Audio Based Bird Species Identification using Deep Learning Techniques

DIT-867 Assignment 3 - written part 1

Author
Erik Rosvall
gusrosver@student.gu.se

Abstract

There have been many attempts to use machine learning to distinguish birds. The different attempts have worked, however the accuracy have not been the best. This new technique that has been represented has proven to be quite effective. This short article will give a general summation of the highlights that made this method so effective.

1 Introduction

Over the years, researchers have used machine learning to train models to distinguish between different spices of birds. This study focuses on how to distinguish between different kinds of birds by training a neural network based on bird sound. The sound is divided into two groups, bird sound and the noise. The study is focused on a large scale sample of birds. While the study was conducted, there were some problems such as background noise, More then one bird species in the sound file, sizes of the sound files; to mention a few. To speed up the tuning of the parameters, the researchers took a small size of the data to test on (Sprengel et al., 2016).

2 Data

This section will highlight the important areas affecting the data, like the data set.

2.1 Data set

The data for the study was provided by the Xeno-Canto foundation, the data that were used were bird sounds. The data set contained 999 different spices of birds. The data set contained over 24 thousand files, approximately 25 files (bird sounds) per bird spices. The data were split into sound and noise categories. The sound file were clear sound of a specific bird and the noise file contained mixed sounds i.e. two diffident bird sounds.

2.2 Data processing

To make the data usable for the model, the sound file were runned through a short-time Fourier transform (STFT). Afterwards they picked in the generated spectrogram based on sizes of rows and columns, the values was converted to a range of [0,1]. By doing these steps, the researchers got a could cherry pick the important parts of the spectrogram. Furthermore, the data was split to chunks of a fixed size to fit the architecture of the neural network. These chunks were then divided into training and test sets. An other benefit of dividing into equal sized chunks is to avoid empty chunks, this means that every sample for the training and testing will be unique (Sprengel et al., 2016).

2.3 Data: errors

To reduce the classification error, the researchers used a vertical shift to the data file, in other words the pitch in the sound were shifted. However, the shift were small, the paper mentions that larger shift did not give any benefits. Furthermore, the researches added two files together, this resulted in that the neural network could see more key patterns at the same time. A consequence of this was a slightly improvement in the models accuracy.

3 Model

This study uses a neural network to classify different birds. The study is using a new technique since there have been multiple attempts to solve this problem with different machine learning techniques. The most common methods used are nearest neighbour, decision trees.

3.1 Tuning parameters

Since the data set contained 999 different types of birds there became a long waiting time between each run. To speed up this process, a subset of the original data set were picked to speed up the fine-tuning, since there was a lot of re-training.

Afterwards the fine-tuning was done, the neural network was trained with a sound file and a combination of different noise files on top to make the neural network find more relevant patterns, by applying this technique, the accuracy did also get a small boost.

The data for that are used to train the model were sorted into batch sizes of 8 or 16, the problem why there were two different sizes of the batches depends on the limitation of the GPU memory (Sprengel et al., 2016).

3.2 Weight functions

To update the weights in the neural network, the Nesterov momentum method where applied. The momentum was 0.9, initial learning rate is 0.1. After a couple of days of training (approximately 100 epochs) the learning rate got reduces to 0.01 (Sprengel et al., 2016).

A way to reduce errors in the training of the neural network, a single bird sound file were trained with a set of multiple different noise files. This resulted in a reduction of errors. Regularization were also tested during the study, both L1 and L2 was tested. However this tuning did not give any better performance on the contrary it slowed down the training.

To improve the model, the authors mentions a few alternatives. One of the alternatives is to modify the cost function also consider background birds and/or just a single birds sound instead of just a fixed sample size. They also mention that there is a possibility to use an ensemble of the existing neural network.

4 Result

4.1 Training

During The training phase, with around 50 different spices the MAP score was approximately 0.84.

4.2 Validation

The test were constructed by a neural network, the neural network had five conventional layers and the last layer was a dense. However, this new technique with the neural network has the highest mean average precision (MAP), the score was 0.69 and the accuracy score of 0.58 (Sprengel et al., 2016).

5 Conclusion

Our final thoughts are expressed here. The article was objectively written i.e. the authors mention both the good and problematic sides respectively. The problem it self is quite interesting, in other words this is not the first thing that comes to mind when reading about machine learning. This kind of research gives a sense of all the possibilities that exist for machine learning.

However, I found it interesting how they came up with clever techniques to fine-tune the model, I thought the model needed a large data set to get accurate values to fine-tune.

The authors mentions that there are a loot of room to improve the model, e.g make an ensemble of the neural network or modify the weight function. Even when they have the highest score for now. They also mention problem with the GPU memory. This is an aspect which can have a huge difference on the speed of the batch sizes and speed of the model. I did not know that a GPU could have this affect on an model.

References

Elias Sprengel, Martin Jaggi, Yannic Kilcher, and Thomas Hofmann. 2016. Audio based bird species identification using deep learning techniques. Technical report.