Emerging Properties in Self-Supervised Vision Transformers

ICLR 2021

Group H: Sharon Barak, Tristan de Blegiers, Jacob O'Quinn, Jacob Pfaffenbichler

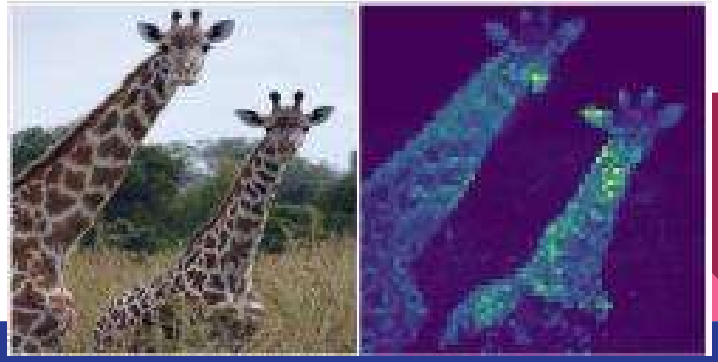
Overview

1. Introduction
2. Related Work
3. Model Architecture
4. Implementation details & figures
5. Results



Introduction

- Transformers have emerged as a solution for CV tasks
- Prior use of transformers have relied on self supervised learning
- SSL has also been used with ConvNets
- Combine self-supervised learning with Vision Transformers
- Knowledge **di**stillation with **no** labels



DINO in Action



Source:
https://www.youtube.com/watch?v=8I1RelnsgMw

Related Topics

- Self-Training and Knowledge Distillations
- Self-Supervised Learning



Self-training & Knowledge Distillation

- Self Training: propagating small set of annotations to large set of unlabeled instances
- Hard vs Soft Labels
- Assignment of soft labels often referred to as knowledge distillation
- Extend Knowledge Distillation to data with no Labels



Self-Supervised Learning

- SSL: Learning from unlabeled data
- Trains models by discriminating images with data augmentations



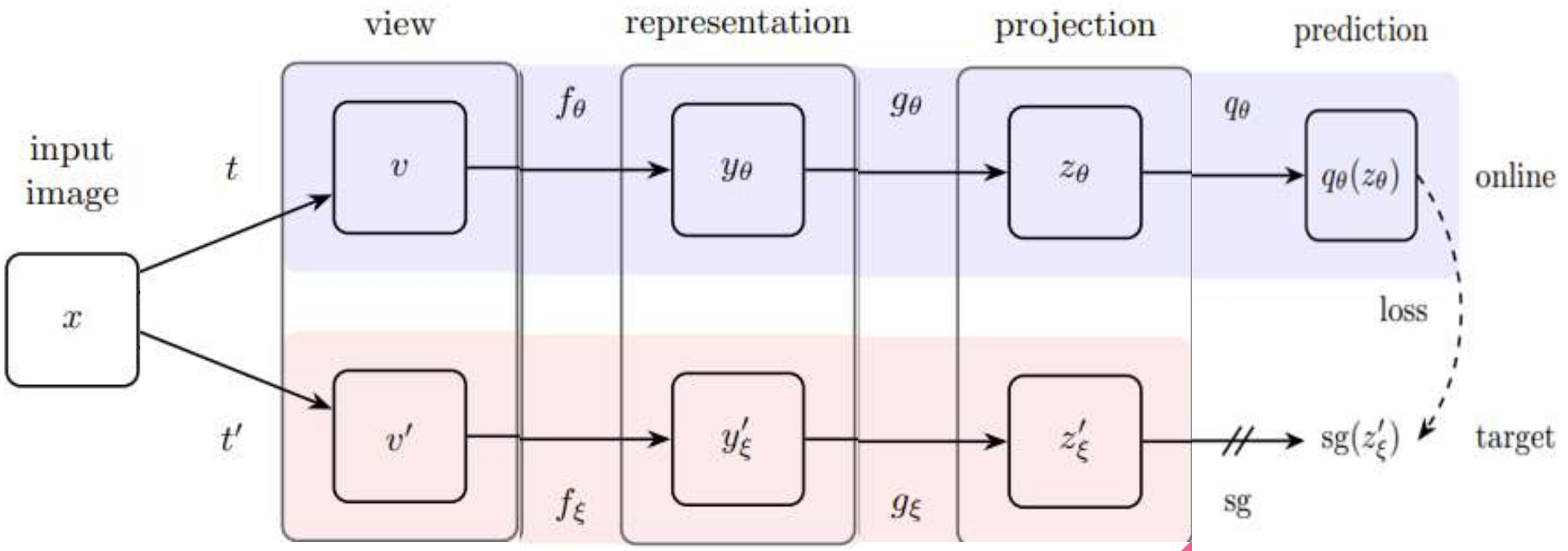
Self-Supervised Learning (cont).

- Recently, it has been found that unsupervised features can be learned without discrimination.

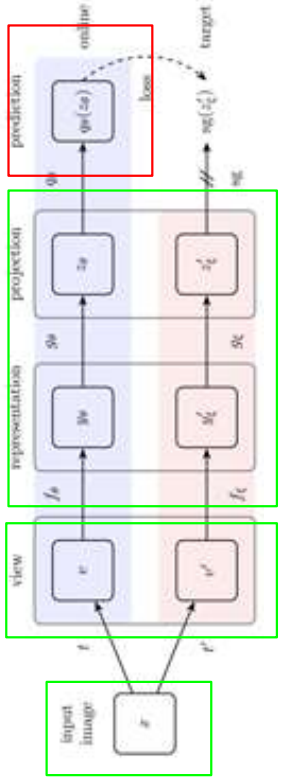
- Compare different data augmentations of the same image

