Humpback Whale Identification

Project in Machine Learning

Riigo Kärner, Timo Tiirats, Hannula-Katrin Pandis

Who are we?

- Riigo Kärner - masters curriculum in Data Science
- Hannula-Katrin Pandis - masters curriculum in Data Science
- Timo Tiirats - masters curriculum in Computer Science

Humpback Whale Identification

- https://www.kaggle.com/c/humpback-whale-identification/overview
- Can you identify a whale by its tail?
- Kaggle competition, ended in February 2019;
- 2129 submissions;
- Happywhale platform providing data.

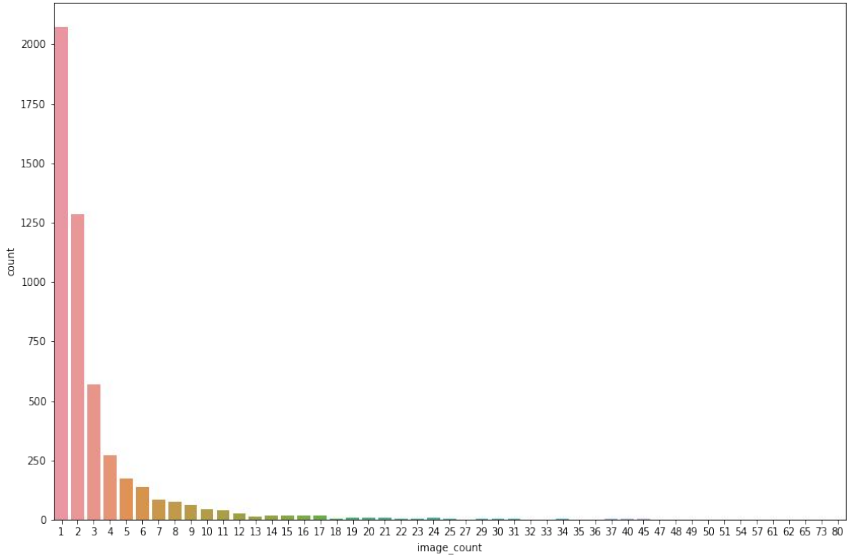
The problem

- Whales are endangered and protected by an international law;
- Scientists use photo surveillance systems to monitor ocean activity - they use the shape of whales' tails and unique marking in the footage to identify the species of the whales;
- For 40 years most of the work has been done manually;
- Aim of this project was to create an algorithm to identify the whales by its tales.

Whale ID: w_f48451c

Data

- 25 361 images, total 5 005 classes for training;
- 9 664 images out of training data have a class “new_whale” - unidentified;
- Remaining 15 697 images for training, in 5 004 classes;
- 7 960 images for test;
- Training data images sizes vary tremendously - from 6KB to 2000KB;

