基于多层 CNN 特征的图像检索技术

邵杰

贺珂珂

钱学林

1. 研究动机

随着带有摄像头的移动设备的普及，图像数据与日剧增，逐渐的，人们对于信息的检索不再局限于文字，更希望通过输入图片的方式直观地检索到目标信息。而图像检索的一个分支，同款服饰图像检索更随着移动电商的发展，在业界展开蓬勃的研究。在以往的图像检索算法中，基于传统特征的 SIFT，和各种 SIFT 的变体起着主体的作用。

自 2012 年起，深度学习在图像、语音、NLP 等领域的工业界取得了巨大成功。而其中的 CNN（Convolutional Neural Network，卷积神经网络）方法在更是计算机视觉的多项任务中（图像分类、目标检测、语义分割等）都取得了惊人的效果。于是，我们不免产生了这种思考，用深度学习中的特征进行图像检索任务的效果会怎样。

在图像检索中，传统的特征大多使用人为定义的特征, 一般基于文本、颜色、形状和纹理的单一或者混合特征取技术, 这种传统的人工特征简单直观, 但是因为依赖于人的逻辑而存在一定的缺陷。而基于 CNN 的特征取算法通过深层卷积神经网络逐层深化有效特征的复杂度, 用高层特征来获取图像更深层的隐藏信息, 从而获得比人工取特征更复杂有效且维数更低的特征, 能够更好地完成对图像信息的有效表征。而对于基于内容的图像检索（CBIR）任务，底层局部特征非常重要，基于 CNN 的高层特征往往包含较高级别的语义信息，局部细节信息比较有限，反而传统特征有着比 CNN 特征更好的效果。

因此，我们想要利用对 CNN 的特征进行全面的实验，重点考察 CNN 特征中的中间层特征，并与传统特征的检索效果进行对比。同时通过对比 CNN 各层（高层、中间层、跨层等）特征（或特征编码）用于图像检索的效果来，对 CNN 的特征有更加深入的理解。我们在 Oxford5k 数据集[2]、Holidays 数据集[8]和 Paris 数据集进行了验证。

2. 图像检索的一般框架

如图 1 所示为图像检索的一般框架，其中图像表示（Image Representation）是图像检索的核心环节，按照特征提取方法的不同可以分为基于传统特征（非 CNN）的图像表示和基于 CNN 特征的图像表示。

