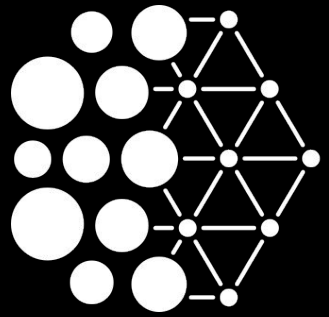
Quebec
Artificial
Intelligence
Institute



Mila

The self-supervised revolution: how to trade compute for labels? Week 2

Gaétan Marceau Caron
gaetan.marceau.caron@mila.quebec

What is a good representation?

- Properties of good representations?
  - Useful
  - Expressive
  - Robust to perturbations
  - Adapt quickly to new data
  - *Interpretable*

- Representations are learned from data in DL:
  - the task is encoded in the labels
  - require a lot a labels to get useful representations

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

What is a good representation?

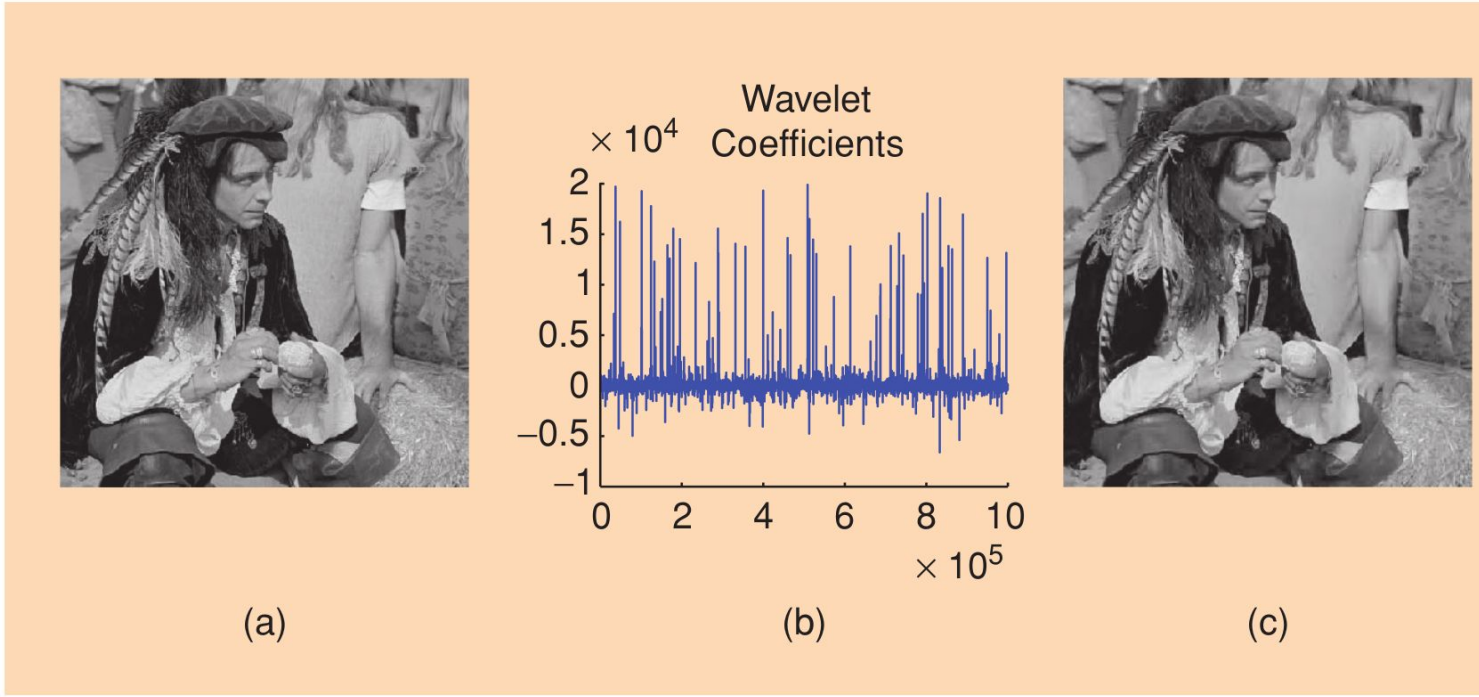


Image credit: Candès, Emmanuel J., and Michael B. Wakin. "An introduction to compressive sampling." IEEE signal processing magazine 25.2 (2008): 21-30.

What is a good representation?

- Properties of good representations?
  - Useful
  - Expressive
  - Robust to perturbations
  - Adap
  - *Inter*

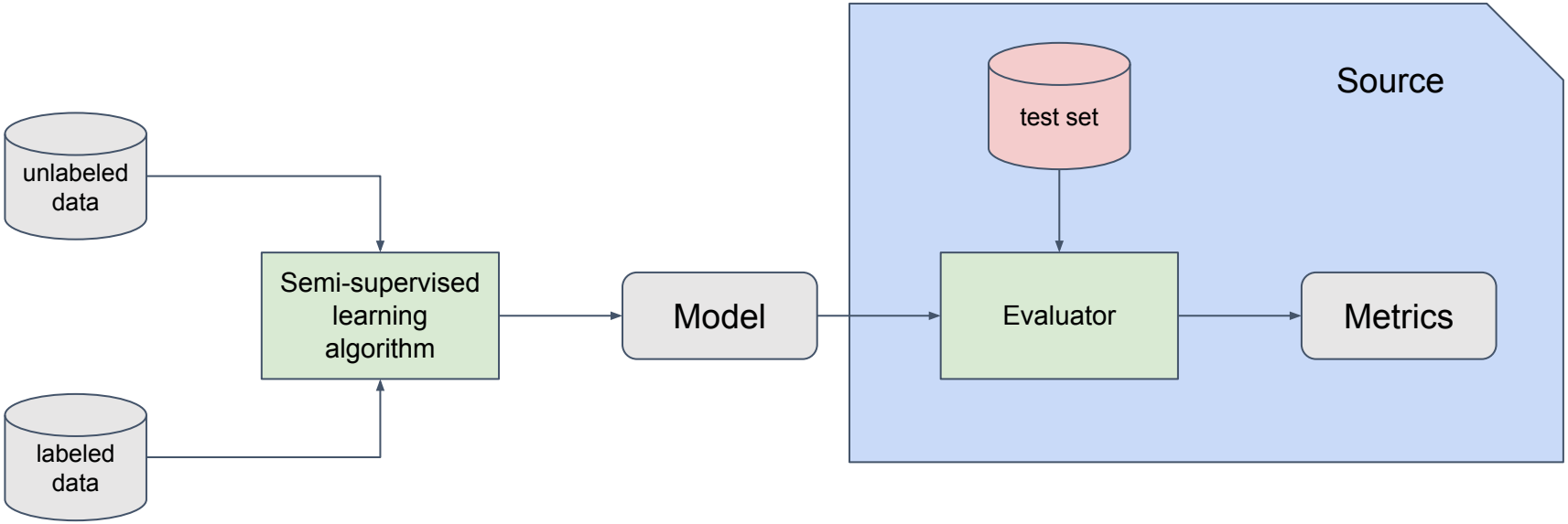
Can we learn good representations without labels?

- Representations are learned from data in DL:
  - the task is encoded in the labels
  - require a lot a labels to get useful representations

Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

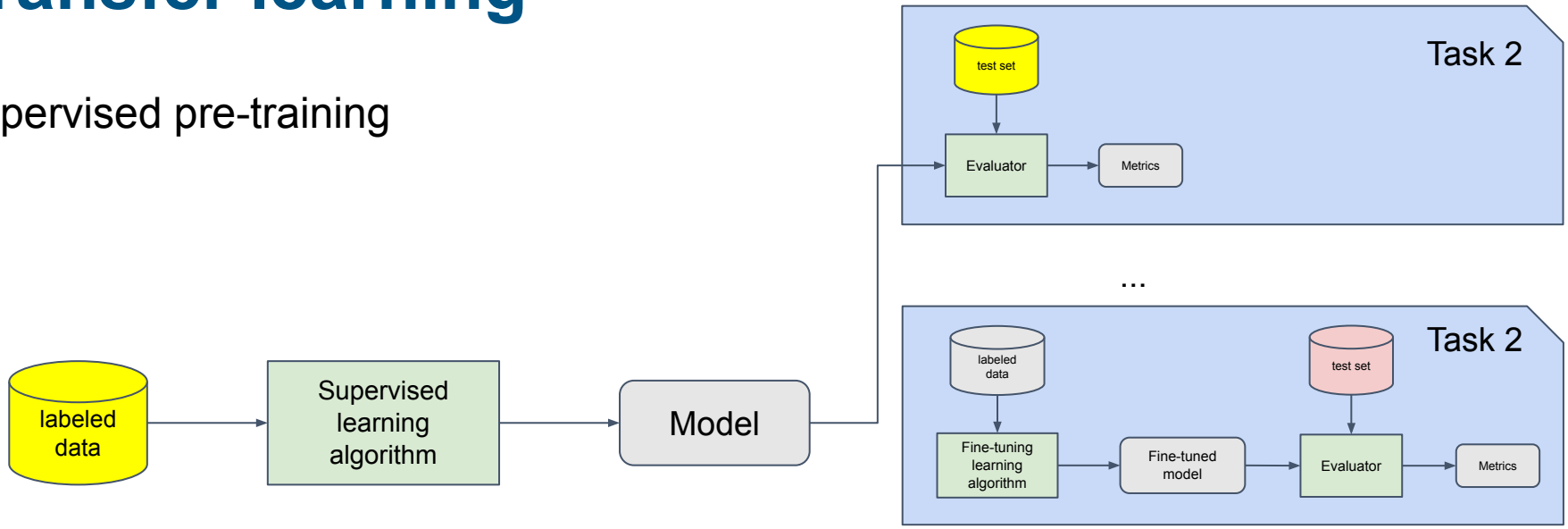
Semi-supervised learning

The downstream task is defined from the beginning



Transfer learning

Supervised pre-training



Self-supervised learning

Goal: learn to encode data with similar structures together.

Common Approaches:

1. Auto-encoding: reconstruct the datum after compressing it
2. Generative adversarial network: learn to generate “realistic” data from noise
3. Pretext tasks: create pseudo-labels from data
4. Contrastive learning: distinguish views from “the same” and “different” datum
5. Clustering: gather data that are similar together

The objectives are to:

- reduce the number of labeled examples for downstream tasks
- improve robustness to noisy labels and out-of-distribution generalization

Self-supervised learning

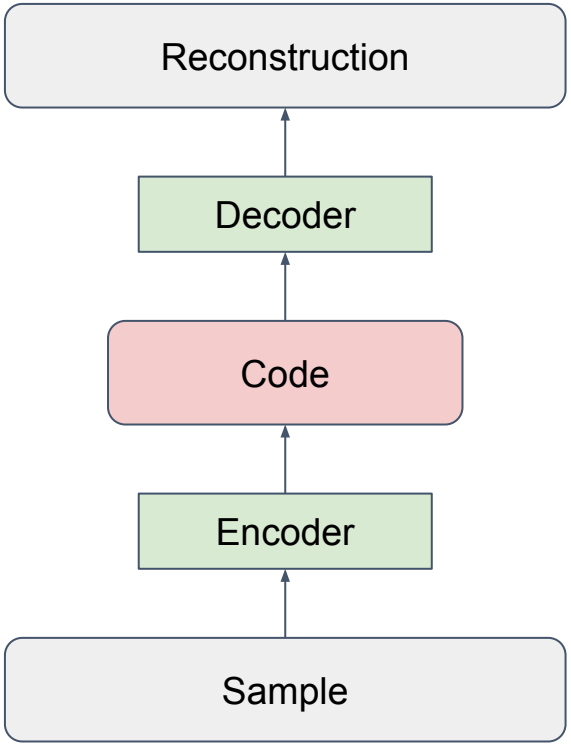
Goal: learn to encode data with similar structures together.

Common Approaches:

1. **Auto-encoding**: reconstruct the datum after compressing it
2. **Generative adversarial network**: learn to generate “realistic” data from noise
3. Pretext tasks: create pseudo-labels from data
4. Contrastive learning: distinguish views from “the same” and “different” datum
5. Clustering: gather data that are similar together

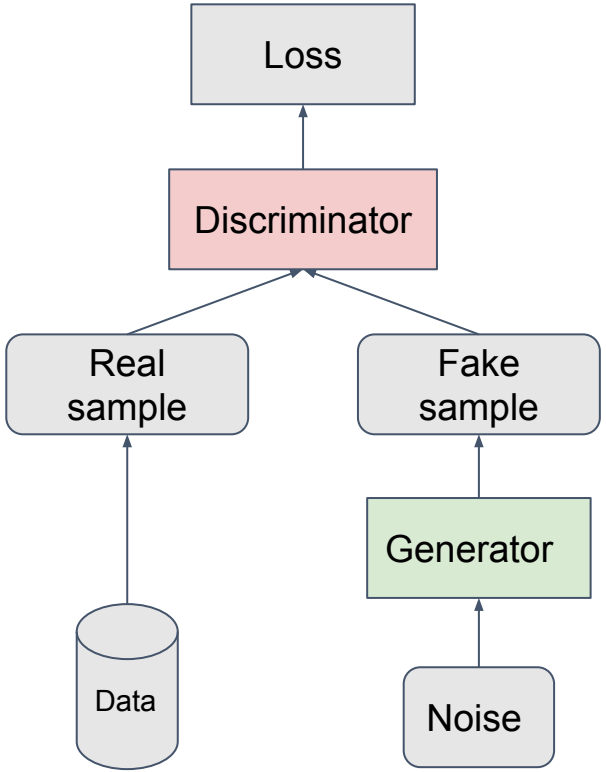
Auto-encoding

- Non-linear Principal Component Analysis
- Implementation of the manifold assumption
- Many variants:
  - Denoising AE
  - Contractive AE
  - Variational AE
  - …
- For transfer learning, a simplified version of denoising AE (BERT) works very well in NLP



Generative adversarial network

- One of the most successful design pattern for generative model
- The generator must use all the details to foul the discriminator



Empirical evidences

- AE and GAN perform well for image denoising and image synthesis

- Representations do not transfer well to other tasks such as image classification

Figure 11. Four hand-picked examples illustrating the image quality and diversity achievable using StylegGAN2 (config F).

