Self Supervised Learning

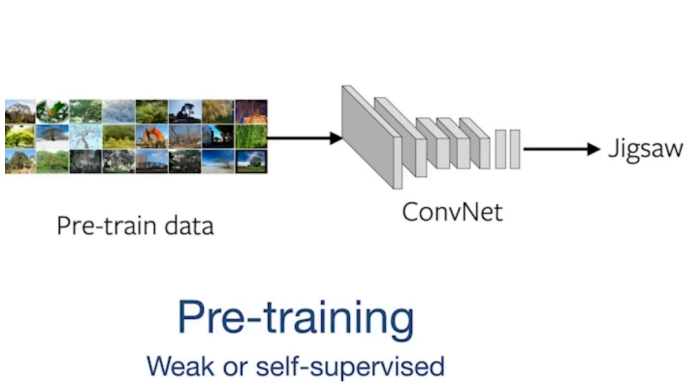
Saturday, 3. July 2021 1:45 PM

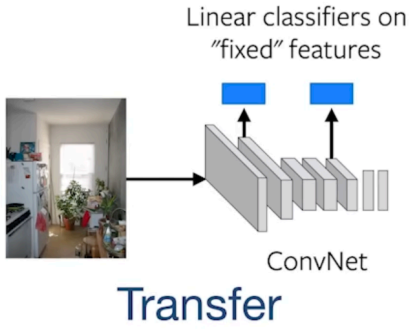
What is "self" supervision?

- Obtain "labels" from the data itself by using a "semi-automatic" process
- Predict part of the data from other parts

=> 1) pre training 2) transfer tasks (classification, detection)

The hope of generalization ... ?



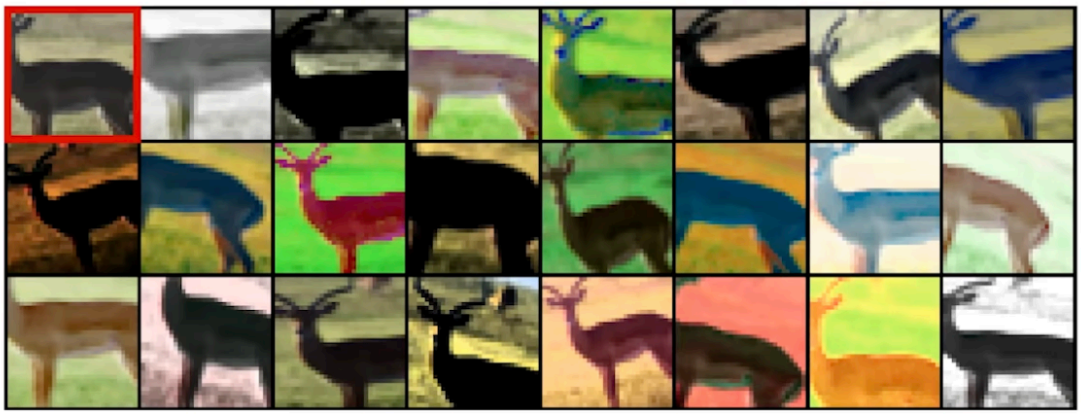


=> solve a pretext task to learn features (using images, videos, audio)
=> for example : patch pose, jigsaw, predicting rotations (n-way classification)

What should pretraining do ?

- represent how images relate and be robust to different conditions

Popular & Common principle for most methods



Learn features such that:

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

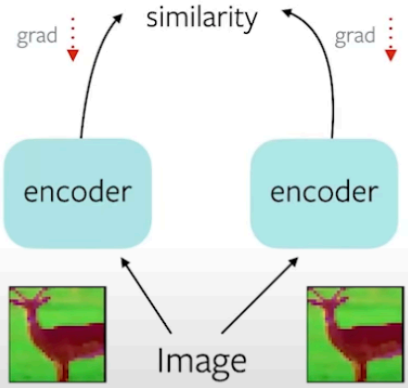
Figure from Dosovitskiy et al., 2014

Often uses a contrastive learning technique :

