ACOUSTIC SCENE AND EVENT DETECTION SYSTEMS SUBMITTED TO DCASE 2018 CHALLENGE

Technical Report

Maksim Khadkevich

Facebook Inc. Menlo Park, CA, USA

ABSTRACT

In this technical report we describe systems that have been submitted to DCASE 2018 [1] challenge. Feature extraction and convolutional neural network (CNN) architecture are outlined. For tasks 1c and 2 we describe transfer learning approach that has been applied. Model training and inference are finally presented.

Index Terms— Acoustic Event Detection, Transfer Learning

1. INTRODUCTION

Audio event and scene classification plays crucial role in multimedia content understanding. The problem of identification of acoustic environment and acoustic events has recently become popular among researchers. Increasing interest in the problem can be partially explained by availability of large-scale data sets that have been released lately. That has attracted many researches from machine learning community, where large-scale data sets can improve performance by a large margin.

2. FEATURE EXTRACTION

Logmel spectrograms are used for training CNNs in all the sumbissions. All audio recordings are downsampled to 16 KHz sampling rate and converted to mono. Window size of 16 ms, window step of 10 ms and 40 mel bands are used to generate logmel features.

3. CNN ARCHITECTURE

Neural network architecture is presented in Fig. 1. CNN model is a VGG [1]-like model with 12 convolutional layers. All the filters in convolutional layers are 3x3. Each convolutional layer is followed by batch-norm and ReLU. After every second convolutional layer features are downsampled using 2x2 max-pooling layer, except for the last one, which is 1x3 max-pooling layer. Logmel features for 1 second of audio, which corresponds to 40-by-100 input size are fed into CNN model. The output of CNN model is an embedding feature of dimension 2048. High-level reasoning is done in classifier via fully connected layers. ReLU layers are inserted after FC1 and FC2. FC1 and FC2 have input and output size of 2048, while FC3 has input size of 2048 and output size equal to the number of classes. Negative log-likelihood is adopted as a loss function.

[1] http://www.robots.ox.ac.uk/∼vgg/research/very_deep/

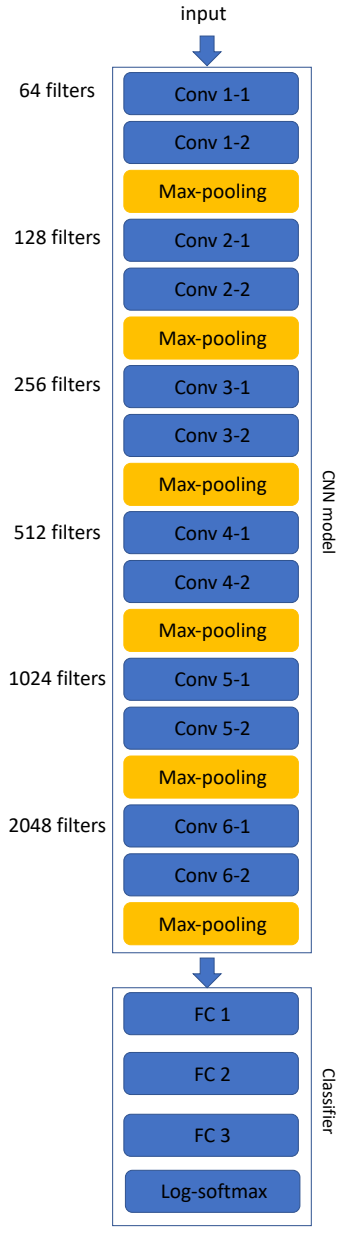

Figure 1: CNN model architecture

4. MODEL TRAINING

There are two training strategies we apply: training from scratch (task 1a) and transfer learning (task 1c and 2).

In transfer learning strategy the model is pre-trained on Audioset [2]. Balanced and unbalanced training splits are merged together and are used as training corpus. Once the model is trained, parameters for all layers from "classifier" block are reinitialized and all parameters from "CNN model" block are frozen. Output size of FC3 block is adjusted to match number of classes in target task (10 for task 1c and 41 for task 2). In the last stage the model is fine-tuned on the training data provided by task organizers.

In training from scratch strategy the model is trained directly on the training data provided by the task organizers.

Since training samples have different length, we randomly select 1 second of audio from each segment on each iteration. Stochastic Gradient Descent (SGD) with momentum is adopted as parameter optimization algorithm. Batch size is set to 1024 samples.

5. INFERENCE

Once the models are trained, final scores are obtained by splitting logmel features for all test samples into 1 second chunks with 0.5 seconds overlap. Final scores for each class are obtained by taking average or max (two different submissions) of scores for all chunks within test sample. For tasks 1a and 1c class with highest score is reported. For task 2 we report three classes with highest average scores.

6. REFERENCES

[1] http://dcase.community/challenge2018/.

[2] https://research.google.com/audioset/