# 《数据科学与工程算法基础》实践报告

报告题目： 利用PCA对图像降维

姓 名：

学 号：

完成日期： 12.12

提供中英文摘要

摘要 [中文]：

本次实验，利用PCA技术对100张图片进行重构，采用两种方法对图片压缩。方法一是对所有图片独立压缩，方法二是对同一文件夹的图片一起压缩。本文分析这两种方法的重构效果，重构误差，压缩率，空间节省的异同。比较在重构时是否将中心加上对于重构结果的影响。比较不同精度下重构误差等的变化情况。

Abstract [英语]

In this experiment,I use PCA technology to reconstruct 100 pictures in two ways. The first method is to compress all pictures independently, and the second method is to compress the pictures in the same folder together. This paper analyzes the similarities and differences of the reconstruction effect, reconstruction error, compression rate, and space saving of these two methods. This paper analyzes whether to add the center to the reconstruction result during reconstruction. This paper compares the changes of reconstruction errors under different accuracy.

1. 项目概述（阐明该项目的科学和应用价值，以及相关工作进展并描述项目的主要内容）

主成分分析方法可以有效地找出数据中最主要的元素和结构，去除噪音和冗余，将原有数据降维，揭露出纷繁复杂的数据背后的简单结构。通过主成分分析降维后的数据可以应用于后续的机器学习和数据可视化的输入数据集，进行进一步数据分析工作。

主成分分析通过利用[正交变换](https://zh.wikipedia.org/wiki/%E6%AD%A3%E4%BA%A4%E5%8F%98%E6%8D%A2" \o "正交变换)来对一系列可能相关的变量的观测值进行线性变换，从而投影为一系列线性不相关变量的值，也即主成分上。通过这样的方法减少数据集的维度，同时最大程度保证压缩后数据和原数据的损失最小。

本项目主要利用PCA方法对图像进行降维处理。图像是典型的高维数据，为了避免维度灾难或者是处于节省空间的考量，对图片进行降维处理是很有必要的。

1. 问题描述（问题定义）

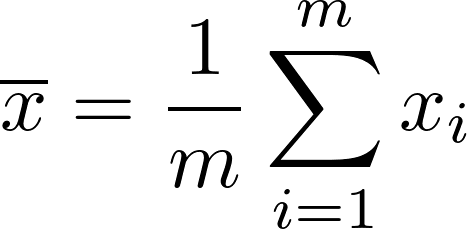
主成分分析（PCA）可用于减少矩阵的维数，在这个项目中我们需要对图像降维并通过投影到新的维度空间中。我们需要重塑所有图像以得到保留原有特征但k权重较小的图像。在给定的数据集中的单个类别中选择100张图像，然后使用PCA压缩图像。

1. 方法（问题解决步骤和实现细节）

我们利用PCA重构2D矩阵。具体解决步骤如下：

1. 对样本进行去中心化。

利用如下公式求解样本中心点。



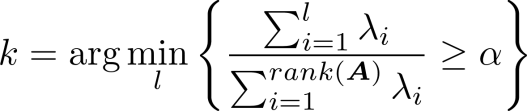
1. 样本协方差矩阵分解

利用python的np.cov计算协方差矩阵。利用np.linalg.eig对得到的协方差矩阵计算特征值和特征向量。

3.选取主成分

对求出协方差矩阵的特征值即奇异值进行升序排序。然后从大大小依次选择奇异值进行累加。

利用如下公式选取主成分个数，取满足精度要求的最小特征值个数。



在代码中，我们通过循环累加，满足条件即break来得到主成分个数。

此处wpsoffice为压缩的精度，用以限定主成分对原始数据的解释能力。在实验中我们通过对比不同wpsoffice取值下的压缩率，重构误差等值分析PCA降维中精度对最后压缩结果的影响。

