Deep Learning Project Proposal : Multi-instrument Classification using Partially Labeled Data and Weakly-supervised Learning

Clément Berger
clement.berger@ensta-paris.fr

Ilyes Er-Rammach
ilyes.er-rammach@ensta-paris.fr

Abstract

While single-instrument recognition has made tremendous progress, multi-instrument recognition is still considered as a hard task. A part of the difficulty comes from the lack of huge strongly labeled datasets. Recently, a large dataset of polyphonic audio clips called OpenMIC has been released. The drawback of the size of the dataset is that it is only weakly-labeled. Previous works have proposed to use an attention mechanism and a Recurrent Neural Network structure called Bidirectional Long-Short Term Memory (BiLSTM) to train on this dataset. Most works in this field use Log-Mel Spectrograms to treat the audio signal before giving it to the network. Here we explore the use of Analytic Wavelet Transform (AWT) to generate scalograms wich are then given to a Convolutional Neural Network (CNN). Such transformations are supposed to be more efficient in giving a representation of the whole signal. We also test different BiLSTM configurations and try some data augmentation techniques. While we do not improve state of the art, our results are encouraging. With more computational power, pretraining our scalogram-CNN structure as feature extractor using a big dataset like YouTube should be able to achieve great results.

1. Introduction

1.1. Motivation

Multi-instrument recognition is a subfield of Music Information Retrieval (MIR) in which, given a list of instruments and an audio clip, one tries to tell if these instruments figure in the clip or not. Such a task is very useful for music providers to make recommandations based on affinity with some instrument, or to create filters for research and so on. An efficient model could also be used as a basis (like a feature extractor) for other MIR tasks such that source separation, or music transcription.

Such a task requires not only machine learning skills, but also signal processing expertise. Great results have been achieved in single-instrument recognition, see e.g. [14]. However, polyphonic sounds are the superposition of multiple instruments with different charasteristics and played differently therefore most of these techniques can not be applied for our task.

Another challenge is the difficulty to create datasets. There are roughly two types of datasets of annotated polyphonic sounds. First there are small datasets but very strongly annotated. We can for example cite MusicNet [13] containing 330 examples. Such datasets face big issues like overfitting. The other type of datasets is huge datasets (at least compared to the previous ones) but only partially labeled. In 2018 has been released a dataset called OpenMIC [6] which belongs to the second category. OpenMIC contains 20 000 audio clips of 10 seconds, sampled at 44,1 kHz. However, given an audio clip, we only know if an instrument is in the clip or not but the offset et onset times are not provided. Practically this means that an audio containing 1 second of violin will have the same label as one completely played with a violin. Moreover, there are some missing labels, meaning that given an audio clip, some instrument labels are not provided.

1.2. Problem definition

1.2.1 Signal processing : different transformations

We provide here a quick overview of two different transformations used in audio processing to create a visual representation of an a signal.

The first one is the genration of a Log-Mel Spectrogram. This is the result of a transformation based on the Short-Time Fourier Transform. The raw signal is divided into a certain number of overlapping frames, before applying a Fourier Transform on each frame. This result in a time frequency representation of the signal, using color variations to represent the magnitude of the Fourier Transforms. However, human sensibility is not homogeneous in frequencies. To account for this problem, we convert the Hertz into what is called a Mel scale [11]. It is a non linear transformation of the frequencies to get a scale which better describe human hearing. The resulting visual representation is called a Log-Mel Spectrogram. Such representations have been

successfully used for music recognition, see e.g. [10].

The second one is the AWT used to generate a scalogram. Given a signal $x(t)$ and a wavelet $\psi(t)$ satisfying $\psi(t) = 0$ for $t < 0$, a function acting as a filter on the signal, the AWT of the signal is $AWT_\psi(t,s) = (1/2\pi)\int_0^\infty \bar{\psi}(s\omega)x(\omega)e^{j\omega}d\omega$. This is a variant of the Continuous Wavelet Transform comonly used in signal processing. We proceed as for the spectrogram to get a visual representation of this transform, which is then called a scalogram.

1.2.2 VGGish network

This section is devoted to briefly present a network created recently for multi-instrument recognition, called VGGish [5]. This name simply comes from the fact that it is derived from the classical VGG model [7] used for image recognition. Amongst other changes, it has been modified to receive a Log-Mel Spectrogram as entry and the end of the network acts now as a compact embedding layer. A precise definition can be found here.

1.2.3 Training using OpenMIC

In this work we try to train a network on the OpenMIC dataset for multi-instrument recognition.

The OpenMIC dataset has been labeled with 20 instruments. For each of the 20 000 audio clips and each instruments, we have a binary value indicating if we have a label for this couple (audio, instrument) or not, and if we have so, we also have the probability that the instrument is effectively in the audio clip. In addition of the raw audios and labels, OpenMIC also provides features extracted using the VGGish. More precisely, a Log-Mel Spectrogram is computed for each raw audio, one frame corresponding to 1 second, and then given to the VGGish. After that, a Principal Component Analysis (PCA) is done on the extracted features for each frame. We end up with 10 vectors (one per second) containing 128 features, for each audio clip.

