Self-supervised Learning

Training Deep Networks from Zero to Hero: avoiding pitfalls and going beyond

Gabriel Cavallari
gabriel.cavallari@usp.br

SIBGRAPI 2021

Context

- Given a task and enough labels, supervised learning can probably solve it really well

- Large amounts of manually labeled data can be:
  - costly, time-consuming, complex and expensive to obtain

- Sometimes real applications require categories that are not present in standard large-scale benchmark datasets

Self-supervised Learning

- Self-supervision is a form of unsupervised learning where the data itself provides the supervision.

- It relies on pretext tasks that can be formulated using only unsupervised data. By producing surrogate labels, those tasks make use of those generated labels to guide the learning process.

- So the pretext task guides us to a supervised loss function. However, we usually don’t care about the final performance of this task.

Self-supervised Learning

- We can think that those models predict part of the data
  - from other incomplete, transformed, distorted or corrupted parts.

- Most models "learn to compare", using some kind of contrastive learning strategy.

Self-supervised Learning

- These models are able to learn useful image representations in order to solve those tasks

- Achieve state-of-the-art performance when we consider methods that rely only on unlabeled images, benefiting almost all types of downstream tasks.

Self-supervised Learning

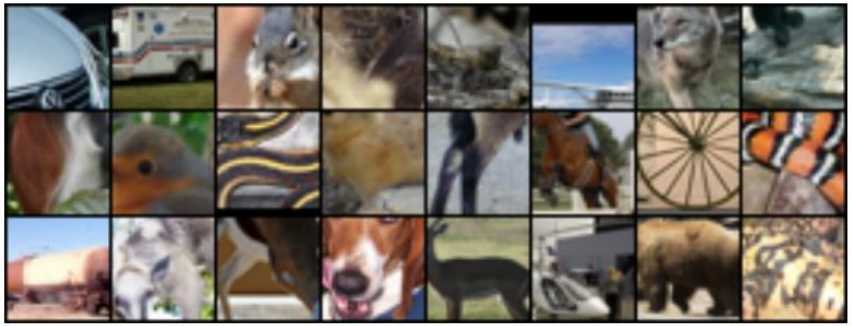
- General overview of methods
  - implementation details can be **very** important here

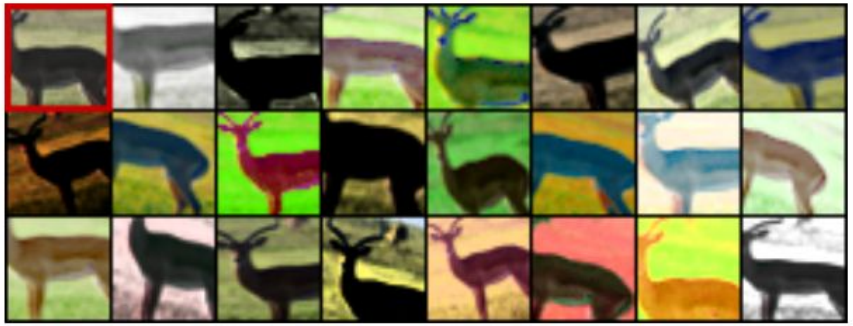
- Most works do use basic structure of CNNs, but very often they modify:
  - the training losses
  - architectures
  - batch manipulation
  - distributed training across multiple GPUs/TPUs

Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks **(Exemplar) - [Dosovitskiy et al., 2015]**

- Train a network to discriminate between a set of surrogate classes.

1. Sample N patches from different images at varying positions and scales
2. Apply a variety of random transformations to each patch.
3. All the resulting distorted patches are considered to belong to the same surrogate class.
4. The pretext task is to discriminate between a set of surrogate classes.

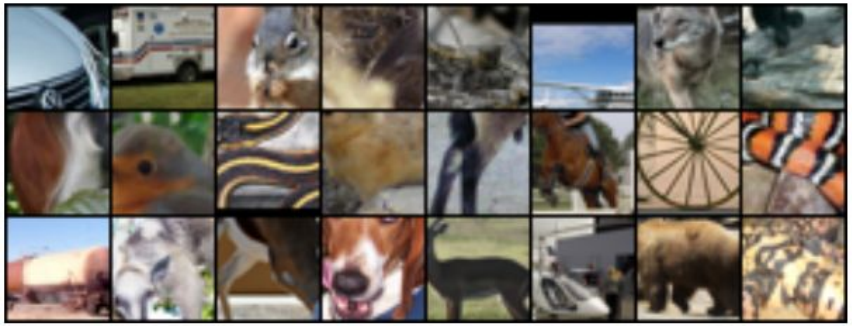


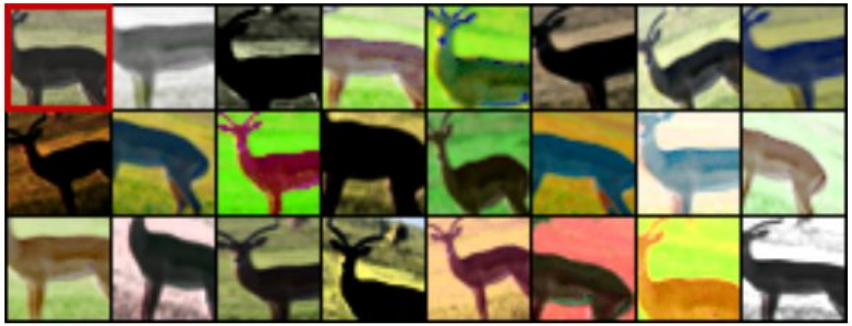


Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks **(Exemplar) - [Dosovitskiy et al., 2015]**

- Things to consider:

