《数据科学与工程算法》项目报告

报告题目： 基于PCA的图片压缩
姓　　名： 高宇菲
学　　号： 10215501422
完成日期： 2023.6.7

摘要 [中文]：

主成分分析（PCA）是一种数据约减或降维技术，它有多种变体，如 robust PCA, sparse PCA, 以及 kernel PCA。然而在图片的有损压缩中，传统的主成分分析最为常用。传统主成分分析中的重要步骤是求解特征值和特征向量。本文首先介绍了两种求特征值和特征向量的方法；其次，用两种方法实现了传统的 PCA，即在特征分解时分别使用了 QR 分解和幂法；然后，将传统 PCA 用于图片压缩，统计了压缩时间，压缩率，重构误差等指标，并得出结论；最后，得出了实验和相关探索得出的结论以及本文可以进一步改进的地方。

Abstract [English]

Principal component analysis (PCA) is a data reduction or dimensionality reduction technique that has several variants such as robust PCA, sparse PCA, and kernel PCA.However, the traditional PCA is most commonly used in lossy compression of images. The important steps in traditional PCA are solving for eigenvalues and eigenvectors. In this paper, we firstly introduce two methods to find the eigenvalues and eigenvectors; secondly, we implement the traditional PCA with two methods, i.e., QR decomposition and power method are used in feature decomposition; then, we use the traditional PCA for image compression, and statistics of compression time, compression ratio, reconstruction error and other metrics are obtained, and conclusions are drawn; finally, conclusions from experiments and related explorations are drawn and this paper can be further improvements.

一、项目概述

主成分分析是一种有效的捕捉数据主要信息，舍弃次要信息的降维方法，因此，PCA 在高维数据降维、减小存储空间方面有重要应用。图片是典型的高维数据。本文用不同方法实现了 PCA 算法，完成了对图片的压缩，并统计了压缩时间，压缩率和重构误差。

1.1 PCA 相关研究

传统的 PCA 算法有多个变种，如 robust PCA,主要用于将矩阵分解成一个低秩矩阵和一个稀疏矩阵相加的形式[1]，sparse PCA 使用 lasso 来产生具有稀疏载荷的修改后的主成分，适用于常规的多元数据和基因表达数组[2]，而 kernel PCA 是一种非线性主成分分析方法[3]。然而，传统的 PCA 算法凭借其简单有效的特点，成为最适合图片压缩和重构的方法。

1.2 图片压缩相关研究

图像压缩分为有损压缩和无损压缩，本文主要关注有损压缩。有损压缩的主要方式有预测编码，变换编码，混合编码，分形图像编码等。然而，本文要实现的 PCA 算法与以上编码方式有所不同。PCA 保留能够最大限度解释数据方差的低维空间，而得到的低维空间在人类看来是很难解释的。

1.3 文章组成

第二部分将给出本文要解决的问题的数学定义；第三部分将首先讲解 PCA 的原理，以及介绍三种实现方法以及实现细节；第四部分将给出实现的方法在 256*256 像素大小图片数据集上的表现，包括压缩时间，压缩率，重构误差等指标，以及两张图片压缩前后的对比；第五部分总结并得出结论。

一、 问题定义（提供问题定义的语言描述与数学形式）

本文想要解决的问题如下：

1. 实现 PCA 算法$\mathcal{F}$。

2. 给定 100 张$M \times N$ 的 RGB 格式图片，每一张压缩前的图片记为$X^{(i)} \in \mathcal{R}^{N \times M}, 0 \leq i < 100$，使用 PCA 算法，选择合适的主成分个数$k$，对其进行压缩，压缩后的图片记为$Y^{(i)} \in \mathcal{R}^{N \times M}, 0 \leq i < 100$。即：

$$\mathcal{F}\left(X^{(i)}, k\right) = Y^{(i)}$$

3. 展示任意两张图片压缩前后的图像，及其重构误差MSE、压缩时间$time_spend$，压缩率$compact_rate$。统计 100 张图片的平均重构误差、平均压缩时间、压缩率。