Image credit: Karras, Tero, et al. "Analyzing and improving the image quality of stylegan." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.

Colorization

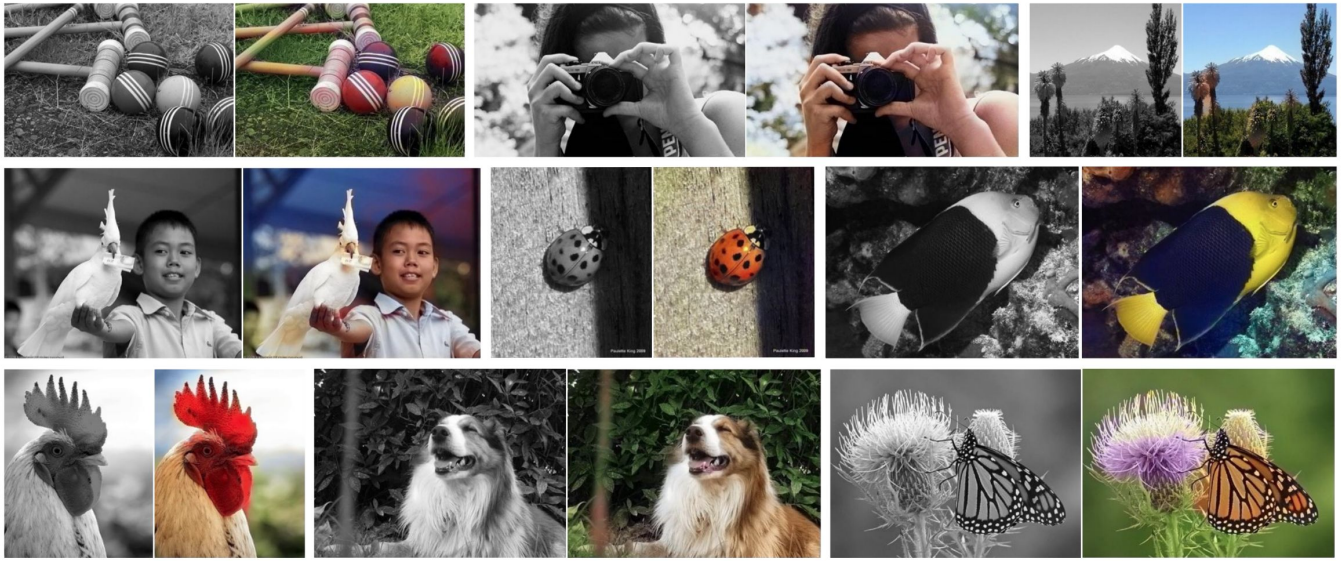
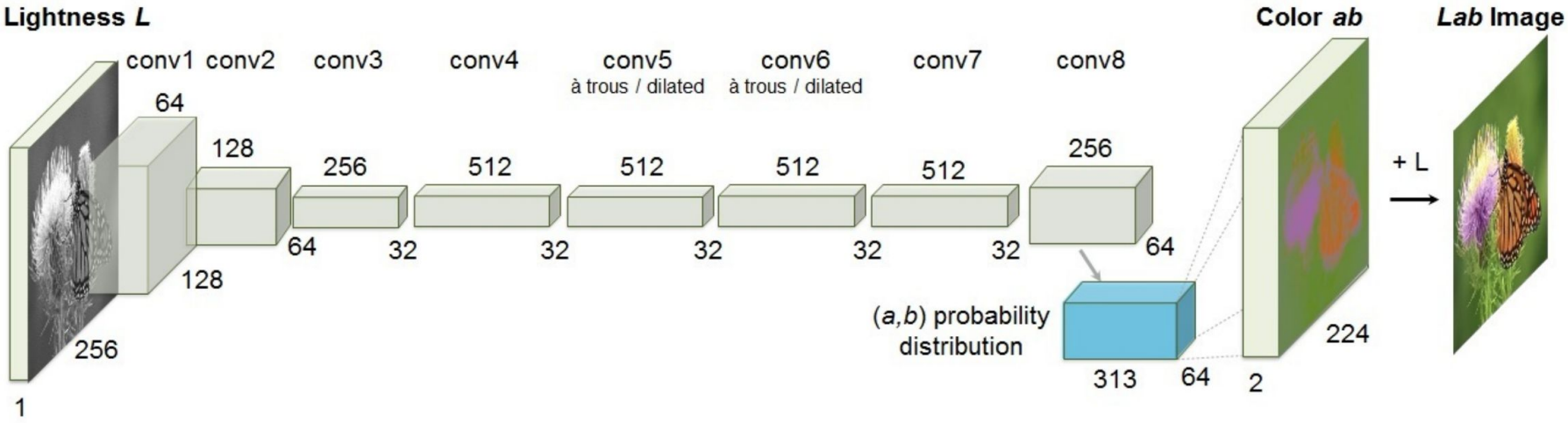


Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).

Image credit: Richard Zhang, Phillip Isola, Alexei A. Efros: Colorful Image Colorization. ECCV (3) 2016: 649-666

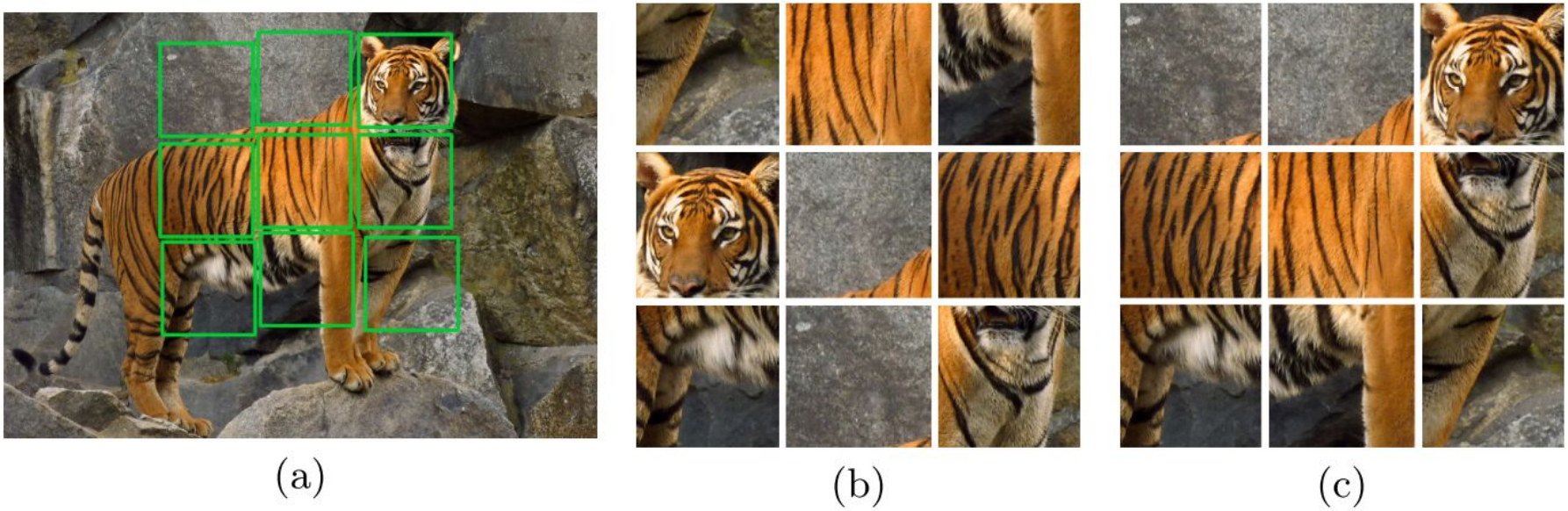
Colorization



Richard Zhang, Phillip Isola, Alexei A. Efros: Colorful Image Colorization. ECCV (3) 2016: 649-666

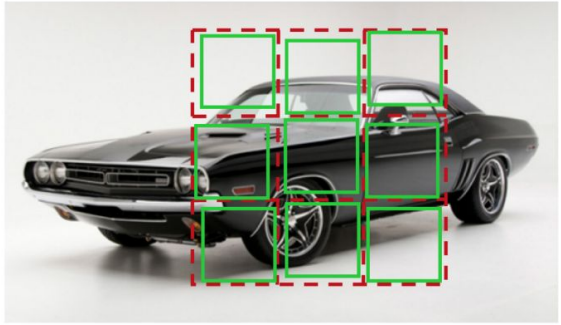
Mila

Solving Jigsaw puzzles



Mehdi Noroozi, Paolo Favaro: Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV (6) 2016: 69-84

Solving Jigsaw puzzles



Permutation Set

index	permutation
64	9,4,6,8,3,2,5,1,7

Reorder patches according to the selected permutation

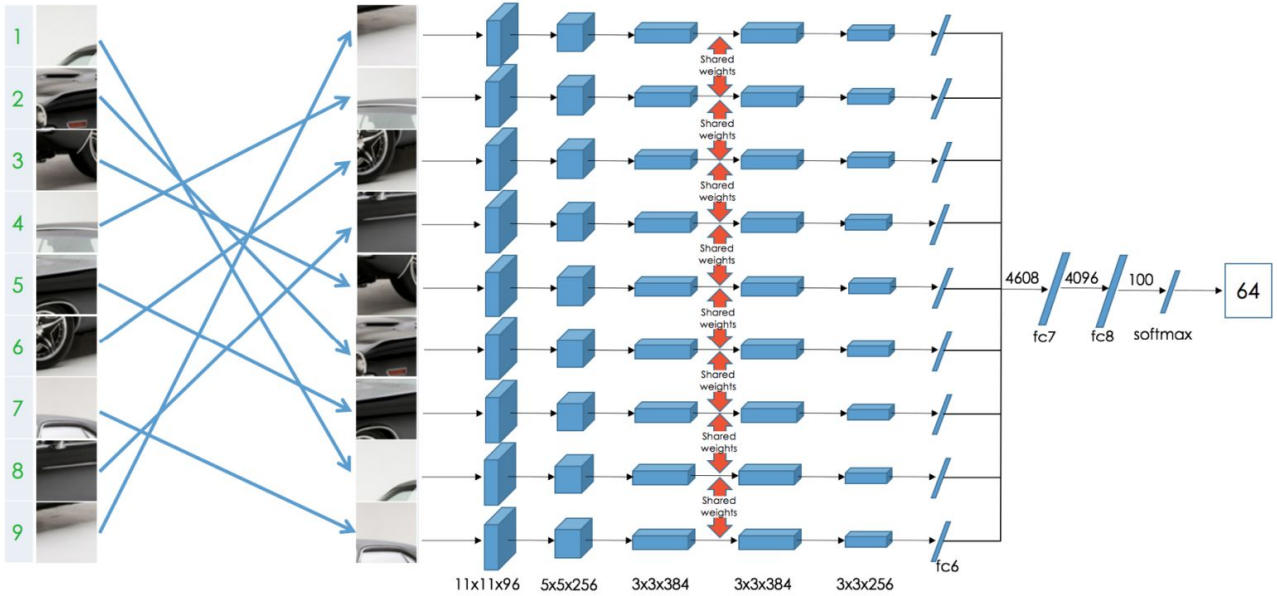
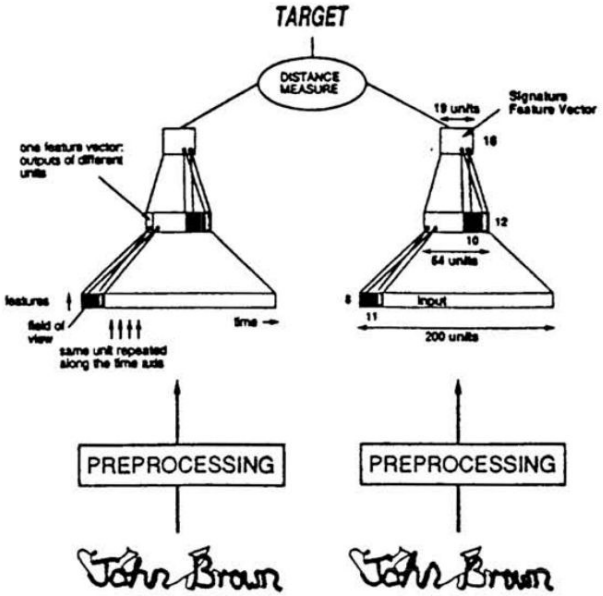


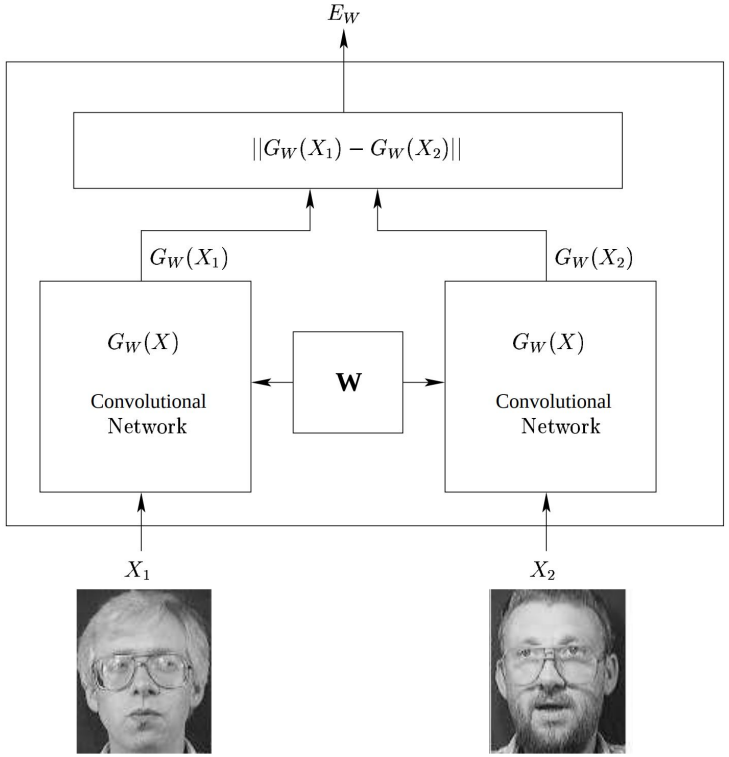
Image credit: Mehdi Noroozi, Paolo Favaro: Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles. ECCV (6) 2016: 69-84

