Comprendre les données visuelles à grande échelle

ENSIMAG
2019-2020

Karteek Alahari & Diane Larlus

17 octobre 2019

Organisation du cours

- **17/10/19** cours 1 - Diane
- **24/10/19** cours 2 - Karteek
- **07/11/19** cours 3 - Karteek
- **14/11/19** cours 4 - Diane
- **28/11/19** cours 5 - Karteek
- **05/12/19** cours 6 - Karteek
- **12/12/19** cours 7 - Diane
- **19/12/19** cours 8 - Diane

Vacances d'hiver

- **09/01/20** cours 9 - Diane + présentation articles 1 & 2 + quizz
- **16/01/20** cours 10 - Diane + présentation articles 3 & 4 + quizz
- **23/01/20** cours 11 - Karteek + présentation articles 5 & 6 + quizz
- **30/01/20** cours 12 - Karteek + présentation articles 7 & 8 + quizz

Attention: la salle change régulièrement

NAVER LABS EUROPE

Cours 4: Recherche d’images - suite et fin

Comprendre les données visuelles à grande échelle
14 novembre 2019

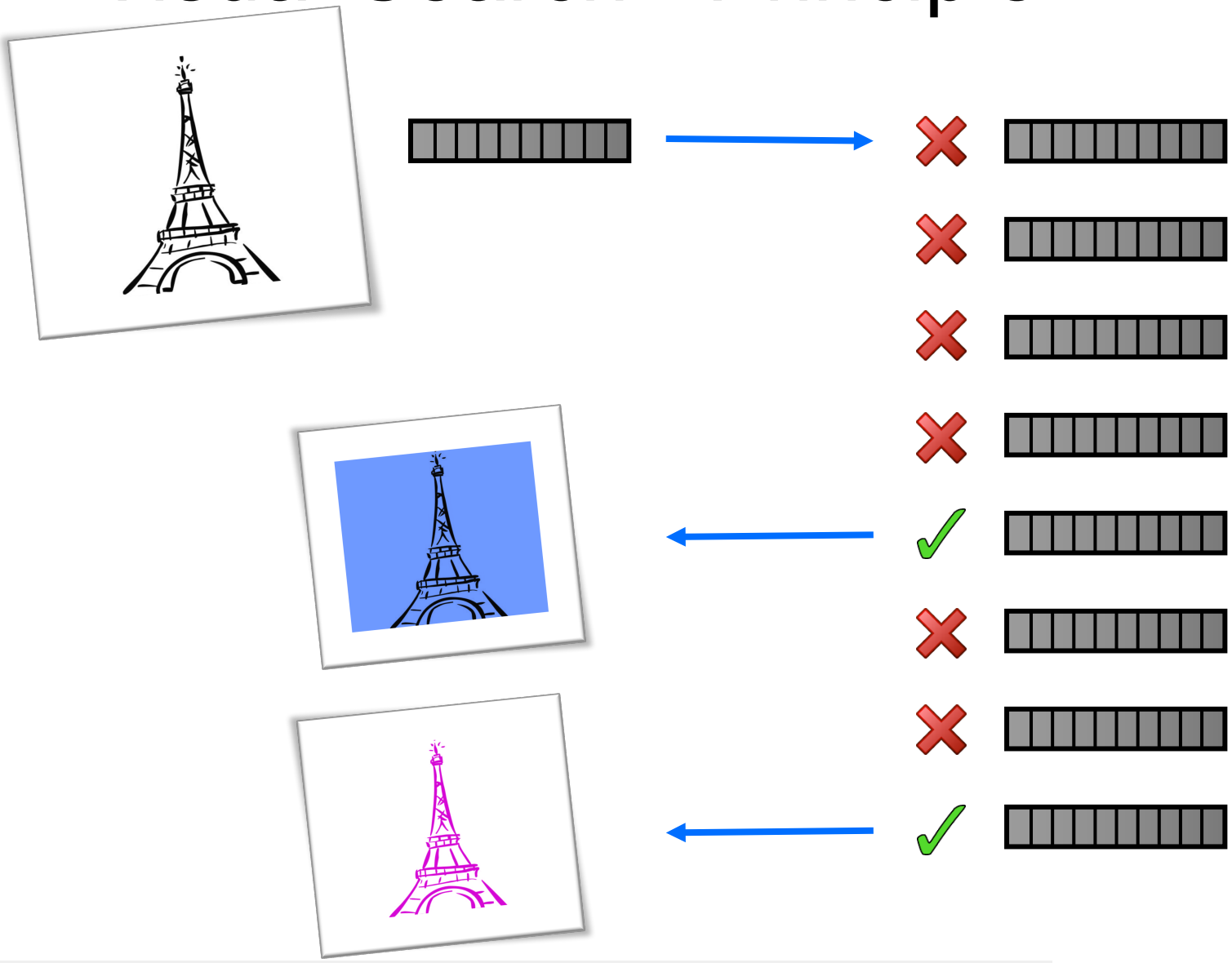
Recherche visuelle d’images – rappel du cours 1

Comprendre les données visuelles à grande échelle

Cours 4: recherche d’images, 14 novembre 2019

NAVER LABS EUROPE

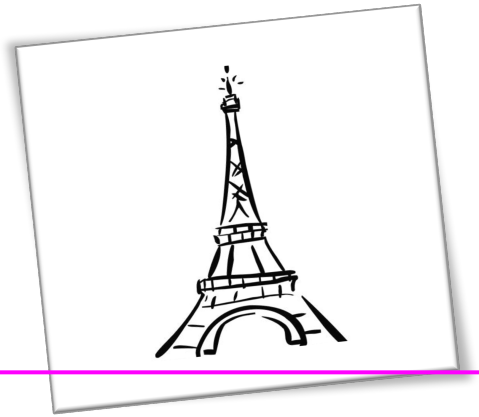
Visual Search - Principle

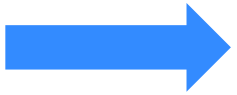


NAVER LABS EUROPE

Inherent ambiguity

What can the user mean with such a single query?





Application dependent!

Cours 1

- Inject prior information
  - Hand-crafting detectors / descriptors to obtain some properties (repeatability, invariance, discrimination, compactness, etc.)
- Leverage training
  - Provided with training data, learn a descriptor appropriate for the task

Ce cours

Rappel cours 1 !

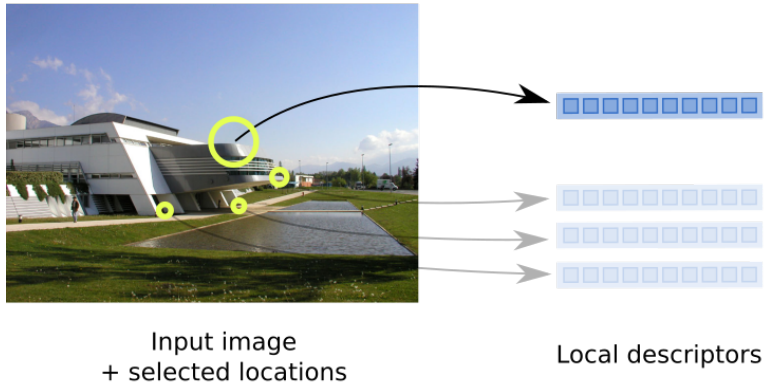
Rappel: variation d'apparence d'une **instance** d'objet donné

NAVER LABS EUROPE

Rappel cours 1 !

Local representations

Summary of the pipeline



Input image
+ selected locations

Local descriptors

Dataset of local desriptors

Bottleneck: pairwise matching

Pairwise matching between,
- local descriptors of the query image
- local descriptors of each gallery

Short list of images with enough pairwise matches

Geometric verification (RANSAC) of the short list of images

Shorter list of retrieved images

NAVER LABS EUROPE

Rappel cours 1 !

Scaling the first selection process

1) Efficient approximate nearest neighbor search on local descriptors

- Randomized K-d trees and variants
  - Priority queue to examine most promising first

[Silpa-Anan & Hartley 2008, Chum et al 2008, Muja & Lowe. 2009]

- Locality-Sensitive Hashing (LSH)
  - Randomized hashing technique
  - Data-dependent variants: spectral hashing, semantic hashing

[Indik & Motwani, 1998, Weiss et al NIPS08, Salakhutdinov & Hinton, SIGIR07]

2) Quantize local descriptor space into a visual codebook

- Uses **inverted files** to store the descriptors
- Efficient image matching

NAVER LABS EUROPE

Rappel cours 1 !

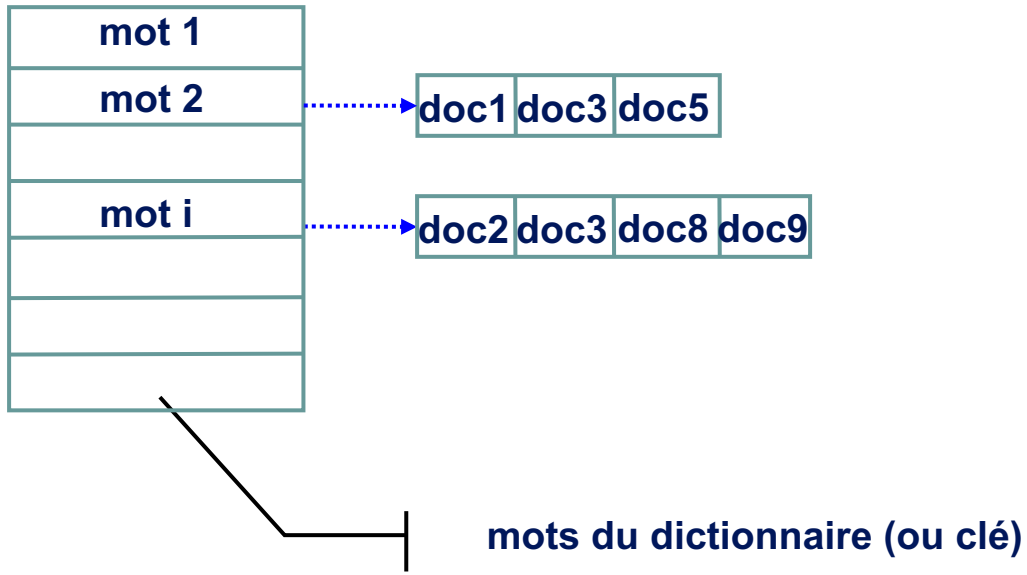
Plan de la suite

- Concept de fichier inversé: motivation
- Exemple d'utilisation d'un fichier inversé dans le cas de la recherche de texte
- Comment créer un vocabulaire visuel pour utiliser cette technique avec des représentations locales d'images
- Application d'un fichier inversé au cas de la recherche d'images

Rappel cours 1 !

Fichier inversé (*inverted file*)

- Cette structure liste des éléments qui ont une valeur donnée pour un attribut
- Exemple courant d'utilisation : requêtes sur documents textuels (mails, …)



- La requête consiste à récupérer la liste des documents contenant le mot
  - coût proportionnel au nombre de documents à récupérer

Rappel cours 1 !