1. Number of surrogate classes
2. Number of Samples per surrogate class
3. Types of transformation

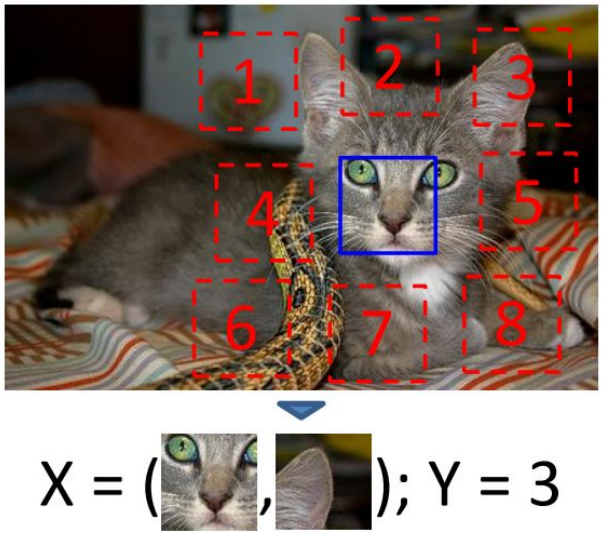




Unsupervised Visual Representation Learning by Context Prediction **(Context Prediction) - [Doersch et al., 2015]**

- Train a network that predicts relative location of two randomly sampled non-overlapping image patches

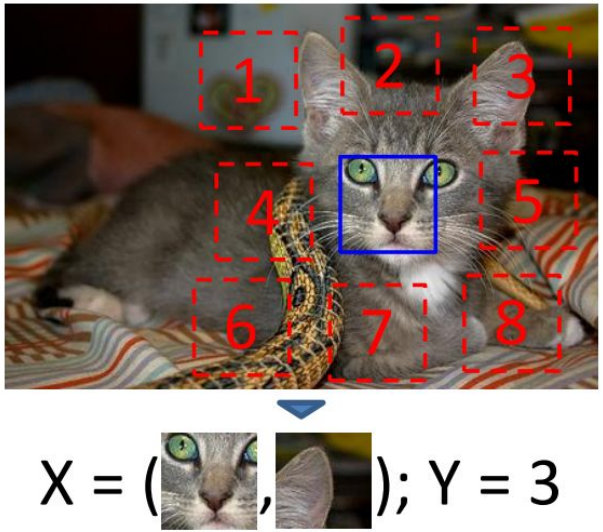
1. Randomly sample the first patch
2. Considering that the first patch is placed in the middle of a 3x3 grid
3. The second patch is sampled from its 8 neighboring locations around it.
4. Predict which one of 8 neighboring locations the second patch is selected from, a classification problem over 8 classes.



Unsupervised Visual Representation Learning by Context Prediction **(Context Prediction) - [Doersch et al., 2015]**

- To avoid the model catching low-level trivial signals:

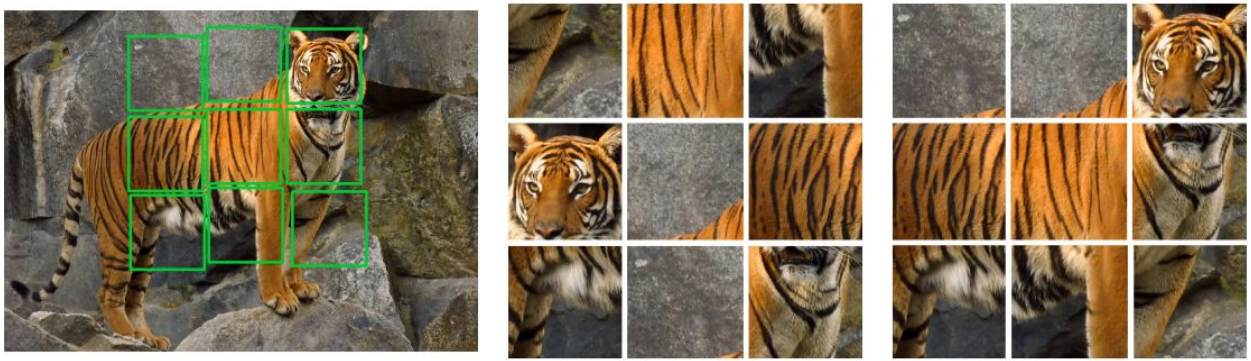
1. Add gaps between patches
2. Small jitters
3. Randomly downsample and upsampling some patches
4. other additional noises



Unsupervised learning of visual representations by solving jigsaw puzzles **(Jigsaw) - [Noroozi and Favaro, 2016]**

- Train a network to place 9 shuffled patches back to the original locations.

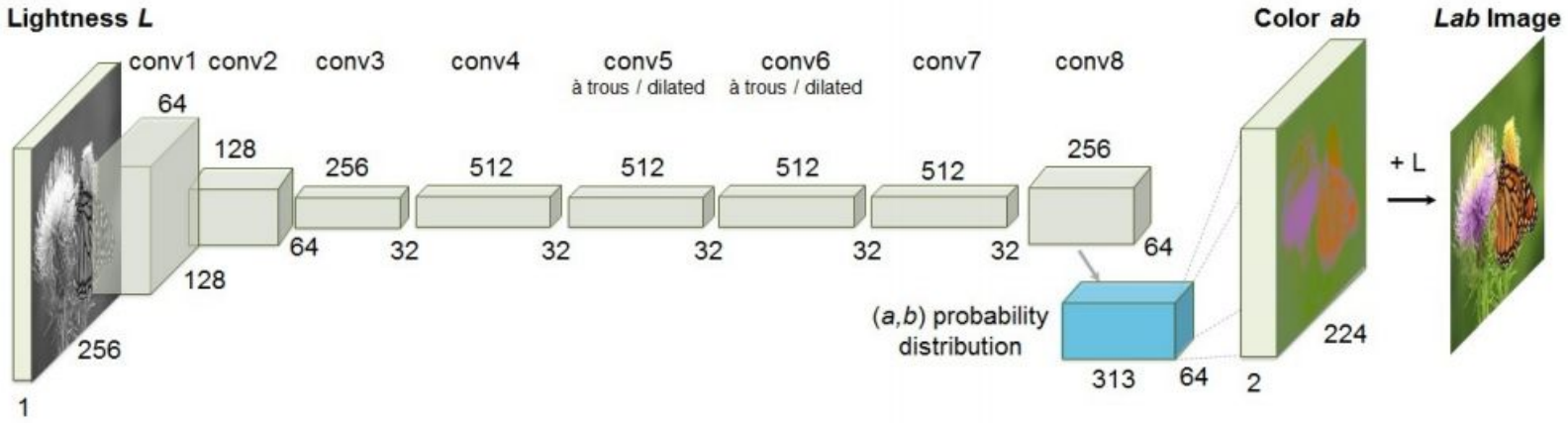
1. define a set of jigsaw puzzle permutations and assign an index to each entry
2. randomly pick one such permutation
3. rearrange the 9 input patches according to that permutation
4. ask the network to return a vector with the probability value for each index



