Apprentissage continu de représentations visuelles

ENSIMAG 2023-2024

Karteek Alahari & Diane Larlus

jeudi 26 octobre 2023

Apprentissage continu de représentations visuelles

- Site Web: https://project.inria.fr/bigvisdata/

- Intervenants:
  - **Karteek Alahari**, Directeur de Recherche @ INRIA Grenoble <karteek.alahari@inria.fr>
  - **Diane Larlus**, Principal Scientist @ NAVER LABS <diane.larlus@naverlabs.com>
- 12 x 1h30 = 18h de cours

- **Évaluation**
  - Examen final écrit
  - Présentations et Quizz sur des articles de recherche

Détails sur le site web

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

Aperçu du cours – Première séance

Cours 1: introduction

- Définition ‘Big Data’, applications, la chaine, lien avec la recherche d’information par le contenu
- List des thèmes abordés: apprentissage supervisé, auto-supervisé, adaptation de domaine, apprentissage continu, problèmes en vidéos
- Définition de l’apprentissage supervisé et de ses étapes majeures: collection des données, choix d’une représentation, choix d’un modèle, apprentissage, sélection d’un modèle
- Collection des données: difficulté et ambiguïté de l’annotation d’images, exemples de grandes bases d’images standard
- Décomposition de l’erreur
- Représentation des données: approches traditionnelles: représentations puis apprentissage, *mid-level representations*, apprentissage de bout-en-bout
- Définition de l’apprentissage profond

Aperçu du cours – Aujourd’hui

Quelques notions utiles

Comment créer des représentations visuelles ?

- Entrainer un modèle sur de grandes quantités de données annotées
- Utiliser ce modèle pour produire une représentation vectorielle pour chaque image proposée en entrée du modèle

Comment réutiliser des représentations visuelles ?

- Utiliser des représentations visuelles directement, optionnellement apprendre un modèle de décision par-dessus
- Utiliser le modèle précédent comme point de départ pour l’apprentissage et l’ajuster (*fine-tuning*) pour la tâche cible
- Appliquer une méthode d’adaptation, par exemple: adaptation de domaine

Comment créer des représentations visuelles quand on n'a pas d’annotations ?

- Introduction à l’apprentissage auto-supervisée

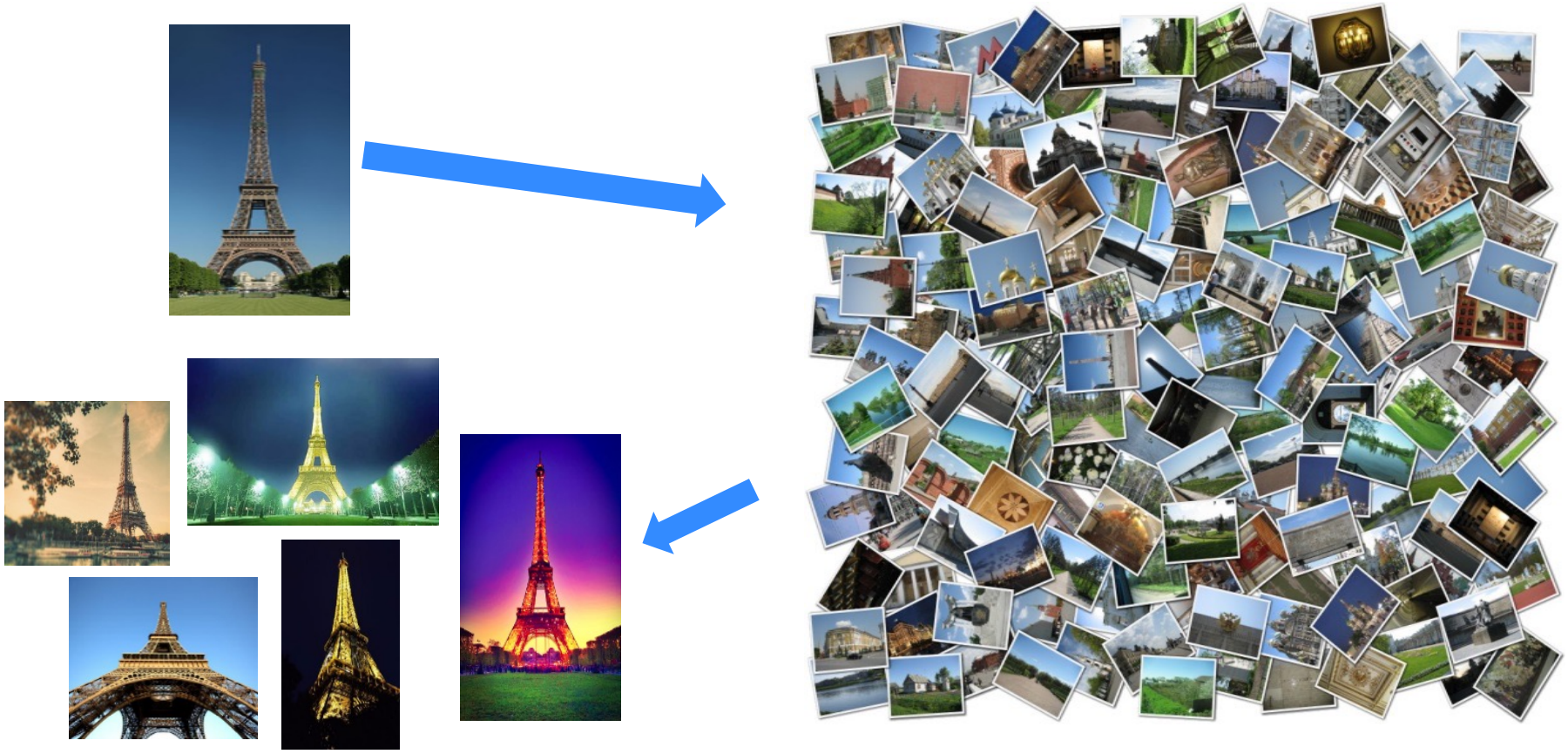
Quelques tâches de vision par ordinateur

2023-2024

Exemples de processus automatisables: **la recherche d'objets / d'instances d'objets**

Principe: recherche d'images par similarité

- Etant donné une image, retrouver les images similaires dans une grande base visuelle



Recherche d'images par similarité

- Utilise la notion de proximité, de similarité, ou de distance entre images
- Généralement, la requête est exprimée sous la forme d'un ou de plusieurs **vecteurs** dans un espace multidimensionnel.
  - définition d'une distance (ou mesure de similarité) sur cet espace
  - recherche des objets dont la distance est minimale
- Les **vecteurs** sont extraits du contenu de l'image

Recherche par le contenu

ou CBIR: *content-based information retrieval*

- Terminologies possibles pour cette tâche
  - *Content-based image retrieval*
  - *Image retrieval*
  - *Image search*
  - *Object search*

Visual Search - Principle

Variation d’apparence d’une **instance** d’objet donné

Exemples de processus automatisables: **la reconnaissance d'objet**

1) Catégorisation d'image

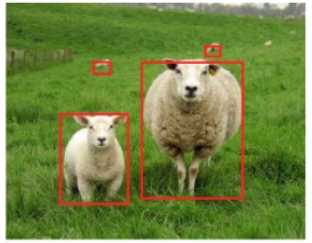
- Catégorie principale associée à l'image, ou réponse oui/non à une liste de catégories connues à l'avance

Sheep ?

yes

2) Détection d'objet

- Boite englobante pour toutes les instances d'une catégorie d'intérêt



Sheep ?

4

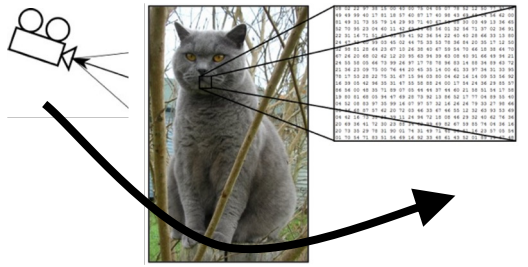
3) Segmentation d'objet, segmentation sémantique

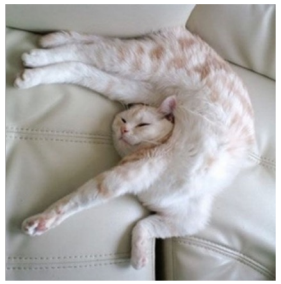
- Localisation précise des objets au niveau du pixel

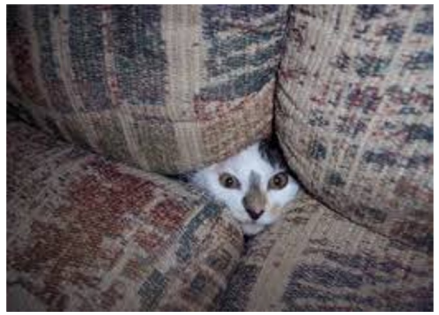
Sheep ?

Difficultés de la modélisation des **catégories** d'objet

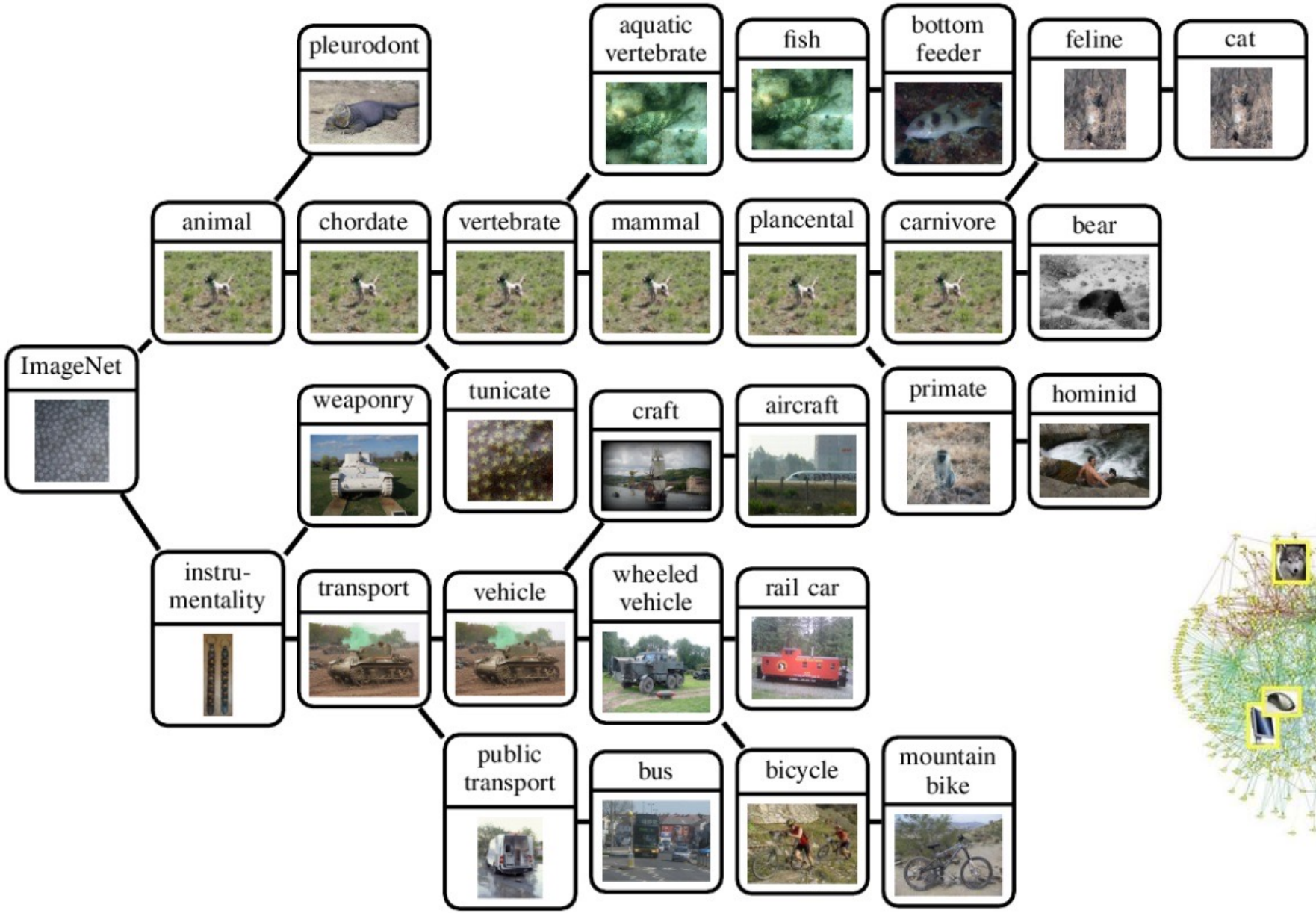
- Illumination, ombres
- Orientation et pose
- Fond texturé, distracteurs
- Occultations
- Variations intra-classe







ImageNet

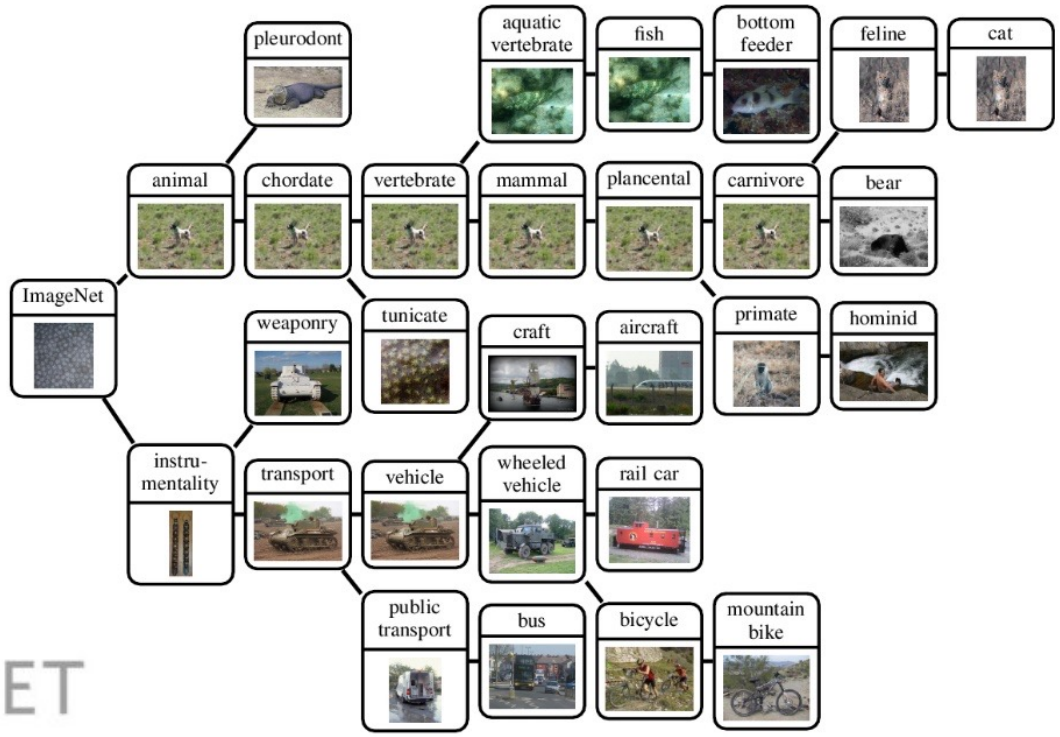


IMAGENET

ImageNet

- ImageNet challenge:
  - Task: Categorize images into 1000 classes

IMAGENET

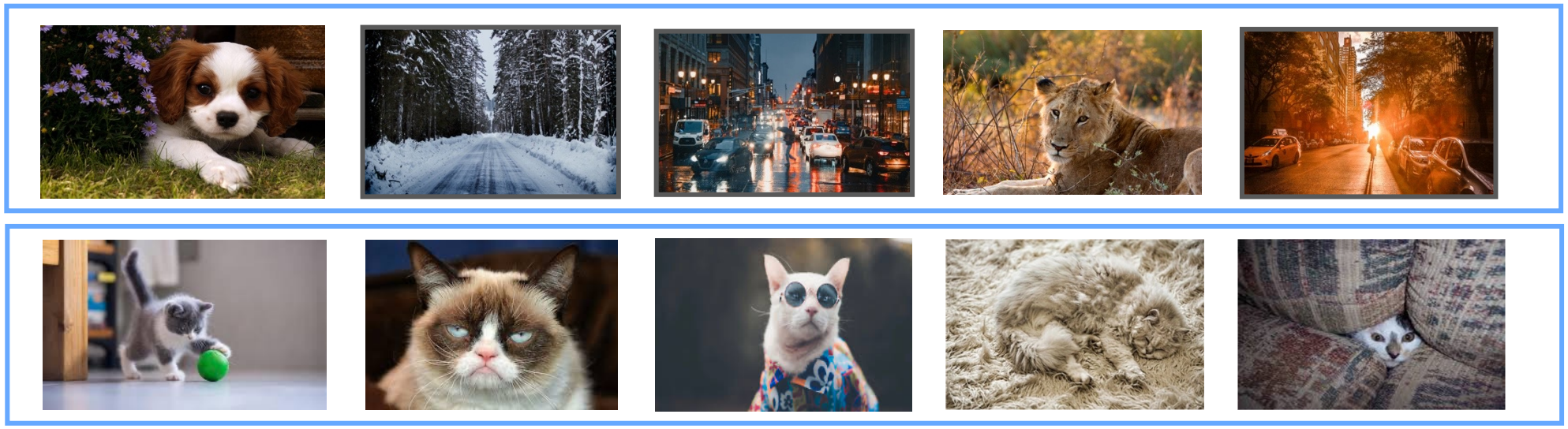


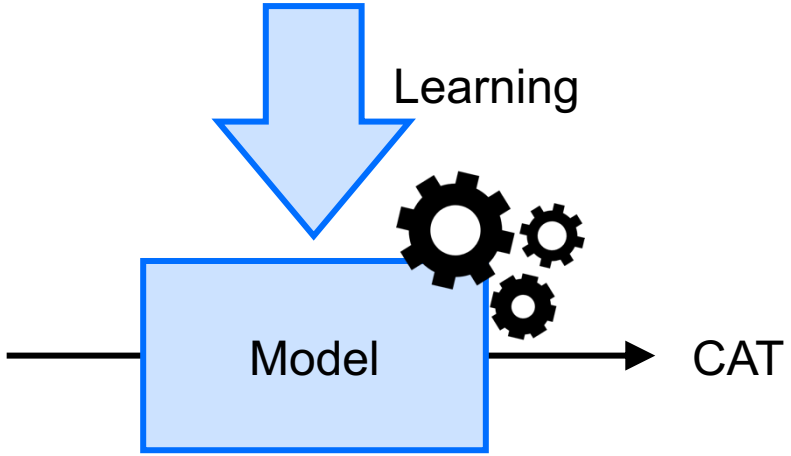
Représentations visuelles

Apprentissage continu de représentations visuelles

2023-2024

Computer vision over the last two decades





Large image collections to train from

4 categories

2005 Pascal VOC

Bicycle

Car

Motorbike

Person

Large image collections to train from

4 categories

2005 Pascal VOC

Ambiguïté de l'annotation d'images

- Difficile de se mettre d'accord sur les annotations
- Exemple: Instructions pour la création d'une vérité terrain pour la compétition PASCAL 2009

What to label	*All objects of the defined categories*, unless: you are unsure what the object is. the object is very small (at your discretion). less than 10-20% of the object is visible. If this is not possible because too many objects, mark image as bad.
Viewpoint	Record the viewpoint of the 'bulk' of the object e.g. the body rather than the head. Allow viewpoints within 10-20 degrees. If ambiguous, leave as 'Unspecified'. Unusually rotated objects e.g. upside-down people should be left as 'Unspecified'.
Bounding box	Mark the bounding box of the visible area of the object (*not* the estimated total extent of the object). Bounding box should contain all visible pixels, except where the bounding box would have to be made excessively large to include a few additional pixels (<5%) e.g. a car aerial.
Truncation	If more than 15-20% of the object lies outside the bounding box mark as Truncated. The flag indicates that the bounding box does not cover the total extent of the object.
Occlusion	If more than 5% of the object is occluded within the bounding box, mark as Occluded. The flag indicates that the object is not totally visible within the bounding box.

➢ Malgré ces instructions, on observe tout de même des incohérences dans les annotations

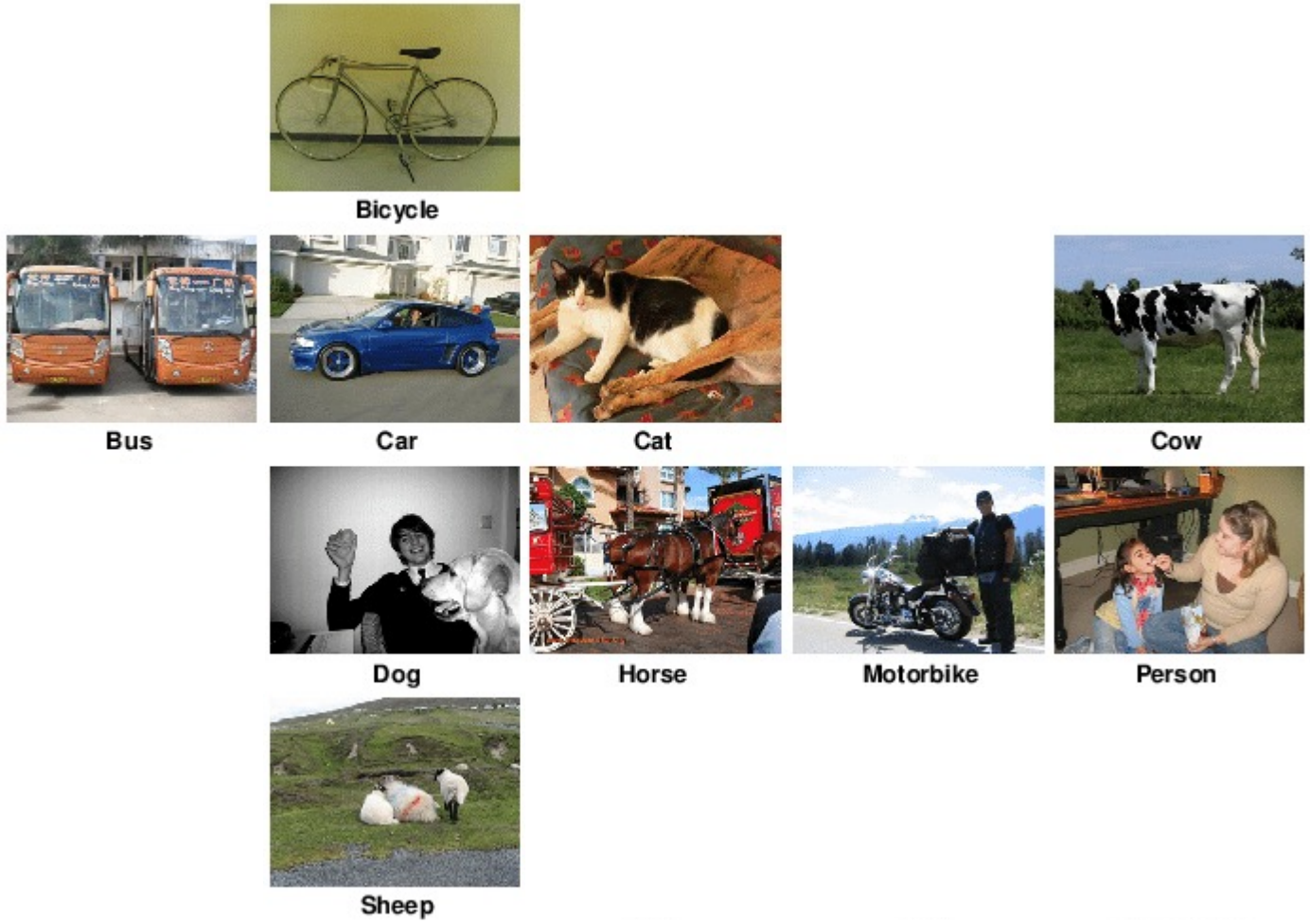
Person

Large image collections to train from

10 categories

2005 Pascal VOC

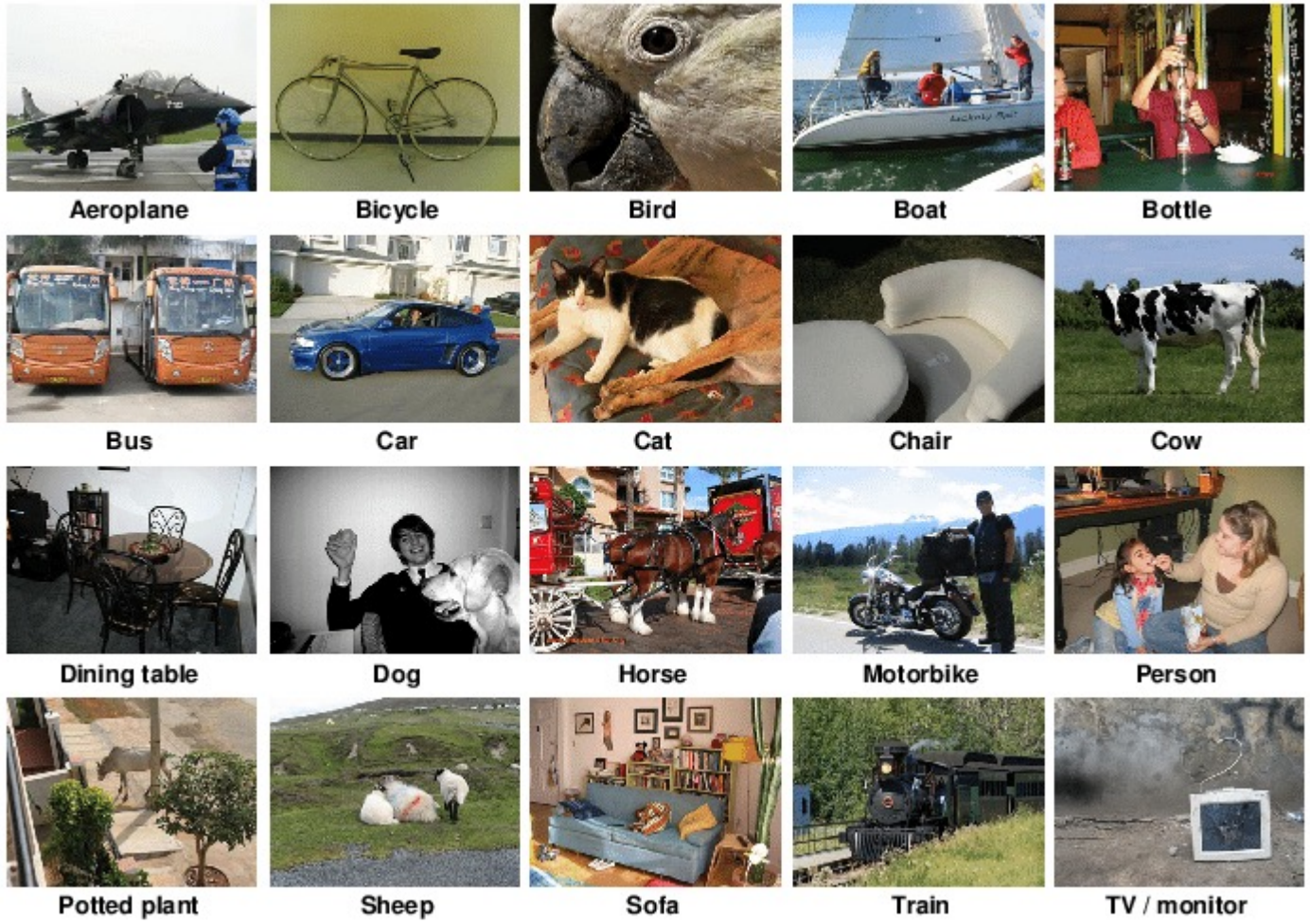
2006 Pascal VOC



Large image collections to train from

20 categories

- 2005 Pascal VOC
- 2006 Pascal VOC
- 2007 Pascal VOC



Large image collections to train from

1000 categories
1 million images

2005 Pascal VOC
2006 Pascal VOC
2007 Pascal VOC

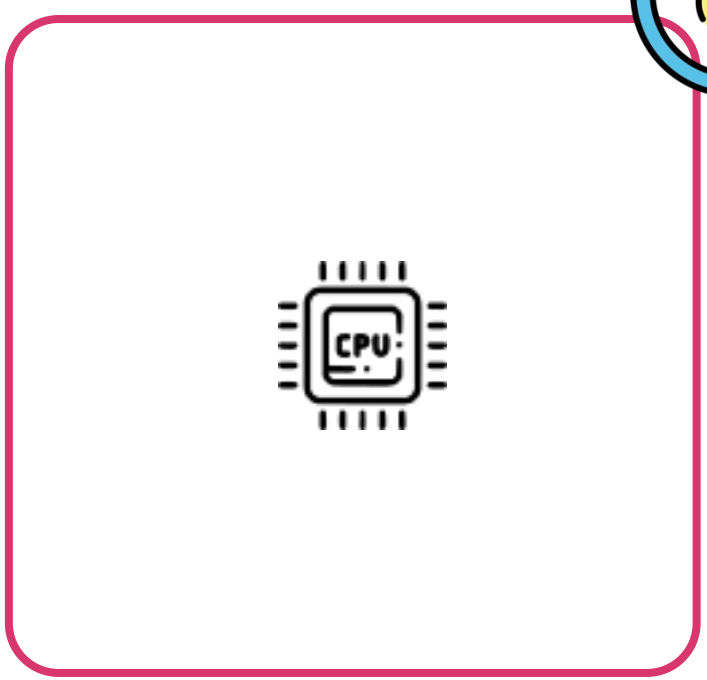
2010 ImageNet

- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET

- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET



2010 Fisher Vectors

2011 Fisher Vectors

[Perronnin et al. CVPR10]

- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET

2010 Fisher Vectors

2011 Fisher Vectors

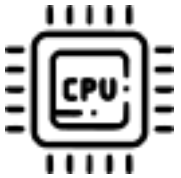
[Perronnin et al. CVPR10]

Traditional Approaches for Recognition

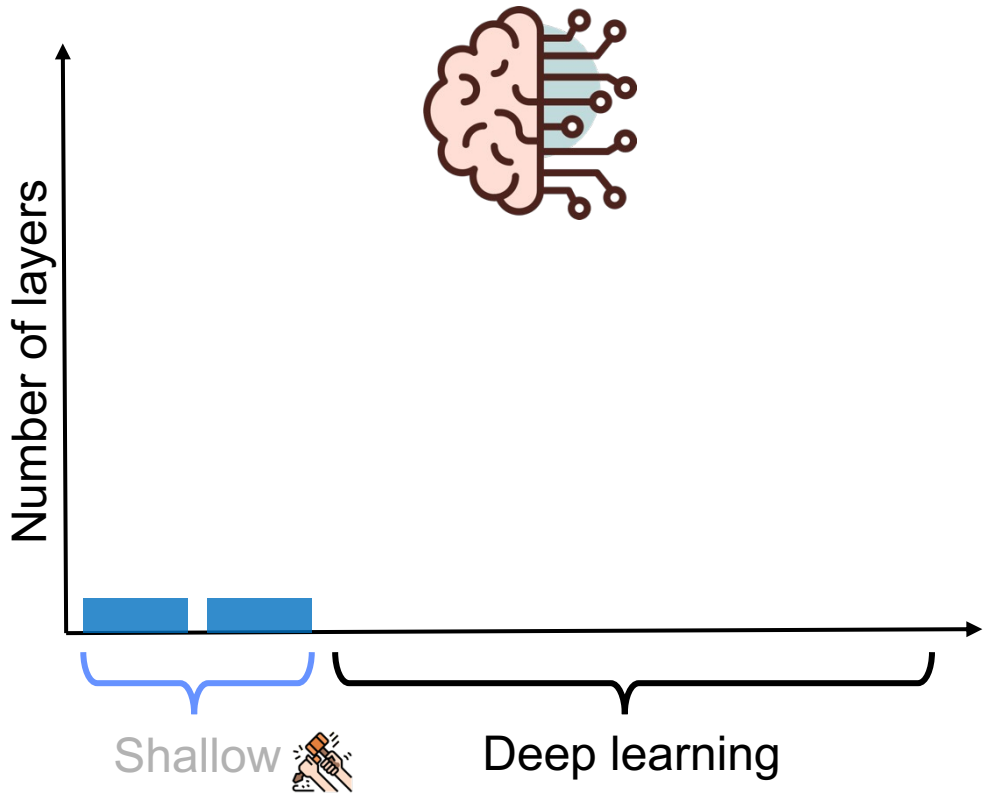
VISION

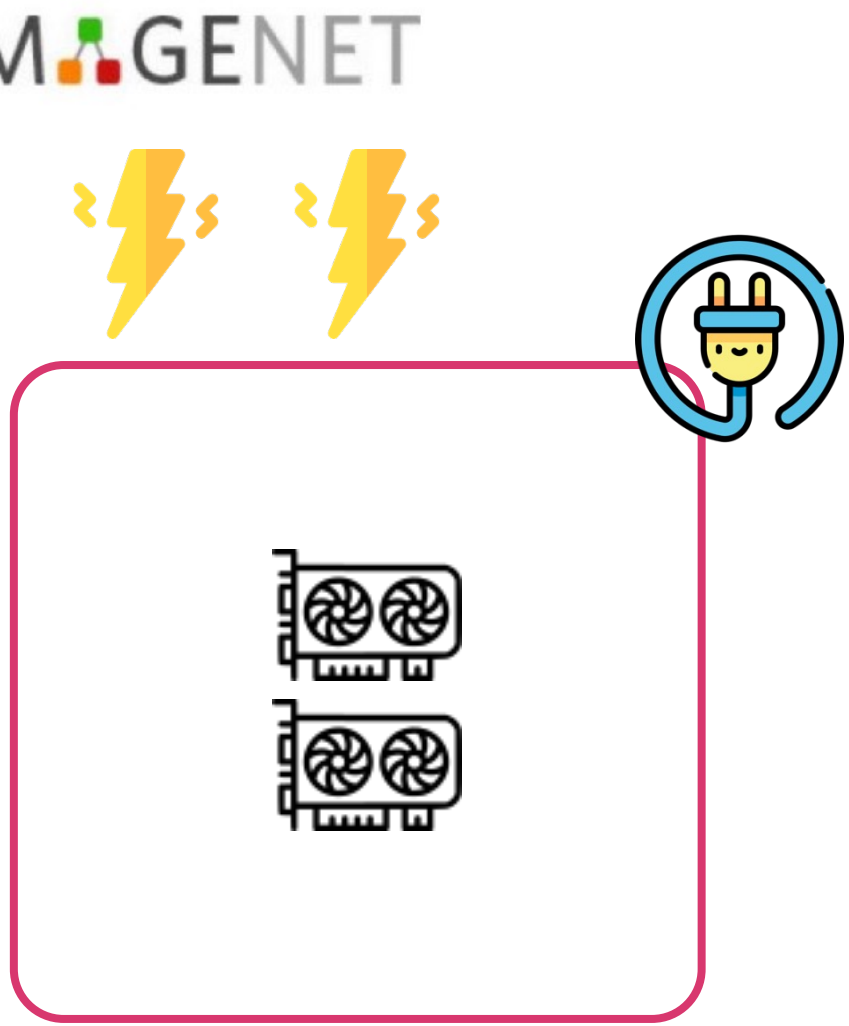
SIFT/HOG → K-Means/ pooling → classifier → "car"

fixed | unsupervised | supervised



- [x] Large image collections to train from
- [x] Deeper models with more parameters





- [x] Large image collections to train from
- [x] Deeper models with more parameters

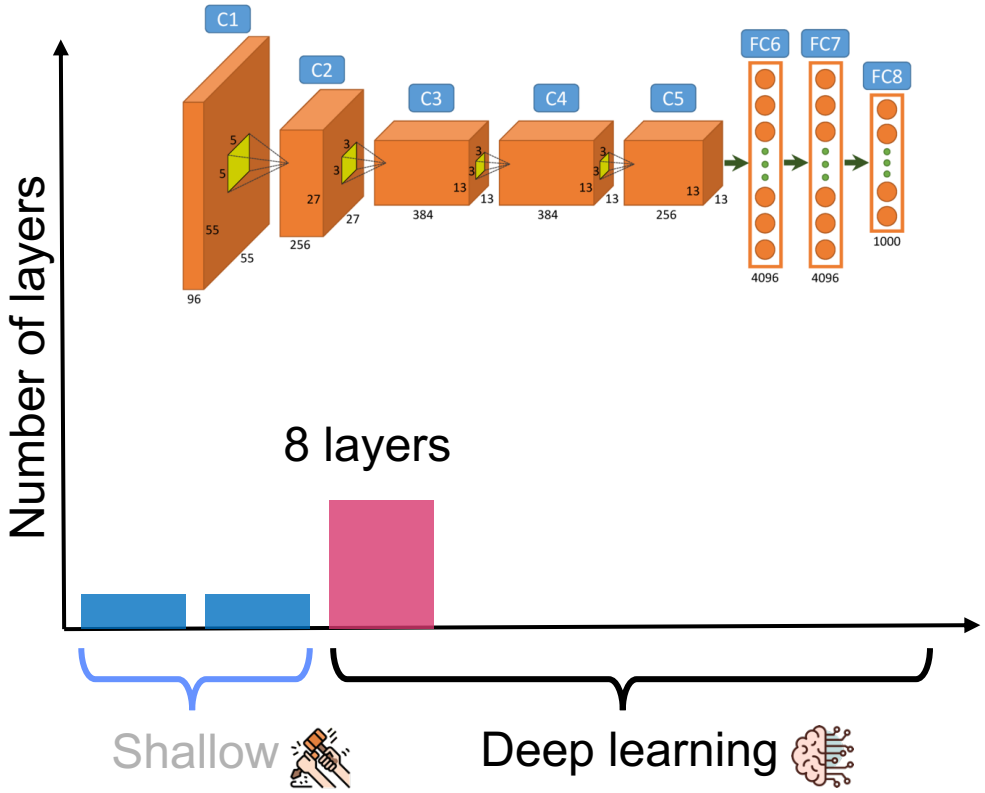
IMAGENET

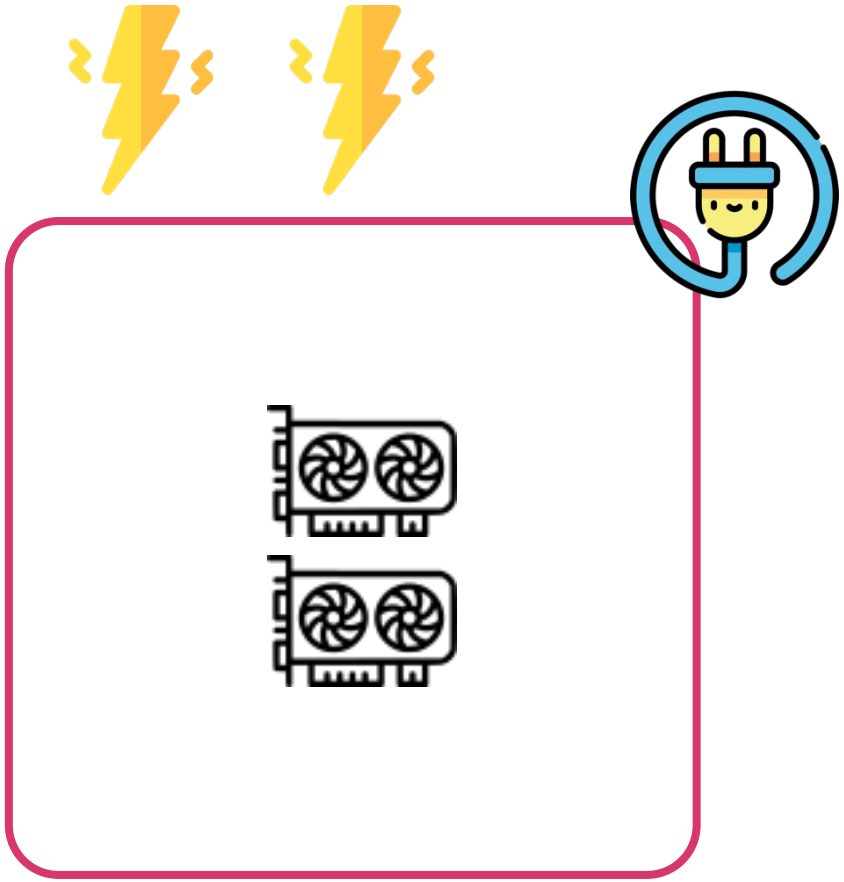
2010 Fisher Vectors

2011 Fisher Vectors

2012 AlexNet

[Krizhevsky@NeurIPS12]

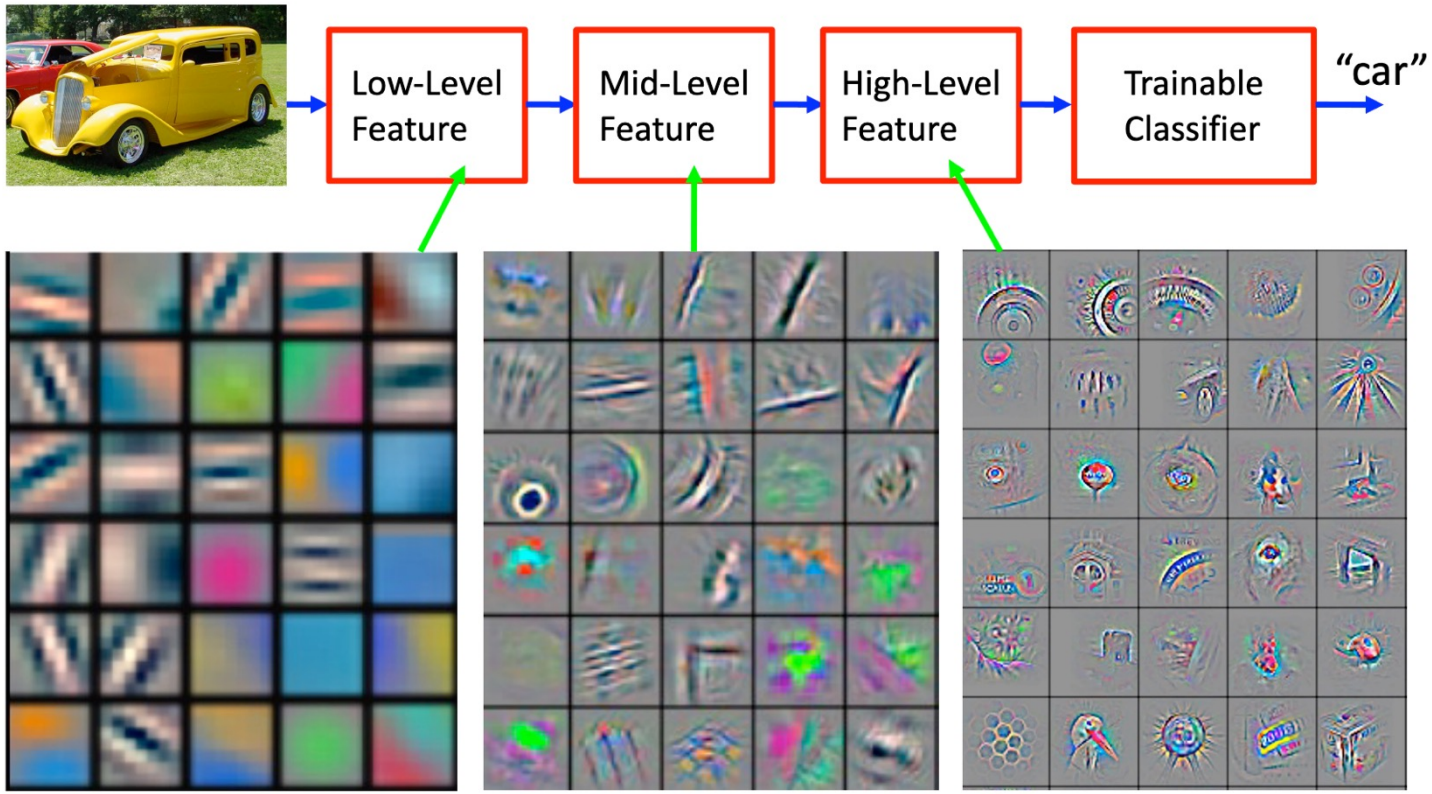




- [x] Large in
- [x] Deeper

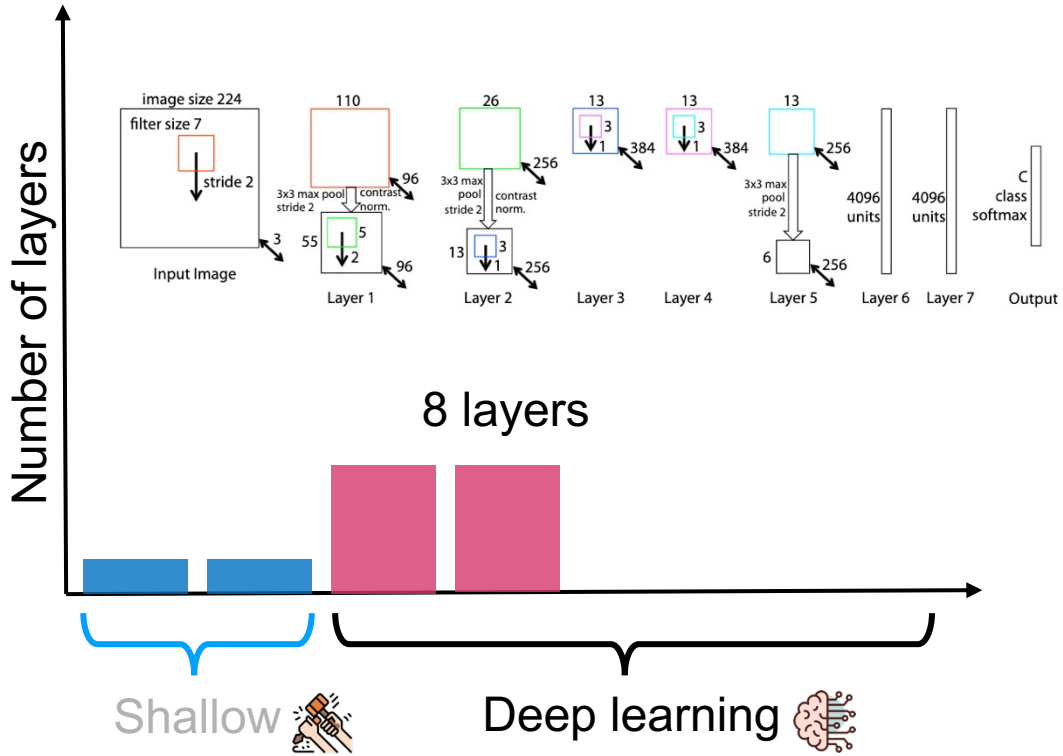
2010 Fishe
2011 Fishe
2012 AlexN
2013 ZF
[Zeiler@ECCV1

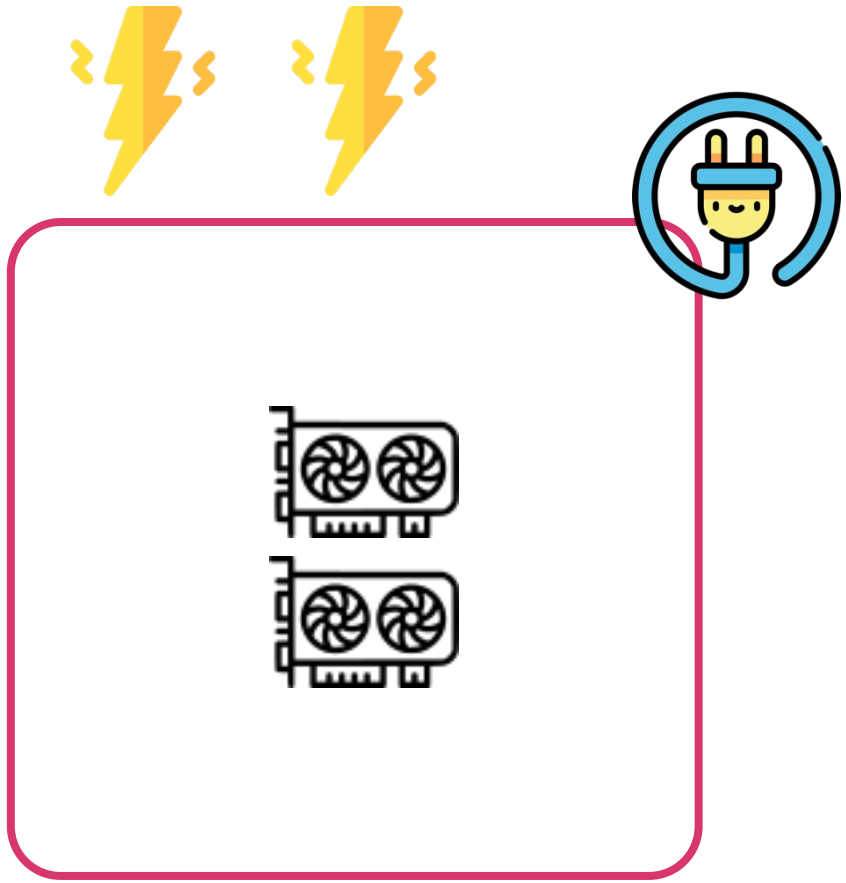
Deep Learning = Hierarchical Compositionality



Feature visualization of convolutional net trained on ImageNet from [Zeiler & Fergus 2013]

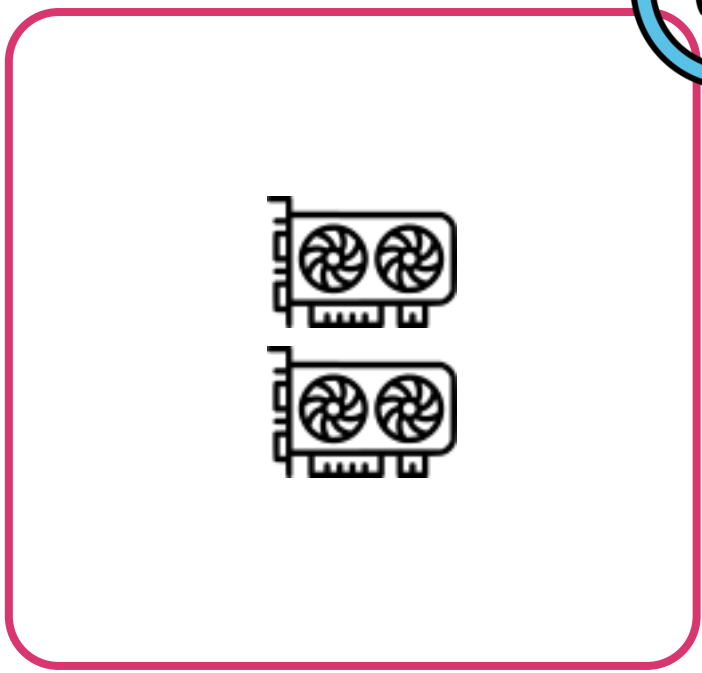
- [x] Large image collections to train from
- [x] Deeper models with more parameters





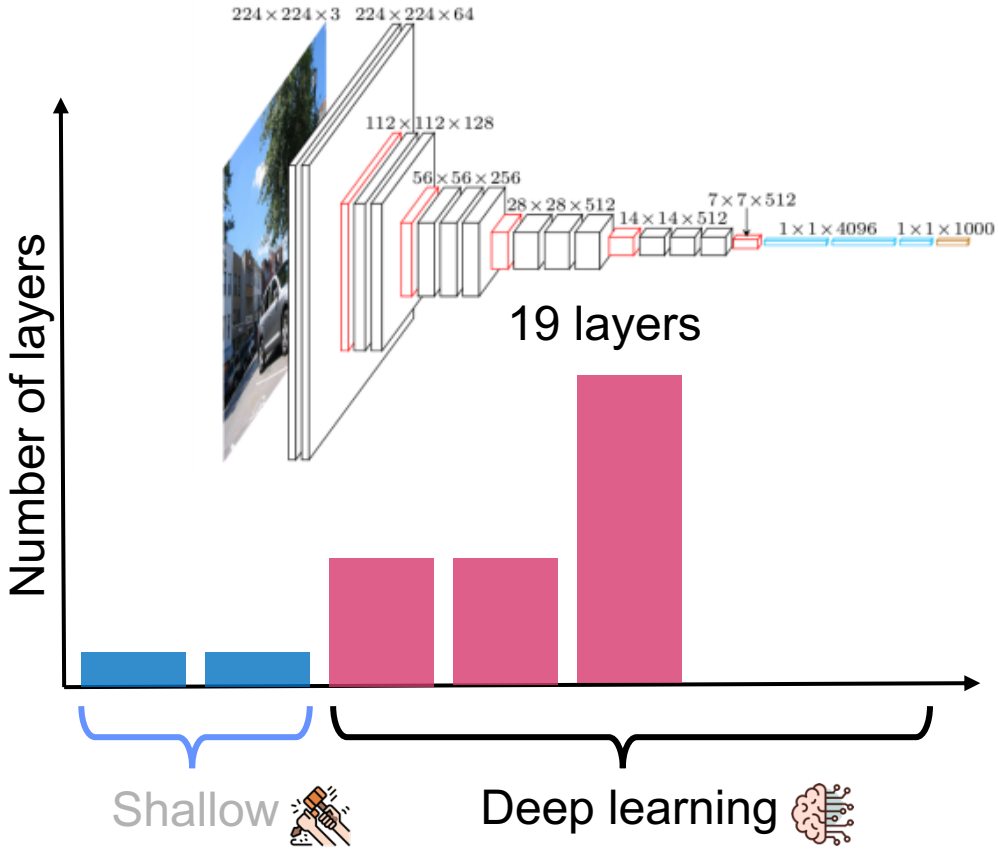
- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET



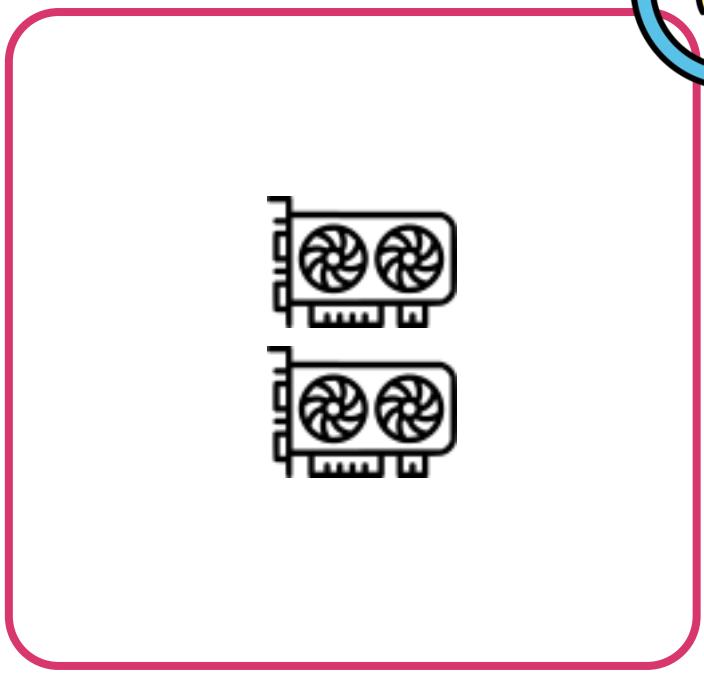
- 2010 Fisher Vectors
- 2011 Fisher Vectors
- 2012 AlexNet
- 2013 ZF
- 2014 VGG

[Simonyan@ECCV14]



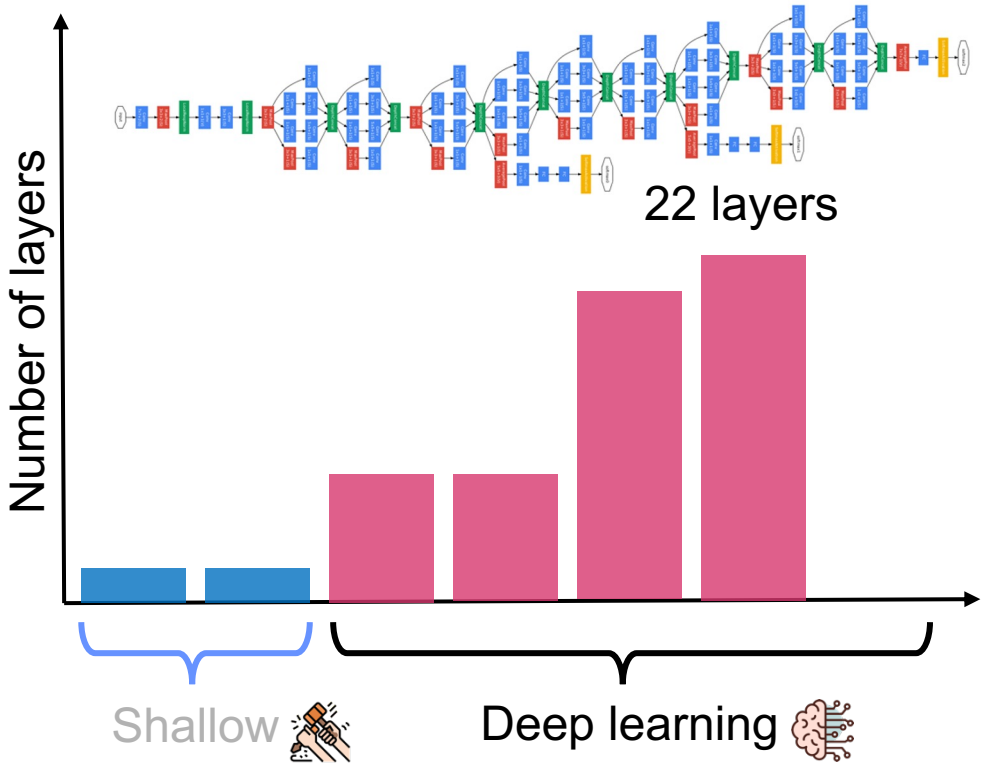
- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET



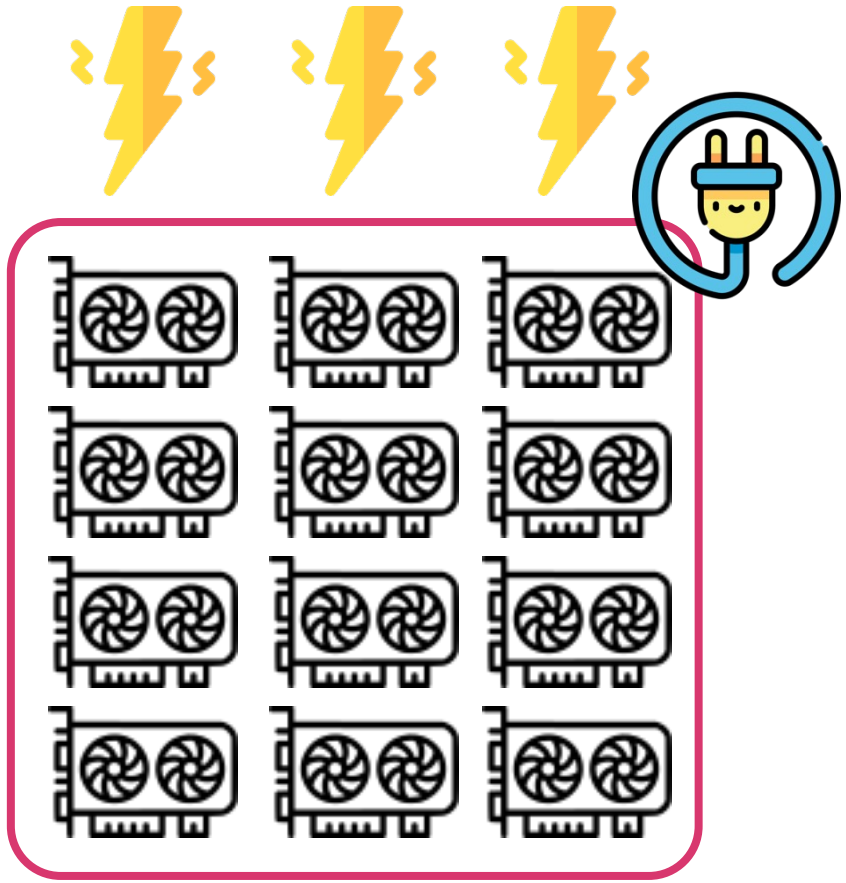
- 2010 Fisher Vectors
- 2011 Fisher Vectors
- 2012 AlexNet
- 2013 ZF
- 2014 VGG
- 2014 GoogLeNet

[Szegedy@CVPR14]



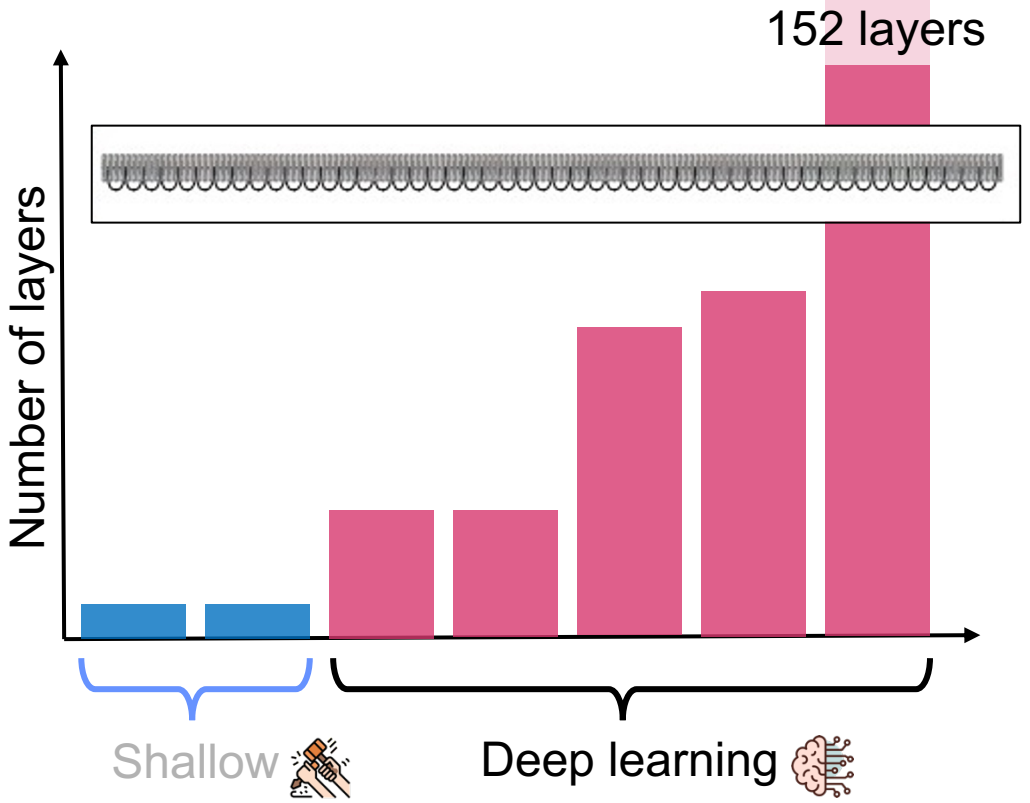
- [x] Large image collections to train from
- [x] Deeper models with more parameters

IMAGENET



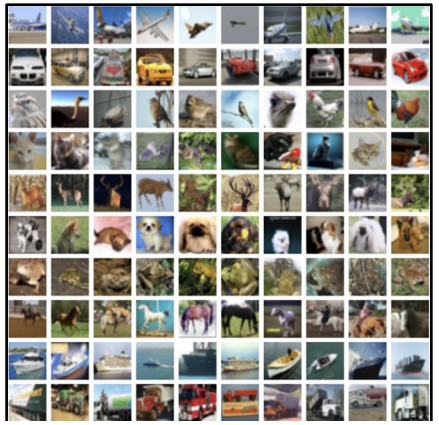
2010 Fisher Vectors
2011 Fisher Vectors
2012 AlexNet
2013 ZF
2014 VGG
2014 GoogLeNet
2015 ResNet

[He@CVPR15]



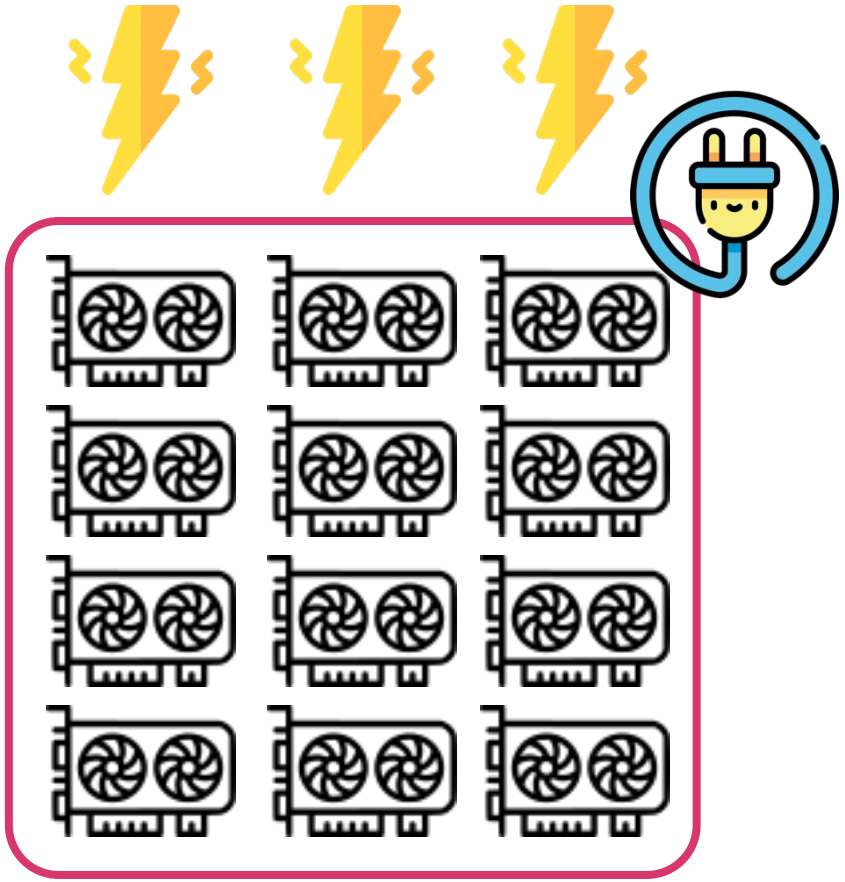
A successful recipe - Summary

- Large image collections to train from
- Deep models with many parameters
- Lots of compute



IMAGENET

[**ImageNet**: Deng@CVPR2009]

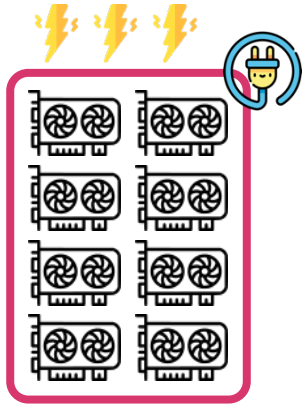


A successful recipe

Task 2

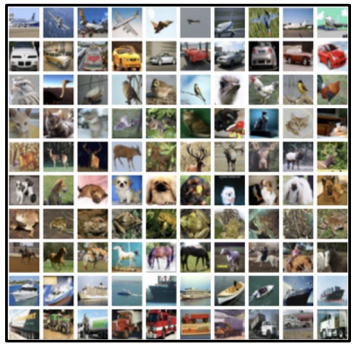
[Weyand@CVPR2020]

Model



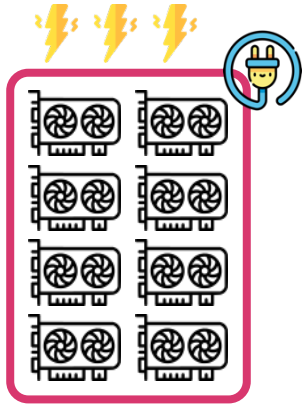
IMAGENET

Task 1



[Deng@CVPR2009]

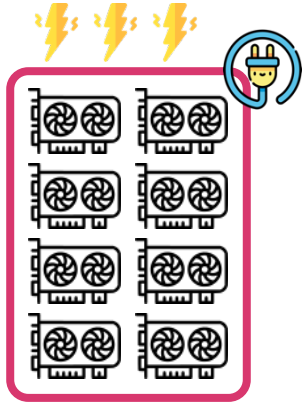
Model



Task 3

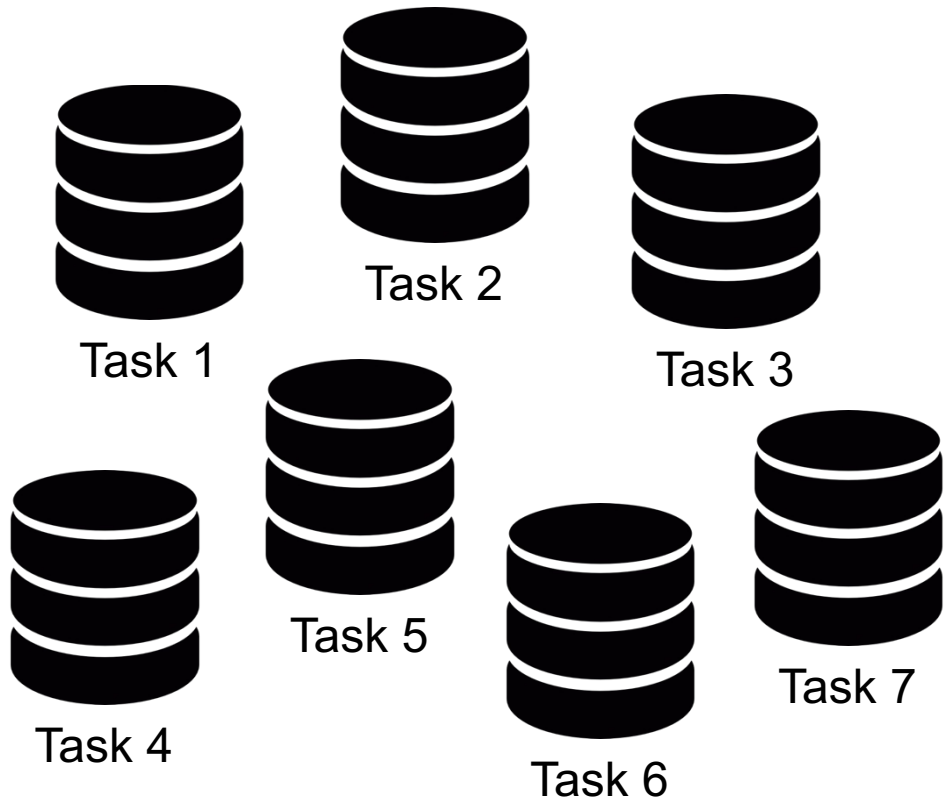
[Minervini@PRL15]

Model

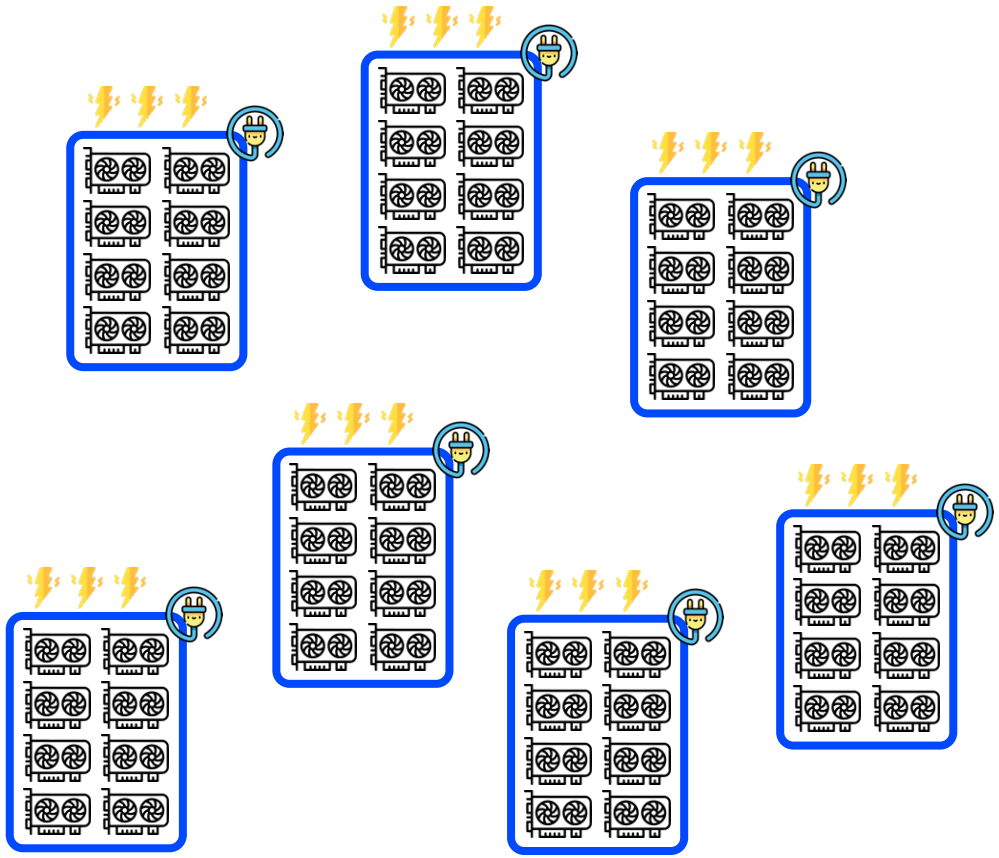


What about
Lifelong learning?

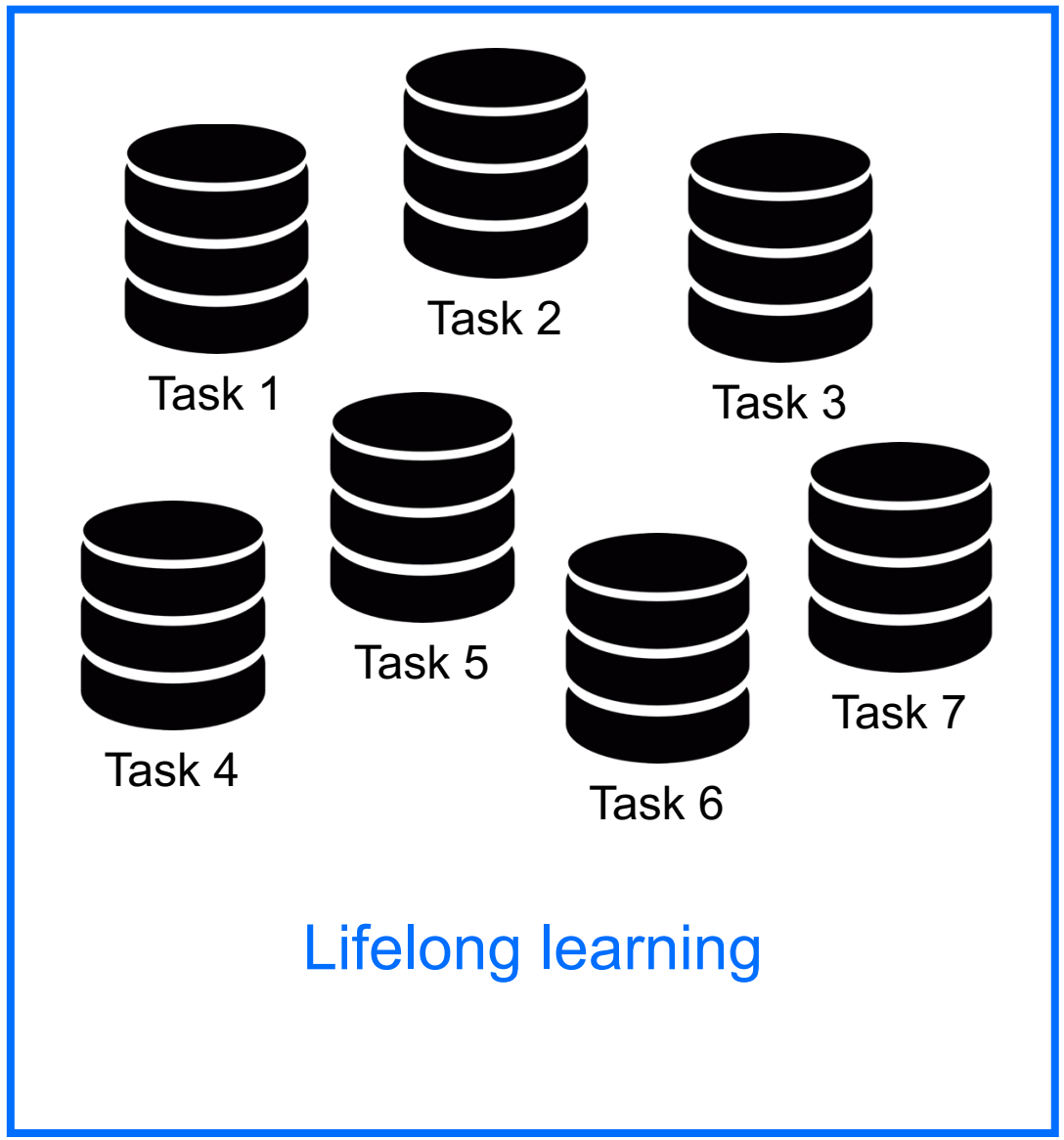
multiple tasks

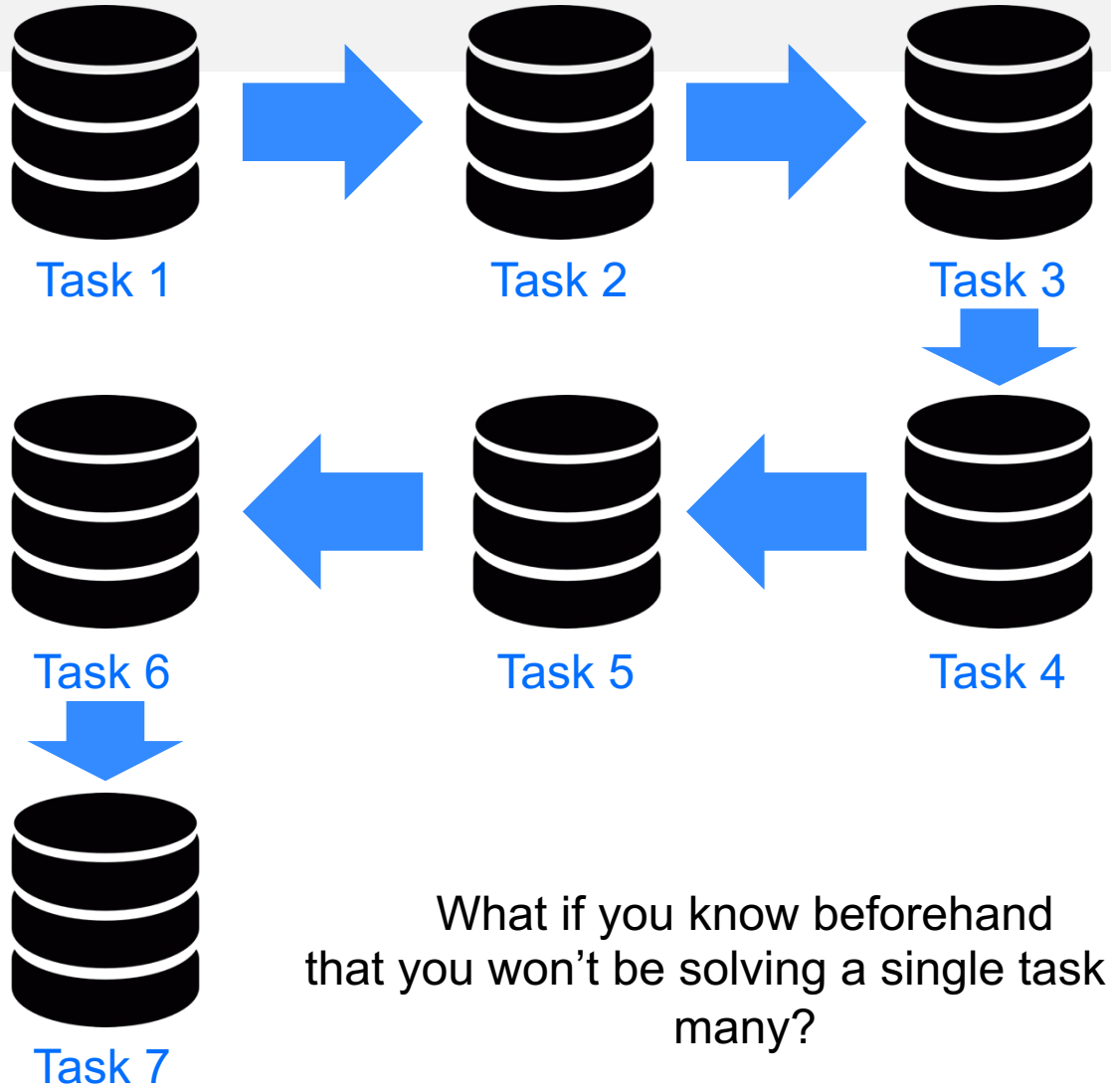


Collection of tasks



multiple tasks

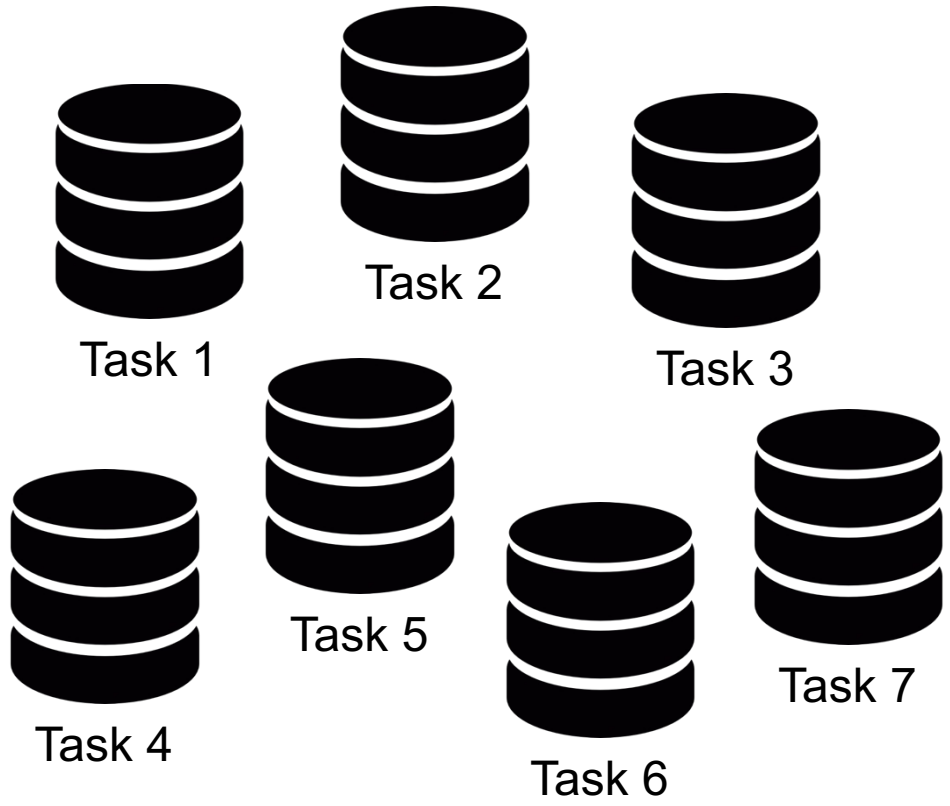




What if you know beforehand
that you won't be solving a single task but many?

