Clustering-style Self-Supervised Learning

Mathilde Caron - FAIR Paris & Inria Grenoble

June 20th, 2021

CVPR 2021 Tutorial:

Leave Those Nets Alone:

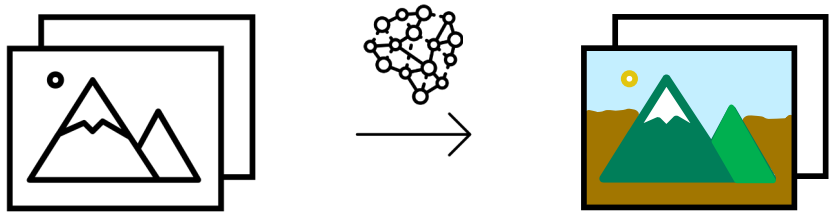
Advances in Self-Supervised Learning

FACEBOOK AI

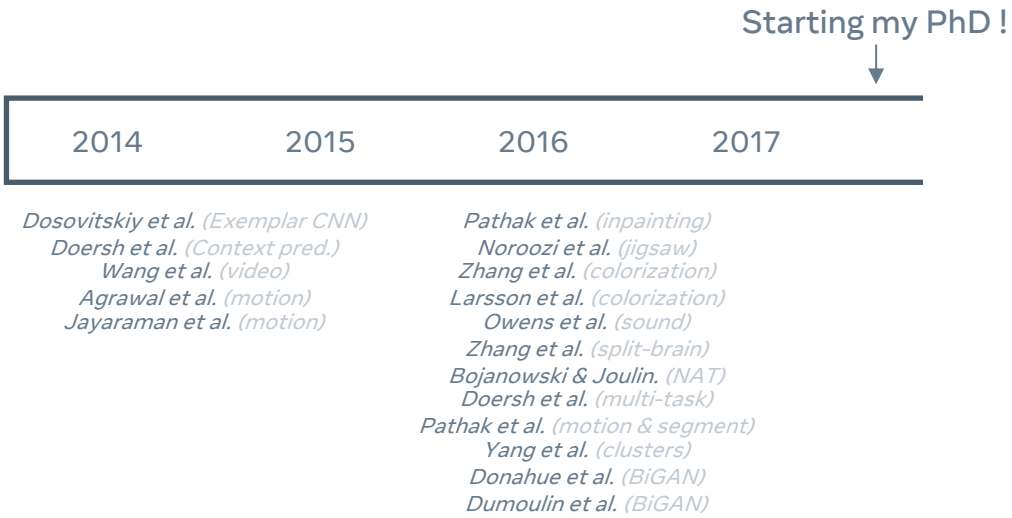
Self-Supervised Learning (SSL)

Designing a learning task that does not rely on human annotations

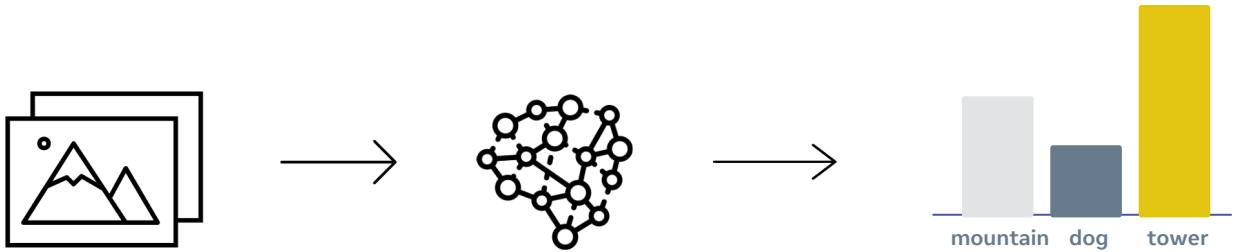
Example: Colorization (Zhang et al. | 2016)



Designing SSL tasks is an active research area



Supervised pre-training: labels classification

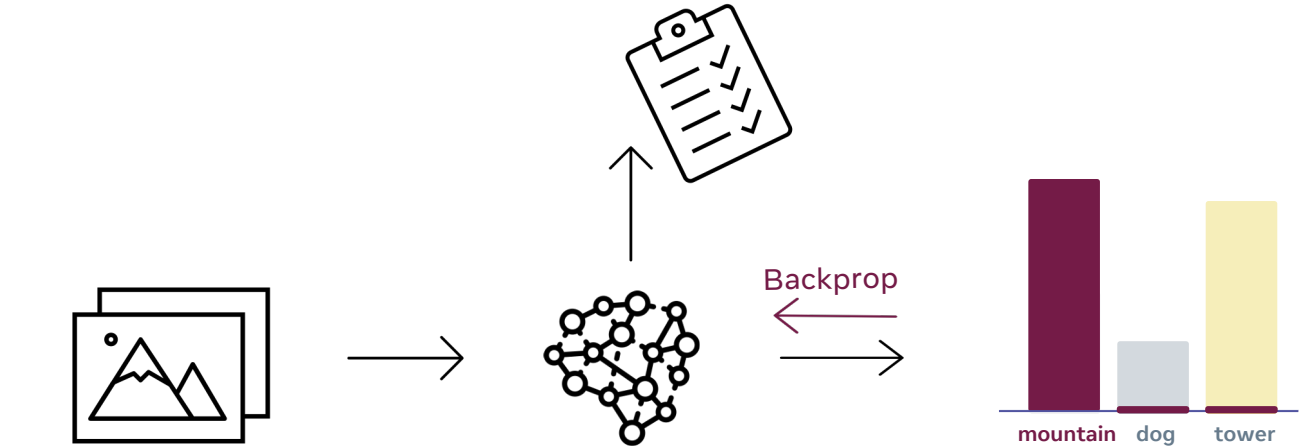


Training images + labels

Neural network

Classification

Supervised pre-training: labels classification

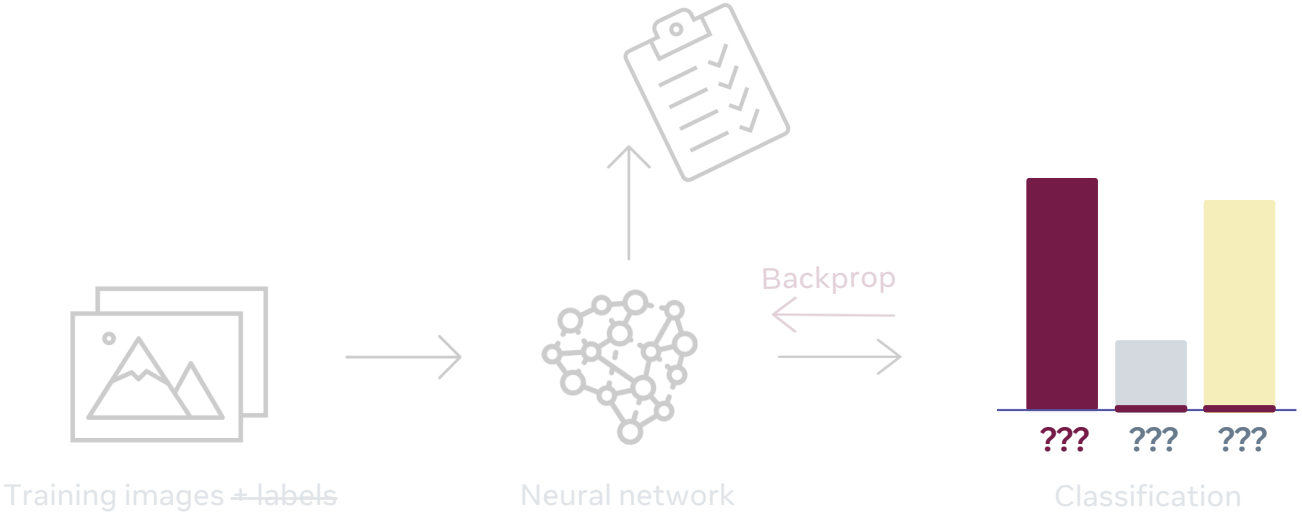


Training images + labels

Neural network

Classification

We do not have labels !



Can we replace labels with clustering ?

DeepCluster: Deep Clustering for Unsupervised Learning of Visual Features

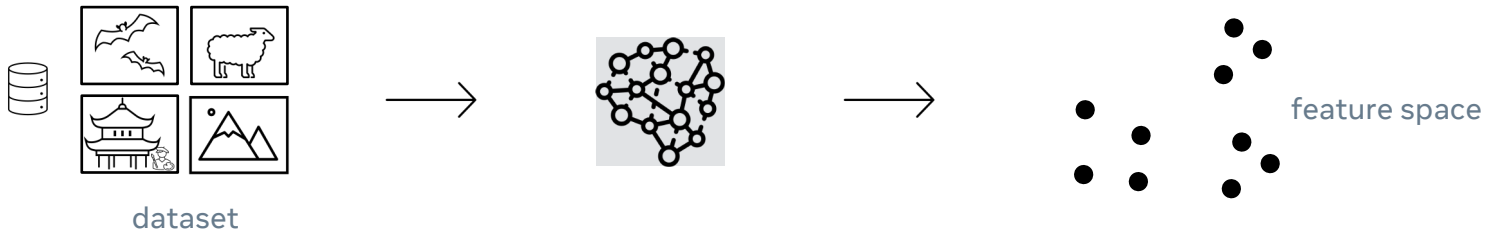
Mathilde Caron, Piotr Bojanowski, Armand Joulin, Matthijs Douze

ECCV 2018

github.com/facebookresearch/deepcluster

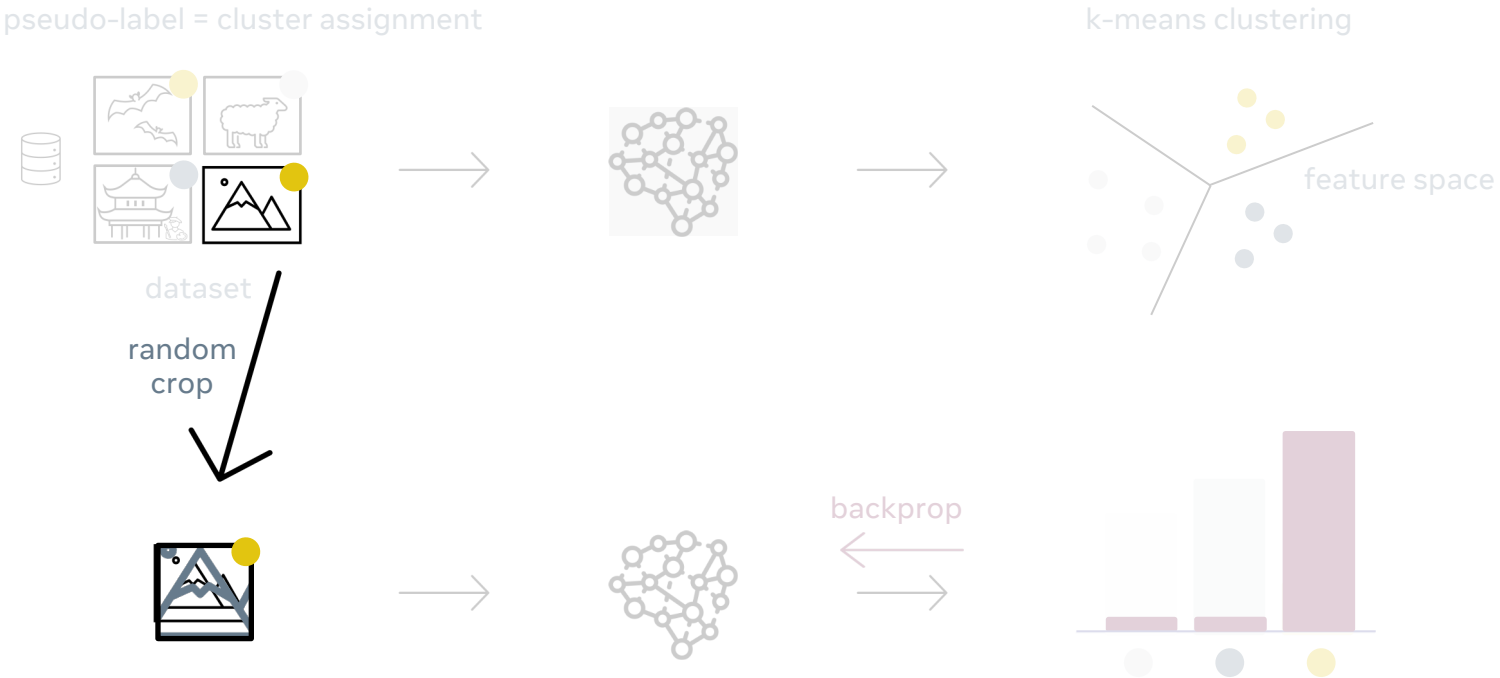
FACEBOOK AI

DeepCluster



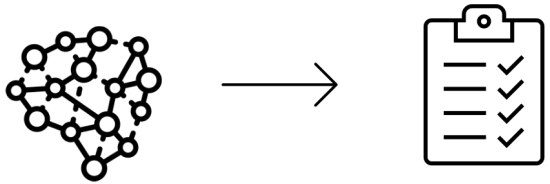
DeepCluster

Invariance to cropping



How to Evaluate Self-Supervised Learning ?

