Contrastive Learning Papers

InstDisc → MoCo → MoCo v2

InvaSpread → SimCLR → SimCLR v2

CPC

CMC

SwAV

No negative samples: BYOL → SimSiam

ViT based: MoCo v3, DINO

InstDisc: **Unsupervised Feature Learning via Non-Parametric Instance-level Discrimination** (2018) 2346

Instance-level classification as the pretext task, to tackle the computational challenges imposed by the large number of instance classes

InvaSpread: **Unsupervised Embedding Learning via Invariant and Spreading Instance Feature** (2019) 388

Instance-level classification as the pretext task

End-to-end learning, without memory bank

CPC: **Representation Learning with Contrastive Predictive Coding** (2018) 4172

learn such representations by predicting the future in latent space by using powerful autoregressive models

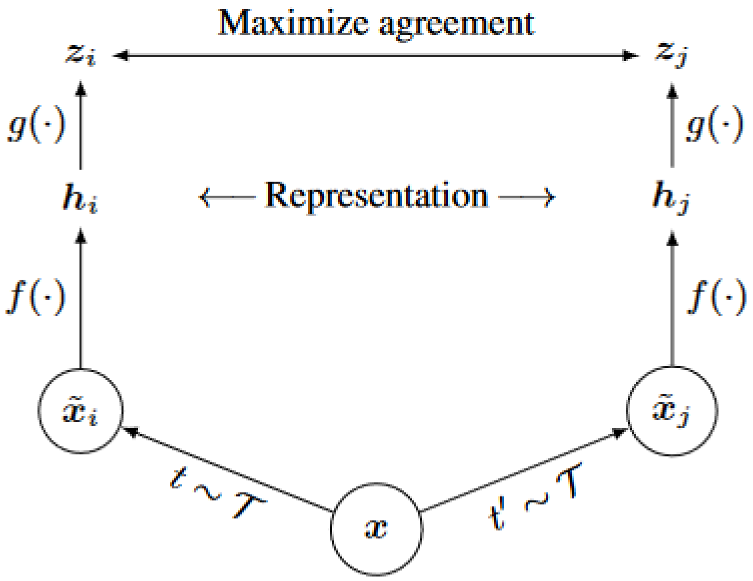
InfoNCE loss: maximize the log-likelihood of correctly identifying a positive pair (two views of the same input) among a set of negative pairs (two views of different inputs)

CMC: **Contrastive Multiview Coding** (2019) 1546

positive samples: different views of the same scene

MoCo (2020) see separate notes

SimCLR: **A Simple Framework for Contrastive Learning of Visual Representations** (2020) 8411



projection head $g(\cdot)$: MLP, only used in training

Composition of multiple data augmentation operations

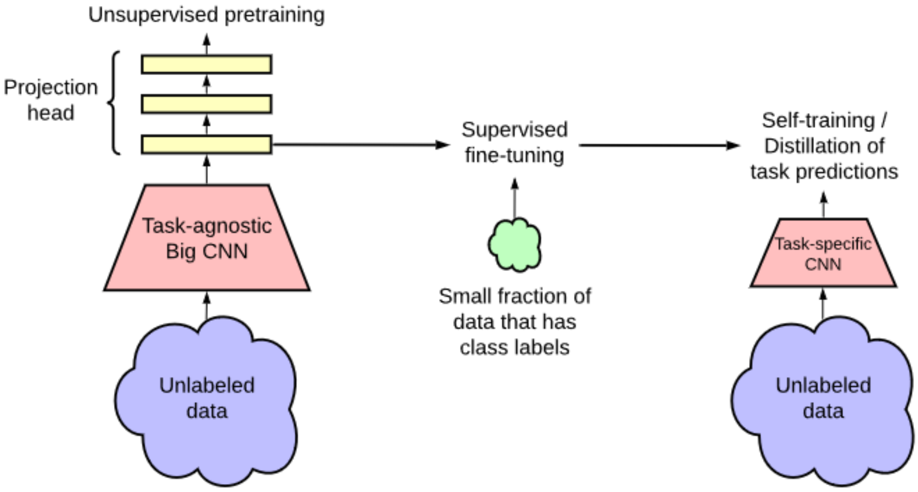
MoCo v2: **Improved Baselines with Momentum Contrastive Learning** (2020) 1686

