

Self-supervised learning in computer vision and medical imaging

Pietro Gori

Associate Professor
Télécom Paris (IPParis)
Paris, France



1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

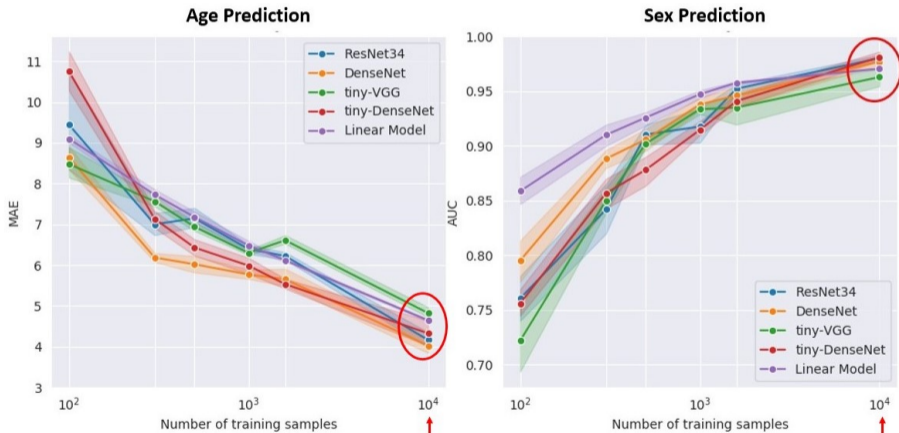
4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- Deep learning (e.g., CNN or ViT) is a lazy and inefficient statistical method that needs millions if not billions of examples to learn a precise task → **data hungry**



- Many specific tasks in Computer Vision, such as object detection¹ (e.g., YOLO), image classification² (e.g., ResNet-50), or semantic segmentation (e.g., U-Net), have reached astonishing results in the last years.
- This has been possible mainly because **large** ($N > 10^6$), **labeled** data-sets were *easily* accessible and *freely* available



¹T.-Y. Lin et al. "Microsoft COCO: Common Objects in Context". In: *ECCV*. 2014.

²J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*. 2009.

- In medical imaging, current **research** datasets are:
 - ▶ **small**: $N < 2k$ for common pathology and $N < 200$ for rare pathology
 - ▶ **biased**: images are acquired in a precise hospital, following a specific protocol with a particular machine (nuisance site effect)
 - ▶ **multi-modal**: many imaging modalities can be available as well as text, clinical, biological, genetic data.
 - ▶ **anonymized, quality checked, accessible, quite homogeneous**
- **Clinical** datasets are harder to analyze since they are usually **not anonymized, not quality checked, not freely accessible, highly heterogeneous**.
- In this talk, we will focus on **research** medical imaging datasets

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

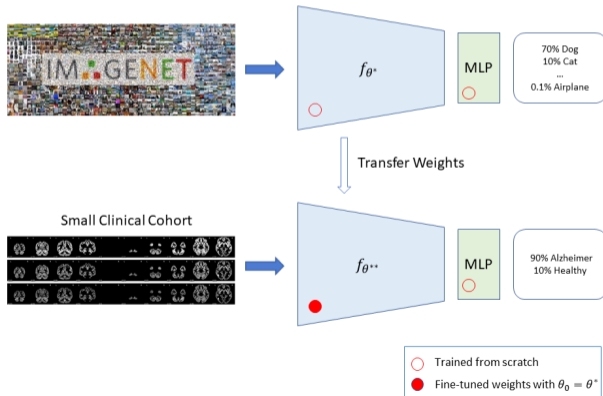
4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- When dealing with small labelled datasets, a common strategy is **Transfer Learning**:
 1. pre-training a model on a *large* dataset and then
 2. fine-tuning it on the *small* target and labelled dataset

- When dealing with small labelled datasets, a common strategy is **Transfer Learning**:
 1. pre-training a model on a *large* dataset and then
 2. fine-tuning it on the *small* target and labelled dataset
- **Supervised pre-training** from ImageNet is common.



- When dealing with small labelled datasets, a common strategy is **Transfer Learning**:
 1. pre-training a model on a *large* dataset and then
 2. fine-tuning it on the *small* target and labelled dataset
- **Supervised pre-training** from ImageNet is common. Its usefulness (that is, feature reuse) increases with³⁴⁵⁶:
 - ▶ reduced target data size (small N_{target})
 - ▶ visual similarity between pre-train and target domains (small FID)
 - ▶ models with fewer inductive biases (TL works better for ViTs than CNN)
 - ▶ larger architectures (more parameters)

³B. Mustafa et al. *Supervised Transfer Learning at Scale for Medical Imaging*. 2021.

⁴C. Matsoukas et al. "What Makes Transfer Learning Work for Medical Images". In: *CVPR*. 2022.

⁵B. Neyshabur et al. "What is being transferred in transfer learning?" In: *NeurIPS*. 2020.

⁶M. Raghu et al. "Transfusion: Understanding Transfer Learning for Medical Imaging". In: *NeurIPS*. 2019.

- When dealing with small labelled datasets, a common strategy is **Transfer Learning**:
 1. pre-training a model on a *large* dataset and then
 2. fine-tuning it on the *small* target and labelled dataset
- **Supervised pre-training** from ImageNet is common. Its usefulness (that is, feature reuse) increases with⁷⁸⁹¹⁰:
 - ▶ reduced target data size (small N_{target})
 - ▶ **visual similarity between pre-train and target domains (small FID)**
 - ▶ models with fewer inductive biases (TL works better for ViTs than CNN)
 - ▶ larger architectures (more parameters)

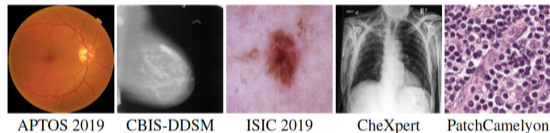
⁷B. Mustafa et al. *Supervised Transfer Learning at Scale for Medical Imaging*. 2021.

⁸C. Matsoukas et al. "What Makes Transfer Learning Work for Medical Images". In: *CVPR*. 2022.

⁹B. Neyshabur et al. "What is being transferred in transfer learning?" In: *NeurIPS*. 2020.

¹⁰M. Raghu et al. "Transfusion: Understanding Transfer Learning for Medical Imaging". In: *NeurIPS*. 2019.

- Natural¹¹ and Medical¹² images can be visually very different ! → **Domain gap**

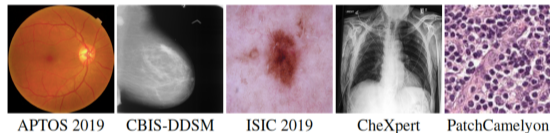


- Furthermore, Medical images can be 3D. ImageNet is 2D.
- Need for **3D, annotated, large** medical dataset

¹¹J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*. 2009.

¹²C. Matsoukas et al. "What Makes Transfer Learning Work for Medical Images". In: *CVPR*. 2022.

- Natural¹¹ and Medical¹² images can be visually very different ! → **Domain gap**



- Furthermore, Medical images can be 3D. ImageNet is 2D.
- Need for **3D, annotated, large** medical dataset → **PROBLEM !**

¹¹J. Deng et al. "ImageNet: A Large-Scale Hierarchical Image Database". In: *CVPR*. 2009.

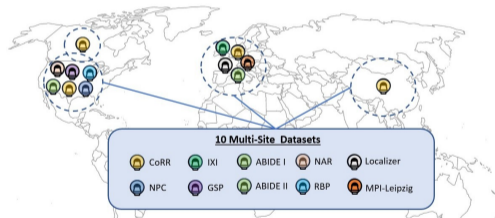
¹²C. Matsoukas et al. "What Makes Transfer Learning Work for Medical Images". In: *CVPR*. 2022.

- Supervised pre-training is a not a valid option in medical imaging. Need for another kind of pre-training.

¹³T. J. Littlejohns et al. "The UK Biobank imaging enhancement of 100,000 participants:" in: *Nature Communications* (2020).

¹⁴B. Dufumier et al. "OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing". In: *NeuroImage* (2022).

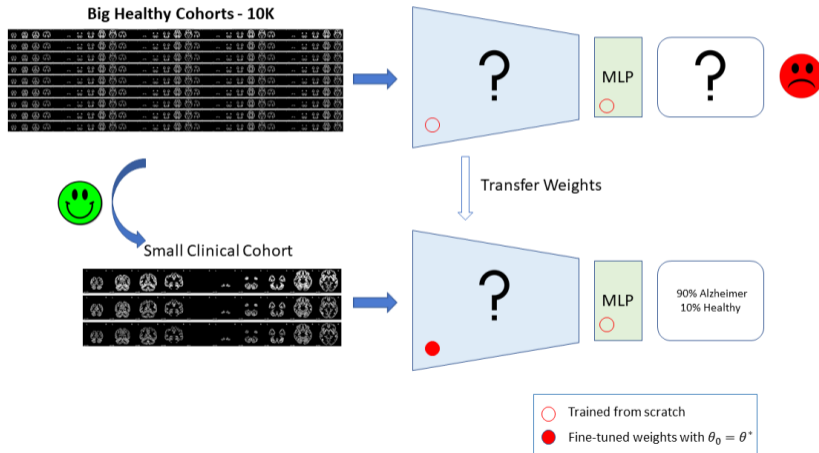
- Supervised pre-training is not a valid option in medical imaging. Need for another kind of pre-training.
- Recently **big, multi-sites** international **healthy data-sets** have emerged, such as UK Biobank¹³ ($N > 100k$) and OpenBHB¹⁴ ($N > 10k$)



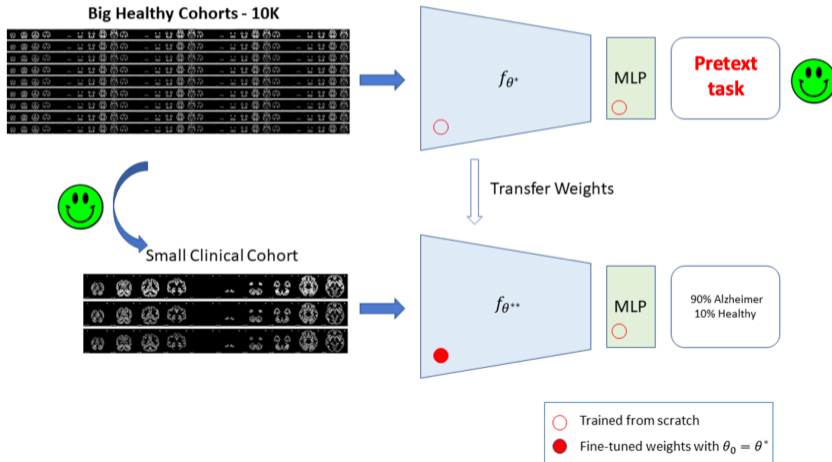
¹³T. J. Littlejohns et al. "The UK Biobank imaging enhancement of 100,000 participants:" in: *Nature Communications* (2020).

¹⁴B. Dufumier et al. "OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing". In: *NeuroImage* (2022).

- How can we employ an **healthy** (thus **unlabeled**) data-set for pre-training ?



- How can we employ an **healthy** (thus **unlabeled**) data-set for pre-training ? → **Self-Supervised pre-training !**



- **Self-supervised pre-training:** leverage an annotation-free *pretext task* to provide a surrogate supervision signal for feature learning.

¹⁵ [C. Doersch et al.](#) “Unsupervised Visual Representation Learning by Context Prediction”. In: *ICCV*. 2015.

¹⁶ [K. He et al.](#) “Masked Autoencoders Are Scalable Vision Learners”. In: *CVPR*. 2022.

¹⁷ [J. Donahue et al.](#) “Large Scale Adversarial Representation Learning”. In: *NeurIPS*. 2019.

¹⁸ [T. Chen et al.](#) “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.

¹⁹ [J.-B. Grill et al.](#) “Bootstrap your own latent: A new approach to self-supervised Learning”. In: *NeurIPS*. 2020.

²⁰ [A. Bardes et al.](#) “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR*. 2022.

- **Self-supervised pre-training:** leverage an annotation-free *pretext task* to provide a surrogate supervision signal for feature learning.
- Pretext task should only use the visual information and context of the images

¹⁵ [C. Doersch et al.](#) “Unsupervised Visual Representation Learning by Context Prediction”. In: *ICCV*. 2015.

¹⁶ [K. He et al.](#) “Masked Autoencoders Are Scalable Vision Learners”. In: *CVPR*. 2022.

¹⁷ [J. Donahue et al.](#) “Large Scale Adversarial Representation Learning”. In: *NeurIPS*. 2019.

¹⁸ [T. Chen et al.](#) “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.

¹⁹ [J.-B. Grill et al.](#) “Bootstrap your own latent: A new approach to self-supervised Learning”. In: *NeurIPS*. 2020.

²⁰ [A. Bardes et al.](#) “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR*. 2022.

- **Self-supervised pre-training:** leverage an annotation-free *pretext task* to provide a surrogate supervision signal for feature learning.
- Pretext task should only use the visual information and context of the images
- Examples of pretext tasks:
 - ▶ Context prediction¹⁵
 - ▶ Generative models^{16,17}
 - ▶ Instance discrimination¹⁸
 - ▶ Teacher/Student¹⁹
 - ▶ Information Maximization²⁰

¹⁵C. Doersch et al. “Unsupervised Visual Representation Learning by Context Prediction”. In: *ICCV*. 2015.

¹⁶K. He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *CVPR*. 2022.

¹⁷J. Donahue et al. “Large Scale Adversarial Representation Learning”. In: *NeurIPS*. 2019.

¹⁸T. Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.

¹⁹J.-B. Grill et al. “Bootstrap your own latent: A new approach to self-supervised Learning”. In: *NeurIPS*. 2020.

²⁰A. Bardes et al. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR*. 2022.

- Pre-text tasks should produce image representations that are:
 1. **Transferable**: we can easily reuse/fine-tune them in different downstream tasks (e.g., segmentation, object detection, classification, etc.)
 2. **Generalizable**: they should not be specific to a single task but work well in several different downstream tasks
 3. **High-level**: representations should characterize the high-level semantics/structure and not low-level features (color, texture, etc.)
 4. **Invariant**: image representations should be invariant to geometric or appearance transformations that do not modify the information content of the image (i.e., irrelevant for downstream task)
 5. **Semantically coherent**: semantically similar images should be close in the representation space

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

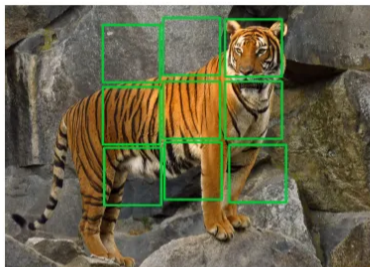
- Given an image, we can divide it into patches and predict their relative position



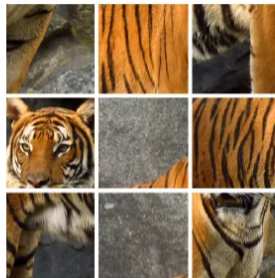
Given a central patch (blue), predict the relative position of the other patch among the 8 possible connected neighbor positions.²¹

- Or we could use all neighbor patches...

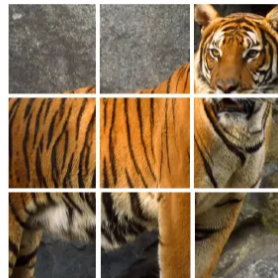
²¹C. Doersch et al. "Unsupervised Visual Representation Learning by Context Prediction". In: *ICCV*. 2015.



(a)



(b)



(c)

First randomly shuffle all 9 patches and then learn to solve a Jigsaw puzzle. Considering all patches together remove ambiguities since placement is mutually exclusive. See central patch and top two left patches.²²

²²M. Noroozi et al. "Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles". In: *ECCV*. 2016.

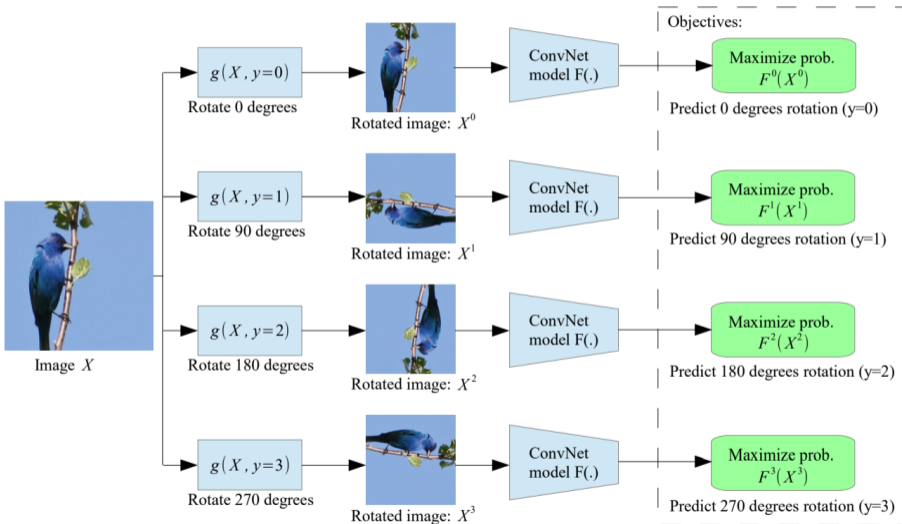
- To avoid trivial shortcuts, low-level trivial signals such as boundary patterns or texture continuing between patches:
 - ▶ Add gaps between patches
 - ▶ Randomly crop each patch
 - ▶ Shuffle several times same image (for jigsaw)
 - ▶ Shift color channels and/or use grayscale images
 - ▶ Add random shifts to each patch
 - ▶ Downsample and then upsample images (robustness to pixelation)

- Instead than predicting the relative position between two patches of an image, predict the geometric transformation applied to the entire image
- Which transformation? → Need for transformations that:
 - ▶ force the network (usually CNN) to recognize all objects depicted in the image and learn high-level semantic features (e.g., class, location, size, pose, etc.)
 - ▶ do not leave easily detectable low-level visual artifacts when applied to images
 - ▶ are well-posed, namely there should not be ambiguity about the original image.

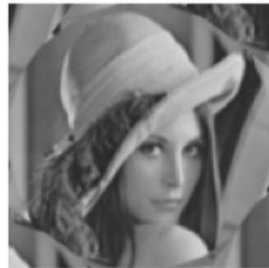
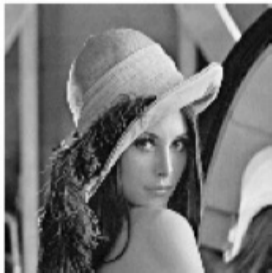
- Instead than predicting the relative position between two patches of an image, predict the geometric transformation applied to the entire image
- Which transformation? → Need for transformations that:
 - ▶ force the network (usually CNN) to recognize all objects depicted in the image and learn high-level semantic features (e.g., class, location, size, pose, etc.)
 - ▶ do not leave easily detectable low-level visual artifacts when applied to images
 - ▶ are well-posed, namely there should not be ambiguity about the original image.
- **Proposed solution: Image rotations by 0, 90, 180, 270 degrees²³**

²³S. Gidaris et al. "Unsupervised Representation Learning by Predicting Image Rotations". In: *ICLR*. 2018.

Transformation prediction



- **No easily detectable low-level visual artifacts:** implementation of rotations only use transpose and flip operations. No interpolations are used which might produce artifacts (note that choice of angles is important !)