Trivial Solutions

$$f_\theta(I) = f_\theta(\text{augment}(I))$$

$$f_\theta(I) = \text{constant}$$



Satisfies the invariance property, but not useful

=> does not produce how images relate to one another

Recent methods categorization of recent self supervision techniques :

Many ways to avoid trivial solutions

Similarity Maximization Objective

- Contrastive learning
  - MoCo, PIRL, SimCLR
- Clustering
  - DeepCluster, SeLA, SwAV
- Distillation
  - BYOL, SimSiam

Redundancy Reduction Objective

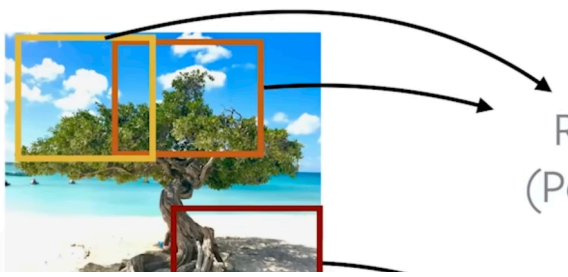
- Redundancy Reduction
  - Barlow Twins

Pretraining :

1) On Imagenet without labels (pretrain ResNet50 initalized randomly) :
2) Transfer to downstream task

Contrastive Learning (MOCO, SImCLR, PIRL)

Nearby patches vs. distant patches of an Image



Related
(Positives)

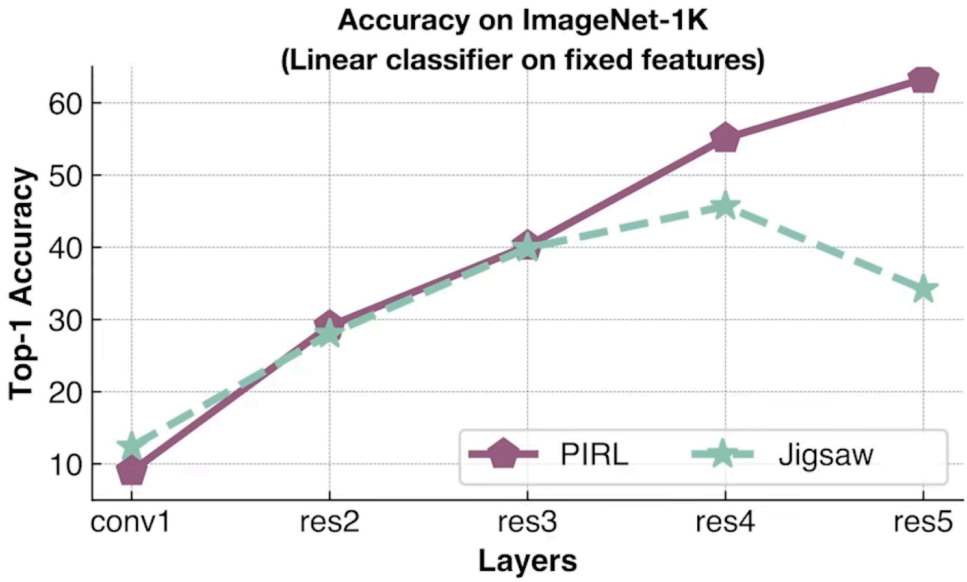
van der Oord et al., 2018,
Henaff et al., 2019

Unrelated
(Negative)

Contrastive Predictive Coding

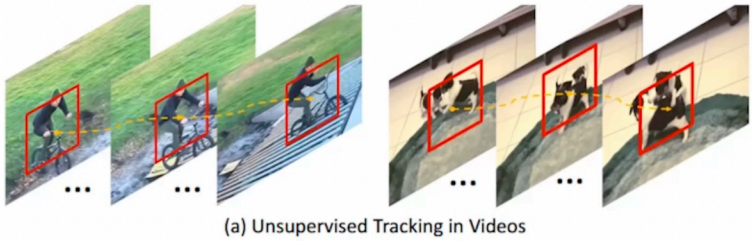
=> improves the semantic features when compared to jigsaw

