Cursos Extraordinarios

verano 2025

“Inteligencia Artificial y Grandes Modelos de Lenguaje: Funcionamiento, Componentes Clave y Aplicaciones”

Zaragoza, del 30 de junio al 02 de julio de 2025

Unsupervised learning

• Reconstruction methods

- Recent uses of reconstruction methods for unsupervised representation learning

• Decoar (Ling 2020)

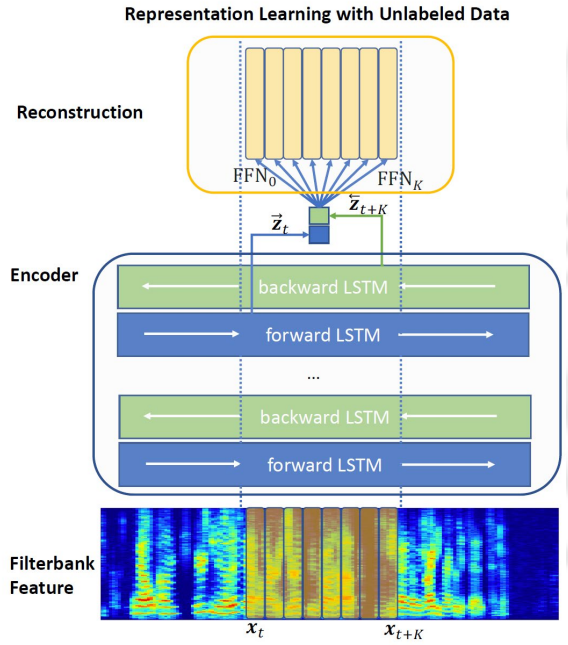
Ling, S., Liu, Y., Salazar, J., & Kirchhoff, K. (2020, May). Deep contextualized acoustic representations for semi-supervised speech recognition. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6429-6433). IEEE

- It also uses a L1 reconstruction to predict missing parts of spectrogram
- Bidirectional RNNs are used:
  - Direct and reverse representations are used to predict
  - **Forward** is used to predict $\boldsymbol{K}$ **frames** from: $\boldsymbol{x}_t$
  - **Backward** is used to predict $\boldsymbol{K}$ **frames** before $\boldsymbol{x}_t + K$
  - The loss is the sum of the prediction loss of future K frames

$$\mathcal{L}_t = \sum_{i=0}^{K} |\mathbf{x}_{t+i} - \mathrm{FFN}_i([\overrightarrow{\mathbf{z}}_t; \overleftarrow{\mathbf{z}}_{t+K}])|$$

  - A small network (different) is used to predict each future frame
    - (same embedding)

$$\mathrm{FFN}_i(\mathbf{v}) = \mathbf{W}_{i,2}\mathrm{ReLU}(\mathbf{W}_{i,1}\mathbf{v} + \mathbf{b}_{i,1}) + \mathbf{b}_{i,2}$$



Unsupervised learning

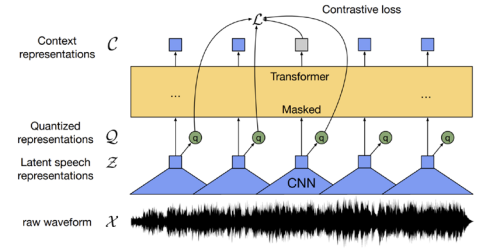
• Prediction of clusters/vq centroids

- The signal is assigned to discrete variables: cluster ids, centroids

• **Wav2vec2.0** (Ling **2020**)

Baevski, A., Zhou, H., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. arXiv preprint arXiv:2006.11477.