Limitations of pretext tasks

- Many pretext tasks were proposed in 2016-2017
- Each required tricks to make them work
- Performances in transfer learning were not convincing
- Also, these tricks were not considered fundamental to learning representations
- Progress in transfer learning for NLP were based on a simpler idea:
  (masked) language modeling

Revisiting siamese network





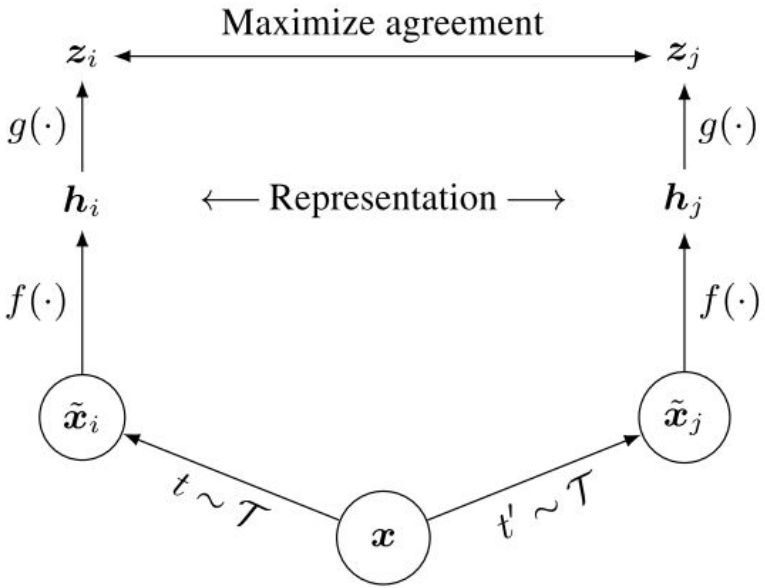
Impostor

Genuine

[1]Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, Roopak Shah: Signature Verification Using a Siamese Time Delay Neural Network. NIPS 1993
[2]Chopra, Sumit, Raia Hadsell, and Yann LeCun. "Learning a similarity metric discriminatively, with application to face verification." 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05). Vol. 1. IEEE, 2005.

Framework for self-supervised learning

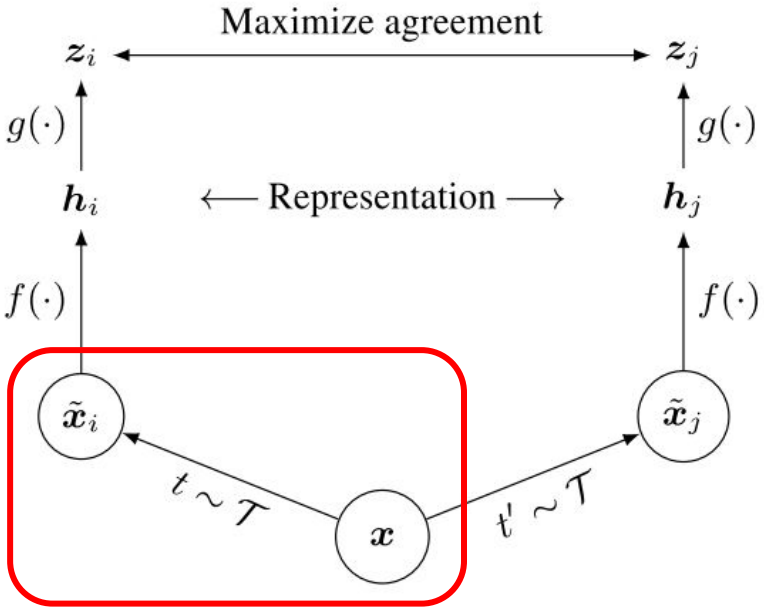
- Data augmentation pipeline
- Encoder architecture
- Similarity measure
- Loss function
- Representation extraction



Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).

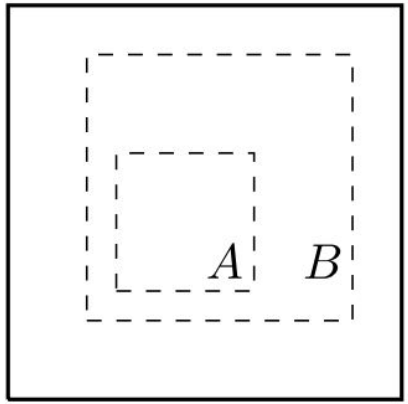
Data augmentation pipeline

Data augmentation pipeline

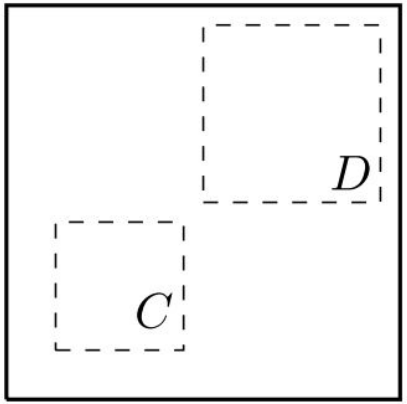


Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Data augmentation pipeline



(a) Global and local views.



(b) Adjacent views.

Figure 3. Solid rectangles are images, dashed rectangles are random crops. By randomly cropping images, we sample contrastive prediction tasks that include global to local view ($B \rightarrow A$) or adjacent view ($D \rightarrow C$) prediction.

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

[2] Caron, Mathilde, et al. "Unsupervised learning of visual features by contrasting cluster assignments." arXiv preprint arXiv:2006.09882 (2020).

Data augmentation pipeline

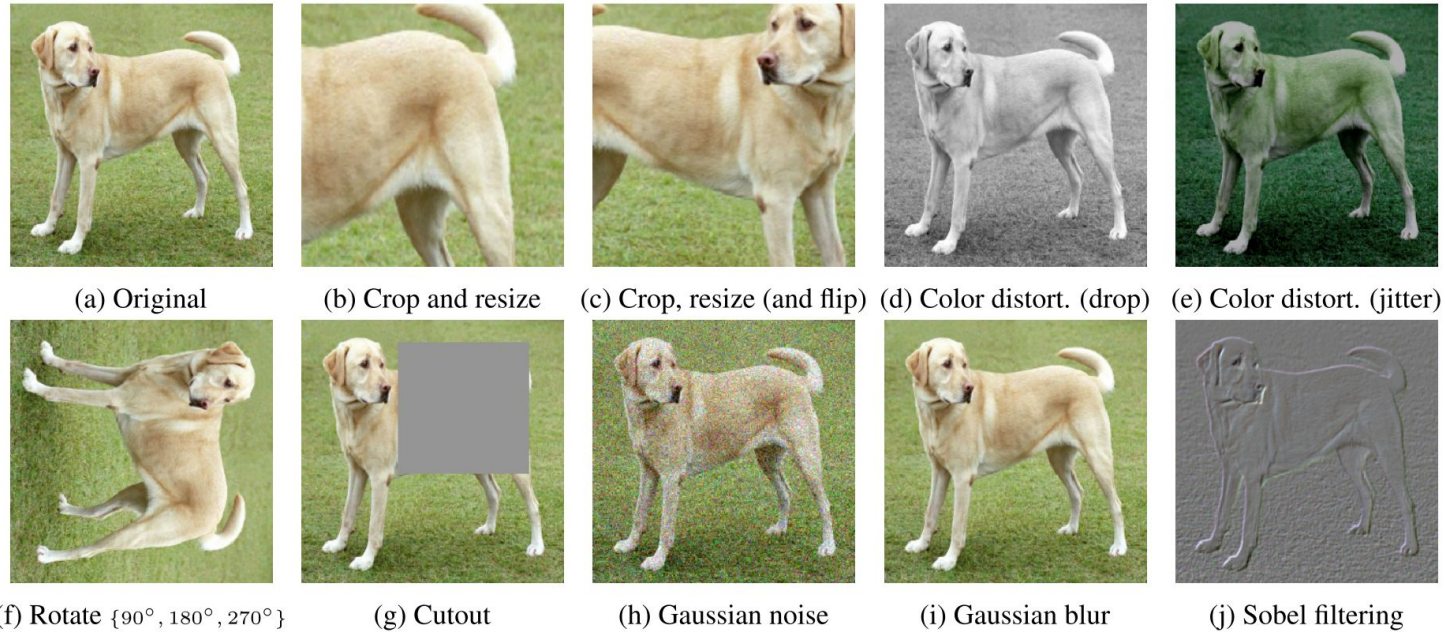
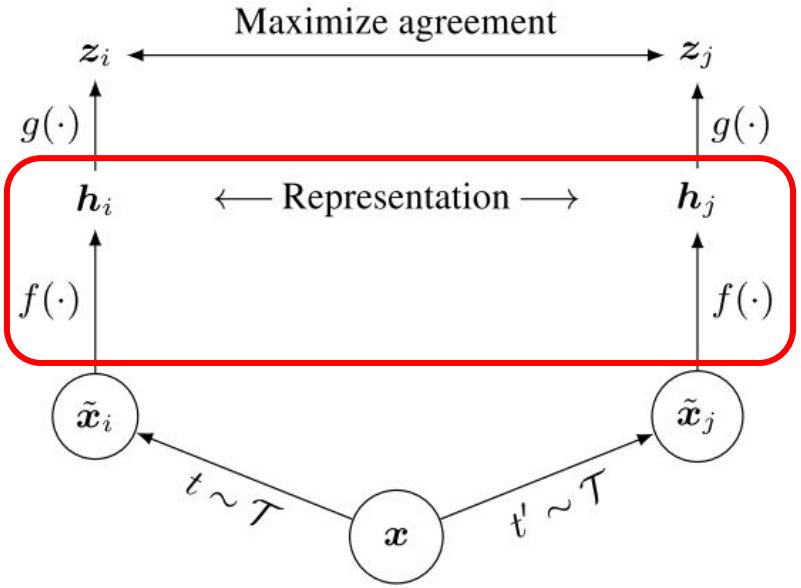


Figure 4. Illustrations of the studied data augmentation operators. Each augmentation can transform data stochastically with some internal parameters (e.g. rotation degree, noise level). Note that we *only* test these operators in ablation, the *augmentation policy used to train our models* only includes *random crop (with flip and resize)*, *color distortion*, and *Gaussian blur*. (Original image cc-by: Von.grzanka)

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

A simple framework

Backbone encoder



Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Backbone encoder

- ResNet-50 or ResNet-161

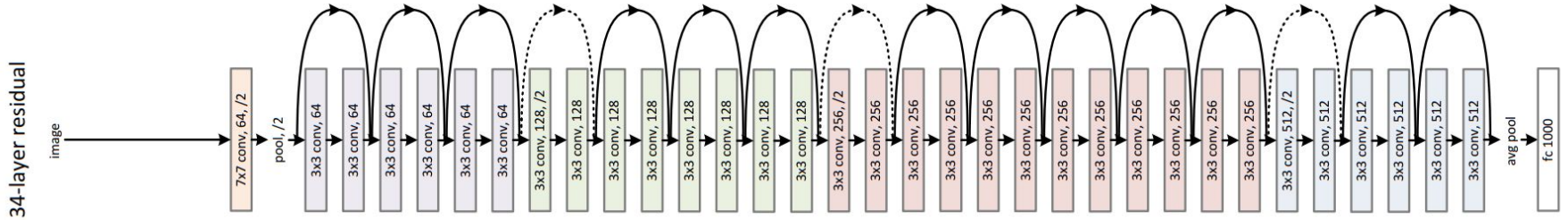
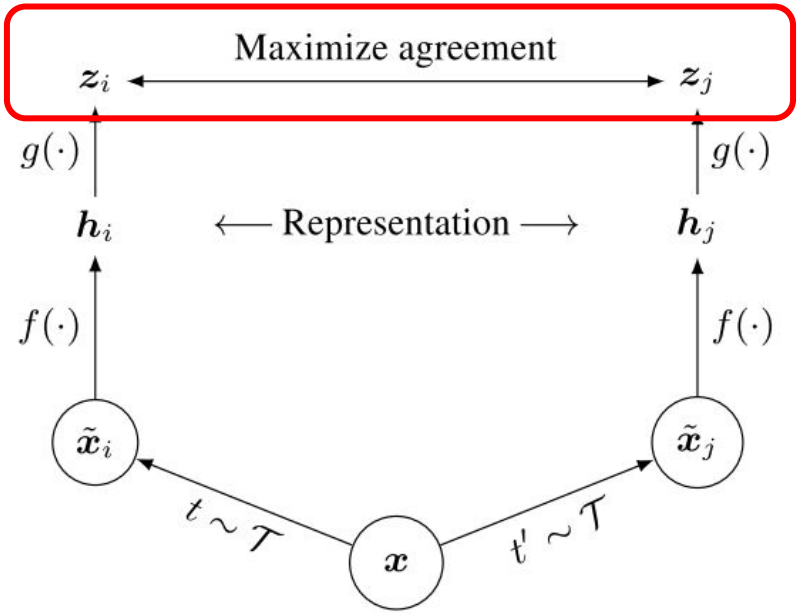


Image credit: Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun: Deep Residual Learning for Image Recognition. CVPR 2016: 770-778

A simple framework

Similarity measure



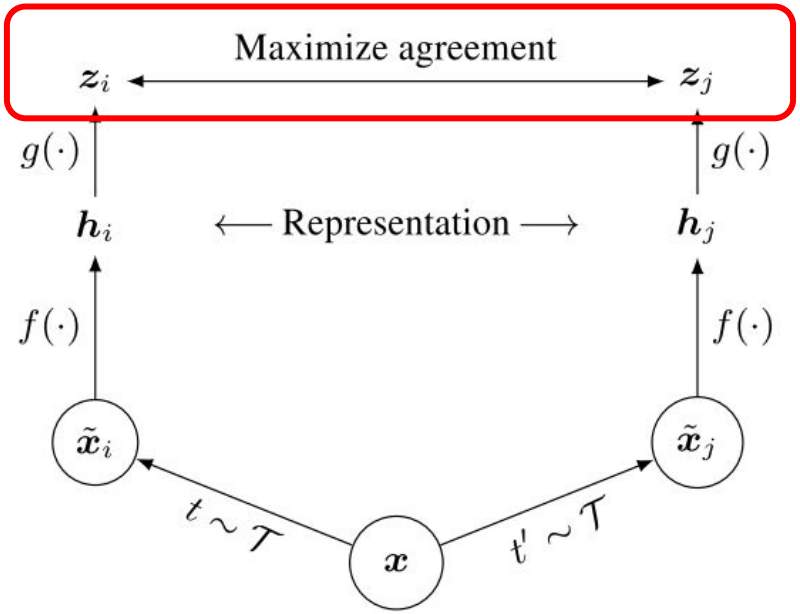
Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Similarity measure

Goal: compare two representations

- dot product
- cosine similarity

The measure can be parametrized.



Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Contrastive learning

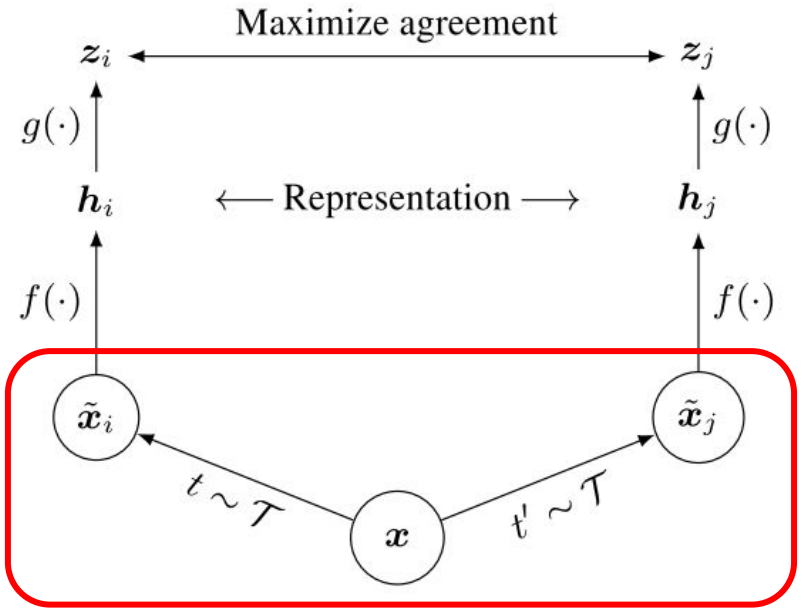
Goal: attract similar examples and repulse dissimilar examples.

Image credit: Falcon, William. "A Framework For Contrastive Self-Supervised Learning And Designing A New Approach." Towards data science (accessed: 2020/12)

Contrastive learning

Generate pairs of positive (similar) and negative (dissimilar) examples:

1. Choose an anchor
2. Generate a positive example
3. Generate a negative example



Data augmentation pipeline

Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).
Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML (2020)

Contrastive loss: NT-Xent

NT-Xent: Normalized Temperature-scaled Cross Entropy

$$\mathcal{L}_\theta(z^a, z^+, Z^-) = -\log \frac{\exp(\Phi(z^a, z^+)/\tau)}{\sum_{z_i^- \in Z^-} \exp(\Phi(z^a, z_i^-)/\tau)}$$

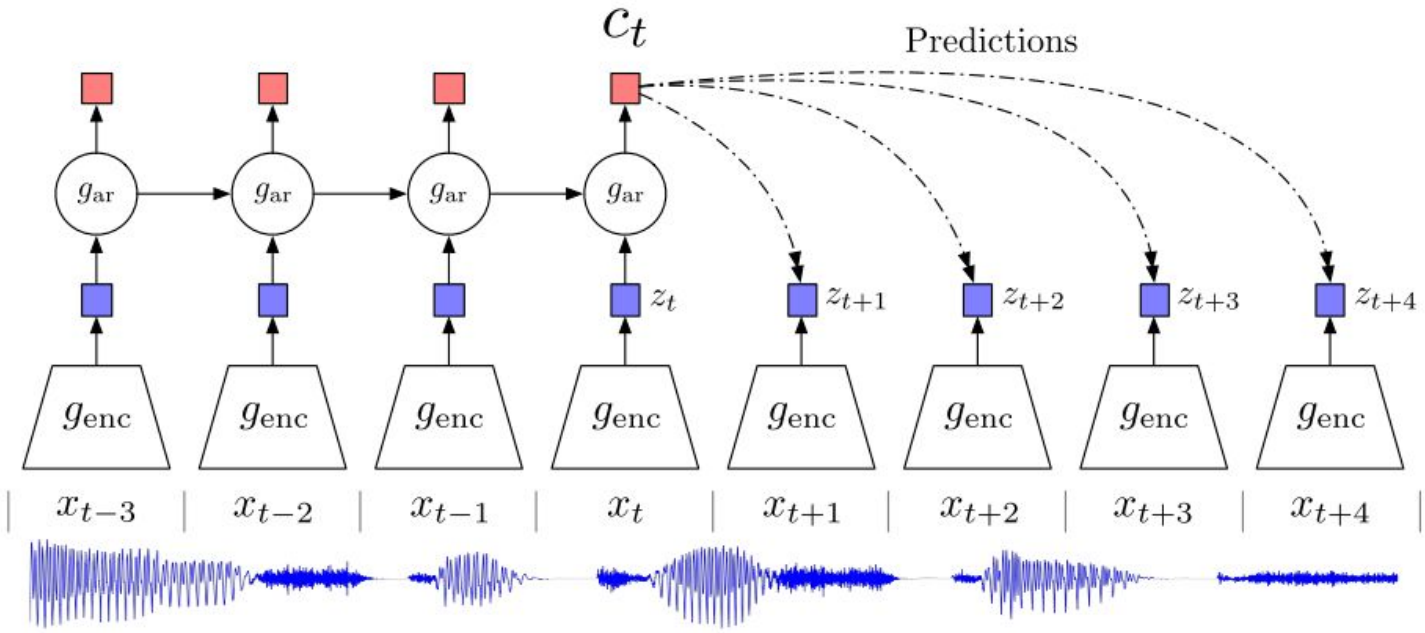
Variants and other losses exist in the litterature:

- Negative contrastive estimation
- Triplet loss
- InfoNCE

Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).
Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML (2020)

Contrastive Predictive Coding (CPC)

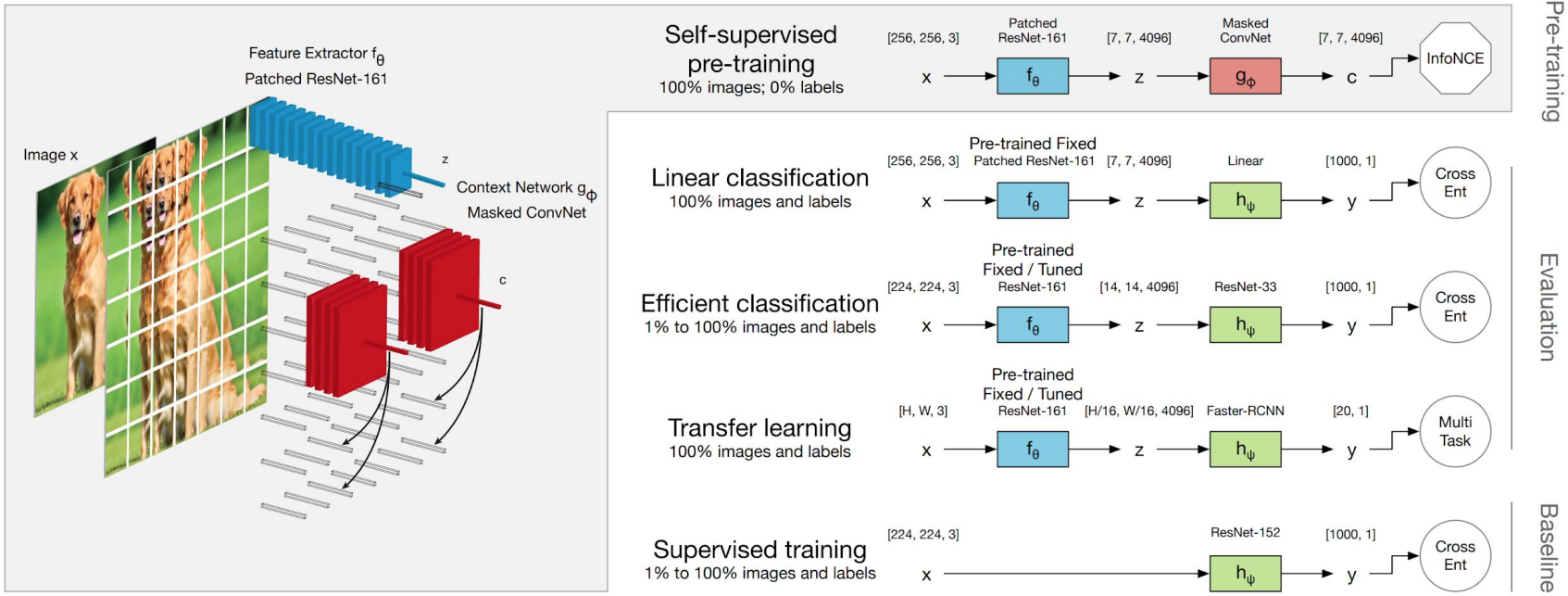
Intuition: autoregressive model with contrastive loss



Oord, Aaron van den, Yazhe Li, and Oriol Vinyals. "Representation learning with contrastive predictive coding." arXiv preprint arXiv:1807.03748 (2018).

Contrastive Predictive Coding (CPCv2)

Intuition: autoregressive model with contrastive loss



Henaff, Olivier et al.. "Data-efficient image recognition with contrastive predictive coding." International Conference on Machine Learning. PMLR (2020).

Negative examples are cumbersome

We need to:

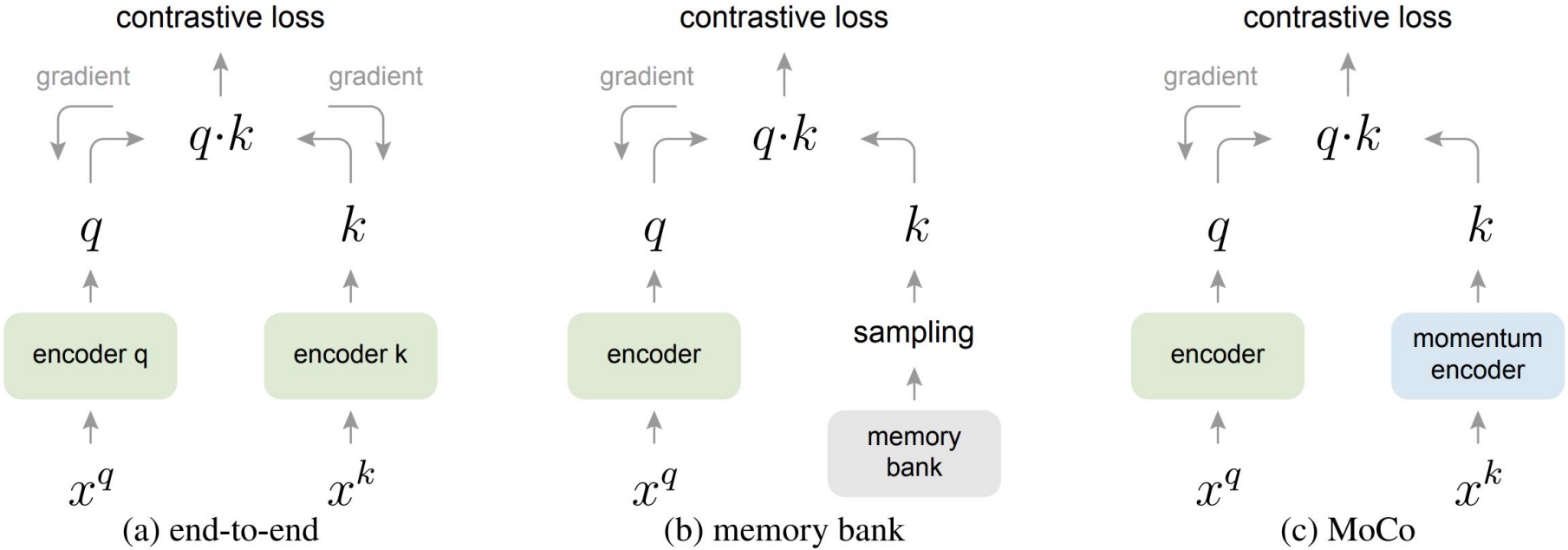
- specify how to sample negative examples
- adapt the optimization algorithm for large batch training (e.g. using LARS)

$$\mathcal{L}_\theta(z^a, z^+, Z^-) = -\log \frac{\exp(\Phi(z^a, z^+)/\tau)}{\sum_{z_i^- \in Z^-} \exp(\Phi(z^a, z_i^-)/\tau)}$$

Dictionary look-up task

Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).

MoCo



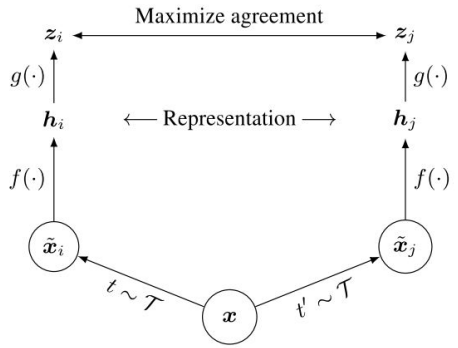
Momentum encoder

$$\theta_{\mathrm{k}} \leftarrow m\theta_{\mathrm{k}} + (1-m)\theta_{\mathrm{q}}$$

Do we really need contrastive learning?

