EECS ELECTRICAL ENGINEERING AND COMPUTER SCIENCE
UNIVERSITY OF MICHIGAN

ICCV'21

EMERGING PROPERTIES IN SELF-SUPERVISED VISION TRANSFORMERS

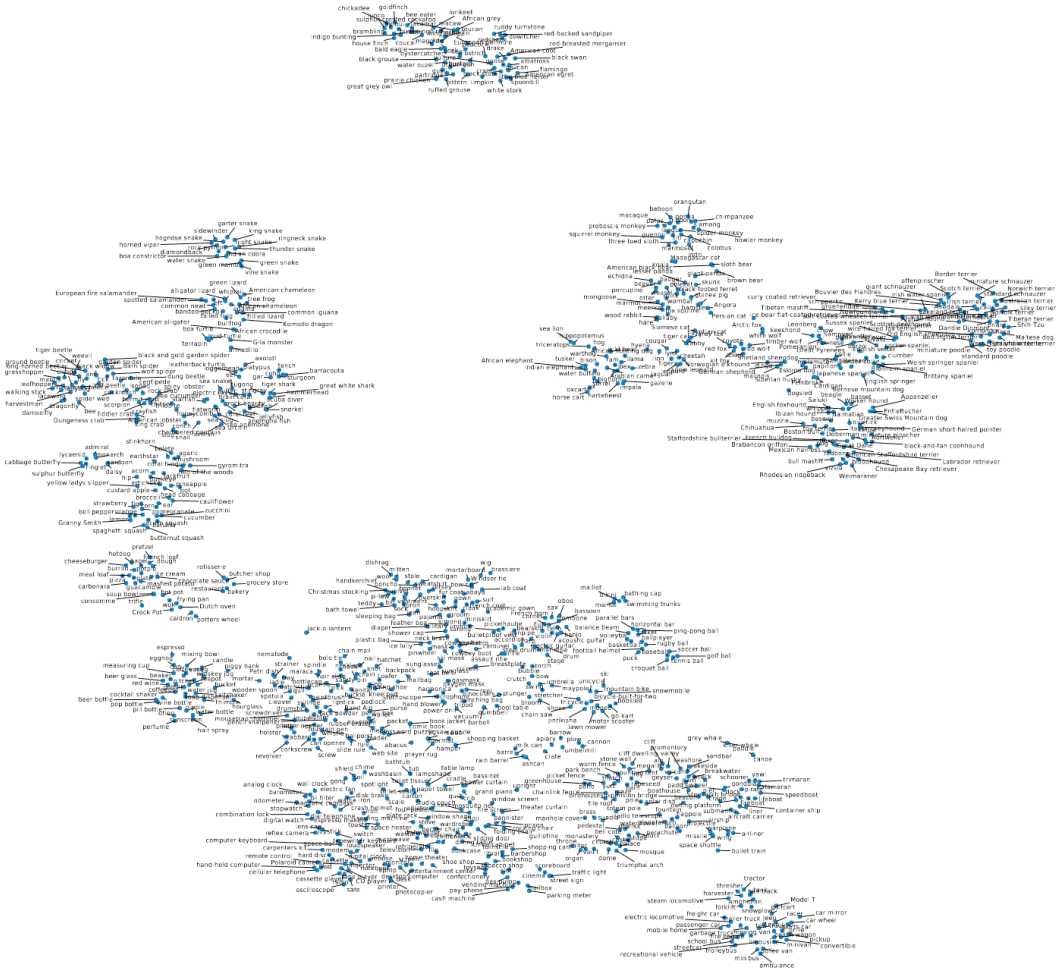
Presented by

Youren Zhang | yourenz@umich.edu

Outline

- Brief intro on representation learning
  - Supervised learning *vs.* representation learning
  - Contrastive learning
- Proposed method - **DINO**
- Experiments

Feel free to ask question any time!