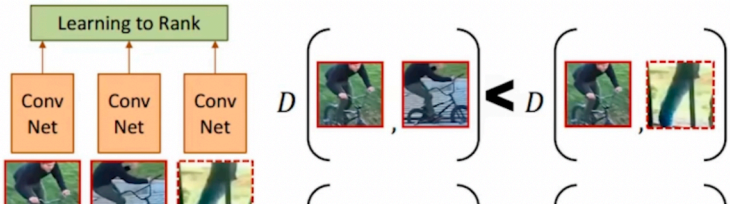
Semantic Features?

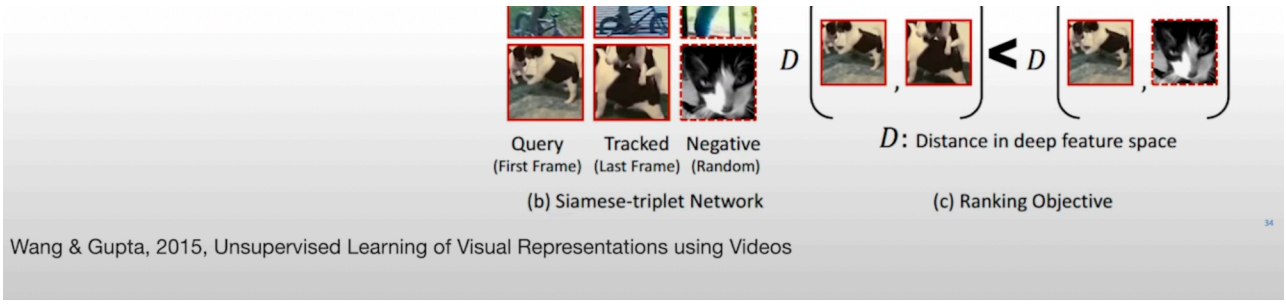


Also for tracking :

Tracking Objects





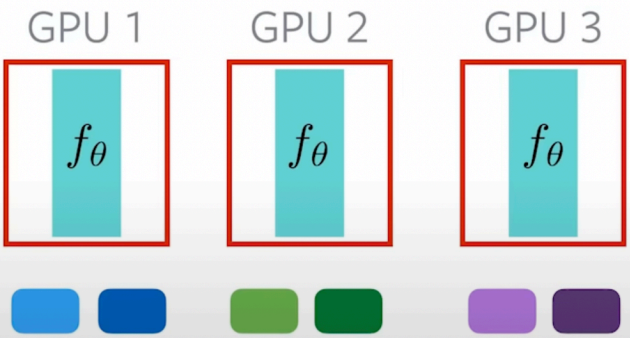


Three possible ways :

Good negartives are very important !!

SimCLR

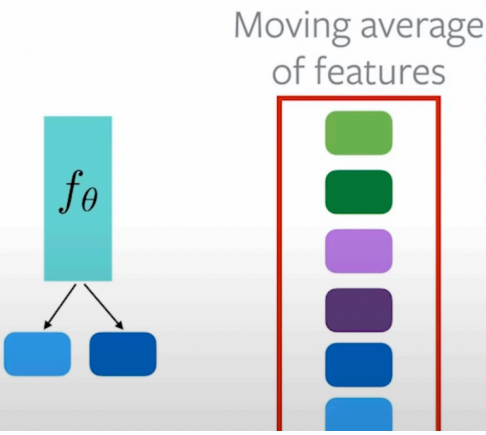
- Large batch size - e.g. in SimCLR
- Pros - Simple to implement
- Cons - Large batch size



