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

Gaétan Marceau Caron

01

Representation learning

What is a representation?

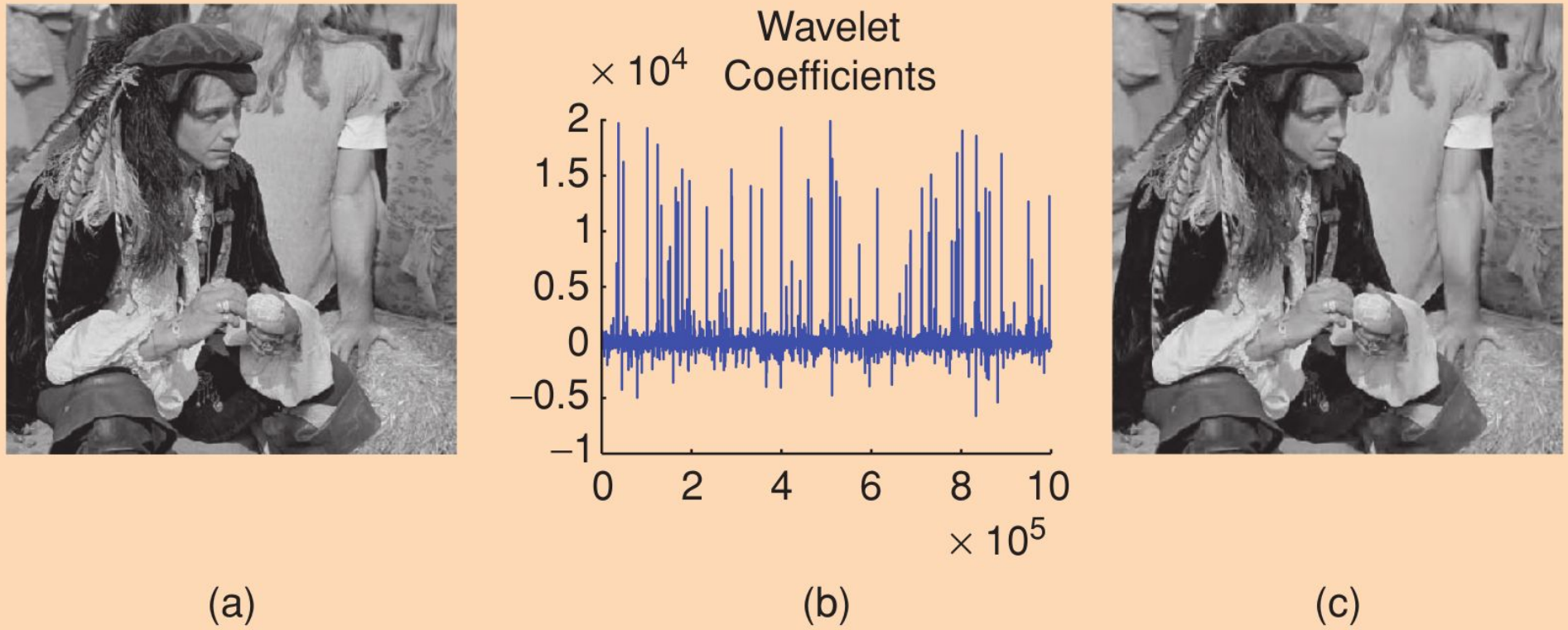


Image credit: Candès, Emmanuel J., and Michael B. Wakin. "An introduction to compressive sampling." IEEE signal processing magazine 25.2 (2008): 21-30.

Mila

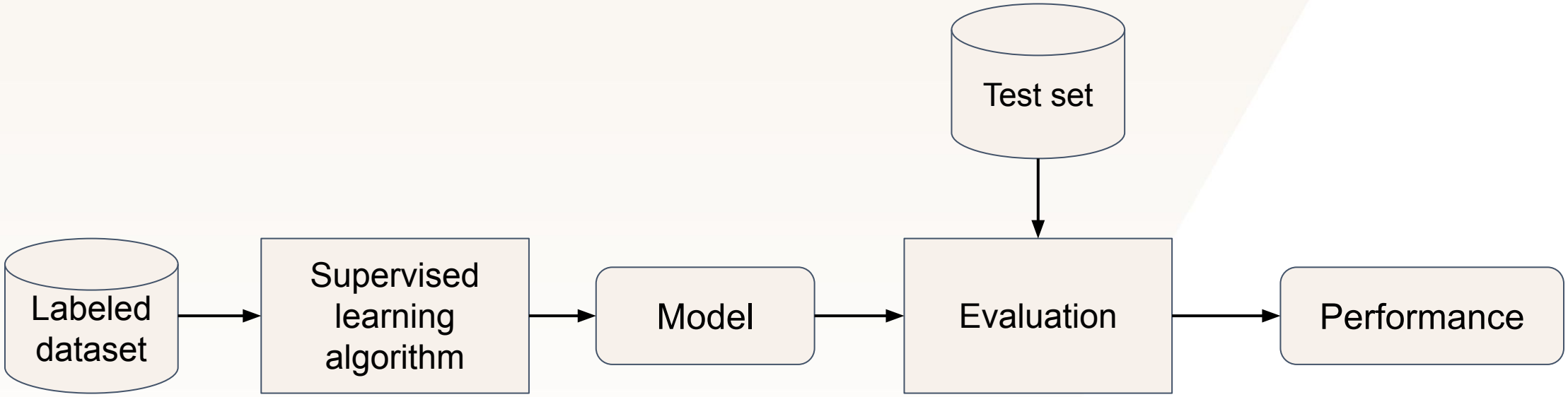
What is a good representation?

Some properties of good representations:

- Compact
- Useful
- Expressive
- Robust to perturbations
- Interpretable

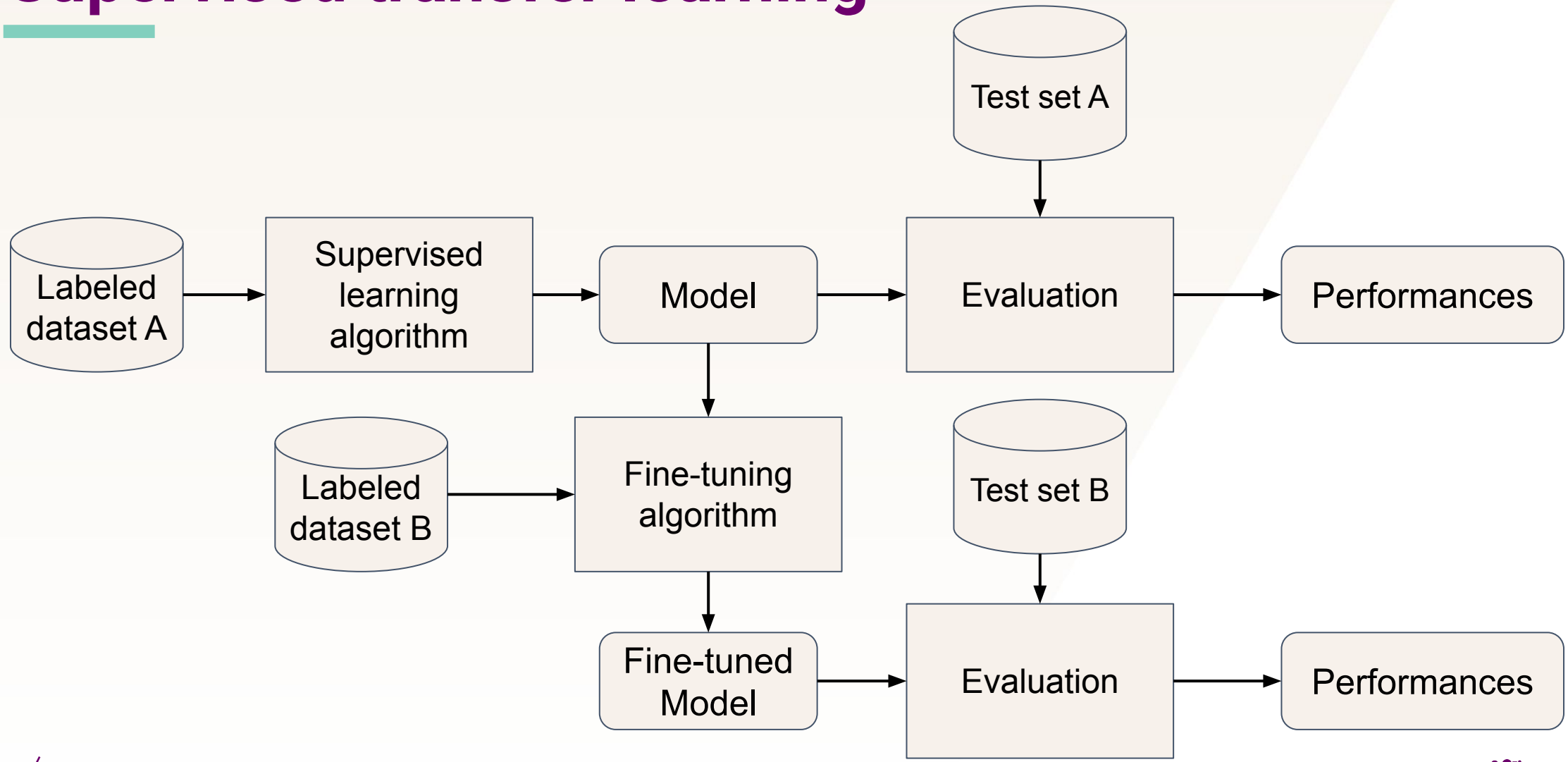
Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

