Non-contrastive Self-Supervised Learning

ZiXi Yan

2022/09/02

Catalog

要分享的论文列表

1. Emerging Properties in Self-Supervised Vision Transformers

* 1 Facebook AI Research 2 Inria∗ 3 Sorbonne University

2. Non-Contrastive Self-supervised Learning for Utterance-Level Information Extraction from Speech

* IEEE

3. C3-DINO: Joint Contrastive and Non-contrastive Self-Supervised Learning for Speaker Verification

* IEEE

self-distillation with no labels(DINO)

Emerging Properties in Self-Supervised Vision Transformers

1 Facebook AI Research 2 Inria∗ 3 Sorbonne University

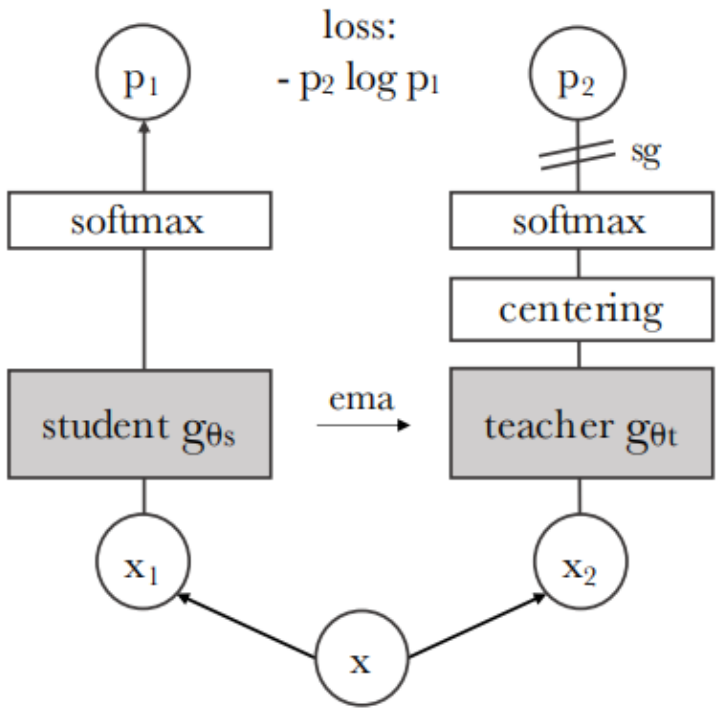


Figure 2: **Self-distillation with no labels.** We illustrate DINO in the case of one single pair of views (x_1, x_2) for simplicity. The model passes two different random transformations of an input image to the student and teacher networks. Both networks have the same architecture but different parameters. The output of the teacher network is centered with a mean computed over the batch. Each networks outputs a K dimensional feature that is normalized with a temperature softmax over the feature dimension. Their similarity is then measured with a cross-entropy loss. We apply a stop-gradient (sg) operator on the teacher to propagate gradients only through the student. The teacher parameters are updated with an exponential moving average (ema) of the student parameters.

Emerging Properties in Self-Supervised Vision Transformers
1 Facebook AI Research 2 Inria∗ 3 Sorbonne University