Effect of 10 successive rotations of 36 degrees to the original image (left) using Nearest Neighbor interpolation (middle) or Linear interpolation (right). Credits: <http://bigwww.epfl.ch/demo/jaffine/index.html>

- **No easily detectable low-level visual artifacts:** implementation of rotations only use transpose and flip operations. No interpolations are used which might produce artifacts (note that choice of angles is important !)
- **Well-posed transformations:** we assume that all images have a usual and standard position. For instance, when using natural images all objects should be “up-standing”

- **No easily detectable low-level visual artifacts:** implementation of rotations only use transpose and flip operations. No interpolations are used which might produce artifacts (note that choice of angles is important !)
- **Well-posed transformations:** we assume that all images have a usual and standard position. For instance, when using natural images all objects should be “up-standing”
- **Detect objects and focus on high-level semantic features:** CNN networks need to focus on high-level features to correctly retrieve the rotation



Input images on the models



(a) Attention maps of supervised model

(b) Attention maps of our self-supervised model

Attention maps (sum of feature activations raised to the power of 1,2,4 respectively).²⁴

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- Generative models use as pretext task the reconstruction of the modified/corrupted/partially observed original image
- The main differences between the methods in the literature are:
 - ▶ how you modify/corrupt the original image
 - ▶ regularization losses (adversarial loss, sub-networks, etc.)
 - ▶ neural network architectural choices (CNN, ViT, etc.)

Here, we will see:

1. Denoising-based autoencoders
2. Colorization-based autoencoders
3. Gan based methods (ALI/ BiGAN / BigBiGAN)

- Generative models use as pretext task the reconstruction of the modified/corrupted/partially observed original image
- The main differences between the methods in the literature are:
 - ▶ how you modify/corrupt the original image
 - ▶ regularization losses (adversarial loss, sub-networks, etc.)
 - ▶ neural network architectural choices (CNN, ViT, etc.)

Here, we will see:

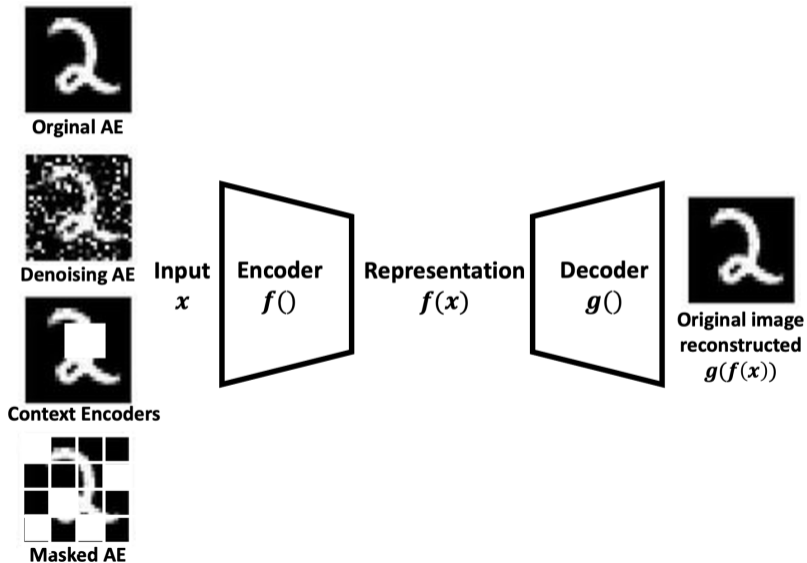
1. Denoising-based autoencoders
2. Colorization-based autoencoders
3. Gan based methods (ALI/ BiGAN / BigBiGAN)

- Usual autoencoders²⁵ are constituted of an encoder $f()$ and a decoder $g()$ and they are trained to minimize the reconstruction error $\|x - g(f(x))\|_2^2$, which amounts to maximizing a lower bound on the mutual information between input x and learnt representation $f(x)$ ²⁶
- To avoid trivial solutions (i.e., $g(f(\cdot)) = I$ identity mapping), the representation $f(x)$ is usually of lower dimension than x (bottleneck), thus compressing the information
- To extract more useful features (i.e., representations), different methods were proposed to reconstruct a “repaired” version of a corrupted original image. The key question is, **how do we corrupt/modify the original image ?**

²⁵G. E. Hinton et al. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* (2006).

²⁶P. Vincent et al. “Stacked Denoising Autoencoders: Learning Useful Representations ...”. In: *JMLR* (2010).

Denoising-based autoencoders



- ▶ **Denoising autoencoders**²⁷: learn to restore a corrupted image where Gaussian/salt-and-pepper noise (or others) has been applied.

- ▶ **Context Encoder/ Inpainting**²⁸: learn to regress pixel values of a large, single missing region using CNN, L2 reconstruction and adversarial losses.

- ▶ **Masked image modeling**²⁹: randomly mask some patches of the image and predict the masked patches from the visible ones using Transformers (e.g., ViT).

²⁷P. Vincent et al. "Extracting and composing robust features with denoising autoencoders". In: *ICML. 2008*.

²⁸D. Pathak et al. "Context Encoders: Feature Learning by Inpainting". In: *CVPR. 2016*.

²⁹K. He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CVPR. 2022*.

- ▶ **Denoising autoencoders**²⁷: learn to restore a corrupted image where Gaussian/salt-and-pepper noise (or others) has been applied.
Problem: localized and low-level corruption, no semantic information needed to repair image
- ▶ **Context Encoder/ Inpainting**²⁸: learn to regress pixel values of a large, single missing region using CNN, L2 reconstruction and adversarial losses.
Problem: 1) image synthesis quality is difficult to evaluate, 2) domain gap between training set (images with holes) and test set (full images), 3) if the mask is not big, no high-level reasoning, just copy low and mid-level neighboring structures
- ▶ **Masked image modeling**²⁹: randomly mask some patches of the image and predict the masked patches from the visible ones using Transformers (e.g., ViT).

²⁷P. Vincent et al. "Extracting and composing robust features with denoising autoencoders". In: *ICML. 2008*.

²⁸D. Pathak et al. "Context Encoders: Feature Learning by Inpainting". In: *CVPR. 2016*.

²⁹K. He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CVPR. 2022*.

- Many methods took inspiration from **Causal language modeling**, such as GPT³⁰, where masked tokens are predicted using only previous (left) tokens, and **Masked language modeling**, such as **BERT**³¹, where masked tokens are recovered using both previous (left) and following (right) tokens
- A direct translation of GPT and BERT from language to vision works well but 1) it is computationally too demanding and 2) it under-performs wrt other contrastive methods³²
- Recent works (BEiT³³, SimMIM³⁴ and **MAE**³⁵) found that the use of ViT, great masking portion and lightweight decoder resulted in SOTA performance.

³⁰A. Radford et al. *Language Models are Unsupervised Multitask Learners*.

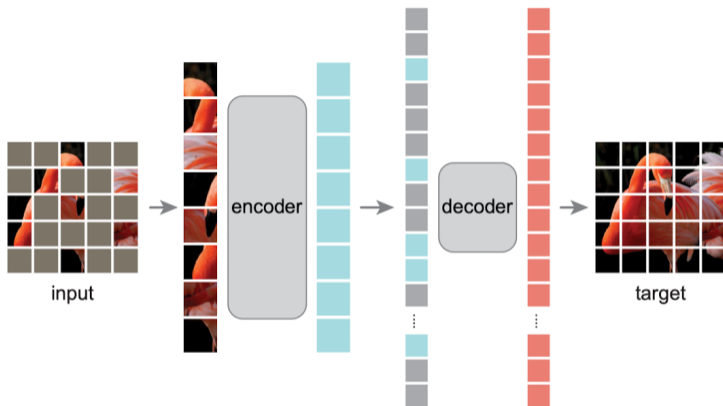
³¹J. Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.

³²M. Chen et al. "Generative Pretraining From Pixels". In: *ICML. 2020*.

³³H. Bao et al. "BEiT: BERT Pre-Training of Image Transformers". In: *ICLR. 2022*.

³⁴Z. Xie et al. "SimMIM: a Simple Framework for Masked Image Modeling". In: *CVPR. 2022*.

³⁵K. He et al. "Masked Autoencoders Are Scalable Vision Learners". In: *CVPR. 2022*.

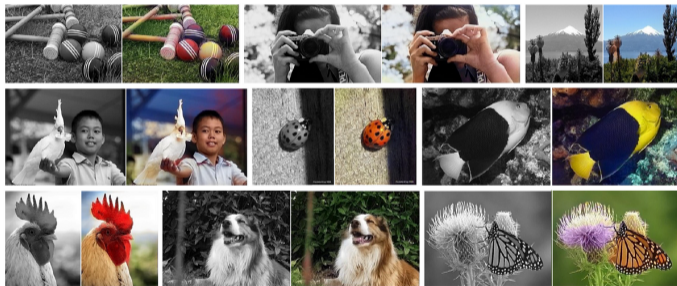


Masked Autoencoders³⁶: encoder ViT takes as input only visible patches. Lightweight decoder takes as input the encoded visible patches and the masked patches (as shared, learned, constant representations) with their positional embeddings.

³⁶K. He et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *CVPR. 2022*.

1. Denoising-based autoencoders
2. Colorization-based autoencoders
3. Gan based methods (ALI/ BiGAN / BigBiGAN)

- Another way to extract useful features in autoencoders is through **colorization**: the encoder takes as input only one channel of the original image (e.g., grayscale/intensity) and the decoder needs to predict the original colors (e.g., RGB/Lab channels)



Choices of training set, color channels, architecture and optimization are important.^{37,38,39}

³⁷R. Zhang et al. "Colorful Image Colorization". In: *ECCV*. 2016.

³⁸G. Larsson et al. "Learning Representations for Automatic Colorization". In: *ECCV*. 2016.

³⁹G. Larsson et al. "Colorization as a Proxy Task for Visual Understanding". In: *CVPR*. 2017.

Pros

- ▶ systematic, and not stochastic, corruption of image remove the pre-training and testing domain gap.
- ▶ Denoising/Inpainting/Masking may use only textural/positional information. Predicting color requires object-level reasoning and thus higher semantic representations.

Cons

- ▶ color does not always carry important semantic information, as in medical imaging
- ▶ Different channels of the input data are not treated equally

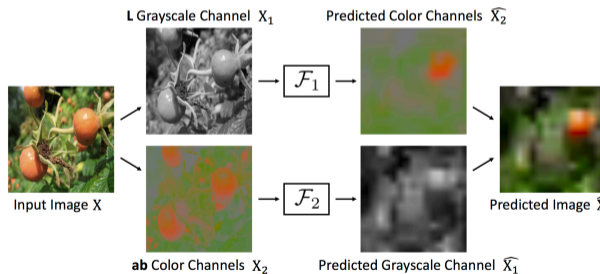
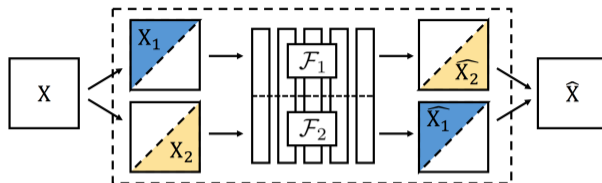
Pros

- ▶ systematic, and not stochastic, corruption of image remove the pre-training and testing domain gap.
- ▶ Denoising/Inpainting/Masking may use only textural/positional information. Predicting color requires object-level reasoning and thus higher semantic representations.

Cons

- ▶ color does not always carry important semantic information, as in medical imaging
- ▶ Different channels of the input data are not treated equally → Possible architectural solution: use multiple sub-networks trained on different but complementary channels^a

^aR. Zhang et al. "Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction". In: *CVPR*. 2017.



Use L channel to predict ab channel and viceversa. Concatenate \hat{X}_1 and \hat{X}_2 to obtain \hat{X}^{40} .

1. Denoising-based autoencoders
2. Colorization-based autoencoders
3. Gan based methods (ALI/ BiGAN / BigBiGAN)

- Autoencoders (AE) and Variational Autoencoders (VAE)⁴¹ are easy to train but tend to produce blurry images
- Other generative models have emerged such as Generative Adversarial Network (GAN)⁴² where a *generator* model G learns a mapping from random variables to input data and a *discriminator* model D learns to distinguish the real data from the fake ones produced by G . The models are trained in a minimax game where G tries to fool D .
- GAN models are known for potentially unstable training and less diversity in generation due to their adversarial training nature. For this reason, other models have been proposed such as Normalizing flows⁴³ and Diffusion models⁴⁴.

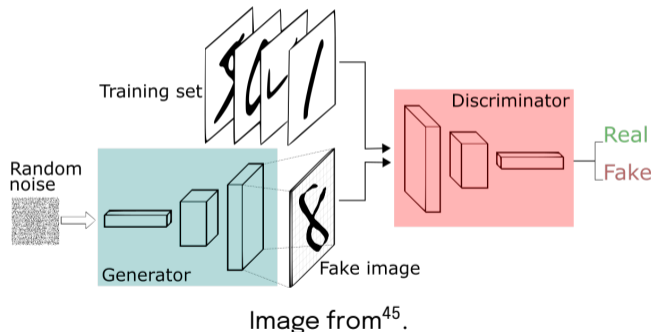
⁴¹D. P. Kingma et al. "Auto-Encoding Variational Bayes". In: *ICLR*. 2014.

⁴²I. Goodfellow et al. "Generative Adversarial Nets". In: *NIPS*. 2014.

⁴³D. J. Rezende et al. "Variational Inference with Normalizing Flows". In: *ICML*. 2015.

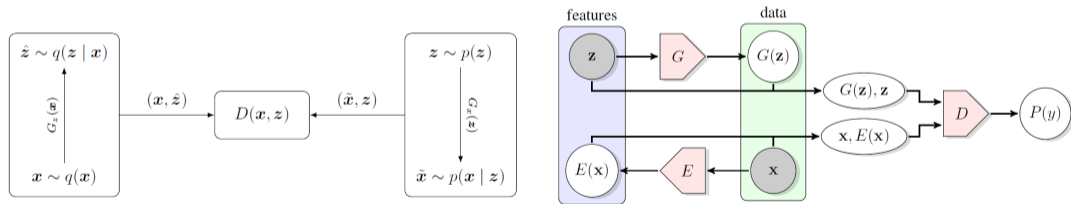
⁴⁴J. Ho et al. "Denosing Diffusion Probabilistic Models". In: *NeurIPS*. 2020.

- However, GANs can not be used to learn rich feature representations in an unsupervised way since they “just” generate images and do not have a bottleneck representation, as in AE/VAE → **How can we use GAN to learn a representation?**



⁴⁵S. Thalles Santos. *A Short Introduction to Generative Adversarial Networks*. URL: <https://sthalles.github.io>.

- However, GANs can not be used to learn rich feature representations in an unsupervised way since they “just” generate images and do not have a bottleneck representation, as in AE/VAE → **How can we use GAN to learn a representation?**
- Two methods, Adversarially Learned Inference (ALI)⁴⁵ and bidirectional GAN (BiGAN)⁴⁶, concurrently proposed to add an encoder E which maps real data to latent representations, the inverse of the mapping learned by the generator G



ALI on the left and BiGAN on the right.

⁴⁵V. Dumoulin et al. “Adversarially Learned Inference”. In: *ICLR*. 2017.

⁴⁶J. Donahue et al. “Adversarial Feature Learning”. In: *ICLR*. 2017.

Algorithm 1 The ALI training procedure.

$\theta_g, \theta_d \leftarrow$ initialize network parameters

repeat

$$\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)} \sim q(\mathbf{x})$$

▷ Draw M samples from the dataset and the prior

$$\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(M)} \sim p(\mathbf{z})$$

$$\hat{\mathbf{z}}^{(i)} \sim q(\mathbf{z} \mid \mathbf{x} = \mathbf{x}^{(i)}), \quad i = 1, \dots, M$$

▷ Sample from the conditionals

$$\tilde{\mathbf{x}}^{(j)} \sim p(\mathbf{x} \mid \mathbf{z} = \mathbf{z}^{(j)}), \quad j = 1, \dots, M$$

$$\rho_q^{(i)} \leftarrow D(\mathbf{x}^{(i)}, \hat{\mathbf{z}}^{(i)}), \quad i = 1, \dots, M$$

▷ Compute discriminator predictions

$$\rho_p^{(j)} \leftarrow D(\tilde{\mathbf{x}}^{(j)}, \mathbf{z}^{(j)}), \quad j = 1, \dots, M$$

$$\mathcal{L}_d \leftarrow -\frac{1}{M} \sum_{i=1}^M \log(\rho_q^{(i)}) - \frac{1}{M} \sum_{j=1}^M \log(1 - \rho_p^{(j)})$$

▷ Compute discriminator loss

$$\mathcal{L}_g \leftarrow -\frac{1}{M} \sum_{i=1}^M \log(1 - \rho_q^{(i)}) - \frac{1}{M} \sum_{j=1}^M \log(\rho_p^{(j)})$$

▷ Compute generator loss

$$\theta_d \leftarrow \theta_d - \nabla_{\theta_d} \mathcal{L}_d$$

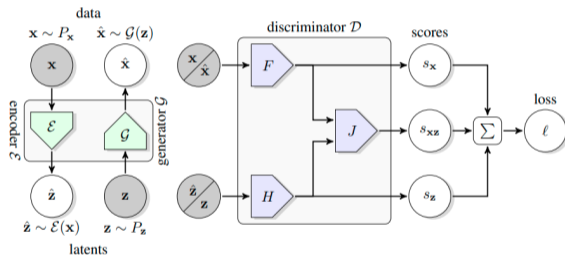
▷ Gradient update on discriminator network

$$\theta_g \leftarrow \theta_g - \nabla_{\theta_g} \mathcal{L}_g$$

▷ Gradient update on generator networks

until convergence

- BiGAN used a simple generator (DCGAN⁴⁷) which can't produce high-quality images and thus can not capture all visual information → poor representation
- BigBiGAN⁴⁸ uses BigGAN⁴⁹ as generator producing SOTA representation results



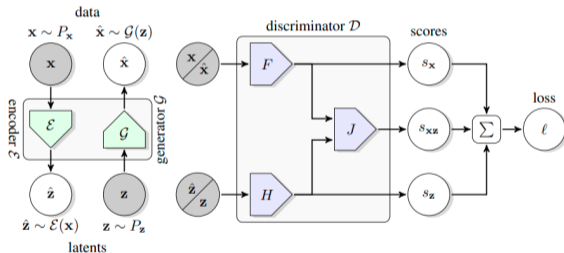
\mathcal{D} has three submodules: two unary $F(x)$, $H(z)$ and a joint one $J(F(x), H(z))$.

⁴⁷A. Radford et al. "Unsupervised Representation Learning with Deep Convolutional GAN". In: *ICLR*. 2016.

⁴⁸J. Donahue et al. "Large Scale Adversarial Representation Learning". In: *NeurIPS*. 2019.

⁴⁹A. Brock et al. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *ICLR*. 2019.

- BiGAN used a simple generator (DCGAN⁴⁷) which can't produce high-quality images and thus can not capture all visual information → poor representation
- BigBiGAN⁴⁸ uses BigGAN⁴⁹ as generator producing SOTA representation results → **More powerful generator can improve the representation quality**



\mathcal{D} has three submodules: two unary $F(x)$, $H(z)$ and a joint one $J(F(x), H(z))$.

⁴⁷ A. Radford et al. "Unsupervised Representation Learning with Deep Convolutional GAN". In: *ICLR*. 2016.

