Apprentissage continu de représentations visuelles

ENSIMAG 2023-2024

NAVER LABS Europe

Inria

Grenoble INP ENSIMAG

Karteek Alahari & Diane Larlus

MIAI Grenoble Alpes
Multidisciplinary Institute In Artificial Intelligence

jeudi 23 novembre 2023

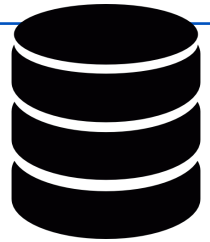
Organisation du cours

19/10/2023	11h15	12h45	ALAHARI Karteek	Intro + cours
26/10/2023	09h45	12h45	LARLUS Diane	Cours
09/11/2023	11h15	12h45	LARLUS Diane	**Articles 1 & 2**
16/11/2023	11h15	12h45	ALAHARI Karteek	Cours + **Article 3**
23/11/2023	11h15	12h45	LARLUS Diane	Cours
07/12/2023	11h15	12h45	LARLUS Diane	Cours
14/12/2023	09h45	12h45	LARLUS Diane + ALAHARI Karteek	**Articles 4, 5, 6** + Cours
21/12/2023	11h15	12h45	ALAHARI Karteek	Cours + **Article 7**
11/01/2024	11h15	12h45	ALAHARI Karteek	Cours + **Article 8**
18/01/2024	11h15	12h45	ALAHARI Karteek	Cours + Révisions

Apprentissage auto-supervisé: rappels

Apprentissage continu de représentations visuelles

2023-2024

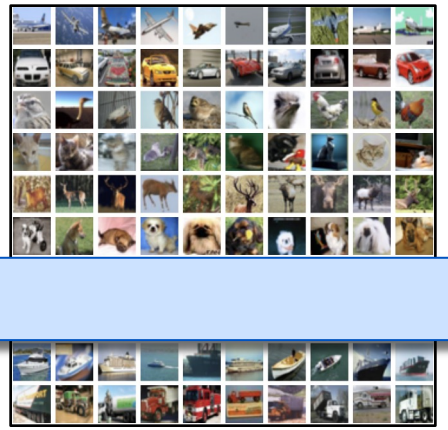


Fully-Supervised Classification

Images + labels

Proxy task

IMAGENET



Reducing annotation cost

Model

Visual representations

Target tasks

Image Classification

Object Detection

Instance Segmentation

Image Retrieval

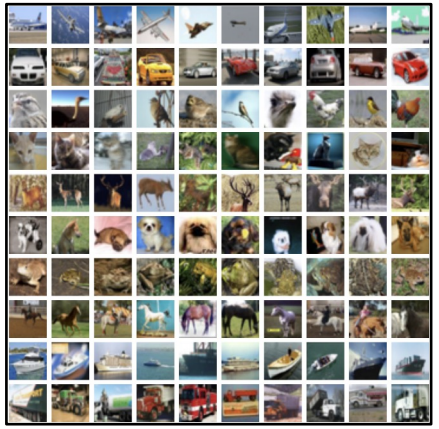
Learning transferable visual representations

No supervision

Reducing annotation cost

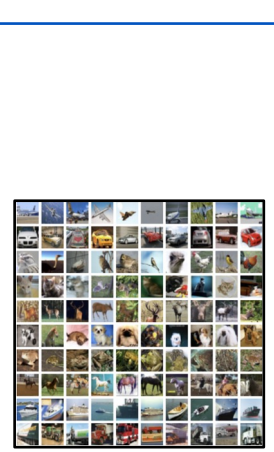
Fully-Supervised
fine-grained annotations
expert knowledge

Self-supervised
annotation-free images
no annotation required



~~labels~~

Apprentissage auto-supervisé: une première étape vers une tache plus concrête



IMAGENET

Images

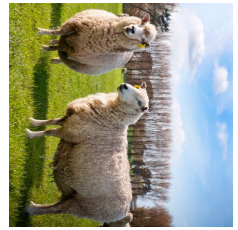
labels

Proxy task
= Rotation prediction

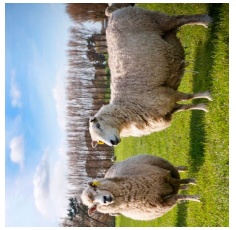
Encoder

Linear classifier

Class 1



Class 2



Class 3

Class 4

Target task = leaf classification

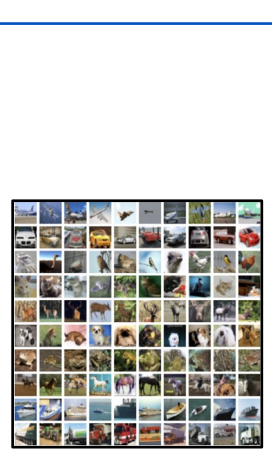
labels

Model

Linear classifier

Leaf class 1
Leaf class 2
…

Apprentissage auto-supervisé: une première étape vers une tache plus concrête



IMAGENET

Images

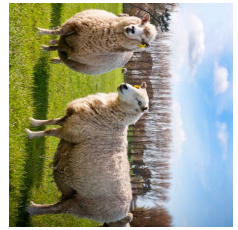
labels

Proxy task
= Rotation prediction

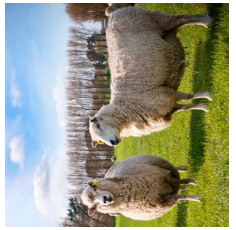
Encoder

Linear classifier

Class 1



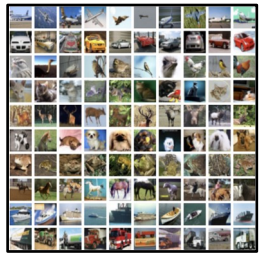
Class 2



Class 3

Class 4

Target task = ImageNet classification



IMAGENET

labels

Model

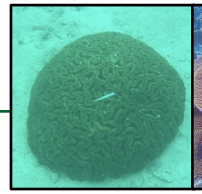
Linear classifier

Eurasian finch

Ruddy turnstone

Dowitcher

Brain coral



Coral reef

Coral fungus

....

Tour d’horizon des méthodes d’apprentissage auto-supervisées

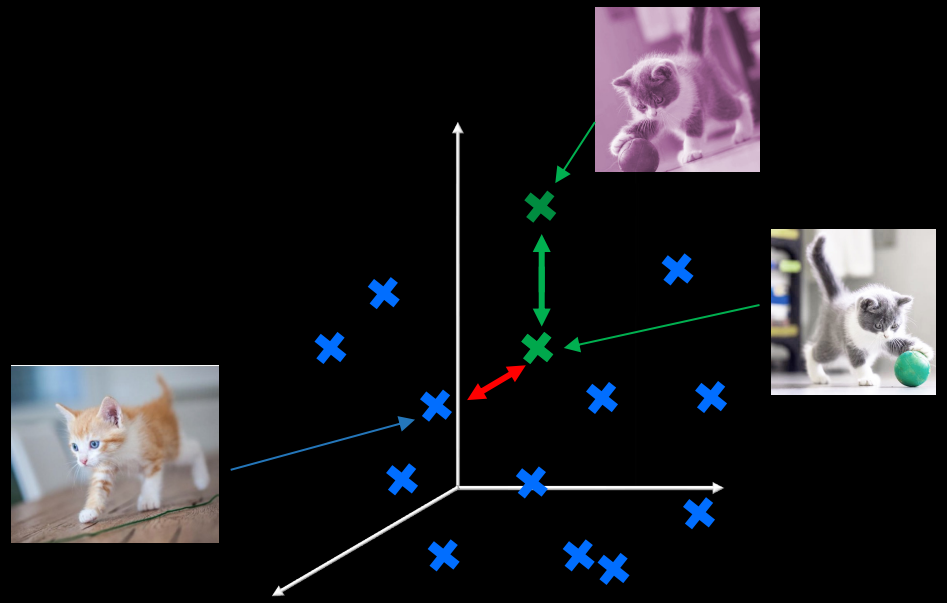
Apprentissage continu de représentations visuelles

2023-2024

TWO MAIN TRENDS

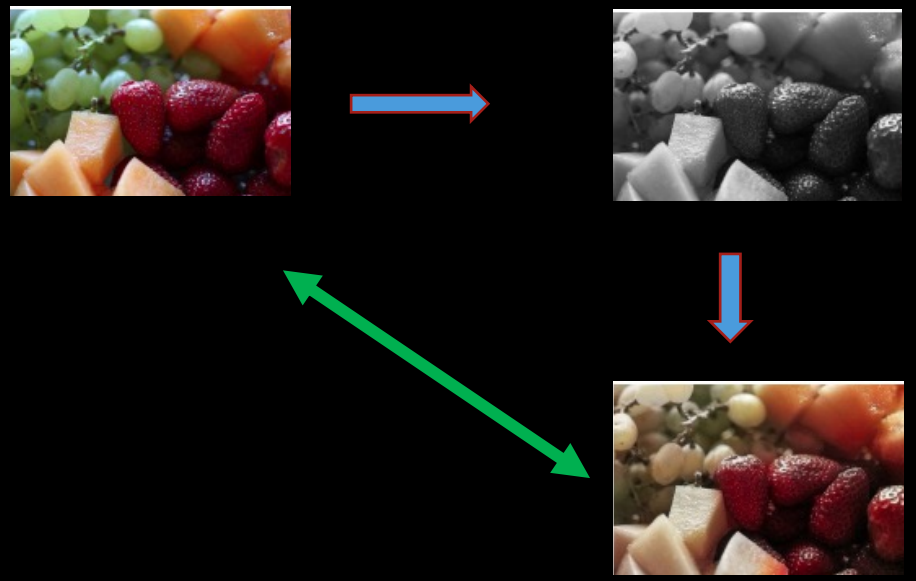
Discriminative SSL

- learn to solve a pretext task (jigsaw puzzle, rotation) or learn instance discrimination (each image is a class)



Generative SSL

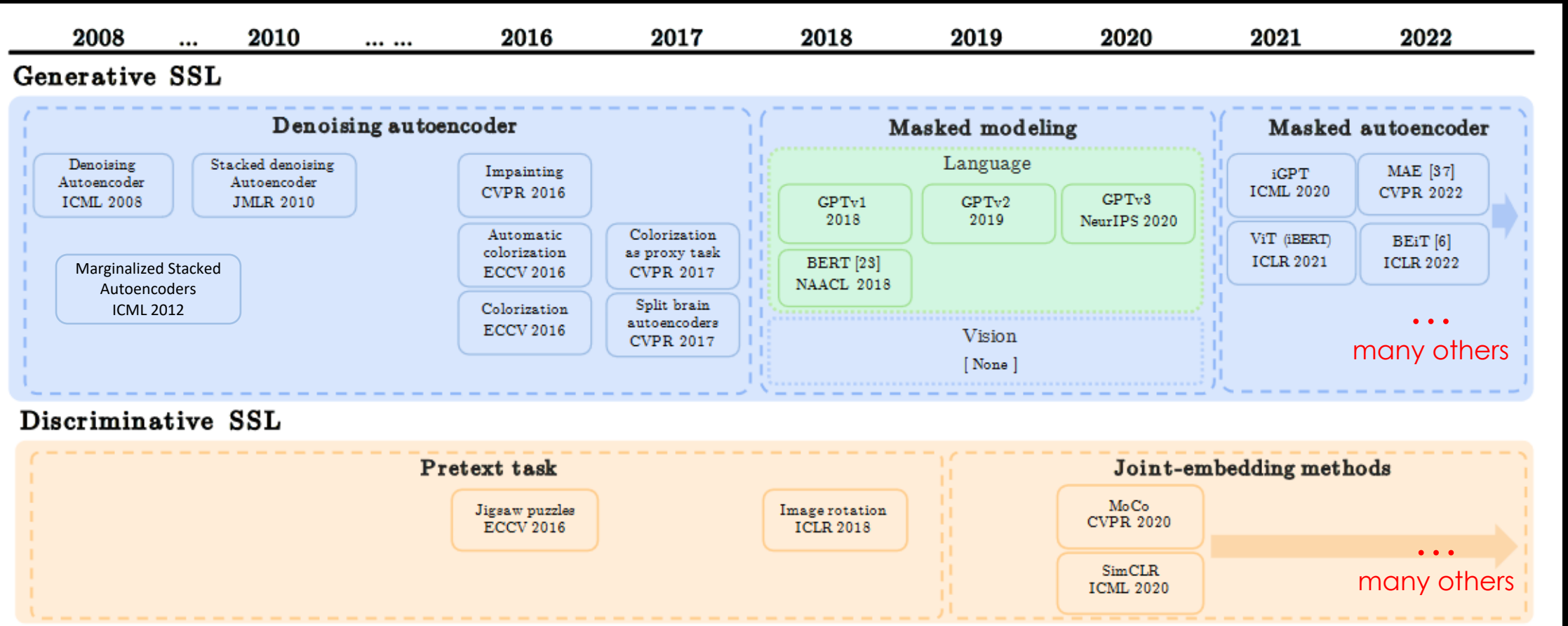
- modify/mask the image and reconstruct the content in general using autoencoders and reconstruction loss



[Slide: Gabriela Csurka]

TIMELINE OF VISUAL SSL

Première "classification" des méthodes SSL



Zhang et al, A Survey on Masked Autoencoder for Self-supervised Learning in Vision and Beyond, ArXiv'22

[Slide: Gabriela Csurka]

Self-supervised learning (or SSL)

Deuxième ”classification” des méthodes SSL

- Train on a **proxy task** (self-supervised)
  - Not (necessarily) an “important” task we care about
  - A task that is defined from the input data alone
  - Should enable us to learn aspects of the visual input/world
  - SSL methods can be divided into **Predictive** or **Contrastive** proxy tasks

[slide borrowed from Yannis Kalantidis]

Les méthodes génératives

- Les **méthodes génératives** pour l'apprentissage de représentations visuelles se focalisent généralement sur l'erreur de reconstruction dans l'espace des pixels.

- Elles utilisent souvent des fonctions de coût appliquées aux pixels (*pixel losses*) et donc ont tendance à se focaliser sur des détails au niveau des pixels plutôt que sur les concepts plus abstraits présents dans les données.

- Ces fonctions de coût basées sur les pixels supposent que les pixels sont indépendants entre eux. Elles ont donc ont des difficultés à capturer des corrélations entre pixels, ainsi que les concepts abstraits.

C'est pourquoi pendant un temps, les méthodes discriminatives, et en particulier les **méthodes contrastives** ont été privilégiées.

Apprentissage auto-supervisé – Méthodes Contrastives

Apprentissage continu de représentations visuelles

2023 -2024

Contrastive Learning

Intuition: Given a set of “similar” and “dissimilar” inputs, learn the ***ranking*** of distances, i.e., learn representations such that

the similarity between “similar” inputs is higher than the similarity between “dissimilar” inputs.

This similarity can be:

Euclidean distance

$$d(x_i, x_j) = \|x_i - x_j\|^2$$
$$= \sum_{k=1}^{d} (x_i^k - x_j^k)^2$$

Cosine similarity

$$cos(x_i, x_j) = \frac{x_i^T x_j}{\|x_i\| \, \|x_j\|}$$

[slide borrowed from Yannis Kalantidis]

Self-supervised learning – Contrastive learning – Overview

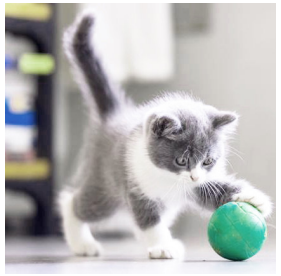
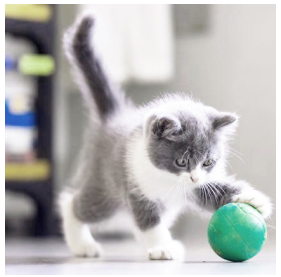


Image Transformations

Self-supervised learning – Contrastive learning – Overview



Model

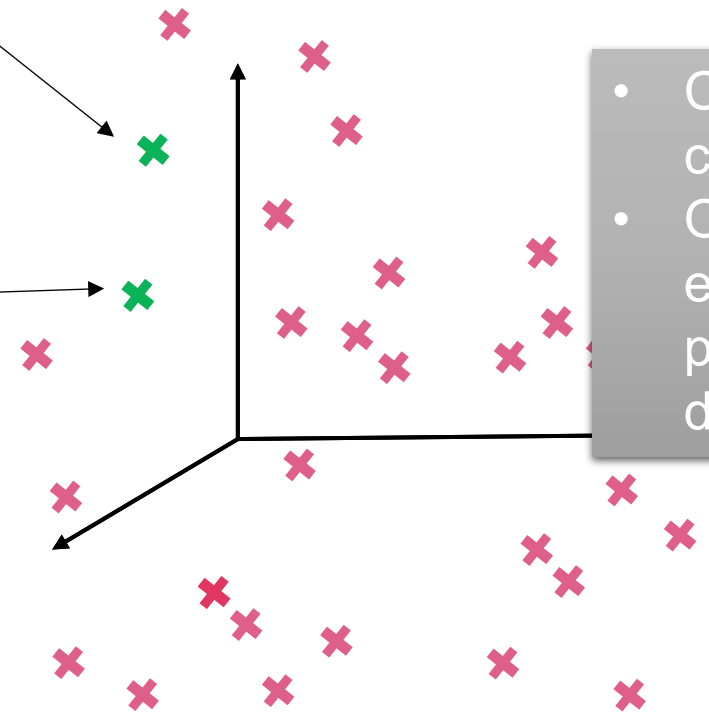
$\mathbf{q}$

Model

$\mathbf{k}$

Model

$\mathbf{n}$

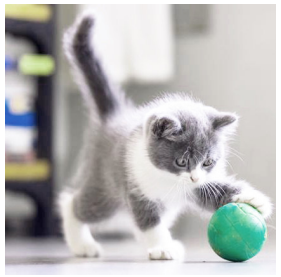


Pretext task = Instance discrimination

- Chaque image est traitée comme sa propre classe
- On regarde les similarités entre images dans le plongement appris = l'espace des représentations d'images

Self-supervised learning – Contrastive learning – Overview

Pretext task = Instance discrimination



Model

$\mathbf{q}$

anchor

Model

$\mathbf{k}$

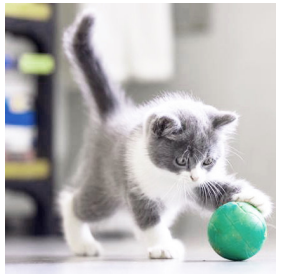
positive sample

Model

$\mathbf{n}$

negative sample

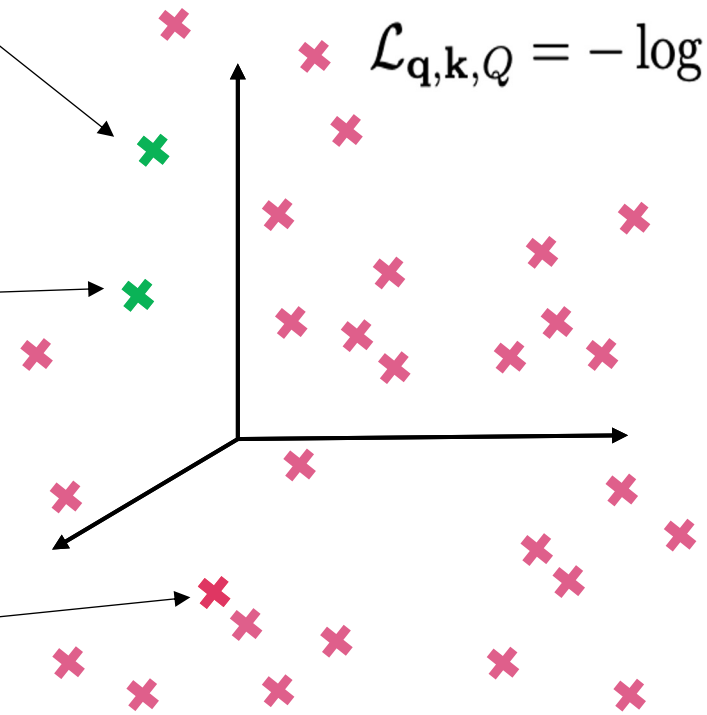
Self-supervised learning – Contrastive learning



