GUNSHOT SOUND DETECTION

Using Deep Convolutional Neural Network

T5 BOOTCAMP
SDAIA AKAEMY
KHULD MANSOUR ALSHAMRANI

1. Project Objective

This project has the objective to construct and train a deep Convolutional Neural Network (CNN) to recognize the gunshot sound among various other sounds.

2. Experimental Setup

We have used python TensorFlow and Keras libraries to construct a CNN model. Librosa library is used to extract the mel-frequency features from audio sounds to feed the network. We have used the UrbanSound8k dataset to train and test the model. This dataset is used as it is. This dataset already contains the 10 folds for cross-validation and we have performed the cross-validation using these folds.

3. Audio Dataset

The dataset is consists of 8732 sound clips. All of these sound clips contain the label information. The files are sorted into 10 folds and folds are created for cross-validation. Following are the categories or classes that are used in the dataset:

0 = air_conditioner
1 = car_horn
2 = children_playing
3 = dog_bark
4 = drilling
5 = engine_idling
6 = gun_shot
7 = jackhammer
8 = siren
9 = street_music

4. Proposed Scheme

The proposed scheme presented in the figure consists of the following steps. The green color boxes present the training phase and the orange color is representing the testing.

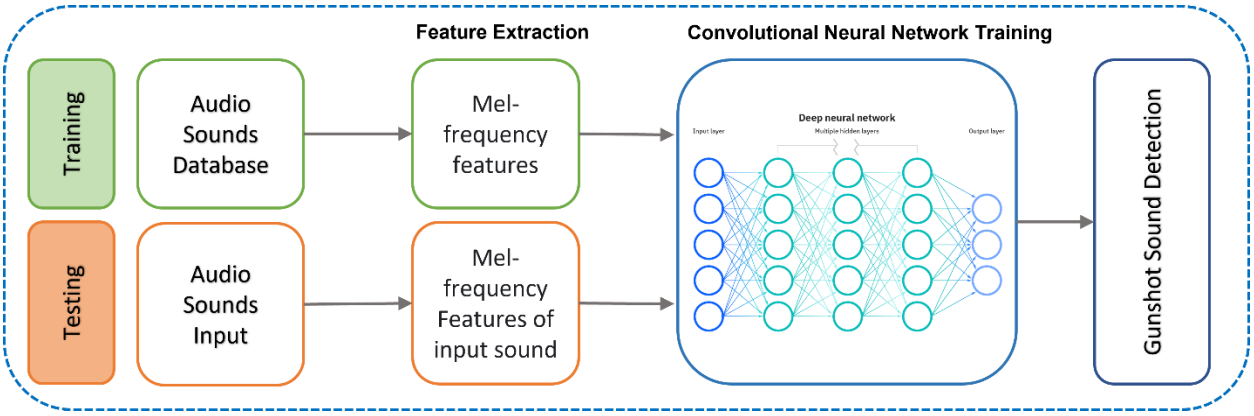

Figure 1: Proposed Scheme

4.1. Feature Extraction

Originally dataset contains 8732 audio sounds from 10 different classes discussed in the dataset section. First, the mel-frequency features are extracted using the python librosa library for all the audio sounds. All the audio sounds are converted to mel-spectrogram and then features are extracted from each mel-spectrogram. In this gunshot detection system, the mel-spectrogram for all the categories of sound is visualized in figure 2.