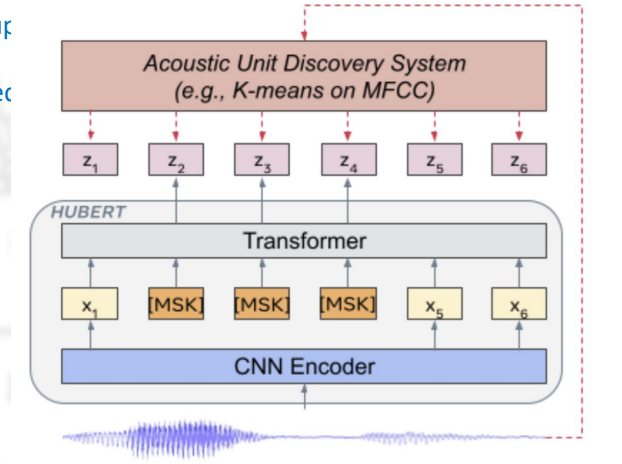
- The centroids of a quantification process are the objective (discrete)
- The training objective requires identifying:
  - the **correct quantized latent audio representation** in a set of distractors
- Improve ASR



• **Hubert** (Hsu **2021**), **Wavlm** (Hsu **2021**)

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., & Mohamed, A. (2021). HuBERT: Self-Sup by Masked Prediction of Hidden Units. arXiv preprint arXiv:2106.07447.
Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., ... & Wei, F. (2022). Wavlm: Large-scale self-supervised processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.

- Iterative clustering -> index of clusters
- Prediction with large transformed (masked), similar to BERT style
  - First iteration: kmeans of standard features, Mel cepstrum
  - Folloing iterations: kmeans of last iteration embeddings
- Improvements on many tasks: ASR, speaker id, emotion recognition
- All these models can be downloaded



Unsupervised learning

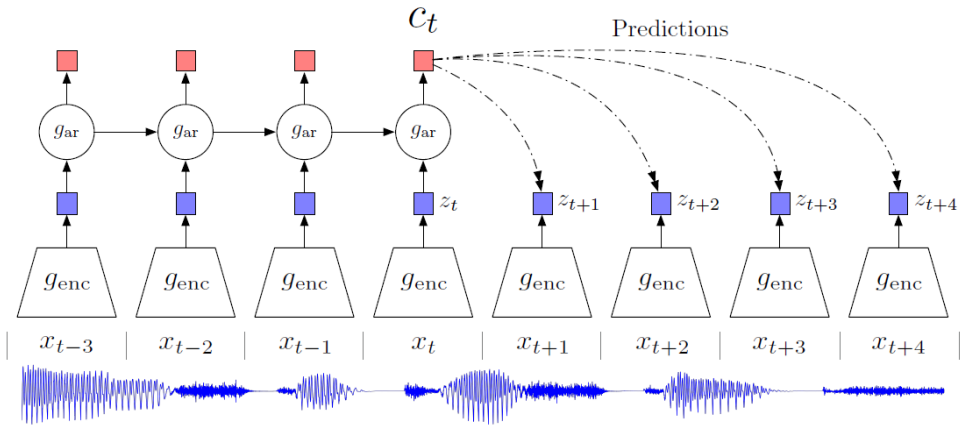
- **Contrastive loss**
- **CPC** (Oord **2018**)

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

- **Wav2vec** (Schneider **2019**)

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862.

- Two stages are combined to learn a feature extractor, input **raw samples** at 16kHz
  - CNN: convolutional network designed to process raw simples
    - The strides are: 5, 4, 2, 2, 2 (product = 160 global downsampling size for 16kHz)
    - The kernel sizes: 10, 8, 4, 4, 4 to approximate 30ms of equivalent reception field
  - RNN: a GRU is used to generate a context embedding that will be used to make predictions



Unsupervised learning

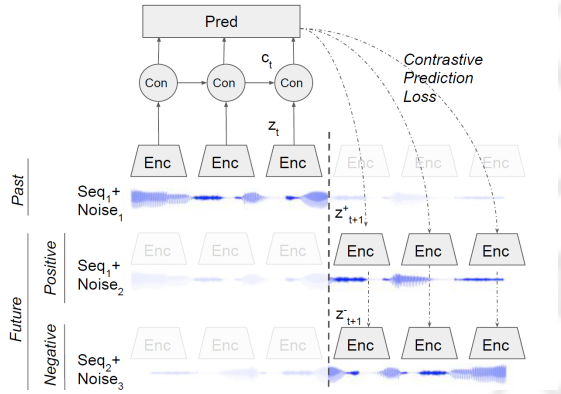
- **Contrastive loss**
- **CPC** (Oord **2018**)

Oord, A. V. D., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv preprint arXiv:1807.03748

- **Wav2vec** (Schneider **2019**)

Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. arXiv preprint arXiv:1904.05862.