Model → $\mathbf{q}$

Model → $\mathbf{k}$

Model → $\mathbf{n}$



$$\mathcal{L}_{\mathbf{q},\mathbf{k},Q} = -\log \frac{\exp(\mathbf{q}^T\mathbf{k}/\tau)}{\exp(\mathbf{q}^T\mathbf{k}/\tau) + \sum_{\mathbf{n}\in Q}\exp(\mathbf{q}^T\mathbf{n}/\tau)}$$

anchor

positive sample

InfoNCE loss

negative sample

La plupart des méthodes contrastives suivent ce schéma

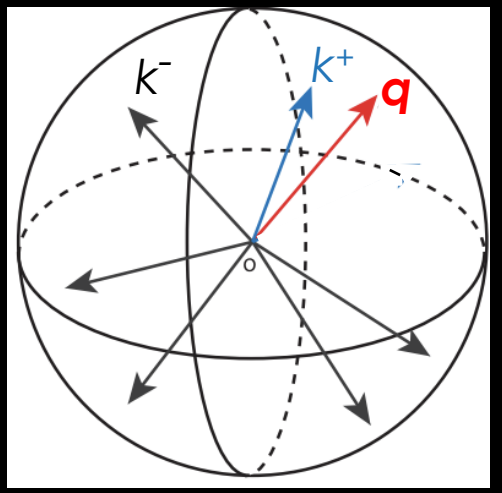
INSTANCE DISCRIMINATION (InstDisc)

[Slide: Gabriela Csurka]

Learn a metric space where

- the distance between two modified instances of a positive samples (q,k^+) is reduced
- the distance with negative samples (k^-) are enlarged

$$\mathcal{L}_{q,k^+,\{k^-\}} = -\log \frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{k^-} \exp(q \cdot k^-/\tau)}.$$



Using the contrastive InfoNCE loss:

- modified from cross entropy loss (normalized temperature-scaled)
- τ is a hyper-parameter (temperature) controlling the concentration level of the distribution

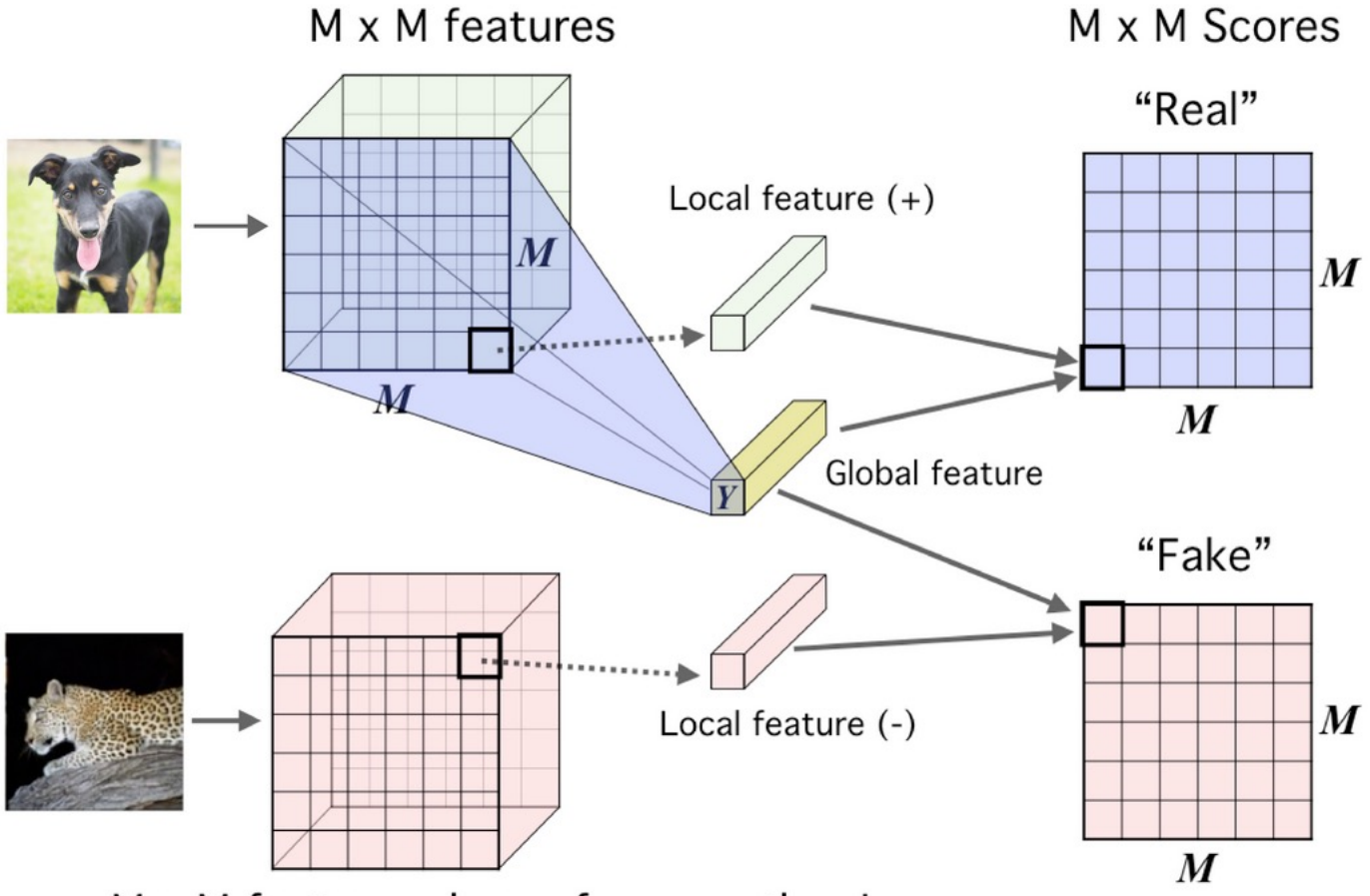
Wu +@CVPR'18

Deep InfoMax

- Contrastive task: classify whether a pair of global and local features are extracted from the same image or not.
- Global features: final output of the CNN
- Local features: output of an intermediate layer
- Loss

$$\mathcal{L}_N = -\mathbb{E}_X \left[\log \frac{\exp\left(f(x)^T f\left(x^+\right)\right)}{\exp\left(f(x)^T f\left(x^+\right)\right) + \sum_{j=1}^{N-1} \exp\left(f(x)^T f\left(x_j\right)\right)} \right]$$

$f(x)$ = global feature
$f(x+)$ = local feature from the same image (positive sample)
$f(x-)$ = local feature from a different image (negative sample)

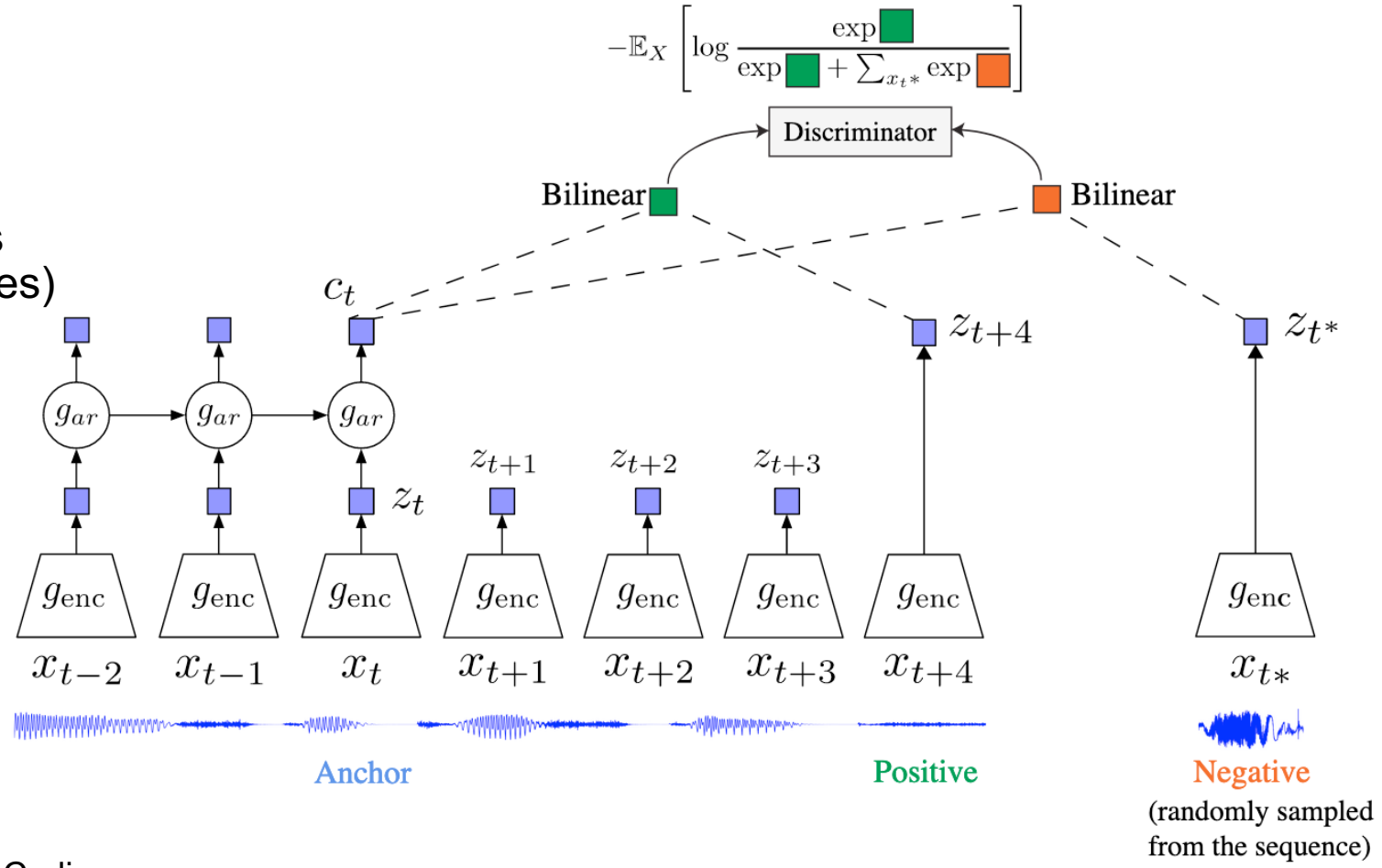


Learning deep representations by mutual information estimation and maximization. ICLR 2019
R Devon Hjelm, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, Yoshua Bengio

Contrastive Predictive Coding

Contrastive Predictive Coding (CPC)

- Contrastive method
- Applicable to any ordered sequence: text, speech, videos, and even images (seen as sequences of pixels or patches)



Representation Learning with Contrastive Predictive Coding
Aaron van den Oord, Yazhe Li, Oriol Vinyals.
Arxiv18

Importance des échantillons négatifs

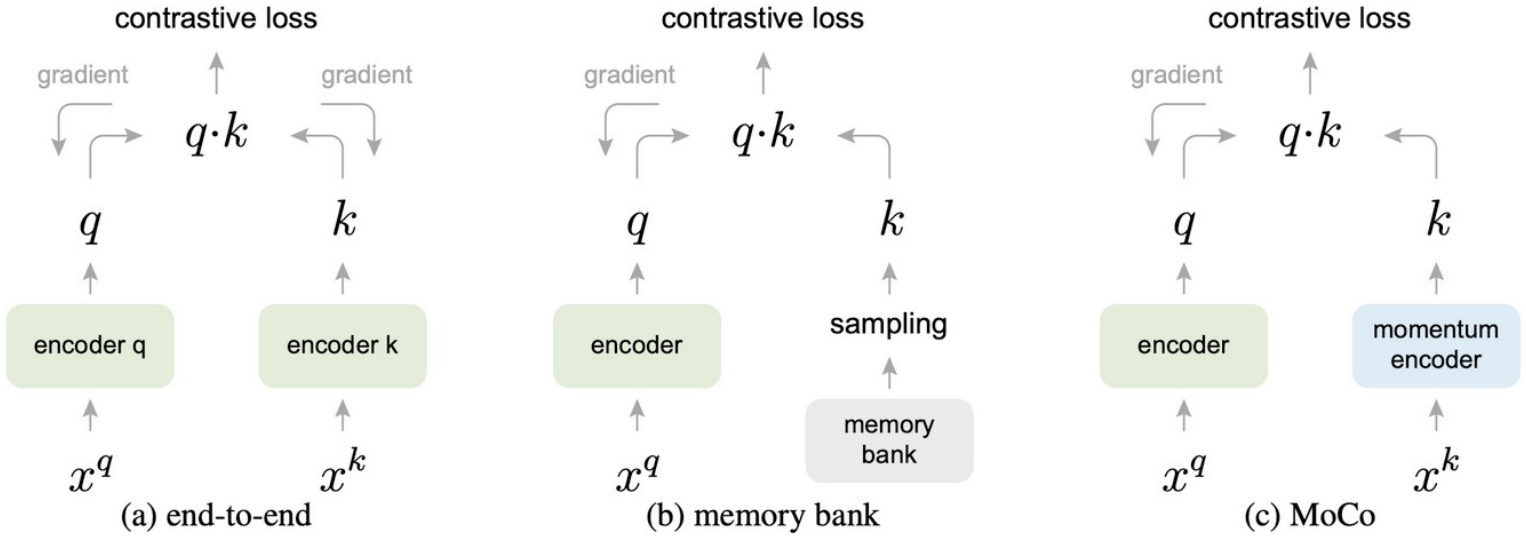
- Les méthodes contrastives donnent de très bons résultats, mais le choix des échantillons négatifs utilisés dans l'apprentissage est crucial.

Première observation: les méthodes contrastives donnent de bien meilleurs résultats quand elles ont accès à un grand nombre d'échantillons négatifs:

- Si les échantillons négatifs sont plus nombreux, ils couvrent mieux la distribution sous-jacente des données, et donc ils proposent un signal d'apprentissage plus fiable.
- Cela permet aussi une meilleure prévention des solutions triviales.

C'est une des intuitions portées par MoCo et SimCLR (présentés ensuite).

MoCo



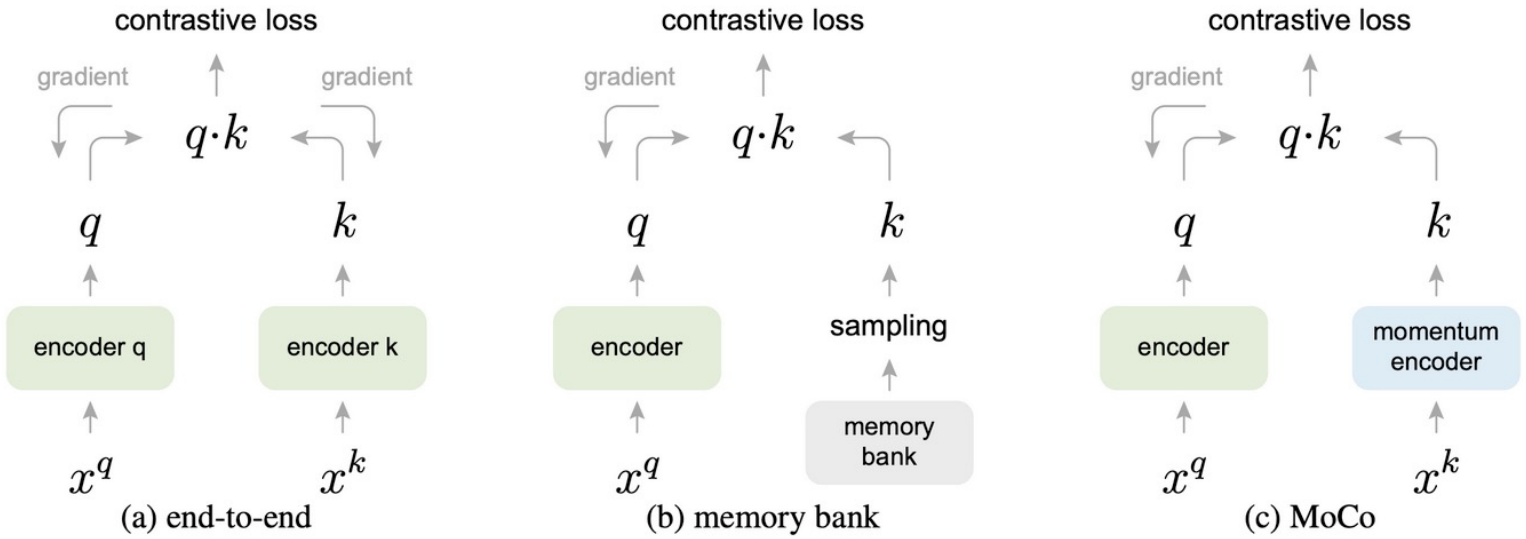
Classiquement, lors d'un apprentissage contrastif de représentations, les gradients circulent à travers les encodeurs pour les échantillons positifs mais aussi pour les échantillons négatifs. Cela signifie que le nombre des négatifs est restreint à la taille du *batch* utilisé pour l'apprentissage.

MoCo contourne cette limitation en maintenant une grande quantité d'échantillons négatifs en mémoire, et en utilisant la rétro-propagation du gradient seulement pour l'encodeur des échantillons positifs.

Un problème survient avec l'utilisation naïve d'une mémoire d'échantillons négatifs: au bout d'un moment les descripteurs stockés ne correspondent plus au modèle courant.

Pour résoudre ce problème, il y a une mise à jour périodique de l'encodeur pour les échantillons négatifs.