(a) (b) (c)

Colorful image colorization **(Colorization) - [Zhang et al., 2016]**

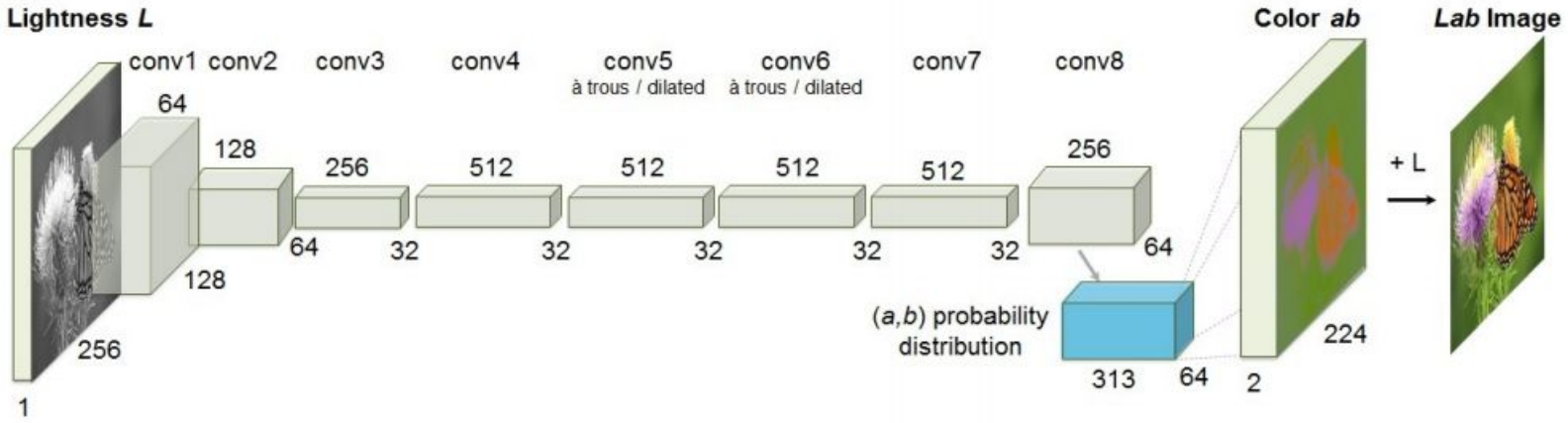
- A model is trained to color a grayscale input image
  - precisely the task is to map this image to a distribution over quantized color value outputs
  - the model outputs colors in the the Lab color space.
  - given the channel L = perceptual lightness
  - the model outputs a and b = red, green, blue, and yellow



Colorful image colorization
(Colorization) - [Zhang et al., 2016]

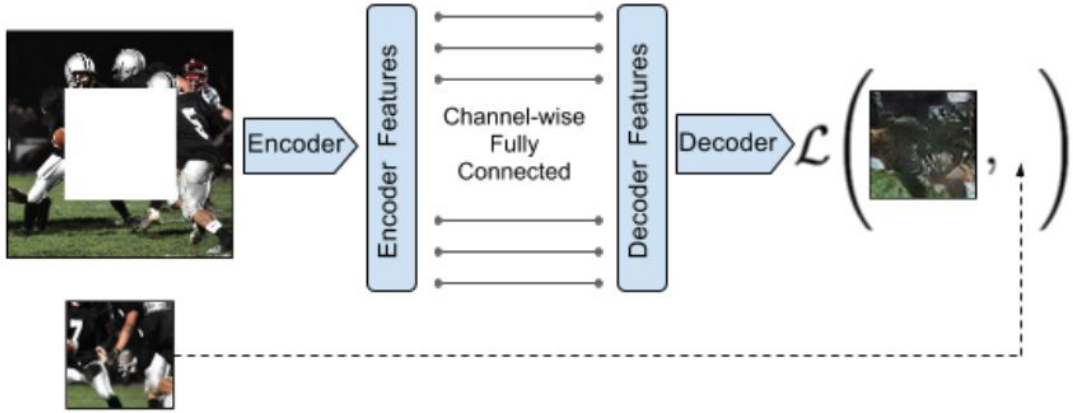
Observations:

