**Hierarchical Clustering of Corals using Image Clustering**

Rudra Sawant

46200264, sawantrudraarun@gmail.com

ABSTRACT

Several approaches have been taken by different scientists over the years to create taxonomy of coral species by looking at their morphology. On molecular examination, the taxonomies created have revealed to have incorrect classifications. In this project we aim to find a relationship between different types of corals and classify them by using image classification and clustering techniques on the QM coral dataset. We use the VGG16[], InceptionV3[] and DenseNet[] models which are pretrained on ImageNet dataset, to train and extract feature embeddings from coral dataset. These embeddings are then clustered using the Agglomerative Heirarchical Clustering[] to obtain a general hierarchy of corals. We show that DenseNet performs the best among the three models on Image classification task and can be used to extract the feature embeddings. Using Agglomerative Heirarchical Clustering with average link criterion on these embeddings, we can generate the general hierarchy of corals.

KEYWORDS

Image Classification, Hierarchical Clustering, Transfer learning

1. Introduction

Since the 1980s, analysis of the skeletal structure and morphology of the corals has been used by different scientists for the construction of coral taxonomy. Most of these methods used skilled human observers to manually establish the relationships between coral species. Although these methods have successful classified many corals, they often have inaccuracies and limitations when compared with the molecular study of the corals [1]. Image classifications algorithms today have proved to be better at pattern recognition tasks than humans and hence can be used to accurately classify the different types of corals [2] based on their morphology and texture. Some CNN-based approaches have been implemented to classify coral images and have achieved good accuracy. In this study we propose the use of state-of-the-art image classification models like VGG16, DenseNet and InceptionV3, to learn and extract important features embeddings from the dataset of coral images is provided by the [Hidden for Double Bind Review]. These embeddings when clustered using hierarchical clustering can represent the relationship between different coral types. This resultant clustering can be then used to get a general hierarchy between the different coral species. The coral dataset we use and contains images of corals belonging to 1006 different categories.

1. Literature Review

Many approaches have been proposed and implemented on the task of classification of corals. In the paper [6], the authors proposed a new type of CNN called MDNet, which used multi-scaled patches on point annotated data and dense connections between its layers to classify the coral images, instead of relying on pixel data. Another approach was proposed in the paper [7], which implemented a two-level classifier using ResNet models. A ResNet model in the first layer identifies whether the image is a texture or structure image and based on that, the coral species is identified using either a ResNet model trained on RSMAS texture dataset or a ResNet model trained on StructureRSMAS dataset. Though this approach assumed the independence of the texture and structural features in classification of corals, it had a good accuracy on classification task. Although the above-mentioned approaches were effective, they required considerable amounts of data to train the model to obtain good accuracy. Like most real-world datasets, the dataset to be used in this project is limited and hence the above approaches would not give a good classification performance.

The paper [8], shows that different transfer learning approaches based on ResNet, with few-shots learning and data augmentation can achieve good accuracy with a dataset having less than 100 labels per class. This shows that transfer learning can work well on image classification with very limited data. The paper [10] proposed a hybrid approach to transfer learning, where handcrafted features were used along with the features obtained from the pretrained VGGNet. There hand-crafted features were colour and texton based features selected from the same patch as the CNN based features. The model could achieve better performance than other benchmark models on MLC coral dataset. An alternative approach mention in paper [11], suggested an iterative approach using hand crafted features at the start. It used transfer learning and unsupervised learning in each iteration to choose reliable labels and optimize them. The above approaches using hand-crafted features face a limitation in cases where obtaining them is difficult, like in the case of our project. A promising solution to this problem was implemented in the paper [12] by adding two new layers to the pretrained CNN models and training them. It implements two different models, one that applied transfer learning from ImageNet dataset to coral texture images dataset and second which applied transfer learning from MLC-2008 coral dataset to coral texture images dataset. These approaches could achieve high accuracy making them candidates to be used in this project.

