

Auto-encodeurs et apprentissage auto-supervisé

IA et Multimédia

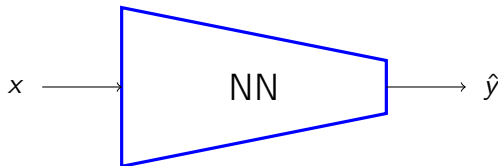
A. Carlier

2026

Apprentissage supervisé

Dans le cadre de l'**apprentissage supervisé**, on cherche à construire des modèles **prédicatifs**. Un modèle doit prédire une classe (*classification*) ou une valeur réelle (*régression*) en réponse à une donnée fournie en entrée.

Le modèle est donc construit pour associer à n'importe quelle entrée x un label prédit $\hat{y}$.

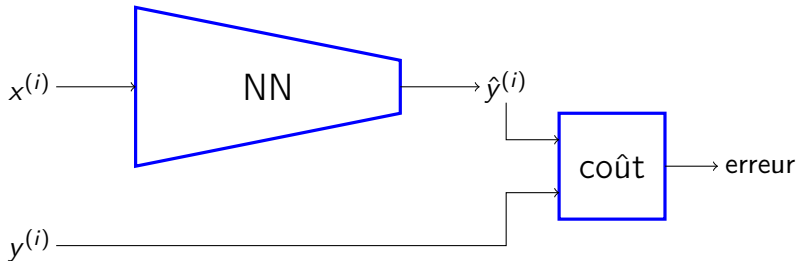


Apprentissage supervisé

Dans le cadre de l'**apprentissage supervisé**, on dispose d'observations (données) et de leurs étiquettes (appelées encore cibles (*target*), catégories ou *labels*) qui constituent un ensemble d'apprentissage. On le note :

$$\mathcal{D} = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots, (\mathbf{x}^{(m)}, y^{(m)})\}.$$

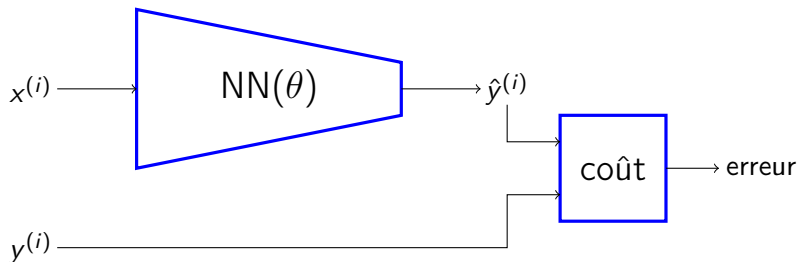
On peut évaluer la performance d'un modèle en comparant sa prédiction $\hat{y}$ au label réel y , au moyen d'une fonction bien choisie que l'on nomme fonction de coût (ou de perte).



Apprentissage supervisé

Les réseaux de neurones sont des fonctions paramétrées (disons par un vecteur θ).

Entraîner un réseau de neurones revient à chercher les meilleurs paramètres θ^* afin de minimiser l'erreur de prédiction commise par le modèle sur l'ensemble d'apprentissage.

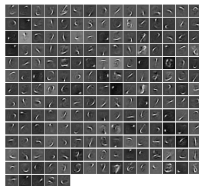
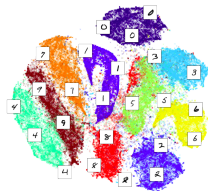


$$\theta^* = \arg \min_{\theta} \frac{1}{m} \sum_{i=1}^m \text{coût}(y^{(i)}, \hat{y}^{(i)})$$

Apprentissage non-supervisé

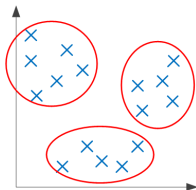
Dans le cadre de l'**apprentissage non supervisé**, on dispose uniquement d'observations

$$\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}.$$

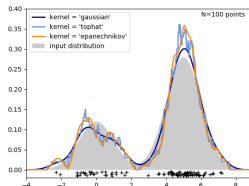


Réduction de dimension

Extraction de caractéristiques



Clustering

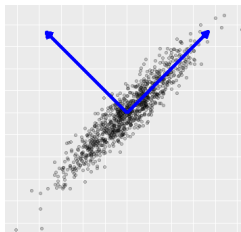


Estimation de densité

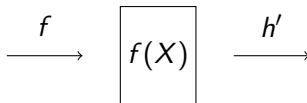
Réduction de dimension : ACP

L'Analyse en Composantes Principales est une technique de **réduction de dimension** dans laquelle on cherche à projeter des données dans un sous-espace de plus faible dimension en maximisant un critère "d'étalement" des données.

La méthode s'appuie sur une diagonalisation de la matrice de variance-covariance. La base de l'espace de projection est obtenue à partir des vecteurs propres associés aux plus grandes valeurs propres.