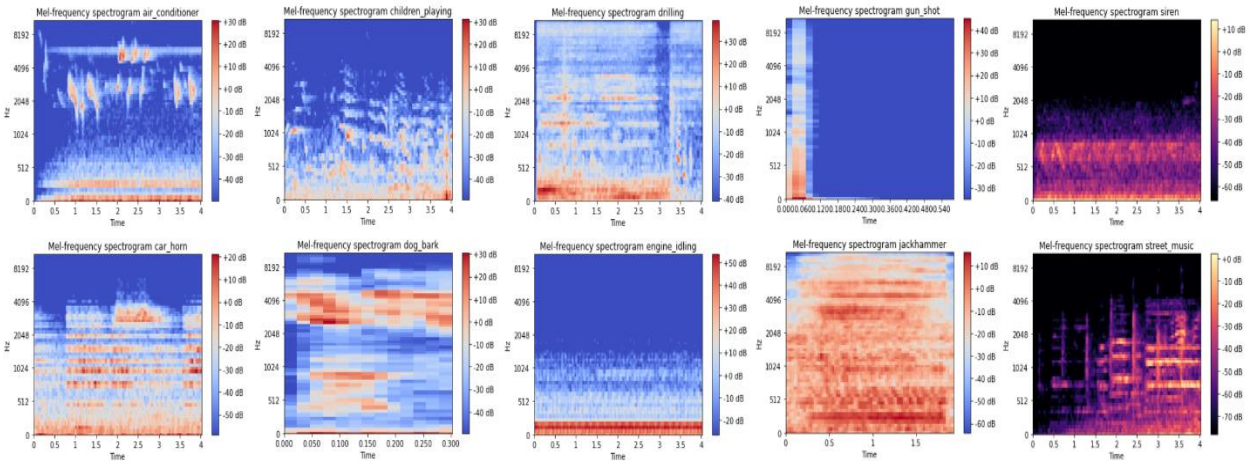

Figure 2: Mel-Spectrom of Sounds

Visually the mel-spectrogram images of each class are very different from each other and can be distinguished with the human eye. We have used the convolutional neural network to feed the Mel spectrogram features to train a model. The original dataset contains 10 folds with almost the same number of audio files. In this system, the first fold is used for testing and the other 9 folds are used to train the model. In other words, the mel-spectrogram features are extracted for training data, and then these features are passed to a deep convolutional neural network model for training purposes. When the CNN model will be trained on these rich features of the mel-spectrogram then it will be able to classify the gunshot sound including all other sounds.

4.2. The CNN Model Training:

The CNN model consisting of the convolutional layer is proved to be very effective to identify the patterns in the images. The following figure shows the CNN model architecture.

```
cnn_model = keras.models.Sequential()
cnn_model.add(keras.layers.Conv2D(24, sizeofkernel, padding="same", input_shape=input_shape))
cnn_model.add(keras.layers.BatchNormalization())
cnn_model.add(keras.layers.Activation("relu"))
cnn_model.add(keras.layers.MaxPooling2D(pool_size=pool_size))

cnn_model.add(keras.layers.Conv2D(32, sizeofkernel,padding="same"))
cnn_model.add(keras.layers.BatchNormalization())
cnn_model.add(keras.layers.Activation("relu"))
cnn_model.add(keras.layers.MaxPooling2D(pool_size=pool_size))

cnn_model.add(keras.layers.Conv2D(64, sizeofkernel, padding="same"))
cnn_model.add(keras.layers.BatchNormalization())
cnn_model.add(keras.layers.Activation("relu"))
cnn_model.add(keras.layers.MaxPooling2D(pool_size=pool_size))

cnn_model.add(keras.layers.Conv2D(128, sizeofkernel, padding="same"))
cnn_model.add(keras.layers.BatchNormalization())
cnn_model.add(keras.layers.Activation("relu"))

cnn_model.add(keras.layers.GlobalMaxPooling2D())
cnn_model.add(keras.layers.Dense(128, activation="relu"))
cnn_model.add(keras.layers.Dense(total_classes, activation="softmax"))
```

Figure 3: The Convolutional Neural Network Layers

We have defined a convolutional neural network with four convolutional layers. Every convolution layer is followed by batch-normalization, activation-relu, and max-pooling layer. At the last convolution layer, there is a global max-pooling layer which is handled by a fully connected layer. In the end, the softmax layer is added to calculate the probabilities of audio classes/ categories. Here, we have used the cross-entropy function as a loss function for training.

5. Evaluation of the Model

We have evaluated the system on the testing data and calculated the average of the prediction to calculate the accuracy. Other than that, we have also calculated the f1-score, precision-score, and recall-score for the testing data.

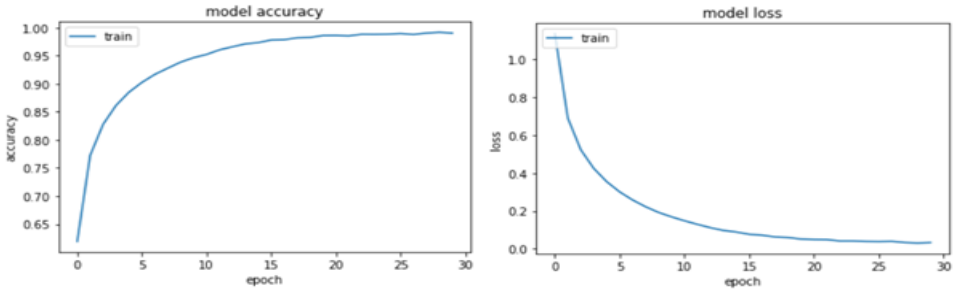

Figure 4: Model Accuracy and Loss

Model accuracy plot provides an insight that the model is efficiently trained on the training data. It shows very promising results for accurately classifying the training data. The loss plot also depicts that loss is converged towards minimum loss.

