Comprendre les données visuelles à grande échelle

ENSIMAG
2018-2019

Karteek Alahari & Diane Larlus

January 10th 2019

Organisation du cours

- Séance du jeudi 10 janvier – 8h15 à 11h15
  - Troisième cours
  - Présentation article 2 + quizz

Cours 3: Représentation d'images avec application à la recherche visuelle - deuxième partie

Comprendre les données visuelles à grande échelle
10 janvier 2019

Représentations globales

Comprendre les données visuelles à grande échelle

Cours 3: représentations d'images, 10 janvier 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

Exemple de description globale : histogramme de couleur

- Chaque pixel est décrit par un vecteur de couleur
  - Par exemple un vecteur RGB $\in R^3$, mais le plus souvent on utilise un autre espace de couleur plus approprié
- L'ensemble des vecteurs de couleurs forme une distribution
  - Description de la distribution avec un histogramme
  - Nécessite la discrétisation de l'espace et la normalisation de l'histogramme
- Comparaison de deux histogrammes par une mesure de dissimilarité, par exemple la « distance » du Khi-2

Color indexing, Swain & Ballard, IJCV 1991

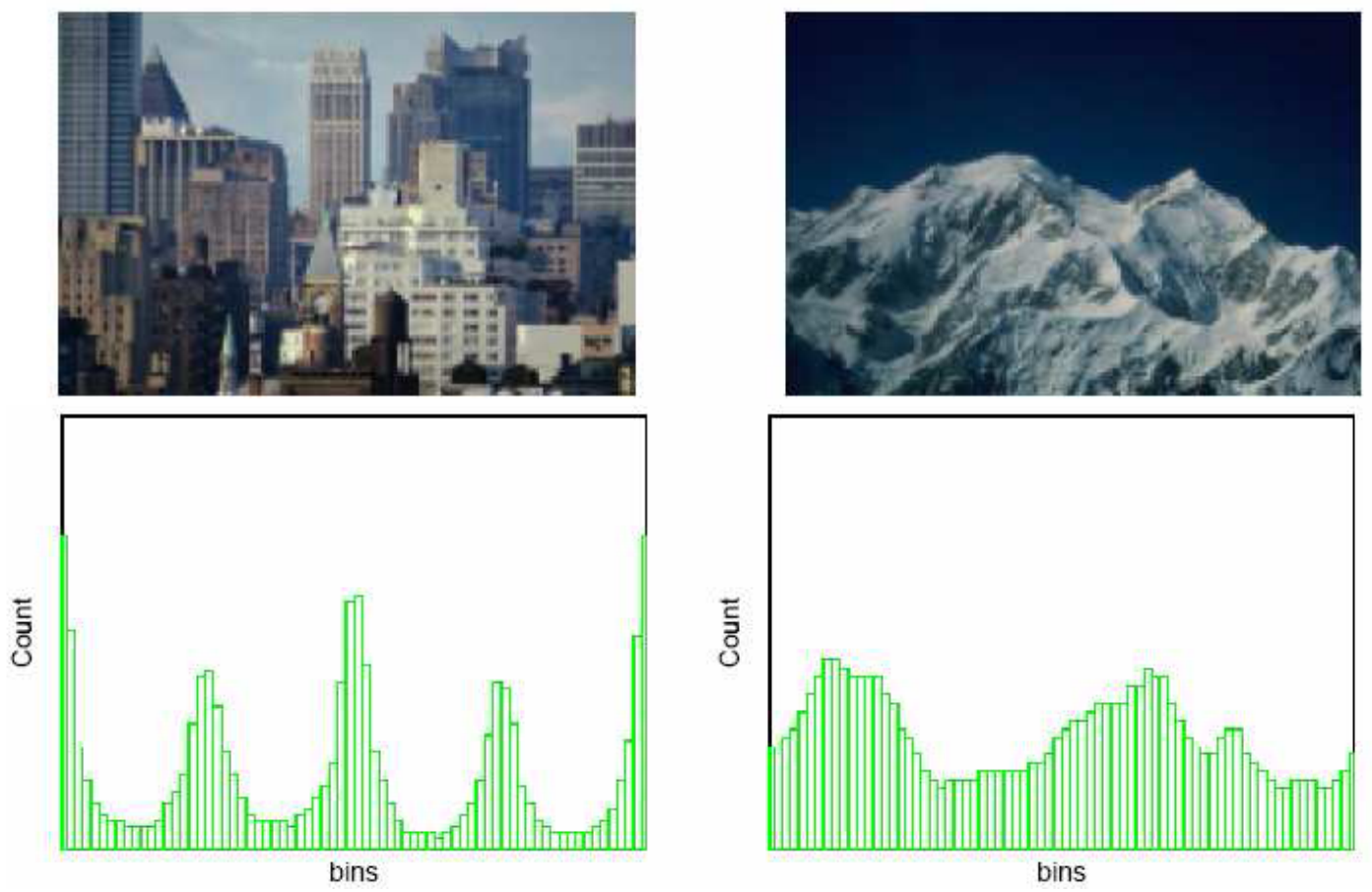
The earth mover's distance, multi-dimensional scaling, and color-based image retrieval
Y Rubner, LJ Guibas, C Tomasi - Proceedings of DARPA Image, 1997

Visualisation des distances pour une représentation basée sur la couleur

NAVER LABS EUROPE

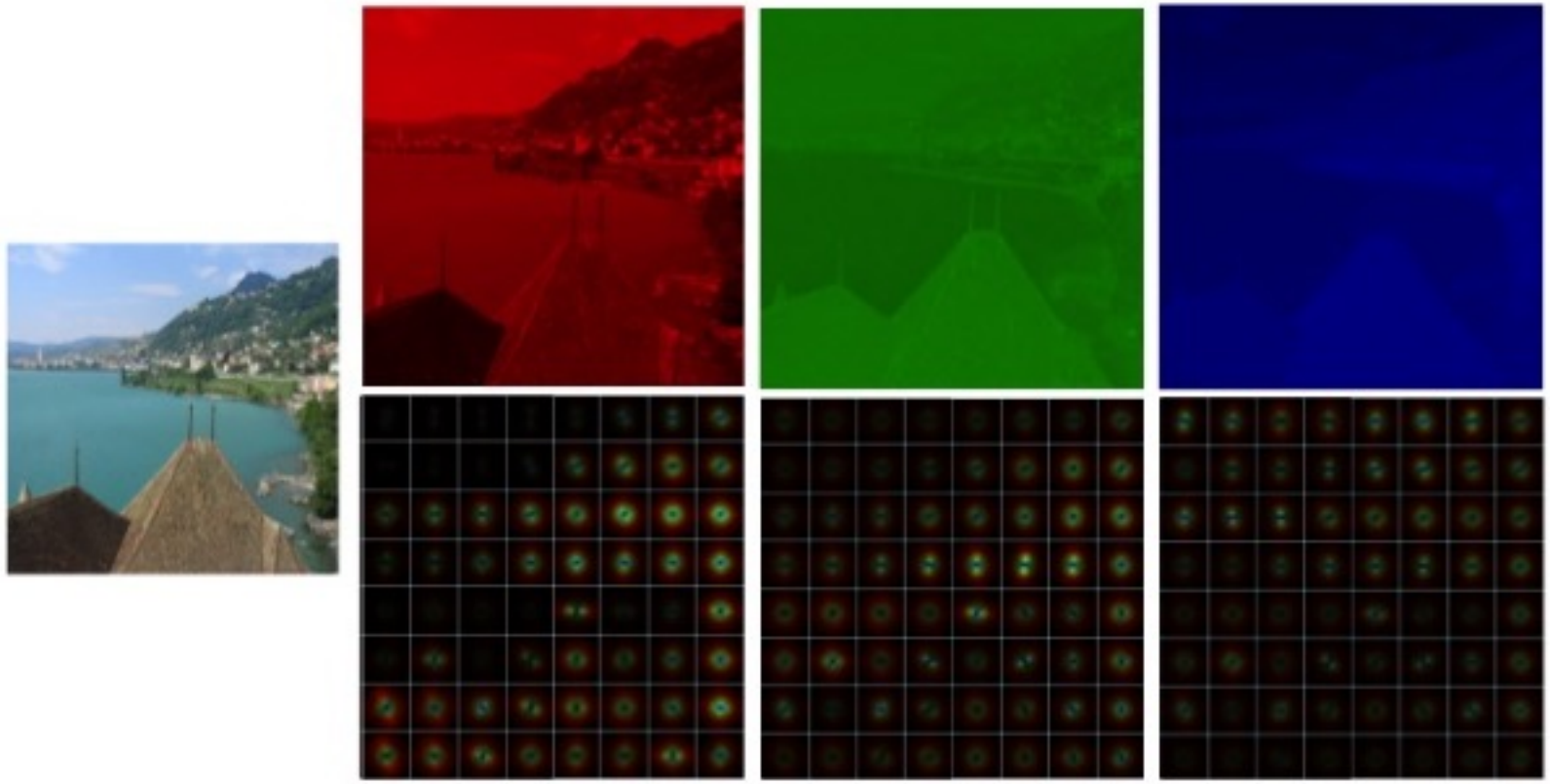
Exemple de description globale : contours

- Exemple de descripteur de texture
- Descripteur = histogramme d'orientation des contours



Apparence globale de l'image : descripteur GIST

- Similaire au descripteur SIFT sur une pyramide d'images, avec image = patch.
- Peut prendre en compte la couleur si appliqué sur les canaux R, G, B séparément.



- [Modeling the shape of the scene: a holistic representation of the spatial enveloppe, Aude Oliva, Antonio Torralba, IJCV 2001], [Gist of the scene, A. Oliva, Neurobiology of attention, 2005]

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
- Nécessité d’un compromis entre ces propriétés
  - Robustes aux transformations (échelle, occultation, éclairage, etc.)
  - Informatifs (bonne description du contenu)
  - Efficaces à calculer, à stocker, à manipuler

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.

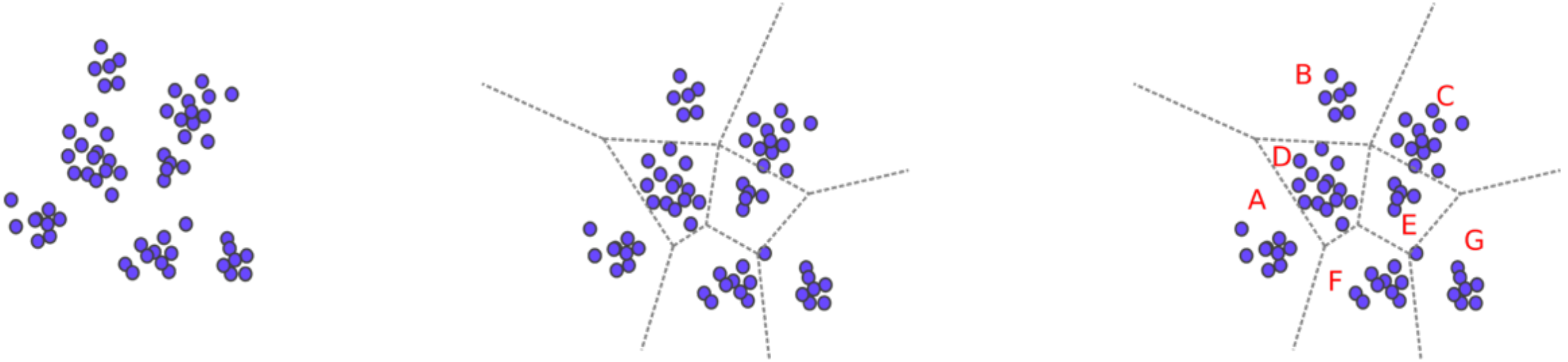
- Quantification: représentation par sac-de-mots,
  ou *bag-of-words* (BoW, aussi appelée *bag-of-visual-words* ou BoV)
  - On suppose une transformation : descripteur local -> index entier
    (par un algorithme de quantification vectorielle = *clustering*)
  - Cette transformation peut être vue comme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

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

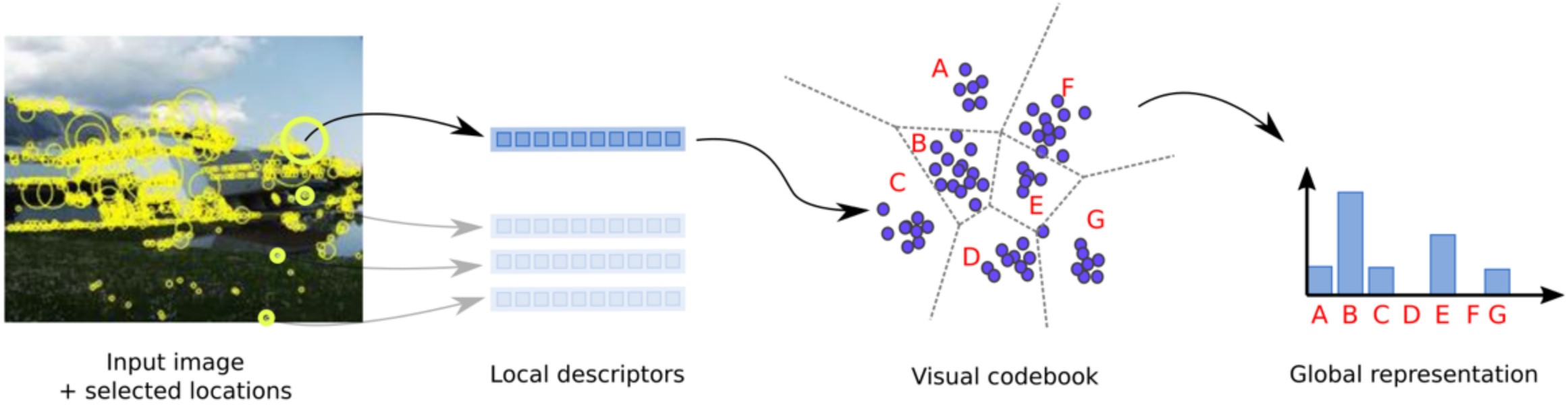


From quantization to bag-of-visual-features

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

Lien avec le cours précédent

La semaine dernière

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

Cette semaine

- Construction d'un vocabulaire visuel afin 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*)

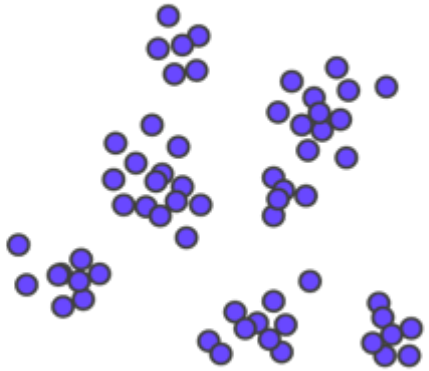
How can we refine this description?

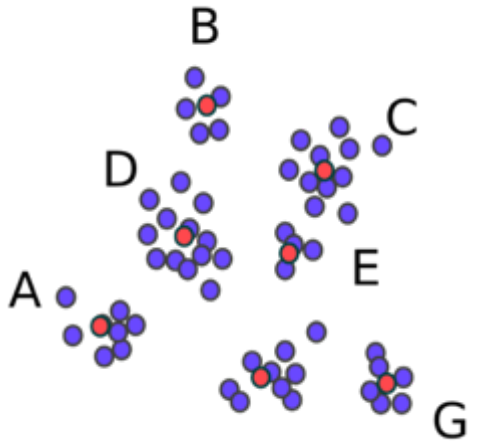
Relatively coarse representation

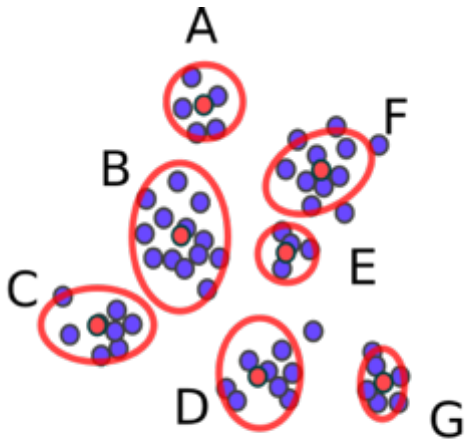
- Solution 1: more entry in the codebook
  - drawback: significant computational cost

[Li et al. ICCV 2009]
[Yang et al. ECCV 2010]

- Solution 2: beyond counting, adding higher order statistics
  - Mean: VLAD
  - Variance: Fisher Vector





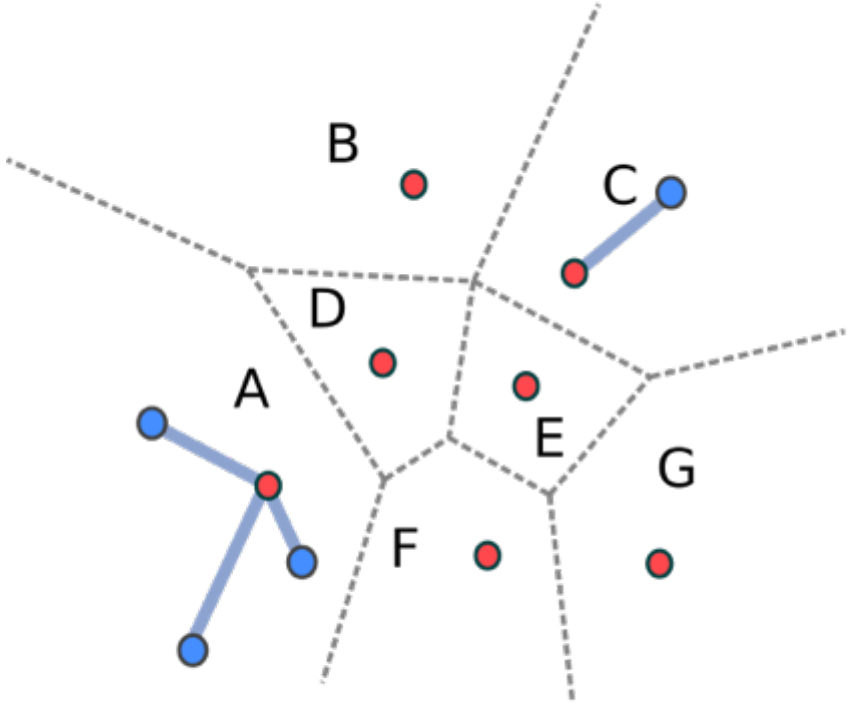


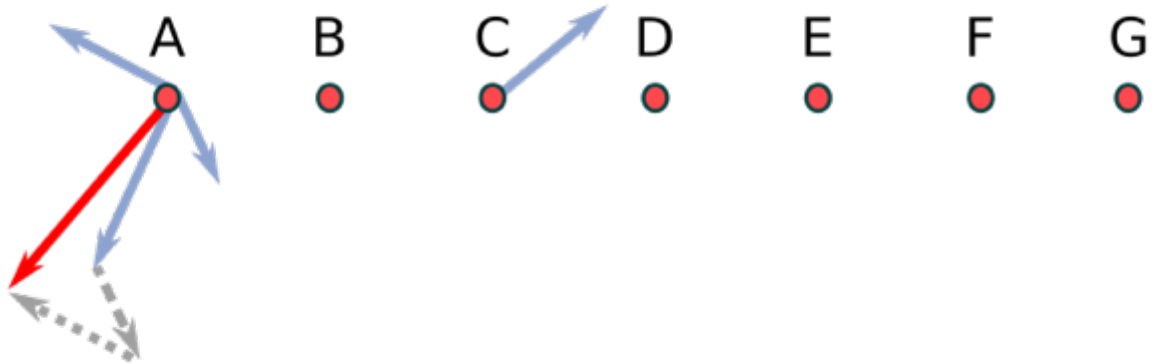
NAVER LABS EUROPE

Extension to higher order statistics

Mean: VLAD (Vector of Locally Aggregated Descriptors)