4.投影到新的空间

计算投影到新空间的空间节省和压缩率。空间节省通过调用python的size函数计算原始矩阵的大小和重构矩阵大小与特征向量大小以及样本中心大小的大小的差值计算。压缩率通过重构矩阵大小与特征向量大小以及样本中心大小的大小的和与原始矩阵大小的比值计算。

1. 重建图像

重构矩阵时，通过先前计算的图片的中心值进行归一化操作来使图像的质量更优。

1. 最后我们导入图片数据集，选择RGB图像，重建RGB分量，堆叠RGB并将数组转化为图片并显示图像。对比压缩后图片与原图片的效果。对于重构后的图像分RGB三个通道分别计算样本集100张图片的重构误差均值。

本次项目采用两种压缩方式进行对比，两种方式的主要区别在于对图像的压缩。方法1：对所有100张图像的每一张单独压缩，并汇总，最后计算得到重构误差、压缩率等的均值。方法2:对所有100张图像一起压缩，并计算重构误差，压缩率等。在方法2中，我们将一个文件夹下的所有图像的矩阵reshape成一维矩阵然后按行拼成一个大的矩阵，对大矩阵求中心，对去中心化后的图片数据求协方差矩阵的特征值和特征向量，按压缩精度选取主元，最后再将得到的重构矩阵按行分割，得到经过PCA降维后的图像。由于发现文件夹中部分图像尺寸不一致，我们统一降维为(65536,1)的图像，并对尺寸不符的图像利用pad函数进行0填充。

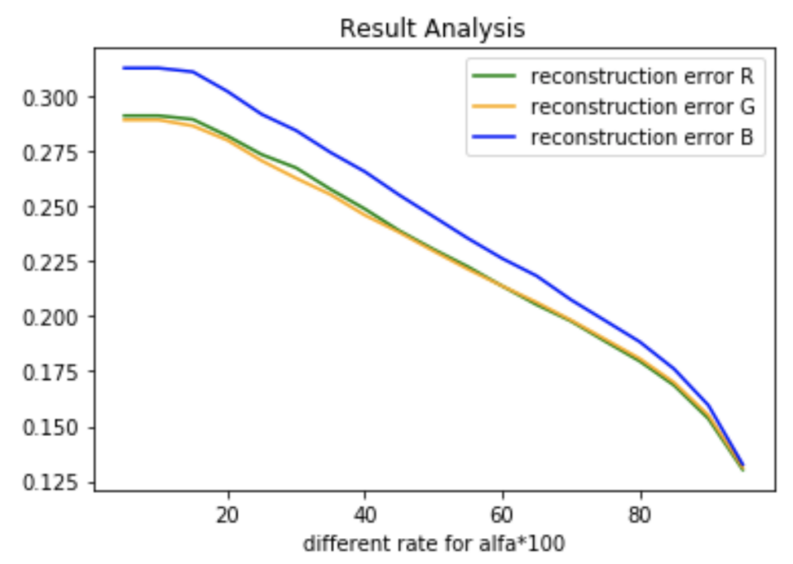
1. 实验结果（验证提出方法的有效性和高效性）

在方法1中，比较不同特征值精度要求得到的重构误差、压缩率、节省空间大小。其中重构误差通过原矩阵和重构后的矩阵的差的F范数与原矩阵F范数的比值来度量。压缩率为压缩后特征向量的大小，中心化向量的大小以及新矩阵的大小的和与原图像矩阵大小的比值。节省空间为原图像矩阵大小减去压缩后特征向量的大小，中心化向量的大小以及新矩阵的大小的和。

