Comprendre les données visuelles à grande échelle

ENSIMAG
2020-2021

Karteek Alahari & Diane Larlus
14 janvier 2021

Organisation du cours

Jeudi 15/10/2020	09h45	11h15	C005-Amphi C (V)	LARLUS Diane
Jeudi 22/10/2020	09h45	11h15	D211 (V)	ALAHARI Karteek
Jeudi 05/11/2020	09h45	11h15	Zoom	ALAHARI Karteek
Jeudi 12/11/2020	09h45	11h15	Zoom	LARLUS Diane
Jeudi 26/11/2020	09h45	11h15	Zoom	ALAHARI Karteek
Jeudi 03/12/2020	09h45	11h15	Zoom	ALAHARI Karteek
Jeudi 10/12/2020	09h45	11h15	Zoom	LARLUS Diane
Jeudi 17/12/2020	09h45	11h15	Zoom	LARLUS Diane
Jeudi 07/01/2021	09h45	11h15	Zoom	LARLUS Diane
Jeudi 14/01/2021	09h45	11h15	Zoom	LARLUS Diane
Jeudi 21/01/2021	09h45	11h15	Zoom	ALAHARI Karteek
Jeudi 28/01/2021	09h45	11h15	Zoom	ALAHARI Karteek

Articles 4 & 5 & 6

Comprendre les données visuelles à grande échelle
14 janvier 2021

Article 4

You Only Look Once:
Unified, Real-Time Object Detection

Joseph Redmon*, Santosh Divvala*†, Ross Girshick¶, Ali Farhadi*†
University of Washington*, Allen Institute for AI†, Facebook AI Research¶
http://pjreddie.com/yolo/

Abstract

We present YOLO, a new approach to object detection. Prior work on object detection repurposes classifiers to perform detection. Instead, we frame object detection as a regression problem to spatially separated bounding boxes and associated class probabilities. A single neural network predicts bounding boxes and class probabilities directly from full images in one evaluation. Since the whole detection pipeline is a single network, it can be optimized end-to-end directly on detection performance.

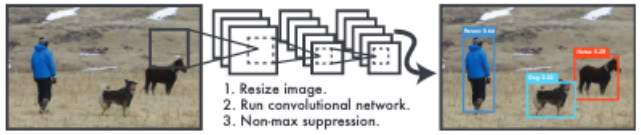


Figure 1: The YOLO Detection System. Processing images with YOLO is simple and straightforward. Our system (1) resizes the input image to 448 × 448, (2) runs a single convolutional network on the image, and (3) thresholds the resulting detections by the model's confidence.

TLDR:

Quelle architecture de réseau de neurones pour qu'un unique réseau de neurones soit capable de

- sélectionner une région dans l'image,
- classifier l'objet qu'elle contient,
- régresser les coordonnées précises de sa boite englobante ?

Questions pour vous

- Pourquoi avez-vous tous choisi cet article ?
- Faut-il pré-entrainer le réseau ? Pourquoi ?
- Quel niveau de précision peut-on obtenir sur ce type d'approche? Quelles sont les facteurs limitants?
- Comment le modèle gère-t-il les occlusions sévères?
- Est-ce facile d'ajouter de nouvelles classes? Qu'est-ce qui doit être modifié ?

Article 5

Learning Two-Branch Neural Networks for Image-Text Matching Tasks

Liwei Wang, Yin Li, Jing Huang, Svetlana Lazebnik

Abstract—Image-language matching tasks have recently attracted a lot of attention in the computer vision field. These tasks include image-sentence matching, i.e., given an image query, retrieving relevant sentences and vice versa, and region-phrase matching or visual grounding, i.e., matching a phrase to relevant regions. This paper investigates two-branch neural networks for learning the similarity between these two data modalities. We propose two network structures that produce different output representations. The first one, referred to as an *embedding network*, learns an explicit shared latent embedding space with a maximum-margin ranking loss and novel neighborhood constraints. Compared to standard triplet sampling, we perform improved neighborhood sampling that takes neighborhood information into consideration while constructing mini-batches. The second network structure, referred to as a *similarity network*, fuses the two branches via element-wise product and is trained with regression loss to directly predict a similarity score. Extensive experiments show that our networks achieve high accuracies for phrase localization on the Flickr30K Entities dataset and for bi-directional image-sentence retrieval on Flickr30K and MSCOCO datasets.

TLDR:

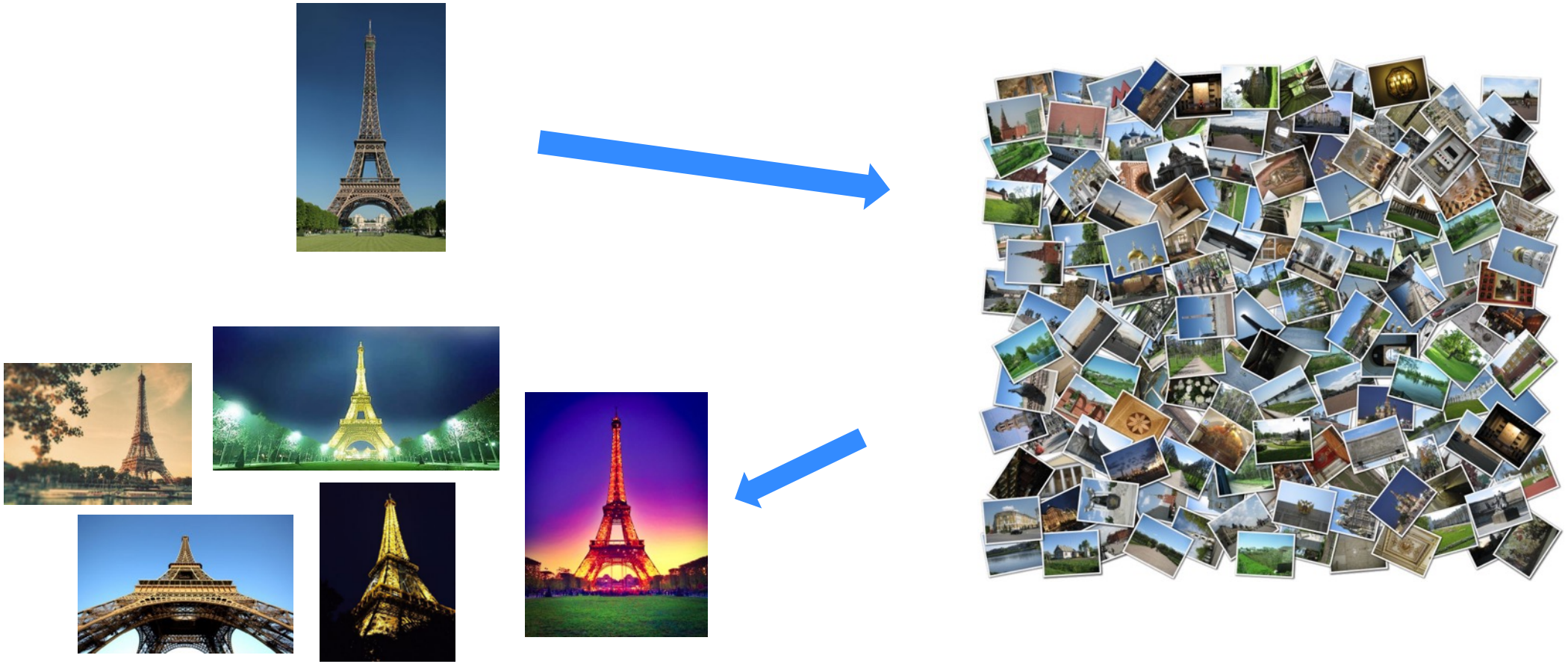
Un bon espace de représentation pour de la recherche cross-modale est un espace qui permet de représenter aussi bien les images et le texte. Mais il est important de conserver les relations naturelles au sein d’une modalité donnée.

Notions abordées dans cet article

- Notion de recherche cross-modale
  - Text-to-image & image-to-text
- Notion de plongement (*embedding*)
  - Plongement : espace de représentation
  - Plongement joint: *(joint embedding)*: espace de représentation convenable pour plusieurs modalités, ex: pour des image & pour leurs légendes
- Apprentissage avec une fonction de coût basée sur des triplets
  - Même principe que ce que nous avons vu dans le cours sur la recherche d'image (*image retrieval*) avec des représentations profondes, mais sinon deux modalités différentes sont considérées dans la même fonction de coût

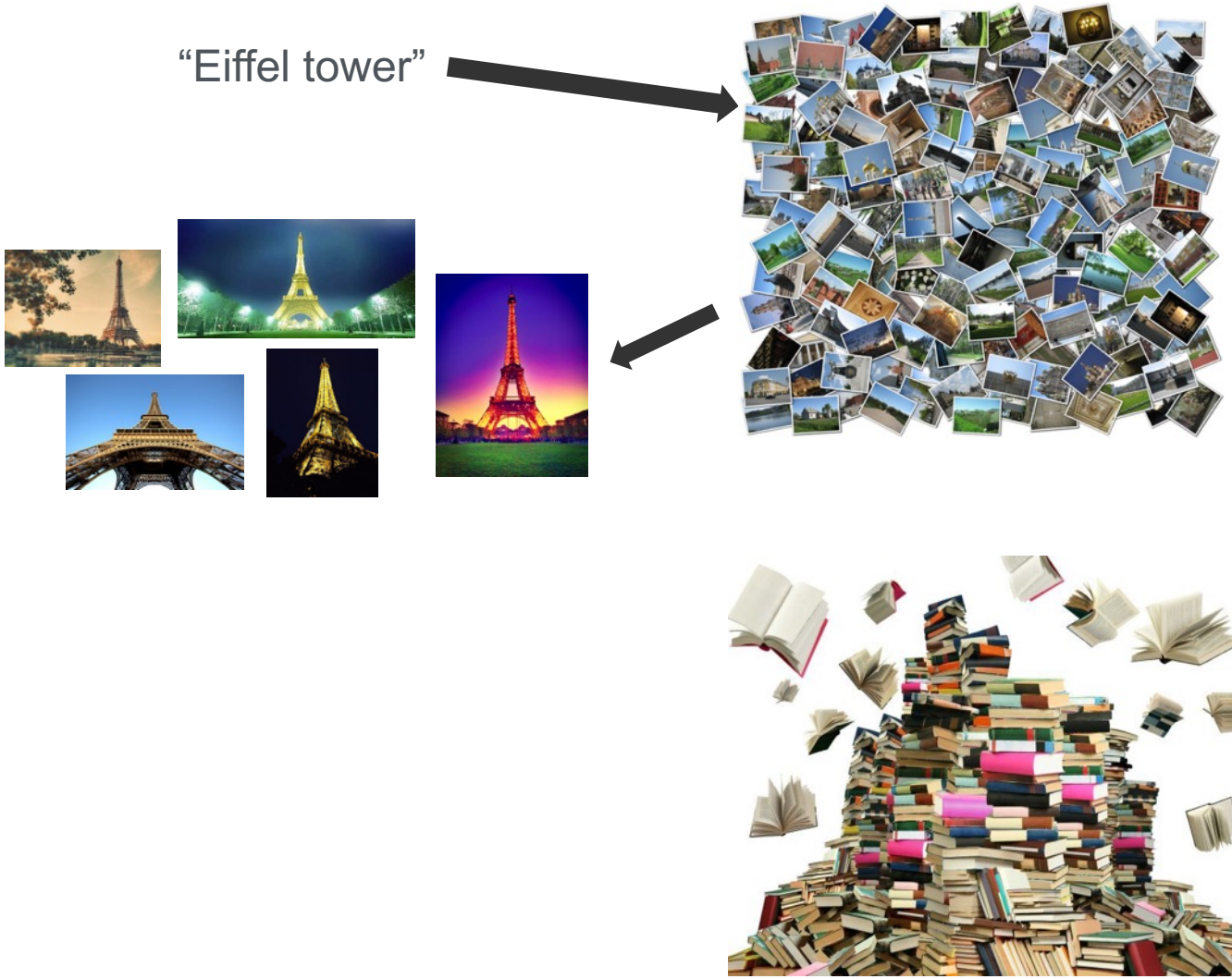
La recherche visuelle classique

En général, la recherche d’images s’intéresse principalement à la recherche d’instances