Recherche de documents textuels (2)

- Exemple
  - vocabulaire = {“vélo”, “voiture”, “déplace”, “travail”, “école”, “Grenoble”}
  - espace de représentation: $\mathbb{R}^6$
  - “Grenoble est une belle ville. Je me déplace à vélo dans Grenoble”

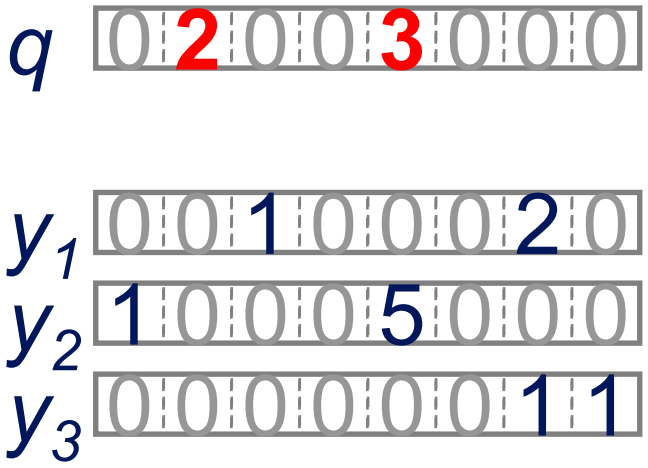
  $f=(1/4,0,1/4,0,0,1/2)^t$

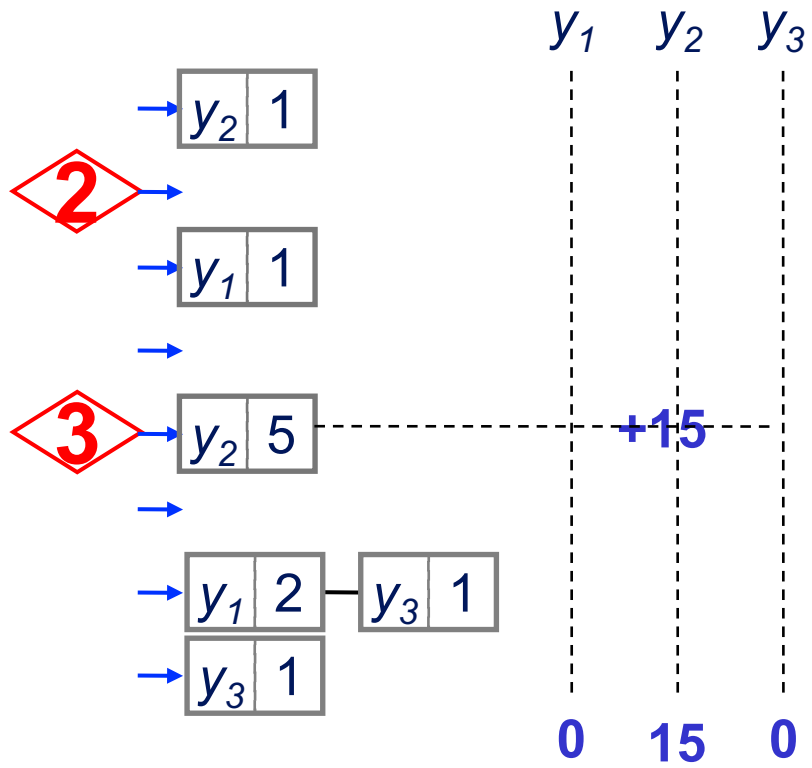
- Recherche de document = trouver les vecteurs les plus similaires à un document requête, représenté par le vecteur *q*
  - au sens d’une mesure de (dis-)similarité, en particulier le produit scalaire

Rappel cours 1 !

Fichier inversé : distances entre vecteurs creux

- La **requête** q et les **éléments de la base** Y sont des vecteurs creux
- Fichier inversé: calcul efficace du produit scalaire (en fait, toute distance L_p)
- Exemple pour le produit scalaire





Rappel cours 1 !

Recherche de documents visuels

- Problème: l'espace des descripteurs locaux est continu
  (par exemple l'espace des SIFT)
  - Il existe potentiellement une infinité de descripteurs locaux possibles
  - Il n'existe pas de concept de vocabulaire = 'ensemble de mots' pour les *patchs* d'images tel que c'est le cas pour le texte

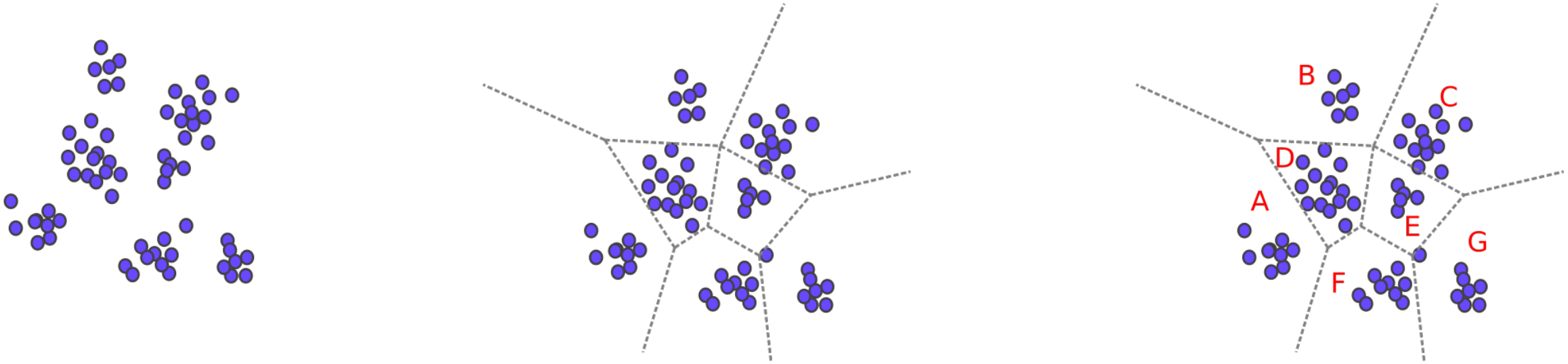
- Solution: créer un **vocabulaire visuel** !
  - Aussi appelé *visual vocabulary* or *visual codebook* dans la littérature
  - Construction par quantification de l'espace des descripteurs locaux

Rappel cours 1 !

Quantization: principle

Principle:

- Discretize the local descriptor space, e.g. SIFT desriptors
- 2 descriptors match i.i.f. they fall in the same bin
- this creates a visual codebook, similar to the textual codebook seen before



NAVER LABS EUROPE

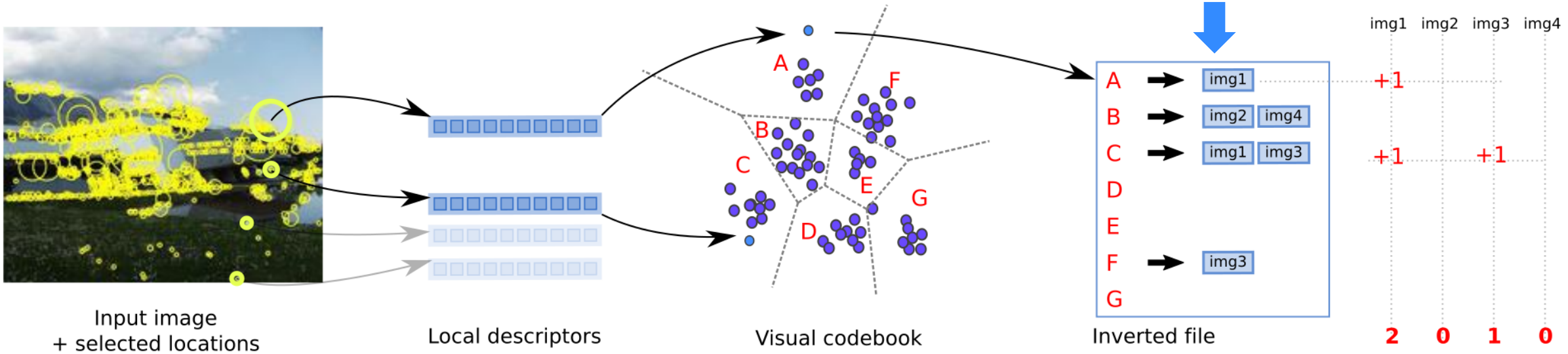
Rappel cours 1 !

Leverage quantization for efficient retrieval

Create an inverted file with all descriptors of the gallery set

Use this for efficient matching of a query image

[Sivic & Zisserman. ICCV 2003]



NAVER LABS EUROPE

Rappel cours 1 !

Description locale des images : résumé et commentaires

- **Les différentes étapes**
  - Sélection de points ou de régions d'intérêt (ex: Harris ou Harris-Affine)
  - Calcul des descripteurs pour chacune de ces régions (ex: SIFT, CS-LBP)
  - Algorithme de mise en correspondance entre images (ex: calcul exhaustif des distances entre paires de descripteurs, utilisation d'un fichier inversé)
  - Vérification de la cohérence géométrique (ex: RANSAC)

Description locale des images : résumé et commentaires

- **Commentaires sur les représentations locales**
  - En pratique: cette approche donne de très bons résultats
  - Principales limitations:
    - il faut apparier les descripteurs d'une image avec tous les descripteurs de la base d'image
      → c'est très lent !
    - Il faut stocker tous les descripteurs locaux des images dans la base
    - la vérification de la cohérence géométrique est relativement lente
      → on ne peut l'utiliser que pour un sous-ensemble d'images *a priori* plus pertinentes

- **Approche alternative**
  - Utilisation de **descripteurs globaux** qui agrègent les descripteurs locaux individuels en un descripteur global par image
    - Les descripteurs locaux ne sont plus stockés, seul le descripteur agrégé l'est
  - Le calcul de la similarité entre images se fait par comparaison directe des descripteurs globaux
    - Possibilité d'utiliser des similarités très efficaces
  - **Descripteurs globaux:**
    - **Sacs-de-mots** (cours 2 + ce cours)
    - **VLAD** et **Fisher Vector** (qui seront présentés plus tard dans le cours)

