# 《数据科学与工程算法基础》实践报告

报告题目： Image Compression

姓 名： 邓淳远

学 号： 10185101155

完成日期： 12/13/2020

提供中英文摘要

摘要 [中文]：

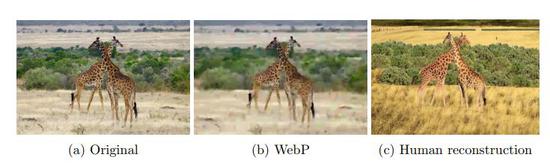
主成分分析可以用来减少矩阵（图片）的维度并且将这些新的维度投影行程新的k维图像。我们利用PCA的技术将将图片投影到新的选择的k个维度上对图片进行一个压缩。之后我们进行一个案例分析以了解不同的情况下图片压缩的情况。因为一共有300张图片，我们会选择100张图片分门别类进行压缩以及300张图片一起压缩以观察不同类型图片压缩的时候压缩率或者重构误差以了解主成分分析的时候一起压缩的图片性质对压缩的影响。此外，我还将研究选取不同特征值的百分率对图片压缩效果的影响。

Abstract [英语]

Principal component analysis (PCA) can be used to reduce the dimensions of the matrix (image) and project those new dimensions to reform the image that retains its qualities but is smaller in k-weight. We choose 100 images from single class from the following dataset, and employ PCA to compress the images. In the final report, I will illustrate the performance of image compression, such as reconstruction error, space saving, compressing rate, and case study, etc.

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）

自媒体技术问世以来，图像技术得到了广泛的研究，出现了多种图像格式和压缩技术，如PNG, JPEG, JPEG 2000, JPEG XR, BPG和WebP等。为了显著缩小图片尺寸，大多数压缩技术允许在压缩图像时有一些损失。然而，所使用的损失函数不对应于人类感知，而且得到图像在高损失水平下会显得模糊和不自然。下图我找到了一个利用WebP进行图像压缩的例子，会导致图像严重的模糊。



在此背景下，利用主成分分析（PCA）进行图像压缩的必要性就体现出来了。

1. 问题描述（问题定义）

图像压缩是数据压缩技术在数字图像上的应用，目的是减少图像数据中的冗余信息，从而用更加高效的格式存储和传输数据。

图像压缩可以是有损数据压缩也可以是无损数据压缩。

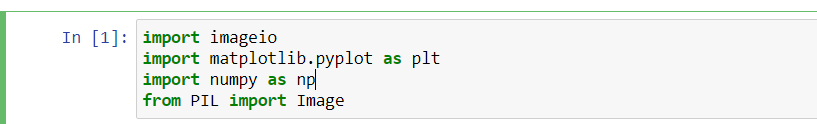
* 无损压缩： 对于如绘制的技术图、图表或者漫画，优先使用无损压缩，对于如医疗图像或者用于存档的扫描图像等，这些有价值的内容的压缩也尽量选择无损压缩方法。常用的无损压缩方法有：游程编码、熵编码法以及LZW这样的自适应字典算法。
* 有损压缩：而有损压缩则非常适合于自然的图像，例如一些应用中图像的微小损失是可以接受的（有时是无法感知的），这样就可以大幅度地减小位速。常用的有损压缩方法有以下几种：
  + 色彩空间：这是化减到图像中常用的颜色。所选择的颜色定义在压缩图像头的调色板中，图像中的每个像素都用调色板中颜色索引表示。这种方法可以与抖动一起使用以模糊颜色边界。
  + 色度抽样：这利用了人眼对于亮度变化的敏感性远大于颜色变化，这样就可以将图像中的颜色信息减少一半甚至更多。
  + 变换编码：这是最常用的方法。首先使用如离散余弦变换（DCT）或者小波变换这样的傅立叶相关变换，然后进行量化和用熵编码法压缩。
  + 分形压缩：这是一种以碎形为基础的图像压缩，适用于纹理及一些自然影像。

图像压缩的目的就是在给定位速或者压缩比下实现最好的图像质量。压缩方法的质量经常使用峰值信噪比来衡量，峰值信噪比用来表示图象有损压缩带来的噪声。但是，观察者的主观判断也认为是一个重要的、或许是最重要的衡量标准。

