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

Gaétan Marceau Caron

Agenda

01 Representation learning
02 Pretext tasks with reconstruction
03 Pretext tasks with pseudo-labels
04 Contrastive learning
05 Autoregressive contrastive learning
06 Contrastive learning w/o negative examples
07 Contrastive learning with clustering
08 Empirical results
09 Discussion

Mila

01

Representation learning

What is a representation?

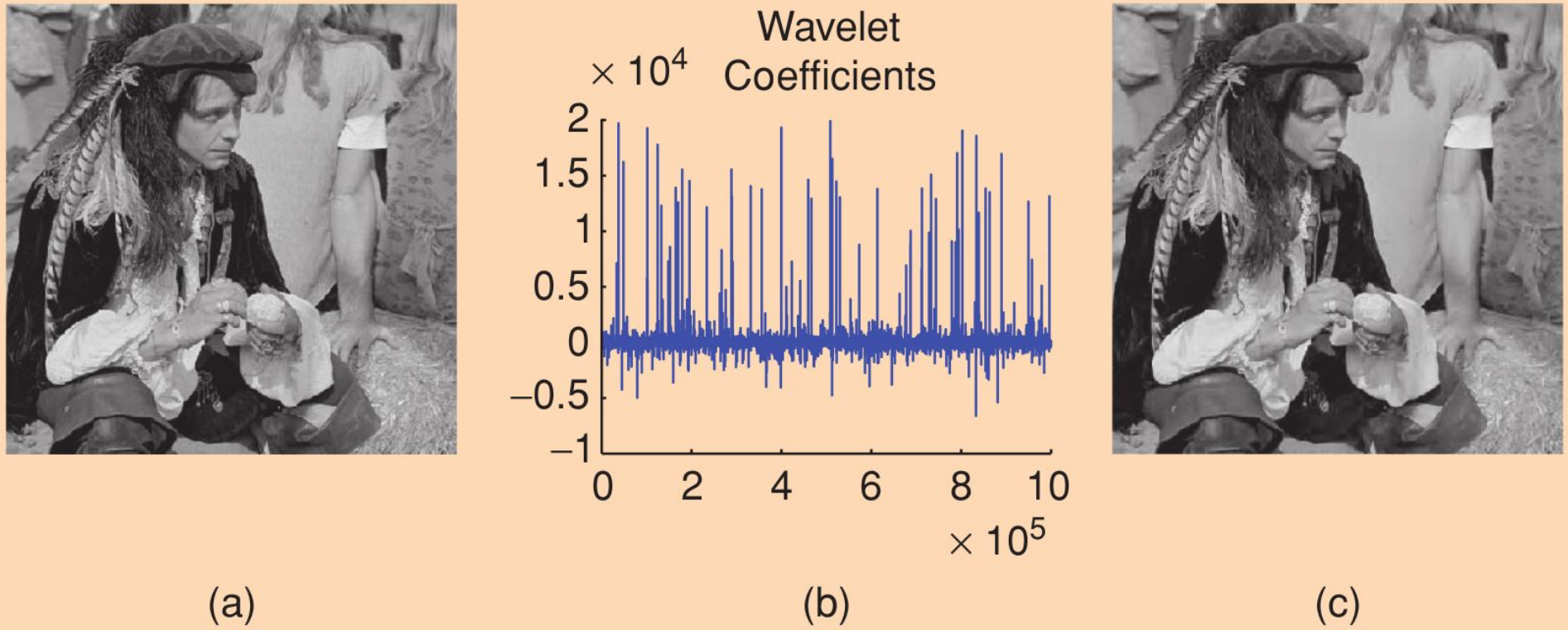


Image credit: Candès, Emmanuel J., and Michael B. Wakin. "An introduction to compressive sampling." IEEE signal processing magazine 25.2 (2008): 21-30.

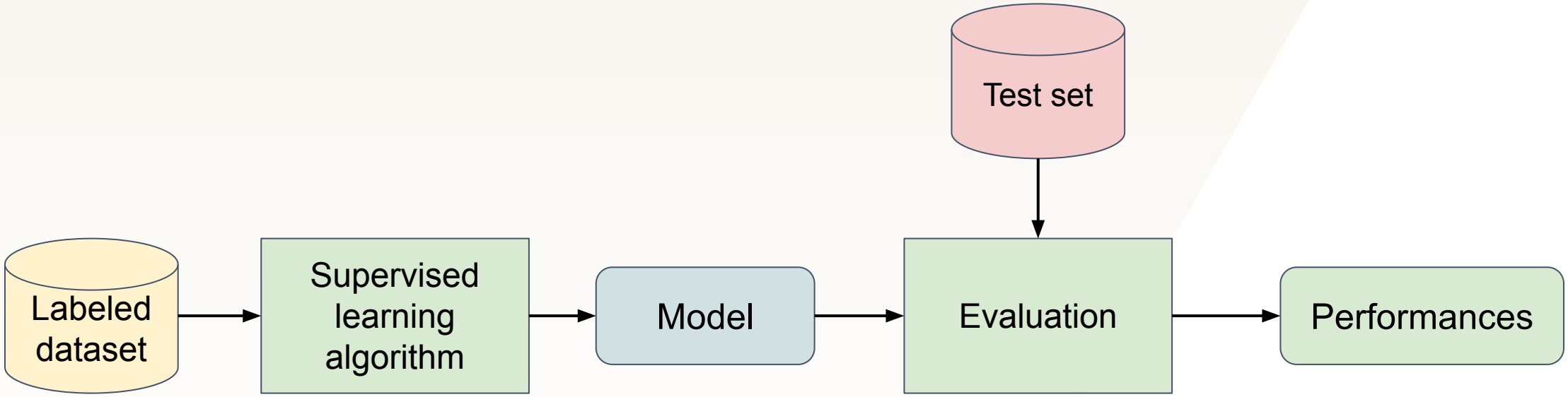
Mila

What is a good representation?

Some (informal) properties of good representations:

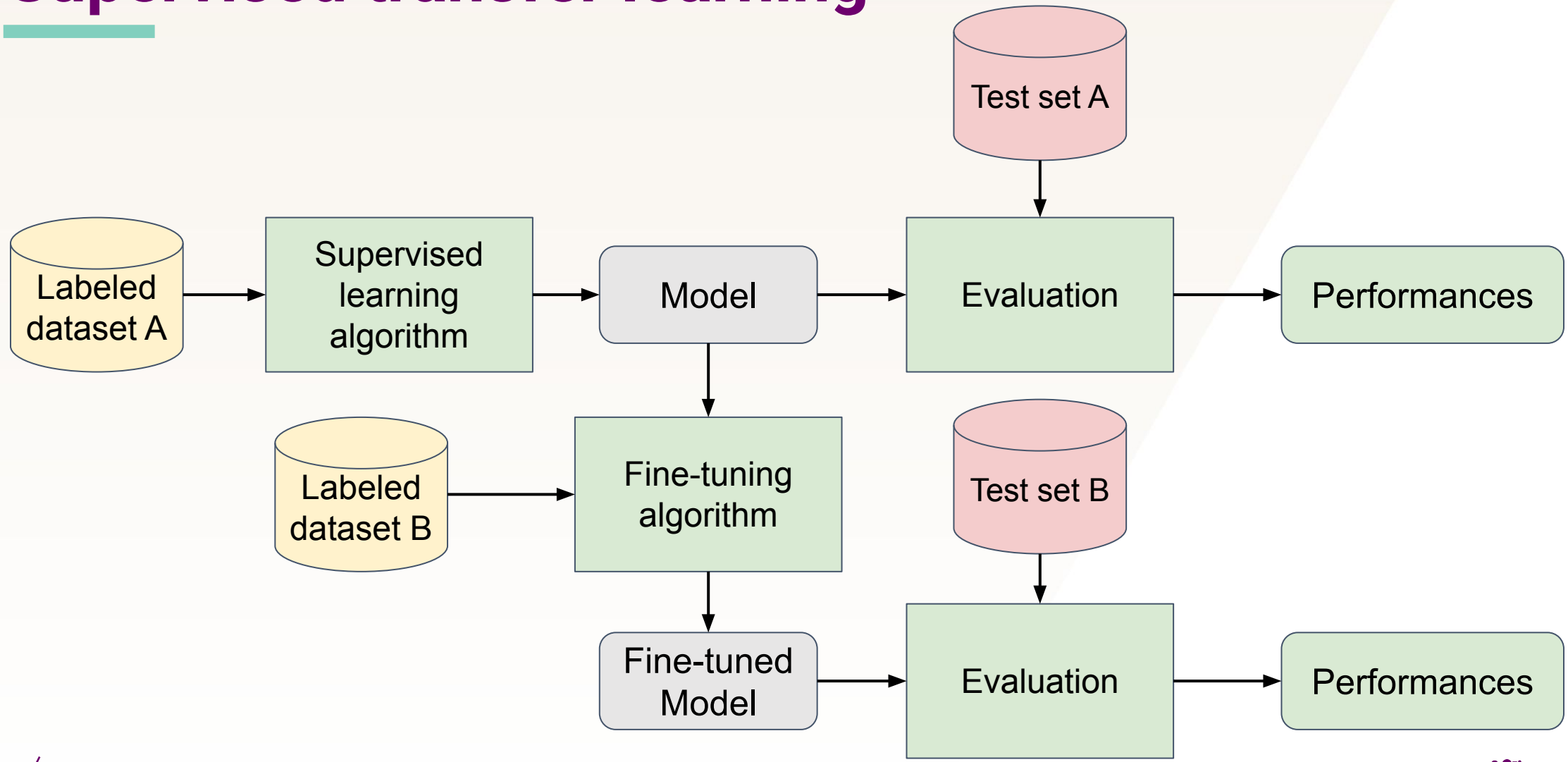
- Useful
- Expressive
- Robust to perturbations
- Interpretable