How can you make sure that each task reuses as much as possible the knowledge that has been acquired when training for other tasks?

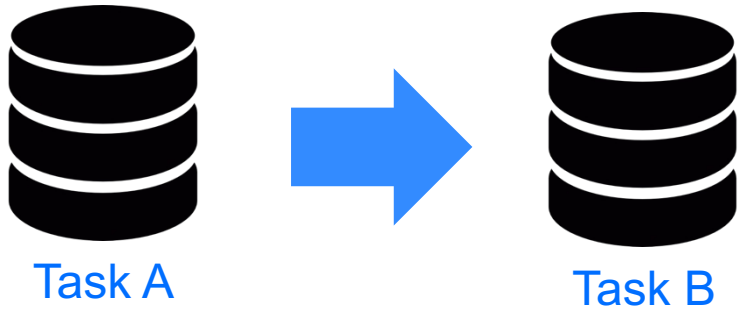
Objectif de ce cours



Apprentissage continu
de représentations

Premier scénario que nous considérons

Utilisation d'une tâche A
pour améliorer une tâche B



Terminologie

Descripteurs / représentations / *features* / caractéristiques

Fonction de coût / Fonction de perte / *Loss function*

Illustrations possibles

Représentations visuelles et transfert de tâche

Apprentissage continu de représentations visuelles

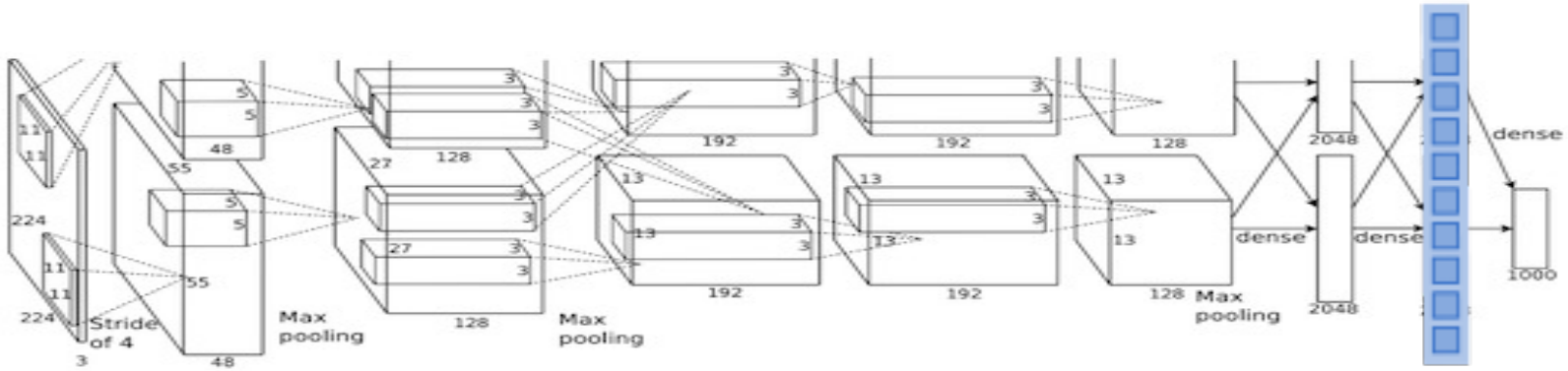
2023-2024

Les réseaux de neurones comme **extracteurs de descripteurs**

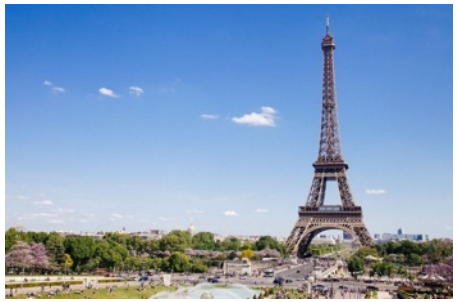
L'architecture CNN pré-entrainée pour la classification peut facilement être utilisée comme un extracteur de représentation.

- Les représentations sont compactes et rapides au moment du test!

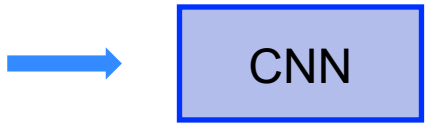
AlexNet



[Krizhevsky et al. NIPS 2012]



Input image



Quel descripteur?

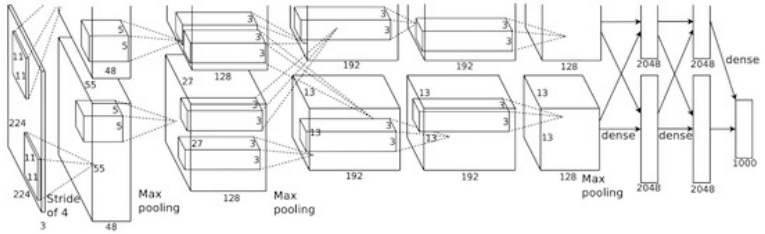
Probabilité sur les 1000 classes d'ImageNet,
ou mieux la sortie de la couche de neurones précédente

Transfert de représentations visuelles

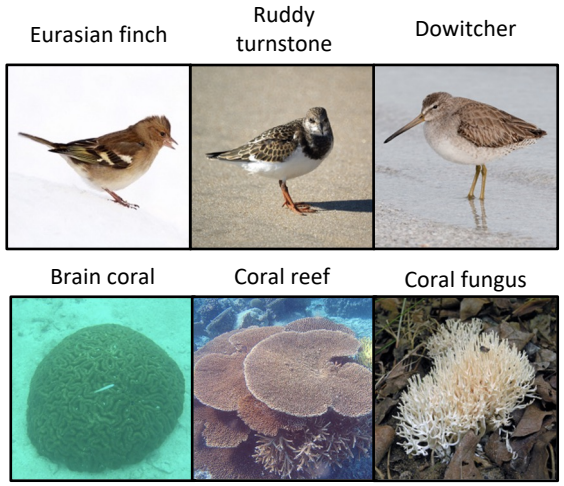
Tâche A

Tâche B

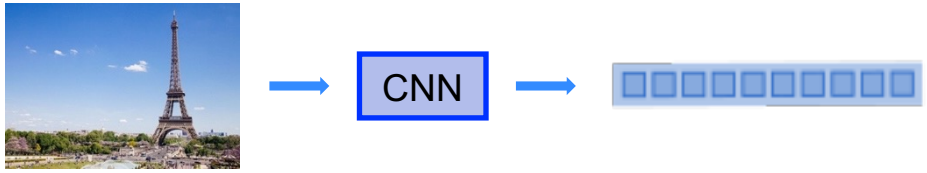
Une architecture de réseaux de neurones, par exemple AlexNet …



…entrainée sur IMAGENET

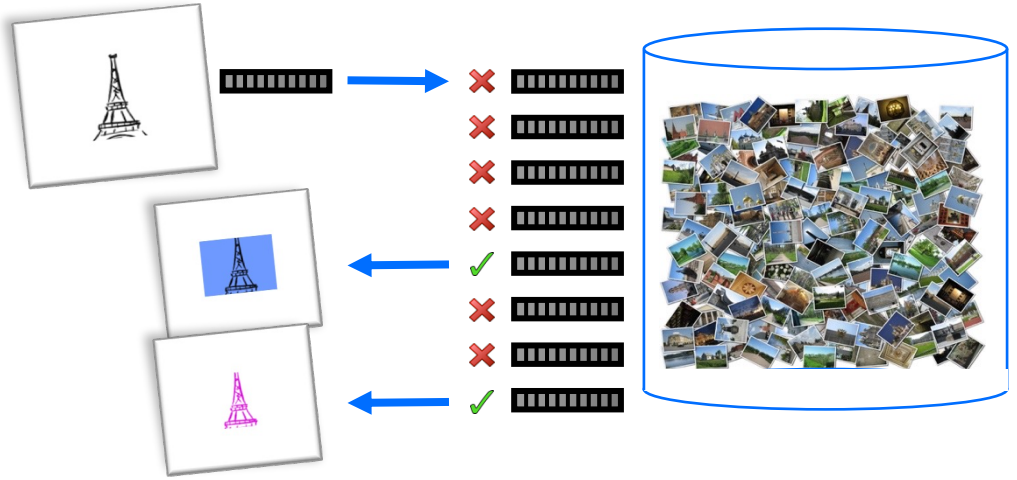


... est classiquement utilisée comme extracteur de représentations visuelles.



Input image

Par exemple pour la tâche de **recherche d'images**,

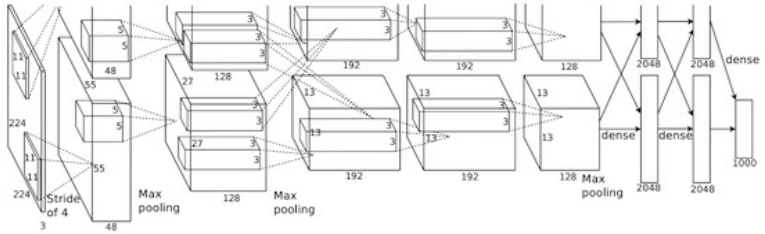


Transfert de représentations visuelles

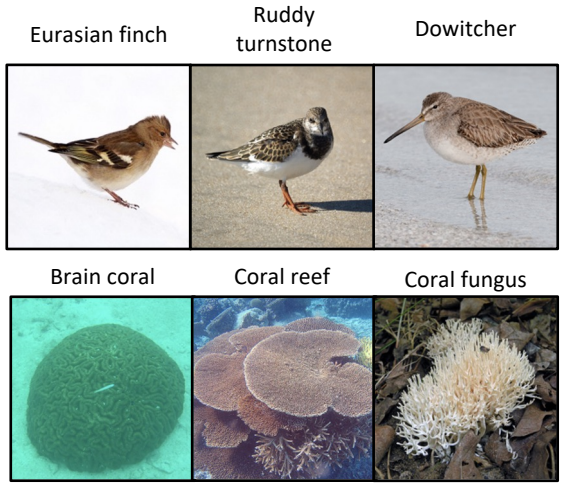
Tâche A

Tâche B

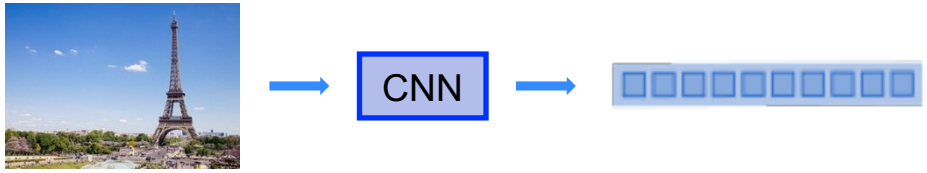
Une architecture de réseaux de neurones, par exemple AlexNet …



…entrainée sur IMAGENET

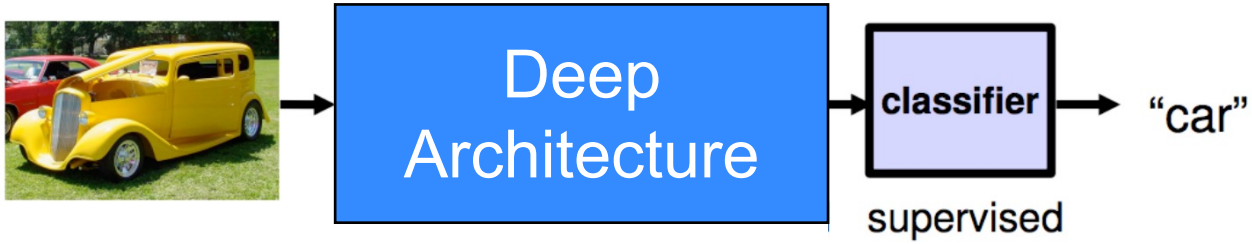


... est classiquement utilisée comme extracteur de représentations visuelles.



Input image

Ou combinée avec un classifieur:
Par exemple pour la **classification d'images**.
(lien avec le cours de la semaine dernière)

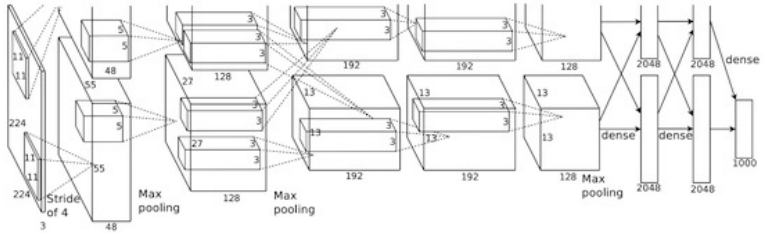


Transfert de représentations visuelles

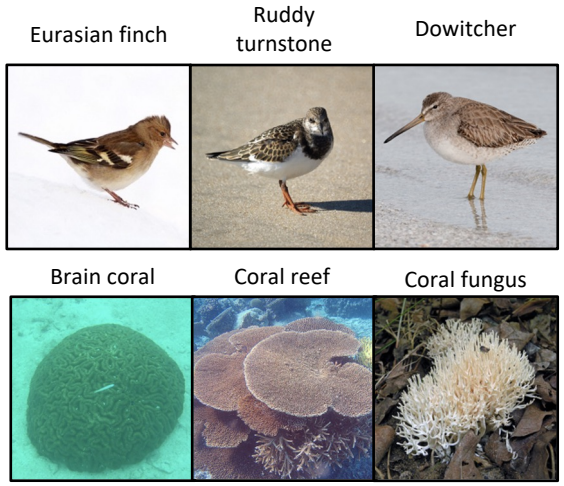
Tâche A

Tâche B

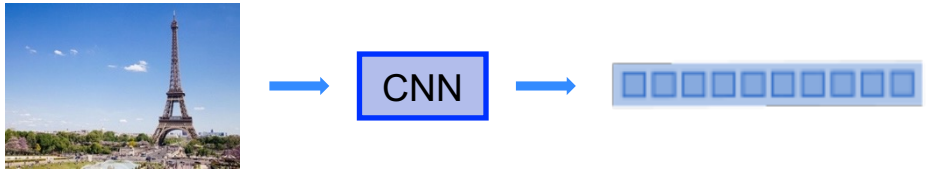
Une architecture de réseaux de neurones, par exemple AlexNet …



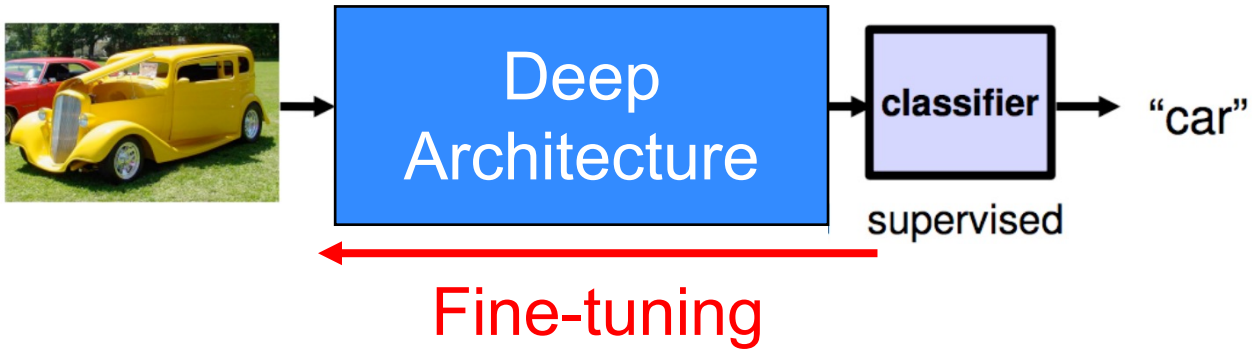
…entrainée sur IMAGENET



... est classiquement utilisée comme extracteur de représentations visuelles.



On peut aussi mettre à jour les poids du réseau pour la nouvelle tâche

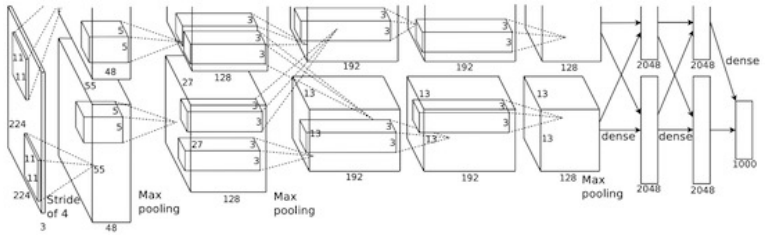


Transfert de représentations visuelles

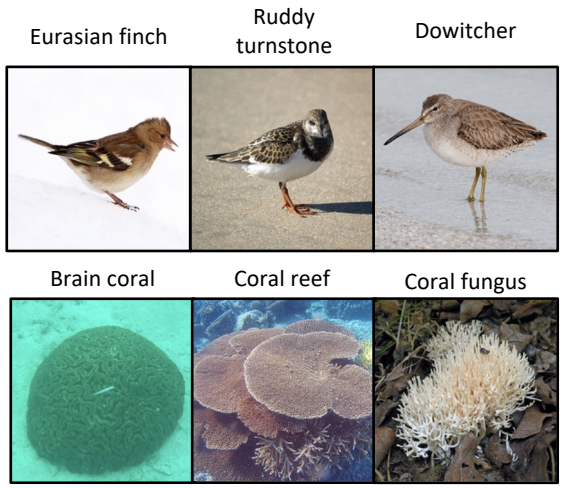
Tâche A

Tâche B

Une architecture de réseaux de neurones, par exemple AlexNet …

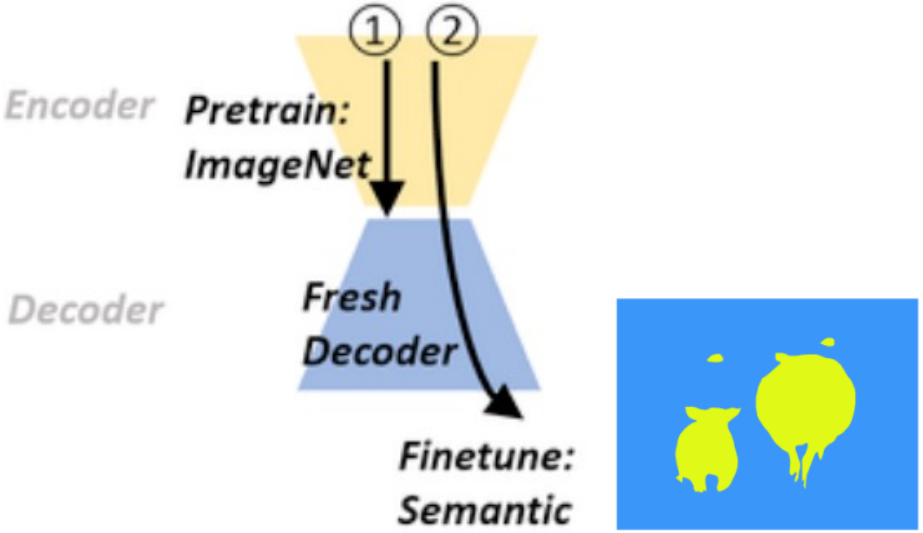


…entrainée sur IMAGENET



... peut être **utilisée de façon totalement différente**

En général: initialisation d'une partie seulement du réseau pour la tâche cible

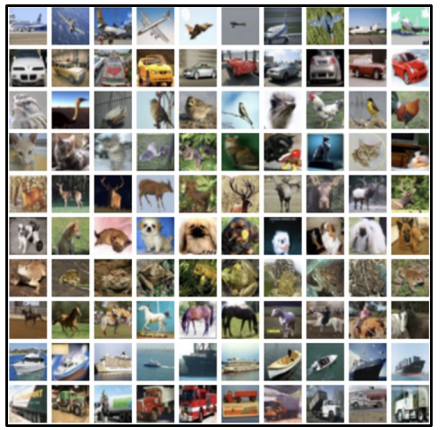


Transfert de représentations visuelles - Résumé

Tâche A

Fully-Supervised Classification

Images + labels



Model

Visual representations

Tâche B

Image Classification

Object Detection

Instance Segmentation

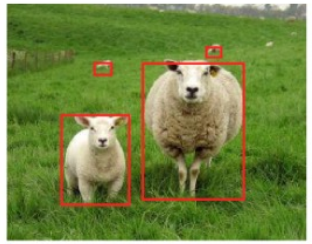
Image Retrieval

Des tâches multiples

Classification

Sheep ?

yes



Sheep ?

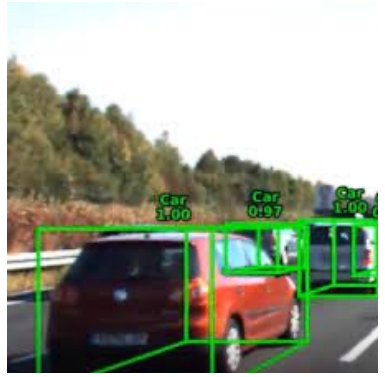
4

Détection

Sheep ?

Segmentation

Estimation de profondeur

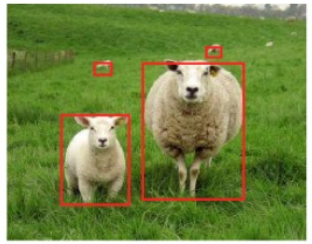


Détection 3D

On suppose que le domaine ne varie pas ou peu
entre les deux tâches A & B.
Dans tous les exemples de ce transparent,
il s'agit d'images naturelles.

Des tâches multiples

Classification

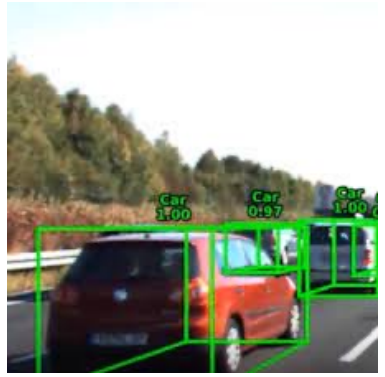


Détection

Segmentation

Supervised Learning

- Input: x (images, text, emails…)
- Output: y (spam or non-spam…)
- (Unknown) Target Function
  - f: X → Y (the “true” mapping / reality)
- Data
  - $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$
- Model / Hypothesis Class
  - g: X → Y
  - $y = g(x) = \text{sign}(w^T x)$
- Learning = Search in hypothesis space
  - Find best g in model class



Détection 3D

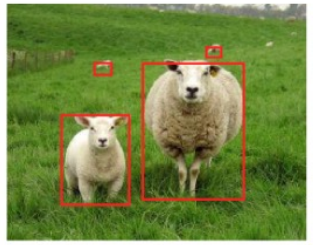
varie pas ou peu
A & B.
e transparent,
relles.

Des tâches multiples

Classification

Sheep ?

yes



Sheep ?

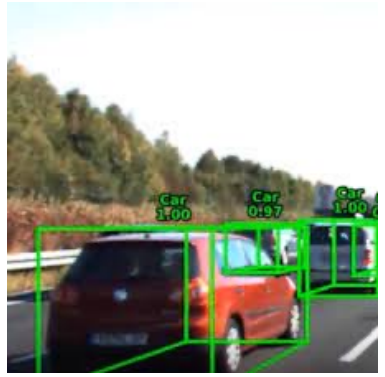
4

Détection

Sheep ?

Segmentation

Estimation de profondeur



Détection 3D

On suppose que le domaine ne varie pas ou peu entre les deux tâches A & B.
Dans tous les exemples de ce transparent, il s'agit d'images naturelles.