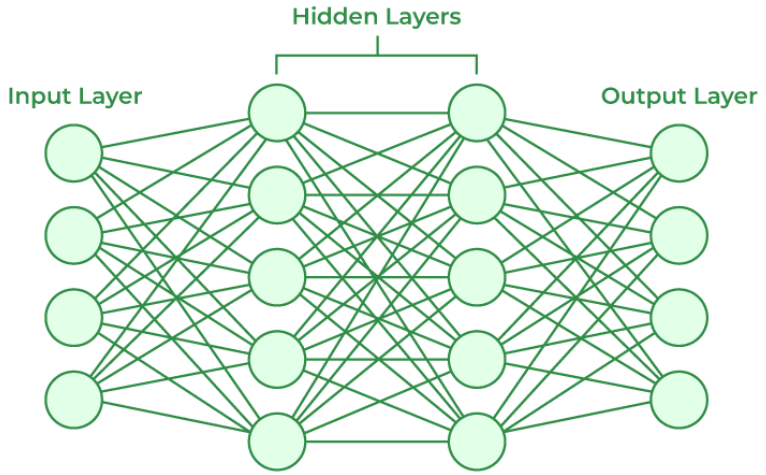


ImageNet classes representation by DINO

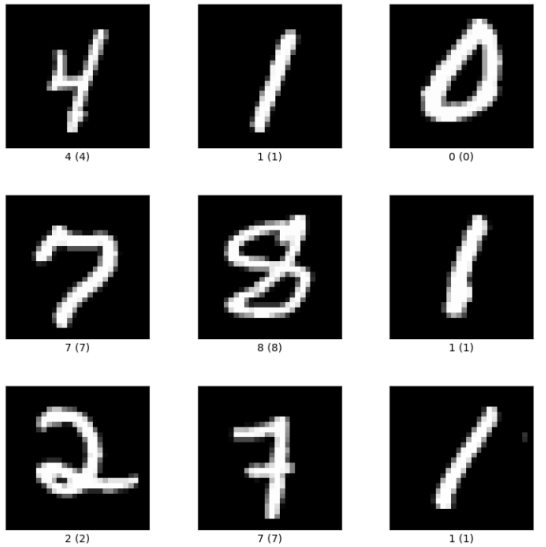
Supervised Learning

Inputs

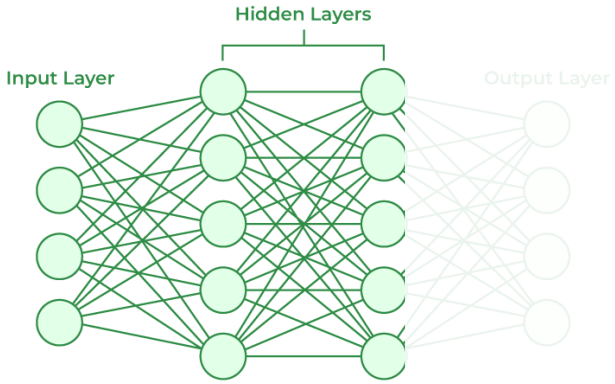
Model



Prediction

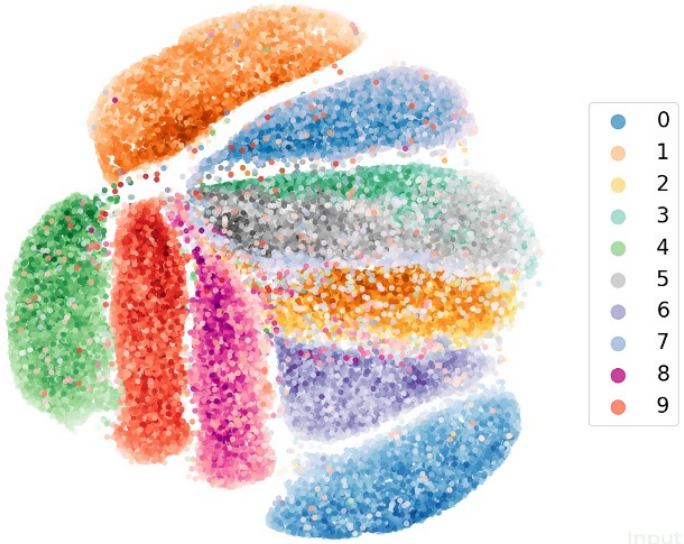


Supervised Learning

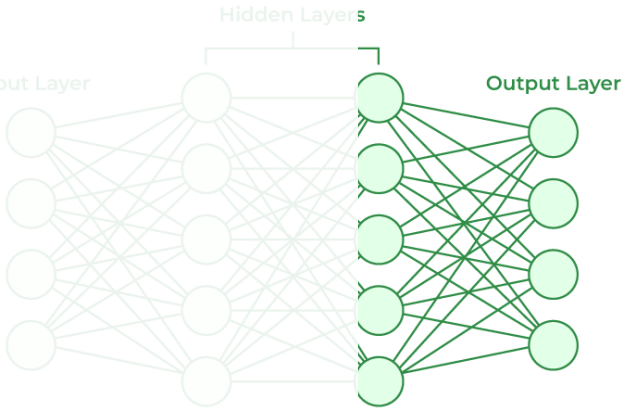


Inputs → Encoder →

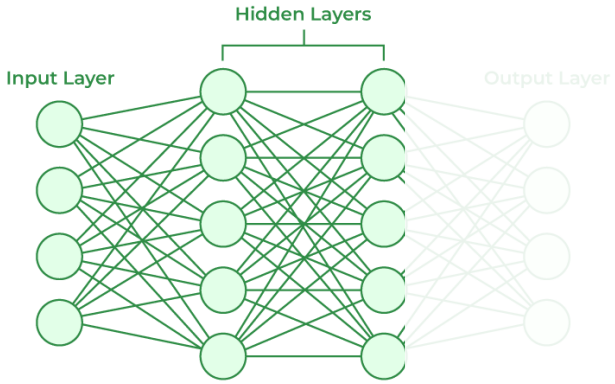
Representations



→ Decoder → Predictions

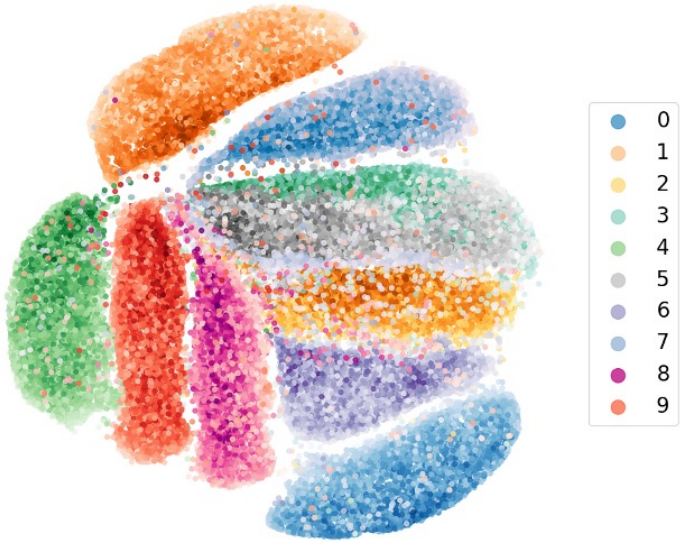


Representation Learning



Inputs → Encoder →

Representations



HOW?