- Aggregate all descriptors assigned to the same visual word





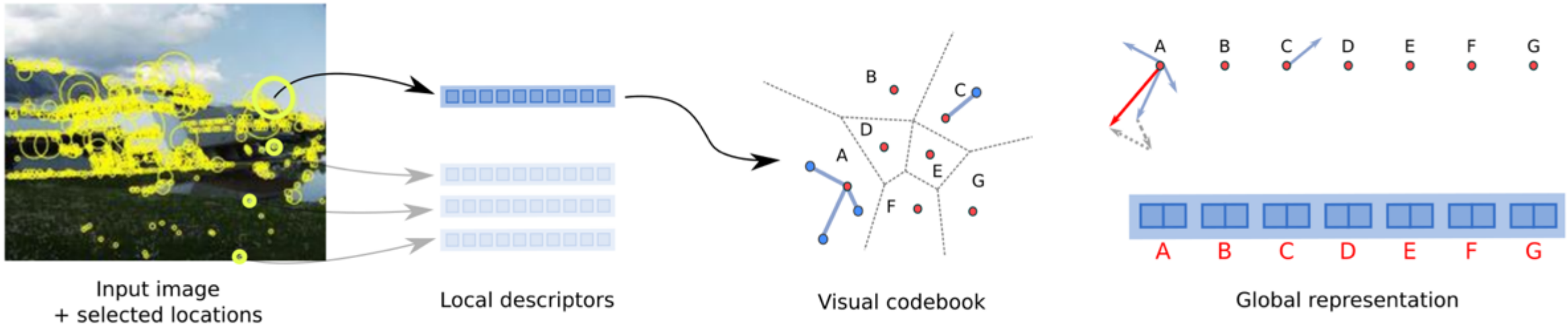
[Jégou et al. CVPR 2010]

NAVER LABS EUROPE

Extension to higher order statistics

Mean: VLAD (Vector of Locally Aggregated Descriptors)

- Aggregate all descriptors assigned to the same visual word
- Concatenate vectors for individual words



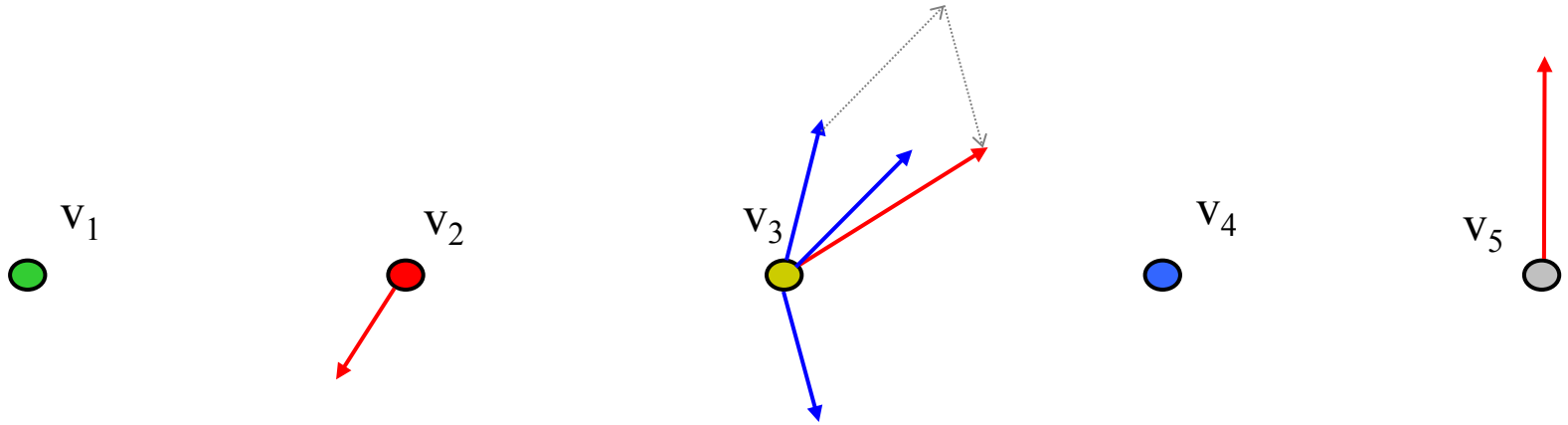
[Jégou et al. CVPR 2010]

NAVER LABS EUROPE

VLAD: Vector of Locally Aggregated Descriptors

En pratique:

- Espace des descripteurs à D-dimensions (SIFT: D=128)
- *k* centres : $c_1,\ldots,c_k$



- Sortie: $v_1 \ldots v_k$ = descripteurs de taille k*D
- L2-normalisation
- Typiquement *k* = 16 or 64 : descripteur de 2048 or 8192 D
- Mesure de similarité = distance L2

VLAD: Vector of Locally Aggregated Descriptors

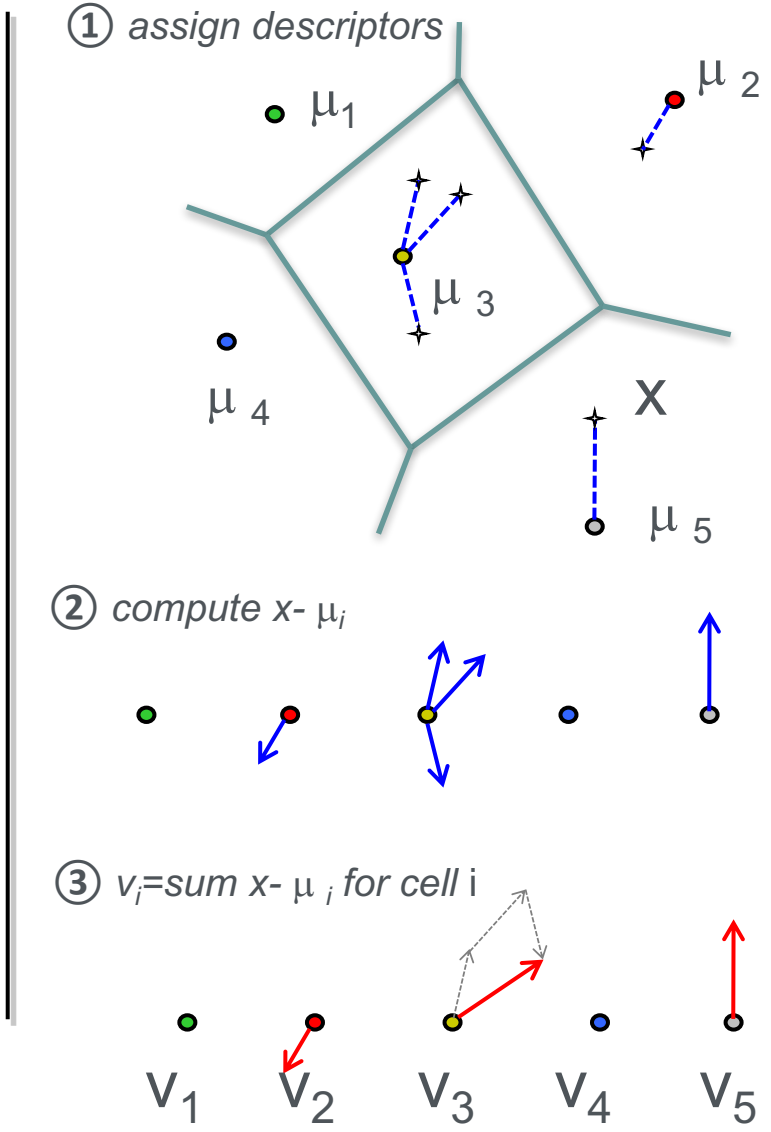
Offline:
apprendre le vocabulaire visuel $\{\mu_1, \ldots, \mu_N\}$,
par exemple avec *K-means*

Online:
Entrée: un ensemble de descripteurs locaux $\mathrm{X} = \{x_1, \ldots, x_T\}$:

- ① affectation des descripteurs au mot le plus proche:

$$NN(x_t) = \arg\min_{\mu_i} \|x_t - \mu_i\|$$

- ②③ calcul de $v_i = \sum_{x_t : NN(x_t) = \mu_i} (x_t - \mu_i)$
- Concaténation des v_i's + ℓ_2-normalisation

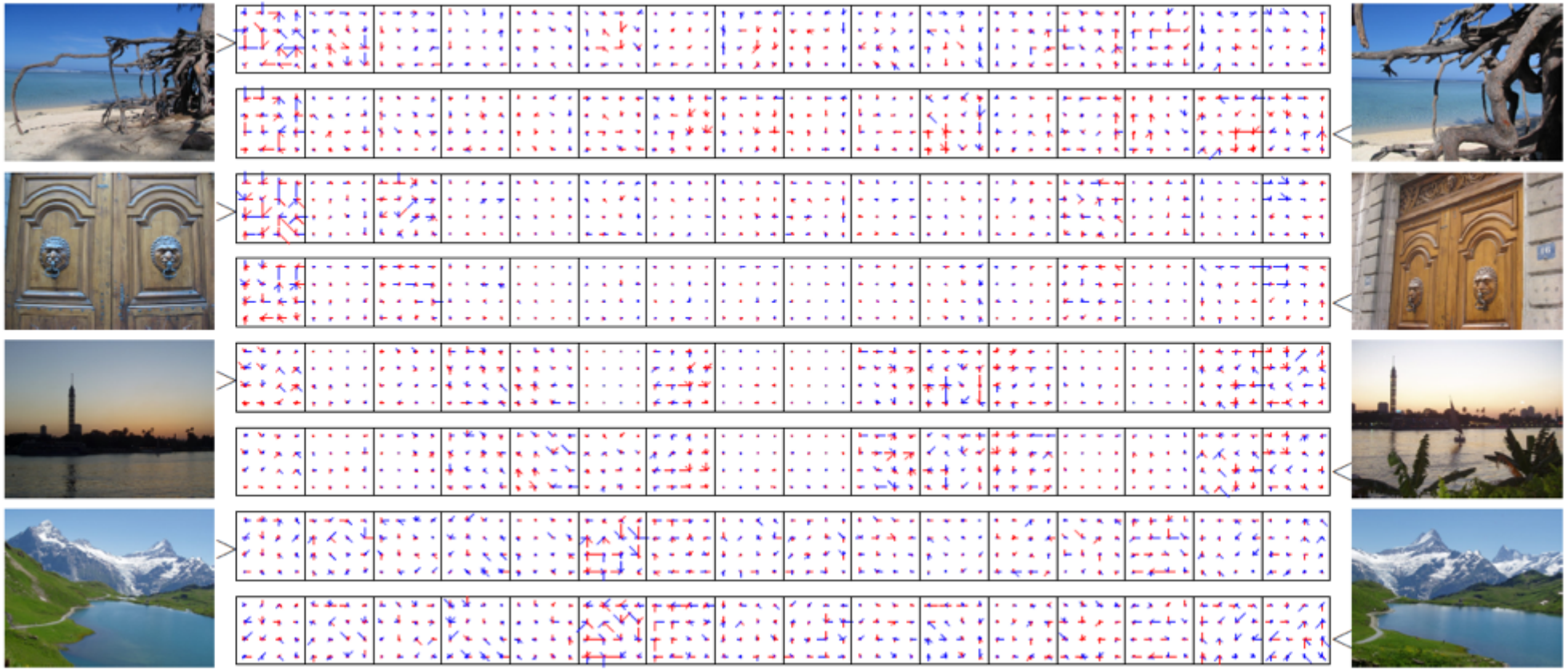


Jégou, Douze, Schmid and Pérez, "Aggregating local descriptors into a compact image representation", CVPR'10.

See also: Zhou, Yu, Zhang and Huang, "Image classification using super-vector coding of local image descriptors", ECCV'10.

VLAD: Vector of Locally Aggregated Descriptors

- Illustration du résultat de l'étape $v_i = \sum_{x_t:NN(x_t)=\mu_i}(x_t - \mu_i)$ dans le cas de descripteurs SIFT

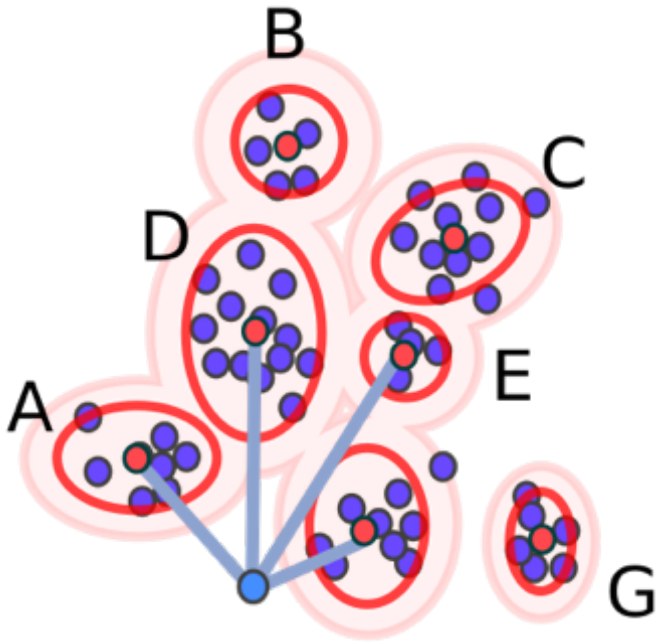


Jégou, Douze, Schmid and Pérez, "Aggregating local descriptors into a compact image representation", CVPR'10.

Extension to higher order statistics

Mean + Variance: Fisher-Vector

- Probabilistic codebook as a mixture of Gaussians
- Descriptors soft-assigned to words
- Compute the gradient of the log-likelihood w.r.t. the parameters of the model



[Perronnin and Dance. CVPR07]
[Perronnin et al. CVPR10]

NAVER LABS EUROPE

Fisher Vector

fonction de score

Principe du vecteur de Fisher

- Etant donné une fonction de vraisemblance (*likelihood function*) u_λ de paramètres λ, la fonction de score d'un échantillon x_t est donnée par :

$$g_\lambda(x_t) = \nabla_\lambda \log u_\lambda(x_t)$$

$\rightarrow$ **dont la dimension dépend du nombre de paramètres**

- Intuition: indique la direction dans laquelle les paramètres λ du modèle devraient être modifiés pour mieux représenter les données

[Slide credit: F. Perronnin, International Computer Vision Summer School. 2014]

Fisher vector

Fisher information matrix

- **Fisher information matrix** (FIM) or negative Hessian:

$$F_\lambda = E_{x \sim u_\lambda}[g_\lambda(x) g_\lambda(x)^T]$$

- Measure similarity between gradient vectors using the **Fisher Kernel (FK)**:

[Jaakkola and Haussler, "Exploiting generative models in discriminative classifiers", NIPS'98]

$$K(x, z) = g_\lambda(x)^T {F_\lambda}^{-1} g_\lambda(z)$$

$\rightarrow$ can be interpreted as a score whitening

- As the FIM is PSD, we can write: ${F_\lambda}^{-1} = {L_\lambda}^T L_\lambda$

and the FK can be rewritten as a dot product between **Fisher Vectors** (FV):

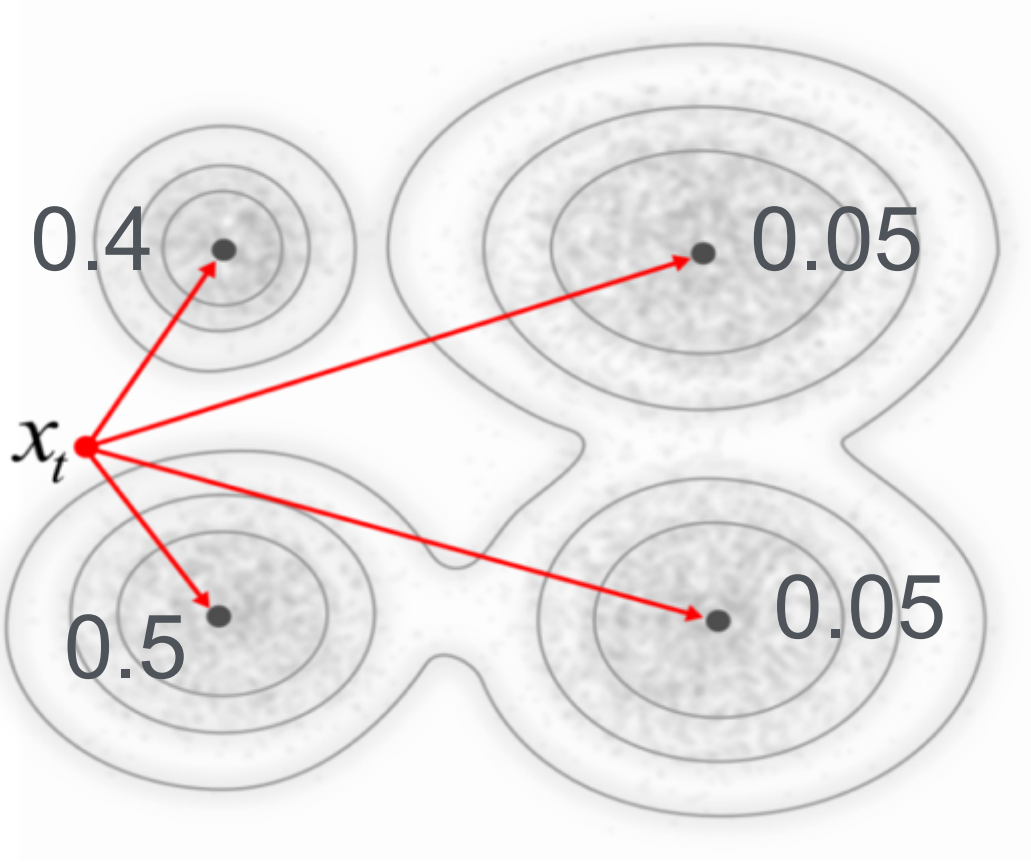
$$\varphi_\lambda^{fv}(x_t) = L_\lambda \, g(x_t)$$

