Proceedings of the Federated Conference on Computer Science and Information Systems pp. 387–390
DOI: 10.15439/2020F82
ISSN 2300-5963 ACSIS, Vol. 21

Retrieving Sound Samples of Subjective Interest With User Interaction

Jan Jakubik
Department of Computational Intelligence
Wroclaw University of Science and Technology
jan.jakubik@pwr.edu.pl

***Abstract*—This paper concerns the retrieval of audio samples with a high degree of user interaction, motivated by a practical use case. We consider an open set recognition scenario in which the goal is to find all occurrences of a subjectively interesting sound selected by a user within a particular audio file. We use only a single starting example and maintain interaction through yes-no answers from the user, indicating whether any new retrieved sound matches the target pattern. We present a small dataset for this task and evaluate a baseline solution based on Nonnegative Matrix Factorization and greedy feature selection.**

***Index Terms*—music information retrieval, matrix decomposition, active learning**

I. Introduction

AUDIO retrieval is a well-established research area with multiple practical use cases. Between music audio [1], sound effect [2], and speech [3] analysis, numerous research problems have been established and tackled with a range of techniques from the areas of signal processing and machine learning. Most novel approaches developed in recent years have leveraged the success of deep learning [4] and more generally, machine learning methods have been deployed in the area for decades.

However, machine learning systems and the methodology of their evaluation can arise concerns about their practicality. The majority of ML systems are evaluated with the implicit assumption of availability of annotated data with a distribution identical to that of the real domain. Quite often, classification tasks are defined with the assumption that the set of classes that need to be recognized does not change over time [5]. Approaches which adress these problems in current ML literature are known as zero-shot and one-shot learning, and the tasks they solve can be described as open-set recognition [6]. The prior knowledge of the target problem is minimized at the time of training and the goal instead is to maximize the system's capability to tackle new problems with as little data as possible.

In this paper, we consider a sound sample retrieval system which requires an open-set recognition scenario. We are motivated by a practical task defined within an R&D project in cooperation with a game development company. Specifically, we attempt to search for sound samples of subjective interest in electronic music, without any prior knowledge of what distinguishing features these sounds may have. To guide the systems' decisions, we are using only a single positive sample and responses resulting from user interactions to narrow down the search. The goal of this system is to allow efficient creation of content synchronized with music audio and as such, it should minimize the user's effort while achieving maximal recall.

The contributions of this paper are as follows: we propose a well-defined, zero-shot active learning scenario, motivated by a practical use case. We provide an evaluation dataset focused on electronic music composed of repetitive samples. The dataset is annotated based on listeners' subjective notion of what constitutes a "sound of interest". Finally, we evaluate an approach that may serve as a simple non-deep learning baseline for this problem and a reference for future work.

II. Related Work

Sound effect retrieval has been considered in several contexts, however, it is usually not in an active learning scenario. The exsiting work [7] focuses on general sound effects against non-musical background. Recent papers have attempted zero-shot learning for music auto-tagging with good results [8].

Active learning techniques have been developed mainly with the goal of lowering annotation costs of full datasets. As such, two types of techniques are typically proposed [9]. The first group consists of approaches based on model uncertainty [10], in which the selection of a new sample is based on the "difficulty" of training samples. A number of criteria for selecting difficult samples have been proposed. The second group of active learning approaches exploits the distribution of samples over the feature space or label space and aims to draw the most representative samples [11]. Approaches of this type can utilize clustering methods to find good representative samples for the entire dataset.

Zero-shot learning has been considered mostly in the area of deep learning, where the ability of artificial neural networks to learn meaningful features from a low-level representation of data can be leveraged [12]. A pre-trained deep network feature extractor allows comparison of samples that emphasizes semantic similarities. This type of extractor can be trained without prior knowledge of classes that need to be recognized.

III. Materials and Methods

In this section, data and methods employed in the study are described. We summarize the problem definition, the dataset gathered for validation of the developed methods and standard

signal processing and machine learning approaches that can be employed to build a baseline system for this task.