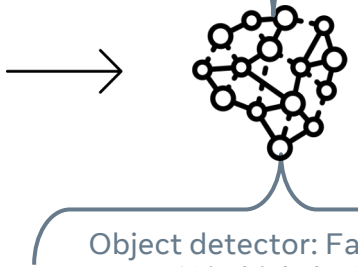
Use learned representations for downstream tasks

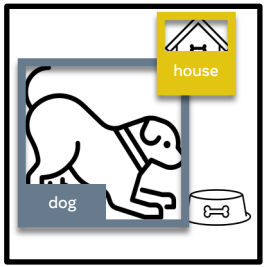


How to Evaluate Self-Supervised Learning ?

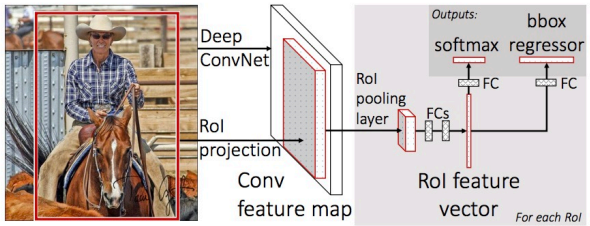
Example: Object detection on Pascal VOC07 dataset

- ❑ Random
- ❑ Supervised
- ❑ Self-Supervised
  - ❑ DeepCluster

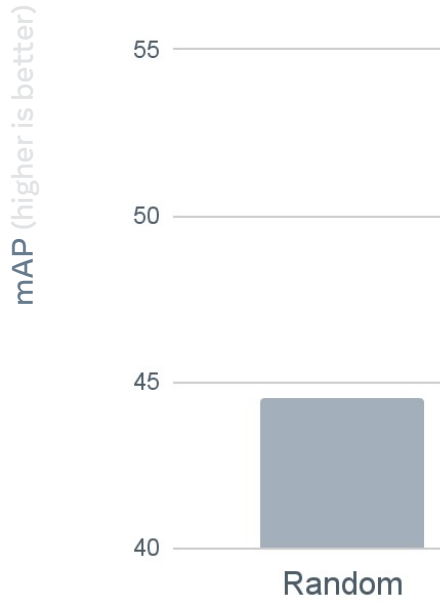




Object detector: Fast R-CNN
(Girshick. | 2015)



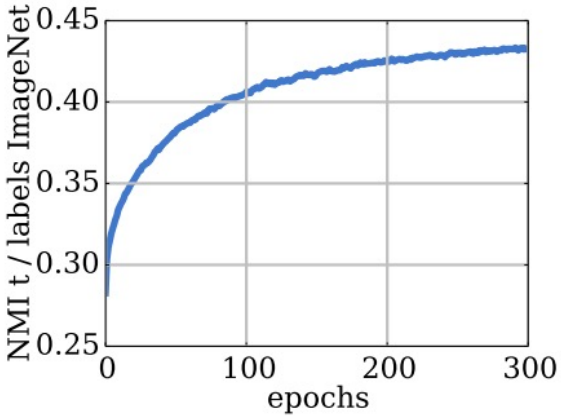
Results on Object Detection on Pascal VOC07



DeepCluster also produces… clusters!

Clustering visualization

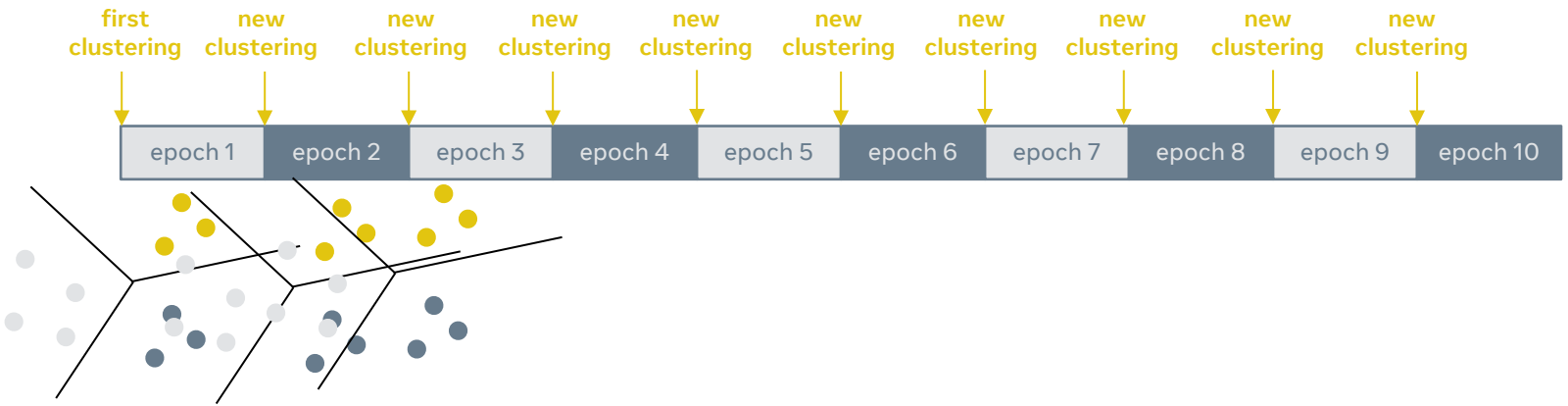
Clustering evaluation



Limitations of DeepCluster

- Does not scale (depends on the dataset size)

The clusters (i.e. pseudo-labels) are refined during training



Limitations of DeepCluster

- Does not scale (depends on the dataset size)

Huge dataset: we can afford only 2 epochs!

Problem: clusters are refined only once…

Limitations of DeepCluster

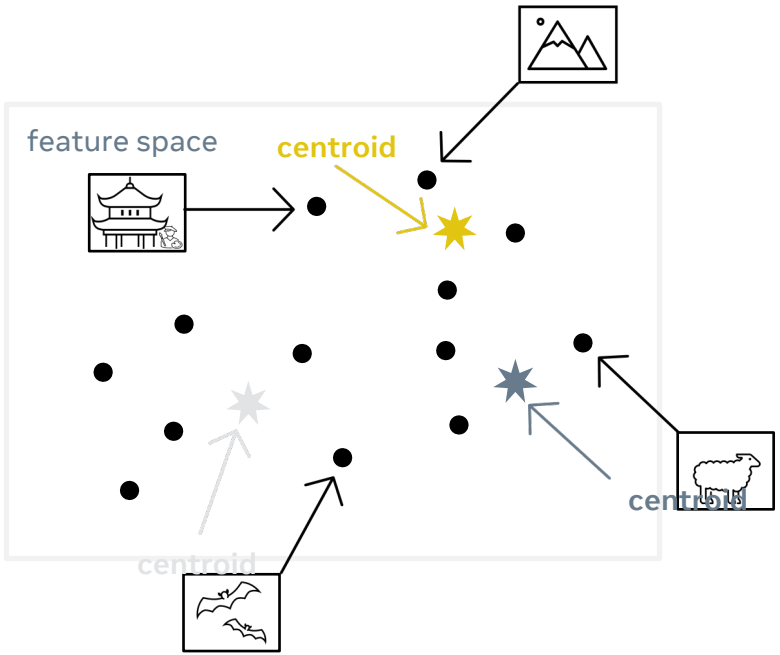
- Does not scale (depends on the dataset size)

Even bigger dataset: we never see an image twice

Problem: the clusters are never refined!

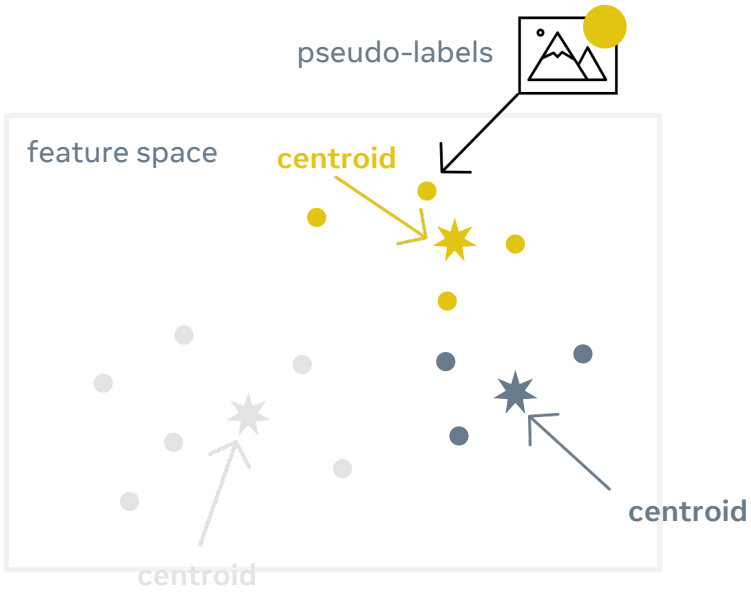
Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?



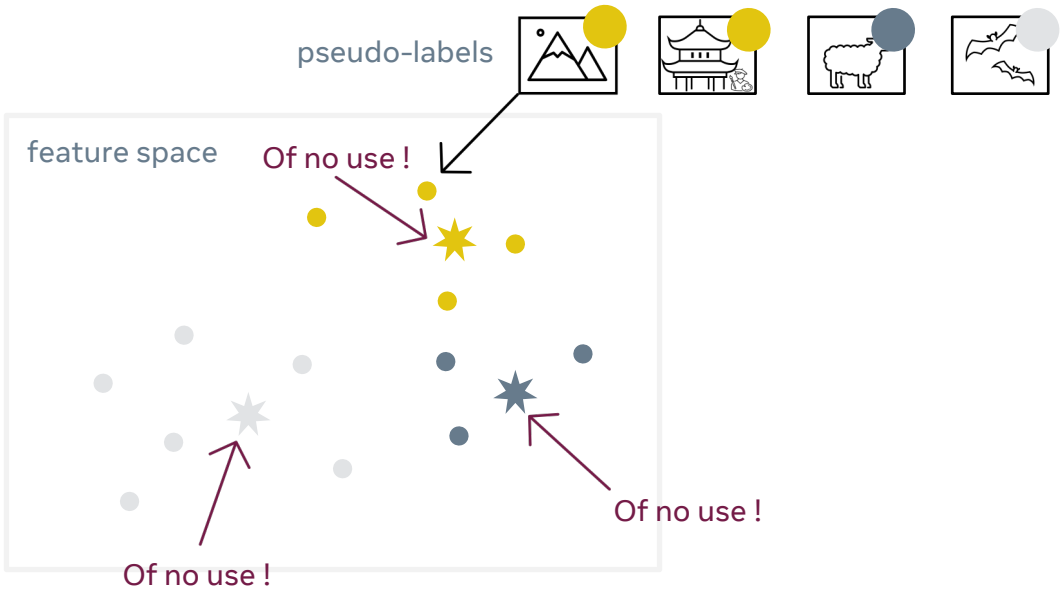
Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?



Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?



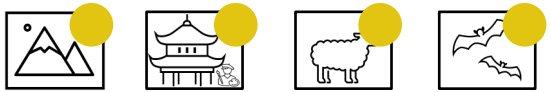
Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse

Collapse of feature space:

Limitations of DeepCluster

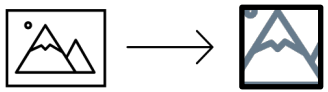
- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse



Collapse of feature space:

Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse
- Importance of random cropping is only implicit



How to overcome these limitations ?

SwAV: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments

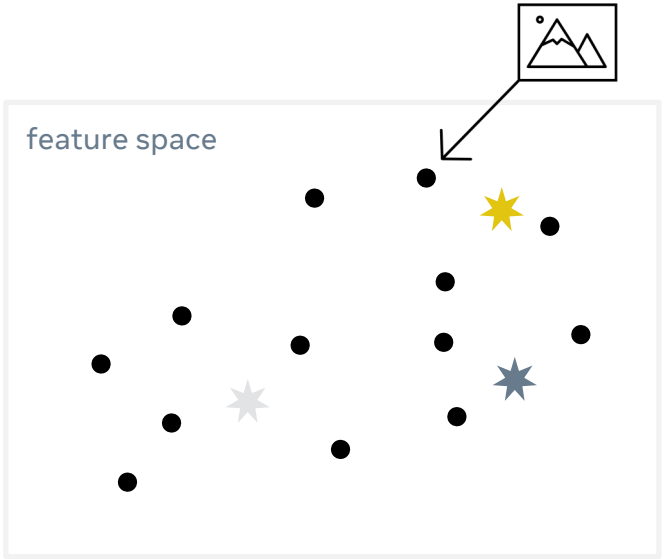
Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin

NeurIPS 2020

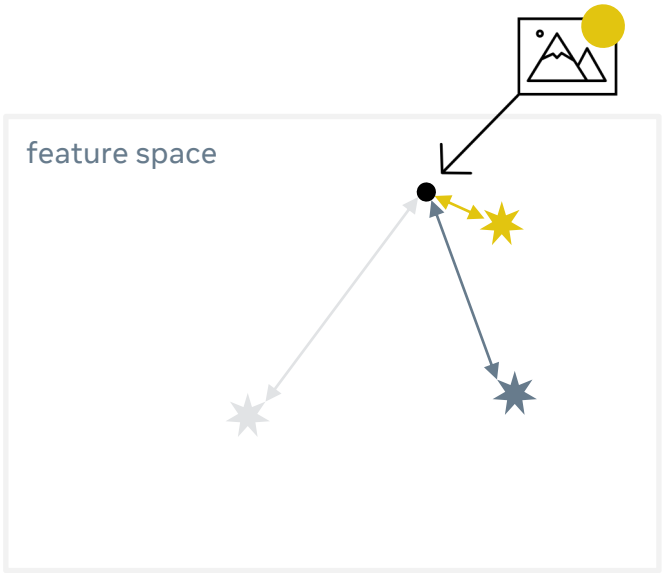
github.com/facebookresearch/swav

FACEBOOK AI

Pseudo-labels in SwAV



Pseudo-labels in SwAV

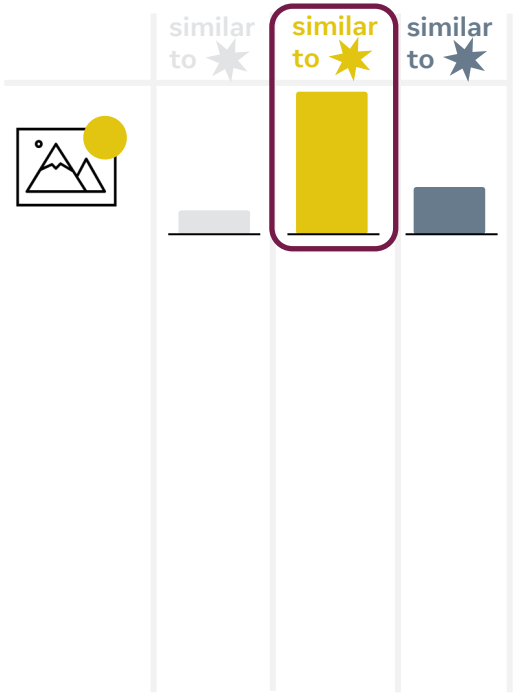


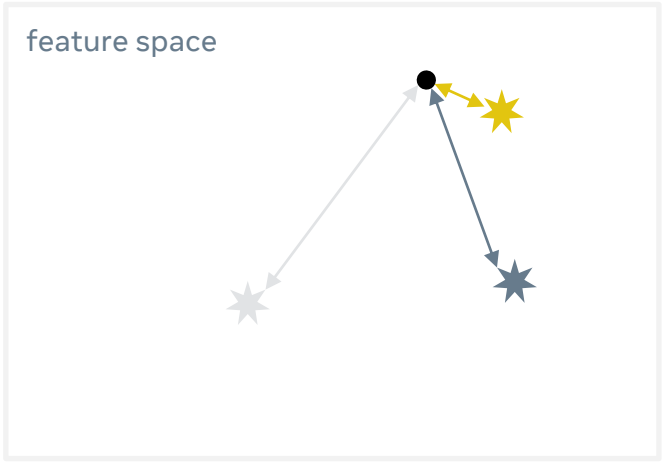
Pseudo-labels in SwAV

We can directly use the neural network to output scores !

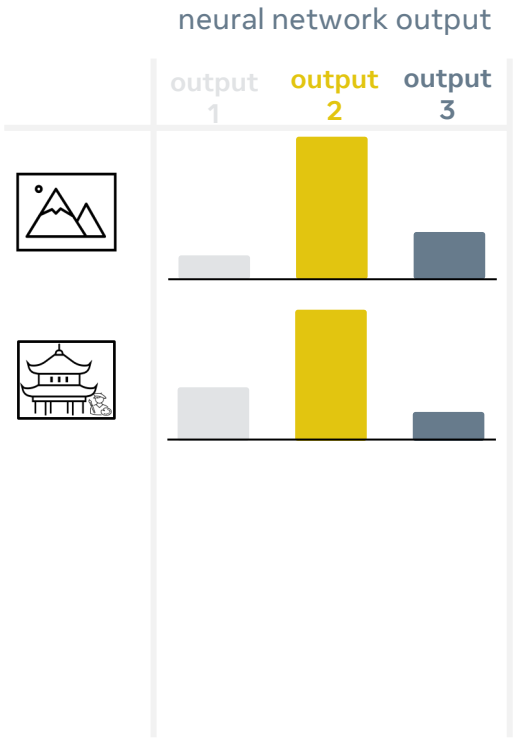
All we need is a code for each cluster !

Most similar





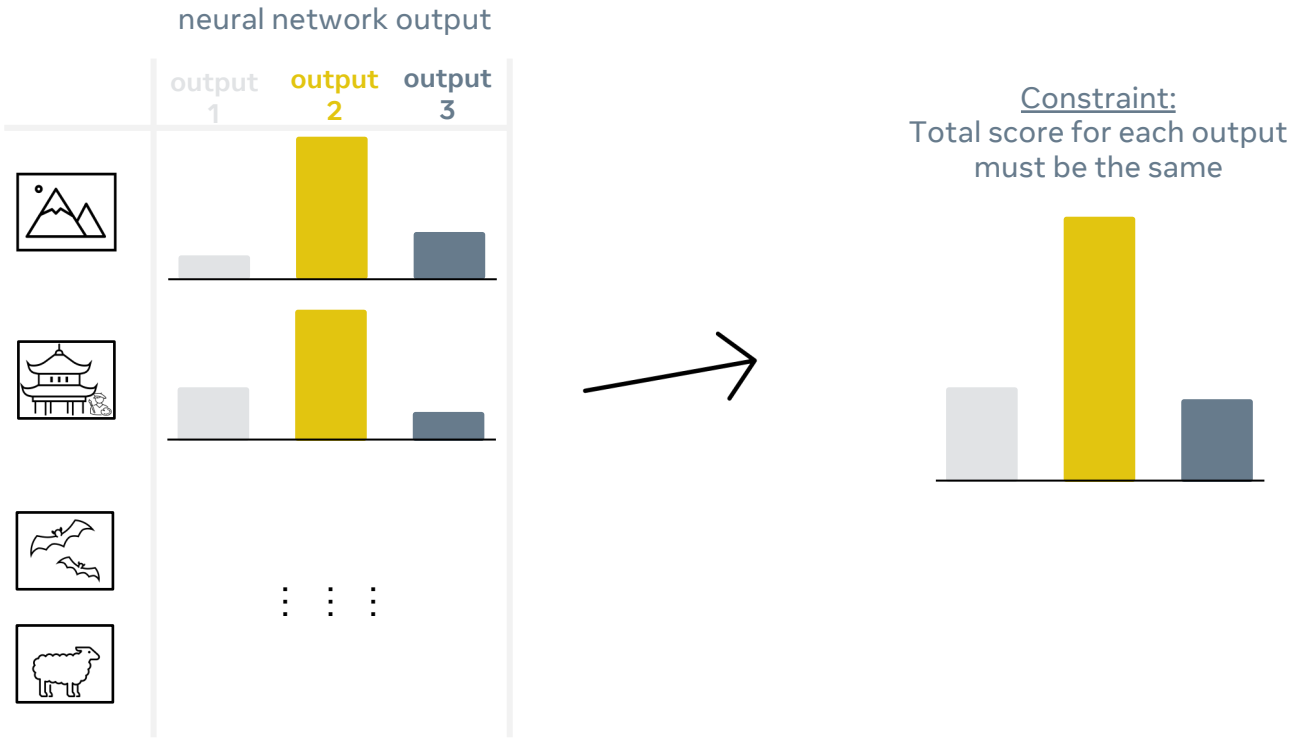
Pseudo-labels in SwAV



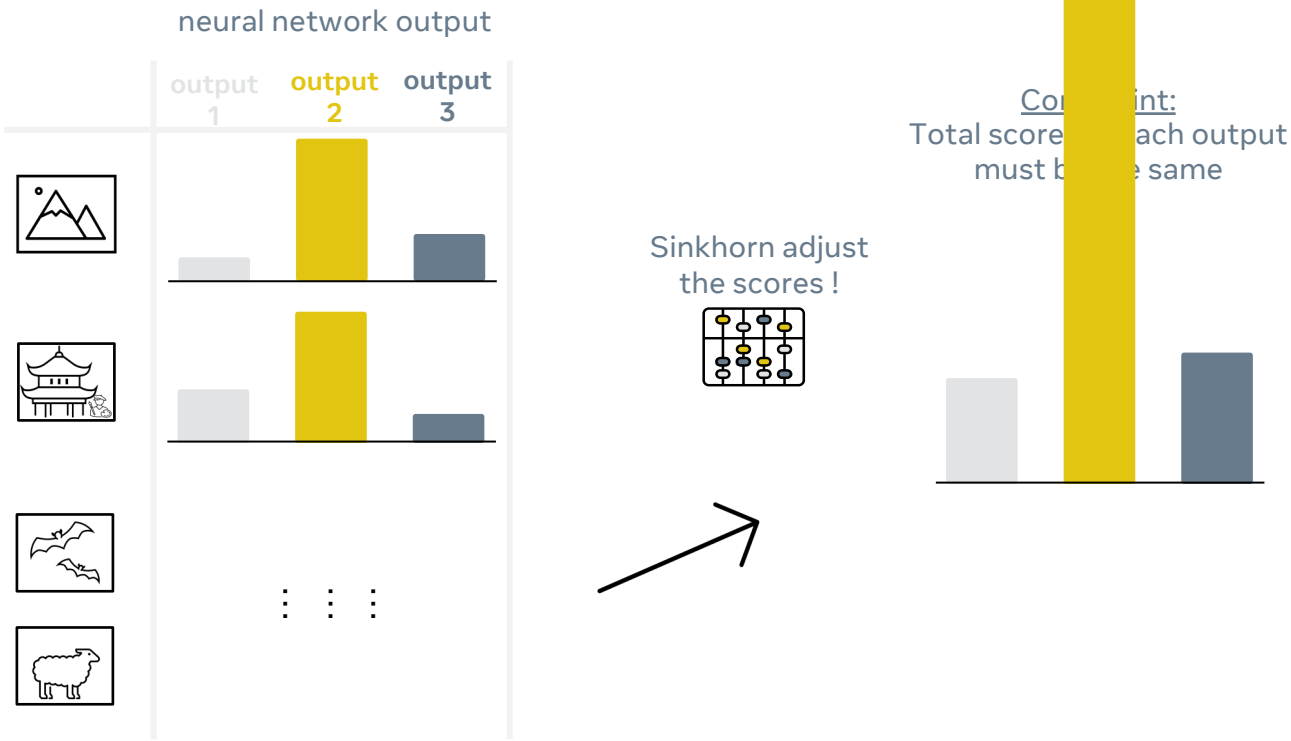
SELA - Asano et al. ICLR 2020
UIC – Chen et al. ECCV 2020