⁴⁸ J. Donahue et al. "Large Scale Adversarial Representation Learning". In: *NeurIPS*. 2019.

⁴⁹ A. Brock et al. "Large Scale GAN Training for High Fidelity Natural Image Synthesis". In: *ICLR*. 2019.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- Usual supervised classification models look for discriminative features that characterize and correctly separate *classes* of objects
- Can we learn a representation where we correctly separate *single instances* (i.e., images) and not classes ?

We'll see four methods:

1. Exemplar-CNN⁵⁰
2. Learning with a non-parametric classifier from a memory bank⁵¹
3. Learning to count⁵²
4. Contrastive Learning⁵³

⁵⁰A. Dosovitskiy et al. "Discriminative Unsupervised Feature Learning with Exemplar CNNs". In: *IEEE TPAMI* (2016).

⁵¹Z. Wu et al. "Unsupervised Feature Learning via Non-parametric Instance Discrimination". In: *CVPR*. 2018.

⁵²M. Noroozi et al. "Representation Learning by Learning to Count". In: *ICCV*. 2017.

⁵³T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

- Usual supervised classification models look for discriminative features that characterize and correctly separate *classes* of objects
- Can we learn a representation where we correctly separate *single instances* (i.e., images) and not classes ?

We'll see four methods:

1. Exemplar-CNN⁵⁰
2. Learning with a non-parametric classifier from a memory bank⁵¹
3. Learning to count⁵²
4. Contrastive Learning⁵³

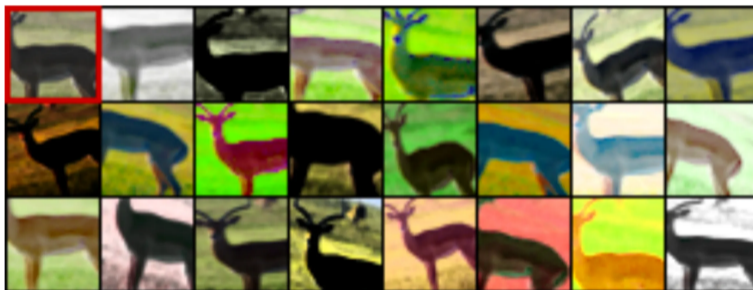
⁵⁰A. Dosovitskiy et al. "Discriminative Unsupervised Feature Learning with Exemplar CNNs". In: *IEEE TPAMI* (2016).

⁵¹Z. Wu et al. "Unsupervised Feature Learning via Non-parametric Instance Discrimination". In: *CVPR*. 2018.

⁵²M. Noroozi et al. "Representation Learning by Learning to Count". In: *ICCV*. 2017.

⁵³T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

- Given a set of different images (or patches) x_i , we first randomly transform K times each image x_i producing a set of transformed images $S_{x_i}^{54}$



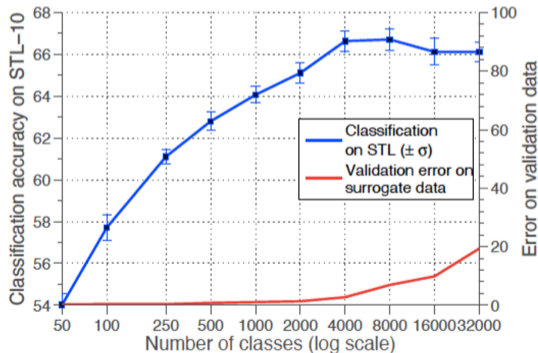
- Given a set of different images (or patches) x_i , we first randomly transform K times each image x_i producing a set of transformed images S_{x_i} ⁵⁴
- Each set S_{x_i} is then considered as a surrogate class with label i and we minimize the classification loss:

$$\mathcal{L}(X) = \sum_i \sum_k \mathbf{I}(i, T_k x_i) \quad (1)$$

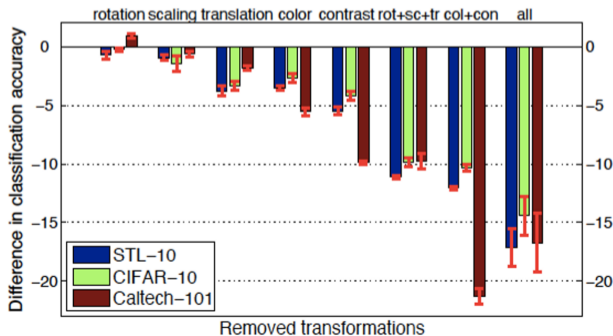
- where T_k is the k -th transformation applied to image x_i and $\mathbf{I}(i, T_k x_i)$ is the cross-entropy (or negative log-likelihood).

⁵⁴A. Dosovitskiy et al. "Discriminative Unsupervised Feature Learning with Exemplar CNNs". In: *IEEE TPAMI* (2016).

- The classification loss of Exemplar-CNN ensures that *different images can be distinguished* and it enforces *invariance to (specified) transformations*
- **Number of training data:** classification accuracy increases until an optimum after which it is likely to draw very similar training images that can be hard to discriminate



- The classification loss of Exemplar-CNN ensures that *different images can be distinguished* and it enforces *invariance to (specified) transformations*
- **Number of training data:** classification accuracy increases until an optimum after which it is likely to draw very similar training images that can be hard to discriminate
- **Influence of Transformations:** classification accuracy varies depending on the used transformations and on the data-set



1. Exemplar-CNN⁵⁵
2. Learning with a non-parametric classifier from a memory bank⁵⁶
3. Learning to count⁵⁷
4. Contrastive Learning⁵⁸

⁵⁵A. Dosovitskiy et al. “Discriminative Unsupervised Feature Learning with Exemplar CNNs”. In: *IEEE TPAMI* (2016).

⁵⁶Z. Wu et al. “Unsupervised Feature Learning via Non-parametric Instance Discrimination”. In: *CVPR*. 2018.

⁵⁷M. Noroozi et al. “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

⁵⁸T. Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.

- In Exemplar-CNN we use a parametric classifier based on a softmax.
- Calling W the weights of the last linear layer and $f(x)$ the input to the last linear layer (i.e., representation features of image x), the output of the last linear layer (before softmax) is their matrix product $Wf(x)$
- The loss of Exemplar-CNN is the cross-entropy, namely $\mathbf{I}(i, x) = -\langle y_i, \log(g(x)) \rangle$, where y_i is the one-hot binary vector ($[0, 0, 1, \dots, 0]$ where 1 is at position i) and $g(x) = \text{softmax}(Wf(x))$
- Each row w_i of the matrix W is a classifier of class (here training sample) i . It can be seen as a template/prototype of class i . The softmax function $g(x)$ gives the probability of image x to be the i -th training sample:

$$P(i|f(x); W) = \frac{\exp(w_i^T f(x))}{\sum_j \exp(w_j^T f(x))} \quad (2)$$

$$P(i|f(x); W) = \frac{\exp(w_i^T f(x))}{\sum_j \exp(w_j^T f(x))} \quad (3)$$

- is a parametric classifier since it depends on W . Can we remove it ?⁵⁹

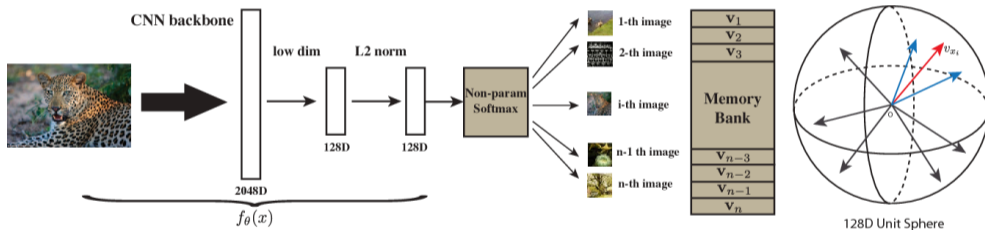
$$P(i|f(x)) = \frac{\exp(f(x_i)^T f(x)/\tau)}{\sum_j \exp(f(x_j)^T f(x)/\tau)} \quad \text{s.t.} \quad \|f(x_t)\|_2 = 1 \quad \forall t \quad (4)$$

- It thus become a non-parametric classifier where:
 - ▶ it does not depend on $W \rightarrow$ no need to compute and store gradient wrt W
 - ▶ we directly compare pair of instances instead than template/prototype
 - ▶ the temperature τ controls the concentration of instances on the unit hyper-sphere \rightarrow important hyper-parameter

⁵⁹Z. Wu et al. "Unsupervised Feature Learning via Non-parametric Instance Discrimination". In: *CVPR*. 2018.

- The loss function is defined as the negative log-likelihood over the training set:

$$\mathcal{L}(X) = - \sum_i \log P(i|f(x_i)) \quad (5)$$



The optimal feature embedding is learned via instance-level discrimination, which tries to maximally scatter the features of training samples over the unit sphere. A memory bank is used to limit computations. At each iteration, only the representation $f(x_i)$ and the network parameters are optimized. The representations $f(x_j)$ of the other samples are kept fixed.⁶⁰

⁶⁰Z. Wu et al. "Unsupervised Feature Learning via Non-parametric Instance Discrimination". In: *CVPR*. 2018.

1. Exemplar-CNN⁶¹
2. Learning with a non-parametric classifier from a memory bank⁶²
3. Learning to count⁶³
4. Contrastive Learning⁶⁴

⁶¹A. Dosovitskiy et al. “Discriminative Unsupervised Feature Learning with Exemplar CNNs”. In: *IEEE TPAMI* (2016).

⁶²Z. Wu et al. “Unsupervised Feature Learning via Non-parametric Instance Discrimination”. In: *CVPR*. 2018.

⁶³M. Noroozi et al. “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

⁶⁴T. Chen et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.

- An image contains a certain number of *visual primitives* that describe and characterize its information content → their number should not vary when applying geometric transformations (scaling, rotation, translation, etc.)

⁶⁵[M. Noroozi et al.](#) “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

- An image contains a certain number of *visual primitives* that describe and characterize its information content → their number should not vary when applying geometric transformations (scaling, rotation, translation, etc.)
- Instead, when dividing the image into half or more parts, the number of visual primitives should decrease

⁶⁵[M. Noroozi et al. "Representation Learning by Learning to Count". In: ICCV. 2017.](#)

- An image contains a certain number of *visual primitives* that describe and characterize its information content → their number should not vary when applying geometric transformations (scaling, rotation, translation, etc.)
- Instead, when dividing the image into half or more parts, the number of visual primitives should decrease
- We look for a representation (i.e., model) f that correctly counts the number of visual primitives of an image

⁶⁵M. Noroozi et al. “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

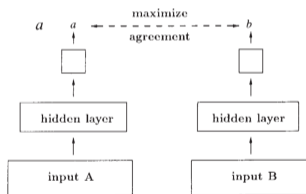
- An image contains a certain number of *visual primitives* that describe and characterize its information content → their number should not vary when applying geometric transformations (scaling, rotation, translation, etc.)
- Instead, when dividing the image into half or more parts, the number of visual primitives should decrease
- We look for a representation (i.e., model) f that correctly counts the number of visual primitives of an image
- If we consider as transformations the down-sampling operator \mathcal{D} and the tiling (regular crop) operator \mathcal{T}_j , we look for a representation f where⁶⁵

$$f(\mathcal{D} \circ x) = \sum_j f(\mathcal{T}_j \circ x) \quad (6)$$

⁶⁵M. Noroozi et al. “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

- In⁶⁶ authors propose to use a **Siamese architecture**, initially proposed in 1992/1993 in^{67,68}, where the same network f maximizes agreement between $f(\mathcal{D} \circ x)$ and $\sum_j f(\mathcal{T}_j \circ x)$ by using as loss $\mathcal{L}(X) = \|f(\mathcal{D} \circ x) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2$.

Do you see any problem? Trivial solution?



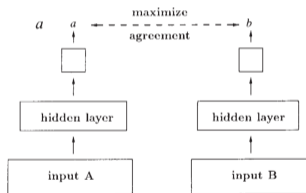
⁶⁶M. Noroozi et al. "Representation Learning by Learning to Count". In: *ICCV*. 2017.

⁶⁷S. Becker et al. "Self-organizing neural network that discovers surfaces in random ...". In: *Nature* (1992).

⁶⁸J. Bromley et al. "Signature Verification using a "Siamese" Time Delay Neural Network". In: *NIPS*. vol. 6. 1993.

- In⁶⁶ authors propose to use a **Siamese architecture**, initially proposed in 1992/1993 in^{67,68}, where the same network f maximizes agreement between $f(\mathcal{D} \circ x)$ and $\sum_j f(\mathcal{T}_j \circ x)$ by using as loss $\mathcal{L}(X) = \|f(\mathcal{D} \circ x) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2$.

Do you see any problem? Trivial solution?



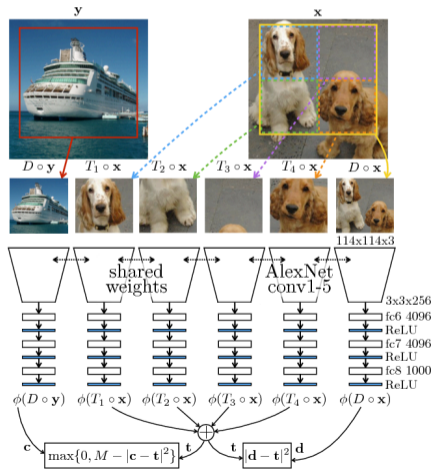
- It may produce the trivial solution $f(x) = 0 \forall x$, which results in the global minimum 0

⁶⁶M. Noroozi et al. "Representation Learning by Learning to Count". In: *ICCV*. 2017.

⁶⁷S. Becker et al. "Self-organizing neural network that discovers surfaces in random ...". In: *Nature* (1992).

⁶⁸J. Bromley et al. "Signature Verification using a "Siamese" Time Delay Neural Network". In: *NIPS*. vol. 6. 1993.

- To this end, a possible solution is using a **pairwise contrastive loss** where we enforce that the number of visual primitives should be different between two different images x and y :



- In the **pairwise contrastive loss**, we force:
 1. transformations of x , that should preserve the number of visual features, to be mapped to the same point in the representation space
 2. transformations of different samples, x and y , to have a distance greater than a positive margin ϵ ($\epsilon \geq 0$) in the representation space: $\|f(\mathcal{D} \circ y) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 > \epsilon$

$$\arg \min_f \mathcal{L}(x, y) = \begin{cases} \|f(\mathcal{D} \circ x) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 \\ \max(0, \epsilon - \|f(\mathcal{D} \circ y) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 \end{cases} \quad (7)$$

- **Problems:**

- ▶ Why just using the down-sampling operator \mathcal{D} and the tiling operator \mathcal{T}_j ?
- ▶ Why using a single “negative” sample y ?

- In the **pairwise contrastive loss**, we force:
 1. transformations of x , that should preserve the number of visual features, to be mapped to the same point in the representation space
 2. transformations of different samples, x and y , to have a distance greater than a positive margin ϵ ($\epsilon \geq 0$) in the representation space: $\|f(\mathcal{D} \circ y) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 > \epsilon$

$$\arg \min_f \mathcal{L}(x, y) = \begin{cases} \|f(\mathcal{D} \circ x) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 \\ \max(0, \epsilon - \|f(\mathcal{D} \circ y) - \sum_j f(\mathcal{T}_j \circ x)\|_2^2 \end{cases} \quad (7)$$

- **Problems:**
 - ▶ Why just using the down-sampling operator \mathcal{D} and the tiling operator \mathcal{T}_j ?
 - ▶ Why using a single “negative” sample y ?
- **Take-home message:** information preserving *transformations*, *instance discrimination* (without class/prototype), *contrastive loss* are three important ingredients

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

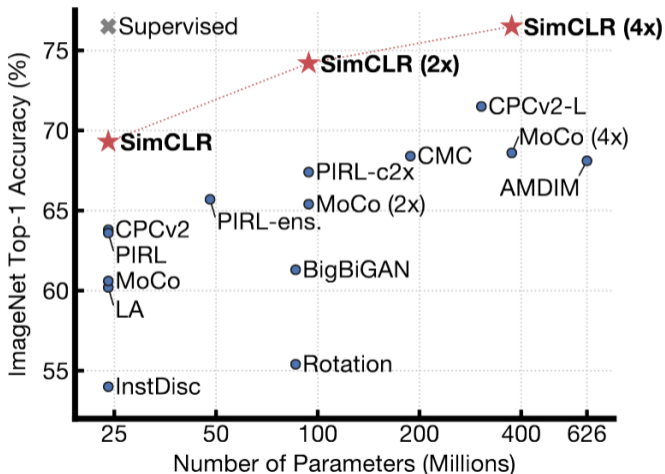
4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

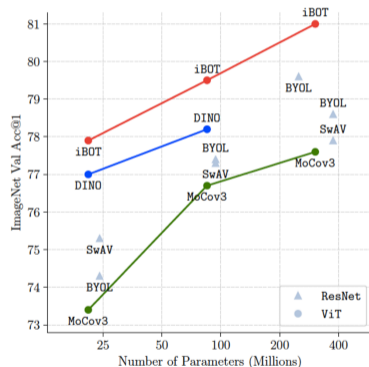
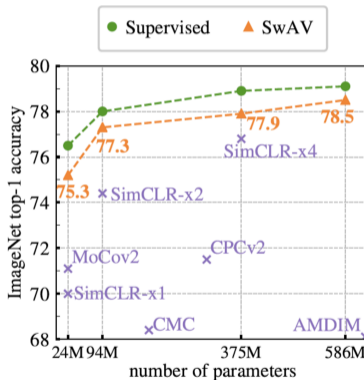
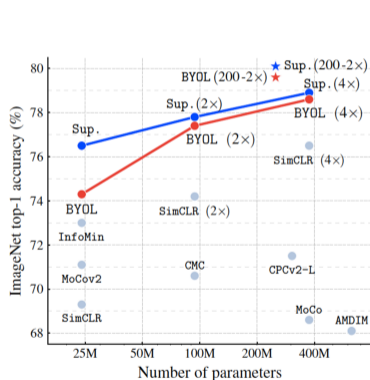
5. Conclusions

- Contrastive learning methods outperform the other pretext tasks⁶⁹



⁶⁹T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

- And recently there has been a plethora of works about it that is closing the performance gap with supervised pretraining^{70,71,72}



⁷⁰J.-B. Grill et al. "Bootstrap your own latent: A new approach to self-supervised Learning". In: *NeurIPS*. 2020.

⁷¹M. Caron et al. "Unsupervised Learning of Visual Features by Contrasting Cluster Assignments". In: *NeurIPS*. 2020.

⁷²J. Zhou et al. "Image BERT Pre-training with Online Tokenizer". In: *ICLR*. 2022.

- **Goal:** given a set of images $x_k \in \mathcal{X}$, learn a mapping function $f_\theta : \mathcal{X} \rightarrow \mathcal{F}$ such that:

if x_a and x_b are semantically similar $\rightarrow f(x_a) \approx f(x_b)$

if x_a and x_b are semantically different $\rightarrow f(x_a) \neq f(x_b)$

- These conditions can be reformulated from a mathematical point using either a **geometric approach**, based on a distance $d(f(x_a), f(x_b))$, or an **information theoretic approach**, based on a statistical dependence measure, such as Mutual Information $I(f(x_a), f(x_b))$.

if x_a and x_b are semantically similar \rightarrow

$$\arg \min_f d(f(x_a), f(x_b))$$

$$\arg \max_f I(f(x_a), f(x_b))$$

if x_a and x_b are semantically different \rightarrow

$$\arg \max_f d(f(x_a), f(x_b))$$

$$\arg \min_f I(f(x_a), f(x_b))$$

Geometric approach (Y. LeCun)

- ▶ Pairwise loss^a
- ▶ Triplet loss^b
- ▶ Tuplet loss^{c,d,e}

^aS. Chopra et al. "Learning a Similarity Metric Discriminatively, with Application to Face Verification". In: *CVPR*. 2005.