[Slide credit: F. Perronnin, International Computer Vision Summer School. 2014]

Fisher vector

Application aux images

- Dans le cas des images, typiquement u_λ est une mixture de gaussiennes ou *Gaussian Mixture Model* (GMM):
  - $u_\lambda(x) = \sum_{i=1}^{N} w_i u_i(x)$



[Slide credit: F. Perronnin, International Computer Vision Summer School. 2014]

Fisher vector

Application aux images

- Dans le cas des images, typiquement u_λ est une mixture de gaussiennes ou *Gaussian Mixture Model* (GMM):
  - $u_\lambda(x) = \sum_{i=1}^{N} w_i u_i(x)$
  - $u_i(x) = \frac{1}{(2\pi)^{D/2}|\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)'{\Sigma_i}^{-1}(x-\mu_i)\right\}$

  → ils constituent les **mots visuels**

  et l'ensemble des paramètres est $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots N\}$

- On suppose généralement que la matrice est diagonale pour des raisons de complexité de calcul

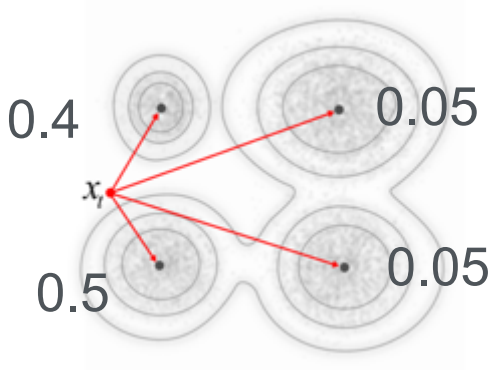
$$\Sigma_i = diag({\sigma_i}^2)$$

→ FV = concaténation des gradients selon les paramètres w_i, μ_i, et σ_i

[Slide credit: F. Perronnin, International Computer Vision Summer School. 2014]

Fisher vector

Résumé



Offline: apprendre un modèle GMM de paramètres $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots N\}$

Online: étant donné un ensemble de descripteurs $\mathrm{X} = \{x_1, \dots, x_T\}$ pour une image

- Pour chaque x_t:
  - Assignement souple à une gaussienne: $\gamma_t(i) = \frac{w_i u_i(x_t)}{\sum_{k=1}^{N} w_k u_k(x_t)}$
  - Accumulation des valeurs

  $\varphi_\mu \; += \left[\dots, \frac{\gamma_t(i)}{\sigma_i \sqrt{w_i}}(x_t - \mu_i), \dots\right]$

  et $\varphi_\sigma \; += \left[\dots, \frac{\gamma_t(i)}{\sqrt{2w_i}}\left(\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1\right), \dots\right]$
- *power-normalization* + normalisation L2

Voir aussi les implémentations:

- INRIA: http://lear.inrialpes.fr/src/inria_fisher/
- Oxford: http://www.robots.ox.ac.uk/~vgg/software/enceval_toolkit/

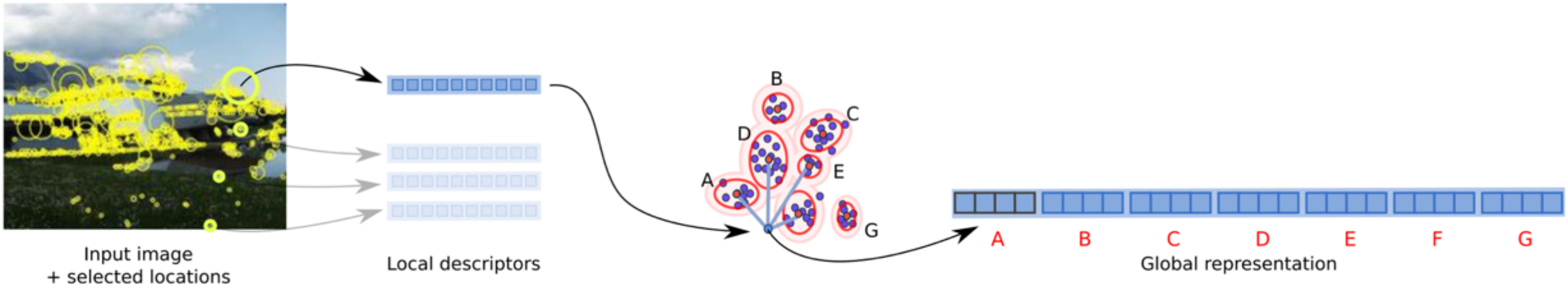
[Slide credit: F. Perronnin, International Computer Vision Summer School. 2014]

Extension to higher order statistics

Mean + Variance: Fisher-Vector

- Aggregate the contribution of all local descriptors
- Concatenate statistics for all the visual words

[Perronnin and Dance. CVPR07]
[Perronnin et al. CVPR10]

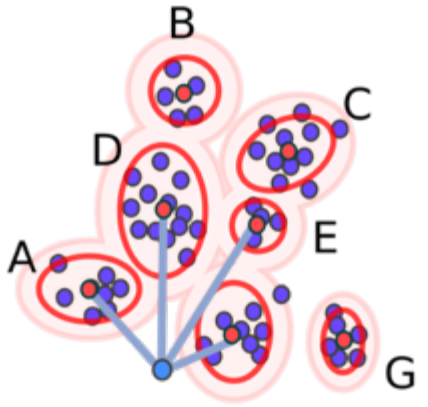


NAVER LABS EUROPE

Fisher vector

En pratique:

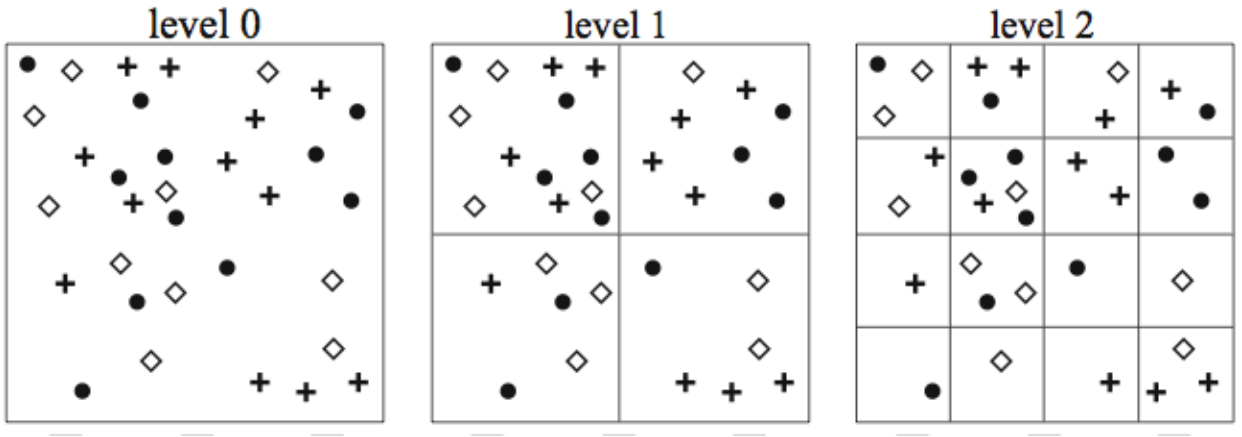
- Espace des descripteurs à d-dimensions (SIFT: d=128)
- Réduction de la dimension par ACP *(PCA)* (e.g. *D*=64)
- *K* gaussiennes de paramètres $\lambda = \{w_i, \mu_i, \Sigma_i, i = 1 \dots K\}$
- Sortie:
  - φ_μ : descripteur de taille *K*D*
  - φ_σ *:* descripteur de taille *K∗D*
- Concaténation et L2-normalisation
- Descripteur final : *2*K*D* dimensions



Comment préserver des informations géométriques ?

Pyramide spatiale

- Problème: les représentations par *« bag-of-patches »* ne gardent aucune information sur la géométrie
- Solution: calculer des agrégats de descripteurs sur des sous-images
  - géométrie rigide
  - taille des vecteurs multipliée (pour l'exemple ci-dessous par 21)



Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, S. Lazebnik, C. Schmid, J. Ponce, CVPR06

Exercice

- Calculer la représentation *Bag-of-Words*
- Calculer la représentation VLAD

(au tableau)

Evaluation on standard benchmarks

Oxford dataset

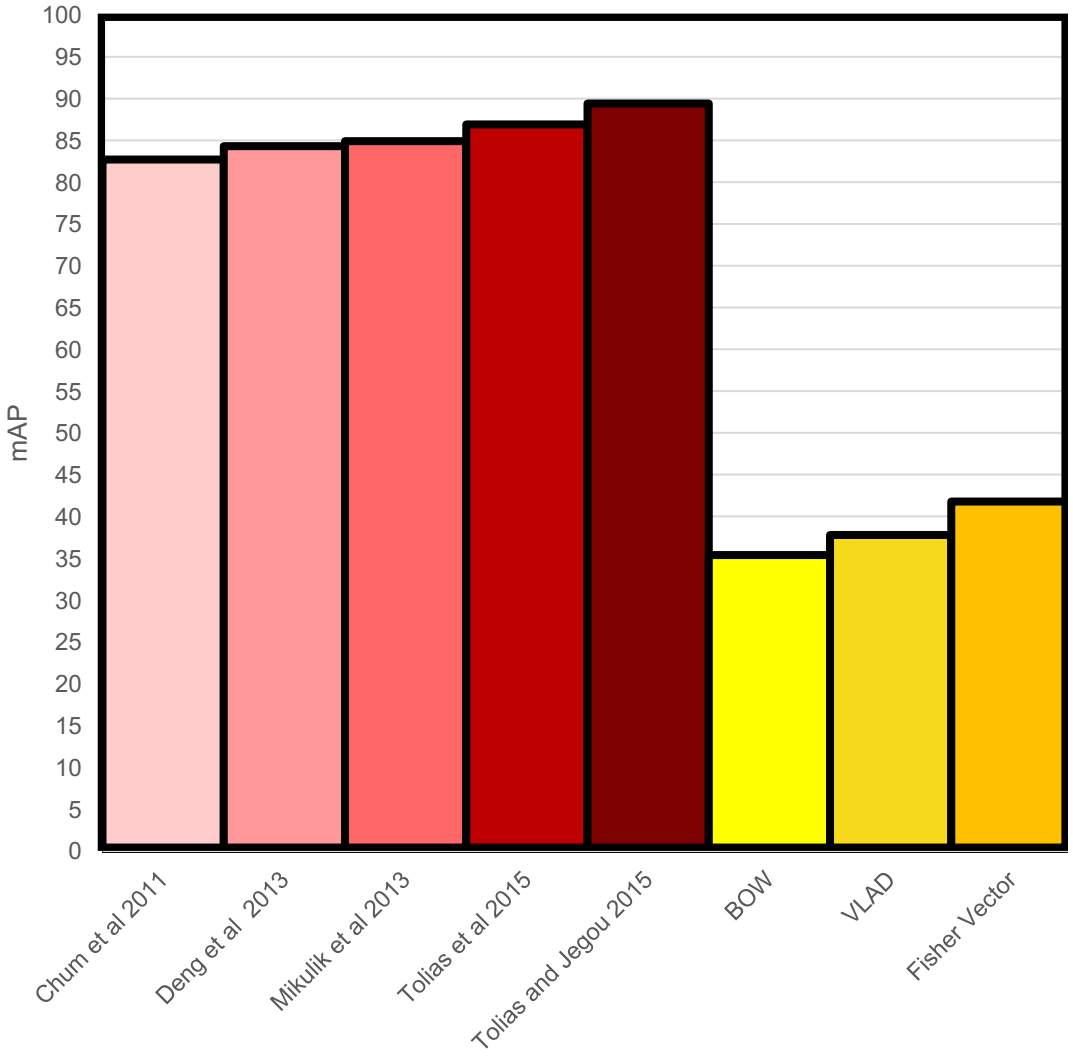
- 5,000 images
- 55 queries
- 11 landmarks

Evaluation

- mean Average Precision (mAP)

[Philbin et al. CVPR07]

Experiments



Oxford 5K

Local representations

Global representations

NAVER LABS EUROPE

Standard approaches - Summary

Matching-based methods – local representations

- highest accuracy
- but high cost of matching and geometry verification

[Philbin et al, 2007, Chum et al 2007, Chum et al 2011, Jegou et al, 2008, Chum et al, 2009, Deng et al, Mikulik et al 2013, Tolias et al 2015, Tolias and Jegou 2015,]

Aggregation methods – global representations

- faster and more efficient
- but lower accuracy than matching ones

[Jegou et al. CVPR 2010, Perronnin et al. CVPR 2010]

NAVER LABS EUROPE

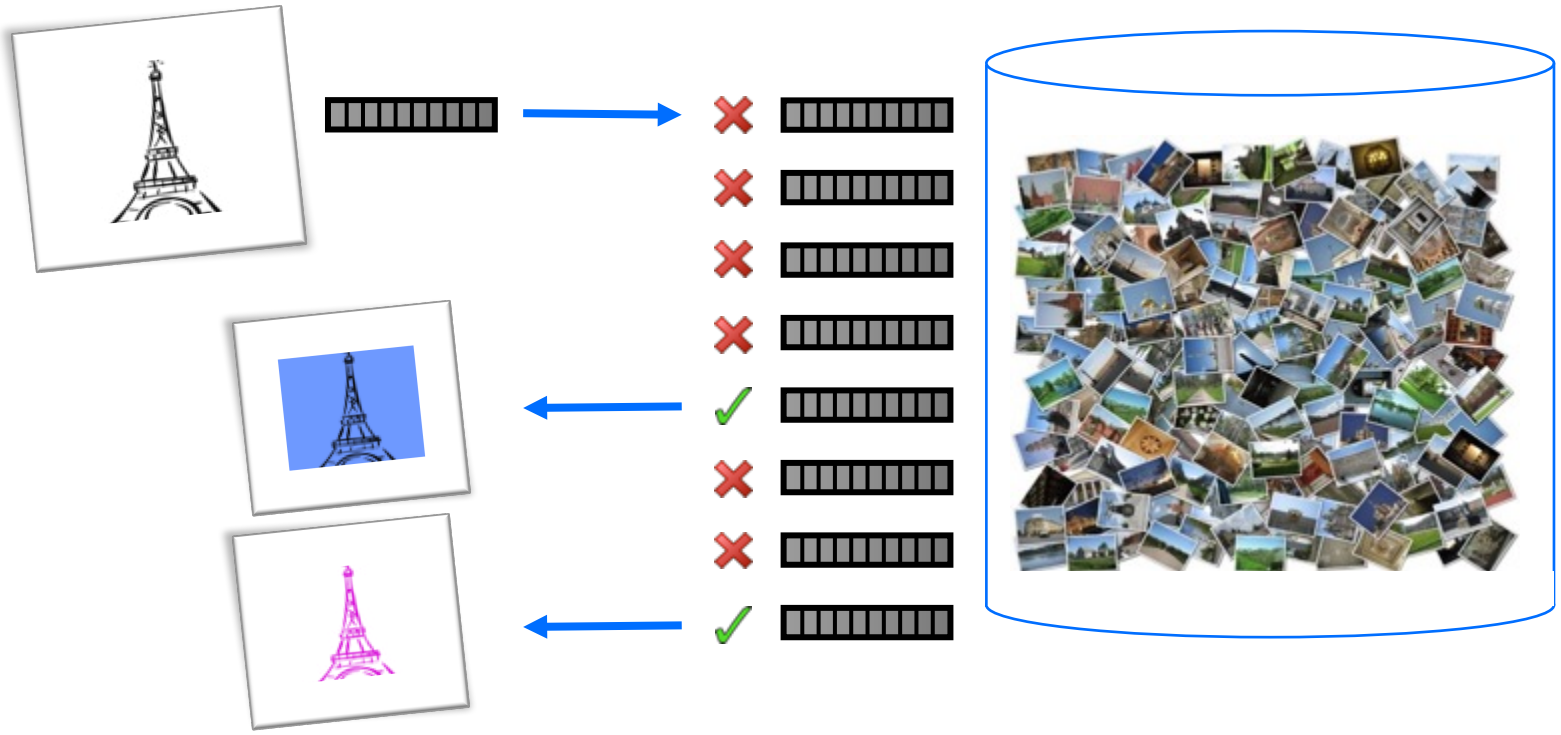
Représentation d’images par apprentissage profond

Comprendre les données visuelles à grande échelle

Cours 3: représentations d’images, 10 janvier 2019

NAVER LABS EUROPE

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



Families of representations

- Early methods
- 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

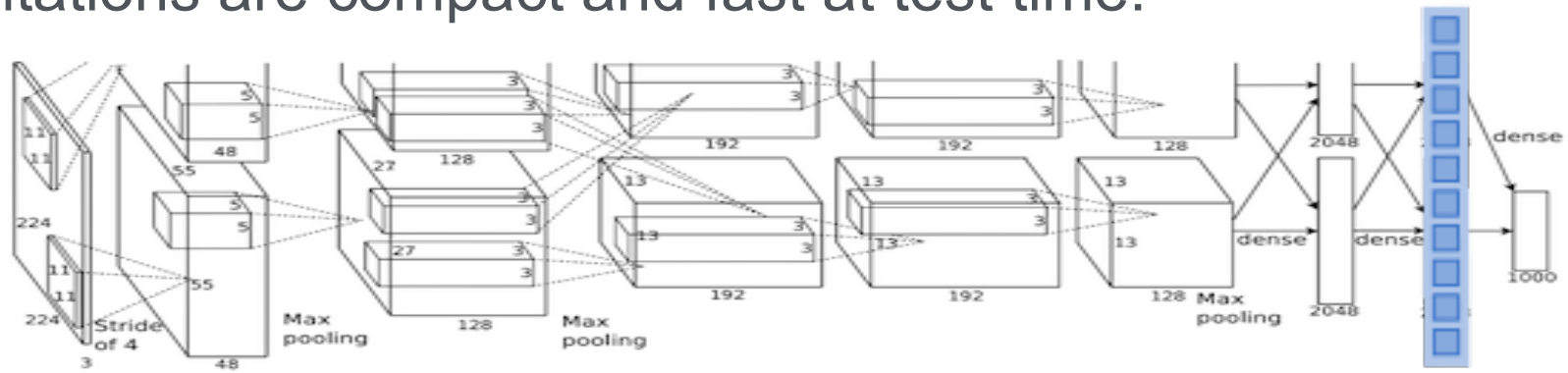
- Un bon modèle de classification capture des informations sémantiques et devrait permettre de bien représenter les images

First deep learning approaches

Pretrain network for classification and use it as a feature extractor

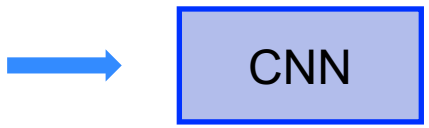
- Representations are compact and fast at test time!

