COGS 118B Final Project

Brian Chu, Julie Hang, Stephanie Kwan, Richard Li, Changchen Liu, and Andrew Truong

3/11/2022

1. Introduction

1.1 *Task*

Through this project, we explored the functionality of principal component analysis (PCA) in image compression. While PCA is a commonly used methodology for computer vision and image compression, we sought to determine its efficacy in simultaneously maintaining important feature while reducing storage cost.

We initially hypothesized that undergoing PCA (*i.e*, reconstructing images with fewer principal components) would preserve the most important features of a given image, while reducing the noise introduced by differing backgrounds. Beyond manual visual inspection, we verified the effects of PCA by performing K-means clustering before and after the dimensionality reduction and reconstruction, in order to have a quantifiable metric of how images would be clustered from their retained features.

1.2 *Dataset*

The dataset we used was the Animal Faces-HQ (AFHQ) dataset from Kaggle. This dataset contains three domain of images; cats, dogs, and wildlife. All images are standardized at $512 * 512$ resolution, and contains roughly 5000 images per domain. We selected this dataset due to having a significant enough sample size to prevent any issues stemming from insufficient data, as well as the standardization of each data sample.

2. Related Works

2.1 *Bishop*, § 9.1 K-Means Clustering and § 12.1 Principal Component Analysis

We referenced sections from Christopher Bishop's *Pattern Recognition and Machine Learning* pertaining to K-Means Clustering and Principal Component Analysis. These provided introductions to these methodologies, as well as means for implementing them.

In prior assignments in this course (Homework 4 & 5, Project 1), we implemented these algorithms

as well. We referenced these implementations in the design of this project.

2.2 *Other Research*

We referenced *K-Means clustering via principal component analysis* (Ding, et al.) for necessary context information for our project. They note the implications dimensionality reduction has for unsupervised learning, as well as means of procuring ideal approximation for optimal clustering. While their research was focused on DNA gene expression rather than images, their work provided direction for our project and confirmation of our approach's efficacy.

3. **Methods**

PCA reduces the dimensionality of the data. K-Means algorithm calculates clusters.

1) Converted colored JPG images containing animal faces into gray scale, which we then represented using a 512 x 512 matrix containing values for each pixel location.

2) Performed KMeans clustering on each individual group in the dataset(cats, dogs, wild), and also combinations of groups(cats + dogs, cats + dogs + wild).

3) Ran PCA on different groups of the dataset: cats, dogs, wild animals, cats and dogs, all animals. For each individual group:

First get the matrix U of all the principal components of the dataset.

Produced scree plot (amount of variation captured with each additional principal component).

Based on scree plot, we chose 150 principal components which still captures a significant amount of variance; in other words, look for points where graph levels off.

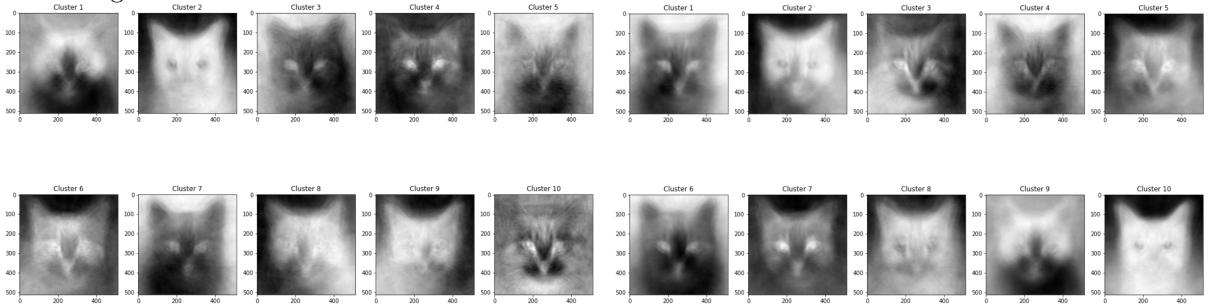
Reconstruct the whole dataset of images using the chosen number of principal components.

4) After reconstruction of the dataset, we performed KMeans clustering again using the original number of clustering.