1. Methodology

Preprocessing

Train the top 2 layers

Train the top layer

Hierarchical Clustering

General Hierarchy

Compare performance

Mean features

Extract features

Hierarchical Clustering

Preprocessing: The dataset consisted of 8745 images of corals belonging to 1006 different categories. The images were taken from different sources, like in laboratory environment, in situ under the water and under microscope. Many of them contained background objects and colors. Some were also of very poor image quality and blur. To ensure uniformity and good image quality across images, only the images of corals taken in laboratory environment were used, since they did not contain any background and were of good image quality. Also, the coral categories containing too few images (less than 3) were removed to ensure that each class in adequately represented. After all this processing, a final dataset containing 3998 coral images belonging to 502 categories was obtained and used for project. Since the number of images per category was still low, different methods of data augmentation like rotation, horizontal and vertical flipping, zooming in, sheering, and changing the brightness were used to increase the dataset.

Timeline

Description automatically generated

Transfer Learning: Three different pretrained image classification models were trained on the coral dataset. These were VGG16[], InceptionV3[] and DenseNet[]. Each of these models have been pretrained on the ImageNet dataset[] and the resultant weights were imported and used as initial weights for training. The top layer of each of these models was replaced with a dense layer. Two approaches were taken in training the models. In the first approach, only the top layer of the models was training on the coral dataset. Whereas, in a second approach the top two layers of the models were trained on the coral dataset. The performance of all models using both the approaches was compared and the model with the best performance was selected for further processing.

Extracting feature embeddings: Once trained, the top layer of the model was removed and feature vectors of the images were extracted by passing each image as input to the new edited model. These embeddings represented the images in such a way that could capture the similarities (or differences) between them. The method of Principal Conponent Analysis (PCA) was used to reduce the dimensions of the embeddings while preserving 99 percent variance among them.

Heirarchical Clustering: The Agglomerative Heirarchical Clustering algorithm was used to cluster the feature embeddings. The clustering was implemented using the Single-link, Complete-link, Average-link and Ward’s criterion. The performance across these different clustering results was tested using the Silhouette score. The same process was repeated by calculating the mean feature embedding for each category and then clustering them. The more the Silhouette score the better the clustering, hence the clustering of the mean embeddings with the best silhouette score was chosen to best represent to general hierarchy between the coral categories.

1. Result

The processed coral dataset of 3998 images belonging to 502 coral categories was used to train each of the models (VGG16, InceptionV3 and Densenet). The performance of these models in shown in Table 1 and 2. Among these, the DenseNet model when trained on its top two layers gave the best performance, hence it was used to extract feature-embeddings from the images. The dimensionality of these features was reduced from 51192 to 428 dimensions using PCA. On performing Agglomerative clustering on these features, it was found that using average-link criterion gives the best silhouette score (see Table 3). The clustering of mean features calculated for each category also had the best silhouette score when average-link Agglomerative clustering was used. The resultant hierarchy of mean features was selected as the general coral hierarchy and is shown in Fig 1.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | VGG16 | InceptionV3 | DenseNet |  |  | VGG16 | InceptionV3 | DenseNet |
| Train Acc. | 46.55% | 76.32% | 70.07% | Train Acc. | 46.55% | 76.32% | 70.07% |
| Train Loss | 3.34 | 15.60 | 8.15 | Train Loss | 3.34 | 15.60 | 8.15 |
| Test Acc. | 20.44% | 24.79% | 28.81% | Test Acc. | 20.44% | 24.79% | 28.81% |

Table 1: Accuracy and loss of models with only Table 2: Accuracy and loss of models with the top

top layer trained. (Acc. stands for Accuracy) two layers trained. (Acc. stands for Accuracy)

|  |  |  |
| --- | --- | --- |
| Criterion | Mean vectors | Feature Vectors |
| Average-link | 0.3861 | 0.0132 |
| Ward’s | 0.0200 | 0.0132 |
| Complete-link | 0.3845 | 0.0137 |
| Single-link | 0.3843 | 0.0029 |