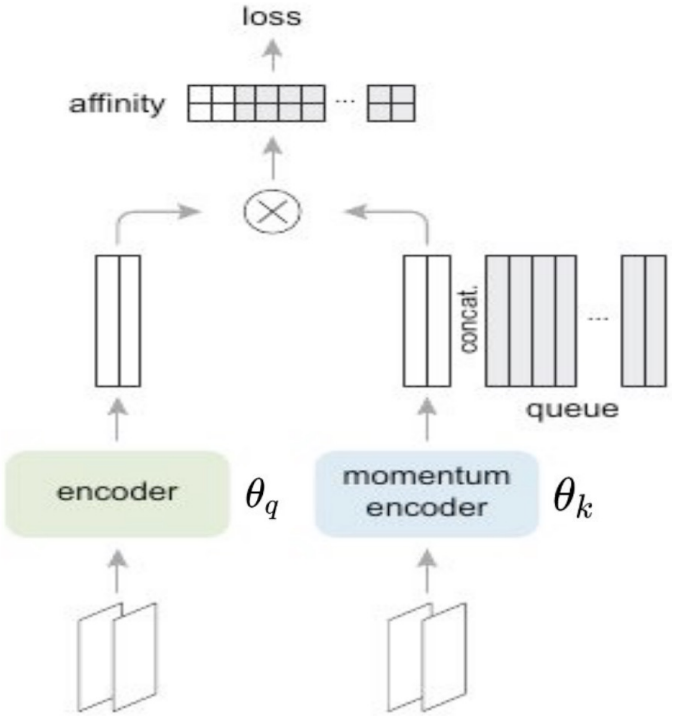
MoCo



- Encodes the keys on-the-fly
- Maintains the queue of keys
- Key encoder update:

$$\theta_k := m \cdot \theta_k + (1 - m) \cdot \theta_q$$

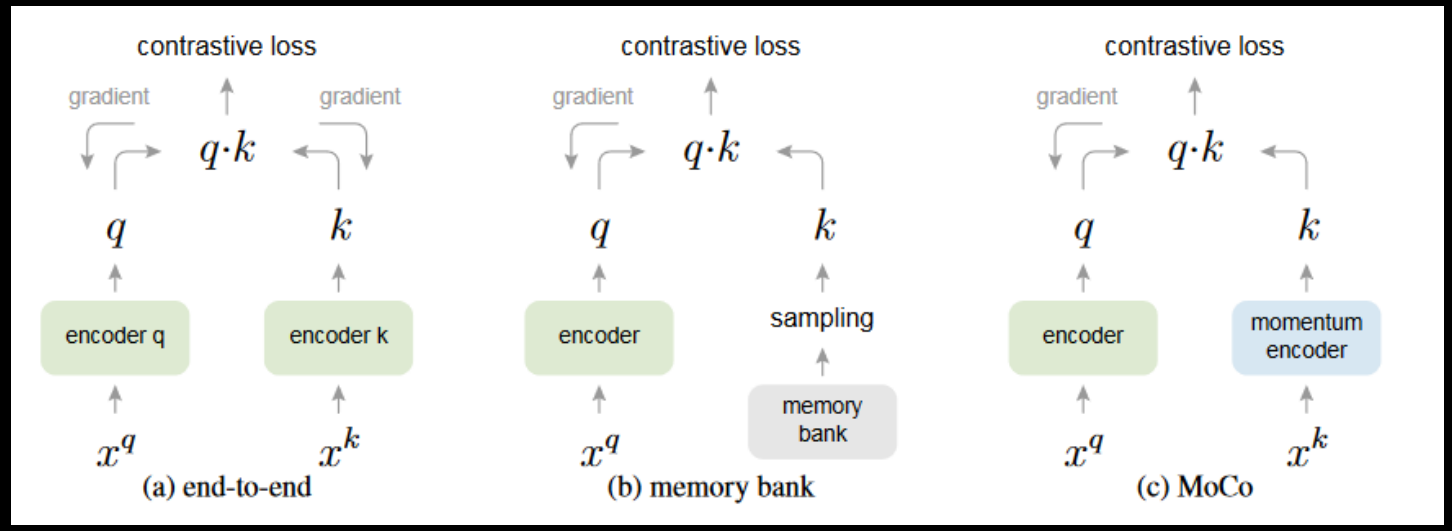
momentum m	0	0.9	0.99	0.999	0.9999
accuracy (%)	*fail*	55.2	57.8	59.0	58.9

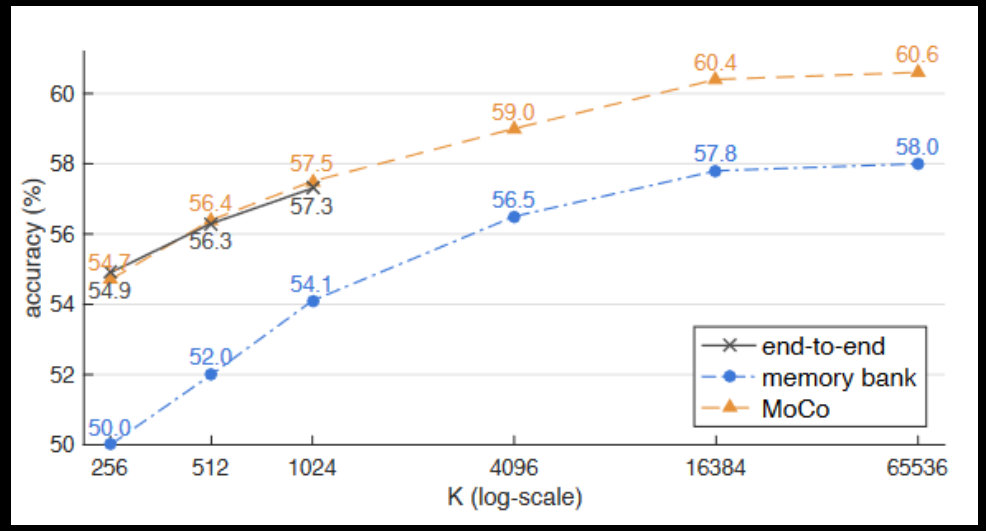


MOMENTUM CONTRASTIVE LEARNING (MoCo)

Keep and encodes keys on-the-fly by a momentum-updated encoder

- maintains a queue that acts as a dictionary (and negatives)





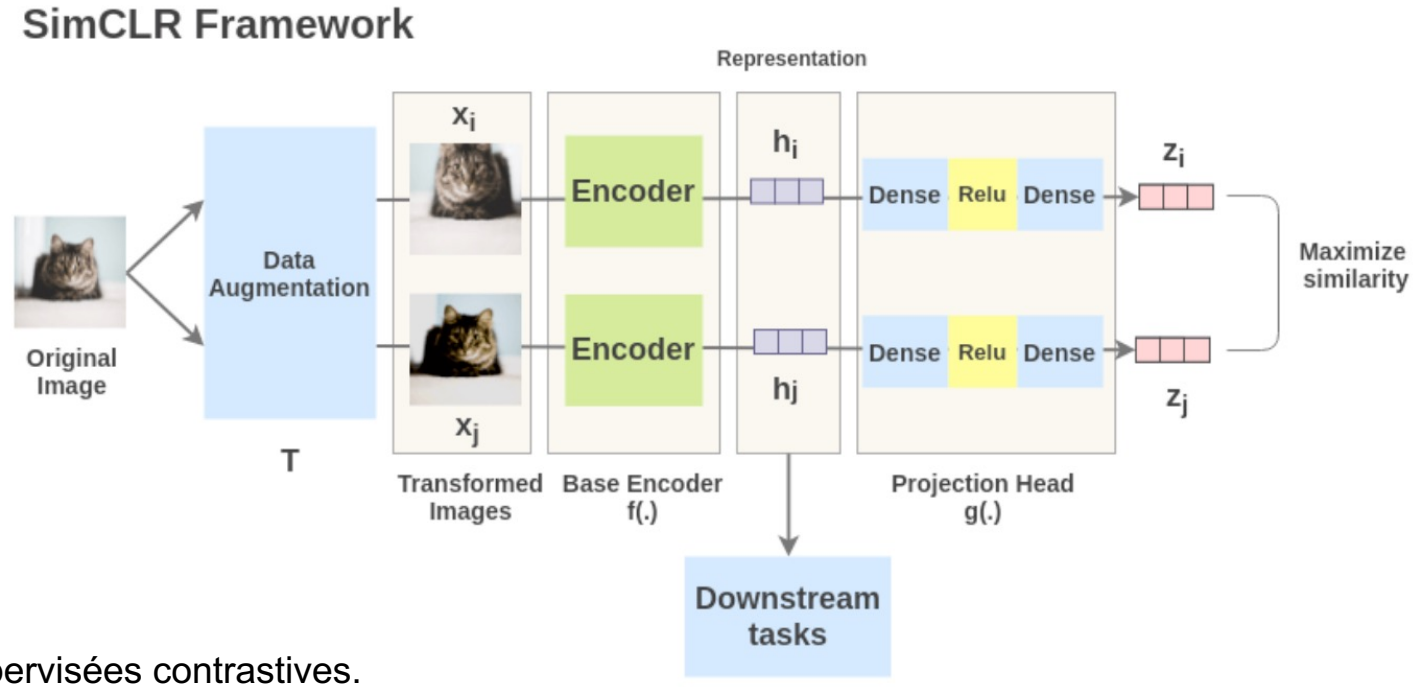
He +@CVPR'20 (MoCo), Chen+@ArXiv'20 (MoCov2), Chen+@ICCV'21 (MoCov3)

[Slide: Gabriela Csurka]

SimCLR

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

(1) Normalized Temperature-scaled Cross-Entropy loss (NT-Xnet)



Cet article propose une large étude des méthodes auto-supervisées contrastives.
Il en résulte un certain nombre de recommandations pour des résultats optimaux :

- Augmentation de données: plus de transformations d'images et des variations plus extrêmes donnent de meilleurs résultats.
- Ajout d'un projecteur "jetable" à la fin de l'apprentissage pour éviter le sur-apprentissage sur la tâche prétexte.
  Cela améliore grandement le transfert es modèles vers d'autres tâches
- Une nouvelle fonction de coût: *Normalized Temperature-scaled Cross-Entropy loss*

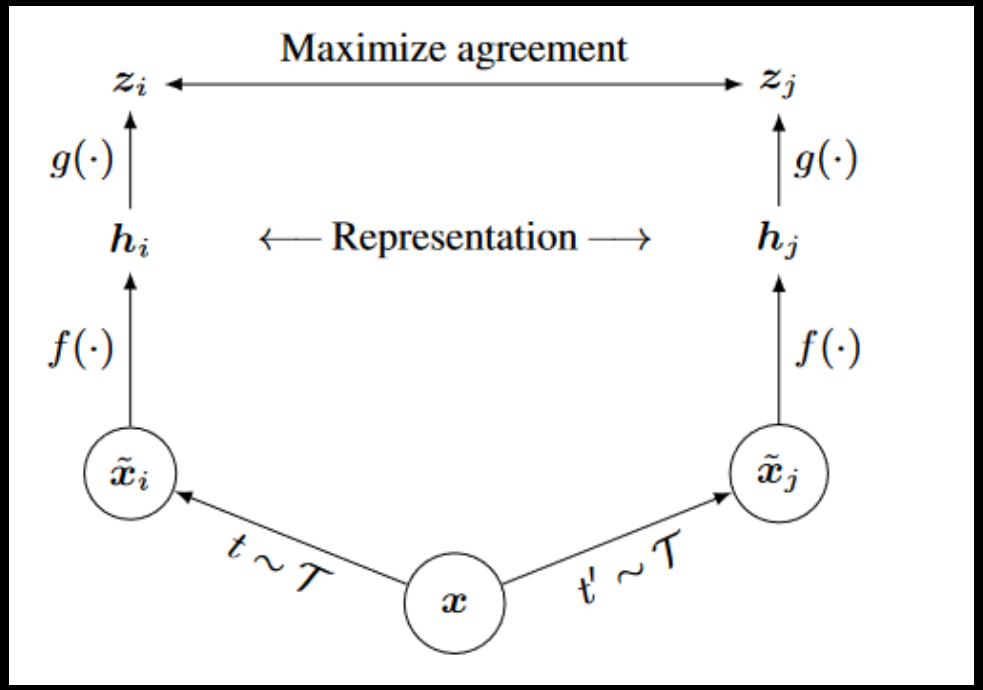
Observation majeure: un plus grand batch (passer de 256 à 8192 images), un réseau plus profond, et un apprentissage plus long (800 *epochs*) ont un très fort impact sur les résultats. Plus que d'autres choix.

Cette méthode obtient de très bons résultats au prix de ressources de calcul que peu d'institutions peuvent se permettre (ici Google)

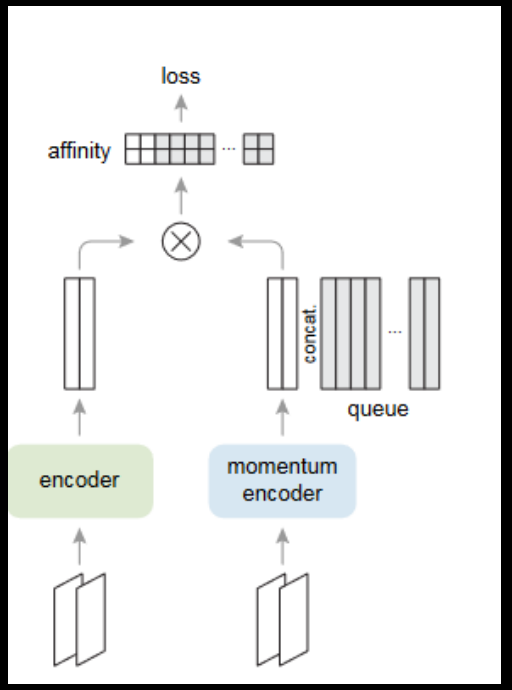
A Simple Framework for Contrastive Learning of Visual Representations
Ting Chen, Simon Kornblith, Mohammad Norouzi, Geoffrey Hinton. ICML 2020

SIMPLE CONTRASTIVE (SimCLR)

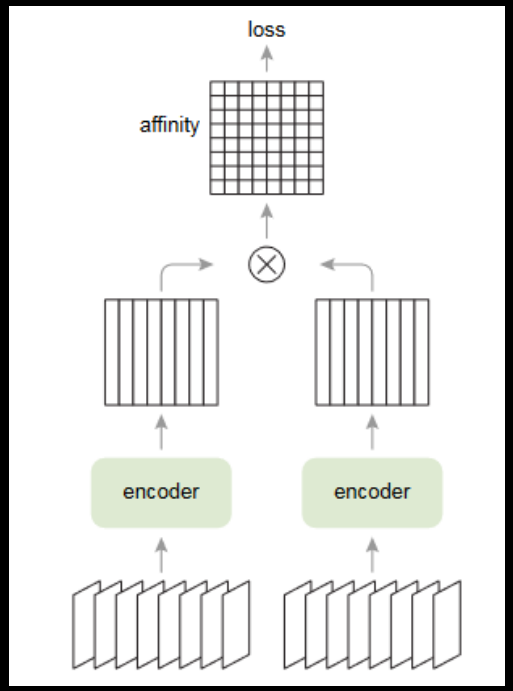
- use a base encoder network *f* and a **projection head** *g* to maximize agreement
- negative keys are from the batch, while in MoCov2 from a stored queue



Chen+@ICML'20

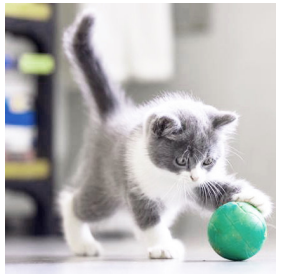


MoCov2
(Chen+@ArXiv'20)



SimCLR
(Chen+@ICML'20)

[Slide: Gabriela Csurka]



Model → q

Model → k

Model → n

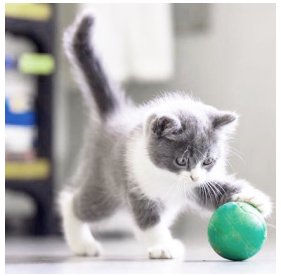
Retour à la formulation de départ:

L'apprentissage contrastif correspond à l'apprentissage d'une représentation qui discrimine les différentes instances (=images).

Une autre façon de formuler le problème:
L'apprentissage contrastif apprend des représentations qui sont **invariantes aux transformations** d'images qui ne changent pas leur sémantique, tout en évitant la solution triviale qui consiste à produire une représentation identique pour toutes les images.

MoCo [He@CVPR20]
SimCLR [Chen@ICML20]

Proxy task: Instance discrimination



Model

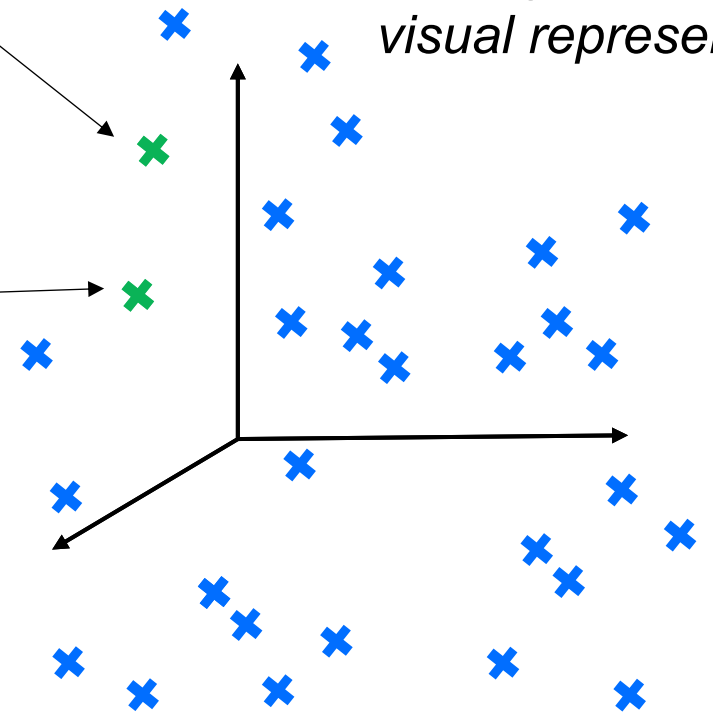
$\mathbf{q}$

Model

$\mathbf{k}$

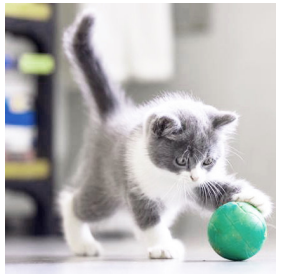
Model

$\mathbf{n}$



Key observation:

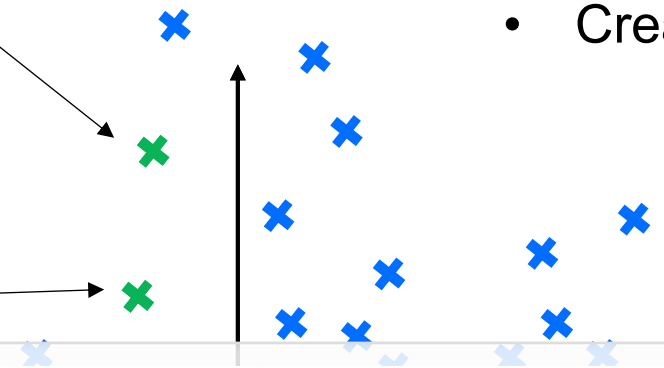
Making the proxy task more difficult leads to visual representations which generalize better



Model → $\mathbf{q}$

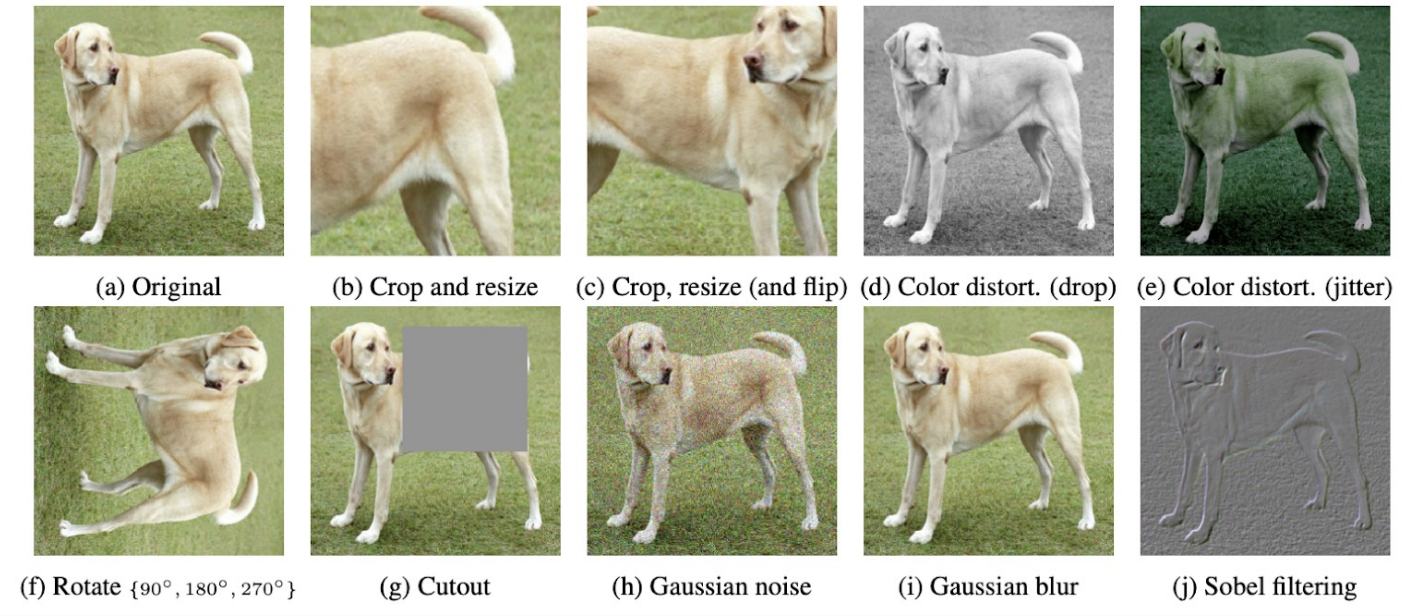
Model → $\mathbf{k}$

Model → $\mathbf{n}$

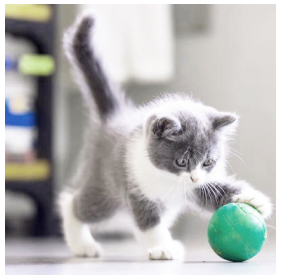


How to make the task harder?