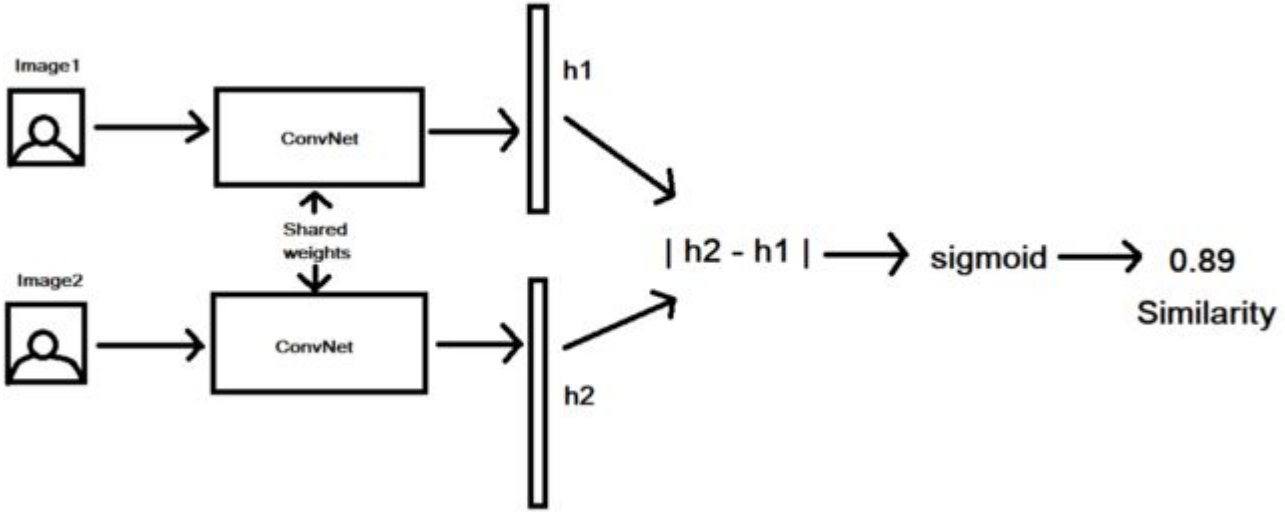


https://www.kaggle.com/kretes/eda-distributions-images-and-no-duplicates

Methods

- Discarded new_whale images from the training (in the beginning);
- Resized other 15k images into 128x128 pixels and 256x256 pixels, used grayscaling;
- Tried “Pure” Convolutional Neural Network (CNN);
- Used Siamese Neural Networks for One-shot Image Recognition (SNN);
- Tried several amount of pairs for one image:
    - Same ID pairs - labeled as 1.
    - Different ID pairs - labeled as 0.
- Trained and validated the model;
- Tested the model;

SNN Architecture

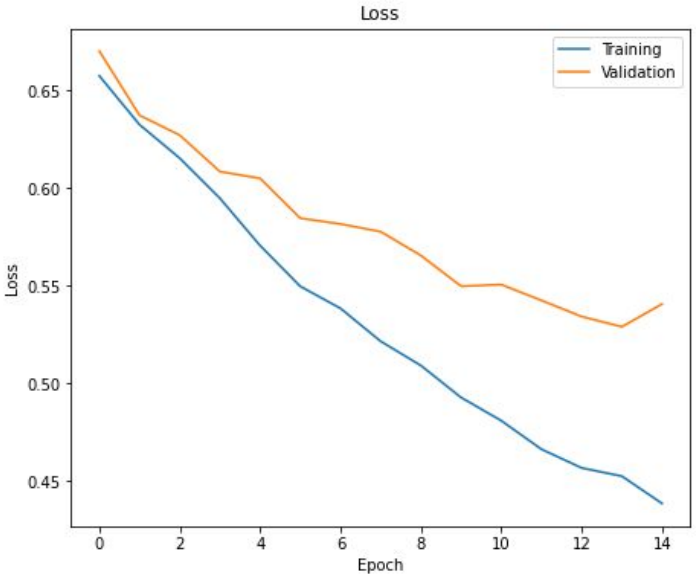


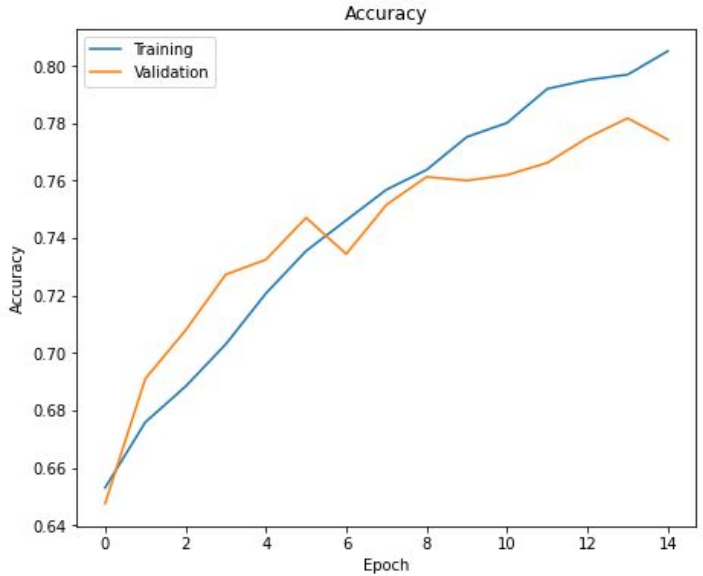
https://medium.com/predict/face-recognition-from-scratch-using-siamese-networks-and-tensorflow-df03e32f8cd0

