DEEP NEURAL NETWORK FOR ACOUSTIC SCENE DETECTION

Alexei Pugachev

ITMO University
Chair of Speech Information Systems,
ul. Krasutskogo 4,
Saint-Petersburg, 196084, Russian Federation
pugachev@speechpro.com

Dmitrii Ubskii

ITMO University
Chair of Speech Information Systems,
ul. Krasutskogo 4,
Saint-Petersburg, 196084, Russian Federation
ubskiy@speechpro.com

ABSTRACT

The DCASE 2016 challenge comprised the task of Acoustic Scene Classification. The goal of this task was to classify test recordings into one of predefined classes that characterizes the environment during the recording.

Index Terms— DCASE 2016, acoustic scene detection, DNN

1. INTRODUCTION

This report explains the stucture of overall system that solves the given task of the challenge.

2. DATA DESCRIPTION

The training data consisted of 1190 audiofiles 30 seconds long each. A label was assigned for each file, showing the environment in which the file was recorded. Evaluation dataset comprised 390 audio files 30 seconds long each and without labels.

3. FEATURES EXTRACTION

The following features were used: 19 MFCC[2] (mel frequency cepstral coefficients)without first coefficient, 20 Δ and 20 $\Delta\Delta$ (acceleration coefficients). Features extracted from each audio file formed a feature set that was later split in 4 parts. By means of cyclic permutation these parts were used to create 4 folders for training. Each folder had data for training (75%) and data for testing (25%).

4. CLASSIFIER

Classification was carried out by Deep Neural Network (DNN)[1] with two hidden Layers. The size of input layer was 59 neurons (19 MFCC + 20 Δ + 20 Δ Δ), hidden layers - 1600 neurons, and output layer - 15 neurons (number of classes). ReLU (rectified linear unit) function was used as neuron activation function:

$f(x) = \max(0, x)$

cross-entropy served as error function:

$H(p, 1-p) = -\sum_x p(x)(1-p(x))$

Training algorithm uses gradient descent[3]. Learning rate was optimized with the aid of cross-validation. Cross-validation set was obtained by splitting train set on new train set (95%) and cross-validation set (5%).

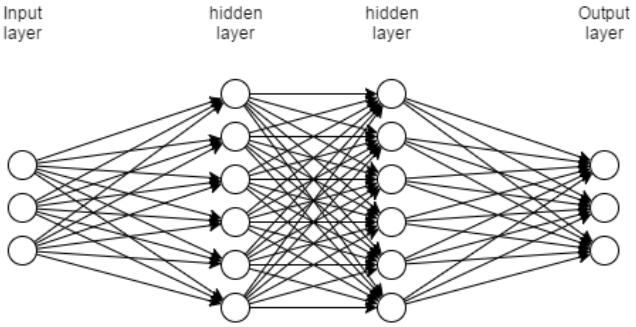

Figure 1: DNN architecture.

5. SCORING

Scoring was done using classification precision function:

$E = \frac{f}{n} * 100$,

where f is the number of samples correctly classified, and n is the total number of samples.

6. RESULTS

Experiments on development set gave the following results:

accuracy	fold1	fold2	fold3	fold4
82.92%	87.9%	87.9%	94.6%	61.3%

4 models were obtained after training on the development dataset. 4 predictions of scenes for evaluation dataset were received based on this models. To obtain the final result, majority voting method was applied. This method helps avoid inaccuracy of models.

7. CONCLUSION

Deep Neural Network outperforms GMM on this task, which is explained by the fact that DNNs are much better at handling large amounts of data. In this case there are some ways to increase accuracy: adding new data to train dataset, or generating more data based on existing dataset (Data augmentation).

8. REFERENCES

[1] Using neural nets to recognize handwritten digits http://neuralnetworksanddeeplearning.com/chap1.html

[2] Mel Frequency Cepstral Coefficient (MFCC) tutorial http://practicalcryptography.com/miscellaneous/machine-learning/guide-mel-frequency-cepstral-coefficients-mfccs

[3] A. Agarwal et. al., "An Introduction to Computational Networks and the Computational Network Toolkit", Microsoft Technical Report MSR-TR-2014-112, 2014.