Apprentissage par représentation

 X Y 

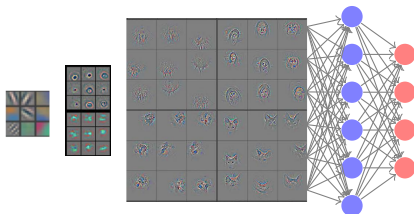
Caractéristiques

Une manière d'interpréter les CNN

 X $f(X)$ h' Y

Extraction de caractéristiques

Prédicteur

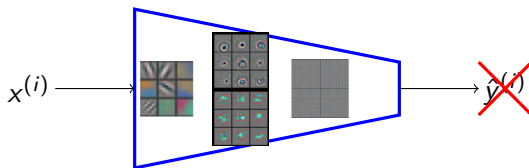


"personne"

On peut interpréter un CNN au regard de l'apprentissage par représentation : la partie convolutive est un extracteur de caractéristiques $f(X)$, et les couches denses de fin constituent notre prédicteur h' .

Apprentissage non-supervisé

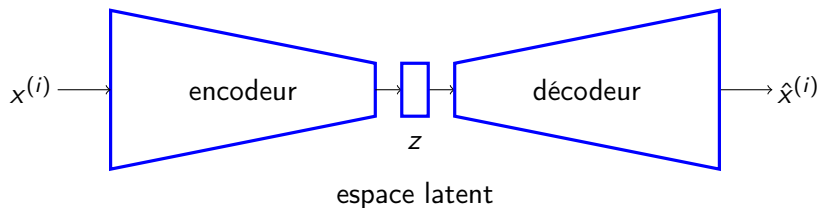
Est-il possible d'apprendre à extraire des caractéristiques, semblables à celles apprises par AlexNet de manière supervisée, mais dans un cadre non supervisé ?



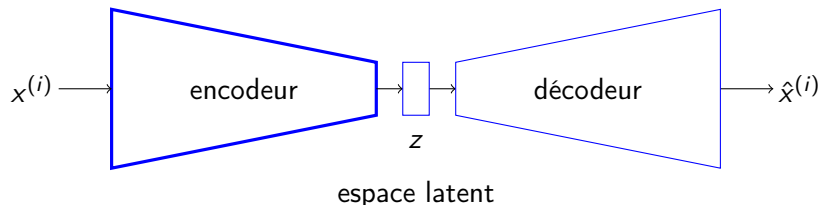
$$\mathcal{D} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\}.$$

Auto-encodeurs

Solution : chercher à prédire les données d'entrée !



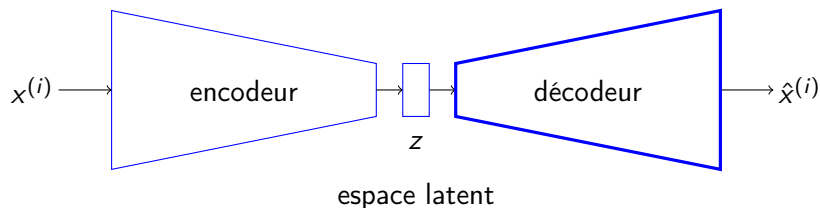
Auto-encodeurs



L'encodeur :

- Extrait les caractéristiques principales (structure) de l'entrée.
- Comprime le signal en réduisant la dimension.
- Conserve suffisamment d'information discriminante pour permettre au décodeur de retrouver l'information initiale de manière convaincante.

Auto-encodeurs

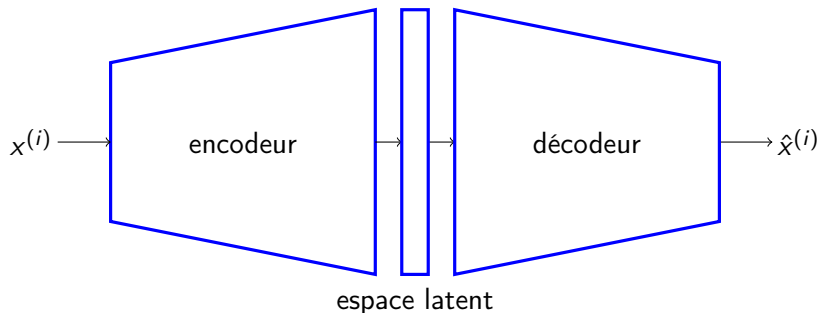


Le décodeur :

- Reconstitue une approximation de l'entrée initiale à partir de sa représentation latente.
- Est une forme de modèle génératif (cf. prochains cours).

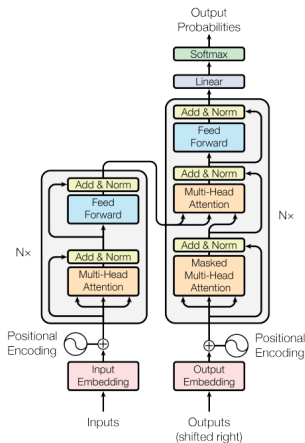
Auto-encodeurs

Il n'y aurait pas vraiment de sens, ni d'intérêt, à construire des auto-encodeurs surconditionnés.



Dans ce cas, l'auto-encodeur n'aurait pas de difficulté à apprendre la fonction identité, ce qui n'a pas d'intérêt.

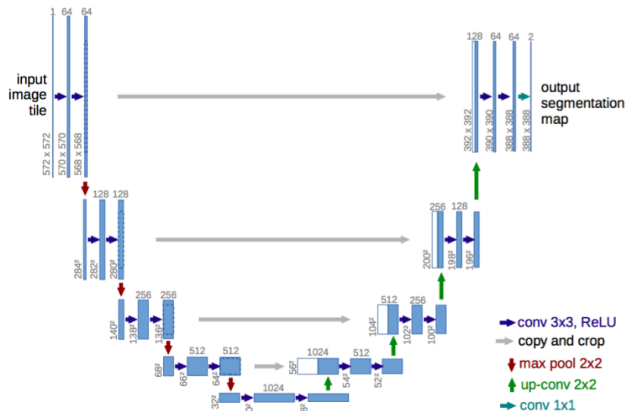
Remarque : *Transformer*



L'architecture du *Transformer* est très proche d'un auto-encodeur.