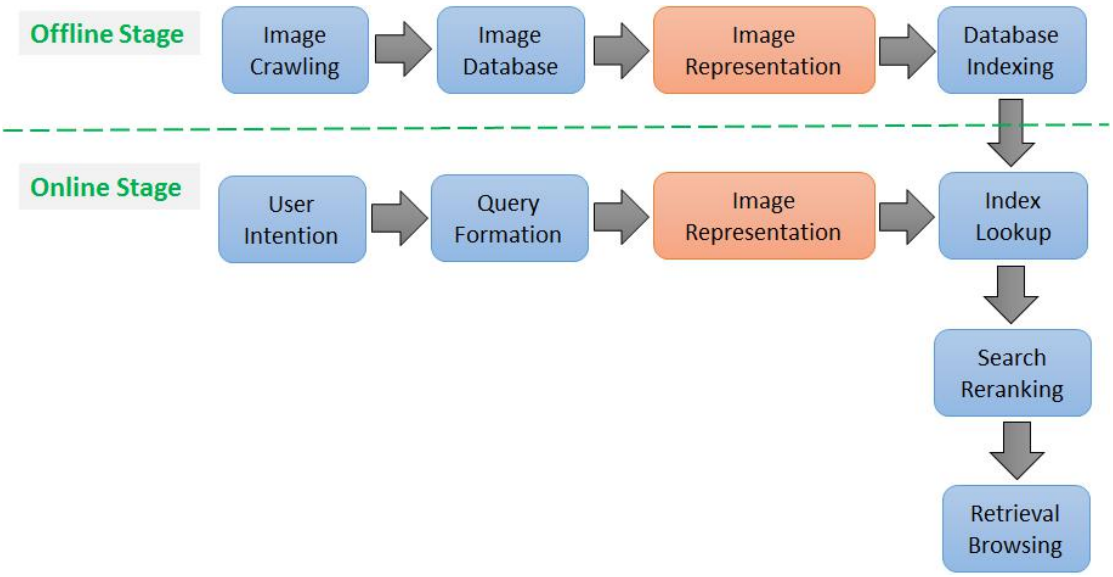

图 1 图像检索的一般框架

3. 基于传统特征（非 CNN）的图像表示

按照图像表示范围的不同，可以分为局部图像表示和全局图像表示。

3.1. 局部图像表示与匹配

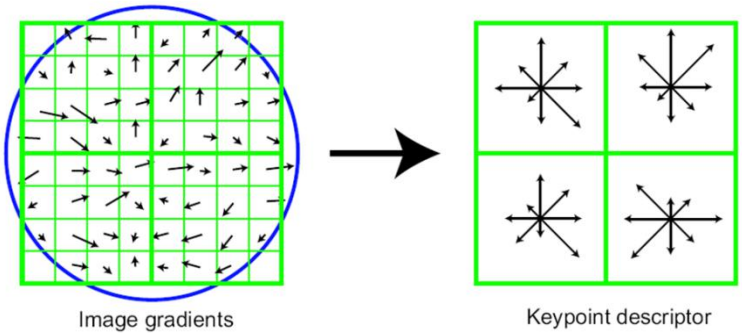

图 2 基于 SIFT 特征的局部图像表示

在图像检索中最常用的局部特征是 SIFT 特征[1]，SIFT 算子是基于尺度不变性和旋转不变性，在图像中检测出若干个关键点，然后提取关键点处的局部“图像块（Patch）”特征（128 维）。