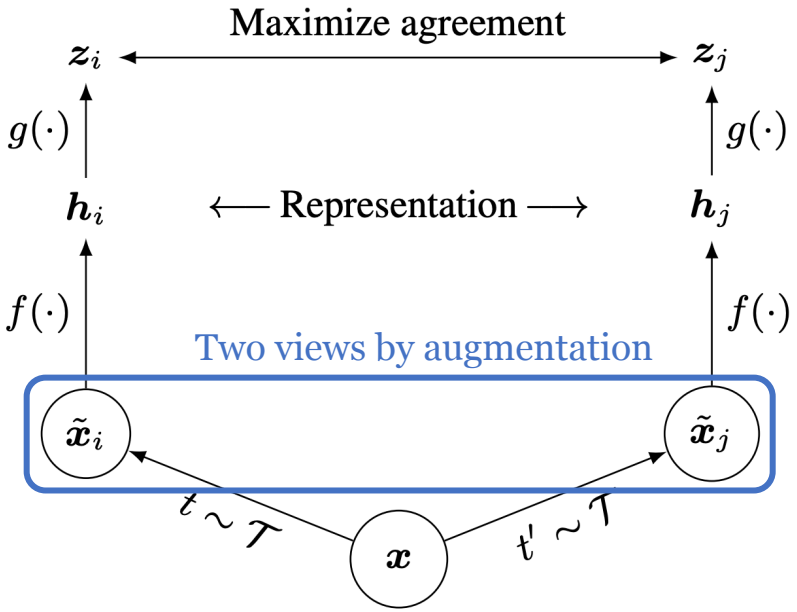
Representation Learning

Self-Supervised Learning

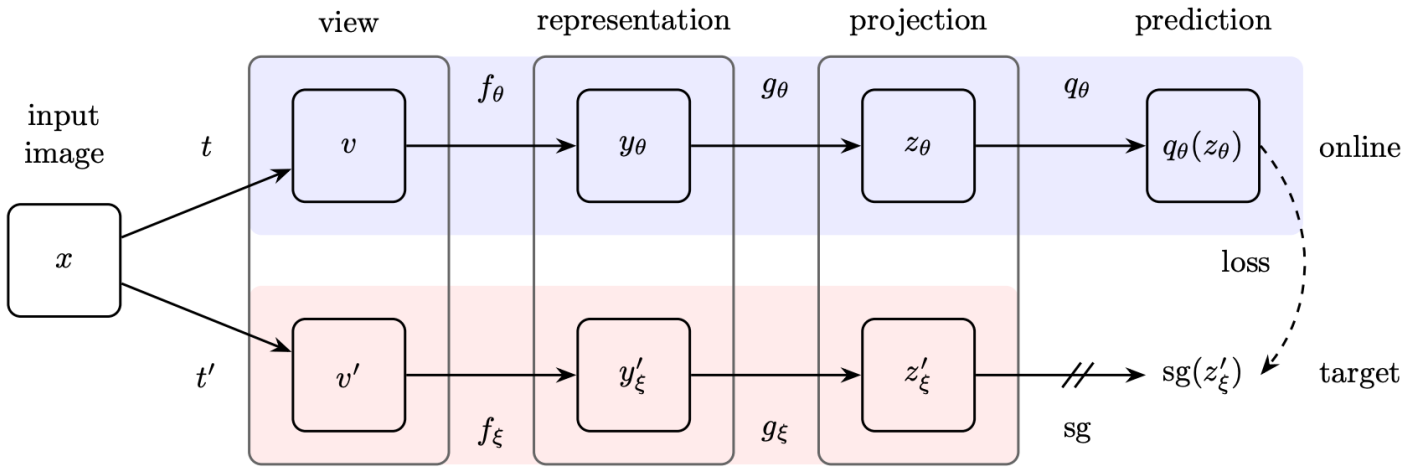
"In self-supervised learning, the system learns to predict **part of its input** from the **other part of input**. In other words, a portion of the input is used as a **supervisory signal** to a predictor fed with the remaining portion of the input."