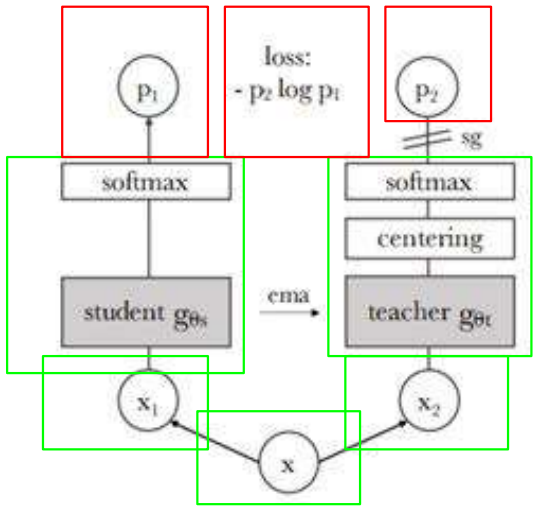


BYOL

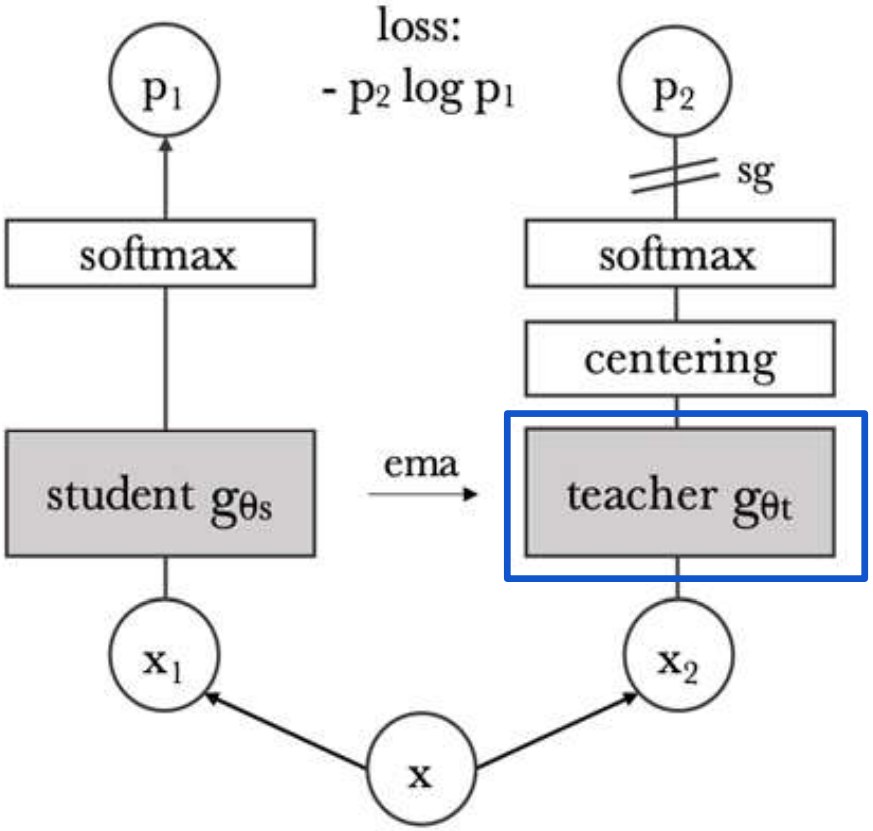

view
representation
projection
prediction
input image
t
t'
x
v
v'
f_θ
f_ξ
y_θ
y'_ξ
g_θ
g_ξ
z_θ
z'_ξ
q_θ
$q_\theta(z_\theta)$
online
loss
$\mathrm{sg}(z'_\xi)$
target
sg

BYOL vs DINO

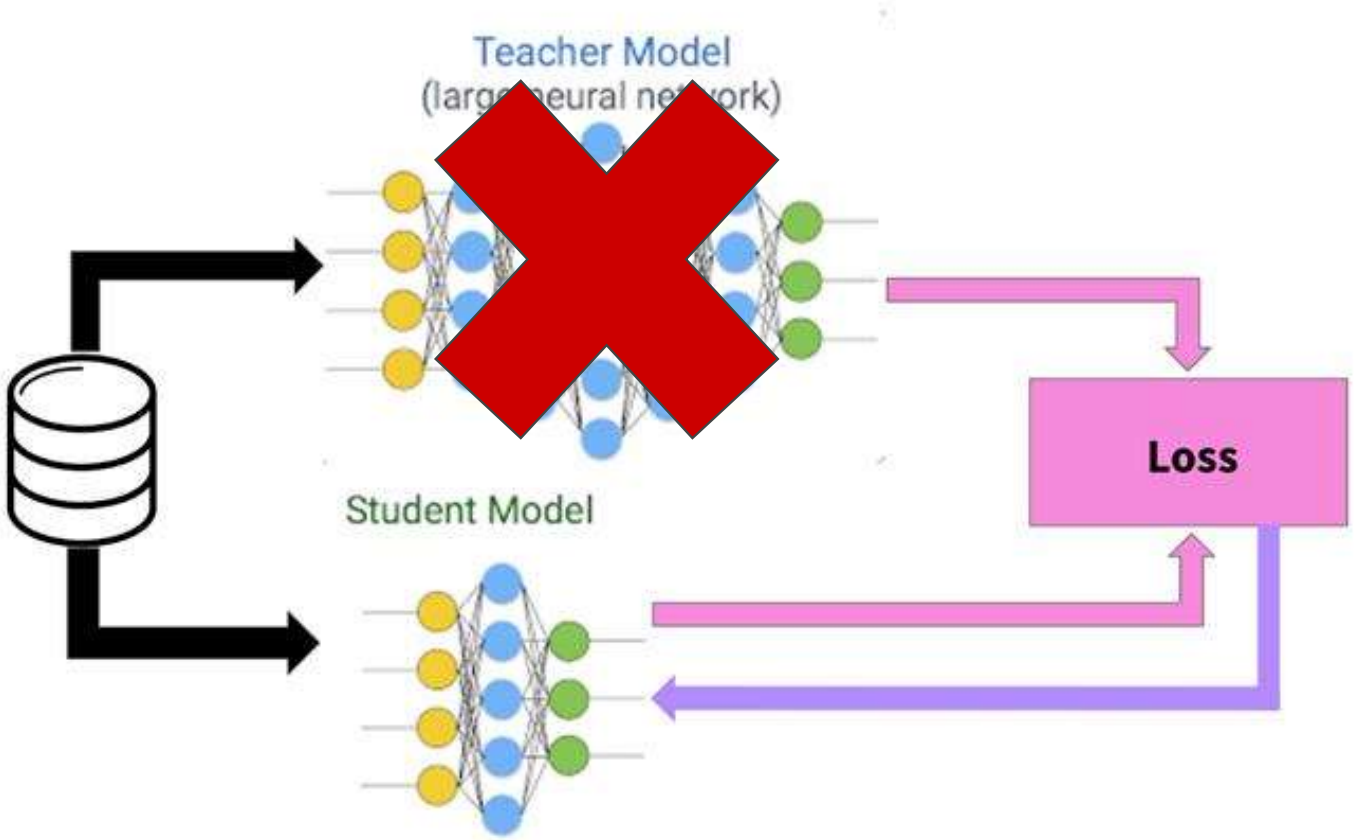

online
target
loss
predictor
projection
representation
view
input image


loss:
- p2 log p1
p1
p2
sg
softmax
softmax
centering
student gθs
ema
teacher gθt
x1
x2
x