Data

- $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$

Cas particulier du transfert de tâche:
Un changement de domaine =
une distribution différente pour les données d'entrée

Data

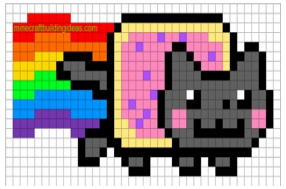
- $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$

Pour un même ensemble de *labels* (étiquettes) cibles, la distribution x~X peut être différente

Domain A

Domain B

Domain C



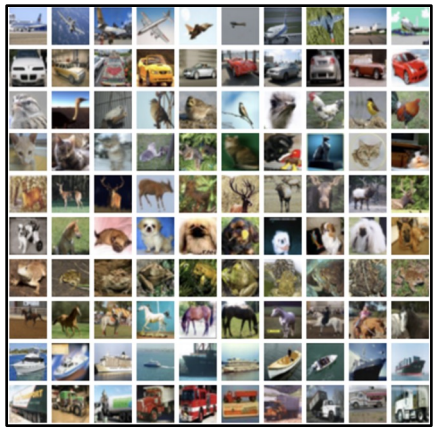
Next: Transfert vers un nouveau domaine

Fully-Supervised Classification

Images + labels

Tâche A

IMAGENET



Model

Visual representations

Target domains

Peintures

Dessins

Images médicales

Images de nuit

Adaptation à un nouveau domaine

Apprentissage continu de représentations visuelles

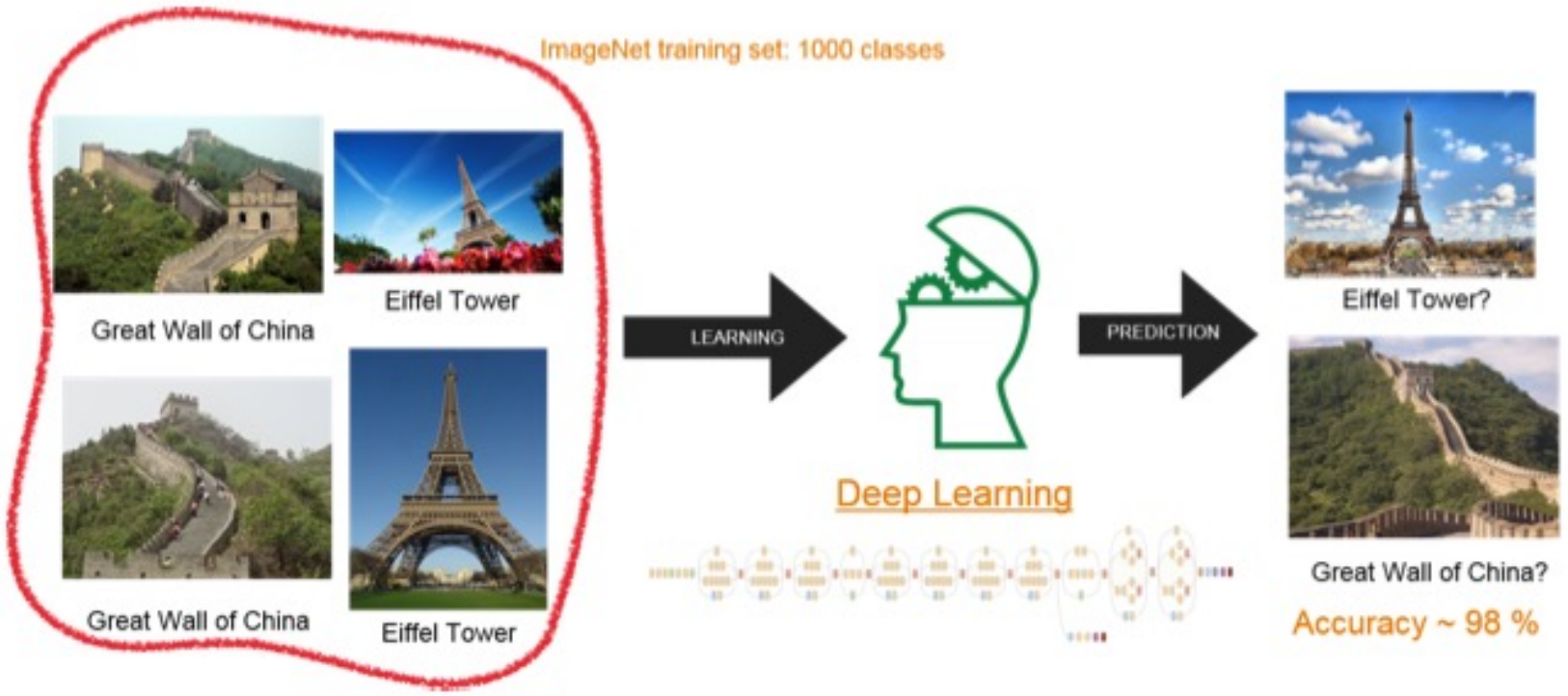
2023-2024

Transparents de Gabriela Csurka

Basés sur un tutoriel à ECCV20

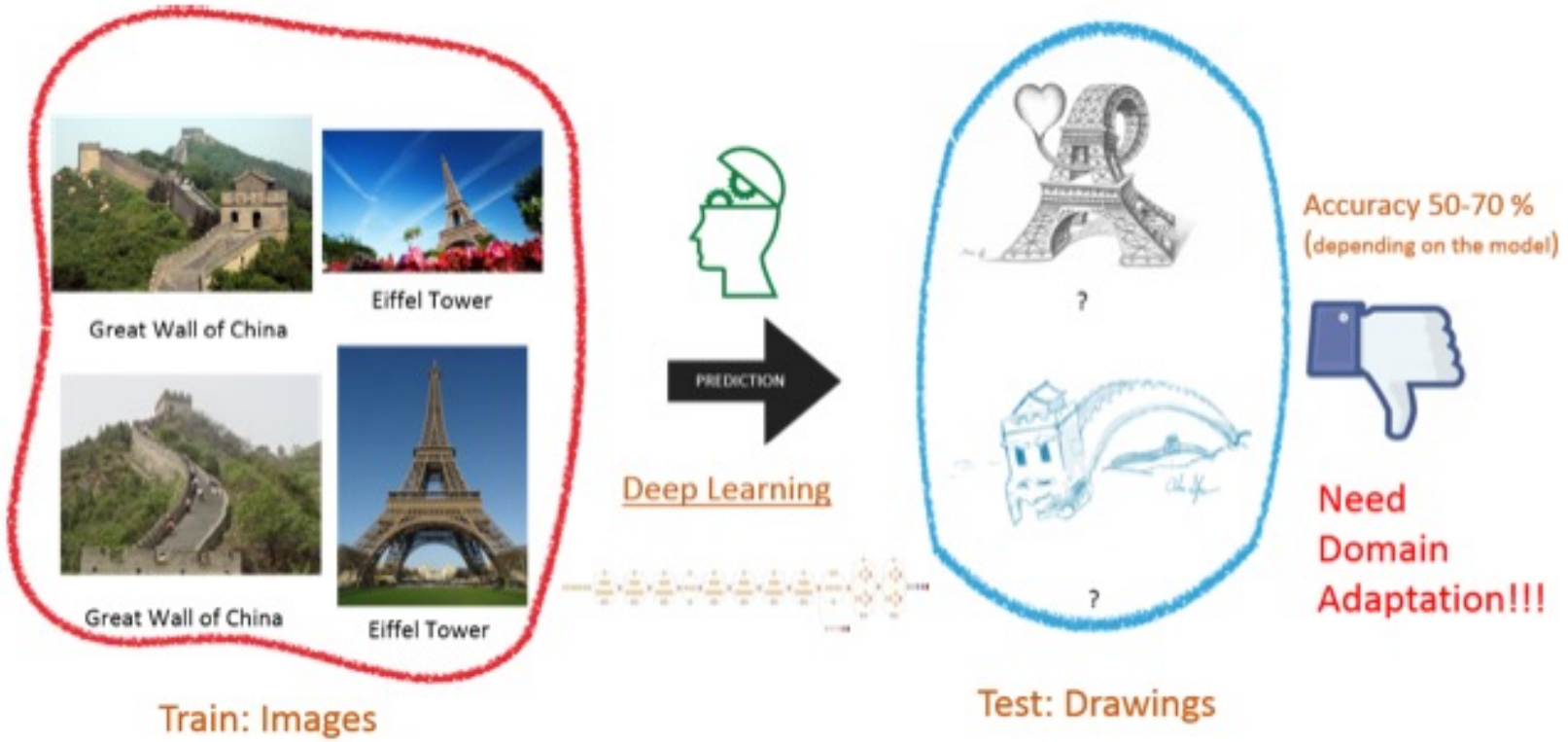
- https://europe.naverlabs.com/eccv-2020-domain-adaptation-tutorial/

Image Classification



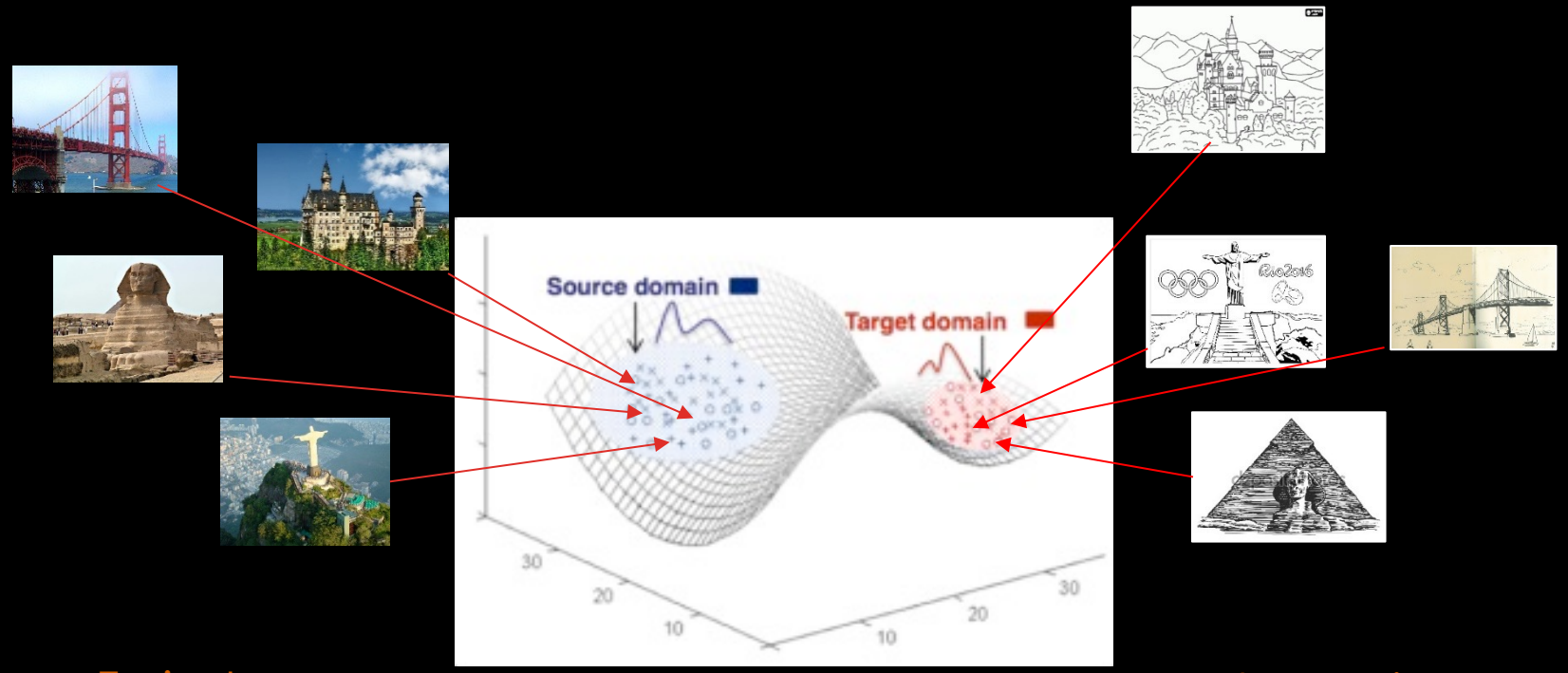
Crédit: Gabriela Csurka

Domain Shift



Crédit: Gabriela Csurka

WHAT HAPPENED?



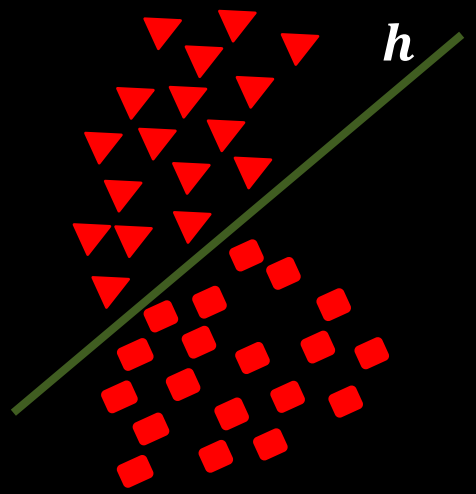
Train: Images

Test: Drawings

The data distributions between training and test sets are different

Crédit: Gabriela Csurka

MORE CONCRETELY - THROUGH A TOY EXAMPLE

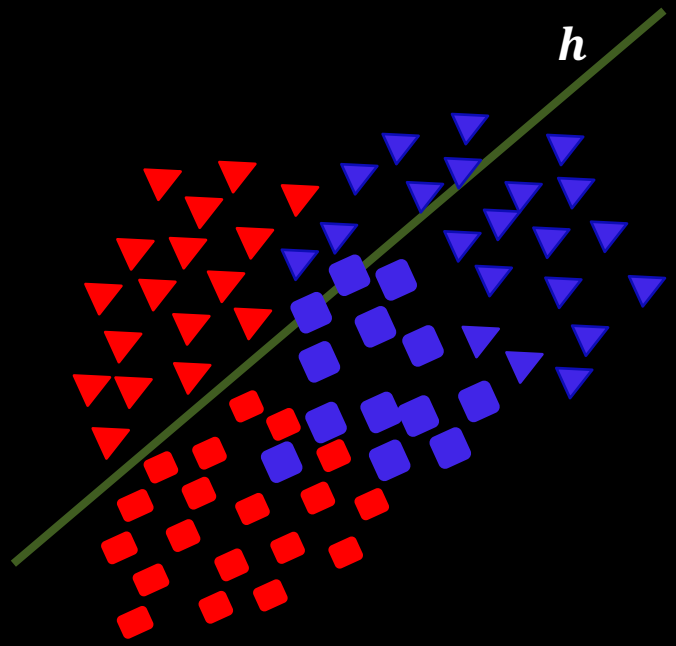


On suppose disposer d'un classifieur binaire entrainé
sur le domaine rouge,
afin de l'utiliser sur le domaine bleu

Let's assume a simple two class case and learn a classifier h to separate them

Crédit: Gabriela Csurka

DOMAIN SHIFT/ DISTRIBUTION MISMATCH

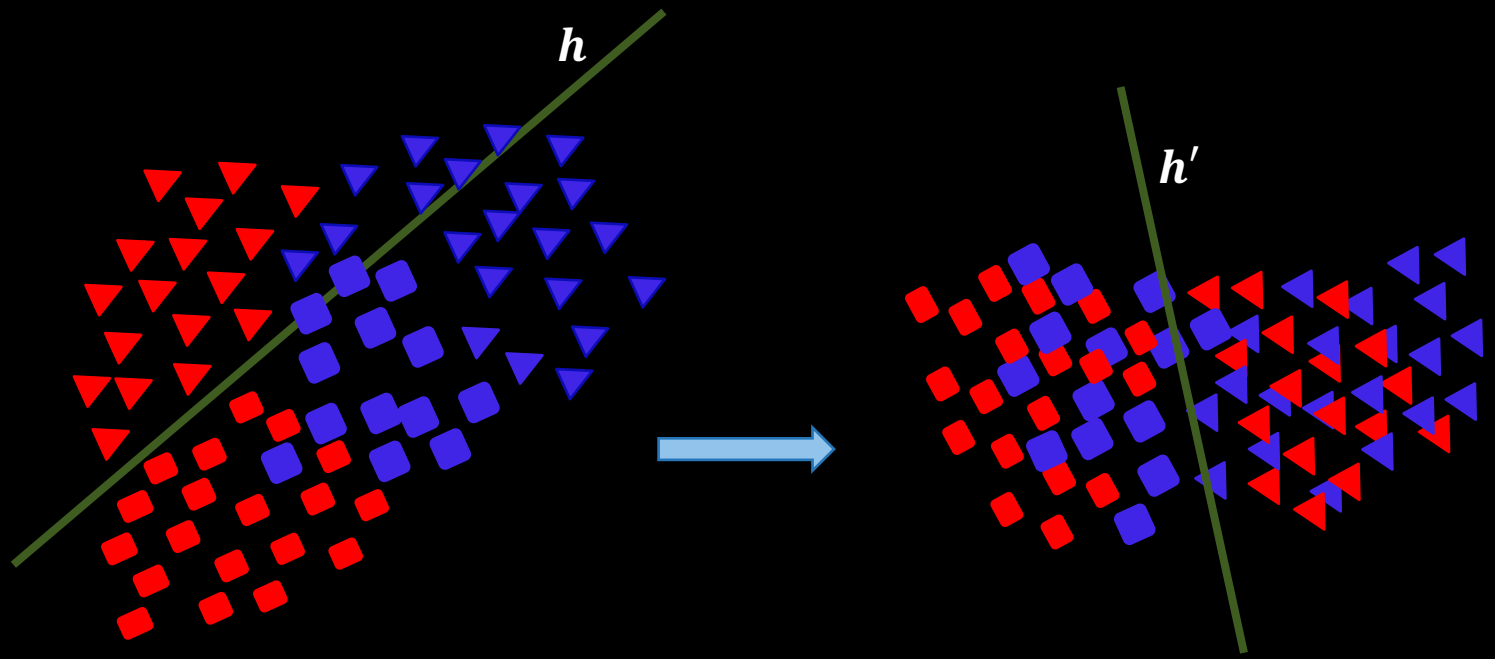


On suppose disposer d'un classifieur binaire entrainé
sur le domaine rouge,
afin de l'utiliser sur le domaine bleu

The classifier h learned on the first domain performs badly on the second one due to a distribution mismatch

Crédit: Gabriela Csurka

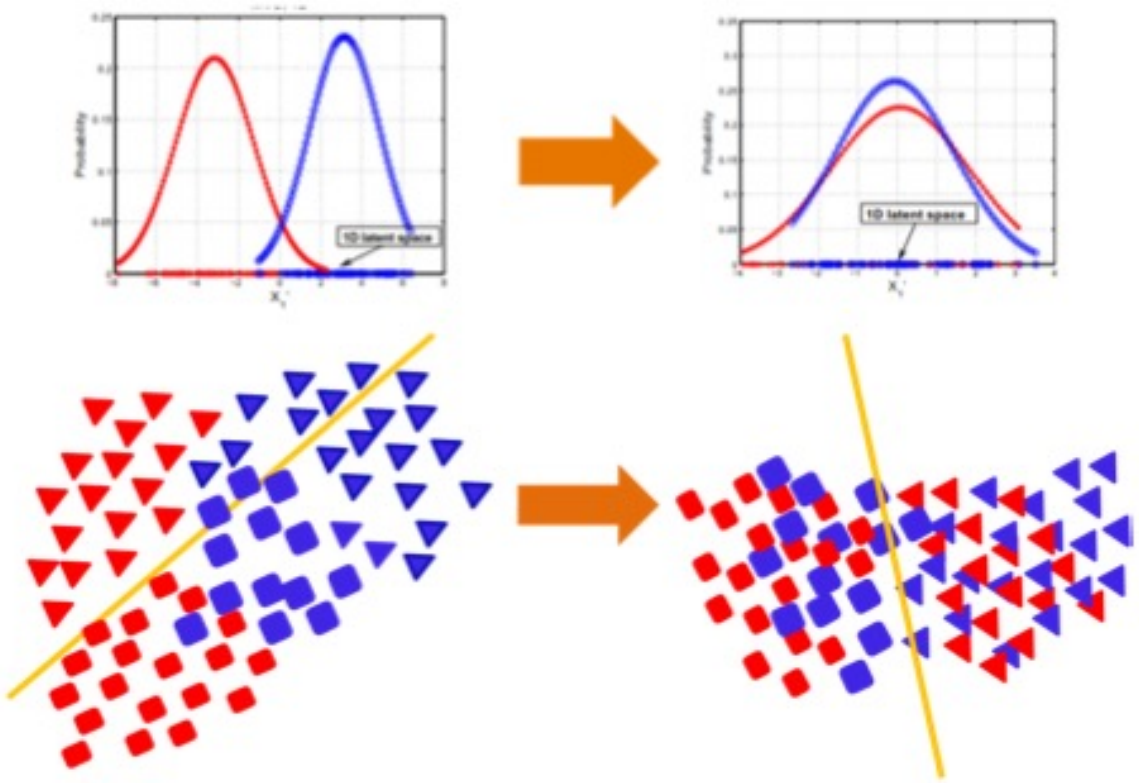
AIM



Learn a new representation space where the data distributions fit

Crédit: Gabriela Csurka

Key idea: solve the distribution mismatch



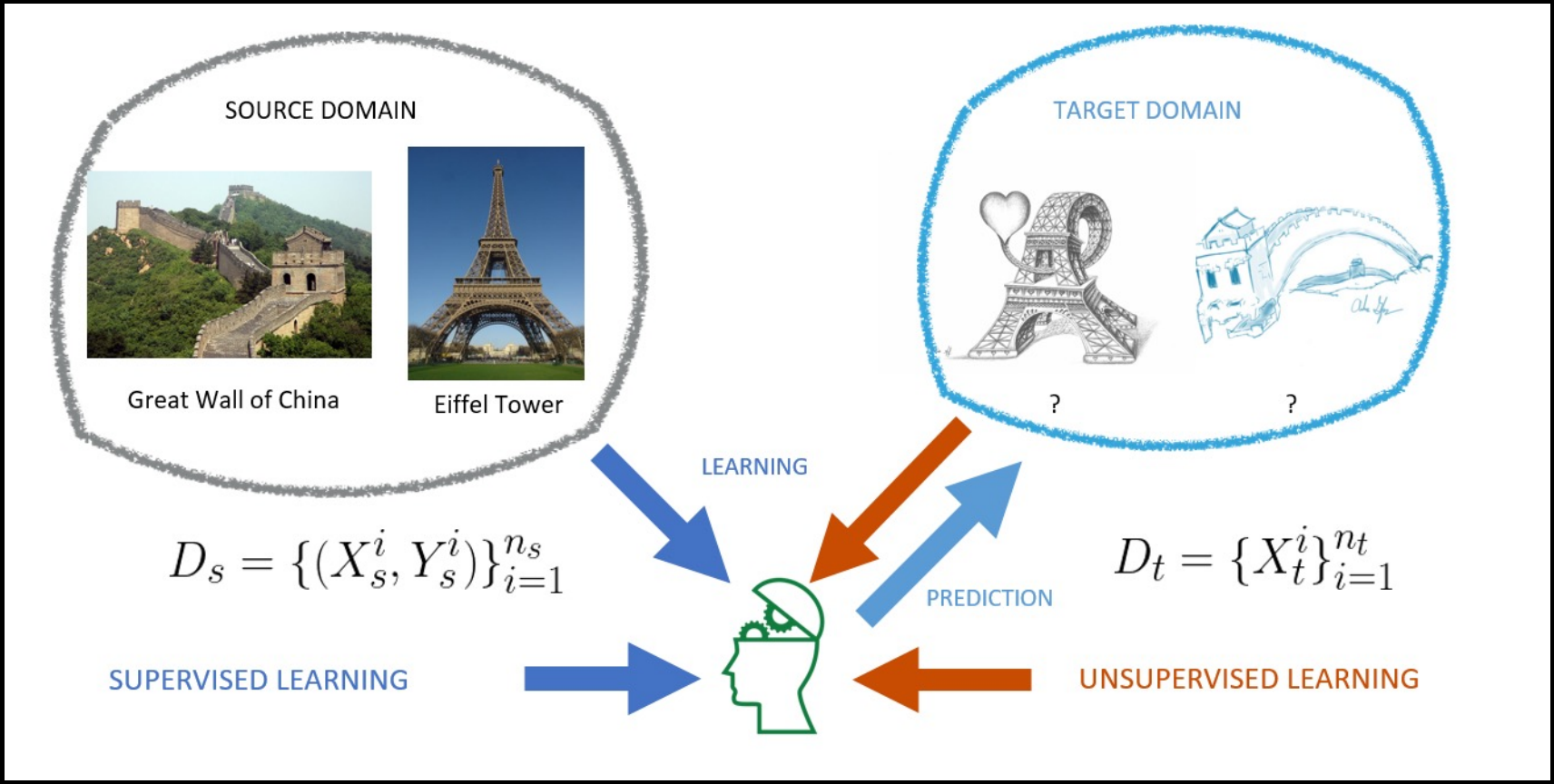
The distribution mismatch, can be measured for example by the Maximum Mean Discrepancy (MMD), Borgwardt +@BIOINFORMATICS'06

On suppose disposer d'un classifieur binaire entrainé sur le domaine rouge, afin de l'utiliser sur le domaine bleu.

L'approche MMD suppose que le changement de domaine affecte tous les descripteurs de la même façon.

DOMAIN ADAPTATION (DA)

Définition de **l'adaptation de domaine**



Leverage labelled data in one domains, referred to as **source**, to learn a classifier for data in a **target domain**

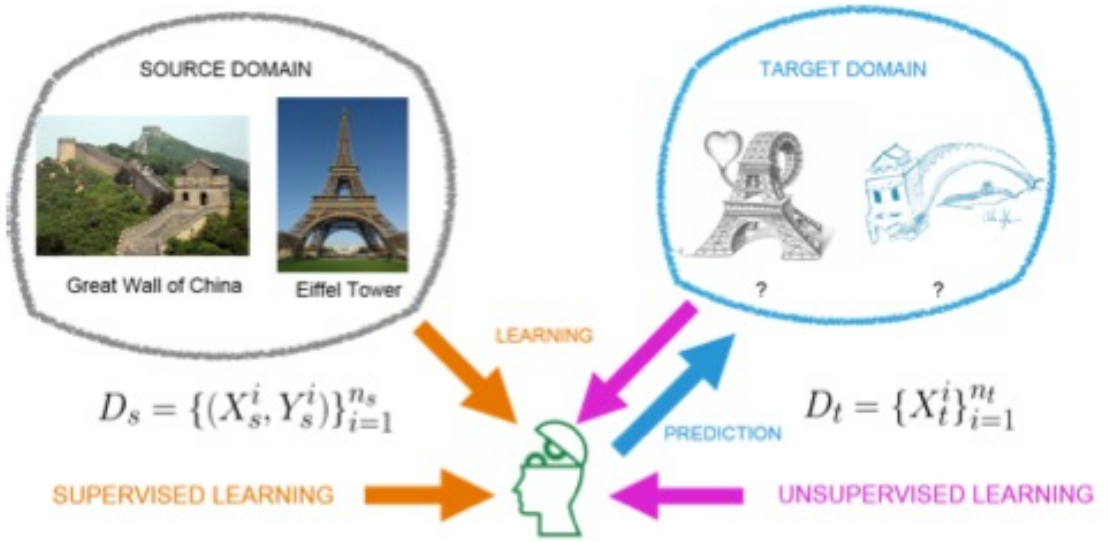
Crédit: Gabriela Csurka

Motivation pour l'adaptation de domaine

On considère un domaine de départ, dit **source**, qui est labélisé / annoté.
On souhaite avoir des bonnes performances pour un domaine dit **cible**.