2. Related work

2.1. Attention mechanism

Multi-instrument recognition is not a new topic, and several works have been published. In 2018, Kong *et al.* [8] introduced an attention mechanism to train a CNN on another weakly-labeled dataset, AudioSet [3]. This mechanism has then been used by Gururani *et al.* [4] on the OpenMIC dataset. Starting from the VGGish features furnished by OpenMIC, an embedding layer is then followed by a prediction layer coupled with the attention mechanism to produce the desired predictions. Let's describe briefly the attention mechanism.

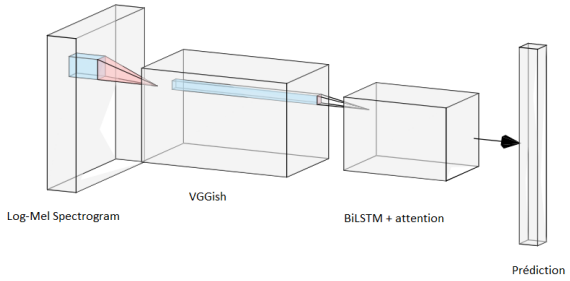

Figure 1. Structure using Log-Mel Spectrogram

In Multi-Instance Label problems, a bag of labels is produced by a score function $S(X) = \mu(f(x)_{x\in X})$, where X is the input composed of multiple instances x, f gives a prediction for each of these instances and μ aggregates these predictions to give the final prediction. The score (the global score S as well as the instance-level scores f) represents in our case the probability of the label (here the probability for the instruments to effectively be part of the audio). Usually a max or an average is used for the aggregation function. However, previous works as Kong *et al.* [8] have shown that learning this function could lead to significant improvements. With this in mind, Gururani *et al.* parametrized this operator as :

$$S(X) = \sum_x w_x f(x)$$

with

$$w_x = \frac{\sigma(v^T h(x))}{\sum_{x'} \sigma(v^T h(x)')}$$

where $h(x)$ is the output of the embedding layer, v is a vector to be learned, and σ is the sigmoid function. Note that the division is here to enforce the sum of coefficients to be 1. The formula also means that the attention layer takes as input the results of the embedding layer and then use this information to aggregate the results of the prediction layer represented by f.

2.2. Using a BiLSTM architecture

2.3. Using scalograms

2.4. Data augmentation

3. Methodology

4. Motivation and problem definition

4.1. Our problem

Given an audio clip and a list of instruments, we aim to determine, for each instrument, if it is played during the audio or not. Our goal is to train a neural network to achieve

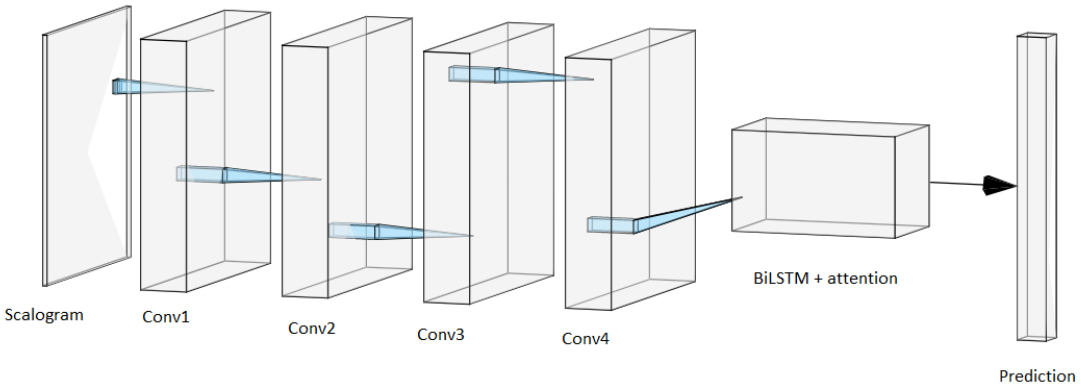

Figure 2. Convolutional network taking scalograms

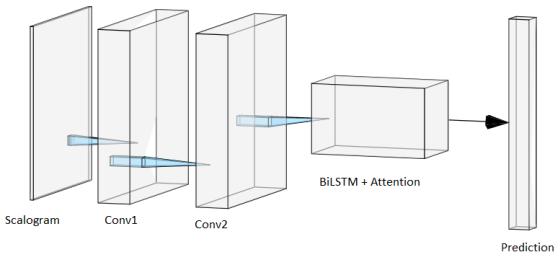

Figure 3. Convolutional network taking scalograms, reduced structure

this task, using the OpenMIC dataset. The underlying goal of using this dataset is to try to find out how to exploit weakly and partially labeled data.

Recently Amir Kenarsari-Anhari [1] got great results training a network on OpenMIC for this purpose. His paper will be the basis of our work, in the sense that we will try to improve his algorithm, as we will discuss in the next section.