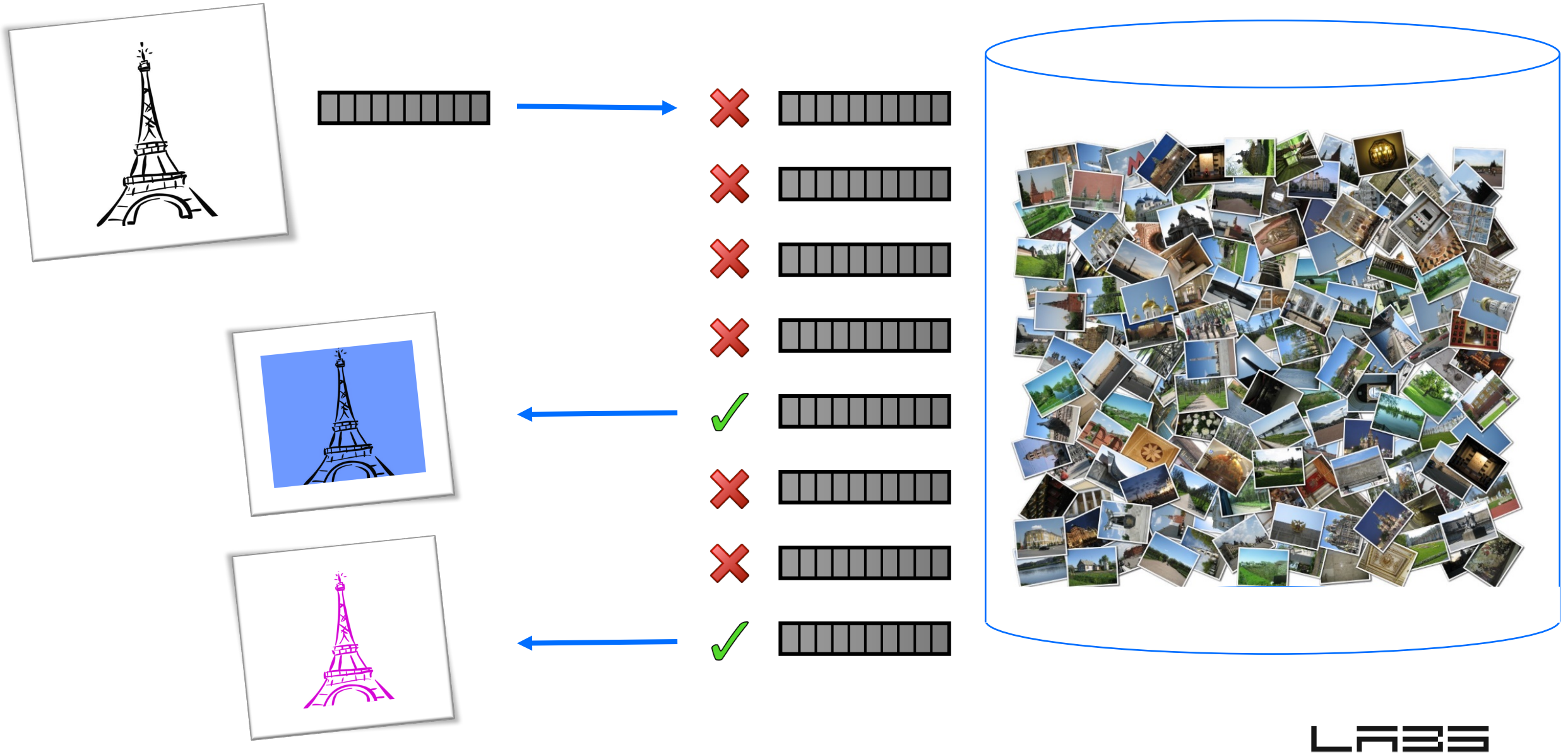
Notions de similarité, de mesures, et d’évaluation

Comprendre les données visuelles à grande échelle

Cours 4: recherche d’images, 14 novembre 2019

NAVER LABS EUROPE

Comment évaluer un système de recherche d'information ?



NAVER LABS EUROPE

Prérequis pour l'évaluation d'un tel système

- Avoir à disposition
  - un ensemble de test (base dans laquelle on recherche)
  - un ensemble de requêtes (peuvent être incluses dans la base de test)
  - une vérité terrain (*ground truth*) pour chaque couple (requête, élément de la base) qui répond à la question : est-ce que l'élément de la base est pertinent pour la requête considérée ?

- Remarques (similaires aux remarques précédentes, mais pour le cas de la recherche d'images)
  - pour comparer deux méthodes, les mêmes ensembles de test et de requêtes doivent être utilisés
    → bases de tests partagées par les chercheurs du domaine
      - Ex: Oxford, Paris (https://github.com/filipradenovic/revisitop)

    → compétition avec introduction de nouvelles bases de test
      - Ex. Landmarks Retrieval 2018 & 2019 (https://www.kaggle.com/c/landmark-retrieval-challenge)
  - la taille de ces ensembles doit être suffisamment grande pour diminuer la variance de l'évaluation
  - attention bases trop faciles / trop difficiles → diminue la sensibilité

Précision/rappel (début)

- Soit *E* un ensemble d'objets (par exemple de textes, d'images, ou de vidéos) muni d'une mesure *q* telle que
  - pour x, y $\in$ *E*, $q(x,y) = 0$ si y est pertinent pour x
    $q(x,y) = 1$ sinon

Remarque: on suppose ici la symétrie de la relation *q*

- Cette mesure est la vérité terrain
- Exemple : x et y sont deux images
  $q(x,y) = 0$ si x et y contiennent le même objet
  $q(x,y) = 1$ sinon

- Soit un ensemble *E'* $\subset$ *E*, et x : x $\in$ *E* et x $\notin$ *E'*
  - *E'* : ensemble dans lequel on effectue la recherche
  - x : la requête

Précision/rappel (suite)

- Le système de recherche est paramétré pour retourner plus ou moins de résultats, entre 1 et *#E'.*
- Compromis :
  - plus on retourne de résultats, plus on a de chance de retourner tous les objets pertinents de la base
  - en général, moins on en retourne, plus le taux d'objets retournés et qui sont pertinents est élevé
- Ces deux notions sont couvertes par les mesures de précision et de rappel

Précision/rappel (suite)

- Soit *R* l'ensemble des résultats retournés, de cardinal #*R*
- Soit *P* l'ensemble des résultats pertinents dans *E'* pour x, c-a-d

$$P = \{ y \in E' / q(x,y) = 0 \}$$

- Soit *A* l'ensemble des résultats retournés et qui sont pertinents

$$A = \{ y \in R / q(x,y) = 0 \}$$

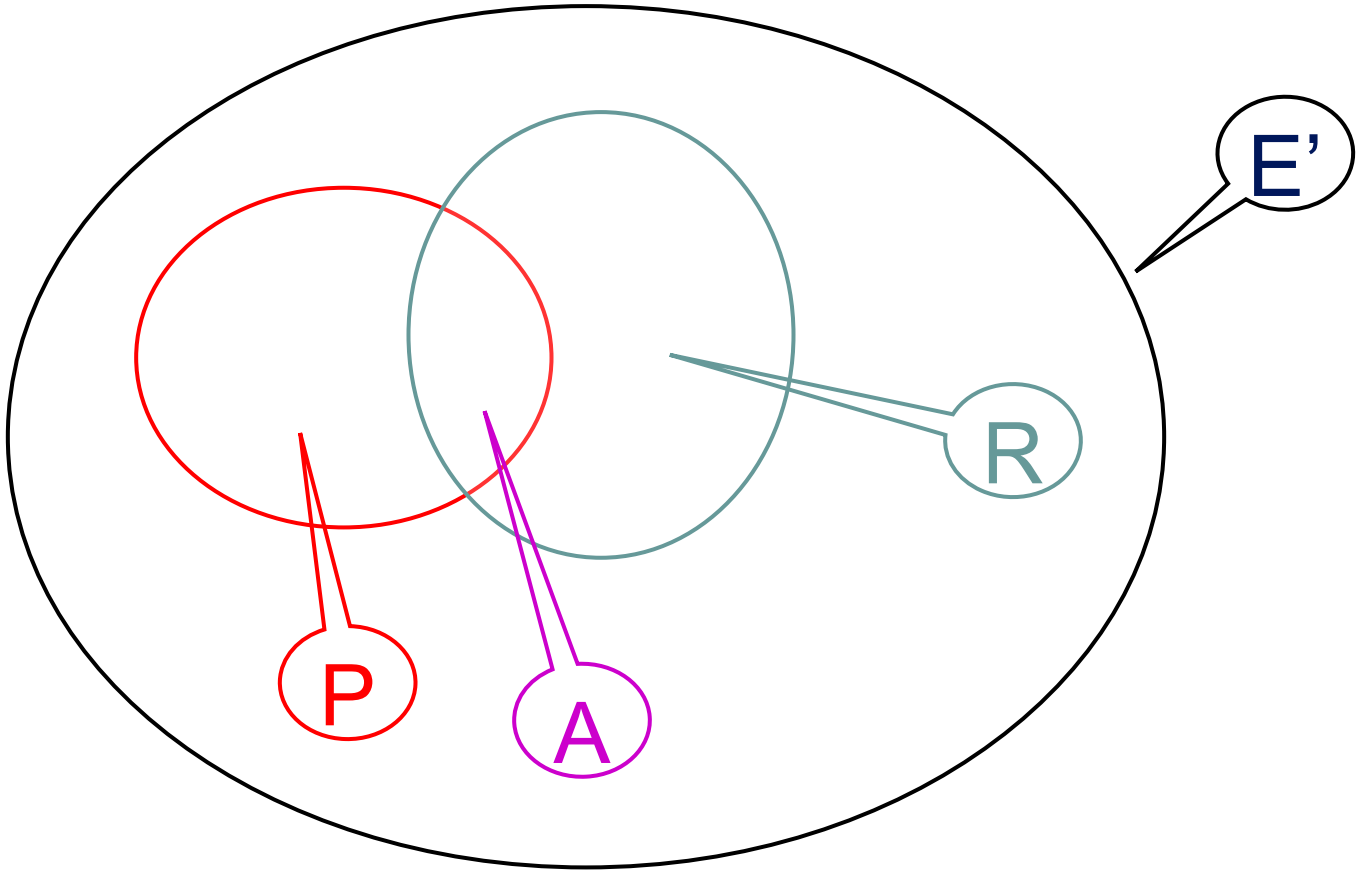
DEFINITION : la **précision** = #*A* / #*R* est le taux d'éléments qui sont pertinents parmi ceux qui sont retournés par le système

DEFINITION : le **rappel** = #*A* / #*P* est le taux d'éléments pertinents qui sont retournés par le système parmi les pertinents

- La performance du système peut être décrite par une courbe précision/rappel, en faisant varier la taille de l'ensemble des résultats retournés

Précision/rappel (suite et presque fin)

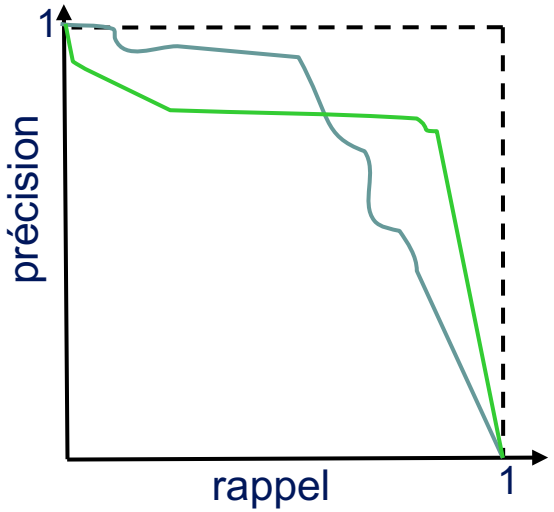
- Remarques :
  - *P* est indépendant de la paramétrisation.
  - *R* varie en fonction de la paramétrisation (qui retourne + ou – de résultats)