Pourquoi ne pas directement entrainer sur le domaine **cible** ?
Souvent: on n'a pas assez d'exemples pour ce domaine.
De plus, on a parfois trop peu, voir aucune annotation pour le domaine cible !

On considère deux types d'adaptation
- l'adaptation **non supervisée**
- l'adaptation **semi-supervisée**



Adaptation à un nouveau domaine

Adaptation supervisée / semi-supervisée

quelques exemples **annotés** sont disponibles dans le domaine cible

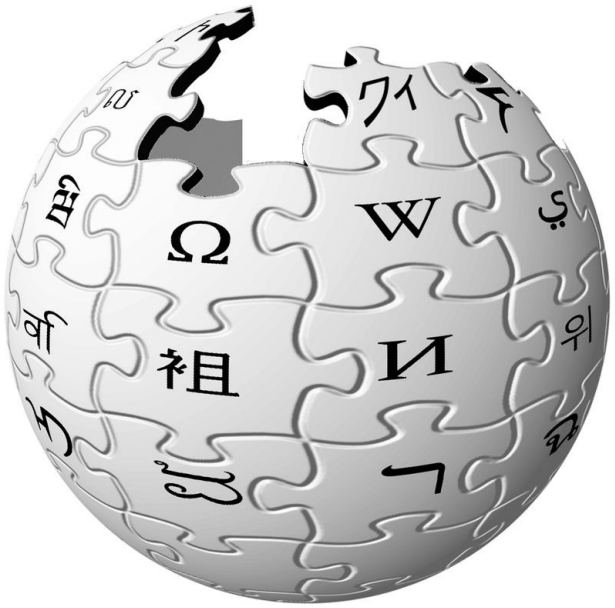
vs

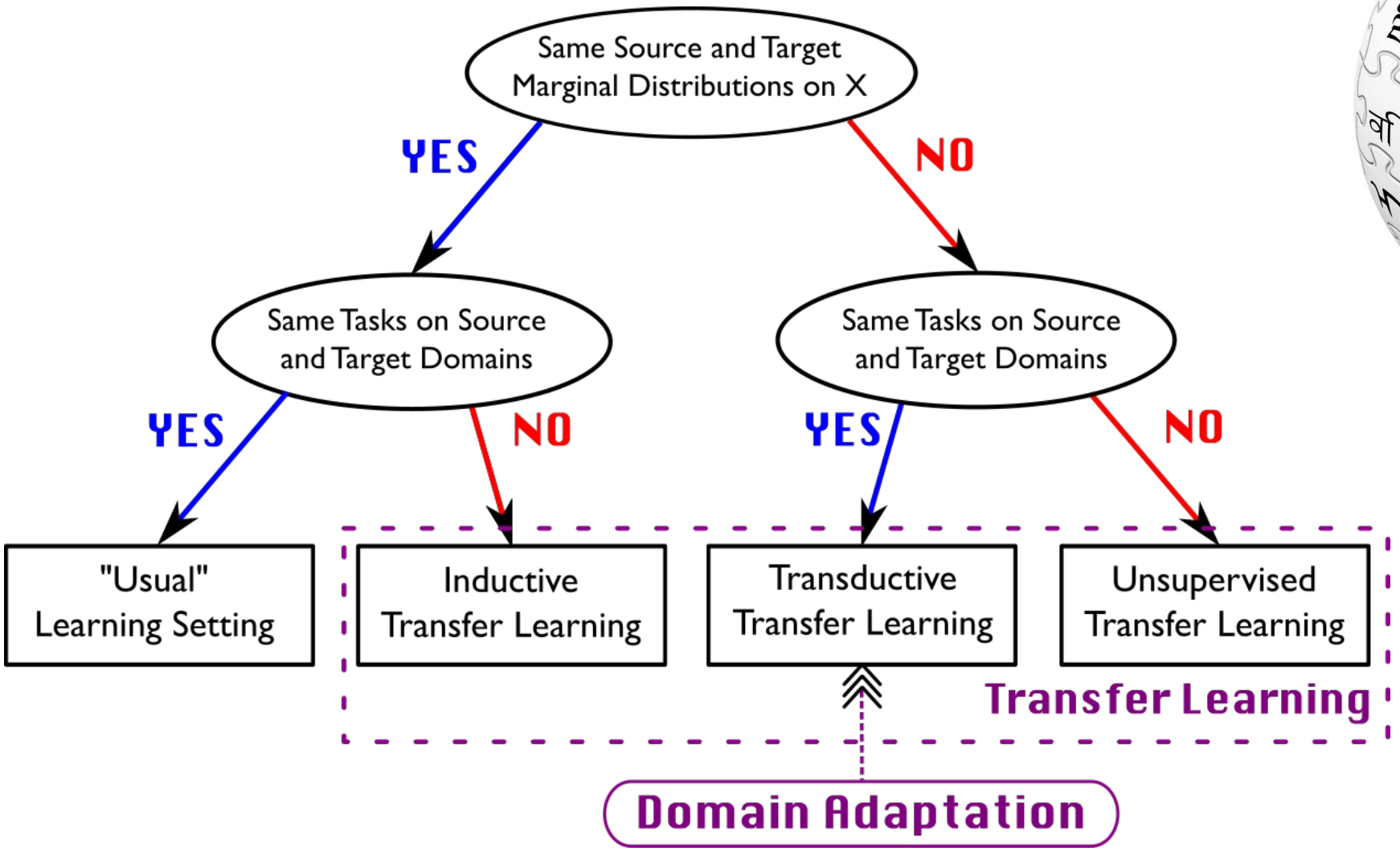
Adaptation non-supervisée

des exemples sont disponibles dans le domaine cible, mais **aucun** n'est annoté

Attention, des exemples (annotés ou non) sont toujours disponibles pour le domaine cible.
Sinon, on ne parle pas d'une tâche d'**adaptation de domaine**,
mais d'une tâche de **généralisation de domaine**.

Positionner l'adaptation de domaine parmi les méthodes de transfert





THERE ARE REAL APPLICATION NEEDS

- Scene-Understanding in Autonomous Driving (AD)
- Long-term visual localization and place recognition
- Robot navigation, control and task learning
- Biomedical imaging
- Biometry and surveillance
- Remote sensing and satellite imagery
- etc

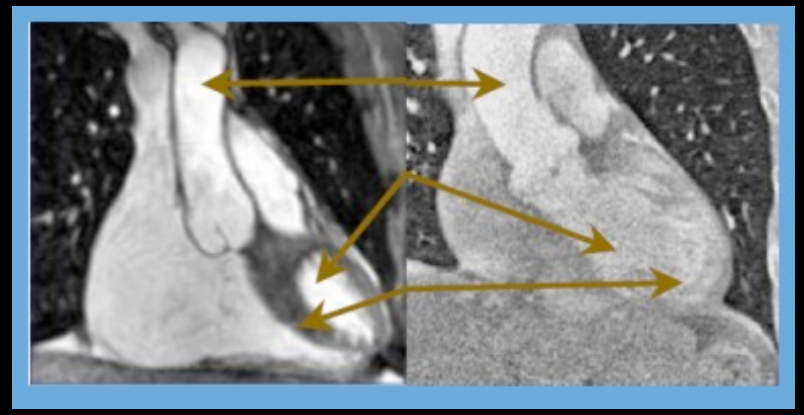
Crédit: Gabriela Csurka

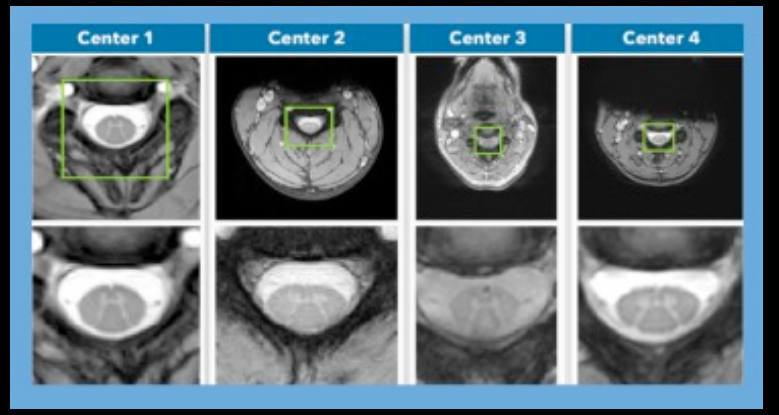
SCENE UNDERSTANDING IN AD

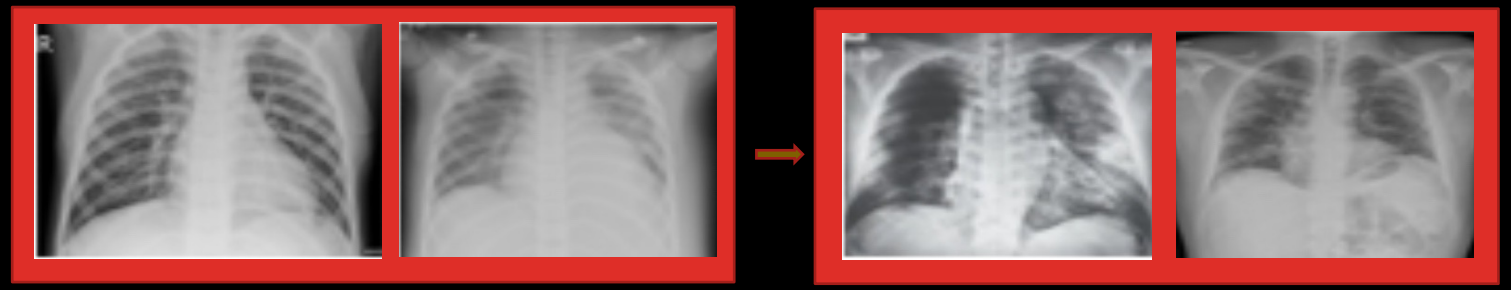
Understanding the traffic scene images is curial for autonomous driving (AD) scenarios as detection errors can be fatal

Crédit: Gabriela Csurka

MEDICAL IMAGING







Adaptation of model between different image modalities (MRI to CT), medical centres, or even pathologies (pneumonia vs COVID)

Crédit: Gabriela Csurka

Example scenarios

Recognition

Detection

Segmentation

Re-identification

Control

Visual localization

Crédit: Gabriela Csurka

Outline

Traditional shallow DA methods

How to exploit deep learning in DA

Main trends in Deep Domain Adaptation

Crédit: Gabriela Csurka

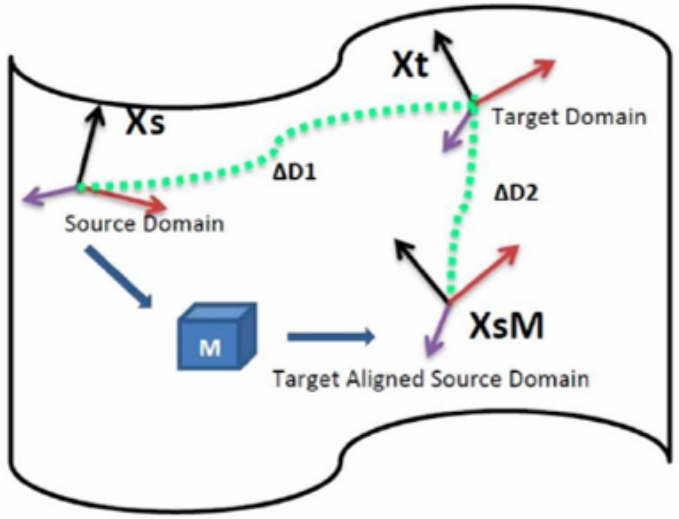
Exemples de méthodes d'adaptation **semi-supervisées**

- Pondération des exemples du domaine source
- Ajustement des paramètres du classifieur (eg: SVM)
- Méthodes basées sur l'apprentissage de métriques
- *Fine-tuning* du modèle source avec les quelques données annotées de la cible (approche principale depuis le *deep learning*)

Exemples de méthodes d'adaptation non-supervisées

- Méthodes d'augmentation des descripteurs
- Alignement des espaces des descripteurs
- Apprentissage d'un nouvel espace de représentation
- Transformations locales (ex: *optimal transport*)

Illustration de l'alignement des espaces



Outline

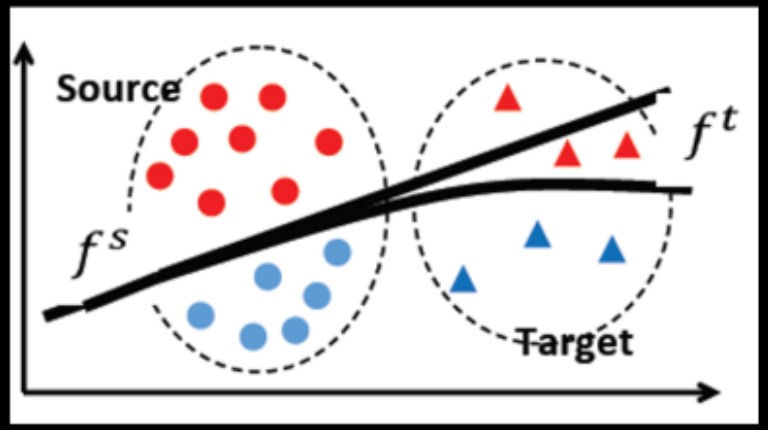
Traditional shallow DA methods

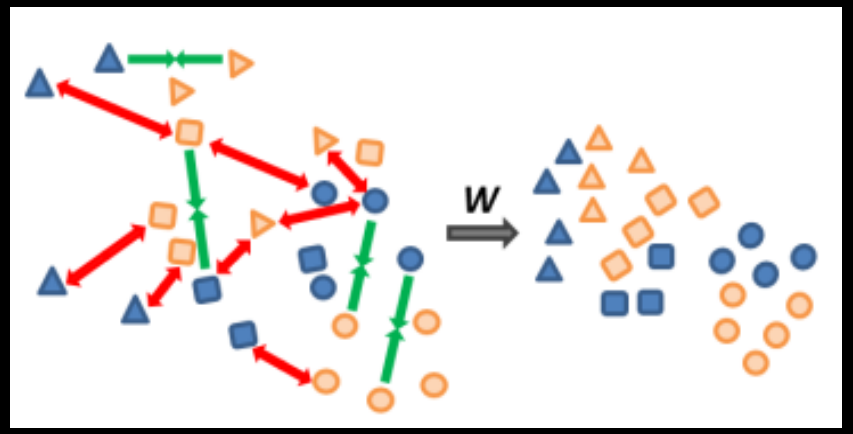
How to exploit deep learning in DA

Main trends in Deep Domain Adaptation

Crédit: Gabriela Csurka

SEMI-SUPERVISED DOMAIN ADAPTATION





Adapting model parameters

- Adaptive SVM (Yang+@MM'07)
- DASVM (Bruzzone+@PAMI'10)
- TrAdaBoost (Dai+@ICML'08)

Metric learning

- ITML (Yang+@ECCV'10)
- NBNN (Tommasi +@ICCV'13)
- DSCM (Csurka+@TaskCV'14)

Crédit: Gabriela Csurka

SUBSPACE REPRESENTATION

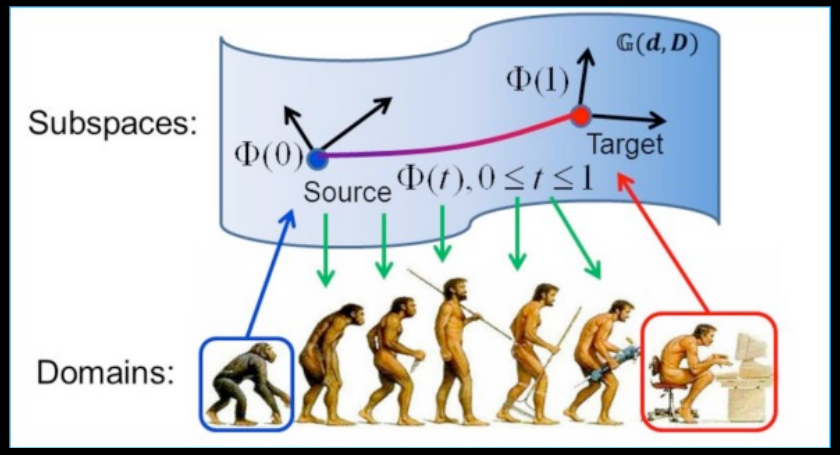


Image courtesy of Boqing Gong

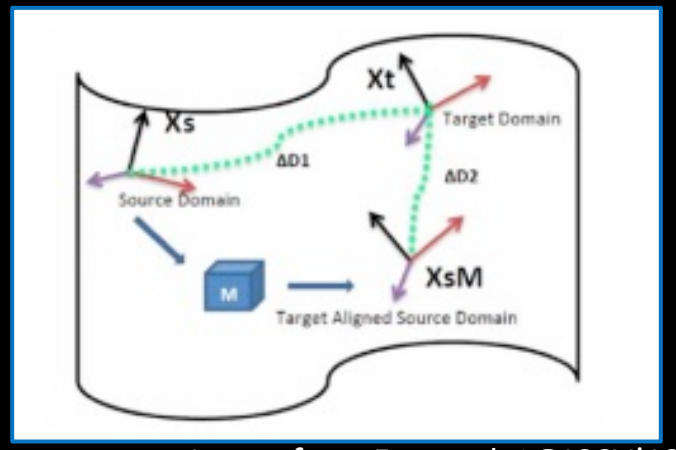


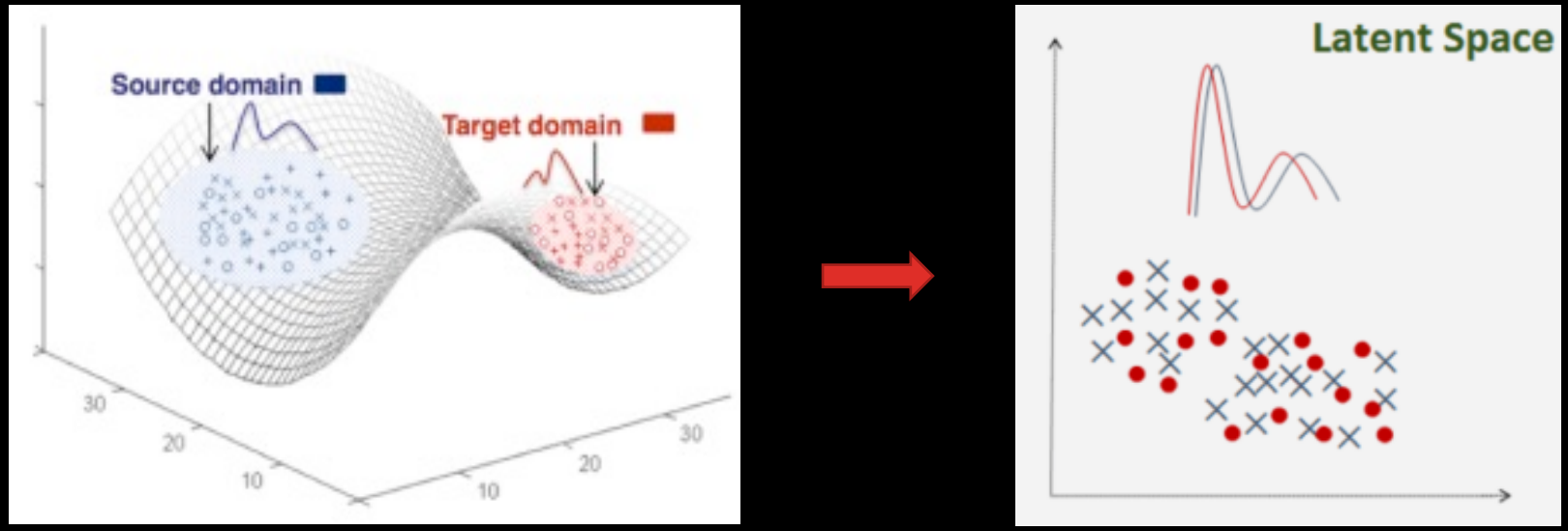
Image from Fernando+@ICCV'13

Project and align data on the subspace manifold

- Geodesic Flow Sampling (Gopalan+@ICCV'11)
- Geodesic Flow Kernel (Gong +@CVPR'12)
- Subspace Alignment (Fernando+@ICCV'13)

Crédit: Gabriela Csurka

FEATURE SPACE TRANSFORMATION



Learn a transformation to a latent space where the distributions matches

- Transfer Component Analysis (Pan+@TNN'11)
- Domain Invariant Projection (Baktashmotlagh+@ICCV'13)
- Transfer Joint Matching (Long +@CVPR'14)
- Optimal Transport (Courty+@PAMI'17) …

Crédit: Gabriela Csurka

Outline

Traditional shallow DA methods

How to exploit deep learning in DA

Main trends in Deep Domain Adaptation

Crédit: Gabriela Csurka

HOW TO EXPLOIT DEEP LEARNING IN DA

Shallow methods using deep features

- use the deep model as feature extractor
- apply any shallow DA method using these features

Using fine-tuned deep architectures

- fine-tune the deep model on the source
- apply the fine-tuned model on the target

Shallow methods using fine-tuned deep features

- fine-tune the deep model on the source
- use the fine-tuned model as feature extractor
- apply any shallow DA method using these features

Deep DA models

- specific deep architectures tailored for domain adaptation
- often initialized with a deep model fine-tuned on the source

Crédit: Gabriela Csurka

Outline

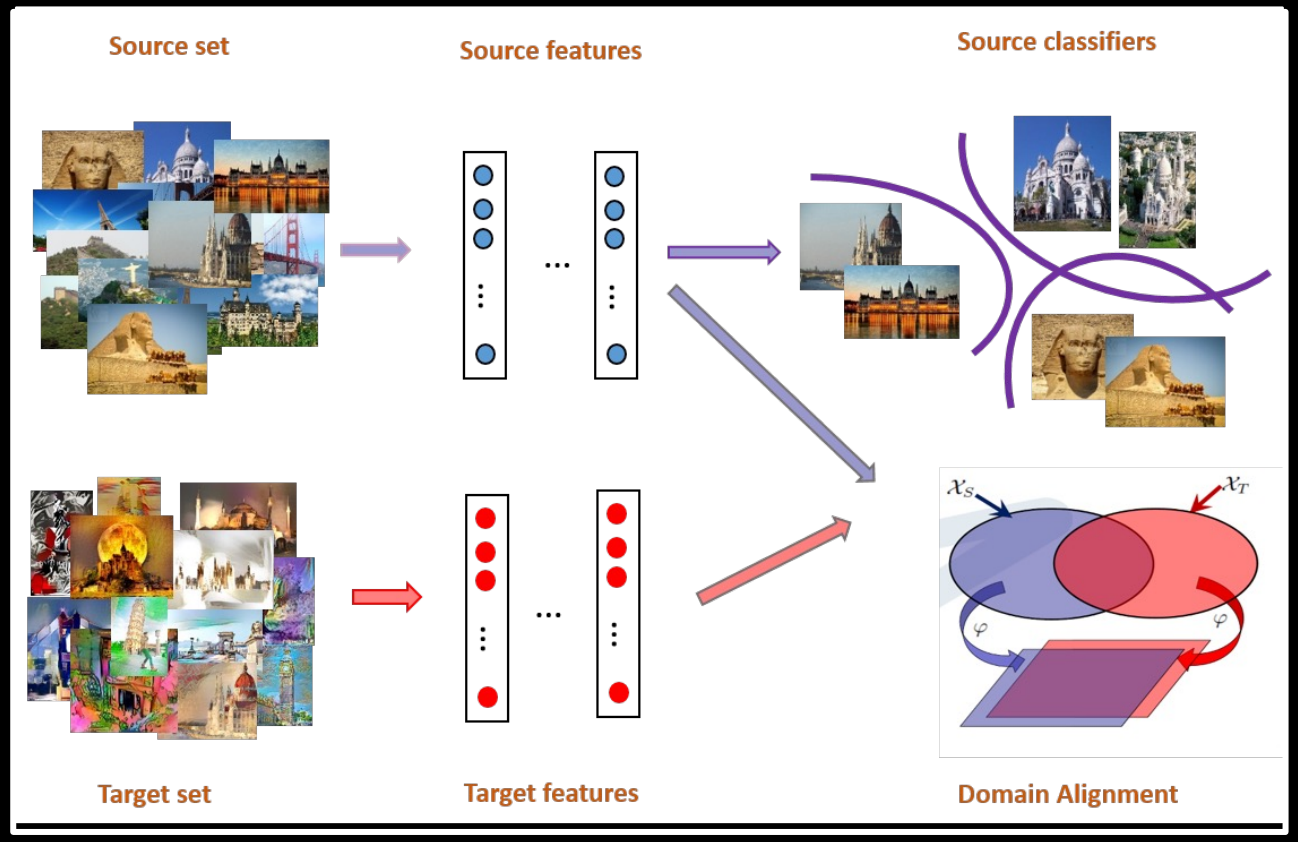
Traditional shallow DA methods

How to exploit deep learning in DA

Main trends in Deep Domain Adaptation

Crédit: Gabriela Csurka

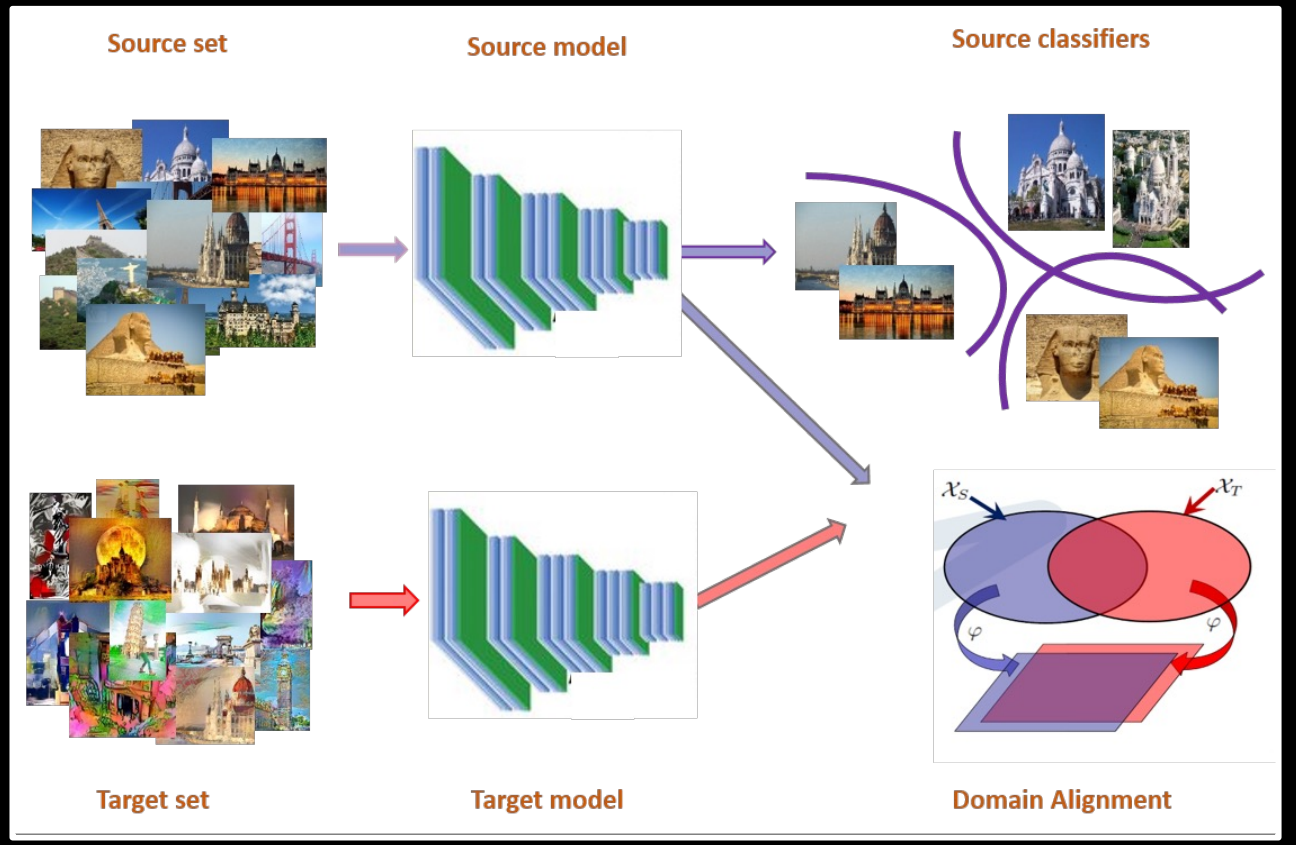
TRADITIONAL DA MODELS



Typically DA is performed on pre-extracted features by minimizing the distribution mismatch