Supervised learning



Mila

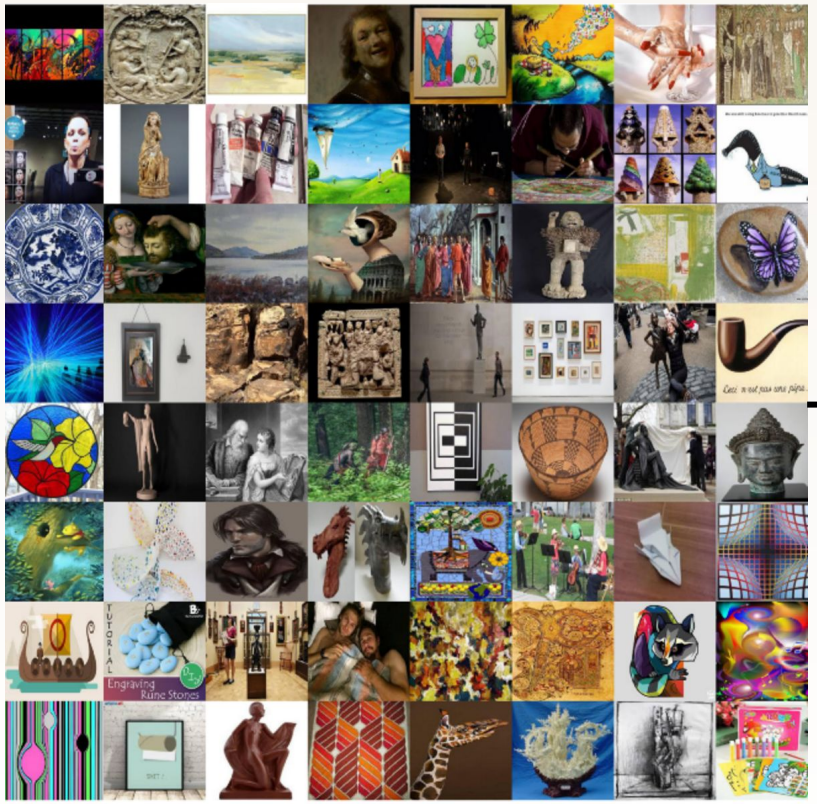
Supervised transfer learning



Mila

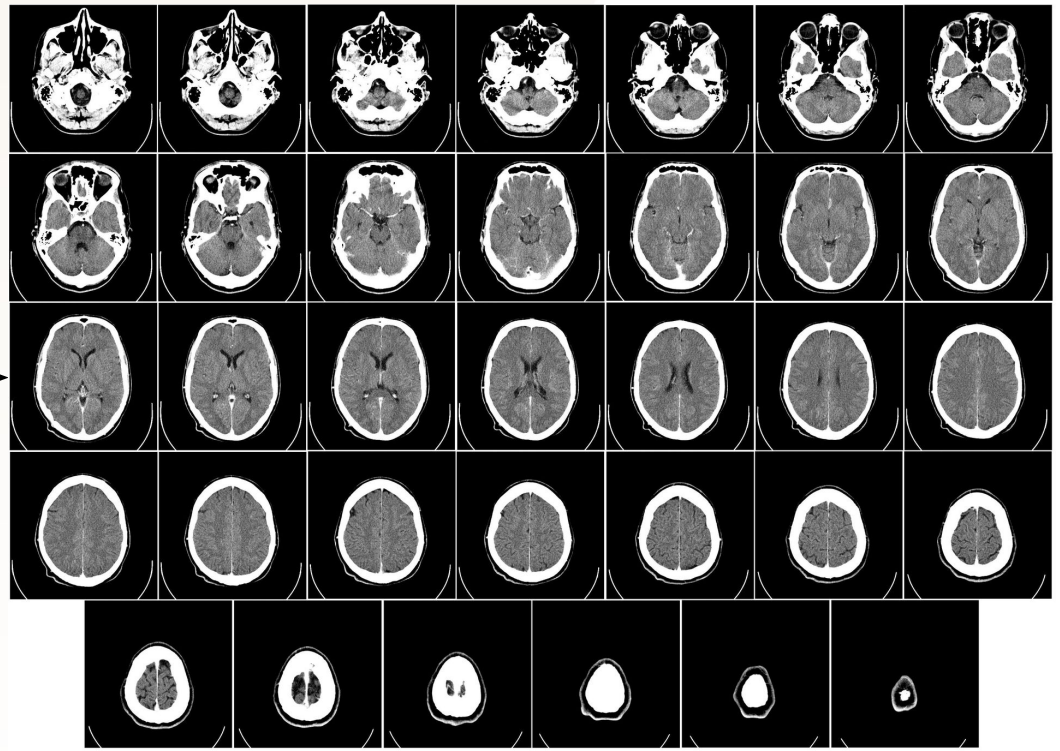
Limitations of supervised pre-training

Pre-training on ImageNet



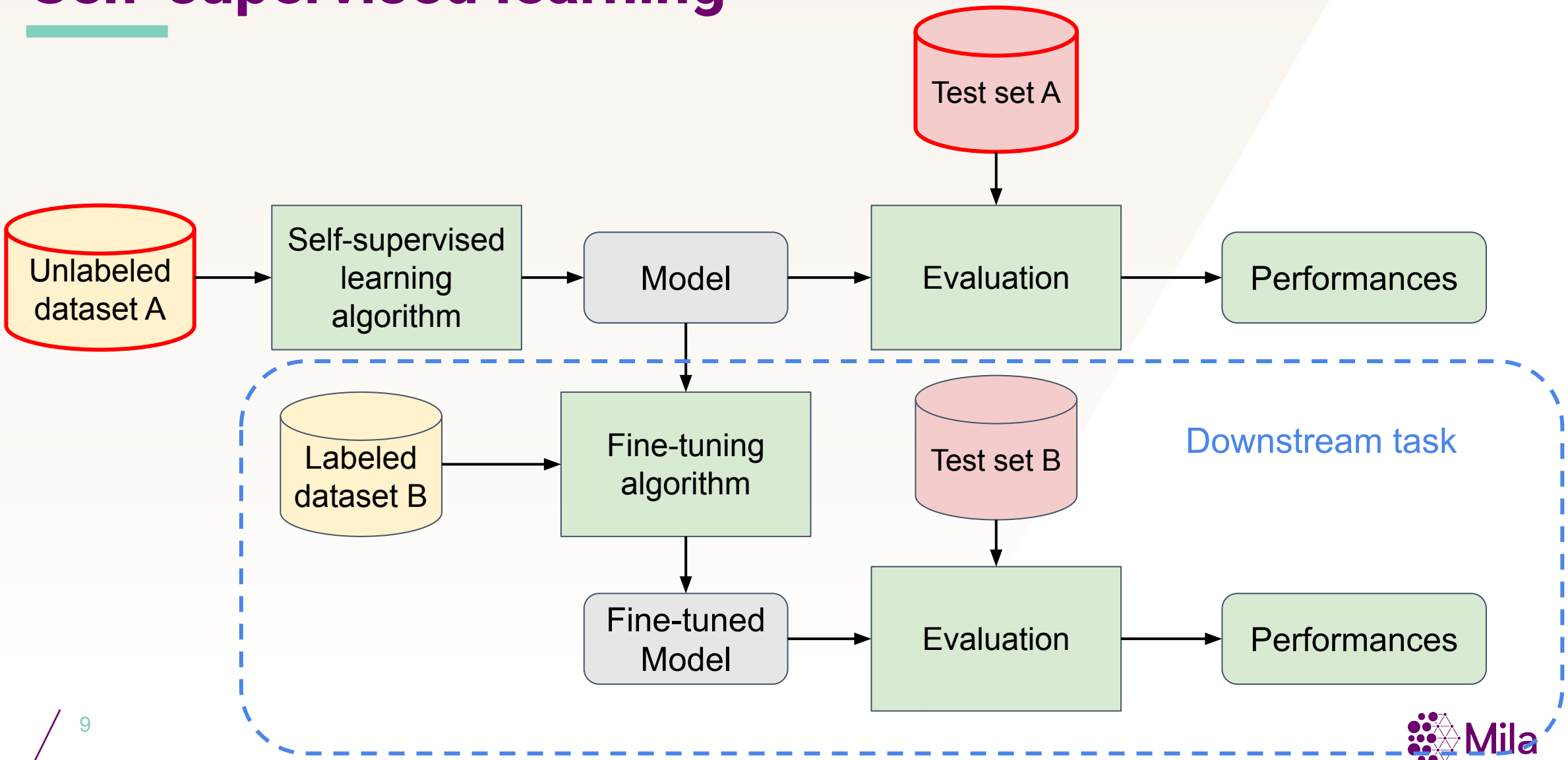
Transfer?

Downstream task on CT scan



Mila

Self-supervised learning



Mila

Objective

Find different techniques that leverage **the massive amount of unlabeled data** to learn representations that can **transfer well to downstream tasks**.

How to learn good representations?

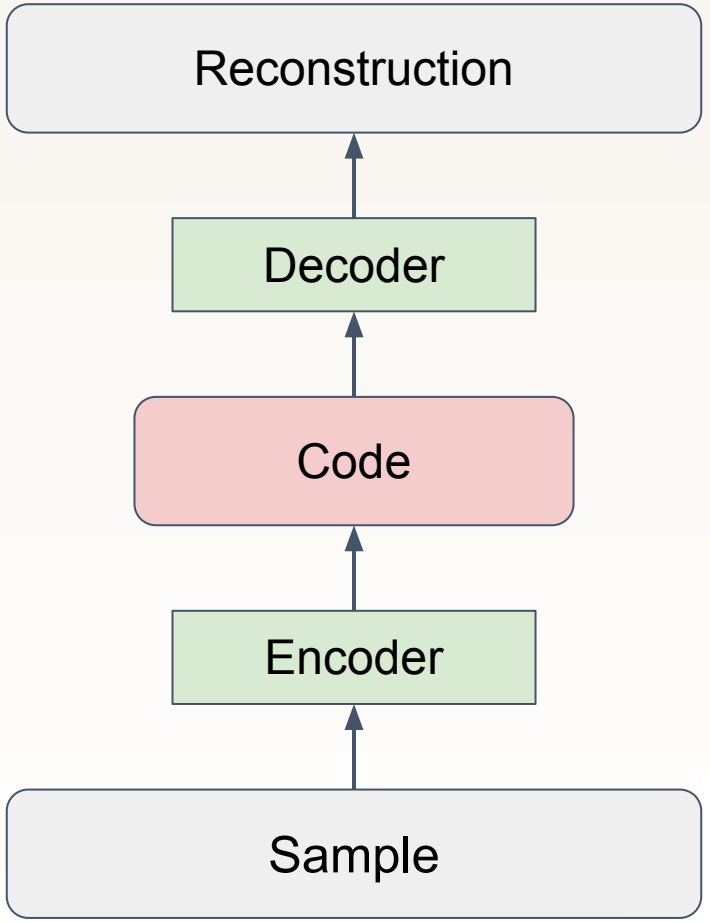
Existing families of models:

- Auto-encoding
- Pretext tasks
- Contrastive learning
- Clustering

02

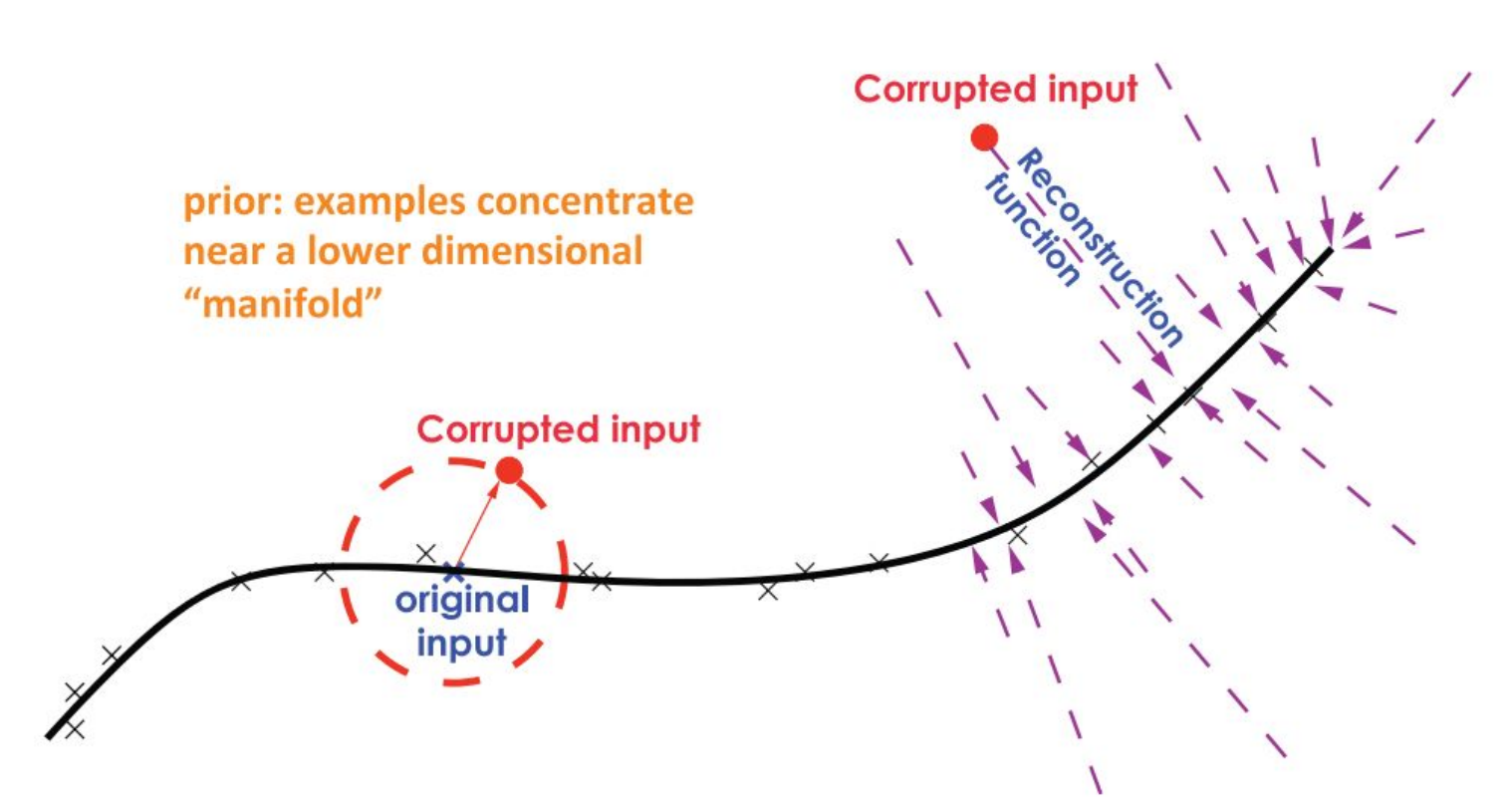
Pretext tasks with reconstruction

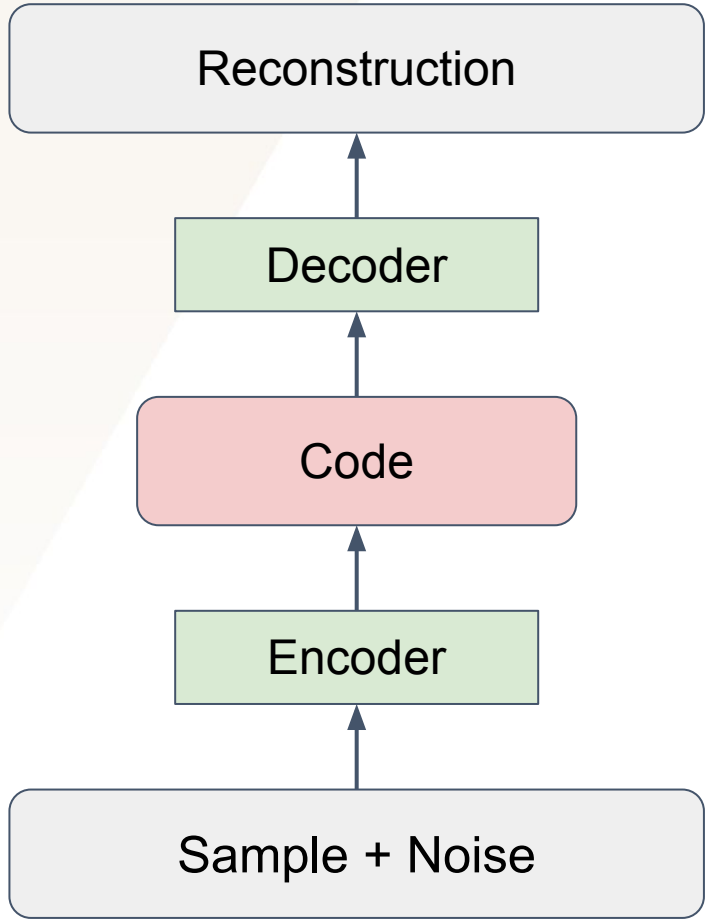
Auto-encoder



Mila

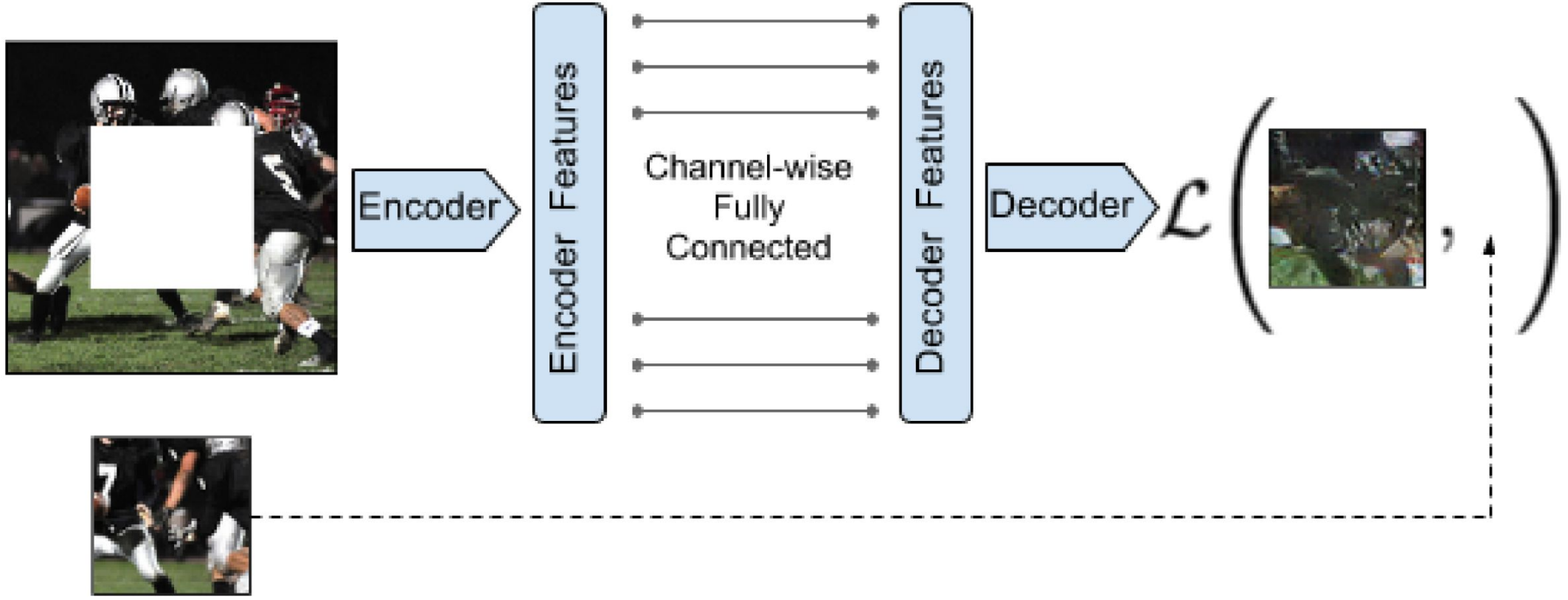
Denoising auto-encoder





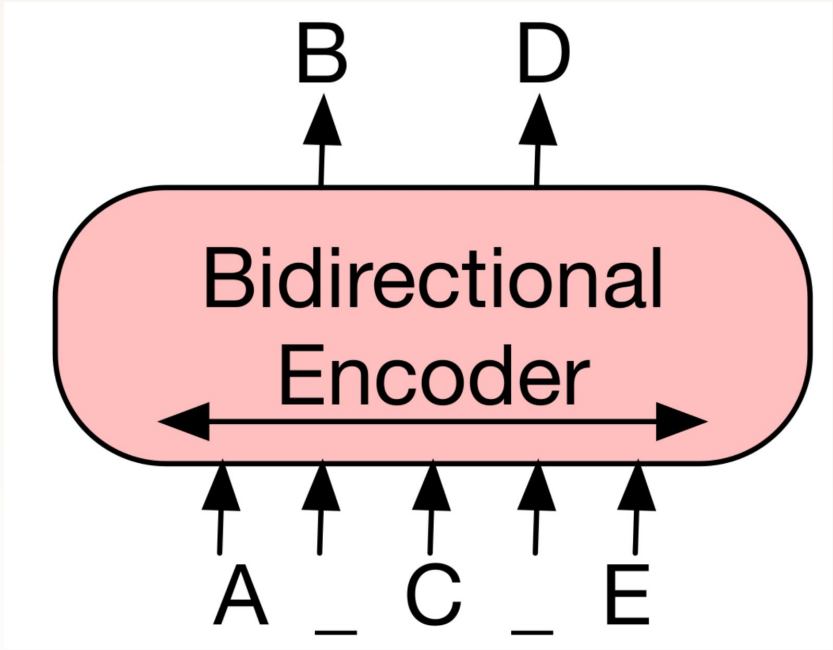
Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