1. The distribution of ab values in natural images is strongly biased towards low ab values
2. Class imbalance problem



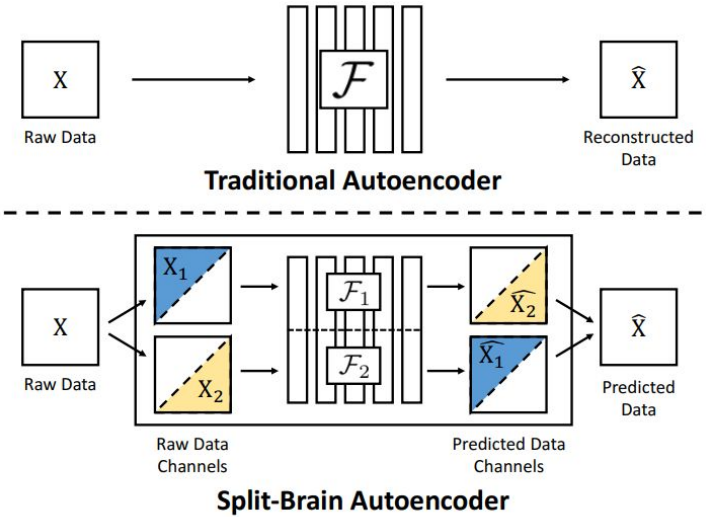
Context encoders: Feature learning by inpainting
(Context encoders) - [Pathak et al., 2016]

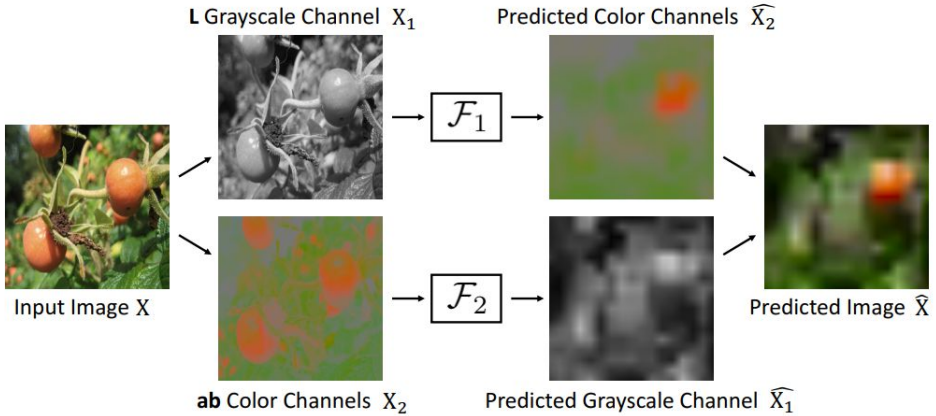
- Propose an in-painting task, where a model is trained to generate a missing region that was arbitrarily removed from the image
  - the model is trained with a combination of a standard pixel-wise reconstruction loss and an adversarial loss



Split-brain autoencoders: Unsupervised learning by cross-channel prediction **(Split-brain) - [Zhang et al., 2017]**

- They use an autoencoder with a split in the architecture, resulting in two separate sub-networks. Each sub-network is trained to predict a subset of the channels of an image in relation to the other channels.



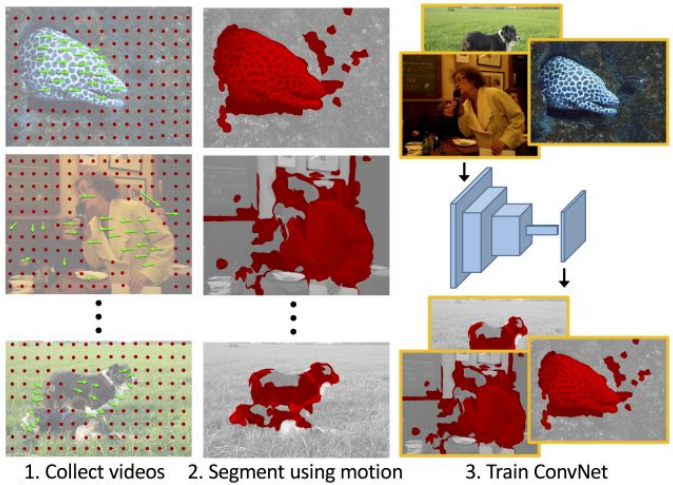


Learning features by watching objects move
[Pathak et al., 2017]

- Explore the task of segmentation using motion on videos.

1. obtain the supervisory signal by segmenting an image into foreground and background considering the optical flow between neighboring frames of a video

2. train a model to predict this segmentation from a single image.



Multi-task Self-Supervised Visual Learning (2017)
[Doersch et al., 2017]

- Combining tasks - even with a naive multihead architecture - always improves performance
  - Tasks used: relative position , colorization , the “exemplar” task and motion segmentation

Pre-training	ImageNet	PASCAL	NYU
RP	59.21	66.75	**80.54**
RP+Col	66.64	68.75	79.87
RP+Ex	65.24	69.44	78.70
RP+MS	63.73	68.81	78.72
RP+Col+Ex	68.65	69.48	80.17
RP+Col+Ex+MS	**69.30**	**70.53**	79.25
INet Labels	85.10	74.17	80.06

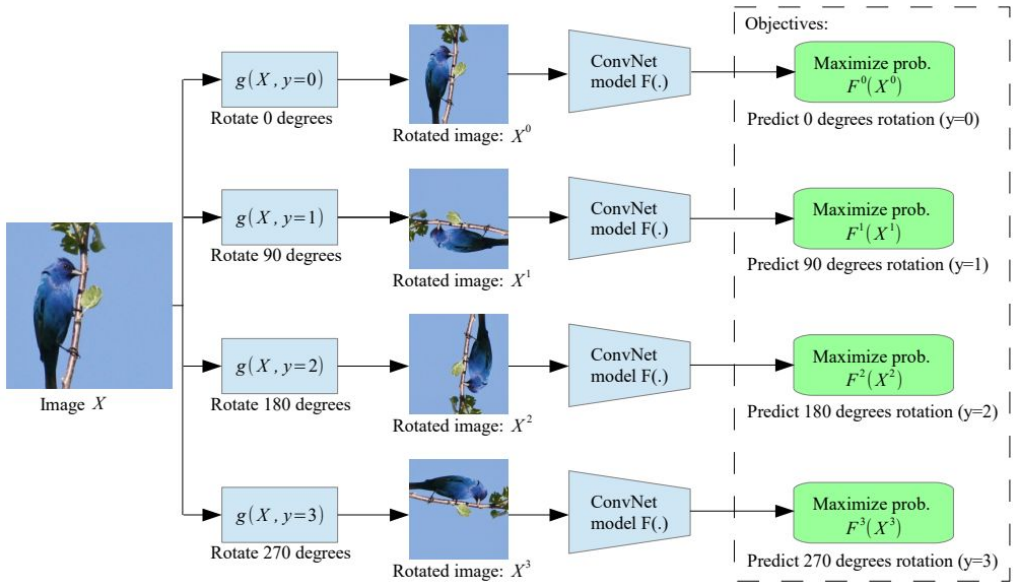
Multi-task Self-Supervised Visual Learning (2017)
[Doersch et al., 2017]