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 1

- Combiner ces architectures avec des méthodes plus standards: méthodes hybrides

Solution 2

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

Plan de la suite

Méthodes hybrides

- NetVlad
  - Une fois les couches convolutionelles appliquées à l'image, cette représentation dense de l'image est agrégée en utilisant des couches qui reproduisent une agrégation par descripteur VLAD
- *Fisher Vector meets Neural Networks*
  - Les couches convolutionnelles sont « remplacées » par une représentation par vecteur de Fisher, seules les couches totalement connectées sont conservées

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

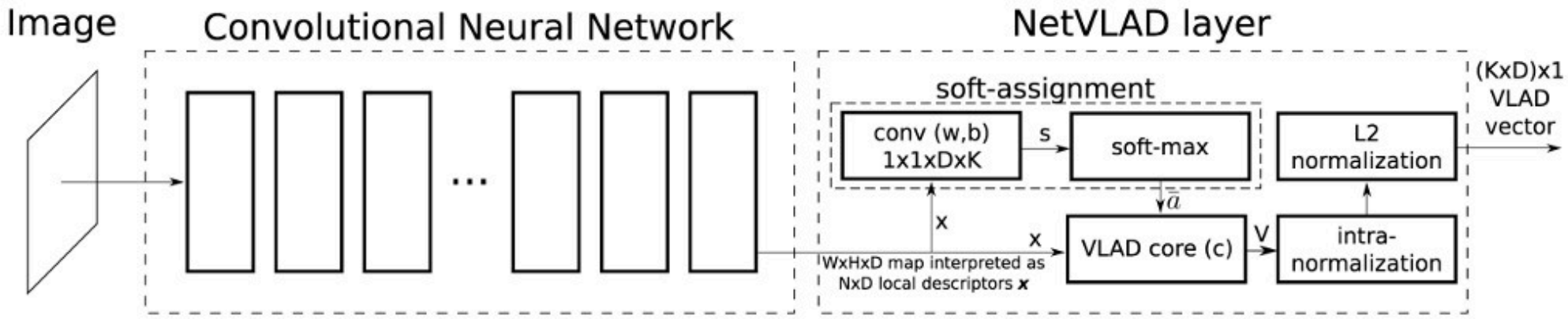
- Une base d'apprentissage appropriée, i.e. labélisée à l'échelle de l'instance et non de la catégorie d'objet. Correction des erreurs d'annotation si besoin.
- Changer l'architecture pour apprendre les détails: architecture qui accepte toute taille et rapport d'échelle d'image, même les images avec 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.

Combining CNN and VLAD: NetVLAD

Principle:

- VLAD block at the end of CNN

[Arandjelovic et al. CVPR16]



Input image

CNN

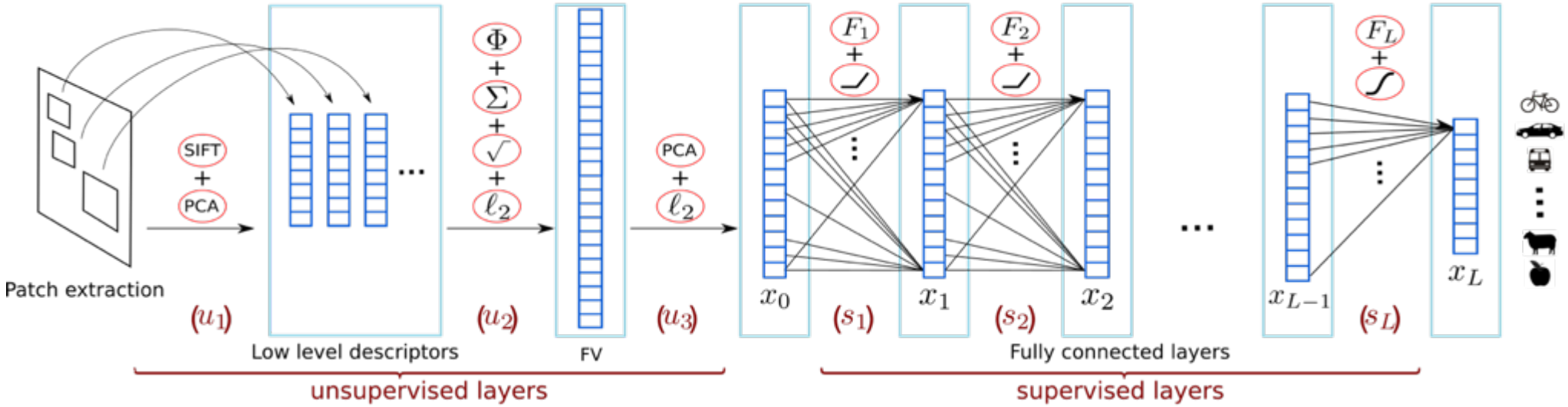
[Gong et al. ECCV14, Paulin et al ICCV15]

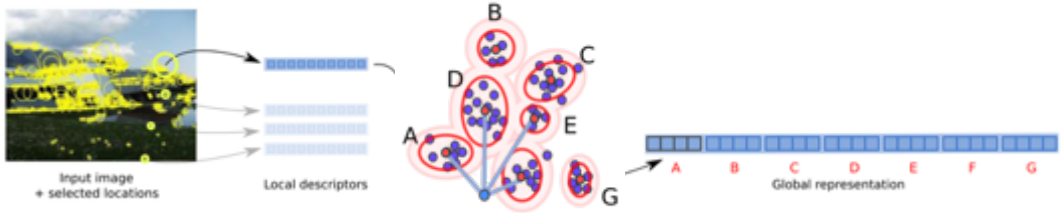
NAVER LABS EUROPE

Combining Fisher Vector and fully connected layers

Principle:

- Fisher Vector representation combined with fully-connected layers

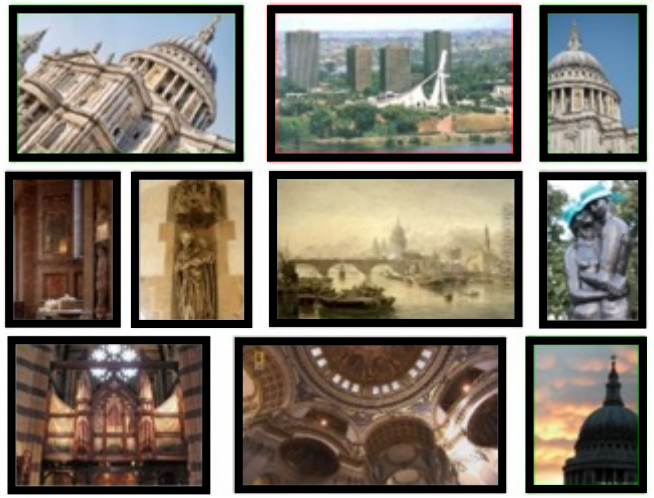




[Gordo et al. CVPR12, Perronnin & Larlus. CVPR15]

NAVER LABS EUROPE

Training for visual search



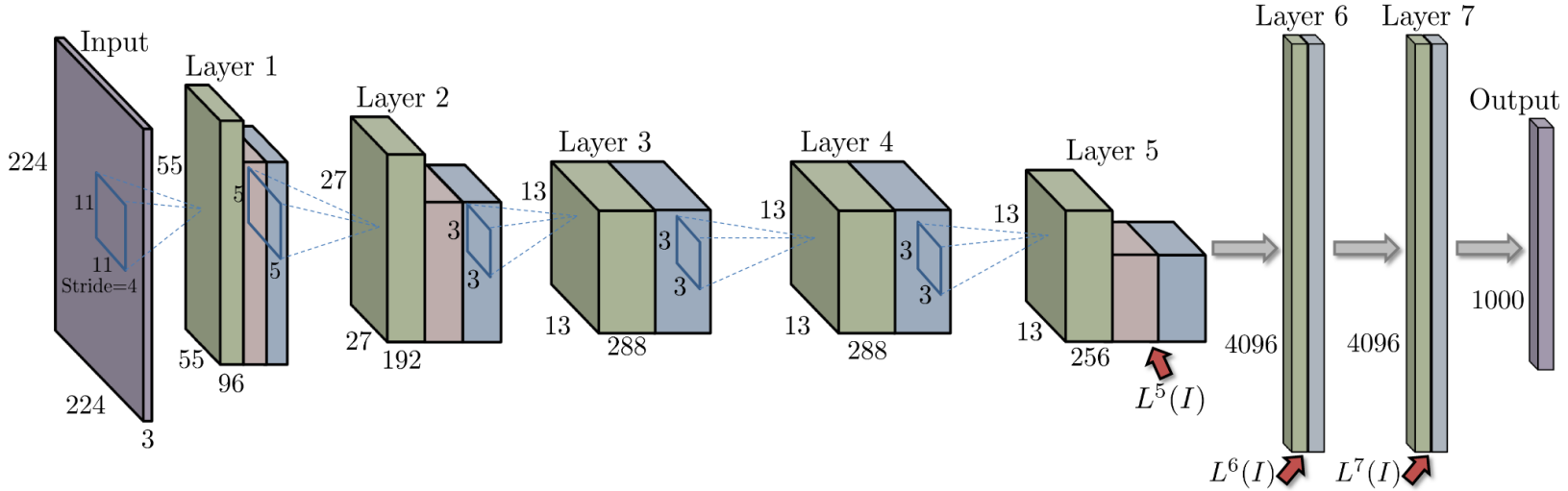
Collect and leverage an appropriate training set

- ~200K images
- ~600 different landmarks

Fine-tune the network

- Images resized to be 224x224
- Softmax classification loss

[Babenko et al, Neural codes @ ECCV14]



NAVER LABS EUROPE

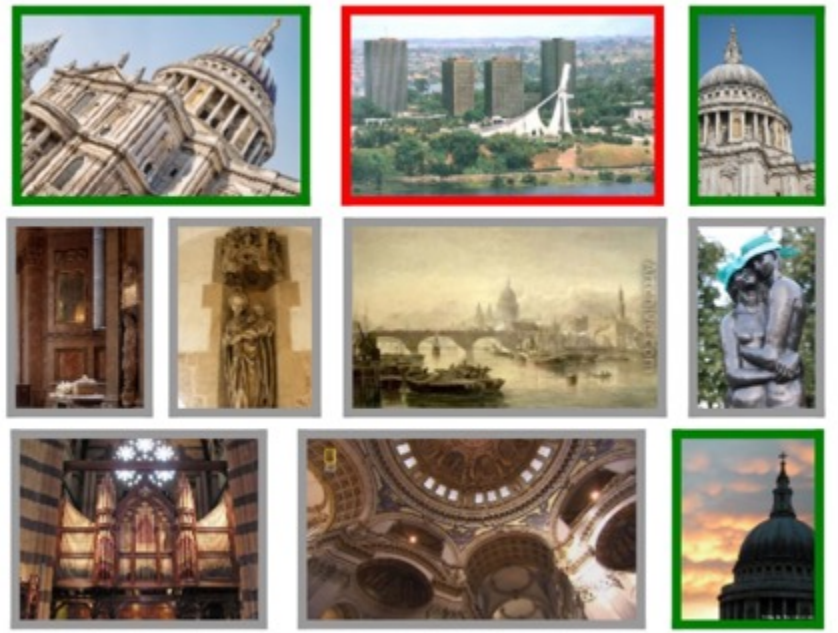
Training for visual search

Limitations

[Babenko et al, Neural codes @ ECCV14]

- Again, images are distorted
- Fine-tune network on relevant images still with a classification loss

- Public landmark dataset is very noisy



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:

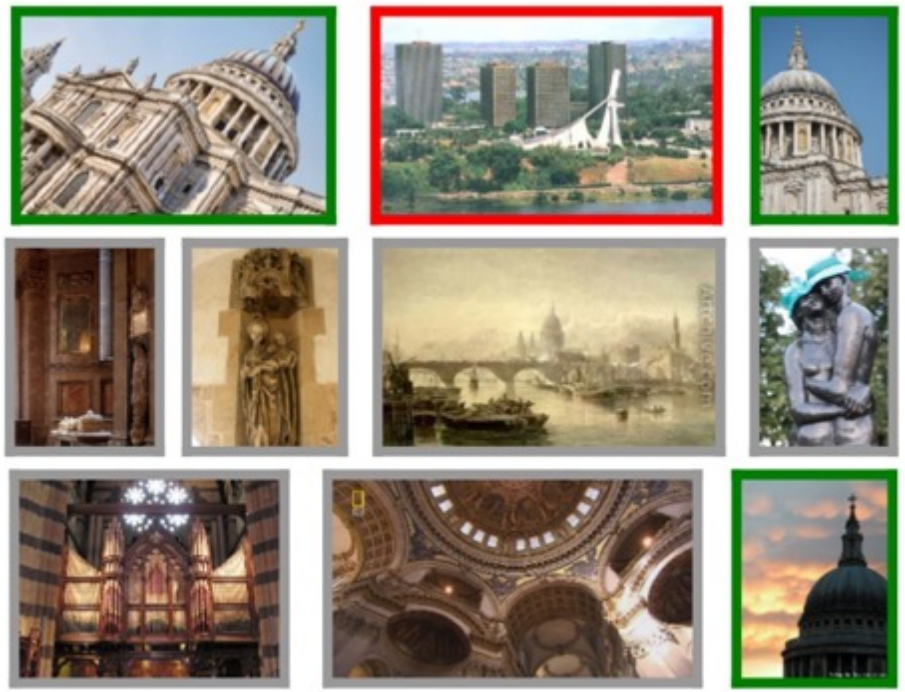
   We should train explicitly for retrieval, not for classification

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]



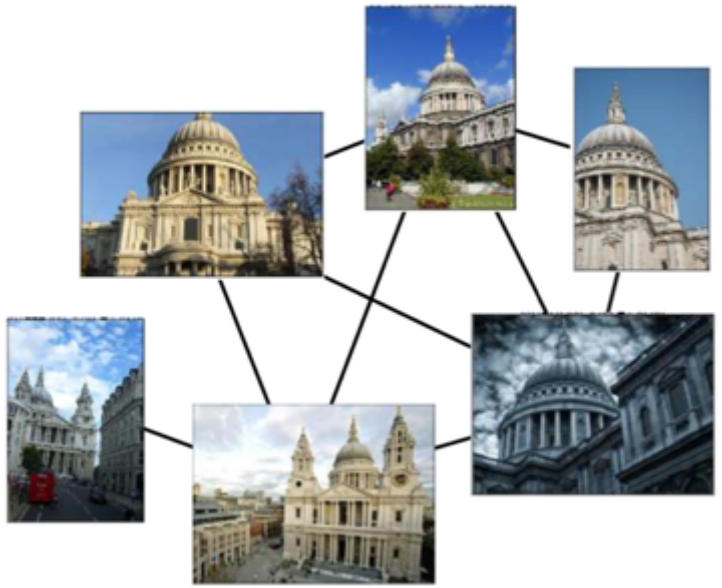
NAVER LABS EUROPE

1. Training Data

Nettoyage automatique de la base d'images

[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 points 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 grande composante connectée par classe est conservée



Résultat

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

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

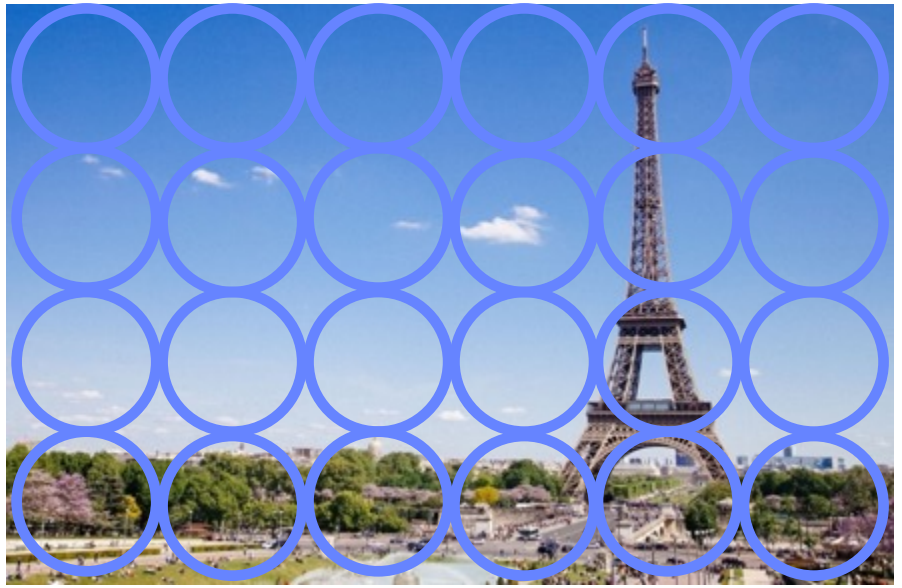
R-MAC descriptor

[Tolias et al, ICLR16]