Context encoders

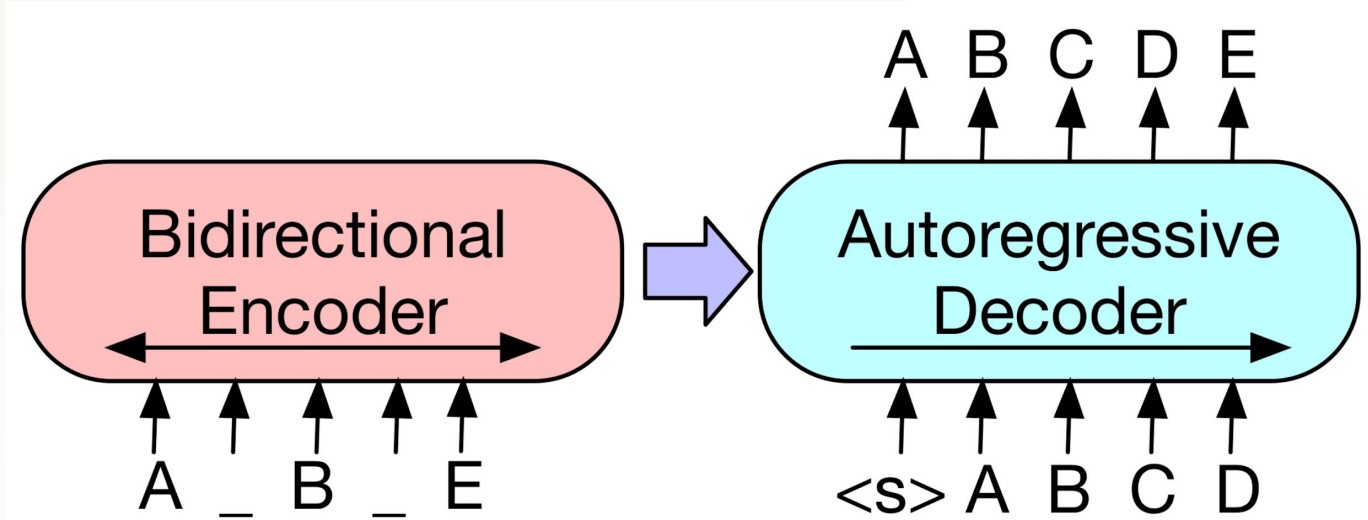


Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, Alexei A. Efros: Context Encoders: Feature Learning by Inpainting. CVPR 2016: 2536-2544

Recent works in NLP



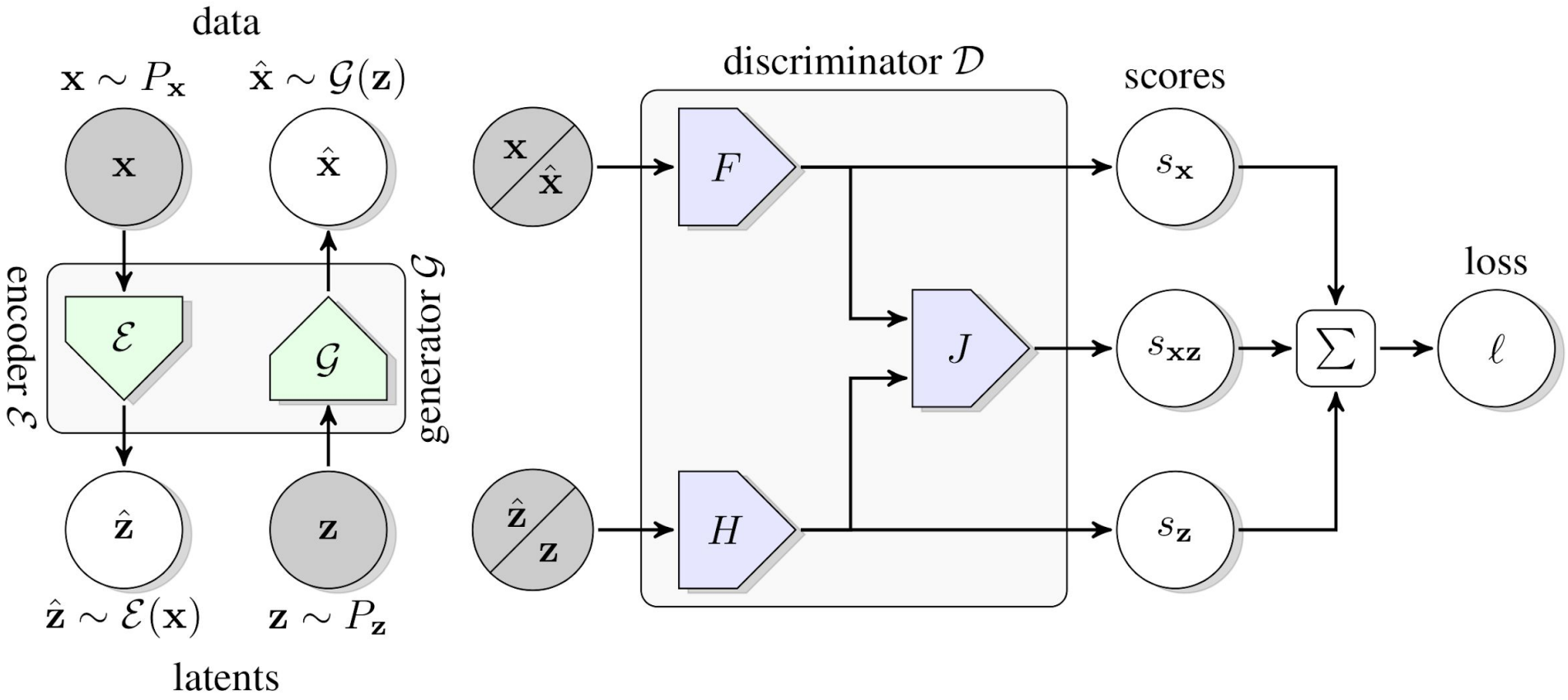
BERT



BART

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, Luke Zettlemoyer: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. ACL 2020: 7871-7880

BigBiGAN



Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, Jingjing Liu: Large-Scale Adversarial Training for Vision-and-Language Representation Learning. NeurIPS 2020

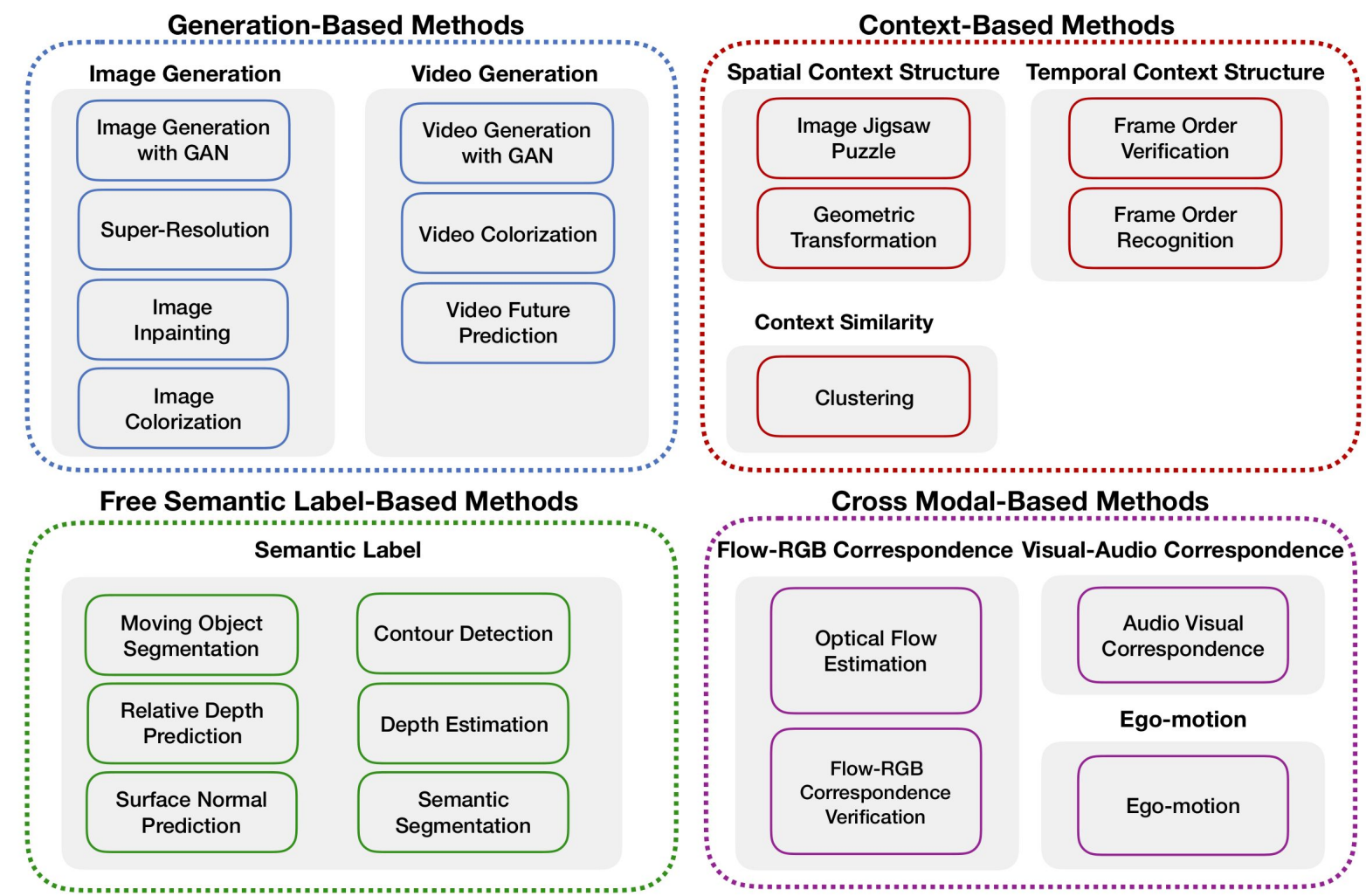
Discussion

- These techniques make predictions in the high-dimensional input space.
- They work better for sequences of symbols than audio signals or images.

03

Pretext tasks with pseudo-labels

Pretext tasks



Longlong Jing, Yingli Tian: Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey. CoRR abs/1902.06162 (2019)

Mila

Colorization

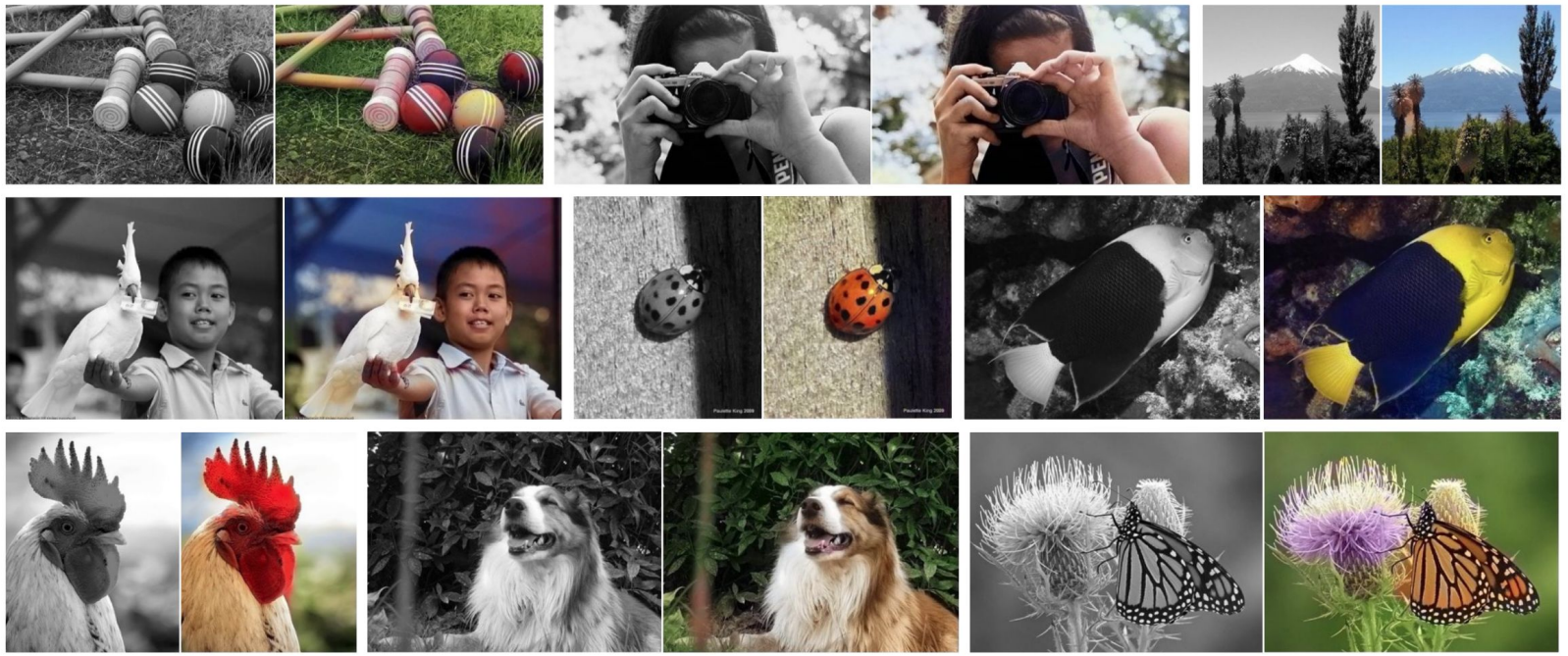
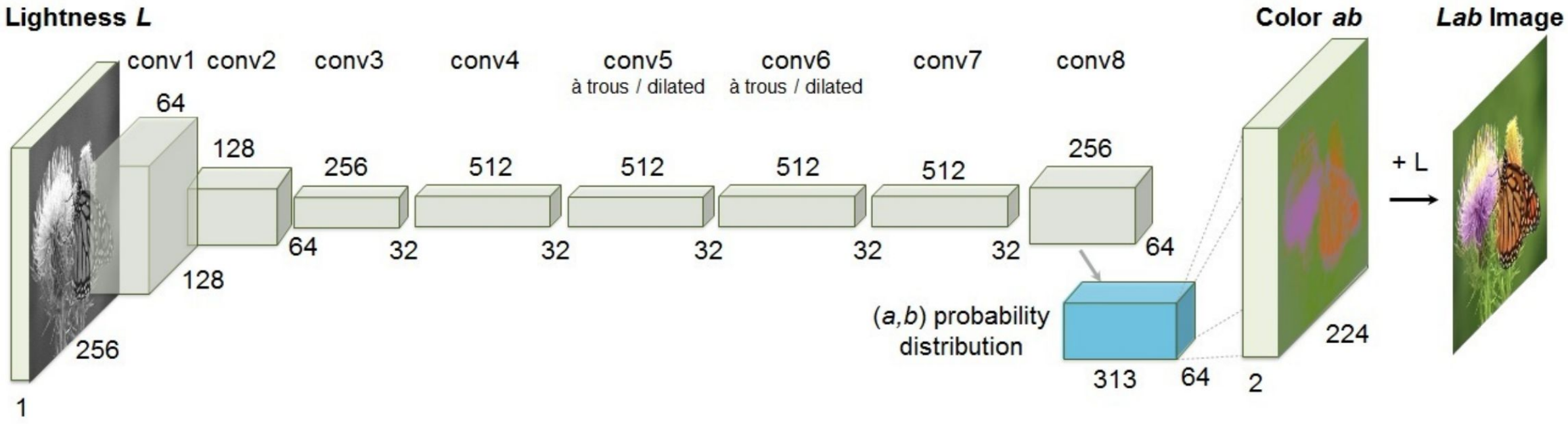


Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).

