哈爾濱工業大學
HARBIN INSTITUTE OF TECHNOLOGY

2020 年春季学期

计算学部《机器学习》课程

Lab 4 实验报告

姓名	李卓君
学号	1180300210
班号	1803104
电子邮件	zhuojunlee724@gmail.com
手机号码	18845636557

目录

1 实验目的与实验要求

1.1 实验目的

实现一个 PCA 模型，能够对给定数据进行降维（即找到其中的主成分）

1.2 实验要求

（1）首先人工生成一些数据（如三维数据），让它们主要分布在低维空间中，如首先让某个维度的方差远小于其它唯独，然后对这些数据旋转。生成这些数据后，用你的 PCA 方法进行主成分提取。

（2）找一个人脸数据（小点样本量），用你实现 PCA 方法对该数据降维，找出一些主成分，然后用这些主成分对每一副人脸图像进行重建，比较一些它们与原图像有多大差别（用信噪比衡量）。

1.3 实验环境

Windows 10, Visual Studio Code, Python 3.8.5

2 实验背景与原理

2.1 实验背景

PCA 的主要功能就是给数据降维，例如给定数据集有两个特征x_1, x_2，正常情况下我们需要在一个平面上去刻画这个数据集，但是这个数据集可能暗含某些信息，使得我们可以将器压缩到一维的特征，从而用一条直线来刻画这个数据集，同样的，对于三维的数据集，我们考虑用一个平面来刻画。

直观地，如果一个三维的数据集的数据刚好分布在空间中某一平面的两侧，那么我们就可以用数据在该平面的投影来降维刻画该数据集，问题在于我们如何寻找到这一低维平面，PCA（主成分分析）就解决了这一问题。

2.2 实验原理

对于原始数据集X，其点x_i在目标低维空间的投影为$W^T x_i$，若要使样本点的投影尽可能的分开，应当使样本点投影后的方差最大化，即令下式最大：

$$\underset{W}{\operatorname{argmax}} \sum_{i=1}^{m} W^T x_i x_i^T W = \underset{W}{arg\max} tr(W^T X X^T W)$$