- Contrastive loss, learns to select ground truth continuation of the audio, among many alternatives (distractors)
  - In theory any audio segment, for efficiency in the same minibatch
    - number of distractors-> batch_size (limited in GPU memory)
  - Augmentations to make the task more difficult: positive match original audio vs. augmented audio



- Pretext task: **correct continuation**
  - Other pretext tasks, from image or text: same file? Correct order of fragments ?

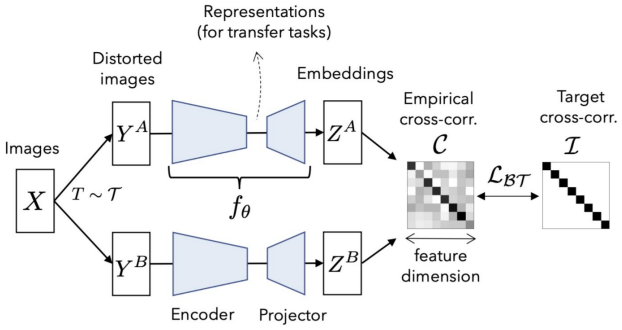
Unsupervised learning

- **Other methods**

- **Barlow twins** (Zbontar **2021**)

Zbontar, J., Jing, L., Misra, I., LeCun, Y., & Deny, S. (2021). Barlow twins: Self-supervised learning via redundancy reduction. arXiv preprint arXiv:2103.03230.SSP) (pp. 6429-6433).

Norm in the batch dimension -> stability



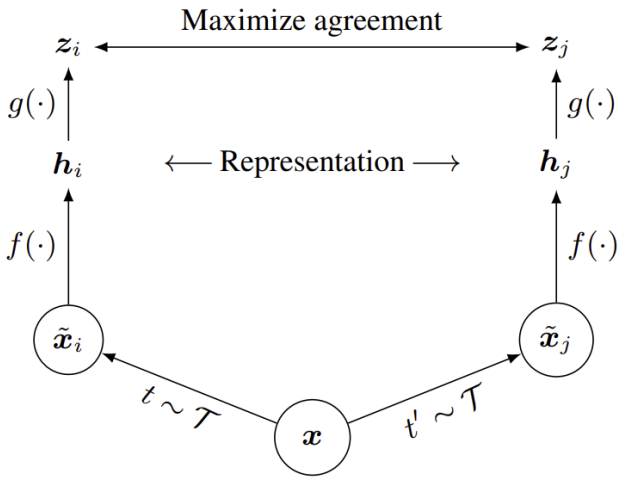
$$\mathcal{L}_{\mathcal{BT}} \triangleq \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} \mathcal{C}_{ij}^2}_{\text{redundancy reduction term}}$$

```
# f: encoder network
# lambda: weight on the off-diagonal terms
# N: batch size
# D: dimensionality of the embeddings
#
# mm: matrix-matrix multiplication
# off_diagonal: off-diagonal elements of a matrix
# eye: identity matrix

for x in loader: # load a batch with N samples
    # two randomly augmented versions of x
    y_a, y_b = augment(x)

    # compute embeddings
    z_a = f(y_a) # NxD
    z_b = f(y_b) # NxD

    # normalize repr. along the batch dimension
    z_a_norm = (z_a - z_a.mean(0)) / z_a.std(0) # NxD
    z_b_norm = (z_b - z_b.mean(0)) / z_b.std(0) # NxD

    # cross-correlation matrix
    c = mm(z_a_norm.T, z_b_norm) / N # DxD

    # loss
    c_diff = (c - eye(D)).pow(2) # DxD
    # multiply off-diagonal elems of c_diff by lambda
    off_diagonal(c_diff).mul_(lambda)
    loss = c_diff.sum()

    # optimization step
    loss.backward()
    optimizer.step()
```