La recherche cross-modale

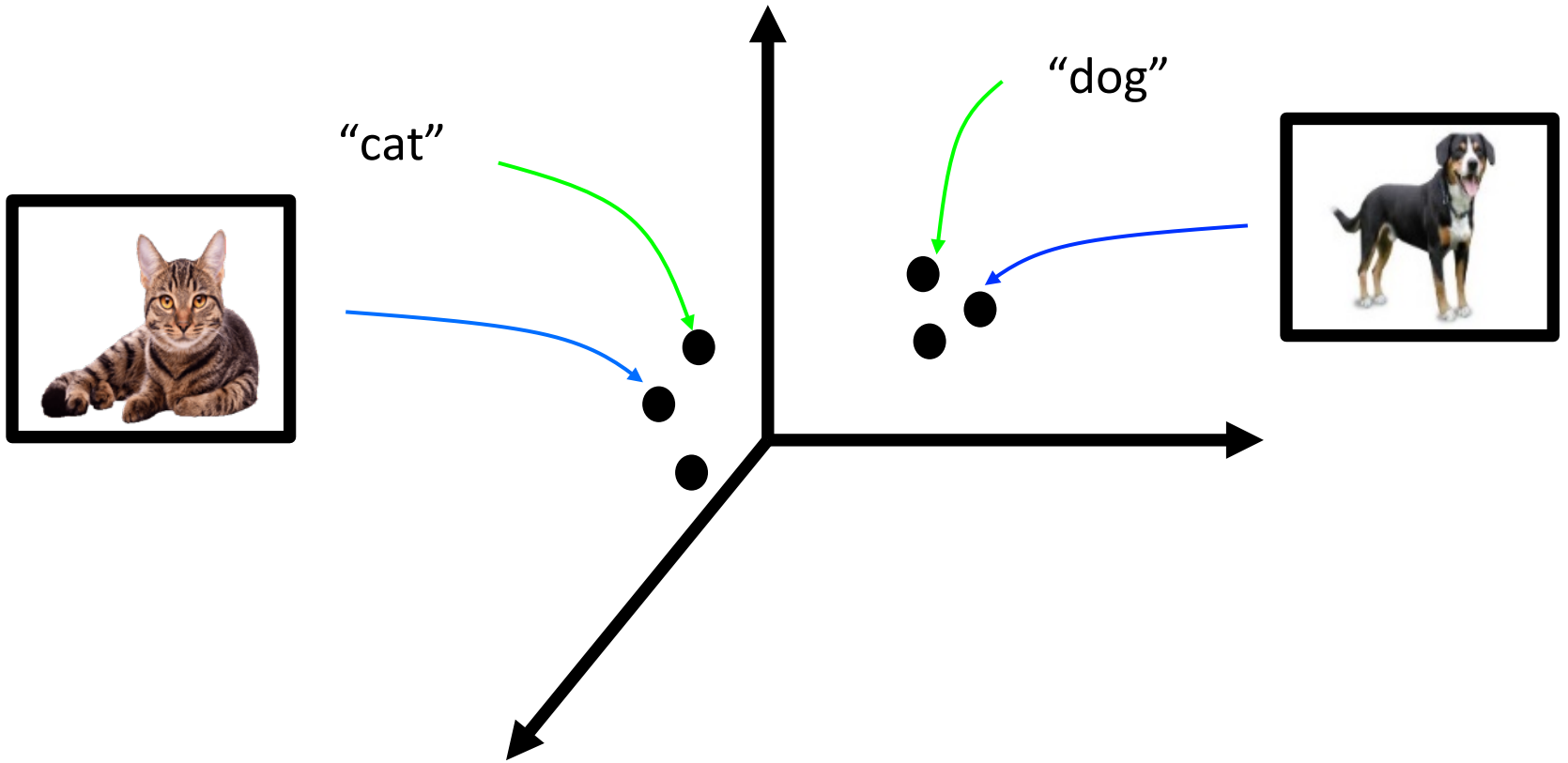
- Exemple of text-to-image and image-to-text retrieval



NAVER LABS Europe

La recherche cross-modale

En général on construit un plongement joint. Ici exemple de deux modalités, et deux « concepts sémantiques » différents: chat & chien.



Exemple de base d'entrainement: (Image,caption) pairs

Flickr 30k dataset

- 30 000 images
- associated to 158 000 captions

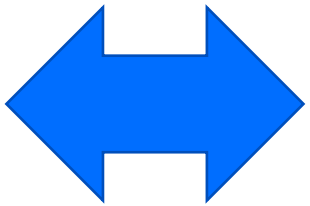
A woman in a green sweater and striped shirt leans out of a window and holds a camera in front of her face , directing it at the photographer .

Woman wearing green jacket taking a picture with a camera .

A woman with glasses staring into the camera at me .

A woman wearing glasses holds a camera to her face .

A woman is taking a picture



Learning an embedding with a triplet loss

Triplet loss – text-to-image

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m - \theta_q^T \phi_+ + \theta_q^T \phi_-)$$

q

People playing hockey on ice

d+

d-

positive

negative

Learning an embedding with a triplet loss

Triplet loss – image-to-text

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m - \phi_q^T \dot{\theta}_+ + \phi_q^T \dot{\theta}_-)$$

q

d+

d-

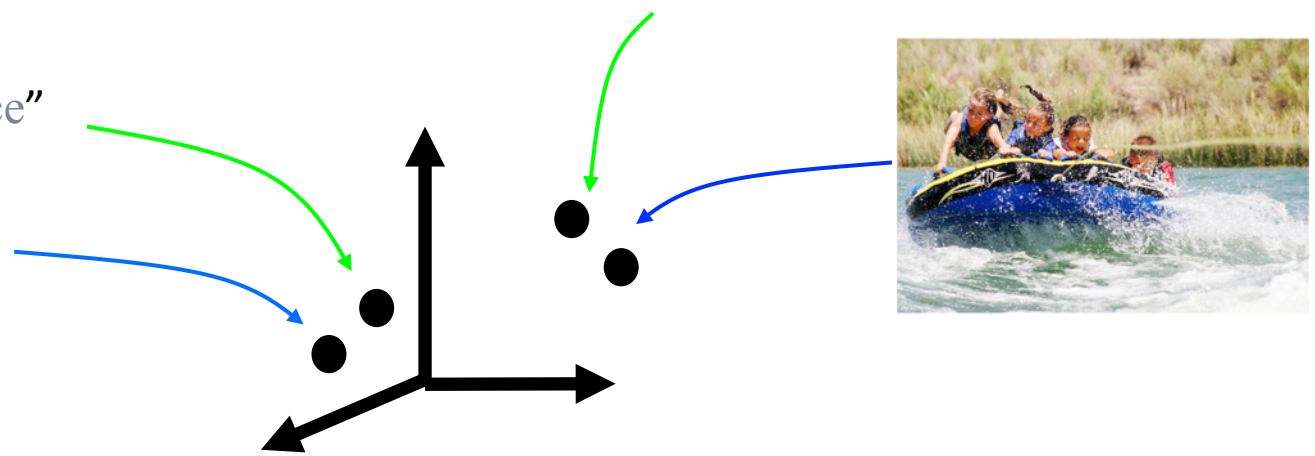
Several girls on a raft floating in rough waters

People playing hockey on ice

Cross-modal retrieval

"Several girls on a raft floating in rough waters"

"People playing hockey on ice"



Cross-modal ranking loss

$$L_{t1}(q, d^+, d^-) = \frac{1}{2}\max(0, m + \phi_q^T\theta_+ - \phi_q^T\theta_-)$$

$$L_{t2}(q, d^+, d^-) = \frac{1}{2}\max(0, m + \theta_q^T\phi_+ - \theta_q^T\phi_-)$$

Within-modal ranking loss

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m + \phi_q^T\phi_+ - \phi_q^T\phi_-)$$

$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m + \phi_q^T\phi_+ - \phi_q^T\phi_-)$$

Standard pour la construction d'un plongement joint

Nouveauté de ce papier

- Fonction de plongement visuelle
- Fonction de plongement textuelle

Article 6

Unsupervised Domain Adaptation by Backpropagation

Yaroslav Ganin GANIN@SKOLTECH.RU
Victor Lempitsky LEMPITSKY@SKOLTECH.RU
Skolkovo Institute of Science and Technology (Skoltech), Moscow Region, Russia

Abstract

Top-performing deep architectures are trained on massive amounts of labeled data. In the absence of labeled data for a certain task, domain adaptation often provides an attractive option given that labeled data of similar nature but from a different domain (e.g. synthetic images) are available. Here, we propose a new approach to domain adaptation in deep architectures that can be trained on large amount of labeled data from the source domain and large amount of unlabeled data from the target domain (no labeled target-domain data is necessary).

As the training progresses, the approach promotes the emergence of "deep" features that are (i) discriminative for the main learning task on the source domain and (ii) invariant with respect to the shift between the domains. We show that this adaptation behaviour can be achieved in almost any feed-forward model by augmenting it with few standard layers and a simple new *gra-*

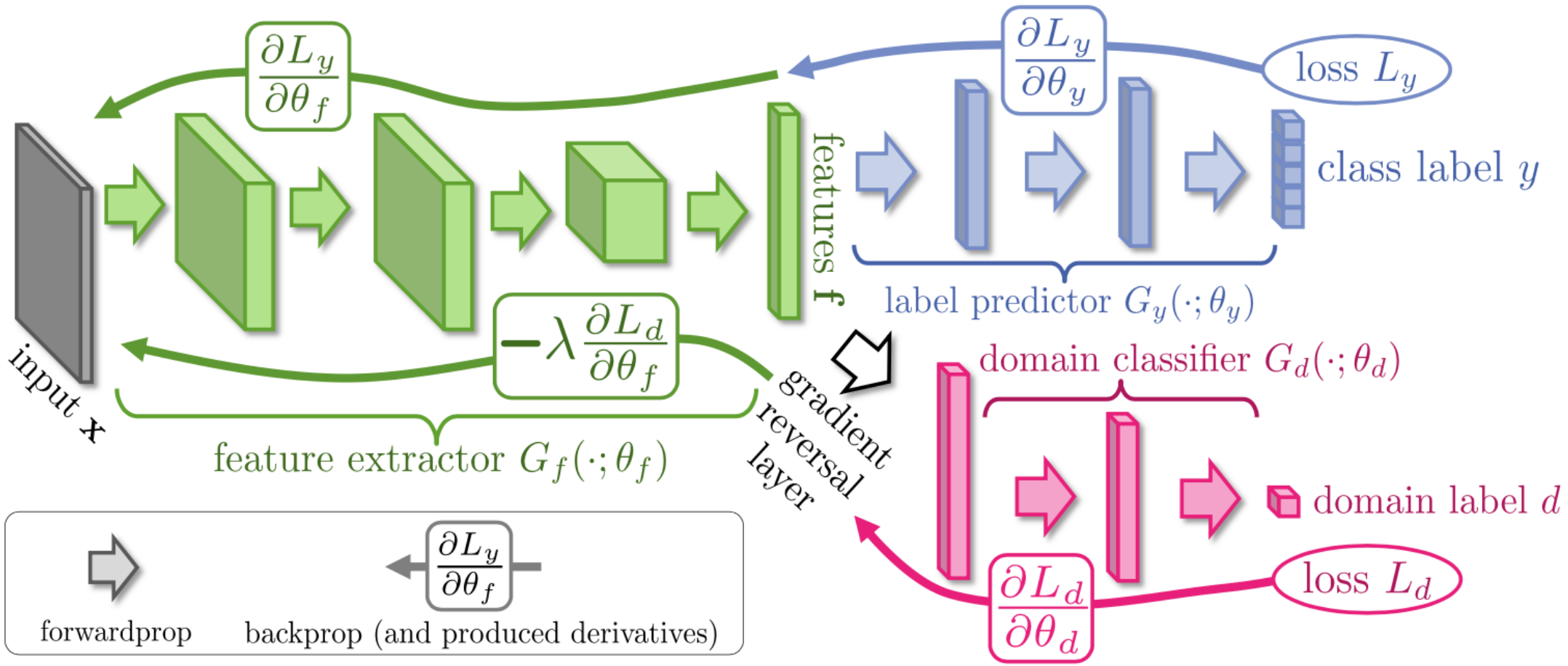
large amount of labeled training data is available. At the same time, for problems lacking labeled data, it may be still possible to obtain training sets that are big enough for training large-scale deep models, but that suffer from the *shift* in data distribution from the actual data encountered at "test time". One particularly important example is synthetic or semi-synthetic training data, which may come in abundance and be fully labeled, but which inevitably have a distribution that is different from real data (Liebelt & Schmid, 2010; Stark et al., 2010; Vázquez et al., 2014; Sun & Saenko, 2014).

Learning a discriminative classifier or other predictor in the presence of a *shift* between training and test distributions is known as *domain adaptation* (DA). A number of approaches to domain adaptation has been suggested in the context of *shallow* learning, e.g. in the situation when data representation/features are given and fixed. The proposed approaches then build the mappings between the *source* (training-time) and the *target* (test-time) domains, so that the classifier learned for the source domain can also be applied to the target domain, when composed with the learned mapping between domains. The appeal of the domain

TLDR:

Pour transférer un modèle de classification d'un domaine vers un autre, il suffit d'entrainer ce modèle à produire des représentations d'images qui ne permettent pas de deviner le domaine auquel elles appartiennent.
Pour cela, on entraine le modèle à tromper un classifieur de domaine

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

Cours 10: Apprentissage auto-supervisé de représentations visuelles

Comprendre les données visuelles à grande échelle
14 janvier 2021

Weakly supervised learning of generic and transferable visual representations

Basé sur une présentation récente. Elle couvre une tâche de vision par ordinateur très populaire récemment. Le premier papier est de 2014, et la majorité ont moins d'un an