Crédit: Gabriela Csurka

DEEP DA MODELS



They learn image representation, domain alignment and the source classifier jointly end-to-end

Crédit: Gabriela Csurka

ADVERSARIAL LEARNING

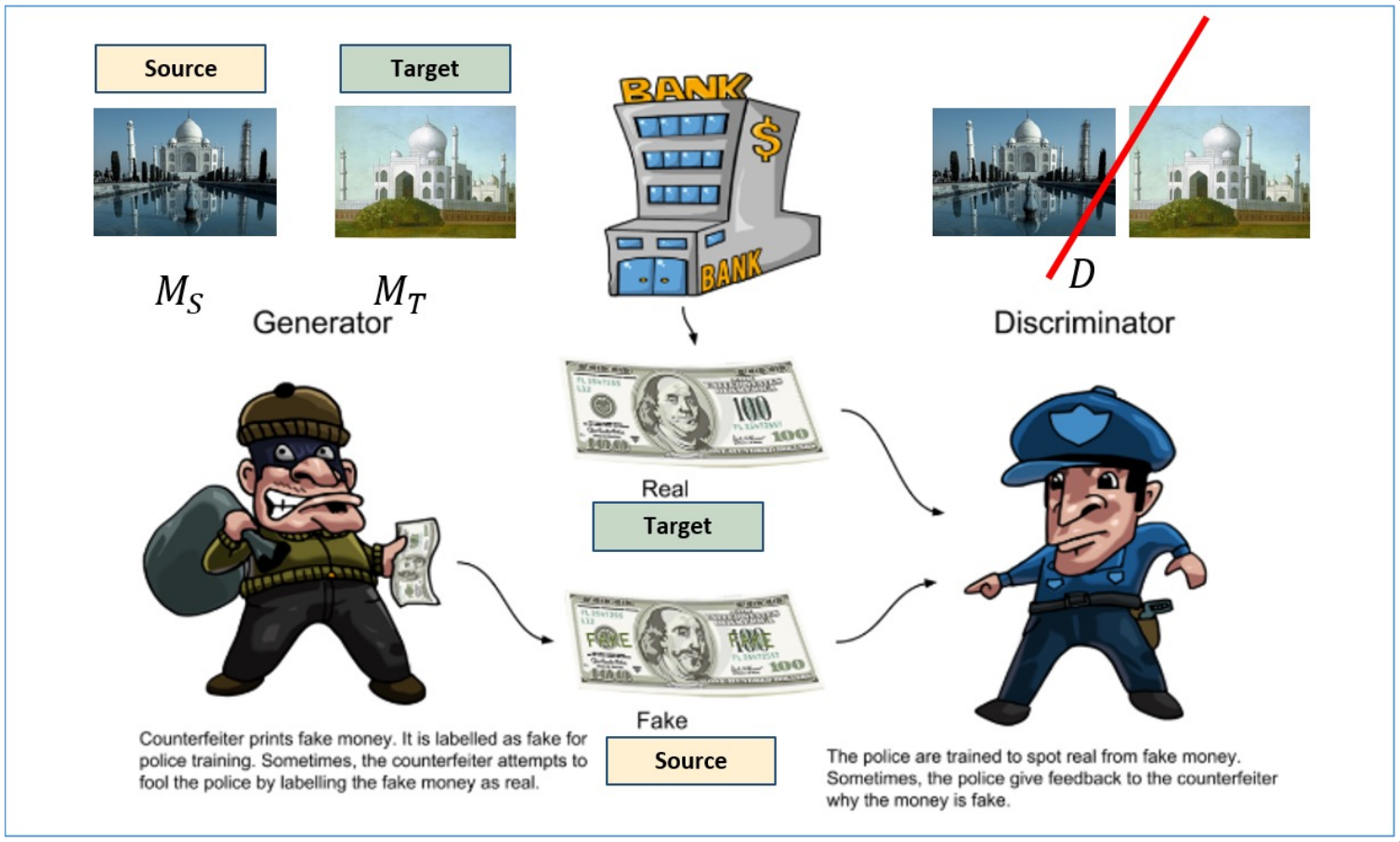


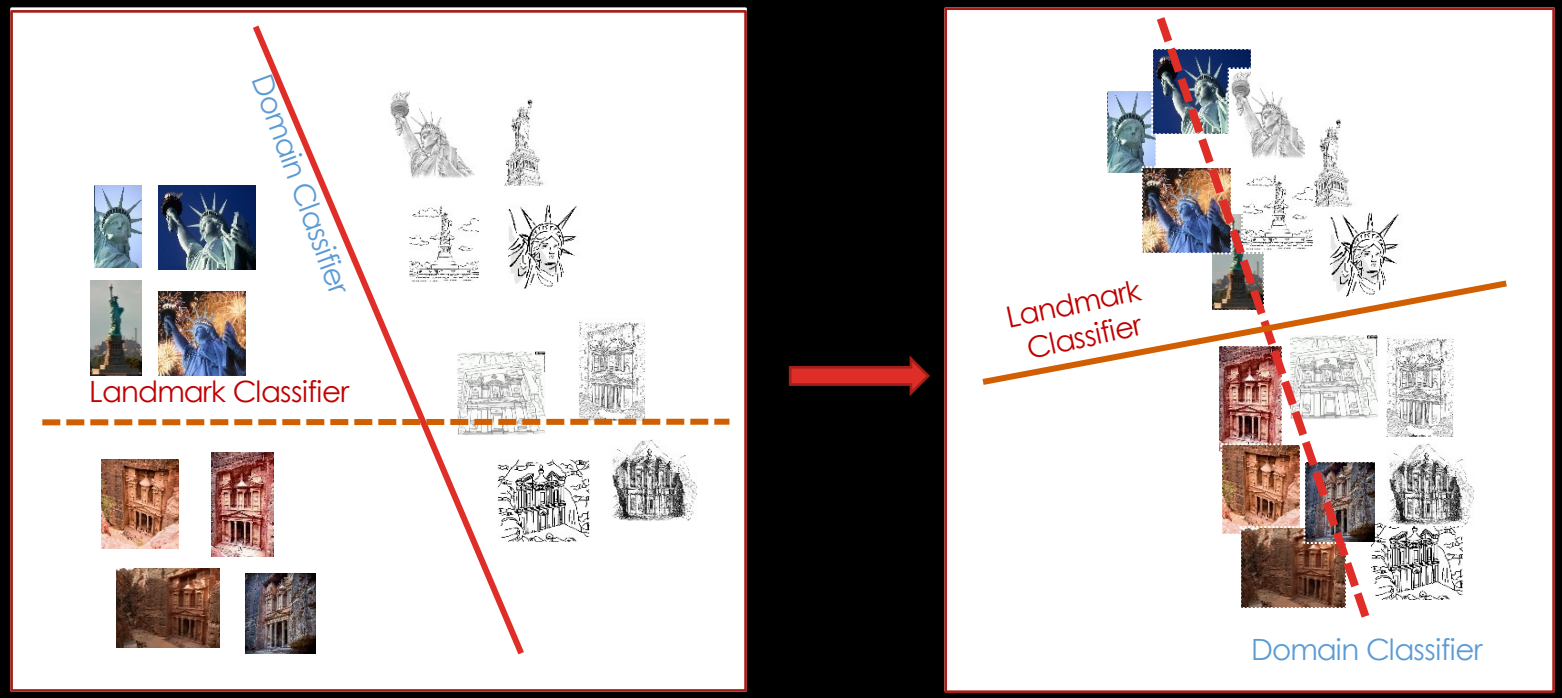
Image courtesy of Richard Gall

Minimizing the Jensen-Shannon divergence between the distributions

- Generative Adversarial Network (GAN), Goodfellow+@NeurIPS'14

Crédit: Gabriela Csurka

DOMAIN CONFUSION



Learn domain classifier and use it to confuse the domains

- Tzeng+@CVPR'17, Arjovsky+@ICML'17, Shen+@AAAI'18

Crédit: Gabriela Csurka

GRADIENT REVERSAL LAYER (GRL)

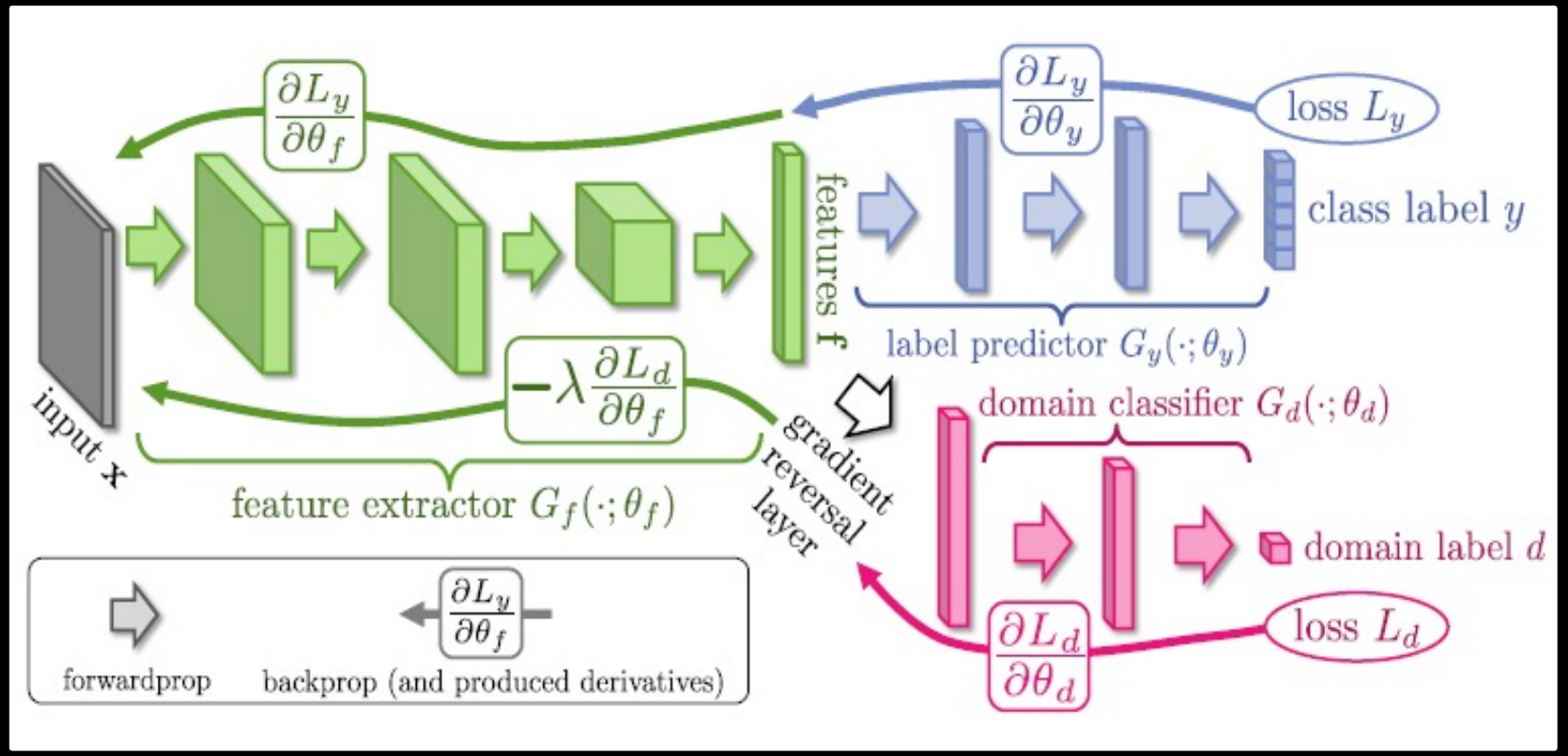


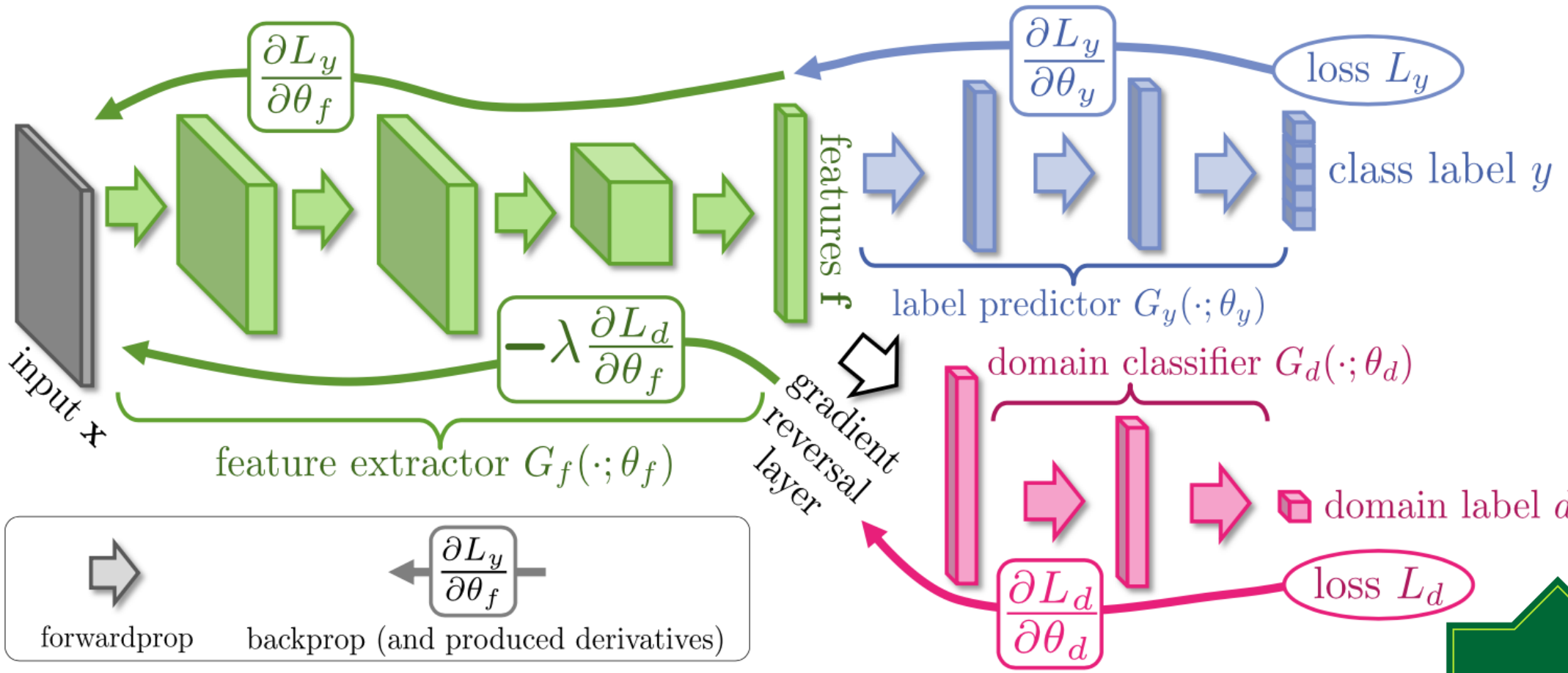
Image from Ganin+@JMLR'16

GRL layer reversing the gradient for the target during backpropagation

- Ganin+@JMLR'16, Bousmalis+@NIPS'16, Pei+@AAAI'18

Crédit: Gabriela Csurka

Unsupervised domain adaptation



Domain-Adversarial Training of Neural Networks
Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, Victor Lempitsky; ICML 2016

Article 1

NETWORK PARAMETER ADAPTATION

Domain specific network weights

- BSW (Rozantsev+@PAMI'18), RPT (Rozantsev+@CVPR'18)

Domain specific batch normalization

- AutoDial (Carlucci+@ICCV'18), AdaBN (Li+@ICLR'18)

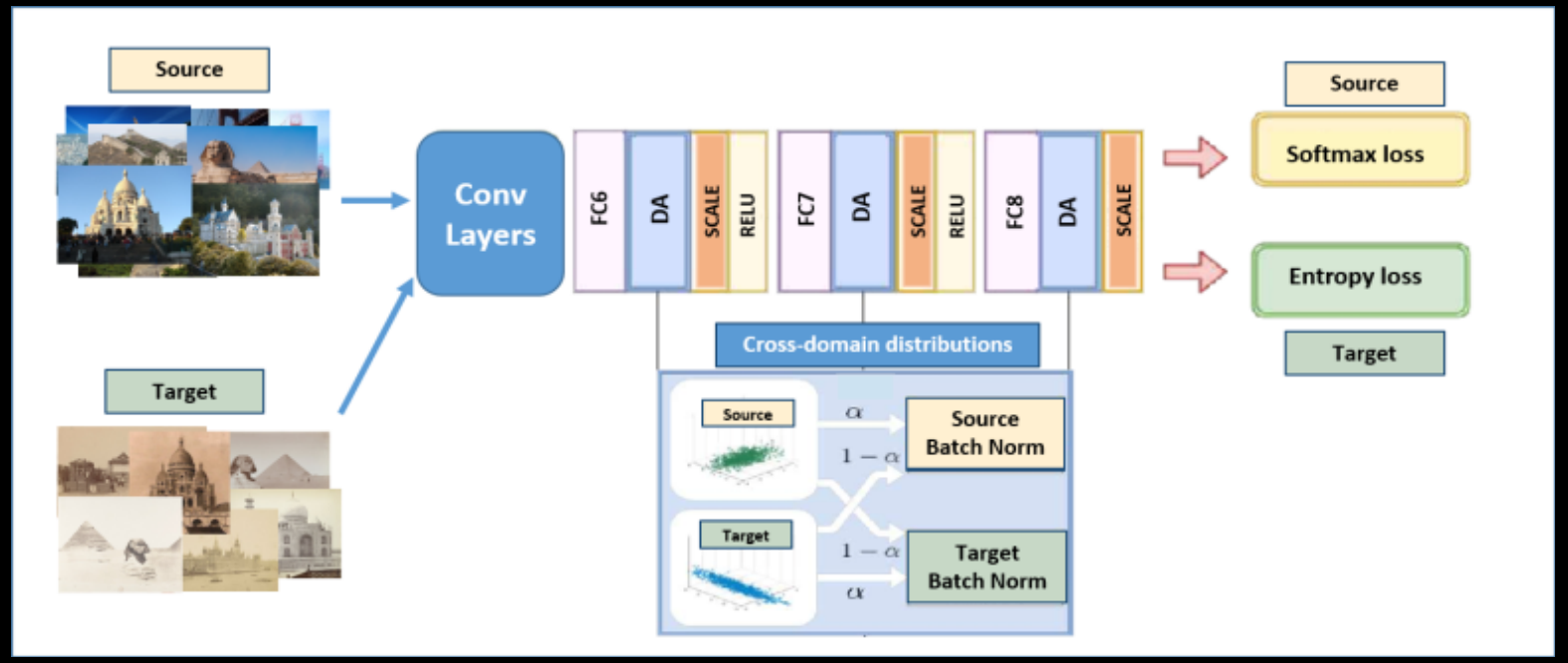


Image based on Carlucci+@ICCV'18

Crédit: Gabriela Csurka

Batch normalization (also known as **batch norm**) is a method used to make training of artificial neural networks faster and more stable through normalization of the layers' inputs by re-centering and re-scaling. It was proposed by Sergey Ioffe and Christian Szegedy in 2015.

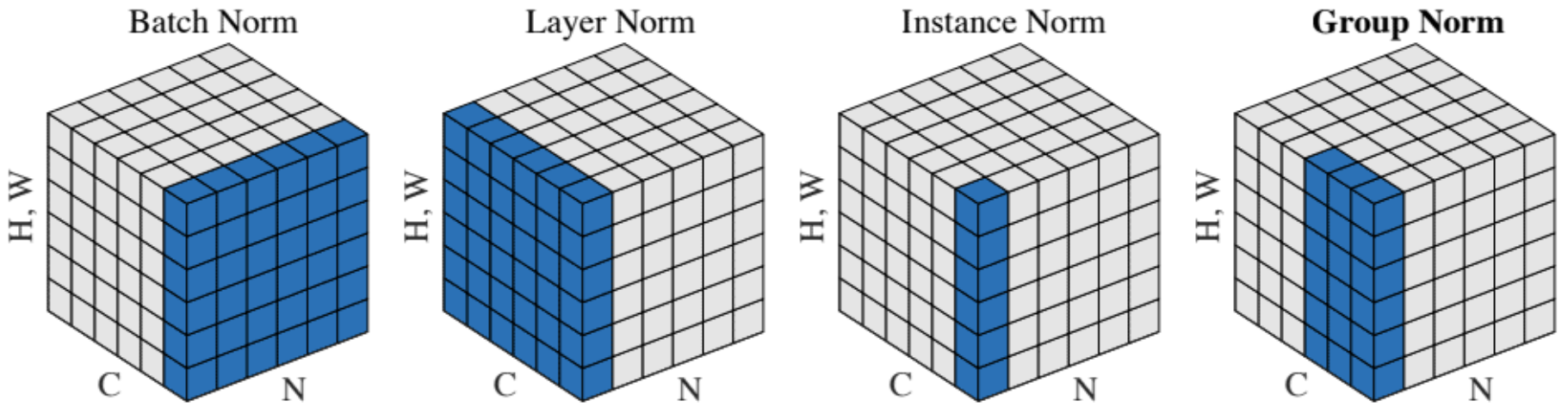


Figure 2. **Normalization methods**. Each subplot shows a feature map tensor, with N as the batch axis, C as the channel axis, and (H, W) as the spatial axes. The pixels in blue are normalized by the same mean and variance, computed by aggregating the values of these pixels.

Image source: https://scortex.io/batch-norm-folding-an-easy-way-to-improve-your-network-speed/

PIXEL LEVEL DOMAIN STYLE TRANSFER

Paired style transfer as pre-processing

- Csurka+@TASKCV'17, Thomas+@ACCV'19, Jackson+@CVPR-WS'19

Unpaired image-to-image style transfer learning

- Yoo+@ECCV'16, Zhu+@ICCV'17, Bousmalis+@CVPR'17, Murez+@CVPR'18

Crédit: Gabriela Csurka

Bilan

- L'apprentissage supervisé d'un modèle à partir de rien (« *from scratch* ») est coûteux
  - en données labelisées
  - en mémoire
  - en temps de calcul

- Cependant, nous n'avons pas toujours accès à de grandes quantités de données pour le domaine qui nous intéresse.

- Le temps de calcul peut également être rédhibitoire.

- Une bonne stratégie est de réutiliser des modèles entrainés à partir de données de tâches voisines (*task transfer*) ou de domaines voisins (*domain adaptation*), afin de réduire le coût de l'apprentissage d'un modèle sur une tâche connexe, et éventuellement d'améliorer les résultats obtenus par ce nouveau modèle.

Dans la suite

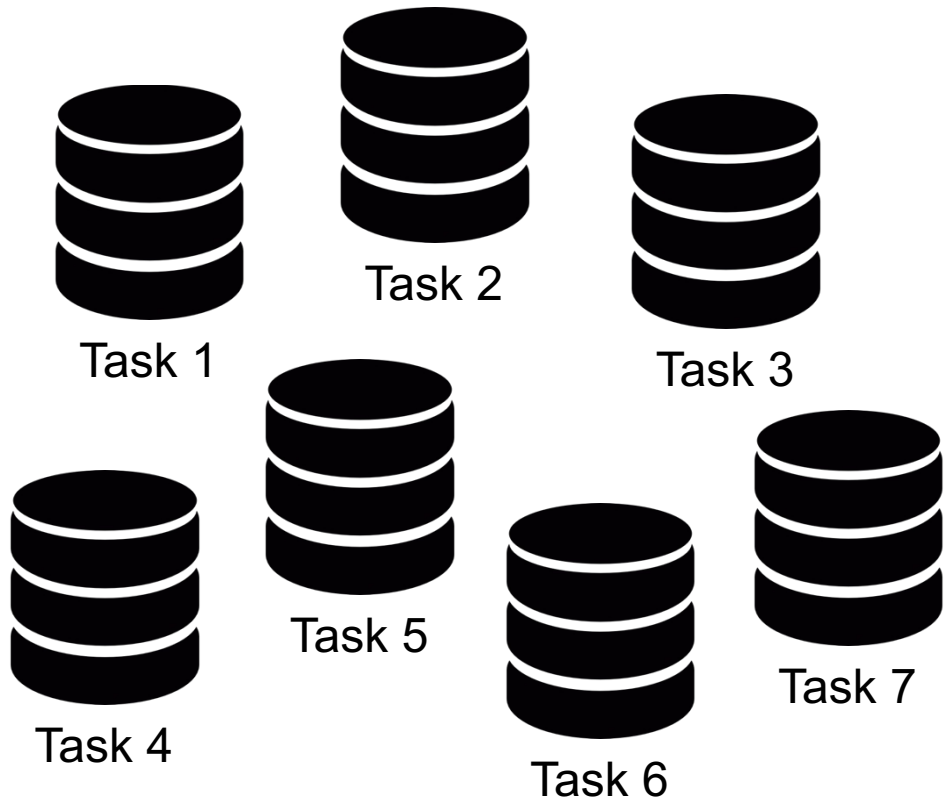
Que faire si nous avons accès à beaucoup de données de pré-entrainement,
mais que celles-ci ne sont pas annotées ?