-- Yann LeCun

Contrastive Learning

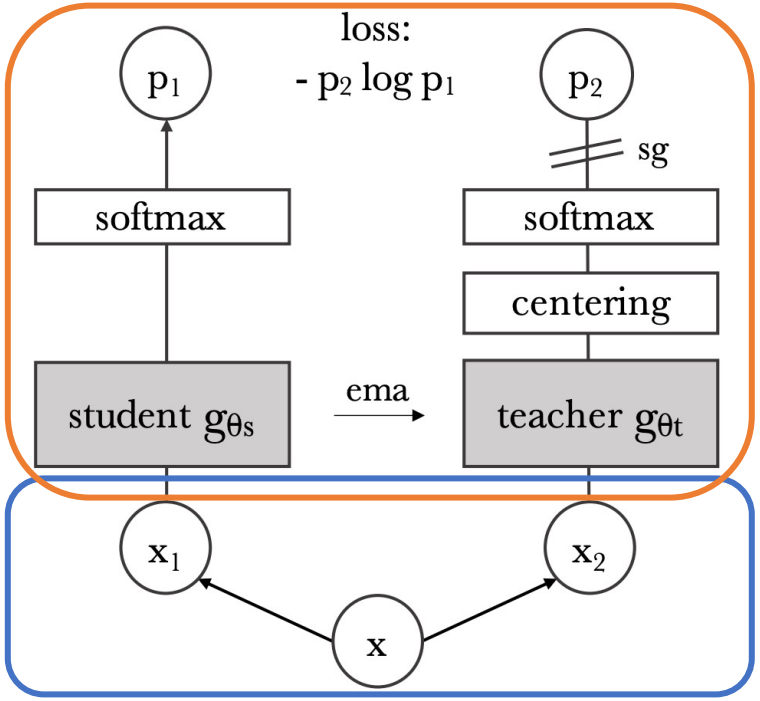


SimCLR (ICML'20)



BYOL (NIPS'20)

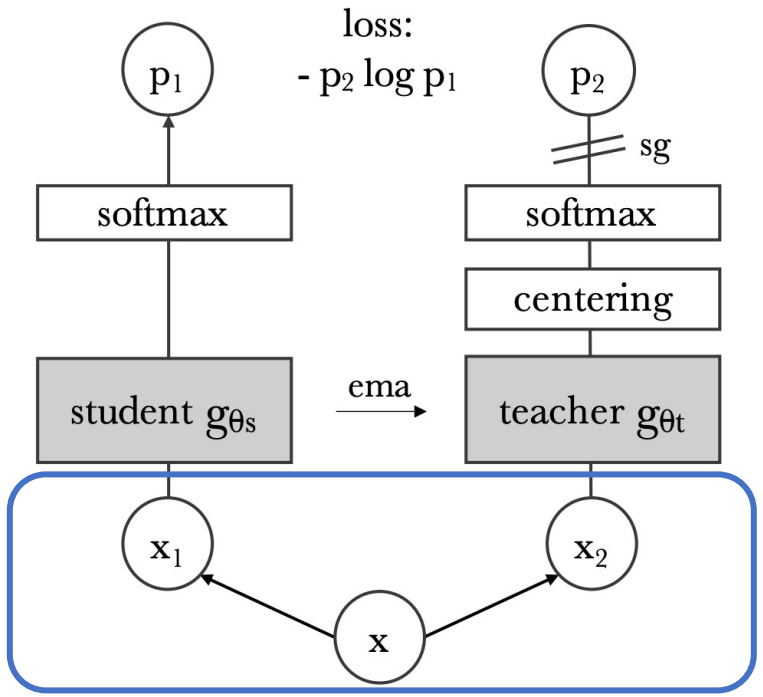
Self-**DI**stillation with **NO** label (It is a **SSL training method**, not a model arch)



Step 2: Self-distillation

Step 1: Augment data with different views

DINO – Step 1

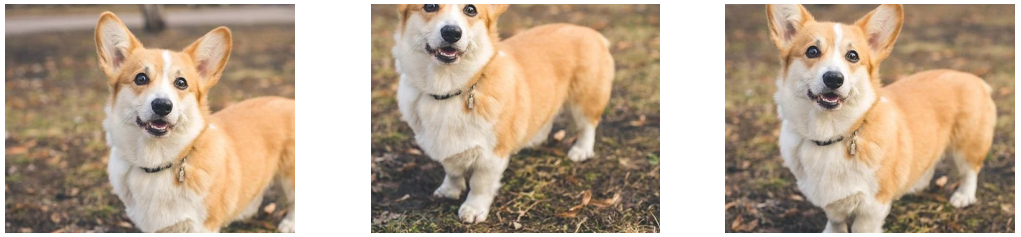


Step 1: Augment data with different views

Multi-Crop Strategy (NIPS'20)

Image Distortion

Global Views



Local Views

DINO – Step 1

Encouraging "local-to-global" **correspondences** ⟵ Multi-Crop Strategy (NIPS'20)

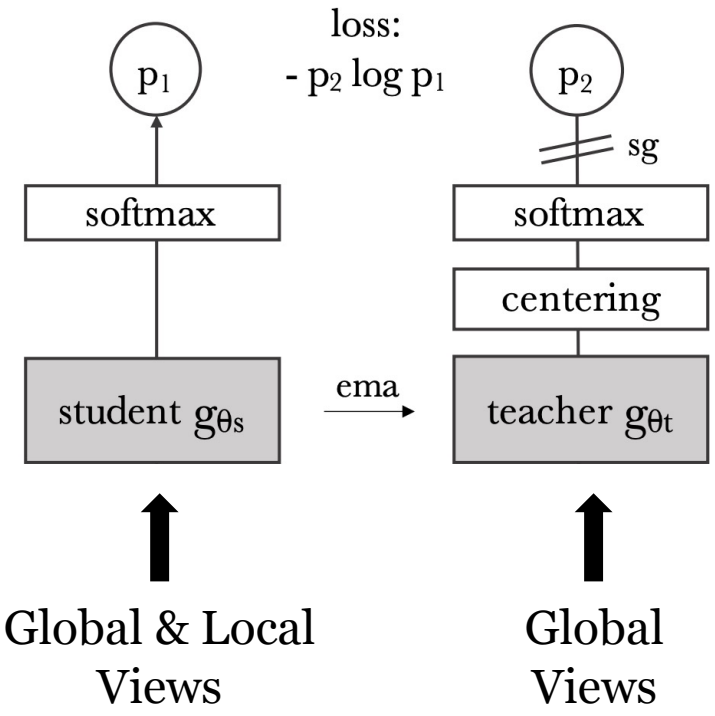
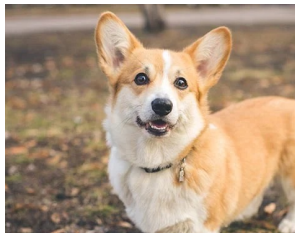


