Unsupervised Learning to Discover Similarities between Bird Species

Ruiming Li, Jacob Zimmerman

rayli@sas.upenn.edu, jdzimmerma@haverford.edu

Abstract

In this project, we aim to find the best ways to cluster images of 230 bird species[1]. The data for the project is from Kaggle and consists of images of birds and labels of their species. We implement feature extraction using pre-trained neural networks such as VGG19 and perform dimensionality reduction using PCA. We experiment with K-means and GMM clustering on the feature extractions of the images. We evaluate the performance of our clustering using 1) measures such as distortion and entropy, 2) visualization and inspection of images in the clusters, and 3) prediction accuracy on unseen images since we have access to the true label of the images. We found that using VGG19 is best for feature extractions and using GMM with 50 clusters and a tied covariance best suits this dataset with appropriate trade off between cluster entropy and algorithm efficiency. We discovered that clustering detects many different patterns uncorrelated to the labels of the images and thus performs poorly for prediction tasks, but can be extremely useful for generalizing similarities across species of birds in biology.

1 Motivation

The dataset consists of images of birds and their corresponding species. The general aim is to classify an unlabelled image into one of the 230 given species. While many supervised methods have been used for classification, we want to investigate the effectiveness of using unsupervised clustering algorithms for classifying bird species. We can do so by setting up at least 230 clusters and assigning images to the nearest cluster and predicting its label by the majority in the cluster. The motivation for using unsupervised clustering is to be able to classify birds by species without the need for labeled data and also to find the clustering best method so we can eventually come up with a smaller number of clusters for the 230 species of birds and discover similarities across species of birds within a cluster.

Investigating unsupervised machine learning methods is valuable for learning from the large amount of unlabeled data available. We can determine how accurate unsupervised clustering can be in differentiating species of the same biological family. Additionally, clustering the 230 species of birds with a smaller number of clusters has important implications in biology because we can examine the similarities of birds across different habitats and hypothesize how evolution created these similarities. It is also helpful to know how unsupervised clustering techniques measure up against supervised techniques if we use them for classification because unlabeled data is abundant while labeled data is scarce.

The challenge for the problem is that we have to figure out the best way to embed each image as a vector and then use that vector to optimally classify and cluster unlabeled data.

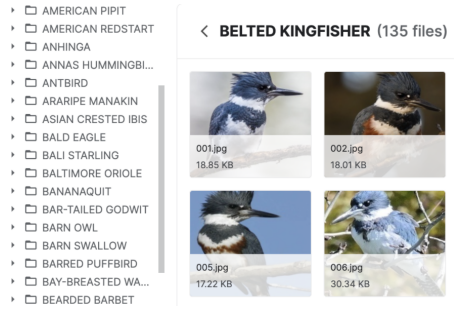


Figure 1: Sample data.

2 Related Work

There are existing Kaggle Kernels(Jupyter Notebooks)[2] that achieve high accuracy using VGG19 or ResNet and fine-tuning additional neural networks as a supervised learning task. We can similarly use VGG16 for feature extraction because it performs well and is widely used in many Kernels. We can do further dimensionality reduction such as PCA on the feature vectors, which has not been considered in other Kernels using neural nets. We believe that doing PCA will be suitable and efficient for our clustering task.

Furthermore, the focus of many Kernels are on classification since this is primarily a supervised classification problem on Kaggle. Aside from looking deeper into classification, we aim to find a novel application of clustering algorithms in discovering similarities between species of birds by using a small number of clusters across the 230 species.

3 Dataset

In this project we are using the "230 Bird Species" dataset provided by Gerry on Kaggle. This dataset consists of the images of birds and their respective species. Below is a summary statistics of the dataset:

Data	Value
Train	32025
Test	1150
Validation	1150
Image size	$224 * 224 * 3$
Number of classes	230

Table 1: summary statistics of dataset.

We visualize the count of images for every species in the training dataset to investigate if there's any class imbalance in the dataset.