Unsupervised learning

- **Other methods**
- Compare pairs of examples and update a loss function to improve generalization of common representation
- **SimCLR**(Chen **2020**)

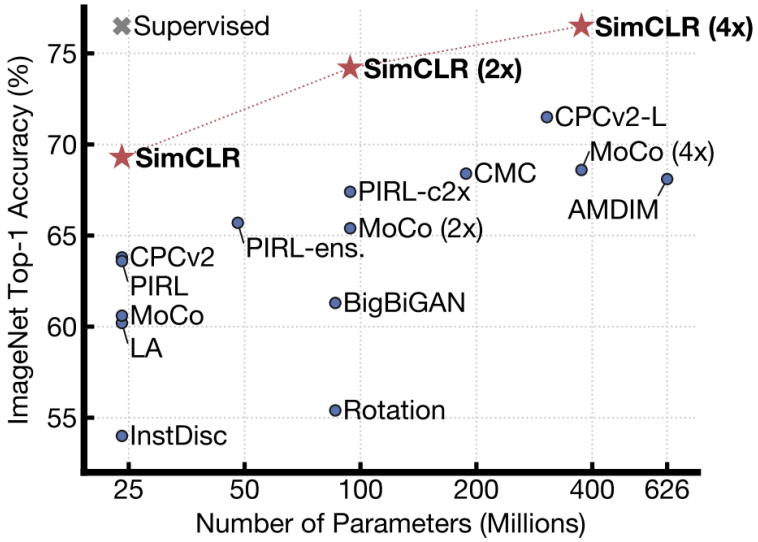
Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." International conference on machine learning. PMLR, 2020..



$f(\boldsymbol{x}_i)$: resnet $g(\boldsymbol{h}_i)$: MLP

$$\boldsymbol{z}_i = g(\boldsymbol{h}_i) = W^{(2)}\sigma(W^{(1)}\boldsymbol{h}_i)$$

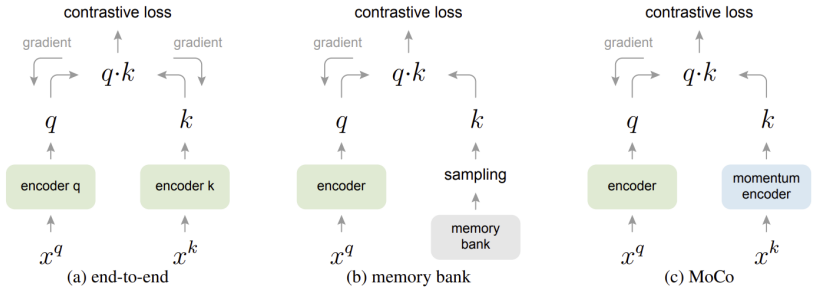
(Linear classifier)



Unsupervised learning

- **Other methods**
- **MoCo**(Grill **2020**)

He, K., Fan, H., Wu, Y., Xie, S., & Girshick, R. (2020). Momentum contrast for unsupervised visual representation learning. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 9729-9738).

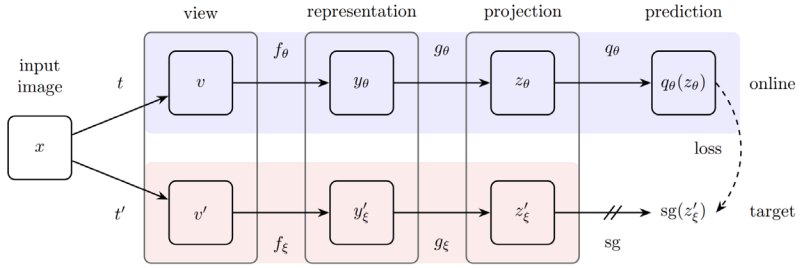


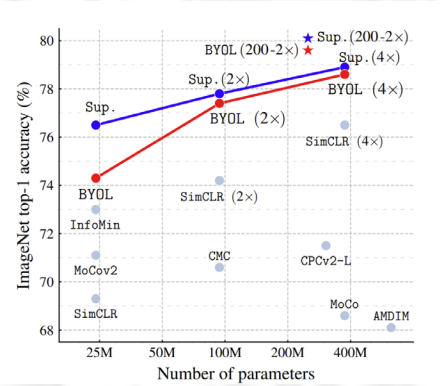
- A network updated using EMA (exponential moving average)
Provides stability to the training

$$\theta_k \leftarrow m\theta_k + (1-m)\theta_q$$

(**BYOL**(Grill **2020**)

