Research Institute for Future Media Computing
未来媒体技术与研究所
Institute of Computer Vision
计算机视觉研究所

多媒体系统导论

Fundamentals of Multimedia System

授课教师：朱映映教授

Email：zhuyy@szu.edu.cn

第十讲

Content-based Image Retrieval

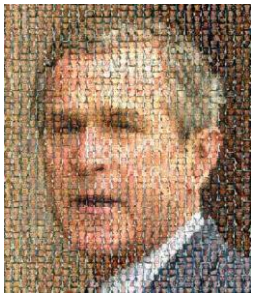
第20章

Background

- Necessity of retrieval
  - *Information is of no use, unless you can actually access it.*

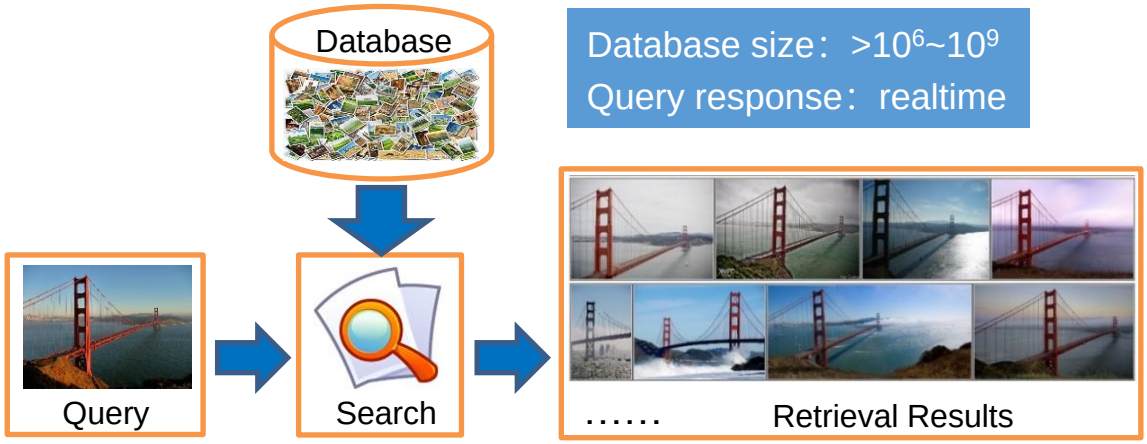
[from the TREC homepage: trec.nist.gov]

- Why do we need image retrieval?
- "A Picture is worth thousand words"
- Not everything can be described in text
- Not everything is described in text



Background

- Content based image retrieval



- Potential applications of content based image retrieval

Image Retrieval

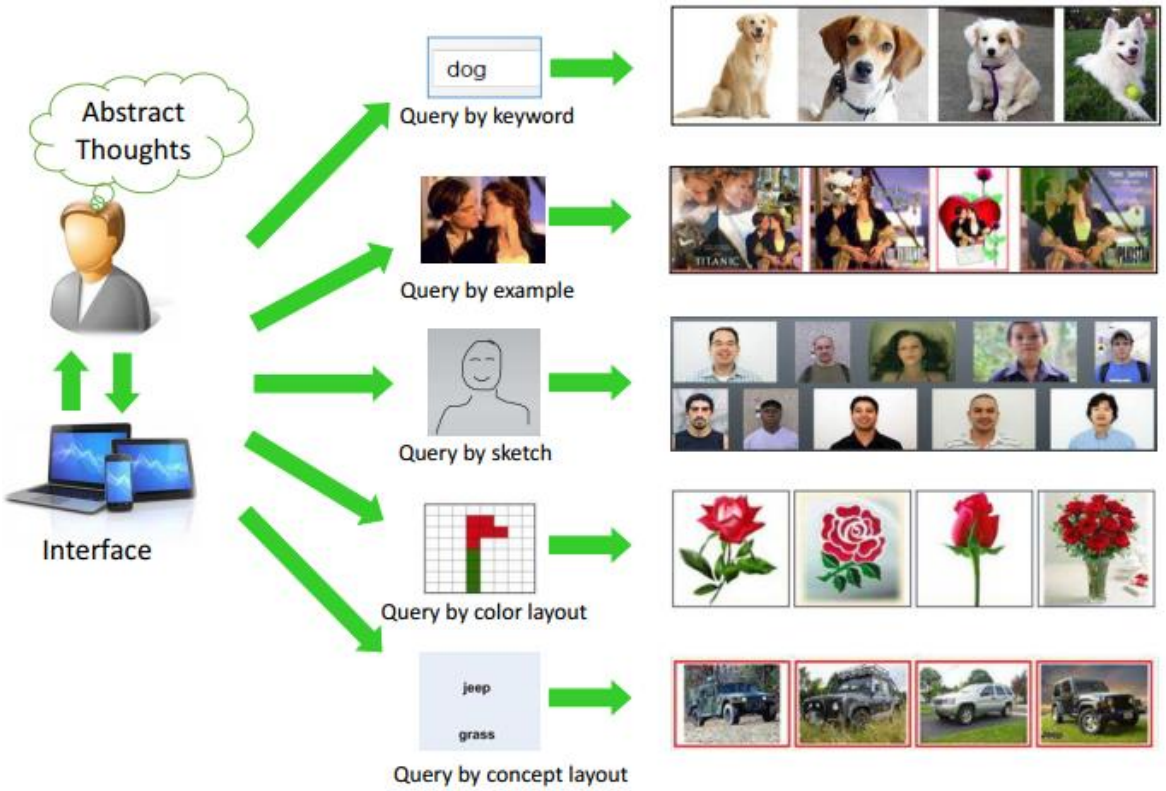
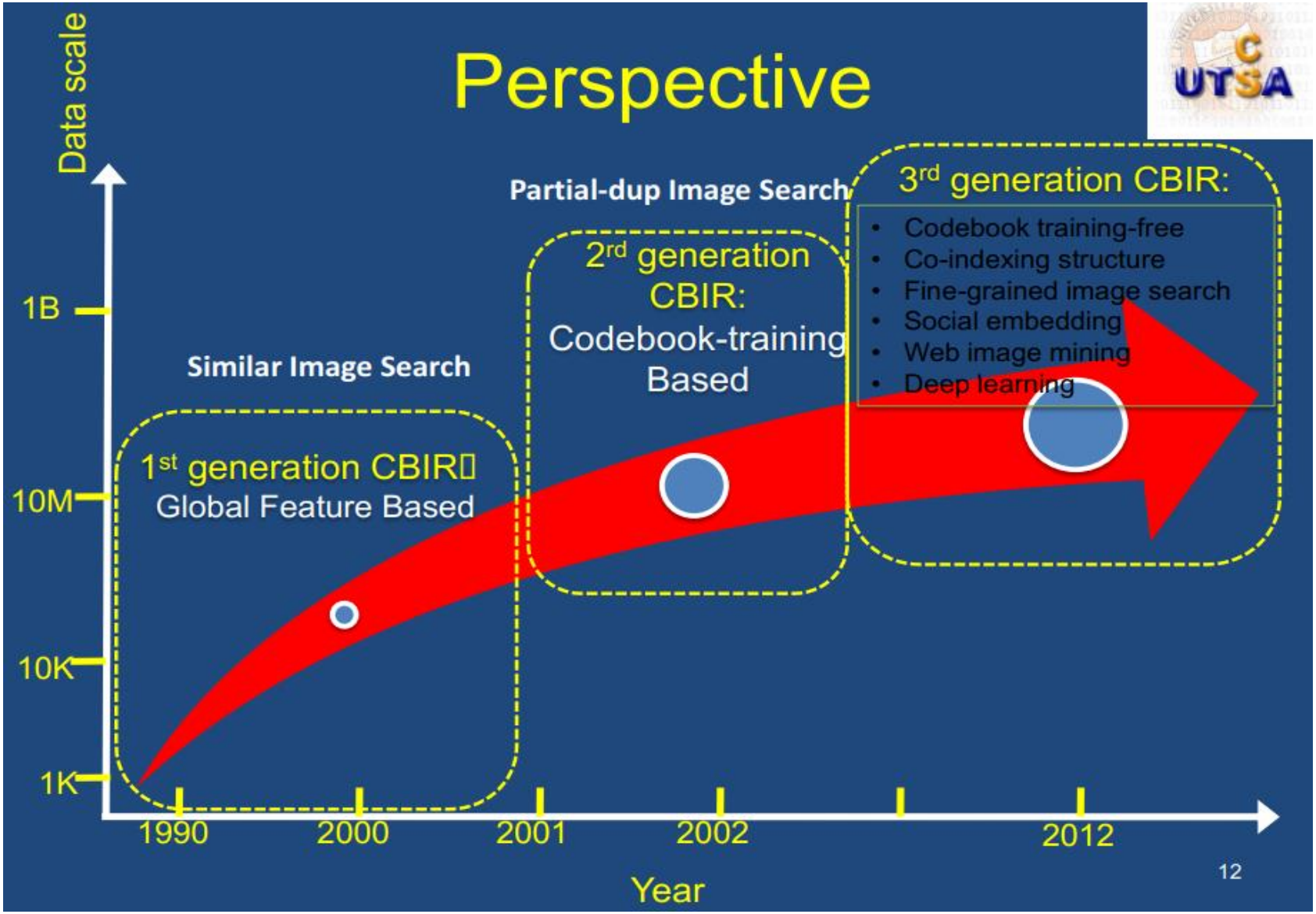


Illustration of different query schemes with the corresponding retrieval results

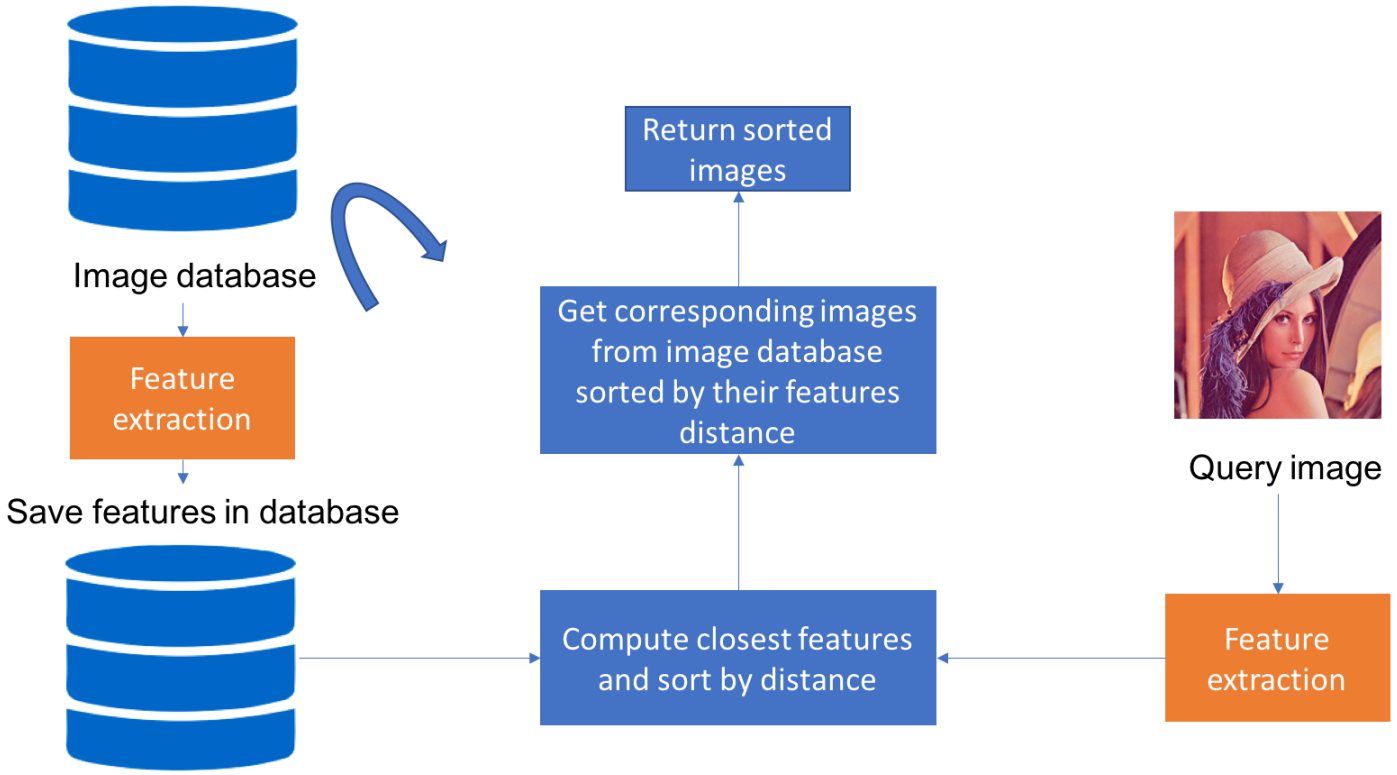
Content-Based on Image Retrieval (CBIR)