Supervised learning



Mila

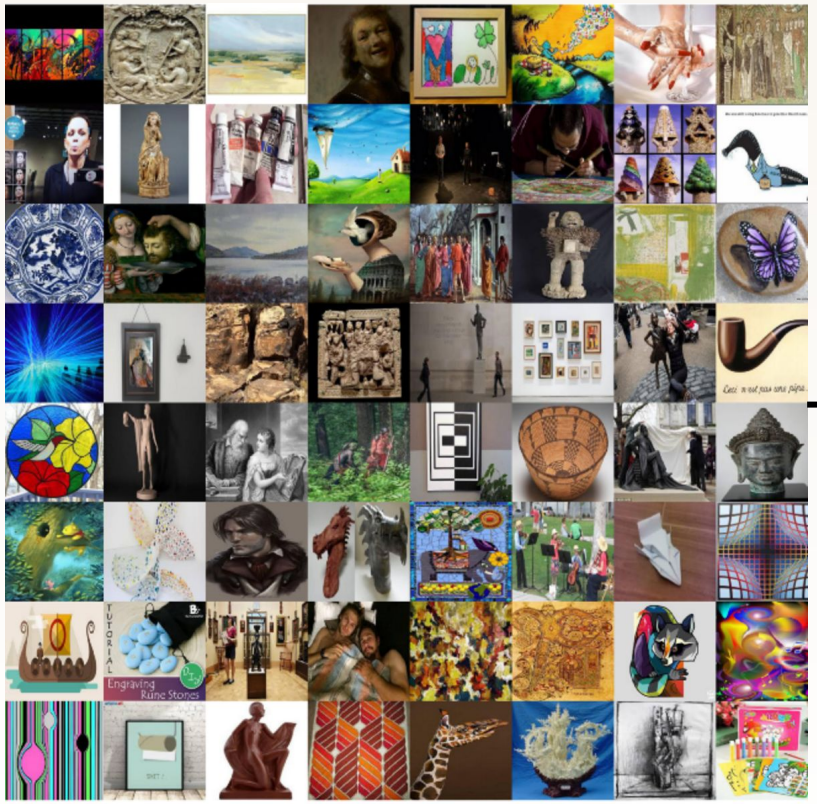
Supervised transfer learning



Mila

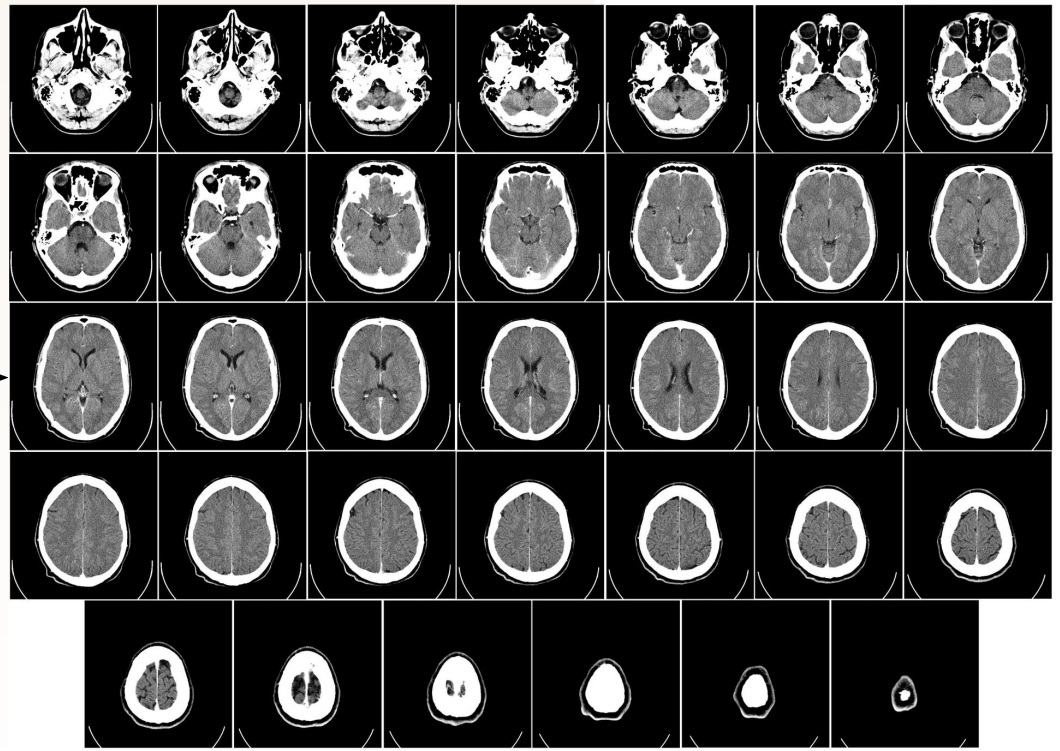
Limitations of supervised pre-training

Pre-training on ImageNet



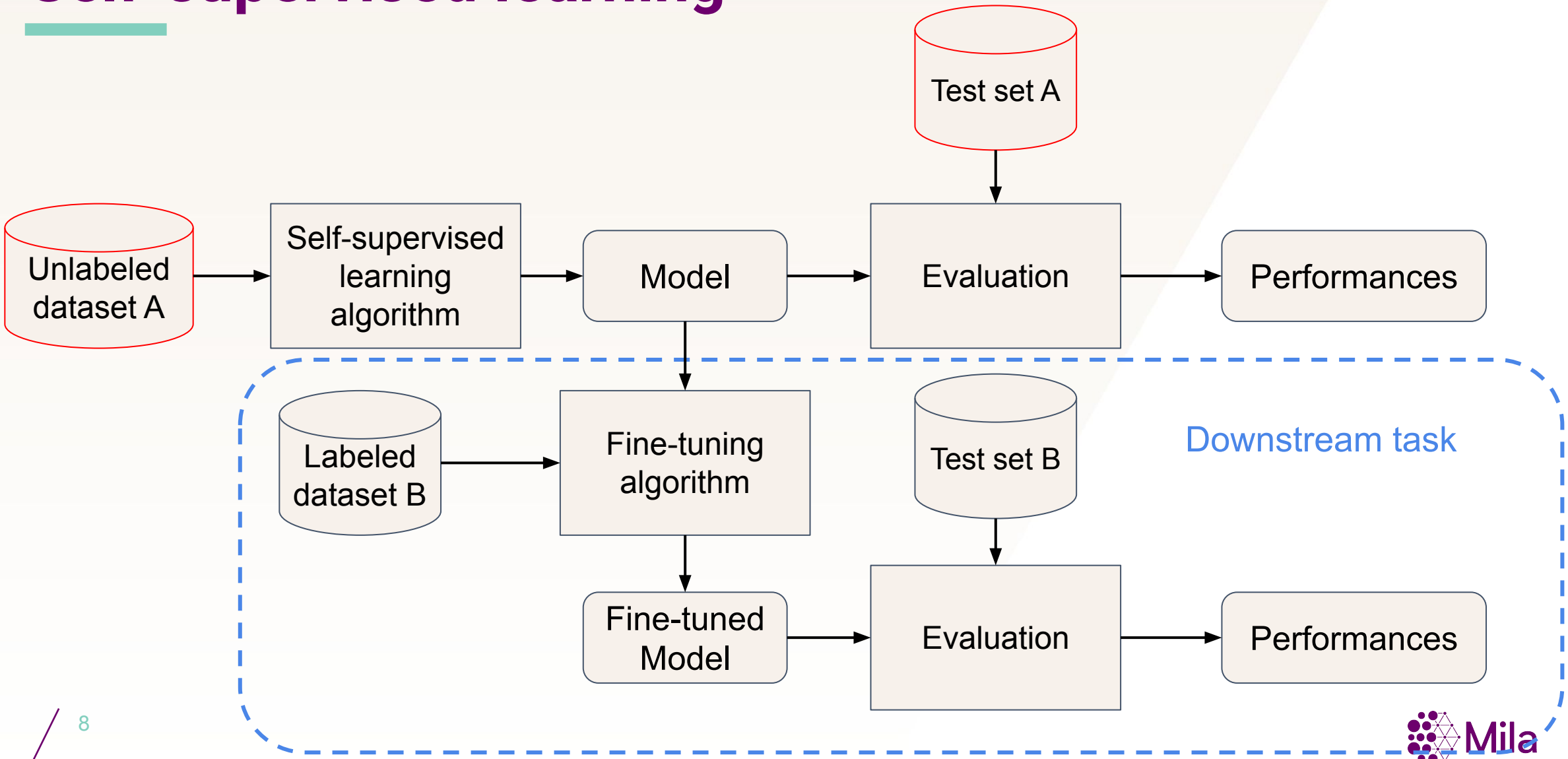
Transfer?

Downstream task on CT scan



Mila

Self-supervised learning



Mila

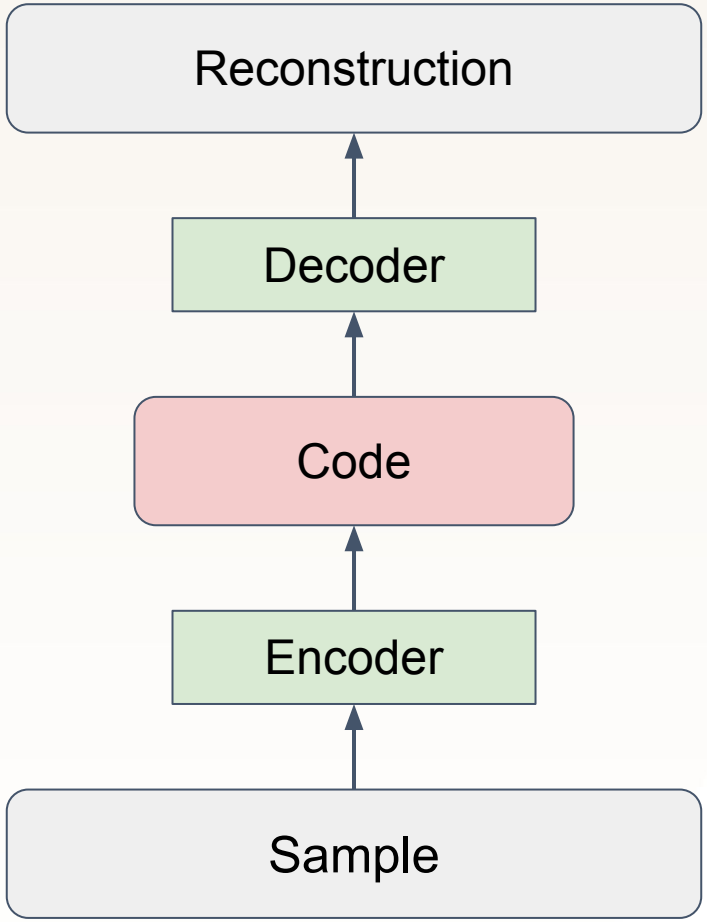
Objective

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

02

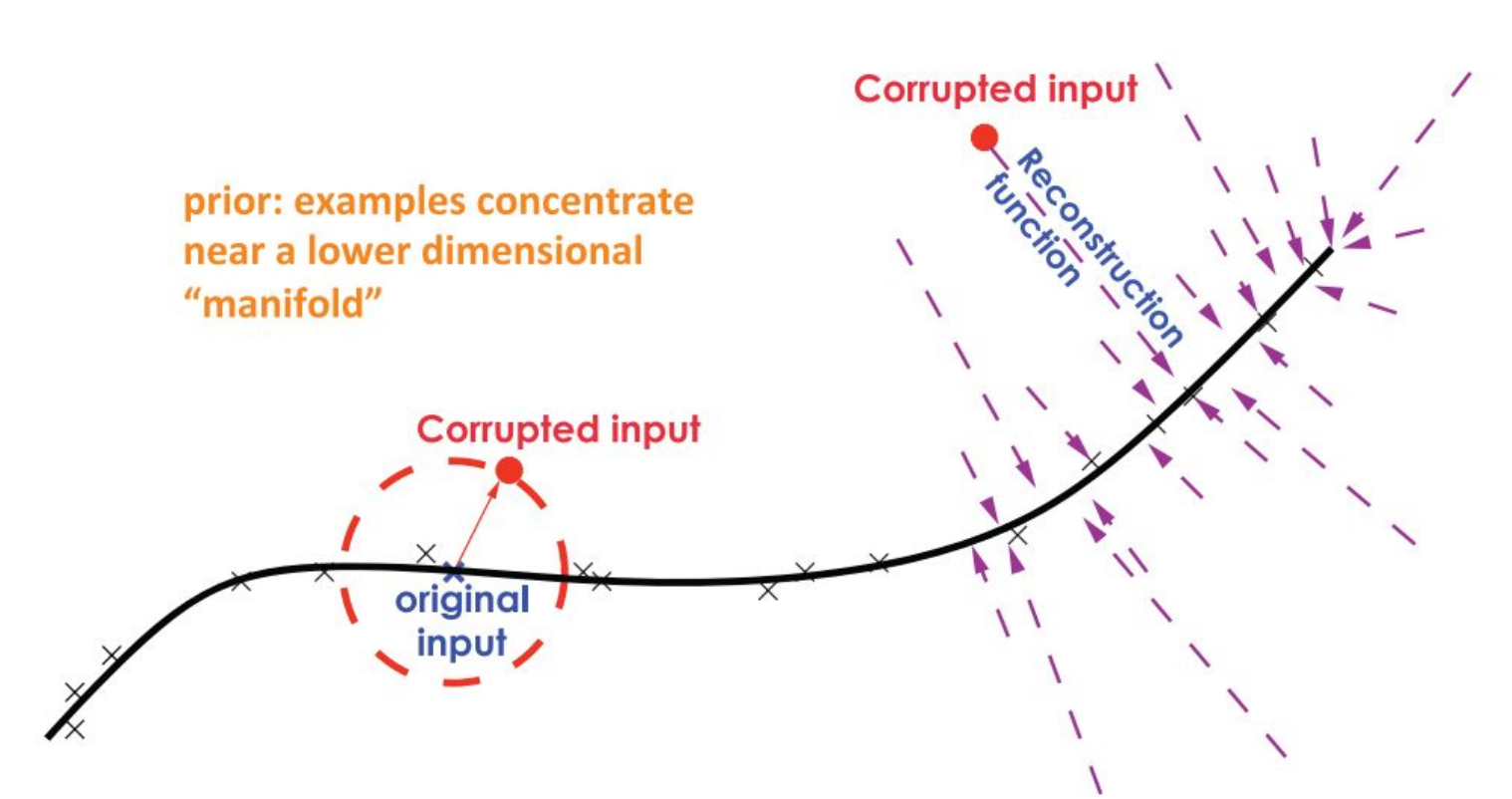
Pretext tasks with reconstruction

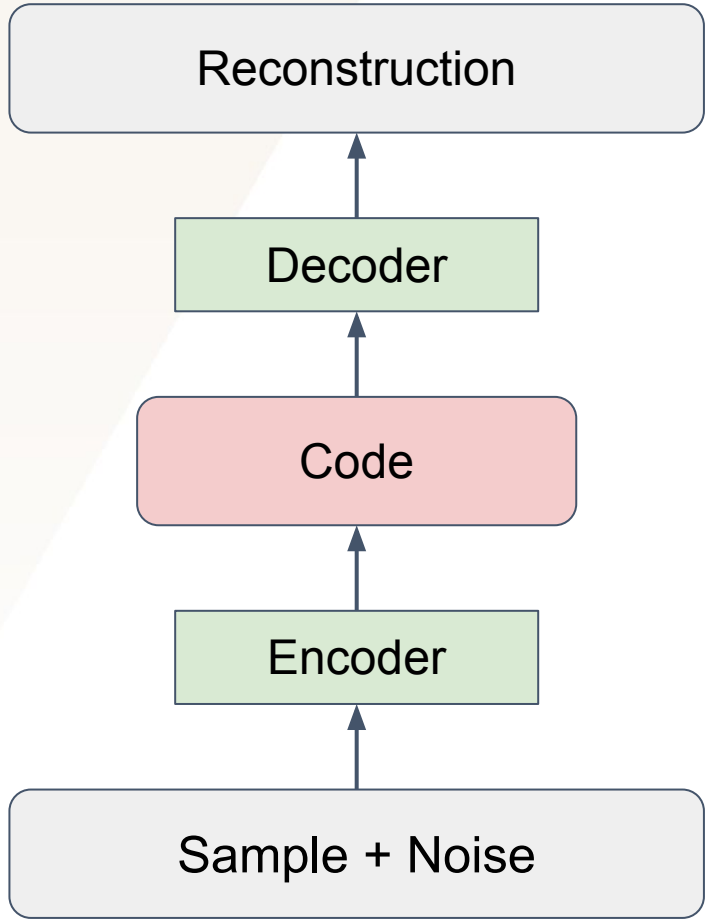
Auto-encoder



Mila

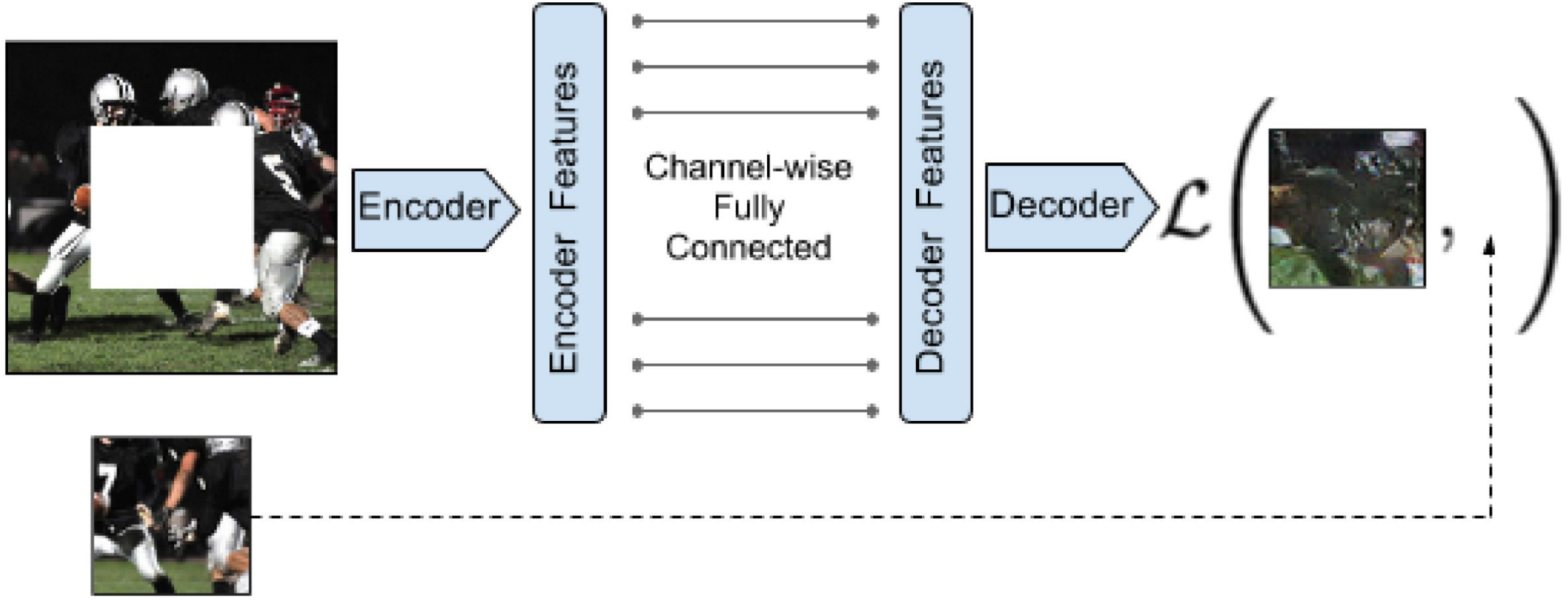
Denoising auto-encoder





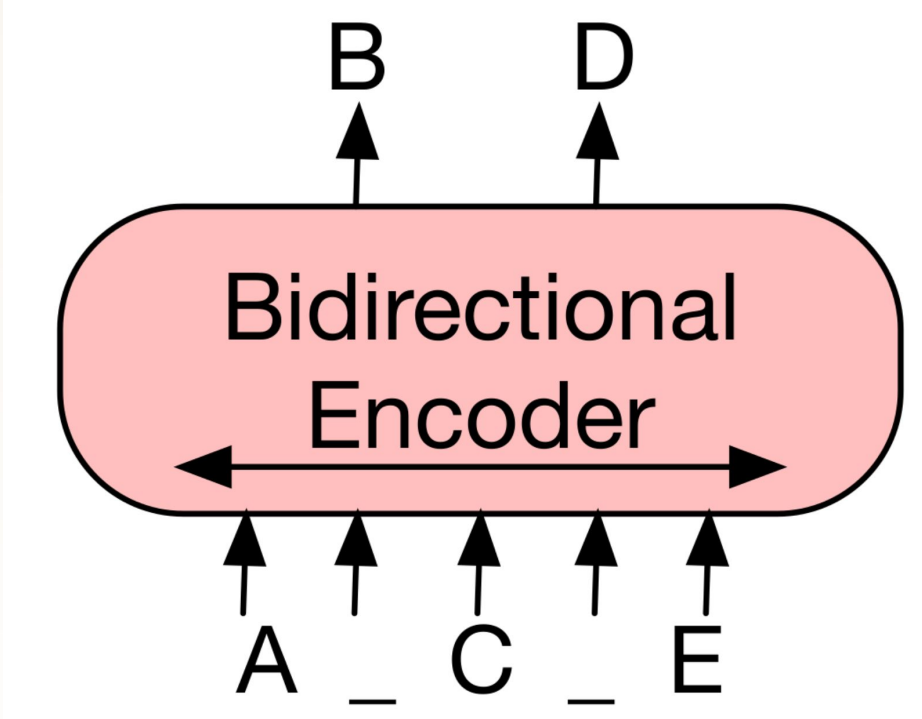
Bengio, Yoshua, Aaron Courville, and Pascal Vincent. "Representation learning: A review and new perspectives." IEEE transactions on pattern analysis and machine intelligence 35.8 (2013): 1798-1828.