$$s.t. W^T W = I$$

PCA 即求$X^T X$的特征值，只需将$X^T X$进行特征值分解并将得到的特征值排序，提取前 K 大的特征值对应的特征向量即可构成变化矩阵W。

具体算法如下：

1.将样本点去中心化：所有样本每一维度减去该维度均值；

2,计算协方差矩阵：

$$\Sigma = \frac{1}{m}\sum_{i=1}^{m} X^T X$$

3.奇异值分解：利用 np.linalg.svd 函数进行奇异值分解，获得特征向量；

4.选取前 K 个特征向量构造降维后样本。

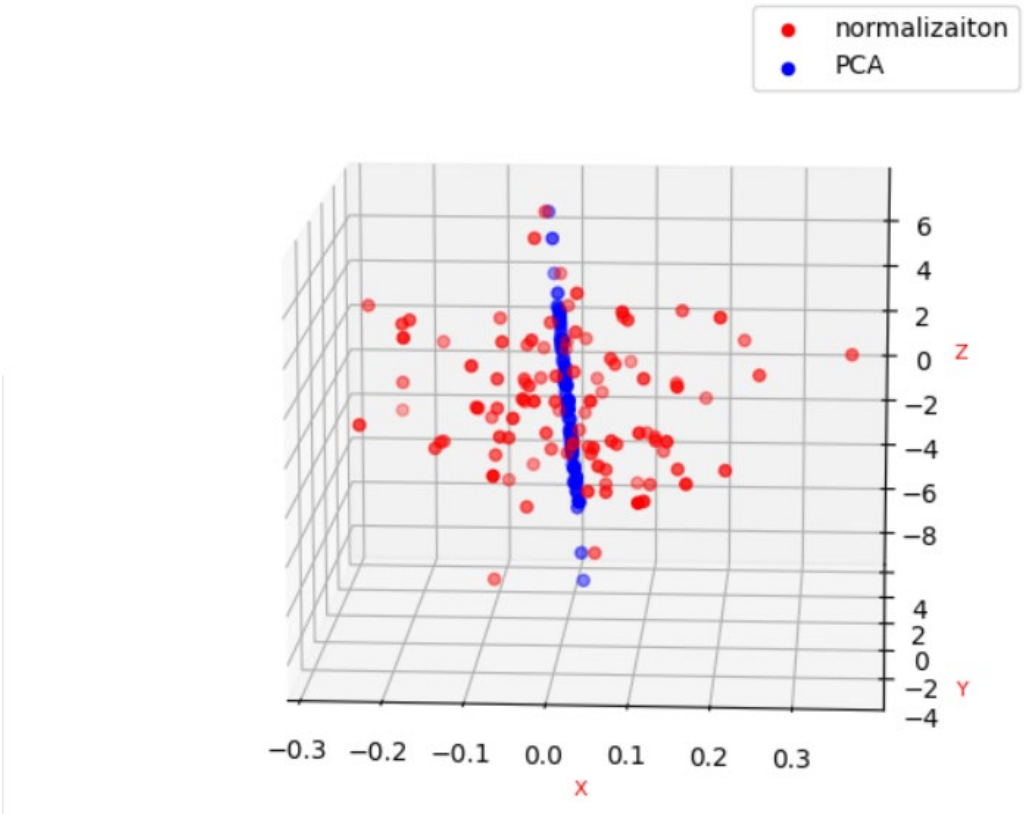
3 实验结果

3.1 手工生成数据

手工生成数据为三维点集，三个维度上的均值均为 4，协方差矩阵为：

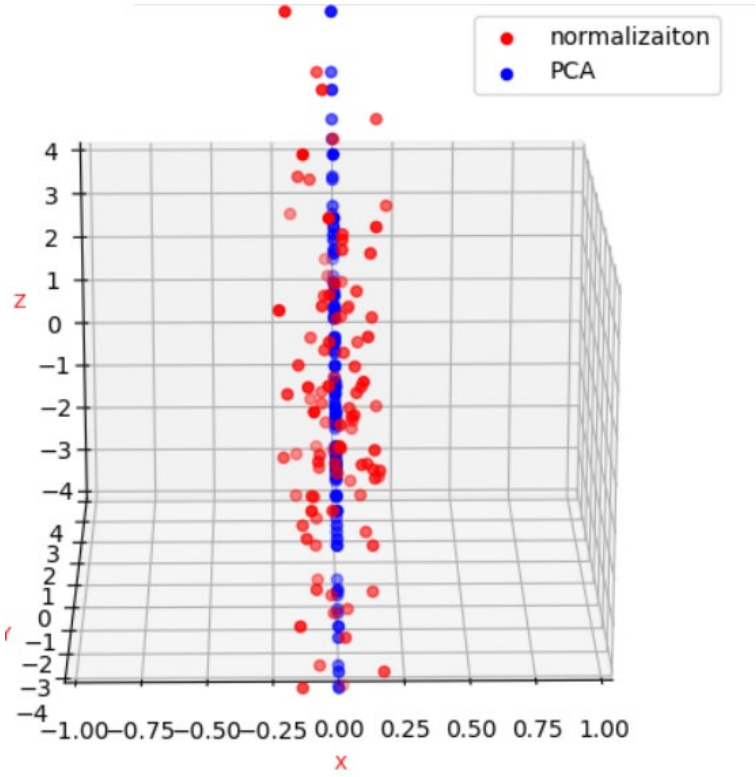
$$\begin{bmatrix} 0.01 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