MoCo + tricks from SimCLR: MLP projection head, more augmentation, more epochs

SimCLR v2: **Big Self-Supervised Models are Strong Semi-Supervised Learners** (2020) 1299

Larger model: 152 layers ResNet

MLP projection head: two layers work best



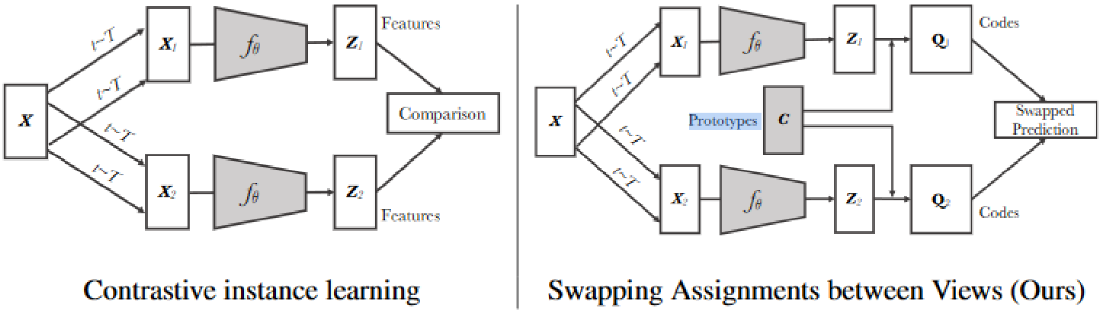
SwAV: **Unsupervised Learning of Visual Features by Contrasting Cluster Assignments** (2020) 1828

Contrastive method + clustering

A “swapped” prediction mechanism where we predict the code of a view from the representation of another view

A new data augmentation strate, multi-crop: use a mix of views with different resolutions in place of two full-resolution

prototype (cluster center)



Contrastive instance learning

Swapping Assignments between Views (Ours)

Figure 1: **Contrastive instance learning (left)** ***vs.*** **SwAV (right).** In contrastive learning methods applied to instance classification, the features from different transformations of the same images are compared directly to each other. In SwAV, we first obtain “codes” by assigning features to prototype vectors. We then solve a “swapped” prediction problem wherein the codes obtained from one data augmented view are predicted using the other view. Thus, SwAV does not directly compare image features. Prototype vectors are learned along with the ConvNet parameters by backpropragation.

BYOL: **Bootstrap your own latent: A new approach to self-supervised Learning** (2020) 2735