[Vaswani et al.] Attention is all you need.

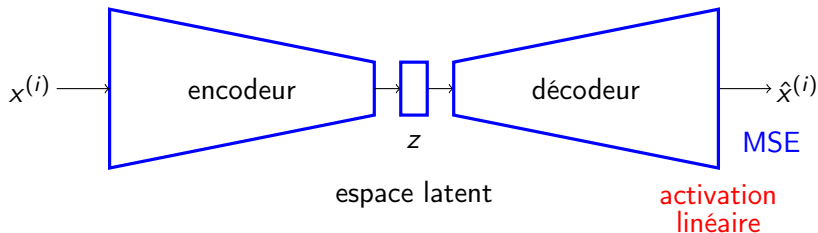
Remarque : U-Net



Le réseau U-Net peut, en altérant légèrement sa forme décrite ci-dessus, être vu comme un auto-encodeur convolutionnel.

[Ronneberger et al.] U-Net : Convolutional Networks for Biomedical Image Segmentation.

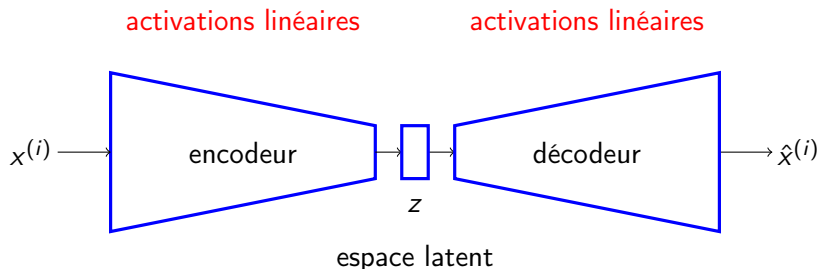
Entraînement d'un auto-encodeur



- Fonction d'activation linéaire sur la couche de sortie.
- Fonction de coût quadratique.
- Fonction objectif à minimiser :

$$J = \sum_{i=1}^n \|x^{(i)} - \hat{x}^{(i)}\|^2$$

Auto-encodeur pour la réduction de dimension



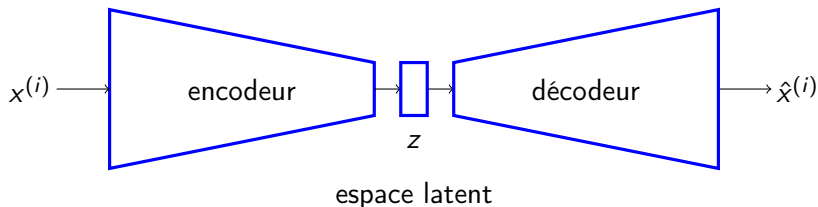
Si toutes les fonctions d'activation des couches cachées de l'encodeur et du décodeur sont linéaires, l'espace latent appris tendra vers l'espace des composantes principales.

En d'autres termes, dans ce cas, l'auto-encodeur est équivalent à l'ACP !

Auto-encodeur pour la réduction de dimension

activations non linéaires

activations non linéaires



A contrario, si les fonctions d'activations de l'encodeur et du décodeur sont non linéaires, nous pouvons voir les auto-encodeurs comme une extension de l'ACP à des projections non-linéaires.

L'encodeur et le décodeur peuvent ainsi bénéficier de la grande capacité de représentation des réseaux de neurones pour projeter les données dans des espaces latents de plus faible dimension, et possédant de meilleures propriétés.

Auto-encodeur pour la réduction de dimension

Attention : il y a un équilibre à trouver entre la capacité de représentation de l'encodeur et du décodeur et la dimension de l'espace latent !

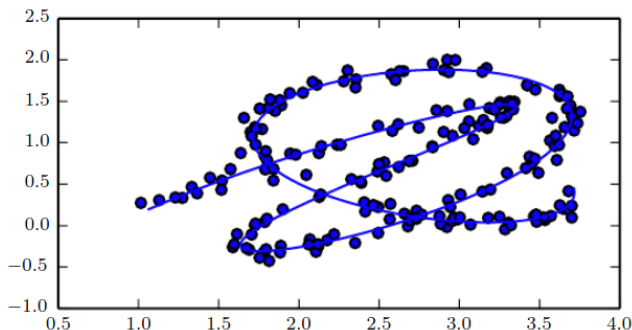
Avec un encodeur et un décodeur de très grande capacité, et un espace latent de dimension 1, on pourrait imaginer un cas dégénéré où l'encodeur apprendrait à associer à chaque donnée $x^{(i)}$ son indice i .

Un tel auto-encodeur n'aurait aucune capacité de généralisation, et serait donc parfaitement inutile.

Variété

Les variétés sont des espaces généralisant les courbes (dimension 1) et surfaces (dimension 2) à des dimensions supérieures.