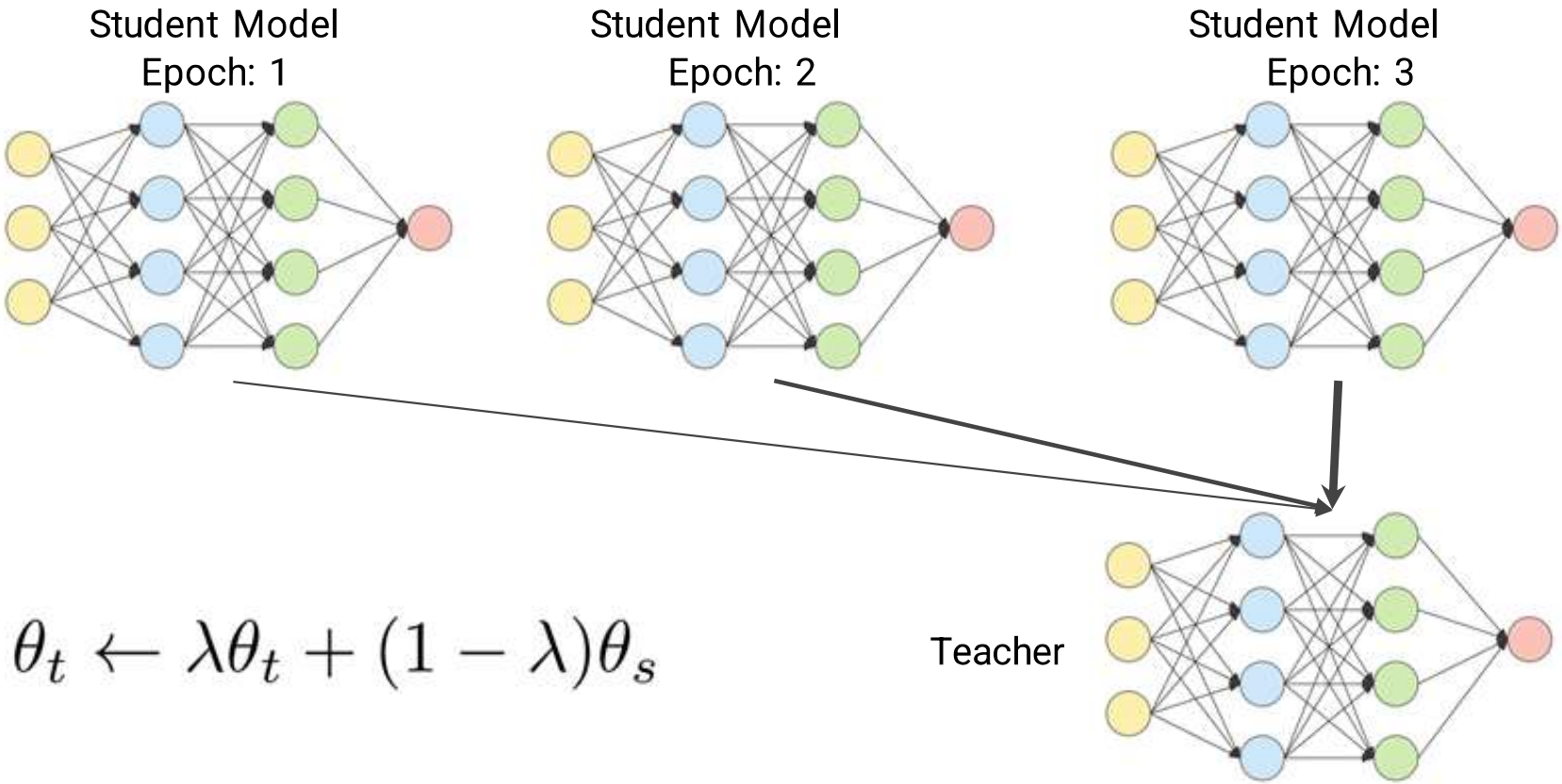
High Level Overview



Teacher

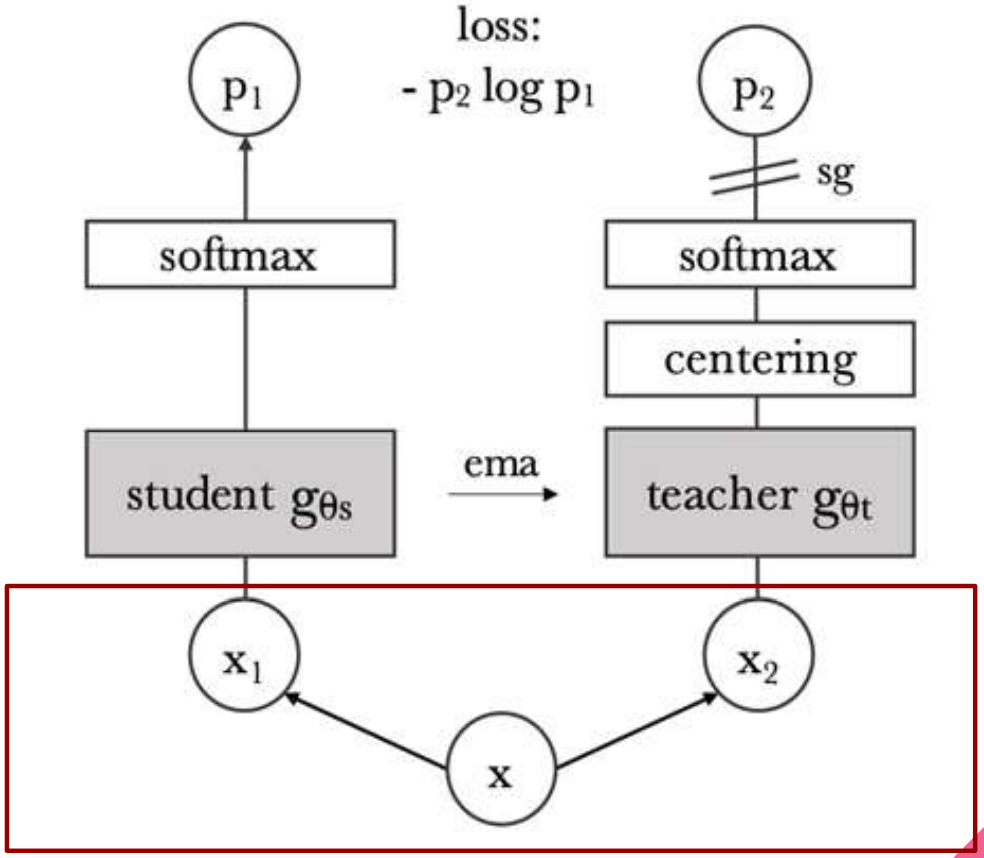


Exponential Moving Average

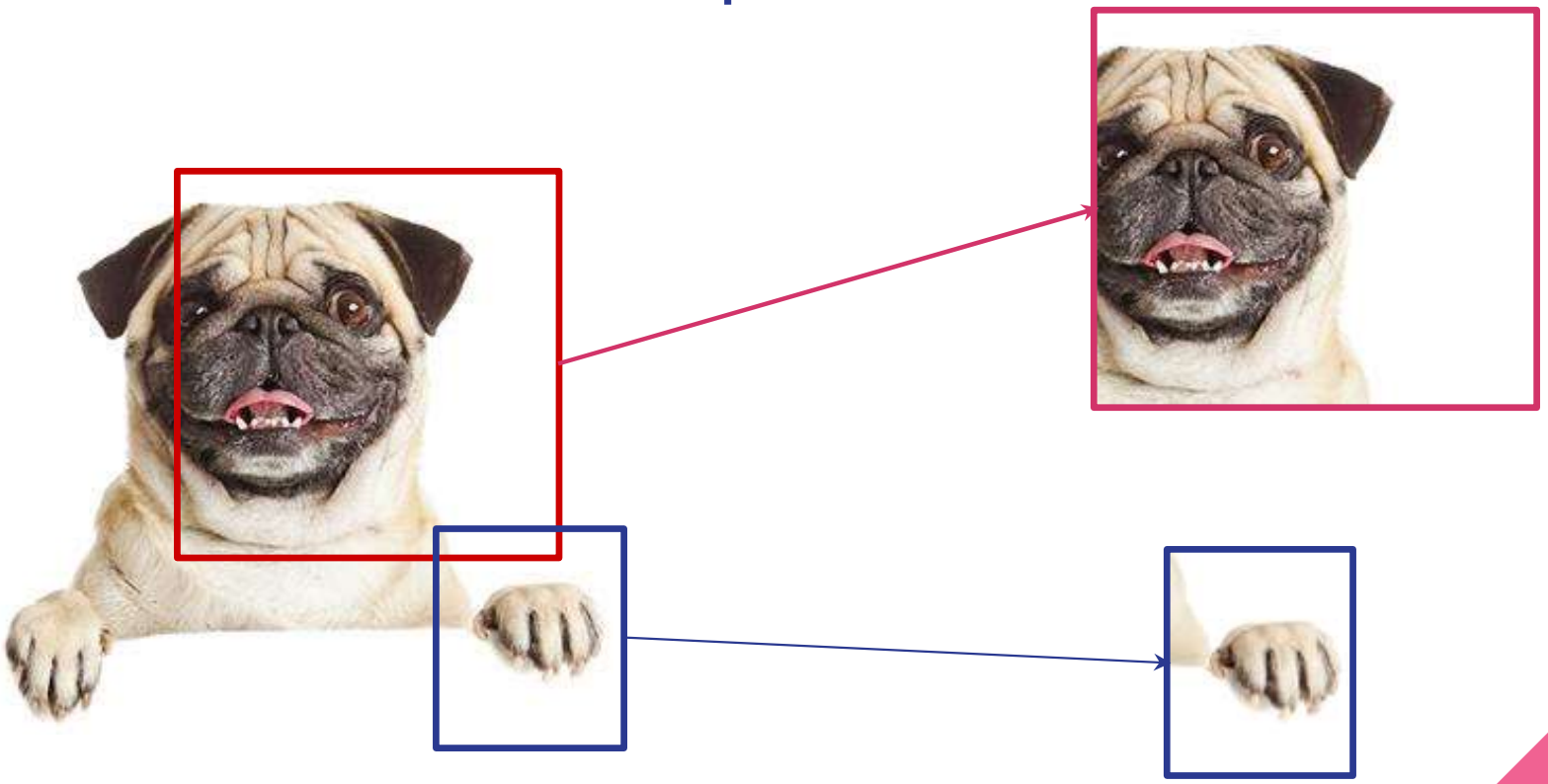


$$\theta_t \leftarrow \lambda\theta_t + (1-\lambda)\theta_s$$