... & Valko, M



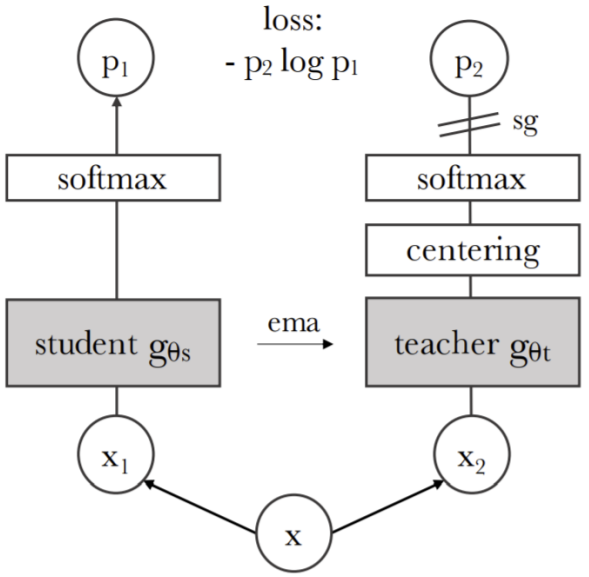


new approach to

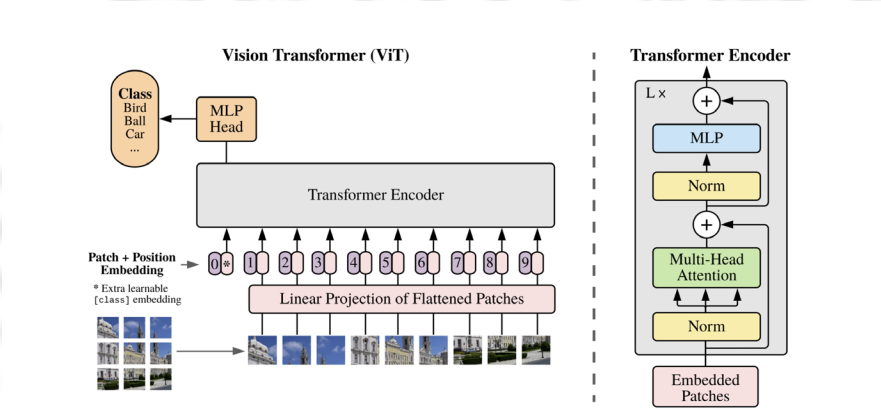
Unsupervised learning

- **Other methods**
- Compare pairs of examples and update a loss function to improve generalization of common representation
- **DINO**(Caron **2021**)

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..



- Two networks teacher student:
  - T(x), S(x)
- ViT architecture, class token

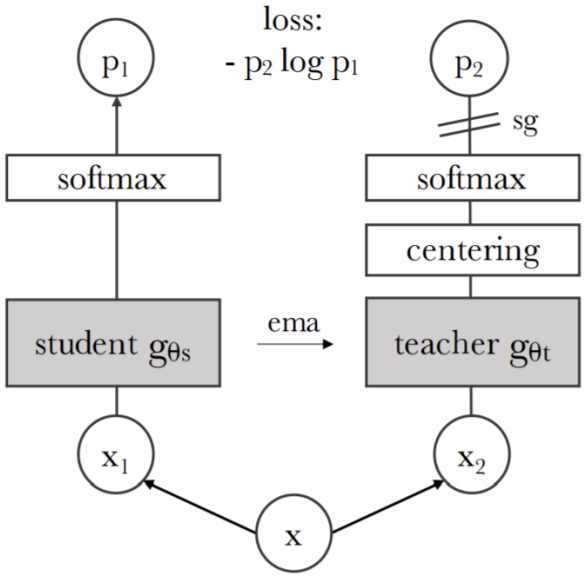


Unsupervised learning

Other methods

- Compare pairs of examples and update a loss function to improve generalization of common representation
- **DINO**(Caron **2021**)

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..