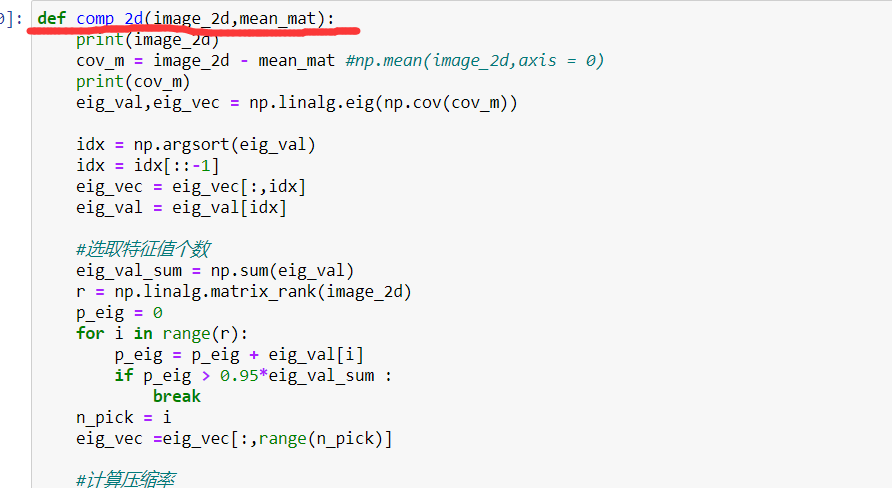
在本文中，我们将使用主成分分析（PCA）技术对图像压缩的重构误差，压缩率，节省空间进行分析。

1. 方法（问题解决步骤和实现细节）

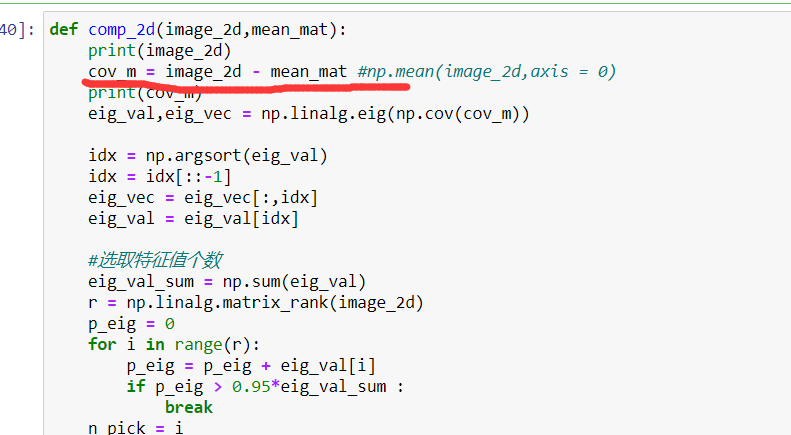
首先介绍使用的库函数，这次实验中我们使用了imageio, matplotlib, numpy以及Image库：