- Create “better” positive pairs



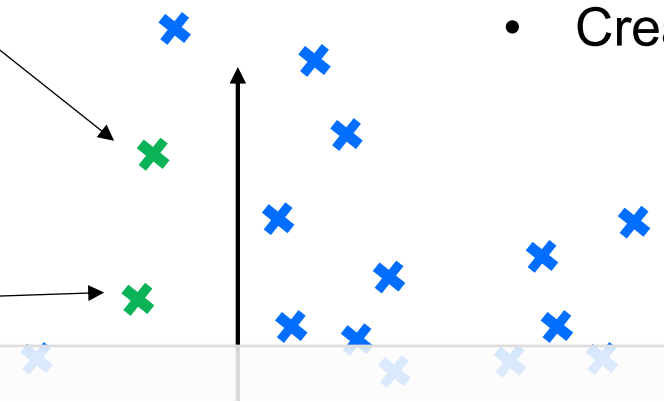
SimCLR [Chen@ICML20]



Model → q

Model → k

Model → n

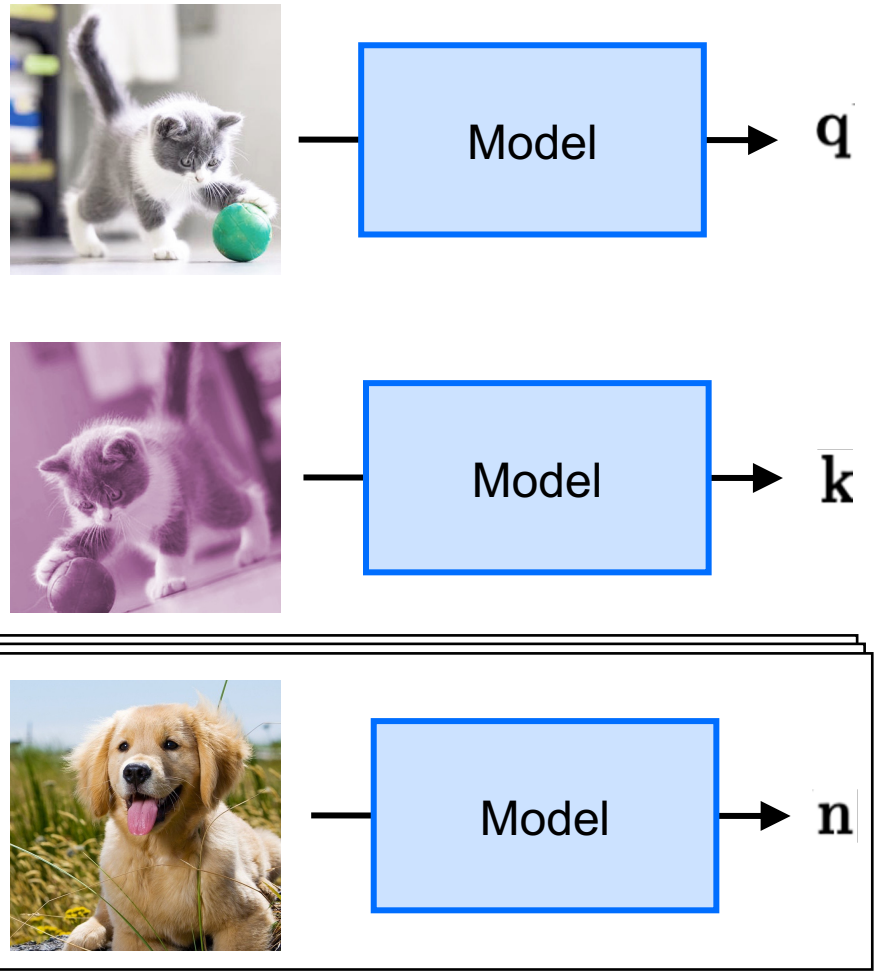


How to make the task harder?

- Create “better” positive pairs

Références qui se sont intéressées à cette question des paires positives plus difficiles:

[MoCo] He, Kaiming, et al. "Momentum contrast for unsupervised visual representation learning." CVPR 2020.
[SimCLR] Chen, Ting, et al. "A simple framework for contrastive learning of visual representations." ICML 2020.
[MoCo-v2] Chen, Xinlei, et al. "Improved baselines with momentum contrastive learning." arXiv preprint arXiv:2003.04297 (2020)
[InfoMin Aug.] Tian, Yonglong, et al. "What makes for good views for contrastive learning." NeurIPS 2020.

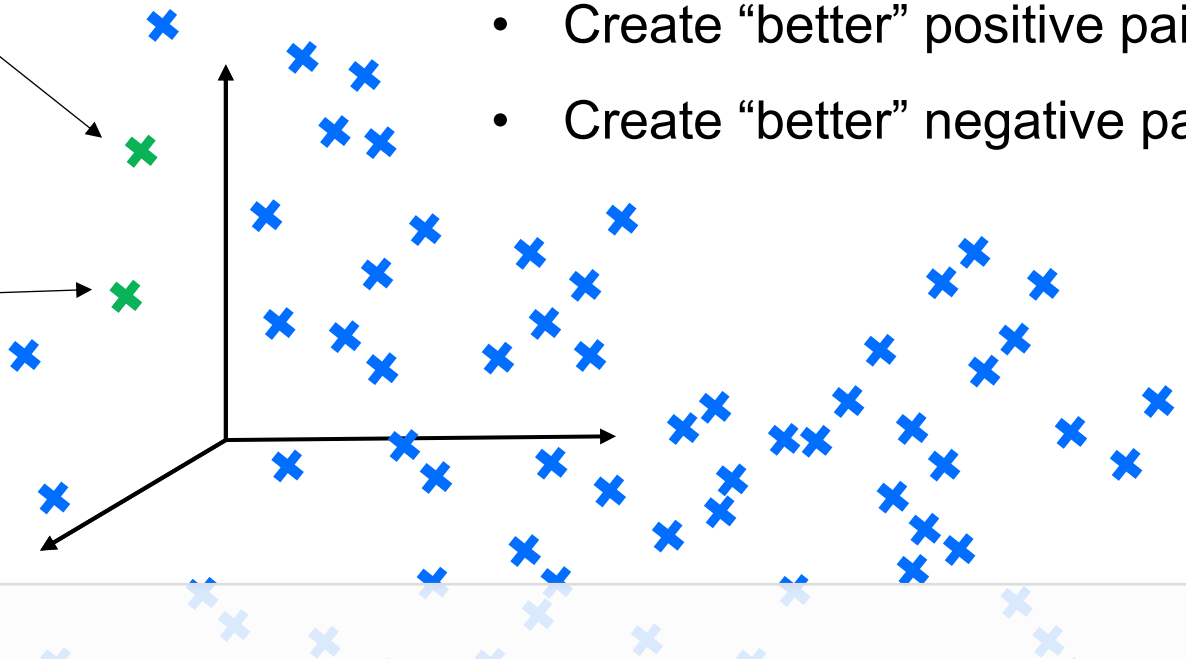


MoCo [He@CVPR20]

SimCLR [Chen@ICML20]

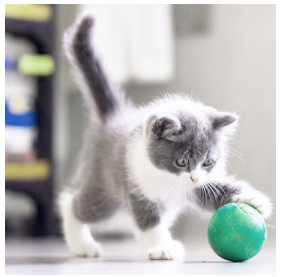
How to make the task harder?

- Create “better” positive pairs
- Create “better” negative pairs



SimCLR: utilise un très grand batch: plus de négatifs, donc plus de négatifs difficiles auxquels comparer l’ancre.

MoCo: augmente la taille de la mémoire, et donc a plus de chance de conserver des négatifs difficiles.



Model → $\mathbf{q}$

Model → $\mathbf{k}$

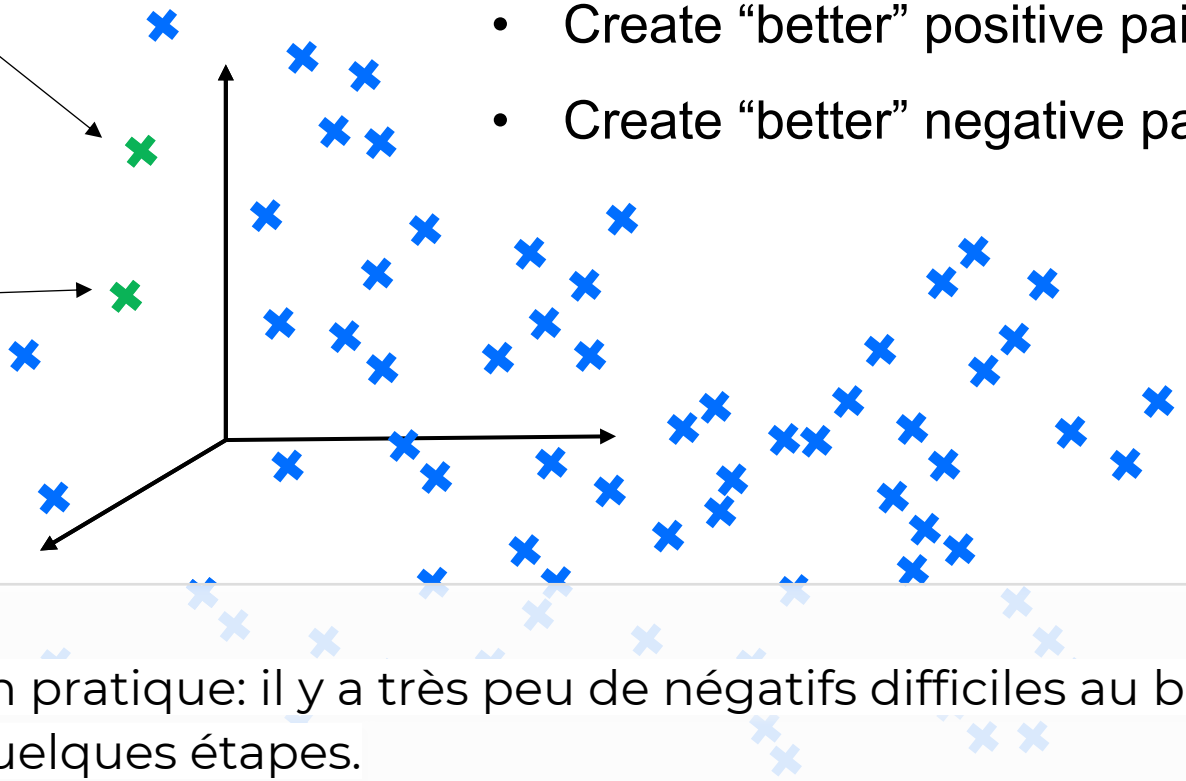
Model → $\mathbf{n}$

MoCo [He@CVPR20]

SimCLR [Chen@ICML20]

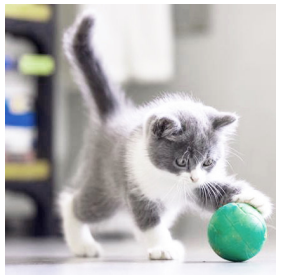
How to make the task harder?

- Create “better” positive pairs
- Create “better” negative pairs



En pratique: il y a très peu de négatifs difficiles au bout de quelques étapes.

Proposition: créer ces négatifs de façon artificielle, directement dans l’espace de représentation.



Model

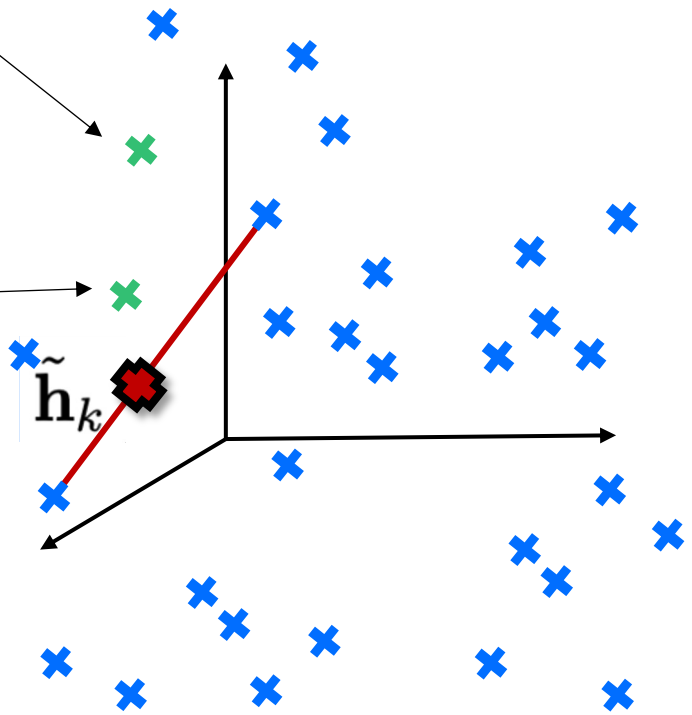
$\mathbf{q}$

Model

$\mathbf{k}$

Model

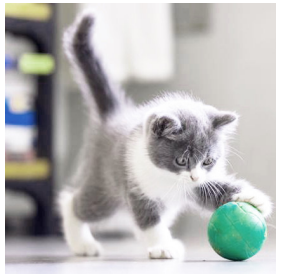
$\mathbf{n}$



What if we mix the **hardest** negatives for each query and synthesize new hard negatives?

by mixing **two negatives**

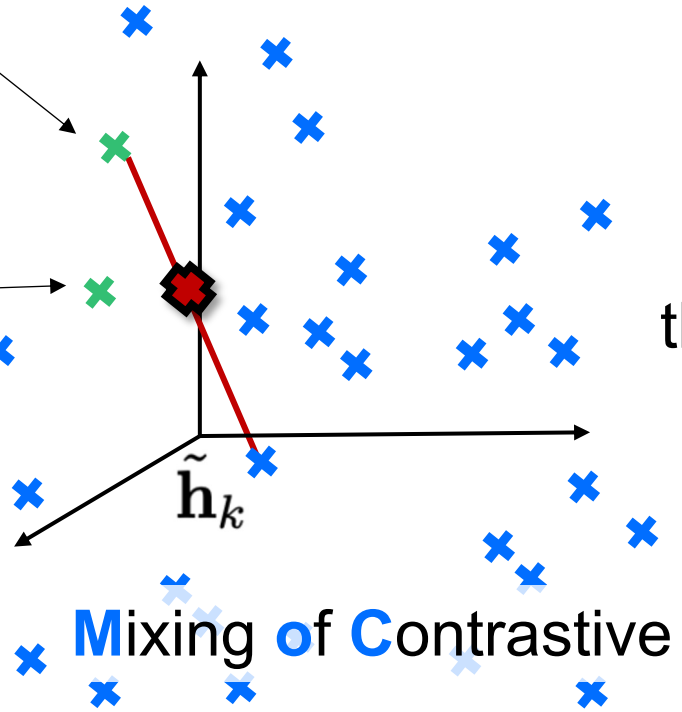
$$\tilde{\mathbf{h}}_k = \alpha_k \mathbf{n}_i + (1 - \alpha_k)\mathbf{n}_j$$



Model → $\mathbf{q}$

Model → $\mathbf{k}$

Model → $\mathbf{n}$



What if we mix the **hardest** negatives for each query and synthesize new hard negatives?

by mixing **two negatives**

$$\tilde{\mathbf{h}}_k = \alpha_k \mathbf{n}_i + (1 - \alpha_k)\mathbf{n}_j$$

the **query** with a **negative**

$$\tilde{\mathbf{h}}'_k = \beta_k \mathbf{q} + (1 - \beta_k)\mathbf{n}_j$$

Mixing **o**f **C**ontrastive **H**ard negati**v**es or **MoCHi**

Hard negative mixing for contrastive learning
Yannis Kalantidis, Bulent Sariyildiz, Noé Pion, Philippe Weinzaepfel, Diane Larlus
Conference on Neural Information Processing Systems (**NeurIPS**) 2020

Vers des méthodes sans négatifs

Rappel: Les méthodes contrastives reposent beaucoup sur les échantillons négatifs, et nous avons vu que leur succès dépend de la qualité de ces échantillons négatifs.

Peut-on apprendre une représentation sans échantillon négatif ?

Des méthodes ont été proposées, qui suivent principalement l'une de ces deux approches:

- Elles utilisent une **fonction de coût spécifique** (par exemple Barlow Twins)
- Elles se basent sur l'**auto-distillation** (par exemple BYOL ou DINO)

[**Barlow Twins**] Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction" ICML 2021.

[**BYOL**] Grill et al. "Boostrap Your Own Latent: A new approach to self-supervised learning" NeurIPS 2020.

[**Dino**] Caron et al. "Emerging Properties in Self-Supervised Vision Transformers" ICCV 2021.

Barlow Twins

Barlow Twins est une méthode conceptuellement simple, qui passe facilement à l'échelle, et qui ne nécessite pas d'échantillons négatifs, grâce à une fonction de coût qui s'intéresse à la **décorrélation** des dimensions de la représentation.

Grâce à cette **fonction de coût**, les solutions triviales sont évitées malgré le fait que cette méthode n'utilise que des paires positives d'échantillons.

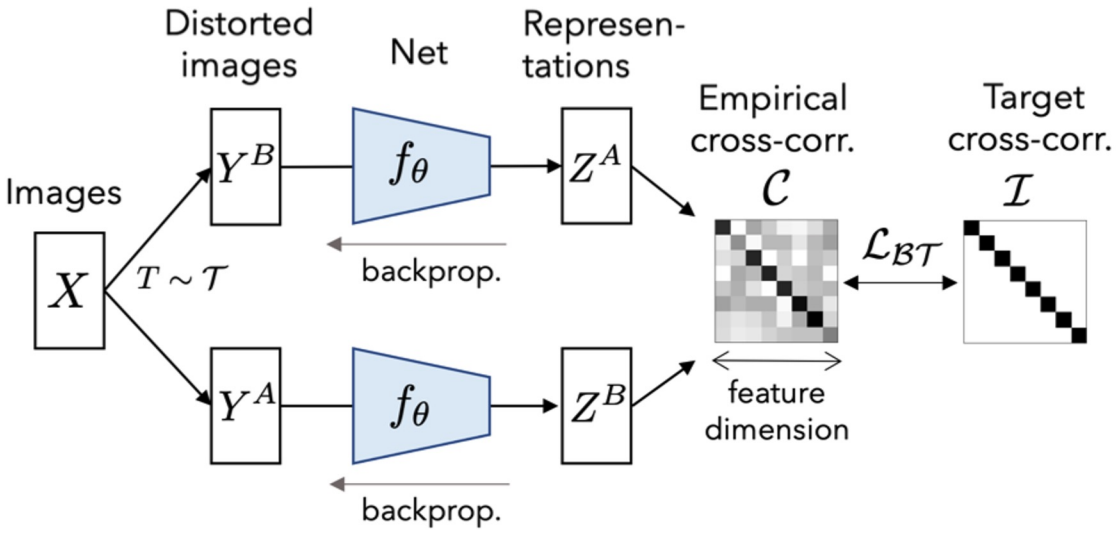
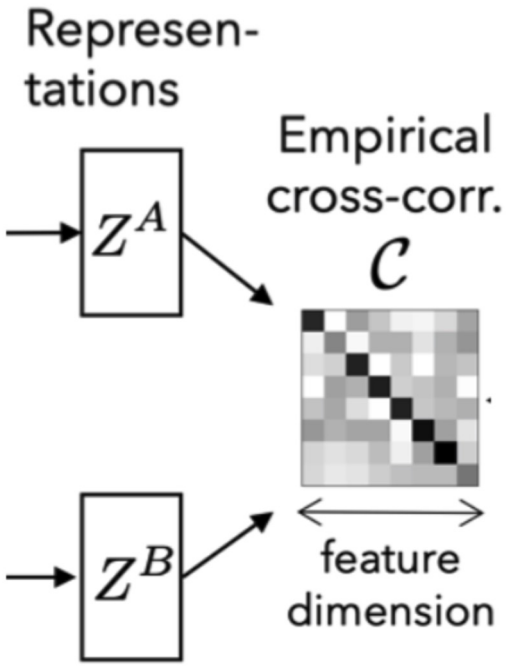


Figure from [**Barlow Twins**]

[**Barlow Twins**] Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction" ICML 2021.