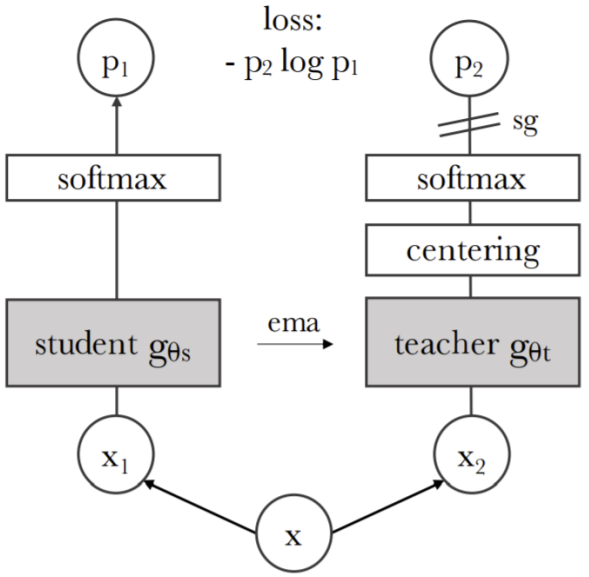
- Two networks T(x), S(x)
  - same arch. (ViT) diff. params
- Two augmentations: x1, x2

- The output of the teacher network T(x2)
- is centered with a mean computed over the batch.
- Each networks outputs a K dim
  - **softmax** (discrete distrib)
- Loss: cross-entropy loss, measure similarity distrib
- The teacher is not updated: stop-gradient (sg)
- The teacher exponential moving average (ema) of the student net

Unsupervised learning

- **DINO**(Caron **2021**)

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..



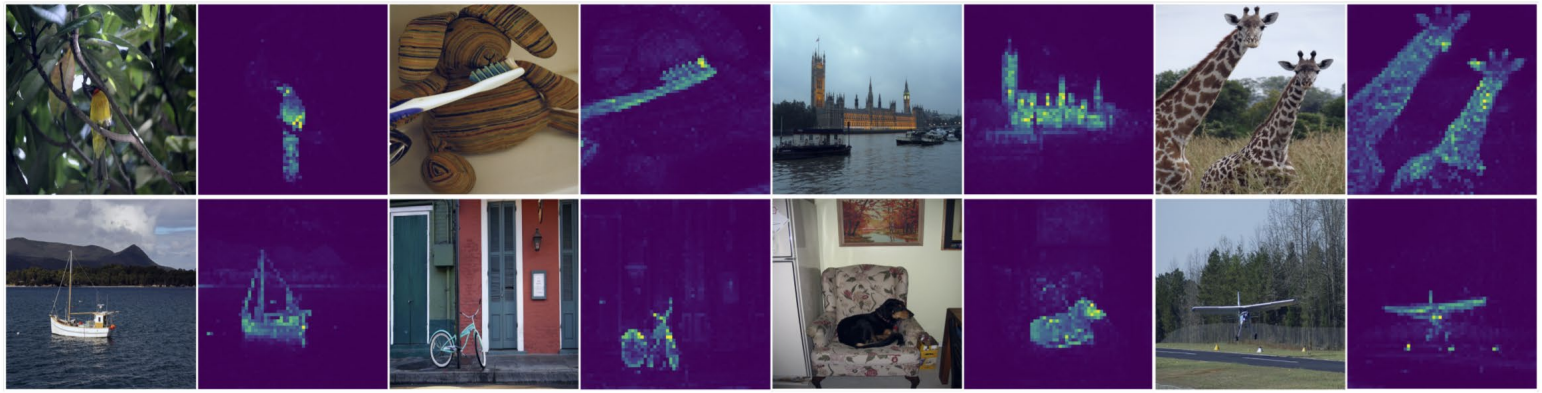
Algorithm 1 DINO PyTorch pseudocode w/o multi-crop.

```
# 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()
```

Unsupervised learning

- **Other methods**
- Compare pairs of examples and update a loss function to improve generalization of common representation
- **DINO**(Caron **2021**)

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



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

Unsupervised learning

- **DINO2**(Oquab **2023**)

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193...

