63.4
DINO	INet	ViT-B/16	62.3	60.7	63.9
DINO	INet	ViT-S/8	**69.9**	**66.6**	**73.1**
DINO	INet	ViT-B/8	**71.4**	**67.9**	**74.9**

04 Results

03 | Discovering the semantic layout of scenes

Figure 3: **Attention maps from multiple heads.** We consider the heads from the last layer of a ViT-S/8 trained with DINO and display the self-attention for [CLS] token query. Different heads, materialized by different colors, focus on different locations that represents different objects or parts (more examples in Appendix).

	Random	Supervised	DINO
ViT-S/16	22.0	27.3	45.9
ViT-S/8	21.8	23.7	44.7

04 Results

04 | Transfer learning on downstream tasks

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
ViT-S/16							
Sup. [69]	**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. [69]	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**

05 Ablation studies

MI2RL Medical Imaging Intelligent Reality Lab

01 | Importance of the Different Components

Table 7: **Important component for self-supervised ViT pre-training.** Models are trained for 300 epochs with ViT-S/16. We study the different components that matter for the k-NN and linear ("Lin.") evaluations. For the different variants, we highlight the differences from the default DINO setting. The best combination is the momentum encoder with the multicrop augmentation and the cross-entropy loss. We also report results with BYOL [30], MoCo-v2 [15] and SwAV [10].

	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

05 Ablation studies

02 | Effect of Patch Size

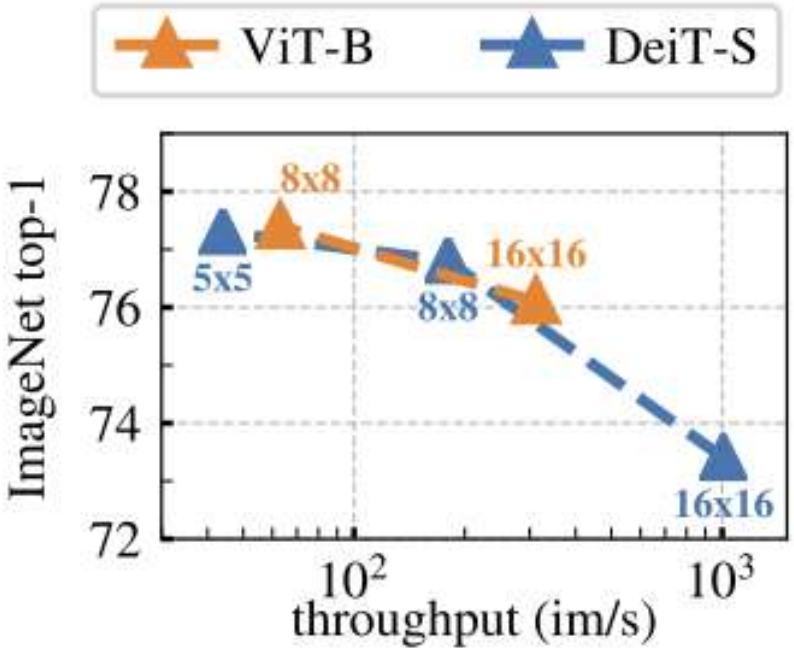
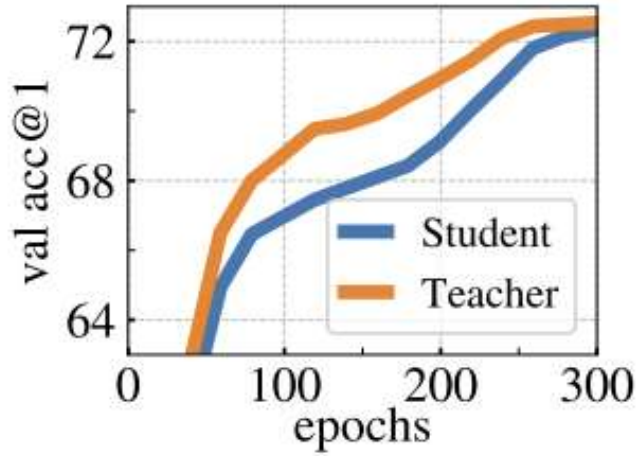


Figure 5: **Effect of Patch Size.** k-NN evaluation as a function of the throughputs for different input patch sizes with ViT-B and ViT-S. Models are trained for 300 epochs.

05 Ablation studies

03 | Impact of the choice of Teacher Network.



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

Figure 6: Top-1 accuracy on ImageNet validation with k-NN classifier. **(left)** Comparison between the performance of the momentum teacher and the student during training. **(right)** Comparison between different types of teacher network. The momentum encoder leads to the best performance but is not the only viable option.

05 Ablation studies

04 | Collapse study.