具体结果如下：

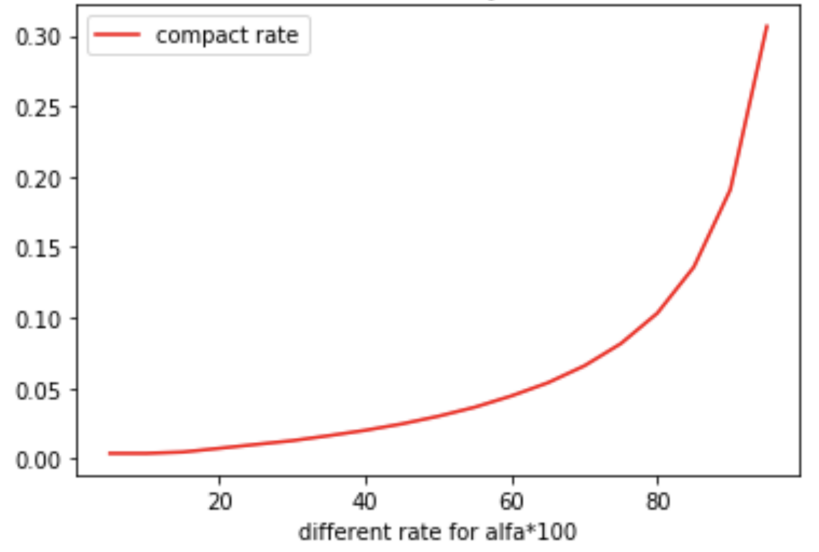
1.计算重构矩阵时加上原先的中心

1）比较不同压缩精度下重构误差的变化情况和趋势



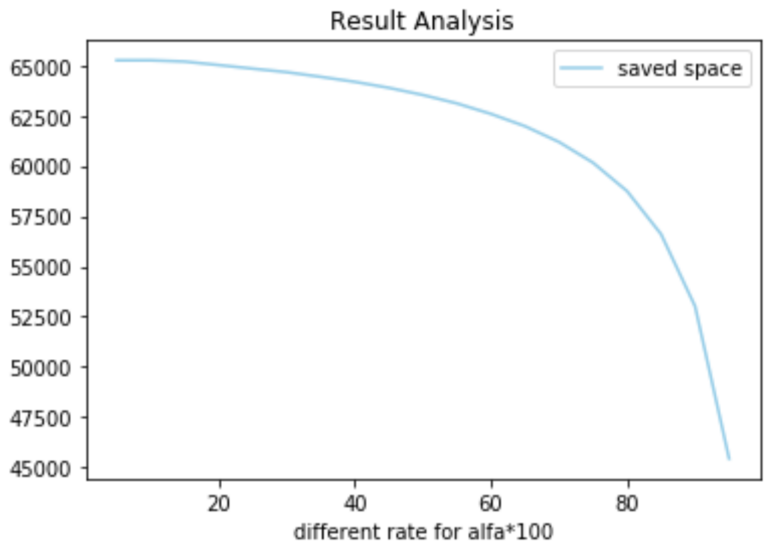
如上图可以看到，随着压缩精度从0.05变化到0.95，重构误差逐渐减小，从0.3减小到0.13。比较不同通道可以发现，B通道的重构误差总体上略高于另两个通道，R和G通道的变化情况较一致。

2）比较不同压缩精度下压缩率的变化情况和趋势



如上图可以看到，随着压缩精度的增加，压缩率逐渐增大。压缩精度从0.05变化到0.95的过程中，压缩率由0.01变化到0.3。此外可以看到压缩精度在0.05到0.45的范围中，压缩率增幅较缓，随机压缩精度的增大，压缩率增速加快，在压缩精度处于0.8到0.95之间时，压缩率猛增，从0.1增至0.3。

