CSCI-SHU 360 Machine Learning: Final Project

# Multi-species Classification for Dolphin and Whales: A Deep Learning Approach

Bale Chen & Yuxuan Xia

## Introduction

In Marine Science, one of the challenges that scientists face is the identification of certain animal species. Most of the researchers still manually identify different species by eyes. Although experts have high accuracy in this task, it would save much time if it can be done by computers.

Inspired by the recent development of the computer vision area, we want to leverage this project as a way to learn and practice the traditional as well as the cutting-edge models in the field of computer vision. We trained multiple Machine Learning and Deep Learning models to complete the Dolphin and Whale species classification task.

More specifically, we picked 6 subspecies belonging to the whale family and the dolphin family. The task is to learn from the images, mostly depicting the fin and lateral body part of the animal, and perform species identification. The real-world usability of our project can be reducing the manual labor of identifying the marine animal species.

## Data Description

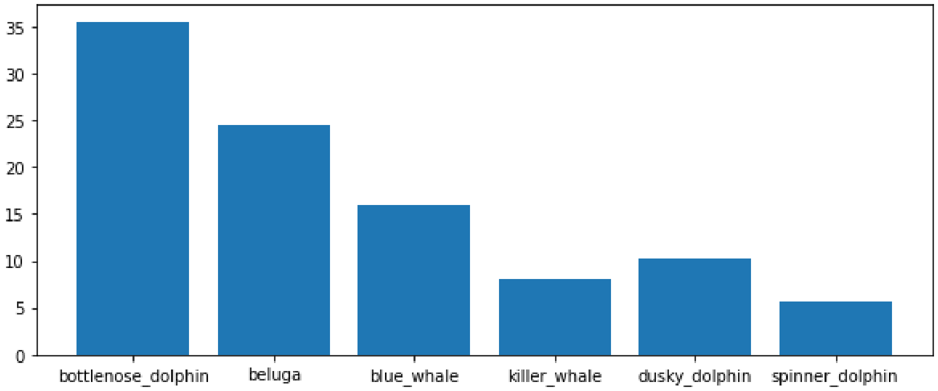
We use the dataset originally from "Happywhale - Whale and Dolphin Identification" Kaggle competition (link). For a more structured and easy-to-implement data, we found the resized whale-dolphin images dataset (contributed by "RDIZZL3") (link). The contributor resized the images to 128×128.

Firstly, we sample a small dataset containing bottlenose dolphins and killer whales. 1500 trains and 500 tests for each category. This is a small dataset just for trial because the small number of samples may not be that representative of the overall performance of the models.

Secondly, we picked the 6 species that have the most number of images. Totally 30348 128x128 RGB standard Images. 6 species are as follows. Dolphin: Bottlenose Dolphin, Dusky Dolphin, and Spinner Dolphin. Whales: Beluga, Blue Whale, Killer Whale.

## Data Preprocessing

In the original dataset, there exists some typos in the species names. Also, the original images include species other than the 6 we selected. Therefore, we cleaned the data when loading them. The training set and validation set has been splitted by 4:1 ratio. For classification tasks with observation data, one of the most common issues is data imbalanceness. With some classes having scarce data compared to others, the models can extract limited information from the training set, which causes inaccuracy. We can see the imbalance of our dataset from the following figure.



Thus, we adopted the Synthetic Minority Over-sampling Technique (SMOTE) to generate a balanced training set. It randomly generates new samples on the line segments from one datapoint to its nearest neighbors. However, this technique was not helpful in our task. On KNN, Logistic Regression, Decision Tree, and ResNet50, the performance dropped by 5% - 10%. We recognize that the imbalance in our dataset might not be very significant and SMOTE creates noisy or irrelevant data points that disturbs the learning process. Another reason can be that the synthetized images are not realistic marine animal images, so that SMOTE adds false data-label pairs to the training set. Therefore, we decided to continue without oversampling.

## Models

We trained a wide range of models. The traditional machine learning models include KNN, Decision Tree, and Logistic regression. The CNN models include AlexNet, VGGNet, ResNet, and Vision Transformer (ViT).

The reason why we chose KNN, Decision Tree and Logistic Regression is that these three represent three types of classification methods in traditional machine learning – Nearest Neighbors, Tree-based Method, Linear Separator. The CNN models we selected are the classical models of the image classification task. Further reasons for selection are here:

1. AlexNet, VGGNet: Among the most foundational and widely-used models in the Computer Vision area.
2. VGGNet is one of the foundation models in the CV area and it is very powerful.
3. ResNet: A very powerful and efficient model of the image classification task. It deals with the problem that the deeper model has worse accuracy. It should be the deeper we have, the more information we got. So ResNet introduced the residual connection to the CNN to gain new state-of-the-art performance on different sets.
4. ViT: A recent model that borrows the idea from the powerful NLP model Transformer. It cuts the image to different 16x16 patches and feeds the patches to the Transformer encoder to do the classification. It has the self-attention mechanism when doing the classification, which might be the potential to reach good performance.