Les données que nous étudions en Multimédia sont en général, dans des espaces de très haute dimension, concentrées autour de variétés (exemple ci-dessous).



Variétés et auto-encodeurs

On cherche à décrire, dans l'espace latent, la variété sur laquelle vivent nos données, i.e. un sous-espace sur lequel on a une haute probabilité de rencontrer des données d'apprentissage.

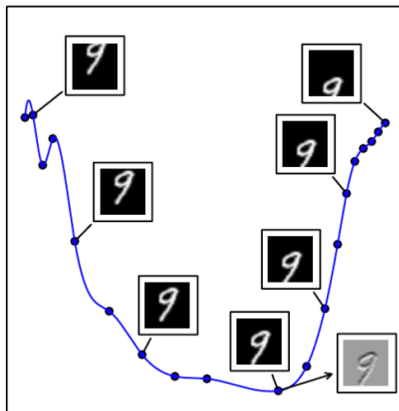


Image de [Goodfellow et al.] Deep Learning

Variétés et auto-encodeurs

Dans l'exemple ci-dessous, deux dimensions de l'espace latent sont illustrées : l'une influe sur l'orientation de la tête, l'autre sur les émotions du visage.



Image de [Kingma et al.] Auto-encoding variational Bayes

Auto-encodeur et détection d'anomalies

Détection d'anomalies dans la base de données Caltech-101 :



0.6372

0.2533

Motorbikes



0.6902

0.2726

Motorbikes



0.5734

0.1781

Watch



0.5947

0.2287

Airplanes

La détection d'anomalies a de nombreuses applications en sécurité, réseau, vidéo-surveillance, maintenance, etc.

Image de [Zhai et al.] Deep Structured Energy Based Models for Anomaly Detection

Auto-encodeurs et régularisation

Les auto-encodeurs sont difficiles à entraîner/calibrer !

- L'espace latent devrait idéalement avoir une dimension semblable à la variété sur laquelle vivent les données ; il est malheureusement impossible de connaître cette dimension à l'avance.
- L'encodeur (et le décodeur) doit avoir une capacité suffisante pour apprendre la fonction qui va de l'espace des données vers l'espace latent (et son inverse).

→ Comme pour les réseaux de neurones classiques, il est intéressant de procéder à une régularisation des auto-encodeurs pour en améliorer l'entraînement.

Auto-encodeur épars

On appelle auto-encodeur épars (*sparse auto-encoder*) un auto-encodeur que l'on entraîne en optimisant la fonction objectif suivante :

$$J = \sum_{i=1}^n ||x^{(i)} - \hat{x}^{(i)}||^2 + \lambda|z|$$

Cette régularisation contraint les variables de l'espace latent ; seules un petit nombre d'entre elles peuvent être actives à la fois.

Les auto-encodeurs épars se sont avérés utile pour l'apprentissage de caractéristiques qui peuvent être ré-utilisées pour la classification.

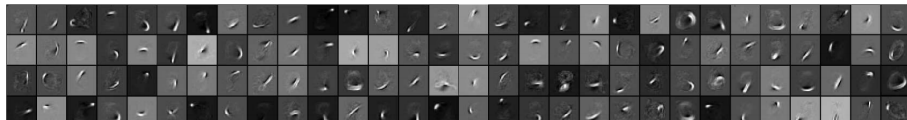
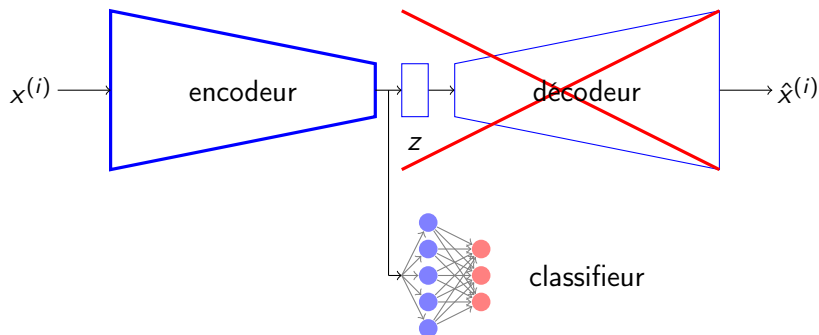


Image de [Makhzani et al.] k-sparse autoencoders

Auto-encodeur et transfer learning



- 1 Entraînement de l'auto-encodeur sur une large base de données non annotée.
- 2 *Fine-tuning* du classifieur sur une petite base de données annotée.

→ une forme d'**apprentissage semi-supervisé** !

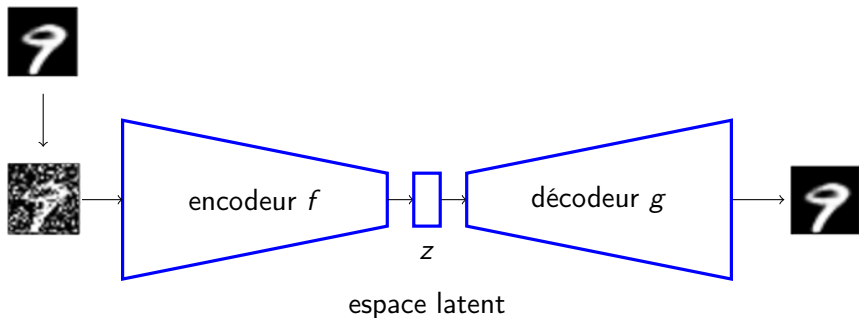
Auto-encodeur épars et recherche d'information

On peut forcer l'espace latent à adopter une représentation épars **binaire**, en affectant une fonction d'activation sigmoïde à la couche latente de l'auto-encodeur.