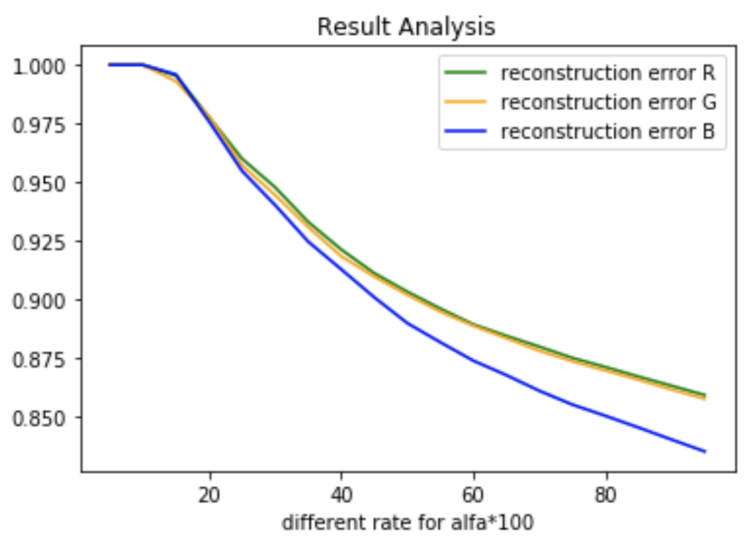
1. 比较不同压缩精度下节省空间的变化情况和趋势



如上图所示，节省空间随着压缩精度的增加而减小。压缩精度从0.05变化到0.95的过程中，节省空间从65257减小到45000。随着压缩精度的增加，节省空间减小的速度增大，在压缩精度在0.8到0.95之间时节省空间减小速度陡然增大。

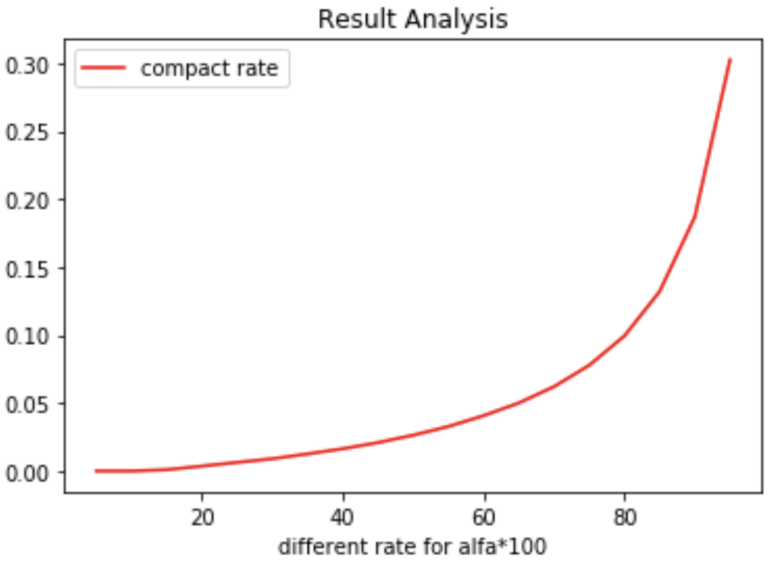
1. 重构矩阵时不加原先的中心（存在系统误差）

1）比较不同压缩精度下重构误差的变化情况和趋势



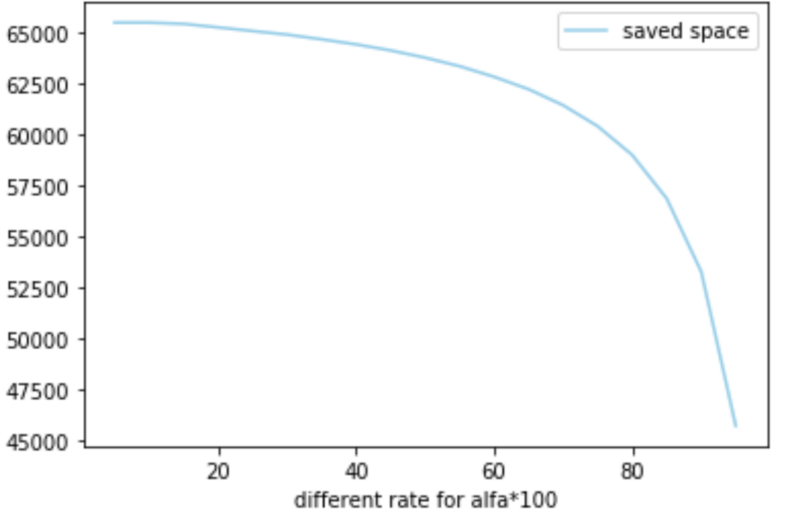
如上图所示，可以明显地发现重构矩阵计算的中没有加上原先的中心导致的重构误差的增大。随着压缩精度从0.05增加到0.95，重构误差从1降低到0.85。从不同通道的变化角度分析，R和G通道的重构误差变化较一致，与重构矩阵计算加上原先中心的情况正好相反的是，B通道的重构误差整体低于R和G通道的重构误差。