A simple framework for contrastive learning - Chen et al., 2020

Memory Bank

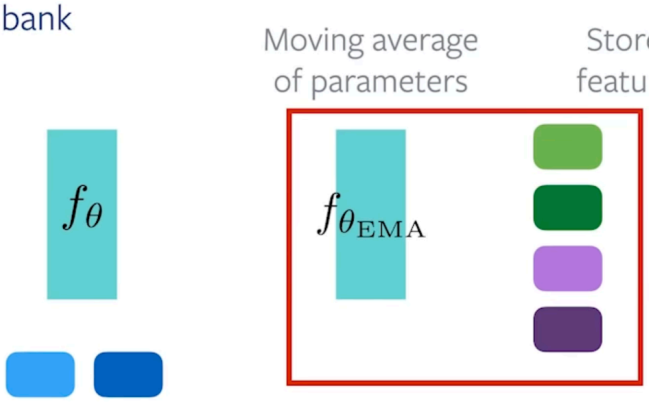
- Maintain a "memory bank" -- momentum of activations
- Pros - compute efficient
- Cons - Needs large memory, not "online"



Non parametric Instance Discrimination - Wu et al., 2018

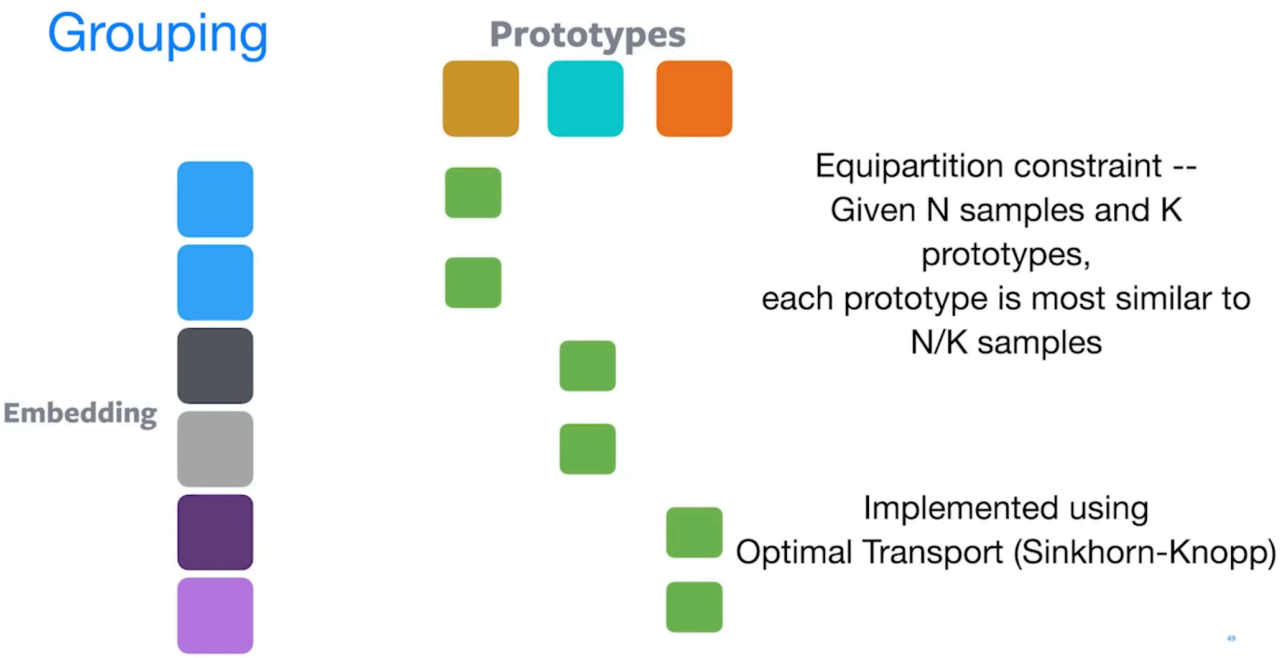
MoCo

- Maintain "momentum" network - MoCo
- Pros - online
- Cons - extra memory for parameters/stored features, extra fwd pass compared to memory bank



Momentum Contrast - He et al., 2019

Clustering Methods (SWAV)