Barlow Twins



$$\mathcal{C}_{ij} = \frac{\sum_b \hat{z}^A_{b,i} \hat{z}^B_{b,j}}{\sqrt{\sum_b (\hat{z}^A_{b,i})^2} \sqrt{\sum_b (\hat{z}^B_{b,j})^2}}$$

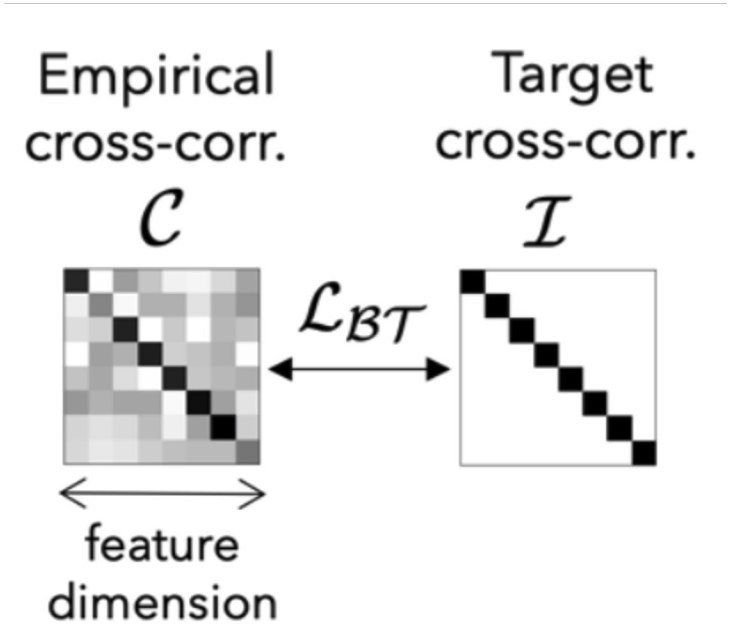
Empirical (batch) cross-corr. matrix $\mathcal{C}$*:*

- across the **feature dimension**
- outer product of the normalized representation for every positive pair
- averaged over the batch

[Slide credit: Yannis Kalantidis]

[**Barlow Twins**] Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction" ICML 2021.

Barlow Twins



The Barlow Twins loss

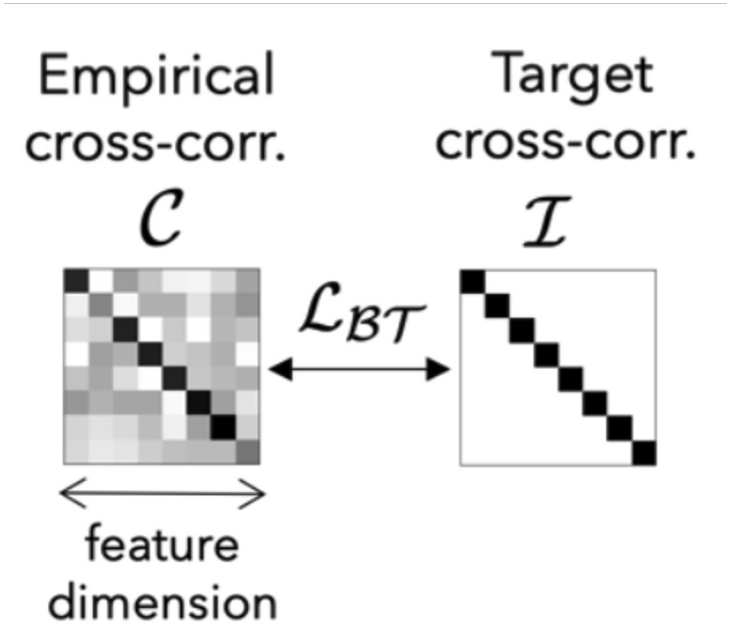
$$\mathcal{C}_{ij} = \frac{\sum_b \hat{z}^A_{b,i} \hat{z}^B_{b,j}}{\sqrt{\sum_b (\hat{z}^A_{b,i})^2} \sqrt{\sum_b (\hat{z}^B_{b,j})^2}}$$

$$\mathcal{L}_{BT} = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{push diagonal elements to 1}} + \lambda \underbrace{\sum_i \sum_{i \neq j} \mathcal{C}^2_{ij}}_{\text{push off-diagonal elements to 0}},$$

[Slide credit: Yannis Kalantidis]

[**Barlow Twins**] Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction" ICML 2021.

Barlow Twins



The Barlow Twins loss

$$\mathcal{C}_{ij} = \frac{\sum_b \hat{z}^A_{b,i} \hat{z}^B_{b,j}}{\sqrt{\sum_b (\hat{z}^A_{b,i})^2} \sqrt{\sum_b (\hat{z}^B_{b,j})^2}}$$

$$\mathcal{L}_{BT} = \underbrace{\sum_i (1 - \mathcal{C}_{ii})^2}_{\text{maximize the dot product of every positive pair}} + \lambda \underbrace{\sum_i \sum_{i \neq j} \mathcal{C}^2_{ij}}_{\text{de-correlate output dimensions}},$$

[Slide credit: Yannis Kalantidis]

[**Barlow Twins**] Zbontar et al. "Barlow Twins: Self-Supervised Learning via Redundancy Reduction" ICML 2021.

Distillation - Définition

En apprentissage automatique, la **distillation de connaissance** (*knowledge distillation*) est le processus qui consiste à **transférer de la connaissance** d'un modèle de grande taille, vers un plus petit.

Motivation:

Des réseaux de neurones plus petits sont moins coûteux surtout au moment de l'inférence, et peuvent être déployés sur des plateformes moins puissances (par exemple des systèmes embarquées).

Bien que les plus grands modèles (les réseaux de neurones les plus profonds ou les ensembles de modèles) ont une plus grande capacité / expressivité que les petits modèles, cette capacité n'est pas forcément utilisée pleinement, mais ils restent plus couteux à déployer, alors qu'un petit réseau pourrait très bien donner des résultats de même qualité, si sa capacité est suffisante.
Si ce n'est pas le cas, une légère dégradation des résultats peut être acceptable au regard de ces contraintes de calcul.

Intuition:

Au lieu d'apprendre directement un réseau compact, l'idée est d'entrainer un réseau plus grand avec plus de poids / paramêtres, et de l'utiliser ensuite pour guider l'apprentissage du réseau compact.

Distilling the Knowledge in a Neural Network
Geoffrey Hinton, Oriol Vinyals, Jeffrey Dean.
NeurIPS15 - IPS Deep Learning and Representation Learning Workshop

Distillation - Définition

En apprentissage automatique, la **distillation de connaissance** (*knowledge distillation*) est le processus qui consiste à **transférer de la connaissance** d'un modèle de grande taille, vers un plus petit.

- Le modèle à la source de la distillation est souvent appelé **modèle « professeur »** (***teacher network***)
- Le modèle qui est appris par distillation est souvent appelé **modèle « étudiant »** (***student network***)

Response-based knowledge distillation:

Quand on parle de distillation de modèle en apprentissage profond, on pense généralement à ce qu'on appelle la **distillation basée sur la réponse** (*Response-based knowledge distillation).*

L'idée est que le réseau « étudiant » essaie de reproduire la prédiction finale du réseau « professeur », c'est-à-dire la probabilité sur les classes dans le cas de la classification.

C'est une méthode simple et effective, utilisée par exemple pour la compression de modèle.

Distillation - Définition

En apprentissage automatique, la **distillation de connaissance** (*knowledge distillation*) est le processus qui consiste à **transférer de la connaissance** d'un modèle de grande taille, vers un plus petit.

Illustration

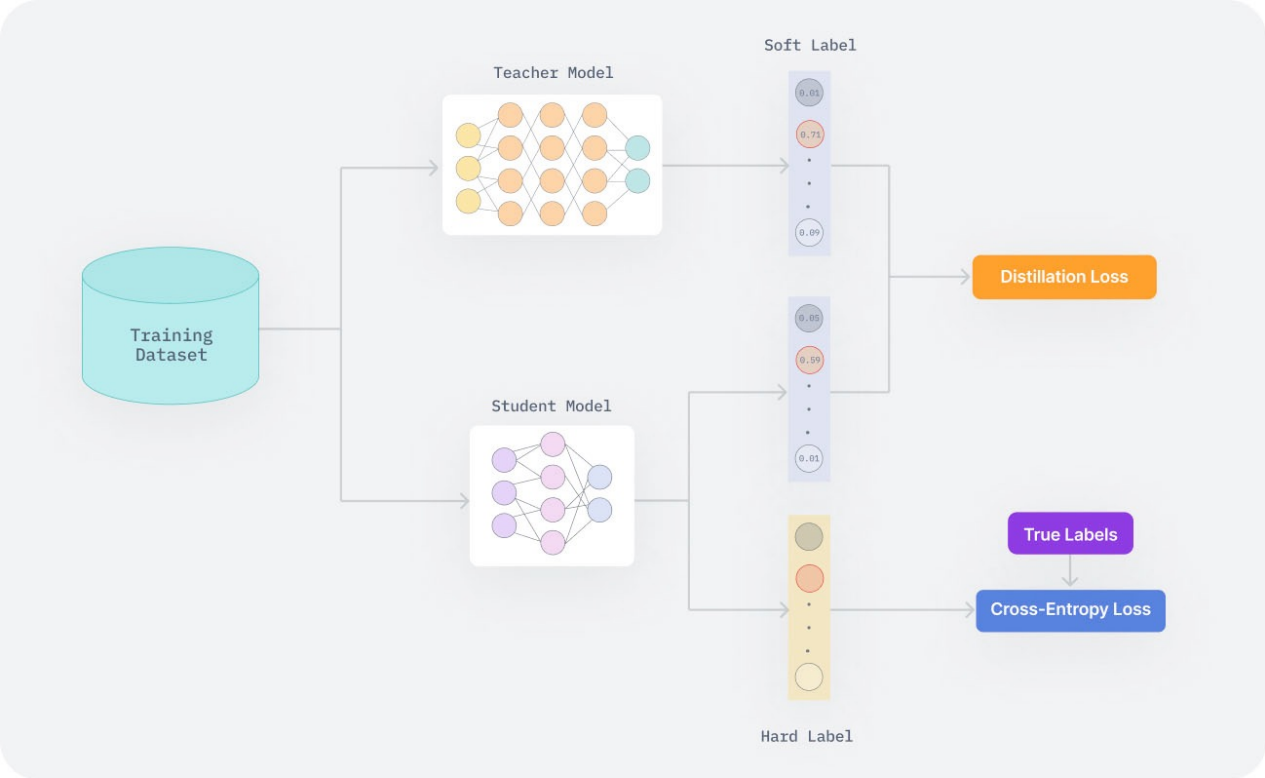


Image Credit: V7

Cette illustration correspond à la distillation telle qu'originalement proposée par (Hinton et al., 2015)

Elle contient deux fonctions de coût

- une qui considère la distillation depuis le réseau « professeur »
- une qui considère la vérité terrain

Distillation - Définition

En apprentissage automatique, la **distillation de connaissance** (*knowledge distillation*) est le processus qui consiste à **transférer de la connaissance** d'un modèle de grande taille, vers un plus petit.

Illustration

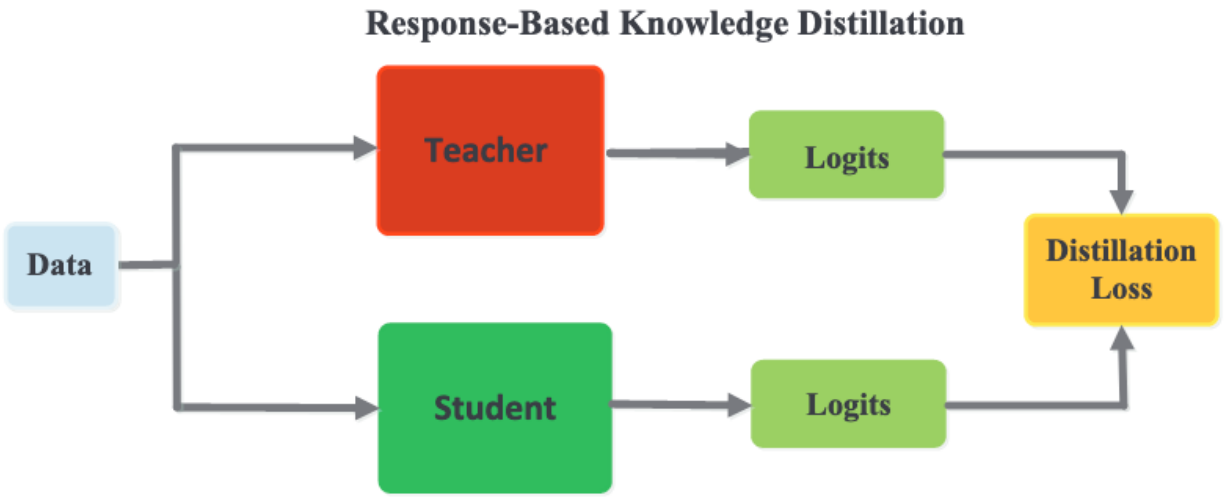


Fig. 4 The generic response-based knowledge distillation.

Image Credit: https://arxiv.org/pdf/2006.05525.pdf

Mais souvent, seule la fonction de coût (*loss*) de distillation est utilisée.

Self-Distillation - Définition

La plupart des méthodes de distillation sont dites hors-ligne (*offline*). La distillation a lieu une fois pour toute, sur un modèle professeur qui ne varie pas, et elle n'a lieu que dans un sens.

Définition:
Dans le cas de **l'auto-distillation (*self-distillation*)**, le même réseau est utilisé pour « le professeur » et pour « l'étudiant ».

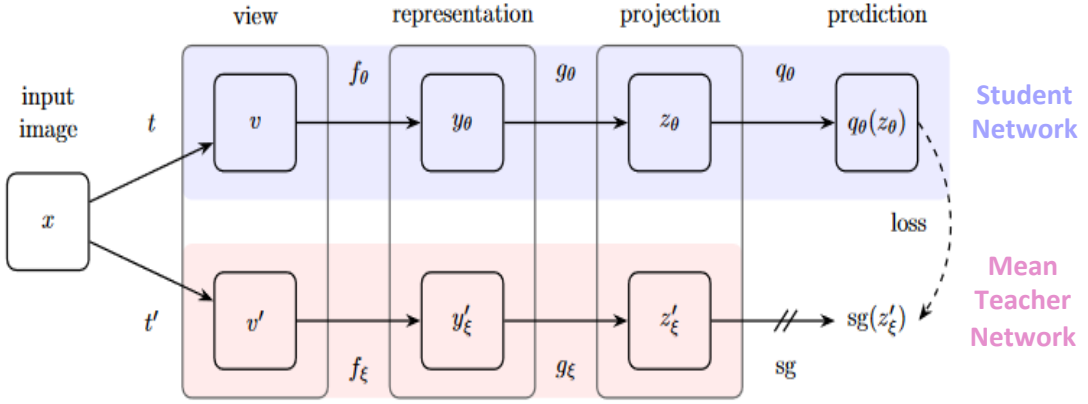
C'est une technique assez populaire en apprentissage auto-supervisé.

BYOL

BYOL est une méthode **auto-supervisée par auto-distillation**.

Elle se repose sur deux réseaux de neurones :

- Le réseau étudiant
- Le réseau professeur



Ces deux réseaux sont appris conjointement et interagissent l'un avec l'autre.

A partir d'une version augmentée d'une image, le réseau étudiant est entrainé à prédire la sortie du réseau professeur sur cette même image, mais modifiée différemment.
Ce réseau professeur produit donc la « vérité terrain ».

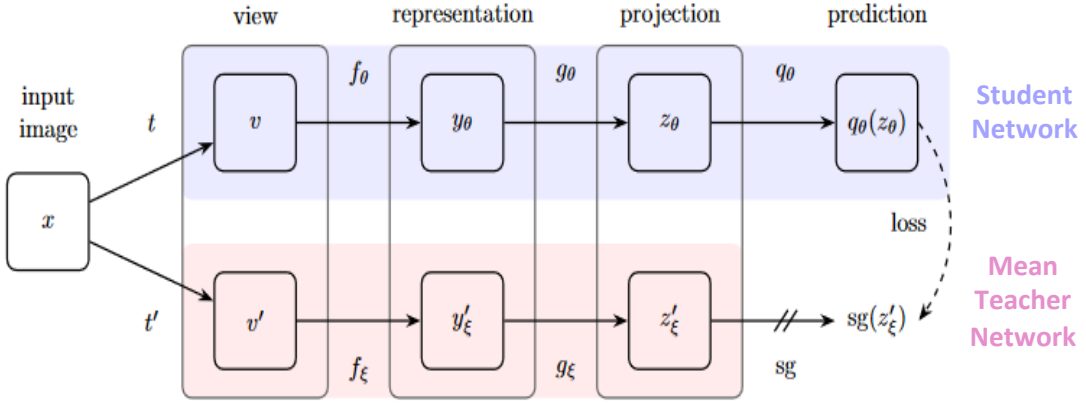
Le réseau professeur est mis à jour de façon incrémentale, en utilisant la méthode dite de ***l'exponential moving average*** (EMA). Cette mise à jour consiste à effectuer une moyenne pondérée entre la valeur courante de ses paramètres, et celles des paramètres du réseau étudiant.

Intuitivement, à chaque époque, le réseau professeur se met à jour de façon similaire au réseau étudiant, mais de façon plus lente.

Bootstrap Your Own Latent: A new approach to self-supervised learning
Jean-bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, Michal Valko
NeurIPS 2020

BYOL

Pourquoi ça marche?