- What is the topic of this image
- What are right keywords to index this image
- What words would you use to retrieve this image ?
- **The Semantic Gap**
  - A picture is worth a thousand words
  - The meaning of an image is highly individual and subjective

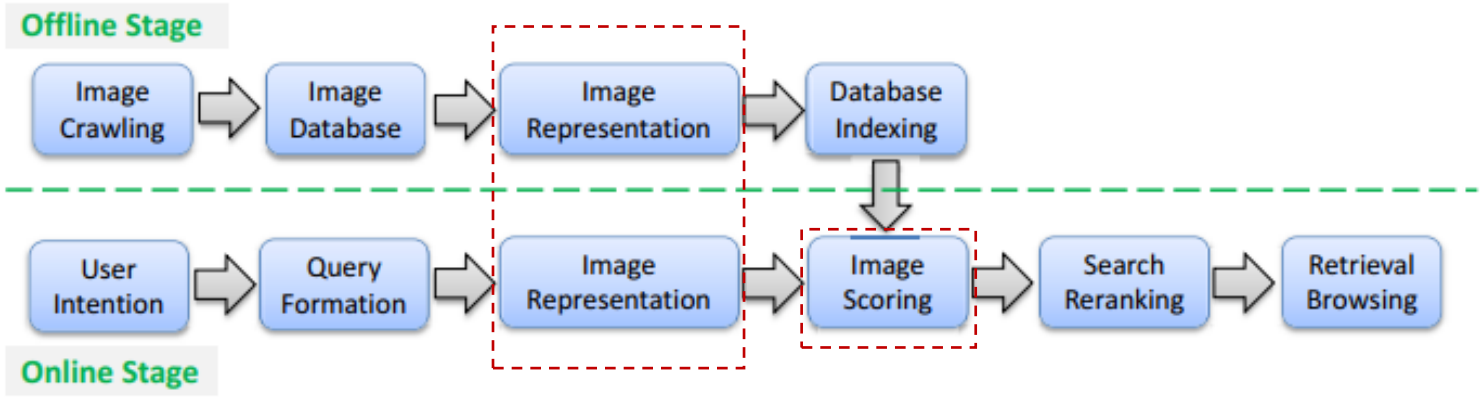
Generations of CBIR



Framework of CBIR

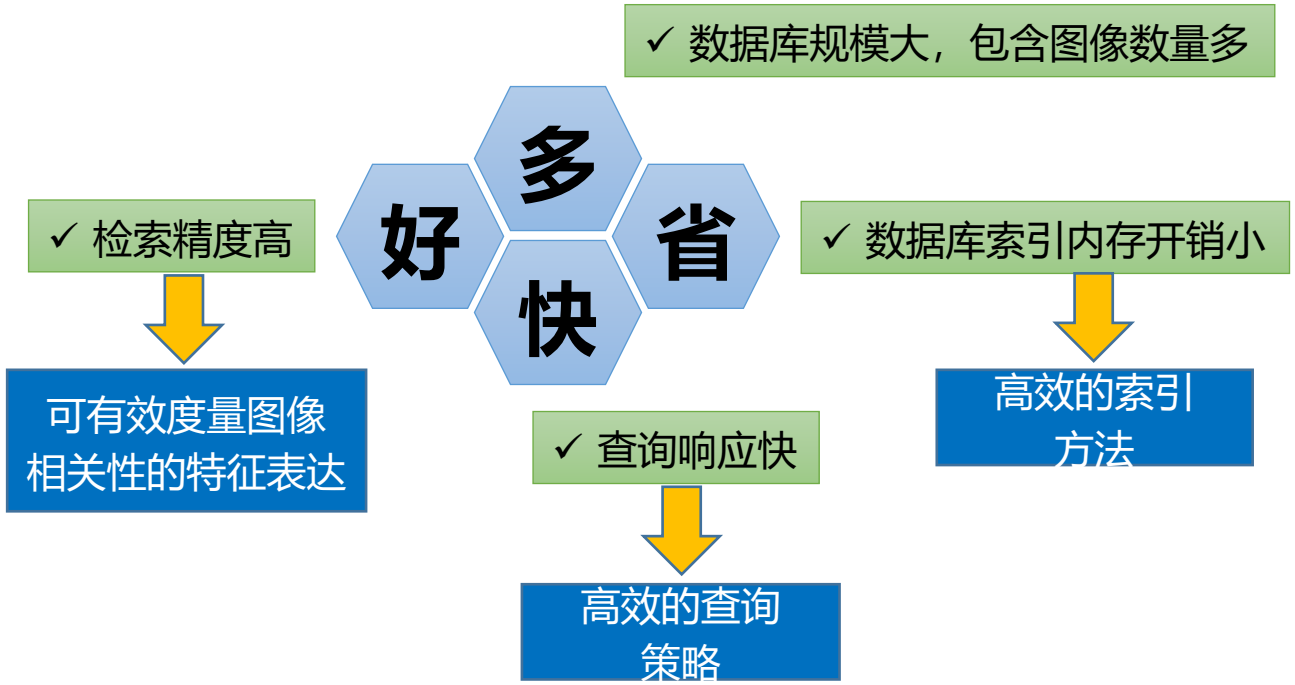


Framework of CBIR



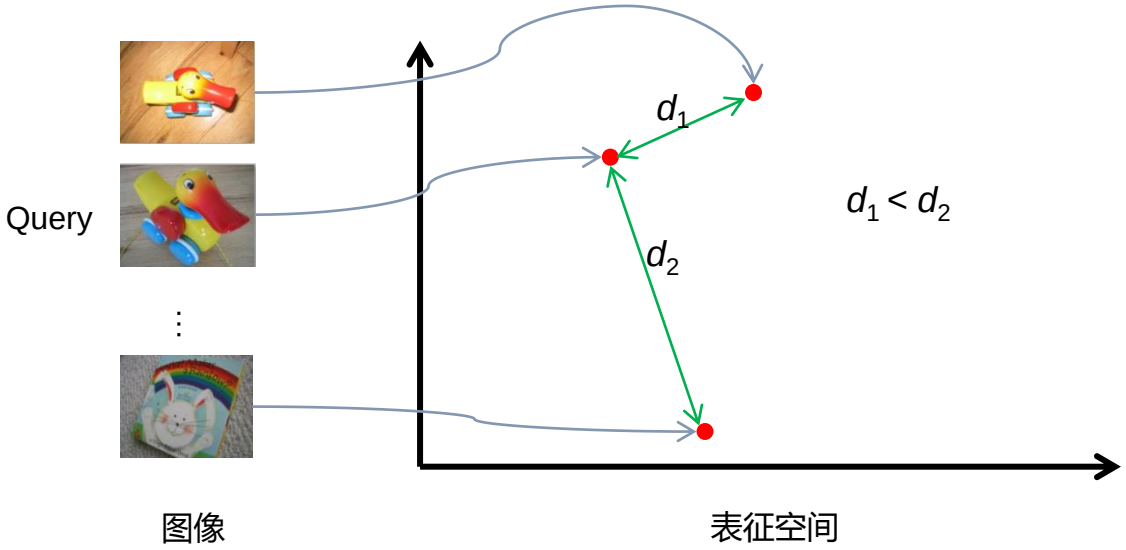
The general framework of content-based image retrieval

Problems with Image Retrieval



Basic Problems

- 图像检索基本问题之一：如何计算图像间的内容相关性?
  - **图像表征**：非结构化图像数据的结构化表达
    - SIFT + BoW/VLAD/FV
    - Activations of the intermediate layers of CNN model
  - **相似性度量**：基于图像表征的相关性计算



Based on Hash

A Simple Example-aHash(均值哈希)

基本思想：比较灰度图每个像素与平均值来实现

步骤：

1. 缩小图片：为了保留结构去掉细节，去除大小、横纵比的差异，把图片统一缩小到 8*8(像素)，共 64 个像素。
2. 转化为灰度图：把缩放后的图片转化为 256 阶的灰度图。
3. 计算平均值： 计算进行灰度处理后图片的所有像素点的平均值。
4. 比较像素灰度值：遍历 64 个像素，如果大于平均值记录为 1，否则为 0。
5. 得到信息指纹：组合 64 个 bit 位，顺序随意保持一致性即可。
6. 对比指纹：计算两幅图片的汉明距离，汉明距离越大则说明图片越不一致，反之，汉明距离越小则说明图片越相似。

A Simple Example-aHash (均值哈希)

结论：

1. 当距离为 0 时，说明完全相同；
2. 通常认为距离>10 就是两张完全不同的图像；
3. 如果汉明距离小于5，则表示有些不同，但比较相近。

优点：

1. 图像放大或缩小，或改变纵横比，Hash值不会改变;
2. 增加或减少亮度或对比度，或改变颜色，对hash值都不会有太大的影响;
3. 计算速度快。

缺点：

丢失了高频信息，丢失了细节

A Simple Example-pHash(感知哈希)

基本思想：使用离散余弦变换(DCT)来获取图像的低频成分

步骤：

1. 缩小尺寸：pHash从小图像开始，但图像大于8*8，32*32是最好的。目的是简化DCT的计算，而不是减小频率；
2. 转化为灰度图像：进一步简化计算量；
3. 计算DCT：计算图像的DCT变换，得到32*32的DCT系数矩阵；
4. 缩小DCT：虽然DCT的结果是32*32大小的矩阵，但我们只要保留左上角的8*8的矩阵，这部分呈现了图像中的最低频率；
5. 计算平均值：如同均值哈希一样，计算DCT的均值；
6. 计算hash值：如同均值哈希一样，根据8*8的DCT矩阵，设置0或1的64位的hash值，大于等于DCT均值的设为”1”，小于DCT均值的设为“0”。组合在一起，就构成了一个64位的整数，这就是这张图像的指纹。