Cette propriété est désirable en **recherche d'information**, car elle permet de calculer la distance entre la représentation de deux données avec un produit scalaire de deux vecteurs creux, ce qui est très rapide.

[Salakhutdinov et al.] Semantic hashing

Auto-encodeur débruiteur



Si on nomme f (resp. g) la fonction décrite par l'encodeur (resp. le décodeur), un auto-encodeur débruiteur cherche à optimiser la fonction objectif suivante :

$$J = \sum_{i=1}^n \|x^{(i)} - g(f(\tilde{x}^{(i)}))\|^2$$

où $\tilde{x}^{(i)}$ correspond à une donnée $x^{(i)}$ bruitée.

Auto-encodeur débruiteur - quelques applications

Débruitage de signal sonore :

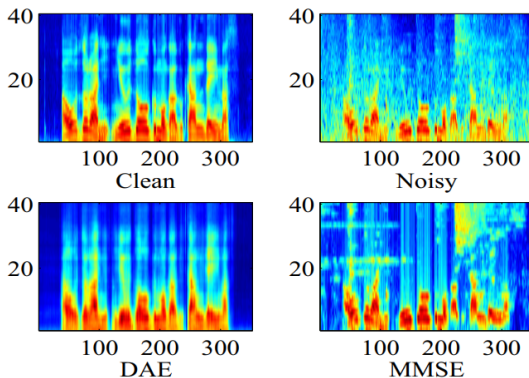


Image de [Lu et al.] Speech Enhancement Based on Deep Denoising Autoencoder

Auto-encodeur débruiteur - quelques applications

Inpainting :

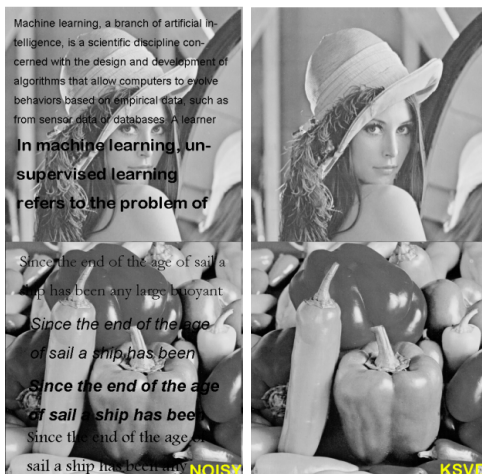


Image de [Xie et al.] Image Denoising and Inpainting with Deep Neural Networks

Une variante : *Context autoencoders*

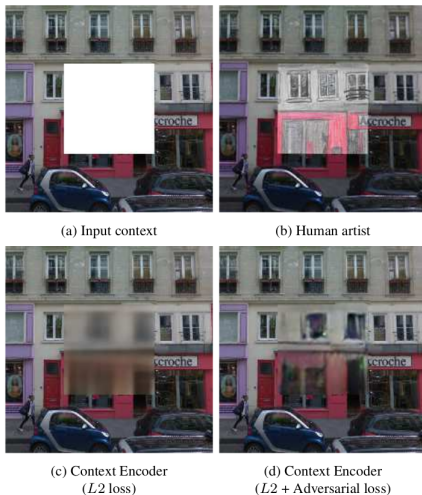
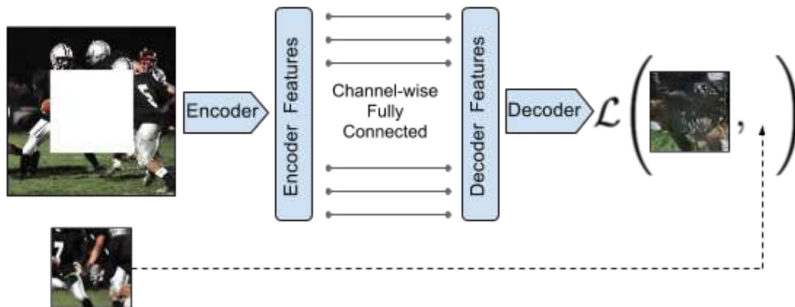


Image de [Pathak et al.] Context Encoders : Feature Learning by Inpainting.

Une variante : *Context autoencoders*



On parle pour cette méthode d'**apprentissage auto-supervisé** (*self-supervised learning*), c'est-à-dire d'un apprentissage où la supervision est assurée par la donnée elle-même. Nous en verrons d'autres exemples dans la suite de ce cours.

Image de [Pathak et al.] Context Encoders : Feature Learning by Inpainting.

Bilan sur les auto-encodeurs

- Les auto-encodeurs forment une classe de réseaux de neurones permettant de faire de **l'apprentissage non supervisé**, ou parfois semi-supervisé.
- Ils sont aussi très utiles pour faire de l'analyse de données, et de la fouille de données, par leur capacité à révéler des variables structurant les données.
- Ils sont utilisés dans de nombreuses applications, et sont d'excellents exemples d'utilisation pratique de l'apprentissage non-supervisé.