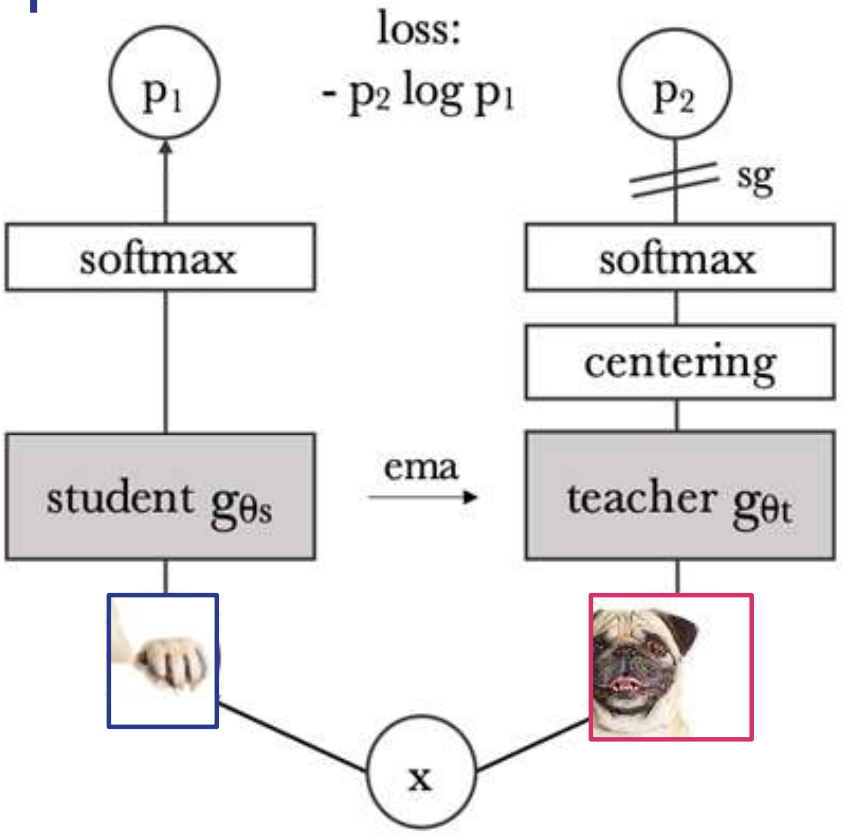
Augmentation



Global vs Local Crop

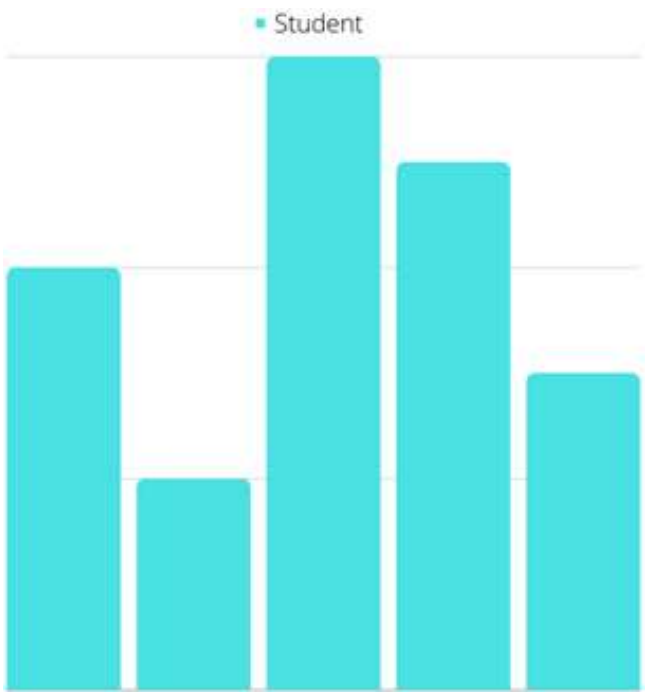


What Is The Output

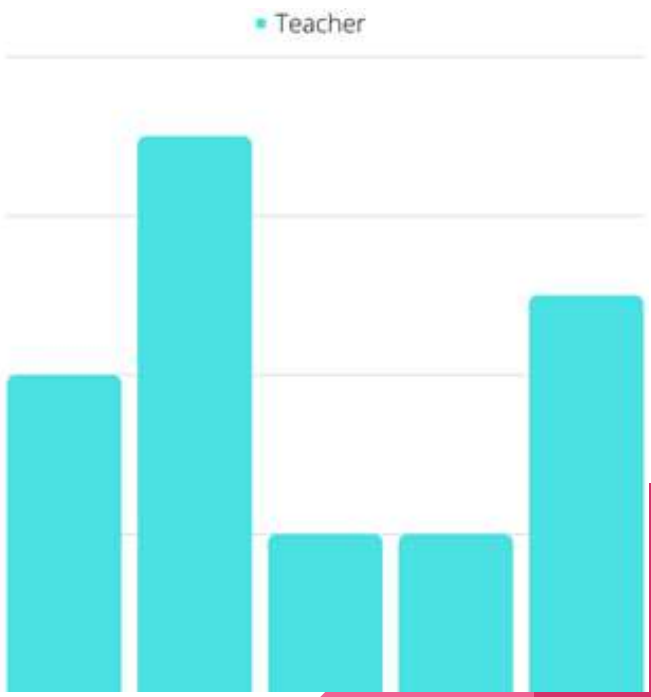


Encoded Vector

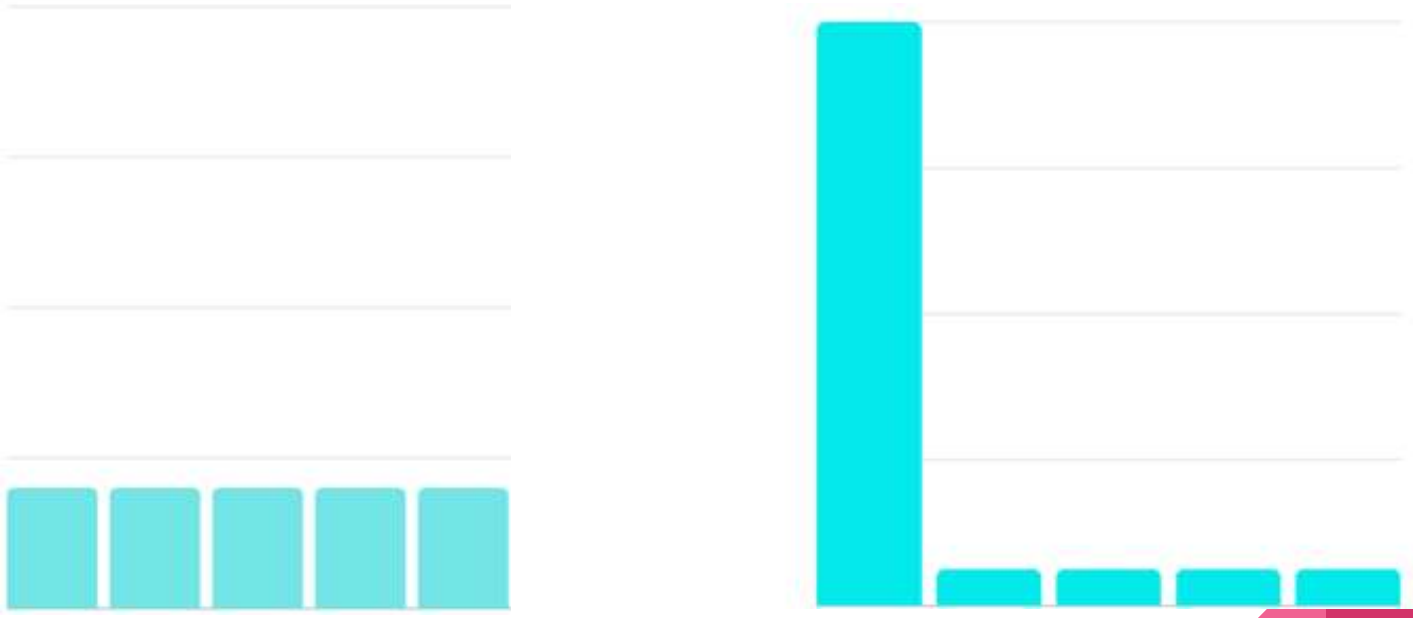
Student => [.2, .1, .3, .25, .15]