^bF. Schroff et al. "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CVPR*. 2015.

^cH. O. Song et al. "Deep Metric Learning via Lifted Structured Feature Embedding". In: *CVPR*. 2016.

^dK. Sohn. "Improved Deep Metric Learning with Multi-class N-pair Loss Objective". In: *NIPS*. 2016.

^eB. Yu et al. "Deep Metric Learning With Tuplet Margin Loss". In: *ICCV*. 2019.

Information theory approach (G. Hinton)

- ▶ Soft Nearest Neighbor^{a,b}
- ▶ Contrastive Predictive Coding (CPC)^c
- ▶ Non-Parametric Instance Discrimination^d
- ▶ Deep InfoMax (DIM)^e

^aR. Salakhutdinov et al. "Learning a Nonlinear Embedding by Preserving Class ...". In: *AISTATS*. 2007.

^bN. Frosst et al. "Analyzing and Improving Representations with the Soft Nearest Neighbor". In: *ICML*. 2019.

^cA. v. d. Oord et al. *Representation Learning with Contrastive Predictive Coding*. 2018.

^dZ. Wu et al. "Unsupervised Feature Learning via Non-parametric Instance Discrimination". In: *CVPR*. 2018.

^eR. D. Hjelm et al. "Learning deep representations by mutual information estimation ...". In: *ICLR*. 2019.

Geometric approach (Y. LeCun)^a

- ▶ Need to define positive (x, x^+) (semantically similar) and negative pairs (x, x^-) (semantically different)
- ▶ Need to define similarity measure (or distance) that is maximized (or minimized)
- ▶ No constraints/hypotheses about negative samples

^aS. Chopra et al. "Learning a Similarity Metric Discriminatively, with Application to Face Verification". In: *CVPR*. 2005.

Information theory approach (G. Hinton)^a

- ▶ Need to define **pdf** of positive $(x, x^+) \sim p(x, x^+)$ and negative pairs $(x, x^-) \sim p(x)p(x^-)$ where $x^- \perp\!\!\!\perp x, x^+$
- ▶ **Maximize Mutual Information** (I) between positive pairs, given **independent** negative pairs: $I(x; x^+) = I(x; x^+, x^-) = \mathbb{E}_{x^- \sim p(x^-)} I(x; x^+)^b$
- ▶ Need to define an **estimator** of I

^aS. Becker et al. "Self-organizing neural network that discovers surfaces in random ...". In: *Nature* (1992).

^bB. Poole et al. "On Variational Bounds of Mutual Information". In: *ICML*. 2019.

- The Information theoretic approach is mathematically sounded and well grounded on the role of Mutual Information (I) estimation in representation learning.
- But ... Large I is **not necessarily predictive** of downstream performance. Good results may depend on architecture choices and inductive biases rather than an accurate I estimation⁷³
- Furthermore, a geometric approach:
 - ▶ is easy to understand and explain
 - ▶ can easily formalize abstract ideas for defining new losses or regularization terms (e.g., data biases)
 - ▶ No need of implausible hypothesis (e.g., negative samples independence).

⁷³M. Tschannen et al. "On Mutual Information Maximization for Representation Learning". In: *ICLR. 2020*.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- ▶ Let $x \in \mathcal{X}$ be a sample (*anchor*)
- ▶ Let x_i^+ be a similar (positive) sample
- ▶ Let x_j^- be a different (negative) sample
- ▶ Let P be the number of positive samples
- ▶ Let N be the number of negative samples
- ▶ Let $f: \mathcal{X} \rightarrow \mathbb{S}^{d-1}$ be the mapping
- ▶ Let $\mathcal{F} = \mathbb{S}^{d-1}$, a $(d-1)$ -sphere

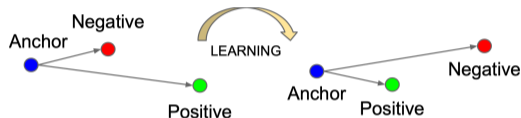


Figure: From Schroff *et al.*^a

^aF. Schroff *et al.* "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CVPR. 2015*.

- ▶ Let $x \in \mathcal{X}$ be a sample (*anchor*)
- ▶ Let x_i^+ be a similar (positive) sample
- ▶ Let x_j^- be a different (negative) sample
- ▶ Let P be the number of positive samples
- ▶ Let N be the number of negative samples
- ▶ Let $f: \mathcal{X} \rightarrow \mathbb{S}^{d-1}$ be the mapping
- ▶ Let $\mathcal{F} = \mathbb{S}^{d-1}$, a $(d-1)$ -sphere

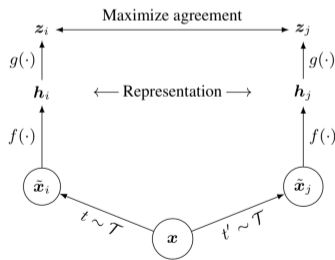


Figure: From Schroff *et al.*^a

^aF. Schroff *et al.* "FaceNet: A Unified Embedding for Face Recognition and Clustering". In: *CVPR*. 2015.

How can we define positive and negative samples ?

- Most methods use a specific architecture:
 - ▶ **Siamese architecture**, initially proposed in 1992/1993 in^{74,75}, where two networks (the same or related as Teacher/Student) maximize agreement between positive samples and minimize agreement between negative samples
 - ▶ Each network is usually divided into two parts: a representation $f(\cdot)$ and a (non-linear) projector $g(\cdot)$ network, which improves performance. Why?



⁷⁴S. Becker et al. "Self-organizing neural network that discovers surfaces in random ...". In: *Nature* (1992).

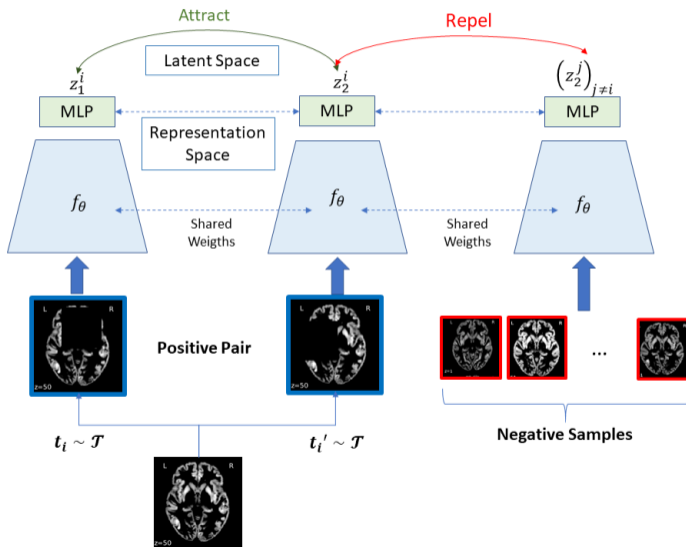
⁷⁵J. Bromley et al. "Signature Verification using a "Siamese" Time Delay Neural Network". In: *NIPS*. vol. 6. 1993.

- Positive samples x_i^+ can be defined in different ways:
 - ▶ **Unsupervised setting (no label):** x_i^+ is a transformation of the anchor x^{76} or a nearest-neighbor from a support set⁷⁷.

⁷⁶T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

⁷⁷D. Dwibedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

Unsupervised setting



- Positive samples x_i^+ can be defined in different ways:
 - ▶ **Unsupervised setting (no label):** x_i^+ is a transformation of the anchor x ⁷⁸ or a nearest-neighbor from a support set⁷⁹.
 - ▶ **Supervised classification setting (label):** x_i^+ is a sample belonging to the same class as x .⁸⁰

⁷⁸T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

⁷⁹D. Dwivedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

⁸⁰P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

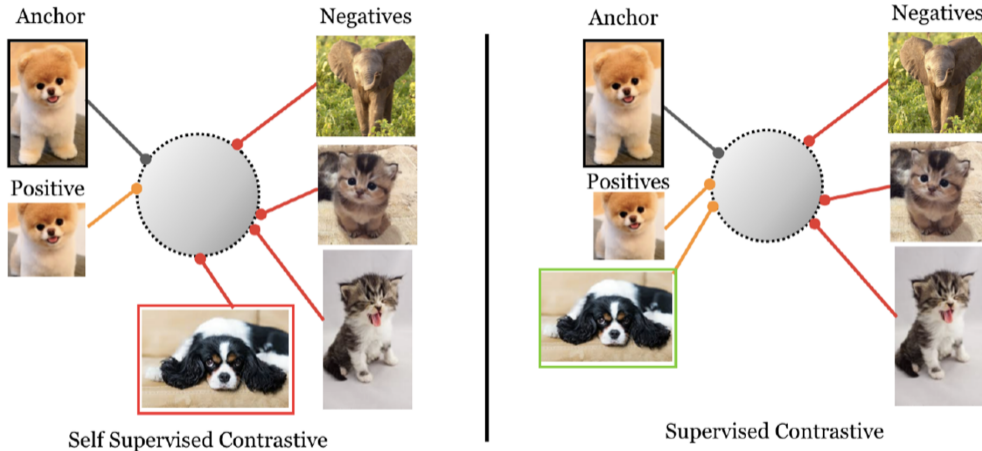


Figure: Image taken from⁸¹

⁸¹P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

- Positive samples x_i^+ can be defined in different ways:
 - ▶ **Unsupervised setting (no label):** x_i^+ is a transformation of the anchor x ⁸² or a nearest-neighbor from a support set⁸³.
 - ▶ **Supervised classification setting (label):** x_i^+ is a sample belonging to the same class as x .⁸⁴
 - ▶ In **regression**⁸⁵ or **weakly-supervised classification**⁸⁶: x_i^+ is a sample with a similar continuous/weak label of x .
- The definition of negative samples x_j^- varies accordingly.

⁸²T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

⁸³D. Dwivedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

⁸⁴P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

⁸⁵C. A. Barbano et al. "Contrastive learning for regression in multi-site brain age prediction". In: *IEEE ISBI*. 2023.

⁸⁶B. Dufumier et al. "Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification". In: *MICCAI*. 2021.

- Positive samples x_i^+ can be defined in different ways:
 - ▶ **Unsupervised setting (no label):** x_i^+ is a transformation of the anchor x^{82} or a nearest-neighbor from a support set⁸³.
 - ▶ **Supervised classification setting (label):** x_i^+ is a sample belonging to the same class as x .⁸⁴
 - ▶ In **regression**⁸⁵ or **weakly-supervised classification**⁸⁶: x_i^+ is a sample with a similar continuous/weak label of x .
- The definition of negative samples x_j^- varies accordingly.

How can we contrast positive and negative samples from a mathematical point of view ?

⁸²T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

⁸³D. Dwibedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

⁸⁴P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

⁸⁵C. A. Barbano et al. "Contrastive learning for regression in multi-site brain age prediction". In: *IEEE ISBI*. 2023.

⁸⁶B. Dufumier et al. "Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification". In: *MICCAI*. 2021.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

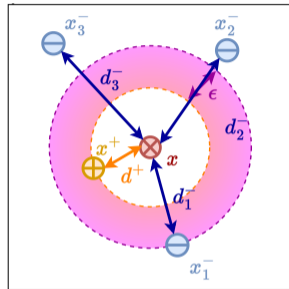
4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- We propose to use an ϵ -margin metric learning point of view⁸⁷

- If we have a single positive x^+ and several negatives x_j^- (e.g., triplet loss), we look for f such that:



$$\underbrace{d(f(x), f(x^+))}_{d^+} - \underbrace{d(f(x), f(x_j^-))}_{d_j^-} < -\epsilon \iff \underbrace{s(f(x), f(x_j^-))}_{s_j^-} - \underbrace{s(f(x), f(x^+))}_{s^+} \leq -\epsilon \quad \forall j$$

- where $\epsilon \geq 0$ is a margin between positive and negative, $s(f(a), f(b)) = \langle f(a), f(b) \rangle_2$.

⁸⁷C. A. Barbano et al. "Unbiased Supervised Contrastive Learning". In: *ICLR*. 2023.

- We propose to use an ϵ -margin metric learning point of view⁸⁷.
- If we have a single positive x^+ and several negatives x_j^- , we look for f such that:

$$\underbrace{d(f(x), f(x^+))}_{d^+} - \underbrace{d(f(x), f(x_j^-))}_{d_j^-} < -\epsilon \iff \underbrace{s(f(x), f(x_j^-))}_{s_j^-} - \underbrace{s(f(x), f(x^+))}_{s^+} \leq -\epsilon \quad \forall j$$

- where $\epsilon \geq 0$ is a margin between positive and negative, $s(f(a), f(b)) = \langle f(a), f(b) \rangle_2$.
- Two possible ways to transform this Eq. in an optimization problem are:

$$\arg \min_f \max(0, \{s_j^- - s^+ + \epsilon\}_{j=1, \dots, N}) \qquad \arg \min_f \sum_{j=1}^N \max(0, s_j^- - s^+ + \epsilon)$$

- when these losses are $= 0$, the condition is fulfilled. Second is lower-bound of first.

⁸⁷C. A. Barbano et al. "Unbiased Supervised Contrastive Learning". In: *ICLR. 2023*.

LogSumExp operator *LSE*

The *LogSumExp* operator *LSE* is a smooth approximation of the **max** function. It is defined as:

$$\max(x_1, x_2, \dots, x_N) \leq LSE(x_1, x_2, \dots, x_N) = \log\left(\sum_{i=1}^N \exp(x_i)\right)$$

- Using *LSE* with the first problem, we obtain the ϵ - *InfoNCE* loss⁸⁸:

$$\arg \min_f \max(0, \{s_j^- - s^+ + \epsilon\}_{j=1, \dots, N}) \approx \arg \min_f \underbrace{-\log\left(\frac{\exp(s^+)}{\exp(s^+ - \epsilon) + \sum_j \exp(s_j^-)}\right)}_{\epsilon\text{-InfoNCE}}$$

⁸⁸C. A. Barbano et al. "Unbiased Supervised Contrastive Learning". In: *ICLR*. 2023.

- When $\epsilon = 0$, we retrieve the InfoNCE⁸⁹, whereas when $\epsilon \rightarrow \infty$ we obtain the InfoL10 (or Decoupled loss⁹⁰).
- It has been shown⁹¹ that these two losses are lower and upper bound of $I(X^+, X)$:

$$\underbrace{\log \frac{\exp s^+}{\exp s^+ + \sum_j \exp s_j^-}}_{\text{InfoNCE}} \leq I(X^+, X) \leq \underbrace{\log \frac{\exp s^+}{\sum_j \exp s_j^-}}_{\text{InfoL10}} \quad (8)$$

- Changing $\epsilon \in [0, \infty)$ can bring to a tighter approximation of $I(X^+, X)$. The exponential function at the denominator $\exp(-\epsilon)$ monotonically decreases as ϵ increases.

⁸⁹A. v. d. Oord et al. *Representation Learning with Contrastive Predictive Coding*. 2018.

⁹⁰C.-H. Yeh et al. "Decoupled Contrastive Learning". In: *ECCV*. 2022.

⁹¹B. Poole et al. "On Variational Bounds of Mutual Information". In: *ICML*. 2019.

- The inclusion of multiple positive samples (\mathbf{s}_i^+) can lead to different formulations (see⁹²). Here, we use the simplest one:

$$\begin{aligned} & \mathbf{s}_j^- - \mathbf{s}_i^+ \leq -\epsilon \quad \forall i, j \\ & \sum_i \max(-\epsilon, \{\mathbf{s}_j^- - \mathbf{s}_i^+\}_{j=1, \dots, N}) \approx \underbrace{- \sum_i \log \left(\frac{\exp(\mathbf{s}_i^+)}{\exp(\mathbf{s}_i^+ - \epsilon) + \sum_j \exp(\mathbf{s}_j^-)} \right)}_{\epsilon\text{-SupInfoNCE}} \end{aligned} \quad (9)$$

- Another formulation is the SupCon loss⁹³, which has been presented as the “most straightforward way to generalize” the InfoNCE loss with multiple positive. However...

⁹²C. A. Barbano et al. “Unbiased Supervised Contrastive Learning”. In: *ICLR*. 2023.

⁹³P. Khosla et al. “Supervised Contrastive Learning”. In: *NeurIPS*. 2020.

- ... it actually contains a **non-contrastive constraint**⁹⁴ on the positive samples:
 $s_t^+ - s_i^+ \leq 0 \quad \forall i, t.$

$$s_j^- - s_i^+ \leq -\epsilon \quad \forall i, j \quad \text{and} \quad s_t^+ - s_i^+ \leq 0 \quad \forall i, t \neq i$$

$$\frac{1}{P} \sum_i \max(0, \{s_j^- - s_i^+ + \epsilon\}_j, \{s_t^+ - s_i^+\}_{t \neq i}) \approx \underbrace{\epsilon - \frac{1}{P} \sum_i \log \left(\frac{\exp(s_i^+)}{\sum_t \exp(s_t^+ - \epsilon) + \sum_j \exp(s_j^-)} \right)}_{\epsilon\text{-SupCon}}$$

- when $\epsilon = 0$ we retrieve exactly \mathcal{L}_{out}^{sup} ⁹⁵.
- One tries to align all positive samples to a single point in the representation space. Thus losing intra-class variability.

⁹⁴C. A. Barbano et al. "Unbiased Supervised Contrastive Learning". In: *ICLR*. 2023.

⁹⁵P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

Table: Accuracy on vision datasets. SimCLR and Max-Margin results from⁹⁶. Results denoted with * are (re)implemented with mixed precision due to memory constraints.

Dataset	Network	SimCLR	Max-Margin	SimCLR*	CE*	SupCon*	ϵ -SupInfoNCE*
CIFAR-10	ResNet-50	93.6	92.4	91.74 \pm 0.05	94.73 \pm 0.18	95.64 \pm 0.02	96.14 \pm 0.01
CIFAR-100	ResNet-50	70.7	70.5	68.94 \pm 0.12	73.43 \pm 0.08	75.41 \pm 0.19	76.04 \pm 0.01
ImageNet-100	ResNet-50	-	-	66.14 \pm 0.08	82.1 \pm 0.59	81.99 \pm 0.08	83.3 \pm 0.06

Table: Comparison of ϵ -SupInfoNCE and ϵ -SupCon on ImageNet-100 in terms of top-1 accuracy (%).