A. Problem Definition

The task in question was defined as the retrieval of interesting sounds, with a focus on particular samples that may be used repeatedly in electronic music. The use scenario was described as follows: an end-user, programmer or game designer, should be able to mark an interesting excerpt within the audio file that contains a "sound of interest". The nature of such a sound is not well-defined. Other occurrences of the sound can be slightly altered. The sound can be easily recognized against a variety of audio backgrounds.

Since a machine learning system for such a task requires a training dataset of positive or negative samples, and the user should not be expected to supply annotations before the retrieval process, we opted for a solution that combines active learning and zero-shot learning approaches. The user only supplies a single positive example and then receives a new sample after each query which they can mark as either positive or negative. This process continues for a limited number of queries, given by a pre-defined budget. The scenario is consistent with the definition of active learning but differs in detail from typically considered AL scenarios. In particular, the standard AL approaches seek to maximize overall performance inprovements, and are considered for entire datasets. Our scenario on the other hand aims to minimize user interaction specifically with negative samples while retrieving all positive samples from a single audio file. This difference is meaningful because many AL approaches base their choice of data for annotation on an uncertainty criterion, i.e., the user would be shown samples that are "equally likely" to be positive or negative.

Algorithm 1 Active Retrieval Procedure

```
function RETRIEVE(t_0, l, b)
    P ← {t_0}
    N ← ∅
    while |N| < b do
        newsample ← GetBestSamples(P, N, l)
        if UserResponse(newsample) = positive then
            P ← P ∪ {newsample}
        else
            N ← P ∪ {newsample}
        end if
    end while
    return P
end function
```

Formally, our input is a sequence of vectors $X = (x_1, x_2, ...x_n)$ which represents the sound file (detailed in subsections C and D), starting point t_0 and length l of the initial positive example, and an answer budget b, which represents the user's patience. We seek a function that given two sets of positive examples P and negative examples N returns the time points at which other occurences of the sound of interest start. The overall search procedure is given by Algorithm 1, in which $GetBestSamples$ is the method used for retrieval (described in subsections E and F), while the function $UserResponse$ is the true-false response from user interaction.

B. The Dataset

The dataset for this study was created based on songs available at sampleswap.com, a website offering a variety of creative commons licensed electronic music. Our goal was to represent the use case of searching for electronic samples of interest, in particular, characteristic and repeating sound effects. The key issue here is that the samples may be present with a variety of different musical backgrounds.

300 audio files were chosen from the sampleswap.com repository and annotated by three people - two trained musicians and one non-musician. The annotations were based on a subjective notion of an interesting sound, with the caveat that the sound must occur multiple times within the file.

Files from four genre categories were annotated - Dubstep, House, Downtempo and Drum'n'Bass. Most audio files in the data were approximately 2 minutes long and the length of "sounds of interests" ranged between 1 and 7 seconds.

The dataset is available on request.

C. Representations of Audio Data

Within the study, we compare three different representations of audio data. We employ the standard approach of representing the sound file as a series of real-space vectors.

Basic approaches in music analysis use spectrogram representations that capture the local distribution of the signal's power over frequency bins. In this study, we include standard Short-Term Fourier Transform (STFT) with linear frequency scale, as well as Constant-Q Transform (CQT), in which frequency bins are spaced logarythmically. The latter corresponds better to the psychoacoustic properties of human hearing, as well as standard western musical scales.

In genre analysis and sound classification, another significant method of audio representation are Mel-Frequency Cepstral Coefficients [13]. MFCC are especially well known for capturing the timbral properties of sound and were commonly used in all types of audio ML tasks before the dominance of deep learning. MFCC can be extracted from short frames of audio and used to create a vector sequence analogous to a spectrogram.

In recent work on music analysis, it is also very common to use learned representations extracted by a pre-trained deep neural network, usually convolutional (CNN). In our preliminary experiments we tested neural representations transferred from other tasks (e.g., CNN autotagging on IRMAS dataset, classification on GTZAN). However, we failed to find one that outperforms NMF on the aformentioned standard features.

