Bird Species Classification on MLSP Dataset

- Deep Modh (140050002)
- Himanshu Agarwal (163050001)
- Prashanth Manjunath (163050043)
- Vivek Kumar Arya (163050087)

Advanced Machine Learning Project, Spring 2018, IIT Bombay

Introduction

- Bird behaviour and population trends have been of interest to wild life scientist.
- This can be helpful in predicting various environment changes.
- Earlier this was mostly done by human efforts.
- Recent trend involve application of machine learning techniques which has helped automate this tedious process.

Introduction

- MLSP 2013 is a relevant kaggle competition, which is a multi-label bird species classification task.
- Given audio samples, consisting of bird sounds that are present in audio clips along with background noise.
- Classify the set of bird species which are present in an audio recording.

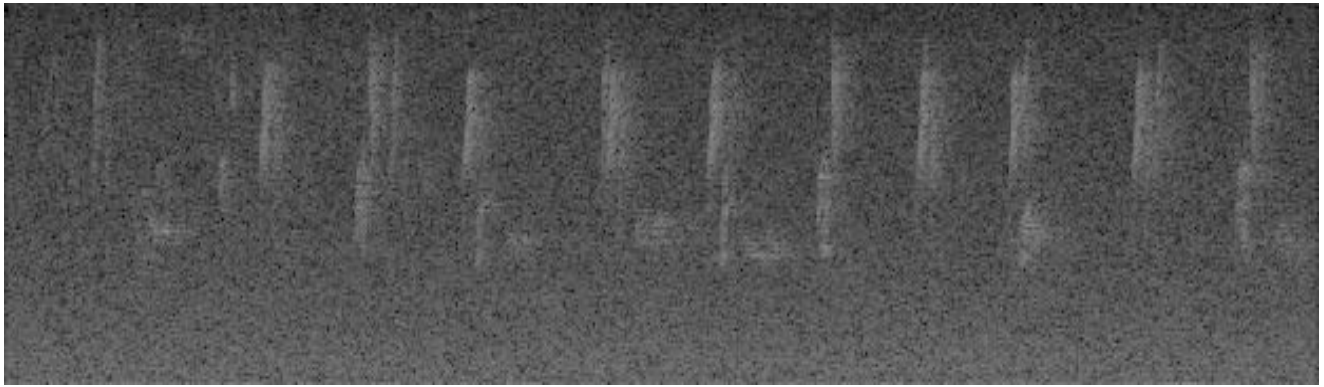
Dataset

- The dataset consists of 645 audio recordings of 10 seconds each.
- Multiple and variable number of bird sounds are present along with background noise.
- Out of these, **322** audio recordings are for **training** and **323** audio recordings are **test** data.
- There are a total of 19 labels corresponding to each bird.
- Need to predict the probabilities of each bird present in the given recording. Hence, this problem becomes a **multi-label supervised classification task**.

Related Literature

- Feature Extraction and audio processing -
  - **Spectrogram Images :**
    - Spectrogram is a visual spectrum of the frequency of an audio as they vary with time. Spectrograms might be three dimensional with 3rd dimension representing the amplitude[1][2].
  - **Raw audio feature extraction:**
    - Extracting various features from the raw audio such as MFCCs, Chromagram, Melspectrogram etc using some audio processing tools[3].

Sampled Spectrogram Image:



Gray spectrogram of PC1_20090606_050012_0020.wav

Related Literature

- Classification models:

  Based on the two described approaches for feature representation, two specialized kind of models can be learned

  - Image Classification Networks:
    - In this approach, audio waveform is feeded to a neural network architectures specialized in image processing like CNNs, LSTMs and variants[1][2].
  - Conventional 1D feature architectures :
    - In this approach, audio features are 1 dimensional vectors, that can be fed to various machine learning models like RF, Decision trees, DNN, CNN[3] etc

Related Literature

- MLSP 2013 winning solution:
  - This competition has a best score of 0.95611
  - It generates the spectrograms, followed by complex image processing, finally to get a 1 dimensional feature vector that captures the important variations in the audio waveform.
  - It then implements a random forest classifier for the bird label classification task[4].
  - Need extreme domain expertise in image processing to get the relevant features.
  - Advantage of deep neural networks is that there is no need for domain expertise and we can concentrate more on relevant neural network architectures.