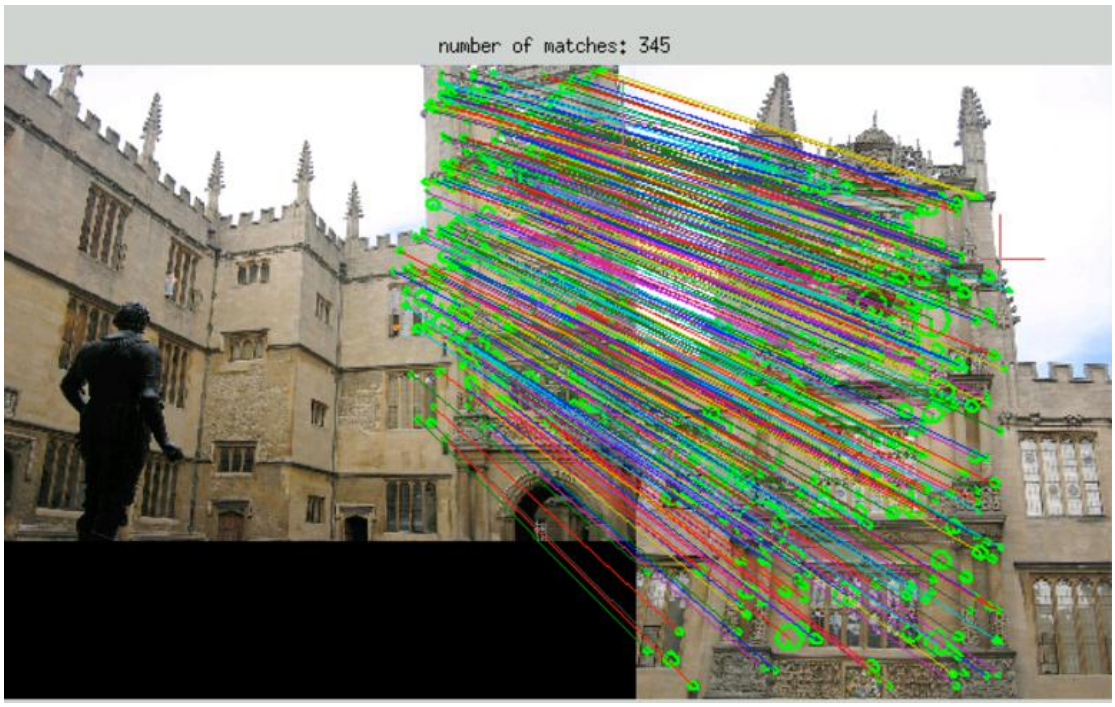

图 3 基于 SIFT 特征的局部匹配效果

3.2. 全局图像表示与匹配

图像检索的目的是在图像数据库中找到与待检索图像最为相近的匹配，上文直接利用了传统的局部特征（例如 SIFT），只能对局部图像进行表示和距离计算，为了能够进一步得到两张图像之间的相似度，还需要根据所提取的局部特征对整张图像进行全局表示。通常采用特征编码的方法，将多个局部特征矢量聚合成一个统一维度的矢量表示：BOW[4]、VLAD[5]、Fisher Vector[6]等，如图 4 所示：

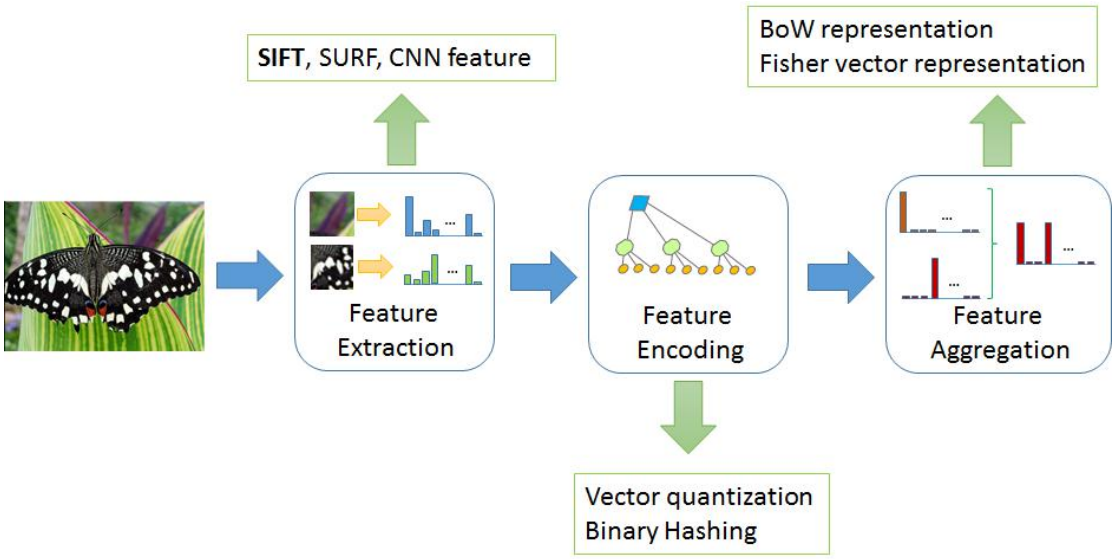

图 4 图像的全局表示框架示意图

上述三种特征编码方法在不同数据集上的效果也各不相同，文献[5]和文献[7]提供的检索结果对比如表 2 所示：

表 2 不同全局图像表示方法的检索结果对比（mAP 值）

	Oxford5k[2]	Holidays[8]
BOW	36.4[5]	54.0[7]
VLAD	55.5[7]	64.6[7]
Fisher Vector	41.8[5]	62.6[5]

在实验中，主要测试了 SIFT+BOW 算法在 Oxford5k 数据集的检索结果。每张图片提取 1976 个特征，5063 张图片共提取约 1000 万个特征，码本大小是 50 万，使得训练的特征总数是码本大小的 20 倍。