Loss	$\epsilon = 0.1$	$\epsilon = 0.25$	$\epsilon = 0.5$
ϵ -SupInfoNCE	83.25 \pm 0.39	83.02 \pm 0.41	83.3 \pm 0.06
ϵ -SupCon	82.83 \pm 0.11	82.54 \pm 0.09	82.77 \pm 0.14

⁹⁶P. Khosla et al. "Supervised Contrastive Learning". In: *NeurIPS*. 2020.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

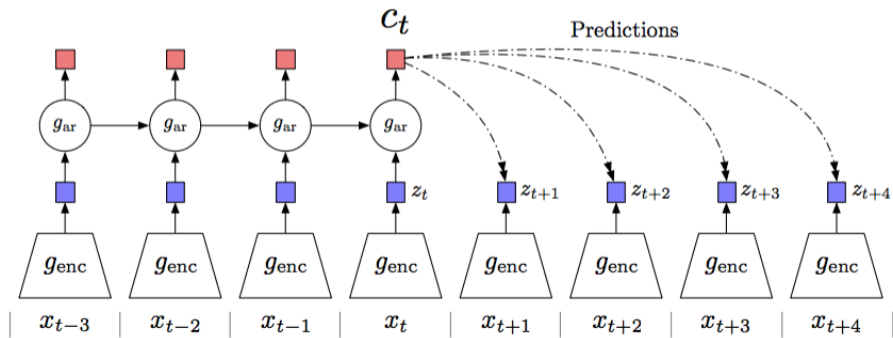
- In a supervised or semi-supervised setting, one can use class labels to define positive and negative samples
- In an unsupervised setting, most methods use a Siamese architecture, the (previously presented) InfoNCE loss⁹⁷ and the fact that positive samples are defined as transformations of the anchor. Methods mainly differ for implementation choices as:
 - ▶ Kind of transformations
 - ▶ Negative/Positive selection
 - ▶ Prototypes/Dictionaryes

⁹⁷A. v. d. Oord et al. *Representation Learning with Contrastive Predictive Coding*. 2018.

- In a supervised or semi-supervised setting, one can use class labels to define positive and negative samples
- In an unsupervised setting, most methods use a Siamese architecture, the (previously presented) InfoNCE loss⁹⁷ and the fact that positive samples are defined as transformations of the anchor. Methods mainly differ for implementation choices as:
 - ▶ Kind of transformations
 - ▶ Negative/Positive selection
 - ▶ Prototypes/Dictionaryes

⁹⁷A. v. d. Oord et al. *Representation Learning with Contrastive Predictive Coding*. 2018.

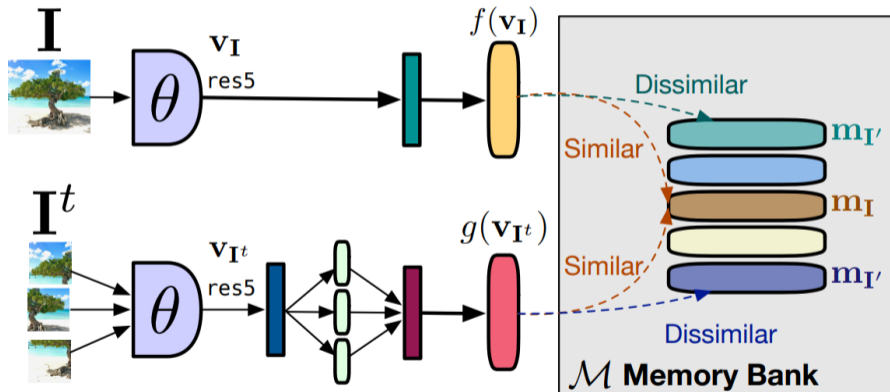
- In Contrastive Predicting Coding (CPCv2)⁹⁸ authors combines context prediction and causal modeling by predicting features of patches using only features from patches that lie above. All images are divided into (overlapping) patches in the same way.



⁹⁸O. Henaff et al. "Data-Efficient Image Recognition with Contrastive Predictive Coding". In: *ICML. 2020*.

Kind of transformations

- In PIRL⁹⁸ authors use Rotation and Jigsaw as transformations.

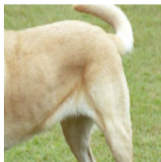


⁹⁸I. Misra et al. "Self-Supervised Learning of Pretext-Invariant Representations". In: *CVPR*. 2020.

- In SimCLR⁹⁸, authors propose a much simpler and more effective (still SOTA) solution using only **random cropping**, **cutout** and **color distortion** (when needed).



(a) Original



(b) Crop and resize



(c) Crop, resize (and flip)



(d) Color distort. (drop)



(e) Color distort. (jitter)



(f) Rotate $\{90^\circ, 180^\circ, 270^\circ\}$



(g) Cutout



(h) Gaussian noise



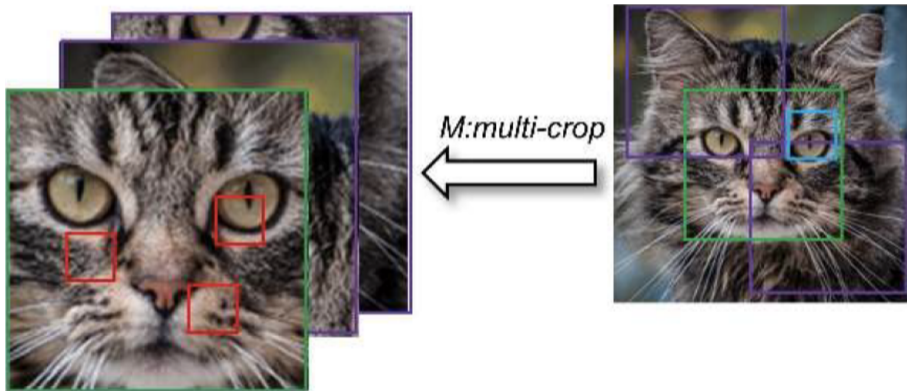
(i) Gaussian blur



(j) Sobel filtering

⁹⁸T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.

- Using multiple cropping (multi-crop^{98,99}) with different sizes increases the number of positives without computational over-head

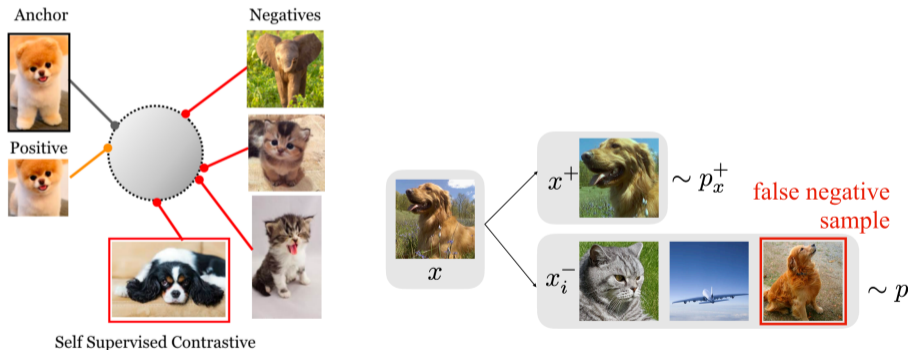


⁹⁸M. Caron et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *NeurIPS*. 2020.

⁹⁹C. Zhao et al. *Multi-crop Contrastive Learning for Unsupervised Image-to-Image Translation*. 2023.

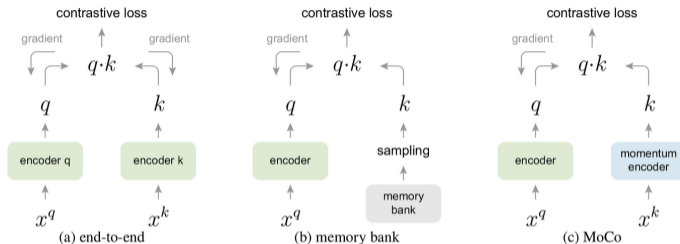
- ▶ Kind of transformations
- ▶ **Negative/Positive selection**
- ▶ Prototypes/Dictionaryes

- Negative selection in an unsupervised setting is very important. Negative samples are usually selected among the other images in the data-set or batch. However, they could actually belong to the same latent class^{100,101}



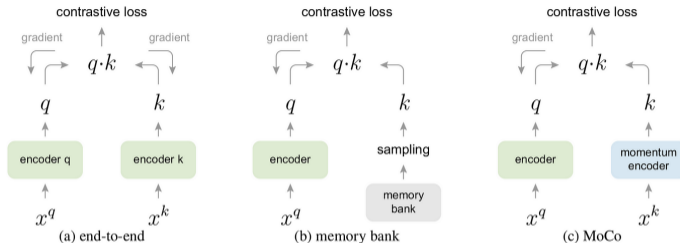
¹⁰⁰P. Khosla et al. “Supervised Contrastive Learning”. In: *NeurIPS*. 2020.

¹⁰¹C.-Y. Chuang et al. “Debiased Contrastive Learning”. In: *NeurIPS*. 2020.



- To limit this issue, since we don't have labels, authors in SimCLR¹⁰² proposed to use very large batch size (and thus very large computational resources) thus increasing the chance of having actual negatives

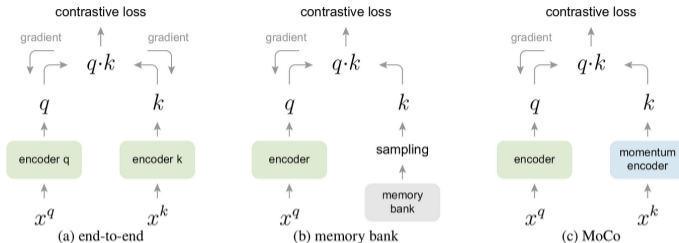
¹⁰²T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML*. 2020.



- Large **memory bank**: representations of *all* samples¹⁰² in the data-set or a subset¹⁰³ (FIFO queue: old batches replaced by new batches).
- At each iteration, a batch is randomly sampled and updated. The representation of a sample in the bank is updated *only* when it is in the batch, otherwise is kept fixed. Representations updated at different moments are not consistent !

¹⁰²I. Misra et al. "Self-Supervised Learning of Pretext-Invariant Representations". In: *CVPR. 2020*.

¹⁰³K. He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CVPR. 2020*.



- **Momentum encoder**¹⁰²: updating via back-propagation the representations of the bank is intractable (too large). Use two encoders (siamese), one for the anchor f_q and one for the bank f_k . The parameters of f_q are updated via back-propagating the gradients of the batch and the parameters of f_k using a momentum m . Only parameters and not representations are updated. More memory-efficient and scalable:

$$\theta_k \leftarrow m\theta_k + (1 - m)\theta_q \quad (10)$$

¹⁰²K. He et al. "Momentum Contrast for Unsupervised Visual Representation Learning". In: *CVPR. 2020*.

- **Importance sampling techniques** (aka Debiasing), as in^{103,104}, use an estimator of the True Negatives (TN) based on (user defined) priors about False Negatives (FN) and on several multiple positives (augmentations of the anchor)
- **Decoupled Contrastive Learning (InfoL10 loss)**¹⁰⁵ removes the coupling term (positive term) at the denominator of the InfoNCE loss which hampers performance when:
 1. a positive sample is very close to the anchor
 2. negative samples are far away from the anchor
 3. there is only a small number of negative samples (i.e., a small batch size)

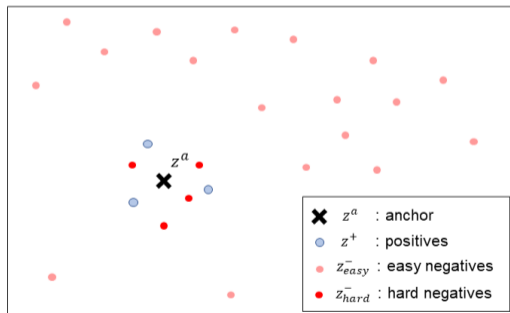
$$\underbrace{\log \frac{\exp s^+}{\exp s^+ + \sum_j \exp s_j^-}}_{\text{InfoNCE}} \rightarrow \underbrace{\log \frac{\exp s^+}{\sum_j \exp s_j^-}}_{\text{Decoupled InfoNCE}} \quad (11)$$

¹⁰³ C.-Y. Chuang et al. “Debiased Contrastive Learning”. In: *NeurIPS*. 2020.

¹⁰⁴ J. Robinson et al. “Contrastive Learning with Hard Negative Samples”. In: *ICLR*. 2021.

¹⁰⁵ C.-H. Yeh et al. “Decoupled Contrastive Learning”. In: *ECCV*. 2022.

- **Hard negatives** are samples that are mapped nearby the anchor, since they are difficult to distinguish from it, but should be far apart¹⁰⁶
- In metric learning^{107,108}, it has been shown that knowing *true* hard negatives can help guide a learning method to correct its mistakes more quickly



¹⁰⁶T. Jang et al. “Difficulty-Based Sampling for Debiased Contrastive Representation Learning”. In: *CVPR*. 2023.

¹⁰⁷F. Schroff et al. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *CVPR*. 2015.

¹⁰⁸H. O. Song et al. “Deep Metric Learning via Lifted Structured Feature Embedding”. In: *CVPR*. 2016.

- Hard negatives can be selected using sampling strategies that are based on the (estimated) similarity at each iteration and that take into account multiple positives¹⁰⁹
- “Semi-hard negative” can be better since easy negatives (far away) do not help learning and too hard negatives might be outliers/noisy¹¹⁰
- The use of a temperature parameter τ in the InfoNCE loss plays a role in controlling the “hardness”¹¹¹. Low τ : focus on the hard negatives. High τ : same importance.

$$\log \frac{\exp(\mathbf{s}^+/\tau)}{\exp(\mathbf{s}^+/\tau) + \sum_j \exp(\mathbf{s}_j^-/\tau)} \quad (12)$$

- Adversarial training can produce more challenging positives and negatives¹¹²

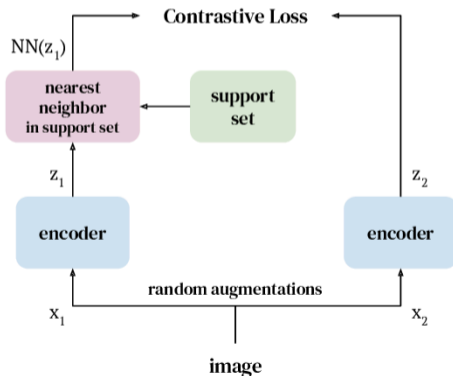
¹⁰⁹J. Robinson et al. “Contrastive Learning with Hard Negative Samples”. In: *ICLR*. 2021.

¹¹⁰C.-Y. Wu et al. “Sampling Matters in Deep Embedding Learning”. In: *ICCV*. 2017.

¹¹¹F. Wang et al. “Understanding the Behaviour of Contrastive Loss”. In: *CVPR*. 2021.

¹¹²C.-H. Ho et al. “Contrastive Learning with Adversarial Examples”. In: *NeurIPS*. 2020.

- In¹¹³, authors use the nearest-neighbour from a support set (FIFO queue) to sample more positives (and not negatives as before).
- The support set must be big enough, is initialized as a random matrix and is updated using the current batch (as a FIFO queue)



¹¹³D. Dwibedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

- Building upon the InfoNCE loss, they propose as loss¹¹⁴:

$$\mathcal{L}_i^{NNCLR} = -\log \frac{\exp(\text{NN}(z_i, Q) \cdot z_i^+ / \tau)}{\sum_j \exp(\text{NN}(z_i, Q) \cdot z_j / \tau)} \quad (13)$$

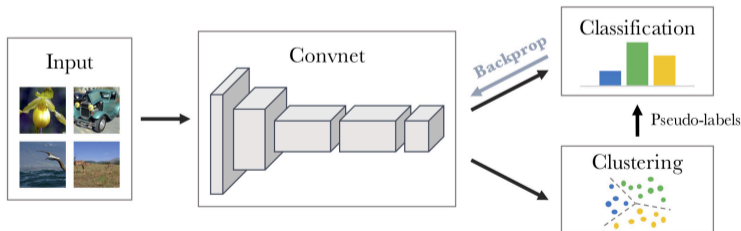
- where Q is the support set, z_i is the representation of the anchor, z_i^+ is an augmentation of the anchor, z_j is an augmentation of a negative sample and $\text{NN}(z_i, Q)$ is the nearest neighbour of the anchor in the (current) representation space:

$$\text{NN}(z_i, Q) = \arg \min_{q \in Q} \|z - q\|_2 \quad (14)$$

¹¹⁴D. Dwibedi et al. "With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning". In: *ICCV*. 2021.

- ▶ Kind of transformations
- ▶ Negative/Positive selection
- ▶ Prototypes/Dictionaries

- Combining clustering and self-supervised learning has been lately studied in^{115,116}
- They combine two steps:
 1. Given pseudo-labels (cluster assignment Q), minimize cross-entropy (classification)
 2. Cluster data given features of encoder f



¹¹⁵M. Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: *ECCV. 2018*.

¹¹⁶Y. M. Asano et al. “Self-labelling via simultaneous clustering and representation learning”. In: *ICLR. 2020*.

- We thus want to learn:
 - ▶ an encoder f , that map data x to feature vectors
 - ▶ a classification head h , which is usually a single linear layer followed by a softmax operator to have class probabilities: $\text{softmax}(h \circ f(x))$
 - ▶ a cluster assignment matrix Q
- It can be seen as a EM optimization where we first fix Q and update the parameters of $h \circ f$ and then fix $h \circ f$ and estimate Q
- In¹¹⁷ authors update Q using the K-means algorithm

¹¹⁷M. Caron et al. "Deep Clustering for Unsupervised Learning of Visual Features". In: *ECCV. 2018*.

- We thus want to learn:
 - ▶ an encoder f , that map data x to feature vectors
 - ▶ a classification head h , which is usually a single linear layer followed by a softmax operator to have class probabilities: $\text{softmax}(h \circ f(x))$
 - ▶ a cluster assignment matrix Q
- It can be seen as a EM optimization where we first fix Q and update the parameters of $h \circ f$ and then fix $h \circ f$ and estimate Q
- In¹¹⁷ authors update Q using the K-means algorithm → **Problem: no well-defined objective function and thus no convergence properties**
- In¹¹⁸, authors propose to partition data in equally-sized subsets (optimal transport problem) → well defined mathematical framework

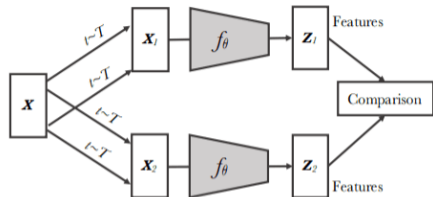
¹¹⁷M. Caron et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: *ECCV*. 2018.

¹¹⁸Y. M. Asano et al. “Self-labelling via simultaneous clustering and representation learning”. In: *ICLR*. 2020.

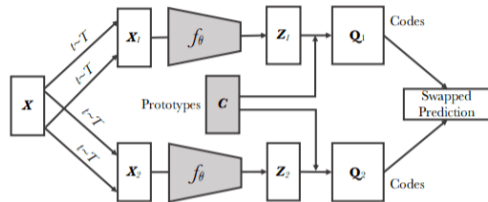
- Clustering methods have two main problems:
 - ▶ They do not scale well with large datasets → need a pass over the entire dataset to obtain cluster assignment
 - ▶ Number of features of penultimate layer of ConvNet is large → need to be PCA-reduced
- To solve this problem one can use prototypes $c^{119,120}$. A code q is obtained for each sample x by mapping its feature $z = f(x)$ to a set of K trainable prototypes c → the dimension of q is K , which is much smaller than the one of z

¹¹⁹M. Caron et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *NeurIPS*. 2020.

¹²⁰J. Li et al. “Prototypical Contrastive Learning of Unsupervised Representations”. In: *ICLR*. 2021.