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

Mila

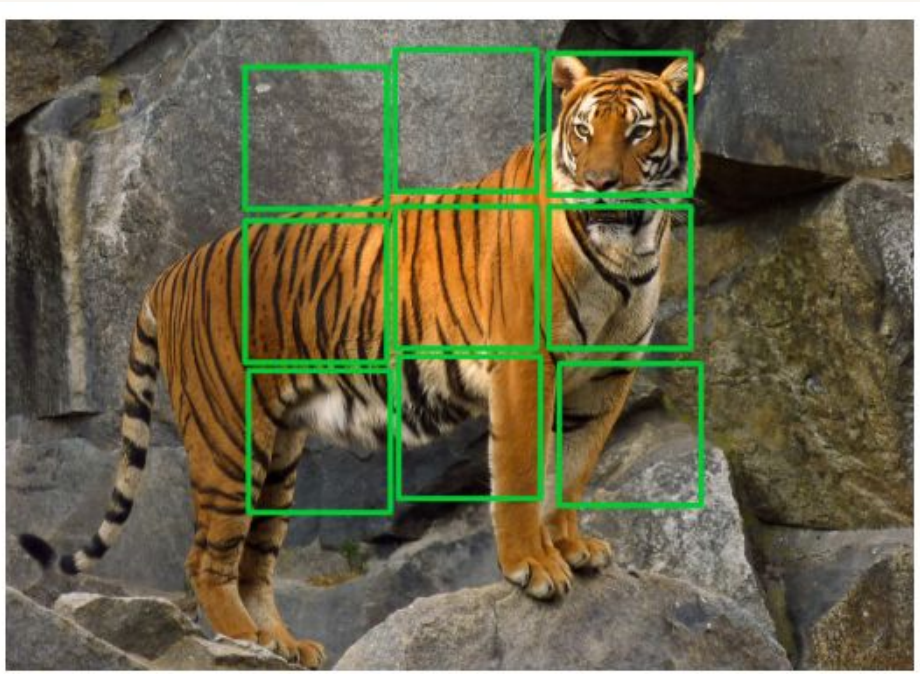
Colorization



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

Mila

Jigsaw puzzle



(a)

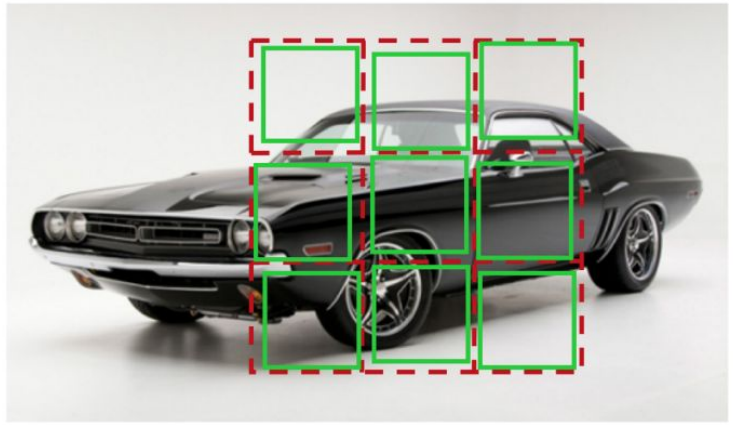
(b)

(c)

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

Mila

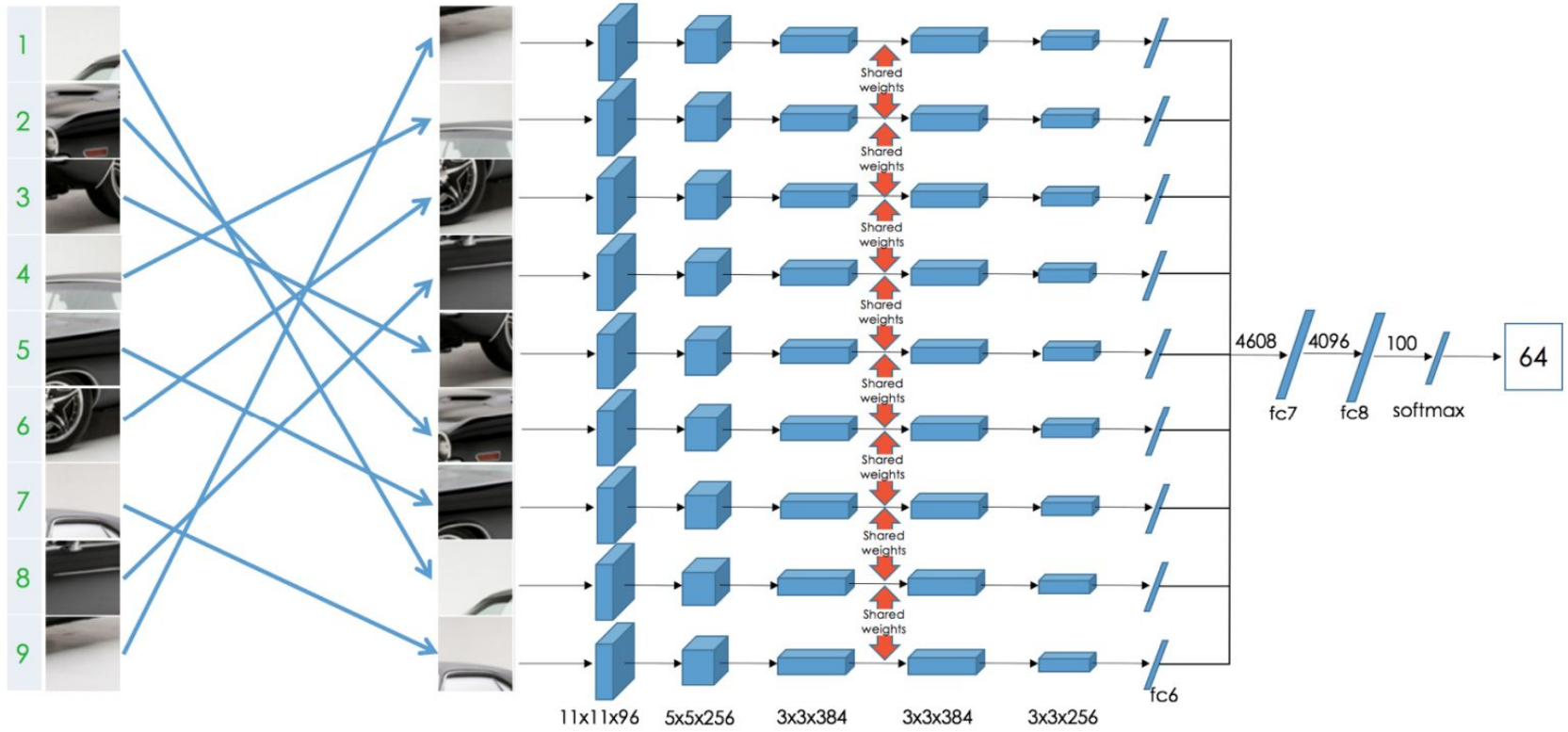
Jigsaw puzzle



Permutation Set

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

Reorder patches according to the selected permutation



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

Mila

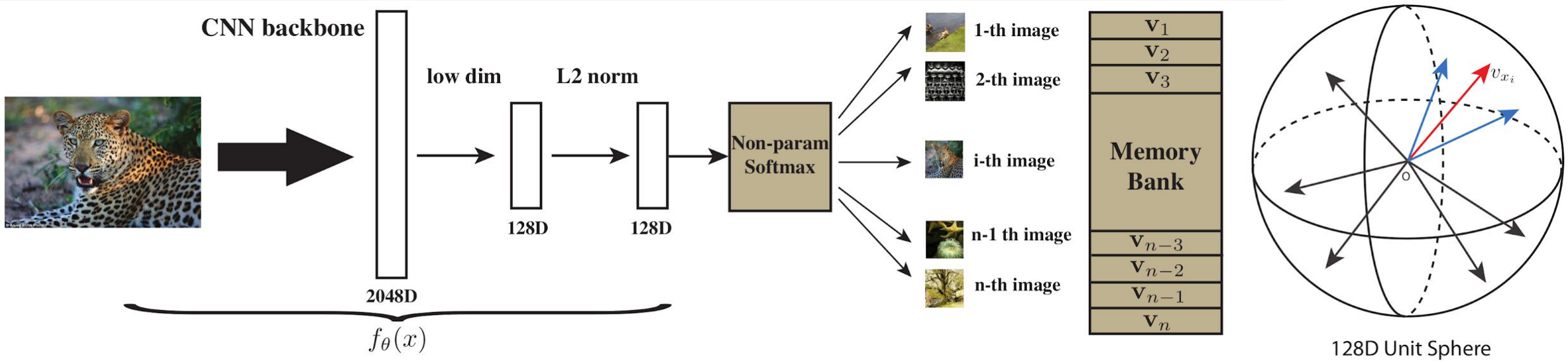
Discussion

- The pseudo-labels are generated with task specific prior knowledge.
- They are not competitive anymore compared to contrastive tasks.

04

Contrastive learning

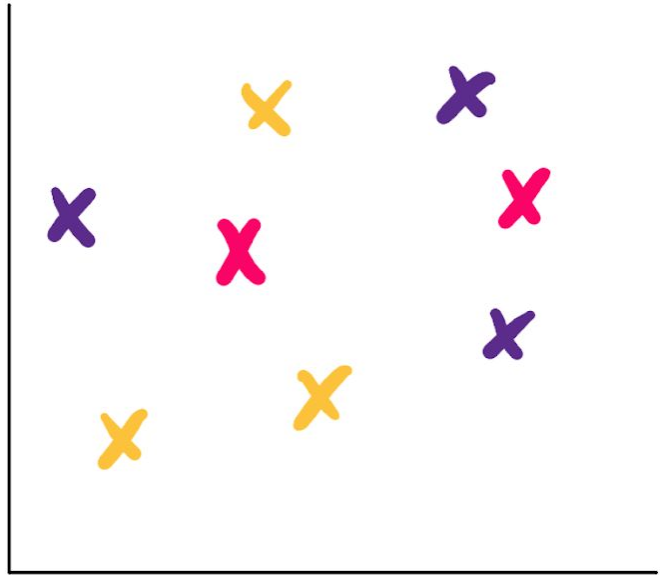
Instance discrimination

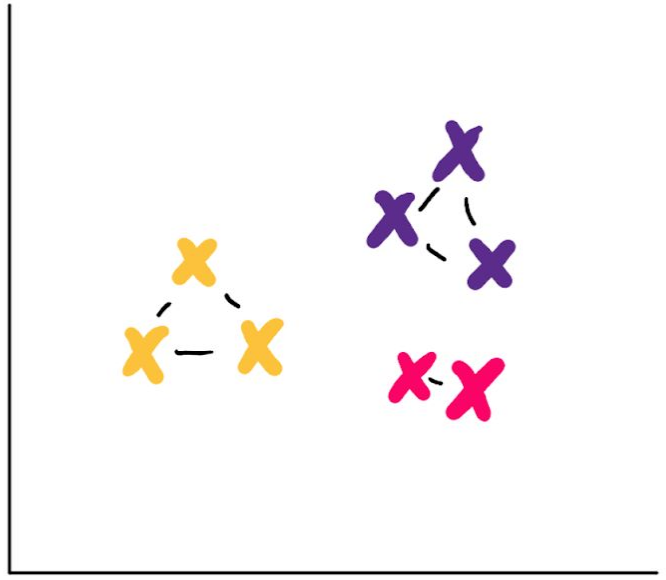


Non-parametric softmax: $$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)}$$

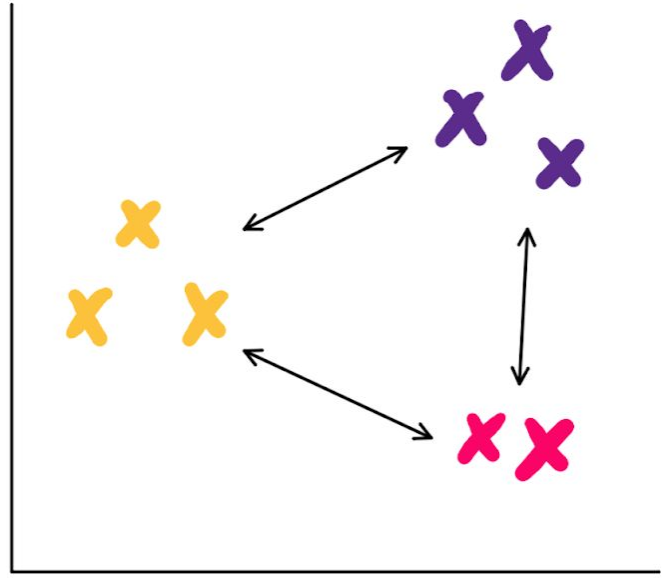
Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742

Contrastive learning





bring closer similar data



move away
dissimilar data

Mila

Normalized Temperature-scaled Cross Entropy

$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Anchor

Positive example

Negative examples

Similarity function

Temperature

$$\tau \in (0, 100]$$

Kihyuk Sohn: Improved Deep Metric Learning with Multi-class N-pair Loss Objective. NIPS 2016: 1849-1857

Why using a temperature hyper-parameter?

Can help the model learn from hard negatives.