Constraint:
Total score for each output must be the same

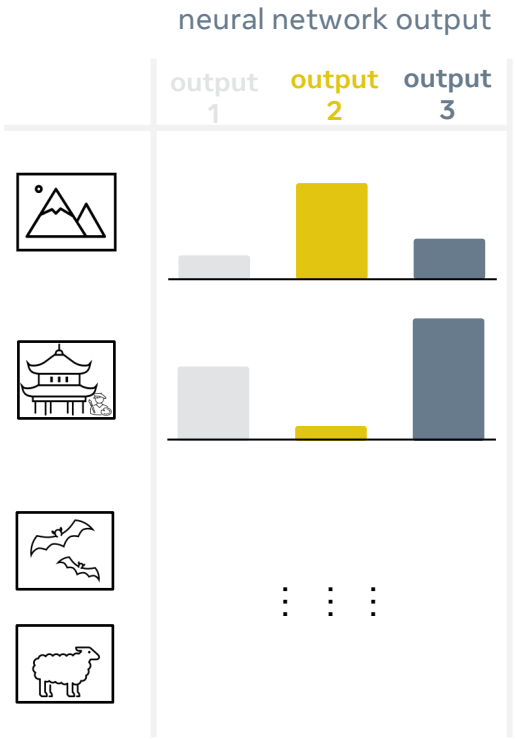
Pseudo-labels in SwAV

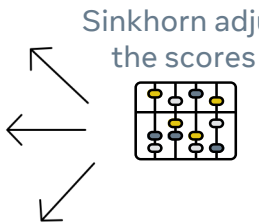


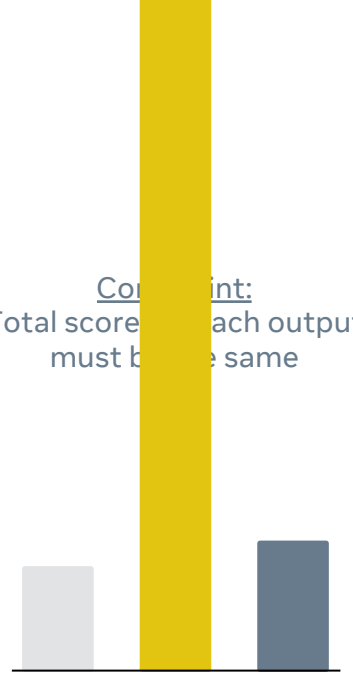
Pseudo-labels in SwAV



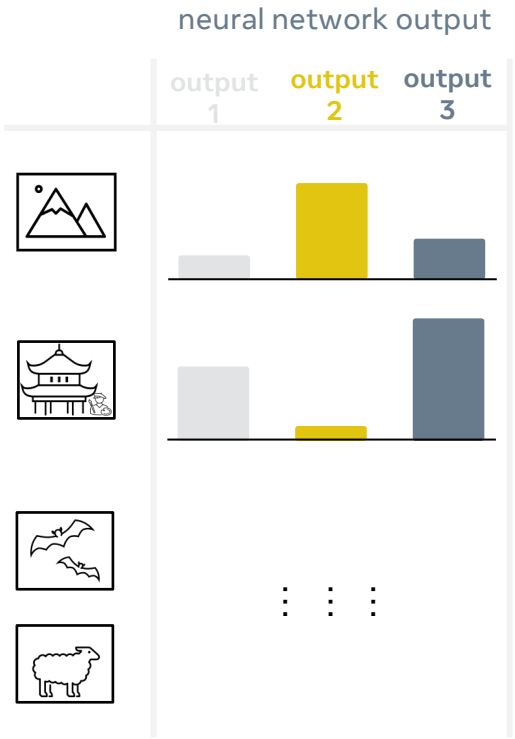
Pseudo-labels in SwAV



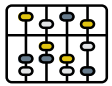




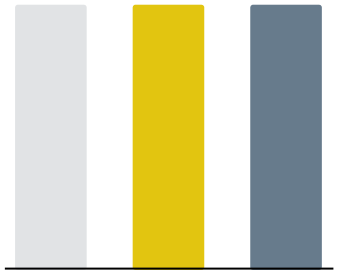
Pseudo-labels in SwAV



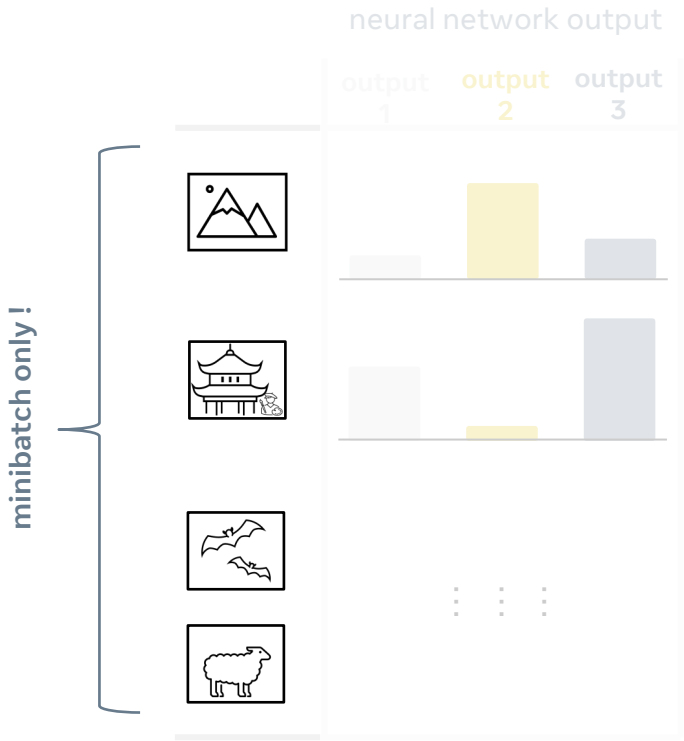
Sinkhorn adjust the scores !



Constraint:
Total score for each output must be the same

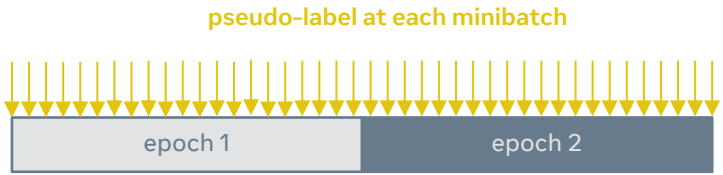


Pseudo-labels in SwAV

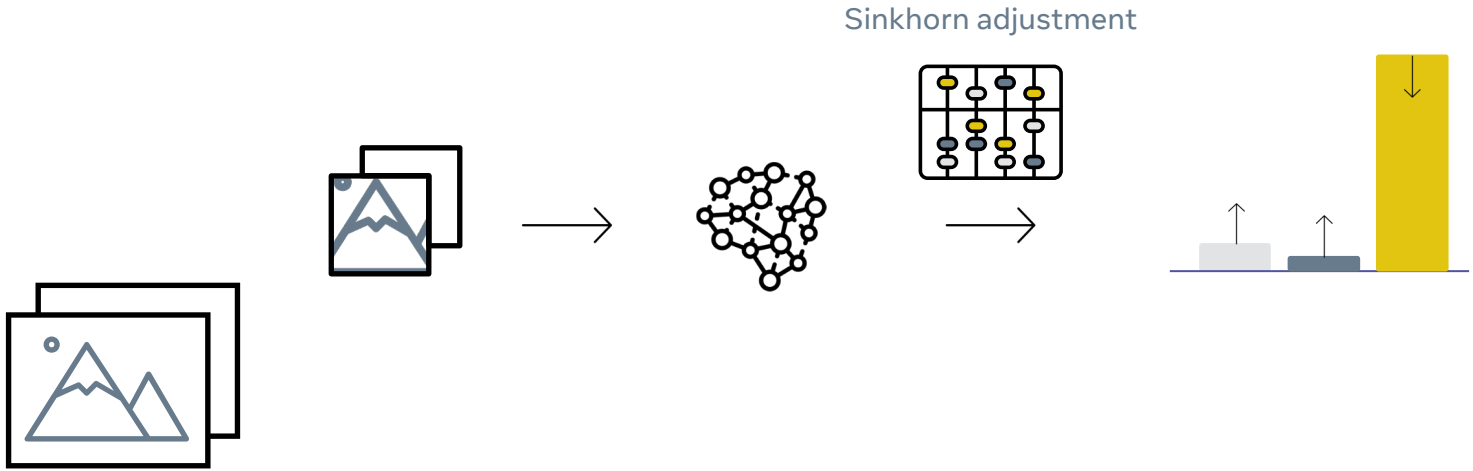


Recap'