实验测试了加入几何验证前后的检索对比结果，如表 3 所示，sift+bow 表示未加入几何验证，sift+bow+gv 表示加入几何验证。

表 3 sift+bow 在 Oxford5k 数据集的实验结果

实验项目	实验结果（mAP）
sift+bow	74.82
sift+bow+gv	83.35

3.3. 问题分析与讨论

通过对比表 1 和表 3 之间的实验结果，可以发现 BOW 算法损失了局部特征的空间信息，因此检索结果出现明显下降（81.79 变为 74.82），然而增加了几何验证后，由于融入空间信息，可以得到好的检索结果（83.35）。

4. 基于 CNN 特征的图像表示

传统的特征大多使用人为定义的特征，一般基于文本、颜色、形状和纹理的单一或者混合特征提取技术，这种传统的人工特征简单直观，但是因为依赖于人的逻辑而存在一定的缺陷，对复杂图像的表达与非刚性物体的泛化能力一般。而基于 CNN 的特征提取算法通过深层卷积神经网络逐层深化有效特征的复杂度，用高层特征来获取图像更深层的隐藏信息，从而获得比人工提取特征更复杂有效且维数更低的特征，能够更好地完成对图像信息的有效表征。

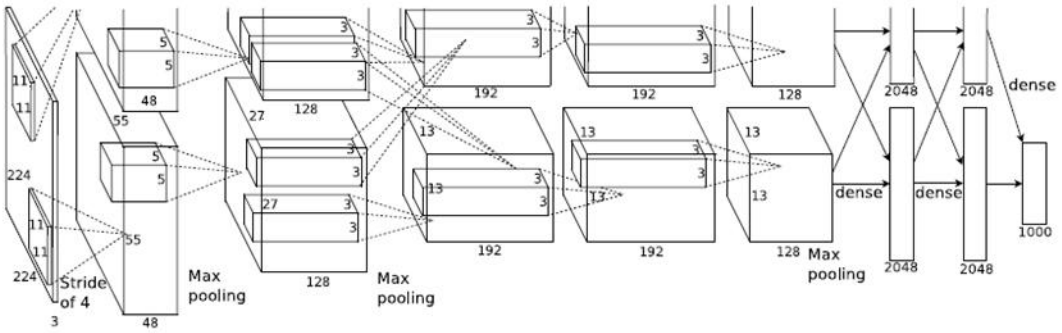

图 5 基于 CNN 特征的图像表示

如图 5 所示，CNN 网络具有不同抽象程度的多个 Layer，较低层得到的是滤波器输出的简单卷积特征（例如 conv1、conv2 等），而较高层得到的是用于分类的高层语义特征（例如 fc6、fc7 等），同样地，按照图像表示范围的不同，可以分为局部图像表示[9][10]和全局图像表示[11]。

4.1. 局部图像表示

在文献[9]中，如图 6 所示，作者将第个卷积层（记作）的输出分解成个 Location 上的局部特征向量，每个特征的维度大小为，即 feature map 个数：