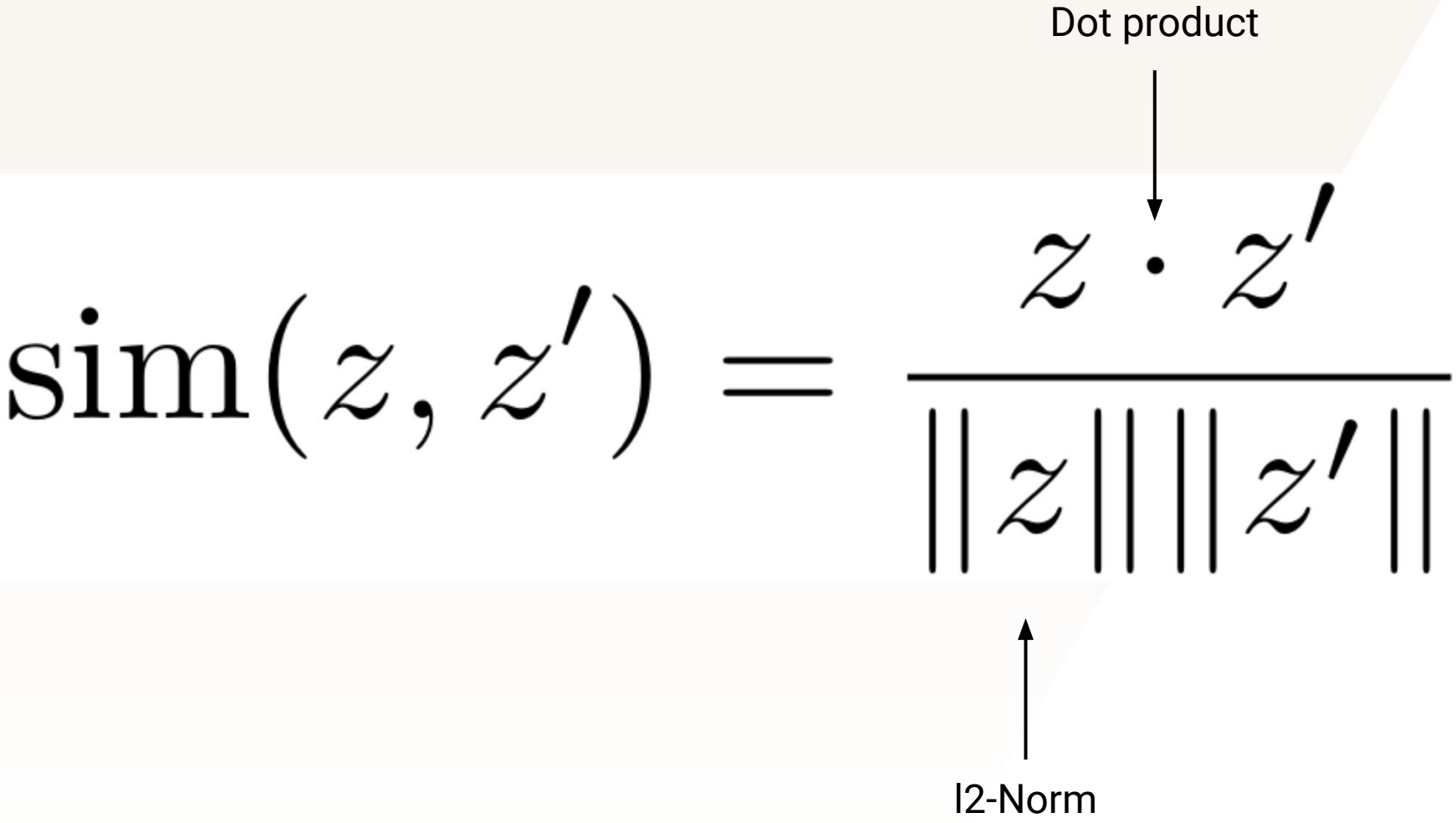
$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Temperature

$$\tau \in (0, 100]$$

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

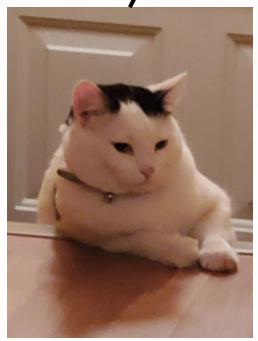
Similarity function



Mila

What are the inputs of the similarity function?

$$\text{sim}(z, z') = \frac{z \cdot z'}{\|z\| \|z'\|}$$

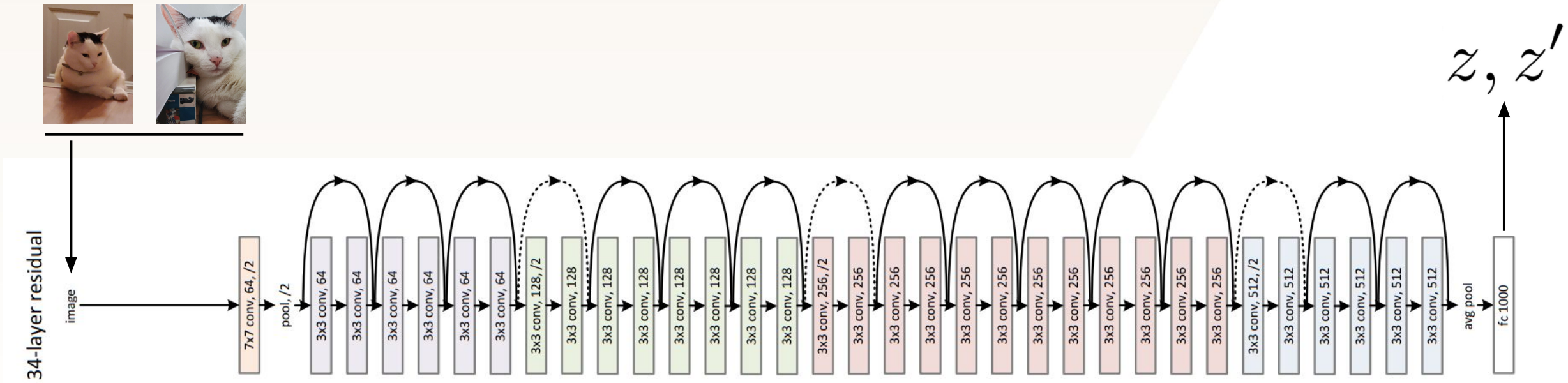


?

Mila

Backbone encoder

ResNet-50 or ResNet-161



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

How to select positive and negative examples?

$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

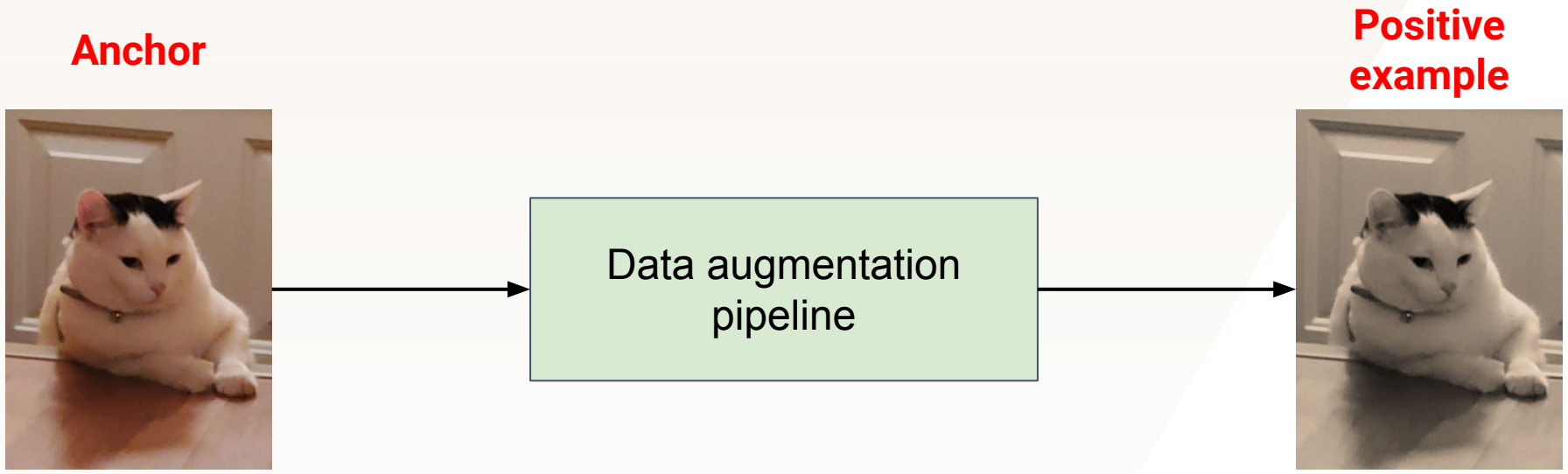
Anchor

Positive example

Negative examples

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

Data augmentation



Mila

Data augmentation

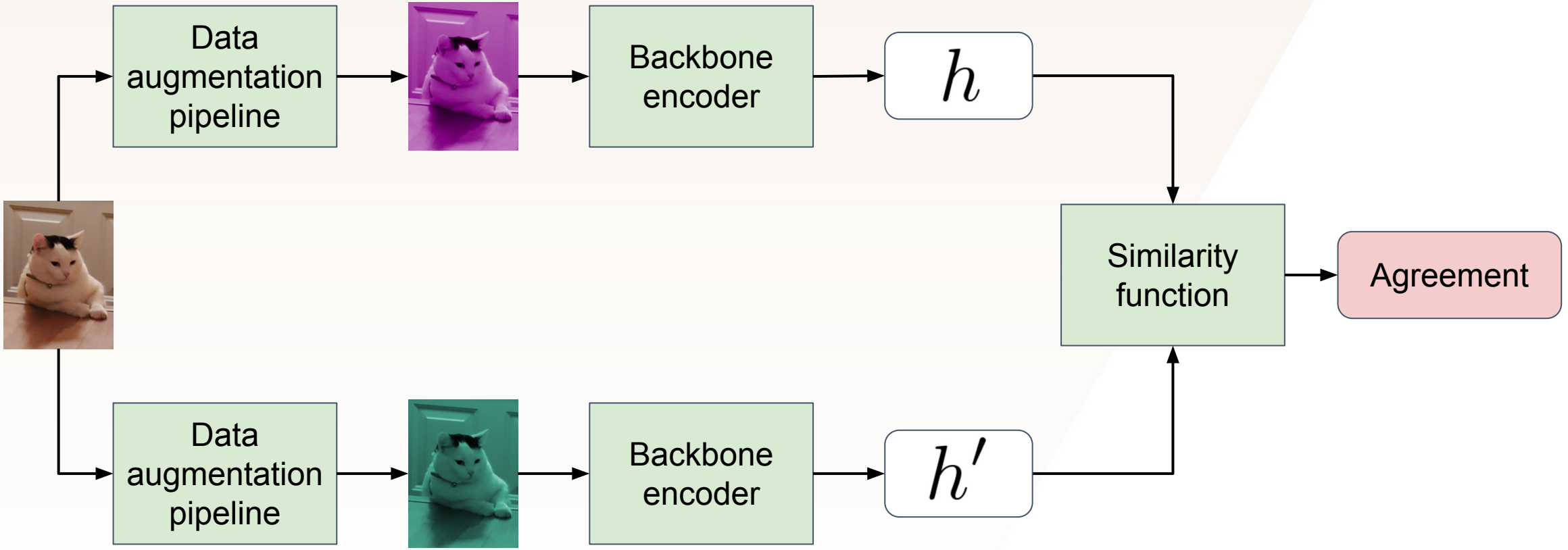
(a) Original (b) Crop and resize (c) Crop, resize (and flip) (d) Color distort. (drop) (e) Color distort. (jitter)

(f) Rotate $\{90°, 180°, 270°\}$ (g) Cutout (h) Gaussian noise (i) Gaussian blur (j) Sobel filtering

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

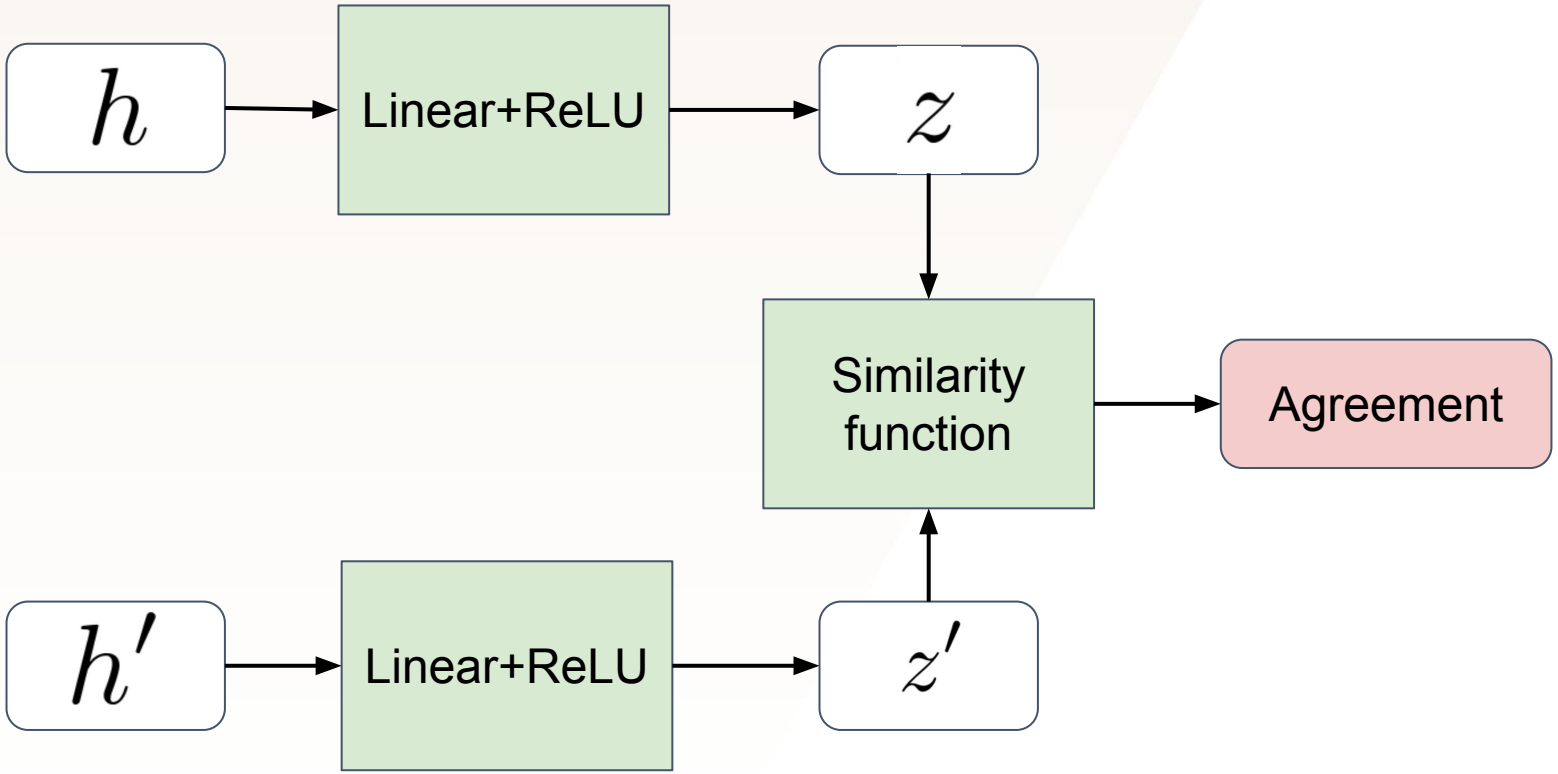
Putting everything together



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

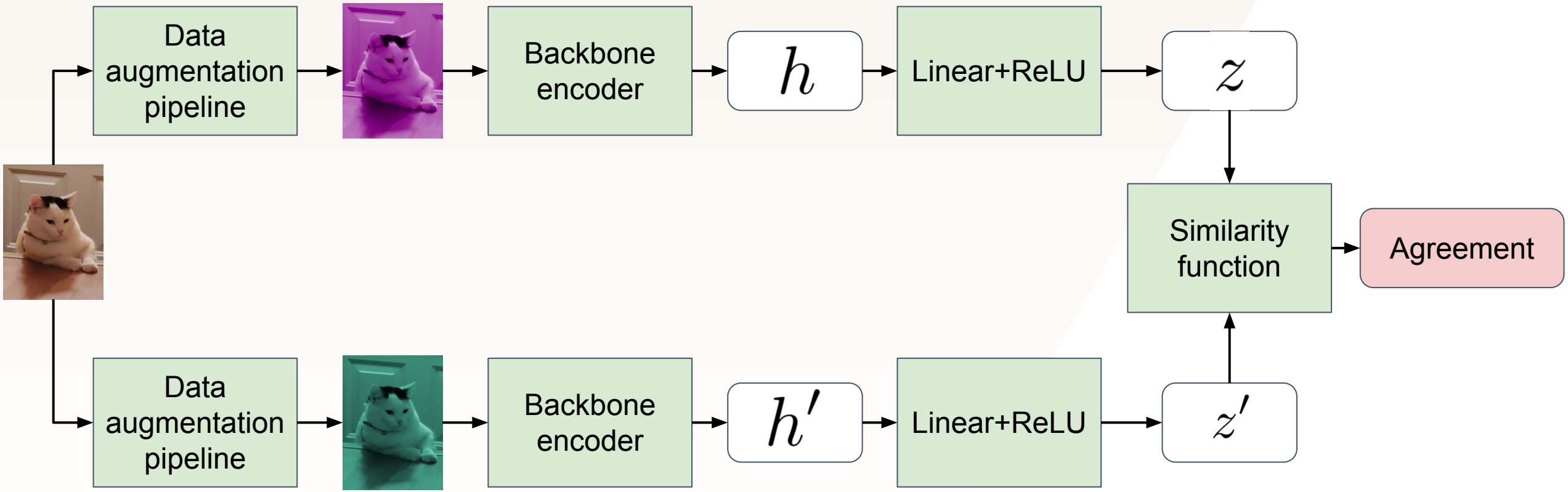
Projection trick

Add a nonlinear projection before computing the similarity



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

SimCLR



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

How to build the set of negative examples?