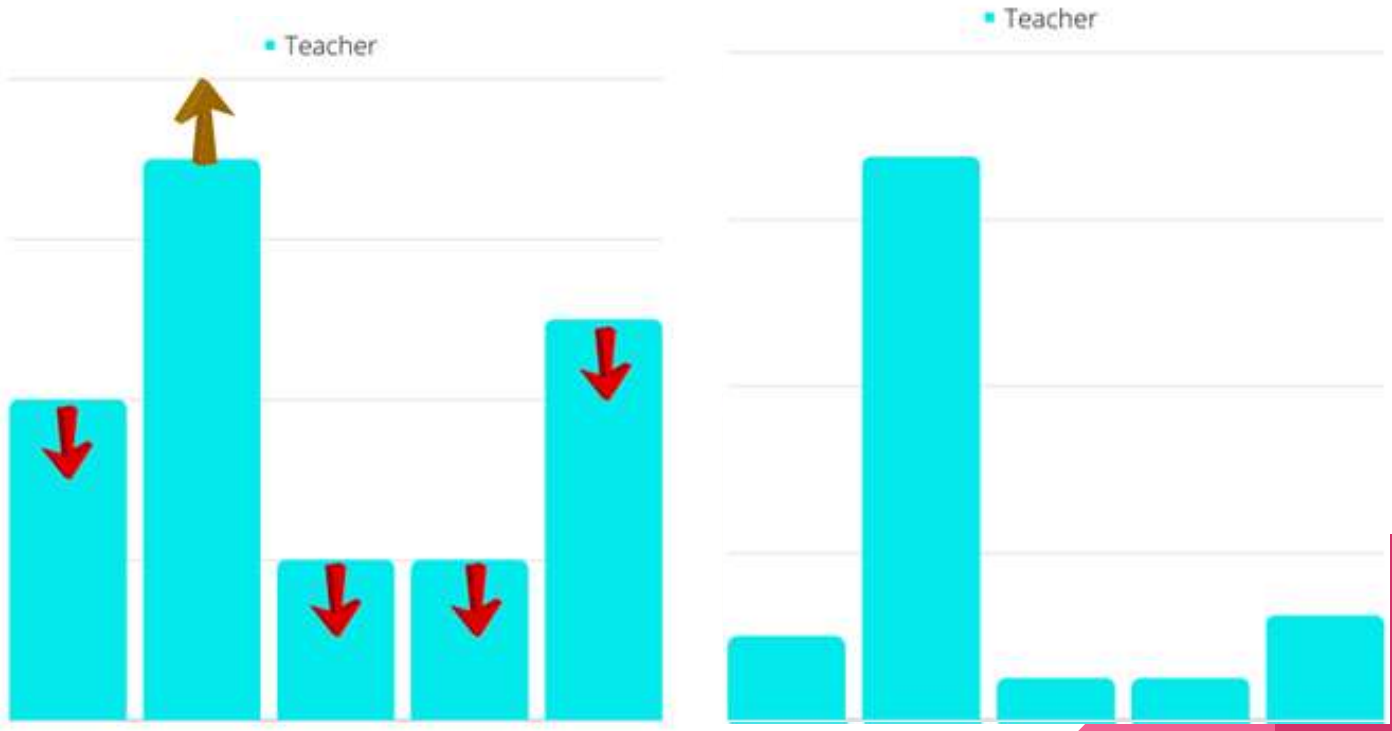
Teacher => [.2, .35, .1, .1, .25]



