SIMPLE CNN AND VGGISH MODEL FOR HIGH-LEVEL SOUND CATEGORIZATION WITHIN THE MAKING SENSE OF SOUNDS CHALLENGE

Patrice Guyot
IRIT, Université de Toulouse, CNRS, Toulouse, France
patrice.guyot@irit.fr

1. INTRODUCTION

This document presents a short description of two systems for sound classification 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*, which were derived from human classification. For this challenge, the organizers provided a development dataset that consists of 1500 audio files of 5 second duration divided into the five categories, each containing 300 files.

The following parts describe the two submitted systems. The first system, called *simplemind*, consists of a Convolutional Neural Network (CNN). The second system, called *nevermind* is based on VGG-like model adapted from [4].

2. THE SIMPLEMIND SYSTEM

2.1. Convolutional Neural Network

In this first system, we use Mel spectrogram as input features. We use the *librosa* library [1] to compute them, with using the default parameters values (nftt=2048, hop size=512). Thus, a spectrogram of 216×128 is created for each audio file of 5 seconds (see Figure 1).

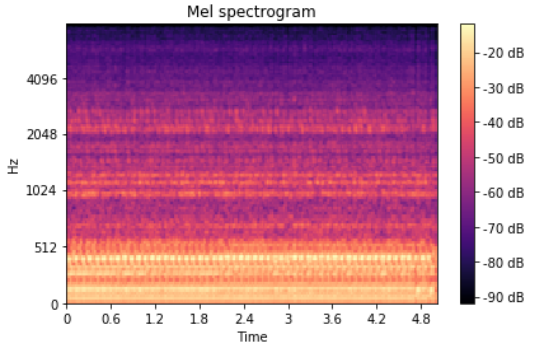

Fig. 1. Example of a Mel spectrogram used as input in the simplemind system.

The *SimpleMind* CNN use convolution layers, as well as max pooling. Drop out is used on the last layers. More precisely, it is composed of the following layers:

- convolution (filter size=3×3, depth=64)
- max pooling (filter size=3×5)
- convolution (filter size=1×3, depth=32)
- convolution (filter size=3×3, depth=32)
- convolution (filter size=1×3, depth=32)
- max pooling (filter size=3×5)
- convolution (filter size=3×3, depth=32)
- max pooling (filter 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

The Adam optimization algorithm was used for training. We used 100 epochs to train the system on the 1500 files of the development dataset.

3. THE NEVERMIND SYSTEM

3.1. VGGish Model for Audioset

The *NeverMind* system is based on the VGGish model for AudioSet [3]. VGGish is a variant of the VGG model described in [5]. In particular it uses 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.

We use the provided VGGish model that is pre-trained on a large YouTube dataset (a preliminary version of what later became YouTube-8M).

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, with 64 mel bins covering the range 125-7500 Hz. 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.

4. RESULTS

The *SimpleMind* and the *NeverMind* was submitted to the *Making Sense of Sound* audio challenge [2]. The systems was evaluated from a new dataset that consists of 500 audio files, 100 files per category.

The results have been published [2] (see Figure 2). A baseline have been provided by the organizers.

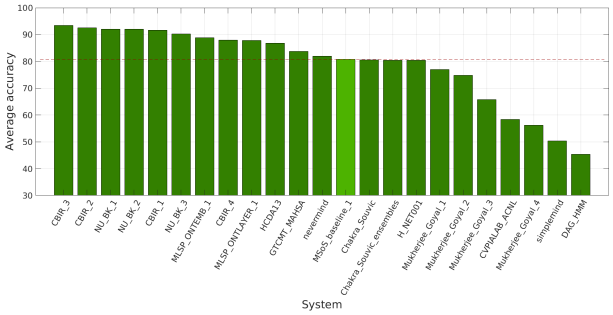

Fig. 2. Results of the Making Sense of Sounds challenge [2].

The *SimpleMind* system shows an accuracy of 50%. The *NeverMind* system reaches 82 % of accuracy and overcome the baseline provided by the organizers (that does not use pre-trained models).

5. CONCLUSION AND PERSPECTIVES

In this document, we presented two systems that was presented to the *Making Sense of Sound* challenge. The *SimpleMind* system is a simple CNN model. It reach only an accuracy of 50%. We suppose that the number of parameters could be to high for the task, with the undesirable consequences of overfitting.

The second system called *NeverMind* reach better results. However, it use external data. With this model, we only use a majority vote based on results for each second of the audio input. In future works, we may improve the temporal modelling of the results. For instance, we could use a more complex system on top of the VGGish model with a recurrent layer.

6. REFERENCES

[1] Librosa - mel spectrogram. `https://librosa.github.io/librosa/generated/librosa.feature.melspectrogram.html`. Accessed: 2019-03-06.

[2] Making sense of sounds data challenge. `https://cvssp.org/projects/making_sense_of_sounds/site/challenge/`. Accessed: 2019-03-06.

[3] Models for audioset: A large scale dataset of audio events. `https://github.com/tensorflow/models/tree/master/research/audioset`. Accessed: 2019-03-06.

[4] 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.

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