Apprentissage auto-supervisé

On parle d'**apprentissage auto-supervisé** (*self-supervised learning*) lorsque la supervision est assurée par la donnée elle-même.

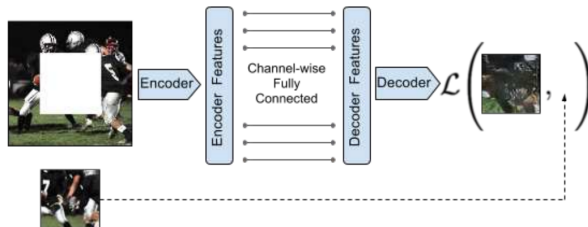


Image de [Pathak et al.] Context Encoders : Feature Learning by Inpainting.

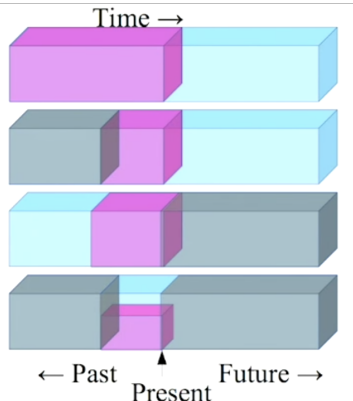
Apprentissage auto-supervisé

On peut regrouper les algorithmes d'apprentissage auto-supervisé en 2 familles de méthodes :

- les méthodes qui cherchent à prédire ou restaurer une partie de la donnée à partir d'une autre partie (*context autoencoders*, auto-encodeurs débruiteurs, etc.),
- et les méthodes **contrastives**.

Tâches prétextes

- ▶ Predict any part of the input from any other part.
- ▶ Predict the **future** from the **past**.
- ▶ Predict the **future** from the **recent past**.
- ▶ Predict the **past** from the **present**.
- ▶ Predict the **top** from the **bottom**.
- ▶ Predict the occluded from the visible
- ▶ **Pretend there is a part of the input you don't know and predict that.**

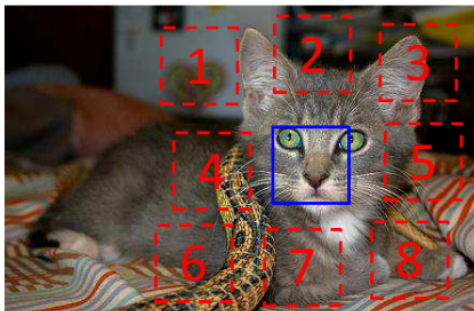


Slide: LeCun

Image de Yann LeCun, *Self-supervised learning : could machines learn like humans ?*

https://www.youtube.com/watch?v=7IOQt7GALV&ab_channel=EPFL

Apprentissage par prédiction du contexte



↓

$$X = (\text{patch}_1, \text{patch}_2); Y = 3$$

Objectif : prédiction de la position relative de deux patches extraits de l'image.

Image de [Doersch et al.] Unsupervised Visual Representation Learning by Context Prediction.

Apprentissage par prédiction du contexte

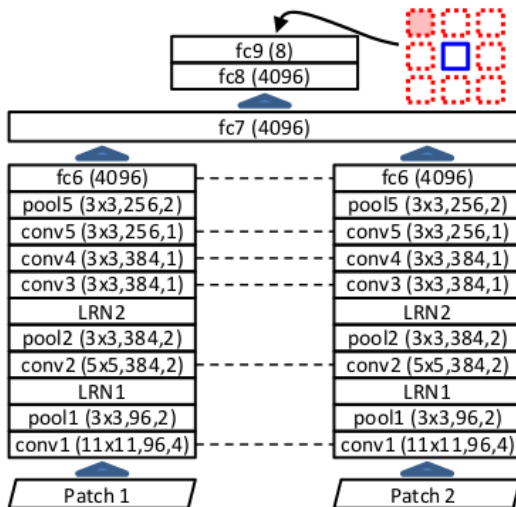


Image de [Doersch et al.] Unsupervised Visual Representation Learning by Context Prediction.