Context encoders



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

Reconstruction in NLP



BERT

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

Colorization

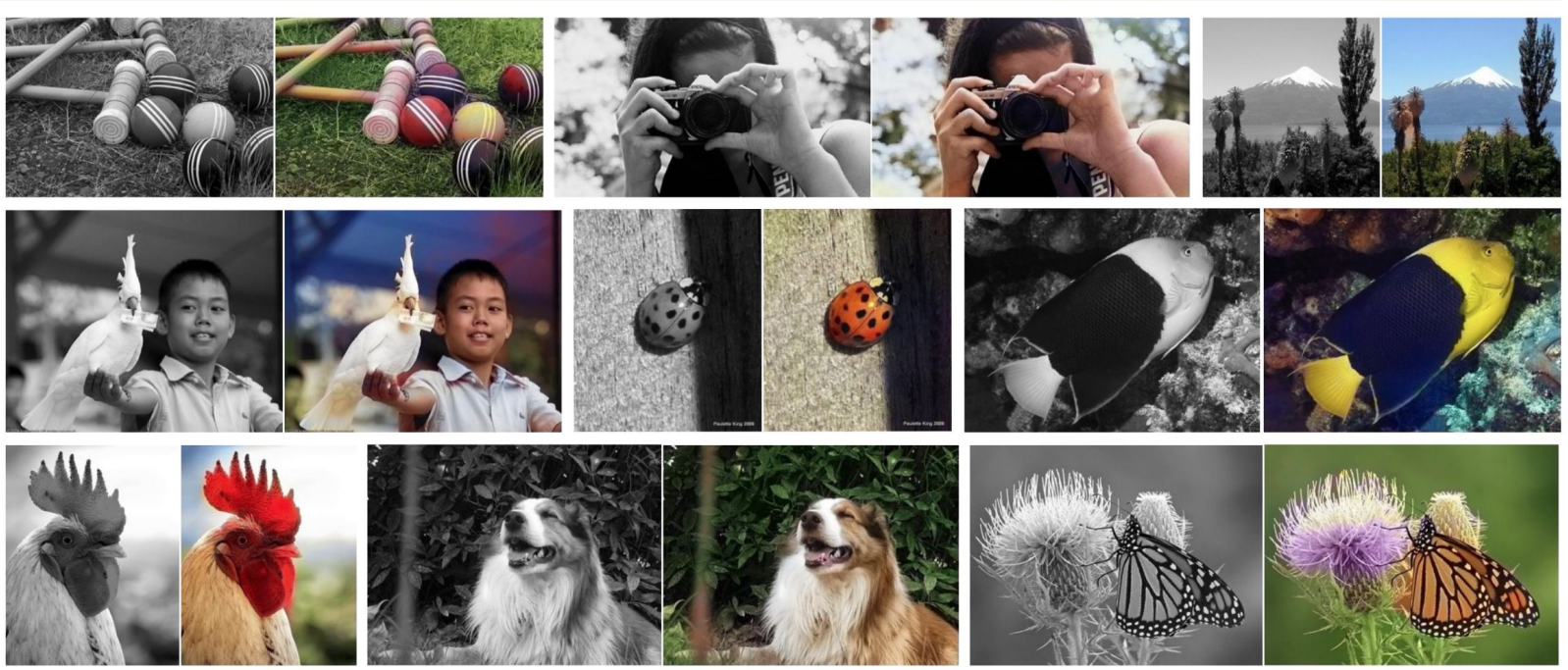
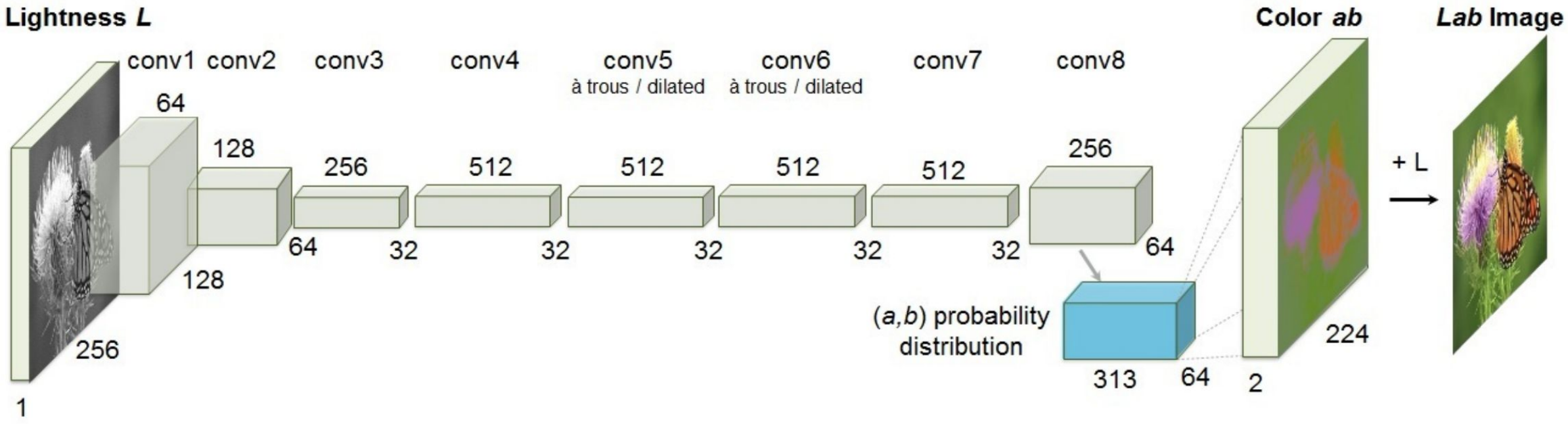


Fig. 1. Example input grayscale photos and output colorizations from our algorithm. These examples are cases where our model works especially well. Please visit `http://richzhang.github.io/colorization/` to see the full range of results and to try our model and code. Best viewed in color (obviously).

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

Colorization



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

Mila

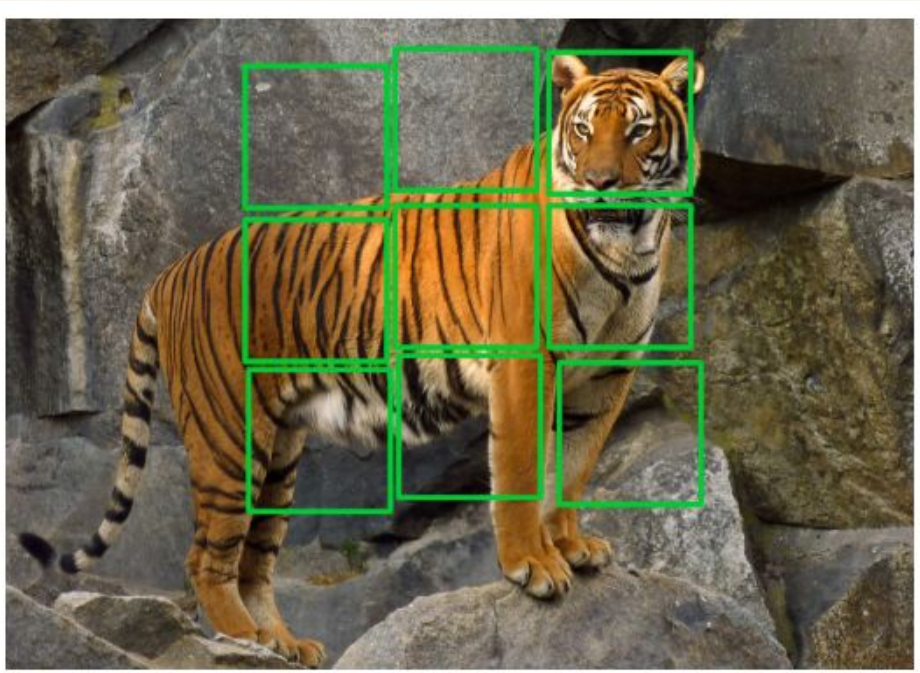
Discussion

- Make predictions in the high-dimensional input space.
- Work better for sequences of symbols than audio signals or images.

03

Pretext tasks with pseudo-labels

Jigsaw puzzle



(a)

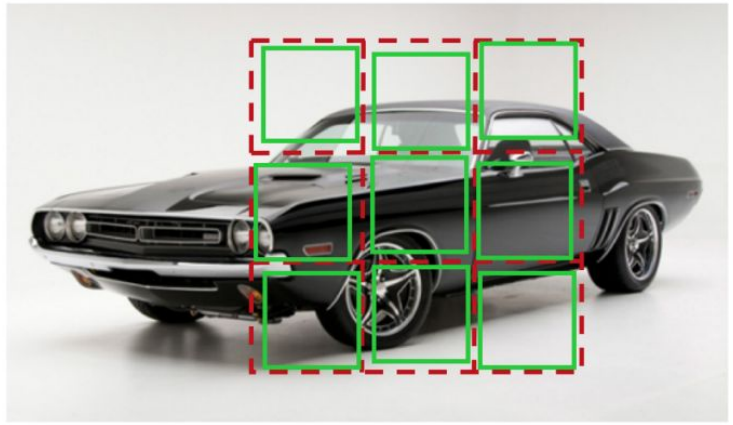
(b)

(c)

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

Mila

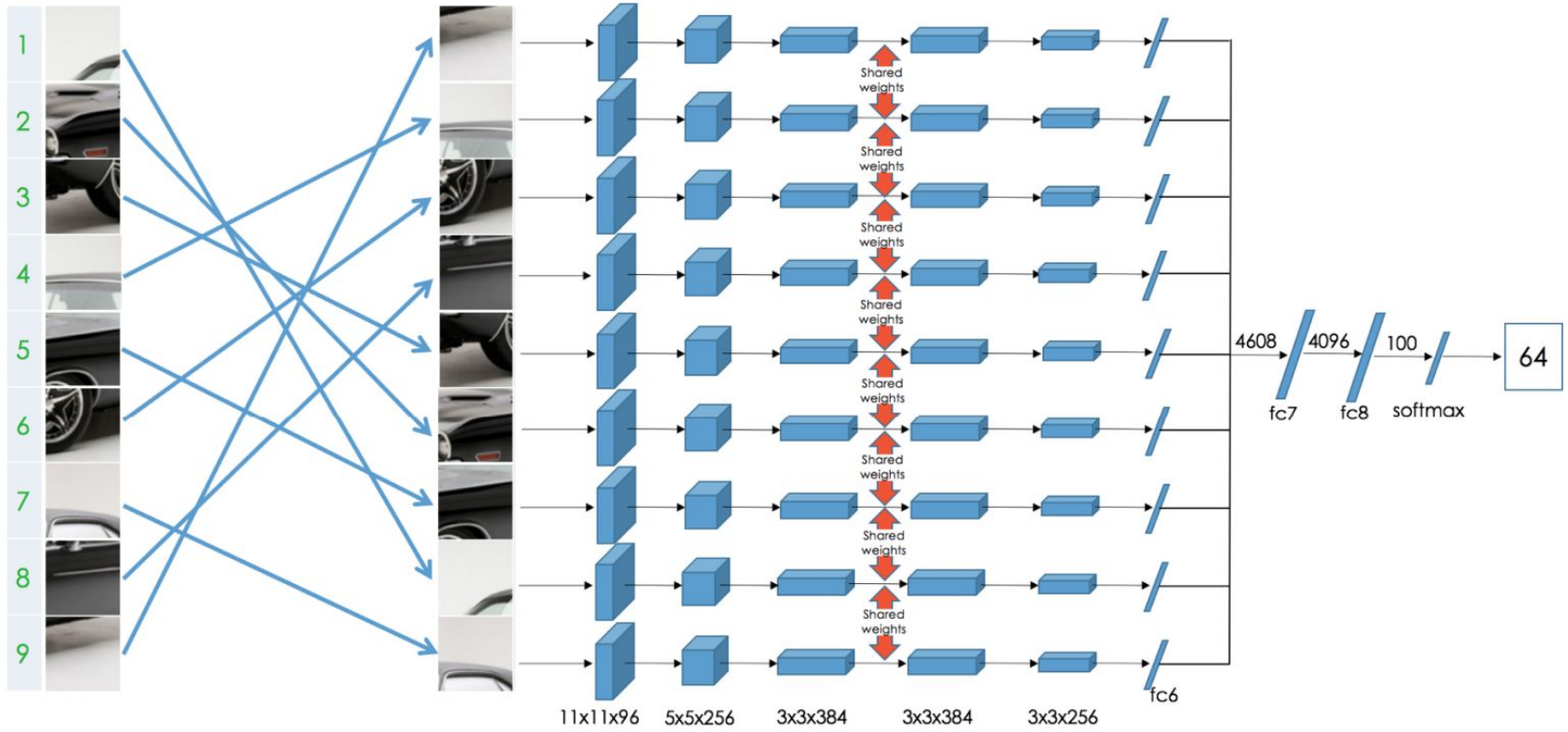
Jigsaw puzzle



Permutation Set

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

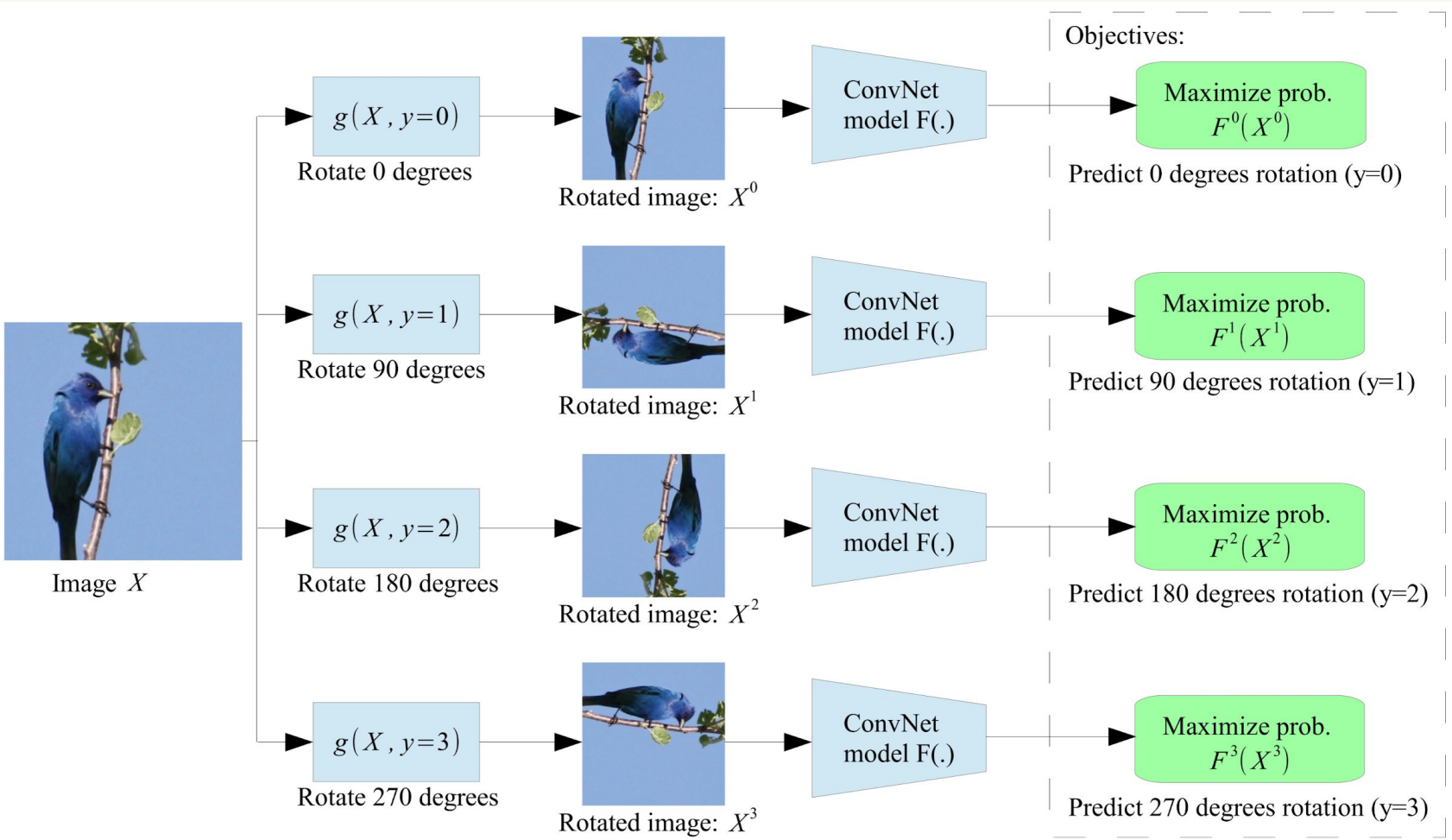
Reorder patches according to the selected permutation