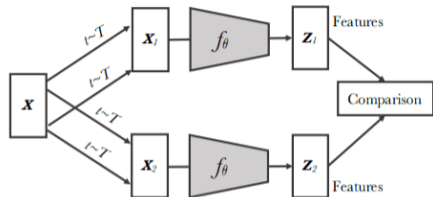
Contrastive instance learning



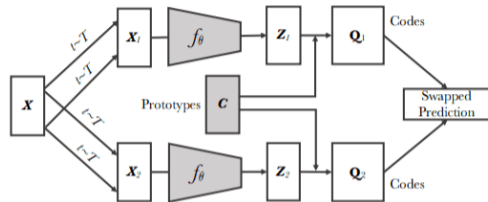
Swapped Assignments between Views (Ours)

- In SwAV¹²¹, authors propose to compute the codes q_s and q_t of two augmentations of the same image x_{is} and x_{it} and predict each code from the feature z of the other image. Given $z_s = f(x_{is})$ and $z_t = f(x_{it})$, we obtain:

$$\mathcal{L}(z_t, z_s) = \ell(z_t, q_s) + \ell(z_s, q_t) \quad (15)$$



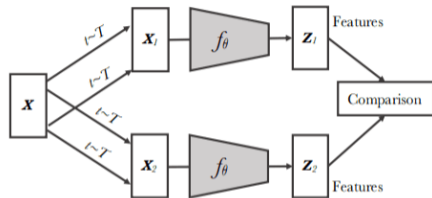
Contrastive instance learning



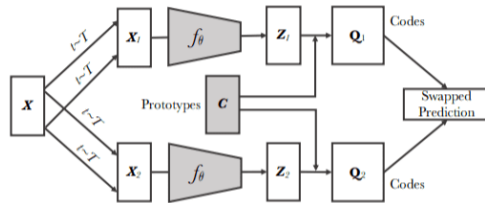
Swapped Assignments between Views (Ours)

$$\mathcal{L}(z_t, z_s) = \ell(z_t, q_s) + \ell(z_s, q_t) \quad (16)$$

- where $\ell(z, q) = -\sum_k q^k \log(p^k)$ measures the fit between the features z and the code q , namely the cross entropy between the code q^k , corresponding to prototype k , and the probability p^k that the feature z belongs to cluster k .

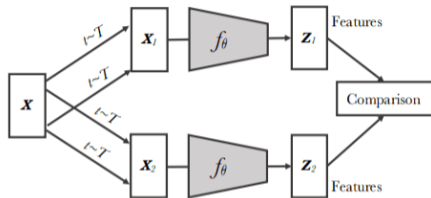


Contrastive instance learning

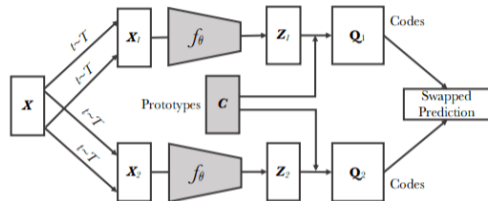


Swapping Assignments between Views (Ours)

- The number of prototypes K is chosen by the user (hyper-parameter) and the soft codes q (no binary assignment) are optimized as to maximize the similarity between the features z and the prototypes c .



Contrastive instance learning



Swapping Assignments between Views (Ours)

- The number of prototypes K is chosen by the user (hyper-parameter) and the soft codes q (no binary assignment) are optimized as to maximize the similarity between the features z and the prototypes c .

Wait a second... and the negatives !? Swav is actually a non-contrastive method ! We'll talk about that in the next section.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

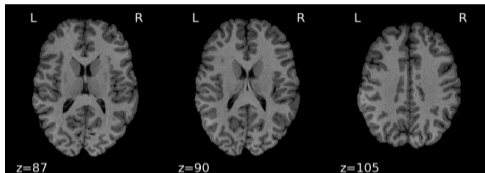
- The previous framework works well when samples are either positive or negative (unsupervised and supervised setting). **But what about continuous/weak labels ?**
- Not possible to determine a hard boundary between positive and negative samples → all samples are positive and negative at the same time

- The previous framework works well when samples are either positive or negative (unsupervised and supervised setting). **But what about continuous/weak labels ?**
- Not possible to determine a hard boundary between positive and negative samples → all samples are positive and negative at the same time
- Let y be the continuous/weak label of the anchor x and y_k of a sample x_k .
- **Simple solution:** threshold d between y and y_k at τ to create positive and negative samples: x_k is x^+ is if $d(y, y_k) < \tau \rightarrow$ **Problem:** how to choose τ ?

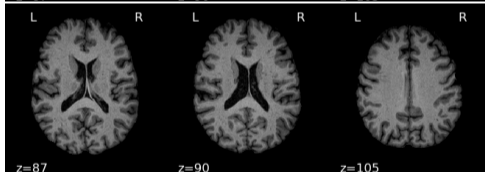
- The previous framework works well when samples are either positive or negative (unsupervised and supervised setting). **But what about continuous/weak labels ?**
- Not possible to determine a hard boundary between positive and negative samples → all samples are positive and negative at the same time
- Let y be the continuous/weak label of the anchor x and y_k of a sample x_k .
- **Simple solution:** threshold d between y and y_k at τ to create positive and negative samples: x_k is x^+ is if $d(y, y_k) < \tau \rightarrow$ **Problem:** how to choose τ ?
- **Our solution:** define a **degree of “positiveness”** between samples using a kernel function $w_k = K_\sigma(y - y_k)$, where $0 \leq w_k \leq 1$.
- **New goal:** learn f that maps samples with a high degree of positiveness ($w_k \sim 1$) close in the latent space and samples with a low degree ($w_k \sim 0$) far away from each other.

Question: Which pair of subjects are closer in your opinion (brain MRI, axial plane) ?

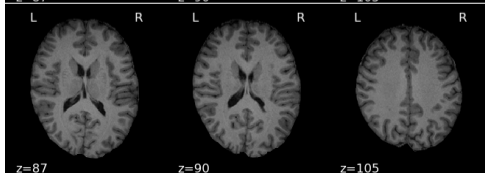
Subject A



Subject B

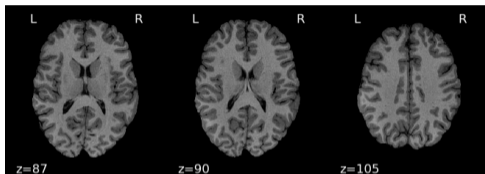


Subject C



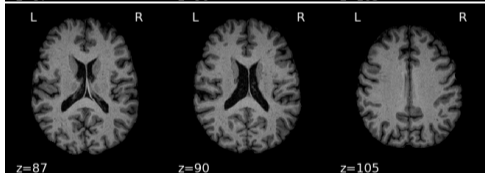
Question: Which pair of subjects are closer in your opinion (brain MRI, axial plane) ?

Subject A



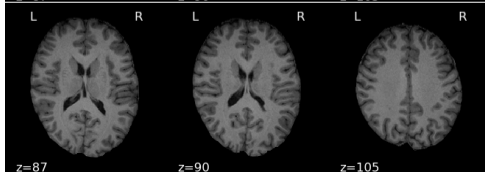
Age=15

Subject B

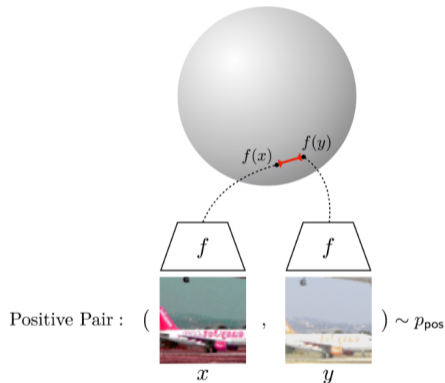


Age=64

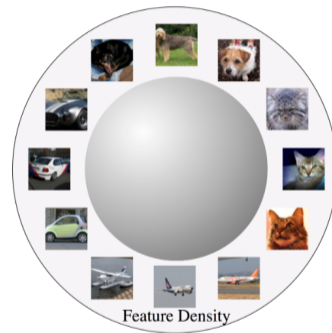
Subject C



Age=20



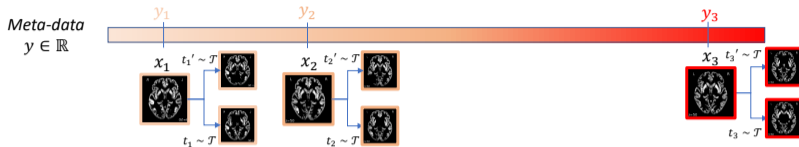
$$\text{Alignment: } \frac{1}{N} \sum_{i=1}^N d_{ii}$$



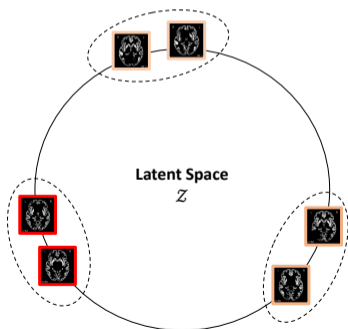
$$\text{Uniformity: } \log \left(\frac{1}{N^2} \sum_{i,j=1}^N e^{-d_{ij}} \right)^{122}$$

¹²²T. Wang et al. "Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere". In: *ICML. 2020*.

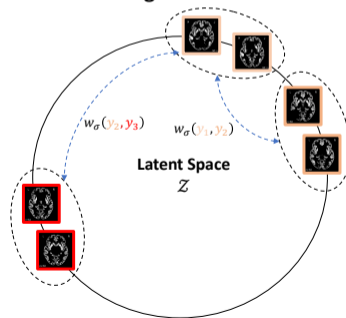
Contrastive Learning - Weakly supervised



SimCLR



y -Aware Contrastive Learning



- In^{123,124}, we propose a new contrastive condition for weakly supervised problems:

$$\frac{w_k}{\sum_j w_j} (s_t - s_k) \leq 0 \quad \forall j, k, t \neq k \in A$$

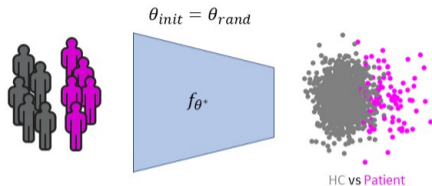
- where A contains the indices of samples $\neq x$ and we consider as positives only the samples with $w_k > 0$, and align them with a strength proportional to w_k .
- As before, we can transform it in an optimization problem obtaining the **y-aware loss**:

$$\arg \min_f \sum_k \max(0, \frac{w_k}{\sum_t w_t} \{s_t - s_k\}_{\substack{t=1, \dots, N \\ t \neq k}}) \approx \mathcal{L}^{y\text{-aware}} = - \sum_k \frac{w_k}{\sum_t w_t} \log \left(\frac{\exp(s_k)}{\sum_{t=1}^N \exp(s_t)} \right)$$

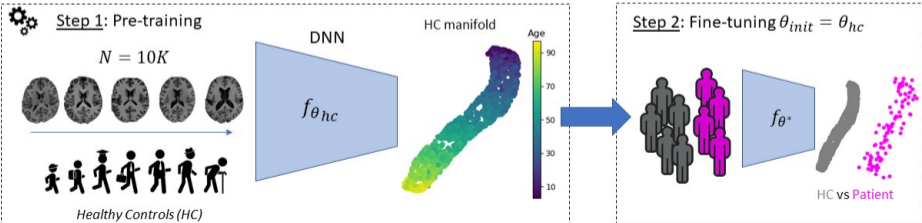
¹²³B. Dufumier et al. “Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification”. In: *MICCAI. 2021*.

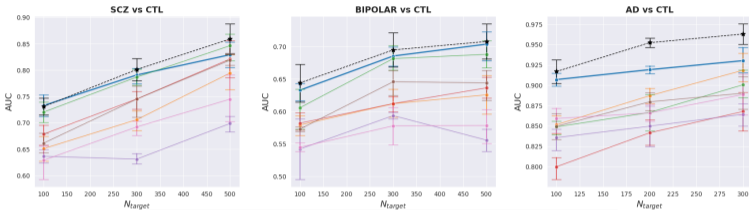
¹²⁴B. Dufumier et al. “Conditional Alignment and Uniformity for Contrastive Learning...”. In: *NeurIPS Workshop. 2021*.

Old Paradigm: supervised learning from scratch

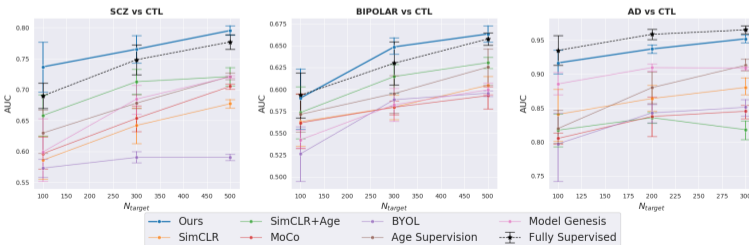


Our New Paradigm



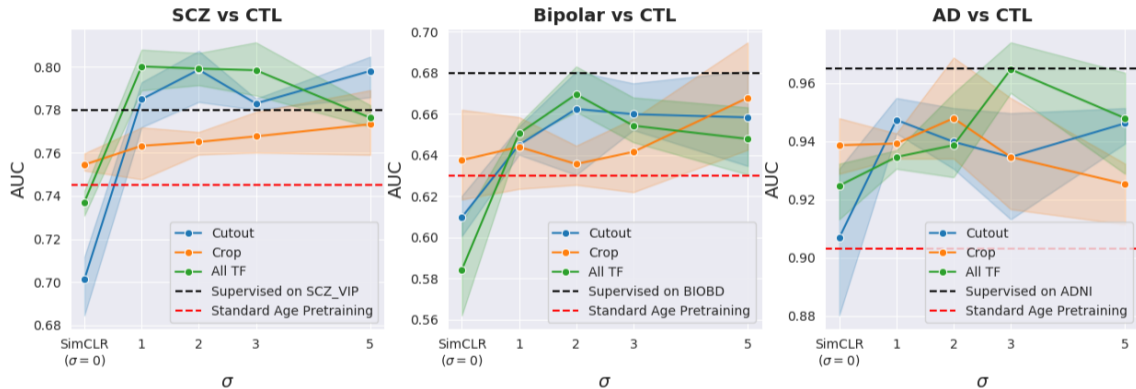


(a) 5-fold CV Stratified on Site.



(b) 5-fold CV Leave-Site-Out

Age-Aware Contrastive Learning Performance vs σ



- ▶ Linear classification performance remains stable for a range $\sigma \in [1, 5]$
- ▶ Adding more transformations improve the representation (in line with SimCLR)
- ▶ Cutout remains competitive while being cost-less computationally

Task	Test Set	Pre-training Strategies					
		Baseline	Weakly Self-Supervised	Self-Supervised		Generative	Discriminative
			Age-Aware Contrastive ¹²⁵	Model Genesis ¹²⁶	Contrastive Learning ¹²⁷	VAE	Age Sup.
SCZ vs. HC \uparrow $N_{train} = 933$	Internal Test	85.27 \pm 1.60	85.17 \pm 0.37	76.31 \pm 1.77	82.31 \pm 2.03	82.56 \pm 0.68	83.05 \pm 1.36
	External Test	75.52 \pm 0.12	77.00 \pm 0.55	67.40 \pm 1.59	75.48 \pm 2.54	75.11 \pm 1.65	74.36 \pm 2.28
BD vs. HC \uparrow $N_{train} = 832$	Internal Test	76.49 \pm 2.16	78.81 \pm 2.48	76.25 \pm 1.48	72.71 \pm 2.06	71.61 \pm 0.81	77.21 \pm 1.00
	External Test	68.57 \pm 4.72	77.06 \pm 1.90	65.66 \pm 0.90	71.23 \pm 3.05	71.70 \pm 0.23	73.02 \pm 2.66
ASD vs. HC \uparrow $N_{train} = 1526$	Internal Test	65.74 \pm 1.47	66.36 \pm 1.14	63.58 \pm 4.35	61.92 \pm 1.67	59.67 \pm 2.04	67.11 \pm 1.76
	External Test	62.93 \pm 2.40	68.76 \pm 1.70	54.95 \pm 3.58	61.93 \pm 1.93	57.45 \pm 0.81	62.07 \pm 2.98

Table: Fine-tuning results.¹²⁸ All pre-trained models use a data-set of **8754 3D MRI of healthy brains**. We reported average AUC(%) for all models and the standard deviation by repeating each experiment three times. Baseline is a DenseNet121 backbone.

¹²⁵B. Dufumier et al. "Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification". In: *MICCAI. 2021*.

¹²⁶Z. Zhou et al. "Models Genesis". In: *Media (2021)*.

¹²⁷T. Chen et al. "A Simple Framework for Contrastive Learning of Visual Representations". In: *ICML. 2020*.

¹²⁸B. Dufumier et al. "Deep Learning Improvement over Standard Machine Learning in Neuroimaging". In: *NeuroImage (under review) ()*.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- We could use $\mathcal{L}^{y\text{-aware}}$ also in regression. But...

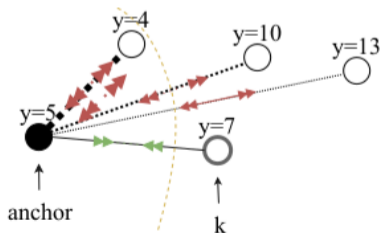
$$\mathcal{L}^{y\text{-aware}} = - \sum_k \frac{w_k}{\sum_t w_t} \log \left(\frac{\exp(s_k)}{\sum_{t=1}^N \exp(s_t)} \right)$$

- ... the numerator aligns x_k , and the denominator *focuses more* on the closest samples in the representation space.

- We could use $\mathcal{L}^{y\text{-aware}}$ also in regression. But...

$$\mathcal{L}^{y\text{-aware}} = - \sum_k \frac{w_k}{\sum_t w_t} \log \left(\frac{\exp(s_k)}{\sum_{t=1}^N \exp(s_t)} \right)$$

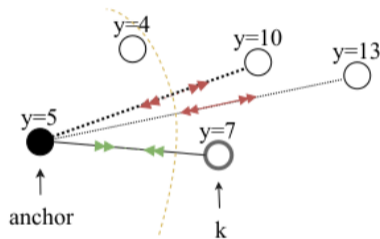
- ... the numerator aligns x_k , and the denominator *focuses more* on the closest samples in the representation space. → **Problem!** These samples might have a greater degree of positiveness with the anchor than the considered x_k



- We thus propose two new losses:

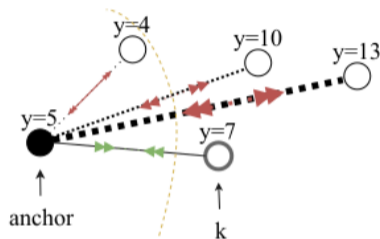
$$w_k(s_t - s_k) \leq 0 \quad \text{if } w_t - w_k \leq 0 \quad \forall k, t \neq k \in A(i)$$

$$\mathcal{L}^{thr} = - \sum_k \frac{w_k}{\sum_t \delta_{w_t < w_k} w_t} \log \left(\frac{\exp(s_k)}{\sum_{t \neq k} \delta_{w_t < w_k} \exp(s_t)} \right)$$



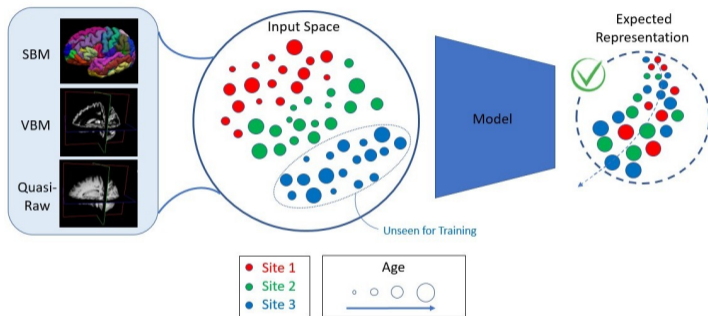
- \mathcal{L}^{thr} repels only the samples that have a y greater than the one of x_k but it still focuses more on the closest samples.

$$w_k[s_t(1 - w_t) - s_k] \leq 0 \quad \forall k, t \neq k \in A(i)$$
$$\mathcal{L}^{exp} = -\frac{1}{\sum_t w_t} \sum_k w_k \log \frac{\exp(s_k)}{\sum_{t \neq k} \exp(s_t(1 - w_t))}$$



- \mathcal{L}^{exp} has a repulsion strength inversely proportional to the similarity between y values, whatever their distance.
- Repulsion strength only depends on the distance in the kernel space. \rightarrow samples close in the kernel space will be close in the representation space.