Apprentissage par prédiction de l'orientation

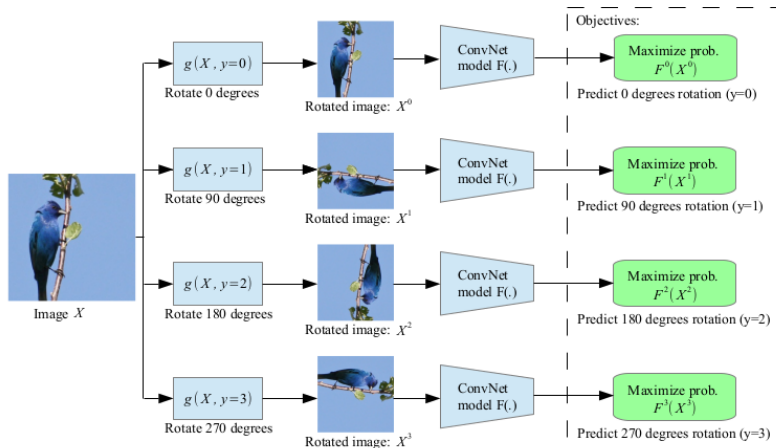
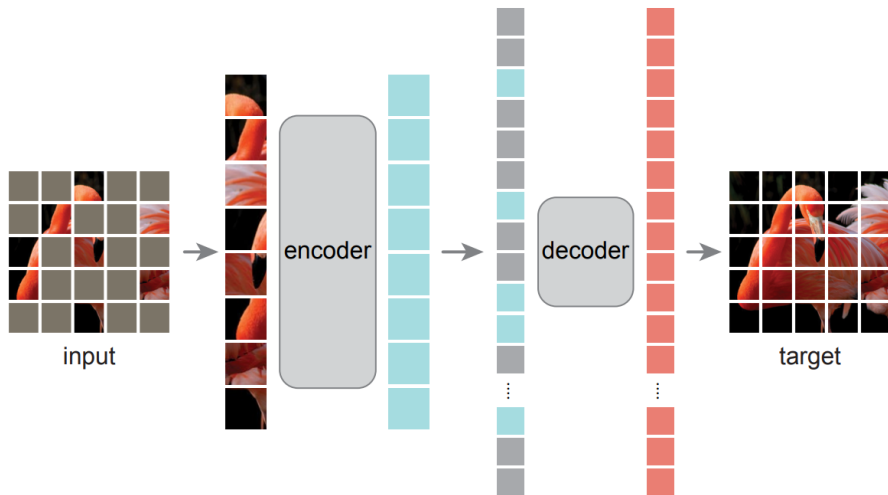


Image de [Gidaris et al.] Unsupervised Representation Learning By Predicting Image Rotations.

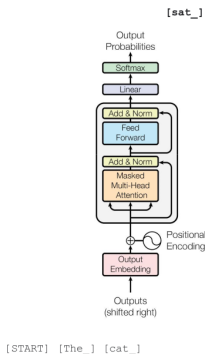
Une variante des *context autoencoders*



[He 2021] Masked Autoencoders Are Scalable Vision Learners

Apprentissage par complétion de texte

Decoder-only GPT



Encoder-only BERT

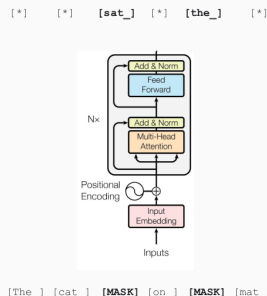


Image de Lucas Beyer.

Apprentissage Contrastif : principe

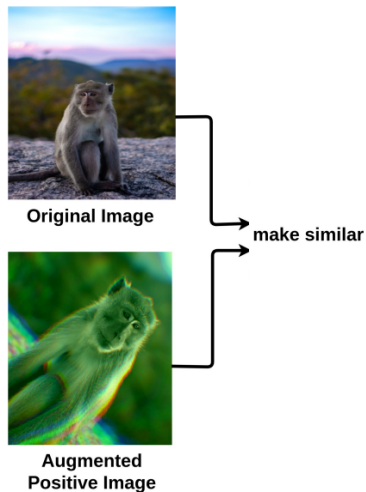
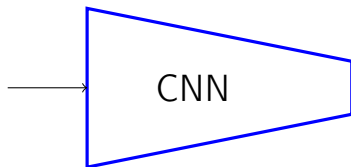
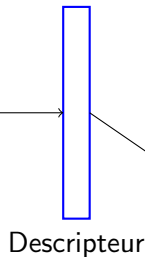
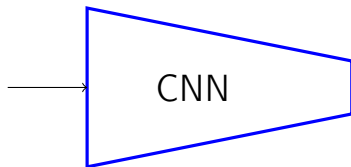


Image de [Jaiswal et al.] A Survey on Contrastive Self-Supervised Learning.

Apprentissage Contrastif : principe



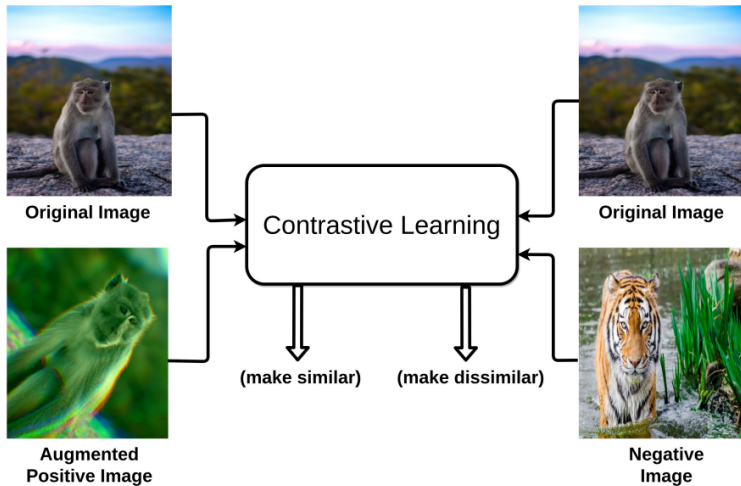
Encodeur



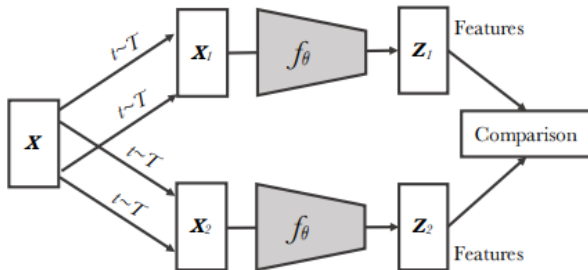
Perte

Apprentissage Contrastif

En procédant de cette manière, on risque d'atteindre une solution triviale (descripteur unique pour toutes les images), c'est pourquoi on introduit des exemples négatifs.