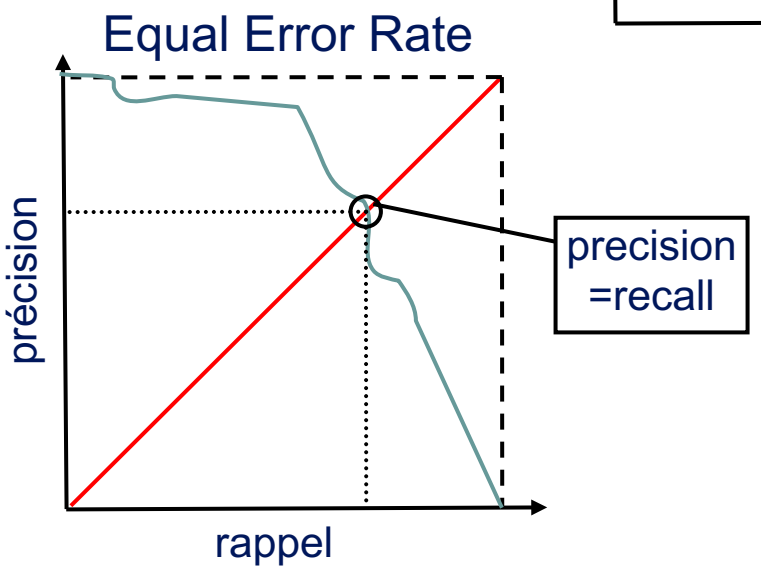
Précision/rappel (suite)

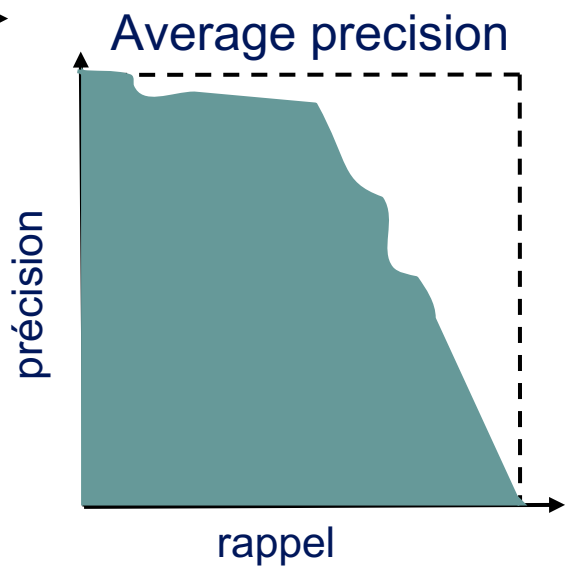
- *Equal Error Rate* et *Average Precision*:

réduire la courbe précision-rappel à une mesure de performance



Quelle est la meilleure méthode ?





NAVER LABS EUROPE

Exercice : système de recherche d'objets

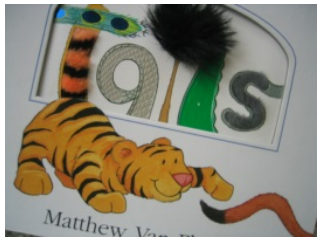
- Pour la requête et les résultats triés suivants : tracer la courbe précision/rappel, calculer la précision moyenne (*Average Precision*)

Proposition: faire l'exercice chez vous – me demander si un problème

1 2 3 4 5




6 7 8 9 10


NAVER LABS EUROPE

Et la pertinence ?

DEFINITION: la pertinence d'un système (pour une paramétrisation donnée) est le taux d'objets qui sont correctement jugées, c'est-à-dire

pertinence = (vrais positifs + vrais négatifs) / taille de la base

- En recherche d'information (image, texte, etc.)
  c'est une **mauvaise** mesure de la qualité du système
  - en général, la plupart des objets ne sont pas pertinents
  - un système qui renverrait systématiquement "négatif" serait quasiment imbattable avec cette mesure
- Intérêt d'avoir des courbes (par exemple PR) pour l'évaluation
  - dépend de l'utilisation : certains utilisateurs cherchent la précision (ex: requête sur Google), d'autres un grand rappel possible (recherche de contenu piraté)
  - *operating point*: choisir le bon point de fonctionnement du système

Mesures et protocoles d'évaluation : conclusion

- Difficulté de trouver une bonne mesure
  - elle doit être adaptée à ce que l'on compare (ex: vecteurs, loi de probabilité, etc.)
  - elle doit répondre à l'objectif recherché

- Évaluation d'un système de recherche dans des données visuelles
  - méthodes identiques à celles utilisées en texte
  - utilisation de courbes plutôt que de scalaires (peuvent être interprétées en fonction du besoin)
  - n'intègrent pas les mesures de similarité, juste leur rang!

Représentations globales

Comprendre les données visuelles à grande échelle

Cours 4: recherche d'images, 14 novembre 2019

NAVER LABS EUROPE

Description globale

- **Une description globale** est une représentation de l'image dans son ensemble, sous la forme d'un vecteur de taille fixe
- Caractéristiques
  - **un** vecteur de description par objet visuel
  - mesure de (dis-)similarité définie sur l'espace de ces descripteurs

Rappel cours 2 !

Agrégation de descripteurs locaux

- Définir un descripteur global à partir de l'ensemble des descripteurs locaux d'une image
  - Compact, et plus facile à manipuler que les descripteurs locaux de départ
- Contraintes
  - Des images similaires doivent avoir des représentations similaires
  - Des images dissimilaires doivent avoir des représentations dissimilaires
- Compromis
  - Robustes aux transformations (échelle, occultation, éclairage, etc.)
  - Informatifs (bonne description du contenu)
  - Efficaces à calculer, à stocker, à manipuler

Rappel cours 2 !

Agrégation de descripteurs locaux

- La base: « *bag-of-features* », « *bag-of-patches* »
  - *bag* = on perd l'ordre, la géométrie.
    On utilise des ensembles non-ordonnés de descripteurs locaux.

- Quantification: représentation par sac-de-mots,
  ou *bag-of-words* (BoW, aussi appelée *bag-of-visual-words* ou BoV)
  - On suppose une transformation : descripteur local -> index entier
    (par un algorithme de quantification vectorielle = *clustering*)
  - Cette transformation peut être vue comme la création d'un vocabulaire visuel
  - Histogramme de ces entiers = descripteur global

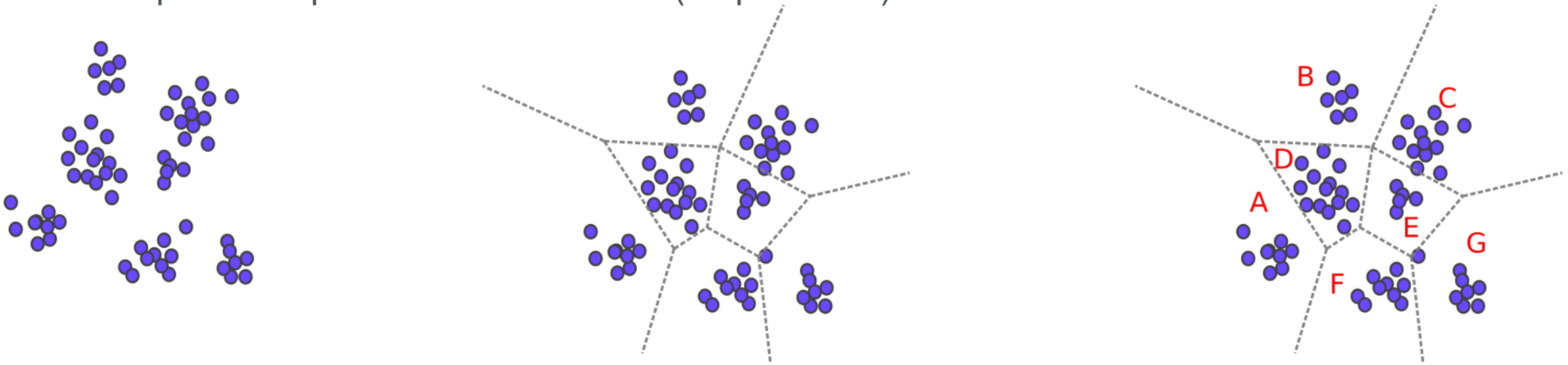
Rappel cours 2 !

Agrégation de descripteurs locaux

Utilisation d'un vocabulaire visuel

Etapes:

- Discrétisation de l'espace des descripteurs, par exemple avec un algorithme de *clustering*
- Chaque descripteur est associé à un (ou plusieurs) mot visuel



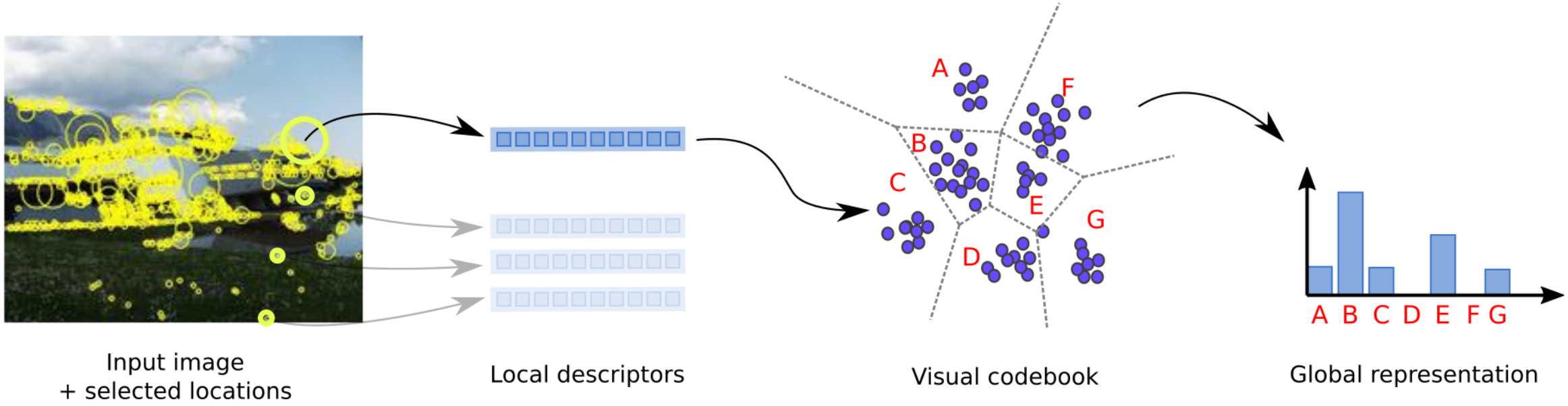
Rappel cours 2 !