优点：

1. 只要图像的整体结构保持不变，hash结果值就不变；
2. 能够避免伽马校正或颜色直方图被调整带来的影响优点：

缺点：

计算速度较Ahash慢

A Simple Example-dHash(差异哈希)

基本思想：基于渐变实现

步骤：

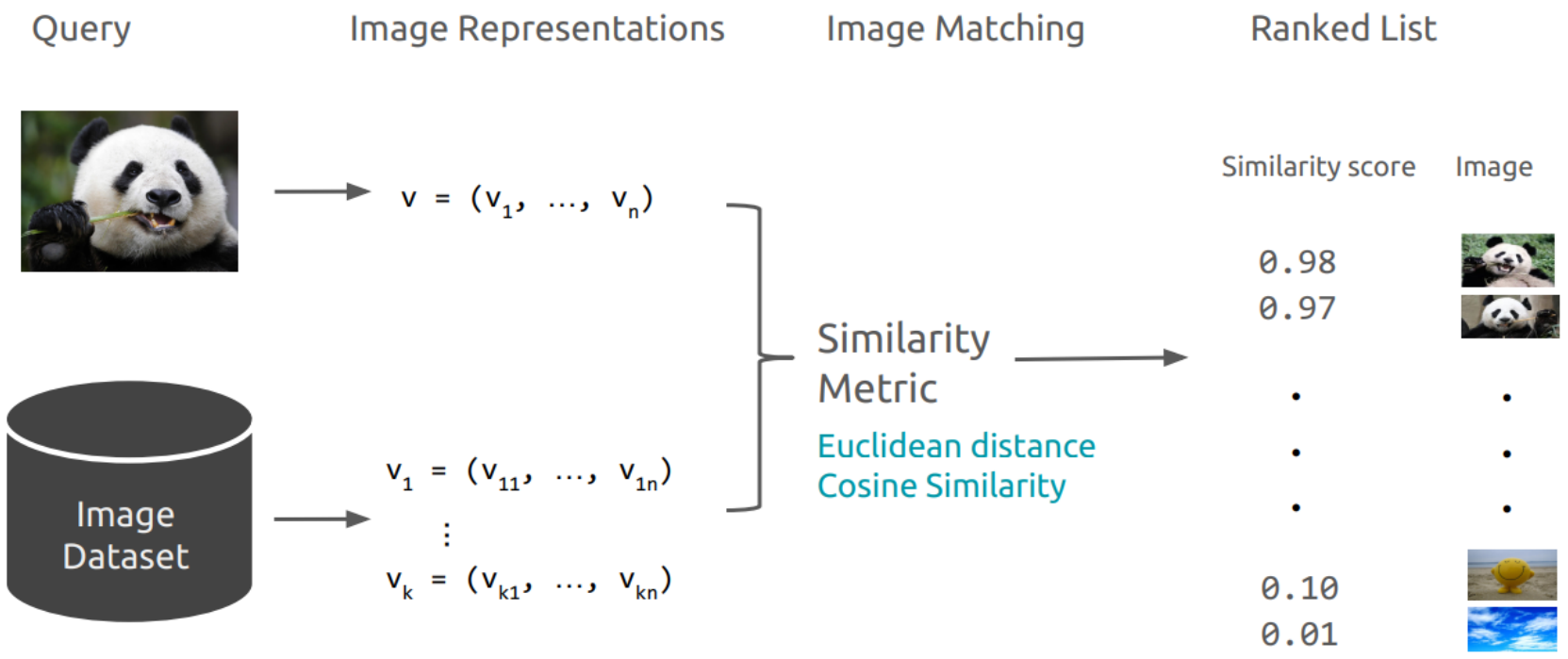
1. 缩小图像：缩小到 9(列)*8(行) 的大小，共 72 个像素点；
2. 转化为灰度图：把缩放后的图片转化为 256 阶的灰度图；
3. 计算差异值： dHash 算法工作在相邻像素之间，这样每行 9 个像素之间产生了 8 个不同的差异，8 行*8，则产生了 64 个差异值；
4. 获得指纹：如果左边像素的灰度值比右边高，则记录为 1，否则为 0；
5. 对比指纹：同平均哈希算法。

优点：

1. 相比 pHash， dHash 的速度要快得多；
2. 相比 aHash， dHash 在效率几乎相同的情况下的效果要更好。

Based on Hand-Crafted Features

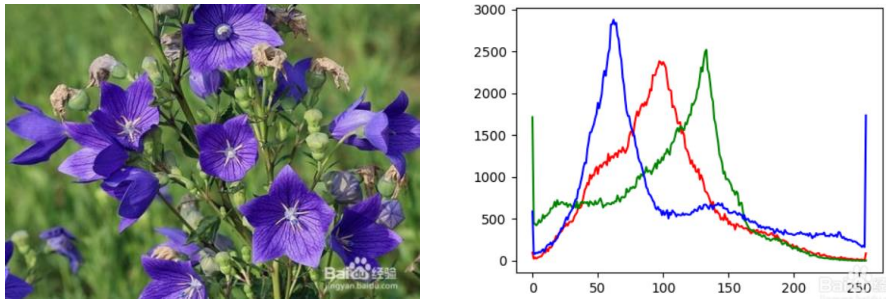
The retrieval pipeline



Global Features

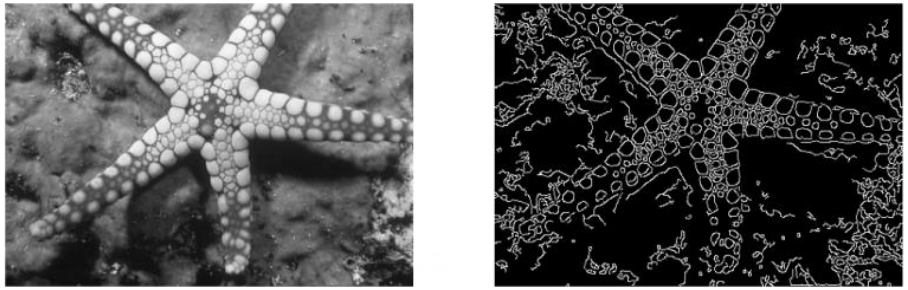
Global features refer to features extracted from the entire image, which describe the overall characteristics and information of the image.

- Color: Color histogram
- Texture: Gray-Level Co-occurrence Matrix（灰度共生矩阵，CLCM），Gabor filters
- Shape: Edge-based Features (Sobel, Canny, Prewitt, Laplacian 等), La Hu moments
- Other: Fourier Transform, Wavelets



Color histogram

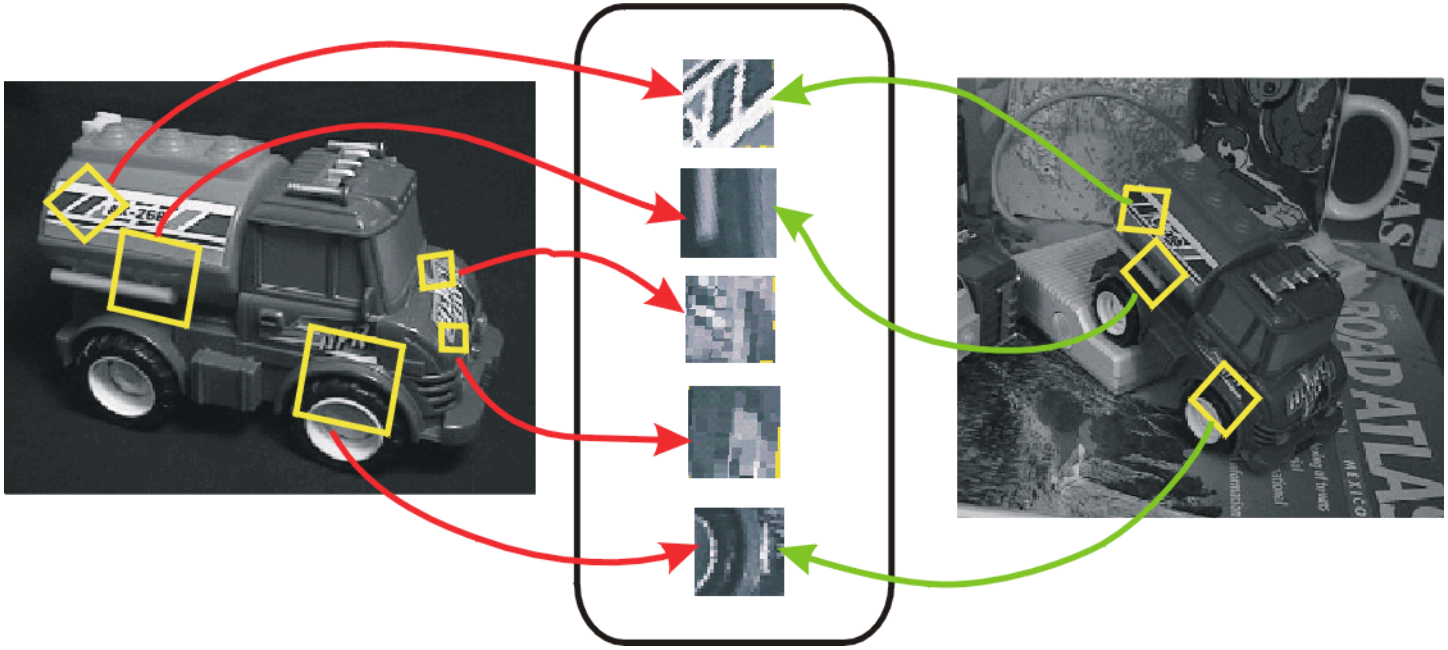
Gabor filters



Edge detetion

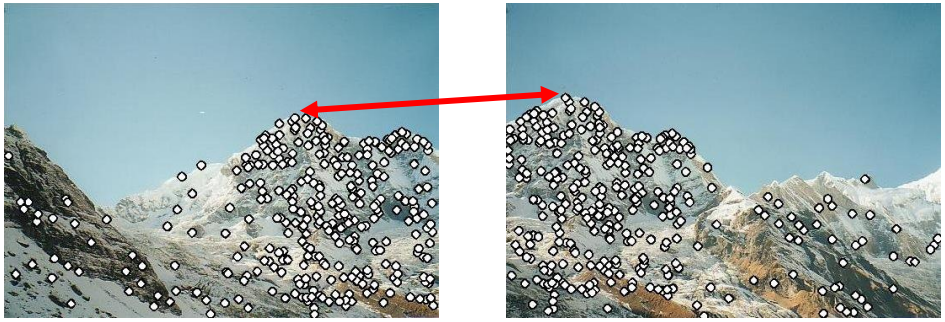
Local Features

- Find features that are invariant to transformations
  - geometric invariance: translation, rotation, scale, affine…
  - photometric invariance: brightness, exposure, …



Local Features

- SIFT
- SURF
- ORB
- BRISK
- HOG
- And so on



Local Feature--SIFT

- SIFT: 尺度不变特征变换（Scale-invariant feature transform，SIFT）[1]
- 该方法于1999年由David Lowe [2] 首先发表于计算机视觉国际会议（International Conference on Computer Vision，ICCV）[3]，2004年再次经David Lowe整理完善后发表于International journal of computer vision（IJCV）[2]
- SIFT特点

① SIFT特征是图像的局部特征，其对旋转、尺度缩放、亮度变化保持不变性，对视角变化、仿射变换、噪声也保持一定程度的稳定性

② 区分性好，适用于在海量特征数据库中进行快速、准确的匹配

③ 多量性，即使少数的几个物体也可以产生大量的SIFT特征向量

④ 高速性，经优化的SIFT匹配算法甚至可以达到实时的要求

⑤ 可扩展性，能够与其它形式的特征向量进行联合

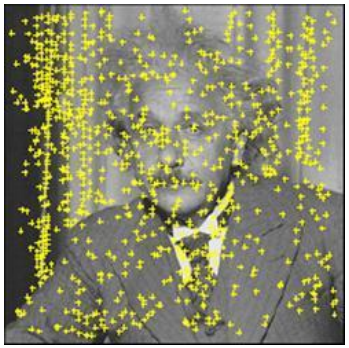
- SIFT算法在一定程度上可解决：目标的旋转、缩放、平移（RST）、图像仿射/投影变换（viewpoint）、光照影响（illumination）、目标遮挡（occlusion）、杂物场景（clutter）、噪声

[1] https://www.cs.ubc.ca/~lowe/keypoints/
[2] David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110.
[3] David G. Lowe, "Object recognition from local scale-invariant features," International Conference on Computer Vision, Corfu, Greece (September 1999), pp. 1150-1157.