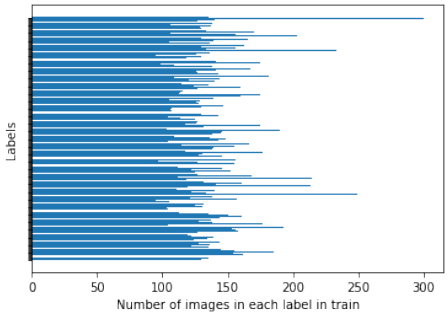


Figure 2: Count of images in each class.

From Figure 1 and further analysis, we concluded that: 1) 70% of the total species has between 100 to 150 number of images. We do not believe that there is serious imbalance towards certain classes, 2) the highest frequency class is the species SORA with 300 images, and the lowest frequency class is SAND MARTIN with 95 images.

3.1 Data Preprocessing

The bird images are already processed into the same size, scale, and resolution, with the subject of interest in the main part of the image. Hence, we do not perform further cropping or rescaling of the data.

To make sure there is a better balance of images in every class, we performed augmentation of the dataset by horizontally flipping images in the dataset and augmenting the number of images in every class to at least 150. We decided to use horizontal flip because images of birds are still meaningful when flipped horizontally, compared to vertical flip or cropping that will create uninformative samples.

4 Problem Formulation

Representation problem: The first step in this project is to convert the images into a vector that can be fed as input into a classifier or used for clustering. Our goal for the representation problem is to find the best representation method that gives the highest accuracy in our classification and clustering tasks later.

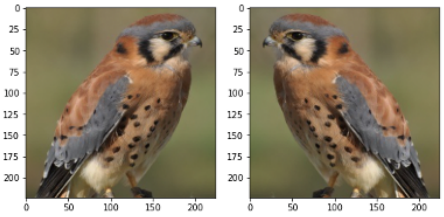


Figure 3: Horizontal flip of images.

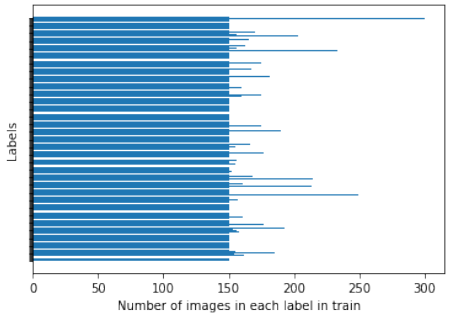


Figure 4: Count of images in each class after augmentation.

This will require either some simple matrix transformation from raw image representation to a 1-d vector, or more complex feature-detection techniques using Convolutions Neural Nets for computer vision. We can compare these methods by looking at accuracy score when feeding various extracted features as input to the classification problem.

We want to further reduce the dimension of the extracted features because the feature space is very large (25088 features) We use the popular Principal Component Analysis because it is common and effective in compressing data into lower dimensions.

Classification problem: We want to classify 230 species of birds correctly, meaning to maximize accuracy score and minimize a loss function between our model prediction and the true label of the images in the validation set. We explore baseline methods such as logistic regression and SVM to train on the labeled dataset as a supervised task and compare results. Their loss functions are the logistic loss and hinge loss.

Our aforementioned goal is that we eventually want to train accurate unsupervised models with unlabeled data. Thus, we try various unsupervised clustering methods for classification. Since we have access to the labels, we can make predictions on the test data. We label each cluster as representing a bird by finding the majority label, then assign each bird in the test data to the closest cluster and see if the label matches. We can then measure the accuracy on our test dataset.

Clustering problem: Aside from classification, our goal is to find the best unsupervised clustering method so we can discover general similarities across different species of birds.

The quality of clustering can be measured by cluster entropy, distortion, and visual inspection of images in the clusters.

Calculating the entropy of the labels in the clusters are helpful since we have access to the cluster labels[3]:

$$H(\Omega) = \frac{N_\omega}{N} H(\omega)$$

where $\Omega = \{\omega_1, \cdots, \omega_k\}$ is the set of clusters, and N_ω is the number of points in cluster ω and N is the total number of points.

Another measurement of clustering quality is the distortion. This is actually the loss function of the k-means algorithm that is optimizing[4]:

$$D(\Omega) = \frac{1}{N} \sum_{i=1}^{N} (x^{(i)} - \mu_{c^{(i)}})^2$$

where μ_c is the centriod for a cluster c that x is assigned to.