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

Mila

RotNet



Spyros Gidaris, Praveer Singh, Nikos Komodakis: Unsupervised Representation Learning by Predicting Image Rotations. ICLR (Poster) 2018

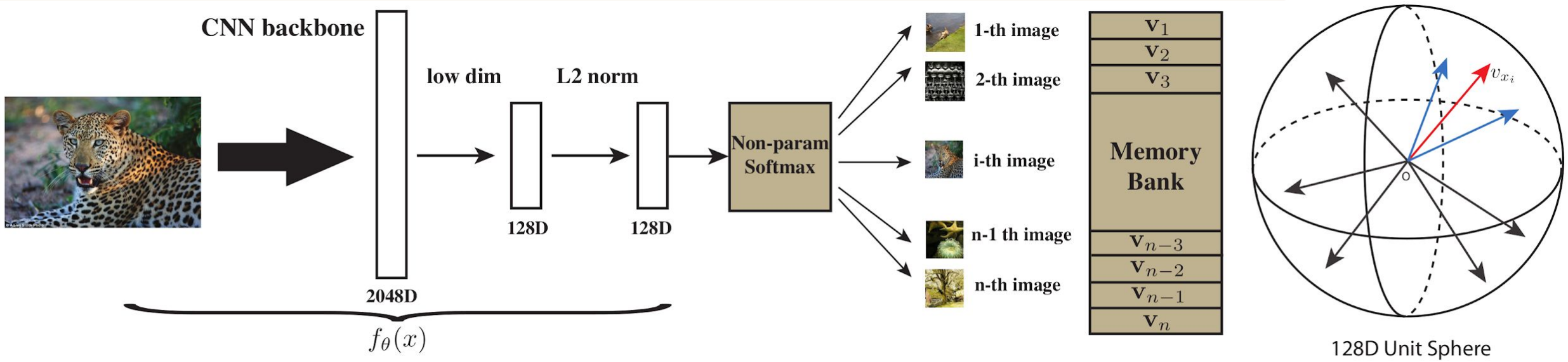
Discussion

- The pseudo-labels are generated with task-specific prior knowledge.
- Not competitive anymore compared to the following methods.

04

Contrastive learning

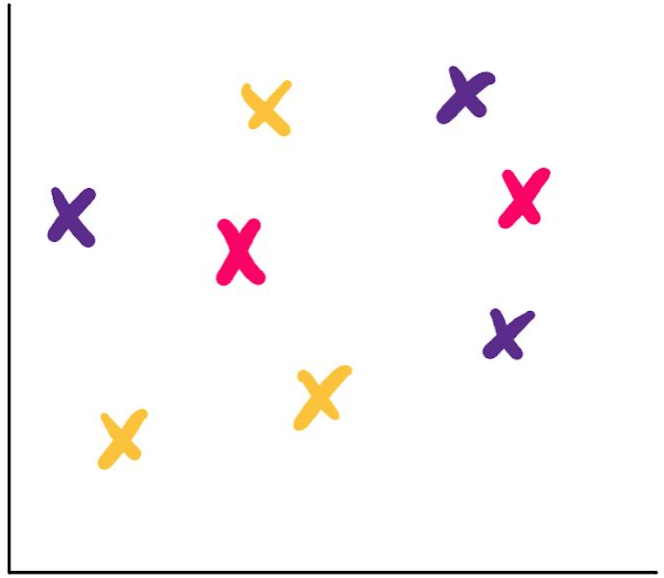
Instance discrimination

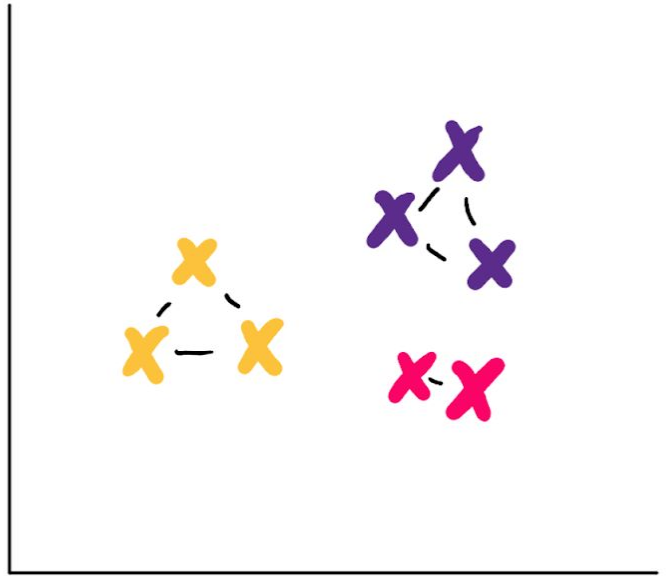


Non-parametric softmax: $$P(i|\mathbf{v}) = \frac{\exp\left(\mathbf{v}_i^T \mathbf{v}/\tau\right)}{\sum_{j=1}^{n} \exp\left(\mathbf{v}_j^T \mathbf{v}/\tau\right)}$$

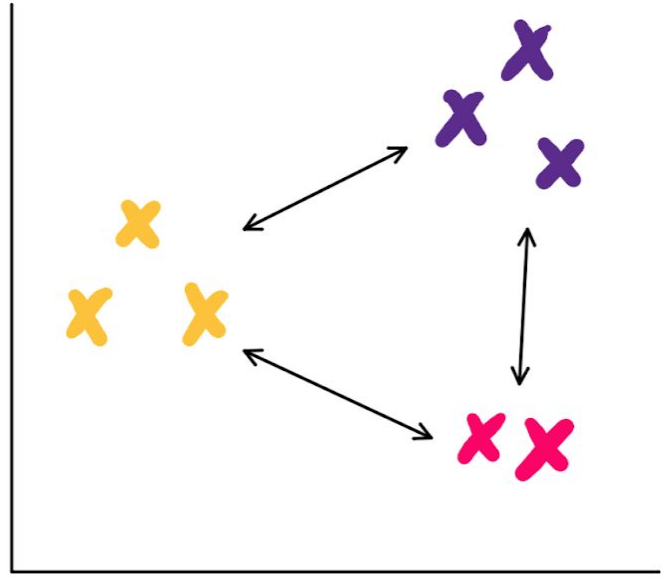
Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742

Contrastive learning





attract similar data



repulse
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

Similarity function

Dot product

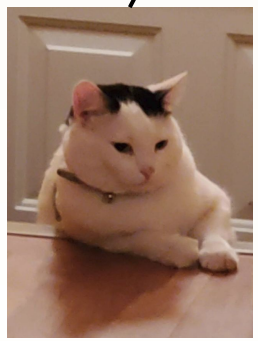
$$\mathrm{sim}(z, z') = \frac{z \cdot z'}{\|z\| \|z'\|}$$

l2-Norm

Mila

What are the inputs of the similarity function?

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

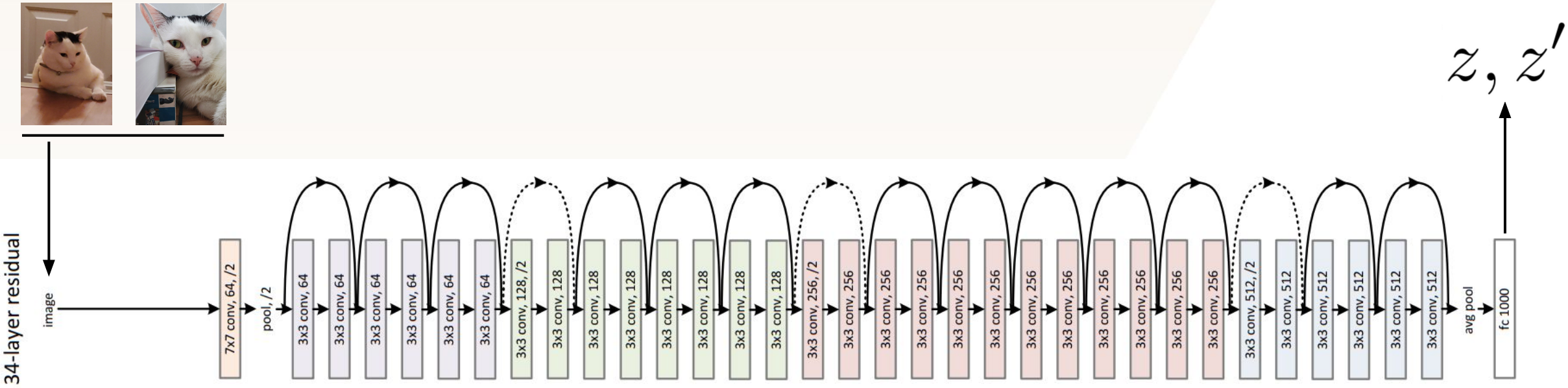


?

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

Mila

How to select positive and negative examples?

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

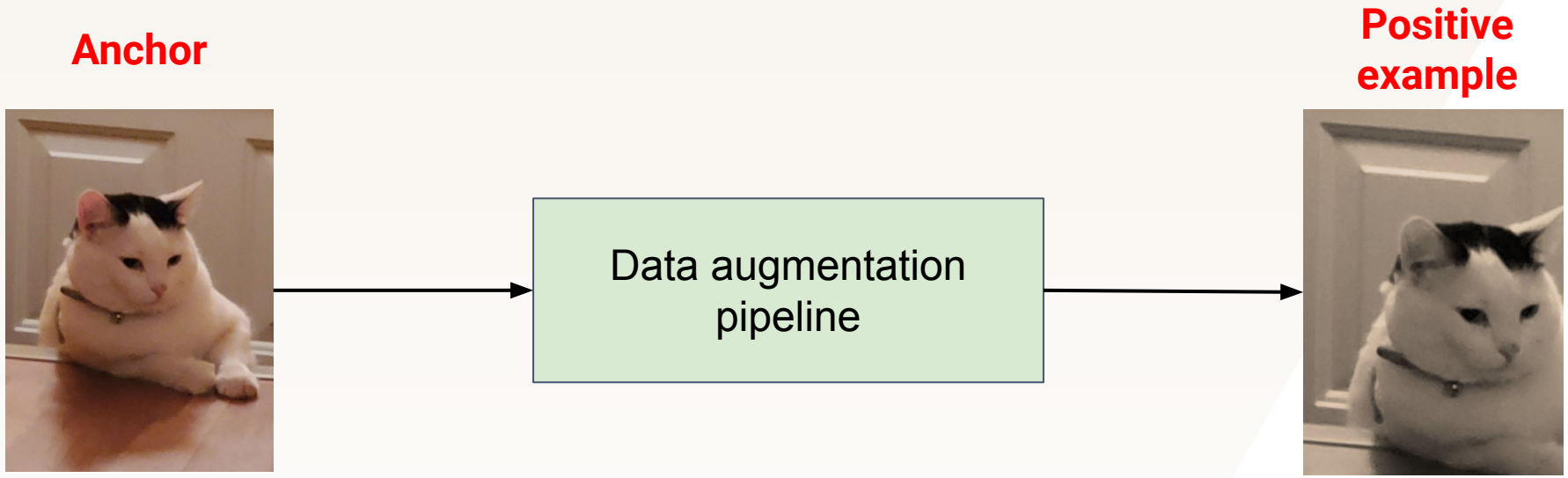
Anchor

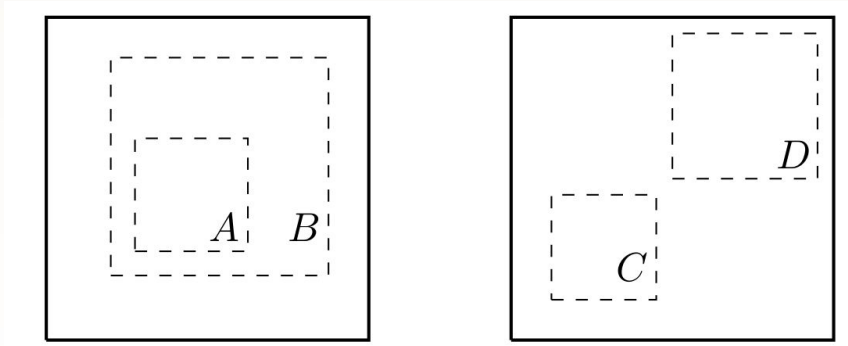
Positive example

Negative examples

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

Data augmentation





(a) Global and local views. (b) Adjacent views.

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

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

Why using a temperature hyper-parameter?