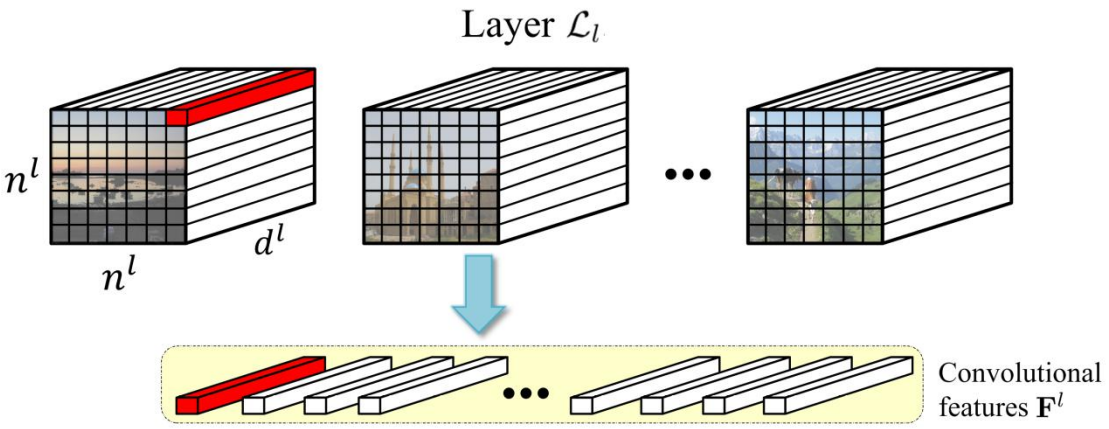

图 6 根据卷积层输出得到局部特征[9]

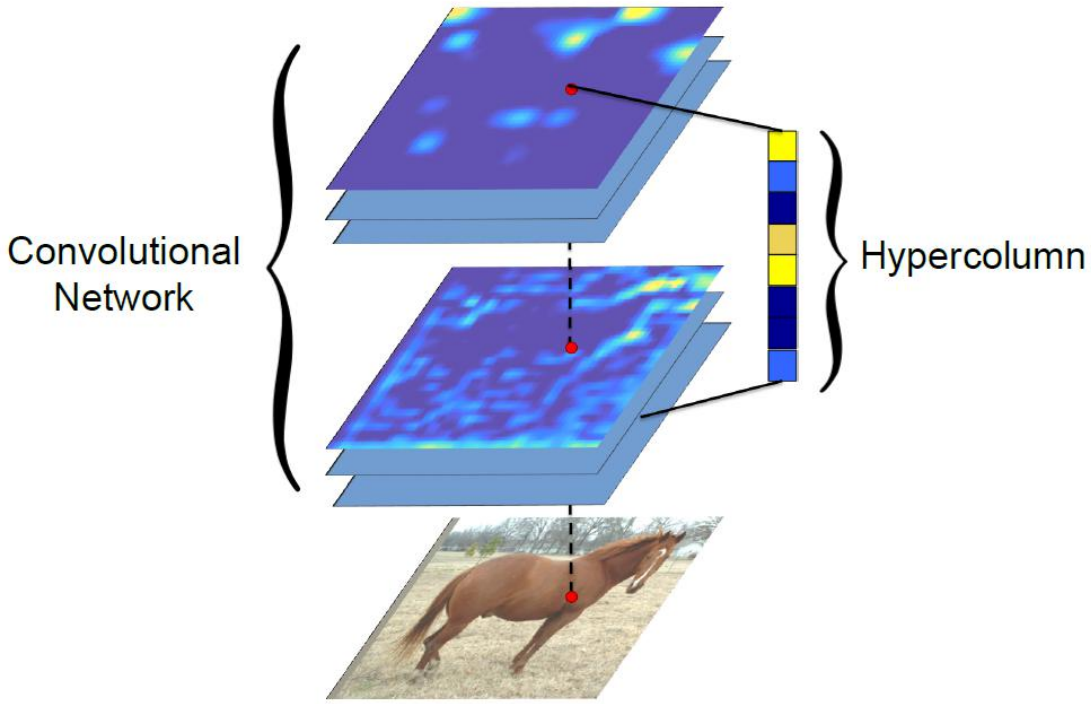

图 7 基于 Hypercolumn 的局部图像表示[10]

由于不同卷积层之间的 feature map 大小可能不同，所以文献[9]中的方法只针对于某一层的 CNN 局部特征。而在文献[10]中，作者提出首先将所有 feature map 缩放至相同大小（例如 50×50），然后将相同 Location 上的特征值纵向连接成一个“超列（hypercolumn）”向量，作为该位置的局部特征，如图 7 所示。

基于 Hypercolumn 的局部图像表示已经被应用于目标分割、目标检测及识别等多个领域，既可以单独利用某一层的 CNN 特征（例如 conv5_3），也可以同时结合多层 CNN 特征（例如 conv5_3 和 conv1_1）。