Avoiding Collapse



Temperature → Sharpening Results

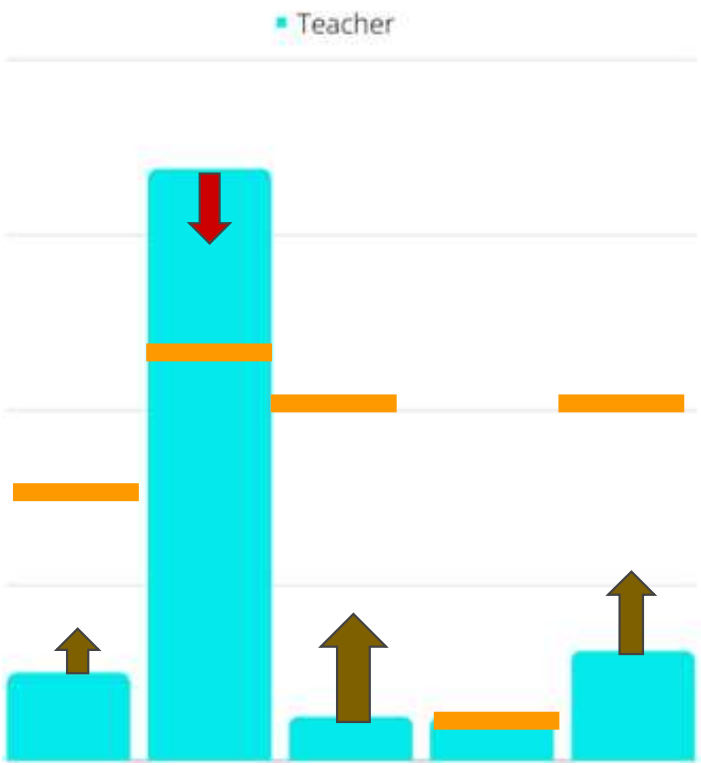


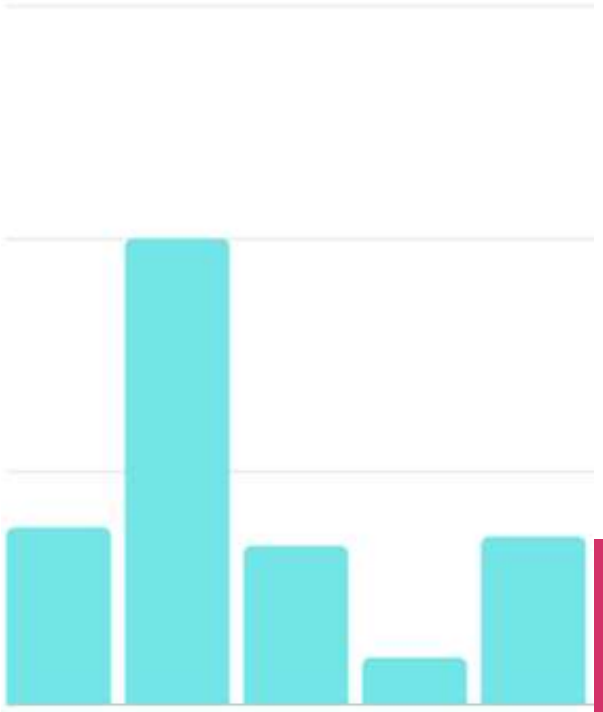
Temperature → Sharpening Results

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$$

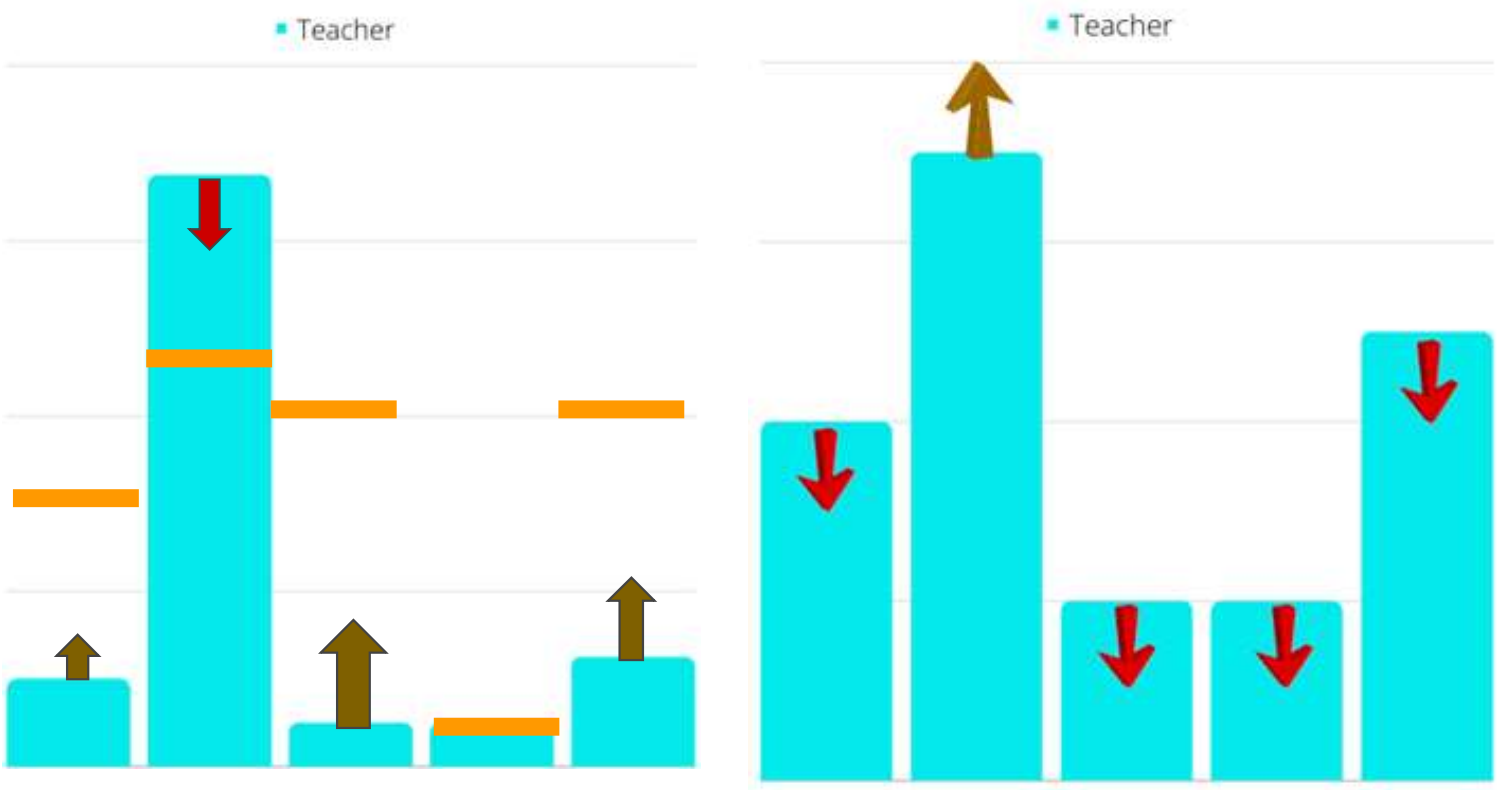
$$P_s(x)^{(i)} = \frac{\exp(g_{\theta_s}(x)^{(i)}/\tau_s)}{\sum_{k=1}^{K} \exp(g_{\theta_s}(x)^{(k)}/\tau_s)}$$

Center → Uniformity





Avoiding Collapse



Center → Uniformity

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

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

Implementation Details

The transformer is exactly equal to DeiT

model	blocks	dim	heads	#tokens	#params	im/s
ResNet-50	–	2048	–	–	23M	1237
ViT-S/16	12	384	6	197	21M	1007
ViT-S/8	12	384	6	785	21M	180
ViT-B/16	12	768	12	197	85M	312
ViT-B/8	12	768	12	785	85M	63

NxN = size of image patches. N = 16 and N = 8.

Implementation Details

A learnable class token is appended to the sequence of image patches.

It does NOT come with a label or supervision.



Implementation Details

Pretrained on ImageNet without labels

Adamw optimizer

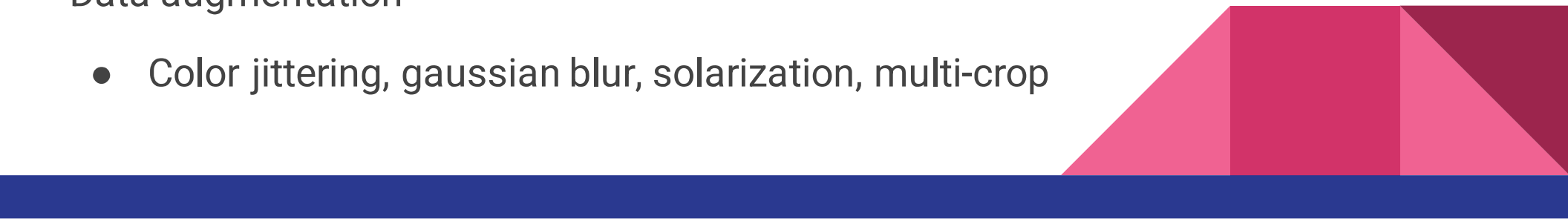
Batch size = 1024

lr = 0.0005*batchsize/256

Weight decay follows a cosine schedule from 0.04 to 0.4

Data augmentation

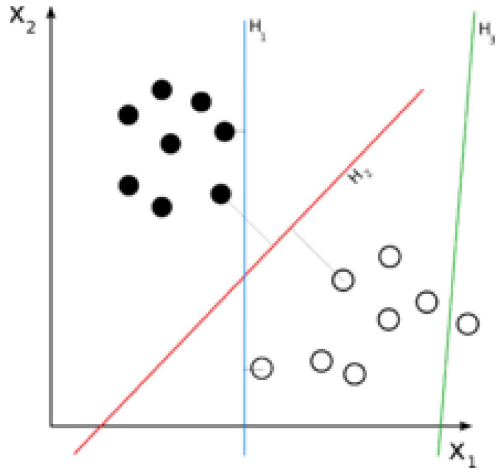
- Color jittering, gaussian blur, solarization, multi-crop



Evaluation Protocols

Learn a linear classifier on frozen features.

- Random resize crops and horizontal flips augmentation during training



Evaluation Protocols

New Method

- Freeze pretrain model and store features of train data on downstream task
- Match image feature to k nearest features
- k = 20