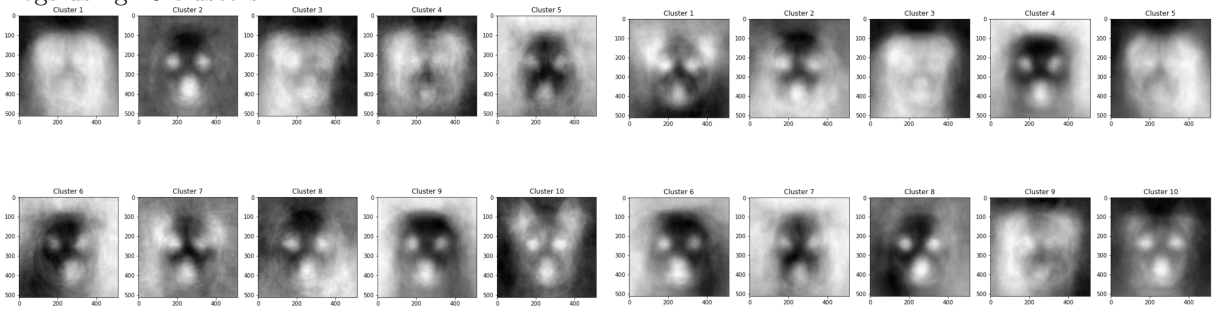
4. **Results**

K-Means clustering of images(left-before PCA, right-after reconstruction with 150 principal components)