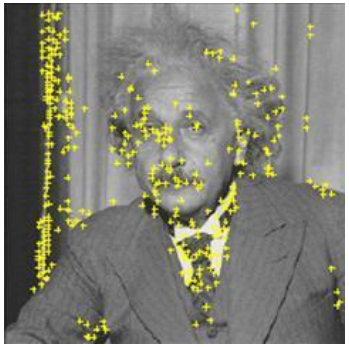
Local Feature--SIFT

- SIFT特征的生成，即从多幅图像中提取对尺度缩放、旋转、亮度变化无关的特征向量。

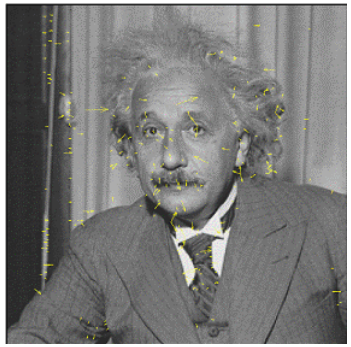
1. 粗检特征点：构建尺度空间，检测极值点，获得尺度不变性



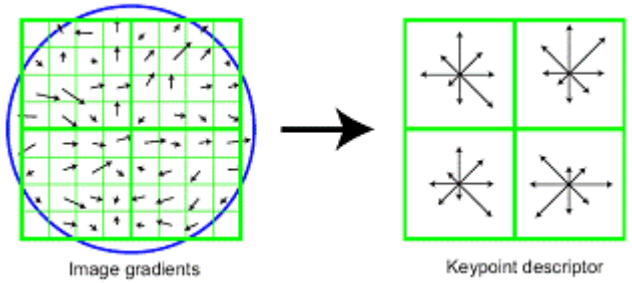
2. 特征点过滤并进行精确定位



3. 为特征点分配方向值



4.生成特征描述子



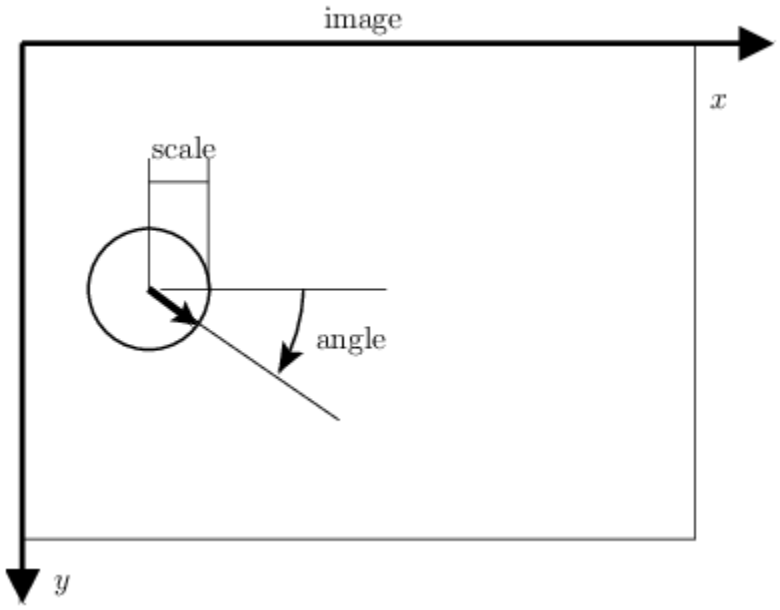
Local Feature--SIFT

- 主要步骤

- **尺度空间的极值检测**：搜索所有尺度空间上的图像，通过高斯微分函数来识别潜在的对尺度和旋转不变的兴趣点。（高斯金字塔->DOG->极值检测）
- **特征点定位**：在每个候选的位置上，通过一个拟合精细模型来确定位置尺度，关键点的选取依据他们的稳定程度。（二阶泰勒展开+边缘点过滤）
- **特征方向赋值**：基于图像局部的梯度方向，分配给每个关键点位置一个或多个方向，后续的所有操作都是对于关键点的方向、尺度和位置进行变换，从而提供这些特征的不变性。
- **特征点描述**：在每个特征点周围的邻域内，在选定的尺度上测量图像的局部梯度，这些梯度被变换成一种表示，这种表示允许比较大的局部形状的变形和光照变换。生成128维的SIFT特征向量。
- **特征匹配：**可以先对128维向量归一化。取图像1中的某个关键点，并找出其与图像2中欧式距离最近的前两个关键点，在这两个关键点中，如果最近的距离除以次近的距离少于某个比例阈值下面的Threshold(也叫ratio)，则接受这一对匹配点。

Local Feature--SIFT

- 一个SIFT keypoint是一个圆形区域并且带有方向，使用4个参数描述该区域的几何结构：
- keypoint的中心位置的坐标(x,y)
- keypoint的scale(圆形区域的半径r)
- keypoint的方向(使用弧度表示的角度θ独特性



Local Feature--SIFT

- 缺点
  - 实时性不高，因为要不断地要进行下采样和插值等操作；
  - 有时特征点较少（比如模糊图像）；
  - 对边缘光滑的目标无法准确提取特征（比如边缘平滑的图像，检测出的特征点过少，对圆更是无能为力）。

- 在一定程度上可解决
  - 目标的旋转、缩放、平移（RST）
  - 图像仿射/投影变换（视点viewpoint）
  - 光照影响（illumination）
  - 目标遮挡（occlusion）
  - 杂物场景（clutter）
  - 噪声

The Classic SIFT Retrieval Pipeline

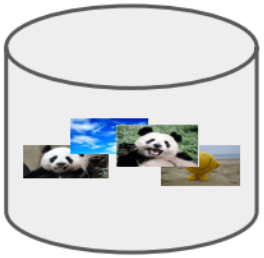
The classic SIFT retrieval pipeline

variable number of
feature vectors per image

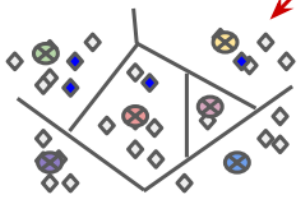
$v_1 = (v_{11}, \ldots, v_{1n})$

$\vdots$

$v_k = (v_{k1}, \ldots, v_{kn})$



N-Dimensional
feature space



M visual words
(M clusters)

Bag of Visual Words

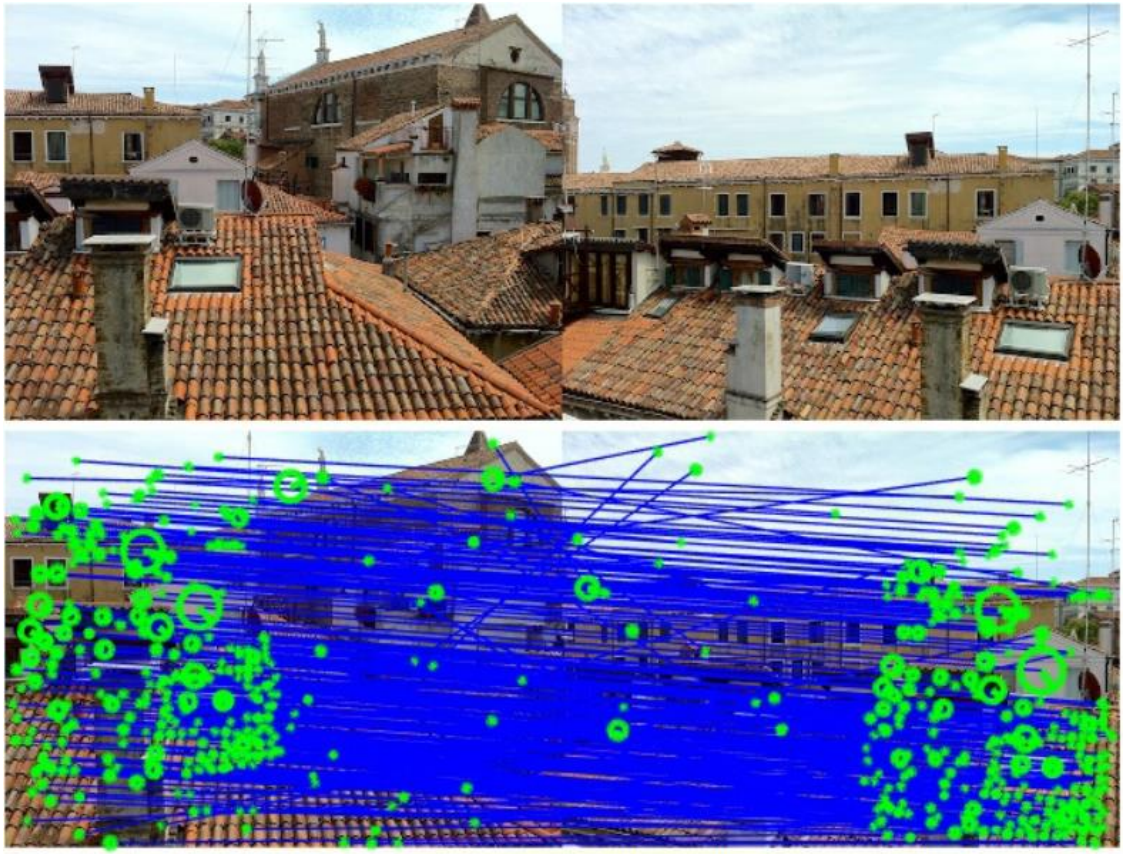
INVERTED FILE

word	Image ID
1	1, 12,
2	1, 30, 102
3	10, 12
4	2,3
6	10

Large vocabularies (50k-1M)
Very fast!
Typically used with SIFT features

Local Feature--SIFT

- 当两幅图像的SIFT特征向量生成以后，下一步就可以采用关键点特征向量的欧式距离来作为两幅图像中关键点的相似性判定度量。取左图的某个关键点，通过遍历找到右图中的距离最近的两个关键点。在这两个关键点中，如果最近距离除以次近距离小于某个阈值，则判定为一对匹配点。