Diane Larlus

Principal Research Scientist
Computer Vision group
NAVER LABS Europe

MIAI Grenoble Alpes
Multidisciplinary Institute In Artificial Intelligence

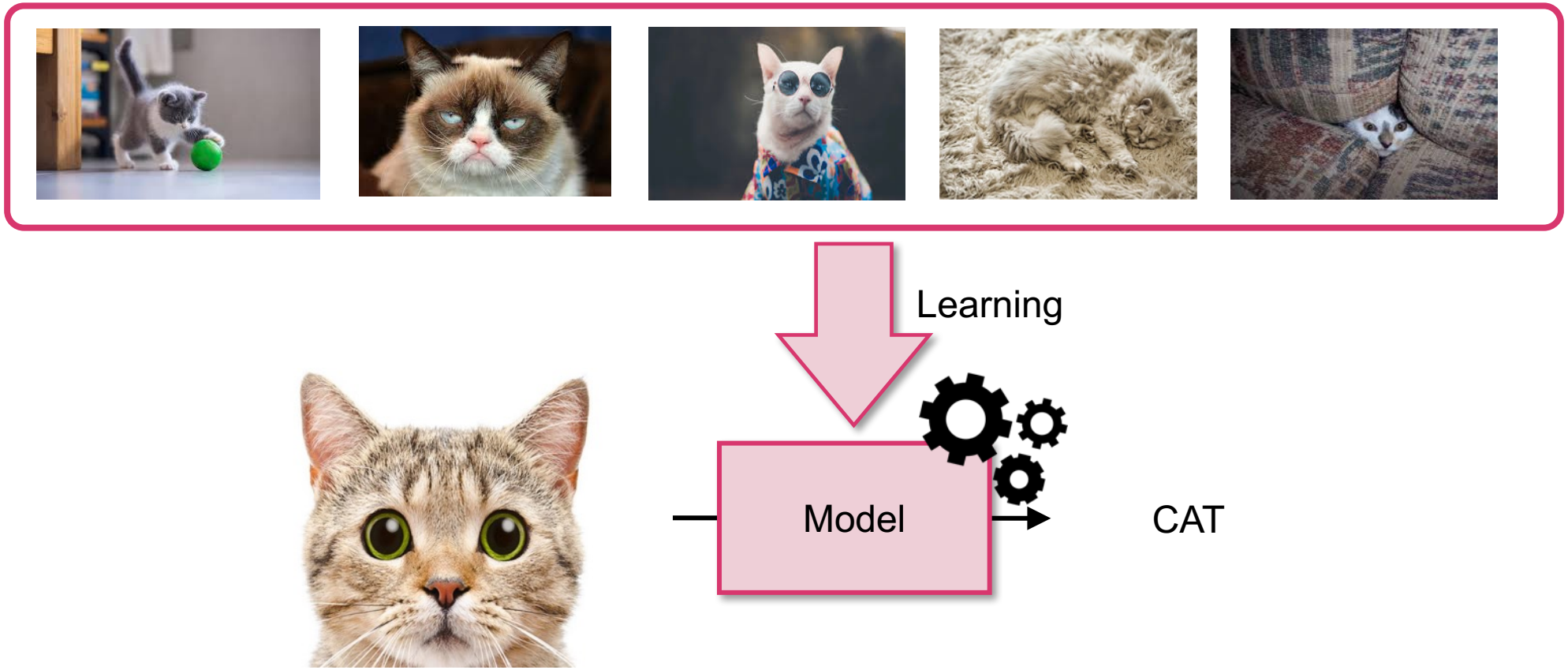
MIAI Chair: Lifelong Representation Learning

October 9th 2020 – MIAI Day – Festival Transfo

NAVER LABS Europe

Computer vision

- Image classification



NAVER LABS Europe

Computer vision over the last two decades

- [x] Large image collections to train from

Computer vision over the last two decades

- [x] Large image collections to train from

4 categories

2005 Pascal VOC

Bicycle

Car

Motorbike

Person

NAVER LABS Europe

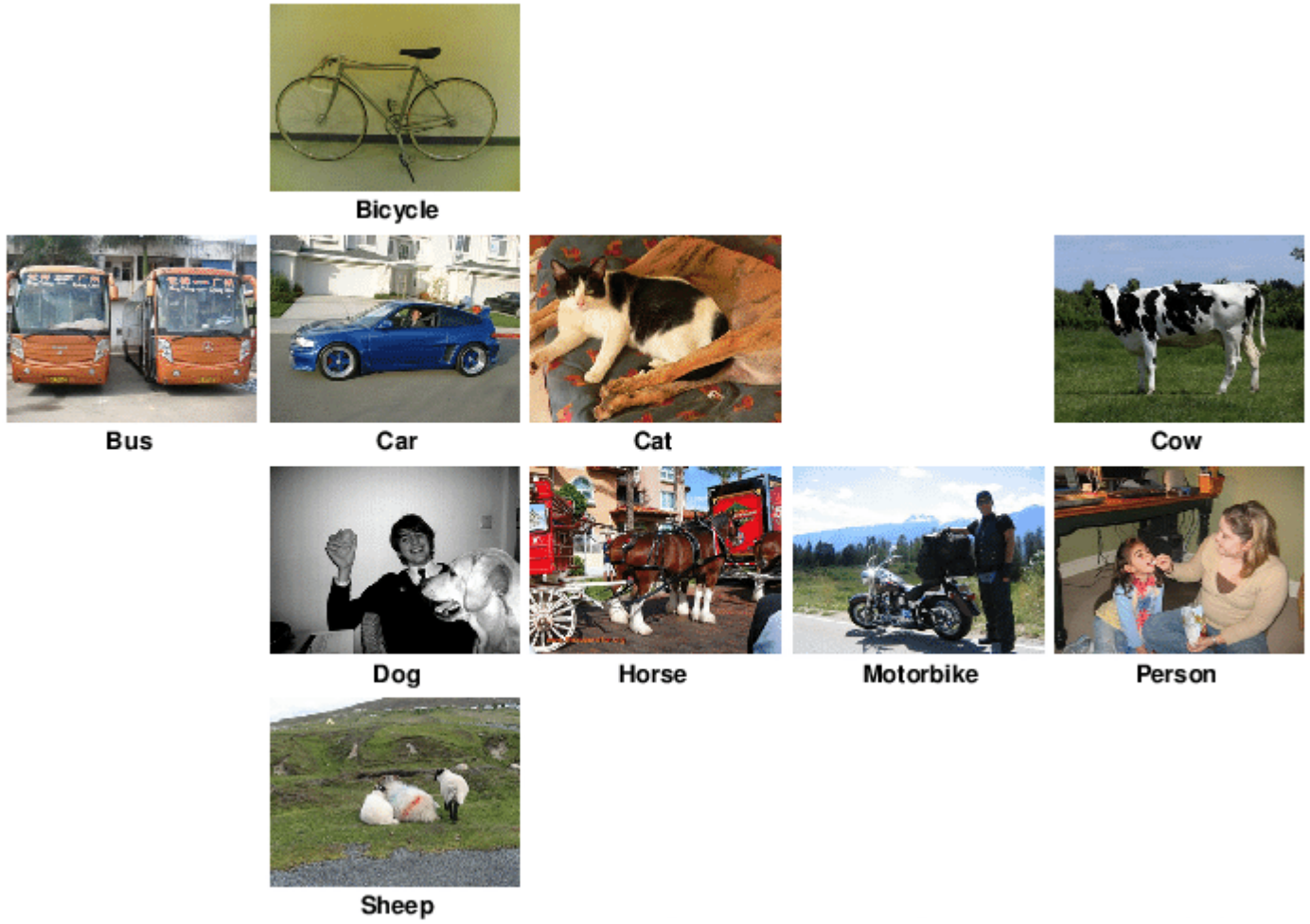
Computer vision over the last two decades

- [x] Large image collections to train from

10 categories

2005 Pascal VOC

2006 Pascal VOC



NAVER LABS Europe

Computer vision over the last two decades

- Large image collections to train from

20 categories

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

NAVER LABS Europe

Computer vision over the last two decades

- [x] Large image collections to train from

1000 categories

2005 Pascal VOC

2006 Pascal VOC

2007 Pascal VOC

2010 ImageNet

NAVER LABS Europe

Computer vision over the last two decades

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

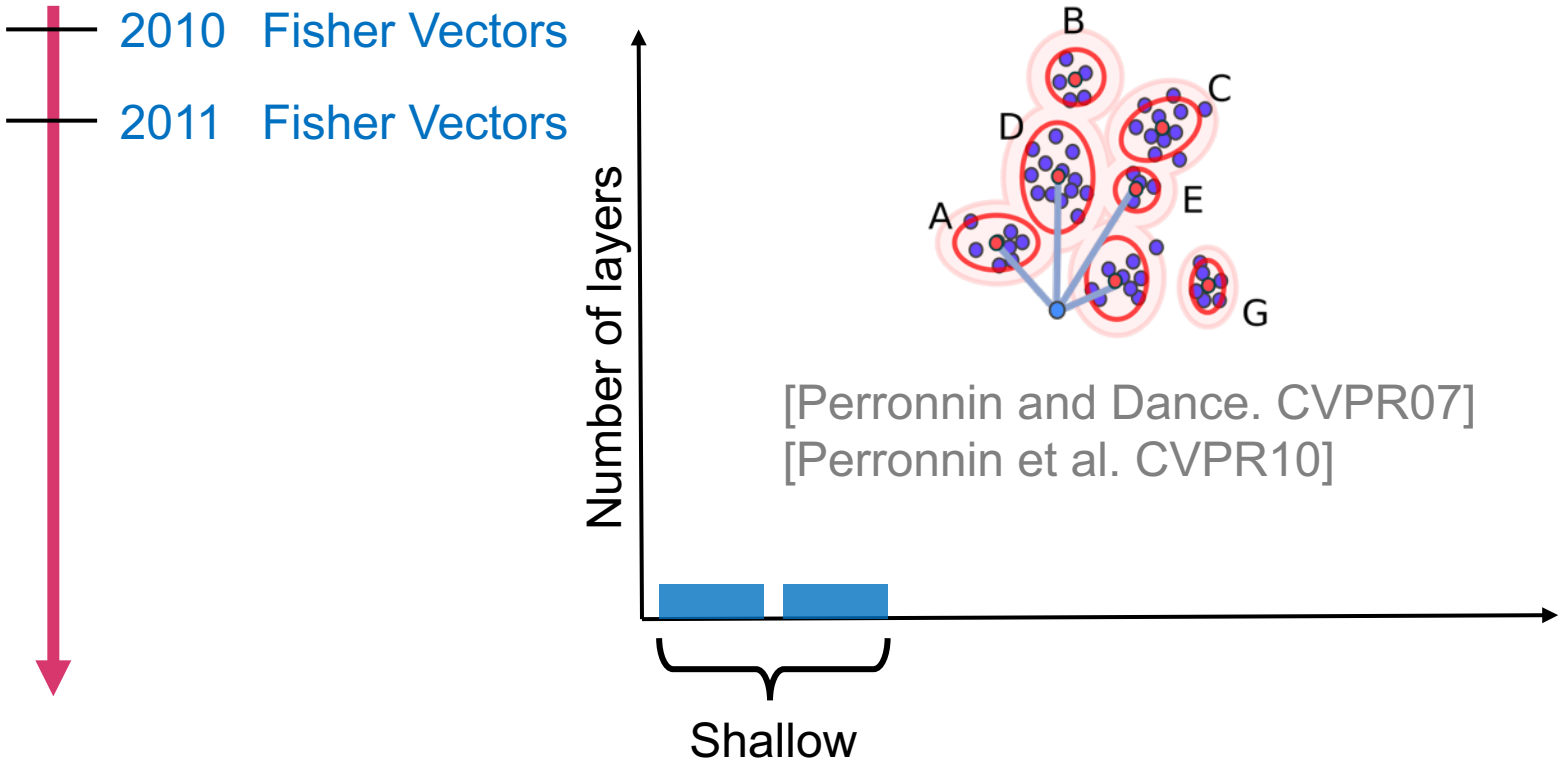
IMAGENET

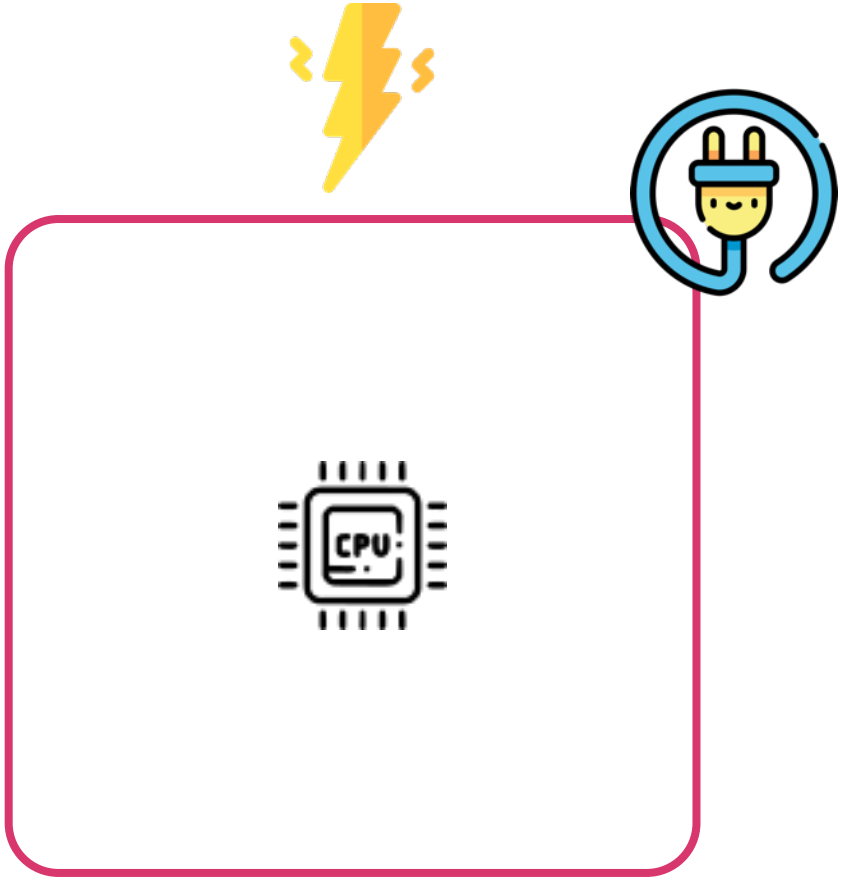
NAVER LABS Europe

Computer vision over the last two decades

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

IMAGENET





NAVER LABS Europe

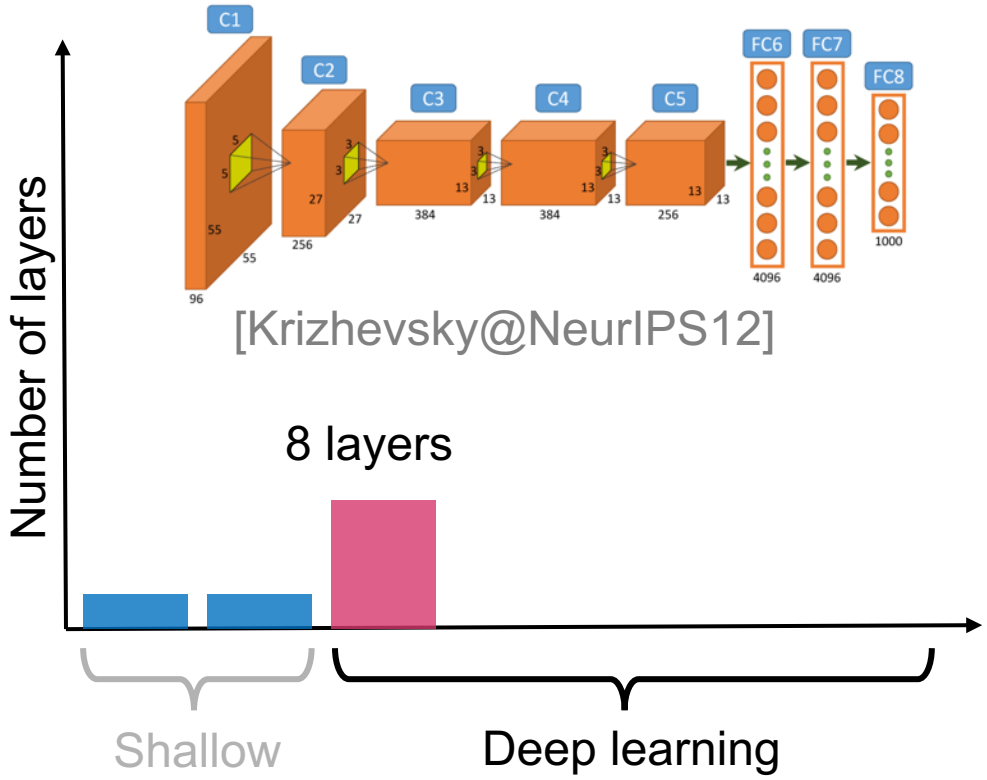
Computer vision over the last two decades