Goal: we want to remove negative examples

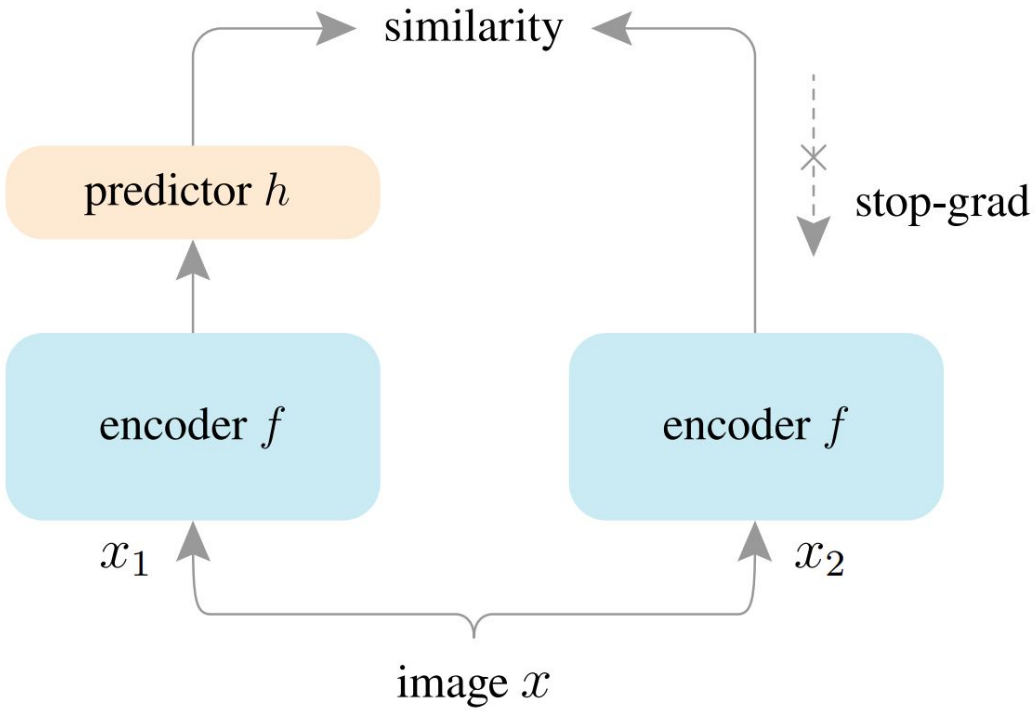
Potential problems: collapse to trivial solutions



$$\mathcal{L}_\theta(z^a, z^+, Z^-) = -\log \frac{\exp(\Phi(z^a, z^+)/\tau)}{\sum_{z_i^- \in Z^-} \exp(\Phi(z^a, z_i^-)/\tau)}$$

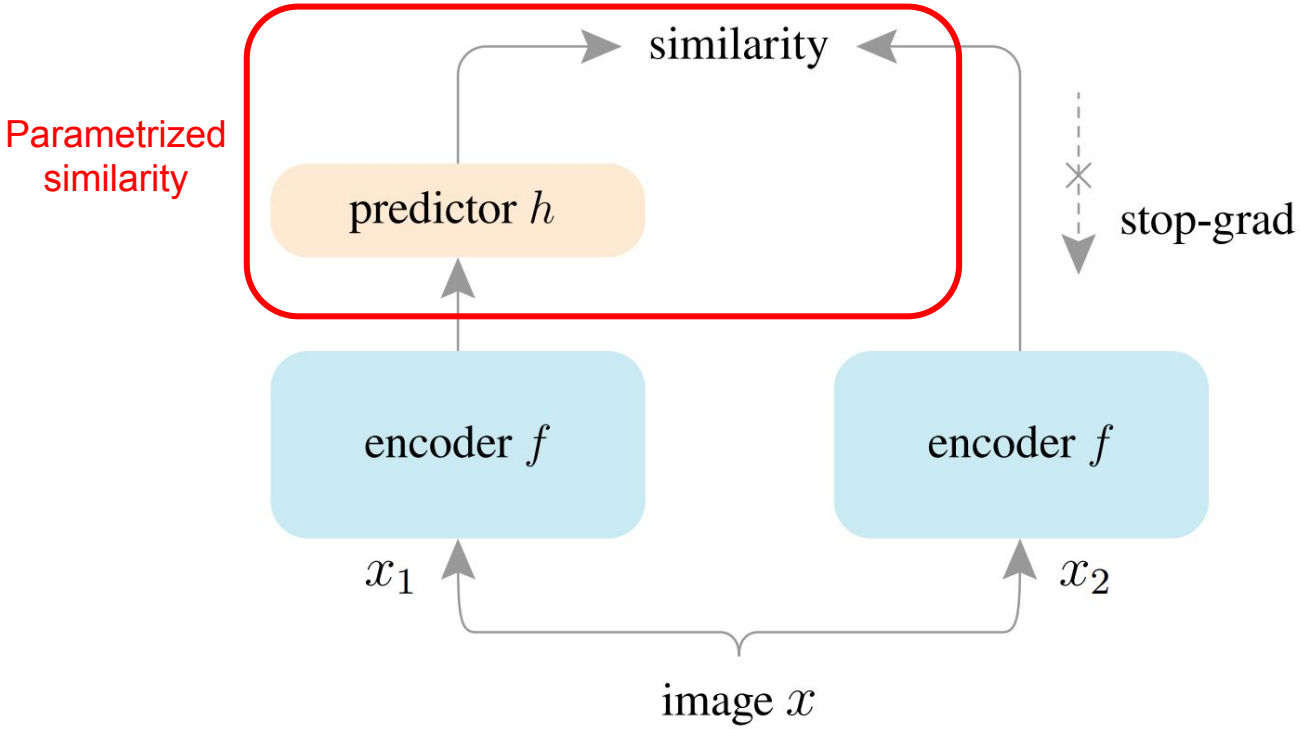
Falcon, William, and Kyunghyun Cho. "A framework for contrastive self-supervised learning and designing a new approach." arXiv:2009.00104 (2020).

Siamese network for self-supervised learning



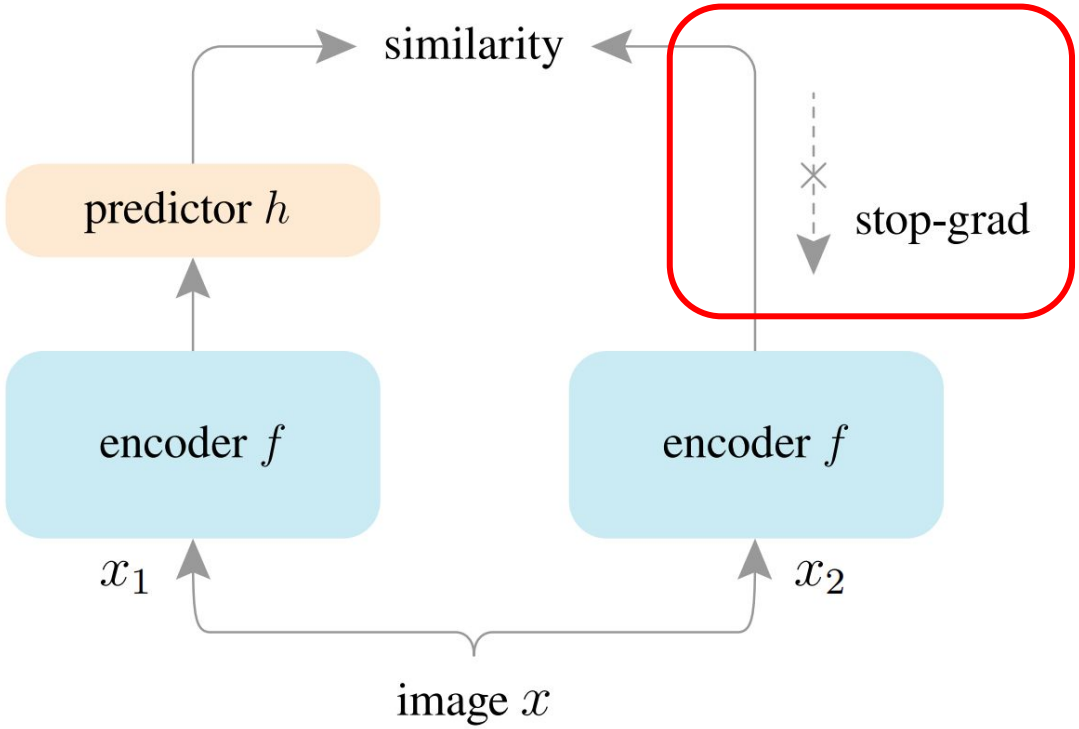
Chen, Xinlei, and Kaiming He. "Exploring Simple Siamese Representation Learning." arXiv preprint arXiv:2011.10566 (2020).

Siamese network for self-supervised learning



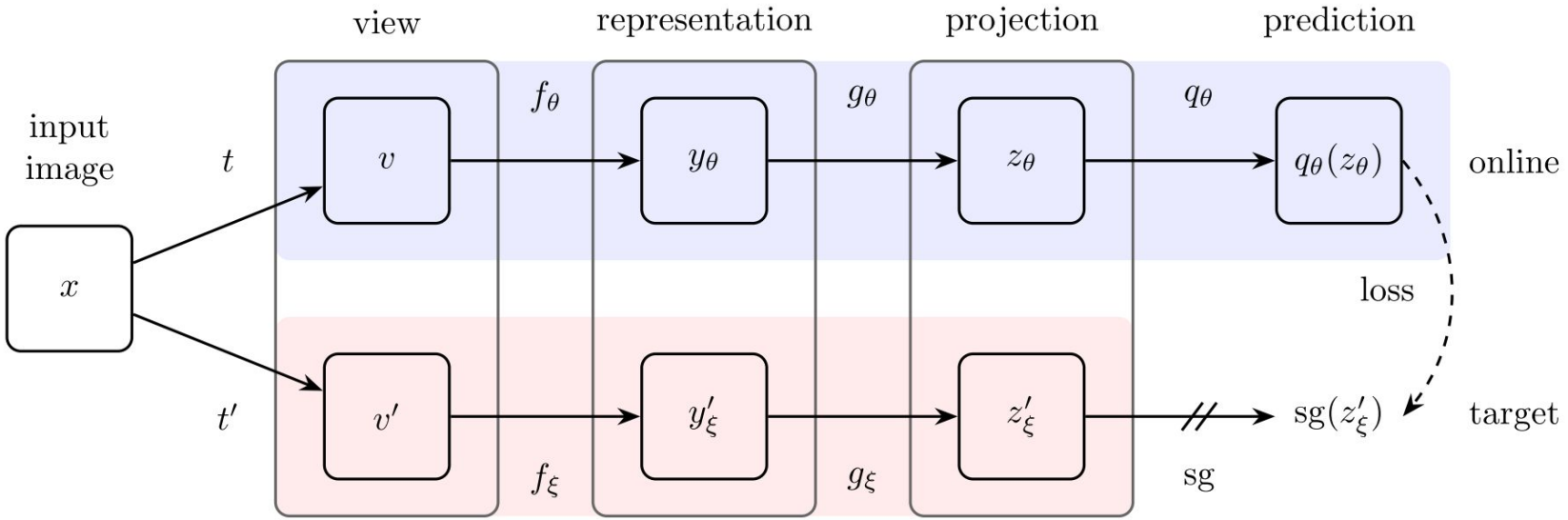
Chen, Xinlei, and Kaiming He. "Exploring Simple Siamese Representation Learning." arXiv preprint arXiv:2011.10566 (2020).

Siamese network for self-supervised learning



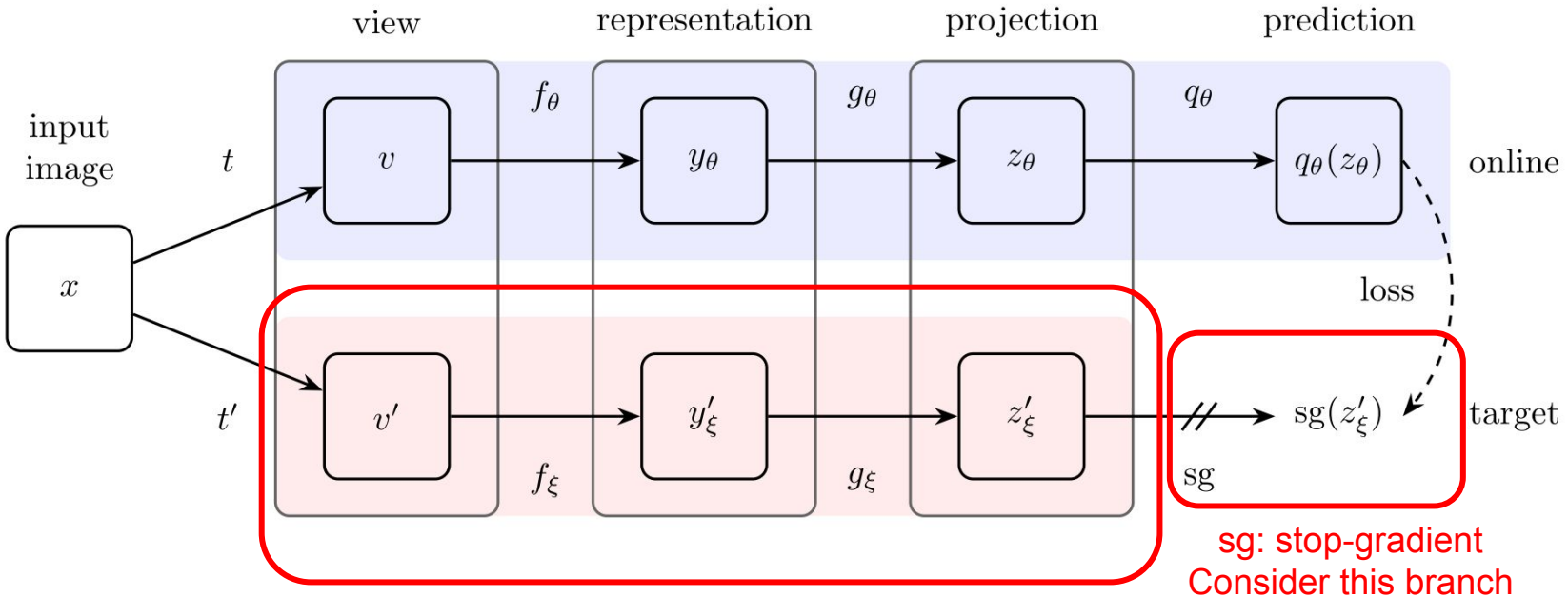
Chen, Xinlei, and Kaiming He. "Exploring Simple Siamese Representation Learning." arXiv preprint arXiv:2011.10566 (2020).