二、 方法（问题解决步骤和实现细节）

3.1 图片预处理

首先将图片大小统一成 256*256 像素。遍历所有图片，检查图片的大小，如果大小不符，则调用 resize 函数调整为目标大小。

3.2 PCA

令$p1, p2, \ldots, pn$为n个正交向量，他们组成方阵P的列向量。$x1, x2, \ldots, xn$为n个样本，他们组成数据矩阵$X \in \mathcal{R}^n$的行向量。假设样本以 0 为均值，1 为方差。

我们希望用新的基底P表示xi：

$$x_i = \alpha_{i1} p_1 + \alpha_{i2} p_2 + \cdots + \alpha_{in} p_n$$

因为我们已经假设 P 是正交基底，我们可以通过以下公式找到α_i：

$$\alpha_{ij} = x_i p_j$$

将α_i记为x_i'，则：

$$x_i' = x_i P$$
$$X' = XP$$

由于X是已经中心化的矩阵，所以$\Sigma = (X^T X)/(n-1)$是协方差矩阵。

转化后数据的协方差矩阵为：

$$X'^T X'/(n-1) = P^T \Sigma P$$

由于Σ是一个对称矩阵，而 P 是正交基底，由特征值分解和正交矩阵的性质可知上面的矩阵是一个对角矩阵。

注意到Σ 是对称矩阵，意味着它有非负的特征值和不相关且正交的特征向量，因此P可以被作为正交基。

为了使得不相关的维度之间方差最大，我们将特征值排序，用前 k 大的特征值表示$k(k < n)$维向量x'：

$$x_i' = \Sigma_{j=1}^{k} x_i p_j$$

由此起到降维效果。

PCA 的速度主要取决于求解特征值的速度。下面使用若干种方法求解特征值。

3.3　QR 分解求解特征值

任何一个非奇异矩阵 A 都可以分解为 A=QR 的形式,其中 Q 为正交矩阵，R 为上三角矩阵。如果要求 R 的所有对角线元素是正的，那么 QR 分解是唯一的。

使用 QR 分解求矩阵特征值的步骤如下：第一步，对矩阵 A 进行 QR 分解,$A = A_1 = Q_1 R_1$，其中 R_1的对角线元素均为正。第二步，计算$A_2 = R_1 Q_1$，第三步，对A_2进行 QR 分解。重复以上步骤 k 次，当 k 足够大时，A_k趋于一个对角元素是 A 的特征值的上三角矩阵。如果 A 是对称正定矩阵，并且特征互不相同，大多数情况下，A_k会收敛到一个对角矩阵，特征值降序排列。而且，这种情况下，我们定义$S_1 = Q_1$，$S_k = S_{k-1} Q_k = Q_1 Q_2 \dots Q_{k-1} Q_k, k > 1$，那么$S_k$将收敛到一个正交矩阵，每一列是 A 的特征向量。

实际上，使用 QR 分解求解矩阵特征值是对幂法的改进。幂法在每一次迭代都将矩阵归一化，但是这样只能得到最大特征值和对应的特征向量。如果在每一次迭代都对矩阵单位正交化，就是 QR 分解，它能一次性得到所有特征值。

3.4　幂法求前 k 大特征值

幂法：

设实矩阵$A = (a_{ij})$有 n 个线性无关的特征向量，且矩阵 A 的特征值为$\lambda_1, \lambda_2, \dots, \lambda_n$，与之对应的特征向量为$p_1, p_2, \dots, p_n$，其中$p_i \in R^n (i = 1,2, \dots, n)$。已知 A 的主特征值$\lambda_1$是实数且不是重根，即满足条件

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \cdots \geq |\lambda_n|$$

其中，λ_1称为矩阵 A 的主特征值。满足以上条件的方阵，可以运用幂法求解主特征值λ_1及其对应的特征向量p_1。

令v_0为非零的初始向量，它可以由正交基$p_1, p_2, \ldots, p_n$,线性表出为

$$v_0 = \alpha_1 p_1 + \alpha_2 p_2 + \cdots + \alpha_n p_n$$

利用矩阵 A 的幂乘构造如下的向量序列：

$$v_1 = Av_0 = \Sigma_{i=1}^{n} \alpha_i \lambda_i p_i,$$
$$v_2 = Av_1 = A^2 v_0 = \Sigma_{i=1}^{n} \alpha_i \lambda_i^2 p_i$$
$$\ldots$$
$$v_{k+1} = Av_k = A^{k+1} v_0 = \Sigma_{i=1}^{n} \alpha_i \lambda_i^{k+1} p_i$$
$$\ldots$$