- Observations:

1. input channels can conflict (grayscale input vs all color channels)
2. learning tasks can conflict

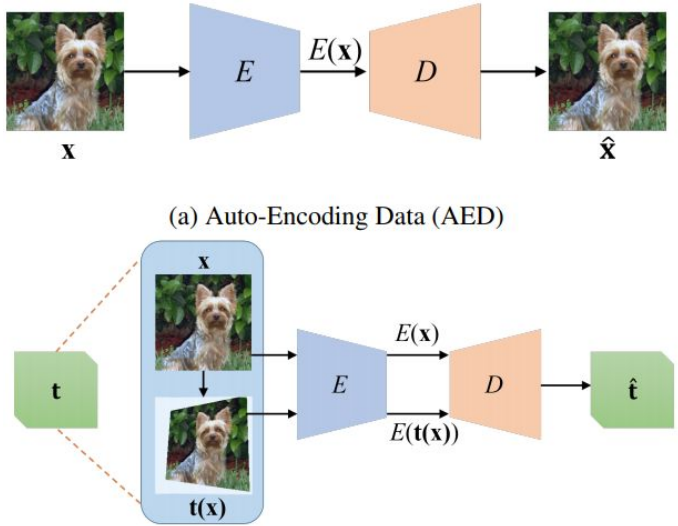
Unsupervised Representation Learning by Predicting Image Rotations **(Rotation / RotNet) - [Gidaris et al., 2018]**

- Predict image rotations
  - Rotate the image into 0º, 90º, 180º or 270º
  - the model is trained to predict which rotation has been applied (4-class classification problem)



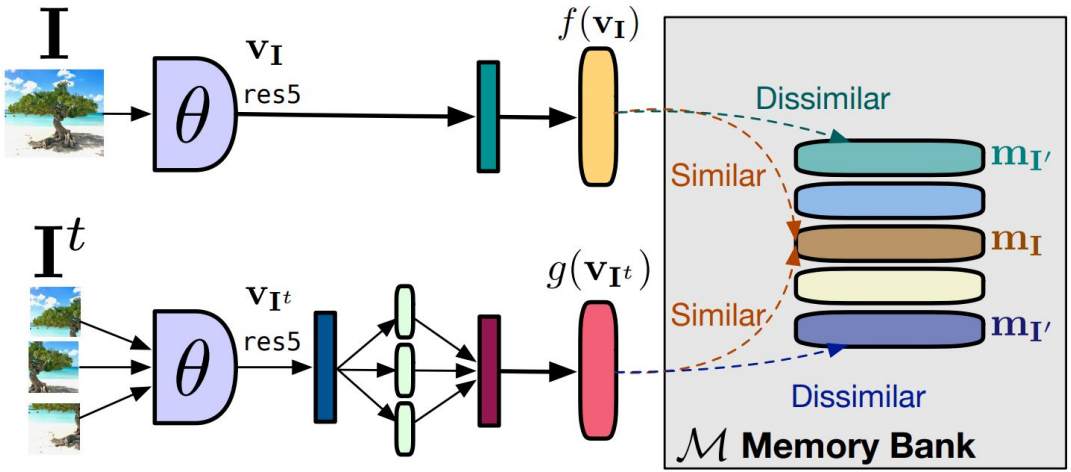
AET vs. AED: Unsupervised representation learning by auto-encoding transformations rather than data **(AET) - [Zhang et al., 2019]**

- perform transformations on the input images through some operators and train autoencoders that are able to directly estimate these operators.



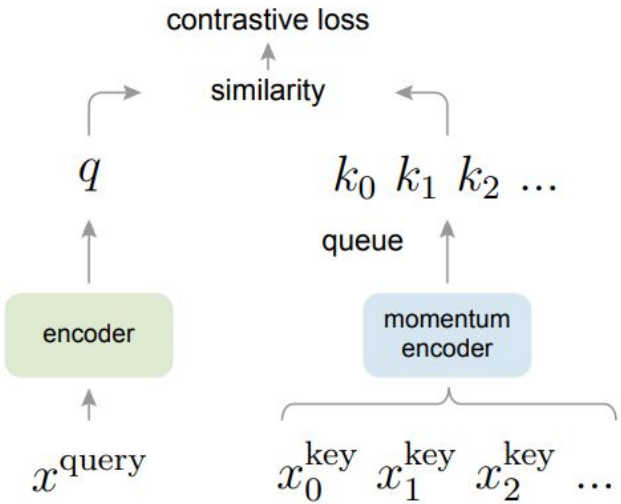
Self-Supervised Learning of Pretext-Invariant Representations (**PIRL**) - **[Misra et al., 2019]**

- Objective is to approximate the representation of the original image with its transformations (with contrastive loss). Use a memory bank of negative examples to compute the loss