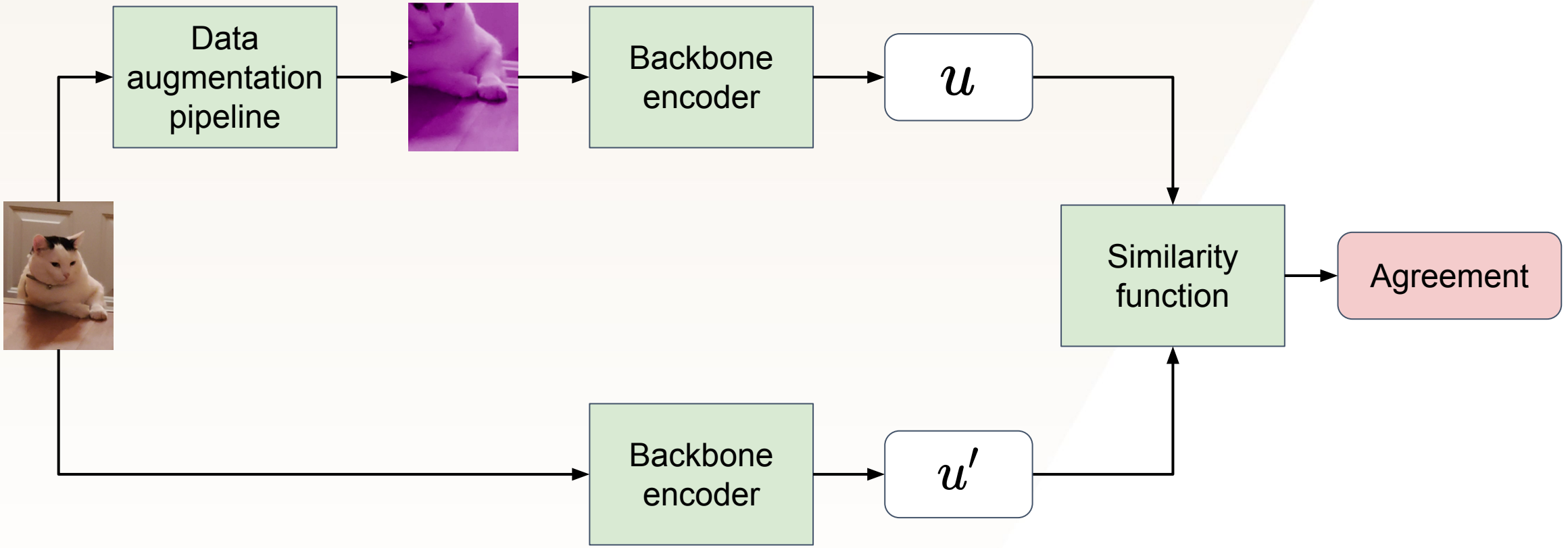
$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

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)

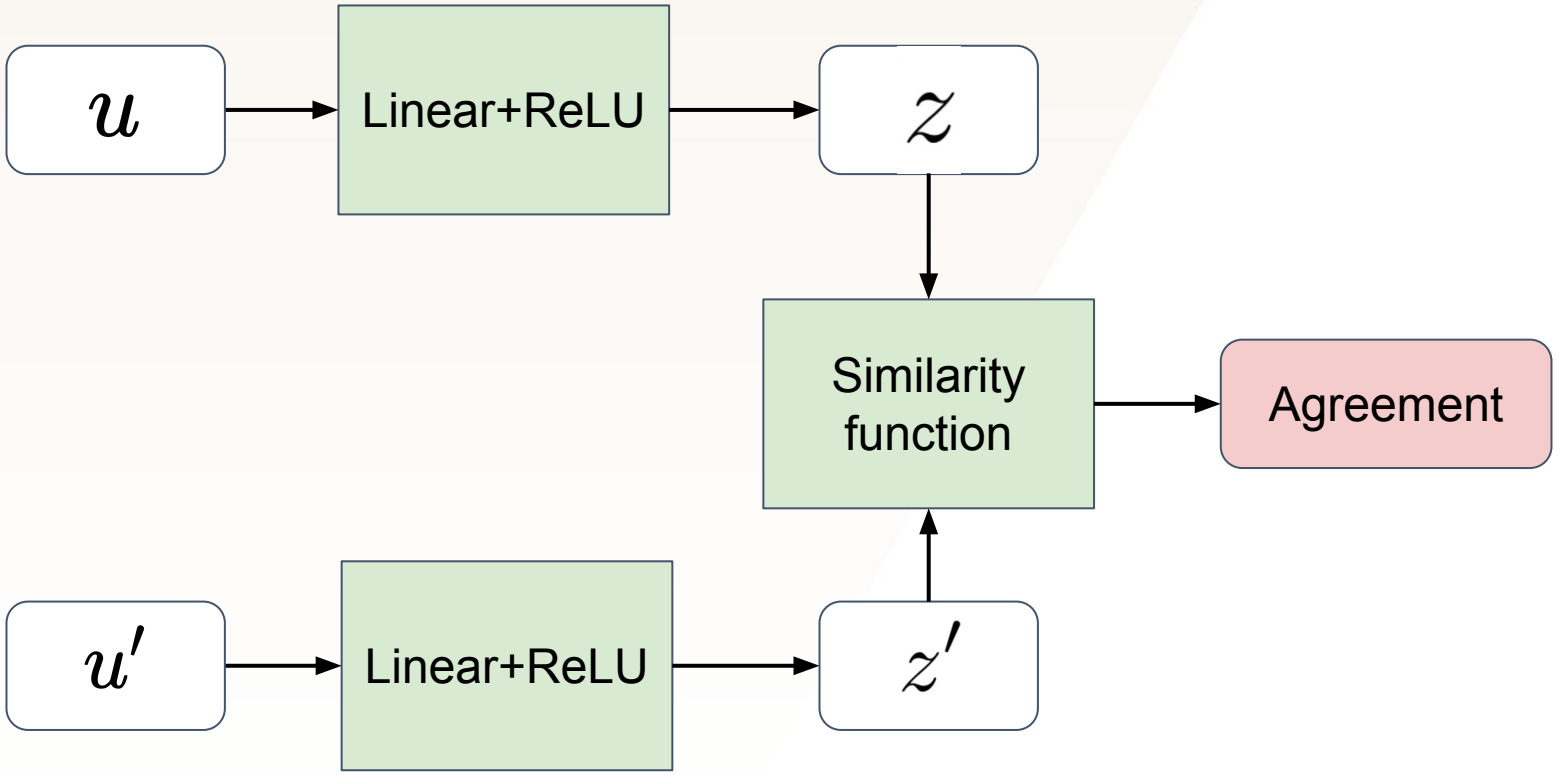
Putting everything together



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

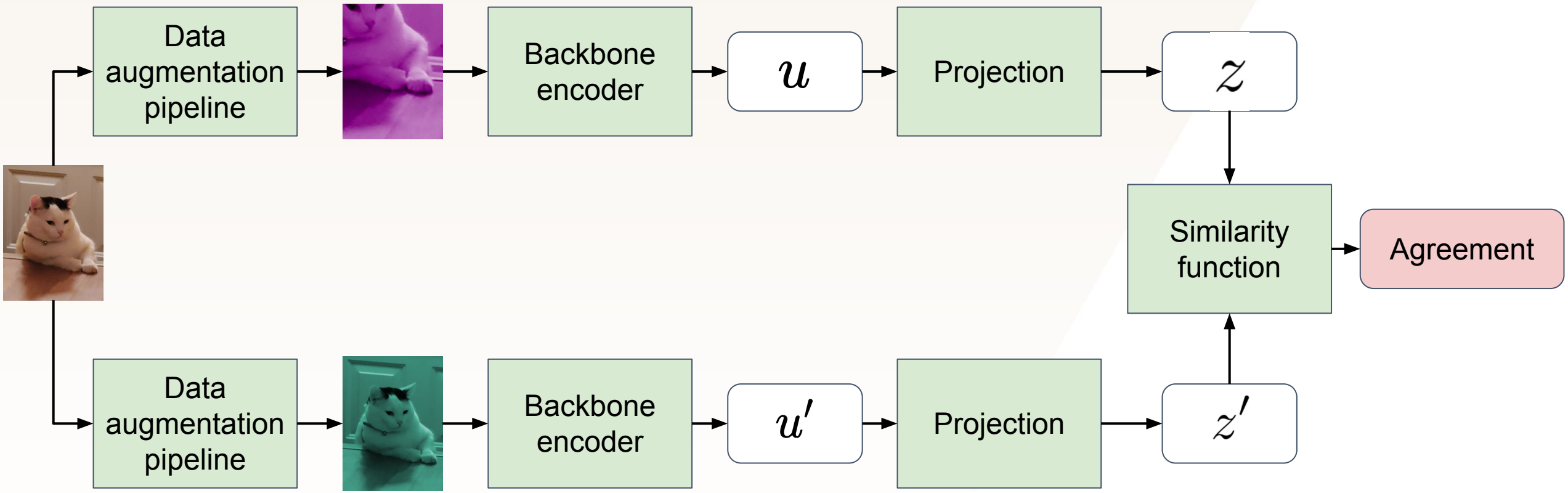
Projection trick

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

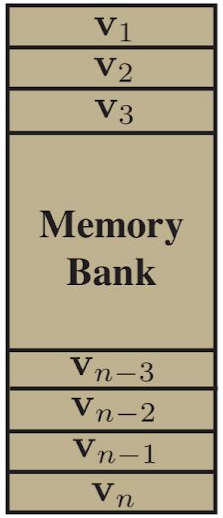
SimCLR



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?

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

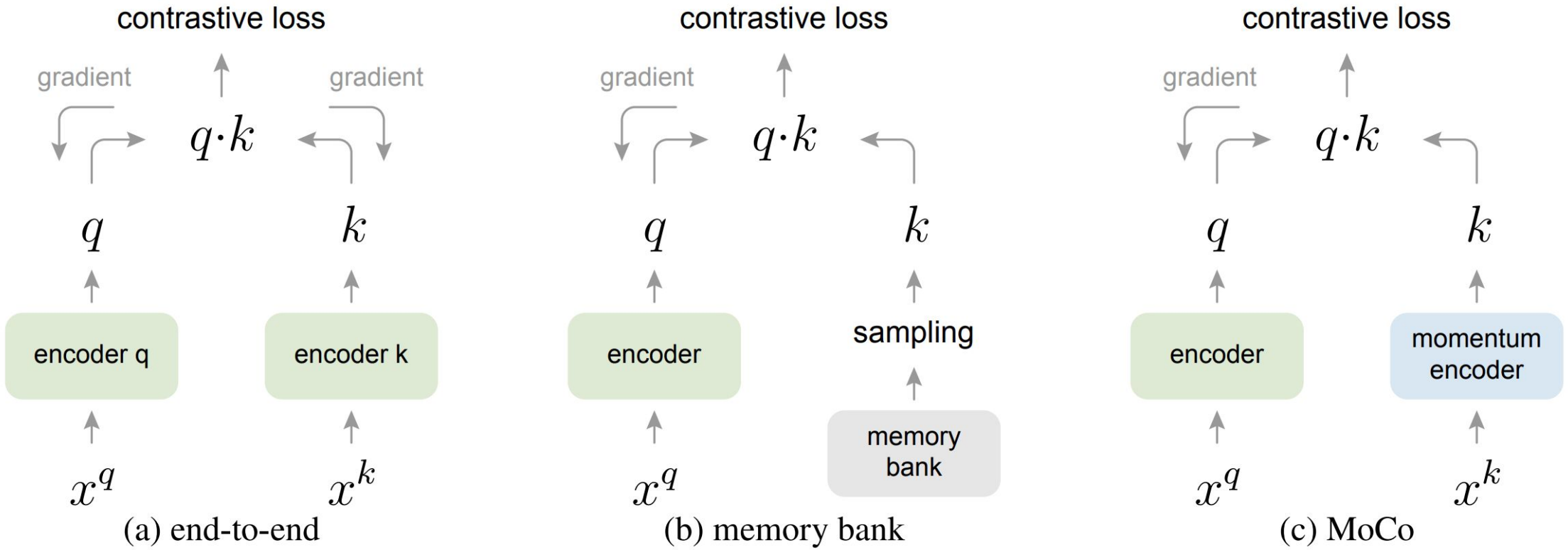


$$\mathcal{L}(z^a, z^+, Z^-) = -\log \frac{\exp(\text{sim}(z^a, z^+)/\tau)}{\sum_{z^- \in Z^-} \exp(\text{sim}(z^a, z^-)/\tau)}$$

Negative examples

[1] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, Dahua Lin: Unsupervised Feature Learning via Non-Parametric Instance Discrimination. CVPR 2018: 3733-3742
[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey E. Hinton: A Simple Framework for Contrastive Learning of Visual Representations. ICML 2020
[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

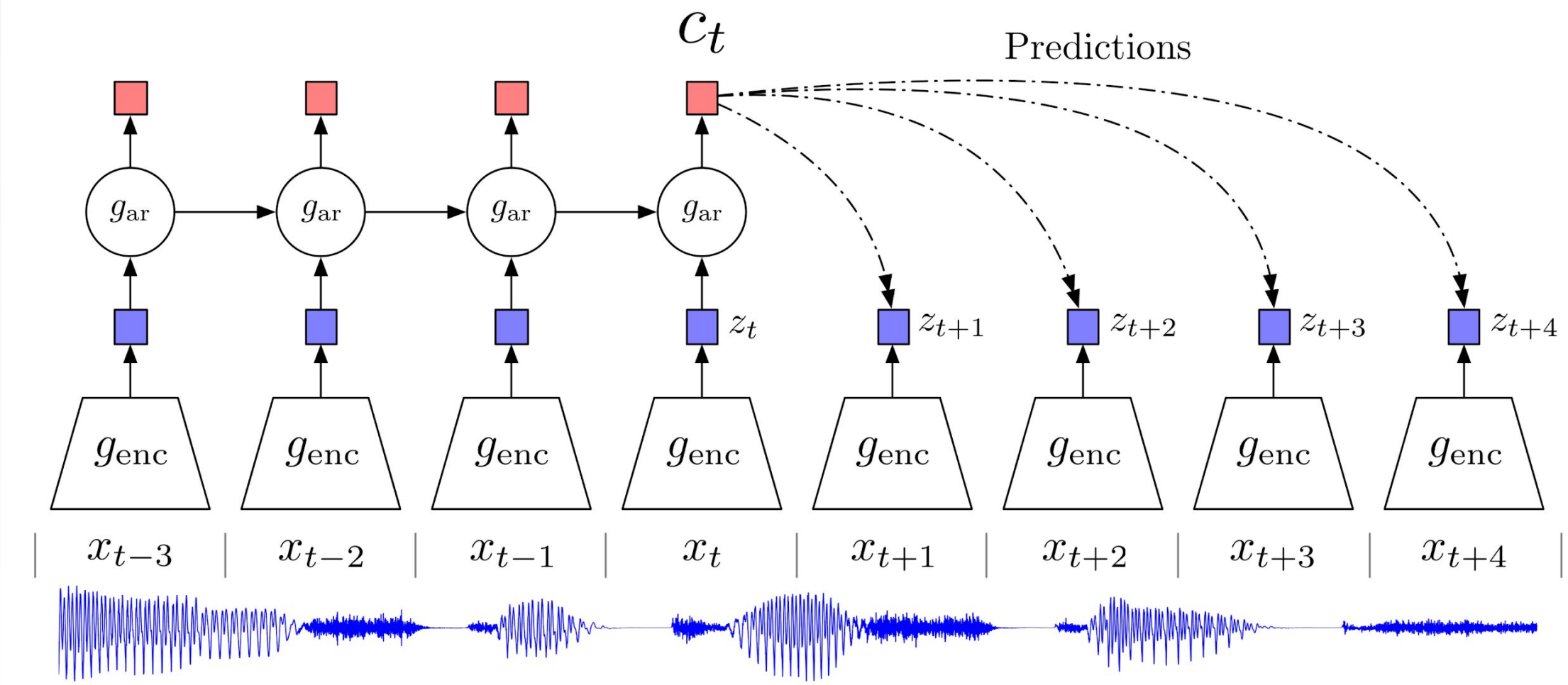
Discussion

- Handling negative examples is difficult.
- Require large batch training.
- They are computationally intensive and work best on huge unlabeled datasets.