其中v_k为第 k 轮的迭代向量。因此，

$$v_k = A^k v_0 = \lambda_1^k \left[\alpha_1 p_1 + \Sigma_{i=2}^{n} \alpha_i \left(\frac{\lambda_i}{\lambda_1} \right)^k p_i \right] = \lambda_1^k (\alpha_1 p_1 + \epsilon_k)$$

当 k趋于无穷时，v_k约等于$\alpha_1 \lambda_1^k p_1$。

关于主特征值λ_1的计算，可以计算

$$\lim_{k=\infty} \frac{(v_{k+1})_i}{(v_k)_i} = \lambda_1$$

为避免上溢或下溢，可以对迭代向量做除以模长的规范化处理。

降阶技术：

矩阵 A 标准化的特征向量构成一组标准正交基， 满足

$$v_i^T \ v_j = \begin{cases} 1 \text{ if } i = j \\ 0 \text{ otherwise} \end{cases}$$

通过正交分解可得$A = \Sigma_{i=1}^{n} \lambda_i v_i \, v_i^T$

求得主特征值后，构造矩阵$A^1 = \Sigma_{i=2}^{n} \lambda_i v_i \, v_i^T$,继续对新的矩阵使用幂法，求出第二大的特征值和对应的特征向量。

3.5 实现细节

QR 分解法和幂法 PCA 都用 Python 实现。

使用 QR 分解求特征值时，将每一张图片转化为 765*256 的二维矩阵，对 256*256 的协方差矩阵使用 QR 分解求特征值和特征向量。 取累积贡献率超过 98%的特征值和对应的特征向量作为新的基底。将图片映射的新的基底上。

使用幂法和降阶技术求解特征值时，考虑到幂法的局限性，要求矩阵是对称矩阵且特征值互不相同，所以将每一张图片转为 25536*3 的二维矩阵，对 3*3 的协方差矩阵使用幂法和降阶技术，保留前两个主成分。