- We don't need k-means
- Explicit constraints to prevent collapse
- Scalable

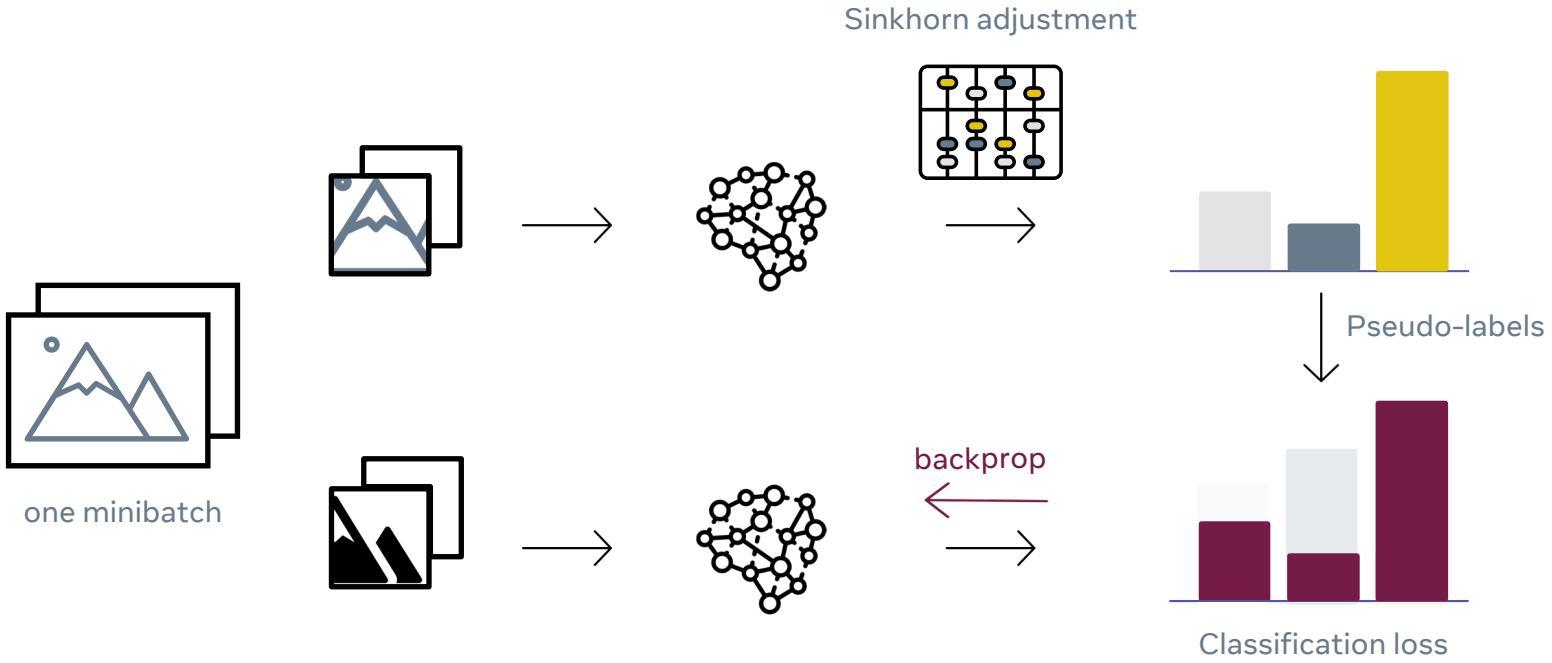


SwAV: the full picture

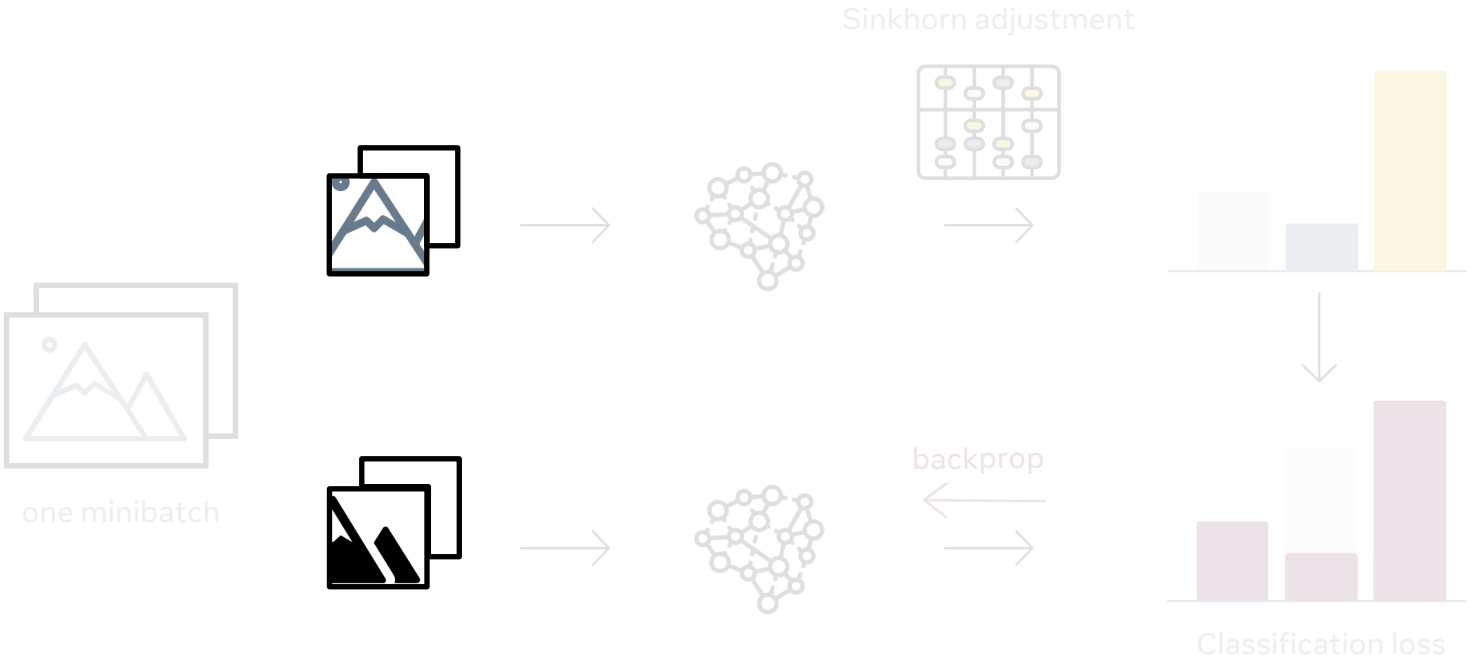


SwAV: the full picture

SimCLR - Chen et al. 2020

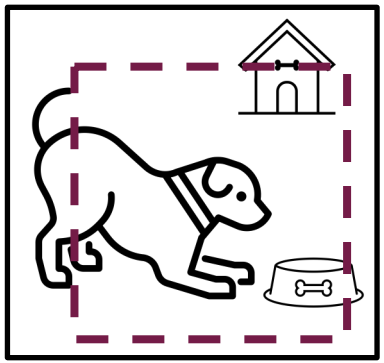


Multi-crop

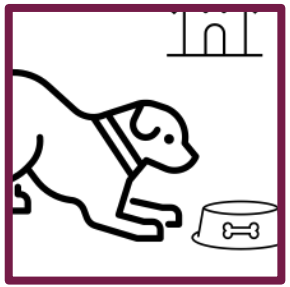

Sinkhorn adjustment
one minibatch
backprop
Classification loss

Multi-crop

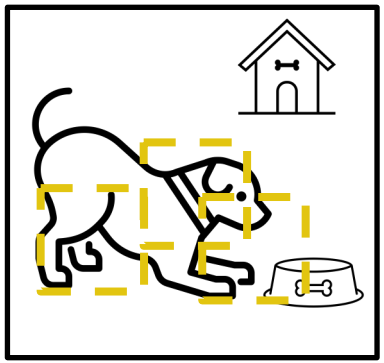
Jigsaw – Noroozi & Favaro. 2016
PIRL - Misra et al. 2020



Global crops



Multi-crop

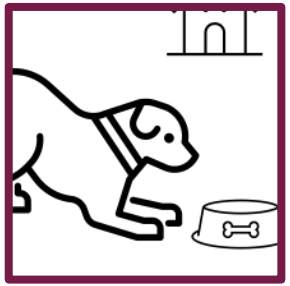


Local crops

Local predict the pseudo-label of global

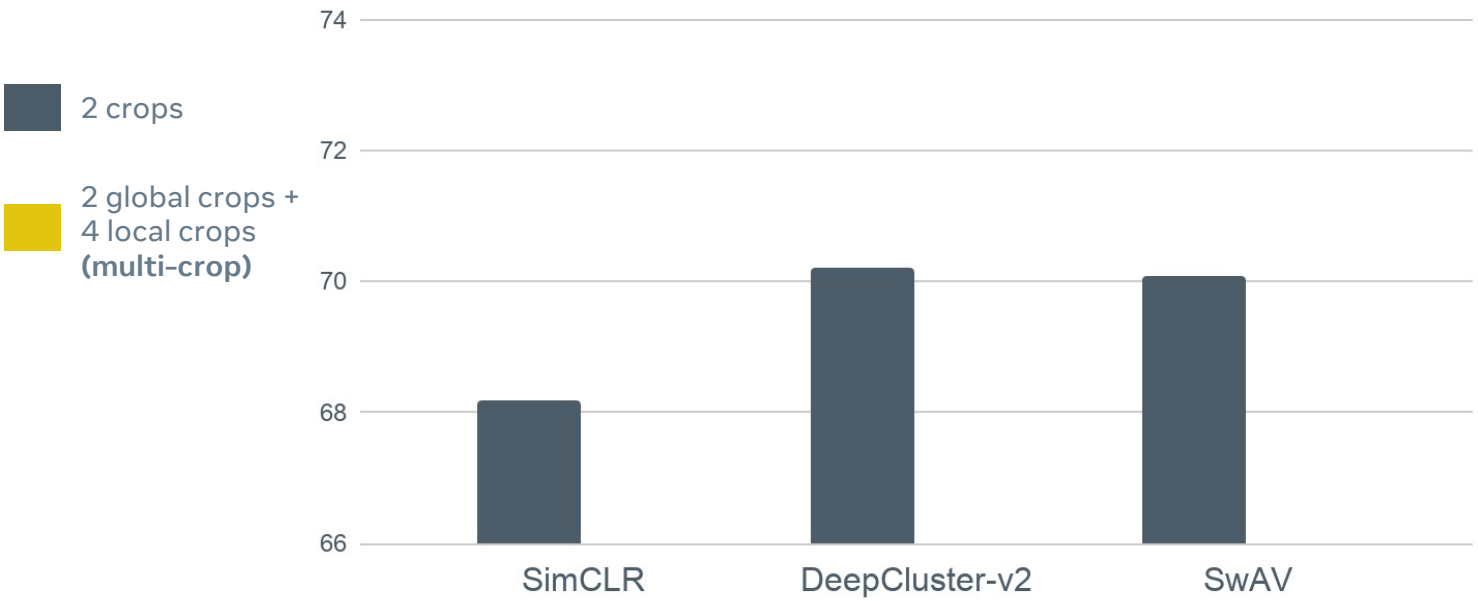
Local-to-global matching

Global crops



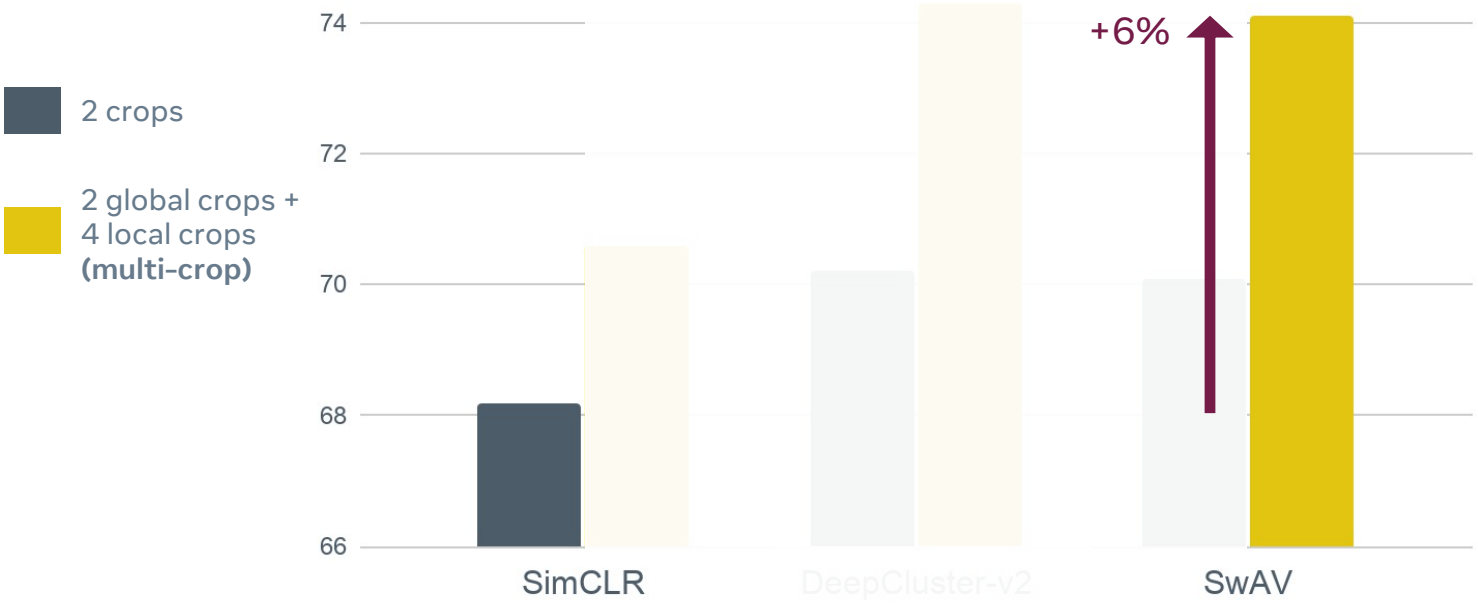
* networks all trained for 400 epochs

Linear benchmark on ImageNet



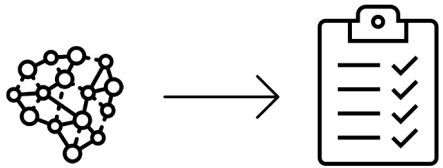
* networks all trained for 400 epochs

Linear benchmark on ImageNet



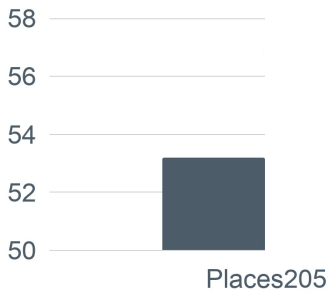
SwAV vs Supervised Pretraining

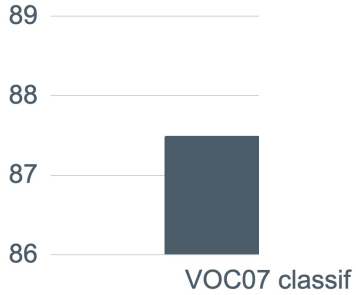
We evaluate representations on different downstream tasks.