然后我们定义一个对2维图片进行操作的函数，

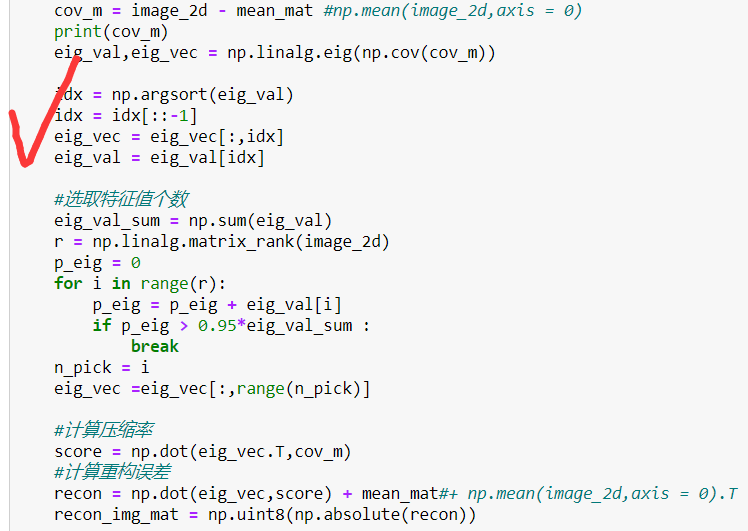


然后我们首先对每一张图片进行一个去中心化操作：



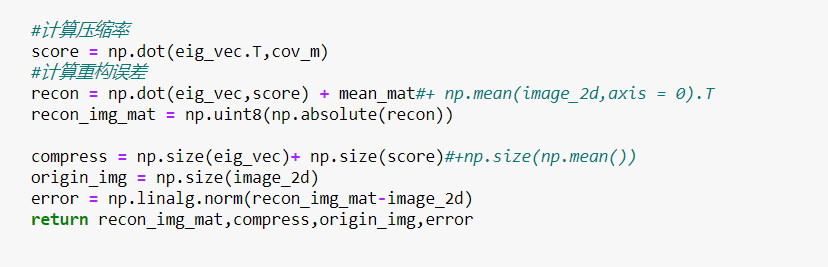
这里的mean\_mat是前面某一种类别的100张图片的平均值。对每一张图片进行一个去中心化操作就可以得到归一化的图片。

然后便是通过协方差矩阵算出矩阵的特征值以及特征向量：



下面我们根据希望选取特征值的百分比选取前k个向量，这里默认的设置是0.95，用于介绍方法，下一部分我会介绍通过调整选取的百分比来影响图像的压缩效果。

接下来就是计算重构误差和压缩率了：



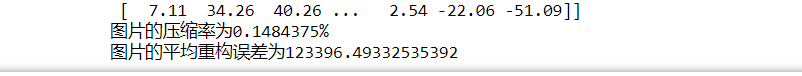
这里计算error我使用的是二范数，返回相应值即可。

1. 实验结果（验证提出方法的有效性和高效性）

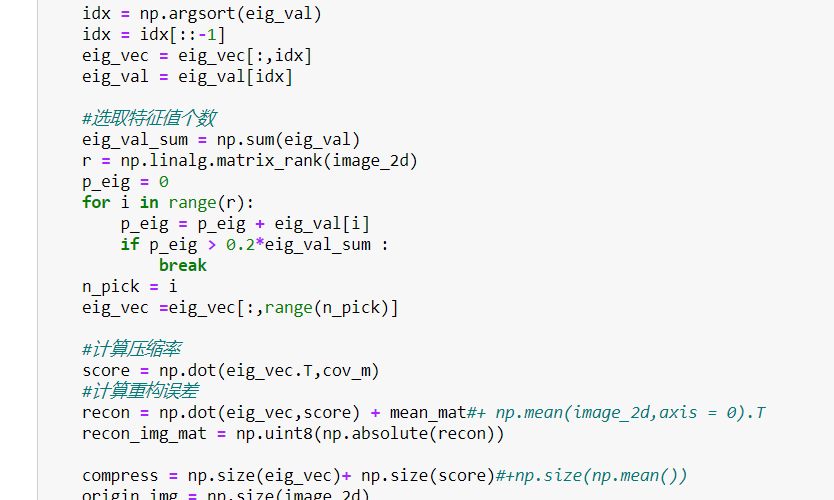
首先我们来研究选取不同百分比对图像压缩的影响：

首先我选取0.1:



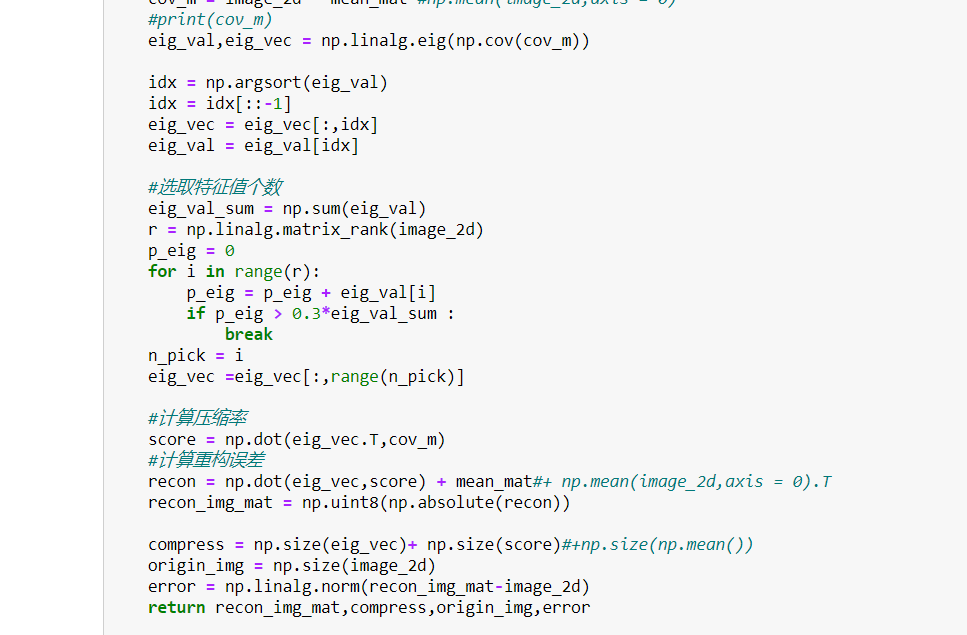


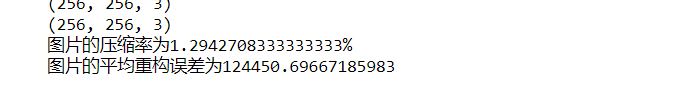
选取0.2:



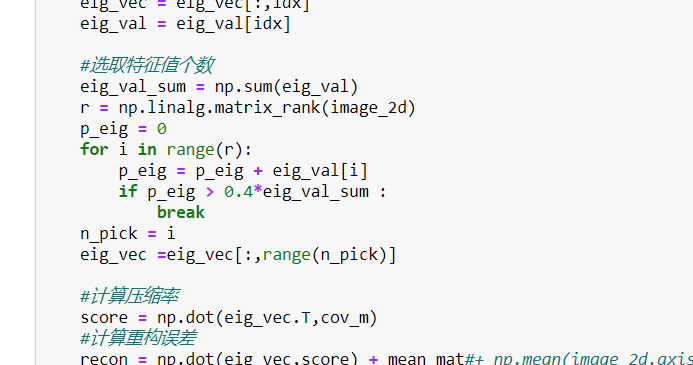


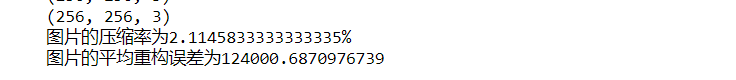
选取0.3：



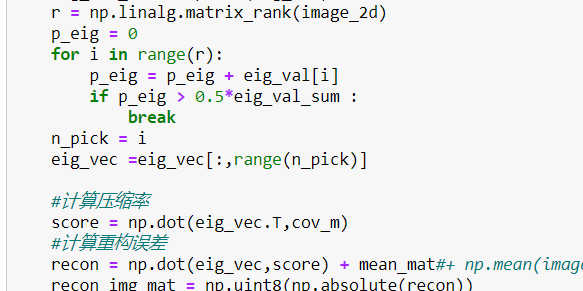


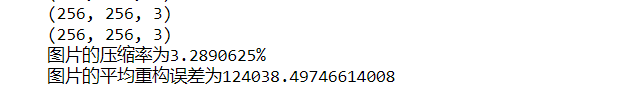
选取0.4：



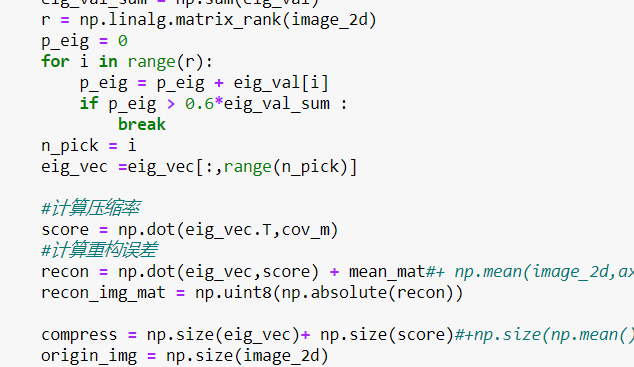


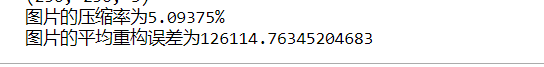
选取0.5：



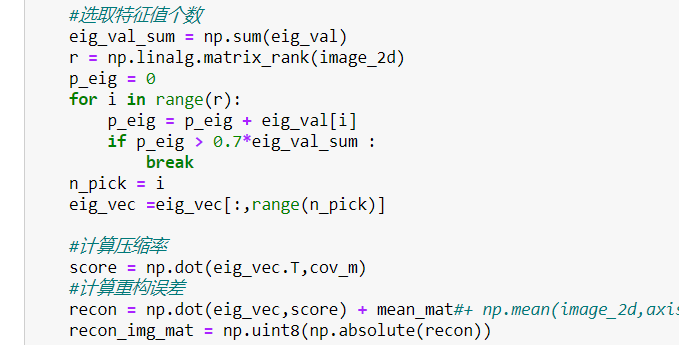


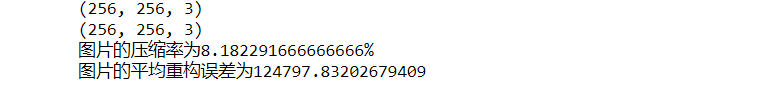
选取0.6：



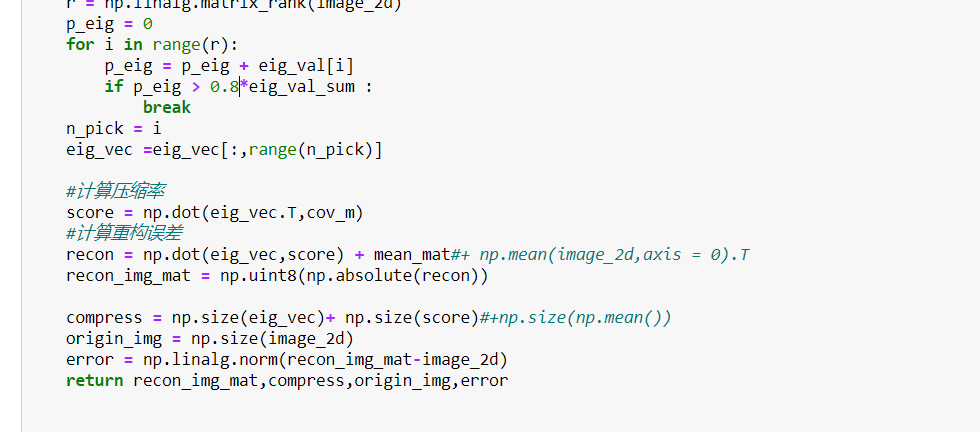


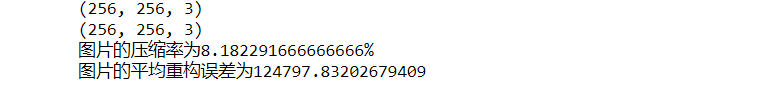
选取0.7：



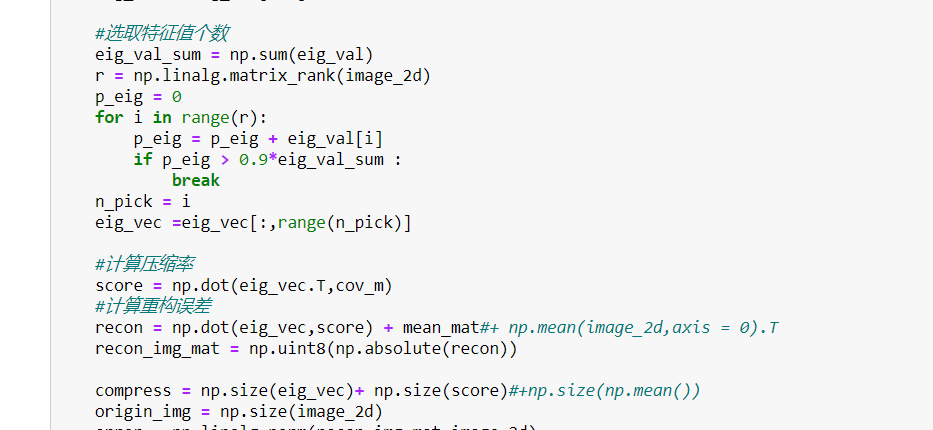


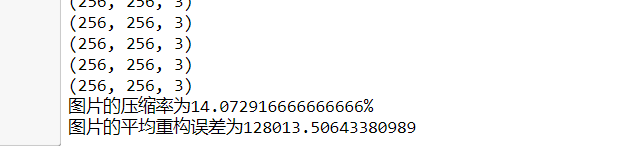
选取0.8：



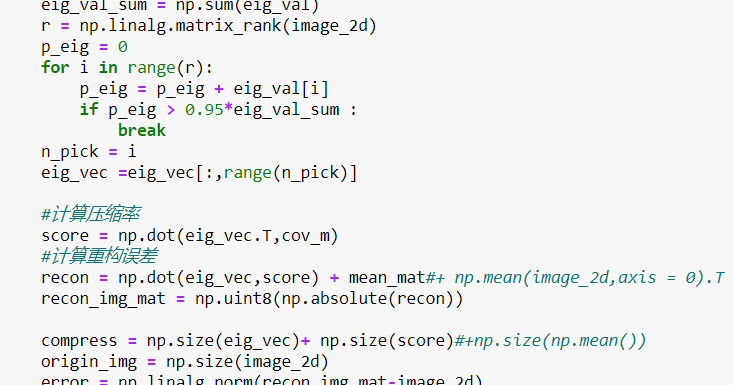