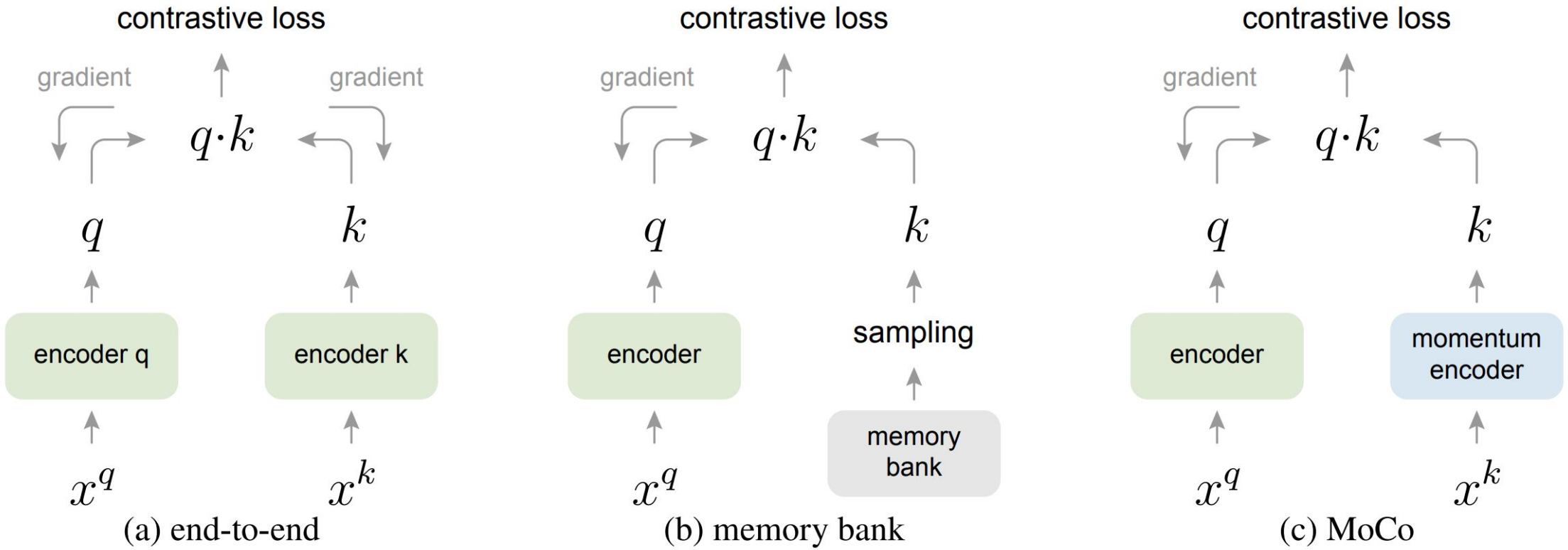
- SimCLR: large mini-batches (~8192) with LARS optimizer [1]
- Memory bank [2]
- Momentum Encoder (MoCo) [3]

$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Negative examples

[1] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
[2] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742
[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross B. Girshick: Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020: 9726-9735

Dictionary look-up interpretation



Momentum encoder

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

Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742
Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, Ross B. Girshick: Momentum Contrast for Unsupervised Visual Representation Learning. CVPR 2020: 9726-9735

SimCLR

Data augmentation pipeline

Backbone encoder

h

Linear+ReLU

z

Similarity function

Agreement

Data augmentation pipeline

Backbone encoder

h'

Linear+ReLU

z'

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

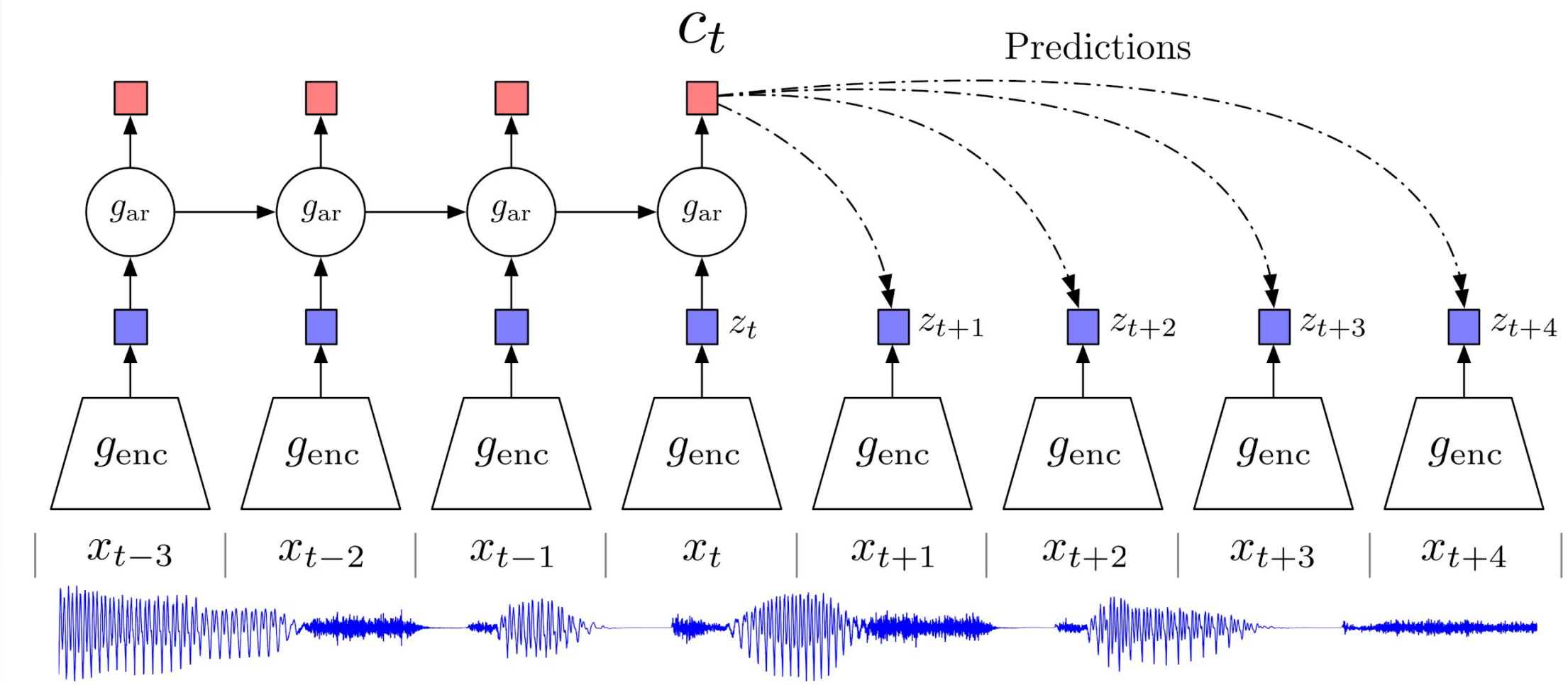
Discussion

- Contrastive techniques achieve better performances than supervised pre-training.
- They are computationally intensive and require huge unlabeled datasets.

05

Autoregressive contrastive learning

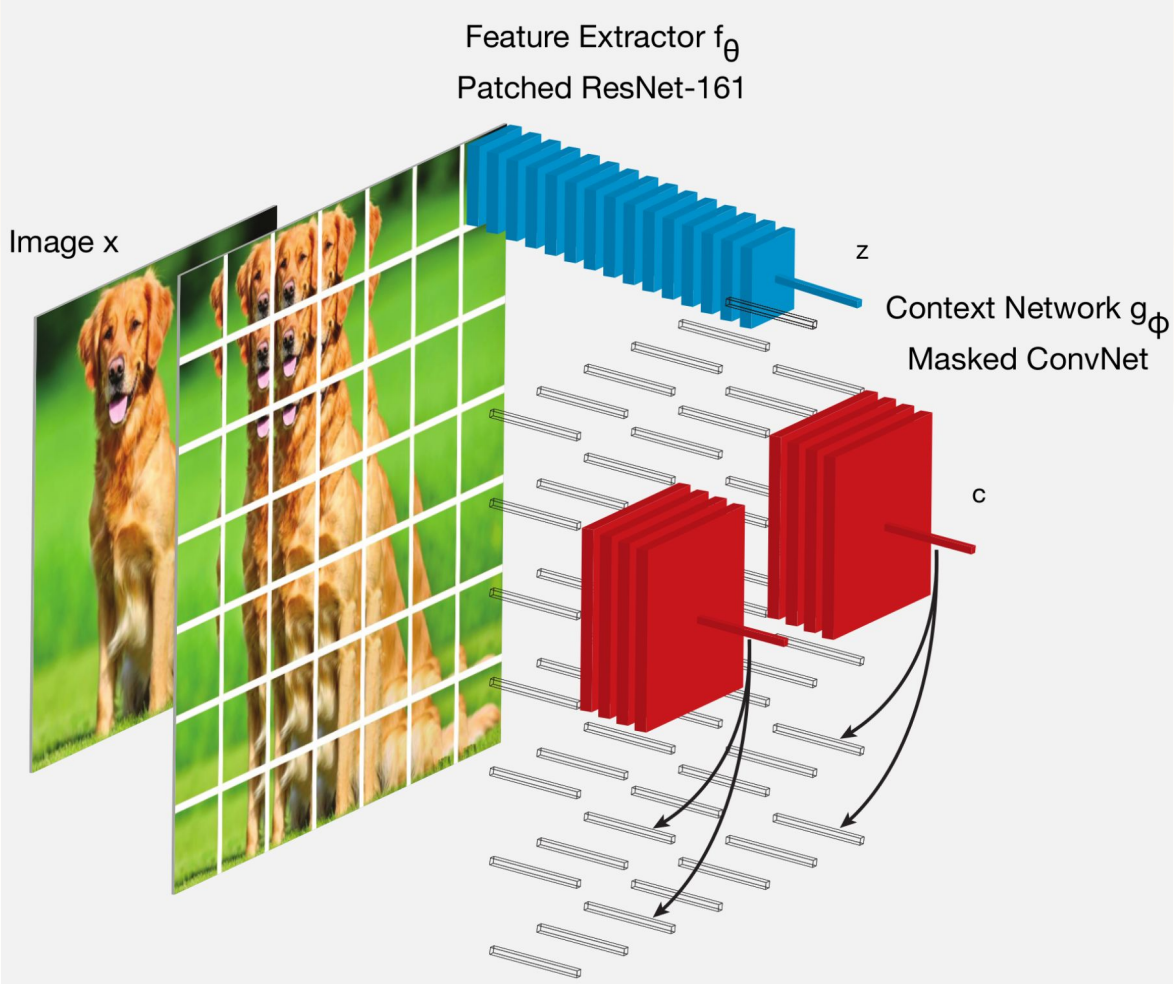
Contrastive Predictive Coding (CPC)



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

Mila

Contrastive Predictive Coding (CPCv2)



Olivier J. Hénaff et al. Data-Efficient Image Recognition with Contrastive Predictive Coding. ICML 2020: 4182-4192

Mila

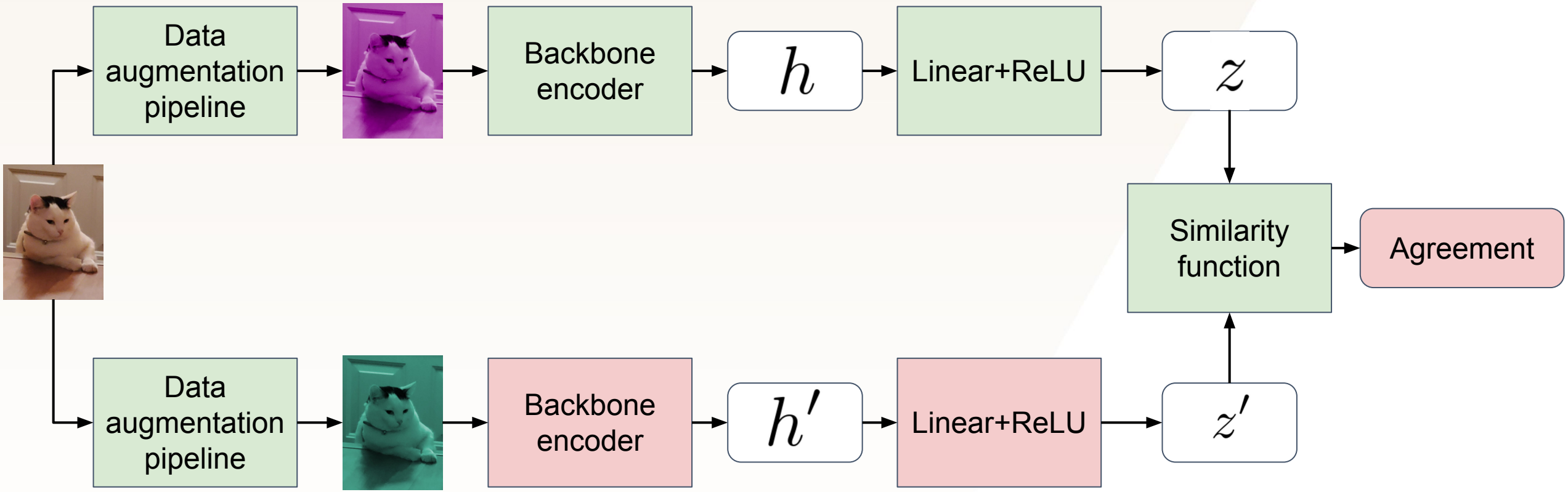
Discussion

- Require the user to fix an order, which may not be intuitive for some types of data such as images.