Key Results

	Linear Classifier (Fixed Features)			Detection (Fine-tuned)	
	ImageNet	Places	iNaturalist	VOC07+12	COCO
Supervised	76.5	53.2	46.7	81.3	40.8
Prior self-supervised	71.1 (-5.4)	52.1	38.9	82.5	42.0
SwAV	75.3 (-1.2)	**56.7**	**48.6**	**82.6**	**42.1**

SEER outperforms this :

SEER: Learning from uncharted images

Method	Pretrain images	Curated	Arch.	Params	ImageNet top-1
Hashtag prediction	1B	Yes	X101-32x8d	91M	82.6
SEER	1B	No	X101-32x8d	91M	81.6

SEER uses no labels and works on random images
Comparable performance to networks trained on curated data with weak supervision

Architecture: ResNeXt-101-32x8d

In video : Audio - Image agreement are good (CMA)

Distillation :

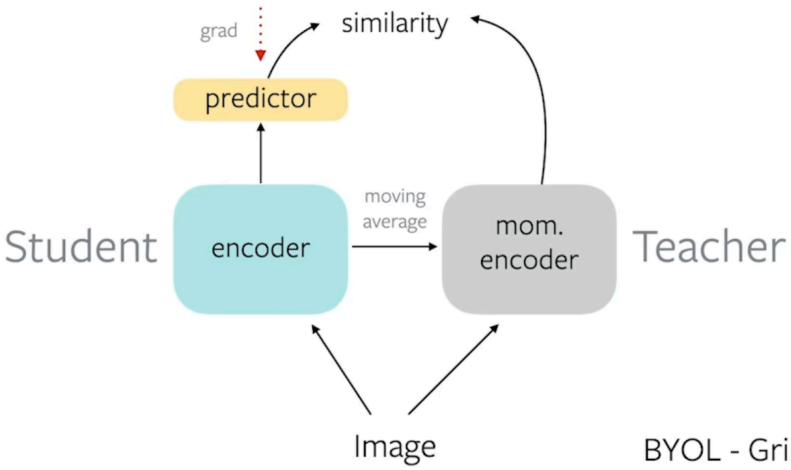
Distillation

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$

- How we do it $f_\theta^{\text{student}}(I) = f_\theta^{\text{teacher}}(\text{augment}(I))$
- Prevent trivial solutions by asymmetry
  - Asymmetric **learning rule** between student teacher
  - Asymmetric **architecture** between student teacher

BYOL

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$
- How we do it $f_\theta^{\text{student}}(I) = f_\theta^{\text{teacher}}(\text{augment}(I))$

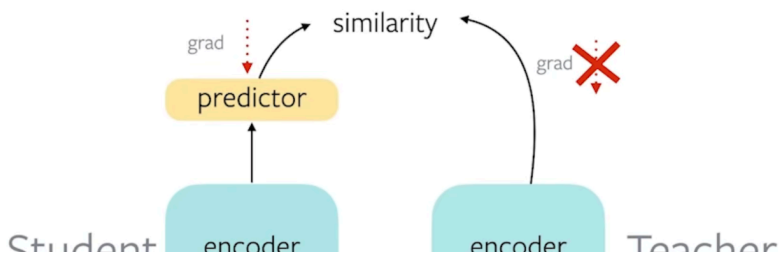


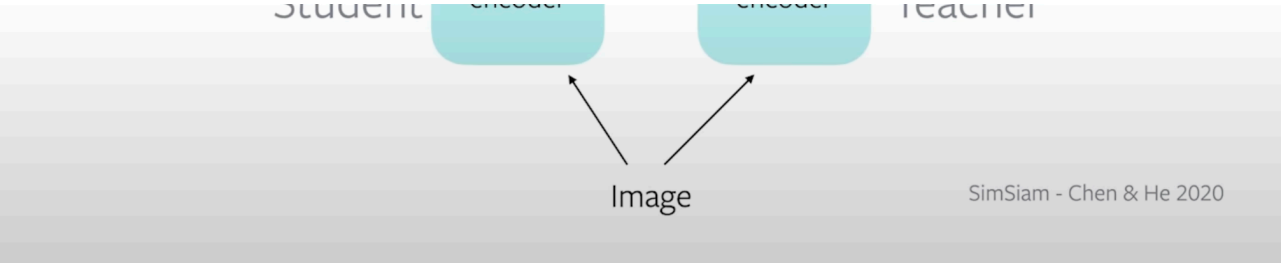
BYOL - Grill et al., 2020

SImSiam does not need the moving average (no seperate set of weights) : Only 2 sources of asymmetry (learning update and prediction head)