4.2. 基于特征编码的全局图像表示

4.2.1. 针对单个卷积层的局部特征[9]

局部特征提取参考了文献[9]提出的方法，然后分别采用 BOW、VLAD 和 Fisher Vector 等算法进行编码，暂时未考虑几何验证。实验分别测试了 VGG-Net[12]在 Oxford5k 数据集[2]、Holidays 数据集[8]和 Paris 数据集[13]上的检索结果，如表 4～表 6 所示：

表 4 Holidays 数据集的实验结果与文献[9]结果对比

实验项目	实验结果（mAP）	文献[9]结果（mAP）
conv4_2+bow	78.47	--

conv4_2+vlad	**84.48**	**about 82**
conv4_2+fv	74.39	--
conv5_1+bow	78.63	--
conv5_1+vlad	84.00	about 79
conv5_1+fv	78.62	--
conv5_3+bow	70.54	--
conv5_3+vlad	77.38	about 75
conv5_3+fv	72.13	--

表 5 Oxford5k 数据集的实验结果与文献[9]结果对比

实验项目	实验结果（mAP）	文献[9]结果（mAP）
conv5_1+bow	53.32	--
conv5_1+vlad	**57.03**	**about 55**
conv5_1+fv	49.06	--
conv5_2+bow	52.11	--
conv5_2+vlad	54.28	about 52
conv5_2+fv	49.27	--
conv5_3+bow	41.79	--
conv5_3+vlad	47.34	about 44
conv5_3+fv	41.58	--

表 6 Paris 数据集的实验结果与文献[9]结果对比

实验项目	实验结果（mAP）	文献[9]结果（mAP）
conv5_1+bow	52.65	--
conv5_1+vlad	**64.28**	**about 62**
conv5_1+fv	53.09	--
conv5_2+bow	49.78	--
conv5_2+vlad	60.57	about 61
conv5_2+fv	52.96	--

conv5_3+bow	47.41	--
conv5_3+vlad	54.59	about 55
conv5_3+fv	48.99	--

算法的实现细节如下：

1） BOW 采用 Kmeans 聚类，并且生成 k-d 树用于近似查询（ANN），码本大小选择 50000；

2） VLAD 用 Kmeans 聚类，码本大小选择 100；

3） Fisher 采用 GMM 聚类，码本大小选择 100；

4） 随机生成数（调用 rand 或 randn 函数）的选择可能导致 mAP 值出现 1 个百分点的偏差，因此实验结果与文献[9]的结果之间的偏差是合理的。

4.2.2. 针对多个卷积层的局部特征[10]

局部特征提取参考了文献[10]提出的方法，然后采用 BOW 算法对 hypercolumn 局部特征进行编码，每张图像用一个固定大小的直方图向量来表示。实验测试了 VGG-Net[12]在 Oxford5k 数据集的检索结果。

这里将每一层特征均 resize 成 30×30，5063 张图像共提取 450 万个特征。分别尝试了多种不同的 Layer 排列组合，得到的实验结果如表 7 所示。

表 7 不同 Layer 排列组合的实验结果对比

实验项目	实验结果（mAP）
conv1_1+conv1_2	10.49
conv1_1+conv5_3	27.2
conv5_3	**44.74**

4.3. 基于 FC 层的全局图像表示

参考文献[11]的做法，直接将 FC 层输出的特征向量作为全局图像表示，例如 fc6 层（4096 维）：

算法 3：基于 CNN-fc6 特征的图像检索（不含 ss 和 aug）

for 图像数据库（*Reference Image Set*）中的每一张图像，

提取该图像的 fc6 层特征

end for

构建特征数据库

for 待检索图像（*Query Image Set*）中的每一张图像，

提取该图像的 fc6 层特征

for 图像数据库（*Reference Image Set*）中的每一张图像，

在特征数据库中，通过索引找到所对应的特征

比较和之间的相似度（和距离成反比），作为的分数

end for

按照的值从高到低进行排序，得到的检索结果

end for

Spatial Search 是从原图中提取出多个不同尺度和位置的图像块，然后根据图像块之间的匹配结果计算 Query Image 和 Reference Image 之间的相似性。