06

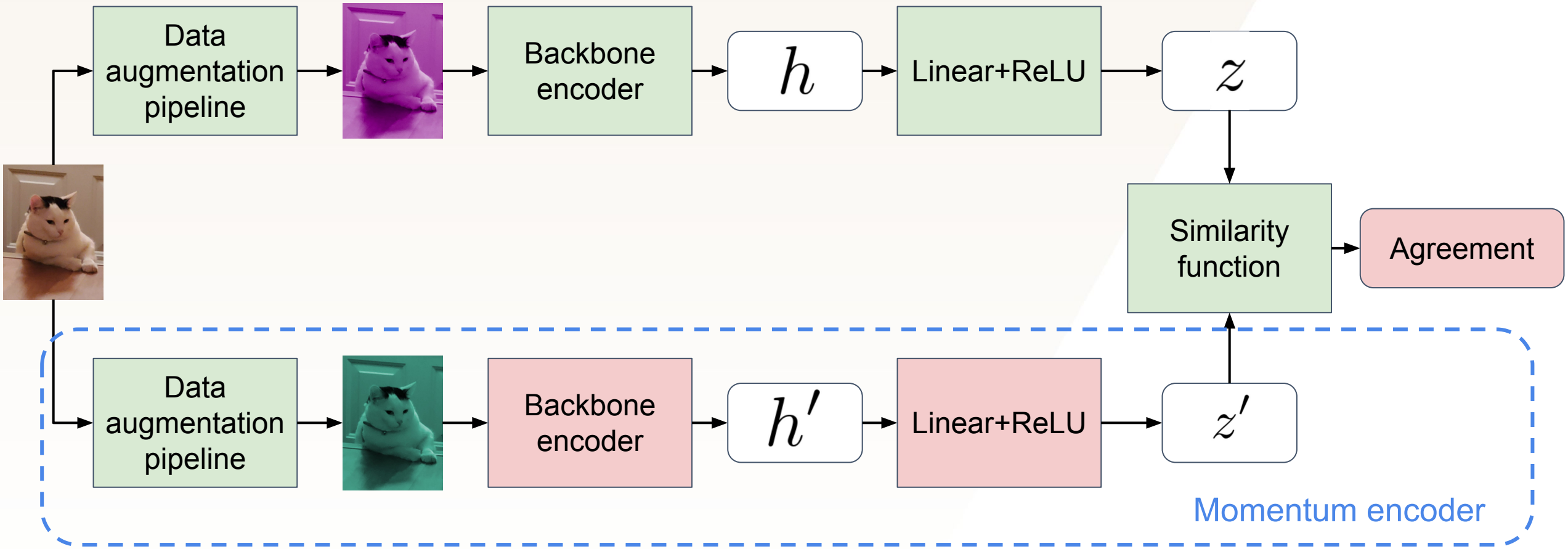
Contrastive learning without negative examples

Bootstrap Your Own Latent (BYOL)



Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko: Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. NeurIPS 2020

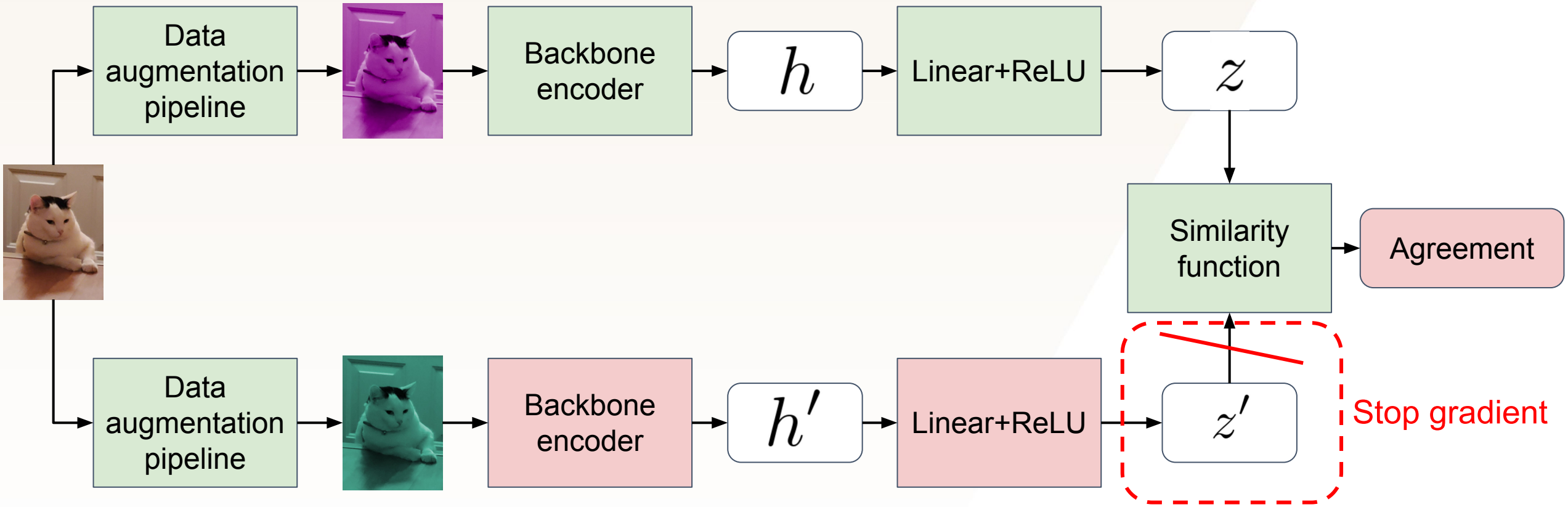
Bootstrap Your Own Latent (BYOL)



Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko: Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. NeurIPS 2020

Mila

Bootstrap Your Own Latent (BYOL)



Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H. Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Ávila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Rémi Munos, Michal Valko: Bootstrap Your Own Latent - A New Approach to Self-Supervised Learning. NeurIPS 2020

Mila

Discussion

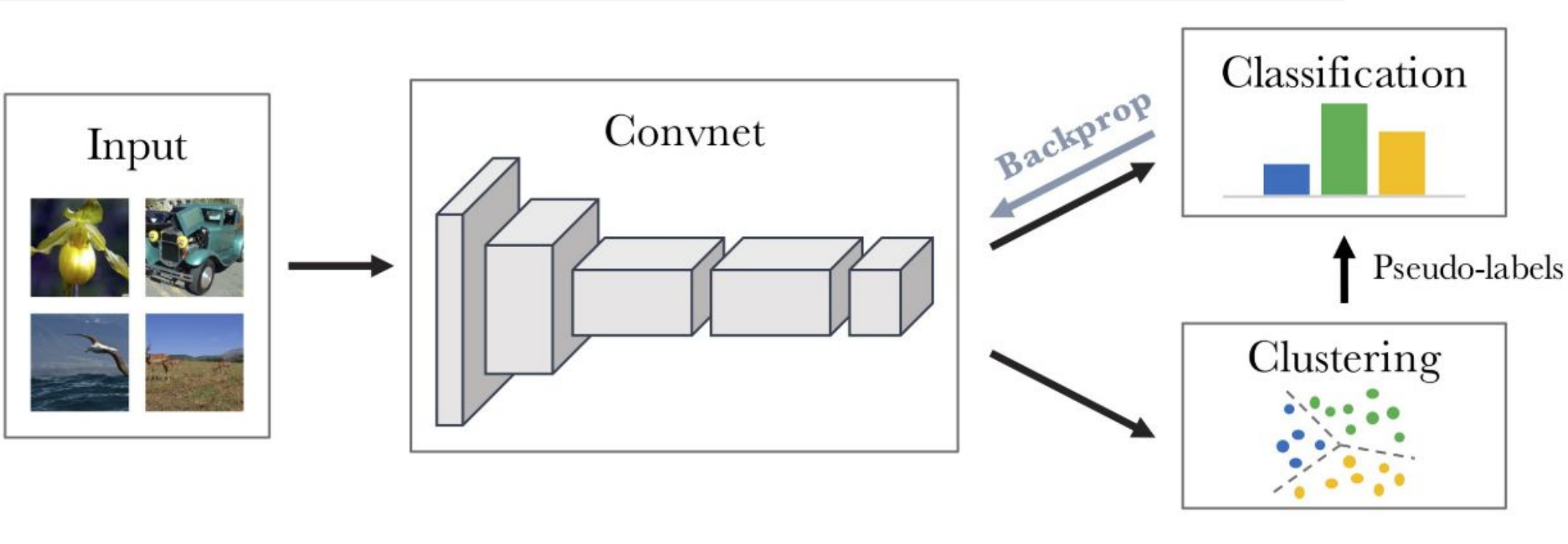
- Avoid the collapse problem with a momentum encoder.
- The momentum hyper-parameter has a huge effect on the representation quality.

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

07

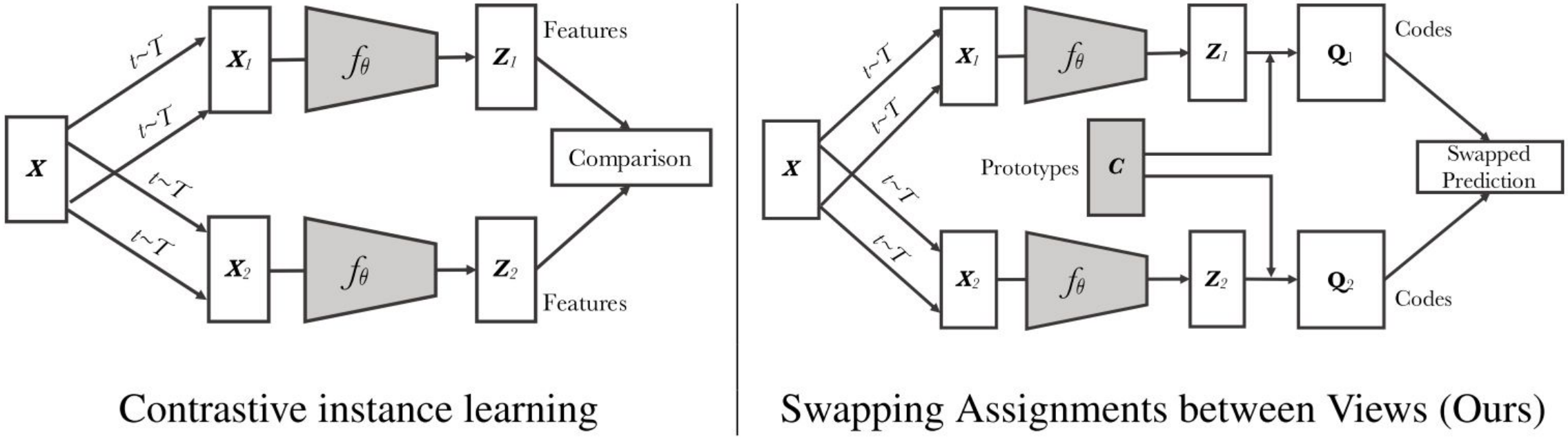
Contrastive learning with clustering

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)