Experiments and results:

- LSTM model:
  - LSTMs have been designed for sequence prediction in spatial inputs such as images, videos and have been observed to perform well on image tasks.
  - They are used in RNNs to effectively model the state using input, output and forget gates.
  - We resize our gray spectrogram images to (623,128), fed as input to an LSTM model.
  - Our model has three layers having 32 hidden nodes each, followed by a 19 dimensional output layer corresponding to 19 bird species.
  - Kaggle scores of our LSTM model are shown below -

Image Size	lstm layer1	lstm layer2	lstm layer3	Kaggle Score
(310, 64)	32	32	32	0.69907
(623, 128)	32	32	32	0.70101

Experiments and results

- Bidirectional LSTM:
  - Based on similar settings, we fed our two dimensional spectrogram images to bidirectional LSTM model
  - We fine tuned the dimensionality of the hidden layers, and spectrograms and three of the configurations are reported in the table below.

Image Size	Bidirectional layer1	Bidirectional layer2	Kaggle Score
(623,128)	32	32	0.70072
(623,128)	64	64	0.68964
(310,64)	32	32	0.69848

Experiments and results

- VGG-like neural network:
  - Vgg Network is a standard cnn architecture from the visual geometry group, which has shown to perform well on images.
  - It comes in two variants 16-layer and 19-layer.
  - We ran a less dimensional variant of vgg network, consisting of 4 convolution layers, 2 max-pooling layers, 1 fully connected dense layer, along with dropout.
  - After fine tuning the final layers were as: Conv2D 64, Conv2D 64, MaxPool 2x2, Dropout-0.25, Conv2D 64, Conv2D 64, MaxPool 2x2, Dropout-0.25, Flatten(), Dense 128, Dropout-0.5, Dense 19, activation sigmoid
  - The best score for this network was observed to be 0.69036

Experiments and results

- Custom CNN Architecture
  - The model takes as input a 2D spectrogram images
  - We follow the CubeRun architecture that was used for the BirdClef 2016 task[5]
  - After fine tuning the final layers were as: BatchNormalization, Conv2D 16, MaxP2D 2x2, BatchNormalization, Conv2D 32, MaxP2D 2x2, BatchNormalization, Conv2D 64, MaxP2D 2x2, BatchNormalization, Conv2D 12, MaxP2D 2x2, BatchNormalization, Conv2D 19, BatchNormalization, Flatten(), Dropout(0.1), Dense 512 Dropout(0.1), Dense 19, activation 'softmax'
  - The best score for this network was observed to be 0.77387

Experiments and results

- Feed forward neural network
  - The model takes as input the same features used to train the winning random forest model
  - After fine tuning we found the below architecture to give the best result
  - Dense 32, Dense 32, Dense 32, Dense 32, Dropout(0.5), Dense 19
  - This model gave a kaggle score of 0.69716

Experiments and results

- CNN model
  - Used the library librosa to get a dense vector of dimension 193, which describes the raw features of the audio file (mfcc-40, chromagram, melspectrogram).
  - Since the feature is a 1 dimensional vector, we feed these as input to 1D convolutional layers and the model architecture is described below
    - Conv1D 32, Conv1D 32, MaxPool1D 4, Conv1D 64, Conv1D 64, GlobalAveragePooling 1D, Dropout(0.5), Dense 19, activation='sigmoid'
  - After fine tuning we found the above architecture to give the best result, giving kaggle score of 0.83212

Image Networks, consolidated results :

Model	Kaggle Score
LSTM	0.70101
Bi-LSTM	0.70072
Variant of VGG network	0.69036
Custom CNN architecture	0.77387

Raw feature Networks, consolidated results:

Model	Kaggle Score
Feed forward neural network	0.69716
CNN	0.83212

THANK YOU!

References

1. http://publications.lib.chalmers.se/records/fulltext/249467/249467.pdf
2. https://arxiv.org/pdf/1804.07177v1.pdf
3. https://github.com/mtobeiyf/audio-classification
4. https://github.com/gaborfodor/MLSP_2013
5. https://github.com/johnmartinsson/bird-species-classification
6. https://keras.io/getting-started/sequential-model-guide/