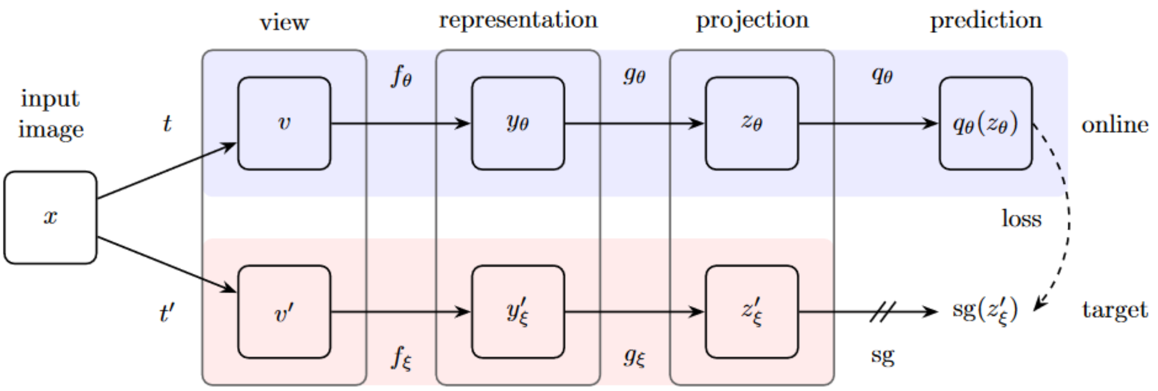
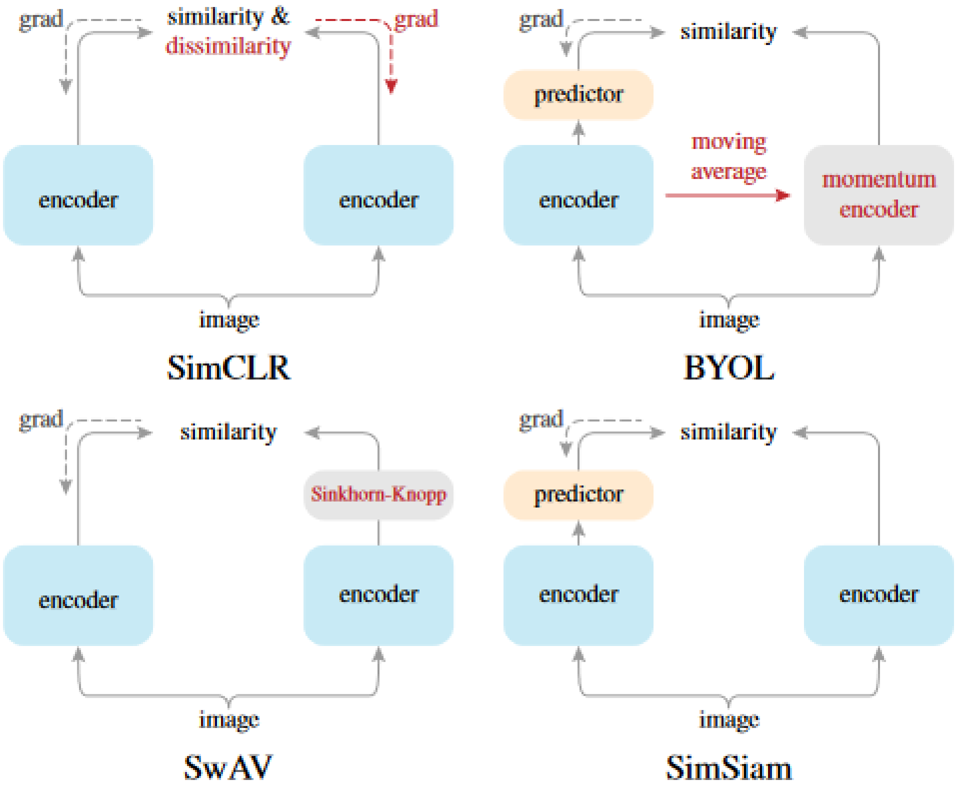


Figure 2: BYOL's architecture. BYOL minimizes a similarity loss between $q_\theta(z_\theta)$ and $\mathrm{sg}(z'_\xi)$, where θ are the trained weights, ξ are an exponential moving average of θ and sg means stop-gradient. At the end of training, everything but f_θ is discarded, and y_θ is used as the image representation.

No negative samples, predicting itself, MSE Loss

Implicit negative samples: mode by batchnorm

SimSiam: **Exploring Simple Siamese Representation Learning** (2020) 1747



MoCo v3: **An Empirical Study of Training Self-Supervised Vision Transformers** (2021) 591

Backbone: ResNet → ViT

Fixed random patch projection: more stable

DINO: **Emerging Properties in Self-Supervised Vision Transformers** (2021) 1201

Self-supervised ViT features contain explicit information about the semantic segmentation of an image, which does not emerge as clearly with supervised ViTs, nor with convnets. These features are also excellent k-NN classifier.