Input image

CNN

Local features



NAVER LABS EUROPE

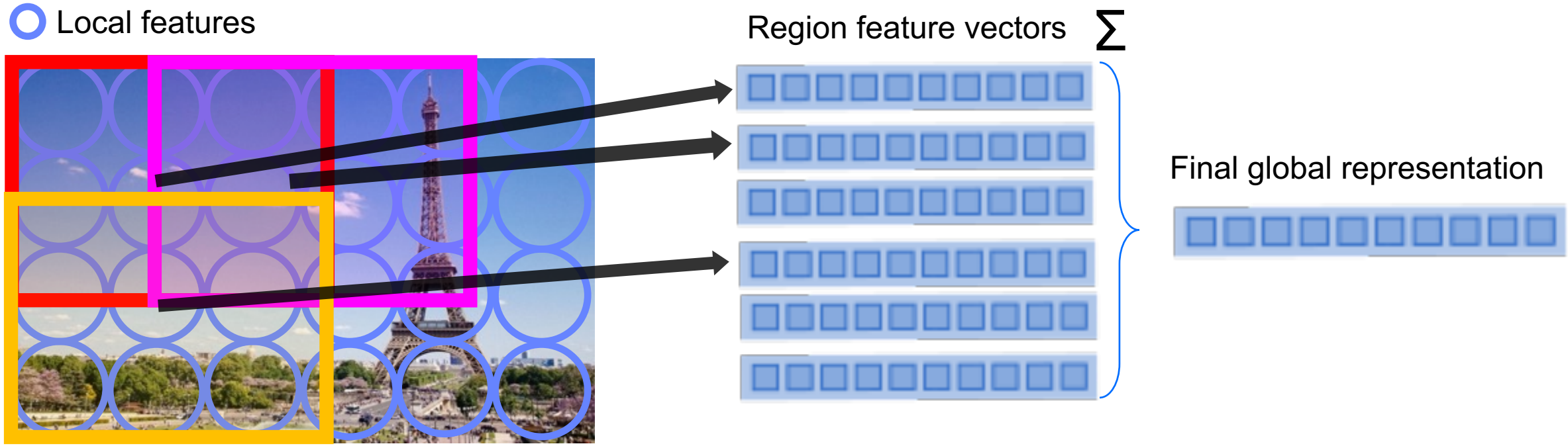
2. Architecture

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

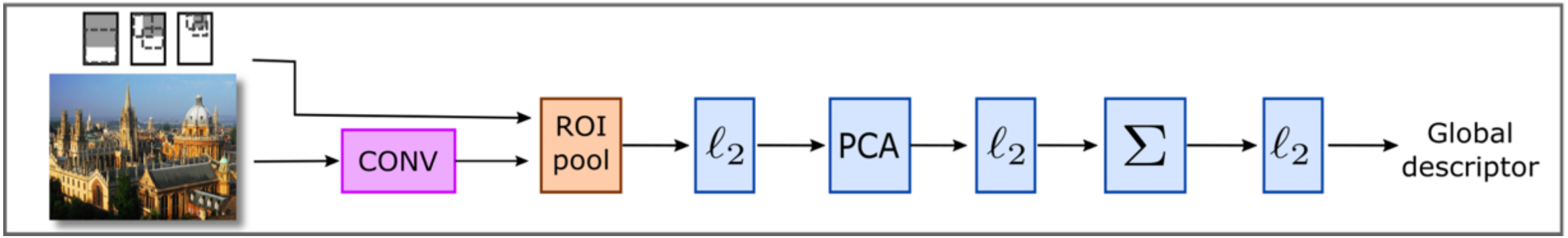
R-MAC descriptor

- 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:

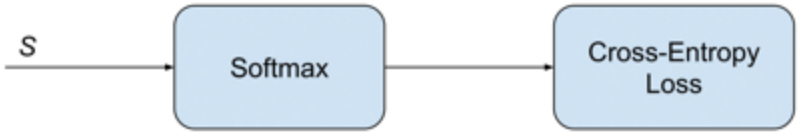


2. Every step is differentiable → **the model can be trained end-to-end!**

3. Training for retrieval

Learning to rank

- Historiquement, les méthodes ont entrainé leurs réseaux de neurones avec une fonction de coût adaptée à la classification (ex: *softmax loss*)



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

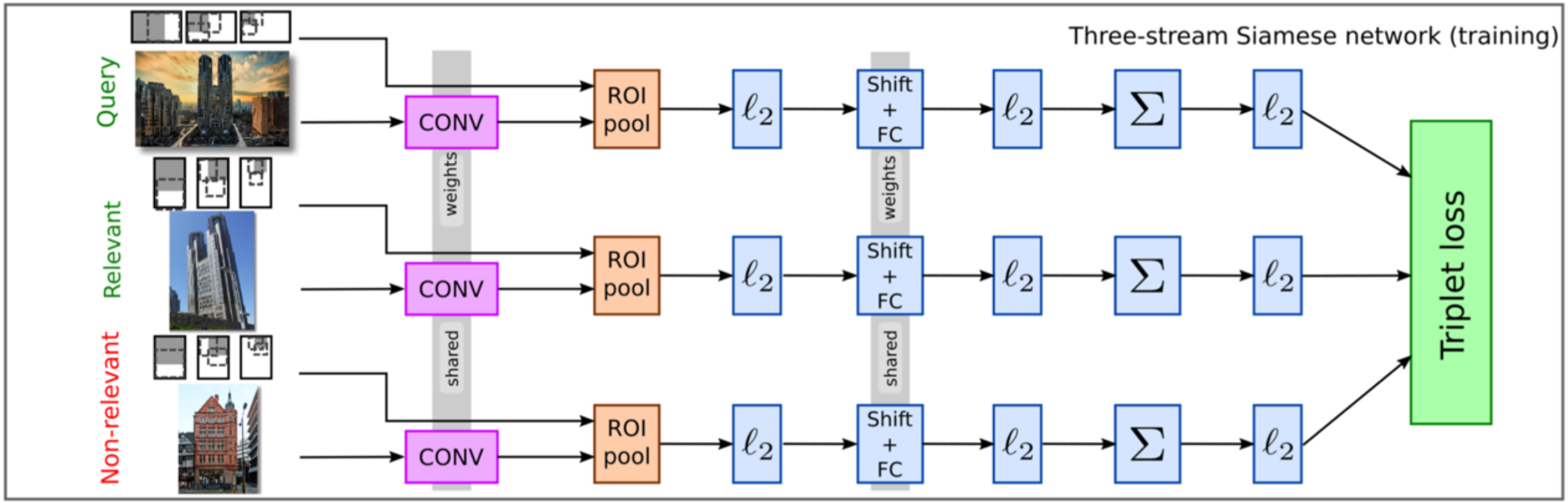
- Problème: on ne souhaite pas reconnaitre la classe, mais, étant donnée une image, ordonner les images de la base de la plus pertinente à la moins pertinente
- Solution: considérer une fonction de coût qui s'intéresse directement au rang des images. Cela demande de modifier l'architecture du réseau d'apprentissage (architecture siamoise)
- Exemple de fonctions 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

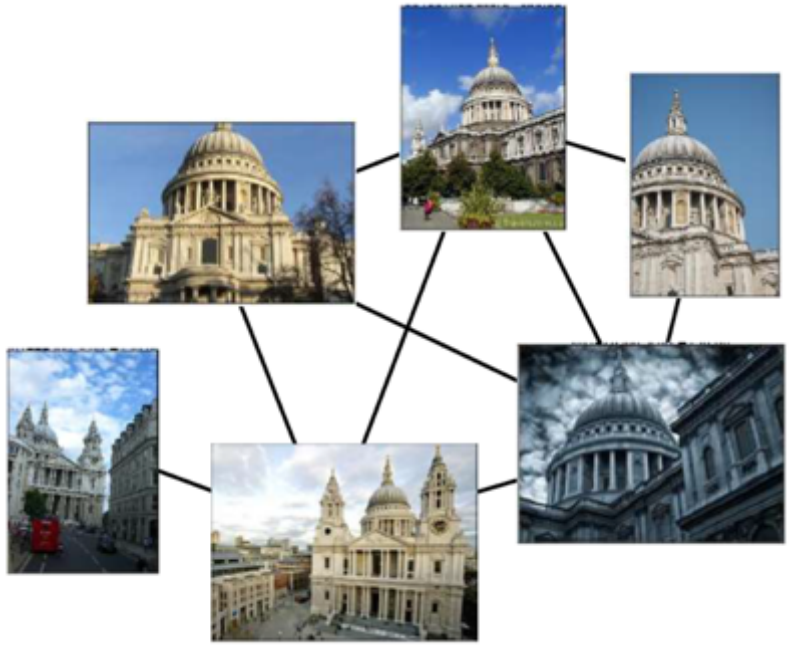
Non-relevant

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

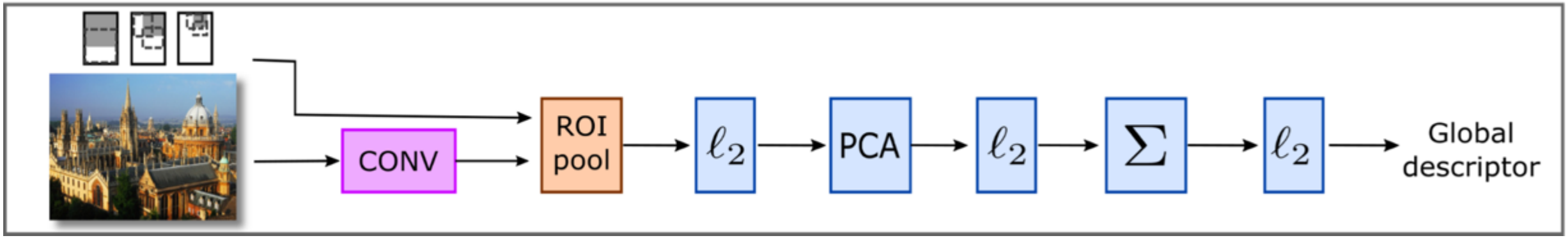
Summary

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

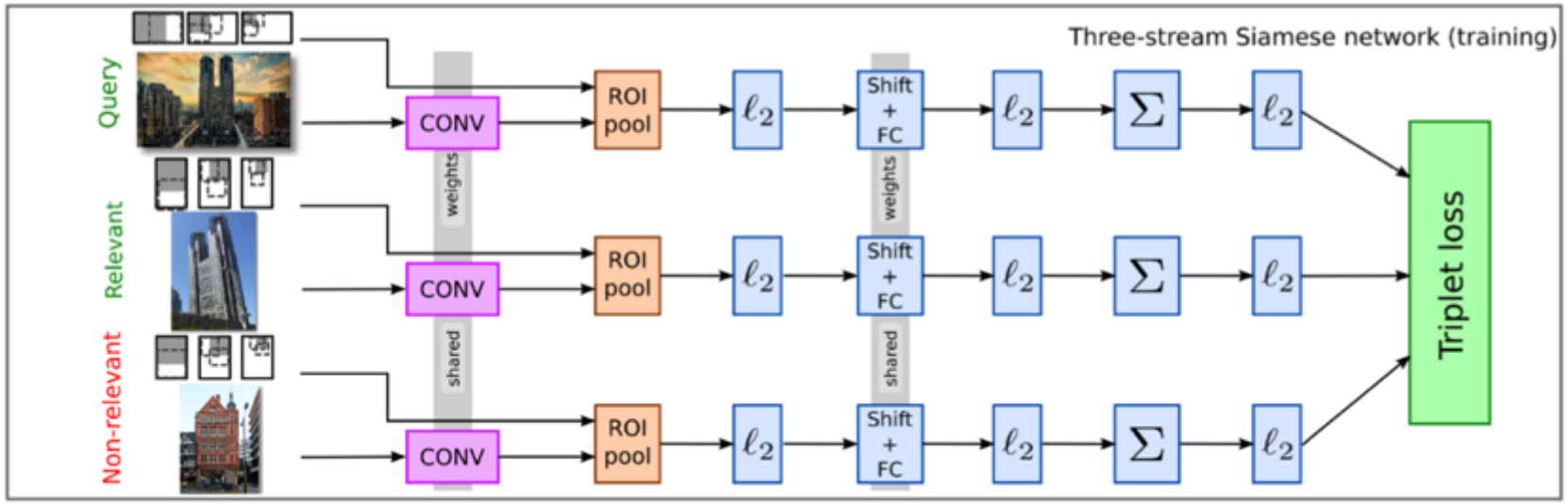
Summary

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

Other aggregation techniques

SPoC

[Babenko and Lempitsky. ICCV15]

- Sum-pooling aggregation over convolutional features
- Center prior for spatial weighting and uniform channel weighting

CroW

[Kalantidis et al. W@ECCV16]

- Non-parametric scheme for spatial- and channel-wise weighting before sum-pooling aggregation
- Boosts the effect of highly activate spatial responses and regulates burstiness

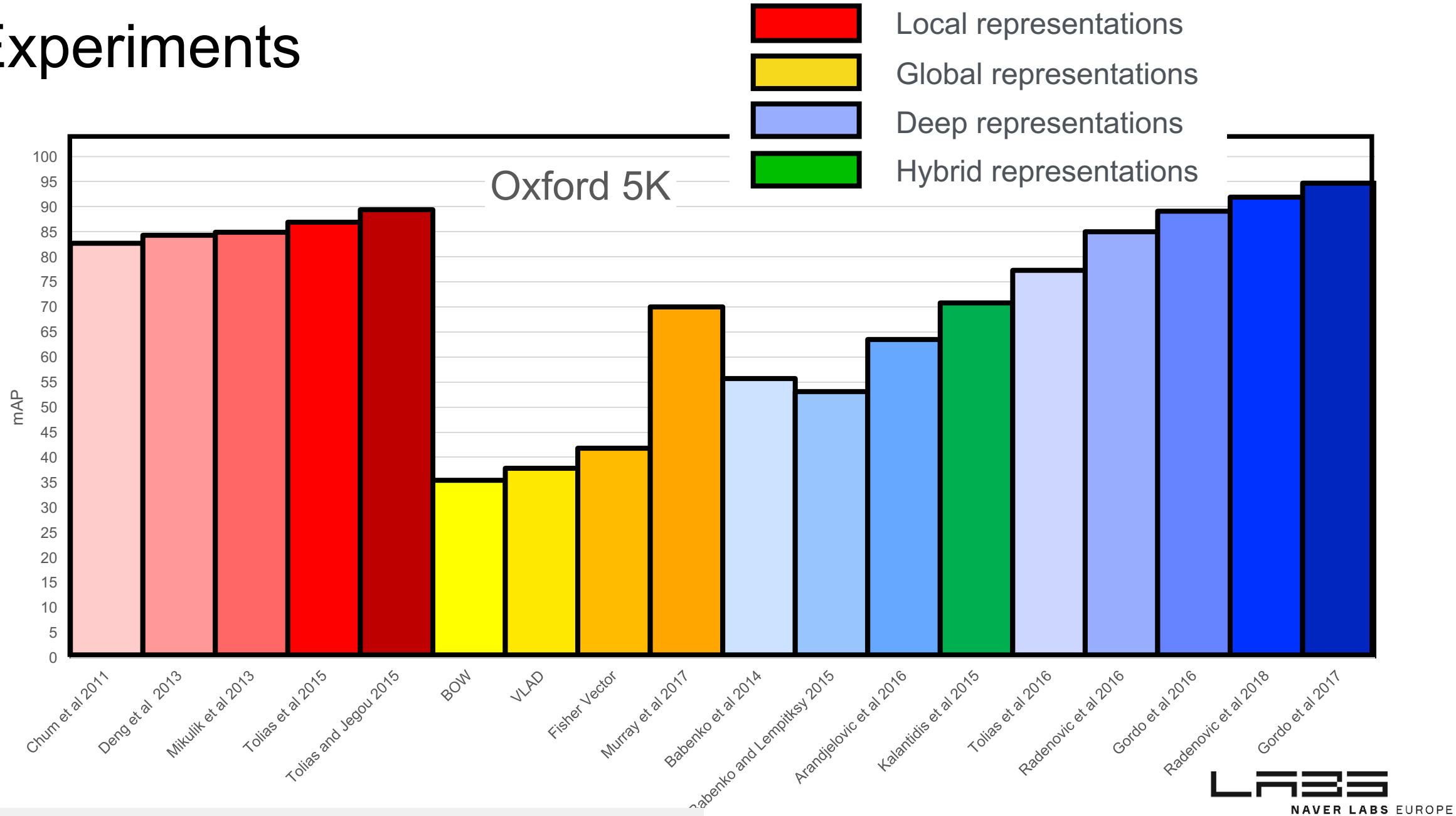
Generalized-Mean pooling

[Radenovic et al. PAMI18]

- Trainable Generalized-Mean pooling layer that generalizes max and average pooling
- Top performing approach

NAVER LABS EUROPE

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]