From quantization to bag-of-visual-features

Principle

[Sivic & Zisserman. ICCV 2003]
[Csurka et al. ECCV SLCV 2004]

- Extract local descriptors
- Convert local descriptors into visual words, using a visual codebook
- Represent images as a histogram of occurrences



NAVER LABS EUROPE

Rappel cours 2 !

Lien avec la partie sur les représentations locales

Début de ce cours

- Construction d'un vocabulaire visuel pour la construction d'un fichier inversé pour une recherche rapide

Ici

- Construction d'un vocabulaire visuel afin de pouvoir d'agréger les descripteurs locaux en une représentation globale
  - Les descripteurs locaux n'ont pas besoin d'être gardés en mémoire
  - Seule la représentation globale est conservée
  - Extrêmement compact, mais
  - pas de vérification géométrique possible,
  - perte totale des informatiques géométriques,
  - quantification = perte d'information -> représentation « grossière » (*coarse*)

Teaser

Dans le cours 7,

nous verrons comment construire des améliorations de la représentation par « sacs-de-mots », qui donnent de meilleurs résultats à un faible coût additionnel

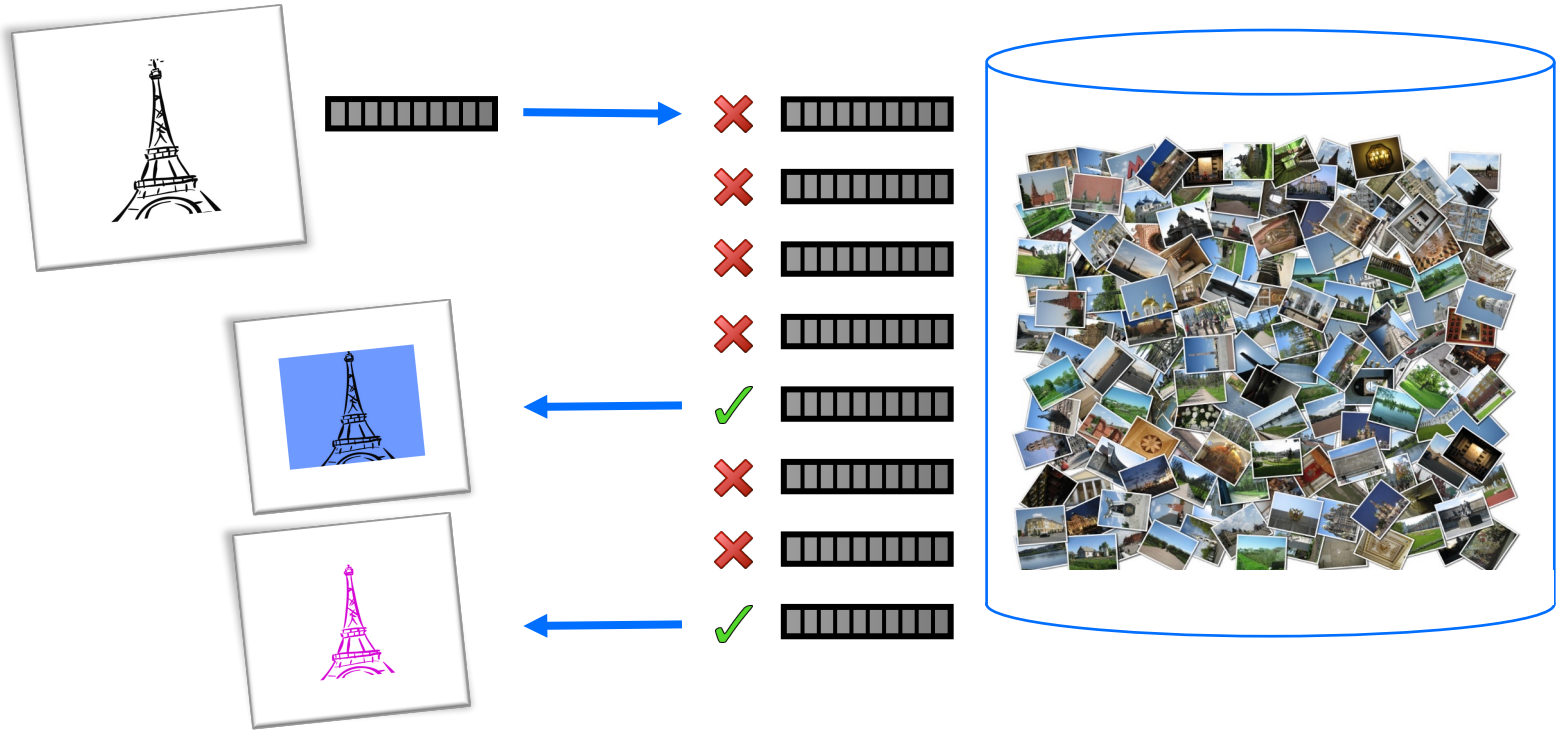
Représentation d’images par apprentissage profond

Comprendre les données visuelles à grande échelle

Cours 4: recherche d’images, 14 novembre 2019

NAVER LABS EUROPE

Object Search *a.k.a. Instance-Level Retrieval*



Families of representations

- Local representations
- Global representations
- **Deep representations**

Corresponding similarity measures

Représentation par apprentissage profond – première version

Principe

- Entrainer un réseau de neurones (ex: CNN) sur une tâche de classification
  - Par exemple pour la classification à 1000 classes sur ImageNet
- Utiliser ce réseau comme une “boite noire” pour extraire des représentations d’images
    - la sortie d’une des couches (traditionnellement l’avant dernière) est utilisée comme représentation
- Eventuellement normaliser ces représentations
- Les comparer avec une mesure de similarité simple, comme le produit scalaire

Motivation

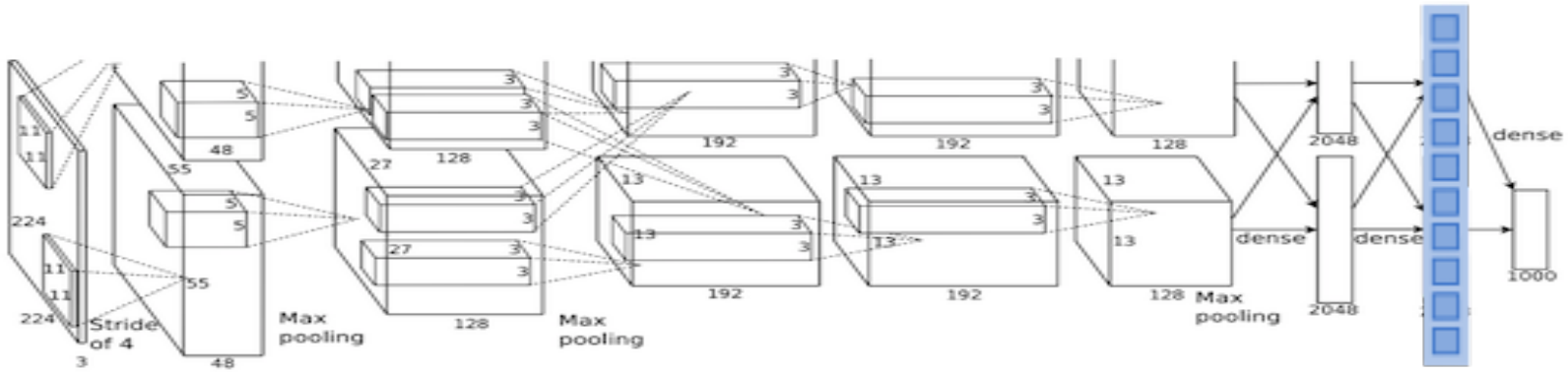
- Un bon modèle de classification capture des informations sémantiques et devrait permettre de bien représenter les images pour la tache de recherche aussi

Représentation par apprentissage profond – première version

L'architecture CNN pré-entrainée pour la classification est utilisée comme un extracteur de représentation

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

AlexNet



[Krizhevsky et al. NIPS 2012]

Input image

NAVER LABS EUROPE

Limitations de cette approche naïve

Limitations évidentes

- Le réseau a été entrainé pour de la classification, sur des catégories génériques, or on veut reconnaitre précisément des objets
  - Généralisation intra-class vs discrimination entre les différentes instances d'une classe
- L'architecture des réseaux utilisés pour la classification prend en général en entrée des images de faible résolution et de taille fixe (distorsion des images)

➢ En pratique, ce type d'approche a donné des résultats décevants

Solution

- Améliorer toutes les étapes et spécialiser l'approche à la tache de recherche d'images: méthodes récentes par apprentissage profond

Repenser les approches par apprentissage profond pour la recherche d'images

- **Une base d'apprentissage appropriée**, i.e. labélisée à l'échelle de l'instance et non de la catégorie d'objet. Correction des erreurs d'annotations si besoin.
- **Changer l'architecture** pour apprendre les détails: architecture qui accepte toute taille et rapport d'échelle d'images, même les images avec une grande résolution
- **Changer la méthode d'apprentissage**: apprendre explicitement pour la recherche d'images plutôt que d'apprendre à classifier. Cela nécessite une modification de l'architecture comme de la fonction de coût

Training for visual search

What can be improved?

1. Training data:

   Public landmark dataset is very noisy, needs to be cleaned automatically

2. Architecture:

   Small details are important for instance level retrieval: need to accommodate high resolution, undistorted images during training

3. Training objective:

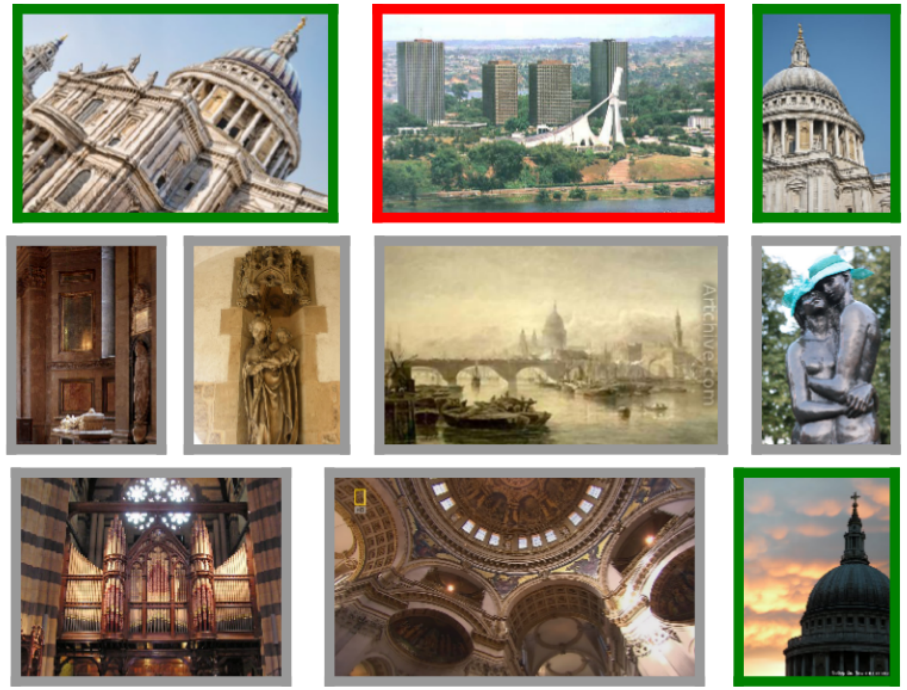
   We should train explicitly for retrieval, not for classification