5. Methodology

5.1. Squeletton of the algorithm

Let's briefly describe the algorithm of [1]. Given an audio clip, we compute its log mel-spectrogram, which consists in a transformation of the signal (in the same sense than a Fourier transform). This is then given to a Convolutional Neural Network (CNN). Its structure is close to the one of the VGG network used in image recognition, reason why it will be referred as the VGGish layer [5]. The next step is to use an attention layer which consists in learning signifiant segments in the clip, see [4]. This ends with a classification layer.

Our work will follow the structure of this model. Our goal will be to modify some of these blocks to try to improve the performances of the algorithm.

5.2. Changing the feature learning

As we discussed, multi-instrument classification and single-instrument classification are pretty different fields. However they both have to deal with the pre processing of the signal. This summer Dutta *et al*. [2] experimented a scalogram different from the mel-scalogram to train a network on single-instrument classification. The idea of their technique is to use features computed on the whole signal while other methods subdivise the signal before processing each part. As they achieved greater results than previous works, our idea would be to try to use this scalogram for our model to get maybe more representative features which would hopefully result in a better training.

Another modification indirectly suggested by this paper would be to even change the VGGish layer and use the one of Dutta *et al*. This would allow us to try different CNN for this part of the algorithm.

We also note that Dutta *et al*. trained only on 14 instruments (compared to 20 in OpenMIC). Experimenting it on our algortihm will so be a double test, trying not only a different task but also different instruments.

5.3. Data augmentation

As Amir Kenarsari-Anhari suggested at the end of his paper, another direction of research to improve his model is to experiment more data augmentation. Our primary goal will be to experiment on changing the feature learning process,

but if possible we plan to try to add different data augmentation techniques, as it is a natural way to try to deal with the small size of the dataset.

A starting point may be a paper from Schlüter and Grill [12] using data augmentation for singing voice detection. Even if voice detection is quite different from instrument classification, this paper presents data augmentation techniques which are a bit more general than just voice detection. Moeover we can also try to take some of the techniques specific to voice detection and to adapt them to our problem, potentially leading to new techniques.

6. Evaluation

We will first try to implement the changes discussed in 2.2., namely changing the scalogram and the VGGish layer. If it's possible, we will then explore some data augmentation techniques. We will experiment different combinations of our changes and compare them at first to the results of [1] to see if we are able to achieve a better performance. We will also make comparisons between them to try to find out the most effective one(s) and try to deduce some interpretation of why some are better than the others. As our purpose is to improve the algorithm of [1], it will constitute our benchmark, meaning that we will focus on comparing our best results with theirs.

We precise that the performances of an algorithm are not measured in terms of the usual accuracy in the field of instrument classification. Indeed, as some instruments are easier than others to recognize, a difference of accuracy doesn't perfectly reflect an improvement in the classification of hard instruments. Therefore we will use the F1 score which is a metric created specificaly for this purpose. Details can be found in [9]. In order to produce a meaningful comparison, we will more precisely use the parameters that are described in [1].

References

[1] Amir Anhari. Learning multi-instrument classification with partial labels. 01 2020. 3, 4

[2] Arindam Dutta, Dibakar Sil, Aniruddha Chandra, and Sarbani Palit. Cnn based musical instrument identification using time-frequency localized features. pages 1–6, 05 2020. 3

[3] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 2

[4] Siddharth Gururani, Mohit Sharma, and Alexander Lerch. An attention mechanism for musical instrument recognition. *CoRR*, abs/1907.04294, 2019. 2, 3

[5] S. Hershey, S. Chaudhuri, D. P. W. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold, M. Slaney, R. J. Weiss, and K. Wilson. Cnn architectures for large-scale audio classification. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135, 2017. 2, 3

[6] Eric J. Humphrey, S. Durand, and B. McFee. Openmic-2018: An open data-set for multiple instrument recognition. In *ISMIR*, 2018. 1

[7] A. Zisserman K. Simonyan. Very deep convolutional networks for large-scale image recognition. *ArXiv*, September 2014. 2

[8] Qiuqiang Kong, Yong Xu, and Mark Plumbley. Audio set classification with attention model: A probabilistic perspective. 04 2018. 2

[9] A. Mesaros, Toni Heittola, and T. Virtanen. Metrics for polyphonic sound event detection. *Applied Sciences*, 6:162, 2016. 4

[10] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. *CoRR*, abs/1703.06697, 2017. 2

[11] E. B. Newman S. S. Stevens, J. Volkmann. A scale for the measurement of the psychological magnitude pitch. *Acoust. Soc. Am.*, 8, 1937. 1

[12] Jan Schlüter and Thomas Grill. Exploring data augmentation for improved singing voice detection with neural networks. 01 2015. 4

[13] John Thickstun, Z. Harchaoui, and Sham M. Kakade. Learning features of music from scratch. *ArXiv*, abs/1611.09827, 2017. 1

[14] M. Lagrange V. Lostanlen, J. Andén. Extended playing techniques: The next mile-stone in musical instrument recognition. *5th International Conference on Digital Libraries for Musi-cology*, 09 2016. 1