Top: A pair of images of the same scene. Bottom: Matching of SIFT descriptors with `vl_ubcmatch`.

sift_match.m

https://www.vlfeat.org/overview/sift.html

Bag of Features

- 特征词袋(BoF，Bag Of Features)借鉴文本处理的词袋（BoW，Bag Of Words)算法。
- **SIFT特征提取:** 提取训练集中所有图像的SIFT特征，设有M幅图像，共得到N个SIFT特征。
- **构建视觉词汇表:** 对提取到的N个SIFT特征进行聚类，得到K个聚类中心，组成图像的视觉词汇表。
- **图像的视觉词向量表示**: 统计每幅图像中视觉词汇的出现的次数，得到图像的特征向量。向量的每个分量表示某个视觉词在图像中出现的次数。即将一幅图像表示为K维的向量（K为聚类中心的个数，也就是视觉词的个数）。在检索时，该特征向量就代表该幅图像。

Brief Summary

- **Feature Extraction**

  Global features；Local features

- **Features Coding**

  BoW[1]: Bag of (Visual) Words；VLAD[2]: Vector of Aggregate Locally Descriptor； FV[3]: Fisher Vector；SPM: Spatial Pyramid Matching[4]

- **Indexing**

  Invert Index；Clustering；Tree

- **Similarity Metric**

  Euclidean distance; Cosine distance；Hamming distance； Manhattan distance

- **ReRanking**

  Geometric Verification

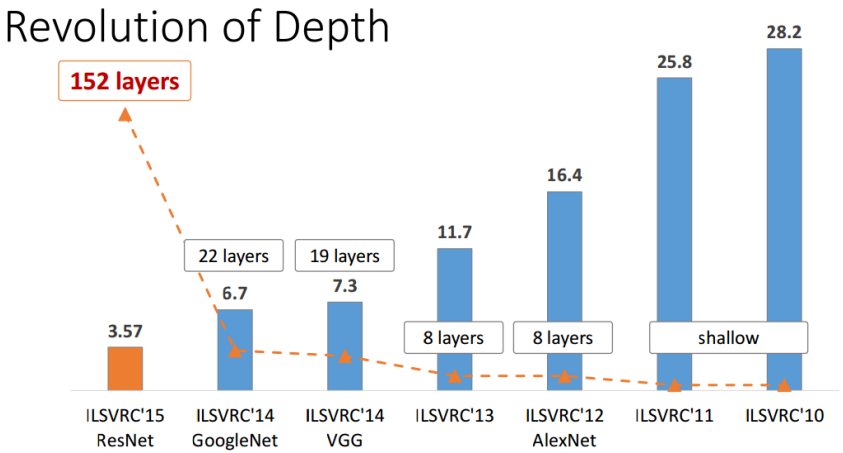
[1] J. Sivic and A. Zisserman, “Video Google: A text retrieval approach to object matching in videos,” in Proc. Int. Conf. Comput. Vis., 2003, Art. no. 1470.
[2] H. Jegou, M. Douze, C. Schmid, and P. Perez, “Aggregating local descriptors into a compact image representation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 3304–3311.
[3] F. Perronnin, J. S anchez, and T. Mensink, “Improving the Fisher kernel for large-scale image classification,” in Proc. Eur. Conf. Comput. Vis., 2010, pp. 143–156.
[4] Lazebnik S , Schmid C , Ponce J . Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories[C]// Computer Vision and Pattern Recognition. IEEE Computer Society, 2006.

传统图像检索方法的局限性

- 特征提取依赖于人工设计的算法，需要专业知识和复杂的调参过程，不能自动适应不同的图像内容和场景。
- 特征描述子通常基于低级的像素或几何信息，难以捕捉高级的语义和概念信息，对光照、角度、视角、遮挡、背景等变化不够鲁棒。
- 特征匹配和相似度计算需要大量的时间和空间资源，对于海量的图像数据，效率和可扩展性较低。
- 特征编码和索引方式往往牺牲一定的精度，为了提高检索速度，可能导致检索结果的质量下降。

Based on Deep Learning

- 深度学习在计算机视觉领域取得巨大成功
  - ImageNet Grand Challenge



- 深度学习：时势造英雄
  - 大规模图像视频数据
  - 强大的计算能力：GPU/TPU

IMAGENET

flickr

60亿张图片，2011年

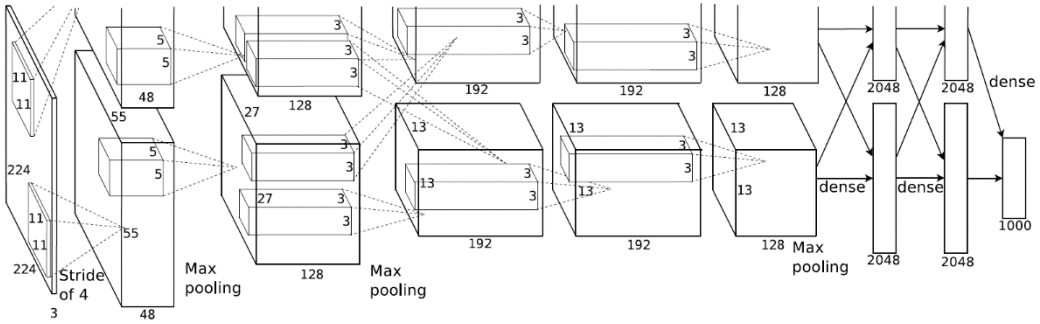
facebook

2500亿张图片，2013年

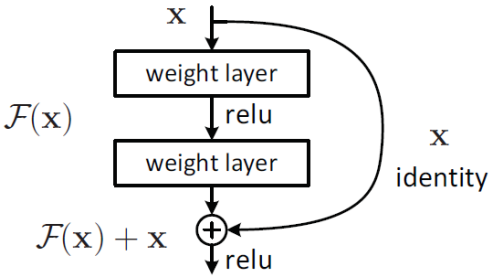
NVIDIA

研究背景：深度学习

- 面向图像分类的深度学习模型

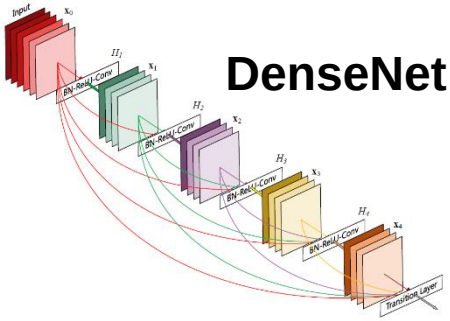


AlexNet, 2012

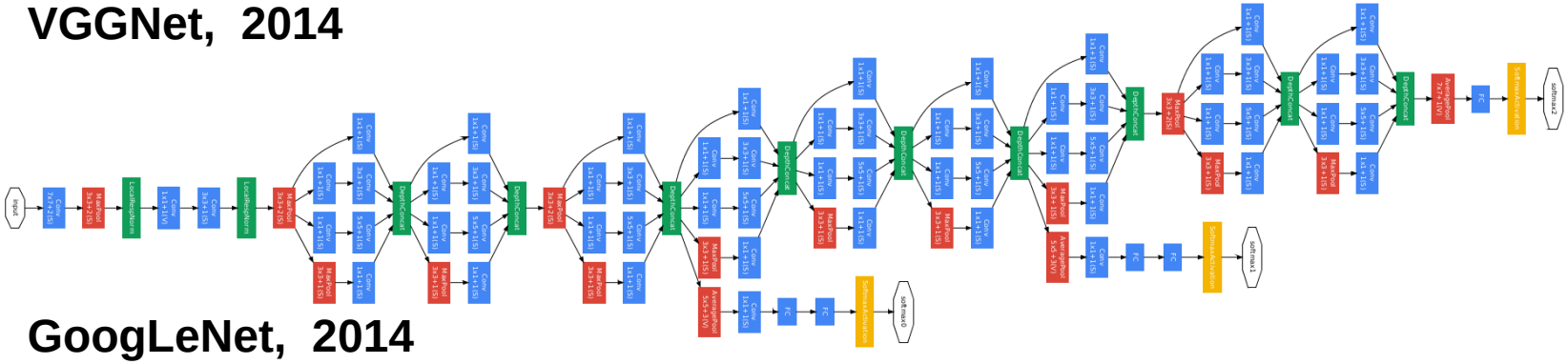


ResNet, 2015

VGGNet, 2014



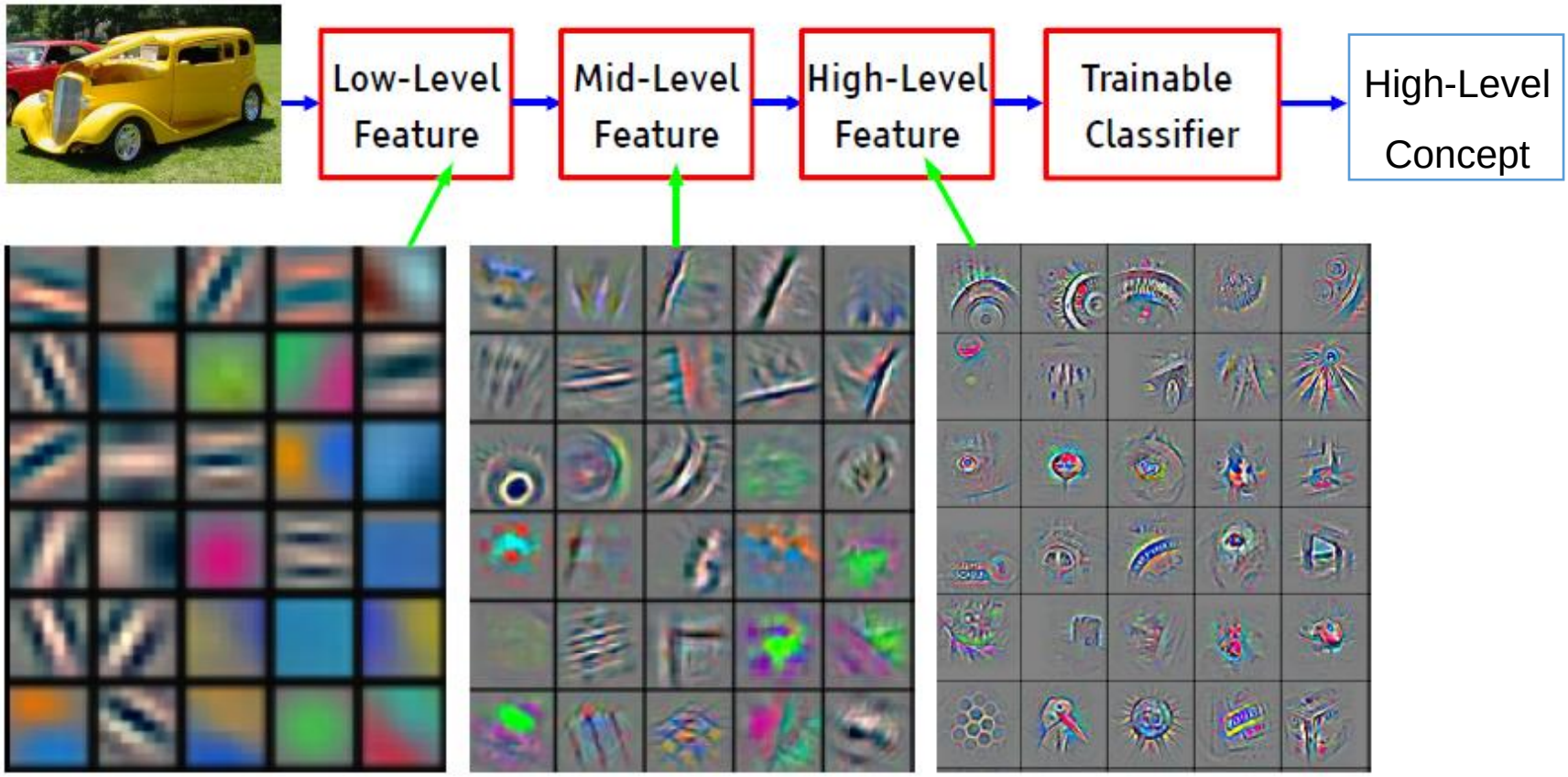
DenseNet, 2017



GoogLeNet, 2014

研究背景：深度学习

- 深度学习本质：层次化的表征学习



[Courtesy of Yann Le Cun]