Image Distortion

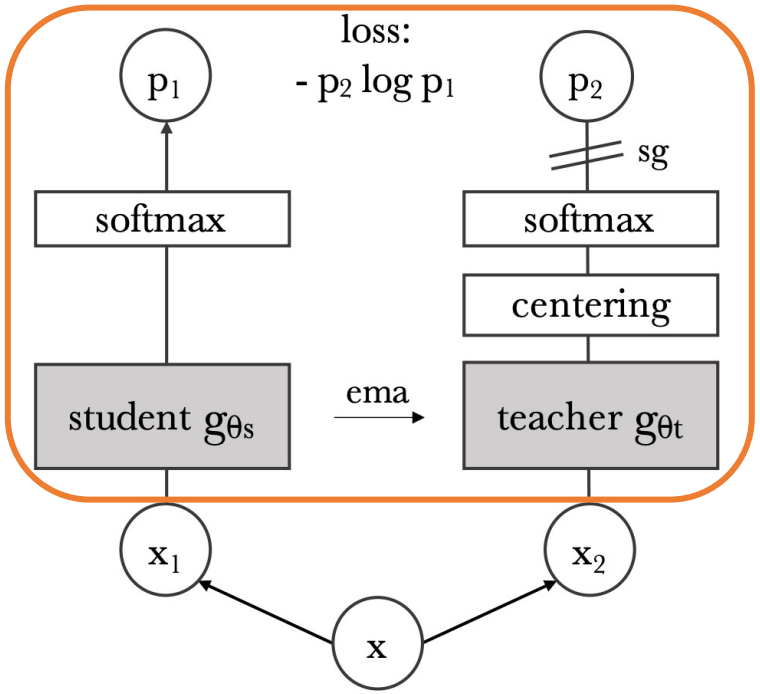
Global Views

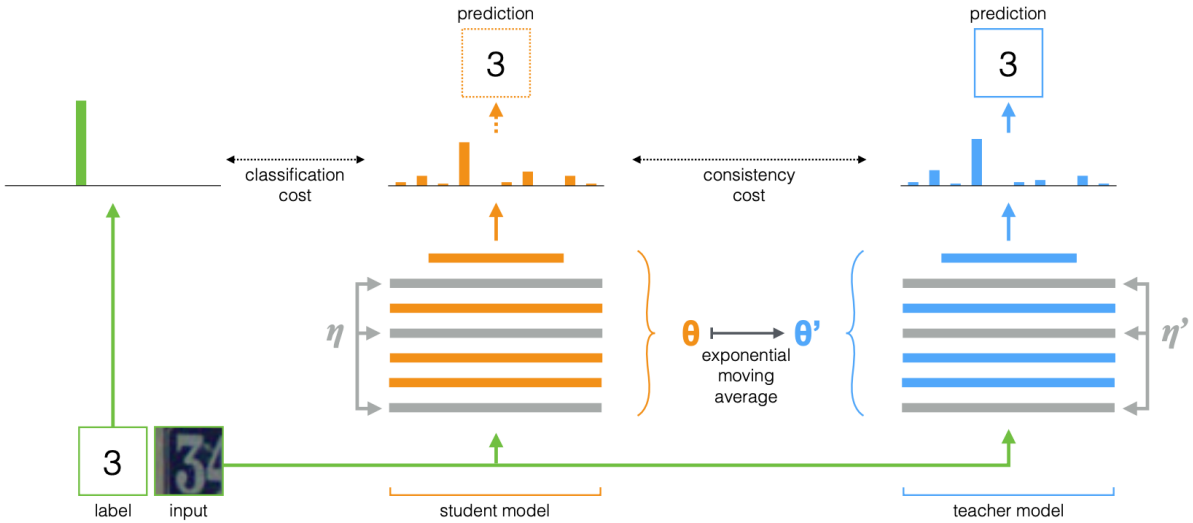


Local Views

DINO – Step 2

Step 2: Self-distillation ➡ Train student network to **match output** of the teacher network

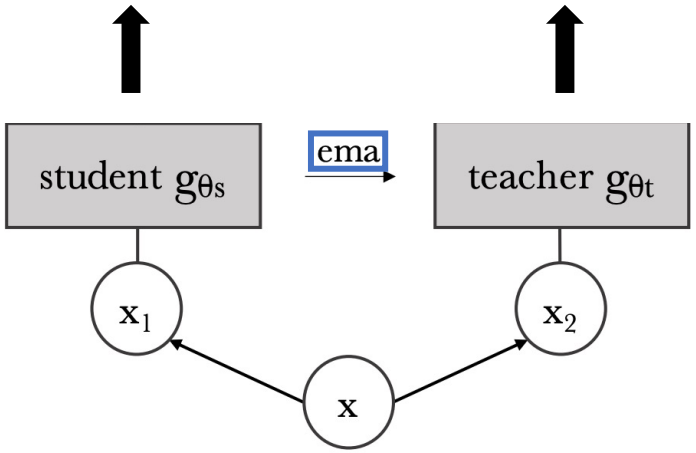




Mean Teacher (NIPS'17)
(as a form of model ensembling)

