Multiple Audio Event Detection

Anonymous Author(s)
Affiliation
Address
email

1 Intoduction

We were provided with the labeled set of 10 Events (1000 samples per event) , of 10000 spec to-grams and we need to predict events present in an spectrograms (not their timestamps). Along with this we have provided with a additional python script "eventroll-to-multihot-vector.py" is given to convert eventrolls in the desired format. The data was already pre-processed we need to just merge with X (events) with the appropriate Y(labels), we just split the training data into training and evaluting data in the ratio of 4:1. We designed the multi-layer Neural Networks consisting of various convolution layers.The model has then tested on the trained model with 2500 samples of unseen data and predicted data were compared with ground truth values to measure of the F1 score.

1.1 Literature Review

Specifically, any research article/paper isn't refered for this assignment.Some, slight reference has been taken from keras, skit-learn documentation for writing the code, and referred to the class material and some online articles to understand working of CNN, and Pooling layer. Then, implemented the code of process the data and designed the CNN model, and pass the processed data as the input.

1.2 Convolutional Neural Network

A convolutional neural network, or CNN, is a deep learning neural network sketched for processing structured arrays of data such as portrayals. CNN are very satisfactory at picking up on design in the input image, such as lines, gradients, circles, or even eyes and faces. This characteristic that makes convolutional neural network so robust for computer vision. CNN can run directly on a underdone image and do not need any preprocessing. A convolutional neural network is a feed forward neural network, seldom with up to 20. The strength of a convolutional neural network comes from a particular kind of layer called the convolutional layer. CNN contains many convolutional layers assembled on top of each other, each one competent of recognizing more sophisticated shapes. With three or four convolutional layers it is viable to recognize handwritten digits and with 25 layers it is possible to differentiate human faces.The agenda for this sphere is to activate machines to view the world as humans do, perceive it in a alike fashion and even use the knowledge for a multitude of duty such as image and video recognition, image inspection and classification, media recreation, recommendation systems, natural language processing, etc.

Convolutional Neural Network Design The construction of a convolutional neural network is a multi-layered feed-forward neural network, made by assembling many unseen layers on top of each other in a particular order. It is the sequential design that give permission to CNN to learn hierarchical attributes. In CNN, some of them followed by grouping layers and hidden layers are typically convolutional layers followed by activation layers. The pre-processing needed in a ConvNet is kindred to that of the related pattern of neurons in the human brain and was motivated by the organization of the Visual Cortex.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

2 Methods and observations

Training Data is divided is given as 1000 examples of each case. The data is passed into a neural networks with series of convolution layers and dense layers. The model 4 CNN and 3 dense layers with suitable kernel sizes and ReLu as the activation function on all the layers, except the last layer i.e., "sigmoid", that allows to make binary classification amongst 10 different events.Similarly,an average pooling is also added after each convolution layer, the pooling layer is responsible for reducing the spatial size of convoluted feature. This is to decrease the computational power required to process the data by reducing the dimensions. There are two type of pooling average pooling and max pooling. But, here average pooling has been used to maintain affect every cell, but needed to reduce the dimension. Finally, created an custom training metric of f1 score by declaring a function and passing it as input. The model is trained by keeping 10 percent data as validation set. figure be low shows the f1 score metric graph vs number of epochs of both train and validation set.

3 Testing the 2500 samples

Each sample has been sliced into partitions, remainder partition after all these partitions, will be duplicated such that. each partition is estimated, by passing each partition into the trained model and finding the maximum predicted argument as the prediction. i partition, that will be duplicated, till it reaches the size limit . And all the predictions are made and added to a .csv file.

4 Result

from the predicted values and true values, the F1 score is 0.50 Precision and recall valued at 0.53 and 0.50 .