The Silhouette Coefficient is another example of evaluation metrics for clustering algorithm, where a higher Silhouette Coefficient score relates to a model with better defined clusters. The Silhouette Coefficient is defined for each sample and is composed of two scores:

a: The mean distance between a sample and all other points in the same class.
b: The mean distance between a sample and all other points in the next nearest cluster.

The Silhouette Coefficient s for a single sample is then given as:

$$s = \frac{b - a}{max(a, b)}$$

Last but not the least, since we have access to the labels of the class, we can look at the accuracy of the clustering by assigning a test set image with the label of the majority class in the closest cluster.

$$accuracy = \frac{\sum_{i=1}^{N} \mathbf{1}(argmax_y(count(y_j \in argmin_c(x_i - \mu_c)^2)) = y_i)}{N}$$

We want to explore different clustering methods and evaluate their performances on each of the metrics.

5 Methods

5.1 Representation

The simplest approach is to use the image pixels directly and flatten the 224*224*3 image into a 1-d array.

However, it is obvious that using a simple 1-d array to represent image is expensive and inefficient. The raw image has many correlated pixel points and many uninformative pixels are not helpful for classification or clustering. Hence, we decided to use transfer learning of pre-trained computer vision neural nets to extract relevant features from each image. We then plan on implementing PCA to reduce the large dimension of the feature vectors.

5.2 Classification

As a baseline we will use logistic regression and SVMs with different kernels to classify the bird species, then determine their effectiveness by measuring their accuracy on a held out test set. We can use this as a baseline comparison for our experimental unsupervised classification clustering techniques.

We will use two popular clustering techniques, K-Means and GMM, and test different hyperparameters. We chose to vary the number of clusters in both clustering methods, as well as vary the covariance matrix types in GMM. We decided on 10, 50, 100, 150, and 230 (230 is the number of bird species, so we use this for classification). As for GMM covariance matrix types, there are four choices: 'full,' 'tied,' 'diag,' and 'spherical.' We initially chose to test all but 'spherical' because that method is very similar to K-Means, which assumes shared spherical variances.

After trying to train full shared covariance matrices on a small training set, we realized that 'full' was taking hours for a only a training set of around 1000. So due to time constraints, we left out 'full.' This would be a great topic for further research.

Our algorithm for clustering classification is as follows:
1) Run clustering with the same number of cluster centers as the number of classes with unlabelled training data.
2) Use the labeled data to find the majority label in a cluster.
3) Assign clusters to test data by calculating the minimum distance from test data to cluster.
4) Label the test data based on the majority class in the cluster

5.3 Method Selection

For K-Means and GMM we created models using different hyperparametes. We measured each K-Means model's effectiveness using accuracy on a test set, entropy, and distortion. We measured each GMM model's effectiveness using accuracy on a test set and entropy.

From these measures, we choose the best model for differentiating the bird species into categories (not species categories, just general categories of similarity).

6 Experiments

6.1 Baseline

We first use baseline logistic regression and SVM to determine the effectiveness of different feature extraction methods.
Similarly, we try the baseline methods on the flattened input

Baseline accuracy with VGG19 (SGD)

Methods	Train Accuracy	Test Accuracy
Logistic Regression	99.9%	96.1%
SVM	99.9%	96.5%

image and obtained a lower training and testing accuracy:

Baseline accuracy without VGG19
(SGD with online learning)

Methods	Train Accuracy	Test Accuracy
Logistic Regression	95.1%	84.4%
SVM	93.2%	86.2%

In order to deal with the large number of input features, we have to implement stochastic learning methods to prevent Colab from crashing due to exceeding usage of RAM. To handle the even larger number of input features without VGG19, we also implemented an online learning method to partial fit parts of the training data and iteratively update the model, instead of fitting the entire dataset at one go.

From the data extraction comparison, we realized that it is too computationally expensive and unhelpful to use the original image array. Extracting it using VGG19 is the best way for classification and clustering problem.