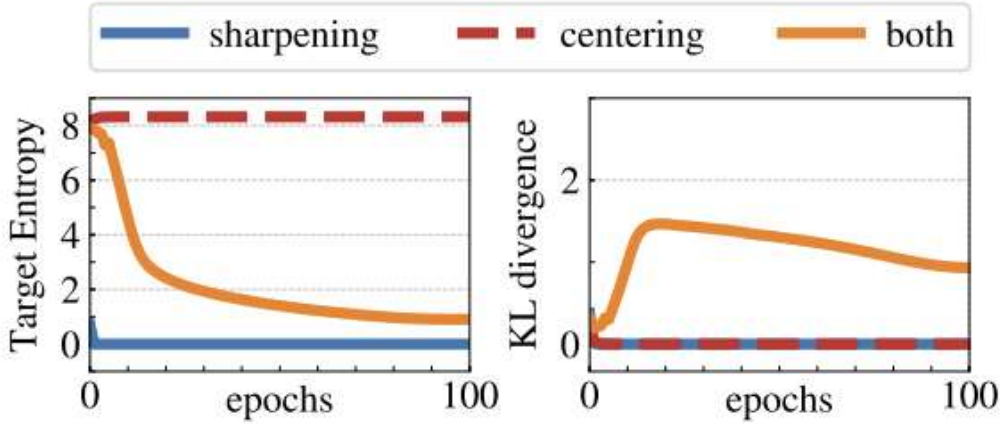


Figure 7: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

$$h(P_t) = -\sum_{i=1}^{K} P_t(x)^{(i)} \log P_t(x)^{(i)}$$

$$H(P_t, P_s) = h(P_t) + D_{KL}(P_t \| P_s).$$

Teacher's target entrophy

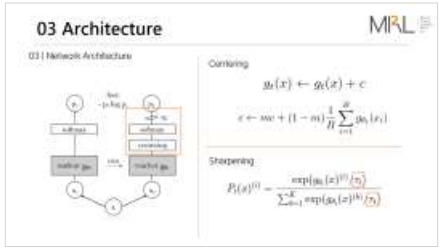
KL divergence

No centering $h(P_t) = 0$

No sharpening $h(P_t) = -\log(1/K)$

05 Ablation studies

04 | Collapse study.



m	0	0.9	0.99	0.999
k-NN top-1	69.1	69.7	69.4	0.1

τ_t	0	0.02	0.04	0.06	0.08	$0.04 \rightarrow 0.07$
k-NN top-1	43.9	66.7	69.6	68.7	0.1	69.7

05 Ablation studies

04 | Compute requirements.

Table 8: **Time and memory requirements.** We show total running time and peak memory per GPU ("mem.") when running ViT-S/16 DINO models on two 8-GPU machines. We report top-1 ImageNet val acc with linear evaluation for several variants of multi-crop, each having a different level of compute requirement.

	100 epochs		300 epochs		
multi-crop	top-1	time	top-1	time	mem.
2×224^2	67.8	15.3h	72.5	45.9h	9.3G
$2\times224^2 + 2\times96^2$	71.5	17.0h	74.5	51.0h	10.5G
$2\times224^2 + 6\times96^2$	73.8	20.3h	75.9	60.9h	12.9G
$2\times224^2 + 10\times96^2$	74.6	24.2h	76.1	72.6h	15.4G

05 Ablation studies

05 | Training with small batches.

bs	128	256	512	1024
top-1	57.9	59.1	59.6	59.9

Table 9: **Effect of batch sizes.** Top-1 with k-NN for models trained for 100 epochs without multi-crop.

05 Ablation studies

04 | Implementation Details.

Optimizer	Adamw
Batch size	1024
Learning rate	lr = 0.0005 ∗ batchsize/256.(first 10 epochs – linear warm up) cosine schedule (after 10 epoch)
weight decay	cosine schedule from 0.04 to 0.4
τ_t	0.04 ~ 0.07 warm up (first 30 epochs)
τ_s	0.1
m	0.9
Multi-crop	O

06 Conclusion

- 자기 지도 학습을 통해 Vision Transformer(ViT)를 사전 훈련하면, 특별한 구조적 설계를 필요로 하는 기존의 컨볼루션 신경망(convnets)과 비교해도 경쟁력 있는 성능을 달성할 수 있다.
- DINO는 ViT 기반 k-NN 분류의 뛰어난 성능과 장면 레이아웃 정보를 포함하는 특성을 가진다.
- 향후 무작위로 수집된 대량의 이미지에서 DINO를 활용한 대형 ViT 모델의 사전 학습을 탐구할 계획

Medical Imaging Intelligent Reality Lab

ASAN Medical Center

Thank You