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

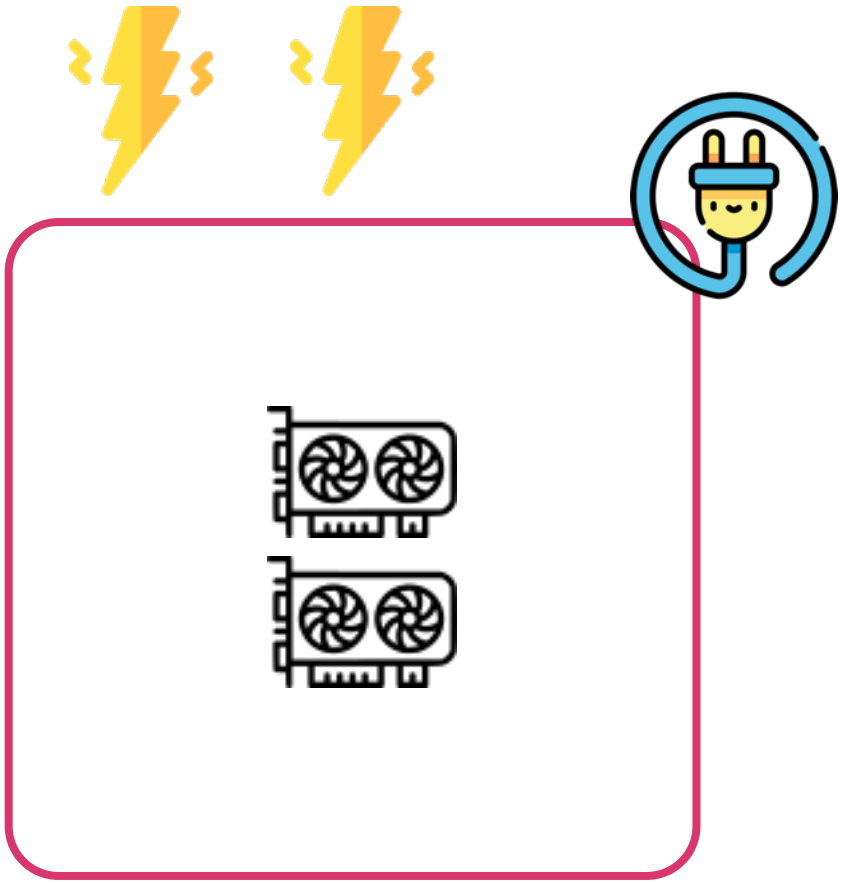
2010 Fisher Vectors

2011 Fisher Vectors

2012 AlexNet



IMAGENET

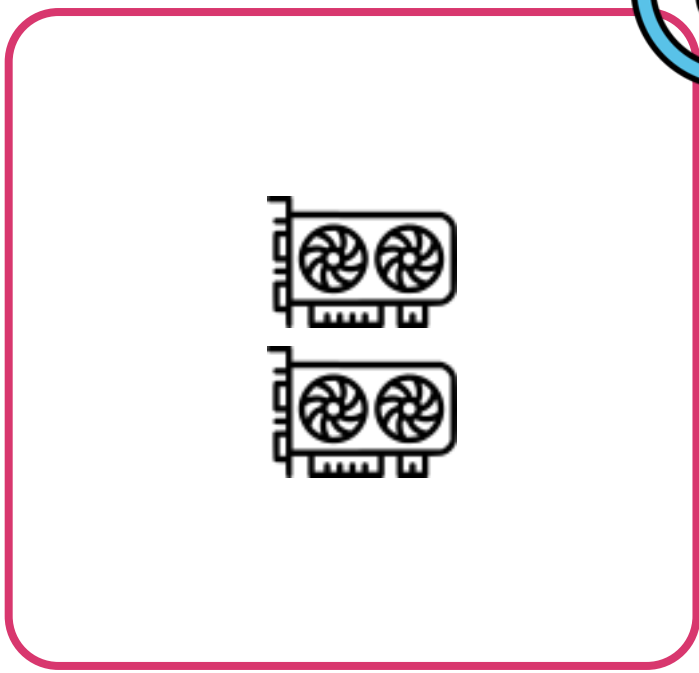


NAVER LABS Europe

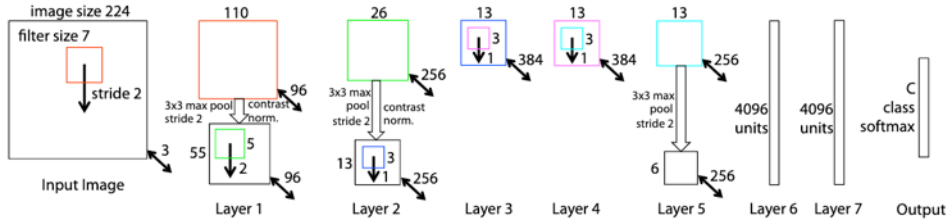
Computer vision over the last two decades

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

IMAGENET



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

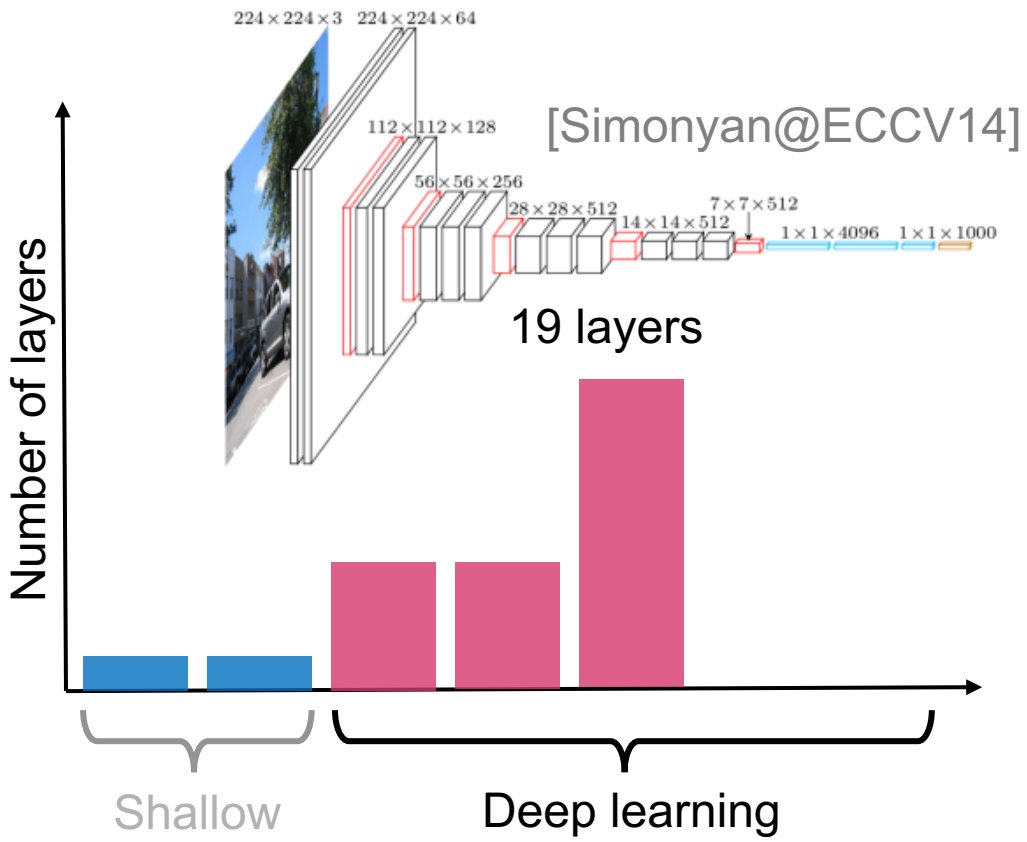


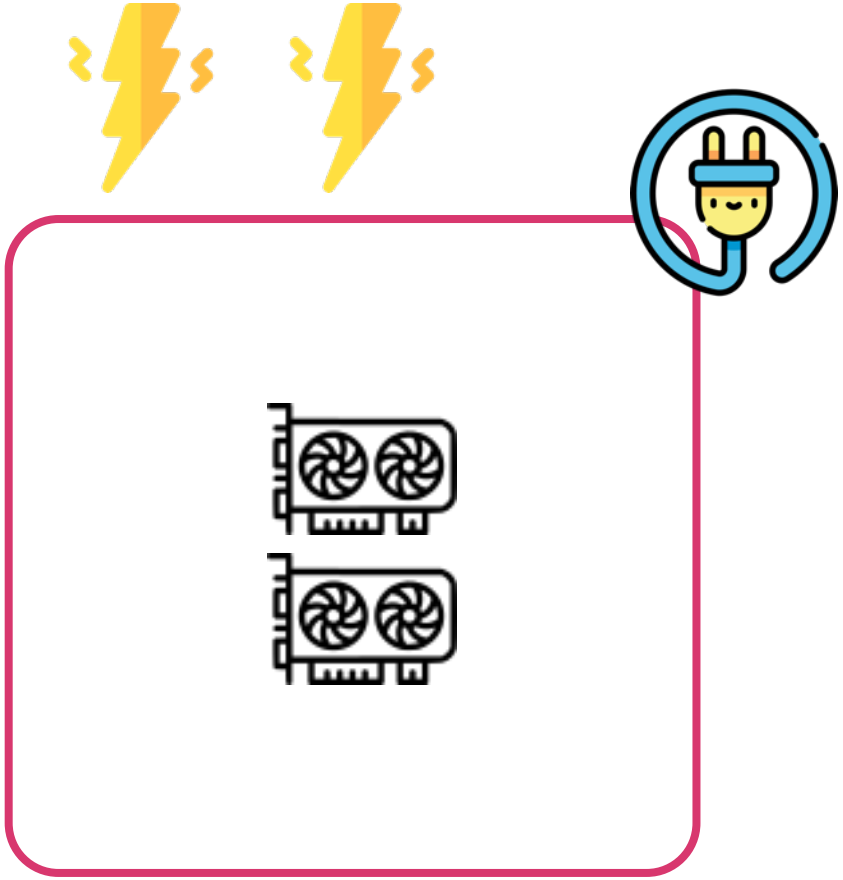
[Zeiler@ECCV14]