## Results and Interpretation

### Small Sample

For the small dataset, we got the accuracy of the test set as follows:

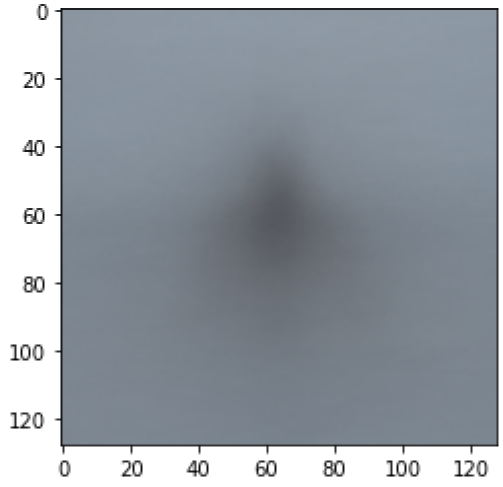
| Models | Accuracy |
|---|---|
| KNN | 80.9% |

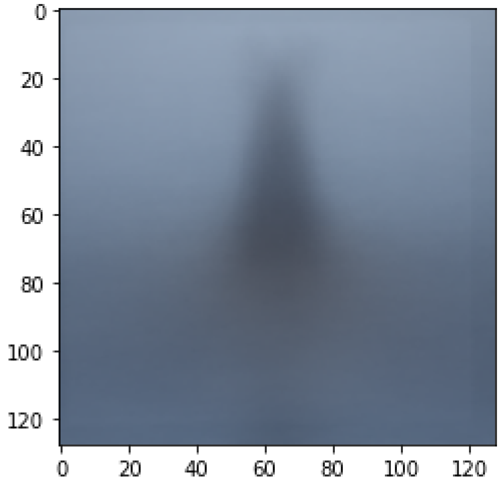
| | |
|---|---|
| Decision Tree | 74.3% |
| Logistic Regression | 82.7% |
| ResNet 50 | 88.4% |
| Vision Transformer | 79.3% |

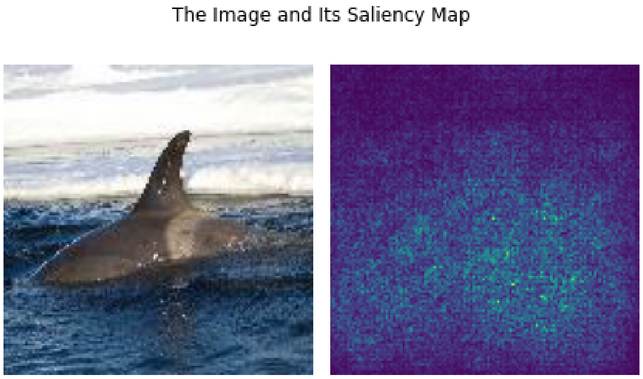
This result is a little bit unexpected.

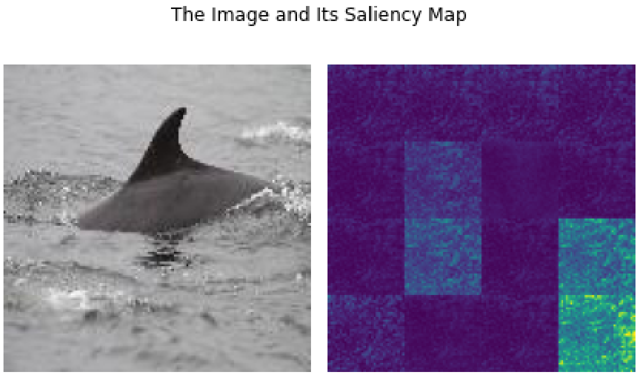
1. The good performance of the non-CNN models.
2. ViT’s performance is way less than the ResNet50

Interpretation of the first unexpected: We can see logistic regression, a linear model can even outperform the ViT model. This is surprising. Here we give the interpretation of this result. I take the summation of all the images of each category in the small dataset and get two average images of the two categories, bottlenose dolphin and killer whale. And here are the two images, the left one is a dolphin and the right one is a killer whale. We can see that the two images stand for the distinguishing feature of these two species. Dolphin’s fin is smaller than that of the killer whale, which is very clear in the average image. And the background is very net, which means the background in the whole dataset is very similar, thus the only difference will be the fin. So, these two kinds of images are very likely to be linear-separable at the pixel level. This is the reason why the non-CNN models can perform so well by just training on the pixel values.