1. 比较不同压缩精度下压缩率的变化情况和趋势



如上图所示，压缩率随着压缩精度从0.05增大到0.95而从0增大到0.3,除了压缩率的初始值为0，与计算重构矩阵时加入原中心得到的结果很类似。

1. 比较不同压缩精度下压缩率的变化情况和趋势

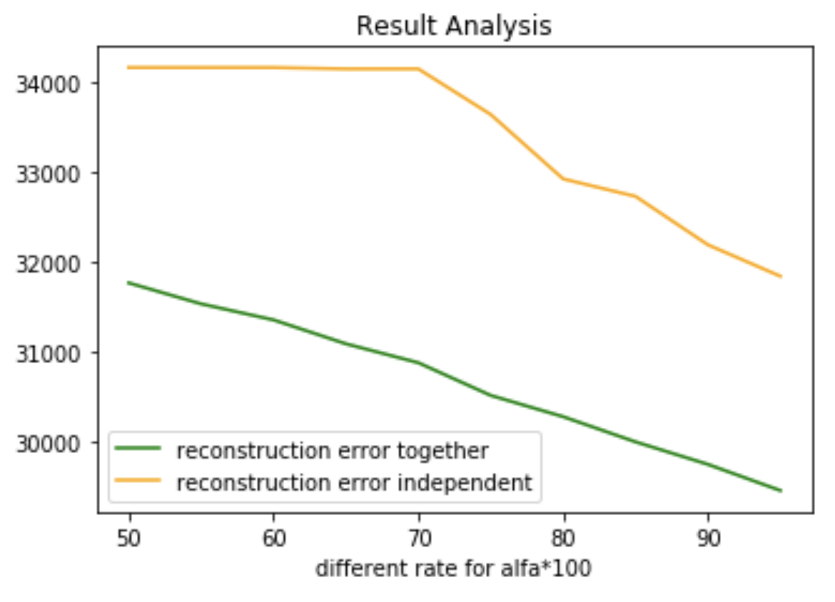


如上图所示，节省空间随着压缩精度从0.05增大到0.95而从65512减小到45686。与计算重构矩阵时加入原中心得到的结果很类似。

在方法二，我们对由大矩阵按行分解重建后的矩阵和原图像矩阵的F范数差来度量重构误差。压缩率为压缩后特征向量的大小，中心化向量的大小以及新矩阵的大小的和与原图像矩阵大小的比值。节省空间为原图像矩阵大小减去压缩后特征向量的大小，中心化向量的大小以及新矩阵的大小的和。此外，对得到的结果与独立压缩的结果进行比较。

