AUDIO TAGGING USING LABELED DATASET WITH NEURAL NETWORKS

Technical Report

Agafonov Iurii

ITMO University
SIS Dept., 49, Kronverkskiy pr.,
Saint-Petersburg, 197101, Russian Federation
agafonov@speechpro.com

Shuranov Evgeniy

Speech Technology Center
4, Krasutskogo st., Saint-Petersburg, 196084,
Russian Federation
shuranov@speechpro.com

ABSTRACT

In this paper, an audio tagging system is proposed. This system uses fusion of 5 Convolutional Neural Network (CNN) and 1 Convolutional Recurrent Neural Network (CRNN) classifiers in attempt to achieve better results. The proposed system reaches 0.95 score in public leaderboard.

Index Terms — audio tagging, neural networks

1. INTRODUCTION

Audio tagging is an application of pattern recognition and machine learning in which an audio signal is mapped to a corresponding sound event. Automatic audio event detection is utilized in a host of applications, including surveillance, speech detection and audio segmentation. This task is also particularly challenging because it involves large amount of classes and multi-label classification.

Most conventional approaches to multi-label classification involve a set of one-vs.-rest binary classifiers, one for each label, their results are then combined, or problem transformation to a single-label classification over the power set of original classes [1]. These approaches do not scale well with increasing number of classes, however, as every new class significantly increases training time and memory requirements.

In this contribution, a system is proposed which uses regression to calculate scores for each possible class for a sample. Proposed solution to the problem of scaling uses only one classifier for all classes that outputs prediction scores for each class. Adding a new class in that case entails only minor adjustments of the system.

Detailed description of the task can be found here [5].

2. SYSTEM OVERVIEW

This section explains the approach to audio classification used in this system.

2.1. General propositions

A simple activity detection algorithm was applied to all audio data to cut silence. It is based on exponentially weighted moving average on signal energy. Then all samples below the selected threshold are discarded.

All audio files were resampled to 16 KHz. Logarithms of mel-spectrogram were used as features with the following parameters: STFT window size 512, hop length 256, number of mels 64, time duration 5 seconds. Audio signals shorter than 5 seconds were sufficiently duplicated to get the correct length.

2.2. Augmentation

Because the training dataset is not balanced by classes, and also is quite small, two types of augmentation were applied to the data: time stretch and pitch shifting [2].

2.3. Neural netrworks' architectures

As was mentioned above proposed system is fusion of 4 CNN classifiers. The architectures of all four models are described below in keras library representation. Categorical cross-entropy was used as loss function in all models. Learning rate was 0.0001, all networks were trained using 4-fold cross-validation.

2.3.1. First model

Layer
Input(None,64,315,1)
Conv2D(64,(7,3),(1,2),"relu")
MaxPool2D((4,1),(2,1))
BatchNormalization()
Conv2D(128,(7,1),(1,1),"relu")
MaxPool2D((4,2),(2,2))
BatchNormalization()
Conv2D(128,(5,1),(1,1),"relu")
BatchNormalization()
Conv2D(128,(1,5),(1,1),"relu")
BatchNormalization()
GlobalMaxPool2D()
Dropout(0.25)
BatchNormalization()
Dense(64,"relu")
Dropout(0.25)
Dense(64,"relu")
Dropout(0.25)

Dense(41,softmax)

2.3.2. Second model

It will be more convenient to describe one block of layers at first:

SecondModelBlock(size)
BatchNormalization()
Activation("relu")
Conv2D(size,(3,3))
BatchNormalization()
Activation("relu")
Conv2D(size,(3,3))

Using this block the entire architecture can be described like that:

Layer
Input(None,64,315,1)
SecondModelBlock(64)
MaxPool2D((3,3))
SecondModelBlock(128)
MaxPool2D((3,3))
SecondModelBlock(64)
GlobalAveragePool2D()
Dense(512)
Dropout(0.25)
Dense(41,softmax)

2.3.3. Third model

Here one block of layers is described:

ThirdModelBlock(size)
Conv2D(size,(3,3))
BatchNormalization()
Activation("relu")
MaxPool2D((4,1))

The entire architecture:

Layer
Input(None,64,315,1)
ThirdModelBlock(64)
ThirdModelBlock(128)
Conv2D(128,(3,3))
Activation("relu")
MaxPool2D((4,1))
Reshape((315,128))
StatisticalPooling()
Dense(128,"relu")
Dropout(0.25)
Dense(64,"relu")
Dropout(0.25)
Dense(41,softmax)

Here StatisticalPooling() layer concatenates mean values and STD values along second axis of input tensor.

2.3.4. Another models

Fourth model is almost the same as the second one except one layer. Here Dense(1024) layer is used instead of Dense(512)
Fifth model is almost the same as SEResNet18 described here [3]. Size of kernel from first convolutional layer was changed to (4,4). First pooling layer was taken with following parameters: MaxPool2D((2, 2), strides=(1, 1)). 128 mels were taken for this model.
Sixth CRNN model is based on model from here [4].

2.4. Fusion

First submission consists of predictions of first, second, third and fourth models described above were fused with the following weights: [0.3, 0.3, 0.2, 0.2]. These weights were trained to reach the best validation MAP@3 score which is described in section 3.
Second submission consists of predictions of first, second, fifth and sixth models which were fused by geometric mean value of predictions.

3. EVALUATION

Submissions are evaluated according to the Mean Average Precision @ 3 (MAP@3):

$$MAP@3 = \frac{1}{U}\sum_{U=1}^{U}\sum_{k=1}^{min(n,3)} P(k),$$

where U is the number of scored audio files in the test data, $P(k)$ is the precision at cutoff k, and n is the number predictions per audio file.
According to MAP@3 the following scores were reached for each model: 0.931, 0.92, 0.913, 0.911, 0.926 and 0.929. Fusion of all these models reached score 0.95 for first submission and 0.949 for second.

4. CONCLUSION

The system proposed in this paper outperforms baseline system by 0.246 on the indoors subset of the provided data when scored using MAP@3. The system also scales considerably better with addition of new classes.

5. REFERENCES

1. G. Tsoumakas, I. Katakis, "Multi-Label Classification: An Overview," in Int J Data Warehousing and Mining, 2007, pp. 1–13.
2. J. Salamon, J. P. BelloDeep, "Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification", IEEE SIGNAL PROCESSING LETTERS, 2016.
3. https://arxiv.org/pdf/1512.03385.pdf
4. http://www.cs.tut.fi/sgn/arg/dcase2017/documents/challenge_technical_reports/DCASE2017_Adavanne_130.pdf
5. Eduardo Fonseca, Manoj Plakal, Frederic Font, Daniel P. W. Ellis, Xavier Favory, Jordi Pons, Xavier Serra.

General-purpose Tagging of Freesound Audio with AudioSet Labels: Task Description, Dataset, and Baseline. Submitted to DCASE2018 Workshop, 2018. URL: https://arxiv.org/abs/1807.09902