NAVER LABS Europe

Computer vision over the last two decades

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



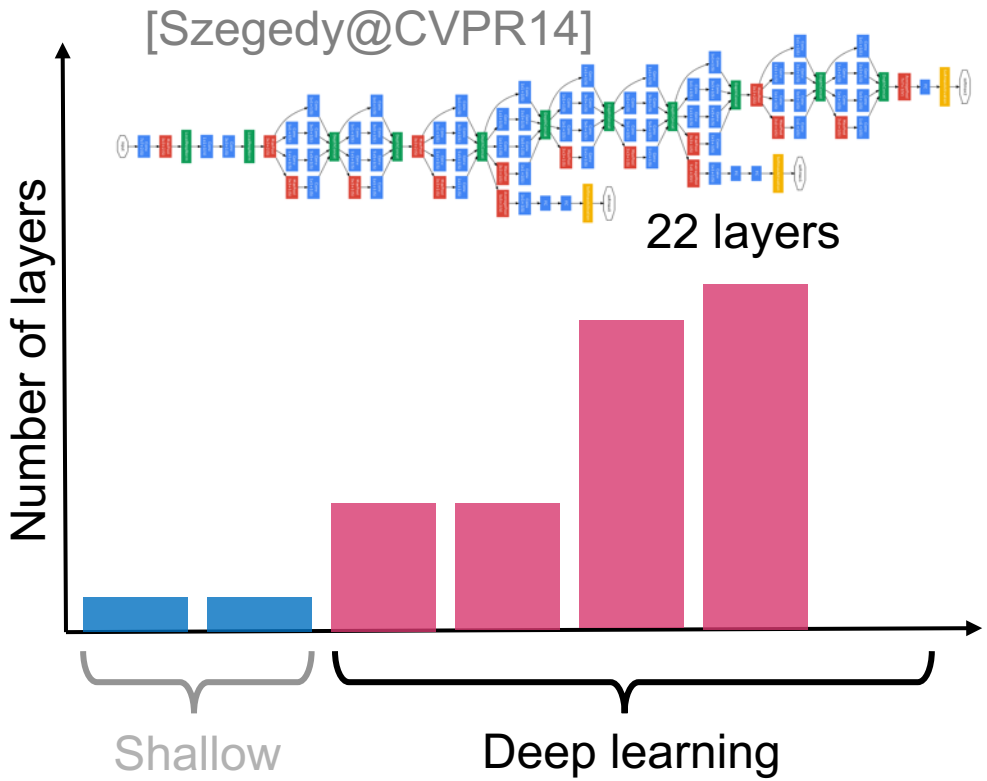


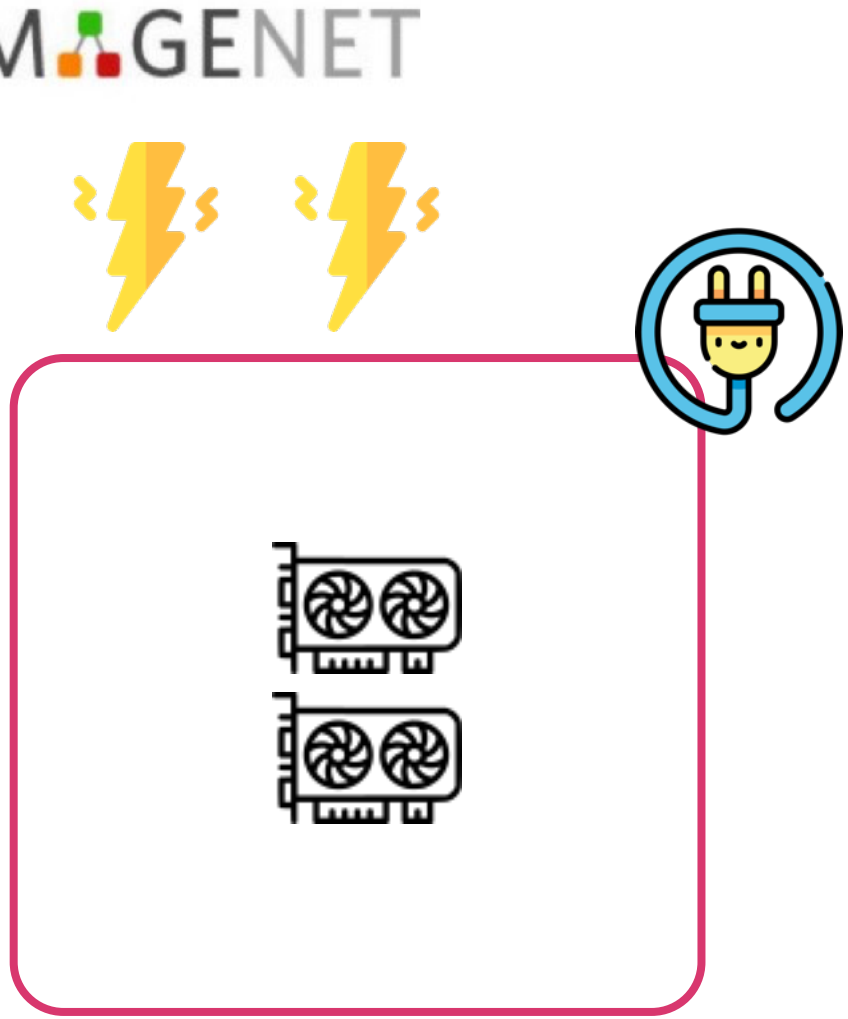
NAVER LABS Europe

Computer vision over the last two decades

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

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

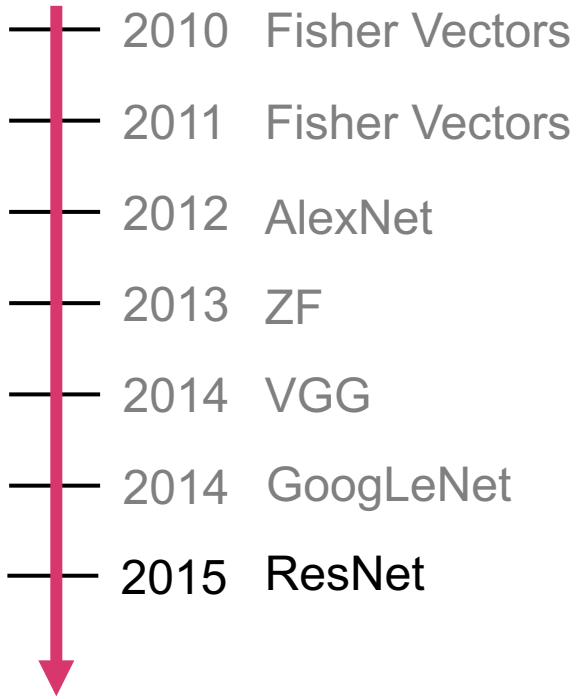


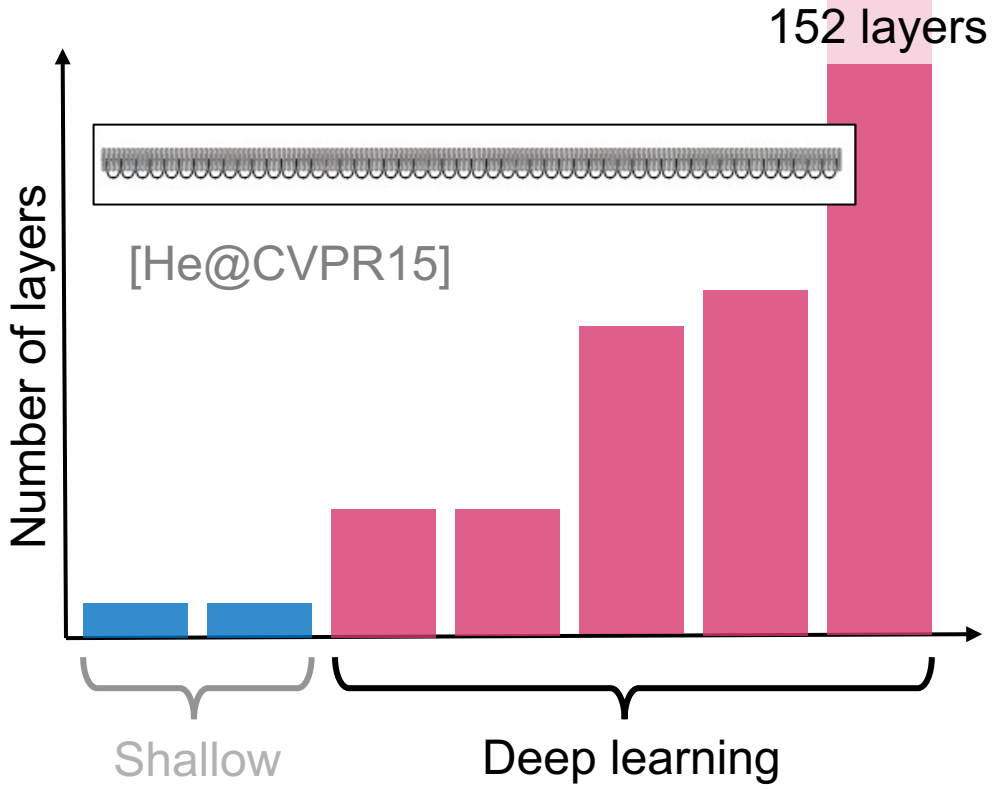


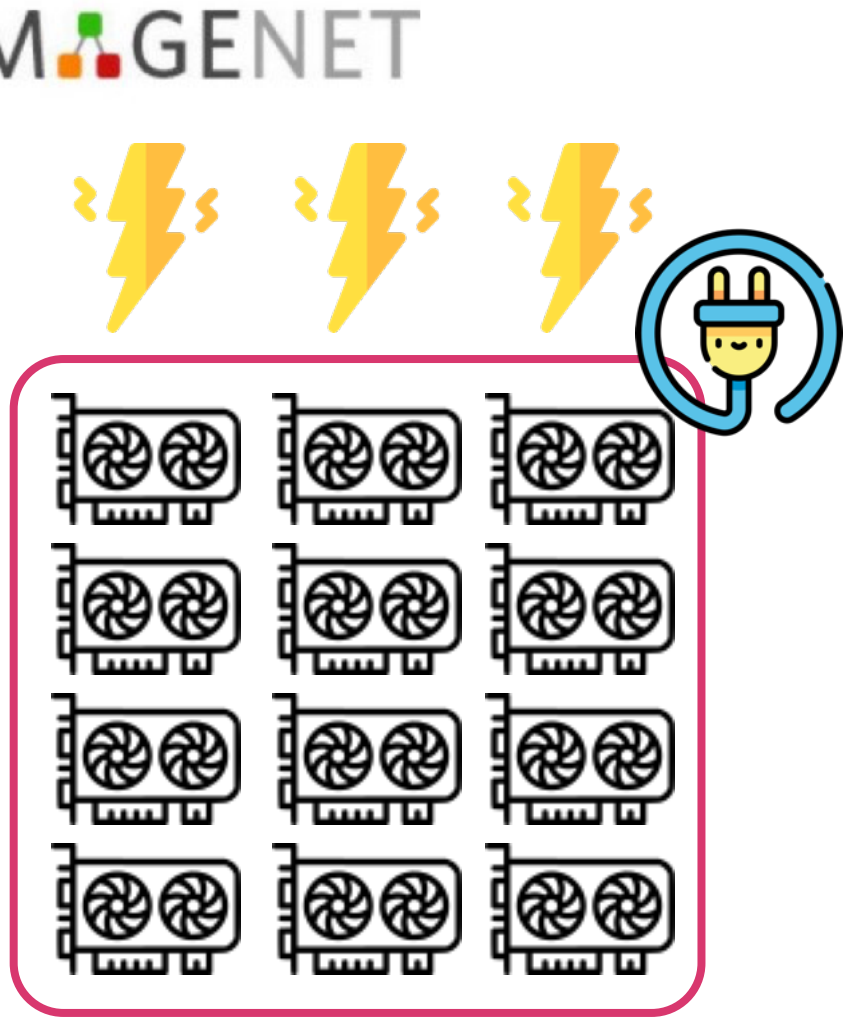
NAVER LABS Europe

Computer vision over the last two decades

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





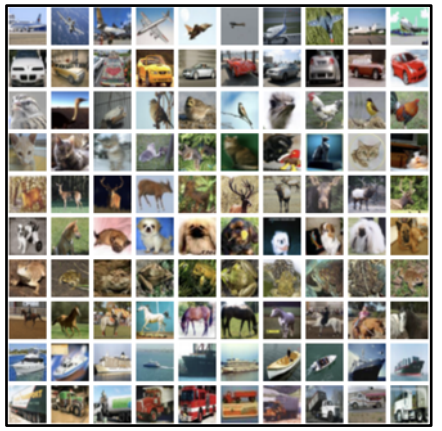


NAVER LABS Europe

Pretraining general-purpose visual representations

Fully-Supervised Classification
Images + labels

IMAGENET



Fully-Supervised Classification
Images + labels

Fully-Supervised Classification
Images + labels

NAVER LABS Europe

Pretraining general-purpose visual representations

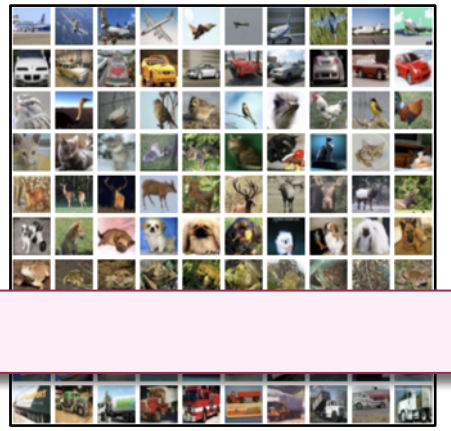
Tâche prétexte

Proxy task

Fully-Supervised Classification

Images + labels

IMAGENET



Reducing annotation cost

Model

Visual representations

Tâches cibles

Target tasks

Image Classification

Object Detection

Instance Segmentation

Image Retrieval

NAVER LABS Europe

Pretraining general-purpose visual representations

No supervision

Pretraining visual representations

Reducing the annotation cost

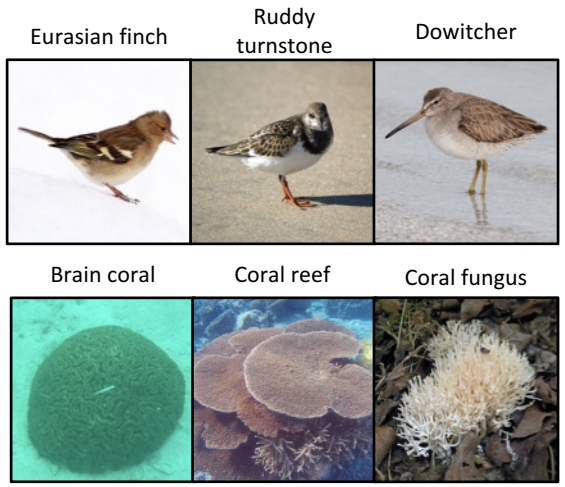
Apprentissage totalement supervisé