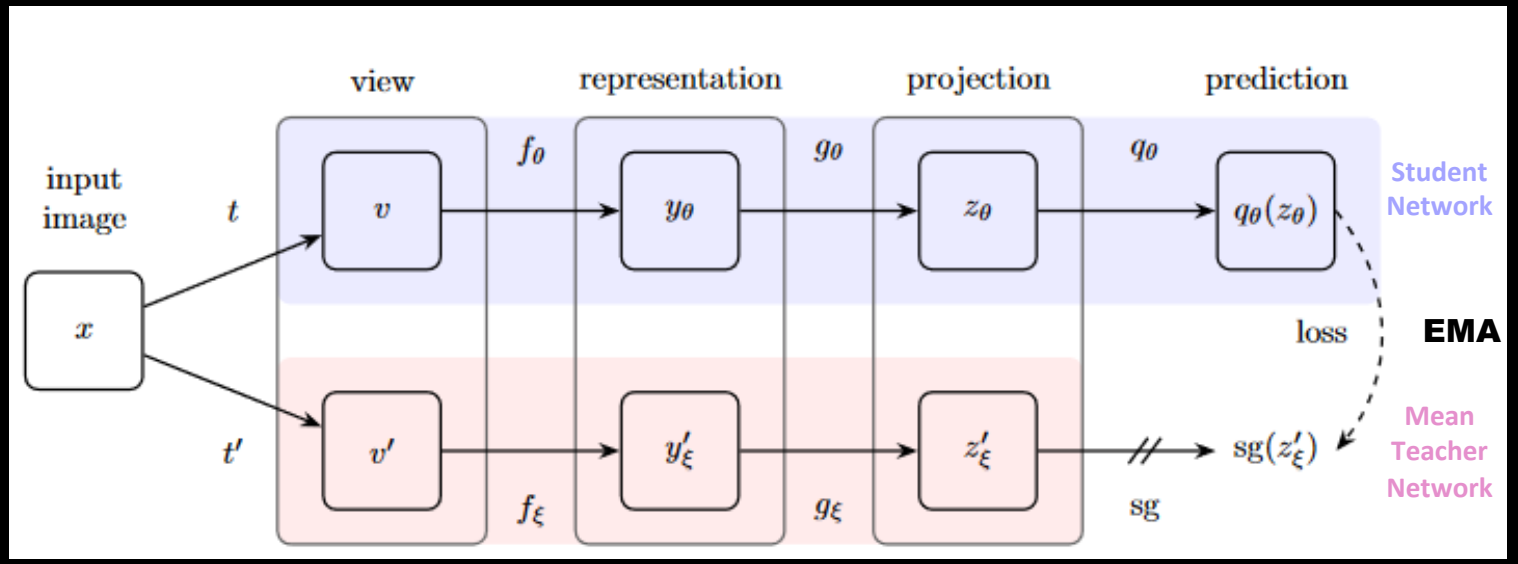


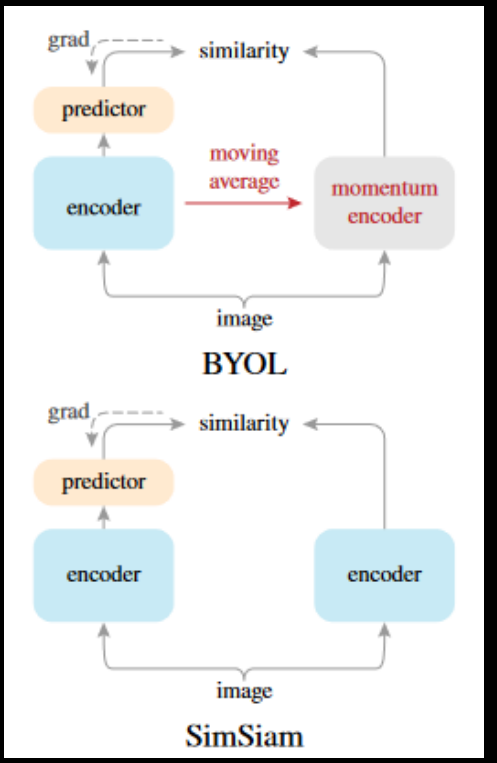
- L'intuition est que BYOL raffine itérativement sa représentation en construisant un réseau « moyen » qui se met à jour lentement.
- BYOL utilise les transformations d'image de SimCLR
- BYOL améliore les résultats par rapport à SimCLR

Bootstrap Your Own Latent: A new approach to self-supervised learning
Jean-bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Daniel Guo, Mohammad Gheshlaghi Azar, Bilal Piot, Koray Kavukcuoglu, Remi Munos, Michal Valko
NeurIPS 2020

BOOTSTRAP YOUR OWN LATENT (BYOL)

- An exponentially moving average (EMA) network (mean teacher) to produce targets for an online student network to stabilize the bootstrap step (stop gradient)
- Apply group normalization (GN) and weight standardization (WS)
- Asymmetric

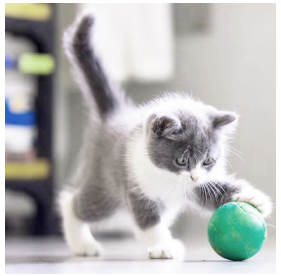




Grill+@NeurIPS'20

[Slide: Gabriela Csurka]

Self-supervised learning – Contrastive learning



Model

$\mathbf{q}$

Model

$\mathbf{k}$

Model

$\mathbf{n}$

Pretext task = Instance discrimination

Rappel

L'apprentissage contrastif vise à

- discriminer les images
- apprendre des représentations qui sont invariantes aux transformations

En pratique

Des groupes distincts sont formés dans l'espace de représentation. Cela fait penser à du clustering

Peut-on directement utiliser du *clustering* pour l'apprentissage auto-supervisé de représentations?

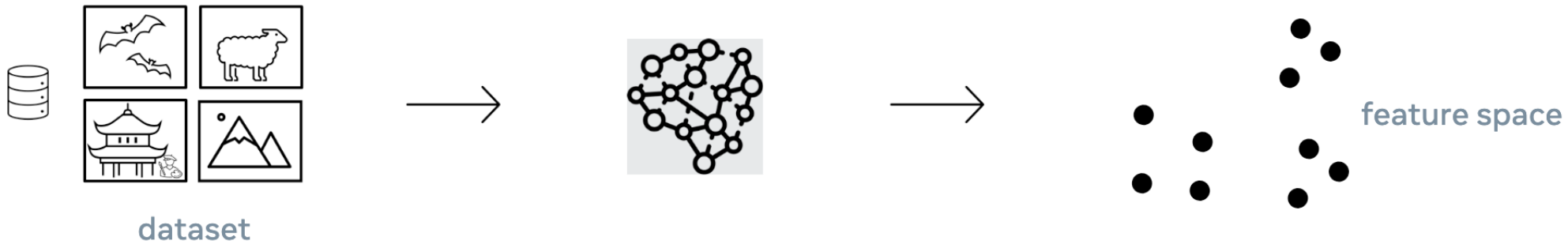
C'est l'intuition qui a guidé:

DeepCluster: Deep Clustering for Unsupervised Learning of Visual Features
Mathilde Caron, Piotr Bojanowski, Armand Joulin, Matthijs Douze
ECCV 2018

(note: tous les prochains transparents proviennent d'une présentation de Mathilde Caron)

DeepCluster

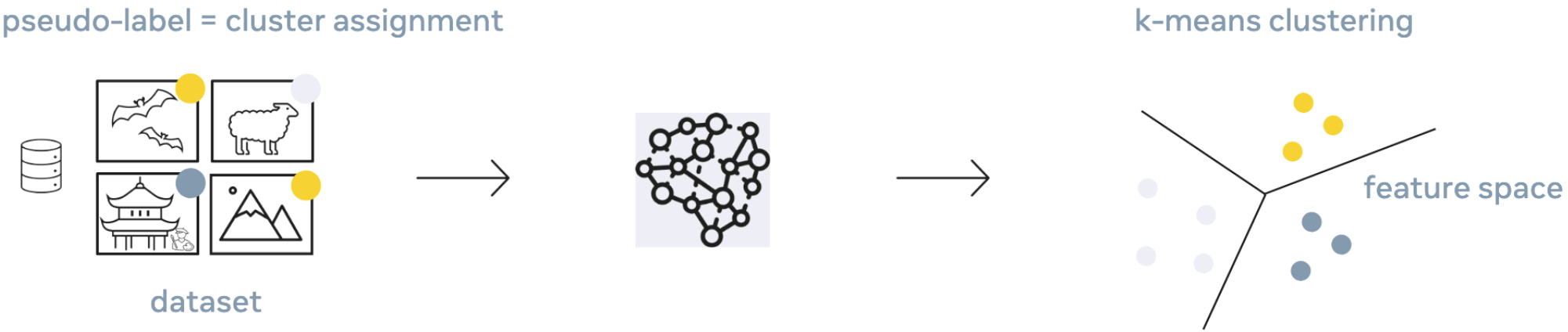
Principe:
Initialiser le modèle, par exemple aléatoirement
Extraire un descripteur (*feature*) pour chaque image à l'aide de ce modèle



[based on Mathilde Caron's slides]

DeepCluster

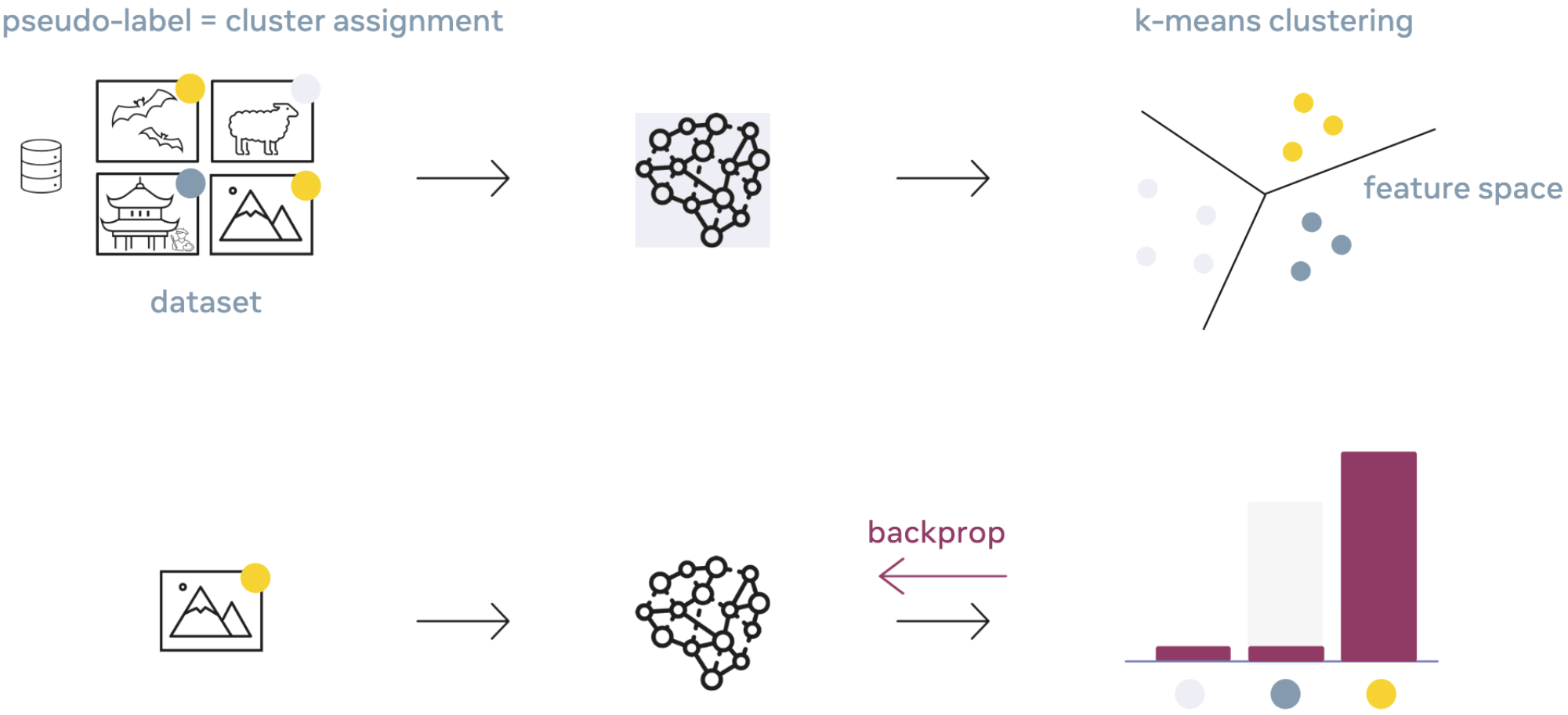
Principe:
Réaliser un clustering dans l'espace de représentation ainsi obtenu.
Utiliser l'affectaction aux clusters comme un label pour les images.



[based on Mathilde Caron's slides]

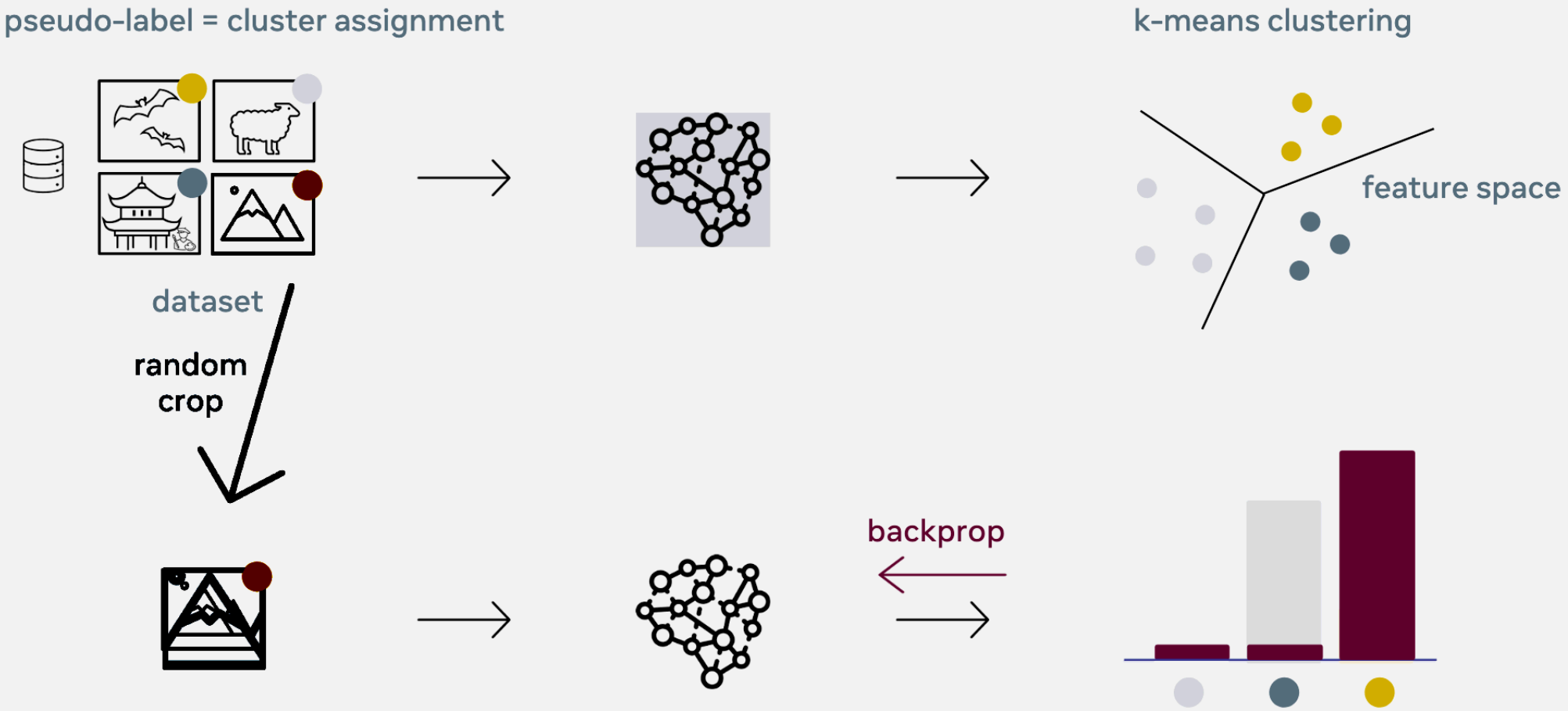
DeepCluster

Principe:
Entrainer le modèle à prédire ce label obtenu par clustering.



[based on Mathilde Caron's slides]

Invariance to cropping

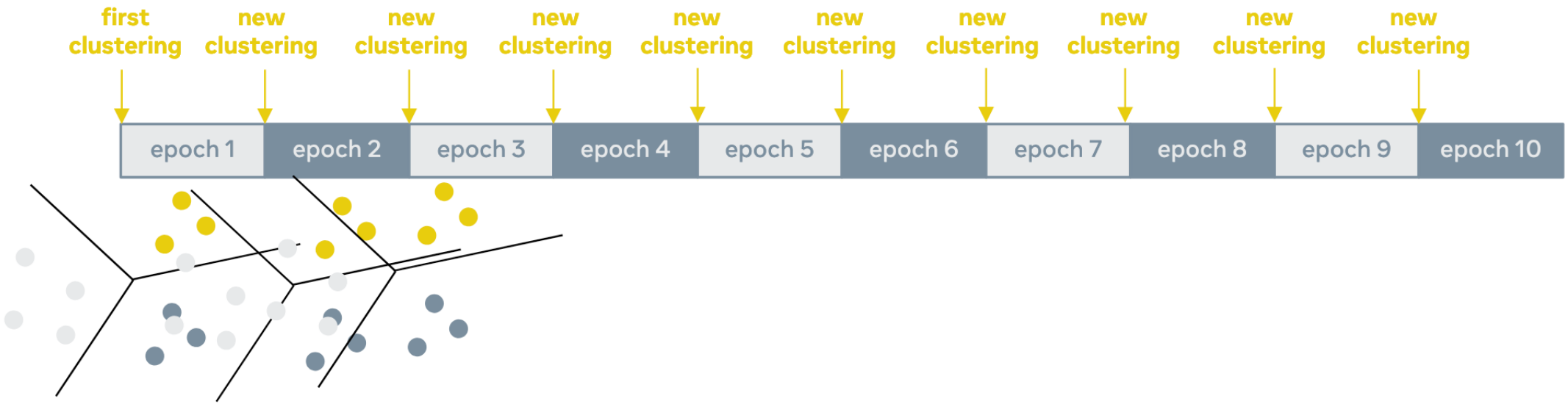


[based on Mathilde Caron's slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)

The clusters (i.e. pseudo-labels) are refined during training



[based on Mathilde Caron's slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)

Huge dataset: we can afford only 2 epochs!

Problem: clusters are refined only once…

[based on Mathilde Caron’s slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)

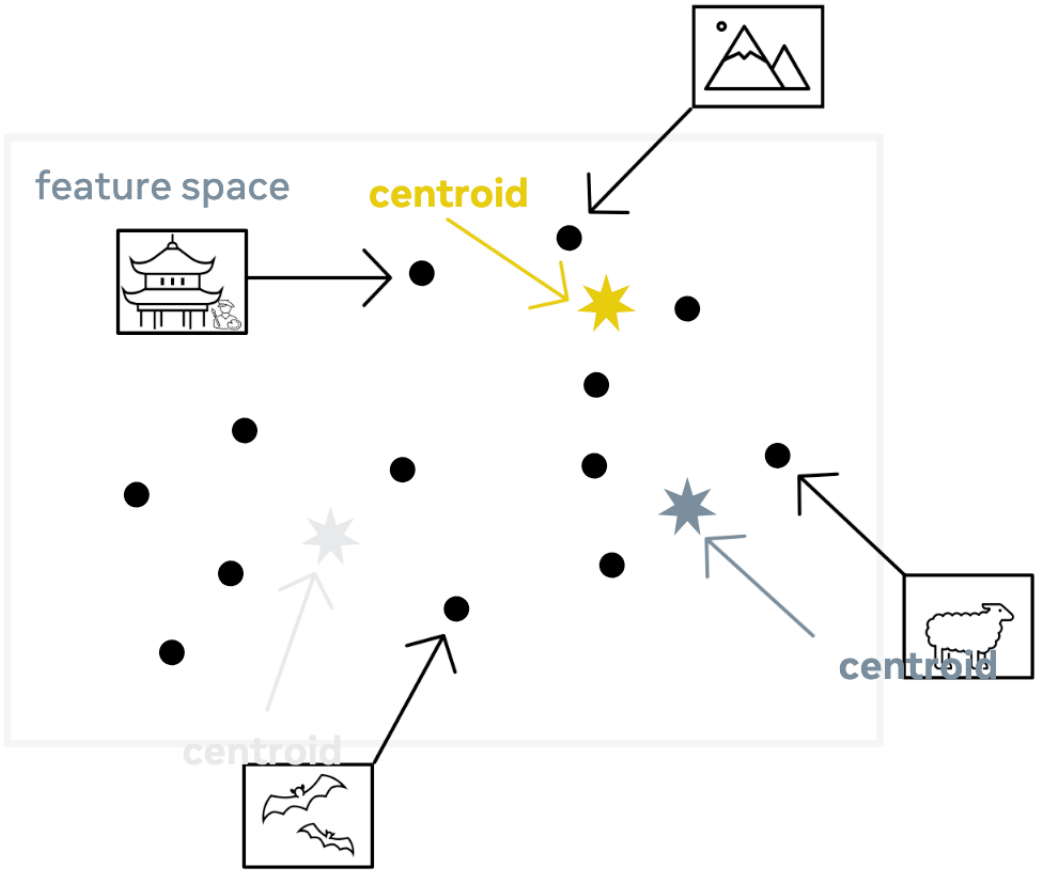
Even bigger dataset: we never see an image twice

Problem: the clusters are never refined!

[based on Mathilde Caron's slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?