Interpretation of the second unexpected: For the CNN models, I check the saliency map. Here are the saliency map. Saliency map is stating which pixel contributes more to the classification scores. The brighter the pixel in the image, the more important the pixel is when deriving the classification result.  The left one is ResNet and the right one is ViT. We can see why ViT's performance is worse than the ResNet. The ViT does not activate the fin or the body part of the dolphin, it mainly activates the bottom right corner, which is the background. However, the ResNet's activation mainly concentrates on the center. This includes the body and the fin part of the dolphin.

## Whole Dataset

On the whole dataset, we firstly set the baseline performance using the three traditional Machine Learning models (KNN, Decision Tree, Logistic Regression). Logistic Regression scores 73.73% of accuracy and ranks the top among the three. Then, we performed Transfer Learning and normal training using AlexNet, VGG16, and ResNet50. Using the pretrained models from pytorch and transferring them to do marine species classification works poorly. AlexNet and VGG16 didn't even pass the baseline set by Logistic Regression. ResNet50 narrowly passes with approximately 75% of accuracy. By contrast, if we train from scratch, the three neural networks score much higher. We can see the exact accuracy of the models in the following table.

| | AlexNet | VGG16 | ResNet50 |
|---|---|---|---|
| Pretrained | 59.427% | 62.894% | 75.097% |

| | | | |
|---|---|---|---|
| Not Pretrained | 87.809% | 88.056% | 85.255% |

Here we encounter two questions: Why Transfer Learning doesn’t work and why AlexNet and VGG16 perform better than ResNet50.

First, Transfer Learning means transferring one trained model to do a similar but different task. In our case, we used the pretrained model from PyTorch to perform dolphin/whale subspecies classification. The PyTorch models are trained on ImageNET, which is a large domain of images. However, if there’s a domain shift, which means our image domain is different from the domain where the transferred model is pre-trained on, the model might perform worse than the training-from-scratch approach.

Second, although ResNet50 is more recent and complex than AlexNet and VGG16, it can also be outperformed by the relatively simpler Neural Networks. One possible reason is that the ResNet might suffer overfitting since it has a much deeper architecture than AlexNet and VGG16. The other possible reason is that ResNet might just not be suitable for dolphin/whale species classification tasks, but rather good at feature extraction. We see this from the pretrained case where ResNet performs the task better than VGG16 and AlexNet. It potentially indicates that ResNet by its nature can extract more valuable information from the images than the simpler models

## Future Improvement

We spent a lot of time changing the ViT model. Inspired by Prof. Mathieu’s idea that the advantage of CNN is that it remains the 2d structure of the image. Because ViT flattens all the patches at the first step, I am wondering whether it will perform better if I remain the 2d structure of all the patches. My model is in the notebook “Code/Small Set/ViT Without Flattening.ipynb”. However, I encountered tons of unexpected bugs when I change the model. Now it can get the output if I input some random tensors. But it has problems with cuda when I want to train the model. This is one of the future improvements that we can make.

Also we can add more species to train the model or try other types of oversample methods. For example, data augmentation. We can feed the image after the rotation, or add filters to the model to make up the unbalanced dataset.

## Conclusion

In our Machine Learning final project, we attempted to address the problem of whale/dolphin subspecies classification using fundamental and cutting-edge neural networks. We found that in small samples with two species, the task wasn't too hard even for traditional Machine Learning models without any convolutional operation. However, when it comes to the whole dataset with 6 species, the traditional models only achieve 62.14% - 73.73% accuracy, which is far from satisfactory. In this case, we have to resort to Deep Learning techniques.

We tried AlexNet, VGG16, and ResNet on the whole dataset using Transfer Learning and normal training. The Transfer Learning technique doesn't work very well potentially due to domain shift issues. Using normal training, VGG16 is the best among the selected models, scoring 88.056% of accuracy rate. We also notice that the relatively more recent model ResNet50 doesn't perform as well, possibly due to overfitting or the fact that it's not suitable for our proposed task.

## Reference

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929, 2020.

B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

Chawla, N., K. Bowyer, Lawrence O. Hall and W. Philip Kegelmeyer. "SMOTE: Synthetic Minority Over-sampling Technique." J. Artif. Intell. Res. 16 (2002): 321-357.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In CVPR, 2016.

Simonyan, Karen, Vedaldi, Andrea, and Zisserman, Andrew. Deep inside convolutional networks: Visualising image classification models and saliency maps. CoRR, 2013.