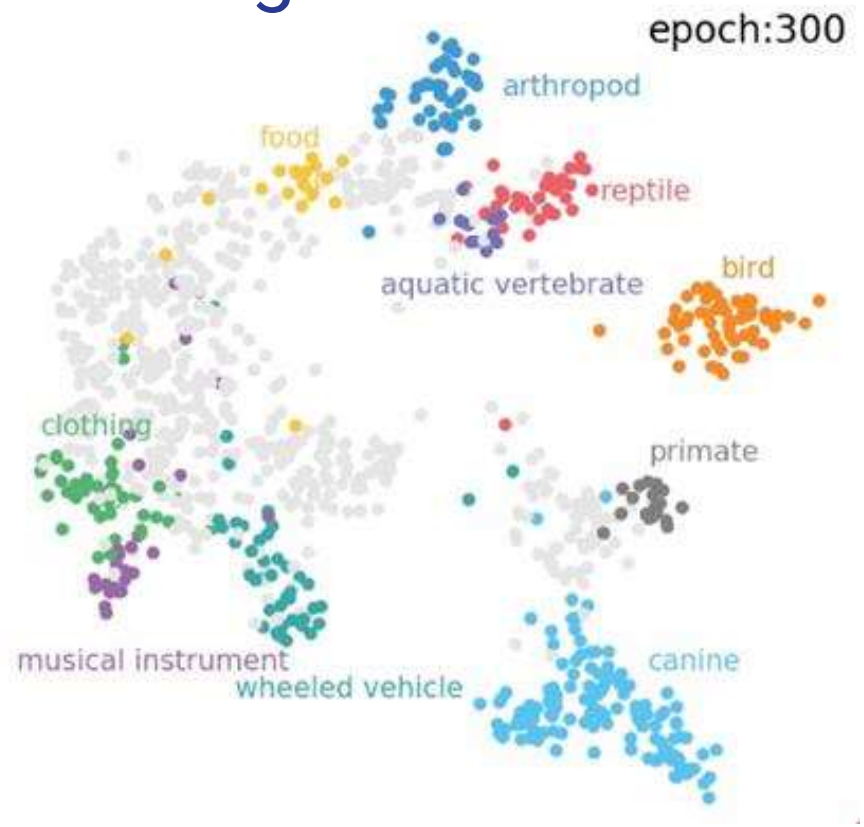


Feature Space Learning

epoch: 0



Feature Space Learning



Note: this is a visualization, the labels are not predetermined in practice

Comparisons

The results reported in this paper of other architectures are run by the authors

Perform k-NN evaluation for models with official released weights



Comparison with Same Architecture - ResNet50

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**

Comparison with Same Architecture - ViT-S

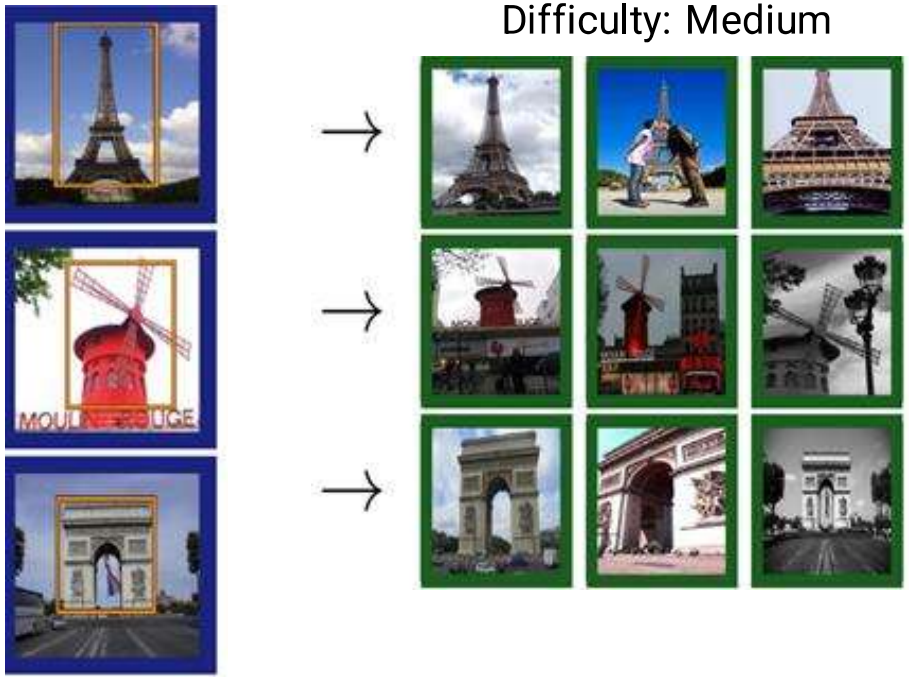
Method	Arch.	Param.	im/s	Linear	k-NN
BYOL* [30]	ViT-S	21	1007	71.4	66.6
MoCov2* [15]	ViT-S	21	1007	72.7	64.4
SwAV* [10]	ViT-S	21	1007	73.5	66.3
DINO	ViT-S	21	1007	**77.0**	**74.5**

Comparison with Different Architectures

Method	Arch.	Param.	im/s	Linear	k-NN
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

Results

Image retrieval:



Results

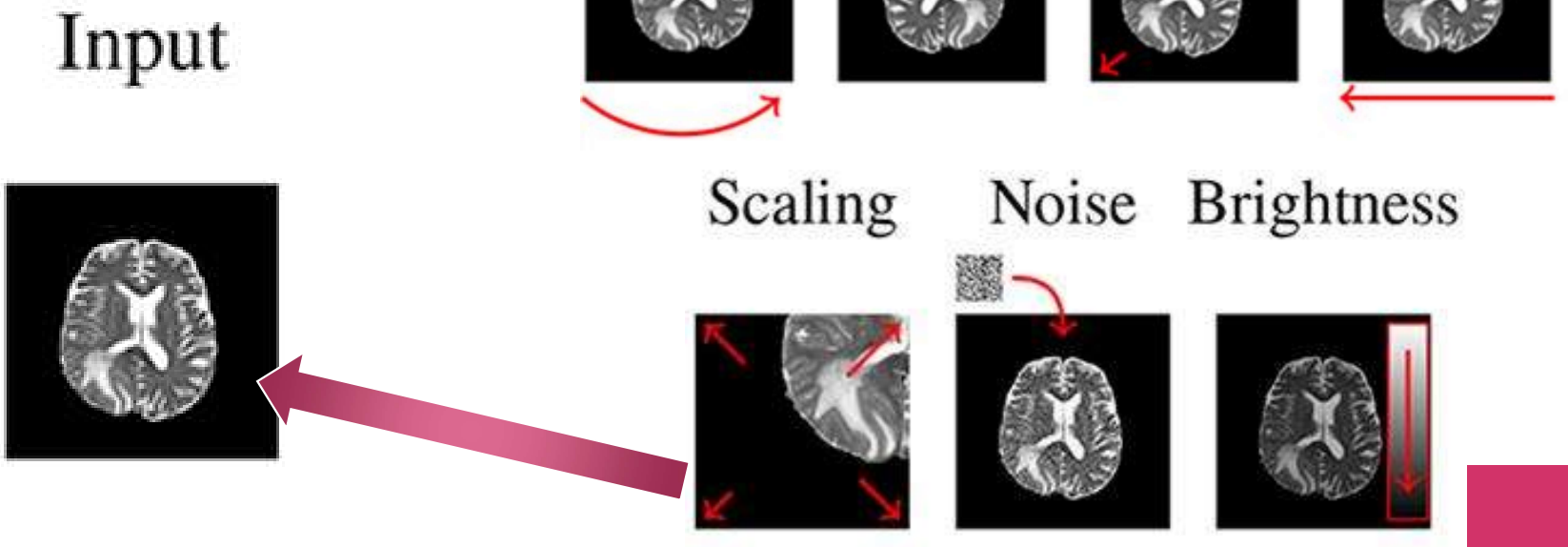
Image retrieval:

Difficulty

Pretrain	Arch.	Pretrain	M	H
Sup. [46]	RN101+R-MAC	ImNet	49.8	18.5
Sup.	ViT-S/16	ImNet	33.5	8.9
DINO	ResNet-50	ImNet	35.4	11.1
DINO	ViT-S/16	ImNet	41.8	13.7
DINO	ViT-S/16	GLDv2	**51.5**	**24.3**

Results

Copy detection:



Results

Copy detection:

Method	Arch.	Dim.	Resolution	mAP
Multigrain [4]	ResNet-50	2048	224^2	75.1
Multigrain [4]	ResNet-50	2048	largest side 800	82.5
Supervised [56]	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**

Results: Segmentation

Supervised



DINO



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

Video segmentation

Scores:

- Region similarity
- Contour accuracy



Results: Segmentation

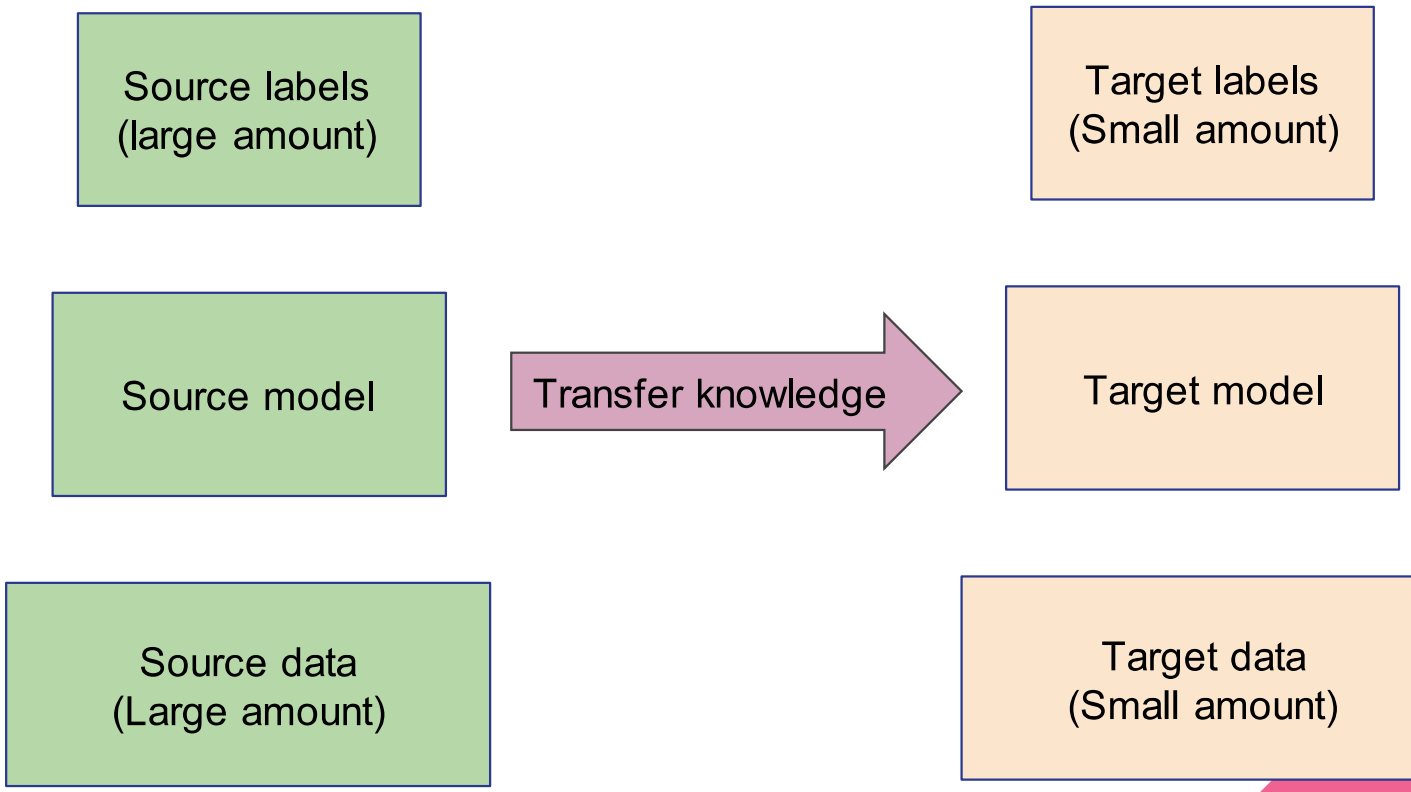
Region

Contour

Video instance segmentation:

Method	Data	Arch.	$\mathcal{J}_m$	$\mathcal{F}_m$	$(\mathcal{J}\&\mathcal{F})_m$
Supervised					
ImageNet	INet	ViT-S/8	63.9	68.1	66.0
STM [39]	I/D/Y	RN50	79.2	84.3	81.8
Self-supervised					
CT [58]	VLOG	RN50	46.4	50.0	48.7
MAST [33]	YT-VOS	RN18	63.3	67.6	65.5
STC [30]	Kinetics	RN18	64.8	70.2	67.6
DINO	INet	ViT-S/16	60.2	63.4	61.8
DINO	INet	ViT-B/16	60.7	63.9	62.3
DINO	INet	ViT-S/8	**66.6**	**73.1**	**69.9**
DINO	INet	ViT-B/8	**67.9**	**74.9**	**71.4**

Results: Transfer Learning



Results: Transfer Learning

	Cifar_{10}	Cifar_{100}	INat_{18}	INat_{19}	Flwrs	Cars	INet
ViT-S/16							
Sup. [56]	**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. [56]	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**

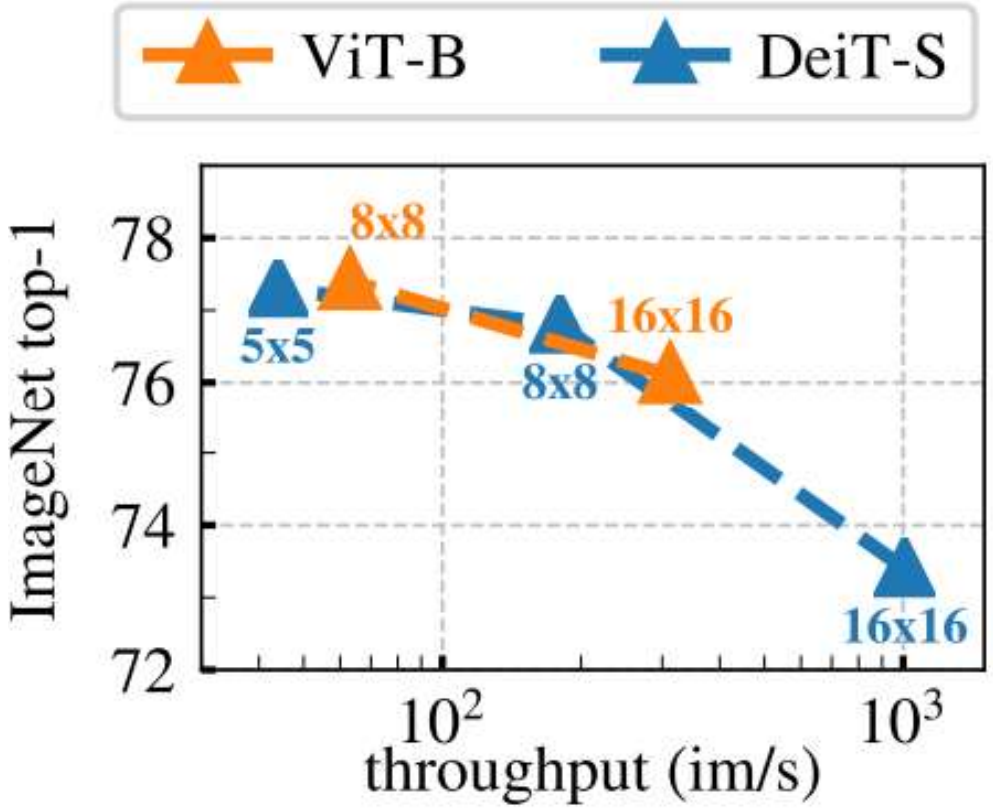
Ablation Study

Updating Teacher - Student Model

Using Global and Local Crops

Method	Mom.	MC	Loss	Pred.	k-NN	Lin.
DINO	✓	✓	CE	✗	72.8	76.1
	✗	✓	CE	✗	0.1	0.1
	✓	✗	CE	✗	67.9	72.5
	✓	✓	MSE	✗	52.6	62.4
	✓	✓	CE	✓	71.8	75.6
BYOL	✓	✗	MSE	✓	66.6	71.4
MoCov2	✓	✗	INCE	✗	62.0	71.6

Comparing Batch Sizes



References

James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. “Lost in quantization: Improving particular object retrieval in large scale image databases”. In CVPR, 2008.

Jerome Revaud, Jon Almazan, Rafael S Rezende, and Cesar ´ Roberto de Souza. “Learning with average precision: Training image retrieval with a listwise loss.” In ICCV, 2019.

Jean-Bastien Grill, Florian Strub, Florent Altche, Corentin ´ Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi ´ Munos, and Michal Valko. Bootstrap your own latent: A new approach to self-supervised learning. In NeurIPS, 2020