算法 4：空间搜索（Spatial Search，SS）

for 待检索图像的每一个子图像块，

for 数据库图像的每一个子图像块，

比较和之间的相似度（和距离成反比）

end for

找到与最相近的图像子块，并将其相似度作为与之间的相似性

end for

计算与之间的相似性

实验测试了 Overfeat-Net[14]、VGG-Net[12]和 Alex-Net[15]在不同数据集下的检索结果，算法的实现细节如下：

1） Spatial Search 操作提取了 6 个 crop（原图、4 个 corner 和 center）和 3 个 rot（旋转 90、180 和 270）；

2） Feature Augmentation 操作是对特征进行后处理：L2 normalization → PCA → Whitening → L2 renormalization，PCA 降维后得到 500 维的特征向量；

3） 实验中分别对比了基本流程（例如 VGG）、引入 Spatial Search（例如 VGG+ss）、引入 Feature Augmentation（例如 VGG+ss+aug）三者的实验

结果。

在 Holidays 数据集[8]上的实验结果如表 8 所示，在 Oxford5k 数据集[2]上的实验结果如表 9 所示，而文献[11]只提供了 OverFeat-Net 的测试结果：

表 8 Holidays 数据集的实验结果与文献[11]结果对比

实验项目	实验结果（mAP）	文献[11]结果（mAP）
Overfeat	64.16	64.2
Overfeat+ss	76.93	76.9
Overfeat+ss+aug	84.33	**84.3**
VGG	72.42	--
VGG+ss	84.86	--
VGG+ss+aug	**87.87**	--
Alex	69.68	--
Alex+ss	81.04	--
Alex+ss+aug	85.53	--

表 9 Oxford5k 数据集的实验结果与文献[11]结果对比

实验项目	实验结果（mAP）	文献[11]结果（mAP）
Overfeat	34.49	32.2
Overfeat+ss	46.41	55.6
Overfeat+ss+aug	57.46	**68.0**
VGG	44.25	--
VGG+ss	56.96	--
VGG+ss+aug	**61.15**	--
Alex	41.38	--
Alex+ss	50.93	--
Alex+ss+aug	53.45	--

4.4. 有选择性的选取局部卷积特征，进行编码

在之前对卷积特征进行编码的时候，所有的卷积特征参与编码，而在实践中发现

卷积特征中有相当一部分都是接近于 0 的，这些特征可以被理解为不显著的。我们进行了实验，去掉这些不显著的特征，只保留关键特征进行编码，是否有助于提高检索结果。

其中的特征筛选方法分为 4 种：

Baseline：从原有局部特征中随机选取 1000 个特征，然后对 1000×N 个特征进行 VLAD 编码，码本大小选择 100，N 表示图片数量；

Filter1：直接求取每个局部特征 norm-1 值，然后按从大到小进行排序，并选取前 1000 个特征；

Filter2：首先对每张 feature map 进行归一化（与最大值的比值），然后按照 Filter2 的方法进行过滤；

Filter3：首先按照 itti 算法求取每张 feature map 的 salience weight，然后分别乘上权值，最后按照 Filter1 的方法进行过滤；

表 10 Holiday 数据集的实验结果

实验项目	Baseline（mAP）	Filter1（mAP）	Filter2（mAP）	Filter3（mAP）
conv2-1	66.82	63.55	64.15	62.99
conv2-2	70.54	69.52	69.78	69.29
conv3-1	79	77.99	78.34	78.59
conv3-2	80.04	80.87	80.27	80.37
conv3-3	80.84	80.34	81.9	82.13

