CONVOLUTIONAL RECURRENT NEURAL NETWORK BASED APPROACH FOR MAKING SENSE OF SOUNDS DATA CHALLENGE

Technical Report

Rajdeep Mukherjee[1]*, Pradhumn Goyal*[1]*, Dipyaman Banerjee*[2]*, Kuntal Dey*[2]*, Pawan Goyal*[1]

[1] Indian Institute of Technology, Kharagpur, India,
{rajdeep1989.iitkgp, pradhumn0708goyal, pawang.iitk}@gmail.com
[2] IBM Research - India, New Delhi, India, {dipyaban, kuntadey}@in.ibm.com

ABSTRACT

Development of an automated system for categorizing sound events like humans is a challenging task. In this paper, we propose a Convolutional Recurrent Neural Network (CRNN) based approach, implemented as a 7-layer Convolutional Neural Network (CNN) followed by a GRU based recurrent layer, for the task of classifying audio data into five broad categories derived from human classification, namely **Nature, Human, Music, Effects** and **Urban**. As part of the Making Sense of Sounds Data Challenge, we were provided with a development dataset consisting of 1500 audio files, divided into the five categories, each containing 300 files. The number of different sound types within each category is however not balanced. The evaluation dataset consists of 500 audio files, 100 files per category. Each audio file is a single channel 44.1 kHz, 16 bit .wav file, of exactly 5 seconds duration. We train our proposed models on 128-band Log-scaled Mel Spectrogram features extracted from the audio data files. Our baseline CNN and CRNN models give an average accuracy score of 66.49% and 69.55% respectively. Data augmentation significantly improves the performance of our proposed models. The CNN and CRNN models when trained with augmented data give average accuracy scores of 87.18% and 91.36% which comfortably beat the current state-of-the-art baseline score of 81% reported on the challenge website.

Index Terms— Making Sense of Sounds, Deep Learning, Convolution, Recurrent, Augmentation

1. INTRODUCTION

Replicating the human capability of recognizing and categorizing sounds in the form of an automated system is a challenging task. As part of the Making Sense of Sounds Data Challenge, our task is to develop an automated system which classifies audio data into five broad categories; **Nature, Human, Music, Effects** and **Urban**. The development dataset provided to us, as part of the challenge, consists of 1500 audio files, divided into the above mentioned five categories, each containing 300 files. The evaluation dataset consists of 500 audio files, 100 files per category. Though both the sets have been collected from the same data source, however the allocation of specific sound types to the development and evaluation sets is not balanced which makes the task even more challenging.

In this report, we present a deep Convolutional-Recurrent Neural Network (CRNN) based approach for the task of audio data classification. We draw motivation for the proposed approach from previous similar works such as [1] and from our success in the Bird Audio Detection task as part of **IEEE DCASE2018 Challenge**. We ranked 7th in the challenge and the technical report of our work can be found here [2]. We first develop a Convolutional Neural Network (CNN) and then extend it to a Convolutional-Recurrent Neural Network by adding a recurrent block on top of the convolutional block. We finally use a softmax activation layer to classify the sound files into the five target categories. While the CNN layer helps us in detecting time and frequency invariant local patterns, the recurrent layer performs temporal context-based analysis. We use 128-band Log-scaled Mel Spectrogram features extracted from the audio data as our input to the deep learning model. We apply *Batch-normalization* after each convolutional layer and *Dropout* after the third, fifth and seventh (final) convolutional layers and final dense layer to prevent our model from overfitting, given that the development dataset size is very small. We finally demonstrate the use of data augmentation by re-sampling each audio file with different rates with respect to the original sampling rate of 44.1 kHz. The details are given in Section 3. We observe a significant improvement of results when our models are trained with the augmented data. The details of the feature extraction process and our proposed neural network architectures are given in Section 2.

2. OUR APPROACH

2.1. Overview and Feature Extraction

Our proposition is based on a deep neural network learning methodology. The input to the training phase of the system comprises of Log-scaled Mel Spectrogram representation of the audio signals. For each audio file under analysis, each of which is of exactly 5 seconds duration, we extract a 128-band Log Mel Spectrogram, that covers the audible frequency range of audio signals, namely 0-$22,050$ Hz, with 23.22 ms windows ($1,024$ samples at 44.1 KHz), and a hop size of 1024 samples thereby not allowing any overlap between two consecutive frames. The obtained segments (128 rows/bands * 216 columns/frames) are provided together with their deltas (computed with default librosa settings) as a 2-channel input to the network. The features are subsequently passed over to the deep neural network for training and testing, in the respective phases.

2.2. The Deep Learning Architecture

For deep learning, we design and compare two different model architectures. The architecture details for the Convolutional Neural Network (CNN) based approach is presented in Table 1. We extend this model with one additional Gated Recurrent Unit (GRU) based recurrent layer to come up with the design of the Convolutional Recurrent Neural Network (CRNN) architecture. The output of the RNN is passed over a fully connected layer, wherein the final class label is generated. The architecture details are provided in Table 2.