1. Training Data

Public dataset of landmark images

- ~200K images
- ~600 different landmarks (Rome colosseum, Big Ben…)
- Many annotation errors

[Gordo et al. ECCV 16, IJCV17]



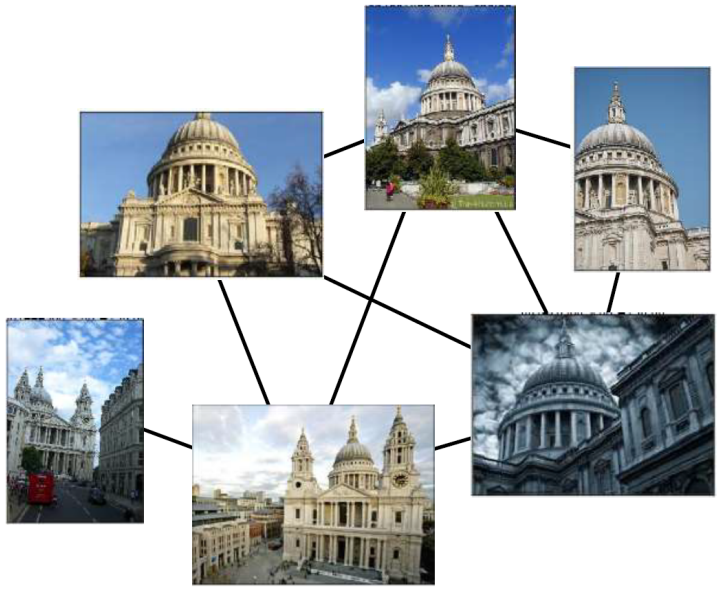
NAVER LABS EUROPE

1. Training Data

Nettoyage automatique de la base d'images existantes

[Gordo et al. ECCV 16, IJCV17]

- Utilisation d'une méthode standard basée sur des descripteurs locaux d'images pour connecter les images pertinentes
  - Détection de point d'intérêt: Hessien-Affine
  - Description des régions sélectionnées avec SIFT
  - Appariement entre les descripteurs d'images
  - Vérification géométrique avec RANSAC
- Construction d'un graphe
  - Seule la plus grosse composante connectée par classe est conservée



Résultat

- Sous-ensemble (40K) d'images spatialement vérifiées
- Annotation approximative de la boite englobante des objets

NAVER LABS EUROPE

2. Architecture

Utilisation d'une architecture qui conserve les détails de l'image

Le R-MAC est un descripteur global d'image qui combine

- l'utilisation d'un réseau pré-entrainé pour la classification pour extraire une description dense des images
- L'extraction de représentations par région
- L'agrégation de ces représentations par région en une représentation finale

[Tolias et al, ICLR16]

2. Architecture

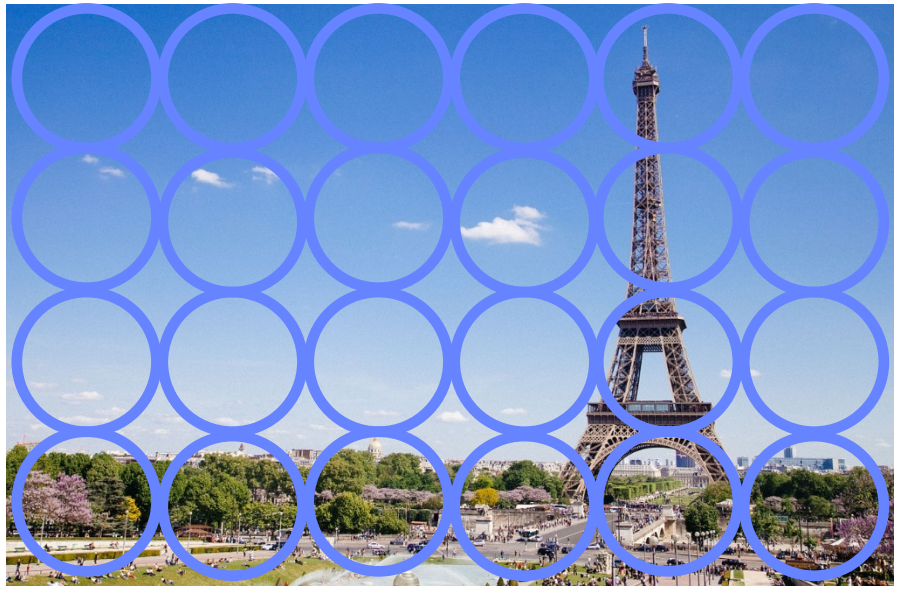
R-MAC descriptor

[Tolias et al, ICLR16]

Input image

CNN

Local features



NAVER LABS EUROPE

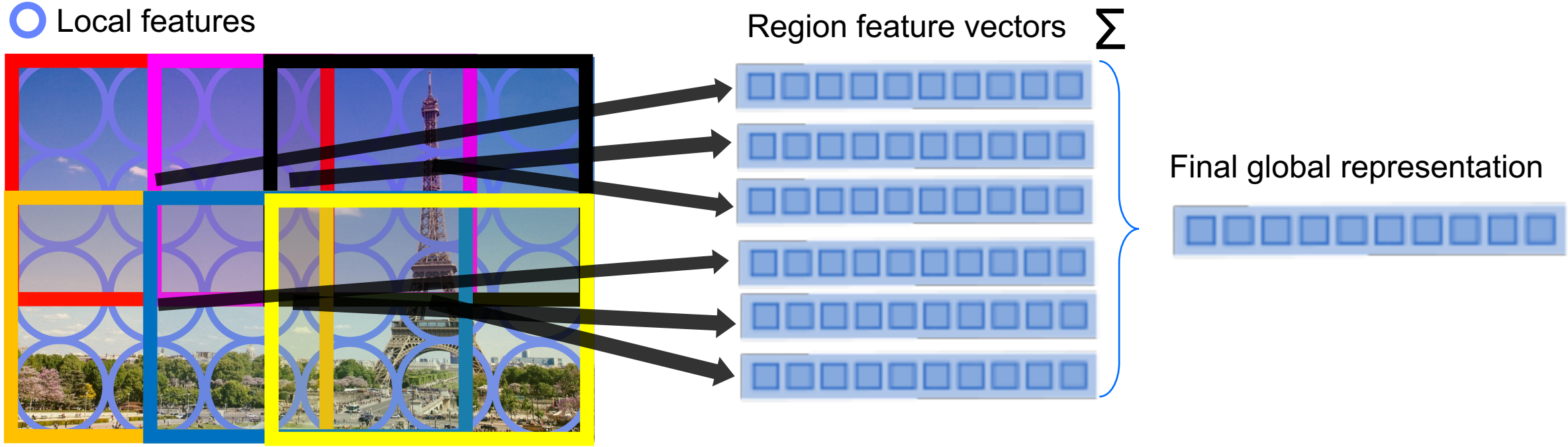
2. Architecture

R-MAC descriptor

[Tolias et al, ICLR16]

Advantages

- no aspect ratio distortion
- can encode high resolution images
- fast comparison with the dot product



NAVER LABS EUROPE

2. Architecture

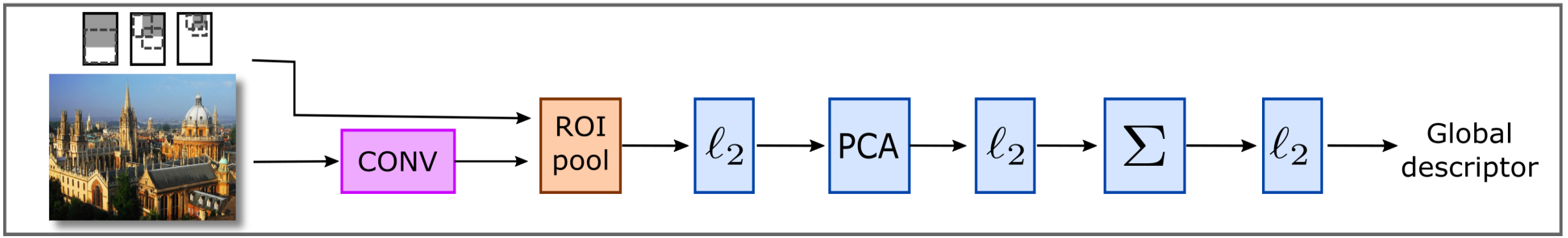
R-MAC descriptor

- CNN as a local feature extractor

Two key observations

[Gordo et al. ECCV 16, IJCV17]

1. The aggregation steps can be integrated inside the network:



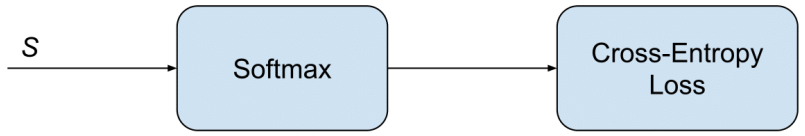
2. Every step is differentiable ⇒ **the model can be trained end-to-end!**

NAVER LABS EUROPE

3. Training for retrieval

Learning to rank

- Historiquement, les méthodes ont entrainé leurs réseaux de neurone avec une fonction de coût appropriée pour la classification (ex: *softmax loss*)



$$f(s)_i = \frac{e^{s_i}}{\sum_j^C e^{s_j}} \qquad CE = -\sum_i^C t_i log(f(s)_i)$$