We then attempted to reduce the dimension of the extracted features further to speed up the training. We tried using PCA on the extracted features and obtained the following results using PCA reduced features for logistic regression:

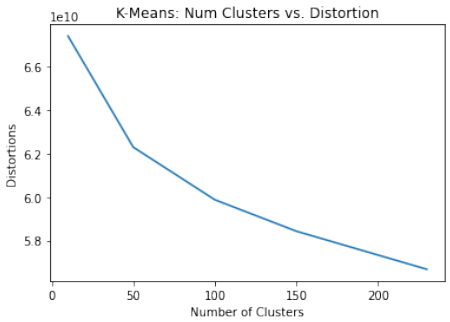
Baseline accuracy with VGG19 and PCA

Methods	Num Comp.	Train Acc	Test Acc
Logistic	5000	96.5%	0.2%
Logistic	10000	Crashed	Crashed
Logistic	20000	Crashed	Crashed

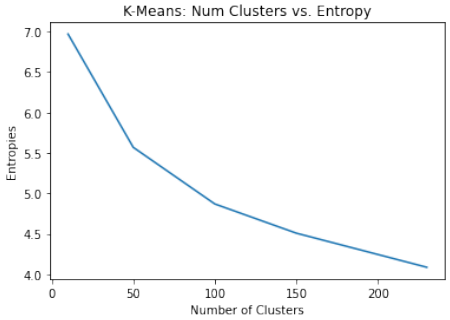
To our surprise, the PCA algorithm either crashes on high number of components or gives extremely low accuracy on lower number of components. We conjecture that the calculation for PCA is too computationally expensive for high number of components, and the remaining features when we use a small number of components are insufficient to distinguish the fine details between the similar images of the birds - we chose a challenging dataset and it is already difficult for human without formal training to classify birds. We also suspect that PCA might not be useful to handle image feature data because there are many sparse, uncorrelated, but equally important spatial features and using a small number of principal components cannot explain the variance in the image sufficiently. As a result, we decided not to use PCA for clustering tasks.

6.2 K-Means clustering

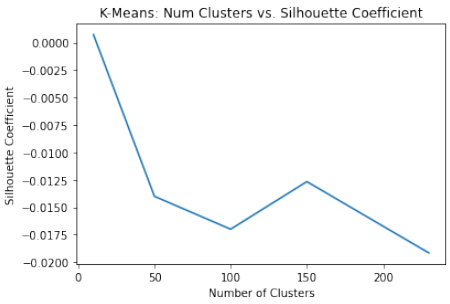
We used scikit-learn to implement K-Means Clustering. We chose to use five different amounts of clusters: 10, 50, 100, 150, and 230 to see which amount gave the best differentiation of birds. Note that 230 was the amount of bird species, so this the clustering algorithm we tested our accuracy on.



The distortion decreases as the amount clusters increases, although it decreases less and less as the number of clusters increases. This is what we expected because the more clusters there are, the greater chance that any observation is close to a cluster center.



The entropy is very similar to the distortion. As expected it decreases, although less and less, as the amount of clusters increase. Again, the more clusters there are, the greater chance that a given point is close to a cluster center.



In short, the silhouette coefficient measures how compact and defined our clusters (higher meaning more compact and defined). In our models, the silhouette coefficient also decreases with the amount of clusters. This makes sense because as we add more clusters, the closer our clusters will be

to other clusters, so the less defined and compact our clusters will be. The exception to this is at 150 clusters. Here, we see an unexpected spike in the silhouette coefficient. This denotes a better model for differentiating the birds into categories of general features.

Finally, our final classification accuracy using K-Means Clustering Algorithm with exactly 230 clusters was 22.4%. This performance of using clustering for classification is poor and we will further investigate the reason later.

6.3 Gaussian Mixture Model

We believe that Gaussian Mixture Models would do a much better job at clustering based on features in general and classifying because the covariances of the clusters are probably not spherical as in K-Means. However, when trying GMM multiple times with and without dimension reduction from PCA, our sessions crashed repeatedly for days leading up to the project deadline. Thus, we were only able to gather the entropy for 10 and 50 clusters.