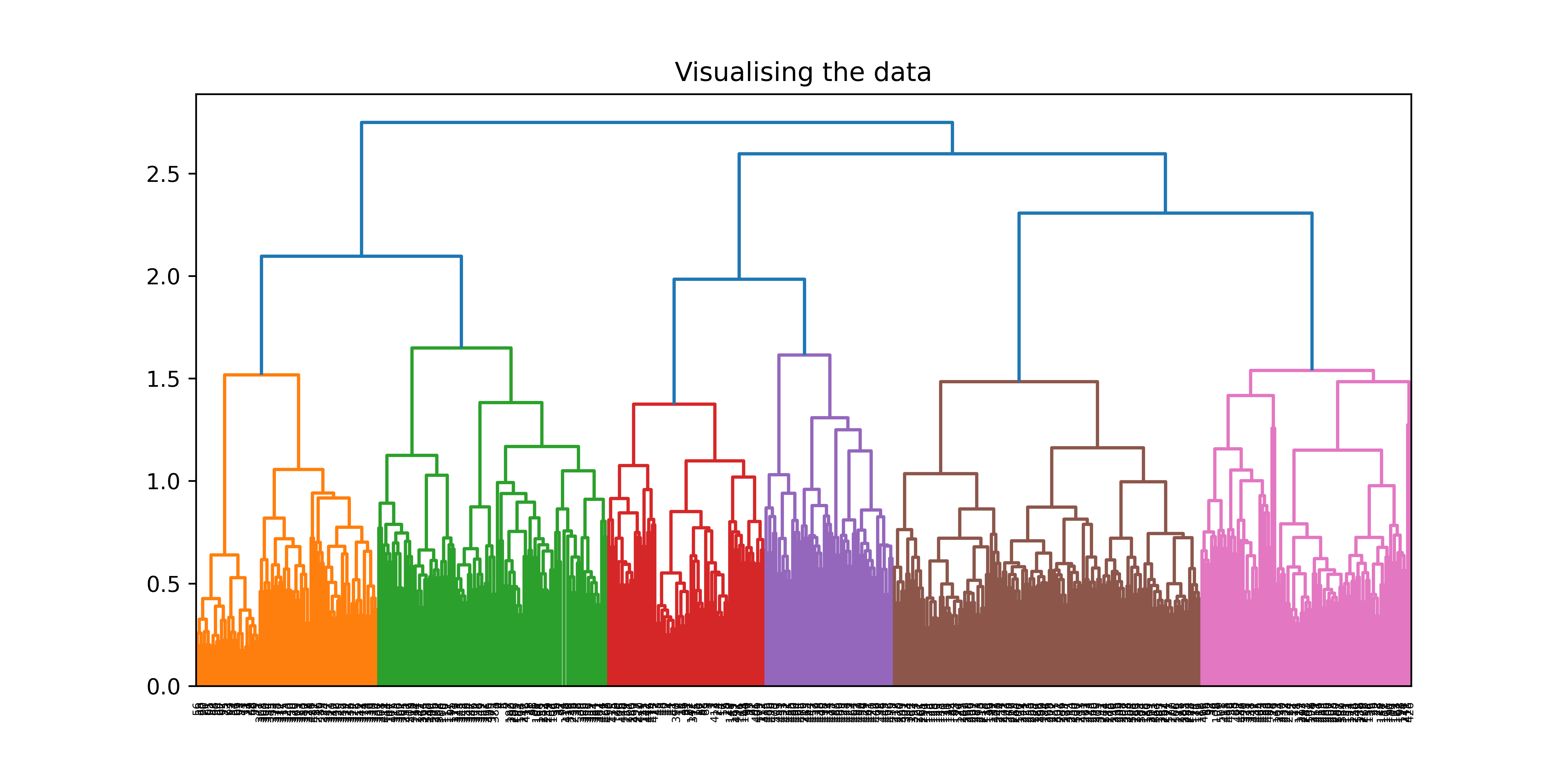
Table 3: Silhouette scores of hierarchical clustering.