DINO – Step 2

sim (Student Representations , Teacher Representations)



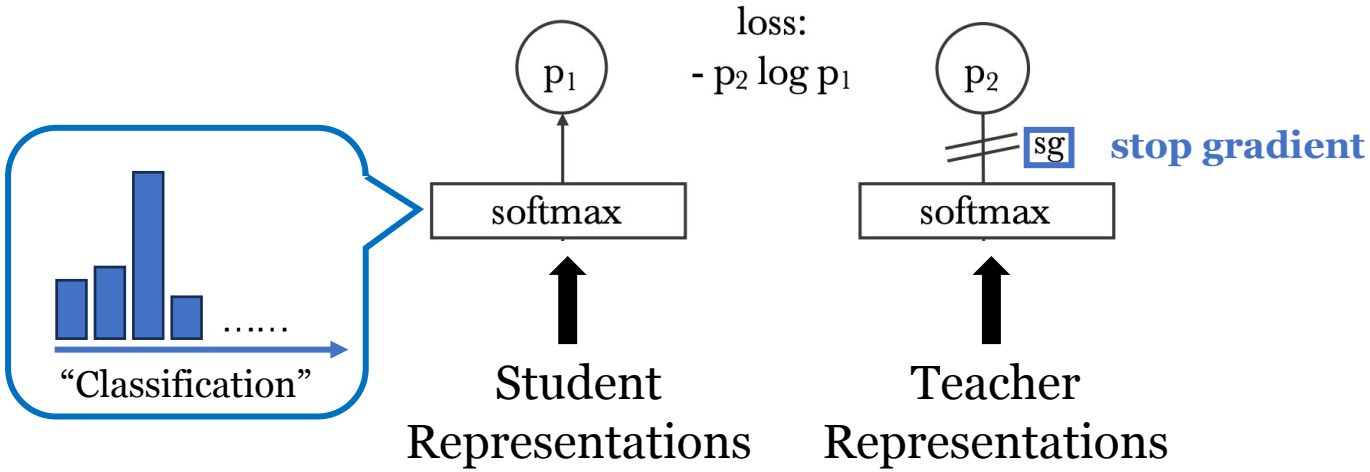
exponential moving average

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

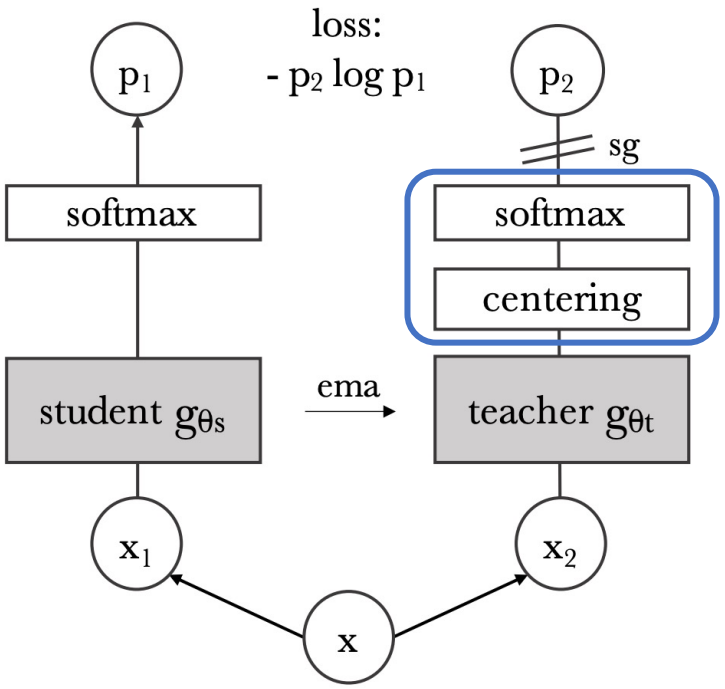
BYOL (NIPS'20) – l_2 distance

$$\left\| \text{Student Representations} , \text{Teacher Representations} \right\|_2$$

DINO – cross entropy



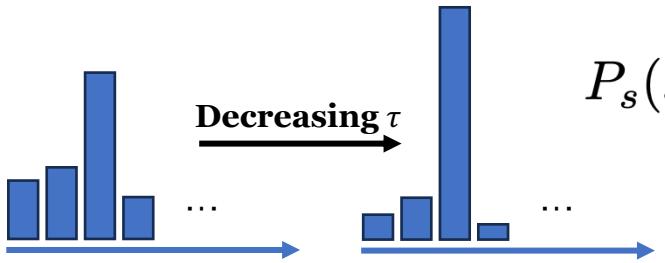
DINO – Step 2



By subtracting mean

(constant representations)