Entropy of GMM

Covariance Type	**Entropy for 10 Clusters**
Diagonal	7.15
Tied	4.00
Covariance Type	**Entropy for 50 Clusters**
Diagonal	6.35
Tied	3.22

Note: "Diagonal" means each cluster has its own diagonal covariance matrix. "Tied" means all the clusters share the same general covariance matrix.

We did not know what to expect in differences of entropy between the two types of covariance matrices, but tied ended up having better entropy. We believe this is because for the small amount of 10 or 50 clusters we used in GMM, having a different diagonal covariance matrix for each cluster did not have much of an effect on optimally separating the data, whereas having a full covariance matrix did a better job of representing the variablility between features in the data.

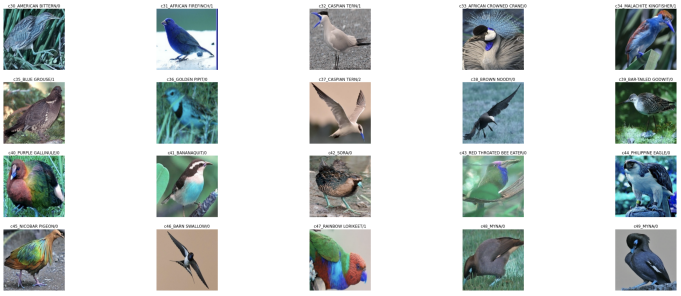
An important observation is that using GMM gives much better cluster entropy than k-means, as the entropy for 10 clusters with GMM is already lower than that for 230 clusters with k-means. However, a critical drawback of GMM is that it is computationally expensive and we can only do it up to 50 clusters. We believe using GMM with 50 cluster centers and a tied covariance type has best balance and trade-off between accuracy and efficiency.

6.4 Visualization

We decided to implement a visualization method that shows the k most representative images in the cluster, measured by the L2 distance from the image feature vector to its assigned cluster center. In this way, we can access the quality of our clustering by visually inspecting any similarities inside clusters and differences across clusters.

First we get the most representative image in each of the 10 clusters from k-mean clustering. As a sanity check, there is a diversity of birds in each cluster and our method is reasonable for on a small number of clusters.

However, having only 10 clusters is too general and visualizing 1 picture is insufficient to discover similarities within the cluster, so we fist visualize the most representative picture for 50 clusters of birds from kmeans.



We continue to observe entirely different labels from the most representative picture in each cluster, meaning our clustering is effective in distinguishing and separating different species of birds. We observe different species, background color, as well as posture of the birds. We can consider these 50 most representative picture from each cluster the sample or model for a set of birds. We can then look at inner-cluster similarity below:

Five most representative images of example cluster (posture,background)

Five most representative images of example cluster (color)

Five most representative images of cluster3 (shape,background)

Five most representative images of example cluster (shape)

From our visual inspection of the 5 most representative pictures in different clusters, we summarized the different types of discoveries in the above sample clusters. We realized that there are many obvious but different types of similarities in different clusters.

For example, in the first example cluster, we found all pictures with birds having the same posture and a branch in the background. In the third example cluster, we see all birds having similar background or habitat with grass, similar body shape, and similar posture of walking on the ground. It would be an interesting and informative starting point to investigate the similarities between these species in biology.

Another qualitative result we come to by our inspection of different types of inner-cluster similarities is that unsupervised learning discover many different types of patterns, such as color, shape, and background objects. It is different from a supervised setting where the algorithm learns to only focus on features that are predictive of the label. It is therefore **not** surprising that our unsupervised clustering has such a lower accuracy on the classification task, since it is not trained to optimize based on the labels and it freely finds patterns that can stem from the angle we capture the picture or the posture of the object, instead of features that directly correlate with the labels of the object.

7 Conclusion and Discussion

7.1 Conclusion

We concluded that the best way to cluster this dataset of images of bird species is to extract features using VGG19 and cluster extracted features using GMM with a tied covariance type and 50 clusters. We confirmed that it is inappropriate to use clustering methods for classification tasks due to the unsupervised nature of clustering and the similarities within the cluster are not representative of the labels of the images. Instead, clustering can be extremely helpful for exploring pattern across images of different classes and labels, such as finding similarities between different species of birds.