Fig 1. Dendogram showing hierarchical relationships between coral categories (on horizontal axis are different coral categories)

1. Conclusion

As seen from the results, using the DenseNet Model on the coral dataset gives us the best performance on the task of classification of coral images into their categories. Although it had a high training accuracy of 88.74%, the test accuracy was low at 39.7%. This could be because of the low number of images available per category (~6 image per category) as well as the low diversity in data (most images in a category are of the same coral sample). The best clustering obtained had a silhouette score of 0.3861 suggesting some overlap between clusters. Although these results are average in terms of accuracy, they might still be very useful for finding the possible closely related categories to a given coral. It could also be used to find the how closely two corals are related. The methodology used proves that transfer learning can be very useful on coral classification tasks. In future work we need to collect more images per category to improve the quality features generated and get better clustering.

1. Acknowledgement

I acknowledge the motivation and guidance of my supervisor Prof. Gianluca Demartini. I also thank the Queensland Museum for providing the dataset and Dr. Tom Bridge for providing the domain knowledge.

* 1. Sub-section

Replace this text with your own

* + 1. Sub sub-section

Replace this text with your own

REFERENCES (ACM Referencing Format)

[1] Patricia S. Abril and Robert Plant. 2007. The patent holder's dilemma: Buy, sell, or troll? *Commun. ACM* 50, 1 (Jan. 2007), 36-44. DOI: https://doi.org/10.1145/1188913.1188915

[2] Sarah Cohen, Werner Nutt, and Yehoshua Sagic. 2007. Deciding equivalences among conjunctive aggregate queries. J. ACM 54, 2, Article 5 (April 2007), 50 pages. DOI: https://doi.org/10.1145/1219092.1219093.

[3] David Kosiur. 2001. *Understanding Policy-Based Networking* (2nd. ed.). Wiley, New York, NY.

[4] Sten Andler. 1979. Predicate path expressions. In *Proceedings of the 6th. ACM SIGACT-SIGPLAN Symposium on Principles of Programming Languages (POPL '79)*. ACM Press, New York, NY, 226-236. DOI: <https://doi.org/10.1145/567752.567774>.

1. J. Veron, “Overview of the taxonomy of zooxanthellate Scleractinia,” Zool. J. Linn. Soc., vol. 169, no. 3, pp. 485–508, 2013, doi: 10.1111/zoj.12076.
2. J. Pavlus, “Computers Now Recognize Patterns Better Than Humans Can,” Scientific American, 2016.
3. M. Modasshir, A. Q. Li, and I. Rekleitis, “MDNet: Multi-Patch Dense Network for Coral Classification,” Ocean. 2018 MTS/IEEE Charleston, Ocean 2018, pp. 1–6, 2019, doi: 10.1109/OCEANS.2018.8604478.
4. A. Gómez-Ríos, S. Tabik, J. Luengo, A. S. M. Shihavuddin, and F. Herrera, “Coral species identification with texture or structure images using a two-level classifier based on Convolutional Neural Networks,” Knowledge-Based Syst., vol. 184, p. 104891, 2019, doi: 10.1016/j.knosys.2019.104891.
5. W. Zhu, B. Braun, L. H. Chiang, and J. A. Romagnoli, “Investigation of Transfer Learning for Image Classification and Impact on Training Sample Size,” Chemom. Intell. Lab. Syst., vol. 211, no. February, p. 104269, 2021, doi: 10.1016/j.chemolab.2021.104269.
6. A. Gómez-Ríos, S. Tabik, J. Luengo, A. S. M. Shihavuddin, B. Krawczyk, and F. Herrera, “Towards highly accurate coral texture images classification using deep convolutional neural networks and data augmentation,” Expert Syst. Appl., vol. 118, pp. 315–328, 2019, doi: 10.1016/j.eswa.2018.10.010.