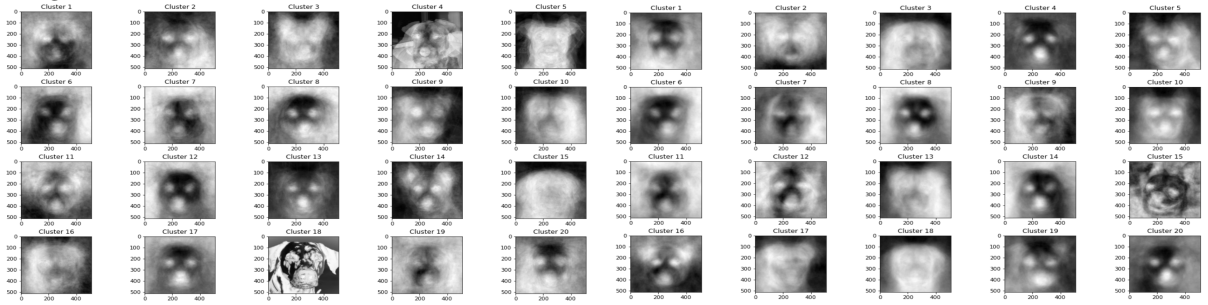
Cats using 10 clusters



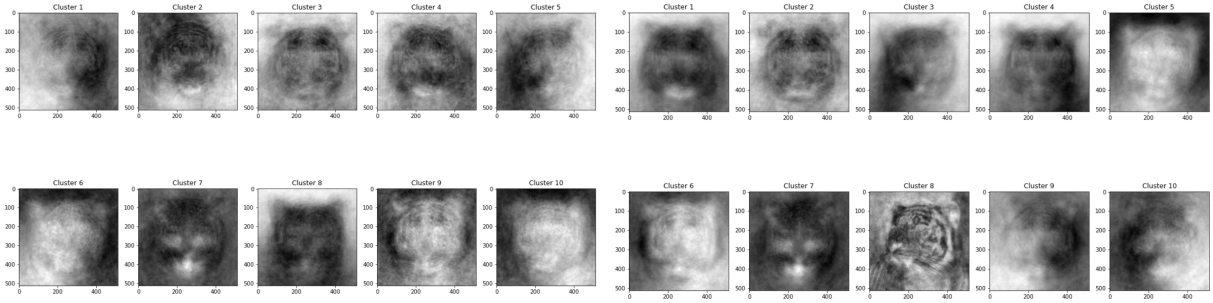
Dogs using 10 clusters



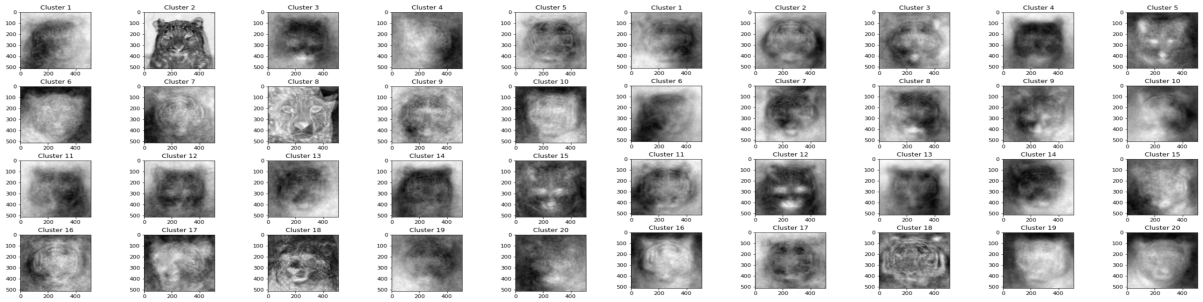
using 20 clusters



Wild Animals using 10 clusters

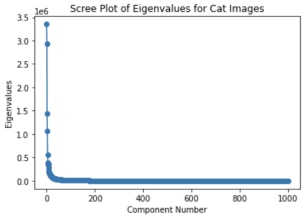


using 20 clusters



We judged the quality of our results based on how well we could relate this image to the real animal we were clustering. The images that performed better, following our expectation, was the dogs dataset using 20 clusters (dogs20), and the wild animal dataset using 10 clusters (wild10). We discovered that the KMeans clustering performed better before the PCA for wild animal dataset using 20 clusters (wild20). This may be due to the randomness like having a better learning path for clustering or because the wild animals had distinct pictures which PCA may have made less distinct. For some others, K-Means produced clusters of similar quality before and after PCA, in the cases of clustering cats using 10 or 20 clusters (cats10, cats20) and dogs using 10 clusters (dogs10), which tells us that in some cases the PCA didn't work that well in improving the K-Means clustering.

We needed to use more principal components than expected. Based on the scree plot, it looked like we just needed 10-20 principal components because that is around where the graph begins to taper off. However, in practice, 10 principal components only achieved very vague results – orientation of the head looked about right, and the ears were generally in the correct area, but there were barely any details beyond that. We used 150 principal components, so that we could at least get details from the clusters that showed lighter/darker spots, or patterns on the animal faces. Our guess as to why using 10-20 principal components did not give good results is because of many inconsistencies and differences in the dataset, including angle of face, species of animal, colors, patterns, etc. Furthermore, we used 1000 images, as compared to the homework where we used 48 images. Because of all these inconsistencies and the large size of the dataset, putting the images together may have caused a lot of noise that erased finer details that could be found in these images.



We had also done K-Means before and after PCA on combinations of the animals. When we did

just dogs and cats, we would expect two distinct clusters, but we didn't get a very definite image as it was very blurry. To solve this, we increased the cluster counts, but tended to get images that looked like a mixed breed of dogs and cats around 10-20 clusters. Finally, we explored combining all the datasets and got very jumbled results for clusters 3, 10, and 20.

Overall, our results are inconclusive due to varied results as cat and dog image datasets do not differ enough in variance or structure to affect the ability for K-Means algorithm to cluster the images. We observed that some images in the reconstructed dog images tended to look like a bunch of dog faces facing different ways clustered on top of each other, which could be due to the dataset containing many different orientations. There were a few exceptions in the wild animal results with 20 clusters before PCA, which gave us an extremely clear image of a tiger.

5. **Discussion**

This project has taught us that logic doesn't always equal results. From the research paper the logic follows that reducing images with PCA(reconstructing images with fewer principal components) would improve the accuracy of KMeans by reducing noise(different backgrounds), while preserving only the most important features, however our results from our dataset do not consistently display this. Our results were inconclusive, since on some datasets, K-Means performed better before PCA, while some after applying PCA for image compression. We found it to be hard to get a good average of an image of an animal due to many different species within the cat or dog dataset and in the wild dataset. In addition there was also inconsistencies in the angles of faces and how much of their body is shown in the picture. Next time, we could possibly find a way to center all images and orientate them the same before processing.

The clustering seemed to not work as well as the MNIST dataset of handwritten digits, which are easier to cluster because the differences in orientation are more obvious, and black and white images are simpler to represent and process. However, if faces were black and white, there would be almost no way to cluster since the orientation of the faces are very similar and important features would be lost.

A few things we could improve our project significantly were to use better datasets which have a more consistent orientation. If they were facing forward and cropped similarly then it could possibly reduce

noise and reduce the inconsistencies that we dealt with throughout the project. In addition to that, we could use GMM clustering to see if there were any difference or improvements. This include the idea of soft clustering versus hard clustering as well as the fact that K-Means algorithm does not account for variance. This could show how the variance would have affected the dataset. If there were more time, we could analyze the dataset to see how many different species are in each dataset, allowing us to use the number of different species to cluster species within each dataset. It may even show us how well it can separate animals into species.

6. **Contributions**

Brian Chu, Julie Hang, Richard Li: Assigned to and primarily focused on implementation on K-Means; worked on intro, related works, and methods for report

Stephanie Kwan, Albert Liu, Andrew Truong: Assigned to and primarily focused on implementation of PCA; worked on results and discussion for report

7. **References**

K-Means and PCA for Image Clustering and Visual Analysis Code Guidance

https://towardsdatascience.com/k-means-and-pca-for-image-clustering-a-visual-analysis-8e10d4abba40

K-Means and PCA Combination Analysis

https://dl.acm.org/doi/abs/10.1145/1015330.1015408

Dataset

https://www.kaggle.com/andrewmvd/animal-faces