SimSiam

- What we want $f_\theta(I) = f_\theta(\text{augment}(I))$





Barlow Twins :

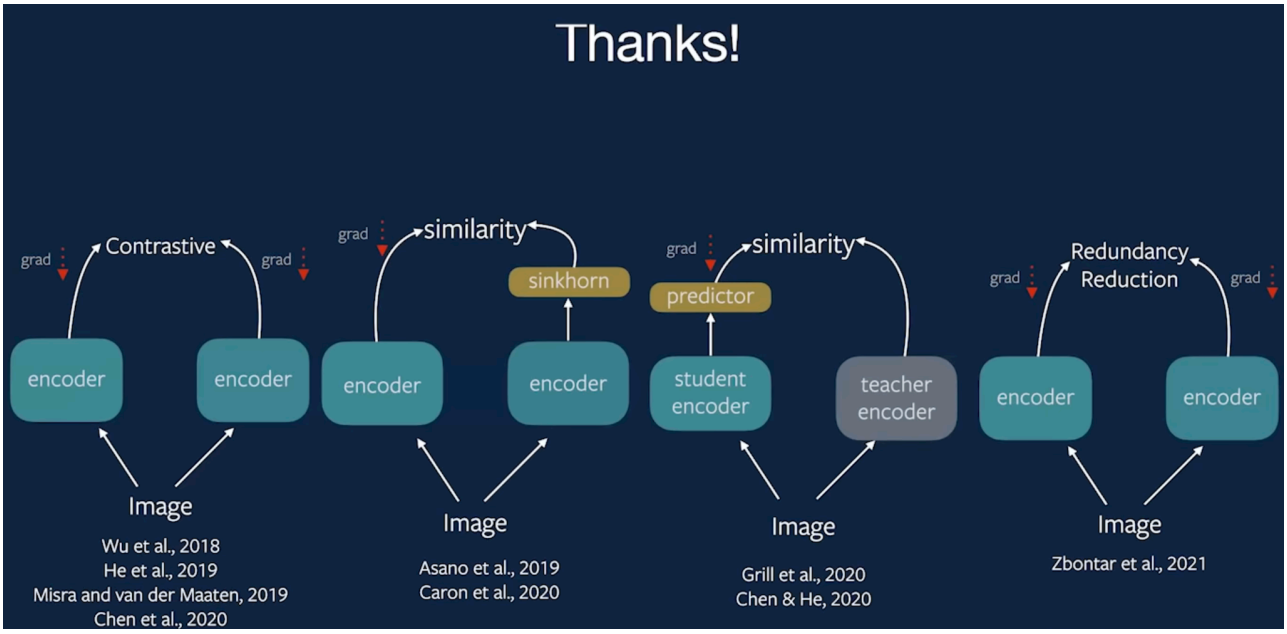
Inspired by Information Theory
Neurons communicate by spiking
Spikes aim to reduce redundancy
Horace Barlows efficient coding hyptothesis

Each Neuron should satisfy :

- **Invariance**
- **Independen**
  - VERY roughly speaking
  - $$f_\theta(I)[i] = f_\theta(\text{augment}(I))[i]$$
    $$f_\theta(I)[i] \neq f_\theta(\text{augment}(I))[j]$$

Other facts :

- **Cetrain loss functions are more depedent on the batch size**



Contrastive	Clustering	Distillation	Redundancy Reduction

Ishan Misra :

General Pros/Cons :

- Clustering based models converge faster (good for small compute)
- Simplicity of implementation : Barlow Twins
- Very different modularities : Contrastive Learning better (two very different encoders/architectures)