Results

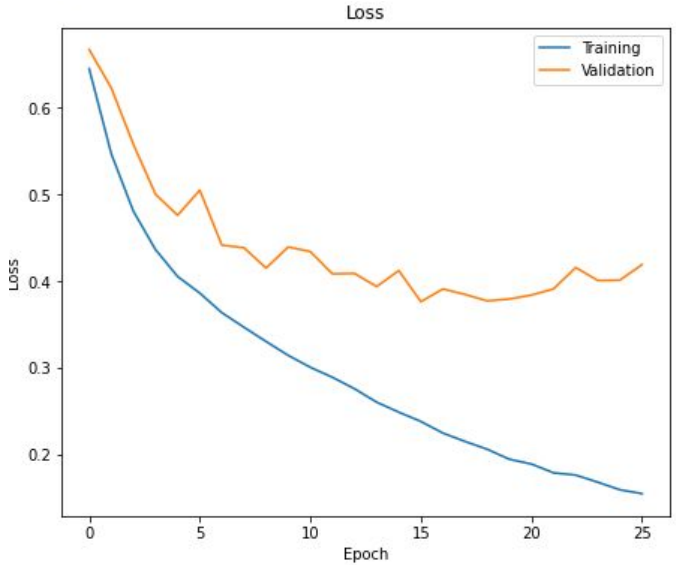
- This challenge was beyond our capabilities - we do not have much experience with neural networks;
- “Pure” CNN did not give any results - this approach is hard for this kind of a task;
- SNN gave some results while training the model, tended to overfit:
  - Tried to tune different parameters (amount of layers, learning rate) and added regularizations - validation loss decreased to 0.55 and validation accuracy increased to 0.78.
  - Used more images (augmented) - val loss decreased to 0.4 and accuracy increased to 0.85.
  - Tried to use bigger images - created memory issues.
  - Used train-on-batch method with random pair making within decreased training set - results similar to smaller images.
- Predictions with our trained models were not accurate:
  - Similarity scores were high for a lot of images, mostly because of the noise in the image - surrounding ocean.
  - ~27% of the test images had a label new_whale, so we had to predict 5004 labels to ~6k images.

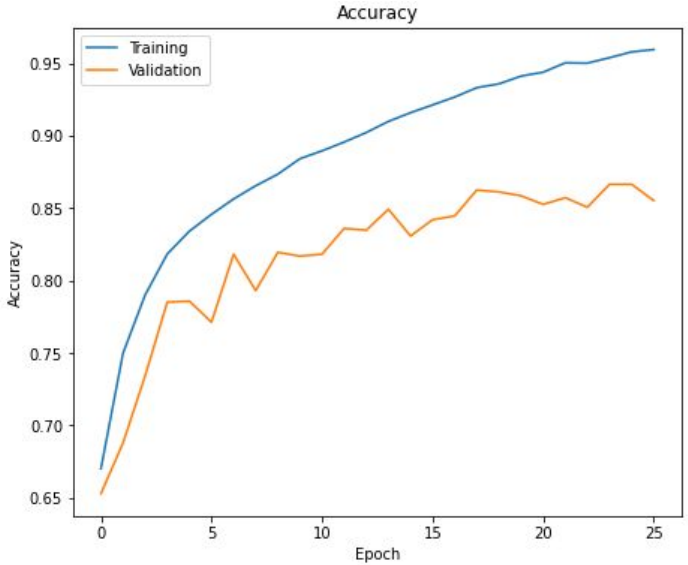
SNN loss and accuracy with 128px images





SNN loss and accuracy with 128px images including augmented ones (~32k images in pairs)





Main lessons

- Training and testing the model takes more time than expected:
  - Neural networks have a lot of parameters to tune and possibilities to use.
- Good coding skills give a great advantage;
- Training big neural networks takes a lot of memory and needs a lot of computing power;

Link

- GIT repository: https://github.com/TiiratsT/HumpbackWhaleIdentification;

THANK YOU!