Machine Learning Engineer Nanodegree

Capstone Proposal

Marcel Schneider March 17th, 2019

Proposal

Domain Background

I will work on an reduced version (due to data size) of the Humpback Whale Identification problem from Kaggle (https://www.kaggle.com/c/humpback-whale-identification (https://www.kaggle.com/c/humpback-whale-identification)). Whales are identified by scientists using photos of their tails. This has been done manually for the last 40 years and was a very time consuming task. As the research on whales has to be improved as they are dying slowly due to global warming and other matters. In this challenge I will try to design a neural network which can identify whales by their thales and thus will give time to scientists to do more important research. I'm personally interested in this project as I'm an environmental activist and as the beautiful wildlife on our planet needs to be protected.

Problem Statement

I'm trying to identify whales by using a relevant sub amount of the pictures given in the kaggle challenge. I will try to reduce the amount of data by excluding severeal whales. One challange of the project is to identify new whales as new whales. Another is to identify whales already existing in the databae. Another problem is that the initial train data set has to be split in train, validation and test data as the original test data is not labelled.

I will measure my sucess rate by using the formula provided by kaggle which is called Mean Average Precision. I will try to reach a mean average precision above 0.85.

Datasets and Inputs

(approx. 2-3 paragraphs)

The training data contains thousands of images of humpback whale flukes. Individual whales have been identified by researchers and given an Id. The challenge is to predict the whale Id of images in the test set. What makes this such a challenge is that there are only a few examples for each of 3,000+ whale Ids.

The following files were originally provided: train.zip - a folder containing the training images train.csv - maps the training Image to the appropriate whale Id. Whales that are not predicted to have a label identified in the training data should be labeled as new_whale. test.zip - a folder containing the test images to predict the whale Id sample_submission.csv - a sample submission file in the correct format.

I will hand in a relevant subset for this projecgt.

My results will be provided by using the sample_submission.csv. I will also try to plot relevant result data.

Solution Statement

I will solve this problem by using deep neural netwoorks for image classification. I will evaluate different network architectures based on transfer learning and being designed from scratch.

Benchmark Model

I will compare to other kernels provided on kaggle.

Evaluation Metrics

I will evaluate the problem by using Mean Average Precision.

$$MAP@5 = \frac{1}{U}\sum_{u=1}^{U}\sum_{k=1}^{min(n,5)} P(k) \times rel(k)$$

where U is the number of images, $P(k)$ is the precision at cutoff k, n is the number predictions per image, and $rel(k)$ is an indicator function equaling 1 if the item at rank k is a relevant (correct) label, zero otherwise.

Once a correct label has been scored for *an observation*, that label is no longer considered relevant for that observation, and additional predictions of that label are skipped in the calculation. For example, if the correct label is `A` for an observation, the following predictions all score an average precision of `1.0`.

```
[A, B, C, D, E]
[A, A, A, A, A]
[A, B, A, C, A]
```

Project Design

Major achivements of this project are:

- Reducing the data set size by removing 66% percent of the images and whales
- Importing data
- Data screening and cleansing
- Implementing multiple neural networks.
- Implementing the evaluation metric using https://www.kaggle.com/pestipeti/explanation-of-map5-scoring-metric (https://www.kaggle.com/pestipeti/explanation-of-map5-scoring-metric)
- evaluation of results (repeat :))

Before submitting your proposal, ask yourself. . .

- Does the proposal you have written follow a well-organized structure similar to that of the project template?
- Is each section (particularly **Solution Statement** and **Project Design**) written in a clear, concise and specific fashion? Are there any ambiguous terms or phrases that need clarification?
- Would the intended audience of your project be able to understand your proposal?
- Have you properly proofread your proposal to assure there are minimal grammatical and spelling mistakes?
- Are all the resources used for this project correctly cited and referenced?