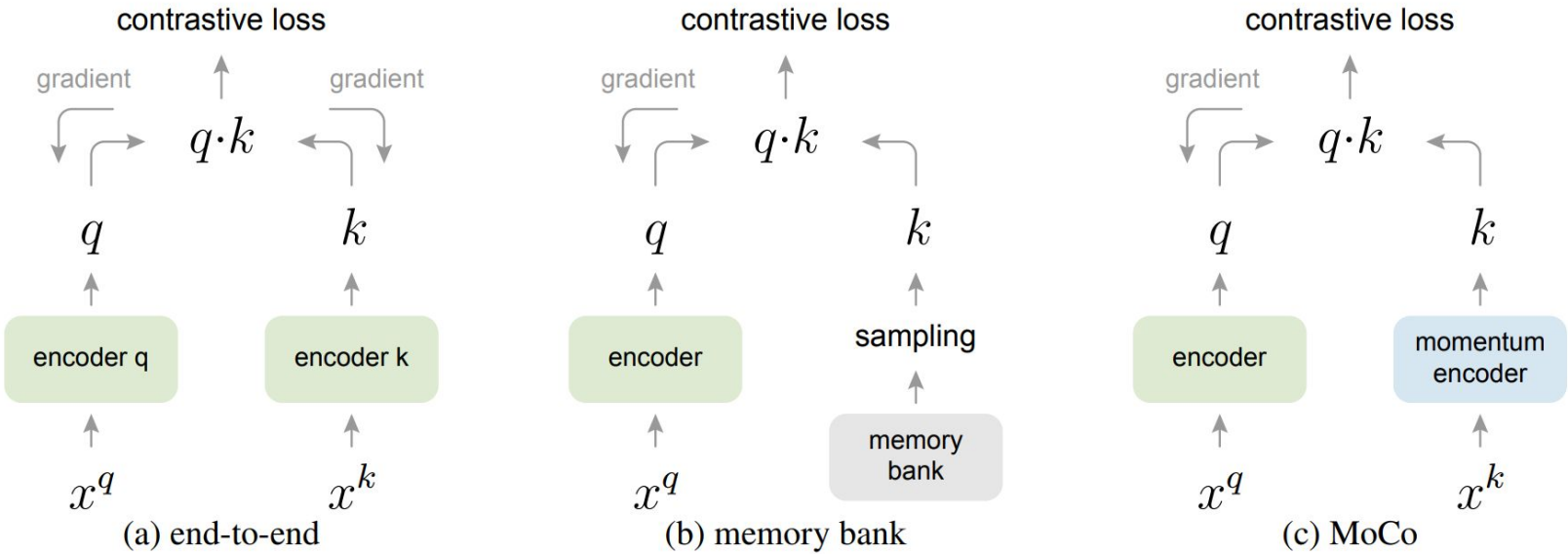


Momentum Contrast for Unsupervised Visual Representation Learning (**MOCO**) - **[He et al., 2020]**

- Views the problem of contrastive learning as a dictionary look-up problem

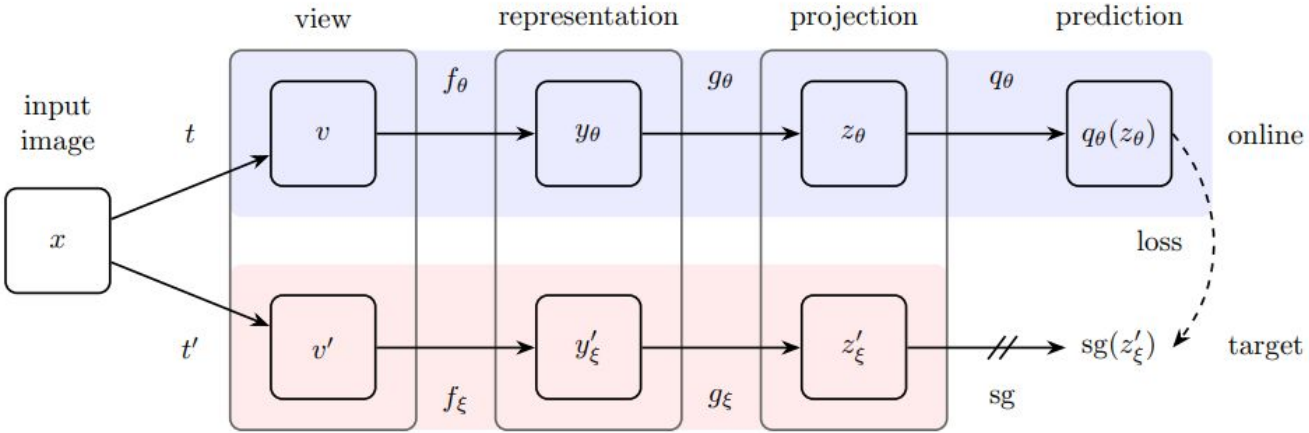


Momentum Contrast for Unsupervised Visual Representation Learning
(**MOCO**) - **[He et al., 2020]**



Bootstrap Your Own Latent A New Approach to Self-Supervised Learning **(BYOL) - [Grill et al., 2020]**

- They use a fixed randomly initialized network to serve as the key encoder



$$\mathcal{L}_{\theta,\xi} \triangleq \left\| \overline{q_\theta}(z_\theta) - \overline{z}'_\xi \right\|_2^2$$