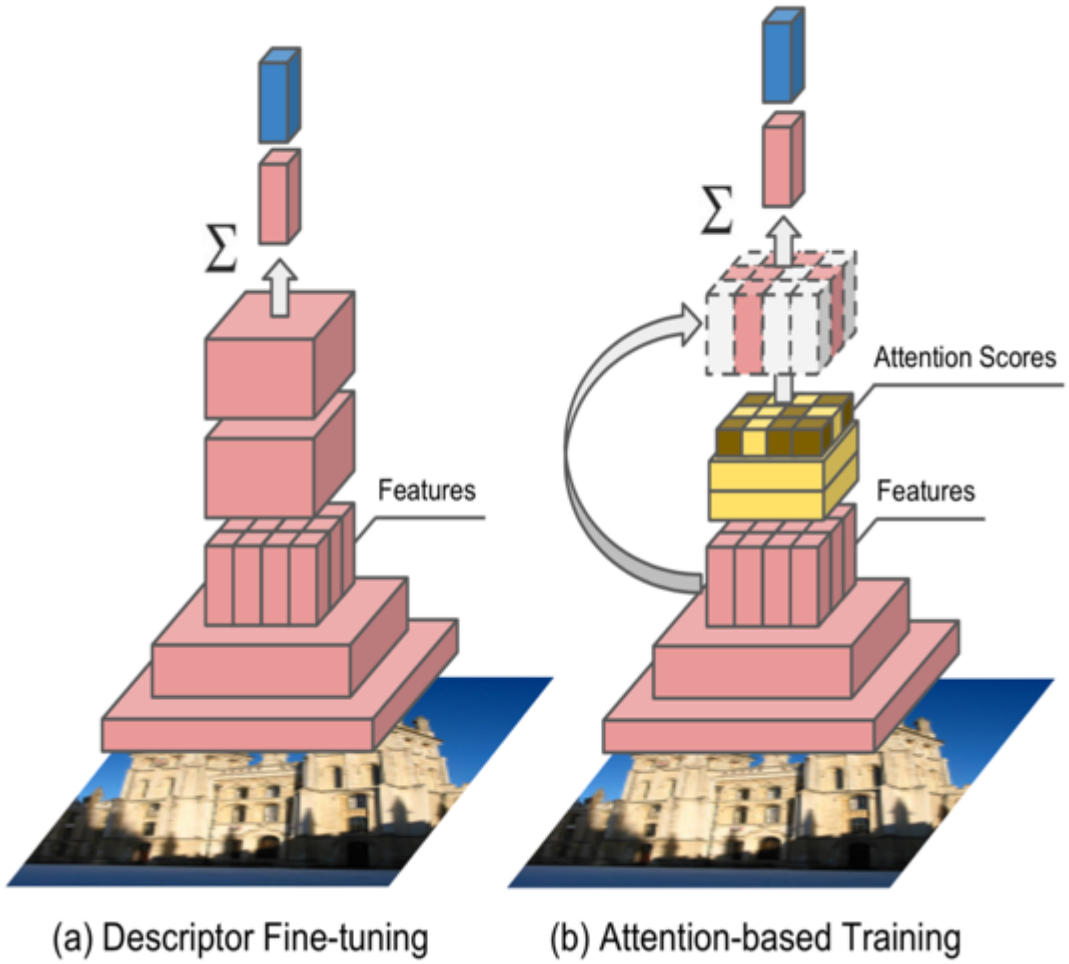
Before

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

Composant clé de la recherche d’images: Indexation

Comprendre les données visuelles à grande échelle

Cours 3: représentations d’images, 10 janvier 2019

NAVER LABS EUROPE

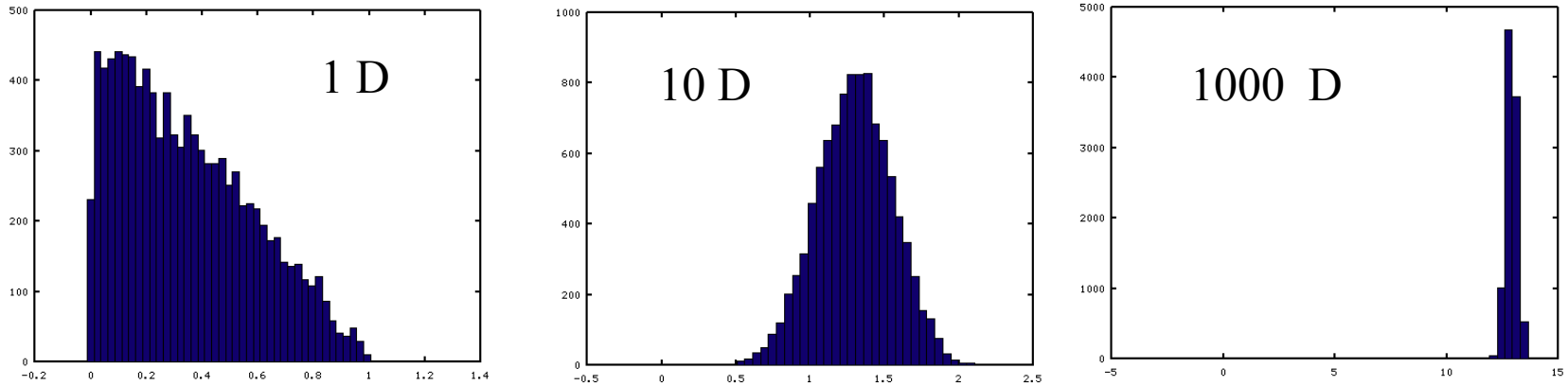
Malédiction de la dimension

- ou « fléau de la dimension » (*curse of dimensionality*)

- Quelques propriétés surprenantes
  - *vanishing variance*
  - phénomène de l'espace vide
  - proximité des frontières

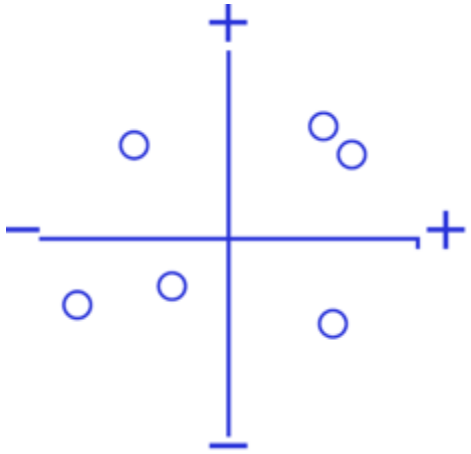
Vanishing variance

- La distance entre paires de points tend à être identique quand *d* augmente
  - le plus proche voisin et le point le plus éloigné sont à des distances qui sont presque identiques
  - exemple avec données uniformes:



- Conséquence
  - le plus proche voisin devient très instable
  - inégalité triangulaire peu efficace pour filtrer
- Les jeux de données uniformes ne peuvent pas être indexés efficacement
  - c'est un peu moins vrai pour des données naturelles (ouf !)

Phénomène de l'espace vide



- Cas d'école : partition de l'espace selon le signe des composantes
- d=100 –> 1.26 10^{30} cellules >> n = le nombre de descripteurs
- Très peu de cellules sont remplies
  - pour une partition pourtant très grossière …

- Ce phénomène est appelé « phénomène de l'espace vide »
  - difficulté pour créer une partition
    - une bonne répartition des points
    - avec une bonne compacité

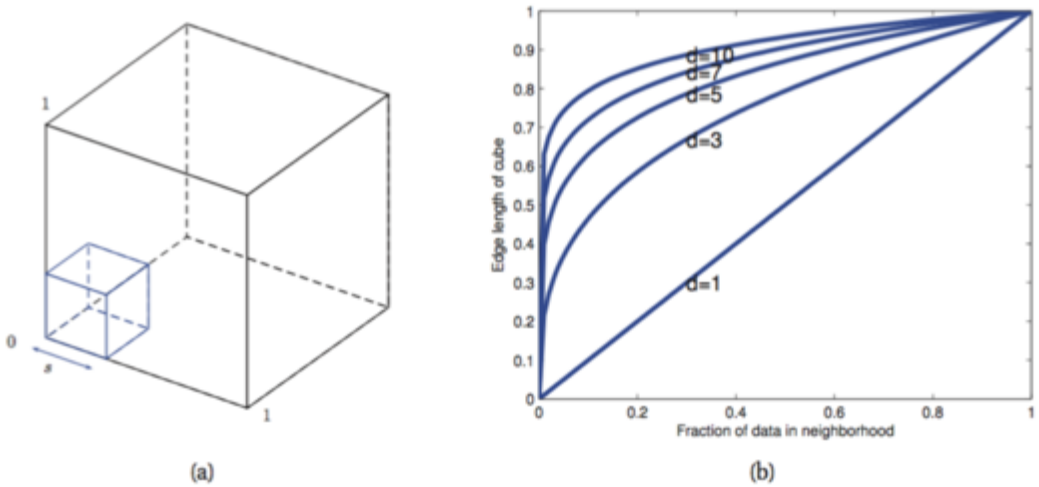


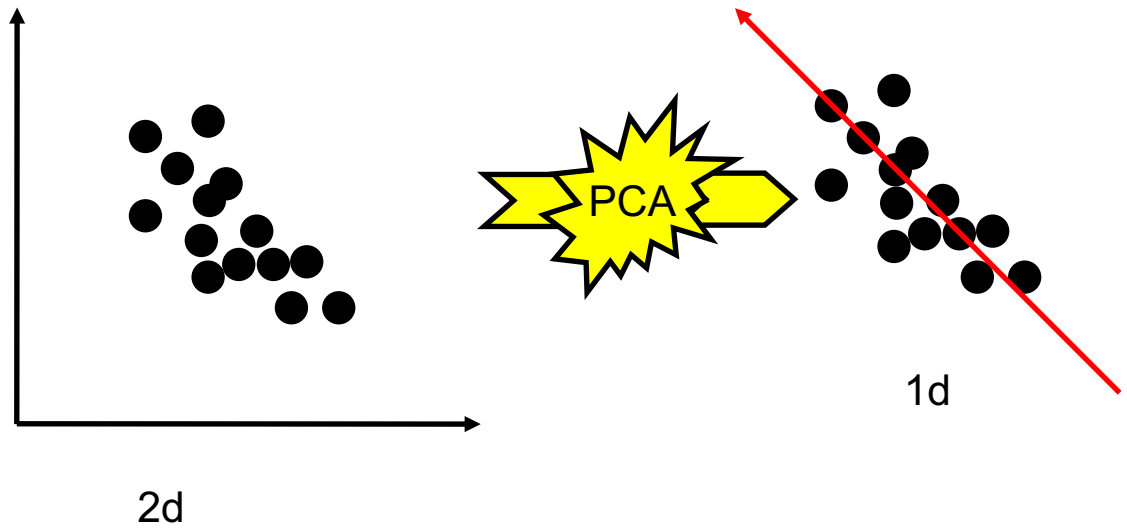
Figure 1.16 Illustration of the curse of dimensionality. (a) We embed a small cube of side s inside a larger unit cube. (b) We plot the edge length of a cube needed to cover a given volume of the unit cube as a function of the number of dimensions. Based on Figure 2.6 from (Hastie et al. 2009). Figure generated by `curseDimensionality`.

Réduction de la dimensionnalité

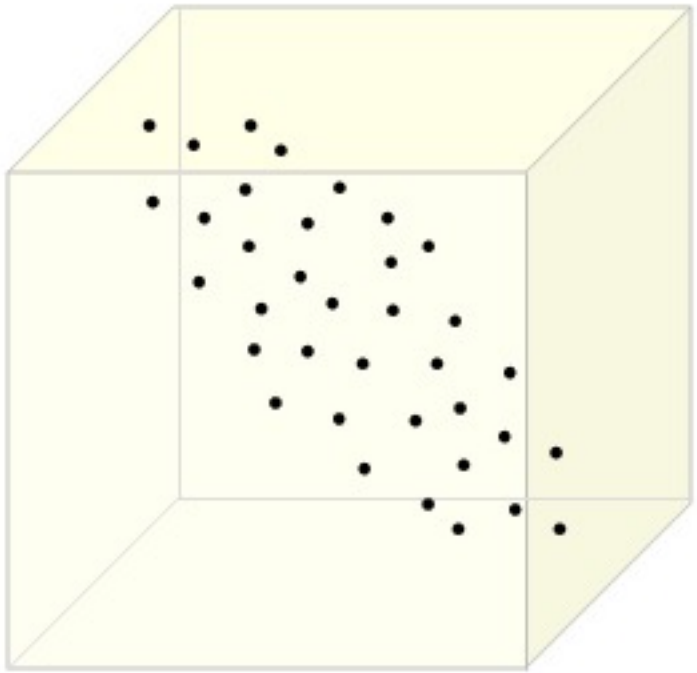
- Approche la plus courante :
  analyse en composantes principales (ACP)
  - PCA en anglais : *Principal Component Analysis*

ACP

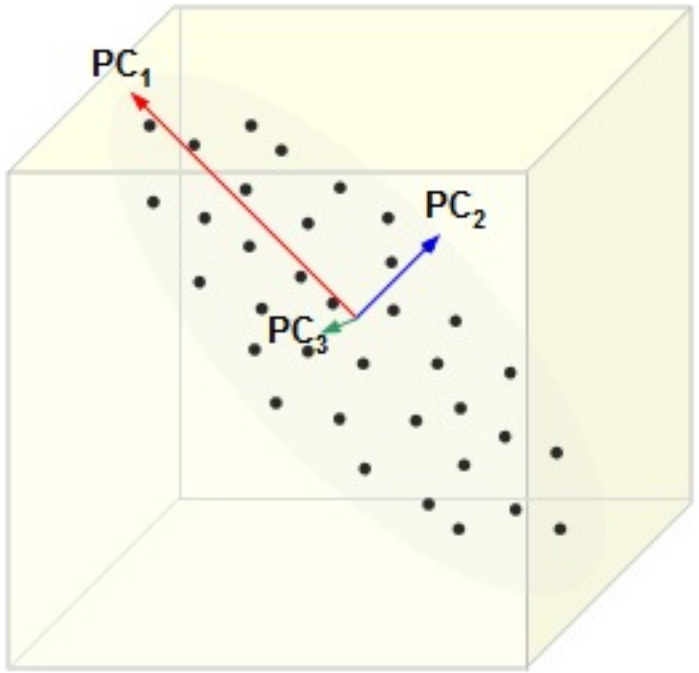
- Analyse des relations statistiques entre les différentes composantes
- Pour être capable de reproduire la plus grande partie de l'énergie d'un vecteur avec un nombre plus faible de dimensions
  - élimination des axes peu énergétiques



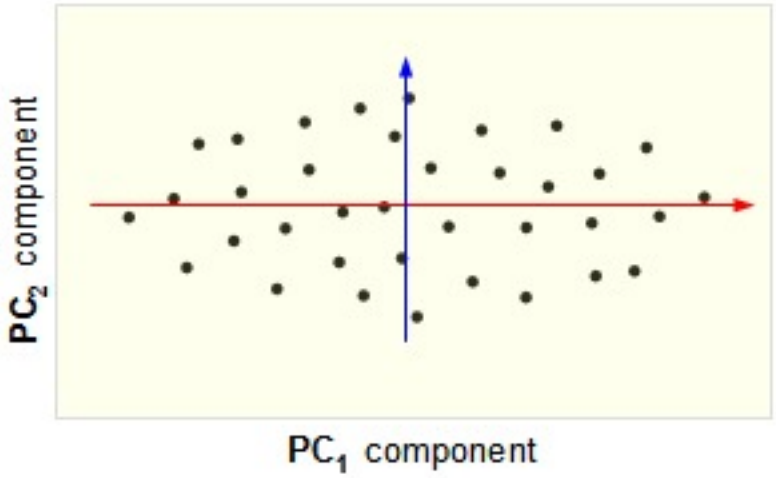
ACP



a



b



c

ACP

- Il s'agit d'un changement de base (translation + rotation)

- On suppose N descripteurs $\{x_1, x_2, x_3, .., x_N\}$
- 4 étapes (*offline*)
  - Calcul de la moyenne des descripteurs
    - $\bar{x} = \frac{1}{N}\sum_1^N x_i$
  - Calcul de la matrice de covariance sur les vecteurs centrés
    - $C = \sum_1^N y_i y_i^T$ avec $y_i = x_i - \bar{x}$
  - Calcul des valeurs propres et vecteurs propres de la matrice de covariance
    - $Ce_i = \lambda_i e_i$
  - Choix des *d'* composantes les plus énergétiques (+ grandes valeurs propres)

- Pour un vecteur, les nouvelles coordonnées sont obtenues par centrage et multiplication du vecteur par la matrice *d' x d* des vecteurs propres conservés

ACP

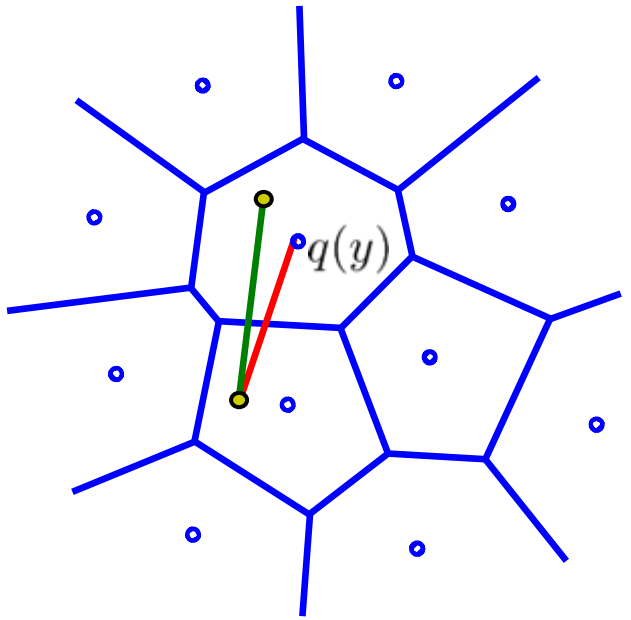
- Moins de dimensions, donc meilleur comportement des algorithmes de recherche
- Si la réduction préserve la plupart de l'énergie
  - les distances sont préservées
  - donc le + proche voisin dans l'espace transformé est aussi le plus proche voisin dans l'espace initial

- Limitations
  - utile seulement si la réduction est suffisamment forte
  - lourdeur de mise à jour de la base (choix fixe des vecteurs propres préférable)
  - peu adapté à certains types de données

Améliorer la finesse de la quantification: le quantificateur produit

$$d(x, y) \approx d(x, q(y))$$

- Problème d'approximation de distances entre requête x et vecteur y de la base
- il faut une quantification plus fine que quelques milliers de centroides
  - *k-means* coûteux
  - assignement coûteux
  - stockage des centroïdes coûteux (même en hiérarchique)
- Solution: le quantifieur produit (*product quantization* ou PQ)