Apprentissage auto supervisé

Fully-Supervised

fine-grained annotations

expert knowledge

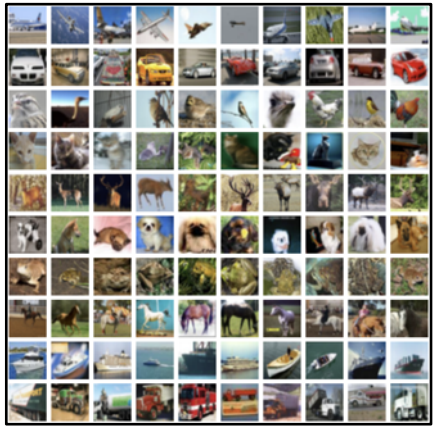


IMAGENET

Self-supervised

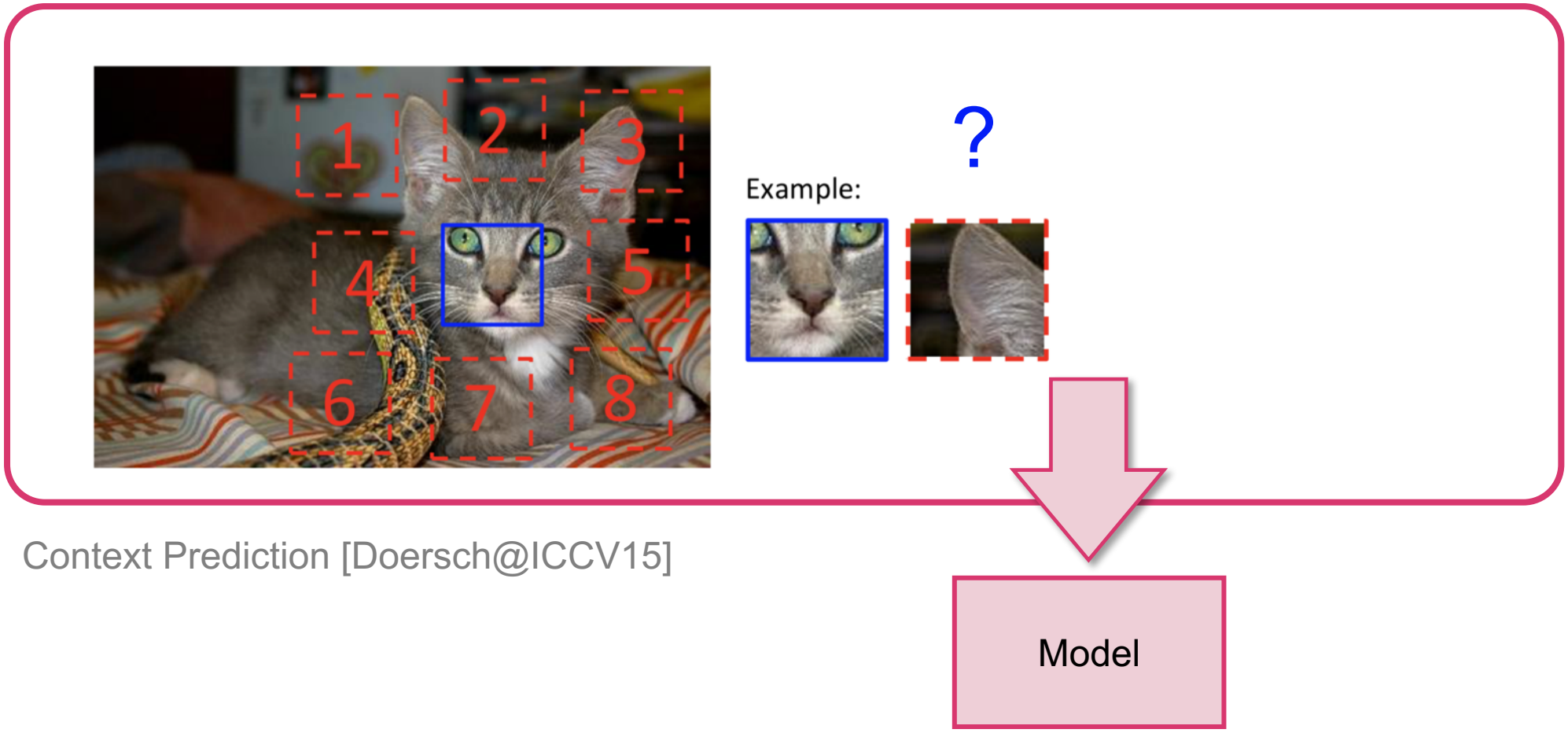
annotation-free images

no annotation required



NAVER LABS Europe

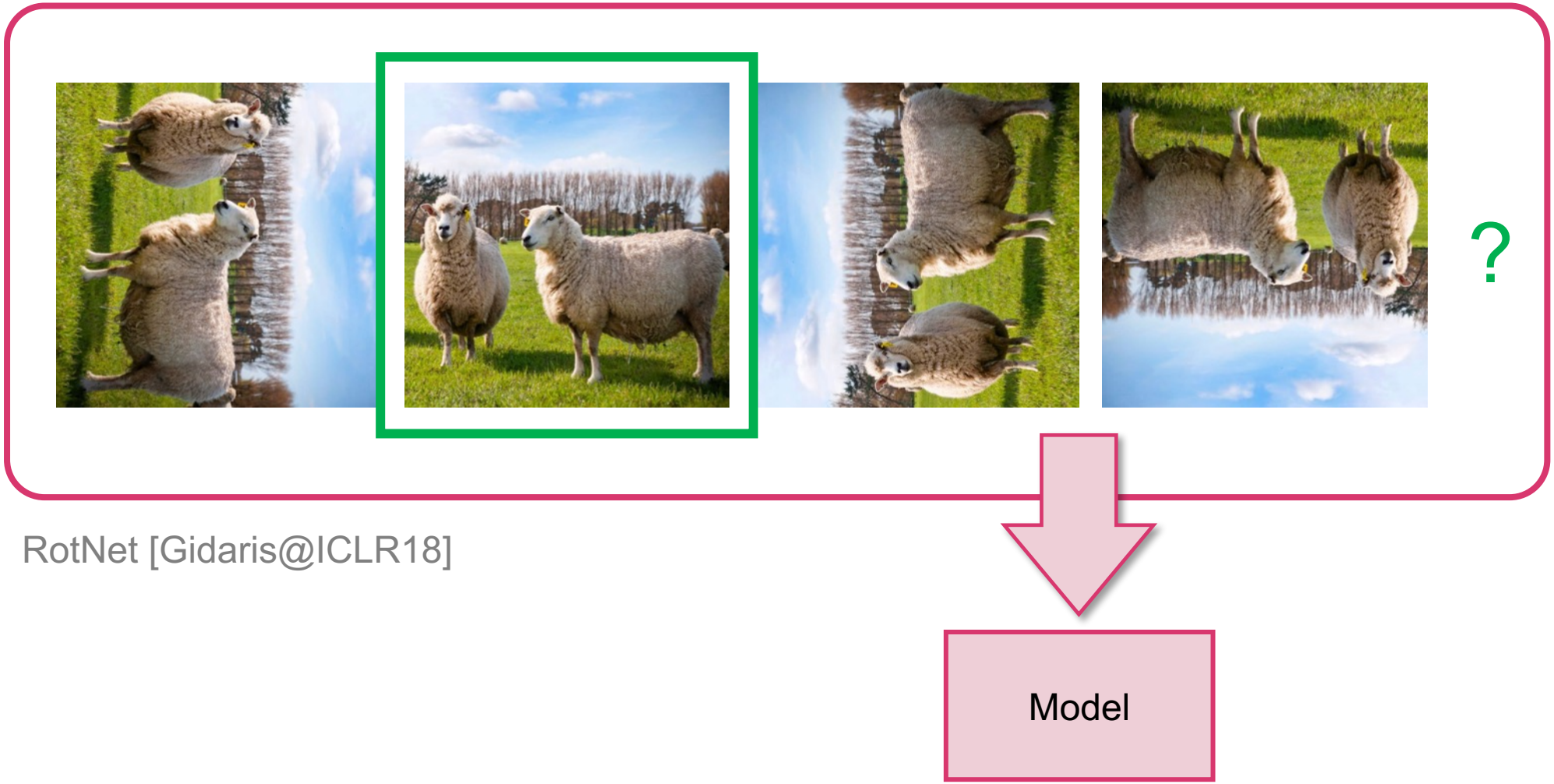
Self-supervised learning



Context Prediction [Doersch@ICCV15]

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

Self-supervised learning



RotNet [Gidaris@ICLR18]

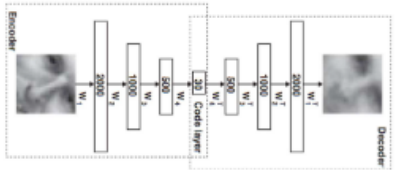
Unsupervised representation learning by predicting image rotations, Spyros Gidaris, Praveer Singh, Nikos Komodakis, ICLR 2018

NAVER LABS Europe

Self-supervised learning

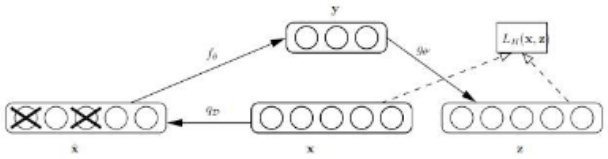
Plus d'info: cours d'Andrew Zisserman

Autoencoders



Hinton & Salakhutdinov. Science 2006.

Denoising Autoencoders

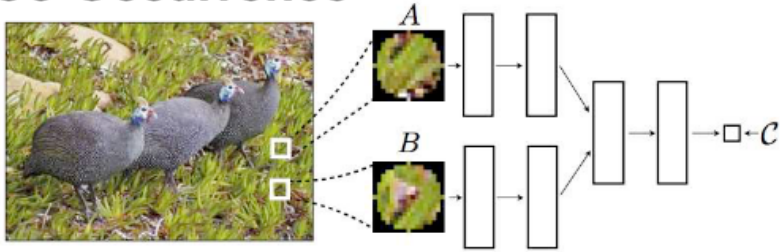


Vincent *et al.* ICML 2008.

Exemplar networks

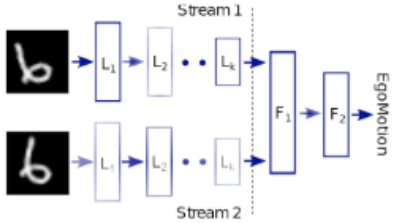
Dosovitskiy *et al.*, NIPS 2014

Co-Occurrence



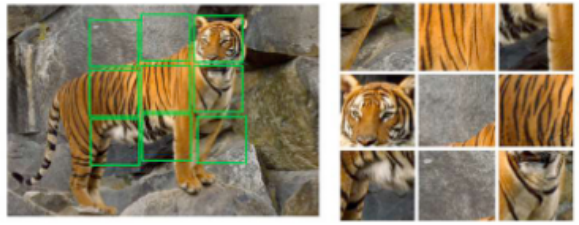
Isola *et al.* ICLR Workshop 2016.

Egomotion



Agrawal *et al.* ICCV 2015 Jayaraman *et al.* ICCV 2015

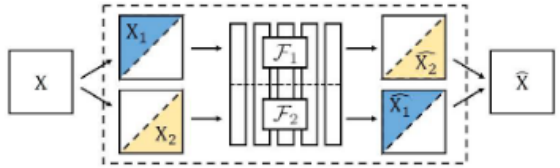
Context



Noroozi et al 2016

Pathak *et al.* CVPR 2016

Split-brain auto-encoders



Zhang *et al.* CVPR 2017

[Slide from Andrew Zisserman]

NAVER LABS Europe

Self-supervised learning – Contrastive learning

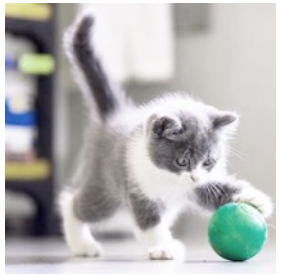
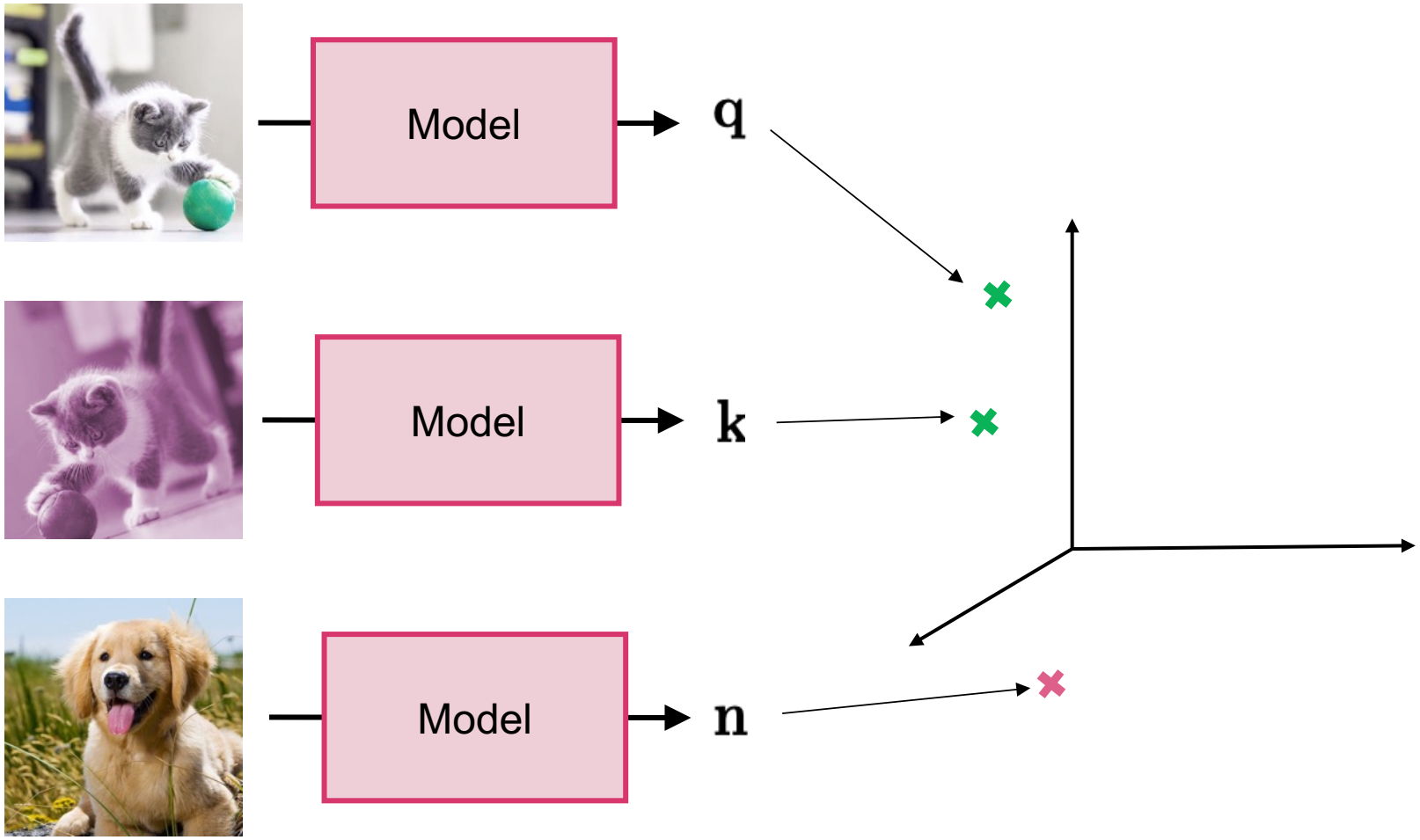