点的个数设置为 100 个，利用 pca 方法降至二维，效果如下

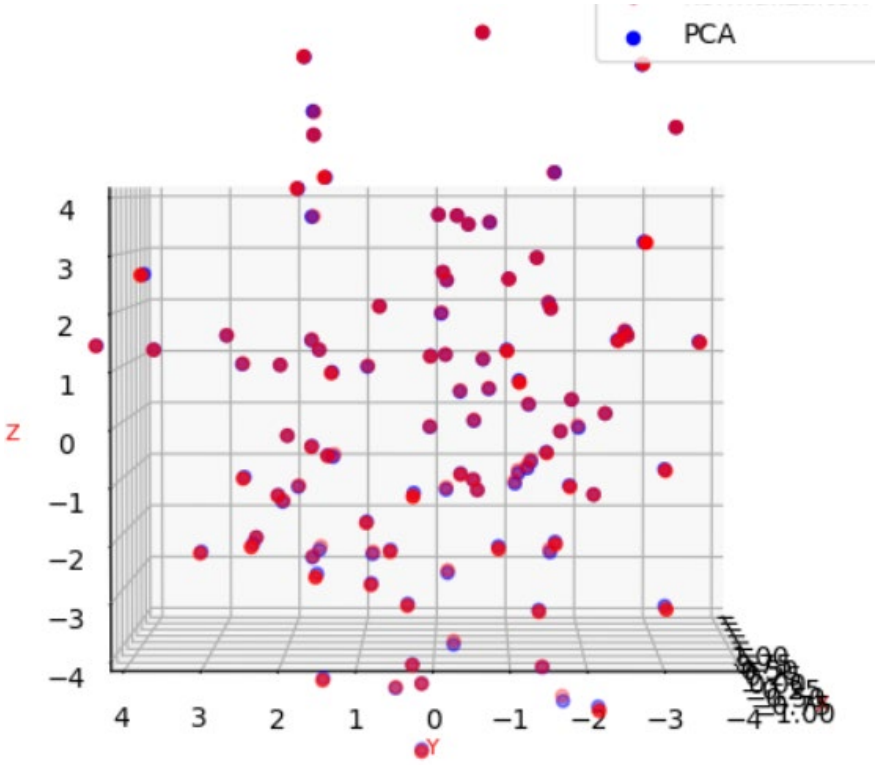


红色点为经去中心化的原始点，其分布在整个空间中，由于各个坐标上的范围不同，原始数据紧密分布在一平面两侧可能不明显，蓝色点位经 pca 后的点，明显分布在空间中一平面