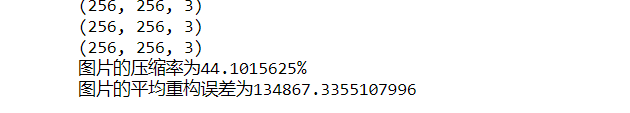
选取0.9:



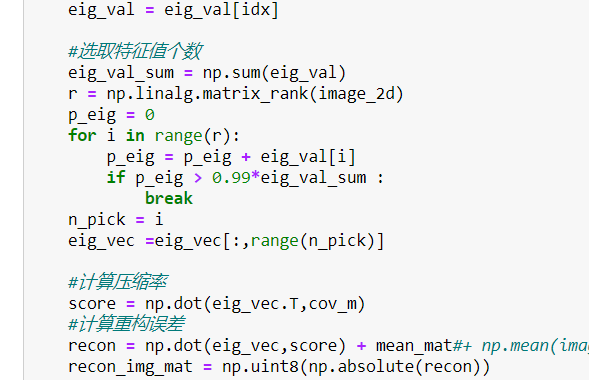


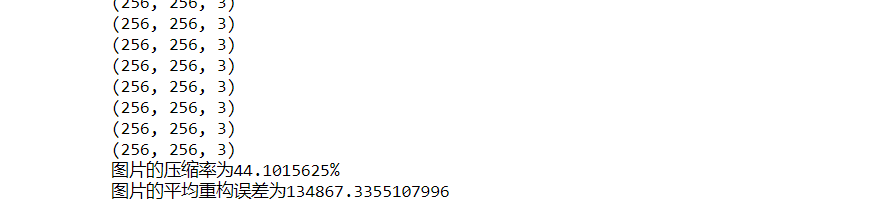
选取0.95：





选取0.99：





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）

通过这次实验，我学会了如何利用主成分分析对图片进行压缩，通过调整不同的类别与压缩图片的个数，我了解了对于PCA来说what matter most。对于PCA特征的选择来说，一定程度上选择更高比例的特征值会有更高的压缩率和更低的重构误差，但这件事也不尽然。因为长尾的存在。当超过一定的值的时候，这个优势就体现的不是那么明显了，在本次图像压缩的时候，我发现在0.95-0.99这个区间差异就不是那么大了，在直觉上可以理解为后面比较小的特征值对于图像整体的重构影响没有那么大。

而对于一类图片进行PCA而言，确保这是“一类”图片是尤为关键的，因为我们去中心化的代码可以看到，我们需要精确在一个pixel上体现出被一类图片归一化了，这也就对于“被PCA的图片”与“一类被PCA的图片”之间的关系提出了很高的要求。

以上就是本次实验的总结以及我对这个lab的一些思考，谢谢高明老师和李助教的耐心指导，感恩。