Avoiding **collapse** by **centering** and **sharpening**

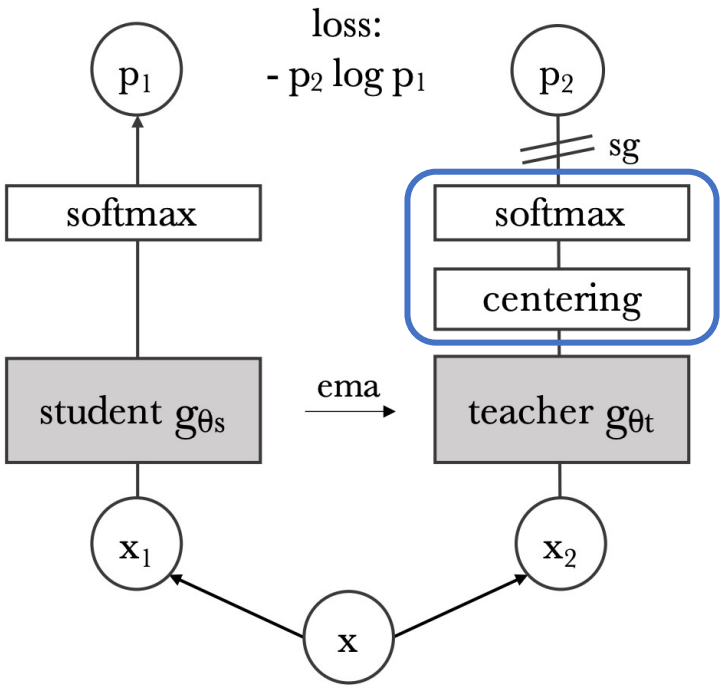


$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)} / \tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)} / \tau_s)},$$

By temperature parameter

Quiz: To sharpen a distribution, τ should increase/decrease.

DINO – Step 2



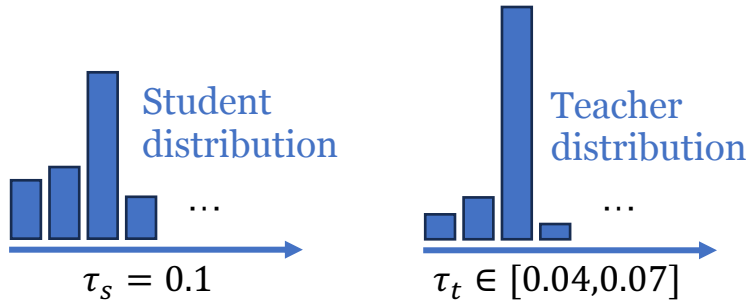
Avoiding occurrence of dominant dimension, but encourages a uniform distribution

(i.e. constant representations)

Avoiding **collapse** by **centering** and **sharpening**

The opposite.

Applying both operations **balances** these effects.



EXPERIMENTS

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

Linear and KNN evaluation on ImageNet

Dataset:

- ImageNet

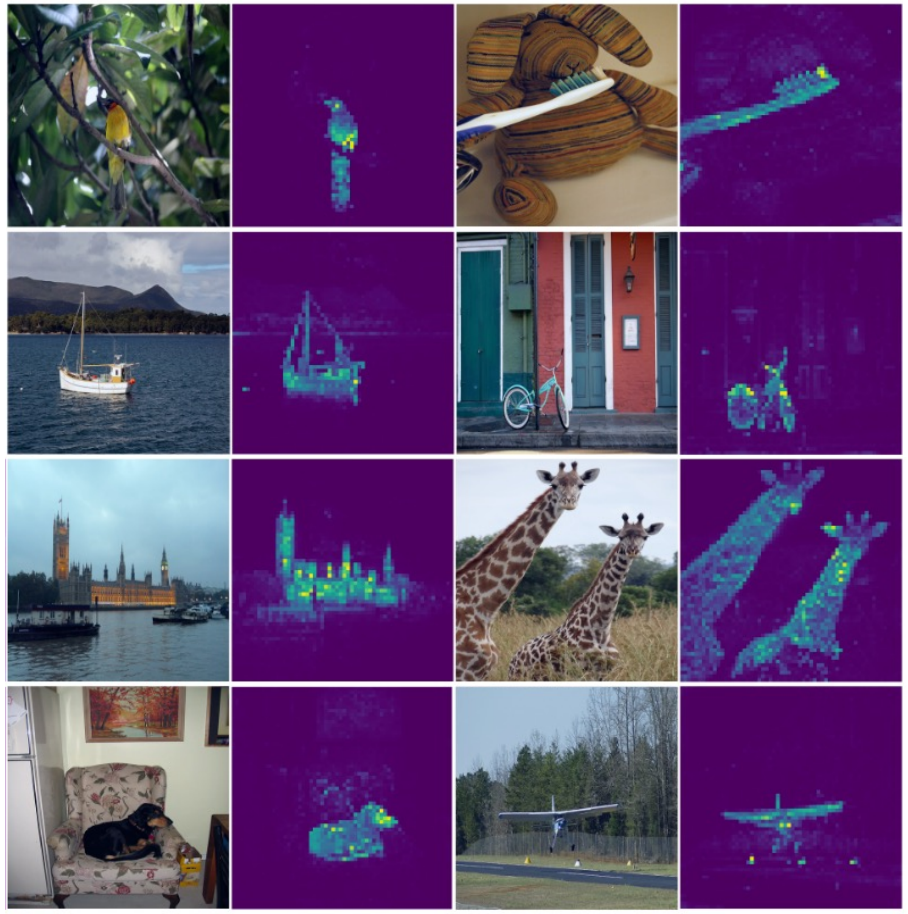
Evaluation Protocols:

- Linear: Train a linear classifier on the learned representation
- KNN evaluation (CVPR'18): Weighted nearest neighbor classifier

Results:

- DINO outperform other SSL learning methods on same model arch.
- The KNN classifier works surprisingly well on ViT-based model.