05

Autoregressive contrastive learning

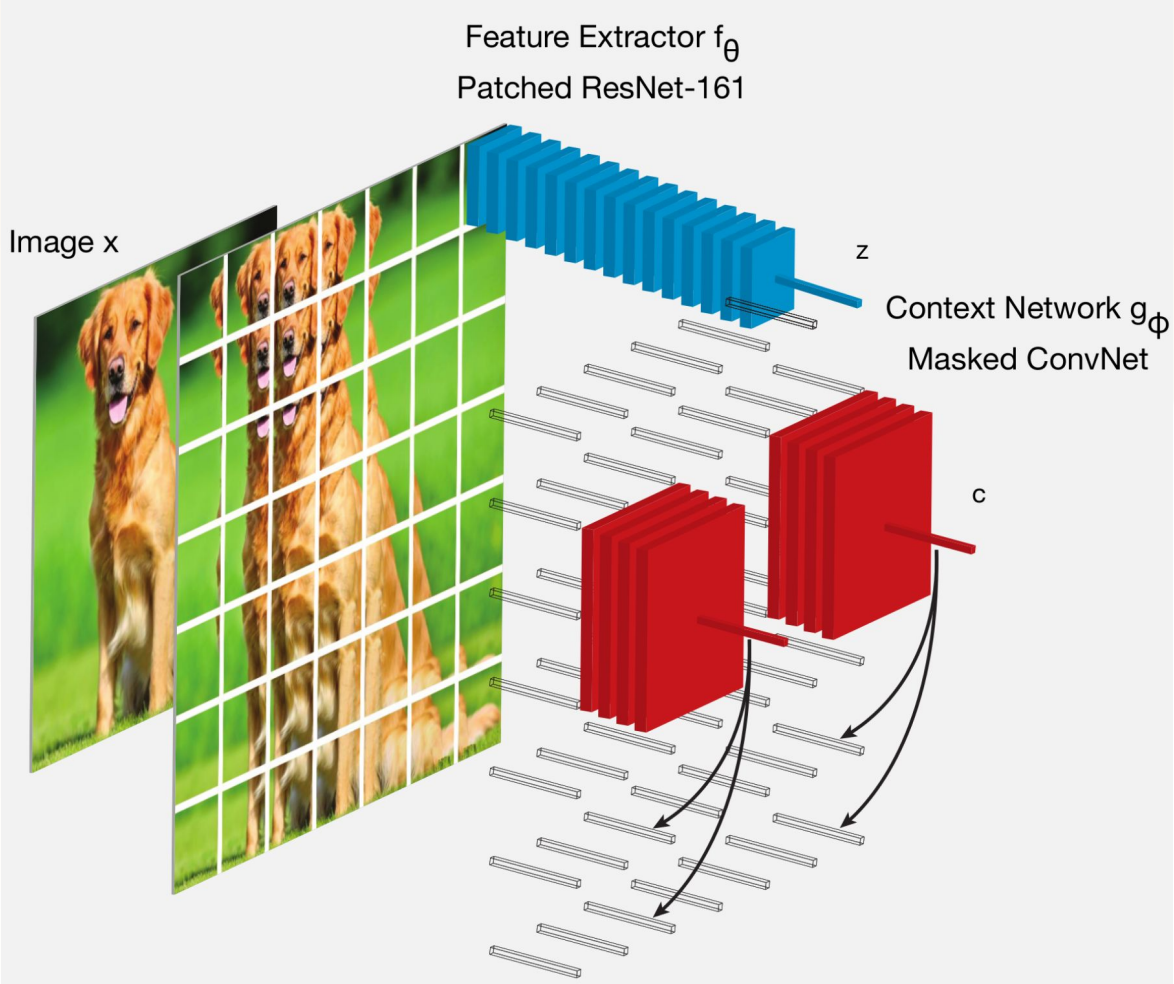
Contrastive Predictive Coding (CPC)



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

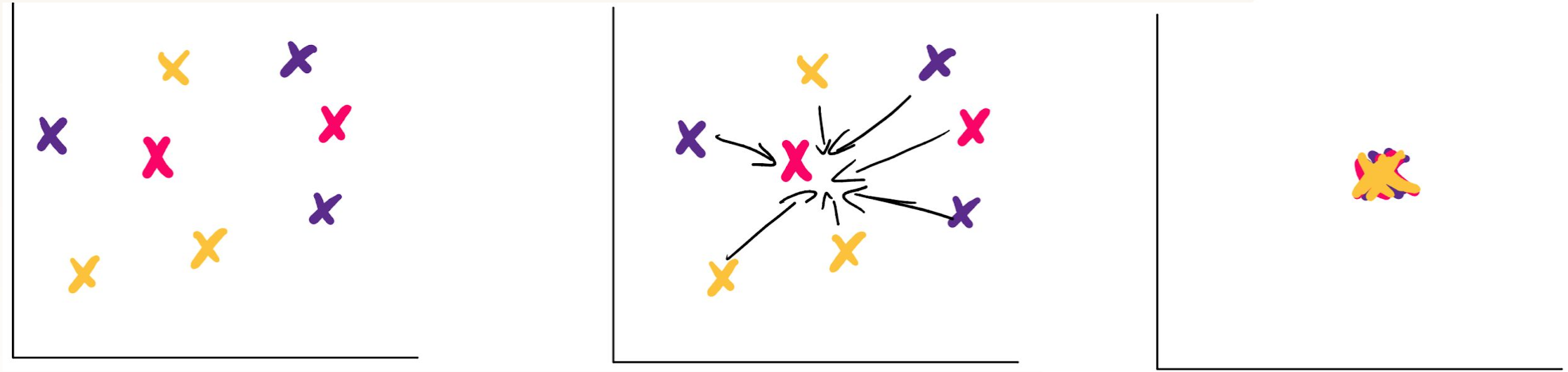
Discussion

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

06

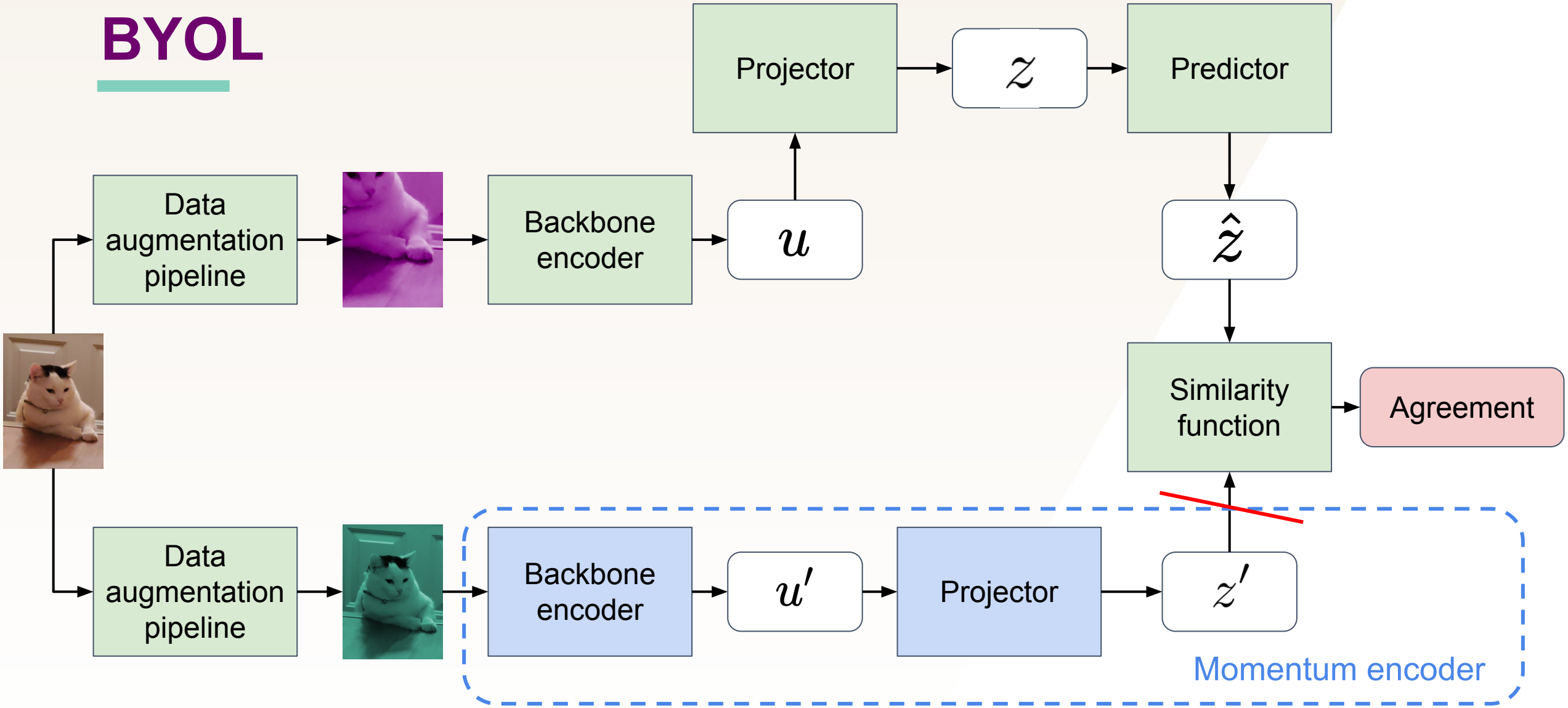
“Contrastive learning” without negative examples

The collapse problem



Mila

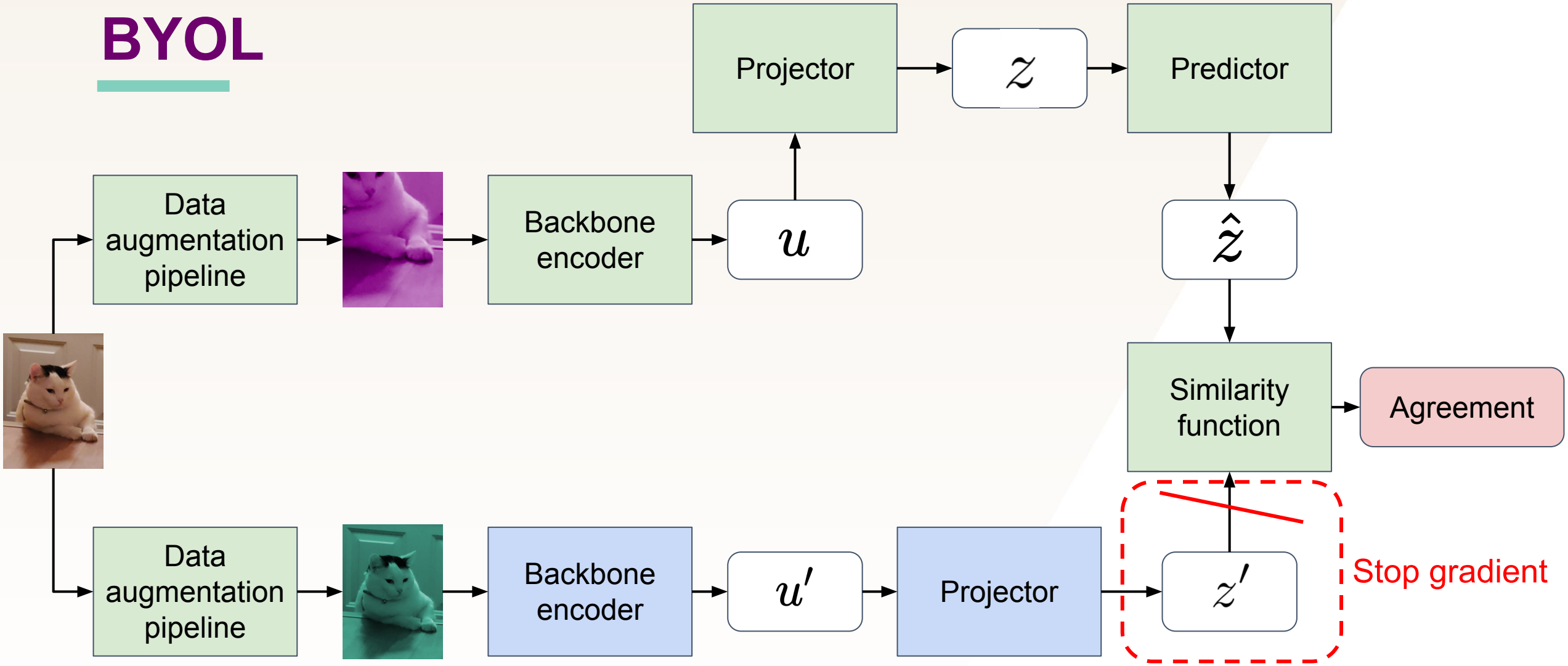
BYOL



Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." NeurIPS (2020).

Mila

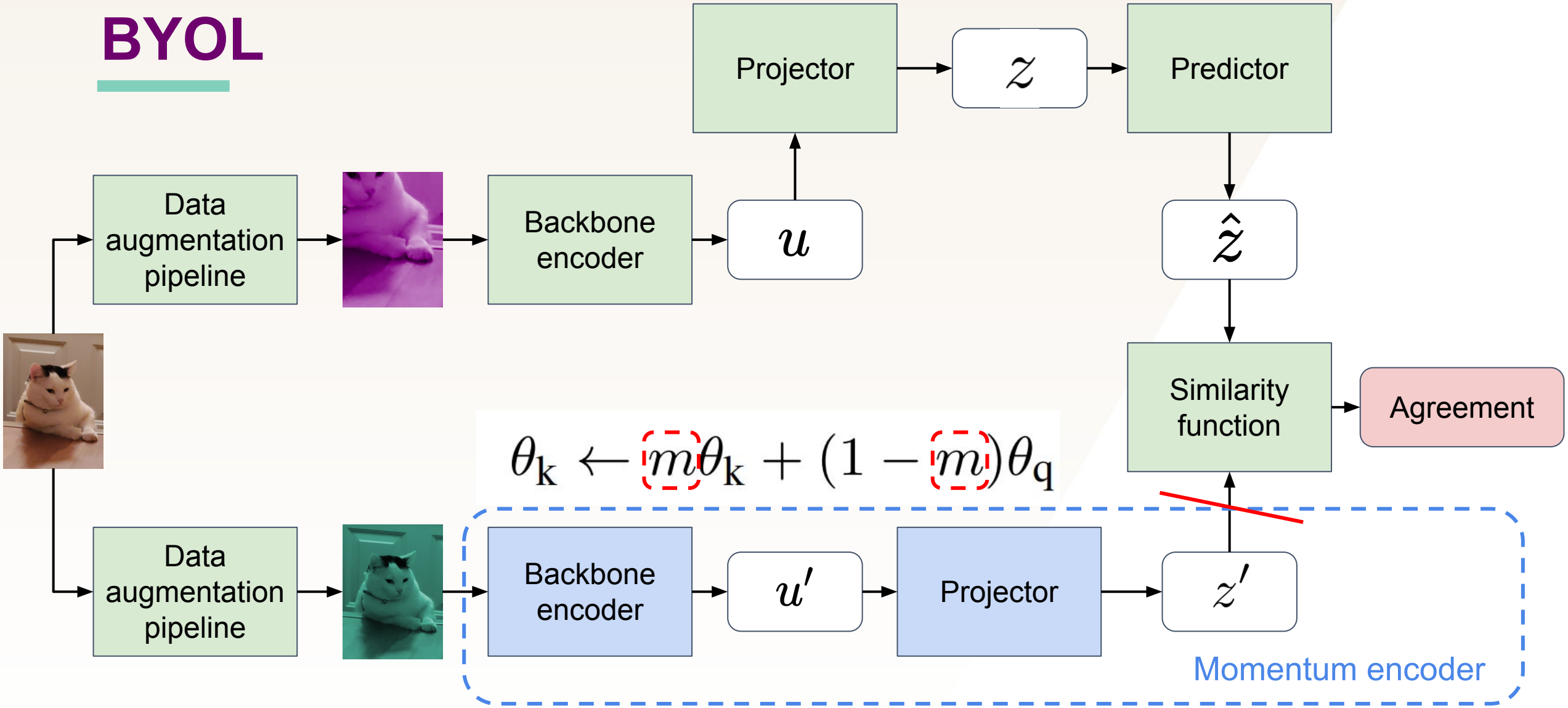
BYOL



Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." NeurIPS (2020).

