Project Title: Humpback Whale Identification Challenge

Team Member: Yuan Meng

Project Summary

The project is to build a deep learning model to identify different whale species by recognizing the shape of whales' tails and unique markings found in footage using over 25,000 images stored in a database called Happy Whale. The human activity of killing whales for oil and food has decreased the population of whales in the world, especially humpback whales. For a long time, oceanographers and marine scientists have photo surveilled ocean activity to keep a log on the population of whales, but this is a slow and work-intensive way of manually keeping a record. Therefore, a faster and more effective way of identifying whale species could help a lot.

Project Description

1.0 Problem definition

The most precise data science field this project relates to is image recognition. That is because each kind of whale has some special physical characteristics. These characteristics can be identified to determine the whale species.

1.1 Dataset description

Collected from the Happywhale Organization [1], the Kaggle [2] dataset involves both training and testing datasets. The training dataset for whale identification includes JPG files and CSV files. The 9580 images are humpback whale flukes' photographs, and the CSV file is to match each image with the relating whale ID. If the individual whale has been identified by researchers, it will have an ID, such as "w_1287fbc". Otherwise, the individual whale will be labeled as a new whale with the ID "new_whale". Among the 9580 images, 8% are new whales and the rest belong to 4250 identified humpback whales, which means that there are only a few examples for each of the whale IDs. The testing data, then, includes 15610 humpback whale fluke images. Therefore, it is large enough to train a deep network, and my goal is to predict such 15610 images' whale ID based on the 9580 labeled images.

1.2 Evaluation metrics

According to the Kaggle [2] evaluation rules, the performance of a method can be evaluated by the Mean Average Precision $MAP = \frac{\sum_{q=1}^{Q} AveP(q)}{Q}$, where Q is the number of queries. And precision is the positive predictive value (PPV), where $Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$. For my prediction result, I will predict 5 labels for the whale ID for each image, so Q is 5 here. The higher the MAP is, the better my method is.

2.0 Network and framework

I will choose Convolutional Neural Network (CNN), one of the most breakthroughs in computer vision used for image detection and analysis, as the deep network for this project.

I plan both to try some standard forms of the network like applying the architectures of ResNet[3], and to try the customized networks built by myself to see which one is a better fit for this project.

As for the framework, I prefer to use Pytorch. Just like what I learned in class, it is easy to implement and understand the code as it's similar to NumPy[4] library. Also, Pytorch is more flexible due to its dynamic computational graph, and it allows to build networks which structure is dependent on the computation itself. Most important, the ResNet networks are built with the Pytorch library, so I would like to choose this one for easier comparison.

3.0 Reference materials

Related work has been adapted and compiled from three sources: one research studies done on humpback whale identification methods at Universidad Politécnica de Madrid in Spain and two done at Microsoft and New York University respectively. These research studies have evaluated neural networks with various image recognition applications to deeper extents and thus closely related to my target problem. I can obtain sufficient background on the projects from the below related work.

1. *Humpback whale identification with convolutional neural networks by* Universidad Politécnica de Madrid*. (2018). [5]*
2. *Deep Residual Learning for Image Recognition by Microsoft Research Group (Kaiming, Xiangyu, Shaoqing, Jian). (2015). [6]*
3. *Learning a Similarity Metric Discriminatively with Application to Face Verification by Courant Institute of Mathematical Sciences at New York University. (2017). [7]*

4.0 Schedule the project

As I am the only person on my team, I do not have to meet with other people and coordinate their time. Here is my rough timetable for completing the project.

Date		Activity
Week 1	04/05/2021 - 04/12/2021	Read related materials Data processing
Week 2	04/12/2021 - 04/19/2021	Building, training, and testing the model Keep trying and adjusting the network and parameters to find the best one
Week 3	04/19/2021 - 04/27/2021	Final testing and arranging all works Report writing Presentation recording GitHub uploading Link submission

References

1. Happywhale. https://happywhale.com/home

2. Kaggle Humpback Whale Identification Challenge.
https://www.kaggle.com/c/whale-categorization-playground/overview

3. ResNet. https://pytorch.org/hub/pytorch_vision_resnet/

4. Numpy. https://numpy.org/doc/stable/user/index.html

5. Can, D. (2018). Humpback whale identification with convolutional neural networks.
https://www.preprints.org/manuscript/201902.0257/v2

6. ResNet paper: Deep Residual Learning for Image Recognition.
https://arxiv.org/pdf/1512.03385.pdf

7. Learning a Similarity Metric Discriminatively with Application to Face Verification.
http://yann.lecun.com/exdb/publis/pdf/chopra-05.pdf