Bootstrap Your Own Latent



Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." Advances in Neural Information Processing Systems 33 (2020).

Bootstrap Your Own Latent



Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." Advances in Neural Information Processing Systems 33 (2020).

Instance classification

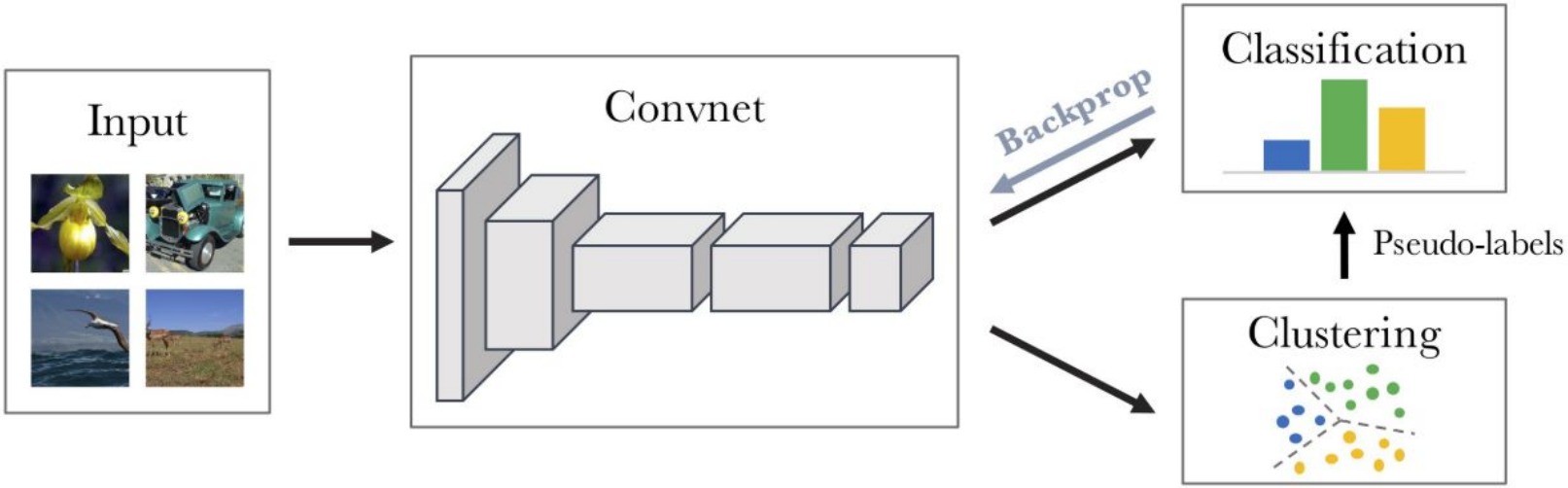
Unsupervised feature learning via non-parametric instance discrimination: use memory bank

Become intractable with dataset size

Alexey Dosovitskiy, Philipp Fischer, Jost Tobias Springenberg, Martin A. Riedmiller, Thomas Brox: Discriminative Unsupervised Feature Learning with Exemplar Convolutional Neural Networks. IEEE Trans. Pattern Anal. Mach. Intell. 38(9): 1734-1747 (2016)

Clustering as pretext task

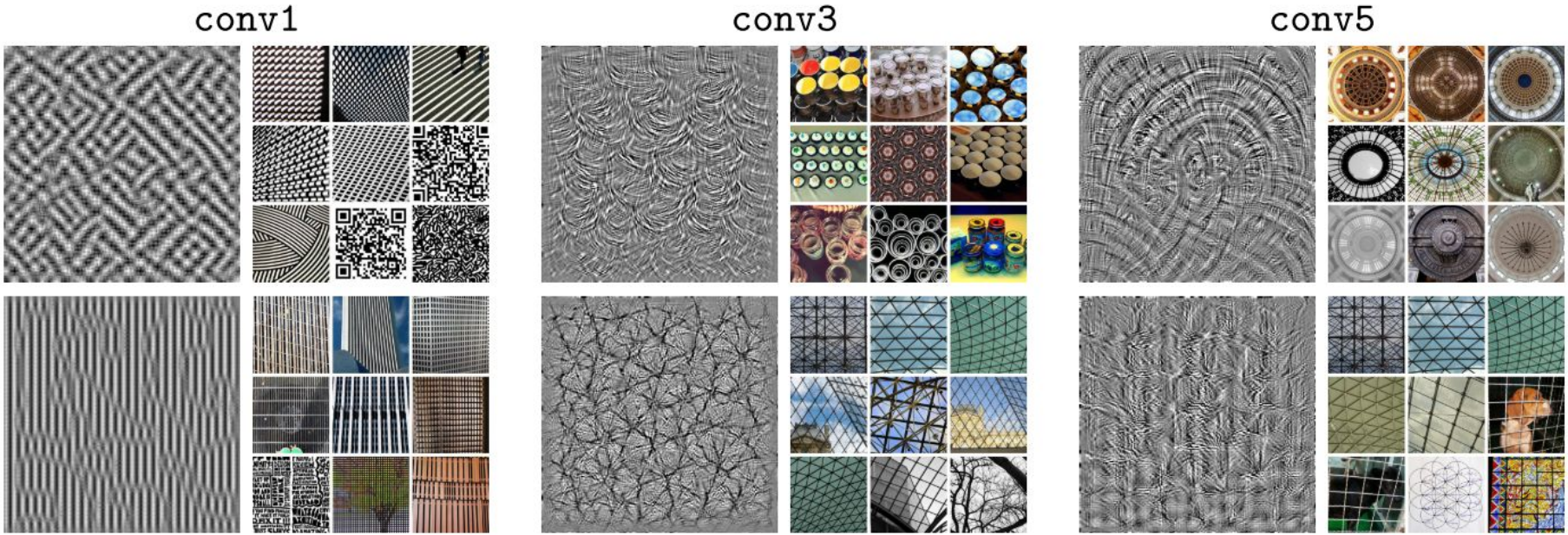
Discriminate between groups of images instead of individual images



Does not scale well with dataset size (e.g. for 1B images)

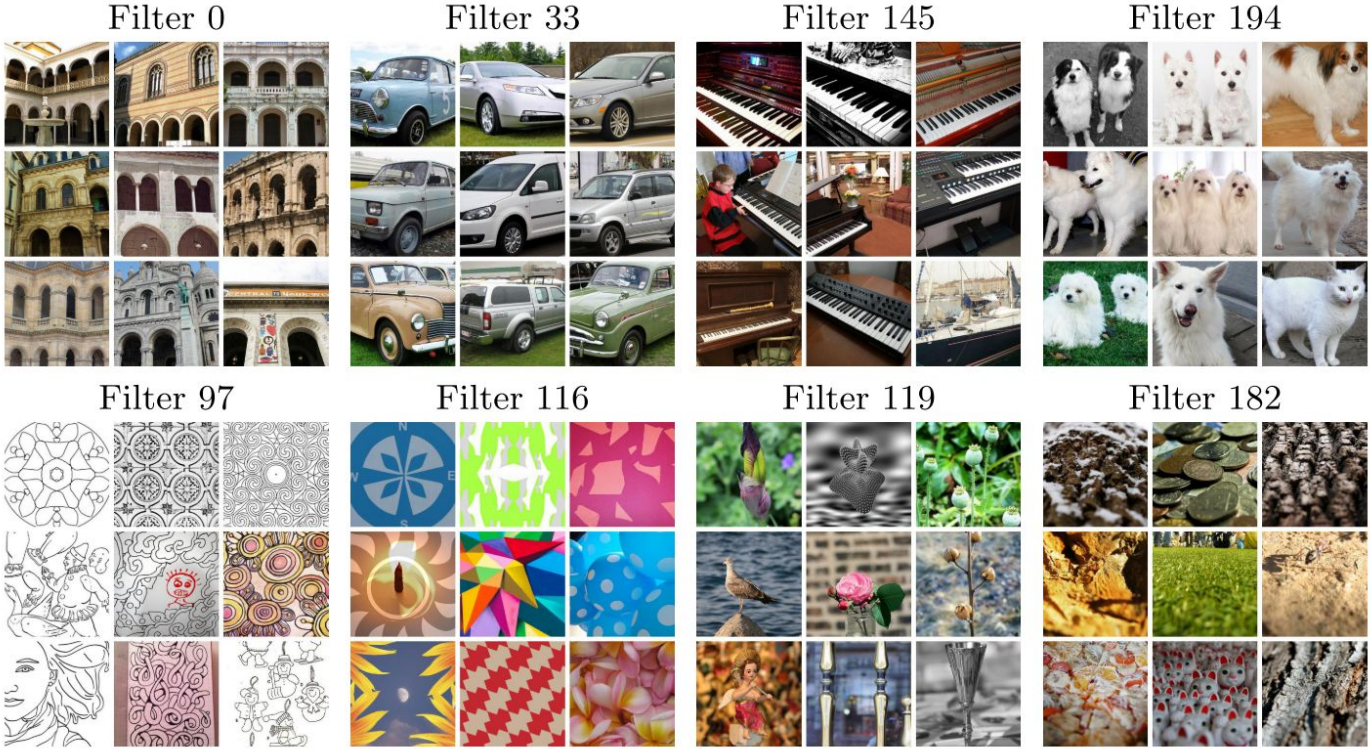
Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV) (2018)

Clustering as pretext task



Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV) (2018)

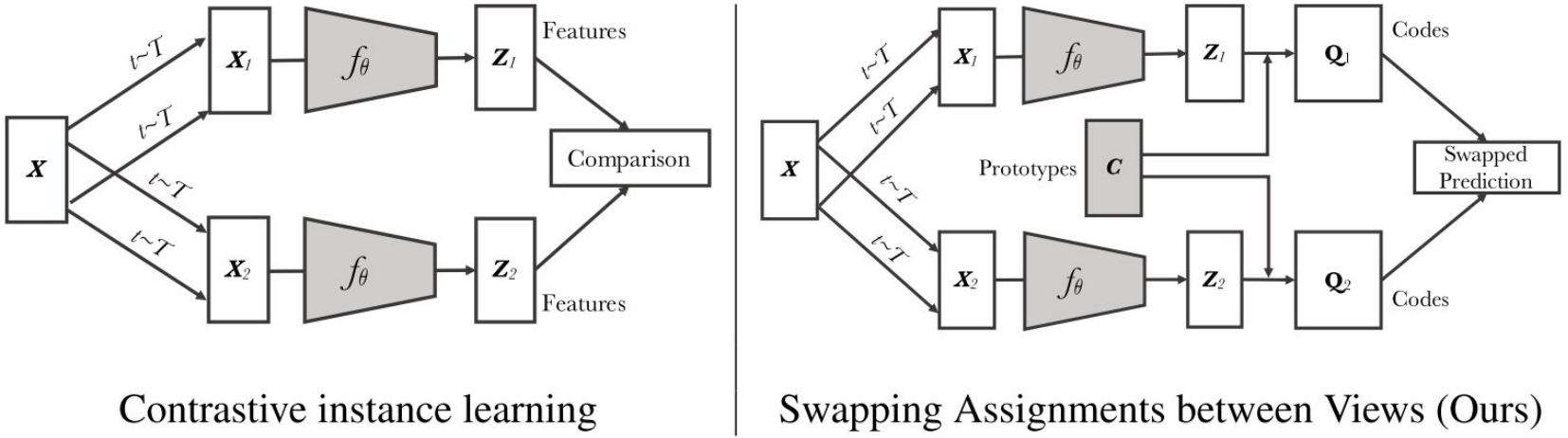
Clustering as pretext task



Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV) (2018)

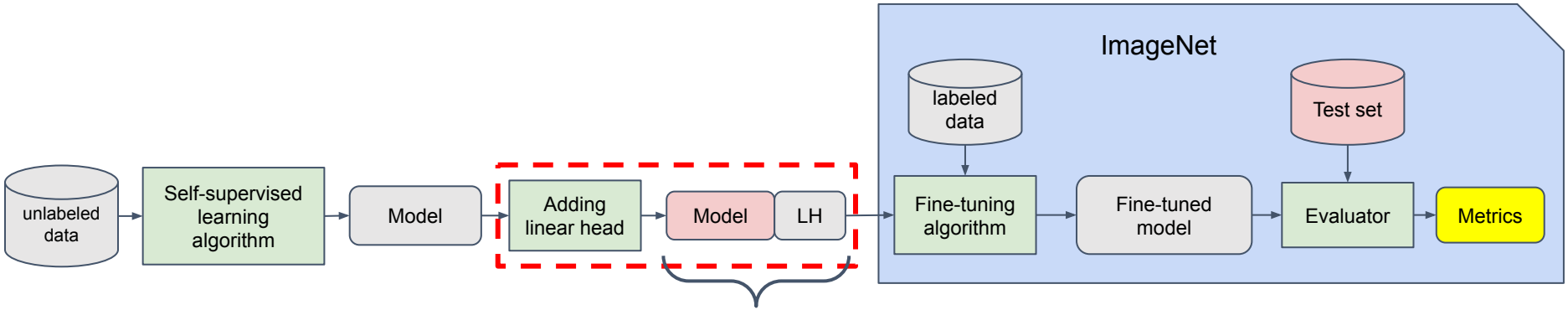
SWAV

Unifying clustering-based approaches with contrastive approaches



Source: Caron, Mathilde, et al. "Unsupervised learning of visual features by contrasting cluster assignments." arXiv preprint arXiv:2006.09882 (2020).