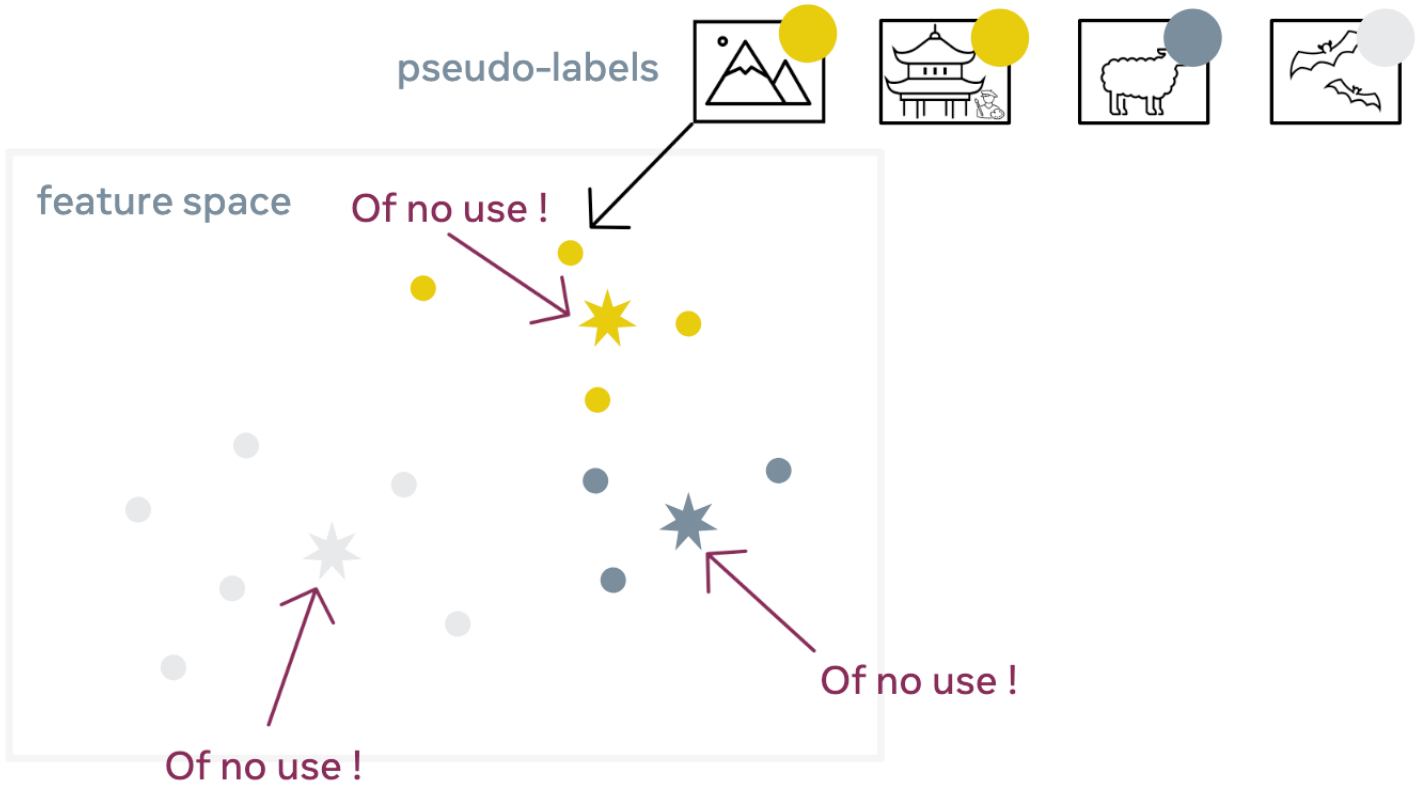


[based on Mathilde Caron's slides]

Limitations of DeepCluster

Le centroïde n'apporte pas d'information, seul l'appartenance aux clusters importe.

- Does not scale (depends on the dataset size)
- Do we really need k-means ?



[based on Mathilde Caron's slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse

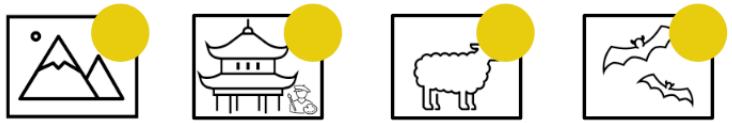
Collapse of feature space:

En pratique, éviter les solutions triviales est assez délicat, et nécessite une implémentation spécifique.

[based on Mathilde Caron’s slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse



Collapse of feature space:

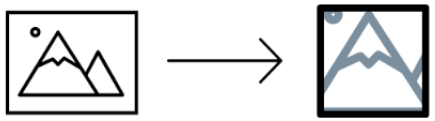
En pratique, éviter les solutions triviales est assez délicat, et nécessite une implémentation spécifique.

Une implémentation naïve résulte en un seul cluster.

[based on Mathilde Caron's slides]

Limitations of DeepCluster

- Does not scale (depends on the dataset size)
- Do we really need k-means ?
- Tricks to avoid collapse
- Importance of random cropping is only implicit



La transformation d'image 'crop' ne joue qu'un rôle implicite.

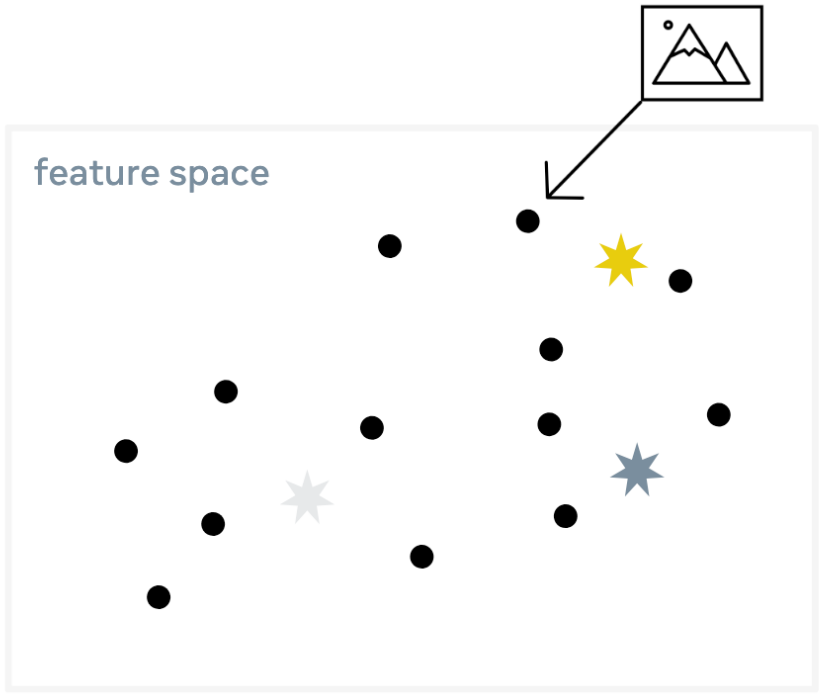
(c'est un point crucial de SWAV, mais surtout de DINO)

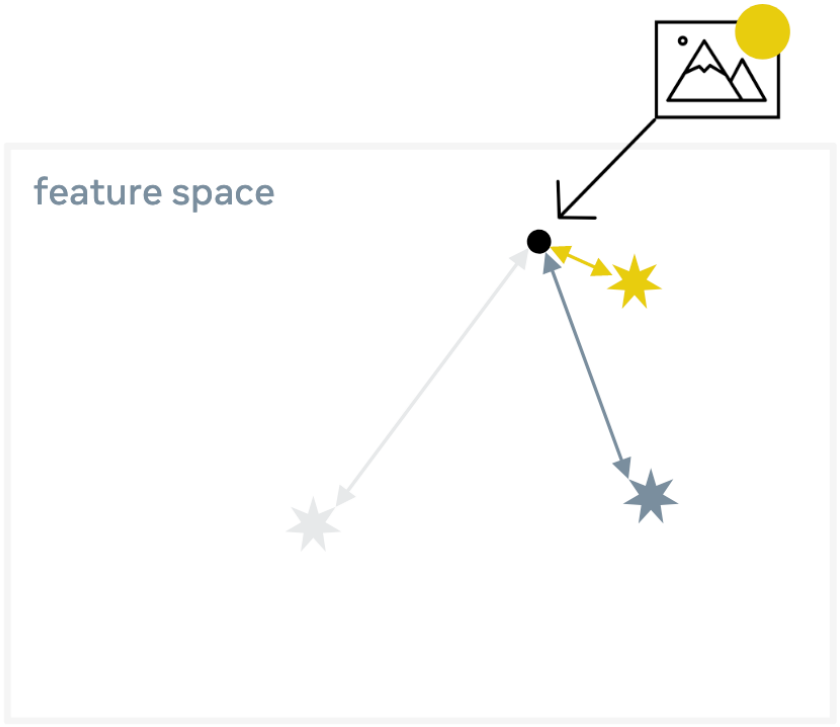
How to overcome these limitations ?

[based on Mathilde Caron's slides]

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

github.com/facebookresearch/swav

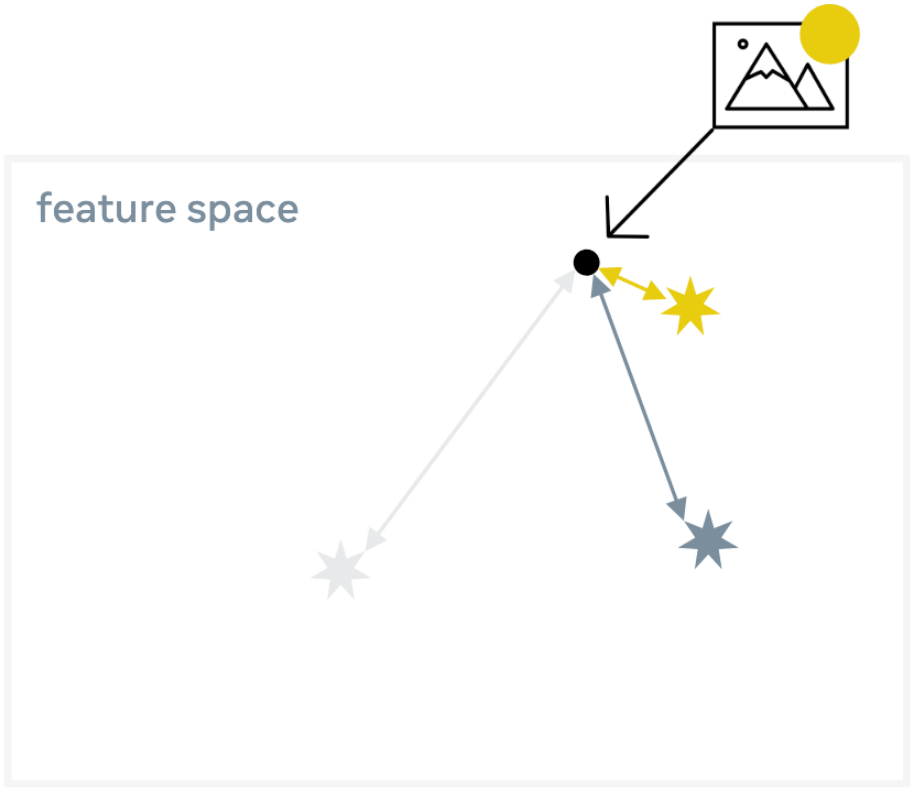


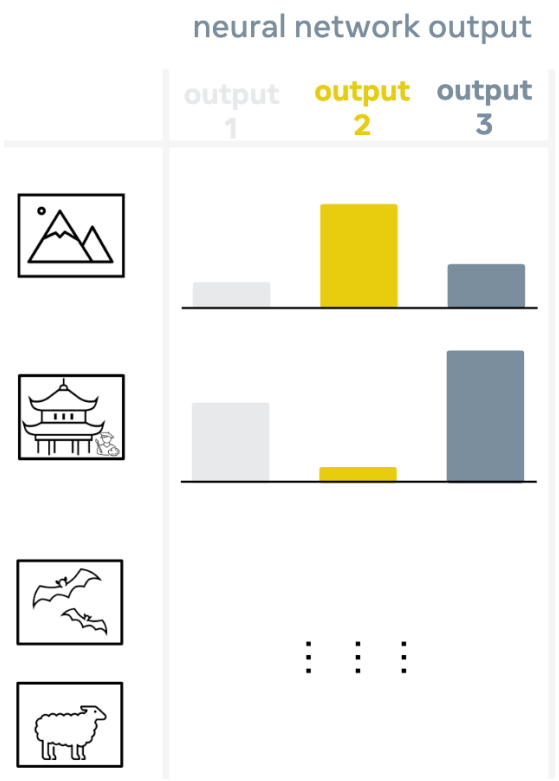


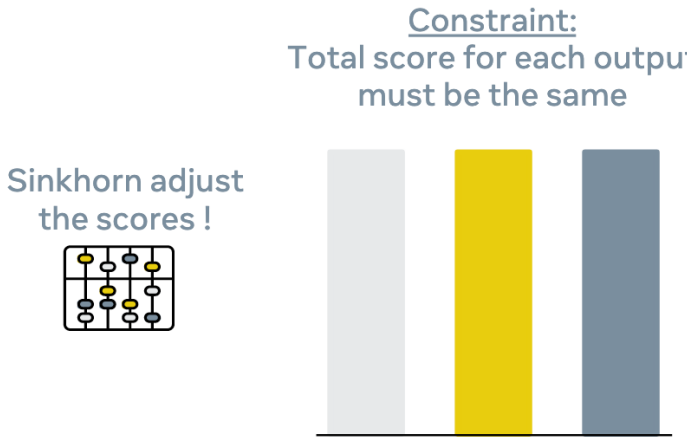
Observations

- For training, all we need is a score for each cluster
- We can directly train the neural network to output scores

[based on Mathilde Caron's slides]







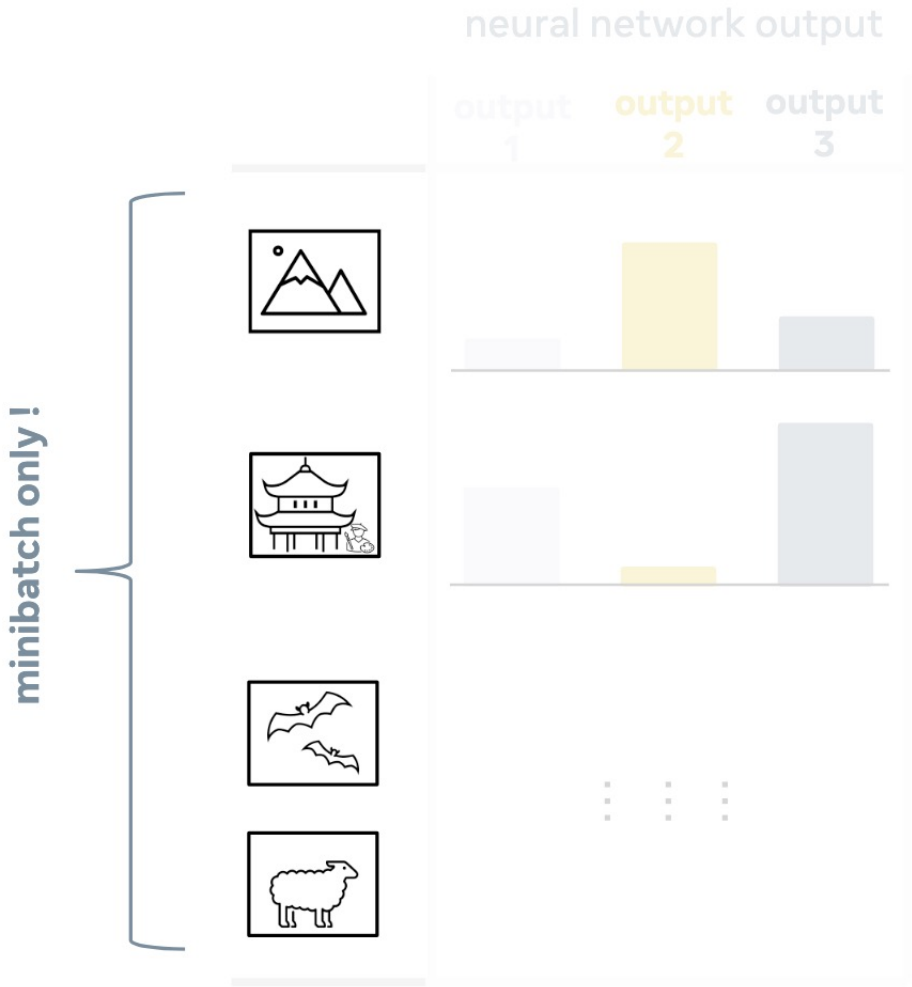
Observations

- For training, all we need is a score for each cluster
- We can directly train the neural network to output scores

Constraints: total score for each output must be the same

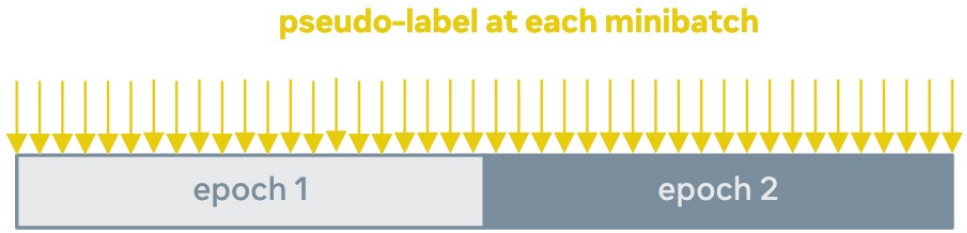
[based on Mathilde Caron's slides]

Pseudo-labels in SwAV



Recap'

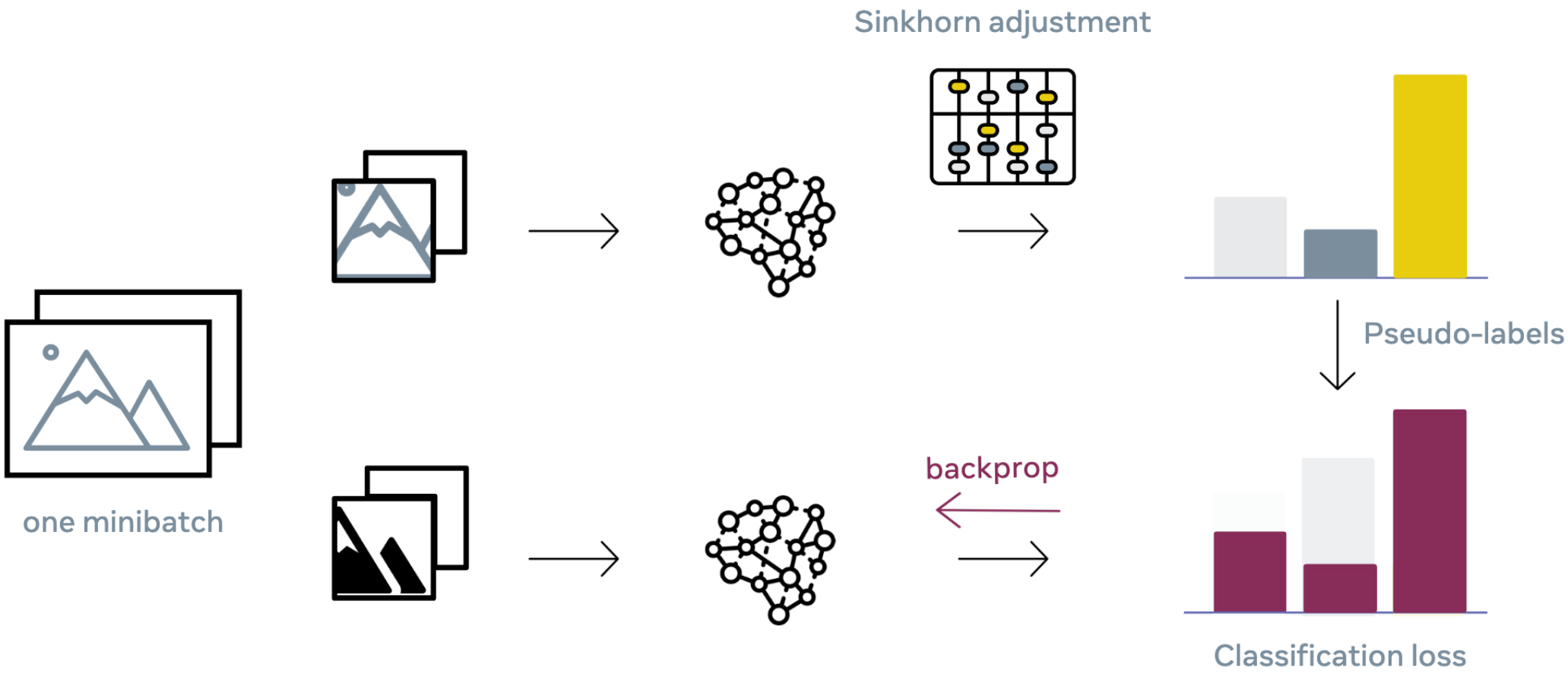
- We don't need k-means
- Explicit constraints to prevent collapse
- Scalable



[based on Mathilde Caron's slides]

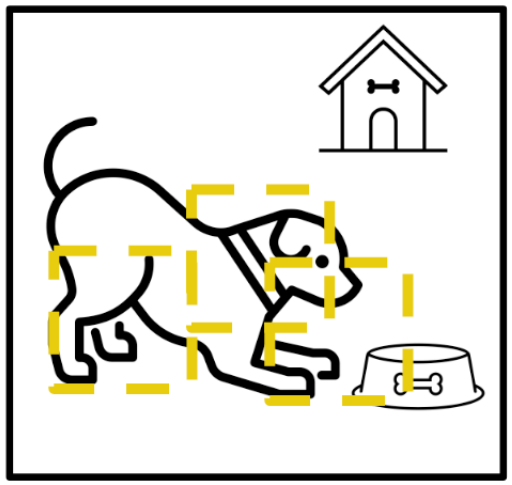
SwAV: the full picture

SimCLR - Chen et al. 2020



[based on Mathilde Caron's slides]

Multi-crop



Local crops

Global crops

Local predict the pseudo-label of global

Local-to-global matching

Encore une fois, cette méthode se base sur l'intuition qu'une bonne représentation (sémantique?) est telle qu'à partir seulement d'une partie d'une image, on peut prédire le descripteur global de toute l'image.