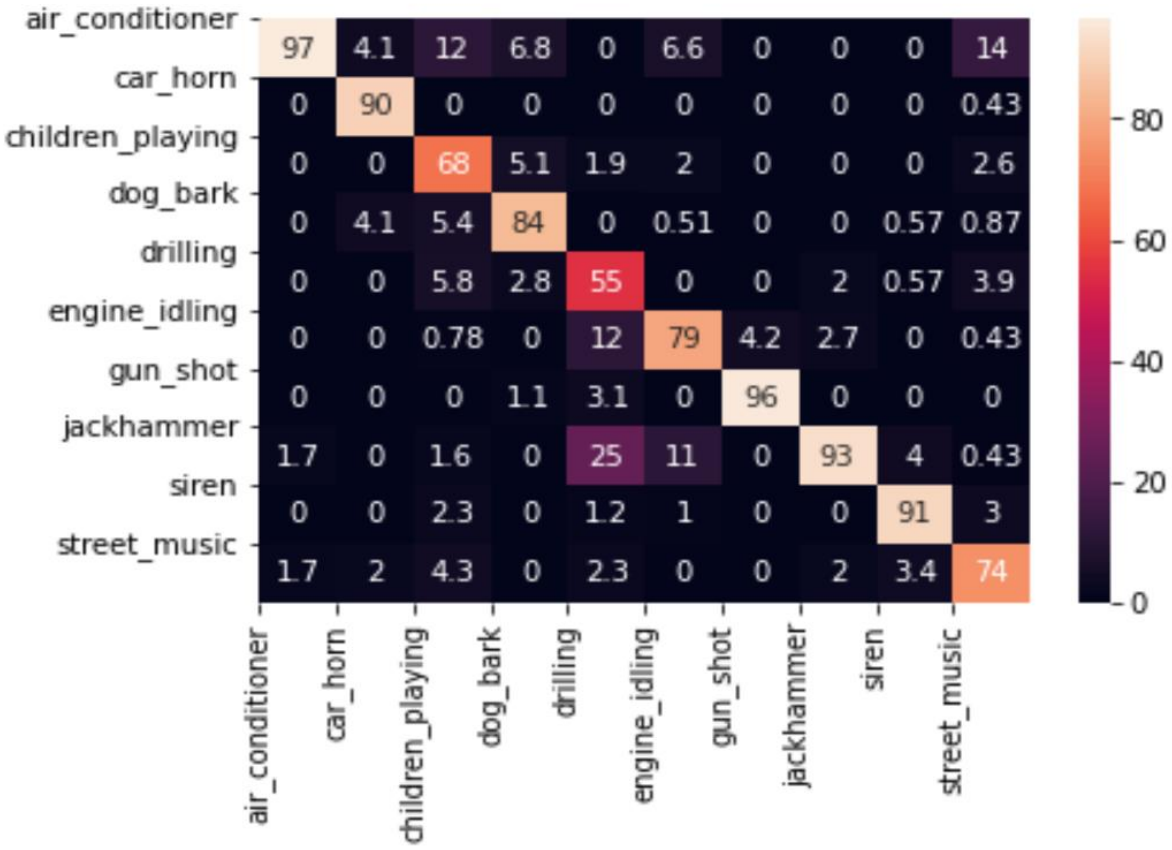

Figure 5: Confusion matrix.

We have drawn the confusion matrix for the testing data to find out the accuracy of gunshot and all other sound classes. This model has efficiently recognized the gunshot sound with 96% correctly identifying the gunshot sounds 3.1% gunshots are wrongly detected as the drilling sound and 1.1% gunshots are detected as dog bark sound. The drilling is very similar to some of the gunshots sound so it increases the probability to be wrongly detected. Although only drilling is the one category that has only a very low percentage of correctly identifying. It is 55% correctly detected and 45% of its sounds are wrongly detected to other classes. This model efficiency can also be noticed that 6 out of 10 classes have 84+% correctly identifying accuracy and 8 out of 10 classes have 74+ percentage in terms of correctly identifying process. The overall accuracy of the system is 78% with an F1 score of 0.80.

Oerall-Score of Model	Percentage
Accuracy	0.78
F1-Score	0.80
Precision	0.82
Recall	0.80

Future Work:

- Combining the gunfire action with gunshot sound detection system.
- 3D location of Southern border of KSA would be covered to the model.
- Use acoustic sensing technology to identify the location of the gunshot