词袋模型 BOVW

Preliminaries

这个 abstract 就是介绍一下 Bag of word 模型，你要是清楚的话就不用看了。

Bag of words 模型常用在信息检索中，可以将文档文本这种抽象的数据转换为计算机擅长理解的向量数据。

词袋模型是需要“训练的”，训练数据就是一大堆的文档，数量记为 M。我们分别将这 M 份文档的所有单词抽取出来，组合成一个大的集合，我们称这个集合为这堆文档的“字典”，其中这个字典包含了 N 个单词。这样我们就可以将每份文档表示为 N 维向量如下例所示，比如我们现在有两个文档如下：

文档 1： Bob likes to play basketball, Jim likes too.

文档 2： Bob also likes to play football games.

基于这两个文档，构造一个词典：

Dictionary = {1: “Bob”, 2. “like”, 3. “to”, 4. “play”, 5. “basketball”, 6. “also”, 7. “football”, 8. “games”, 9. “Jim”, 10. “too”}。

我们可以用向量形式表示这两个文档:

1：[1, 2, 1, 1, 1, 0, 0, 0, 1, 1]

2：[1, 1, 1, 1 ,0, 1, 1, 1, 0, 0]

当每份文档都变成 N 维向量后，这些向量数据对于计算机而言就变得非常好处理了。

向量中每个元素表示词典中相关元素在文档中出现的次数，bag of words 模型其实也可以认为是一种统计直方图，可以通过该模型很方便的计算词频。

以上就是 Bag of words 模型，就是将文档集合想办法用向量的形式进行表示，其向量特征就是单词在文档集合中出现的频率。

SIFT 算法

经过 SIFT 算法，我们可以将一副图像映射为一个局部特征向量集； SIFT 特征得到的结果是，对于图像上的每一个兴趣点都得到一个 128 维的特征向量（图上有多少兴趣点，兴趣点的分布都由算法本身决定。）这些向量都具有平移，缩放，旋转不变性，同时对光照变化，仿射能让投影变换也有一定的不变性。

BoVW

先来谈谈什么叫视觉词汇(visual words)，视觉词汇的提出是基于 bag of words 模型的。首先对于数据集中所有的图片提取 sift 特征（就是一堆 128 维的特征向量，注意 SIFT 不是必须的，你也可以用别的算法提取特征。）。然后通过 K-Means 对这一堆特征向量进行聚类，我们聚了 K 个类，有 K 个聚类中心，每个聚类中心都是一个 128 维的特征向量，这 K 个特征向量就是我们的 “视觉单词” ，注意聚类中心才是“视觉单词” 啊。

我们给这 K 个聚类特征向量分别取一个标号，这样就有 K 个标号，比如从 1,2,3....K，每个标号就分别代表了该类簇中心的特征向量，每类的聚类特征向量和它对应的标号就能组成一

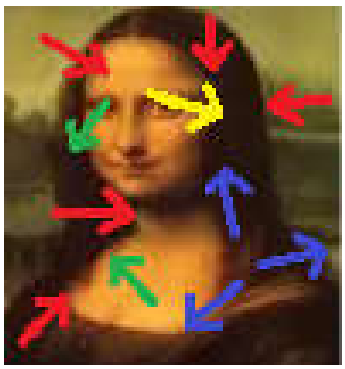
个像 BoW 模型里面提到的 “词典”，例如：

Dictionary: {1: [128 维向量] , 2: [128 维向量], 3 : [128 维向量] , ... , K: [128 维向量] }

这些聚类中心代表了一类具有共同特点的特征。因此，描述一张图上的视觉单词的分布也是描述这张图的方法之一。

BoVW 模型就是提取数据集中的所有图片的 visual words，然后将所有的 visual words 构造一个词典（这就类比于，BoW 模型提取所有文档的词汇组成一个集合，构造词典是一样的道理）。

再构造出词典之后，我们就要基于这个词典（假设我们通过聚类得到 K=4 个聚类中心，即词典中有 4 个视觉词汇，我们这里拿蓝，绿，红，黄作为这 K=4 个聚类向量的标号，如图 B 所示），将每张图片转换为向量。我们现在举例，尝试对图片 A（就这一张图片），将图片 A 基于词典，转换为特征向量。具体做法如下



图片 A

图 B：视觉词典: {蓝：[128 维向量，绿：[128 维向量]，红：[128 维向量]，黄：[128 维向量]}

我们现在对图片 A 使用 SIFT 算法，提取出一堆特征点，每个特征点都是一个 128 维特征向量，如图上的小箭头所示，先忽略小箭头的颜色，假装他们都是黑色的！！。假如我们提取出了 N=11 个特征点（对应图上 11 个箭头），我们对每一个特征点，依次求取这个特征点的特征向量和这 K=4 个聚类特征向量的距离，并找到这个特征点距离哪个聚类特征向量最近，就将这个特征点归为该聚类簇中的那类。

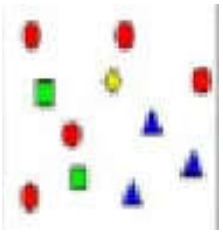


图 C： 将每个特征点根据与 K 个聚类特征向量的距离将其进行分类。

对每个特征点都执行这个分类操作，就相当于统计图片 A 的 K=4 个类特征在图片 A 出现的频次。我直接上可视化图感受一下。

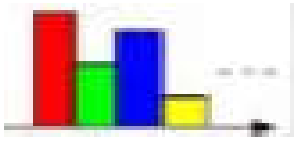


图 D： 统计 K=4 个 “视觉词汇”在图 A 出现频次！

所以最终我们得到的图片 A 的向量为[3, 2, 5, 1] ([蓝,绿,红,黄])

Conclusion

要理解 BoVW 必须要清楚 BoW 模型的细节，在理解过程中注意二者的类比!