- Algorithm

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()
```

Emerging Properties in Self-Supervised Vision Transformers
1 Facebook AI Research 2 Inria∗ 3 Sorbonne University

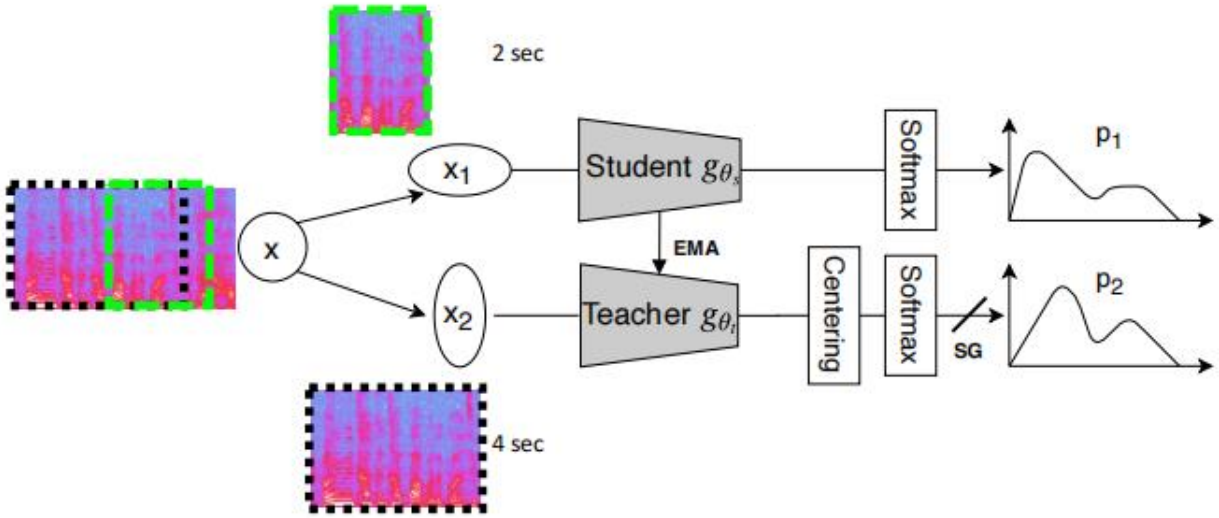
- Algorithm

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

Non-Contrastive Self-supervised Learning for Utterance-Level Information Extraction from Speech
* IEEE

- Motivation
  - we cannot be sure if all the negative samples belong to the classes different from the positive sample's class.
  - non-contrastive methods, do not require negative samples, so they are free from the issue above.



- Methods
  - DINO
    - adapted a non contrastive self-supervised learning technique, DINO, first proposed in computer vision, to the speech domain for the self-supervised pre-trained model

Non-Contrastive Self-supervised Learning for Utterance-Level Information Extraction from Speech
* IEEE

- Experiments

TABLE V
COMPARISON OF DINO TO THREE CONTRASTIVE LEARNING METHODS IN SV ON VOXCELEB1 *test*.

Method	EER(%)
DINO	**4.83**
AP+AAT [18]	8.65
MoCo (ProtoNCE) [19]	8.23
CSSL w/ $L_{ap} + L_{ch(mse)}$ [20]	8.28

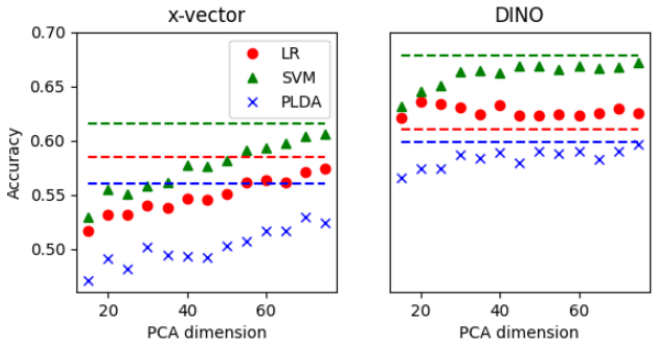


Fig. 3. Frozen SER. Dotted lines are accuracies when PCA is not used

TABLE VI
OVERVIEW OF FINE-TUNING EXPERIMENTS FOR SV. $FT2_{AAM}$ MEANS THAT THE AAM LOSS WAS USED IN THE FINE-TUNING WHILE *FT2* USES A CROSS ENTROPY LOSS. X-VECTOR (-) MEANS THAT THE PRE-TRAINING OF THE X-VECTOR SYSTEM DID NOT USE VOXCELEB2 *dev*

	Pretraining(-Finetuning)	labeled data (pre-training, fine-tuning)	back-end	labeled data (backend)	EER (%)	MinDCF(p=0.01)
1	DINO	none, none	CS	none	4.83	0.463
2			PLDA	VoxCeleb1 *dev*	2.38	0.289
3	DINO-*FT2*	none, VoxCeleb1 *dev*	CS	none	3.41	0.373
4			PLDA	VoxCeleb1 *dev*	2.20	0.278
5	DINO-$FT2_{AAM}$	none, VoxCeleb1 *dev*	CS	none	2.51	0.241
6			PLDA	VoxCeleb1 *dev*	2.01	0.255
7	x-vector (-)	VoxCeleb1 *dev*, none	CS	none	3.32	0.345
8			PLDA	VoxCeleb1 *dev*	2.75	0.290