EXPERIMENTS



Attention map from a ViT/8 trained with no supervision

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**

DINO shows competitive result even though the training objective nor our architecture are designed for dense tasks

Ablation – Collapse Study

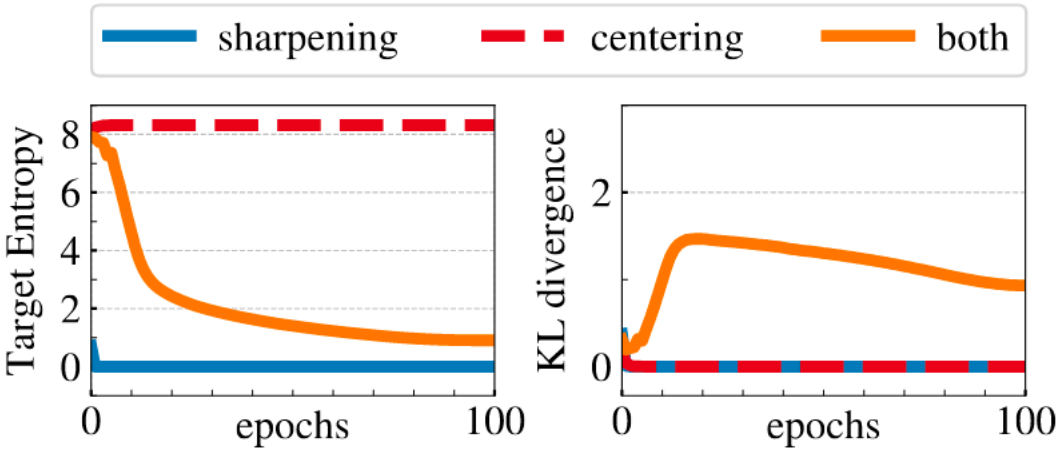
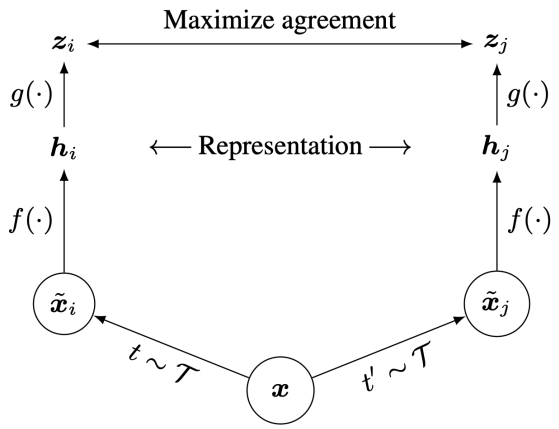


Figure 7: **Collapse study. (left)**: evolution of the teacher's target entropy along training epochs; **(right)**: evolution of KL divergence between teacher and student outputs.

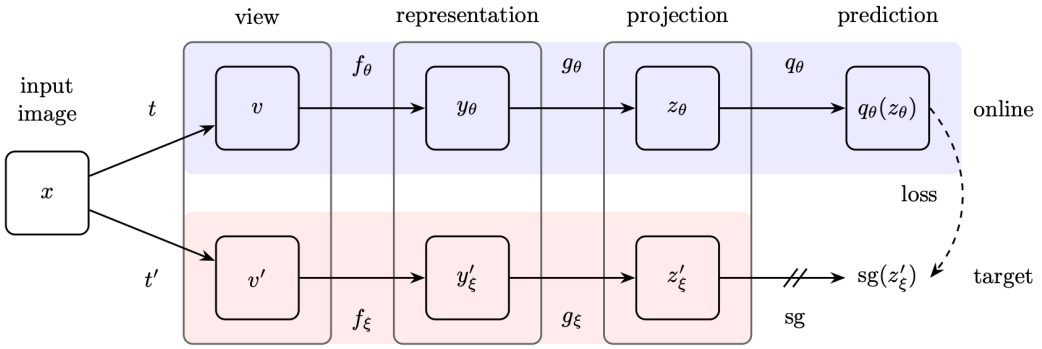
The entropy converges to different values with no centering or no sharpening, indicating that both operations induce different form of collapse. Applying both operations **balances** these effects

More about Collapse

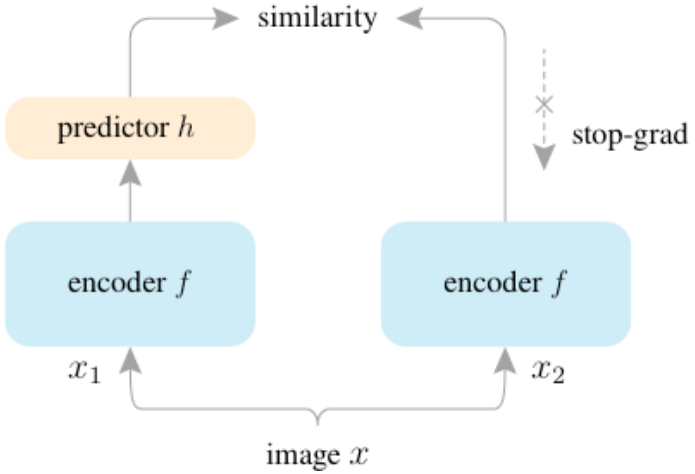
To prevent **collapsing** to a constant …



SimCLR (ICML'20)



BYOL (NIPS'20)



SimSiam (CVPR'21)

Negative Pairs

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

Momentum Encoder

$$\xi \leftarrow \tau\xi + (1-\tau)\theta.$$

Stop-grad matters!

THANKS!