两张图片 src, dir 的重构误差使用 MSE 衡量

$$MSE = \frac{\Sigma_i^n \Sigma_j^n \left(src_{ij} - dir_{ij} \right)^2}{n^2}$$

压缩率使用压缩后的大小减少量与压缩前大小的比值来衡量。

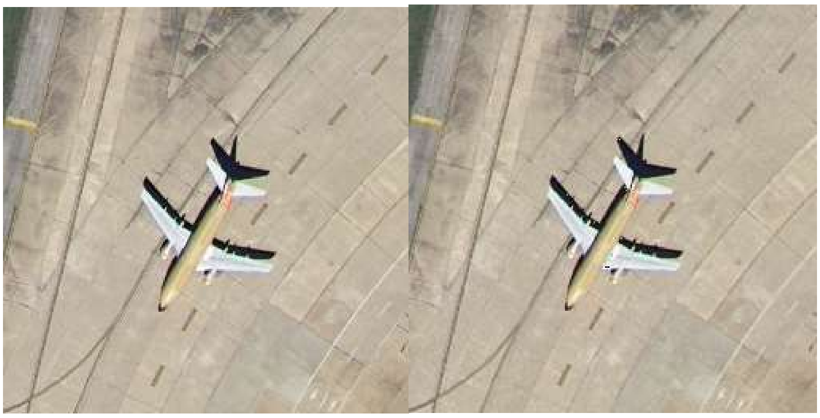
三、 实验结果（验证提出方法的有效性和高效性）

	平均重构误差（MSE）	压缩时间（s）	压缩率
QR 分解	29.7	0.47	5.6%
幂法+降阶技术	93.29	0.013	5.6%

airplane15.tif 压缩前后对比：（QR 分解法）

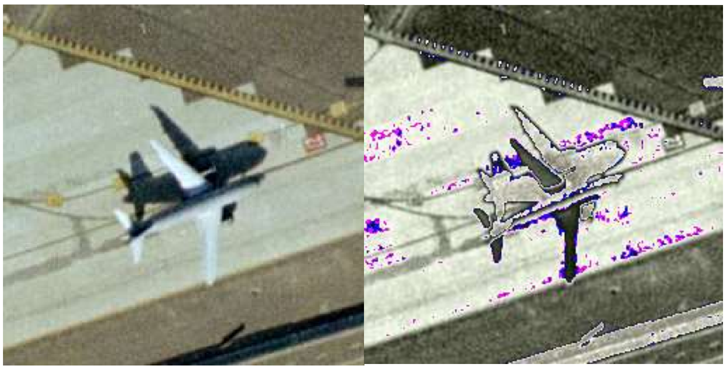
airplane15.tif 压缩后重构误差(MSE):31.9019
airplane15.tif 压缩时间:0.51s
airplane15.tif 压缩率:5.39%

airplane22.tif 压缩前后对比：（QR 分解法）



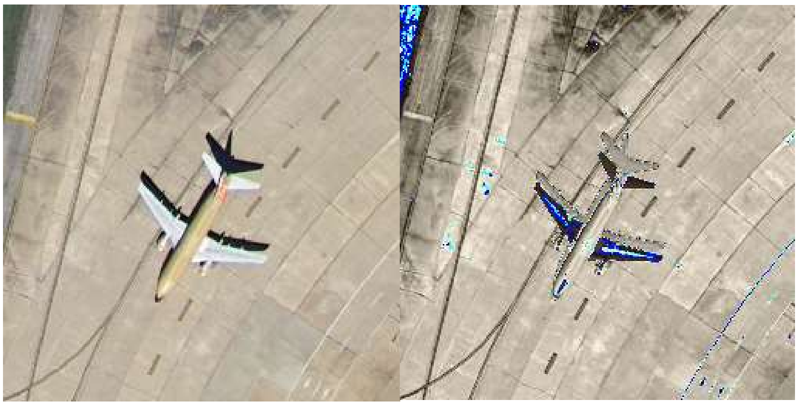
Airplane22.tif 压缩后重构误差(MSE):14.0067
Airplane22.tif 压缩时间:0.33s
Airplane22.tif 压缩率:5.35%

airplane15.tif 压缩前后对比：（幂法+降阶技术）



airplane15.tif 压缩后重构误差(MSE):118.1567
airplane15.tif 压缩时间:0.014s
airplane15.tif 压缩率:5.39%

airplane22.tif 压缩前后对比：（幂法+降阶技术）



airplane22.tif 压缩后重构误差(MSE):81.3814
airplane22.tif 压缩时间:0.011s
airplane22.tif 压缩率:5.35%

五、结论

经过探索和实验，得到以下结论：

1. 幂法求特征值有其适用范围，在图片压缩中幂法局限性明显，压缩效果在重构误差上不如 QR 分解法。

2. kernel pca 和 sparse pca 不适用于图片压缩任务，虽然他们可以将数据降维，但是重构非常复杂。

3. 正交优化[4]也可以用来求解前 k 个特征值，

$$min: F(x) = tr(X^T BX); X \in R^{n\times p}$$
$$s.t. X^T X = I$$

但是由于这篇文章的代码由 MATLAB 写成，在 Python 中调用 MATLAB 函数需要传入函数作为参数，较为复杂，本文没有实现。

4. 经探索，PCA 图片压缩最好使用传统的 PCA，求解特征值特征向量使用 QR 分解，速度也不到一秒，简单实用。

5. 使用传统 PCA 压缩图片，压缩率并不高，只有 5%左右。如果想要更高的压缩率，需要使用其他更有效的算法。

参考文献：
[1]Robust Principal Component Analysis. Emmanuel J. Cand`es
[2]Sparse Principal Component Analysis. Hui ZOU
[3]Kernel Principal Component Analysis. Bernhard SchSlkopf
[4] A Feasible Method for Optimization with Orthogonality Constraints. Zaiwen Wen