Input: 128 × 216 × 2
3 × 3 Convolution 2D (32)(Padding='same') + ReLu 3 × 3 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 Convolution 2D (64) + ReLu Batch Normalization 4 × 4 MaxPooling 2D (Strides 4 × 4) + Drop-Out (0.3)
3 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 MaxPooling 2D (Strides 3 × 3) + Drop-Out (0.3)
3 × 3 Convolution 2D (64) + ReLu Batch Normalization 5 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 MaxPooling 2D (Strides 3 × 3) + Drop-Out (0.5)
Dense(5) + L2 Weight_Regularizer (0.001) + Drop-Out (0.5)
Activation (Softmax)

Table 1: CNN Model Architecture.

Input: 128 × 216 × 2
3 × 3 Convolution 2D (32)(Padding='same') + ReLu 3 × 3 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 Convolution 2D (64) + ReLu Batch Normalization 4 × 4 MaxPooling 2D (Strides 4 × 4) + Drop-Out (0.3)
3 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 MaxPooling 2D (Strides 3 × 3) + Drop-Out (0.3)
3 × 3 Convolution 2D (64) + ReLu Batch Normalization 5 × 5 Convolution 2D (64) + ReLu Batch Normalization 3 × 3 MaxPooling 2D (Strides 3 × 3) + Drop-Out (0.5) ReShape (3, 64)
GRU (128, return_sequences=False)
Dense(5) + L2 Weight_Regularizer (0.001) + Drop-Out (0.5)
Activation (Softmax)

Table 2: CRNN Model Architecture.

3. EXPERIMENTS

We train each of our proposed models on the 128-band Log Mel Spectrogram features extracted from the audio files as explained in Section 2. For data augmentation, we re-sample each input audio file with the following rates: 0.55, 0.75, 0.9, 1.2, 1.4, with respect to the original sampling rate of 44.1 KHz. This introduces variance into our dataset, by stretching the sound file temporally and also by lowering and raising the pitch of the audio file. The data augmentation process significantly helped us in improving the performance of our systems. *Base_CNN* represents our baseline CNN architecture as detailed in Table 1. *Base_CRNN* represents the CRNN architecture as detailed in Table 2. We represent the models trained with augmented datasets with a suffix *_Aug*. We split the development dataset, consisting of 1500 audio files, into training set and validation set. We train our base models on the training data consisting of 1200 audio files and obtain the True Positive and F scores for each of the five categories on the validation data consisting of 300 audio files. For training with augmentation, 80% of the entire augmented dataset is considered for training and the remaining 20% for validation and obtaining the scores. Each model is trained for 100 epochs. The results are reported in Table 3 and Table 4 respectively.

	True Positive Rates				
Method	**Effects**	**Human**	**Music**	**Nature**	**Urban**
Base_CNN	81.67	52.63	77.27	68.52	52.38
Base_CNN_Aug	88.43	90.50	90.96	89.09	76.92
Base_CRNN	76.67	52.63	84.85	79.63	53.97
Base_CRNN_Aug	89.32	90.78	95.63	92.47	88.59

Table 3: True Positive Rates comparison of various methods

	F Score				
Method	**Effects**	**Human**	**Music**	**Nature**	**Urban**
Base_CNN	66.67	52.63	85.00	66.67	61.11
Base_CNN_Aug	81.76	89.01	93.27	88.63	82.86
Base_CRNN	70.23	58.82	90.32	63.70	62.96
Base_CRNN_Aug	86.49	91.55	96.90	92.59	89.30

Table 4: F Score comparison of various methods

4. CONCLUSION

In this paper we present a deep convolutional-recurrent neural network (CRNN) architecture for classifying audio data into five broad categories; namely Effects, Human, Music, Nature and Urban; as part of the Making Sense of Sound Data Challenge 2018. We first use a convolution neural network and then extend it to add a recurrent layer on top. Both the models achieve comparable results. However, the CRNN architecture slightly outperforms the CNN model. Data augmentation significantly helps in improving the performance of both the systems. The average accuracy score of 87.18 % for *Base_CNN_Aug* and 91.36 % for *Base_CRNN_Aug* comfortably beat the current state-of-the-art baseline score of 81 % as reported on the challenge website. In future, we would like to apply attention on top of our proposed models.

5. REFERENCES

[1] D. Stowell, Y. Stylianou, M. Wood, H. Pamuła, and H. Glotin, "Automatic acoustic detection of birds through deep learning: the first bird audio detection challenge," *Methods in Ecology and Evolution*, 2018.

[2] R. Mukherjee, D. Banerjee, K. Dey, and N. Ganguly, "Convolutional recurrent neural network based bird audio detection," DCASE2018 Challenge, Tech. Rep., September 2018.