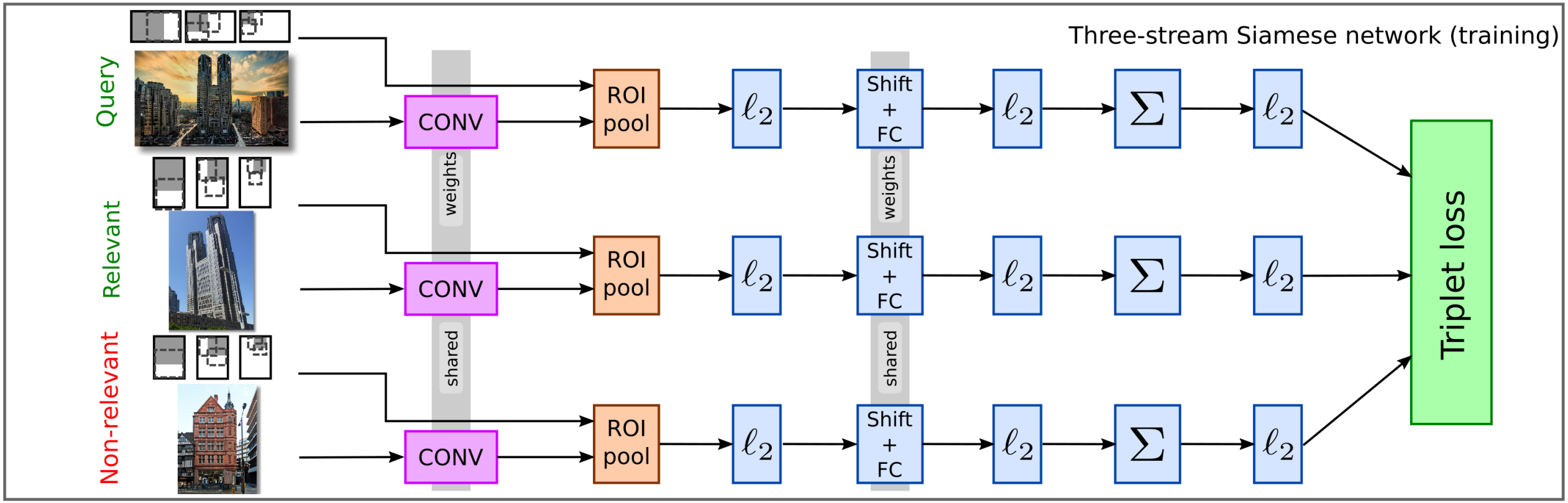
- Problème: on ne souhaite pas reconnaitre la classe, mais, étant donnée une image, ordonnée les images de la base de la plus pertinente à la moins pertinente
- Solution: considérer une fonction de coût qui s'intéresse directement au rang des images. Cela demande de modifier l'architecture du réseau d'apprentissage (architecture siamoise)
- Exemple de fonction de coût de ce type
  - **Contrastive** (perte contrastive) [Radenovic et al. ECCV 16, PAMI18]
  - **Triplet** (fonction d'erreur de triplets) -> exemple choisi dans la suite [Gordo et al. ECCV 16, IJCV17]

NAVER LABS EUROPE

3. Training for retrieval

[Gordo et al. ECCV 16, IJCV17]

Learning to rank



NAVER LABS EUROPE

3. Training for retrieval

[Gordo et al. ECCV 16, IJCV17]

Triplet loss

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

Query

Relevant

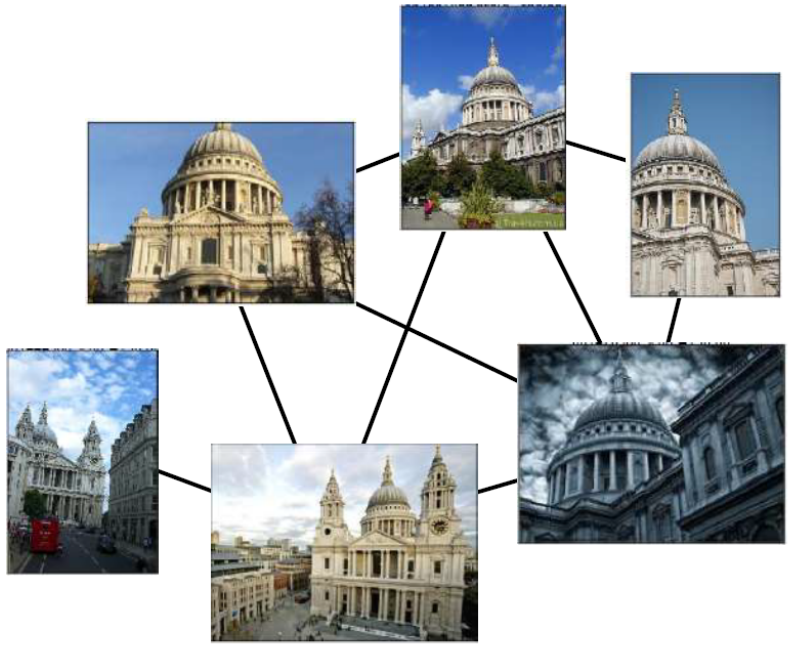
Non-relevant

NAVER LABS EUROPE

Summary

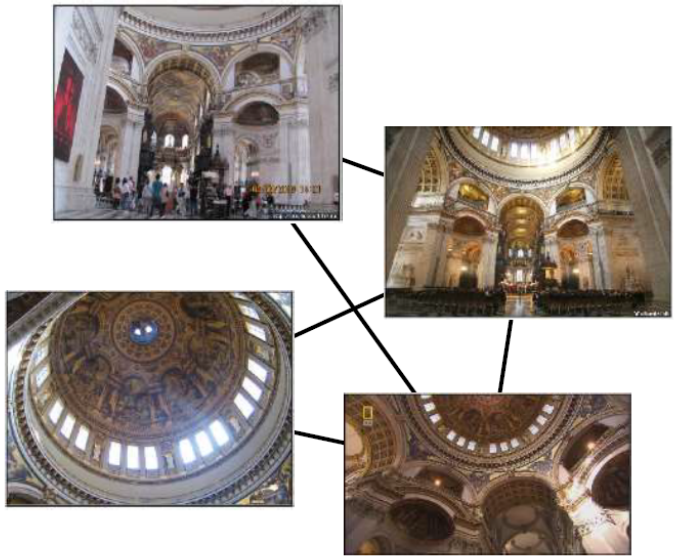
Problems:

1. Features from ImageNet are not good at intra-class discrimination



Solutions:

1. Train on an **automatically cleaned** dataset of landmarks



NAVER LABS EUROPE

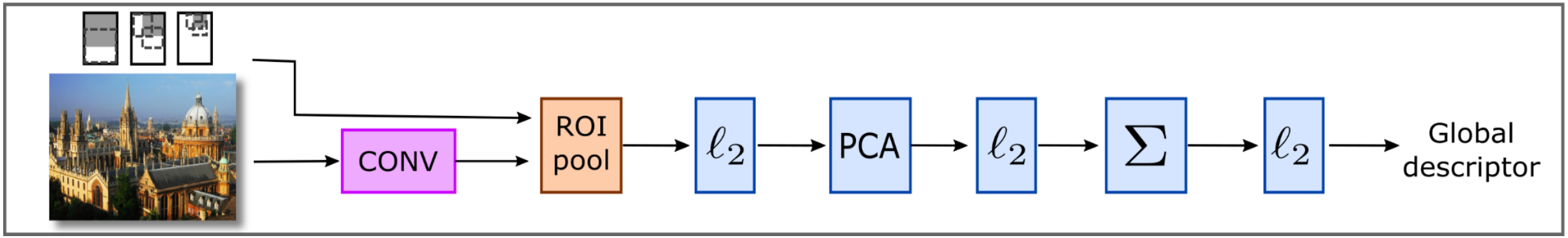
Summary

Problems:

1. Features from ImageNet are not good at intra-class discrimination
2. Usual architectures work on small crops of the image and at low resolution

Solutions:

1. Train on an **automatically cleaned** dataset of landmarks
2. Use an **architecture that preserves image details**



NAVER LABS EUROPE

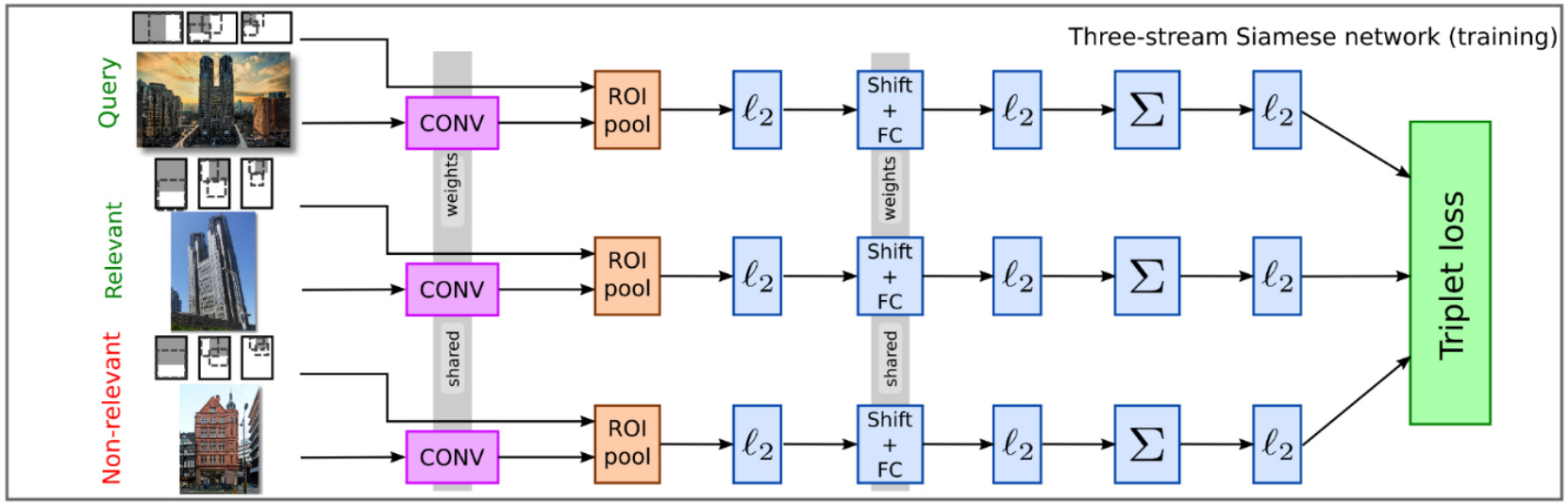
Summary

Problems:

1. Features from ImageNet are not good at intra-class discrimination
2. Usual architectures work on small crops of the image and at low resolution
3. Networks are typically trained for classification

Solutions:

1. Train on an **automatically cleaned** dataset of landmarks
2. Use an **architecture that preserves image details**
3. Train network with a **ranking loss**