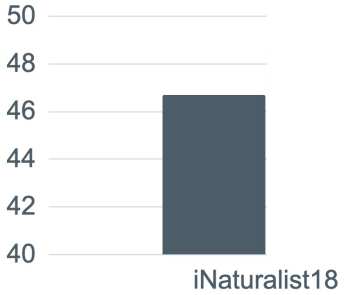


SwAV vs Supervised Pretraining

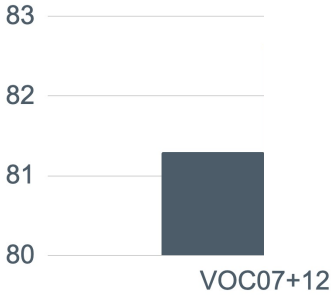
Classification – Linear

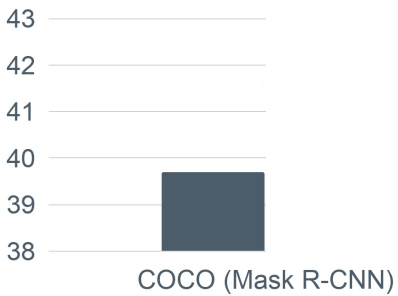


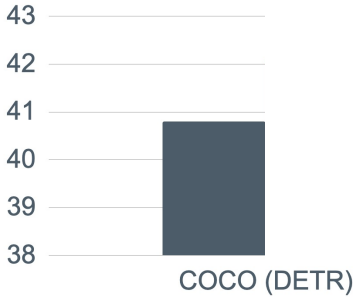




Object Detection – Full finetuning







Great milestone for SSL in 2020

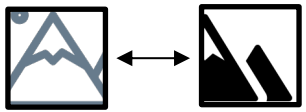
SSL outperform supervised pre-training in transfer learning

Excellent performance on ImageNet
e.g. SimCLR-v2 (Chen et al) and BYOL (Grill et al) > 79% top-1 !!

Great milestone for SSL but...

Recent SSL methods are very similar to each other (simsiam Chen & He 2020)

→ performance saturation



Let us seek progress in an orthogonal direction !

ViT Dosovitskiy et al. 2020
DeiT Touvron et al. 2020

Can we improve SSL by using Vision Transformers ?

DINO: Emerging Properties in Self-Supervised Vision Transformers

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, Armand Joulin

Under review

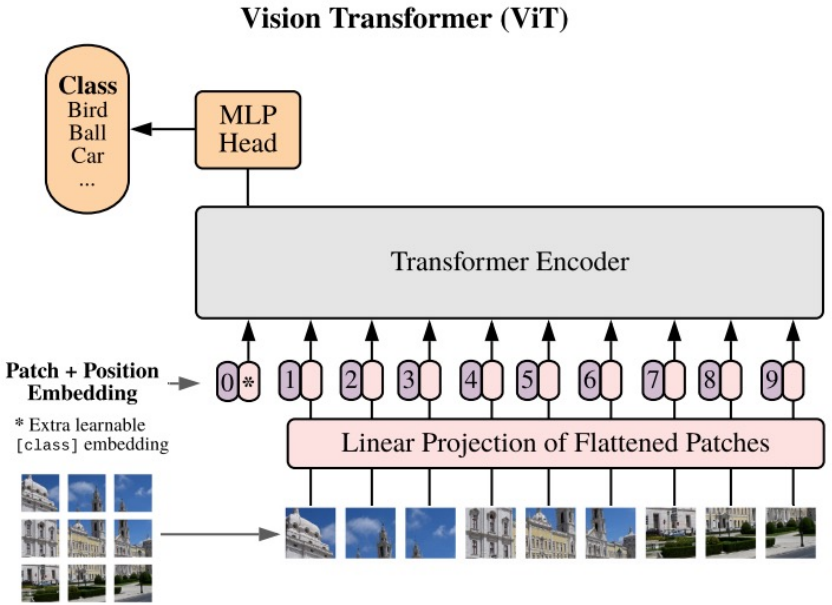
github.com/facebookresearch/dino

FACEBOOK AI

ConvNets & Vision Transformers

ConvNets is de facto architecture for images.

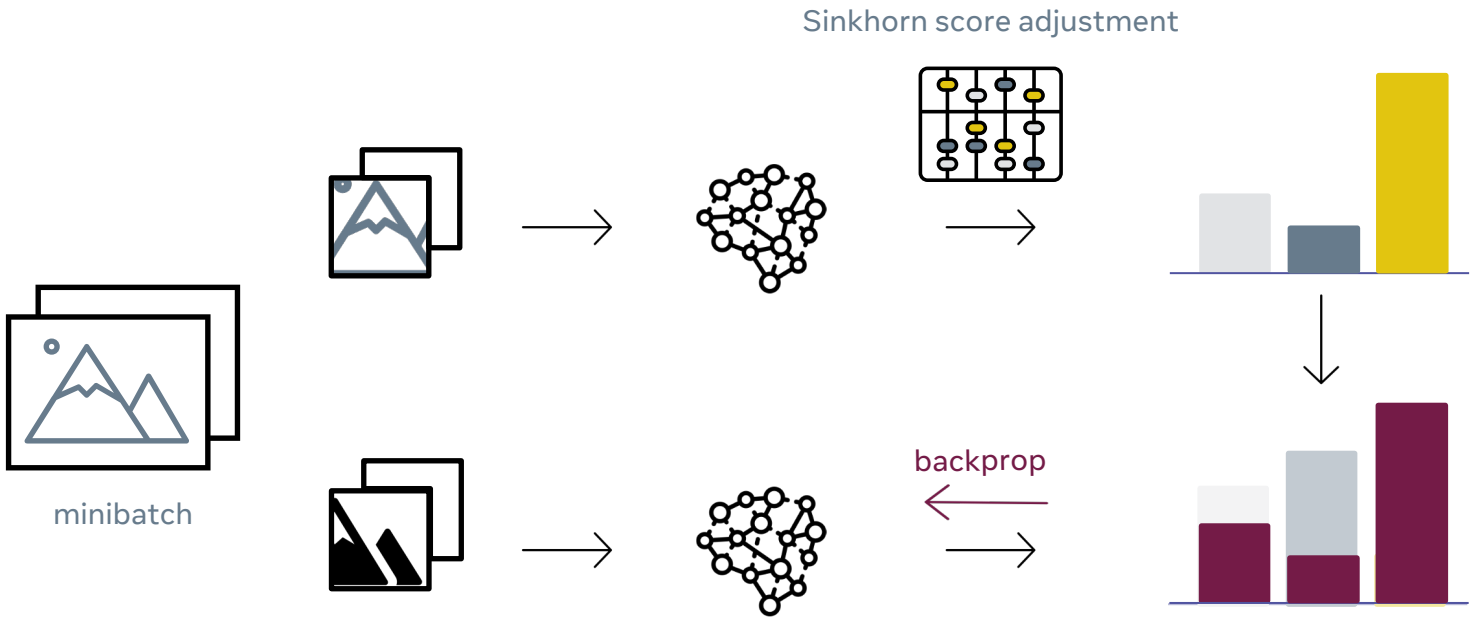
Recently, Vision Transformers (**Dosovitskiy et al. 2020**) have emerged as an alternative to ConvNets.



SwAV: ConvNet VS ViT

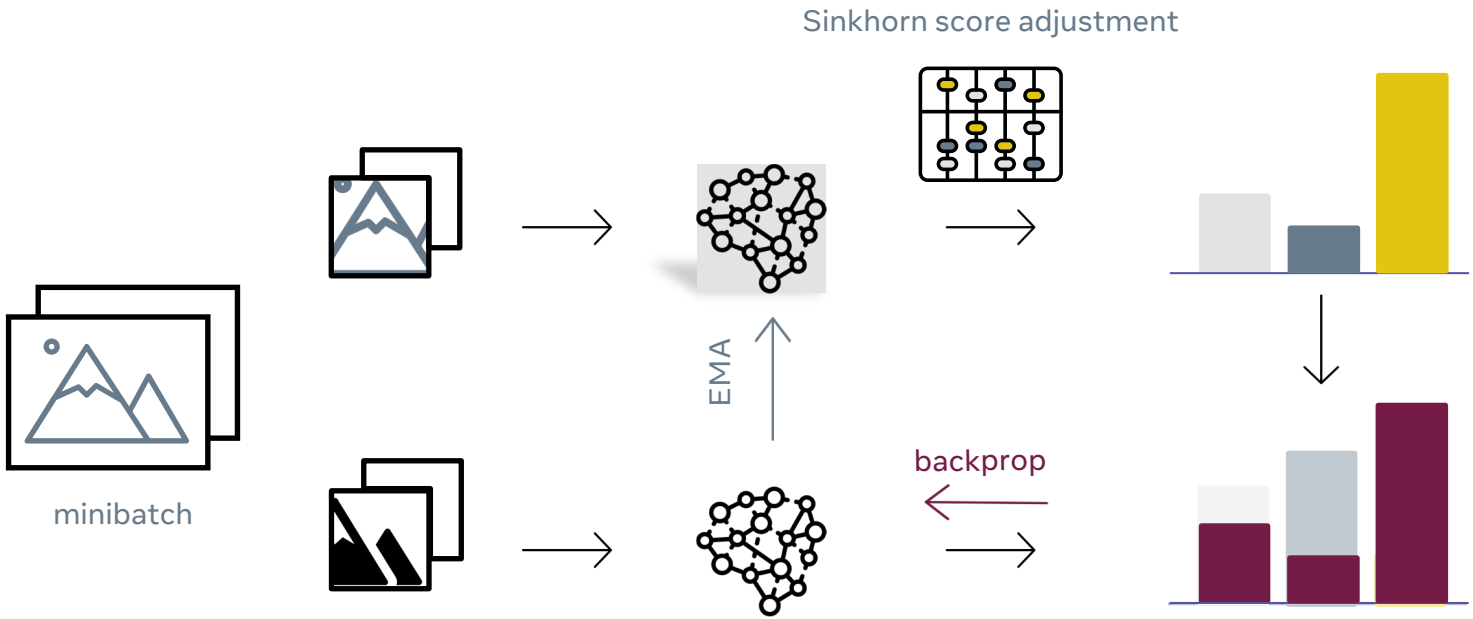
From SwAV to DINO

Mean Teacher – Tarvainen et al. 2017
MoCo - He et al. CVPR 2020
BYOL – Grill et al. NeurIPS 2020