OpenBHB Challenge: Representation Learning for Age Prediction with Site Effect Removal



- **OpenBHB Challenge:** age prediction with site-effect removal \rightarrow Brain age \neq chronological age in neurodegenerative disorders !
- N_{train} : 5330 3D brain MRI scans (different subjects) from 71 acquisition sites.
- Two private test data-sets (internal and external)
- To participate https://ramp.studio/problems/brain_age_with_site_removal108/144

Method	Int. MAE ↓	BAcc ↓	Ext. MAE ↓	\mathcal{L}_c ↓
$\mathcal{L}^{y\text{-aware}}$	<u>2.66</u> ±0.00	6.60±0.17	4.10±0.01	1.82
\mathcal{L}^{thr}	2.95±0.01	<u>5.73</u> ±0.15	4.10±0.01	1.74
\mathcal{L}^{exp}	2.55 ±0.00	5.1 ±0.1	3.76 ±0.01	1.54

Table: Comparison of contrastive losses.

Method	Model	Int. MAE ↓	BAcc ↓	Ext. MAE ↓	\mathcal{L}_c ↓
Baseline (ℓ_1)	DenseNet	2.55±0.01	8.0±0.9	7.13±0.05	3.34
	ResNet-18	2.67±0.05	6.7±0.1	4.18±0.01	1.86
	AlexNet	2.72±0.01	8.3±0.2	4.66±0.05	2.21
ComBat	DenseNet	5.92±0.01	2.23±0.06	10.48±0.17	3.38
	ResNet-18	4.15±0.01	4.5 ±0.0	4.76±0.03	1.88
	AlexNet	3.37±0.01	6.8±0.3	5.23±0.12	2.33
\mathcal{L}^{exp}	DenseNet	2.85±0.00	5.34±0.06	4.43±0.00	1.84
	ResNet-18	2.55 ±0.00	5.1±0.1	3.76 ±0.01	1.54
	AlexNet	2.77±0.01	5.8±0.1	4.01±0.01	1.71

Table: Final scores on the OpenBHB Challenge leaderboard using a 3D ResNet-18. **MAE**: Mean Absolute Error. **BAcc**: Balanced Accuracy for site prediction. **Challenge score**: $\mathcal{L}_c = \text{BAcc}^{0.3} \cdot \text{MAE}_{\text{ext}}$.

- Contrastive learning is more robust than traditional end-to-end approaches, such as cross-entropy, against noise in the data or in the labels¹²⁹.
- What about data bias, such as the site-effect ?

Method	Model	Int. MAE ↓	BAcc ↓	Ext. MAE ↓	\mathcal{L}_c ↓
Baseline (ℓ_1)	ResNet-18	2.67±0.05	6.7±0.1	4.18±0.01	1.86
ComBat	ResNet-18	4.15±0.01	4.5±0.0	4.76±0.03	1.88
\mathcal{L}^{exp}	ResNet-18	2.55±0.00	5.1±0.1	3.76±0.01	1.54

- \mathcal{L}^{exp} shows a small overfitting on internal sites but also a **low debiasing capability** towards site effect → BAcc should be equal to random chance: $1/n_{\text{sites}} = 1/64 \sim 1.56$
- Need to include debiasing regularization terms, such as FairKL¹³⁰. Please have a look !

¹²⁹F. Graf et al. “Dissecting Supervised Contrastive Learning”. In: *ICML*. 2021.

¹³⁰C. A. Barbano et al. “Unbiased Supervised Contrastive Learning”. In: *ICLR*. 2023.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- Self-distillation methods use a Siamese architecture with *two* different neural networks: online (student) and target (teacher).
- They are non-contrastive methods → → **only positives, no negatives are used !**
- They avoid collapse using *asymmetric architectures* and *different optimization procedures* (e.g., EMA, stop-gradients, diverse learning rates/weight decay) for the two networks^{131·132·133·134}

¹³¹C. Zhang et al. “How Does SimSiam Avoid Collapse Without Negative Samples?” In: *ICLR. 2022*.

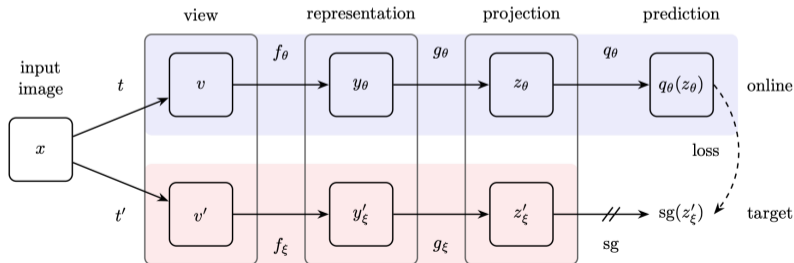
¹³²Q. Garrido et al. “On the duality between contrastive and non-contrastive self-supervised”. In: *ICLR. 2023*.

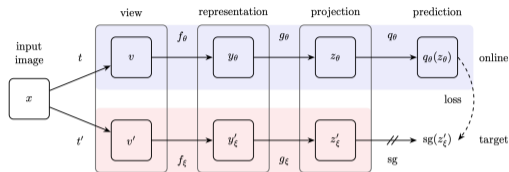
¹³³Y. Tian et al. “Understanding self-supervised learning dynamics without contrastive pairs”. In: *ICML. 2021*.

¹³⁴M. S. Halvagal et al. *Predictor networks and stop-grads provide implicit variance regularization in BYOL/SimSiam. 2022*.

Bootstrap Your Own Latent (BYOL)

- BYOL¹³⁵ is the first method that introduced self-distillation to avoid collapse of Siamese architecture (constant output).
- Online is composed of: encoder f_θ , projector g_θ and predictor q_θ .
- Target is composed of: encoder f_ξ and projector g_ξ . The parameters ξ are different from θ and they are updated using an exponential moving average (i.e., momentum encoder): $\xi \leftarrow \tau\xi + (1 - \tau)\theta$



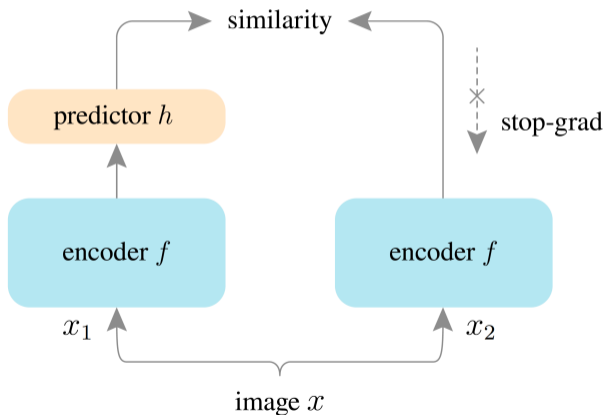


- The idea of BYOL is to maximize agreement between the outputs of the two networks that take as input augmentations of the **same** image
- Agreement is defined as the symmetric mean squared error between normalized outputs (v and v' are fed to both networks):

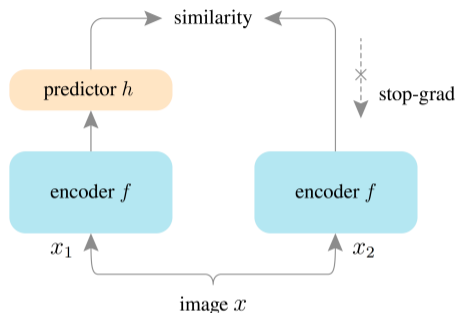
$$\begin{aligned} \mathcal{L}_{\theta, \xi} &= \frac{1}{2} \left\| \frac{q_\theta(z_\theta)}{\|q_\theta(z_\theta)\|_2} - \frac{z'_\xi}{\|z'_\xi\|_2} \right\|_2^2 + \frac{1}{2} \left\| \frac{q_\theta(z'_\theta)}{\|q_\theta(z'_\theta)\|_2} - \frac{z_\xi}{\|z_\xi\|_2} \right\|_2^2 \\ &\approx -\frac{\langle q_\theta(z_\theta), z'_\xi \rangle}{\|q_\theta(z_\theta)\|_2 \|z'_\xi\|_2} - \frac{\langle q_\theta(z'_\theta), z_\xi \rangle}{\|q_\theta(z'_\theta)\|_2 \|z_\xi\|_2} \end{aligned} \quad (17)$$

Simple Siamese networks (SimSiam)

- SimSiam¹³⁶ removes the EMA (momentum encoder) using simply a stop-grad and... it magically works without collapsing !

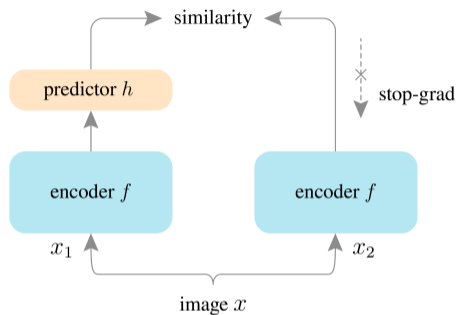


¹³⁶X. Chen et al. "Exploring Simple Siamese Representation Learning". In: *CVPR*. 2021.



- SimSiam uses an even simpler architecture with one *shared* encoder f and a predictor h only for the student network. Two views (i.e., augmentations) of the same image x are then matched minimizing the symmetric loss:

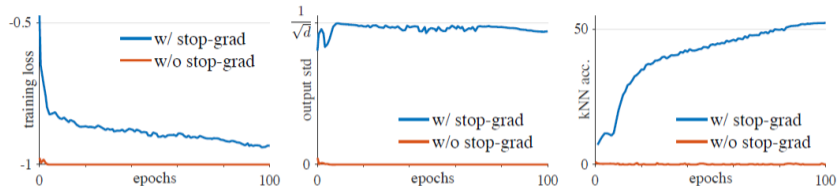
$$\mathcal{L}_{f,h}(x_1, x_2) = -\frac{\langle h(f(x_1)), f(x_2) \rangle}{\|h(f(x_1))\|_2 \|f(x_2)\|_2} - \frac{\langle h(f(x_2)), f(x_1) \rangle}{\|h(f(x_2))\|_2 \|f(x_1)\|_2} \quad (18)$$



- SimSiam does not collapse thanks to the stop-gradient: $f(x_2)$ and $f(x_1)$ are treated as constants in the loss and the encoder f receives no gradients from them

$$\mathcal{L}_{f,h}(x_1, x_2) = -\frac{\langle h(f(x_1)), f(x_2) \rangle}{\|h(f(x_1))\|_2 \|f(x_2)\|_2} - \frac{\langle h(f(x_2)), f(x_1) \rangle}{\|h(f(x_2))\|_2 \|f(x_1)\|_2} \quad (19)$$

Simple Siamese networks (SimSiam)



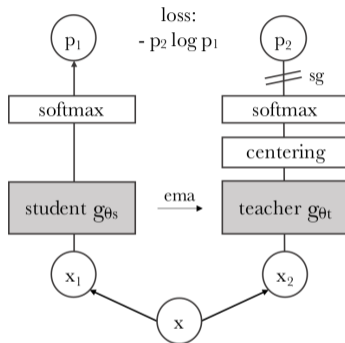
	acc. (%)
w/ stop-grad	67.7±0.1
w/o stop-grad	0.1

case	proj. MLP's BN		pred. MLP's BN		acc. (%)
	hidden	output	hidden	output	
(a) none	-	-	-	-	34.6
(b) hidden-only	✓	-	✓	-	67.4
(c) default	✓	✓	✓	-	68.1
(d) all	✓	✓	✓	✓	unstable

pred. MLP h		acc. (%)
baseline	lr with cosine decay	67.7
(a)	no pred. MLP	0.1
(b)	fixed random init.	1.5
(c)	lr not decayed	68.1

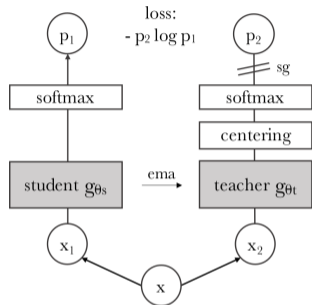
- The stop-gradient and the architecture design (presence of h , different learning rates for f and h , batch normalization, etc) are very important... but, they don't explain **How does SimSiam avoid collapse without negative samples?** → still open question¹³⁷

¹³⁷C. Zhang et al. "How Does SimSiam Avoid Collapse Without Negative Samples?" In: *ICLR*. 2022.



- In¹³⁸ authors propose DINO which differs from BYOL and SimSiam: 1) same architecture for teacher and student, 2) knowledge distillation, 3) centering and sharpening the teacher output, 4) use of ViT, 5) multi-crops

¹³⁸M. Caron et al. "Emerging Properties in Self-Supervised Vision Transformers". In: *ICCV*. 2021.



- Teacher and student use the same architecture: $g = f \circ h$ where f is an encoder (ViT or ResNet) and h is a MLP projection head. **No predictor is used for the student!**
- When using ViT no Batch Norms, Group Normalization, Weight decay/standardization are used (thus less hyper-parameters to tune) \rightarrow important for BYOL¹³⁹

¹³⁹P. H. Richemond et al. *BYOL works even without batch statistics*. 2020.

- Generalized knowledge distillation¹⁴⁰ (KD) matches the soft labels of student and teacher instead than hard labels. Soft labels allow to uncover relations between classes that would be difficult to detect with hard labels.
- In DINO the vector of K features z is transformed into soft-labels (each dimension i becomes a probability) using the softmax function:

$$\begin{aligned} P_s(x)^i &= \frac{\exp(g_{\theta_s}(x)^i/\tau_s)}{\sum_{k=1}^K \exp(g_{\theta_s}(x)^k/\tau_s)} = \frac{\exp(z_s^i/\tau_s)}{\sum_{k=1}^K \exp(z_s^k/\tau_s)} \\ P_t(x)^i &= \frac{\exp(g_{\theta_t}(x)^i/\tau_t)}{\sum_{k=1}^K \exp(g_{\theta_t}(x)^k/\tau_t)} = \frac{\exp(z_t^i/\tau_t)}{\sum_{k=1}^K \exp(z_t^k/\tau_t)} \end{aligned} \quad (20)$$

- where $z = g(x)^i$ indicates the i -th dimension of z , and τ controls the sharpness of the output distribution (the smaller, the sharper)

¹⁴⁰D. Lopez-Paz et al. "Unifying distillation and privileged information". In: *ICLR*. 2016.

- Given the parameters of the two networks (θ_t and θ_s), we compute the symmetric loss (as in BYOL and SimSiam):

$$\mathcal{L}_{\theta_s, \theta_t}(x_1, x_2) = \frac{1}{2}H(P_s(x_1), P_t(x_2)) + \frac{1}{2}H(P_s(x_2), P_t(x_1)) \quad (21)$$

- where $H(a, b) = -a \log(b)$ is the cross-entropy.
- As in BYOL, the parameters of the teacher are not updated via back-propagation (i.e., stop-gradient) but using an EMA:

$$\theta_t \leftarrow \lambda \theta_t + (1 - \lambda) \theta_s \quad (22)$$

- Authors use a **multi-crop strategy**: x_1 is a *global* view while x_2 is a *local* view with smaller resolution

- DINO can have two forms of collapse:
 - ▶ the model output is uniform along all the dimensions ($z^i = z^j \forall i, j$)
 - ▶ the model output is dominated by one dimension ($z^i \gg z^j \forall j \neq i$)

- DINO can have two forms of collapse:
 - ▶ the model output is uniform along all the dimensions ($z^i = z^j \forall i, j$) \rightarrow **sharpening** P_t using low τ_t
 - ▶ the model output is dominated by one dimension ($z^i \gg z^j \forall j \neq i$) \rightarrow **centering** P_t
- However, sharpening and centering avoids one collapse but encourage the other, **they need to be used together**
- Centering can be seen as adding a bias term to the teacher: $z_t \leftarrow z_t + c$, and it is updated with an EMA:

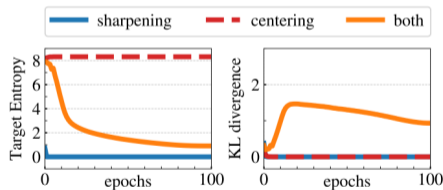
$$c \leftarrow mc + (1 - m) \frac{1}{B} \sum_{i=1}^B g_{\theta_t}(x_i) \quad (23)$$

- where $m > 0$ is a rate hyper-parameter and B is the batch size.

	Method	Loss	multi-crop	Center.	BN	Pred.	Top-1
1	DINO	CE	✓	✓			76.1
2	-	MSE	✓	✓			62.4
3	-	CE	✓	✓		✓	75.6
4	-	CE		✓			72.5
5	MoCov2	INCE			✓		71.4
6	-	INCE	✓		✓		73.4
7	BYOL	MSE			✓	✓	71.4
8	-	MSE			✓		0.1
9	-	MSE		✓			52.6
10	-	MSE	✓		✓	✓	64.8

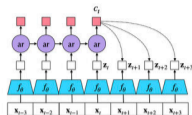
	Method	Mom.	SK	MC	Loss	Pred.	k-NN	Lin.
1	DINO	✓	✗	✓	CE	✗	72.8	76.1
2	-	✗	✗	✓	CE	✗	0.1	0.1
3	-	✓	✓	✓	CE	✗	72.2	76.0
4	-	✓	✗	✗	CE	✗	67.9	72.5
5	-	✓	✗	✓	MSE	✗	52.6	62.4
6	-	✓	✗	✓	CE	✓	71.8	75.6
7	BYOL	✓	✗	✗	MSE	✓	66.6	71.4
8	MoCov2	✓	✗	✗	INCE	✗	62.0	71.6
9	SwAV	✗	✓	✓	CE	✗	64.7	71.8

SK: Sinkhorn-Knopp, MC: Multi-Crop, Pred.: Predictor
CE: Cross-Entropy, MSE: Mean Square Error, INCE: InfoNCE

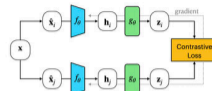


- **Take home message:** Non-contrastive teacher-student methods do not need negatives and avoid collapse using architectural/optimization solutions. → still not clear what's the best strategy and why it works !