Contrastive and clustering together



Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS 2020

Mila

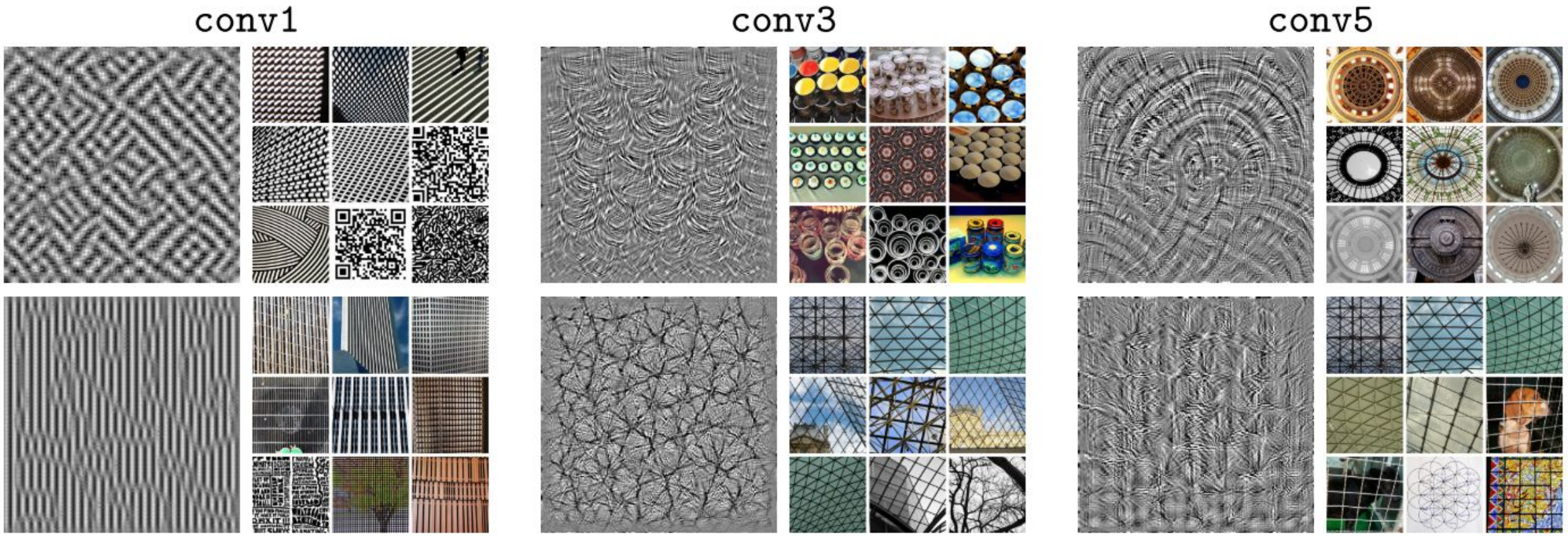
08

Empirical results

Qualitative assessment

Mila

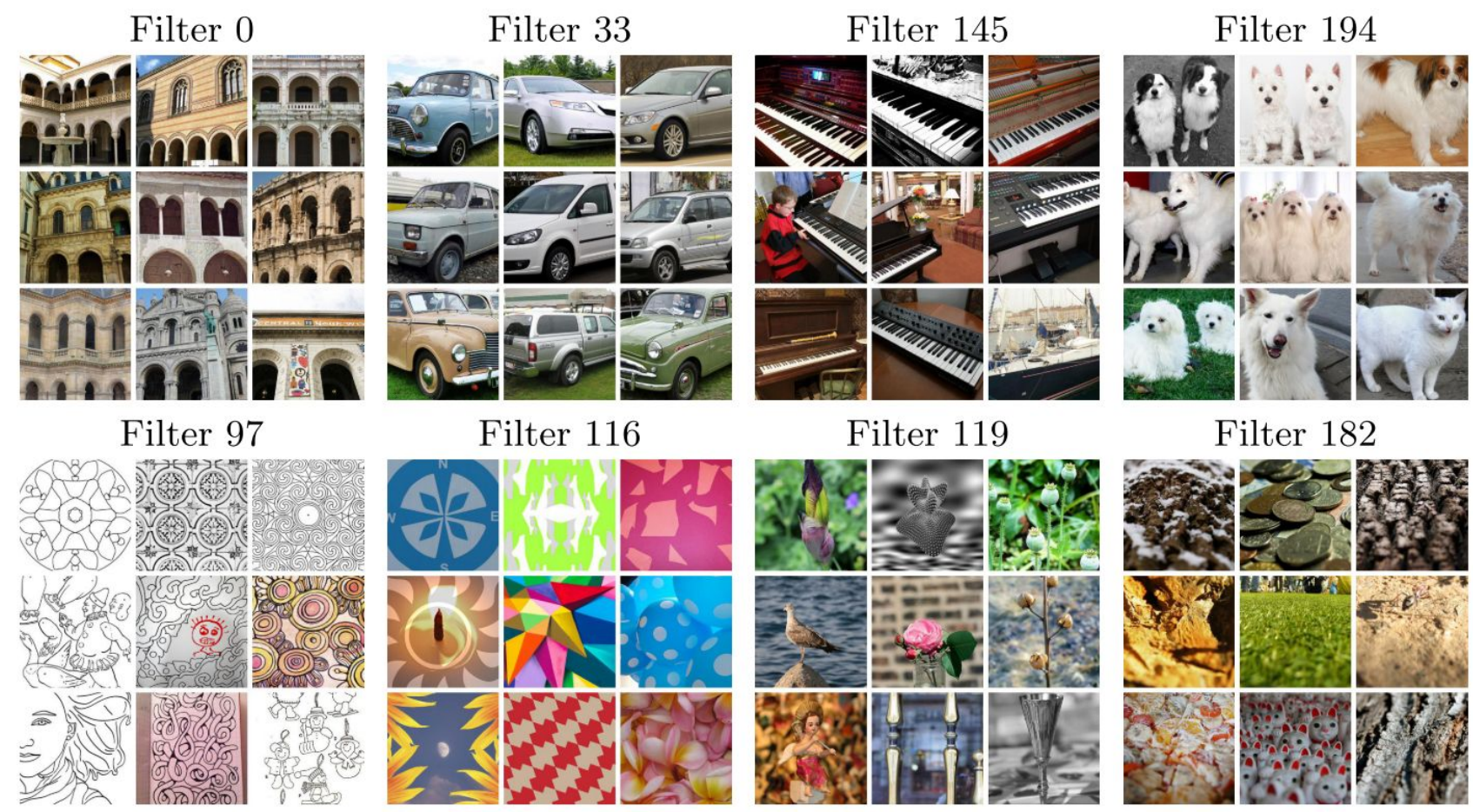
Learned representations



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

Mila

Learned representations

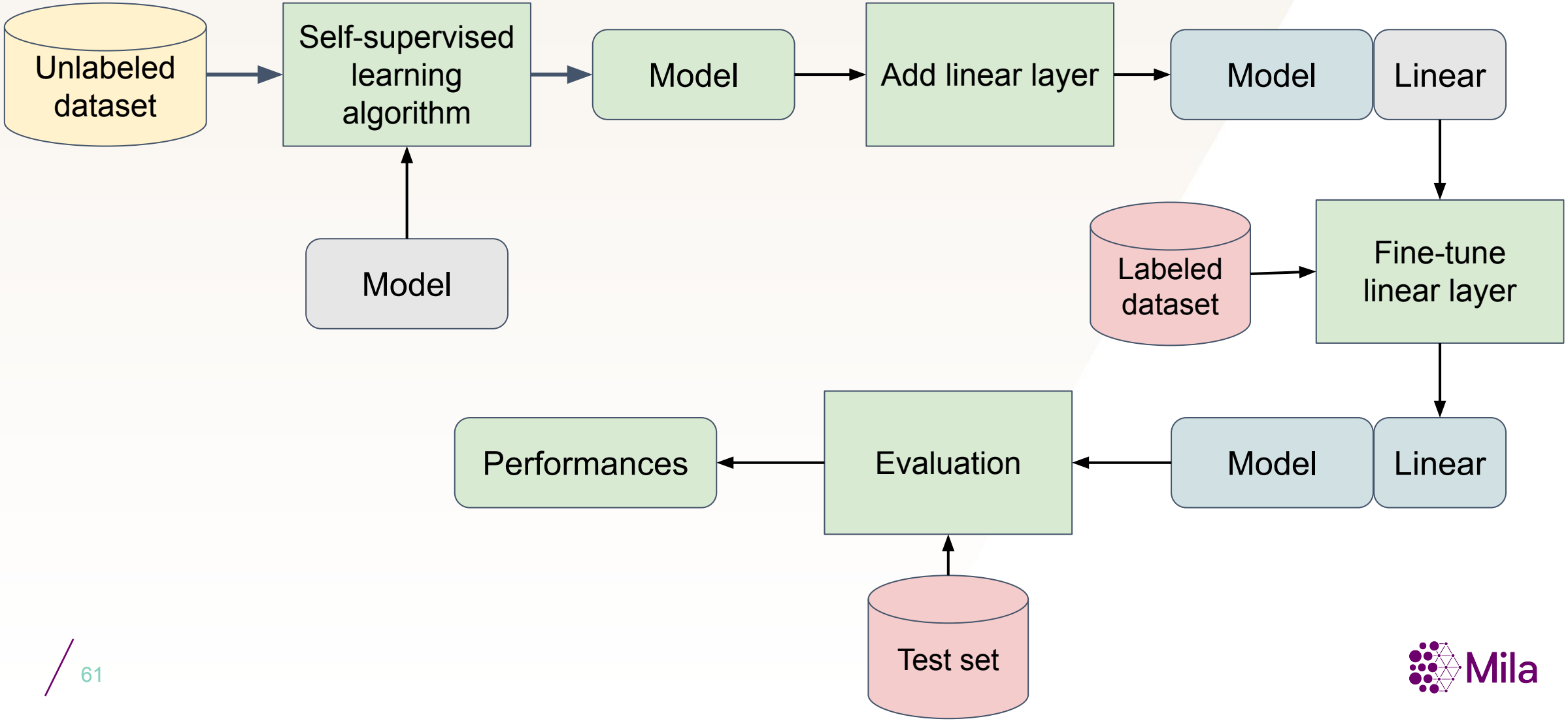


Caron, Mathilde, et al. "Deep clustering for unsupervised learning of visual features." Proceedings of the European Conference on Computer Vision (ECCV) (2018)

Mila

Linear evaluation protocol

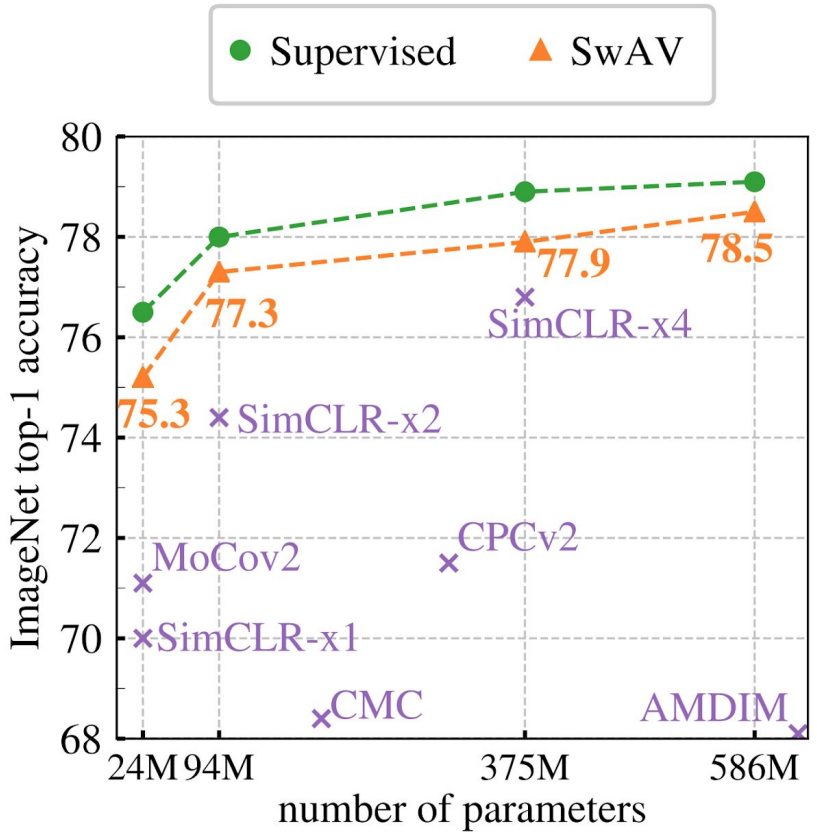
Linear evaluation



Mila

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**

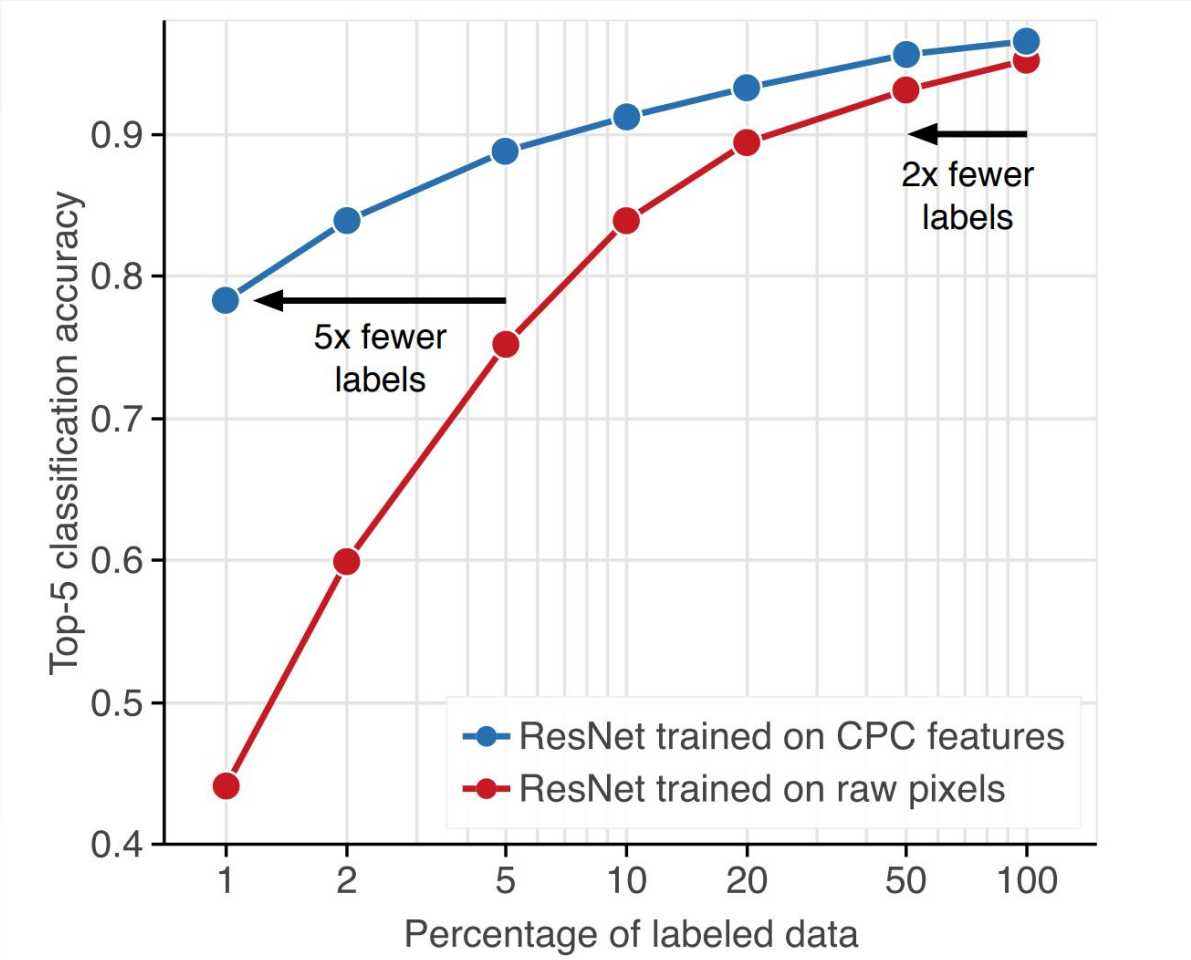


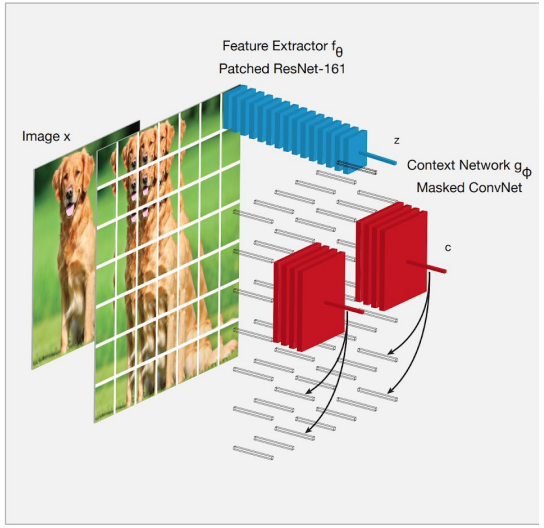
Method	Top-1	Top-5
Local Agg.	60.2	-
PIRL [35]	63.6	-
CPC v2 [32]	63.8	85.3
CMC [11]	66.2	87.0
SimCLR [8]	69.3	89.0
MoCo v2 [37]	71.1	-
InfoMin Aug. [12]	73.0	91.1
BYOL (ours)	**74.3**	**91.6**

(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).

Empirical results: CPC v2





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

Mila

Semi-supervised learning

Mila

Empirical results: BYOL

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.

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

Mila

Empirical results: BYOL

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

- Several works are inspired from AMDIM.
- Computation budget and datasets can be huge.
- Pytorch lightning implements a SSL framework for fast prototyping.
- The field is moving very fast. Stay up-to-date!
  - SSL methods in vision depend heavily on data augmentations.
  - How to automate data augmentation discoveries?
- SSL has a significant impact on democratizing DL.

Philip Bachman, R. Devon Hjelm, William Buchwalter: Learning Representations by Maximizing Mutual Information Across Views. NeurIPS 2019: 15509-15519
https://github.com/PyTorchLightning/pytorch-lightning

09

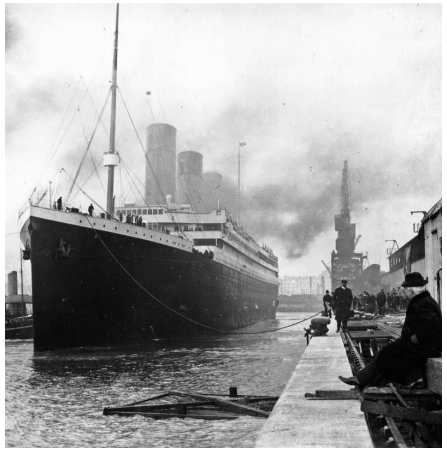
Questions

Tutorial 2: categorical data with MLP

Task: predict whether or not a passenger survived the Titanic disaster based on passenger data only.

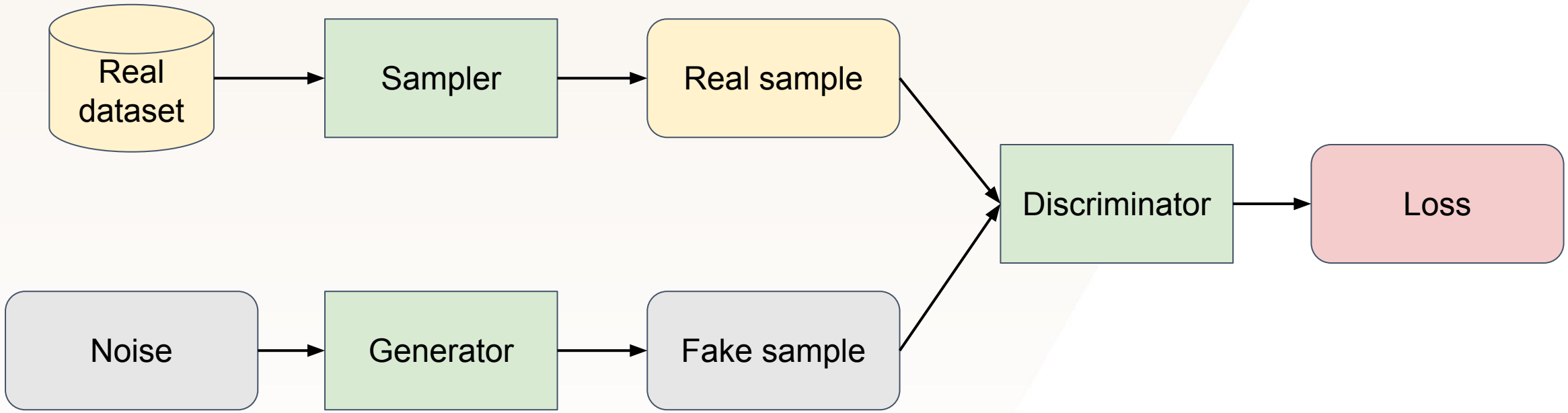
Objectives:

- identify and preprocess the relevant features for the task,
- define a multilayer perceptron model,
- train the model and analyze the results.



Source: Wikimedia commons

Generative adversarial network



Mila