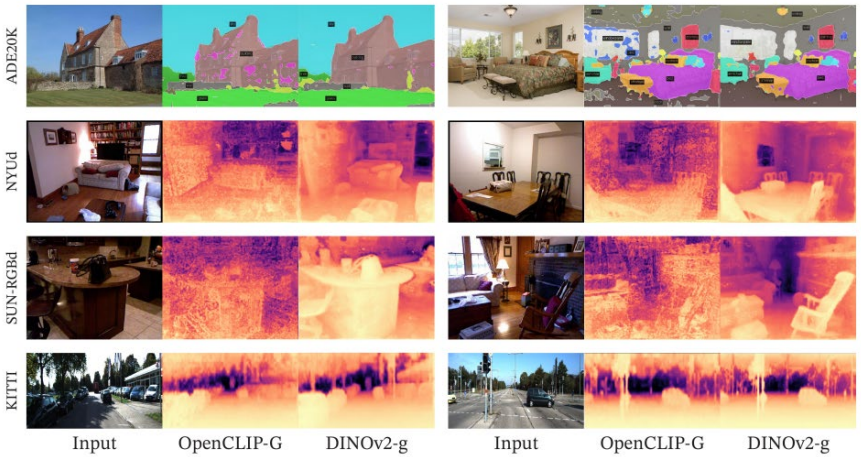
Method	Arch.	Data	Text sup.	kNN val	linear val	linear ReaL	linear V2
Weakly supervised							
CLIP	ViT-L/14	WIT-400M	✓	79.8	84.3	88.1	75.3
CLIP	ViT-L/14_{336}	WIT-400M	✓	80.5	85.3	88.8	75.8
SWAG	ViT-H/14	IG3.6B	✓	82.6	85.7	88.7	77.6
OpenCLIP	ViT-H/14	LAION	✓	81.7	84.4	88.4	75.5
OpenCLIP	ViT-G/14	LAION	✓	83.2	86.2	89.4	77.2
EVA-CLIP	ViT-g/14	custom*	✓	**83.5**	86.4	89.3	77.4
Self-supervised							
MAE	ViT-H/14	INet-1k	×	49.4	76.6	83.3	64.8
DINO	ViT-S/8	INet-1k	×	78.6	79.2	85.5	68.2
SEERv2	RG10B	IG2B	×	–	79.8	–	–
MSN	ViT-L/7	INet-1k	×	79.2	80.7	86.0	69.7
EsViT	Swin-B/W=14	INet-1k	×	79.4	81.3	87.0	70.4
Mugs	ViT-L/16	INet-1k	×	80.2	82.1	86.9	70.8
iBOT	ViT-L/16	INet-22k	×	72.9	82.3	87.5	72.4
DINOv2	ViT-S/14	LVD-142M	×	79.0	81.1	86.6	70.9
	ViT-B/14	LVD-142M	×	82.1	84.5	88.3	75.1
	ViT-L/14	LVD-142M	×	**83.5**	86.3	89.5	78.0
	ViT-g/14	LVD-142M	×	**83.5**	**86.5**	**89.6**	**78.4**

Unsupervised learning

- **DINO2**(Oquab **2023**)

Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193...

Linear layer trained from features



Out of distribution examples

Unsupervised learning

- **JEPA, V-JEPA, V-JEPA2**()

Assran, M., Bardes, A., Fan, D., Garrido, Q., Howes, R., Muckley, M., ... & Ballas, N. (2025). V-JEPA 2: Self [illegible] Understanding, Prediction and Planning. arXiv preprint arXiv:2506.09985.

Model predicts missing information from representations