Image Transformations

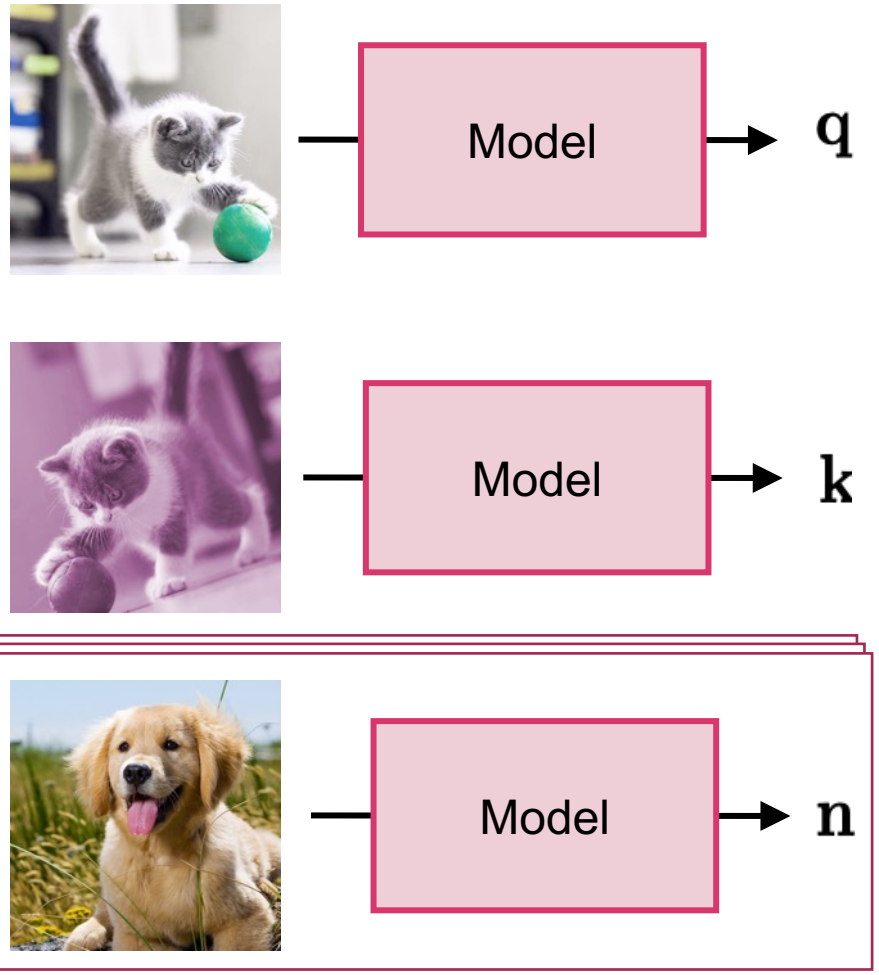
Type d'apprentissage auto-supervisé le plus populaire en ce moment: **L'apprentissage contrastif**

Self-supervised learning – Contrastive learning



NAVER LABS Europe

Self-supervised learning – Contrastive learning



Tâche prétexte:
La discrimination d'instance

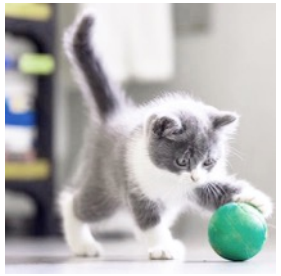
- Chaque image est traitée comme sa propre classe
- On regarde les similarités entre images dans le plongement appris = l'espace de représentation d'images

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

Proxy task: Instance discrimination

NAVER LABS Europe

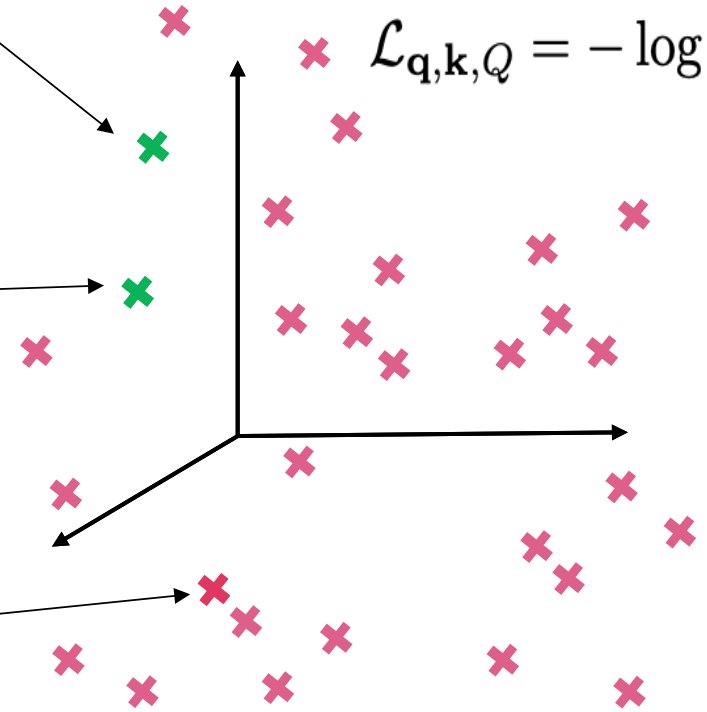
Self-supervised learning – Contrastive learning



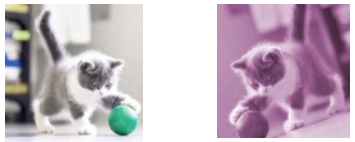
Model → $\mathbf{q}$

Model → $\mathbf{k}$

Model → $\mathbf{n}$



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

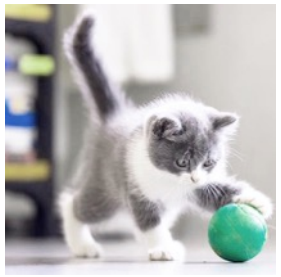


MoCo [He@CVPR20]

Proxy task: Instance discrimination

NAVER LABS Europe

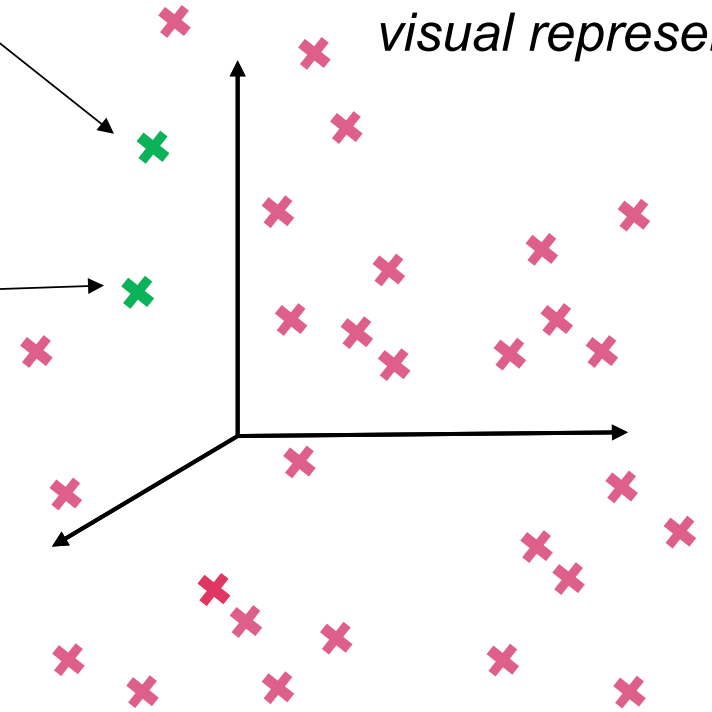
Self-supervised learning – Contrastive learning



Model → q

Model → k

Model → n



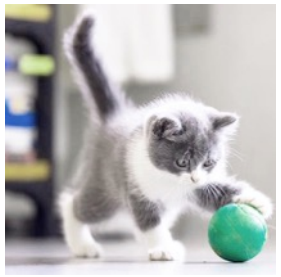
Key observation:

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

MoCo [He@CVPR20]

Proxy task: Instance discrimination

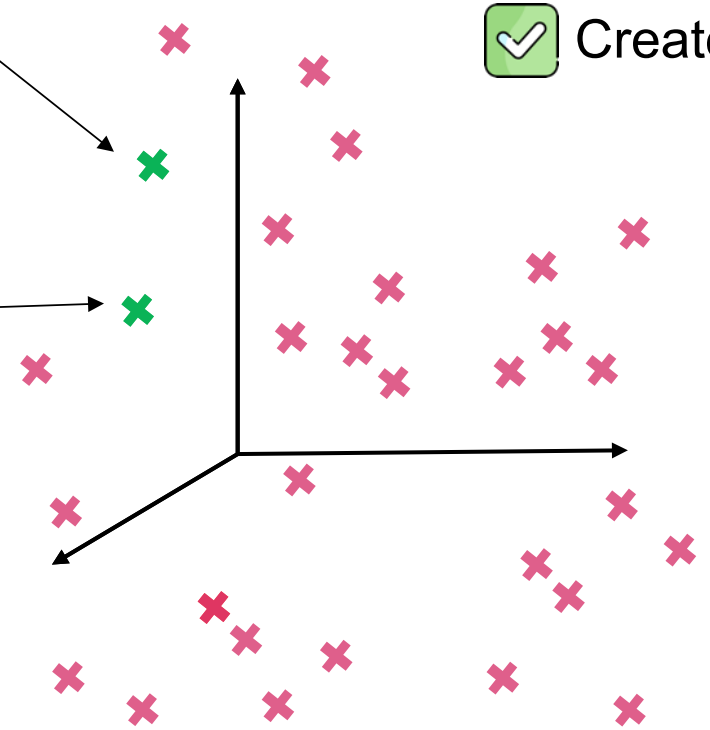
Self-supervised learning – Contrastive learning



Model → $\mathbf{q}$

Model → $\mathbf{k}$

Model → $\mathbf{n}$

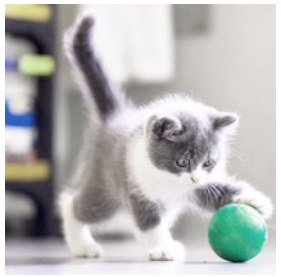


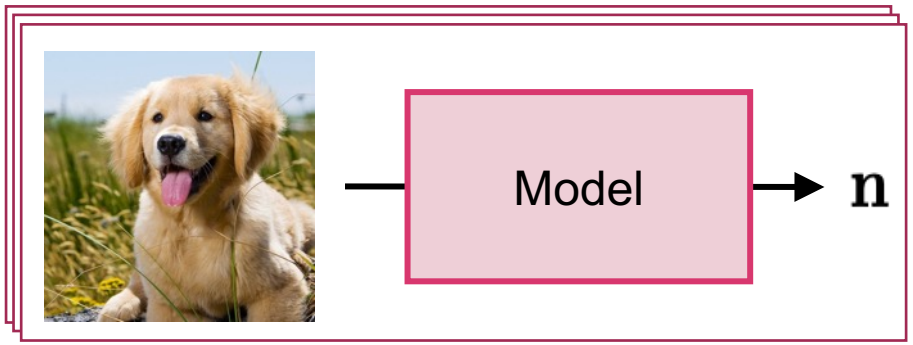
How to make the task harder?

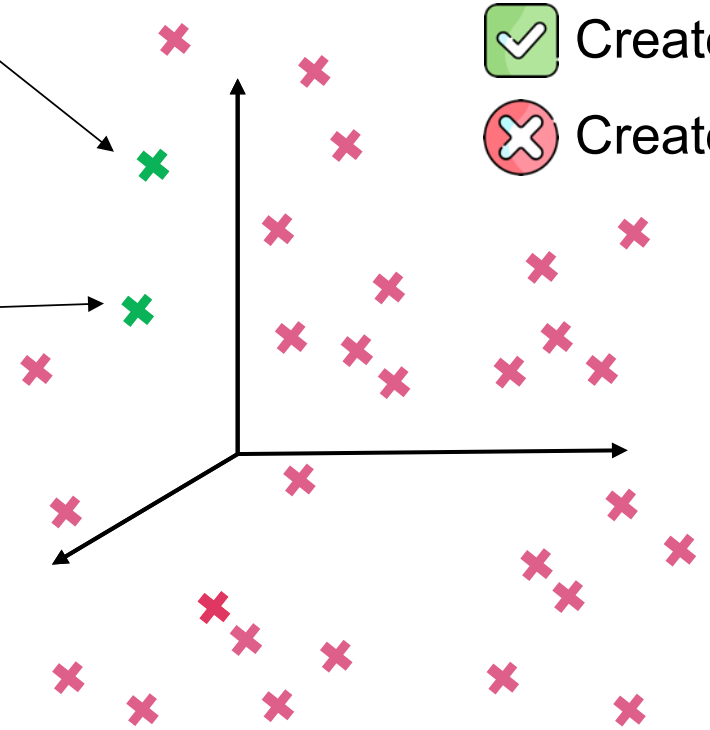
- [x] Create “better” positive pairs

Proxy task: Instance discrimination

Self-supervised learning – Contrastive learning







How to make the task harder?