9	x-vector	VoxCeleb2 *dev*, none	CS	none	2.18	0.205
10			PLDA	VoxCeleb1 *dev*	1.87	0.211
11	x-vector-*FT2*	VoxCeleb2 *dev*, VoxCeleb1 *dev*	CS	none	2.46	0.298
12			PLDA	VoxCeleb1 *dev*	1.98	0.215
13	x-vector-$FT2_{AAM}$	VoxCeleb2 *dev*, VoxCeleb1 *dev*	CS	none	2.21	0.215
14			PLDA	VoxCeleb1 *dev*	1.98	0.250

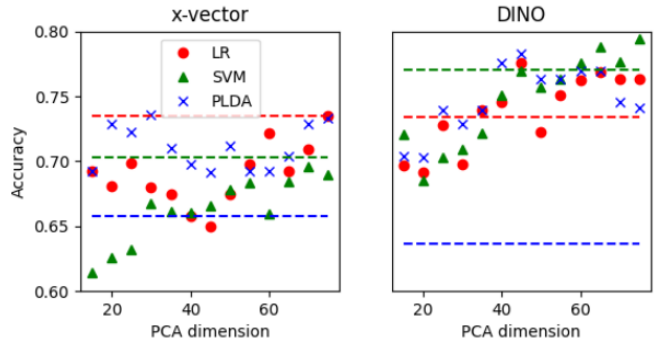


Fig. 7. Frozen AD detection. Dotted lines are accuracies when PCA is not used

C3-DINO: Joint Contrastive and Non-contrastive Self-Supervised Learning for Speaker Verification
* IEEE

- Methods
  - MoCo speaker embedding system
  - DINO speaker embedding system

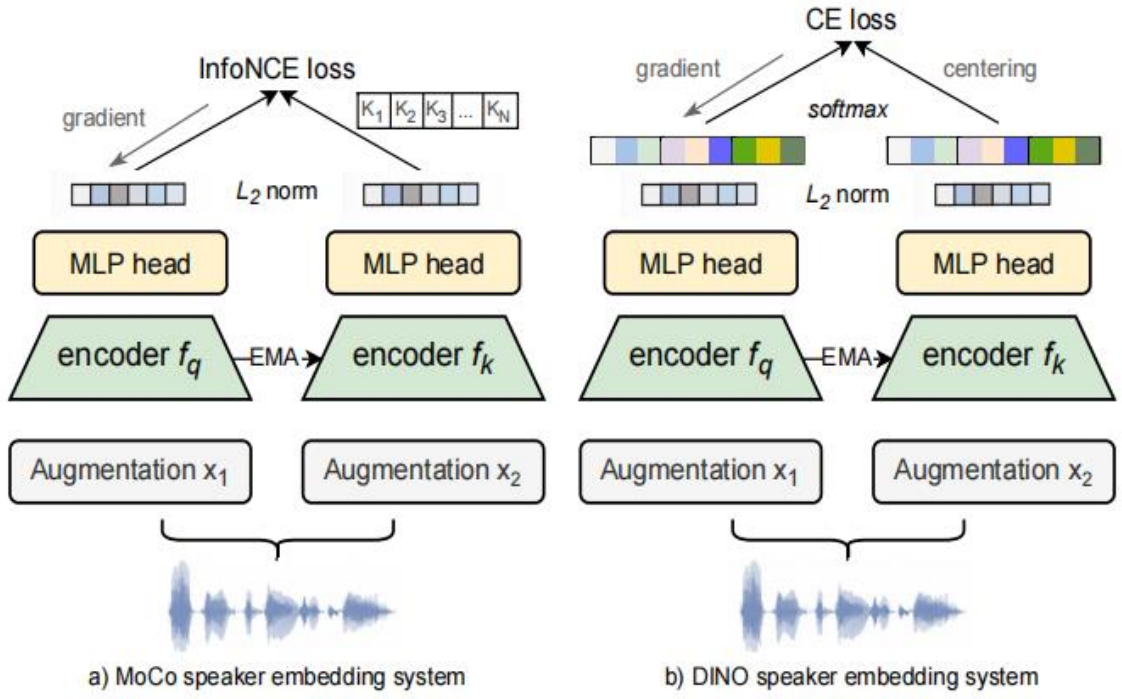


Fig. 1: The baseline MoCo and DINO training diagrams.

- Methods
  - MoCo speaker embedding system

$$\mathbf{v}_{qi} \cdot \mathbf{v}_{kj} > 0.8 \times \mathbf{v}_{qi} \cdot \mathbf{v}'_{ki}, j \in [1, K], \tag{7a}$$

$$\mathbf{v}_{qi} \cdot \mathbf{v}'_{ki} > 0.4, \tag{7b}$$

C3-DINO: Joint Contrastive and Non-contrastive Self-Supervised Learning for Speaker Verification
* IEEE

- Experiments

TABLE III: The SV performances of MoCo and DINO baseline systems w.r.t. three main stream architectures on Voxceleb1 test set. We use R34L and E-TDNN to represent ResNetSE34L and ECAPA-TDNN to save some space.

Training Data		EER(in%)			minDCF		
		TDNN	R34L	E-TDNN	TDNN	R34L	E-TDNN
Vox1	MoCo	11.3	10.8	**9.8**	0.76	0.72	**0.67**
Vox2	MoCo	8.5	8.3	**7.3**	0.63	0.62	**0.61**
Vox1	DINO	7.2	6.8	**6.1**	0.76	0.72	**0.52**
Vox2	DINO	5.6	4.6	**4.0**	0.51	0.50	**0.48**

TABLE V: System comparisons on Vox1 test set (EER in%).

supervision	method	Vox1 test
self-sup	i-vector [42]	15.3
self-sup	AP+AAT [42]	8.6
self-sup	MoCo+ProtoNCE [29]	8.2
self-sup	MoCo (E-TDNN) [52]	7.3
self-sup	Siaseme+SSR [39]	7.0
self-sup	C3-MoCo (system 3)	**6.4**
self-sup	DINO [44]	4.8
self-sup	C3-DINO$_2$ (system 8)	**2.5**
self-sup	C3-DINO$_2$ (system 8 + LDA)	**2.2**
semi-sup	GCL+PLDA (15% label) [43]	6.0
semi-sup	MoCo+SupCon (15% label) [29]	4.3
sup	x-vector [46]	3.1
sup	ECAPA-TDNN [11]	**0.9**

Thank