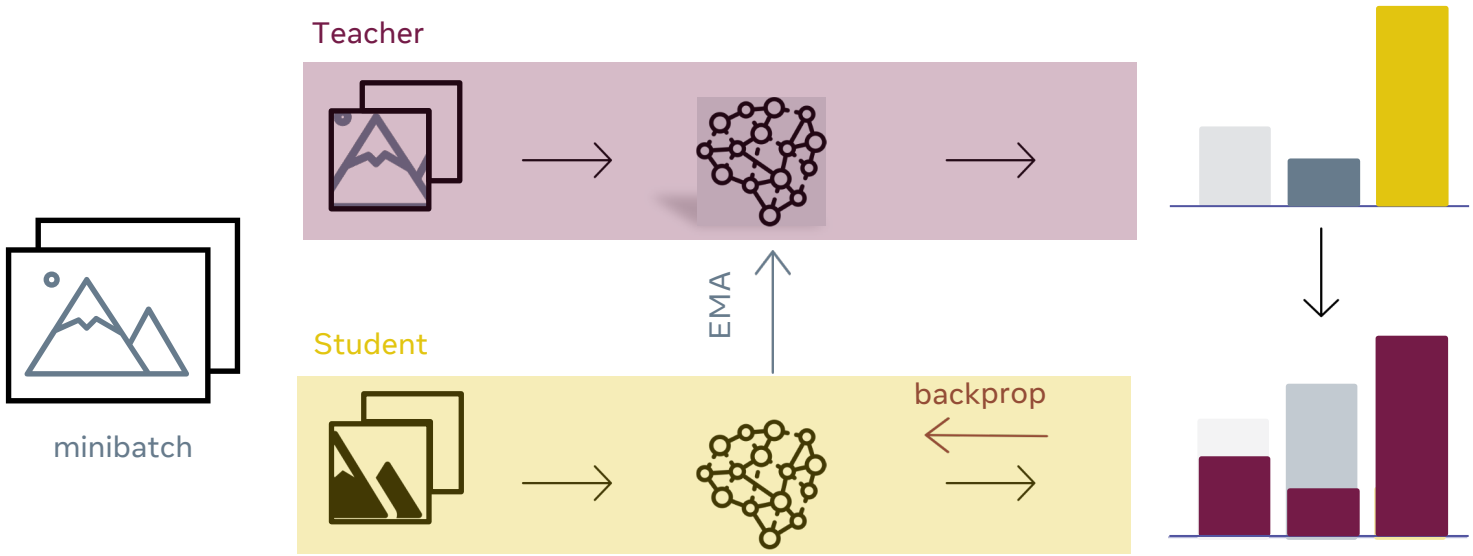


From SwAV to DINO

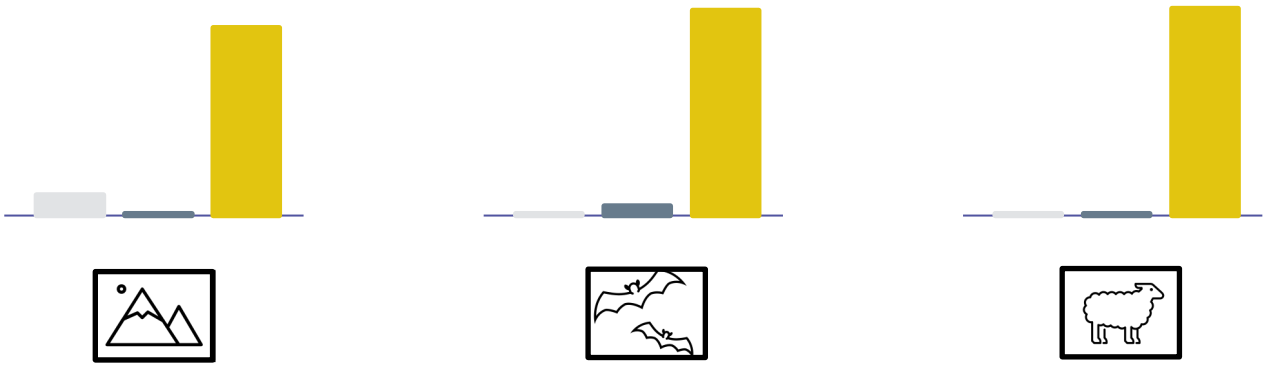
Mean Teacher – Tarvainen et al. 2017
MoCo - He et al. CVPR 2020
BYOL – Grill et al. NeurIPS 2020