Considering the extremely large amount of bird species (230) that are often indifferentiable to even humans and usually share many features, an accuracy of 22.4% denotes a decent unsupervised classifier. However, this accuracy was nowhere near the accuracy of the classic supervised algorithms like Logistic Regression and SVMs. One of the main drawbacks to our clustering algorithm that we did not anticipate was the significant use of the shapes of the birds and the background shapes/colors in clustering. Although similar birds are likely to be in similar environments or positions, it is not a very accurate classification feature for each and every bird because, for example, many birds have pictures of them flying or sitting on a perch, but that does not mean they are the same species. Supervised learning algorithms have a great advantage because they are able to determine which features are important in classifying different species. Furthermore, if we used this technique to classify animals of different biological families, such as dogs, cats, and horses, we would expect a very high accuracy because their features would be very different. Additionally, if we fewer numbers of bird species, say 10, it would be much easier to differentiate based on features because there would be fewer confusions between birds with similar features, and we would expect a much higher accuracy. Overall, we conclude that classification using unsupervised clustering algorithms is best avoided when there are many, very similar objects to classify. However, it is a viable solution if the objects are noticeably different or if there are only a few of them.

7.2 Limitations and Reflections

Our dataset is very large and there are 230 classes to be classified, and therefore our project is mainly limited by computing power, time, and manpower (we only have a 2-people team). Here's a few problems we encountered along the project:

1) Initially, we tried to reduce the extracted features further using PCA to speed up our machine learning process. However, the testing accuracy drops drastically to almost 1% after keeping 5000 components. We realize there are too many nuances between the bird species that the further "detail-oriented" components represent, and it is impractical to scale down the extracted features because it traded off accuracy too much. We believe this is because many birds share the same general features, so it is up to the details to differentiate them.

2) Since we do not reduce the number of components to a small number, the computation of our algorithms is extremely expensive and Colab keeps crashing on all algorithms. To improve the performance of the algorithm, I changed all the baseline logistic and SVM methods to stochastic gradient descent using logistic loss and hinge loss. The session does not crash and returns favorable results.

3) Running clustering algorithms on the entire dataset was time-consuming and often crashed, thus we broke the 36,045 training set into 20,000 training 8,022 testing and 8,023

validation. Due to crashing over the course of a couple days, we fell short on time so we were unable to do all the cluster sizes for GMM and trained the model on only 10 clusters and 230 clusters.

7.3 Future work

We believe there is immense potential of using clustering algorithm to discover similarities between images. A further impact we can make using the clustering is to collaborate with experts in biology to analyze the quality of our clusters using the biological scientific methods, on top of our visualization of representative pictures in each cluster.

We can collect more data about the geographical location, species relations, evolution, and habitats information of the birds and run clustering with multiple datasets, so we can better generalize results and make it more applicable in scientific discoveries.

As noted in section 5.2, using a full covariance matrix with GMM, while computationally expensive, could produce drastically better results because we believe that each cluster varies in many dimensions.

We can also use more advanced deep learning frameworks in image clustering, such as disentangling features to discover similarities in terms of birds posture, shapes, and background colors separately.

We are also interested in investigating interpretability techniques in clustering, such as LIME and SHAP, to explain the clustering results and highlight the important regions that determine the assigned cluster of an images.

We want to explore the different ways of representing image data rather than RGB, such as Hue-Saturation Value. This could yield more accurate clusters to classify bird species.

We would like to try clustering on different families of species, such cats, dogs, and horses, rather than just birds because it is likely easier to differentiate between different families of species rather than species within the same family. We believe this will result in a better classification accuracy.

References

[1] Gerry, "230 bird species," Dec 2020.

[2] aditya276, "Bird cnn vgg16 99 accuracy on test set," Jun 2020.

[3] Snives, "How to calculate clustering entropy? a working example or software code," Apr 2016.

[4] S. Shams, "K-means clustering," Apr 2018.