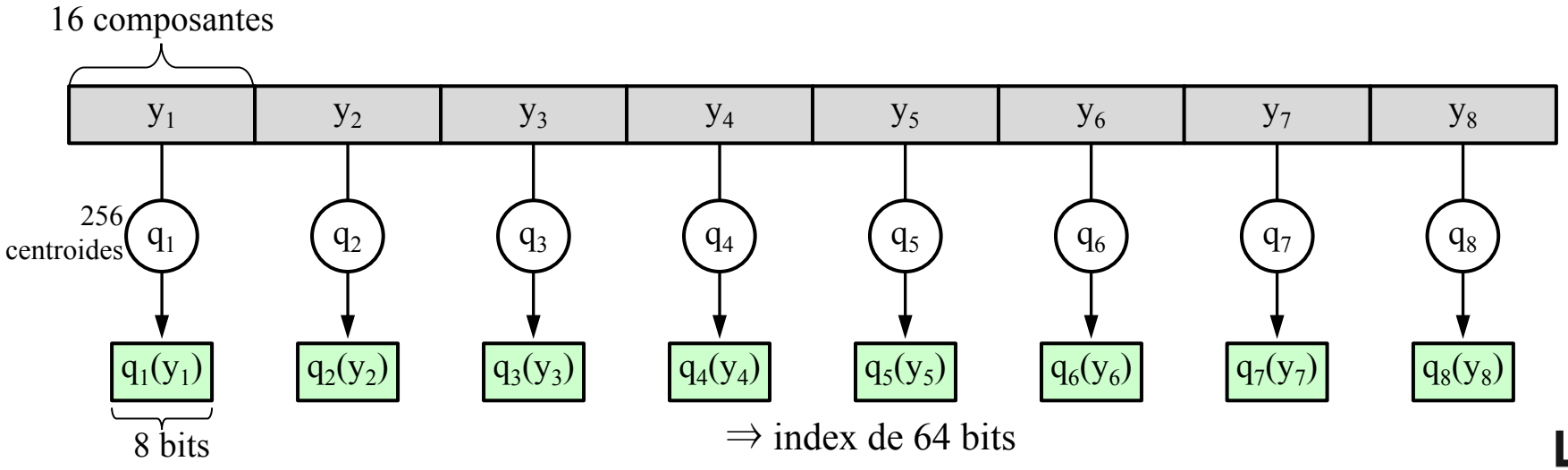


Quantificateur produit pour la recherche des plus proches voisins

- Vecteur coupé en *m* sous-vecteurs: $y \rightarrow [y_1|\ldots|y_m]$
- Sous-vecteurs quantifiés séparément: $q(y) = [q_1(y_1)|\ldots|q_m(y_m)]$

  chaque *qi* a un petit vocabulaire

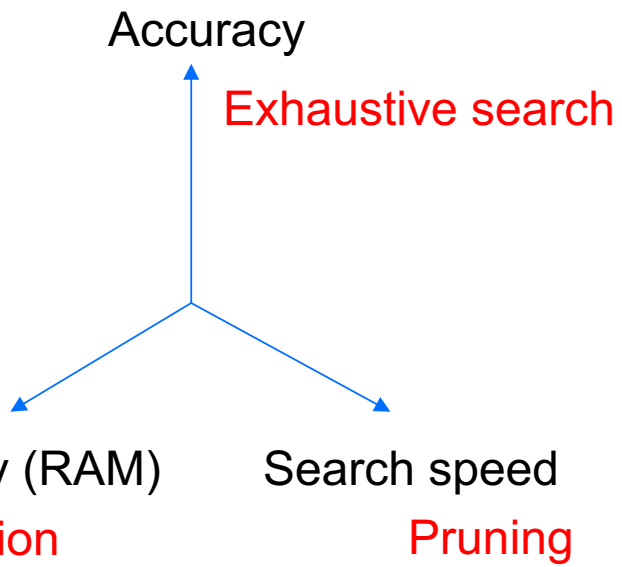
- Exemple de SIFT: y possède 128 dimensions. Il est découpé en 8 sous-vecteurs de dimension 16



NAVER LABS EUROPE

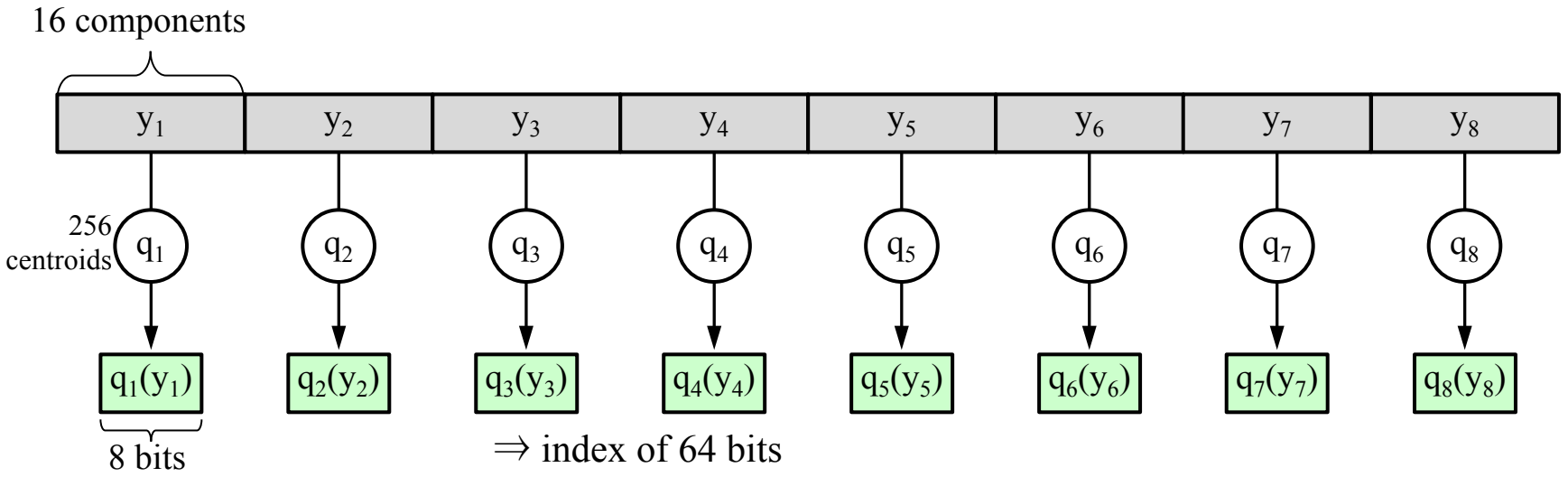
Ingredients of a full system 1/2

Trade-offs in similarity search



Compression

- Product Quantization (PQ)



[Hervé Jegou's tutorials]

NAVER LABS EUROPE

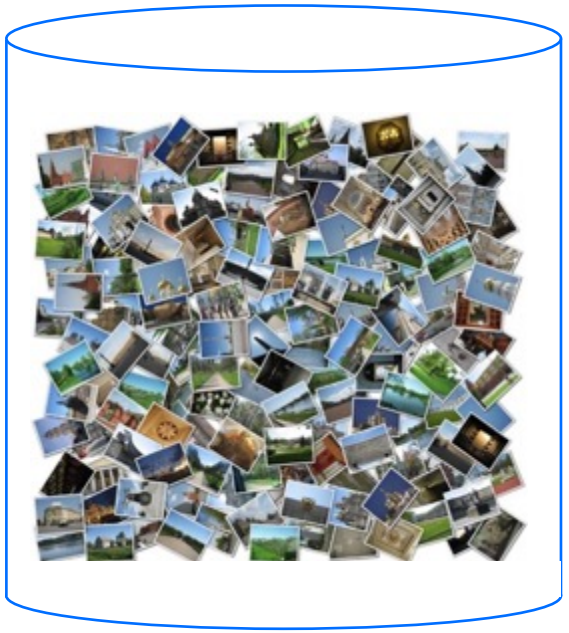
Ingredients of a full system 2/2

Leveraging the topology of the gallery set

- Query expansion

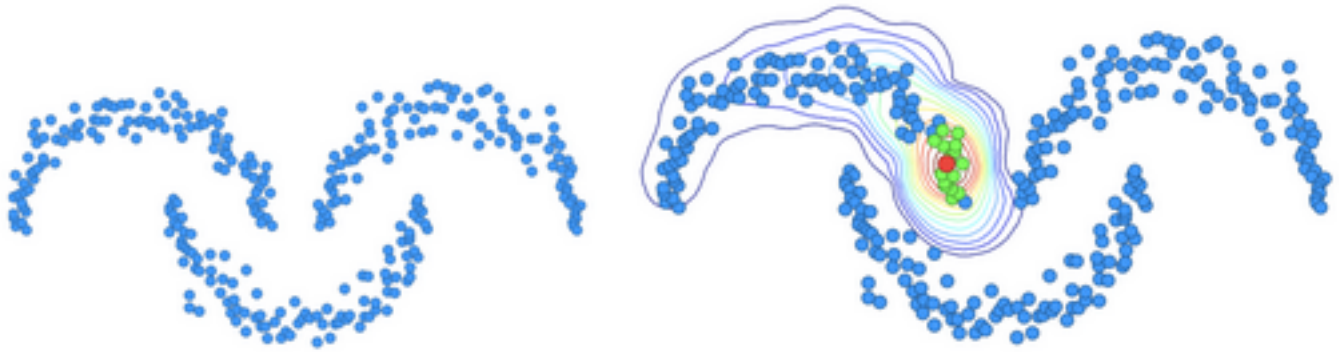
[Chum et al. ICCV07, Chum et al CVPR11]

  - Process the list of results (e.g. spatially verify)
  - If some images are good, use them to process some other augmented queries



- Diffusion

[Iscen et al. CVPR17, CVPR18]



NAVER LABS EUROPE

Landmark Recognition Challenge @ CVPR18

Top-performing techniques:

- CNN-based global features
  - Including RMAC, Differentiable RMAC, Generalized-Mean pooling
- Some combine with local features as well
  - E.g. SIFT, DELPH, for re-ranking and query expansion
- Query expansion and/or Diffusion
- Some extra points with:
  - Multi-resolution, ensembling

https://www.kaggle.com/c/landmark-retrieval-challenge/leaderboard

More about the workshop + slides [https://landmarkscvprw18.github.io]

NAVER LABS EUROPE

Requêtes complexes: cross-modalité et recherches sémantiques

Comprendre les données visuelles à grande échelle

Cours 3: représentations d'images, 10 janvier 2019

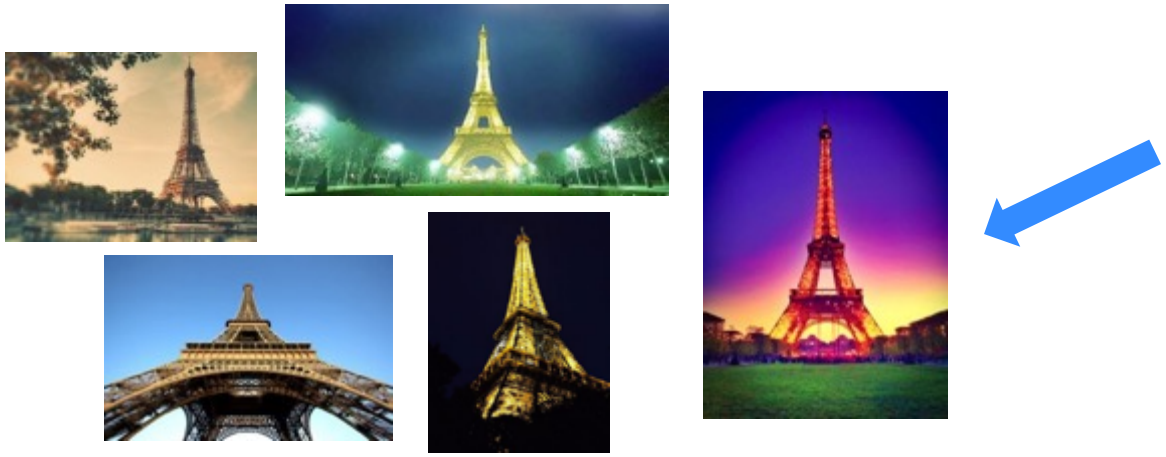
NAVER LABS EUROPE

Plan de la suite

- Recherche visuelle sémantique
  - Principe
  - Apprentissage d’une représentation dédiée
- Recherche cross-modale
  - Notion de **plongement** (*Embedding*)

Traditional Image Retrieval

The instance-level retrieval task



NAVER LABS EUROPE

Beyond Standard Image Retrieval

Complex queries that involve the full scene

Semantic Image Retrieval

[*Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. Gordo & Larlus. CVPR17*]

NAVER LABS EUROPE

Semantic image retrieval

Is the semantic retrieval task well-defined?

Human annotations

%?

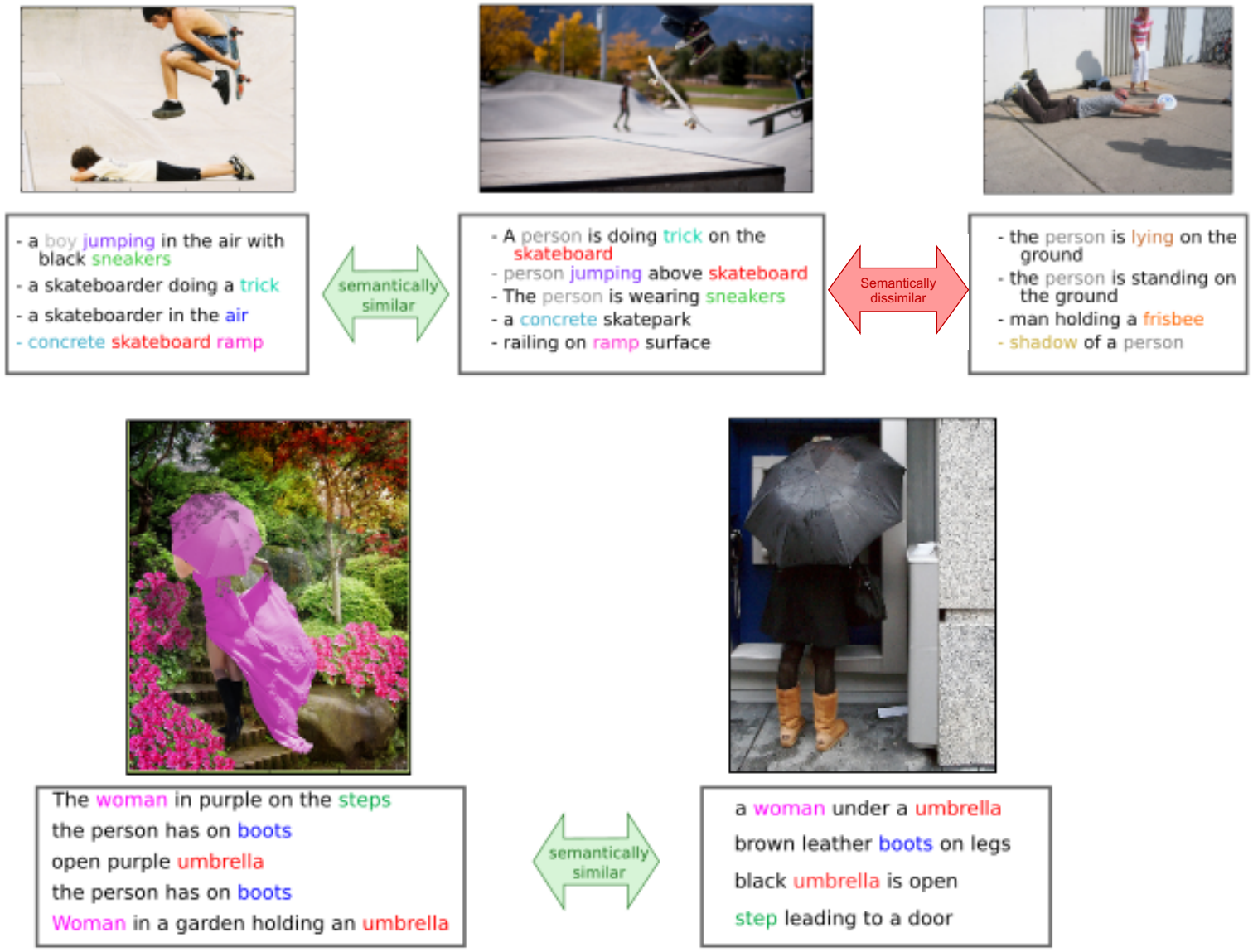
%?