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

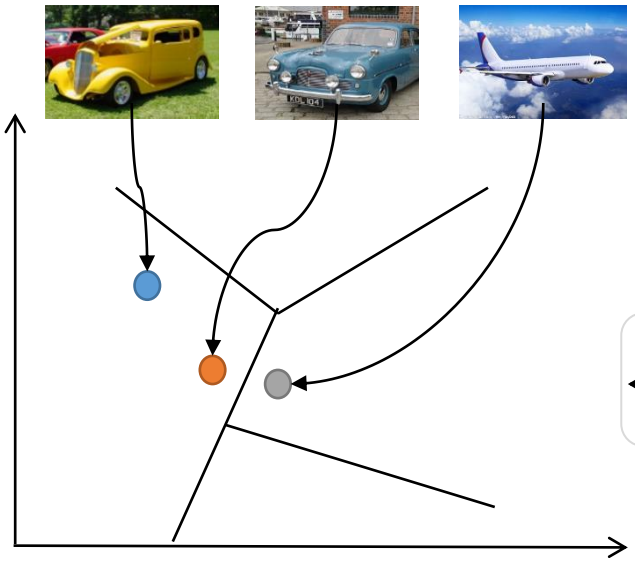
研究背景：深度学习

- 图像分类 vs. 图像检索

Low-Level Feature → Mid-Level Feature → High-Level Feature → Trainable Classifier → High-Level Concept

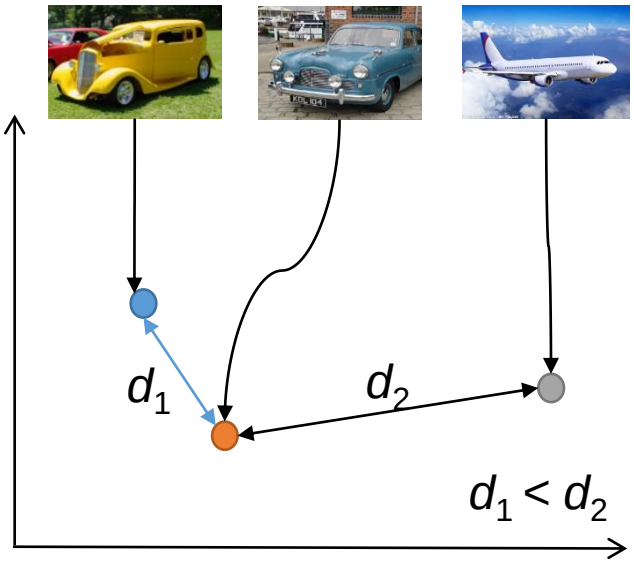
Feature Extraction

Classification



已知 图像类别 未定义

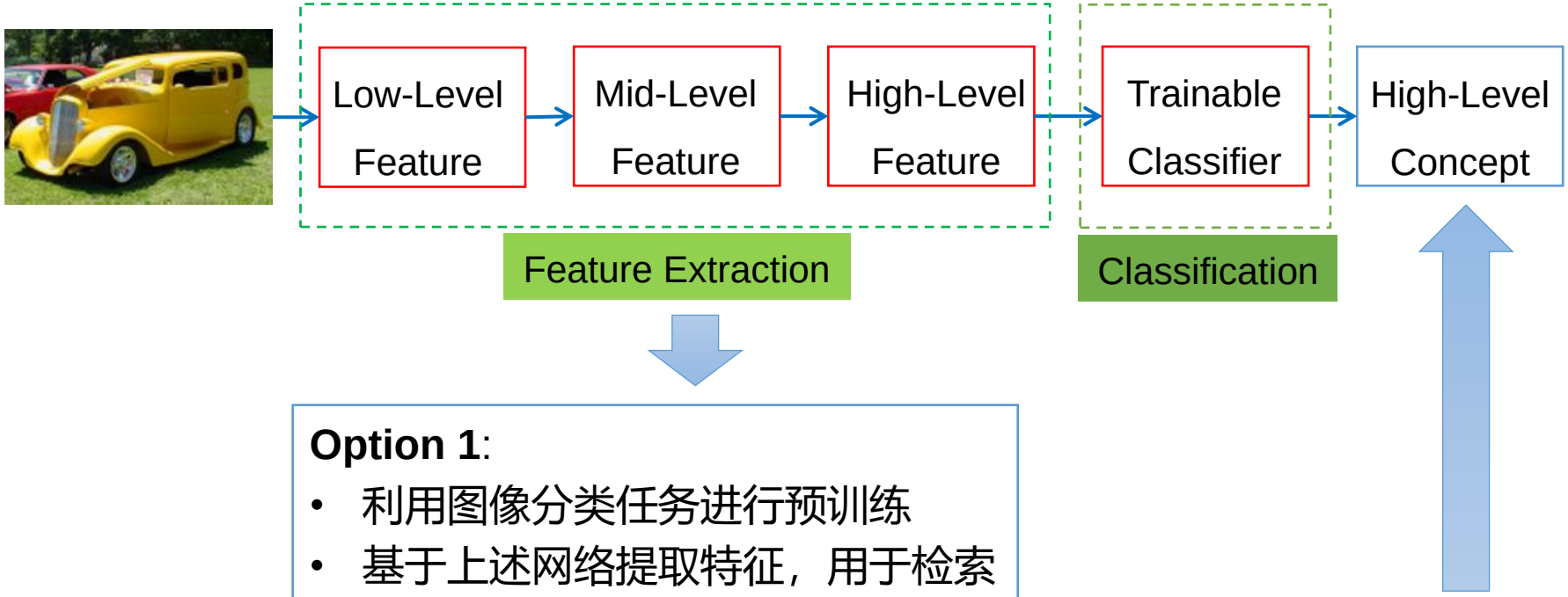
图像分类：特征空间划分



图像检索：相关度度量

如何将深度学习用于图像检索?

- 关键：大规模的标注的训练数据



Option 1:
- 利用图像分类任务进行预训练
- 基于上述网络提取特征，用于检索

Option 2:
- 面向检索任务，设计监督信息
- fine-tune预训练好的CNN模型

卷积神经网路（Convolutional Neural Network）

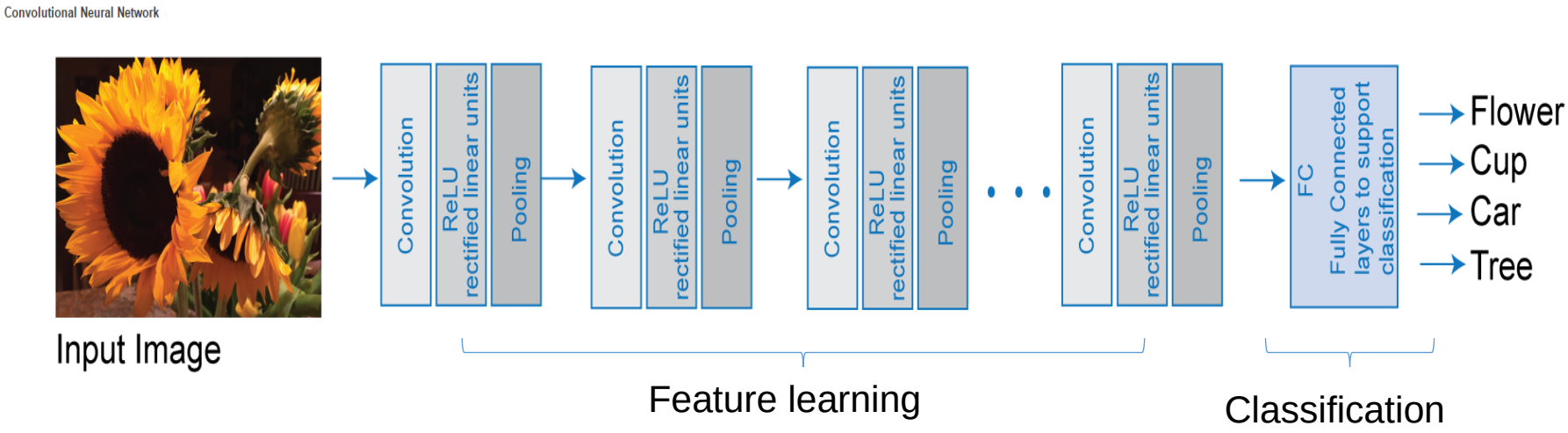
使用 CNN 实现深度学习变得越来越流行的三大因素：

- CNN 消除了手动提取特征的需要 — CNN 直接学习特征。
- CNN 可以产生先进的识别结果。
- CNN 可以重新训练以完成新的识别任务，能够在现有性能优秀的神经网络模型基础上构建自己的模型。