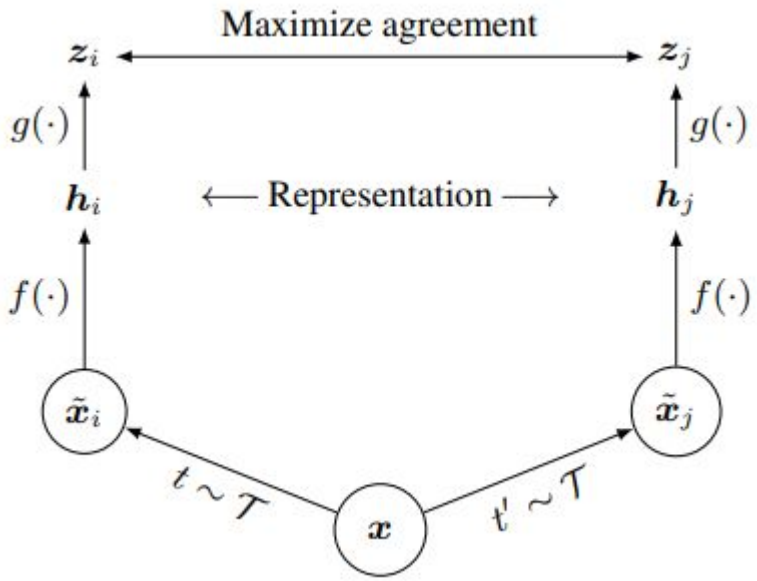
Bootstrap Your Own Latent A New Approach to Self-Supervised Learning **(BYOL) - [Grill et al., 2020]**

- Authors point out that it does not collapse because:

1. The addition of a predictor to the online network
2. The use of a slow-moving average of the online parameters as the target network encourages encoding more and more information within the online projection and avoids collapsed solutions.

A Simple Framework for Contrastive Learning of Visual Representations
(SimCLR) - [Chen et al. 2020]

- Maximize agreement between differently augmented views of the same sample via a contrastive loss in the latent space



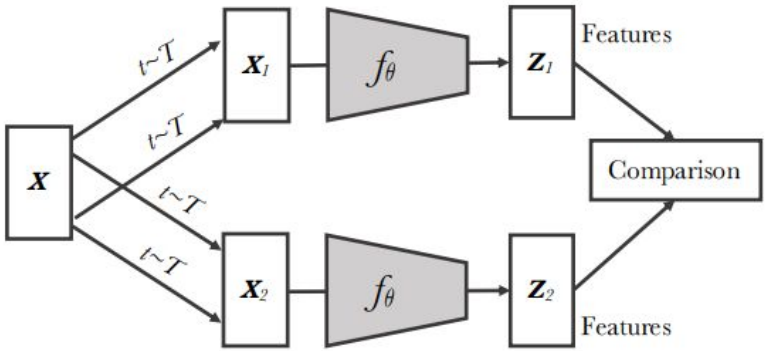
A Simple Framework for Contrastive Learning of Visual Representations **(SimCLR) - [Chen et al., 2020]**

1. randomly sample a minibatch of N examples
2. this will give us a total of 2N data points (augmented examples) derived from the minibatch
3. given a positive pair, treat the other 2(N-1) augmented examples within a mini-batch as negative examples

Note: SimCLR needs a large batch size to incorporate enough negative samples to achieve good performance (size: 4096)

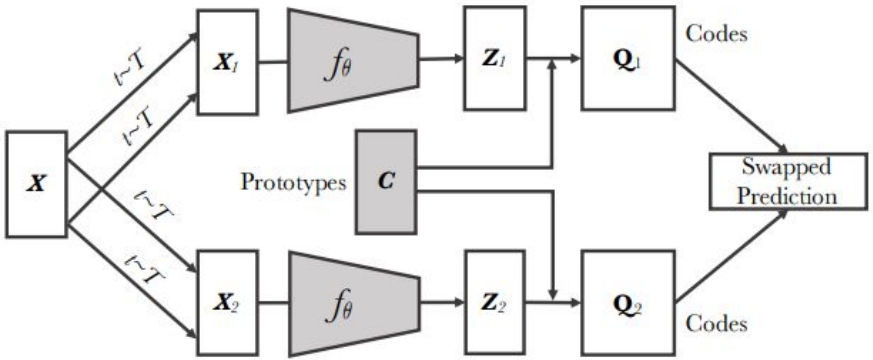
Unsupervised learning of visual features by contrasting cluster assignments **(SwAV) - [Caron et al., 2020]**

- Proposes a swapped prediction contrastive objective to deal with multi-view augmentation.



Contrastive instance learning

Unsupervised learning of visual features by contrasting cluster assignments - **(SwAV) - [Caron et al., 2020]**



Swapping Assignments between Views (Ours)

Unsupervised learning of visual features by contrasting cluster assignments -
(SwAV) - [Caron et al., 2020]

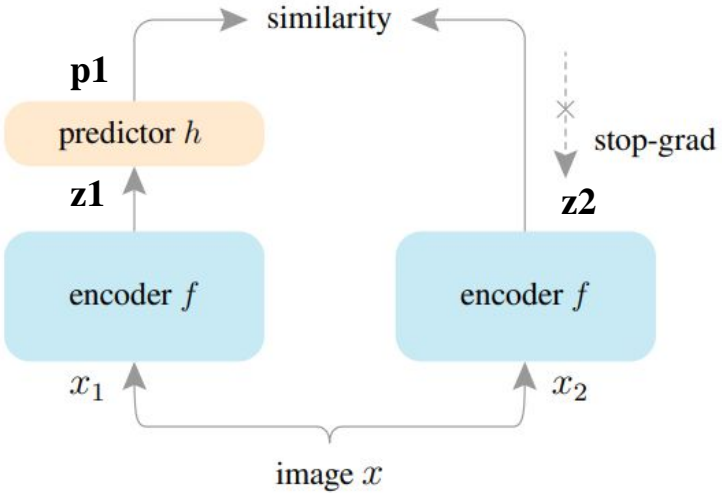
Unsupervised learning of visual features by contrasting cluster assignments - **(SwAV) - [Caron et al., 2020]**

- The big idea behind this work:

1. Typical clustering-based methods are offline
2. Here they compute the codes online allowing this method to scale to potentially unlimited amounts of data.