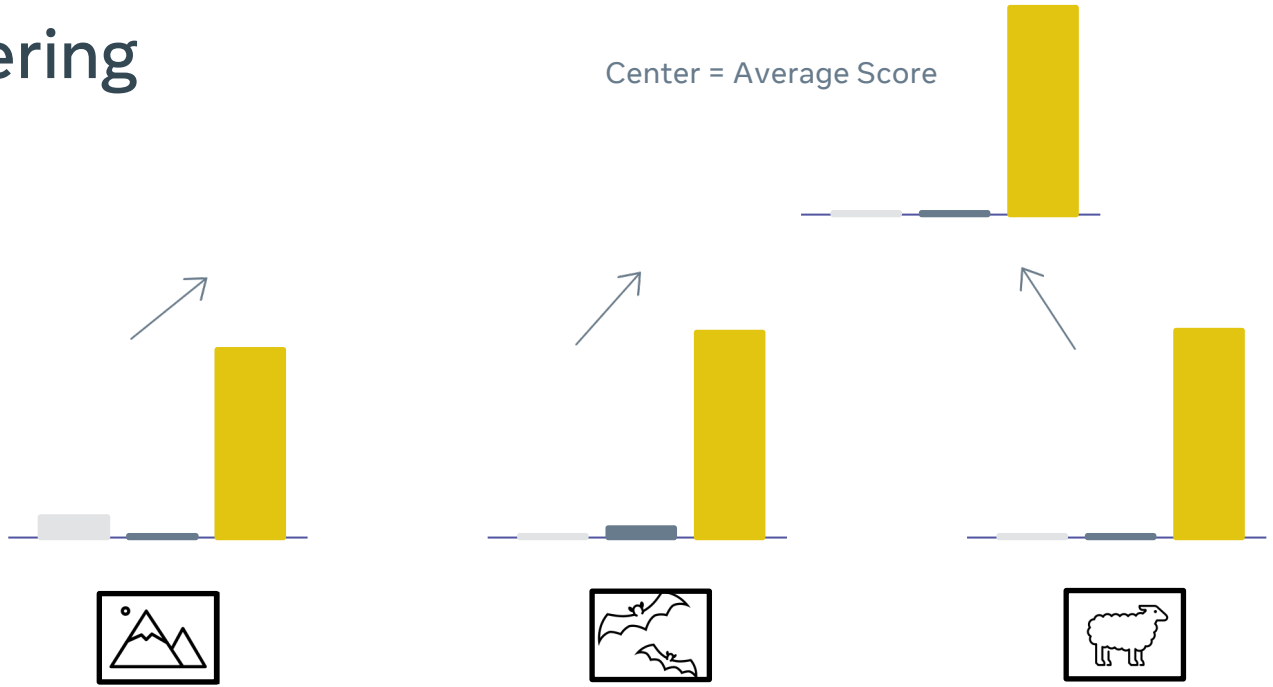
DINO: Self-Distillation with No Labels



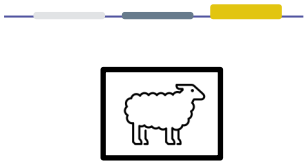
Collapse to one unique dimension



Centering

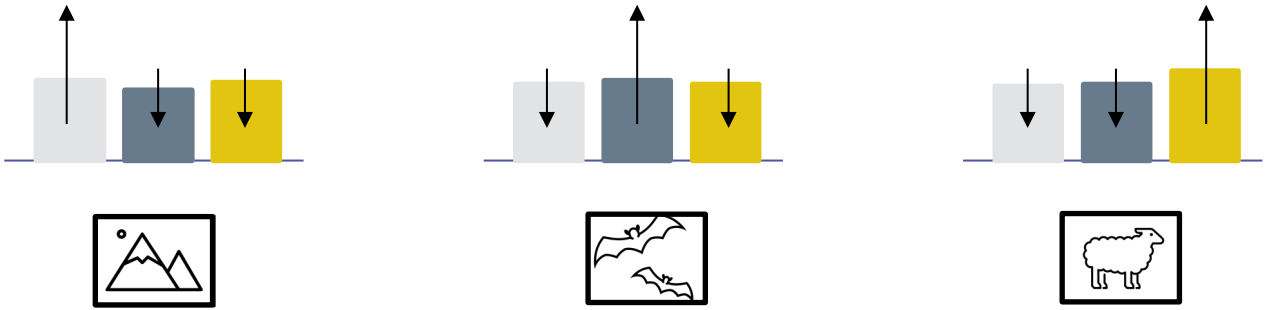


Centering

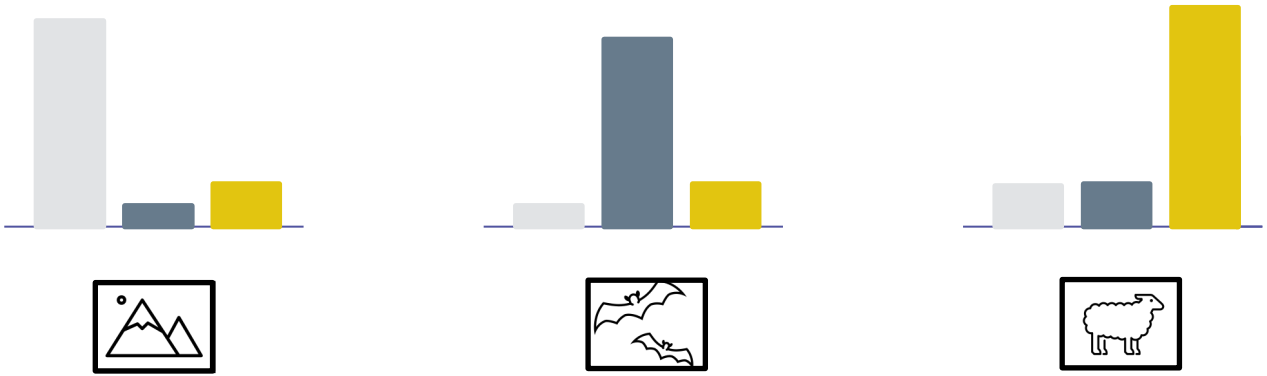


Centering

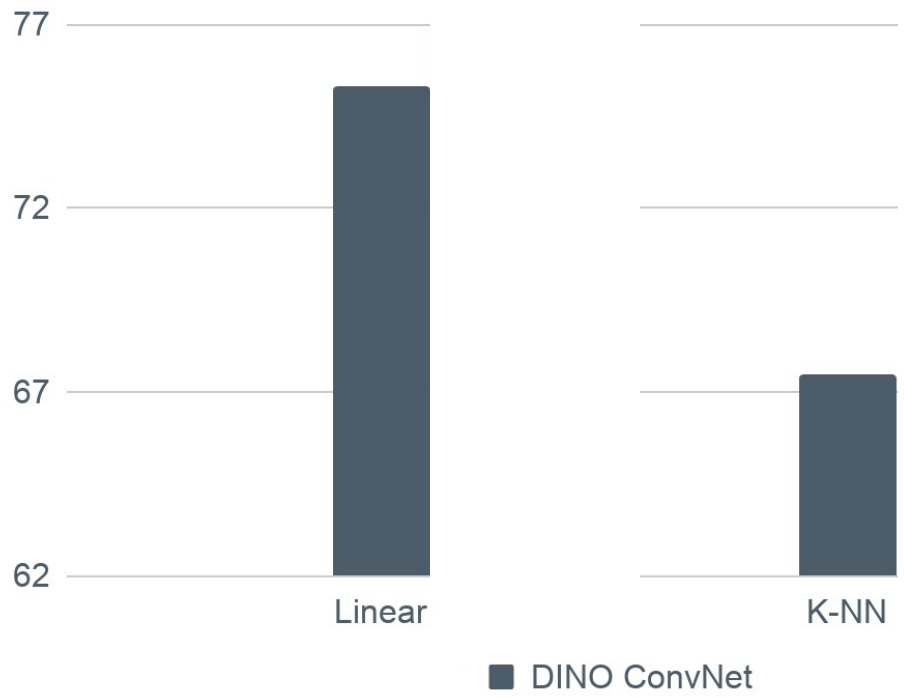
Collapse to uniform assignment



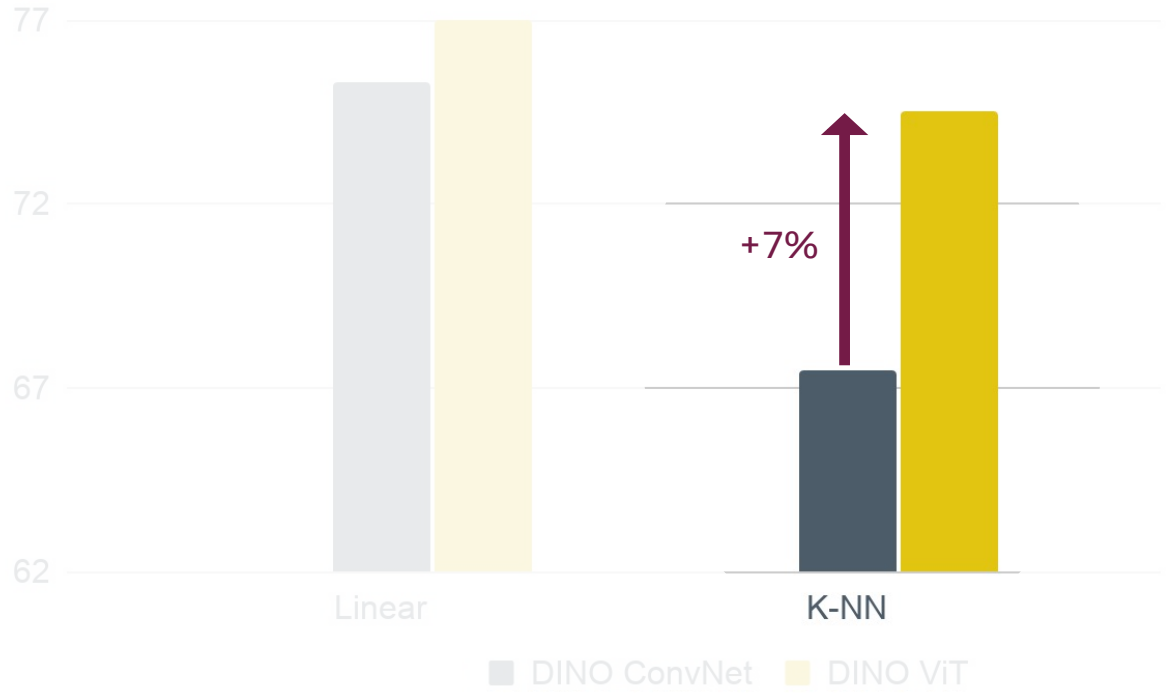
Centering + Sharpening



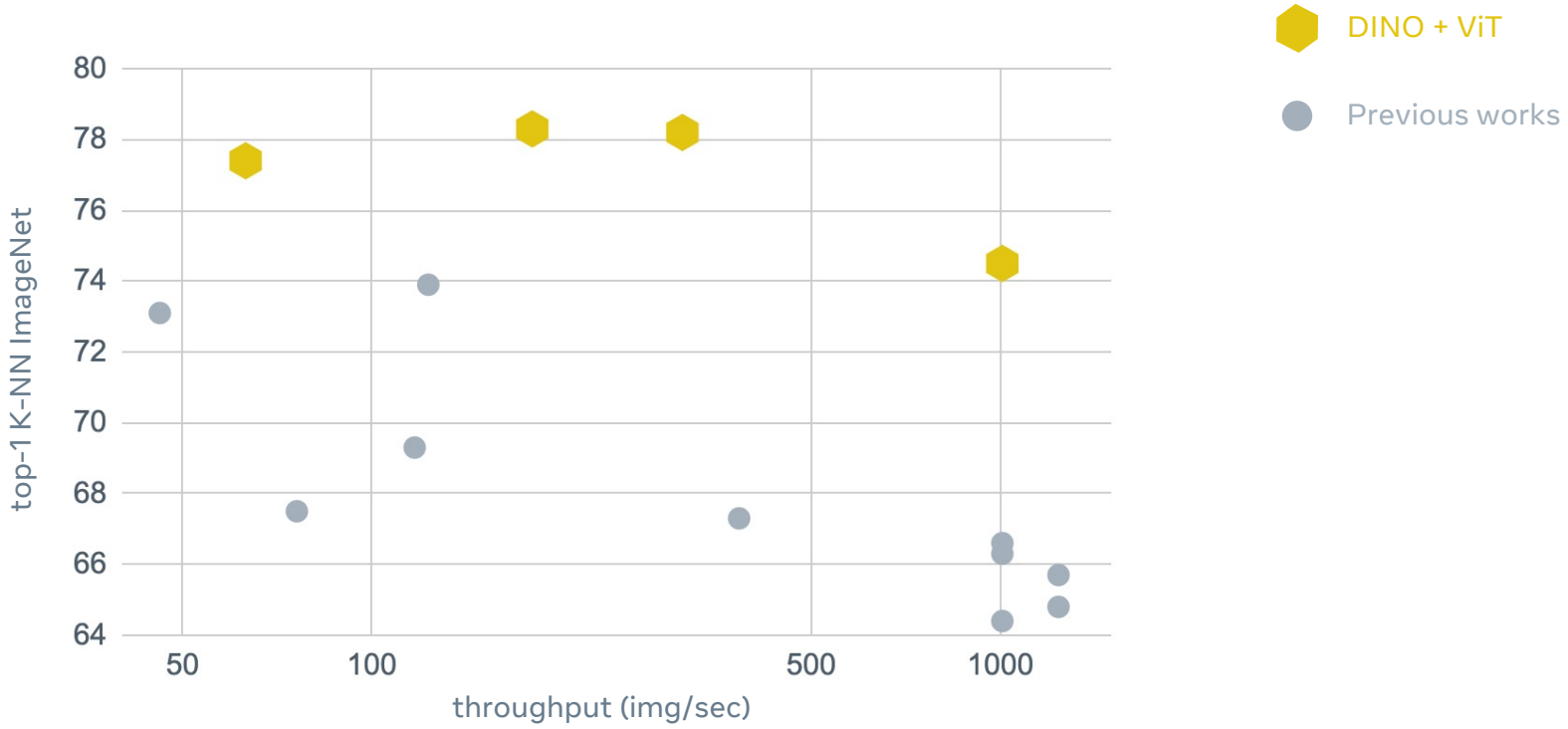
DINO: ConvNet VS ViT



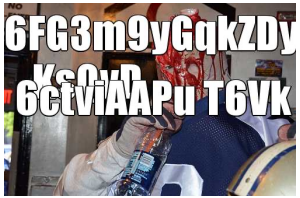
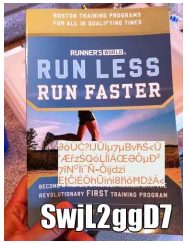
DINO: ConvNet VS ViT



DINO + ViT: excellent K-NN performance



Application to copy detection

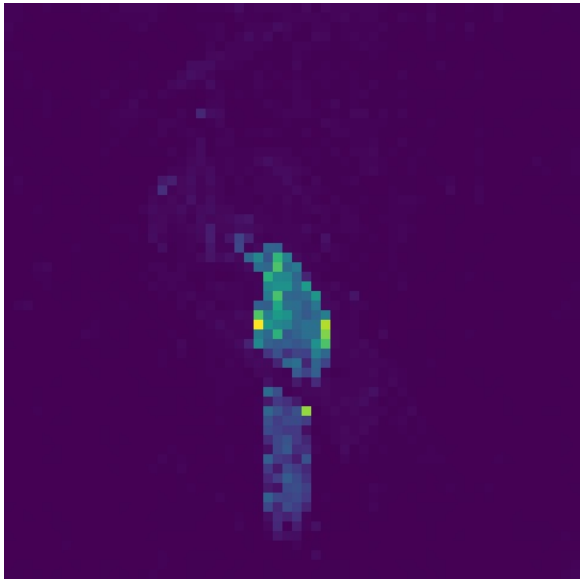
Query	DINO Average Precision: 85.5%	Supervised ViT Average Precision: 76.4%	Multigrain architecture Average Precision: 82.5%
			
			

DINO & ViT: Recap’

- ☑ DINO trains to high performance with ViTs
- ☑ k-NN performance ++
  → Applications to copy detection and image retrieval
- ❑ Interpretability

Self-Attention visualizations

- We look at the self-attention of the [CLS] token of the last block



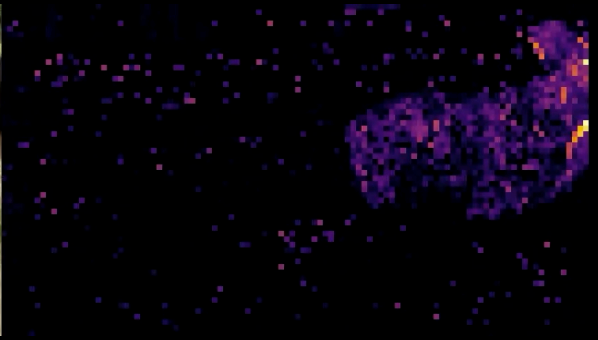
Self-Attention visualizations

- We look at the self-attention of the [CLS] token of the last block

Method	mIoU with GT
DINO	45.9
Supervised	27.3

We train ViT with other SSL methods:

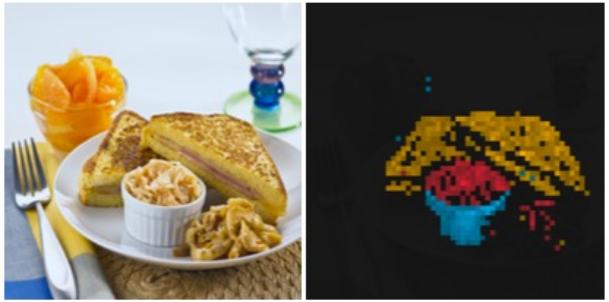
Dino w/o multicrop	45.1
MoCo-v2	46.3
BYOL	47.8
SwAV	46.8



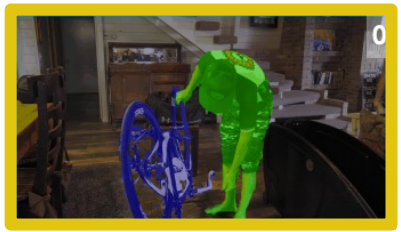
supervised

DINO applied per-frame
to a video

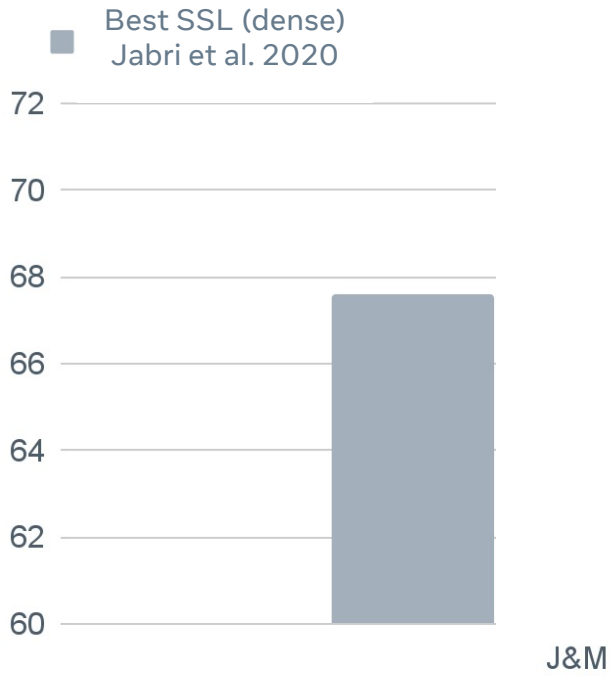
Different attention heads focus on different parts



Application to video object segmentation on DAVIS17



Application to video object segmentation on DAVIS17



Thank You

FACEBOOK AI