Mila

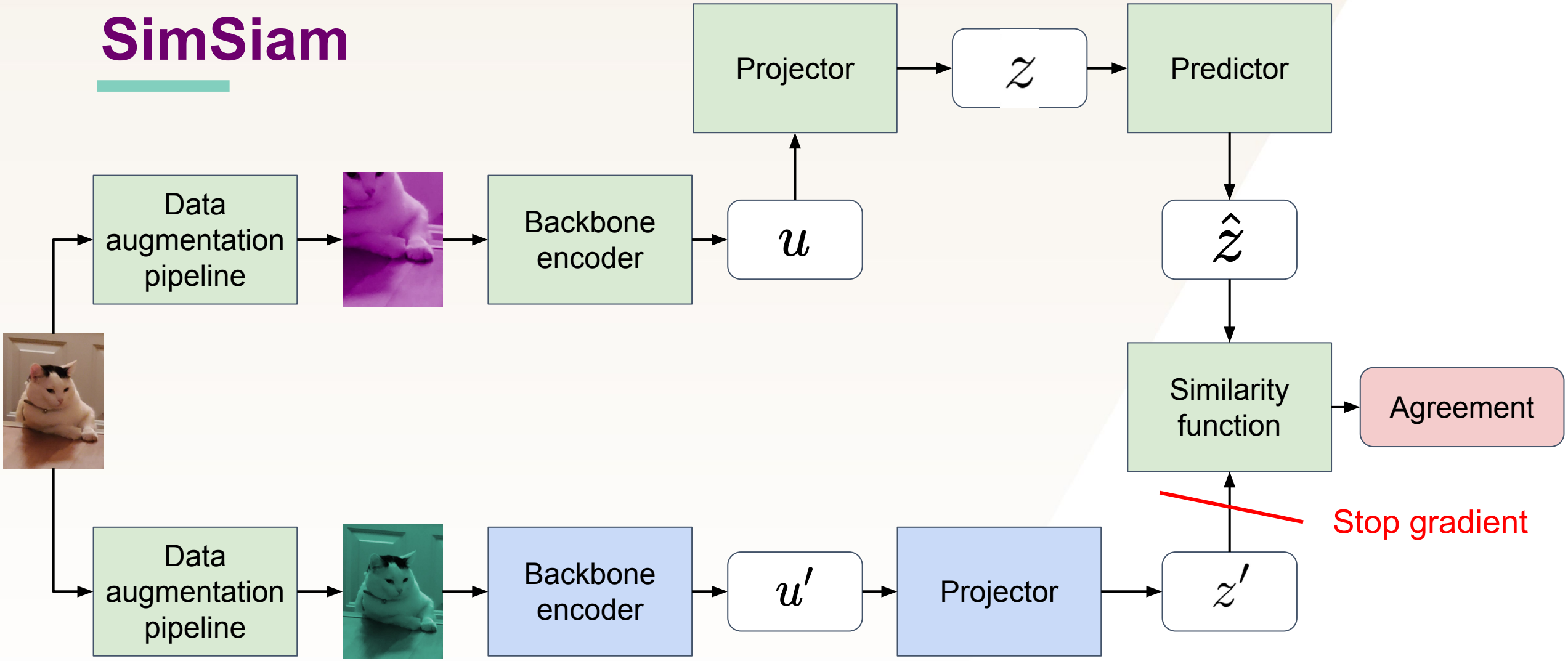
BYOL



Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." NeurIPS (2020).

Mila

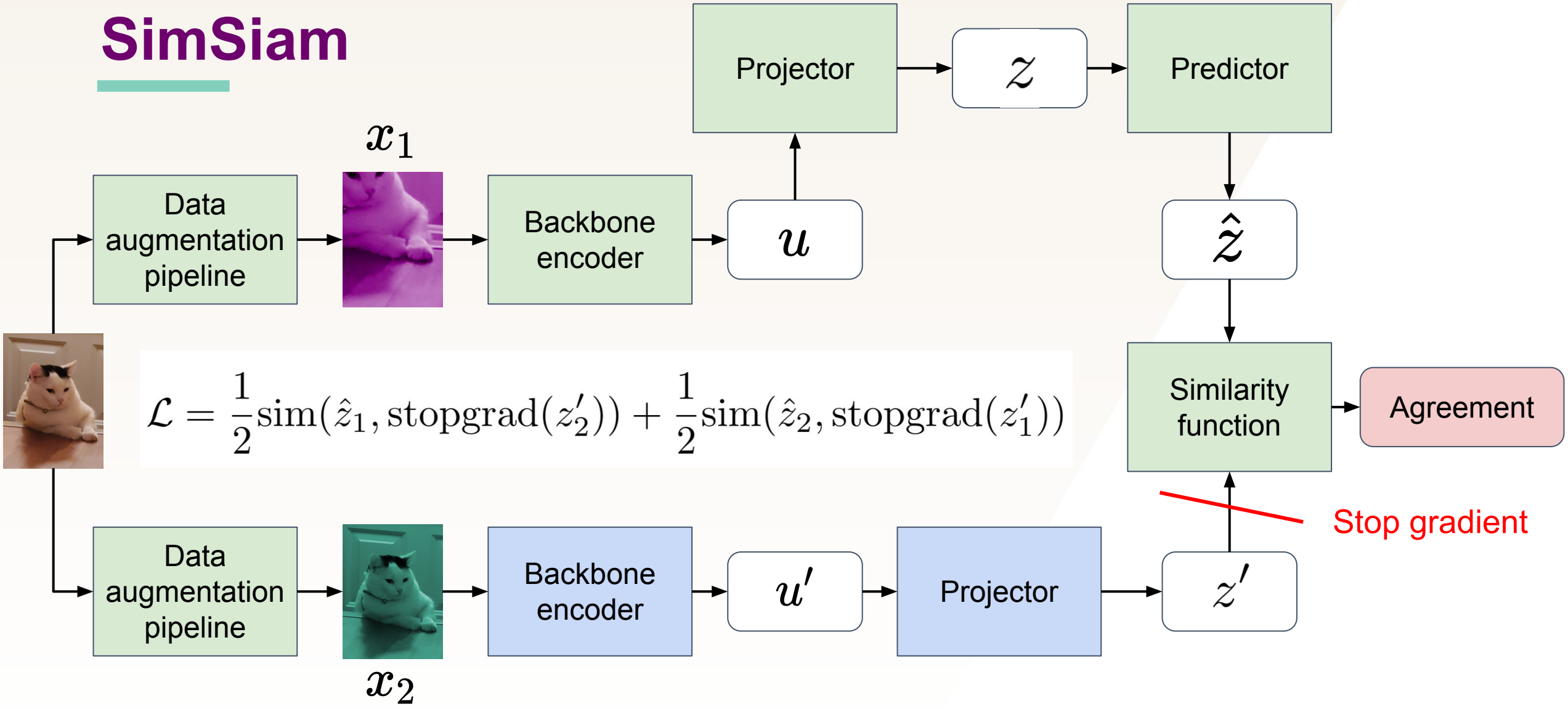
SimSiam



Xinlei Chen, Kaiming He: Exploring Simple Siamese Representation Learning. CoRR abs/2011.10566 (2020)

Mila

SimSiam



Xinlei Chen, Kaiming He: Exploring Simple Siamese Representation Learning. CoRR abs/2011.10566 (2020)

Mila

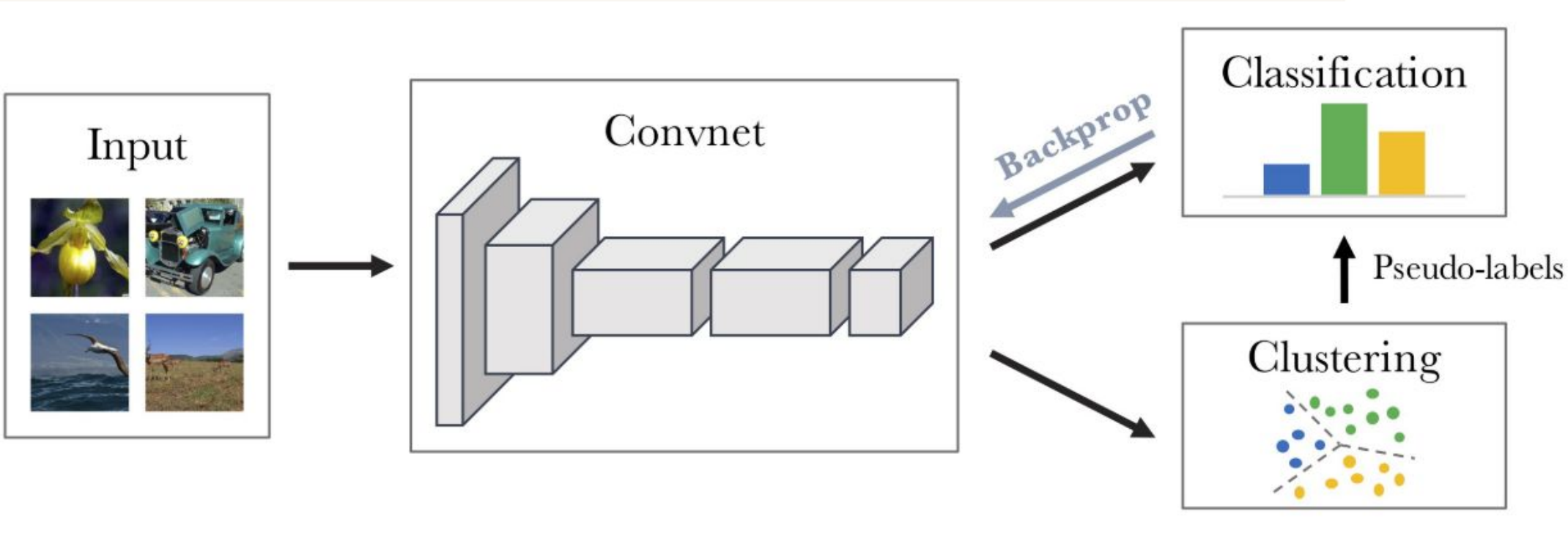
Discussion

- Different solutions to avoid the collapse problem.
- Competitive results compared with purely contrastive methods.

07

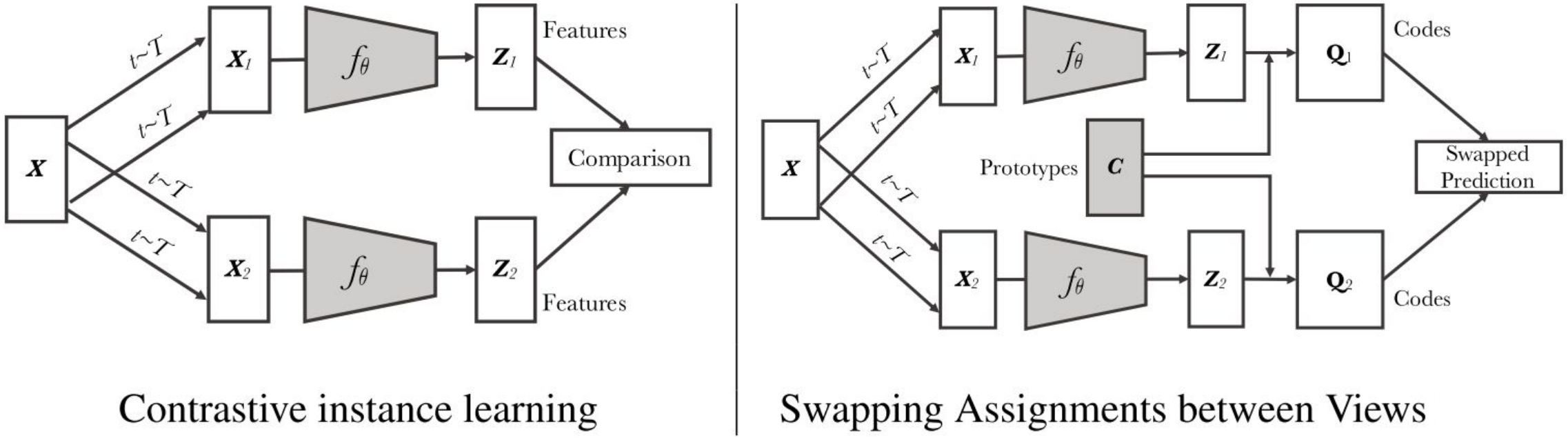
Contrastive learning with clustering

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)