[*Gordo & Larlus. CVPR17*]

NAVER LABS EUROPE

Semantic image retrieval

Human-generated captions



[*Gordo & Larlus. CVPR17*]

NAVER LABS EUROPE

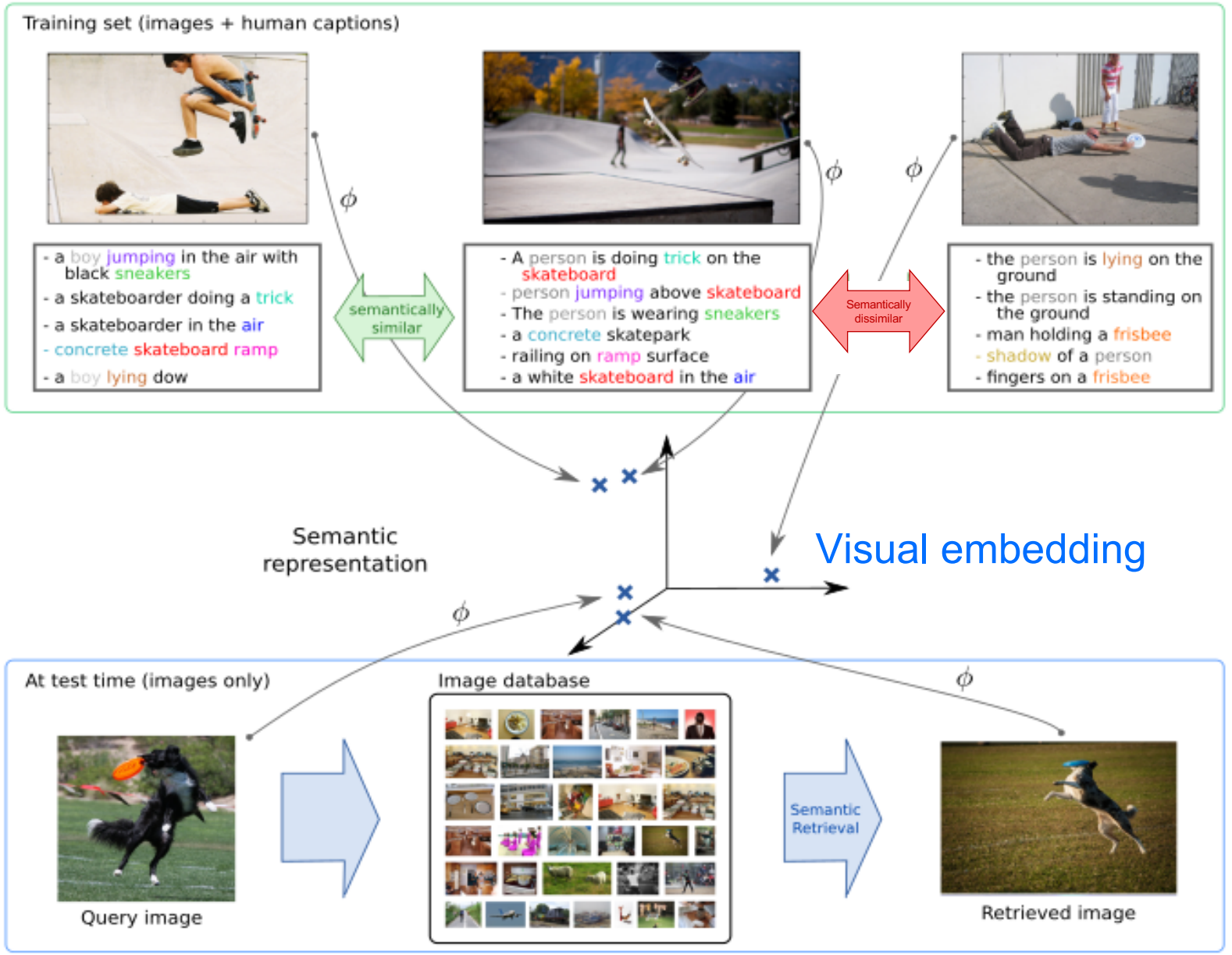
Semantic image retrieval

Building visual representations:

Intuition

For images with similar captions:

Visual representations close in the **semantic embedding space**

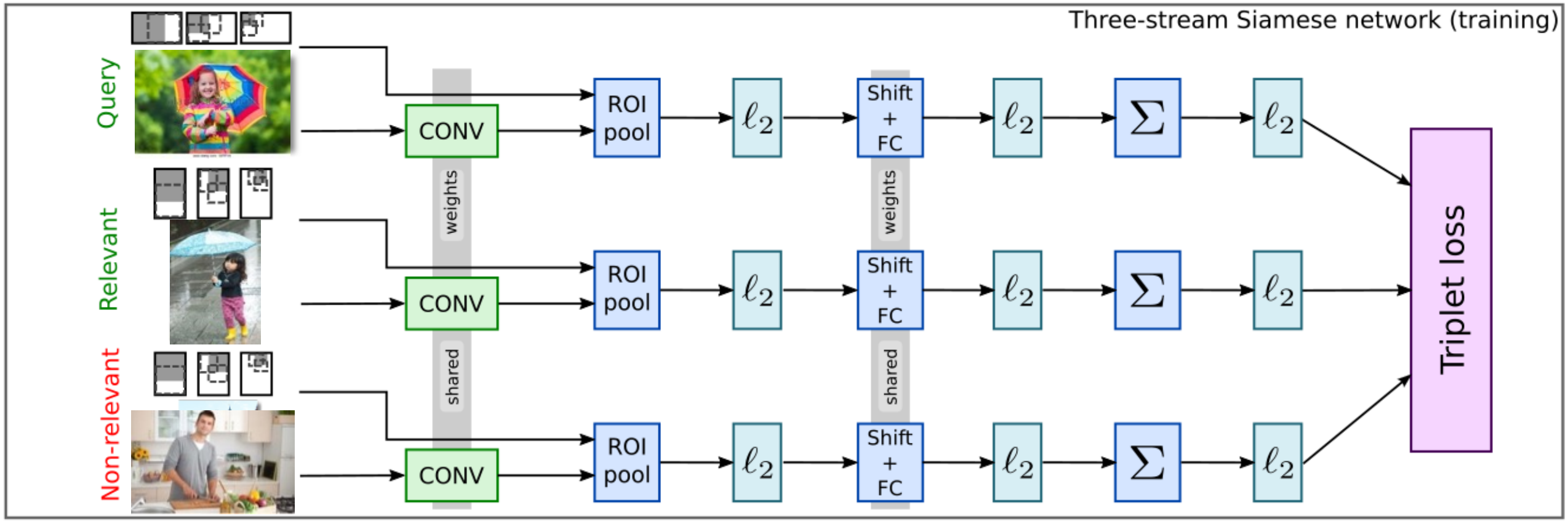


[*Gordo & Larlus. CVPR17*]

Learning a semantic embedding

- Learning to rank
  - Three-stream Siamese Network

[Gordo@ECCV16, Gordo@IJCV17]



Learning a semantic embedding

Visual loss:

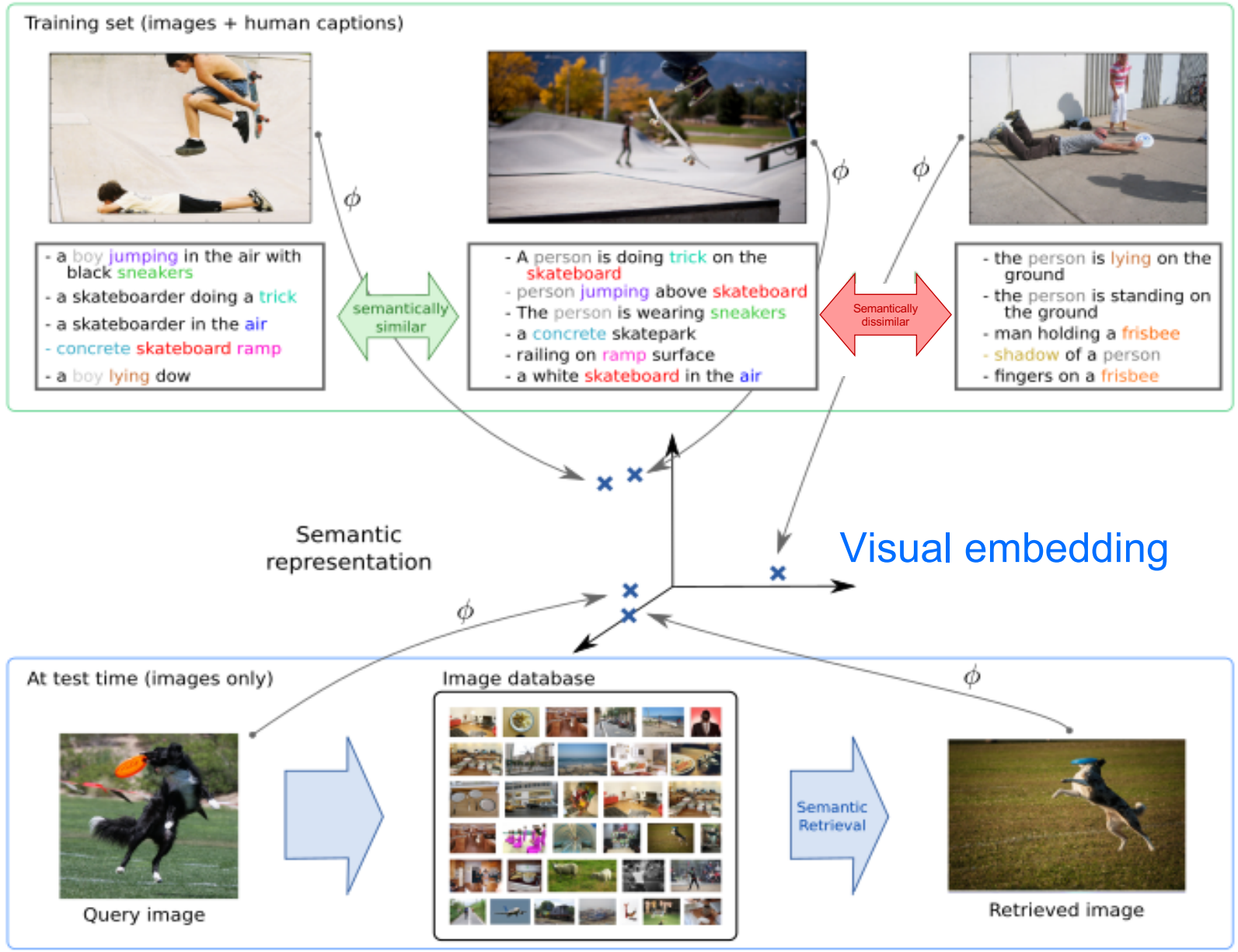
$$L_v(q, d^+, d^-) = \frac{1}{2}\max(0, m - \phi_q^T\phi_+ + \phi_q^T\phi_-)$$

[*Gordo & Larlus. CVPR17*]

Semantic image retrieval

Training the model

The visual representations of images with similar captions are close in the **semantic embedding space**



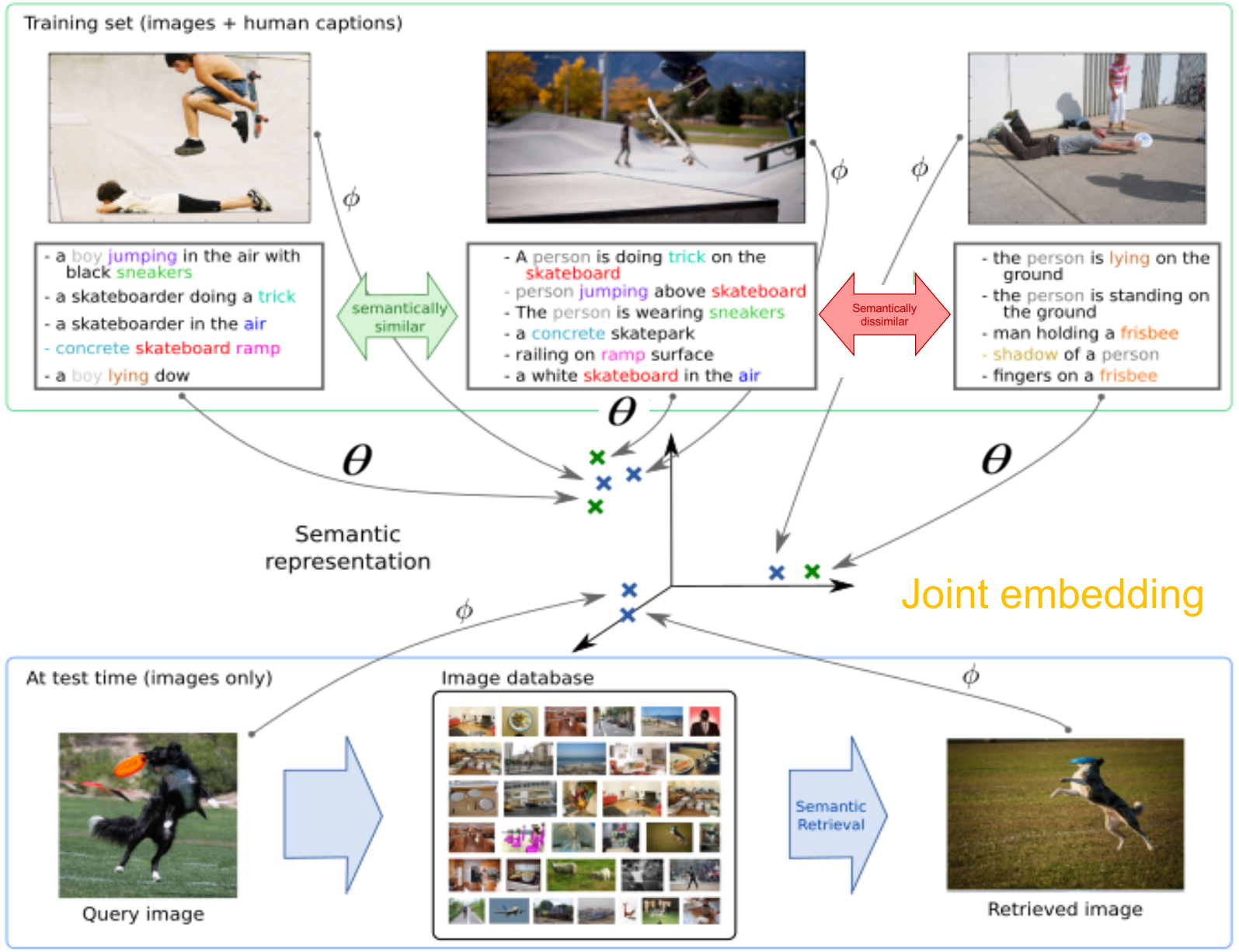
[*Gordo & Larlus. CVPR17*]

Semantic image retrieval

Training the model

The visual representations of images with similar captions are close in the **semantic embedding space**

The textual representations of corresponding captions are close in the **semantic embedding space**

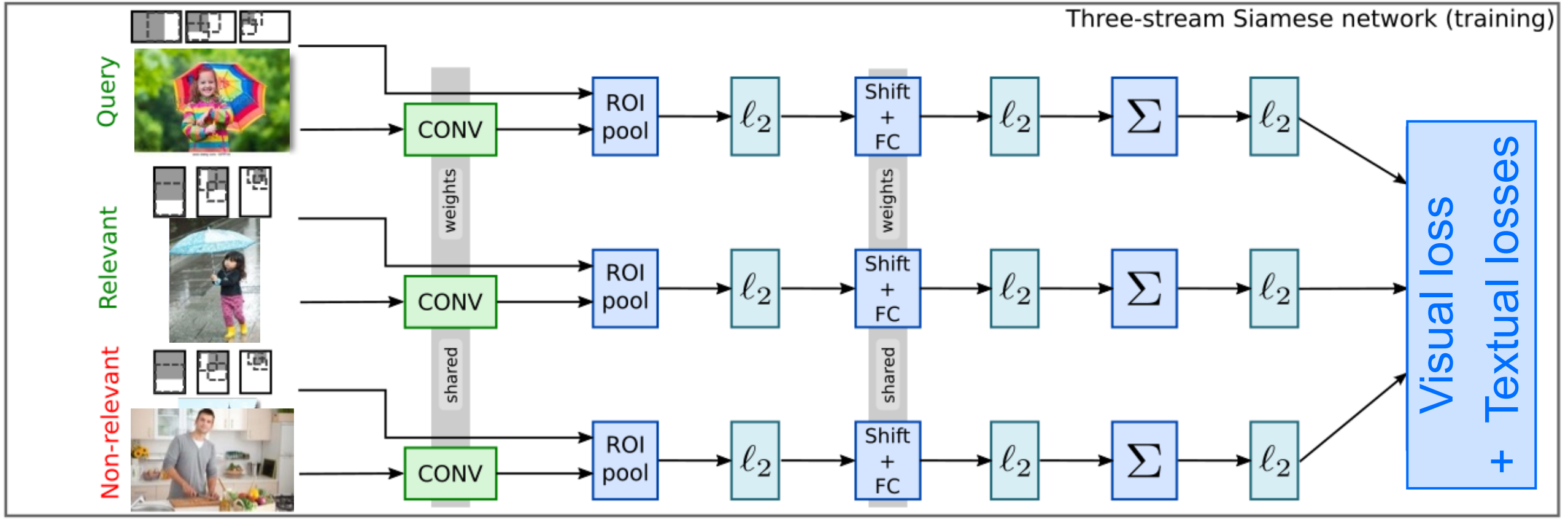


[*Gordo & Larlus. CVPR17*]

Learning a semantic embedding

[Gordo@ECCV16, Gordo@IJCV17]

- Learning to rank
  - Multiple losses are used to learn a good visual representation



Learning a semantic embedding

Visual loss:

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

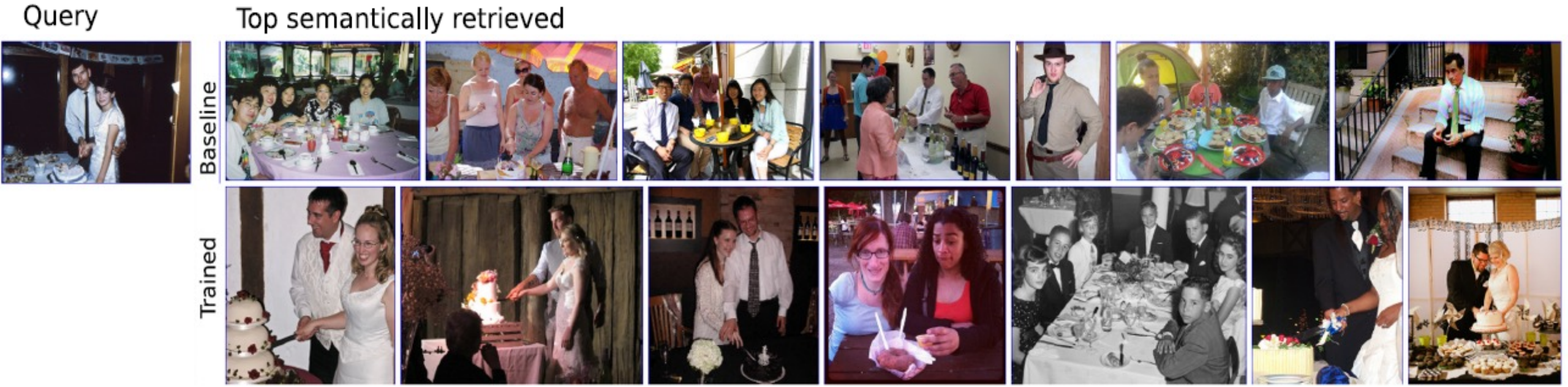
Textual losses:

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

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

[*Gordo & Larlus. CVPR17*]

Semantic retrieval: qualitative results



[*Gordo & Larlus. CVPR17*]

NAVER LABS EUROPE

Semantic retrieval: leveraging the joint embedding

The joint embedding allows for multimodal queries

[*Gordo & Larlus. CVPR17*]

NAVER LABS EUROPE

Text-to-image and Image-to-text retrieval

Offline: build a joint text+image representation

The semantic embedding space is good for cross-modal searches

Online: embed the query, and compare with the embeddings of the elements in the data set