Deep learning for high-level sound categorization

Patrice Guyot

IRIT, Université de Toulouse, CNRS, Toulouse, France

patrice.guyot@irit.fr

1 Introduction

This document presents short descriptions of the systems submitted at the data challenge "Making Sense of Sounds" in 2018. The aim of this challenge is to classify audio files into five different classes (*Nature, Human, Music, Effects, Urban*).

The followings part describe the submitted systems. The first system, which is called *simplemind*, consists of a simple Convolutional Neural Network (CNN). The second system, which is called *nevermind* consist of a more elaborate VGG-like model adapted from [1].

2 The simplemind system

2.1 Network

This system is based on a Convolutional Neural Network. It has been implemented with Keras. It is composed by the following layers:

- convolution (size=3×3)
- max pooling (size=3×5)
- convolution (size=1×3)
- convolution (size=3×3)
- convolution (size=1×3)
- max pooling (size=3×5)
- convolution (size=3×3)
- max pooling (size=2×2)
- dropout (rate=0.2)
- dense (128 units)
- dropout (rate=0.2)
- dense (5 units)

All activation functions are *relu* with the exception of the last layer which is based on a *softmax* activation function.

2.2 Experiment

We used mel spectrogram as input features (nftt=2048, hop size=512). Thus, an image of dimension 128 × 216 is produced for each recording. We used 100 epochs to train the system.

3 The nevermind system

3.1 VGGish Model for Audioset

This system is based on the VGGish Model for AudioSet[1]. VGGish is a variant of the VGG model [2], in particular Configuration A with 11 weight layers. The input size was changed to 96x64 for log mel spectrogram audio inputs. The last group of convolutional and maxpool layers is removed, so there are only four groups of convolution/maxpool layers instead of five. Instead of a 1000-wide fully connected layer at the end, there is a 128-wide fully connected layer. This acts as a compact embedding layer.

3.2 Additional layers

On the top of the VGGish Model for AudioSet, we added six fully connected layers with respectively 100, 80, 60, 40, 20, and 5 units.

3.3 Experiment

The audio features was identical to the VGGish Model for Audioset. They consist of log mel spectrograms. The audio files was split in five parts, so the system was trained on 7500 labeled audio experts. To produce the inferences, we simply used a majority vote on the five parts. We used 10 epochs to train the system.

References

[1] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al. Cnn architectures for large-scale audio classification. In *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*, pages 131–135. IEEE, 2017.

[2] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[1] `https://github.com/tensorflow/models/tree/master/research/audioset`