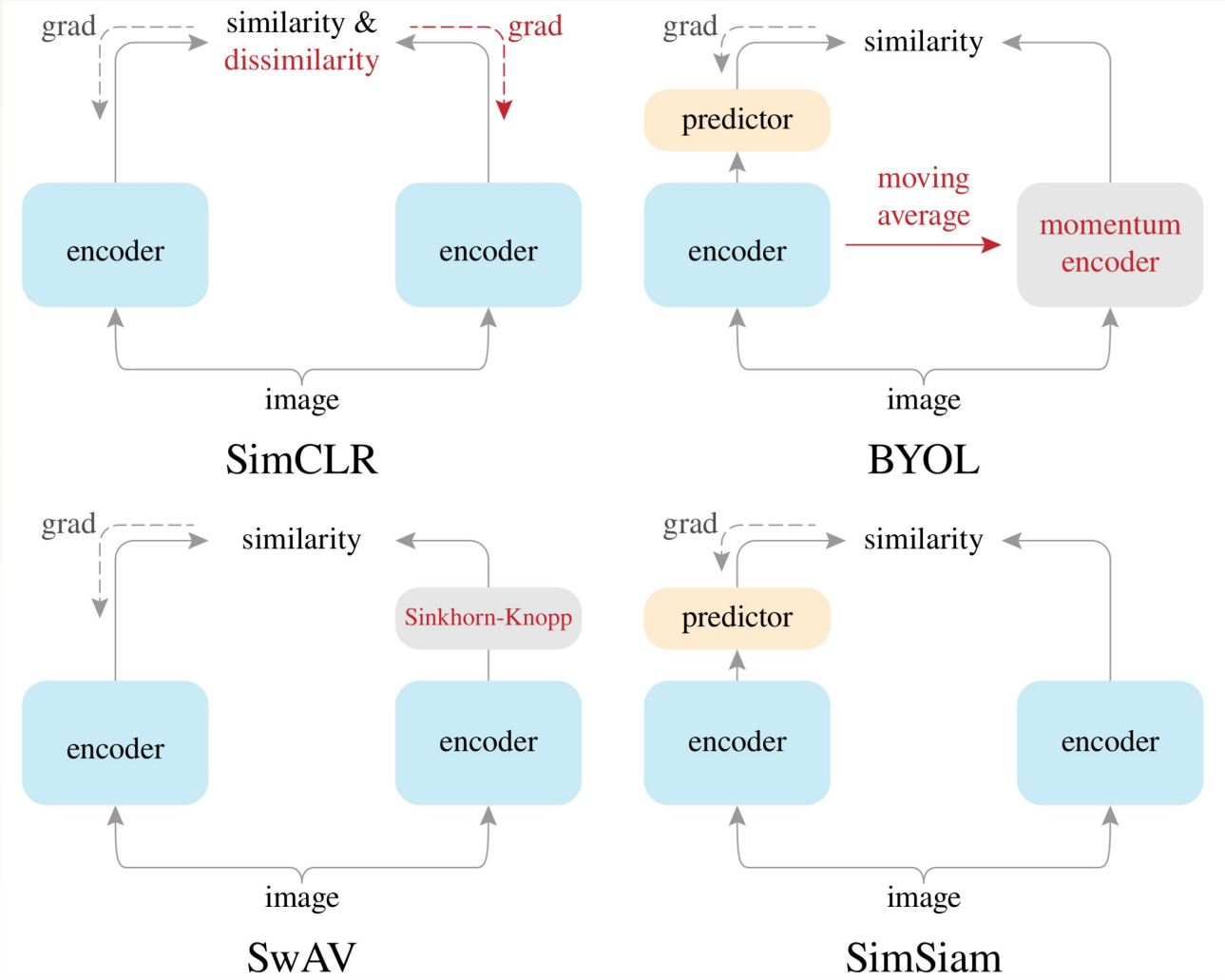
Contrastive and clustering together



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

Mila

Siamese representation learning



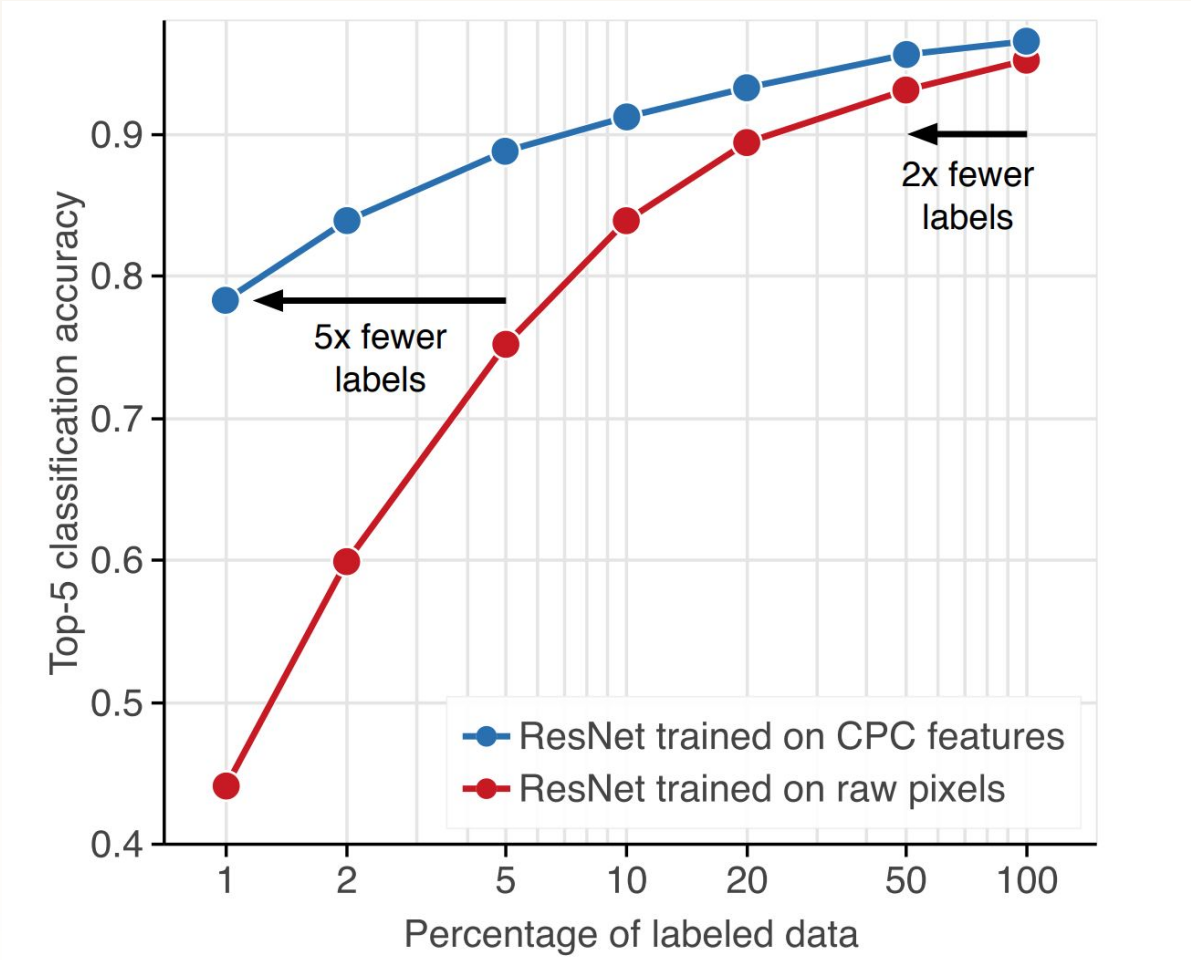
Xinlei Chen, Kaiming He: Exploring Simple Siamese Representation Learning. CoRR abs/2011.10566 (2020)

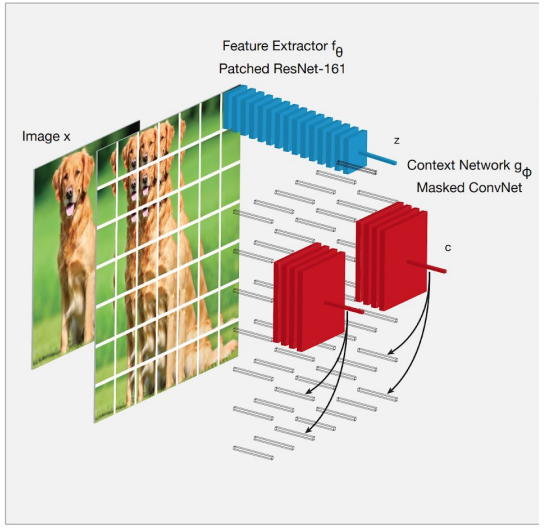
Mila

08

Empirical results

Empirical results: CPC v2





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

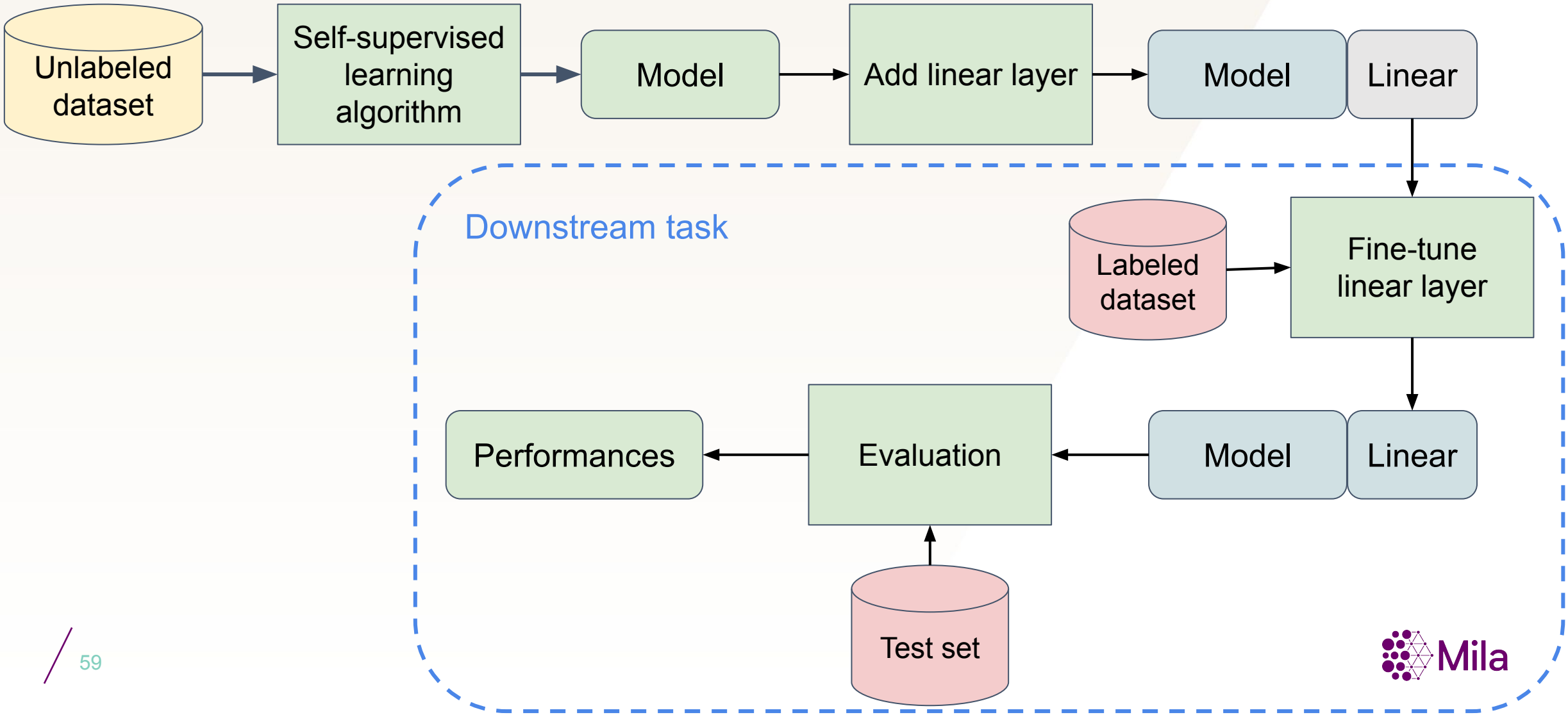
Mila

Empirical results: BYOL

Method	1% labels		10% labels	
	Top-1	Top-5	Top-1	Top-5
Supervised	25.4	48.4	56.4	80.4
SimCLR [10]	48.3	75.5	65.6	87.8
SwAV	**53.9**	**78.5**	70.2	**89.9**

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

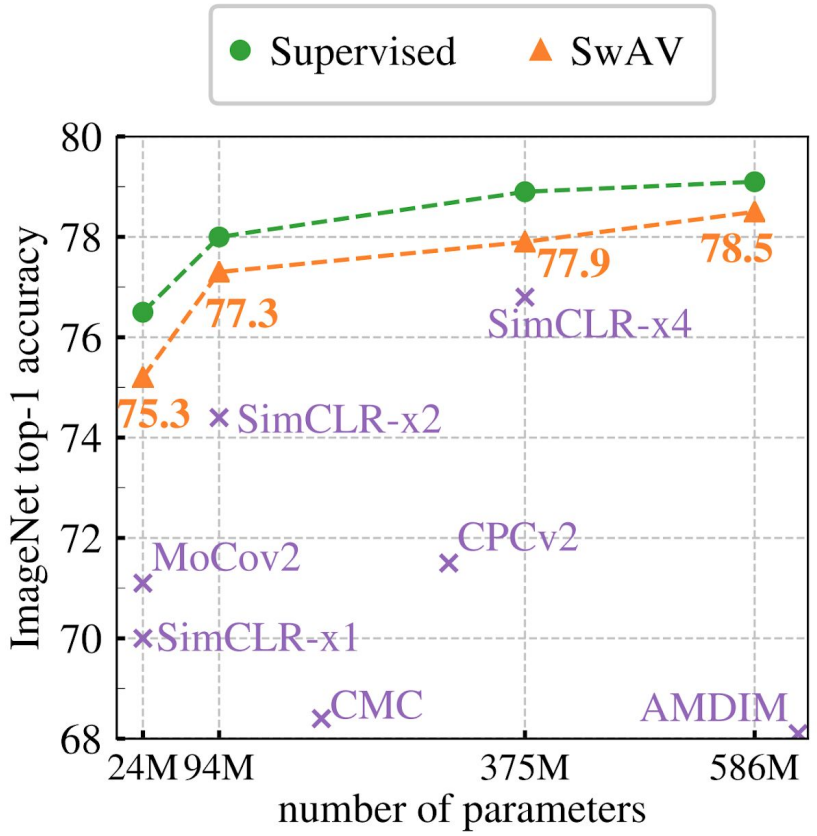
Linear evaluation protocol



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
Local Agg.	60.2
PIRL [35]	63.6
CPC v2 [32]	63.8
CMC [11]	66.2
SimCLR [8]	69.3
MoCo v2 [37]	71.1
InfoMin Aug. [12]	73.0
BYOL (ours)	**74.3**

(a) ResNet-50 encoder.

Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, Armand Joulin: Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. NeurIPS (2020)
Grill, Jean-Bastien, et al. "Bootstrap Your Own Latent-A New Approach to Self-Supervised Learning." NeurIPS (2020).

Mila

SSL on uncurated data

Pre-training on an uncurated dataset of 1 billion random public images from Instagram.

Method	Frozen	Finetuned
Random	15.0	76.5
MoCo	-	77.3*
SimCLR	60.4	77.2
SwAV	**66.5**	**77.8**

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

Discussion

- Siamese network design pattern is now common in SSL.
- Computation budget and datasets can be huge.
- SSL methods in vision depend heavily on data augmentations.
  - How to automate data augmentation discoveries?
- Pytorch lightning implements a SSL framework for fast prototyping.
- SSL has a direct impact on democratizing DL.

Discussion

- We covered a tiny subset of the literature.
- The field is moving very fast.
- Stay up to date!

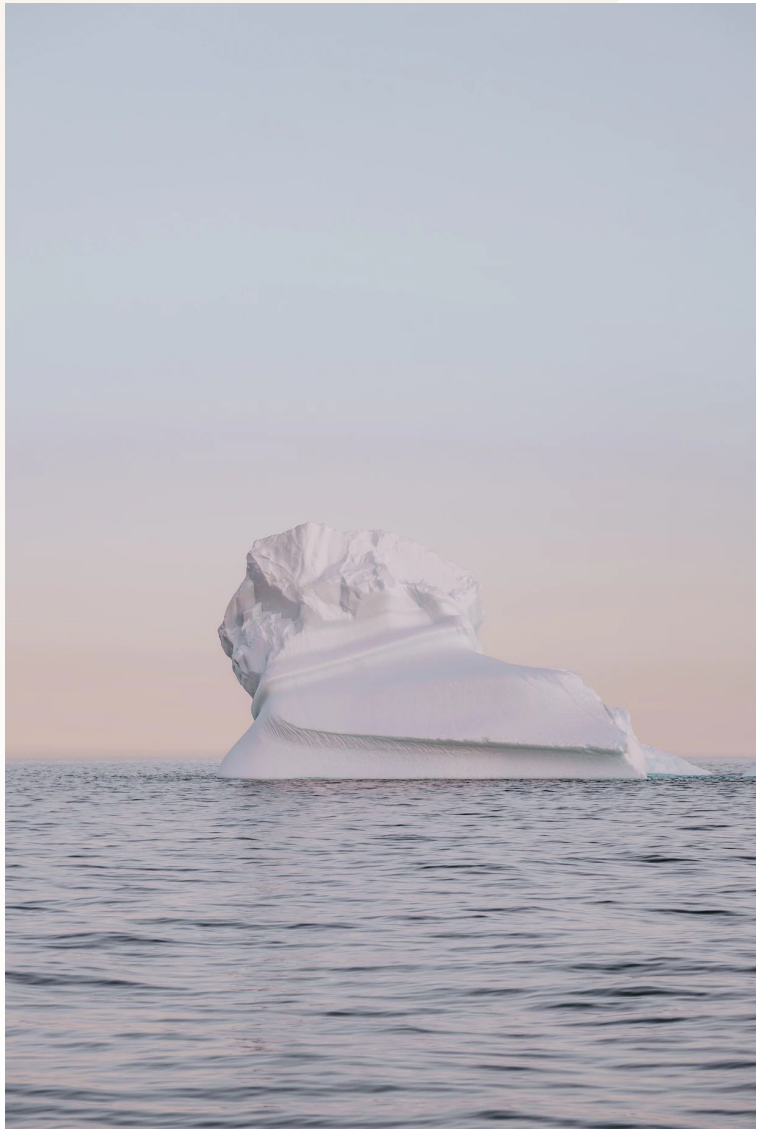


photo credit: Annie Spratt (Unsplash)

Mila

09

Questions

Thank you!