L’apprentissage auto-supervisé

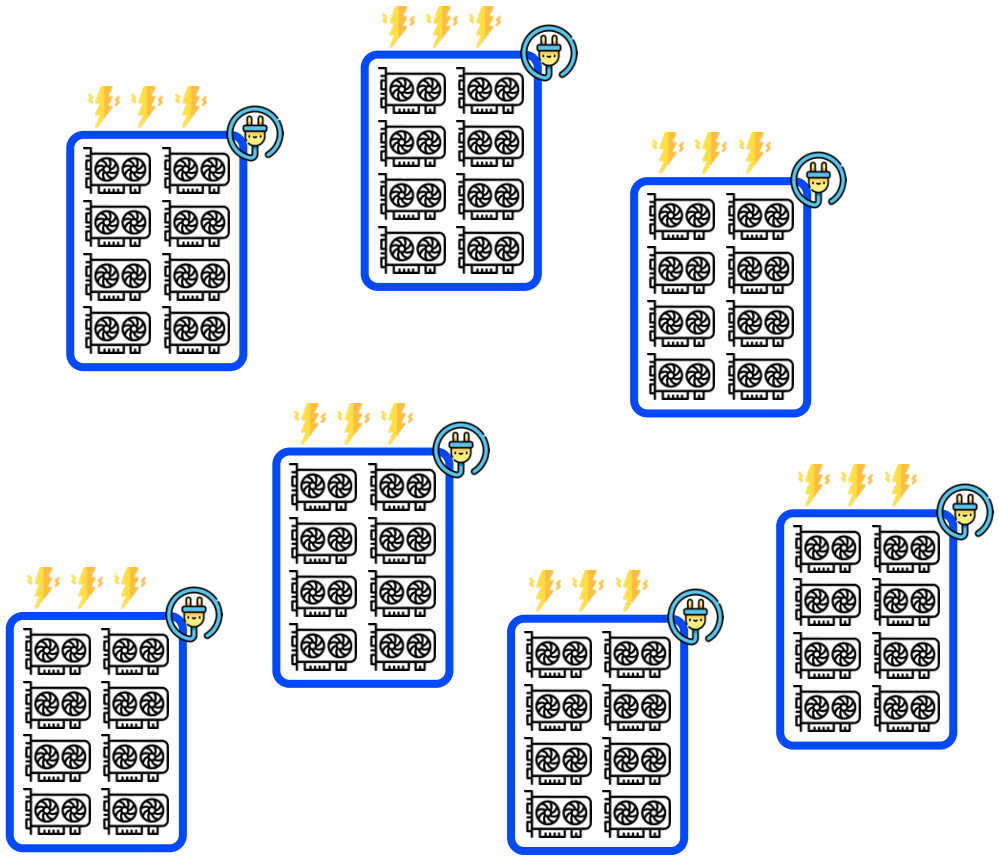
Apprentissage continu de représentations visuelles

2023-2024

multiple tasks

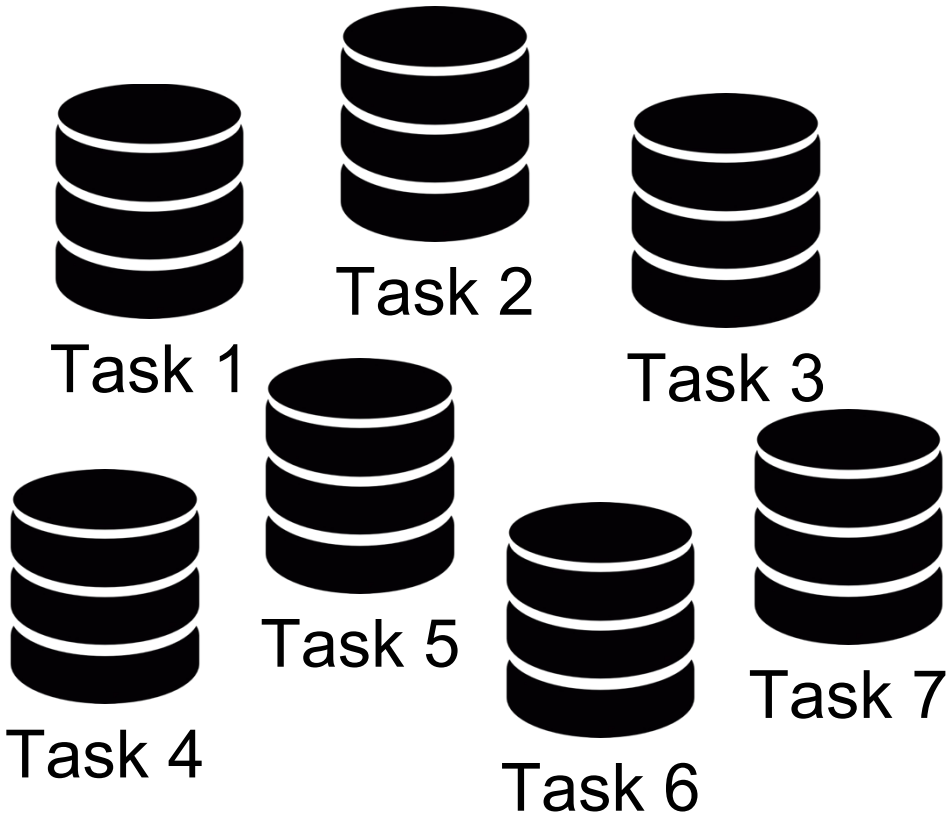


Collection of tasks



Waste of compute

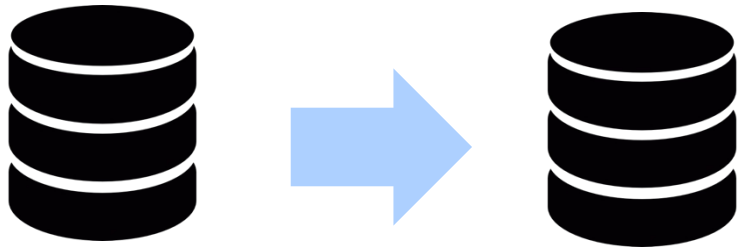
Lifelong learning



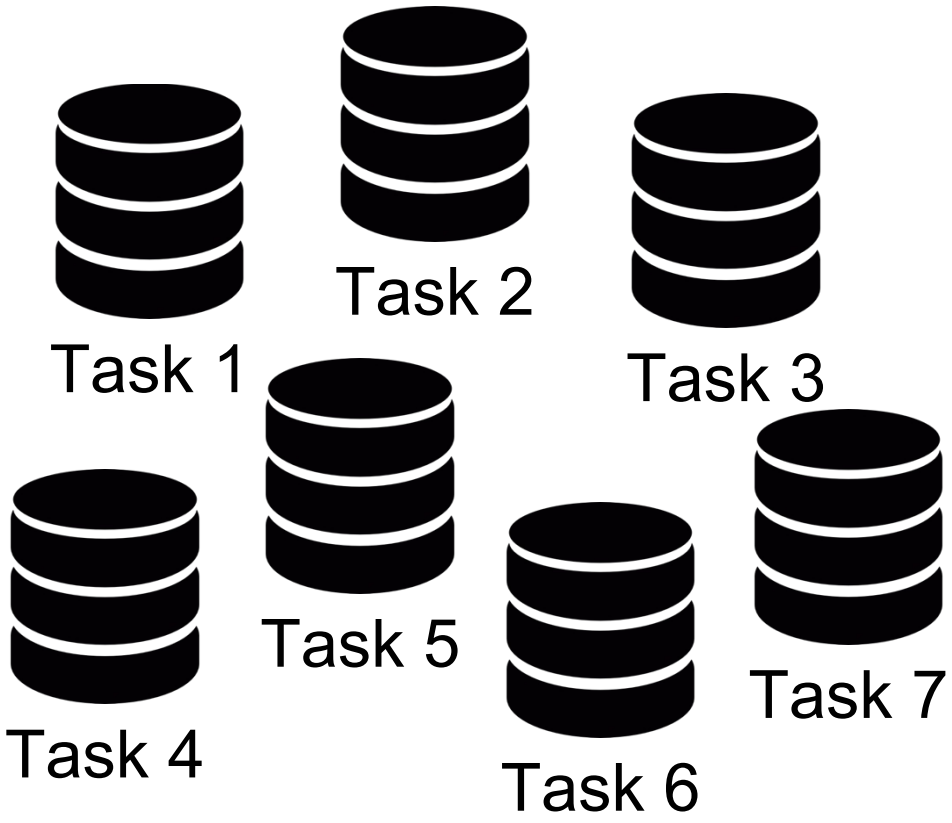
What about Lifelong learning?

Most simple case

Is Task A useful for Task B?



Lifelong learning of generic visual representations

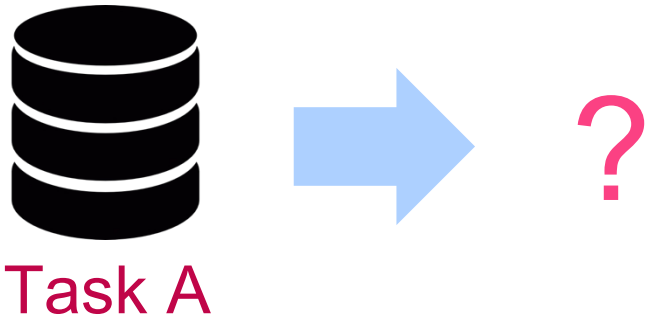


What about Lifelong learning?

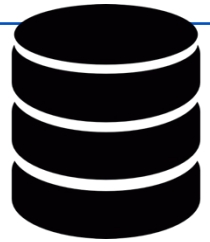
What if we don't know in advance what the target task will be?

This is what **pretraining** is about

- How should we train on Task A so it will transfer better later



- Task A becomes a **proxy / pretext task**

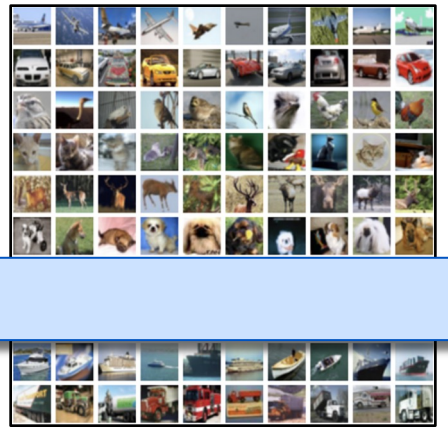


Fully-Supervised Classification

Images + labels

Proxy task

IMAGENET



Reducing annotation cost

Model

Visual representations

Target tasks

Image Classification

Object Detection

Instance Segmentation

Image Retrieval

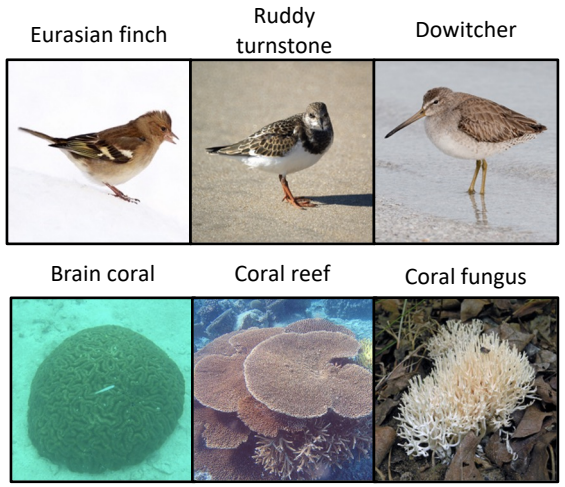
Learning transferable visual representations

Reducing annotation cost

Fully-Supervised

fine-grained annotations

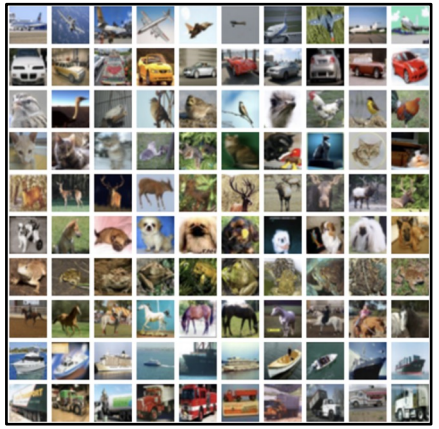
expert knowledge



Self-supervised

annotation-free images

no annotation required



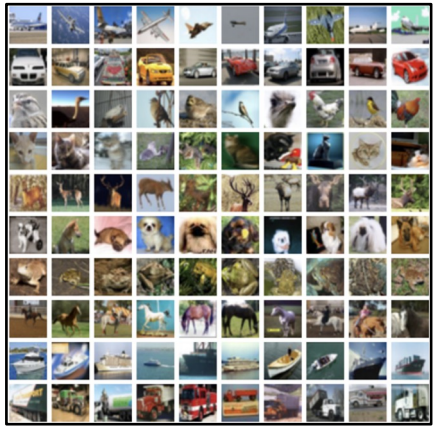
Learning transferable visual representations

No supervision

Reducing annotation cost

Fully-Supervised
fine-grained annotations
expert knowledge

Self-supervised
annotation-free images
no annotation required



~~labels~~

Définition

Pour Wikipedia

- L'apprentissage auto-supervisé est une méthode d'apprentissage automatique en intelligence artificielle. Cette méthode d'apprentissage est devenue le standard dans le domaine du traitement langage naturel.

- Yann Le Cun, lauréat du Prix Turing 2018 pour ses travaux sur l'apprentissage profond compare cette méthode à du « remplissage de trou » et donne pour cela un exemple simple : elle permet d'analyser les images de début et de fin d'une séquence vidéo et d'entraîner un algorithme à prédire les images intermédiaires.

Définition de l’apprentissage auto-supervisé

Ensemble de techniques d’apprentissage non-supervisées, qui empruntent aux techniques du registre de l’apprentissage supervisé, mais remplacent la vérité terrain traditionnellement obtenue par des annotations manuelles par une vérité terrain produite automatiquement à partir des données elles-mêmes. L’apprentissage vise à apprendre à prédire ces annotations automatiques. On parle alors de tâche prétexte.

Self-supervised learning (or 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 still be a hard task
  - Should enable us to learn aspects of the visual input/world
- No annotations required
  - Scalability: use any image/video - no need for labels
  - Flexibility: find the data that fits your downstream task

[slide borrowed from Yannis Kalantidis]

Self-supervised learning (or 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 still be a hard task
  - Should enable us to learn aspects of the visual input/world
- No annotations required
  - Scalability: use any image/video - no need for labels
  - Flexibility: find the data that fits your downstream task

Does this mean that I don’t need
to care about data acquisition anymore?

[slide borrowed from Yannis Kalantidis]

Be extra careful of what data you use

Does this mean that I don't need
to care about data acquisition anymore?

You need to care even more!!

The myth of "un-curated" datasets

- SSL amplifies biases in the dataset (as much, if not more as supervised learning)
- The absence of labels makes practitioners look at data even less
- There is no such thing as "un-curated datasets", working without labels is not an excuse

[slide borrowed from Yannis Kalantidis]

Compléments de motivation pour l'apprentissage auto-supervisé

- Problèmes avec **l'apprentissage supervisé**, lorsque celui-ci est utilisé pour le **pré-entrainement** de modèles ou de représentations pour une autre tâche cible:
  - La collection ou la création d'annotations manuelles représente un coût non négligeable, souvent rédhibitoire (e.g. Pascal VOC, ImageNet).
  - Les données brutes ont une structure bien plus riche que les annotations fournies. De nombreux aspects de cette structure sont totalement ignorés par les annotations.
  - Les algorithmes purement supervisés sont souvent surspécialisés, et les modèles obtenus trop spécifiques aux annotations fournies. Cela entrave leurs propriétés de généralisation.
- **L'apprentissage "auto-supervisé"** semble être une alternative prometteuse pour le pré-entrainement, car les données elles-mêmes sont utilisées pour la supervision de l'algorithme d'apprentissage, et remplacent un ensemble d'annotations plus ou moins arbitraire.

Le pari de la crème glacée (*Gelato bet*)

En 2014, le détecteur d'objets R-CNN obtient des résultats impressionnants sur la base PASCAL VOC.
Cela représente une grande réussite pour le *deep learning* à l'époque.

Une des raisons de cette réussite:
La base PASCAL VOC, avec ses quelques milliers d'images, est trop petite pour entrainer un réseau convolutionnel *from scratch.*
Le réseau a donc besoin d'être préalablement pré-entrainé sur ImageNet, puis ensuite ajusté (*fine-tuned*) sur PASCAL VOC.

La grande différence entre les labels des bases ImageNet et PASCAL VOC a inspiree à **Alyosha Efros** la suggestion suivante:
Et si la partie importante du pré-entrainement était liée uniquement aux images d'ImageNet et pas à ses labels?

Il fit donc le pari suivant:

"If, by the first day of autumn (Sept 23) of 2015, a method will exist that can match or beat the performance of R-CNN on Pascal VOC detection, without the use of any extra, human annotations (e.g. ImageNet) as pre-training, Mr. Malik promises to buy Mr. Efros one (1) gelato (2 scoops: one chocolate, one vanilla)."

Il a perdu son pari, mais l'apprentissage auto-supervisé est tout de même en train de jouer un rôle crucial en vision par ordinateur, et dans d'autres domaines de l'apprentissage automatique.

Source: https://people.eecs.berkeley.edu/~efros/gelato_bet.html

Apprentissage auto-supervisé – premières méthodes

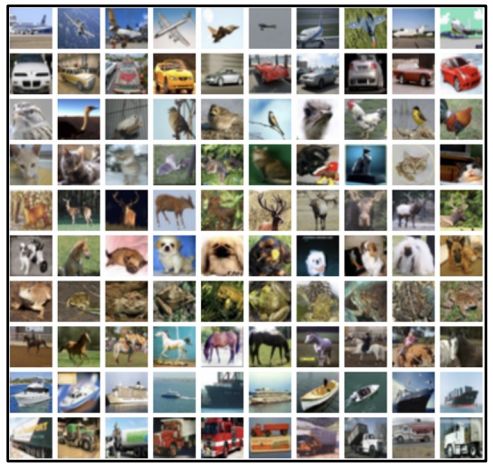
Apprentissage continu de représentations visuelles

2023-2024

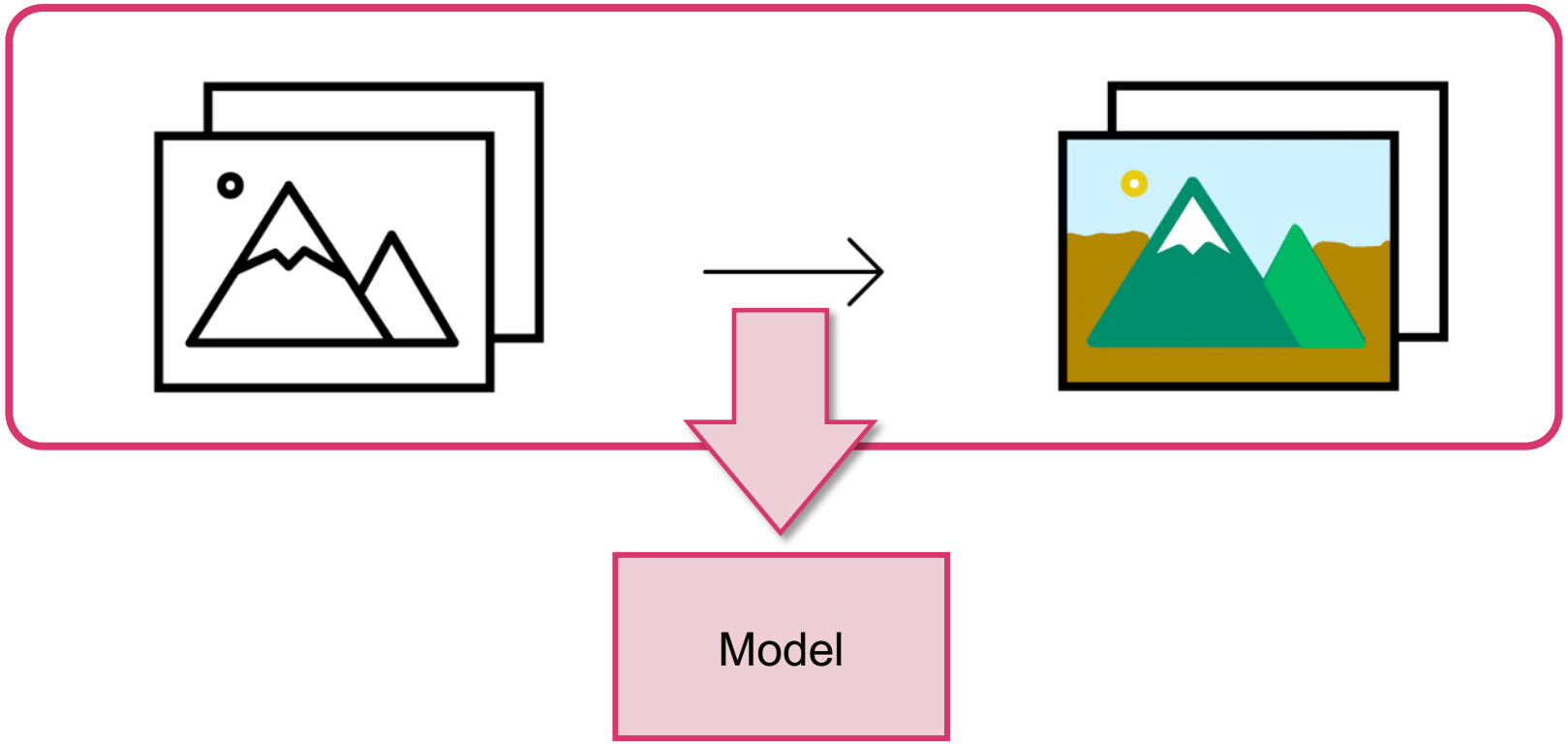
Self-supervised learning – early methods

Pretext task = Image colorization

One of the first generative pretext tasks



~~labels~~

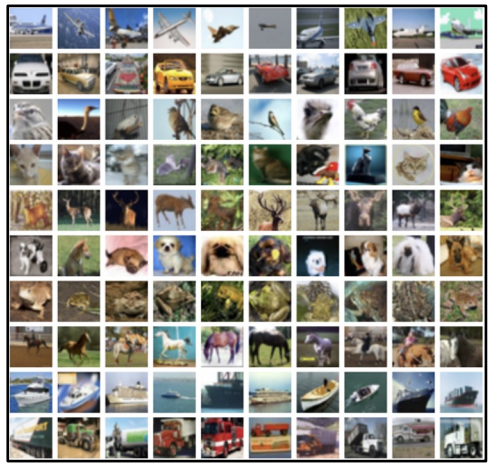


Zhang et al. 2016

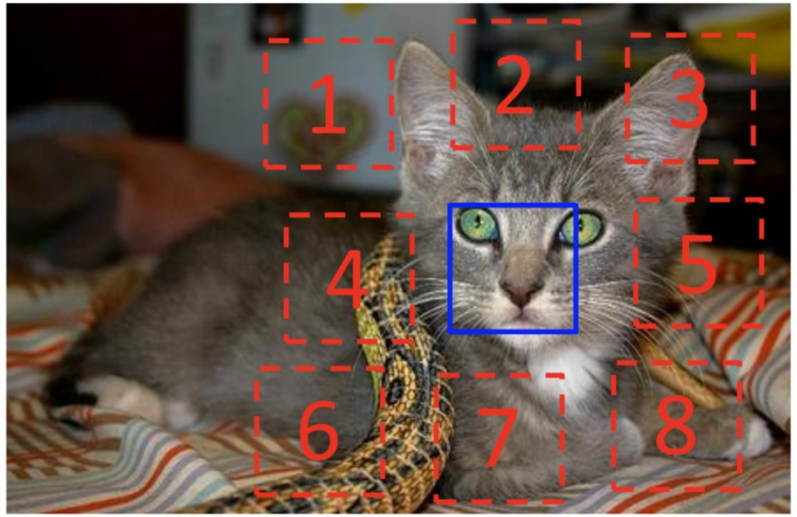
Self-supervised learning – early methods

Pretext task = Context prediction

One of the first discriminative pretext tasks



~~labels~~



?

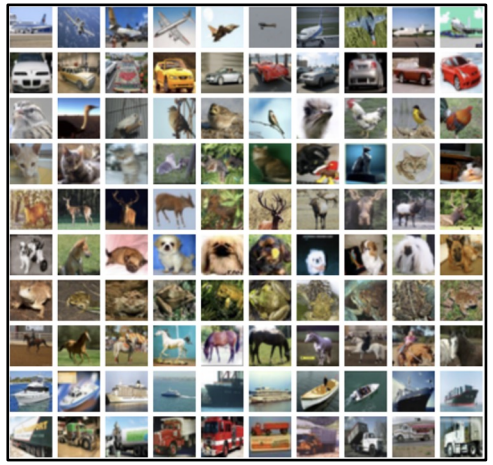
Example:

Model

Unsupervised Visual Representation Learning by Context Prediction
Carl Doersch, Abhinav Gupta, and Alexei A. Efros. ICCV 2015

Self-supervised learning – early methods

Pretext task = Rotation prediction



labels

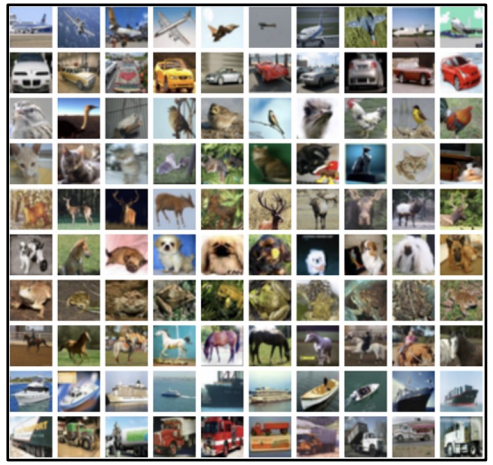
?

Article 2

Model

Unsupervised representation learning by predicting image rotations.
Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

Self-supervised learning – early methods



~~labels~~

2014	2015	2016	2017
Dosovitskiy et al. (Exemplar CNN)		*Pathak et al. (inpainting)*	
Doersh et al. (Context pred.)		*Noroozi et al. (jigsaw)*	
Wang et al. (video)		*Zhang et al. (colorization)*	
Agrawal et al. (motion)		*Larsson et al. (colorization)*	
Jayaraman et al. (motion)		*Owens et al. (sound)*	
		Zhang et al. (split-brain)	
		Bojanowski & Joulin. (NAT)	
		Doersh et al. (multi-task)	
		Pathak et al. (motion & segment)	
		Yang et al. (clusters)	
		Donahue et al. (BiGAN)	
		Dumoulin et al. (BiGAN)	

Slide credit: Mathilde Caron