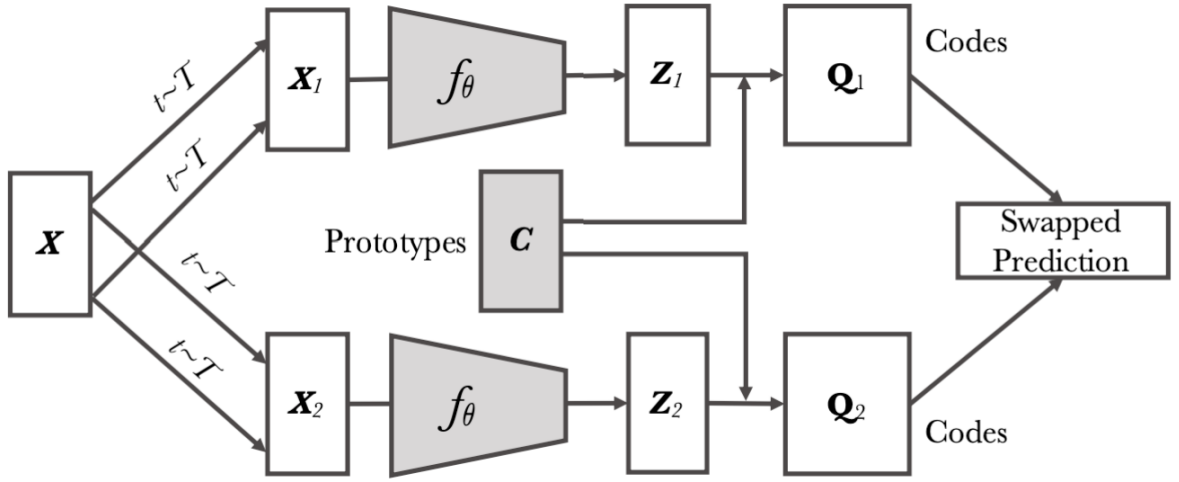
[based on Mathilde Caron's slides]

Swapping Assignments between multiple Views (SwAV)

Résumé

Différences fondamentales avec les méthodes par apprentissage contrastif:

- Les augmentations
- Une étape de clustering au lieu du projecteur utilisé par exemple dans SimCLR



Architecture finale

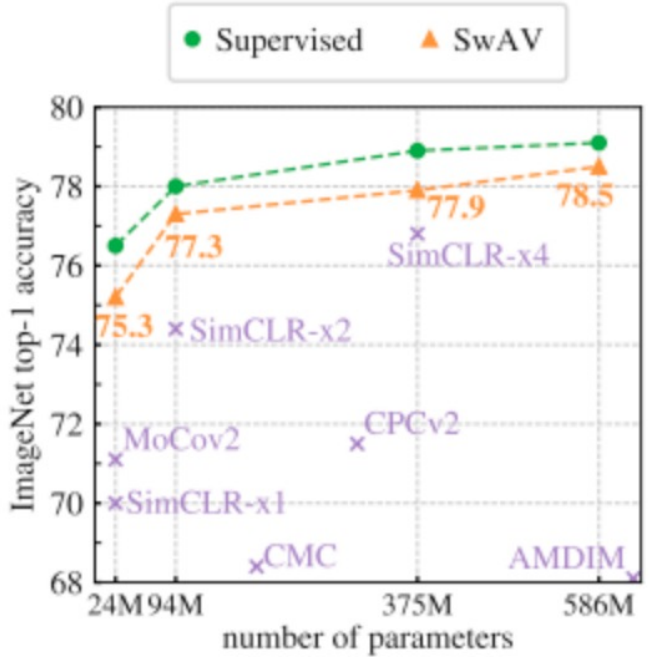
- Augmentation “multi-crop” en plus des augmentations classiques
- Clustering
- Fonction de coût basée sur un problème de prédiction d’échange (*swap*)

    L(z,q) mesure l’accord entre

        q: le ”code” (résultat du clustering) de la première vue et
        z: la représentation de la deuxième vue

Les “codes” peuvent être considérés comme la vérité terrain.

Hypothèse: des vues (i.e. transformations) différentes de la même image devraient contenir la même information, donc il devrait être possible de prédire le “code” d’une vue à partir de la représentation d’une autre vue



Next

Une des premières méthodes qui utilise les transfomeurs dans de l'apprentissage auto-supervisé pour la vision par ordinateur.

Une amélioration de SwAV par les même auteurs.

DINO: Emerging Properties in Self-Supervised Vision Transformers
Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, Armand Joulin
ICCV21

github.com/facebookresearch/dino

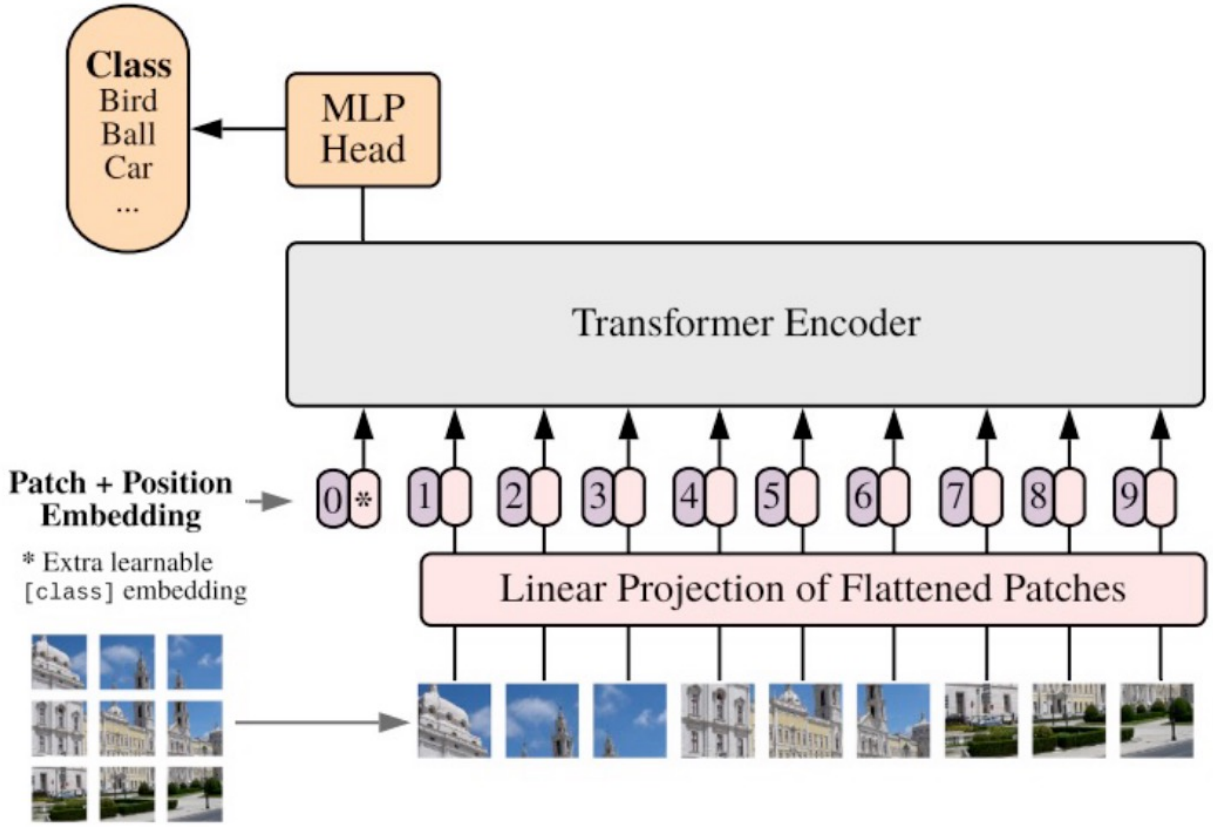
Les transformeurs en vision par ordinateur

Vision Transformer or ViT

Les transformeurs modélisent les relations entre paires de “token”.

En vision, l’unité de base est le pixel, mais considérer des paires de pixels aurait un coût trop élevé.

Proposition des ViTs: ce sont les patchs qui sont traités comme des “token”.

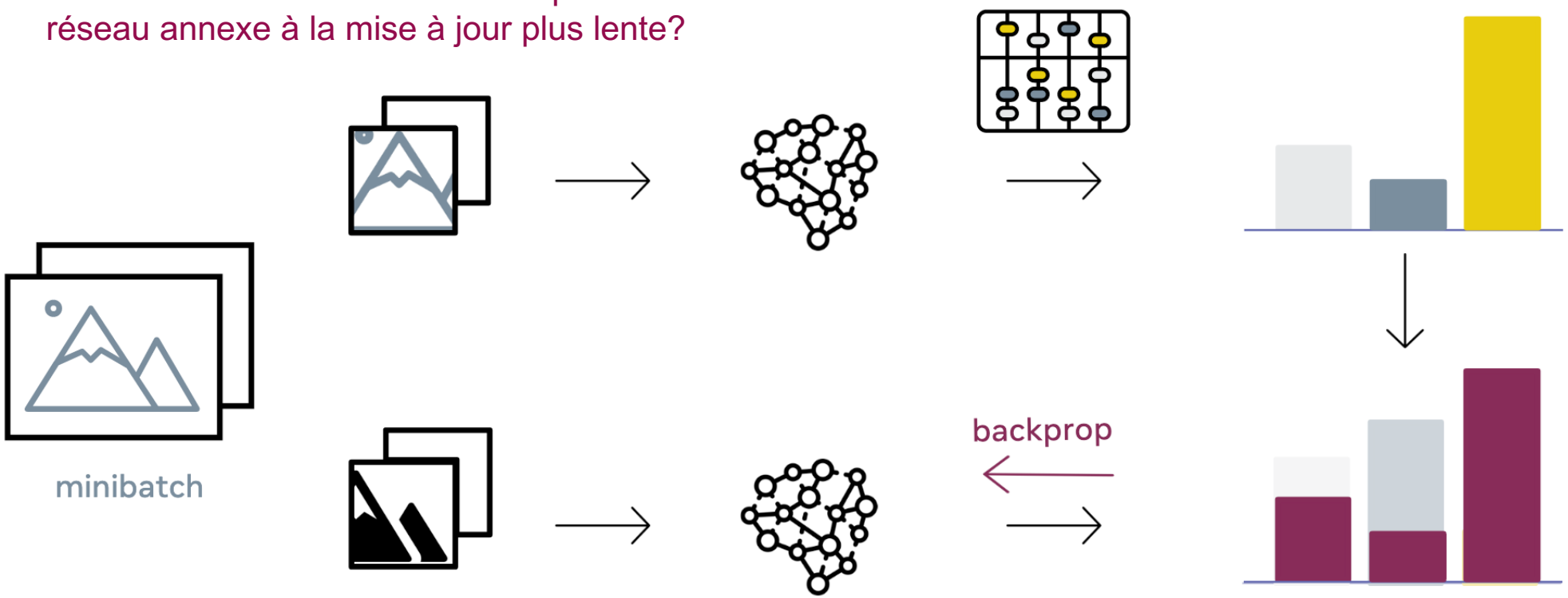


An image is worth 16x16 words: Transformers for image recognition at scale.
Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, Neil Houlsby.
ICLR21

From SwAV to DINO

Mean Teacher – Tarvainen et al. 2017
MoCo - He et al. CVPR 2020
BYOL – Grill et al. NeurIPS 2020

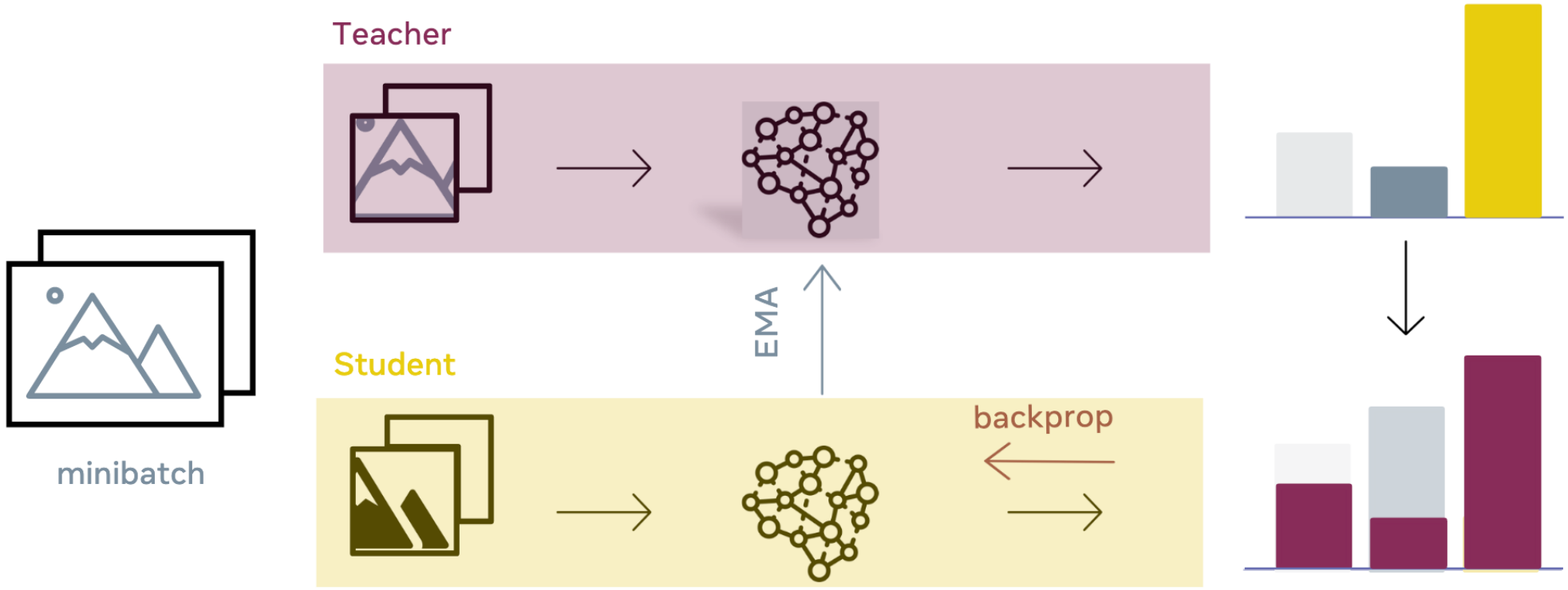
Ci-dessous: SwAV
Question: Pourquoi ne pas s'inspirer des très bons résultats de Moco & BYOL qui utilisent un réseau annexe à la mise à jour plus lente?



[based on Mathilde Caron's slides]

DINO: Self-Distillation with No Labels

Ci-dessous: DINO
Réponse: Revisiter SwAV avec un méchanisme d'auto-distillation

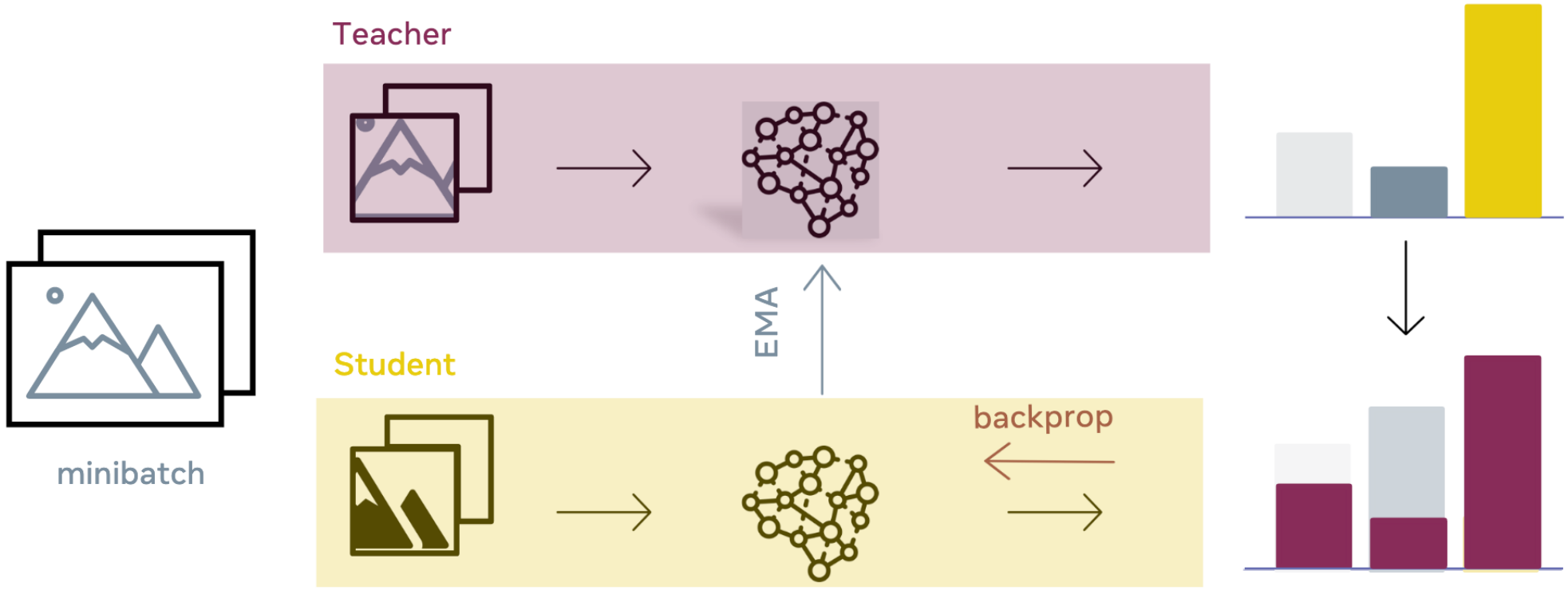


[based on Mathilde Caron's slides]

DINO: Self-Distillation with No Labels

Autres différences:

- Architecture ViT et non basée sur des convolutions
- Nécessité de centrer et “d’aiguiser” (*sharpen*) pour éviter les solutions triviales



[based on Mathilde Caron’s slides]