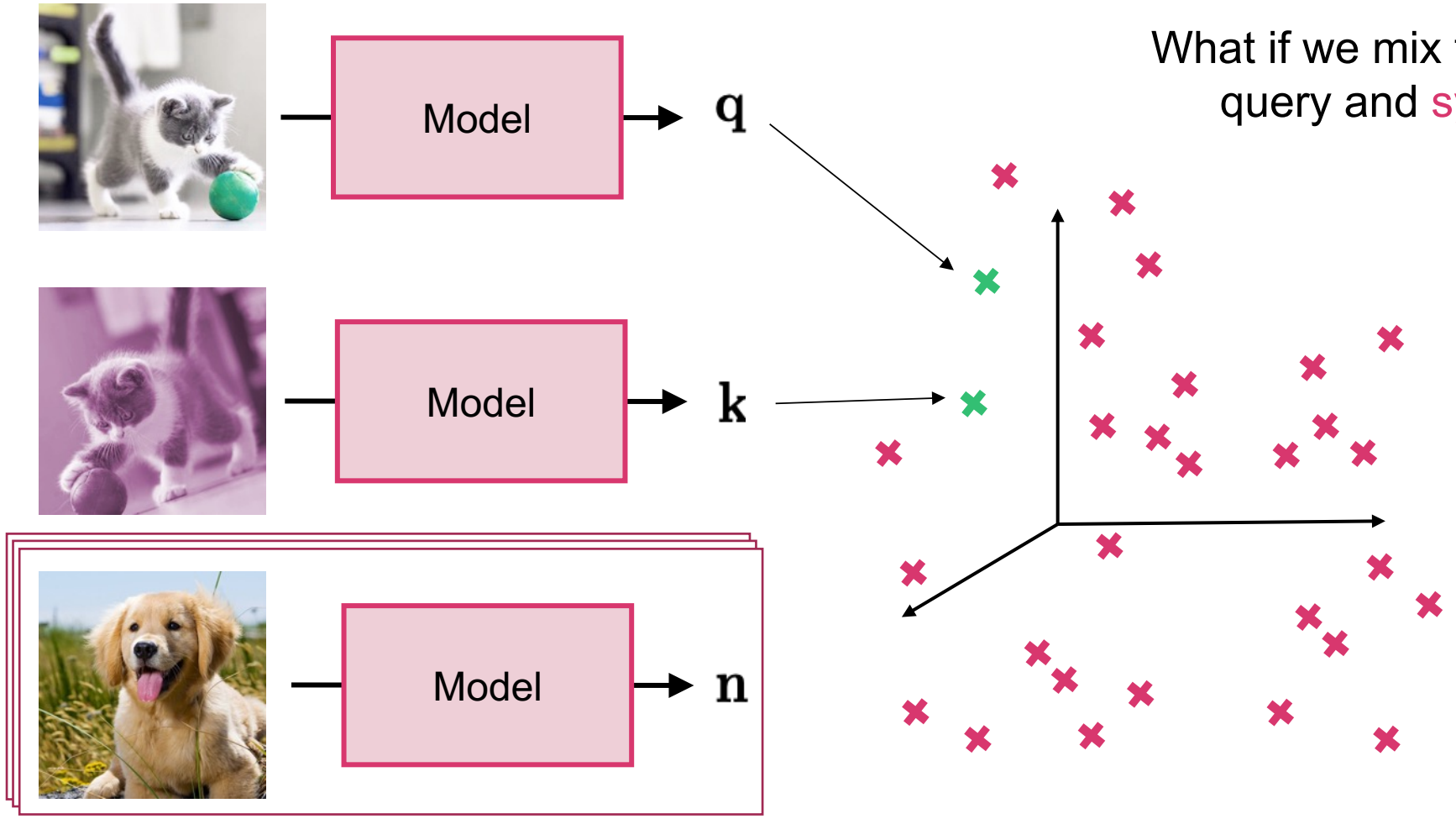
- Create “better” positive pairs
- Create “better” negative pairs

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

Proxy task: Instance discrimination

NAVER LABS Europe

Self-supervised learning – Contrastive learning

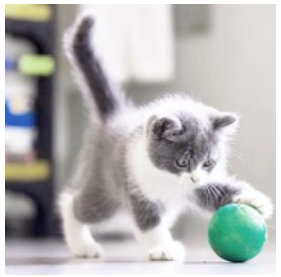


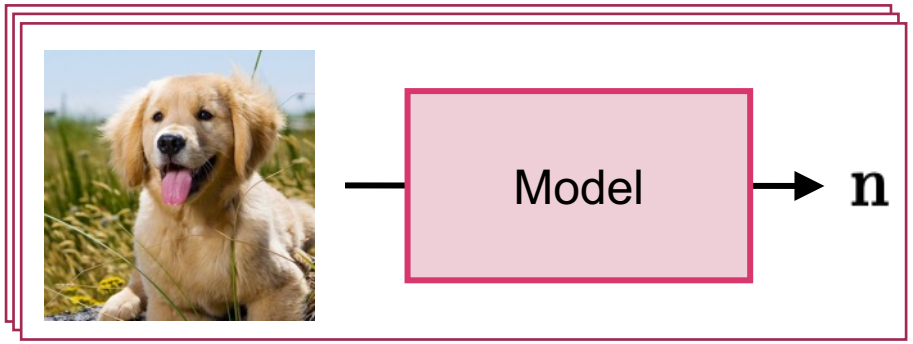
What if we mix the **hardest** negatives for each query and synthesize new hard negatives?

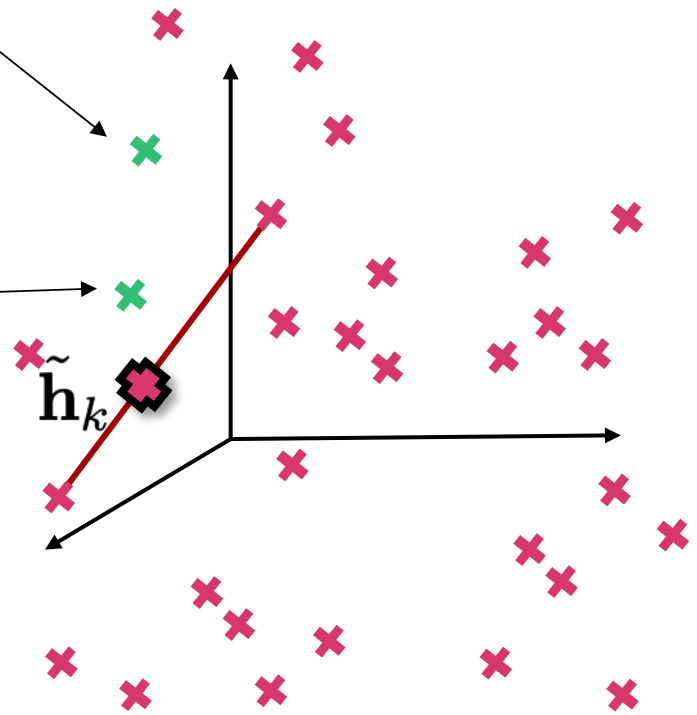
MoCo [He@CVPR20]

NAVER LABS Europe

Self-supervised learning – Contrastive learning







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

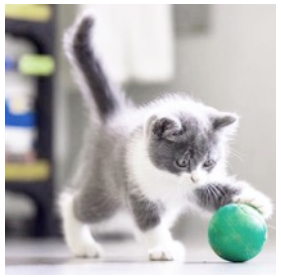
by mixing **two negatives**

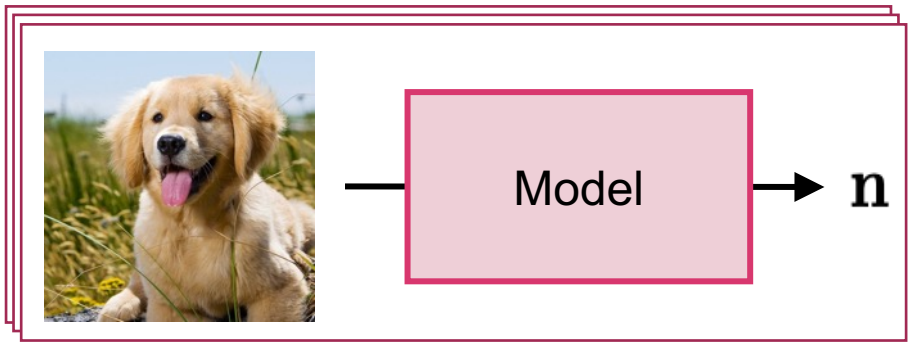
$$\tilde{\mathbf{h}}_k = \alpha_k \mathbf{n}_i + (1 - \alpha_k)\mathbf{n}_j$$

MoCo [He@CVPR20]

NAVER LABS Europe

Self-supervised learning – Contrastive learning





MoCo [He@CVPR20]

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**

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

MoCHi [Kalantidis@NeurIPS20]

NAVER LABS Europe

Pretraining general-purpose visual representations

Weak annotations

Pretraining visual representations

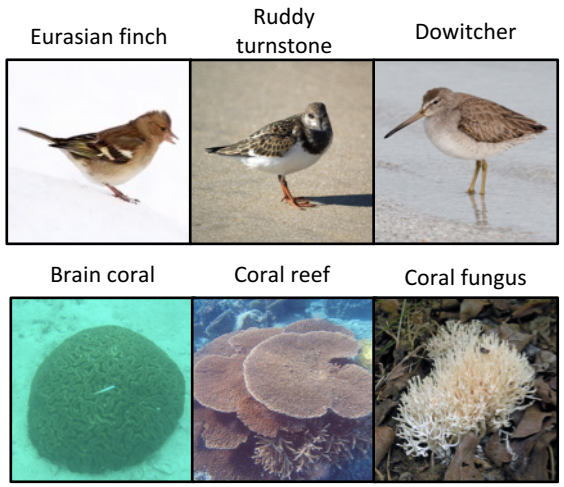
Reducing the annotation cost

Apprentissage totalement supervisé

Fully-Supervised

fine-grained annotations

expert knowledge



Apprentissage faiblement supervisé

Caption-supervised

side information

smaller sets

a statue of a man stands in front of an old red bus.
a big and red bus with many displays for people to watch.
a red double decker bus parked next to a statue.
the double decker bus is beside a statue near restaurant tables.
a view of a bus sitting in front a small wooden statue.

a busy street with cars and trucks down it
an intersection with a view that looks towards a small downtown area.
cars parked on the side of the street and traveling down the road
an intersection with a stop light on a city street.
a street filled with lots of traffic under a traffic light.

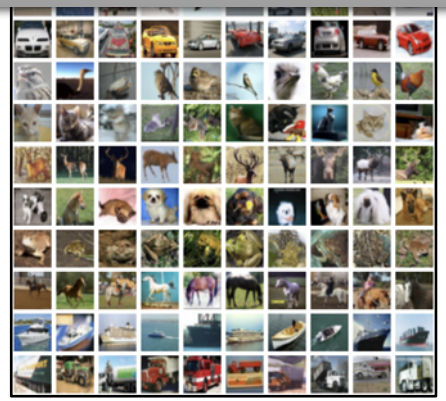
Apprentissage auto supervisé

Self-supervised

annotation-free images

computationally demanding

Nécessite des millions voir des centaines de millions d'images



NAVER LABS Europe

Masked language modeling task

Input:	Text

BERT model
[Delvin *et al.* 2018]

"Little girl holding red umbrella"

Mask a token

"Little girl holding red [MASK]"

Language Model

[MASK] = Umbrella

Image-Conditioned Masked Language Modeling Task

Input: Image	Caption

"Little girl holding red umbrella"

↓

Mask a token

↓

"Little girl holding red [MASK]"

Multi-modal network =
ConvNet + Language Model + Auxiliary modules

[MASK] = Umbrella

ICMLM [Sariyildiz@ECCV2020]

- ✓ Encode both input modalities
- ✓ Align the representations of the semantic concepts
- ✓ Focus on the image regions corresponding to [MASK]

NAVER LABS Europe

Résumé – Take home message

- Les représentations visuelles apprises sur la base annotée ImageNet (avec ses étiquettes) se transfèrent très bien à toute sorte d'autres tâches de vision par ordinateur comme la classification d'autres catégories, la détection ou la segmentation d'objets, la recherche d'image, etc. C'est une pratique standard. Elles est à peine mentionnée dans les articles, mais vous la trouverez toujours.
- Pour relaxer la contrainte des étiquettes, des méthodes dites d'apprentissage auto-supervisé ont été proposées. En général, elles utilisent des techniques d'apprentissage supervisé mais génèrent automatiquement des étiquettes / annotations à partir des données elles-mêmes, sans nécessiter une vérité terrain fournie par des annotateurs. C'est pour cela qu'elles appartiennent à la famille des méthodes d'apprentissage non-supervisées.
- Parmi les méthodes auto-supervisées les plus performantes, les méthodes basées sur l'apprentissage contrastif utilisent des transformations d'images, souvent extrêmes, et considèrent que chaque image correspond à sa propre classe.
- Les méthodes d'apprentissage auto-supervisées les plus performantes demandent de grandes quantités d'images. Souvent toute la base ImageNet. Parfois des bases encore plus grandes.
- Une façon de réduire le nombre d'images nécessaires pour l'apprentissage tout en ne demandant pas d'étiquette précise par image, est de réintroduire des annotations faibles, comme des descriptions textuelles. Une tâche prétexte est par exemple la prédiction de mots masqués, comme utilisé par les modèles de l'état de l'art en traitement automatique des langues (NLP), mais conditionnée par l'image.