上，下图为调整坐标范围的效果图：



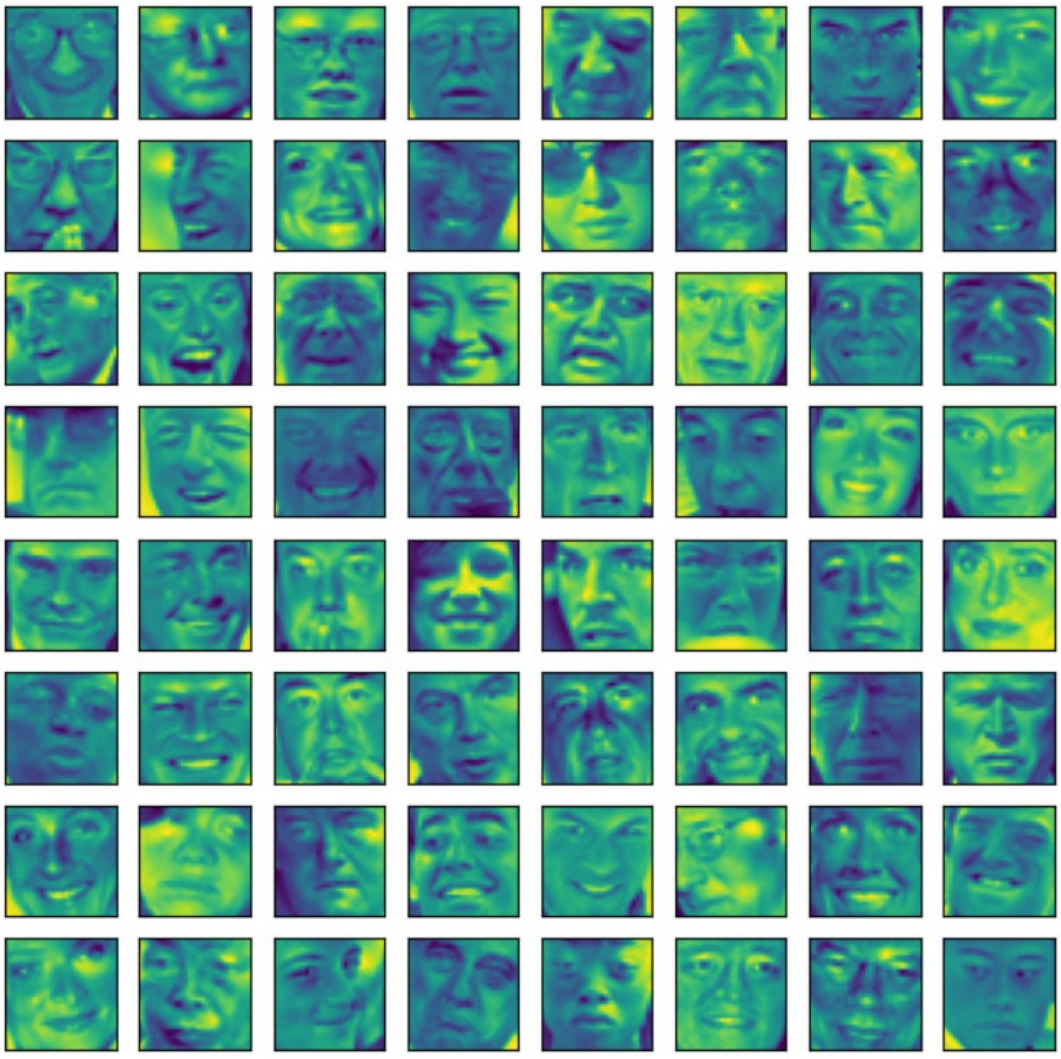
拖动图像可以看出，蓝点与红点在蓝点所在平面的投影基本重合：



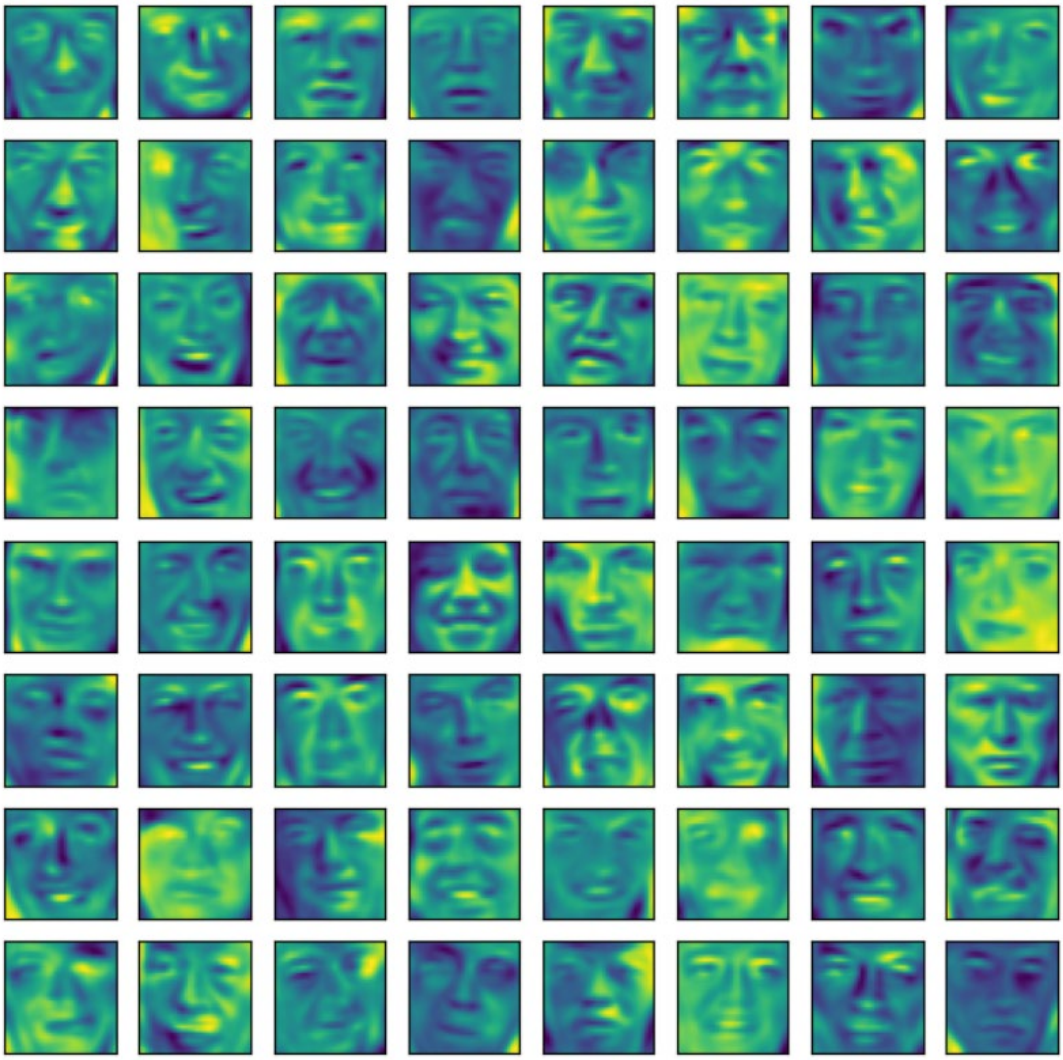
3.2 人脸数据集

人脸数据集选用吴恩达在 Coursera 课程实验中使用过的数据集 ex7faces.mat，其中包括 5000 张人脸图，每张图用 32*32=1024 大小的向量表示，由于是灰度图，所以向量上每个点的取值为 0-255。数据集实际上就是一个 5000*1024 大小的矩阵。

截取前 64 张人脸，经去中心化后如下：



接下来考虑用 pca 进行压缩，将原有的 32*32 个像素表示的人脸压缩为 10*10=100 个像素表示，效果如下：



对比可以看出，压缩至 10*10 的人脸失真较大，但仍能辨认，接下来对于 30*30、25*25、15*15、10*10、8*8、6*6、4*4 的 pca 压缩维度进行测试，输出信噪比(单位：dB)：