(a) CPC



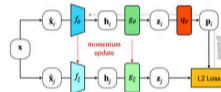
(b) SimCLR



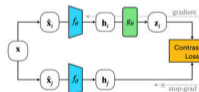
(c) MoCo V2



(d) BYOL



(e) SimSiam



(f) SwAV

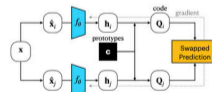


Image from¹⁴¹

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

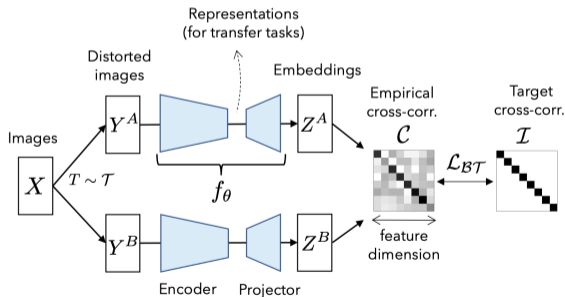
4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

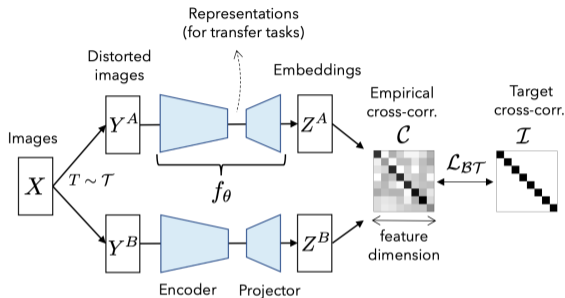
4.2 Information Maximization

5. Conclusions

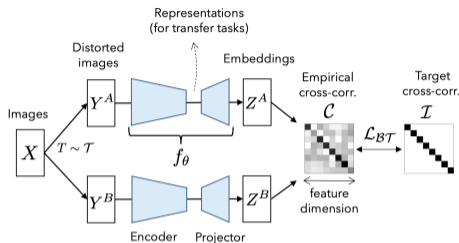
- Another category of non-contrastive methods maximizes the information content of the embeddings by reducing the redundancy
- Barlow Twins¹⁴² makes the normalized cross-correlation matrix computed from twin embeddings as close to the identity matrix as possible



¹⁴²J. Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction". In: *ICML*. 2021.



- Barlow Twins does not use: 1) large batch, 2) asymmetric networks (no predictor), 3) EMA, 4) clustering/prototypes, 5) stop-gradient → **Easier to optimize!**
- Two views, y^A and y^B , of the same image x are fed to the same network f producing $z^A = f(y^A)$ and $z^B = f(y^B)$.
- The batch matrix Z with all z (views of multiple images) is then batch-normalized along each dimension i ($\hat{Z}_i = \frac{Z_i - \mu_i}{\sigma_i}$).



- We first define the normalized cross-correlation matrix C between the random vectors $\mathbf{z}^A = [\mathbf{z}_1^A, \mathbf{z}_2^A, \dots, \mathbf{z}_D^A]^T = [D, 1]$ and $\mathbf{z}^B = [\mathbf{z}_1^B, \mathbf{z}_2^B, \dots, \mathbf{z}_D^B]^T = [D, 1]$, where D is the dimension of the embeddings (i.e., number of elements):

$$C = [D, D] \triangleq \mathbb{E}[\mathbf{z}^A (\mathbf{z}^B)^T] = \begin{bmatrix} \mathbb{E}[\mathbf{z}_1^A \mathbf{z}_1^B] & \mathbb{E}[\mathbf{z}_1^A \mathbf{z}_2^B] & \dots & \mathbb{E}[\mathbf{z}_1^A \mathbf{z}_D^B] \\ \dots & \dots & \dots & \dots \\ \mathbb{E}[\mathbf{z}_D^A \mathbf{z}_1^B] & \mathbb{E}[\mathbf{z}_D^A \mathbf{z}_2^B] & \dots & \mathbb{E}[\mathbf{z}_D^A \mathbf{z}_D^B] \end{bmatrix} \quad (24)$$

- We can approximate the \mathbb{E} with an average on the (normalized) batch samples $\hat{Z}^A = [N, D]$ and $\hat{Z}^B = [N, D]$, where N is the batch size, and define $C = ((\hat{Z}^A)^T \hat{Z}^B) / N$.
- Each elements of C has values comprised between -1 (anti-correlation) and 1 (correlation) and is defined:

$$C_{ij} = \frac{\sum_{t=1}^N z_{ti}^A z_{tj}^B}{\sqrt{\sum_{t=1}^N (z_{ti}^A)^2} \sqrt{\sum_{t=1}^N (z_{tj}^B)^2}} \quad (25)$$

- where t is a batch index, i and j refers to the dimension of the embeddings (i.e., $1 \leq i, j \leq D$)

- Intuitively, since the two networks A and B take as input two views of the same image, we would like the embeddings z^A and z^B to be **invariant** to the augmentations and thus that homologous elements of z^A and z^B have **similar values** and **same sign**
- At the same time, we would like that our representations z^A and z^B do not contain **redundant information**, which means that different elements of z^A and z^B should have different and unrelated values
- The loss of Barlow Twins is thus:

$$\mathcal{L}_{BT} = \underbrace{\sum_i (1 - C_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} C_{ij}^2}_{\text{redundancy reduction term}} \quad (26)$$

- where $\lambda > 0$ is a hyper-parameter to tune

- Intuitively, since the two networks A and B take as input two views of the same image, we would like the embeddings z^A and z^B to be **invariant** to the augmentations and thus that homologous elements of z^A and z^B have **similar values** and **same sign**

→ $C_{ij} = 1 \forall i$ **perfect correlation**

- At the same time, we would like that our representations z^A and z^B do not contain **redundant information**, which means that different elements of z^A and z^B should have different and unrelated values

→ $C_{ij} = 0 \forall i, j$ **perfect de-correlation**

- The loss of Barlow Twins is thus:

$$\mathcal{L}_{BT} = \underbrace{\sum_i (1 - C_{ii})^2}_{\text{invariance term}} + \lambda \underbrace{\sum_i \sum_{j \neq i} C_{ij}^2}_{\text{redundancy reduction term}} \quad (26)$$

- where $\lambda > 0$ is a hyper-parameter to tune

- VICReg¹⁴³ is similar to Barlow Twins but instead than using a normalized cross-correlation, they use a loss with three different terms
 - ▶ **Mean squared distance** between the embedding vectors z_t^A and z_t^B → invariance to the augmentations
 - ▶ **Variance regularization term**: the standard deviation (over a batch) of each dimension of the embedding is forced to be above a given threshold. → it prevents a collapse (possible in BT) where all dimensions have the same value
 - ▶ **Covariance term**: it attracts the off-diagonal coefficients of the covariance matrices of Z^A and Z^B to be close to 0. → it reduces redundancy, as in BT, without using cross-correlation $C = ((\hat{Z}^A)^T \hat{Z}^B) / N$ but single covariance matrices $C^A = ((\hat{Z}^A)^T \hat{Z}^A) / N$, $C^B = ((\hat{Z}^B)^T \hat{Z}^B) / N$
- Results between BT and VicReg are very similar...

¹⁴³A. Bardes et al. "VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning". In: *ICLR. 2022*.

1. Introduction

1.1 Transfer Learning

2. Self-supervised Learning

2.1 Context prediction

2.2 Generative models

2.3 Instance discrimination

3. Contrastive Learning

3.1 A geometric approach

3.2 ϵ -margin metric learning

3.3 Efficient implementations

3.4 Weakly supervised

3.5 Regression

4. Non-contrastive learning

4.1 Teacher/Student or Self-distillation

4.2 Information Maximization

5. Conclusions

- Different methods of self-supervised learning exist. Recently contrastive and non-contrastive methods have emerged obtaining SOTA results and, in some cases, on-pair or very similar to fully supervised methods
- Architecture (e.g., siamese networks, CNN/ResNet/ViT, BN, temperature τ , etc.) and optimization (e.g., losses, lr, etc.) choices are important. But probably the most important factor in unsupervised SSL is the transformations/augmentations¹⁴⁴ → it depends on the downstream task (e.g., classification, recognition, segmentation) and data (e.g., medical, multimodal)
- Recent non-contrastive methods (BYOL, SwAV) do not need negatives (small batch, smaller pre-training datasets, less memory and less computing power) **BUT**, they can be harder to train (e.g., SimSiam) and their underlying mechanisms are still poorly understood → geometric approach as in¹⁴⁵ or¹⁴⁶ ?

¹⁴⁴I. Bendidi et al. *No Free Lunch in Self Supervised Representation Learning*. 2023.

¹⁴⁵C. Zhang et al. "How Does SimSiam Avoid Collapse Without Negative Samples?" In: *ICLR*. 2022.

¹⁴⁶Q. Garrido et al. "On the duality between contrastive and non-contrastive self-supervised". In: *ICLR*. 2023.



Florence Carton



Matthis Maillard



Rebeca Vetil



Emma Sarfati



Camille Ruppli



Giammarco La Barbera



Ali Mammadov



Robin Louiset

Institutions & Partners



UNIVERSITÀ
DI TORINO



- Asano, Y. M. et al. “Self-labelling via simultaneous clustering and representation learning”. In: *ICLR*. 2020.
- Bao, H. et al. “BEiT: BERT Pre-Training of Image Transformers”. In: *ICLR*. 2022.
- Barbano, C. A. et al. “Contrastive learning for regression in multi-site brain age prediction”. In: *IEEE ISBI*. 2023.
- Barbano, C. A. et al. “Unbiased Supervised Contrastive Learning”. In: *ICLR*. 2023.
- Bardes, A. et al. “VICReg: Variance-Invariance-Covariance Regularization for Self-Supervised Learning”. In: *ICLR*. 2022.
- Becker, S. et al. “Self-organizing neural network that discovers surfaces in random ...”. In: *Nature* (1992).
- Bendidi, I. et al. *No Free Lunch in Self Supervised Representation Learning*. 2023.
- Brock, A. et al. “Large Scale GAN Training for High Fidelity Natural Image Synthesis”. In: *ICLR*. 2019.
- Bromley, J. et al. “Signature Verification using a “Siamese” Time Delay Neural Network”. In: *NIPS*. Vol. 6. 1993.
- Caron, M. et al. “Deep Clustering for Unsupervised Learning of Visual Features”. In: *ECCV*. 2018.
- Caron, M. et al. “Emerging Properties in Self-Supervised Vision Transformers”. In: *ICCV*. 2021.
- Caron, M. et al. “Unsupervised Learning of Visual Features by Contrasting Cluster Assignments”. In: *NeurIPS*. 2020.
- Chen, M. et al. “Generative Pretraining From Pixels”. In: *ICML*. 2020.
- Chen, T. et al. “A Simple Framework for Contrastive Learning of Visual Representations”. In: *ICML*. 2020.
- Chen, X. et al. “Exploring Simple Siamese Representation Learning”. In: *CVPR*. 2021.

- Chopra, S. et al. “Learning a Similarity Metric Discriminatively, with Application to Face Verification”. In: *CVPR*. 2005.
- Chuang, C.-Y. et al. “Debiased Contrastive Learning”. In: *NeurIPS*. 2020.
- Del Pup, F. et al. *Applications of Self-Supervised Learning to Biomedical Signals: where are we now*. 2023.
- Deng, J. et al. “ImageNet: A Large-Scale Hierarchical Image Database”. In: *CVPR*. 2009.
- Devlin, J. et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*.
- Doersch, C. et al. “Unsupervised Visual Representation Learning by Context Prediction”. In: *ICCV*. 2015.
- Donahue, J. et al. “Adversarial Feature Learning”. In: *ICLR*. 2017.
- Donahue, J. et al. “Large Scale Adversarial Representation Learning”. In: *NeurIPS*. 2019.
- Dosovitskiy, A. et al. “Discriminative Unsupervised Feature Learning with Exemplar CNNs”. In: *IEEE TPAMI* (2016).
- Dufumier, B. et al. “Conditional Alignment and Uniformity for Contrastive Learning...”. In: *NeurIPS Workshop*. 2021.
- Dufumier, B. et al. “Contrastive Learning with Continuous Proxy Meta-data for 3D MRI Classification”. In: *MICCAI*. 2021.
- Dufumier, B. et al. “Deep Learning Improvement over Standard Machine Learning in Neuroimaging”. In: *NeuroImage (under review)* ().

- Dufumier, B. et al. “OpenBHB: a Large-Scale Multi-Site Brain MRI Data-set for Age Prediction and Debiasing”. In: *NeuroImage* (2022).
- Dumoulin, V. et al. “Adversarially Learned Inference”. In: *ICLR*. 2017.
- Dwibedi, D. et al. “With a Little Help from My Friends: Nearest-Neighbor Contrastive Learning”. In: *ICCV*. 2021.
- Frosst, N. et al. “Analyzing and Improving Representations with the Soft Nearest Neighbor”. In: *ICML*. 2019.
- Garrido, Q. et al. “On the duality between contrastive and non-contrastive self-supervised”. In: *ICLR*. 2023.
- Gidaris, S. et al. “Unsupervised Representation Learning by Predicting Image Rotations”. In: *ICLR*. 2018.
- Goodfellow, I. et al. “Generative Adversarial Nets”. In: *NIPS*. 2014.
- Graf, F. et al. “Dissecting Supervised Contrastive Learning”. In: *ICML*. 2021.
- Grill, J.-B. et al. “Bootstrap your own latent: A new approach to self-supervised Learning”. In: *NeurIPS*. 2020.
- Halvagal, M. S. et al. *Predictor networks and stop-grads provide implicit variance regularization in BYOL/SimSiam*. 2022.
- He, K. et al. “Masked Autoencoders Are Scalable Vision Learners”. In: *CVPR*. 2022.
- He, K. et al. “Momentum Contrast for Unsupervised Visual Representation Learning”. In: *CVPR*. 2020.
- Henaff, O. et al. “Data-Efficient Image Recognition with Contrastive Predictive Coding”. In: *ICML*. 2020.
- Hinton, G. E. et al. “Reducing the Dimensionality of Data with Neural Networks”. In: *Science* (2006).
- Hjelm, R. D. et al. “Learning deep representations by mutual information estimation ...”. In: *ICLR*. 2019.

- Ho, C.-H. et al. “Contrastive Learning with Adversarial Examples”. In: *NeurIPS*. 2020.
- Ho, J. et al. “Denoising Diffusion Probabilistic Models”. In: *NeurIPS*. 2020.
- Jang, T. et al. “Difficulty-Based Sampling for Debaised Contrastive Representation Learning”. In: *CVPR*. 2023.
- Khosla, P. et al. “Supervised Contrastive Learning”. In: *NeurIPS*. 2020.
- Kingma, D. P. et al. “Auto-Encoding Variational Bayes”. In: *ICLR*. 2014.
- Larsson, G. et al. “Colorization as a Proxy Task for Visual Understanding”. In: *CVPR*. 2017.
- . “Learning Representations for Automatic Colorization”. In: *ECCV*. 2016.
- Li, J. et al. “Prototypical Contrastive Learning of Unsupervised Representations”. In: *ICLR*. 2021.
- Lin, T.-Y. et al. “Microsoft COCO: Common Objects in Context”. In: *ECCV*. 2014.
- Littlejohns, T. J. et al. “The UK Biobank imaging enhancement of 100,000 participants:” in: *Nature Communications* (2020).
- Lopez-Paz, D. et al. “Unifying distillation and privileged information”. In: *ICLR*. 2016.
- Matsoukas, C. et al. “What Makes Transfer Learning Work for Medical Images”. In: *CVPR*. 2022.
- Misra, I. et al. “Self-Supervised Learning of Pretext-Invariant Representations”. In: *CVPR*. 2020.
- Mustafa, B. et al. *Supervised Transfer Learning at Scale for Medical Imaging*. 2021.
- Neyshabur, B. et al. “What is being transferred in transfer learning?” In: *NeurIPS*. 2020.
- Noroozi, M. et al. “Representation Learning by Learning to Count”. In: *ICCV*. 2017.

- Noroozi, M. et al. “Unsupervised Learning of Visual Representations by Solving Jigsaw Puzzles”. In: *ECCV*. 2016.
- Oord, A. v. d. et al. *Representation Learning with Contrastive Predictive Coding*. 2018.
- Pathak, D. et al. “Context Encoders: Feature Learning by Inpainting”. In: *CVPR*. 2016.
- Poole, B. et al. “On Variational Bounds of Mutual Information”. In: *ICML*. 2019.
- Radford, A. et al. *Language Models are Unsupervised Multitask Learners*.
- Radford, A. et al. “Unsupervised Representation Learning with Deep Convolutional GAN”. In: *ICLR*. 2016.
- Raghu, M. et al. “Transfusion: Understanding Transfer Learning for Medical Imaging”. In: *NeurIPS*. 2019.
- Rezende, D. J. et al. “Variational Inference with Normalizing Flows”. In: *ICML*. 2015.
- Richemond, P. H. et al. *BYOL works even without batch statistics*. 2020.
- Robinson, J. et al. “Contrastive Learning with Hard Negative Samples”. In: *ICLR*. 2021.
- Salakhutdinov, R. et al. “Learning a Nonlinear Embedding by Preserving Class ...”. In: *AISTATS*. 2007.
- Schroff, F. et al. “FaceNet: A Unified Embedding for Face Recognition and Clustering”. In: *CVPR*. 2015.
- Sohn, K. “Improved Deep Metric Learning with Multi-class N-pair Loss Objective”. In: *NIPS*. 2016.
- Song, H. O. et al. “Deep Metric Learning via Lifted Structured Feature Embedding”. In: *CVPR*. 2016.
- Thalles Santos, S. *A Short Introduction to Generative Adversarial Networks*. URL:
<https://sthalles.github.io>.
- Tian, Y. et al. “Understanding self-supervised learning dynamics without contrastive pairs”. In: *ICML*. 2021.

- Tschannen, M. et al. “On Mutual Information Maximization for Representation Learning”. In: *ICLR*. 2020.
- Vincent, P. et al. “Extracting and composing robust features with denoising autoencoders”. In: *ICML*. 2008.
- Vincent, P. et al. “Stacked Denoising Autoencoders: Learning Useful Representations ...”. In: *JMLR* (2010).
- Wang, F. et al. “Understanding the Behaviour of Contrastive Loss”. In: *CVPR*. 2021.
- Wang, T. et al. “Understanding Contrastive Representation Learning through Alignment and Uniformity on the Hypersphere”. In: *ICML*. 2020.
- Wu, C.-Y. et al. “Sampling Matters in Deep Embedding Learning”. In: *ICCV*. 2017.
- Wu, Z. et al. “Unsupervised Feature Learning via Non-parametric Instance Discrimination”. In: *CVPR*. 2018.
- Xie, Z. et al. “SimMIM: a Simple Framework for Masked Image Modeling”. In: *CVPR*. 2022.
- Yeh, C.-H. et al. “Decoupled Contrastive Learning”. In: *ECCV*. 2022.
- Yu, B. et al. “Deep Metric Learning With Triplet Margin Loss”. In: *ICCV*. 2019.
- Zbontar, J. et al. “Barlow Twins: Self-Supervised Learning via Redundancy Reduction”. In: *ICML*. 2021.
- Zhang, C. et al. “How Does SimSiam Avoid Collapse Without Negative Samples?” In: *ICLR*. 2022.
- Zhang, R. et al. “Colorful Image Colorization”. In: *ECCV*. 2016.
- . “Split-Brain Autoencoders: Unsupervised Learning by Cross-Channel Prediction”. In: *CVPR*. 2017.
- Zhao, C. et al. *Multi-crop Contrastive Learning for Unsupervised Image-to-Image Translation*. 2023.
- Zhou, J. et al. “Image BERT Pre-training with Online Tokenizer”. In: *ICLR*. 2022.

Zhou, Z. et al. “Models Genesis”. In: *Media* (2021).