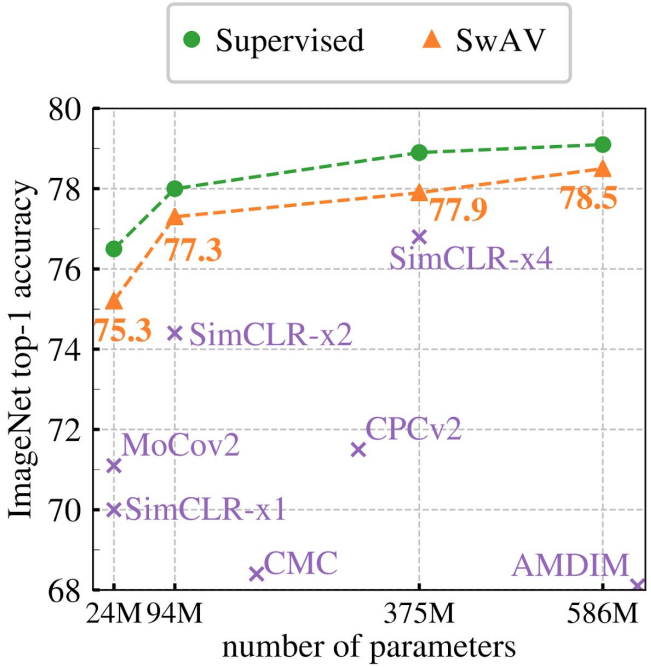
Quality of representations: linear evaluation



The model parameters are frozen, only the linear head (LH) parameters are fine-tuned.

Empirical results

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [64]	R50	24	39.6
Jigsaw [45]	R50	24	45.7
NPID [57]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [67]	R50	24	58.8
NPID++ [43]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [43]	R50	24	63.6
CPC v2 [27]	R50	24	63.8
PCL [36]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	**75.3**

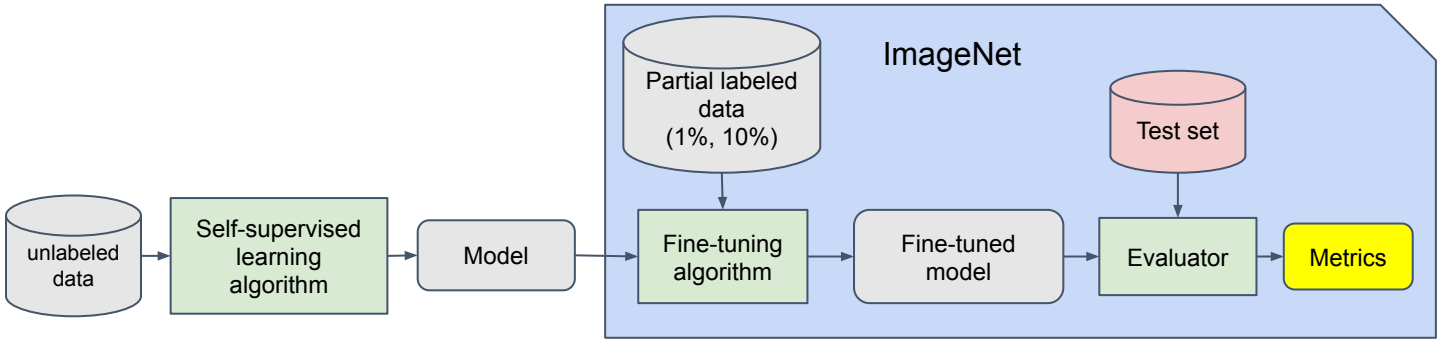


Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

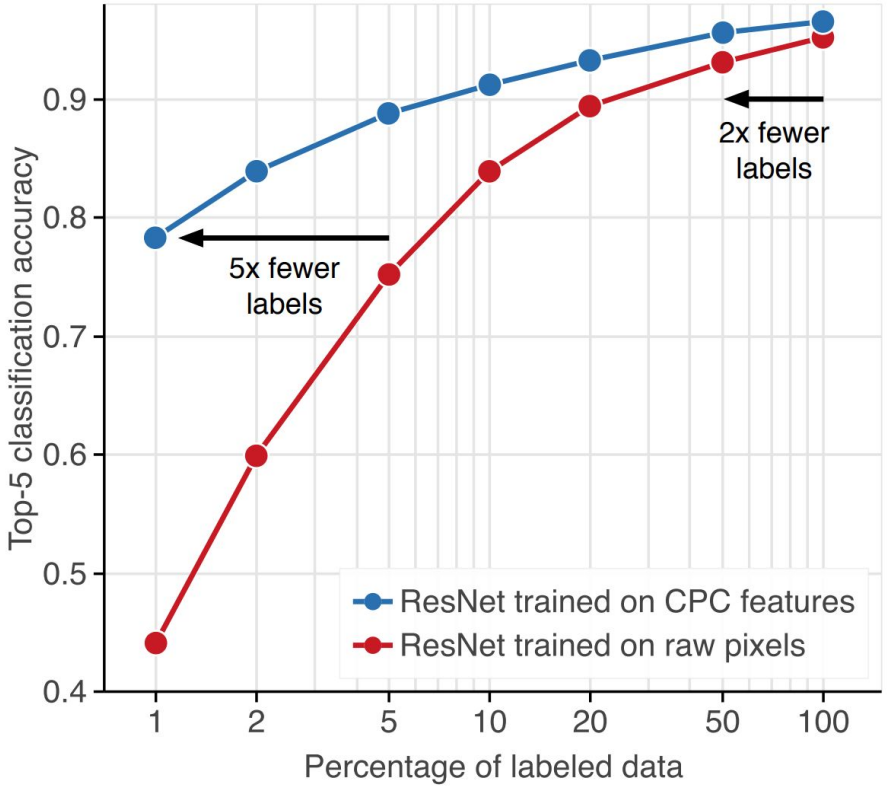
(a) ResNet-50 encoder.

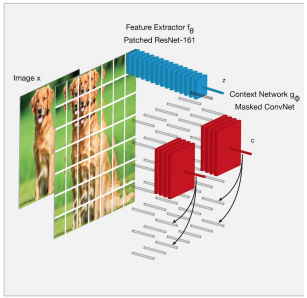
Caron, Mathilde, et al. "Unsupervised learning of visual features by contrasting cluster assignments." arXiv preprint arXiv:2006.09882 (2020).
Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." Advances in Neural Information Processing Systems 33 (2020).

Semi-supervised learning



Empirical results for CPC v2





Henaff, Olivier et al.. "Data-efficient image recognition with contrastive predictive coding." International Conference on Machine Learning. PMLR, 2020.

Empirical results

Method	Top-1		Top-5	
	1%	10%	1%	10%
Supervised [77]	25.4	56.4	48.4	80.4
InstDisc	-	-	39.2	77.4
PIRL [35]	-	-	57.2	83.8
SimCLR [8]	48.3	65.6	75.5	87.8
BYOL (ours)	**53.2**	**68.8**	**78.4**	**89.0**

(a) ResNet-50 encoder.

Method	Architecture	Param.	Top-1		Top-5	
			1%	10%	1%	10%
CPC v2 [32]	ResNet-161	305M	-	-	77.9	91.2
SimCLR [8]	ResNet-50 (2×)	94M	58.5	71.7	83.0	91.2
BYOL (ours)	ResNet-50 (2×)	94M	62.2	73.5	84.1	91.7
SimCLR [8]	ResNet-50 (4×)	375M	63.0	74.4	85.8	92.6
BYOL (ours)	ResNet-50 (4×)	375M	69.1	75.7	87.9	92.5
BYOL (ours)	ResNet-200 (2×)	250M	**71.2**	**77.7**	**89.5**	**93.7**

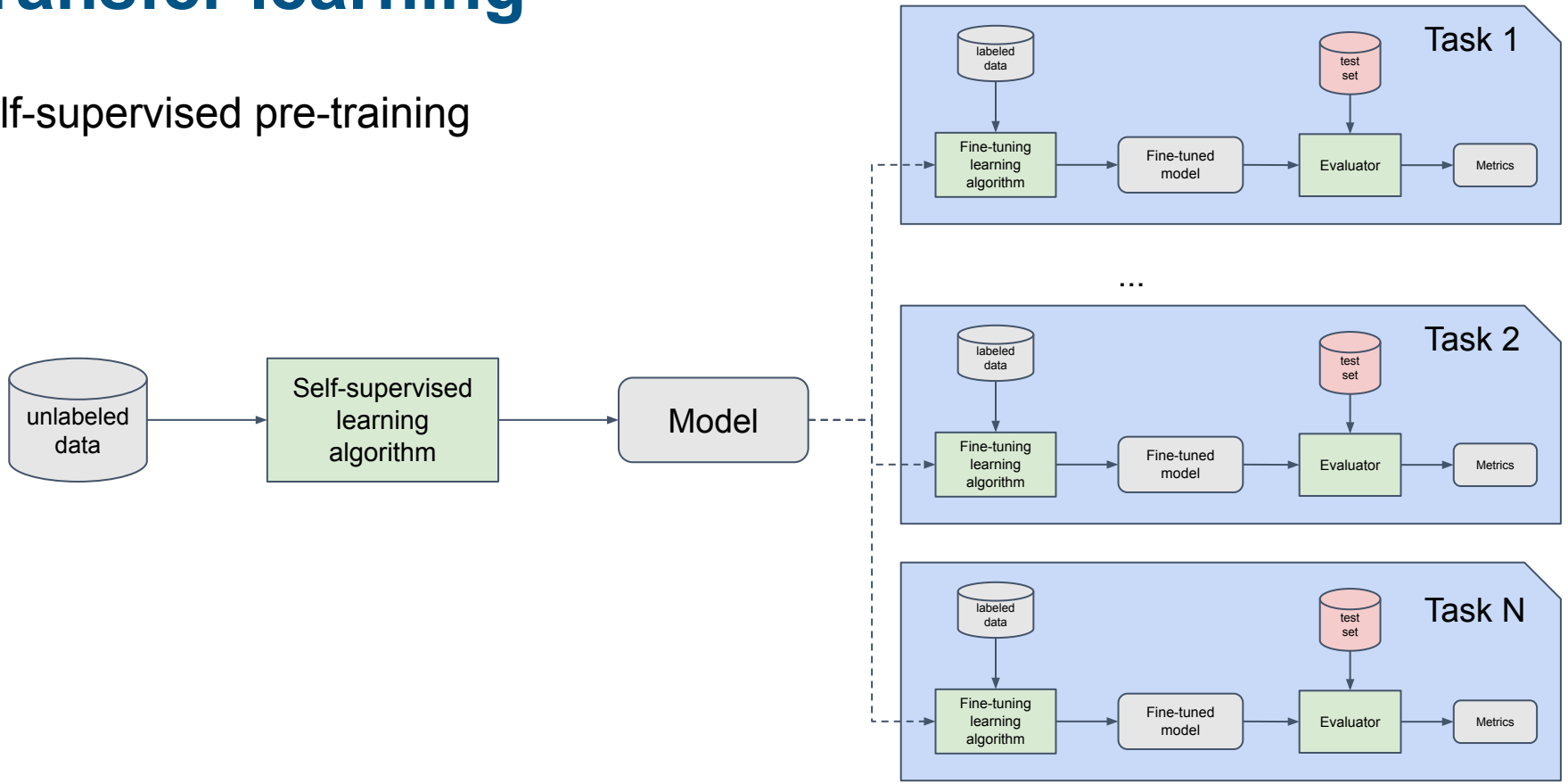
(b) Other ResNet encoder architectures.

Table 2: Semi-supervised training with a fraction of ImageNet labels.

Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." Advances in Neural Information Processing Systems 33 (2020).

Transfer learning

Self-supervised pre-training



Empirical results

Method	Food101	CIFAR10	CIFAR100	Birdsnap	SUN397	Cars	Aircraft	VOC2007	DTD	Pets	Caltech-101	Flowers
Linear evaluation:												
BYOL (ours)	**75.3**	91.3	**78.4**	**57.2**	**62.2**	**67.8**	60.6	82.5	75.5	90.4	94.2	**96.1**
SimCLR (repro)	72.8	90.5	74.4	42.4	60.6	49.3	49.8	81.4	**75.7**	84.6	89.3	92.6
SimCLR [8]	68.4	90.6	71.6	37.4	58.8	50.3	50.3	80.5	74.5	83.6	90.3	91.2
Supervised-IN [8]	72.3	**93.6**	78.3	53.7	61.9	66.7	**61.0**	**82.8**	74.9	**91.5**	**94.5**	94.7
Fine-tuned:												
BYOL (ours)	**88.5**	**97.8**	86.1	**76.3**	63.7	91.6	**88.1**	**85.4**	**76.2**	91.7	**93.8**	97.0
SimCLR (repro)	87.5	97.4	85.3	75.0	63.9	91.4	87.6	84.5	75.4	89.4	91.7	96.6
SimCLR [8]	88.2	97.7	85.9	75.9	63.5	91.3	88.1	84.1	73.2	89.2	92.1	97.0
Supervised-IN [8]	88.3	97.5	**86.4**	75.8	**64.3**	**92.1**	86.0	85.0	74.6	**92.1**	93.3	**97.6**
Random init [8]	86.9	95.9	80.2	76.1	53.6	91.4	85.9	67.3	64.8	81.5	72.6	92.0

Table 3: Transfer learning results from ImageNet (IN) with the standard ResNet-50 architecture.

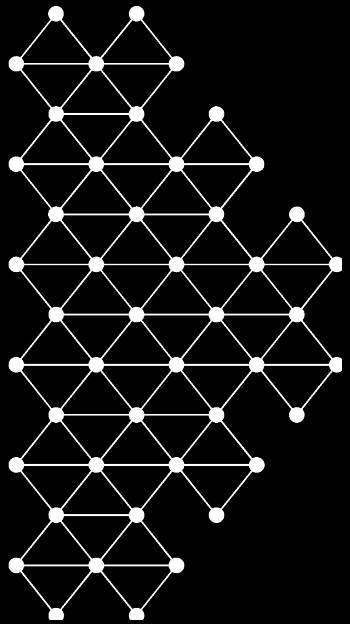
Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." Advances in Neural Information Processing Systems 33 (2020).