NAVER LABS EUROPE

Evaluation on standard benchmarks

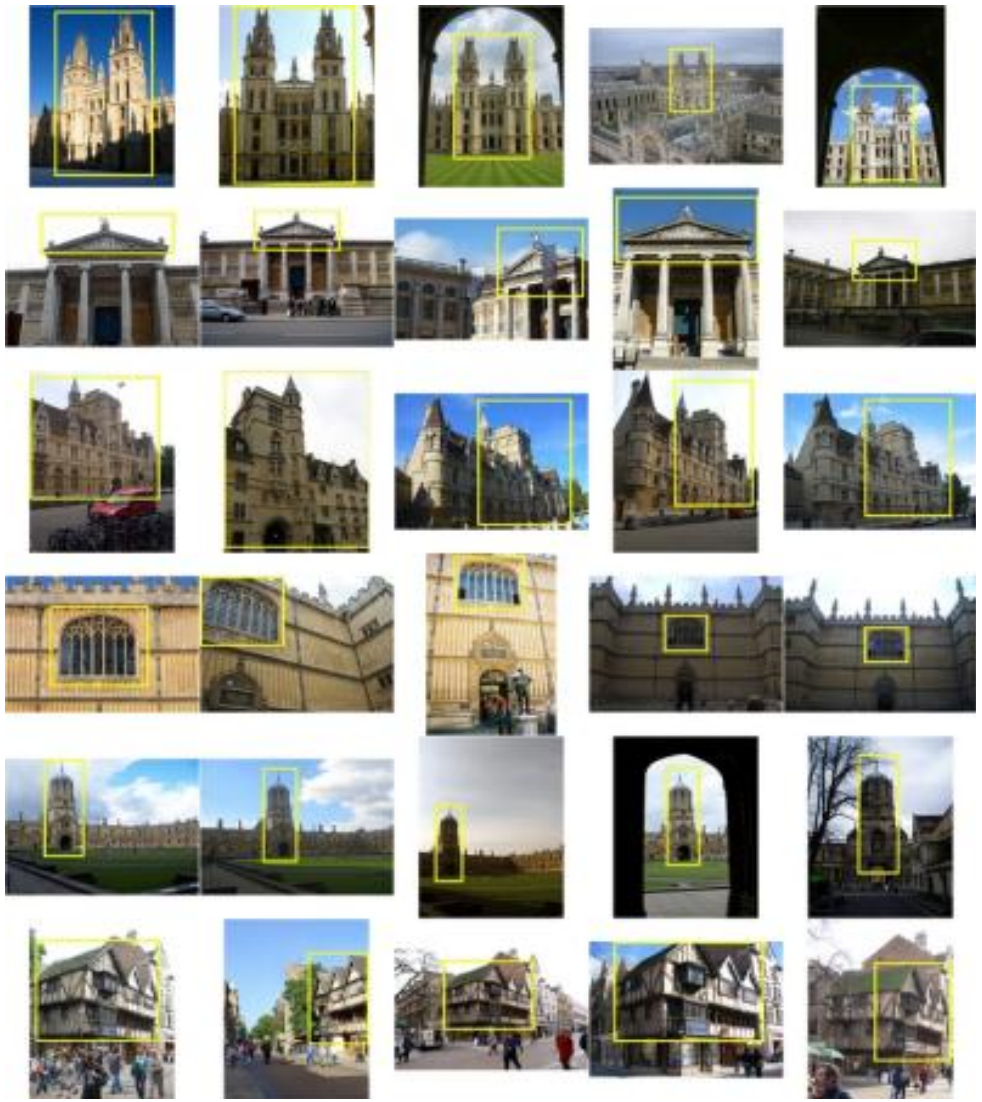
Oxford dataset

- 5,000 images
- 55 queries
- 11 landmarks

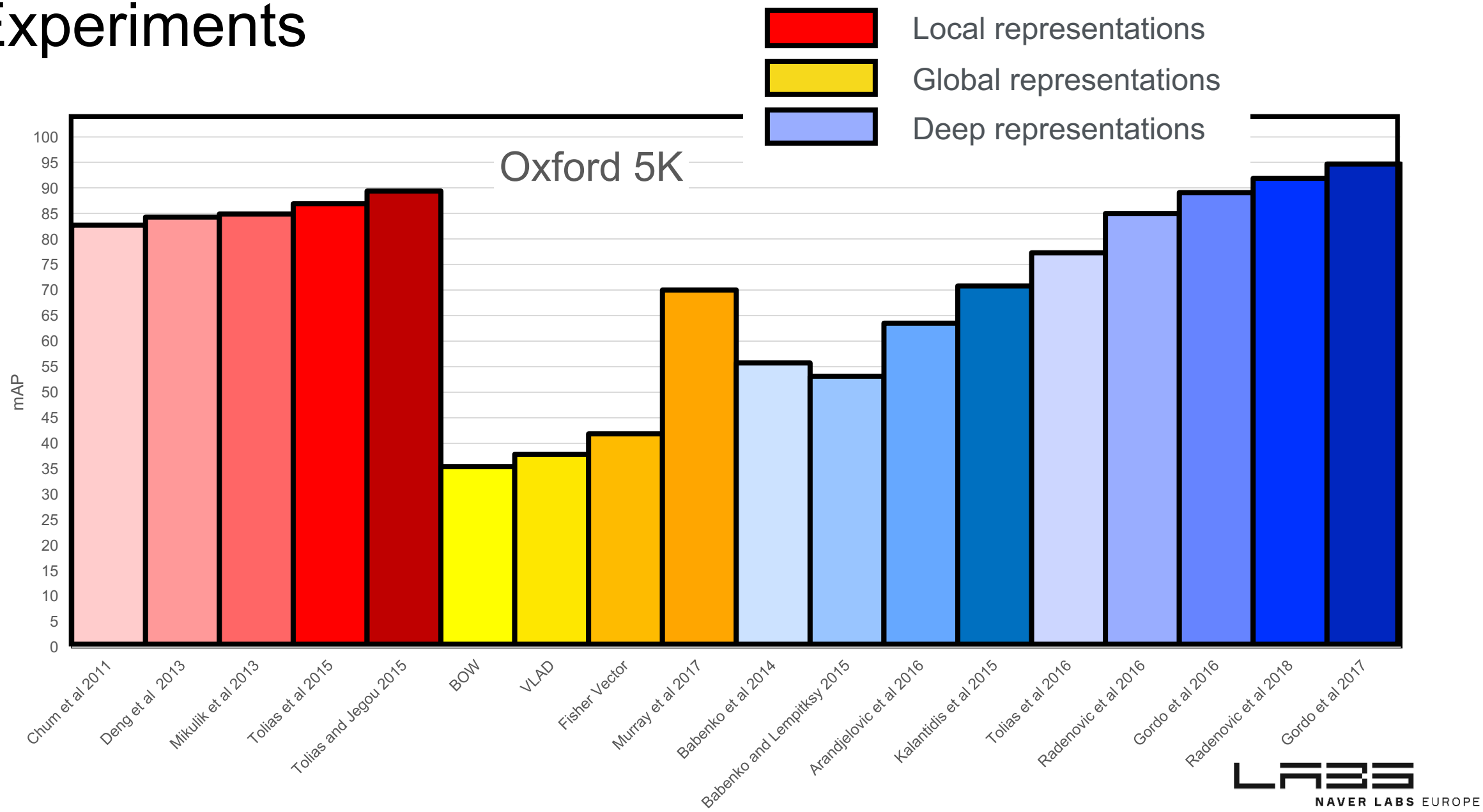
Evaluation

- mean Average Precision (mAP)

[Philbin et al. CVPR07]



Experiments



NAVER LABS EUROPE

Qualitative results

[Gordo et al. ECCV 16, IJCV17]

NAVER LABS EUROPE

Impact of fine-tuning an ImageNet model

Where do conv5 neurons fire? [Gordo et al. ECCV 16, IJCV17]

Before

After

NAVER LABS EUROPE

Impact of fine-tuning an ImageNet model

Where do conv5 neurons fire? [Gordo et al. ECCV 16, IJCV17]

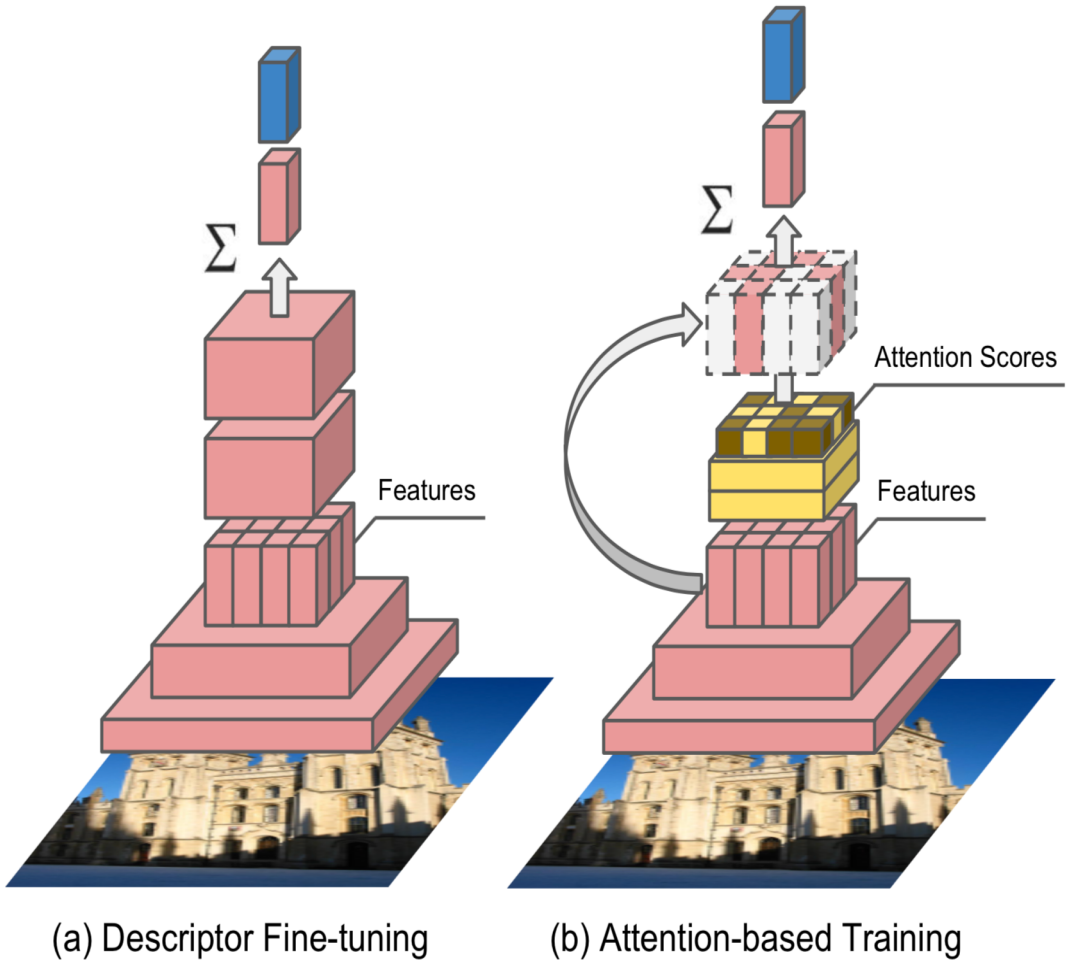
Before

After

NAVER LABS EUROPE

Représentations locales par apprentissage profond

[Noh et al. ICCV17]



Principle

- Dense Localized Feature Extraction: fully convolutional network, trained for the task
- Attention-based Keypoint Selection: a technique to effectively select a subset of the features
- Note: keypoint selection comes after descriptor extraction
- Selected regions can be used for geometrical verification

Advantages

- “deep version” of local approaches: drop-in replacement for keypoint detectors and descriptors
- Good results on large-scale datasets

NAVER LABS EUROPE