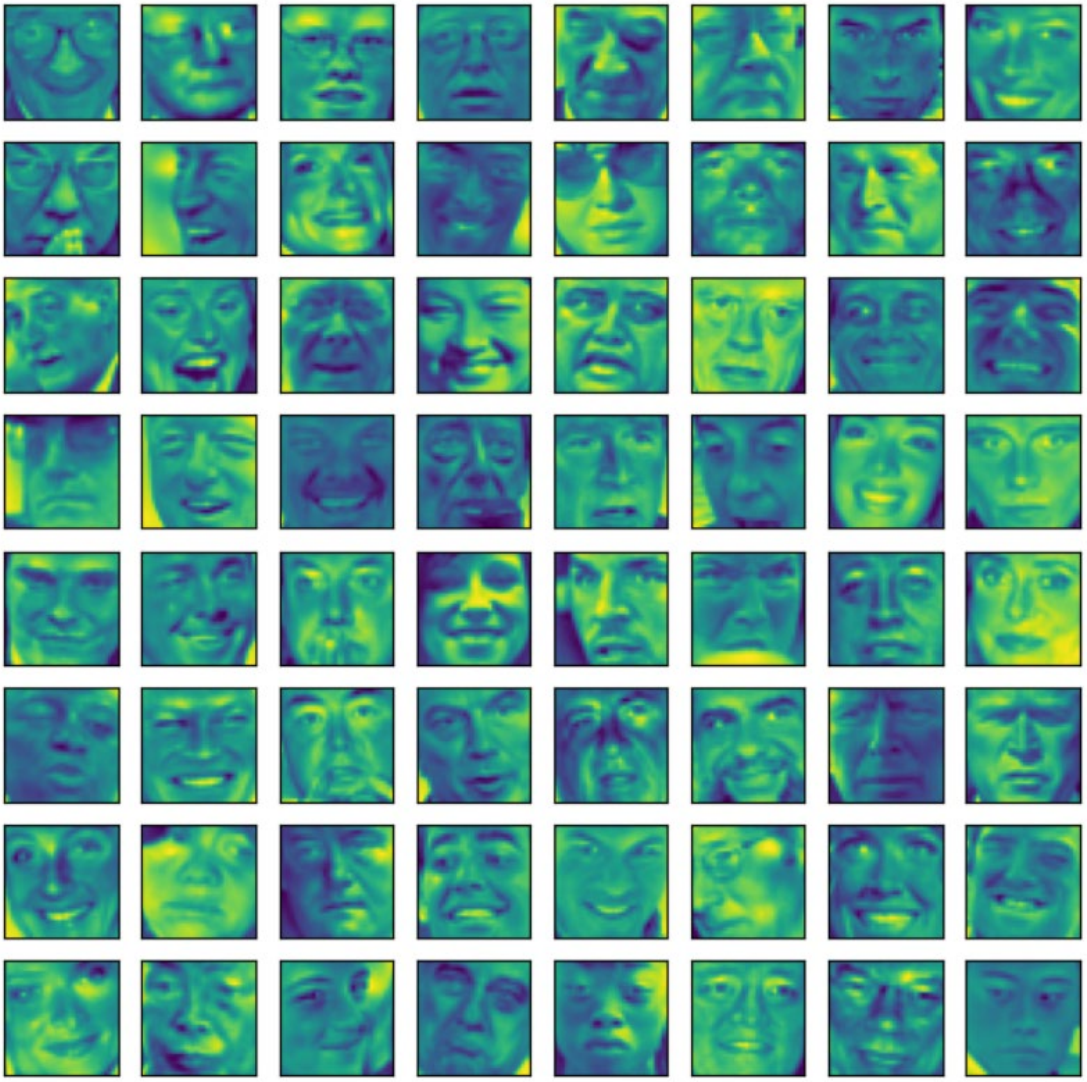
```
信噪比：  55.64833091395126
信噪比：  44.44123481777649
信噪比：  32.60434926871815
信噪比：  27.68678148991536
信噪比：  25.731406579371043
信噪比：  23.730705124600668
信噪比：  21.734731574418827
```