CNN工作原理

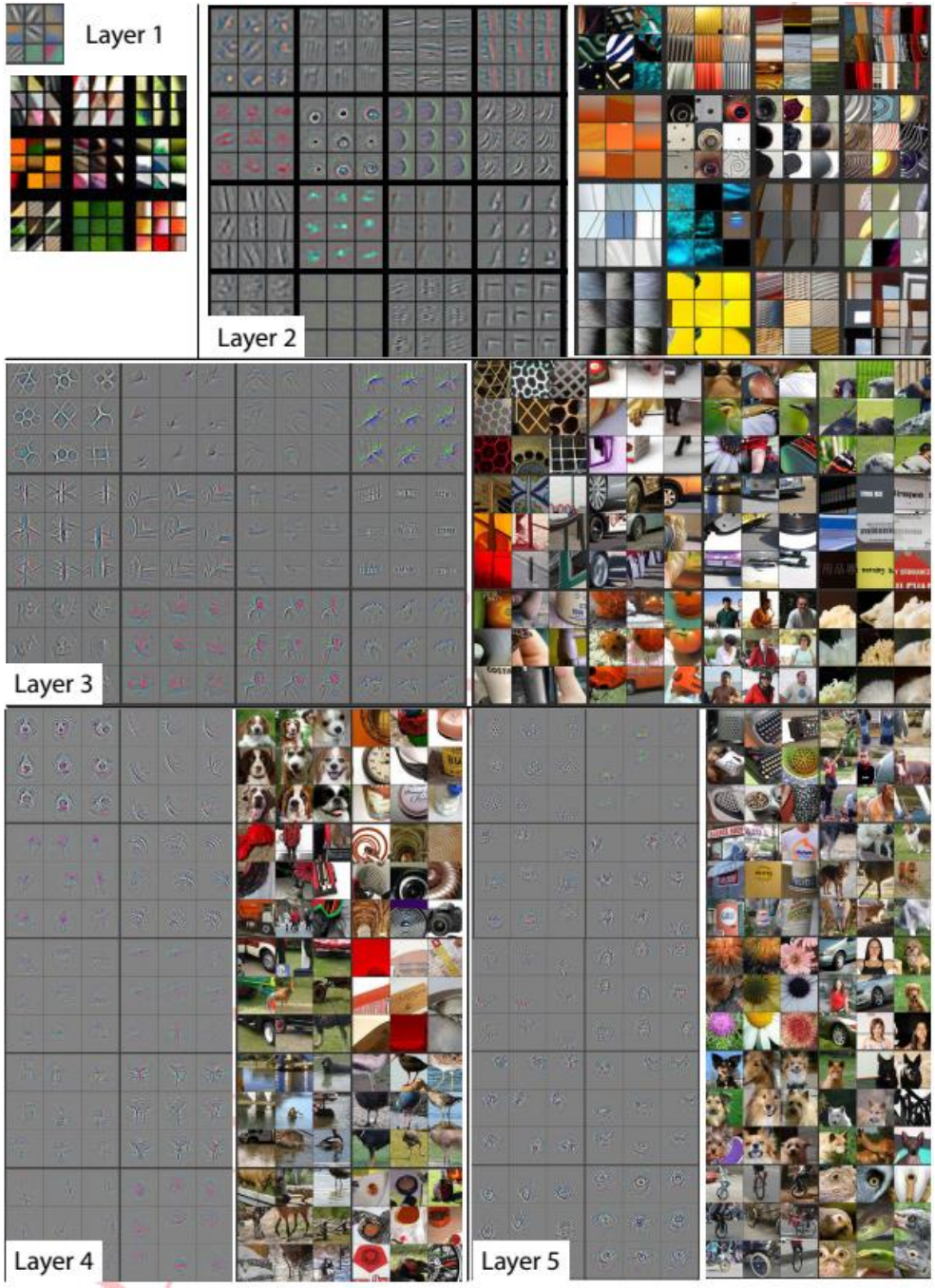
- 卷积神经网络可能有数十个甚至数百个层，每个层学习检测图像的不同特征。卷积层滤波器会应用到不同分辨率的各个训练图像，且每个卷积图像的输出会用作下一层的输入。滤波器最初可以是非常简单的特征，例如亮度和边缘，然后增加复杂度，直至可以唯一地确定目标特征。
- CNN 进行图像、文本、声音和视频的特征识别和分类。

CNN Framework



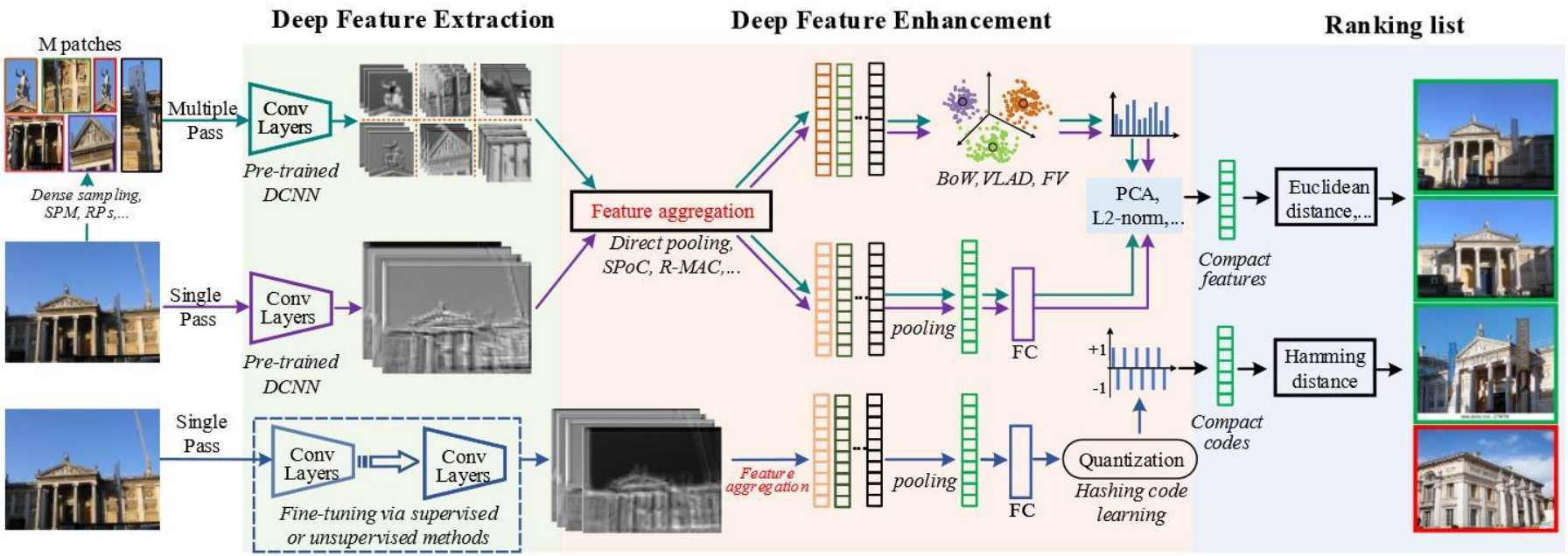
CNN Framework

- CNN最常见的三个层是：卷积、激活（如 ReLU） 以及池化。这些层执行可修改数据的操作，旨在学习特定于数据的特征。
- 卷积将输入图像放进一组卷积滤波器，每个滤波器激活图像中的某些特征。
- ReLU层通过将负值映射到零和保持正数值，实现更快、更高效的训练。有时候这称为激活，因为只有激活的特征才能转移到下一层。
- 池化通过执行非线性下采样，减少网络需要学习的参数个数，从而简化输出。
- 这些操作在几十层甚至几百层上反复进行，每一层都学习识别不同的特征。

