Multiple Audio Event Detection / Audio Tagging

Ayush Kumar
200246
ayushk20

Abstract

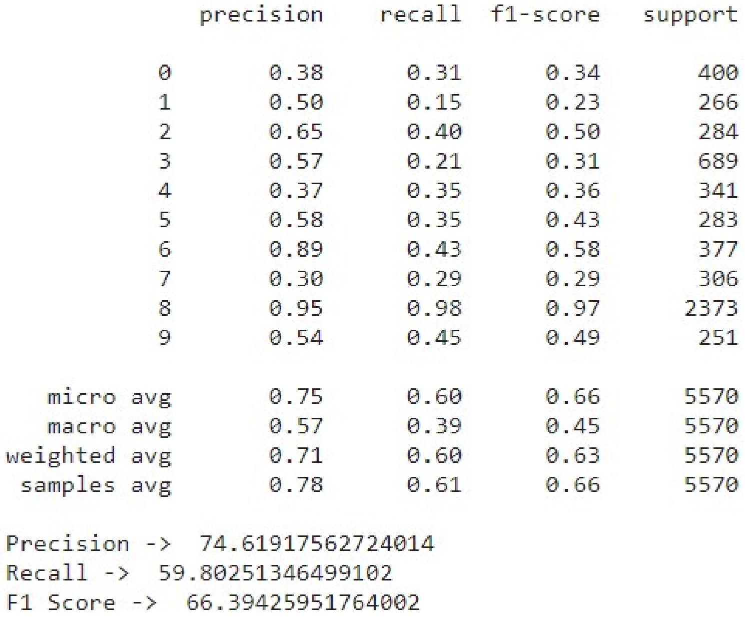
In this assignment we mainly focused on Audio Event Detection. We had 10000 .npy files as dev & validiation set data. For each of these files we need to output the voices present in the file. The basic outline and motive of this assignment are to pre-process the data efficiently and model the data in deep learning models.

1 My Methods

1.1 Convolutional Neural Network (CNN)

We used Conv2D layer in this assignment.This layer creates a convolution kernel that is convolved with the layer input to produce a tensor of outputs.The main advantage of CNN compared to its predecessors is that it automatically detects the important features without any human supervision. For example, given many pictures of cats and dogs it learns distinctive features for each class by itself. CNN is also computationally efficient.

Results by CNN:



```
              precision    recall  f1-score   support

           0       0.38      0.31      0.34       400
           1       0.50      0.15      0.23       266
           2       0.65      0.40      0.50       284
           3       0.57      0.21      0.31       689
           4       0.37      0.35      0.36       341
           5       0.58      0.35      0.43       283
           6       0.89      0.43      0.58       377
           7       0.30      0.29      0.29       306
           8       0.95      0.98      0.97      2373
           9       0.54      0.45      0.49       251

   micro avg       0.75      0.60      0.66      5570
   macro avg       0.57      0.39      0.45      5570
weighted avg       0.71      0.60      0.63      5570
 samples avg       0.78      0.61      0.66      5570

Precision ->  74.61917562724014
Recall ->  59.80251346499102
F1 Score ->  66.39425951764002
```

F1 score = 66.394
Precision = 74.619
Recall = 59.802

1.2 Dense Neural Network (DNN)

In any neural network, a dense layer is a layer that is deeply connected with its preceding layer which means the neurons of the layer are connected to every neuron of its preceding layer. This layer is the most commonly used layer in artificial neural network networks.

Results by DNN:
F1 score = 45.341
Precision = 54.619
Recall = 38.757

1.3 Recurrent-Convolutional Neural Network (RCNN)

Classification of audio with variable length using a CNN + LSTM architecture.

Results by RCNN:
F1 score = 38.654
Precision = 44.619
Recall = 34.095

1.4 Recurrent Neural Network (RNN)

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Results by RNN:
F1 score = 40.378
Precision = 50.619
Recall = 33.584

1.5 K Nearest Neighbours (KNN)

KNN algorithm assumes the similarity between the new case/data and available cases and put the new case into the category that is most similar to the available categories. K-NN algorithm stores all the available data and classifies a new data point based on the similarity. This means when new data appears then it can be easily classified into a well suite category by using K- NN algorithm.

Results by KNN:
F1 score = 58.705
Precision = 66.619
Recall = 52.471

1.6 Support Vector Machine (SVM)

The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

Results by SVM:
F1 score = 40.378
Precision = 74.619
Recall = 36.873