Apprentissage contrastif standard (SimCLR)



Utilisation de réseaux siamois pour apprendre des descripteurs z_i proches pour des images similaires et éloignés pour des images différentes.

[Chen et al.] A Simple Framework for Contrastive Learning of Visual Representations

Image de [Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.

Apprentissage Contrastif - augmentation de données

Le succès de ces algorithmes repose beaucoup sur l'augmentation de données utilisée.

Image augmentation	Top-1 SimCLR (repro)
Baseline	67.9
Remove flip	67.3
Remove blur	65.2
Remove color (jittering and grayscale)	45.7
Remove color jittering	63.7
Remove grayscale	61.9
Remove blur in $\mathcal{T}'$	67.5
Remove solarize in $\mathcal{T}'$	67.7
Remove blur and solarize in $\mathcal{T}'$	67.4
Symmetric blurring/solarization	68.1
Crop only	40.3 $\pm$ 0.3
Crop and flip only	40.2
Crop and color only	64.2
Crop and blur only	41.7

Image de [Grill et al.] Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

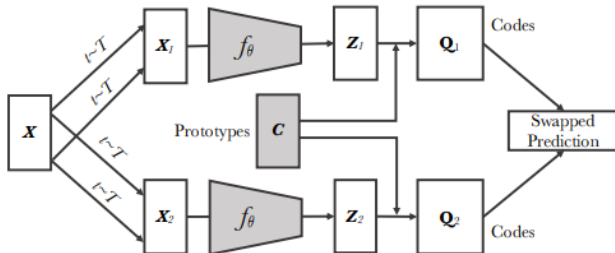
Apprentissage Contrastif - taille des batches

Ces algorithmes nécessitent de grandes tailles de batch size pour échantillonner suffisamment d'exemples négatifs et éviter la divergence.

Batch size	Top-1 SimCLR (repro)
4096	67.9
2048	67.8
1024	67.4
512	66.5
256	64.3 $\pm$ 2.1
128	63.6
64	59.2 $\pm$ 2.9

Image de [Grill et al.] Bootstrap Your Own Latent A New Approach to Self-Supervised Learning

Apprentissage contrastif par cluster (SwAV)



Avec une idée similaire à SimCLR, cette méthode passe par un clustering intermédiaire dans l'espace de descripteurs et cherche à assigner le même cluster à des vues (augmentations) différentes de la même image.

Image de [Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.

Résultats

Method	Arch.	Param.	Top1
Supervised	R50	24	76.5
Colorization [65]	R50	24	39.6
Jigsaw [46]	R50	24	45.7
NPID [58]	R50	24	54.0
BigBiGAN [15]	R50	24	56.6
LA [68]	R50	24	58.8
NPID++ [44]	R50	24	59.0
MoCo [24]	R50	24	60.6
SeLa [2]	R50	24	61.5
PIRL [44]	R50	24	63.6
CPC v2 [28]	R50	24	63.8
PCL [37]	R50	24	65.9
SimCLR [10]	R50	24	70.0
MoCov2 [11]	R50	24	71.1
SwAV	R50	24	75.3

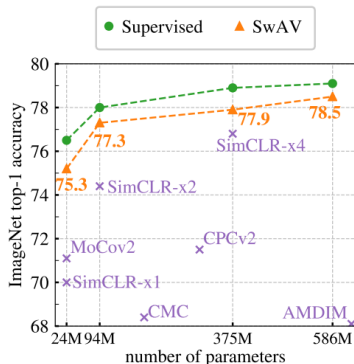
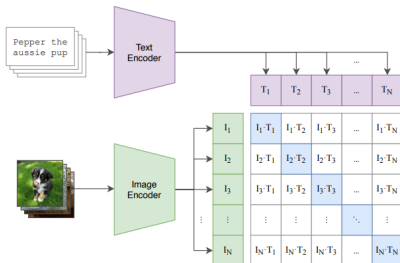


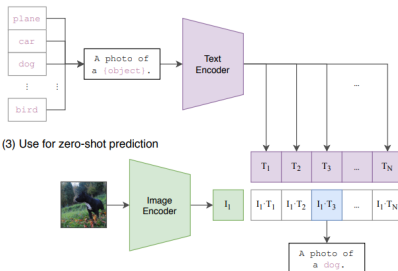
Image de [Caron et al.] Unsupervised Learning of Visual Features by Contrasting Cluster Assignments.

Modèles de fondation : CLIP

(1) Contrastive pre-training



(2) Create dataset classifier from label text



Apprentissage contrastif multi-modal et bonnes performances en Zero-shot Transfer Learning

Image de [Radford et al.] Learning Transferable Visual Models From Natural Language Supervision.

Et en TP :

Nous manipulerons les auto-encodeurs sur la base de données MNIST :

- 1 Entraînement d'un auto-encodeur simple
- 2 Lien entre dimension de l'espace latent et erreur de reconstruction.
- 3 Lien entre capacité de l'auto-encodeur et erreur de reconstruction.
- 4 Application de *morphing* via une interpolation dans l'espace latent.