Exploring Simple Siamese Representation Learning **(SimSiam) - [Chen et al., 2020]**

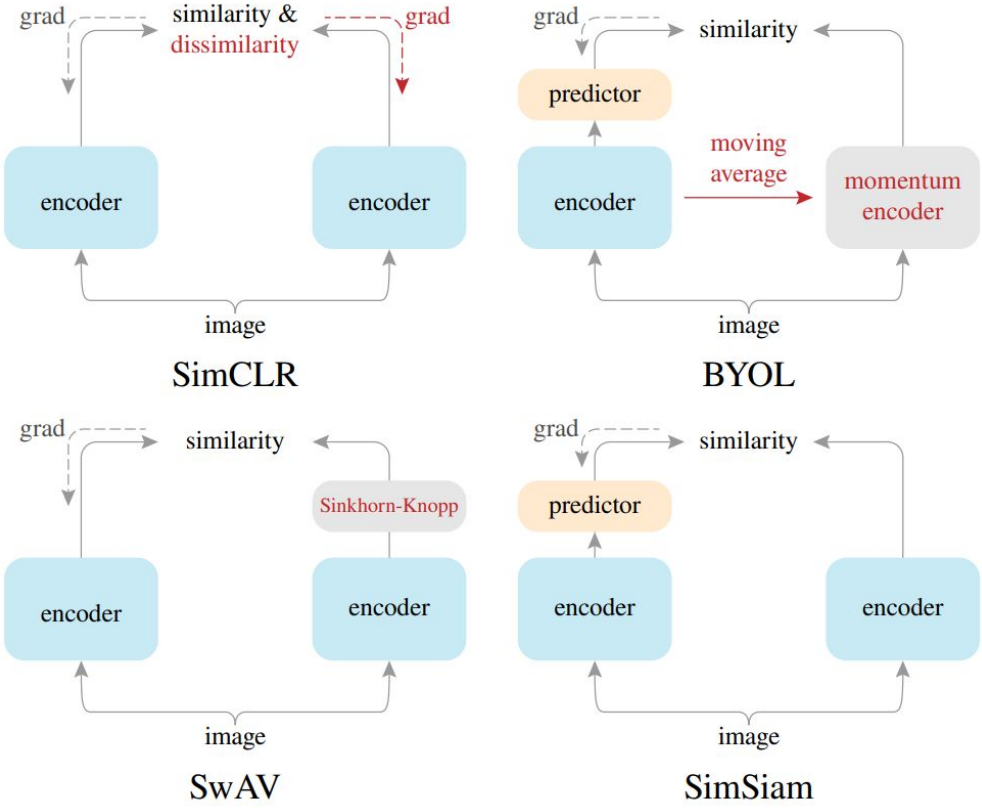
1. Two augmented views of one image are processed by the same encoder network f
2. Then a prediction MLP h is applied on one side, and a stop-gradient operation is applied on the other side.
3. The model maximizes the similarity between both sides.



$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2},$$

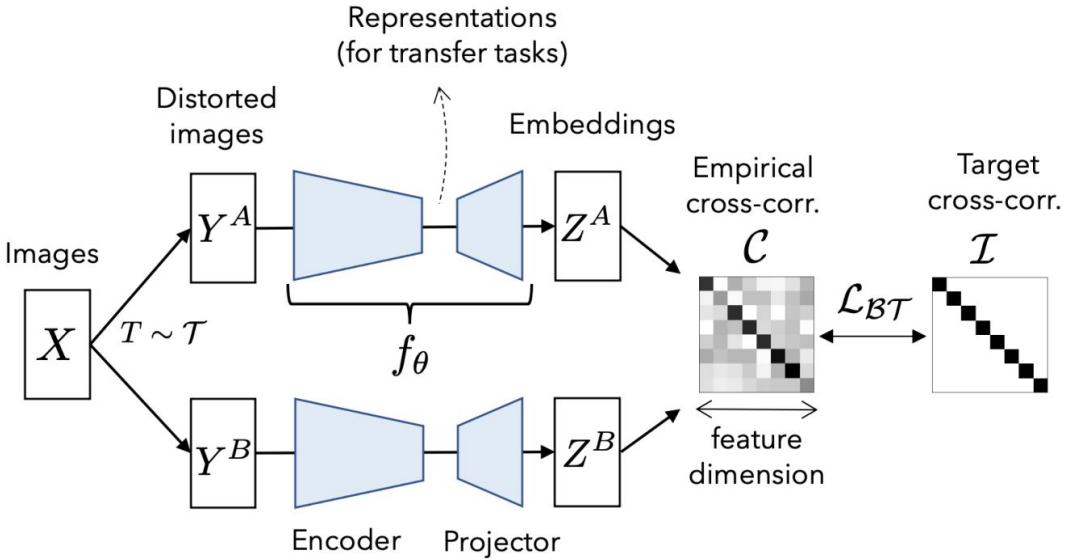
$$\mathcal{L} = \frac{1}{2}\mathcal{D}(p_1, \texttt{stopgrad}(z_2)) + \frac{1}{2}\mathcal{D}(p_2, \texttt{stopgrad}(z_1)).$$

Exploring Simple Siamese Representation Learning
(SimSiam) - [Chen et al., 2020]



Barlow Twins: Self-Supervised Learning via Redundancy Reduction (**Barlow Twins**) - **[Zbontar et al., 2021]**

1. feed two distorted versions of samples into the same network to extract features
2. compute the cross-correlation matrix between the embeddings
3. try to make this matrix close to the identity.



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

Want to learn more?

- L. Jing and Y. Tian, **"Self-supervised visual feature learning with deep neural networks: A survey,"** IEEE transactions on pattern analysis and machine intelligence, 2020

- X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, **"Self-supervised learning: Generative or contrastive,"** IEEE Transactions on Knowledge and Data Engineering, 2021

- https://github.com/jason718/awesome-self-supervised-learning

Want to learn more?

- Let’s see a toy example with the rotation-prediction task on the Fashion-MNIST dataset