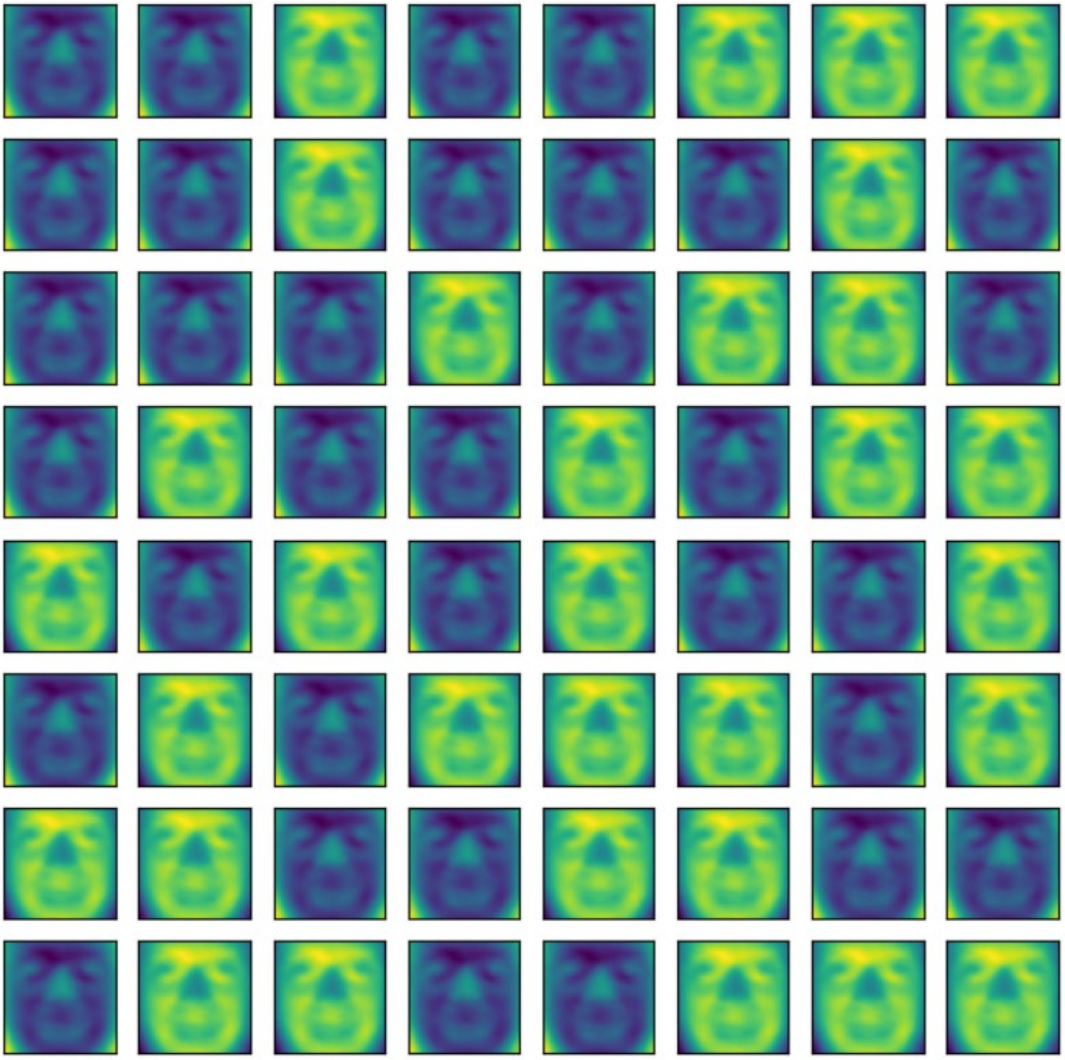
其中信噪比计算公式如下：

$$MSE = \frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}||I(i,j) - K(i,j)||^2$$

$$PSNR = 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) = 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right)$$

按照网络上的说法，当 PSNR 高于 40dB 时说明图像质量极好，在 30-40dB 时失真可以察觉但可接受，在 20-30dB 时图像质量差，低于 20dB 时图像不可接受，下面分别给出 30*30 和 1*1 的图像:





可以看出压缩至 30*30 时与原图几乎没有差别，而压缩至 1*1 时知能呈现两种人脸，无法区分。

4 实验结论

PCA 算法降低了训练数据的维度的同时保留了主要信息，如果下降的维度适当，实际上是对于原有样本噪声的消除。但被舍去的信息不一定不重要，只不过未在训练集上表现，PCA 也有可能加重了过拟合。

由于 PCA 的特性，它可以用于数据压缩以减少样本维度，提取了主要特征，减少噪声可以在后续计算中提高速度；另外对于一些较高维度的数据也可以通过 PCA 降维以实现可视化，使数据研究更加直观。