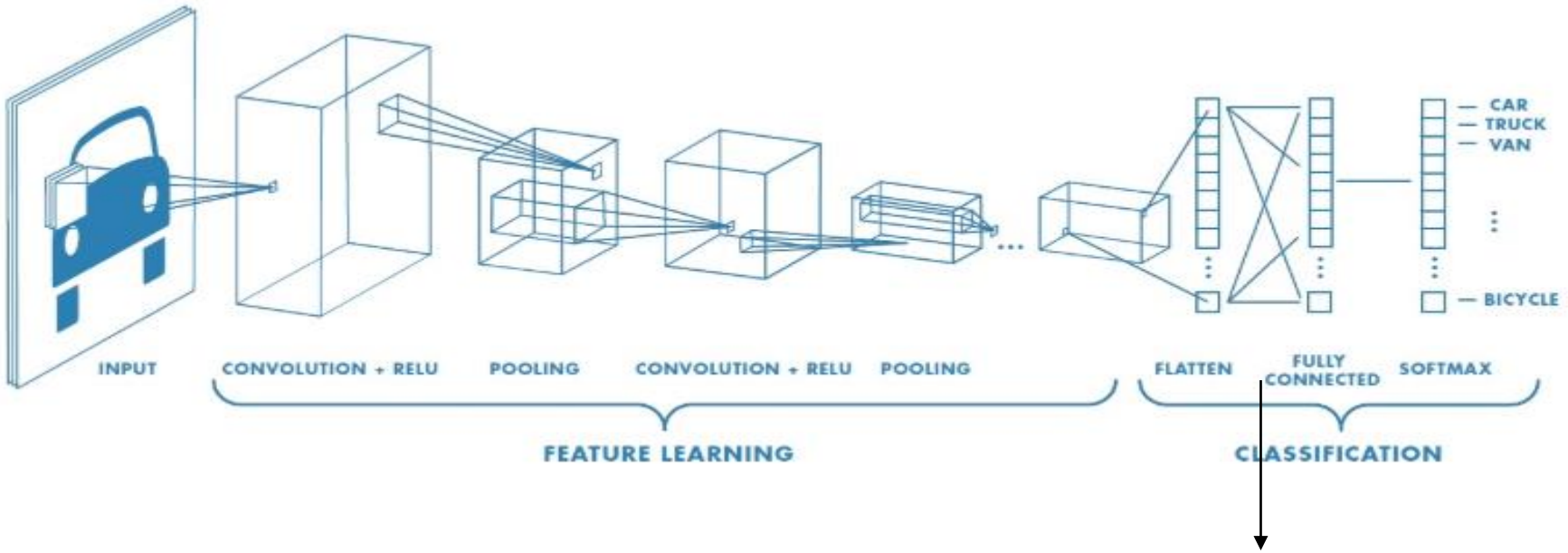


卷积神经网络深度特征的层次特性

CNN-based Pipeline of Image Retrieval



Deep Features



Deep Features

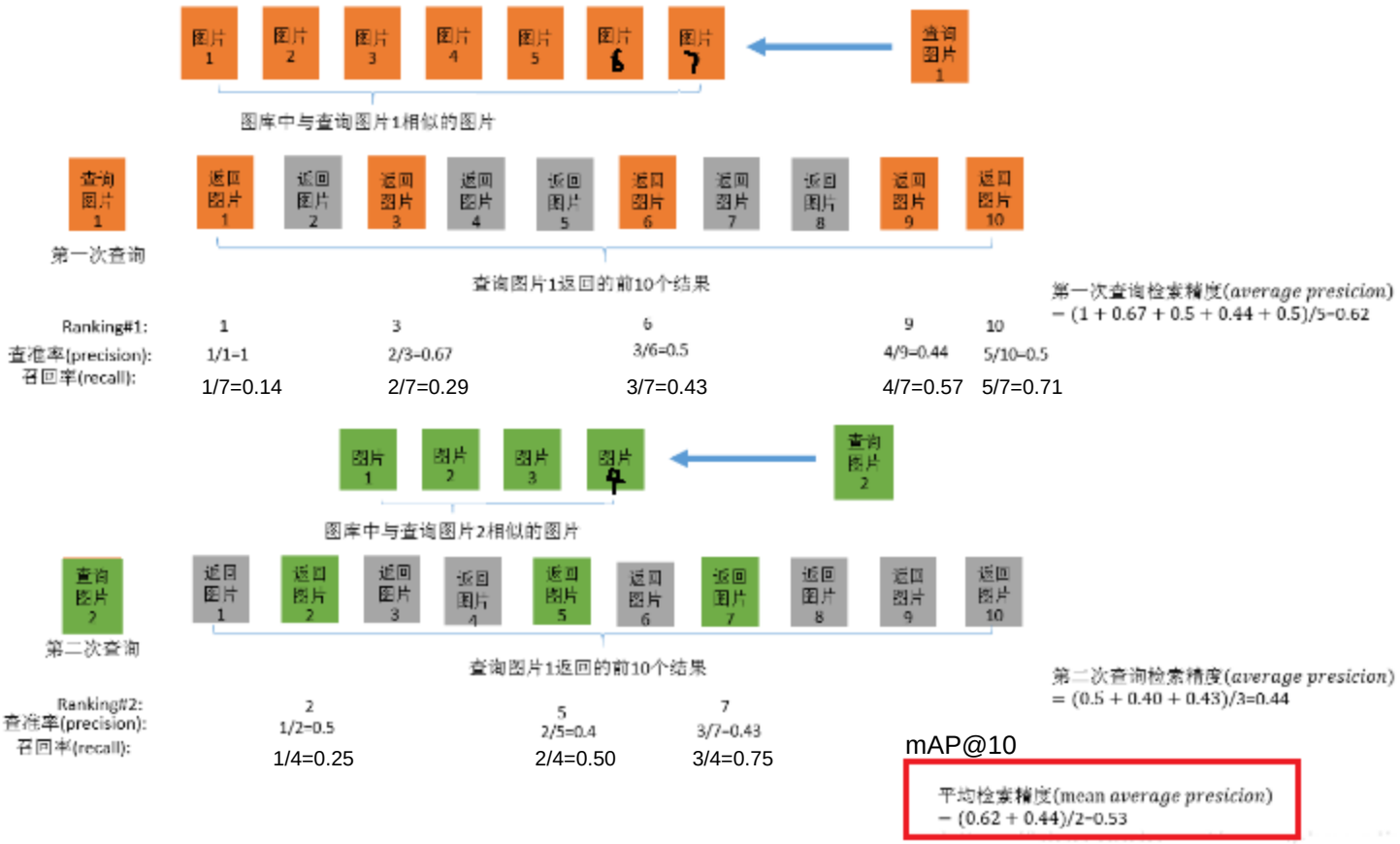
评价指标

假设返回K个检索结果，其中C个是相似（相关）的图像，检索数据集实际有N个相似图像。

- 准确率（Precision）：Precision@K = C / K
- 召回率（Recall，又叫查全率）：Recall@K = C / N
- 平均检索精度（Average Precision，AP）代表不同召回率下准确率的平均值，**是一种同时考虑召回率和准确率的评价指标**
- $AP = \frac{1}{C}\sum_{k \in S} P@k$, 其中$C$为相关的图像数目，$k$为相似的图像所在的排序位置，S为所有相似图像排序位置的集合
- 平均检索精度均值（mAP）是查询多张图像的检索结果AP的平均值。如Q张查询图像的检索精度AP分别为AP_1,AP_2,...,AP_Q，那么平均检索精度均值$\mathrm{mAP} = \frac{1}{\mathrm{Q}}\sum_{\mathrm{i}}^{\mathrm{Q}} \mathrm{AP}_\mathrm{i}$
- 平均检索精度均值（mean Average Precision，mAP）在利用mAP的评估的时候，需要知道：

  1. 每个Query有多少个相似的图像，计算AP时需要考虑所有相似的图像;
  2. 排序结果中这些相似图像的位置；
  3. 相似的定义（一般事先会定义）。

评价指标


图片1
图片2
图片3
图片4
图片5
图片6
图片7
查询图片1
图库中与查询图片1相似的图片
查询图片1
返回图片1
返回图片2
返回图片3
返回图片4
返回图片5
返回图片6
返回图片7
返回图片8
返回图片9
返回图片10
第一次查询
查询图片1返回的前10个结果
第一次查询检索精度(average presicion)
= (1 + 0.67 + 0.5 + 0.44 + 0.5)/5=0.62
Ranking#1:
查准率(precision):
召回率(recall):
1
1/1=1
1/7=0.14
3
2/3=0.67
2/7=0.29
6
3/6=0.5
3/7=0.43
9
4/9=0.44
4/7=0.57
10
5/10=0.5
5/7=0.71
图片1
图片2
图片3
图片4
查询图片2
图库中与查询图片2相似的图片
查询图片2
返回图片1
返回图片2
返回图片3
返回图片4
返回图片5
返回图片6
返回图片7
返回图片8
返回图片9
返回图片10
第二次查询
查询图片1返回的前10个结果
第二次查询检索精度(average presicion)
= (0.5 + 0.40 + 0.43)/3=0.44
Ranking#2:
查准率(precision):
召回率(recall):
2
1/2=0.5
1/4=0.25
5
2/5=0.4
2/4=0.50
7
3/7=0.43
3/4=0.75
mAP@10
平均检索精度(mean average presicion)
= (0.62 + 0.44)/2=0.53

Dataset

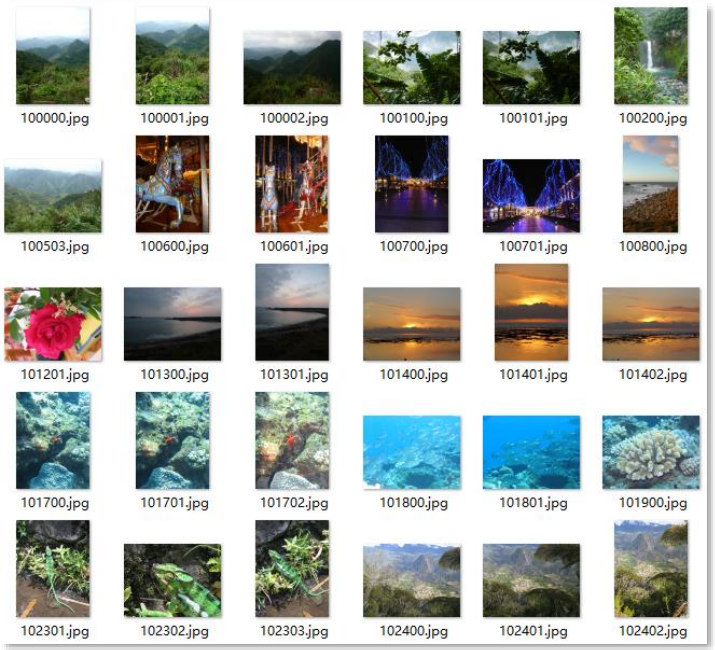
Holidays dataset There are 1,491 images from 500 groups in the Holidays dataset10. Images in each group are taken on a scene or an object with various viewpoints. The first image in each group is selected as query for evaluation.

URL: https://lear.inrialpes.fr/~jegou/data.php.html

Oxford Building dataset (Oxford-5K) The Oxford Buildings Dataset consists of 5062 images collected from Flickr by searching for particular Oxford landmarks. The collection has been manually annotated to generate a comprehensive ground truth for 11 different landmarks, each represented by 5 possible queries. This gives a set of 55 queries over which an object retrieval system can be evaluated. Some junk images are mixed in it as distractor.

URL: http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/

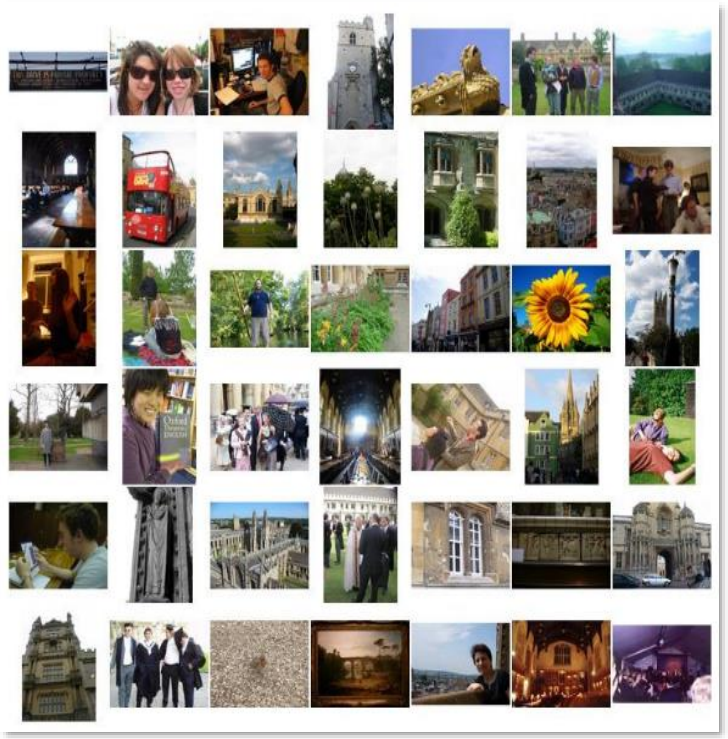
Holidays Dataset



- Holidays数据集包含1491张高分辨率图像。这些图像大多是从假期旅行照片中选取的，涵盖了各种场景和物体。
- 数据集中有500个图像组。每一组都代表一个不同的场景。每组的第一张图像是查询图像，正确的检索结果是该组的其他图像。

[3] Jegou H, Douze M, Schmid C. Hamming embedding and weak geometric consistency for large scale image search[C]//European conference on computer vision. Springer, Berlin, Heidelberg, 2008: 304-317.

Oxford 5K Dataset



- **Oxford 5k** 数据集是一个广泛用于图像检索和视觉识别研究的基准数据集。它由牛津大学的Visual Geometry Group（VGG）发布，主要用于评估图像检索算法的性能。
- Oxford 5k 数据集包含5062张高分辨率(1024*768)图像。这些图像主要拍摄于牛津市的各种标志性建筑和景点。
- 数据集中包含11个不同的建筑物或景点，每个景点有5个查询图像，总计55个查询图像。
- 对每个查询图像，提供了一些相关的检索结果的地面真值（ground truth），这些结果标记为相关（relevant）、不相关（non-relevant）或干扰（junk）。

Good - A nice, clear picture of the object/building.
OK - More than 25% of the object is clearly visible.
Bad - The object is not present.
Junk - Less than 25% of the object is visible, or there are very high levels of occlusion or distortion.

[1] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//2007 IEEE conference on computer vision and pattern recognition. IEEE, 2007: 1-8.

Dataset

NUS-WIDE dataset The dataset includes: (1) 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags; (2) six types of low-level features extracted from these images, including 64-D color histogram, 144-D color correlogram, 73-D edge direction histogram, 128-D wavelet texture, 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions; and (3) ground-truth for 81 concepts that can be used for evaluation. Based on this dataset, we identify several research issues on web image annotation and retrieval. We also provide the baseline results for web image annotation by learning from the tags using the traditional k-NN algorithm. The benchmark results show that it is possible to learn models from these data to help general image retrieval.

URL: https://lms.comp.nus.edu.sg/wp- content/uploads/2019/research/nuswide/NUS-WIDE.html

Google Landmarks Dataset v2 （Large-Scale Benchmark for Instance-Level Recognition and Retrieval） This is the second version of the Google Landmarks dataset (GLDv2), which contains images annotated with labels representing human-made and natural landmarks. The dataset can be used for landmark recognition and retrieval experiments. This version of the dataset contains approximately 5 million images, split into 3 sets of images: train, index and test.

URL: https://paperswithcode.com/dataset/google-landmarks-dataset-v2