Discussion

- A lot of research has happened in the past few years.
- Some central ideas have been proposed more than 20 years ago.
-

Quebec
Artificial
Intelligence
Institute

Mila



Questions?

What is a good representation?

Signal processing question

“What are the different representations of a signal, their properties and the complexity of the algorithms to compute them?”

Machine learning question

“Is there a universal task with which we can learn good representations for all downstream tasks and all datasets?”

What is a good representation?

Signal processing question

“What are the different representations of a signal, their properties and the complexity of the algorithms to compute them?”

Machine learning question

“Is there a universal [illegible] which we can learn good [illegible]ions for all downstre[illegible]d all datasets?”

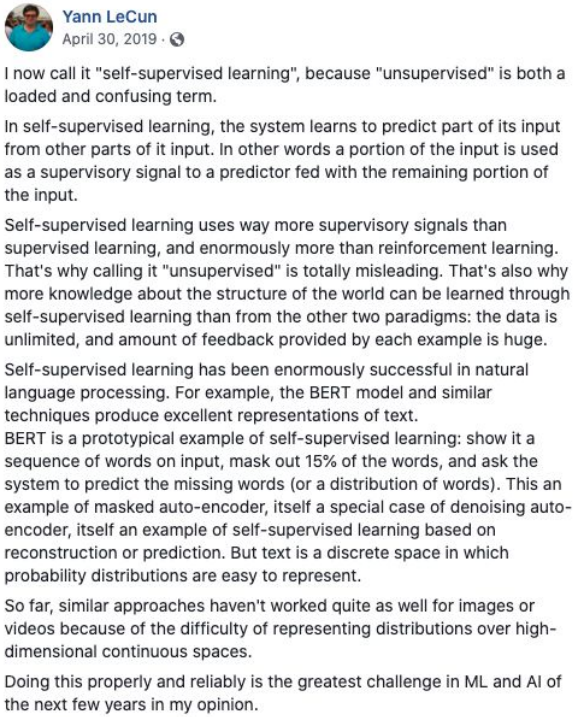
No free lunch theorem

Plan

- TODO: Using Self-Supervised Learning Can Improve Model Robustness and Uncertainty
-

Unsupervised learning

- Learning without human annotations
- It includes:
  - Latent variable models
  - Generative models
  - Self-supervised learning
  - ...
- The term is now *loaded* and *confusing*



Yann LeCun
April 30, 2019 ·

I now call it "self-supervised learning", because "unsupervised" is both a loaded and confusing term.

In self-supervised learning, the system learns to predict part of its input from other parts of it input. In other words a portion of the input is used as a supervisory signal to a predictor fed with the remaining portion of the input.

Self-supervised learning uses way more supervisory signals than supervised learning, and enormously more than reinforcement learning. That's why calling it "unsupervised" is totally misleading. That's also why more knowledge about the structure of the world can be learned through self-supervised learning than from the other two paradigms: the data is unlimited, and amount of feedback provided by each example is huge.

Self-supervised learning has been enormously successful in natural language processing. For example, the BERT model and similar techniques produce excellent representations of text.
BERT is a prototypical example of self-supervised learning: show it a sequence of words on input, mask out 15% of the words, and ask the system to predict the missing words (or a distribution of words). This an example of masked auto-encoder, itself a special case of denoising auto-encoder, itself an example of self-supervised learning based on reconstruction or prediction. But text is a discrete space in which probability distributions are easy to represent.

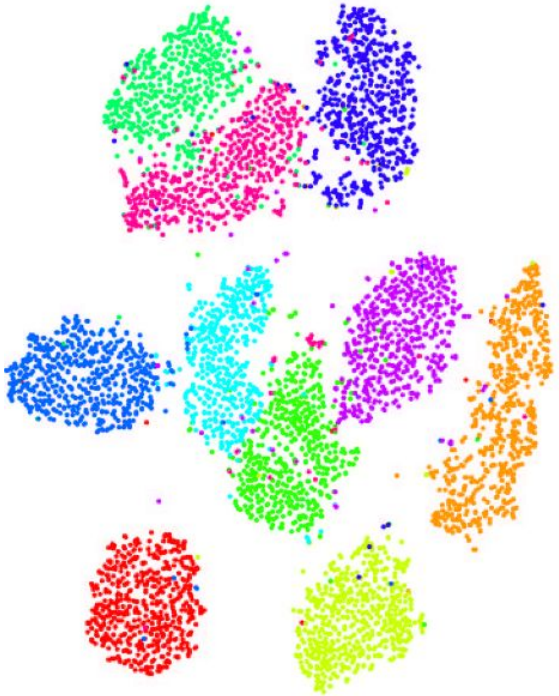
So far, similar approaches haven't worked quite as well for images or videos because of the difficulty of representing distributions over high-dimensional continuous spaces.

Doing this properly and reliably is the greatest challenge in ML and AI of the next few years in my opinion.

What is a good representation?

Examples of transformations over data:

- Fourier/Wavelet transform
- Polynomial representation
- Principal component analysis/T-SNE
- Representation learning



MNIST cluster by T-SNE

Image credit: Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

Data augmentation pipeline

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

SimCLR results

- Composition of operations is crucial for learning good representations

- Contrastive learning needs stronger data augmentation than supervised learning

Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020

Computational efficiency

Robustness

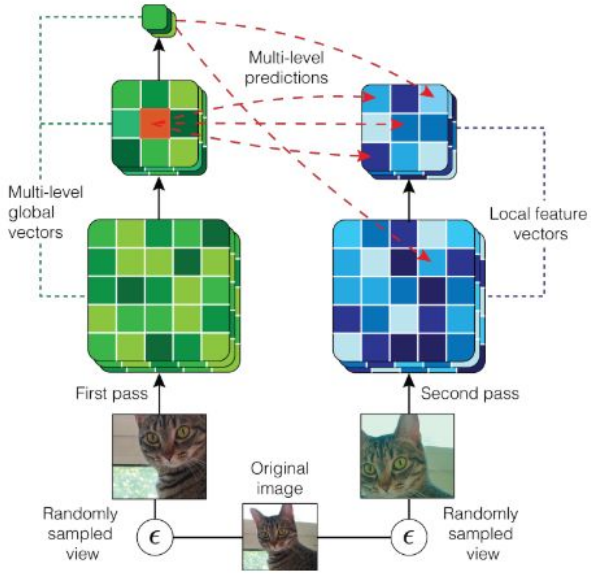
Acknowledgement

- Inspired partly from IFT6268 “Self-supervised Representation Learning” course by Aaron Courville.

Augmented Multiscale Deep InfoMax (AMDIM)

Data augmentation is random horizontal flip on original image, then one of [crop, color jitter, to grayscale]

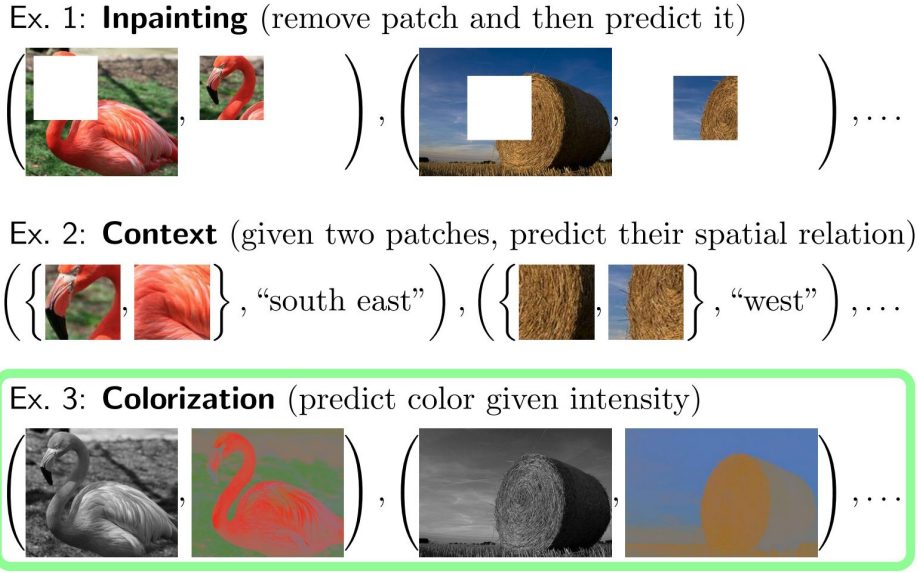
Maximize MI between different layers of the encoder (a resnet)



Bachman, P., Hjelm, R. D., & Buchwalter, W. (2019). *Learning Representations by Maximizing Mutual Information Across Views*.

Pretext tasks (early works)

Examples of pretext tasks for vision: inpainting, colorization, deblurring

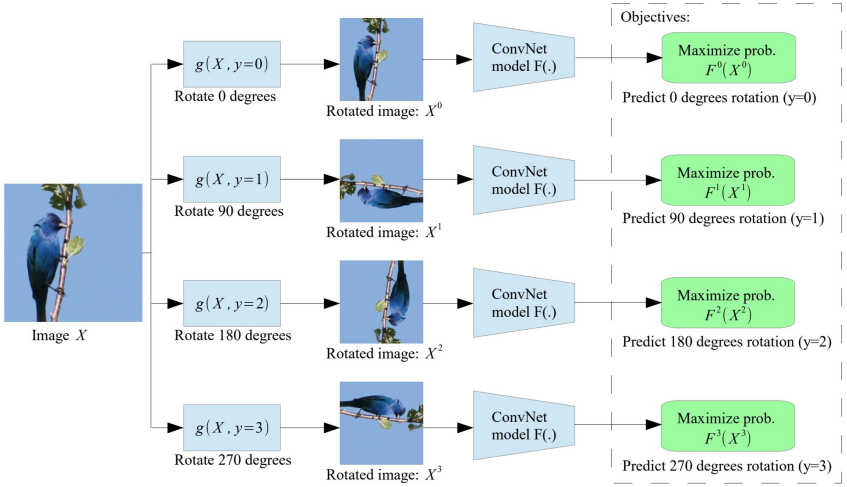


https://arxiv.org/abs/1703.04044

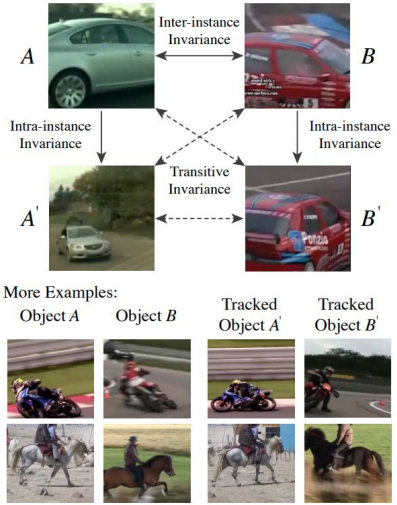
- Became really popular as generative models became more robust (encoder = feature extractor)
- **Sadly**: learned representations do not transfer well across datasets, and are not that great. **Common hypothesis**: tasks that require "higher level" knowledge to be solved are better.

Exemplar-based pretext tasks

Each image has its own class, generate more examples via data augmentation (for images) or tracking (videos) and predict those



https://arxiv.org/abs/1803.07728



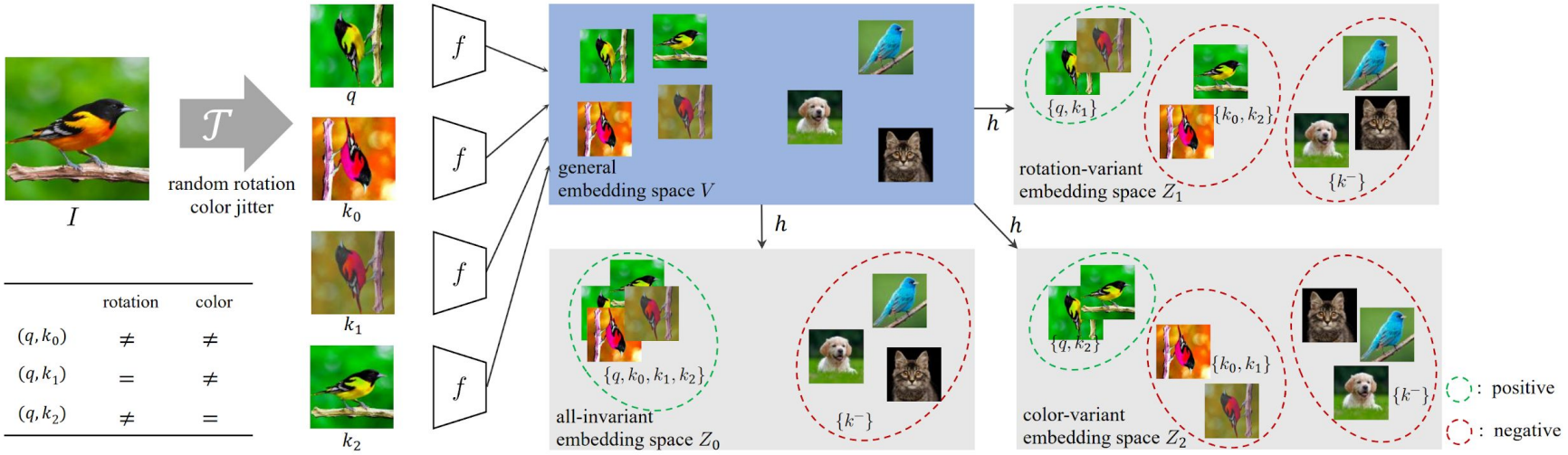
https://arxiv.org/abs/1708.02901

Note: most contrastive approaches (coming up) use this idea!

Contrastive methods

“Leave-one-out” contrastive learning (LooC)

Simple idea: learn per-transform embedding spaces, combine in downstream task!



	rotation	color
(q, k_0)	$\neq$	$\neq$
(q, k_1)	$=$	$\neq$
(q, k_2)	$\neq$	$=$