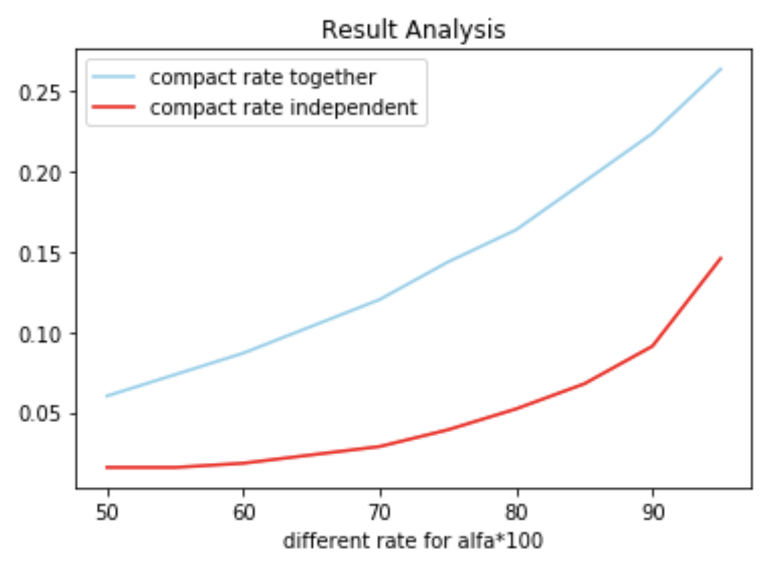
具体结果如下

1）比较不同精度下重构误差的变化情况：



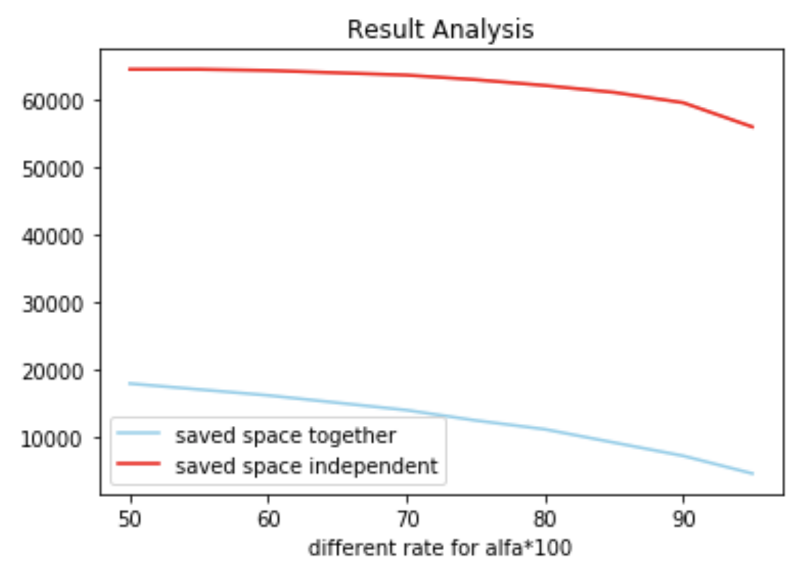
对比一起压缩和独立压缩的重构误差。图中一起压缩的重构误差变化由绿色线条显示，独立压缩的重构误差变化由黄色线条显示。可以看到拼合成大矩阵一起压缩的效果更好。

2）比较不同压缩精度下压缩率的变化情况：



对比一起压缩和独立压缩的压缩率。图中一起压缩的压缩率变化由蓝色线条显示，独立压缩的压缩率化由红色线条显示。可以看到拼合成大矩阵一起压缩的的压缩率整体比独立压缩的压缩率更高。

3）比较不同精度下空间节省的变化情况：



对比一起压缩和独立压缩的空间节省。图中一起压缩的空间节省变化由蓝色线条显示，独立压缩的空间节省化由红色线条显示。可以看到拼合成大矩阵一起压缩的的空间节省整体比独立压缩的空间节省更高。

1. 结论（对使用的方法可能存在的不足进行分析，以及未来可能的研究方向进行讨论）

在本次实验中，我利用对图像独立压缩和一起压缩分析比较重构误差，压缩率，空间节省的变化情况。将所有图片一起压缩中，对于缺失值的处理的更好解决方法值得继续研究。此外，本次实验未涉及对于不同文件夹的图像，放在一起压缩，并研究结果，可以比较不同文件夹图片一起压缩的效果与单个文件夹压缩效果。