4.5. 问题分析与讨论

1. 对于 Layer 数较多的 CNN 网络，最后一个卷积层（例如 conv5_3）的局部特征由于主要包含与类别相关的高层语义，检索效果不好，反而某些中间层（例如 conv4_2、conv5_1 等）特征更适用于目标检索，另外 CNN 局部特征适合进行特征编码，尤其是 VLAD 编码效果最好；
2. 通过分析表 7 的实验结果，可以发现较低层特征的使用会导致图像检索结果下降；
3. Spatial Search 和 Feature Augmentation 两种方法均可以显著改善 fc 层特

征的检索结果。

4. 对于高层的 CNN 特征，筛选出较为显著的局部特征，进行编码，能够改善检索结果。

5. 总结及展望

在此次实验中，我们将各层的 CNN 特征进行编码，用于图像检索，在 Oxford5k 数据集[2]、Holidays 数据集[8]和 Paris 数据集进行了验证。其中，重点考察 CNN 特征中的中间层特征，并与传统特征的检索效果进行对比，结果显示 CNN 的中间层特征效果略接近于传统特征。同时通过对比 CNN 各层（高层、中间层、跨层等）特征（或特征编码）用于图像检索的效果，结果显示中间层特征明显好于高层特征和底层特征。

我们也发现，CNN 特征结合 Spatial Search 等改进方法后，有较大的提高空间。且对于高层的 CNN 特征，筛选出较为显著的局部特征，进行编码，能够改善检索结果。

如何更好地融合 CNN 的各层特征应用于图像检索，有待于进一步研究。

6. 参考文献

[1] D. G. Lowe, Distinctive image features from scale-invariant keypoints. IJCV, vol. 2, no. 60, pp. 91-110, 2004.

[2] Philbin J, Chum O, Isard M, et al. Object retrieval with large vocabularies and fast spatial matching[C]//Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8.

[3] https://github.com/vedaldi/visualindex

[4] Sivic J, Zisserman A. Video Google: A text retrieval approach to object matching in videos[C]//Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on. IEEE, 2003: 1470-1477.

[5] Jégou H, Perronnin F, Douze M, et al. Aggregating local image descriptors into compact codes[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 2012, 34(9): 1704-1716.

[6] Perronnin F, Dance C. Fisher kernels on visual vocabularies for image categorization[C]//Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on. IEEE, 2007: 1-8.

[7] Arandjelovic R, Zisserman A. All about VLAD[C]//Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on. IEEE, 2013: 1578-1585.

[8] Jégou H, Douze M, Schmid C. Hamming embedding and weak geometry consistency for large scale image search-extended version[J]. 2008.

[9] Ng J Y H, Yang F, Davis L S. Exploiting Local Features from Deep Networks for Image Retrieval[J]. arXiv preprint arXiv:1504.05133, 2015.

[10] Hariharan B, Arbeláez P, Girshick R, et al. Hypercolumns for object segmentation and fine-grained localization[J]. arXiv preprint arXiv:1411.5752, 2014.

[11] Razavian A S, Azizpour H, Sullivan J, et al. CNN features off-the-shelf: an astounding baseline for recognition[C]//Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on. IEEE, 2014: 512-519.

[12] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[13] Philbin J, Chum O, Isard M, et al. Lost in quantization: Improving particular object retrieval in large scale image databases[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.

[14] Sermanet P, Eigen D, Zhang X, et al. Overfeat: Integrated recognition, localization and detection using convolutional networks[J]. arXiv preprint arXiv:1312.6229, 2013.

[15] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[C]//Advances in neural information processing systems. 2012: 1097-1105.

[16] Azizpour H, Razavian A S, Sullivan J, et al. From generic to specific deep representations for visual recognition[J]. arXiv preprint arXiv:1406.5774, 2014.

[17] Razavian A S, Sullivan J, Maki A, et al. Visual instance retrieval with deep convolutional networks[J]. arXiv preprint arXiv:1412.6574, 2014.