D. Matrix Decomposition Approach

Matrix decomposition methods are a standard approach in audio signal processing. Through matrix factorization methods, every spectrogram frame can be expressed as a linear

combination of a number of base vectors corresponding to commonly occurring "sound components". Our baseline approach uses a factorized representation to identify the key components of the sound of interest and search for other occurrences within the audio file. Unlike deep learning feature extractors, MF representation can be trained on the level of a single audio file and only separate components relevant to that particular file, which makes it a good baseline without the requirement of training on a large dataset.

A nonnegative matrix factorization (NMF) is a decomposition that represents a given matrix X as a product of two nonnegative matrices W and H. As an optimization problem, NMF is obtained by solving Eq. 1:

$$\arg \min_{W>0,H>0} \|X - WH\|_F^2 \tag{1}$$

Additional constraints can be imposed to induce sparsity on matrices W, H, or minimize their Frobenius norm. For our baseline, we use the implementation in the scikit-learn library that allows both Frobenius norm and L1 norm regularization. The exact optimization problem in the scikit-learn implementation is formulated as follows (Eq. 2):

$$\begin{aligned} \arg \min_{W>0,H>0} ||X - WH||_F^2 + \alpha l(||W||_1 + ||H||_1) + \\ + \alpha(1-l)(||W||_F^2 + ||H||_F^2) \end{aligned} \tag{2}$$

The NMF representation allows for better retrieval of sounds of interest when other background sounds are present. The simple model can separate different additive components of the data matrix, and our expectation is that this will help separate different sound sources, including one corresponding to the sample of interest.

E. Feature Selection

Our baseline idea is to use NMF to separate the sound of interest from its background. It is particularly useful to employ a selection mechanism that chooses only the NMF components that give the best separation between positive and negative samples. This selection is performed after every user decision. Formally, the criterion is defined in Eq. 3:

$$\arg\max_i \min_{x \in P, y \in N} \|f_i(x) - f_i(y)\|_2 \tag{3}$$

where $f_i(x)$ indicates the i-th feature value of the sample x. The feature selection is greedy and the size of the selected subset of features is a hyperparameter we tune experimentally.

F. Best Sample Retrieval

Feature selection is repeated after every user decision, and the actual retrieval procedure is then based on simple nearest neighboor. Using the selected features, we choose an excerpt of given length with the lowest distance from its closest neighboor set of positives that does not overlap with any of the already returned excerpts.

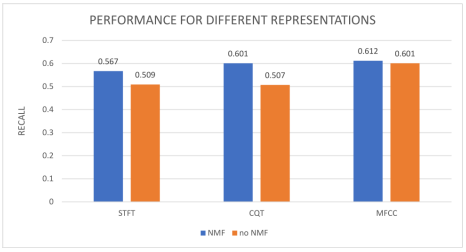

Fig. 1. Results of retrieval depending on audio representation

IV. Results

The experiments were performed using librosa [14] and scikit-learn [15] libraries. Implementations of standard STFT, CQT, and MFCC extraction in librosa were all used with default parameters. The data was then decomposed using a version of NMF provided in scikit-learn.

The classifier is a simple nearest neighbor method that returns the new sample that minimizes l2 distance to the closest positive sample while not overlapping with any of the samples in either positive or negative set. We use an answering budget of $b = 10$ and use recall as our main figure of merit. Since annotations are not perfectly timed due to human limitations, we consider every retrieved excerpt that overlaps in time with a ground truth excerpt for more than half of its duration as a true positive.

A. Comparison of Audio Representations

The first experiment compares the results achieved with different audio representations. MFCC, CQT, and STFT representations are compared in two variants: without NMF and after NMF. Results in terms of the recall are presented in Fig. 1.

There is a clear improvement resulting from the use of NMF to separate the basic components of the sound. In addition, the use of MFCC appears to be preferable to the use of standard frequency-domain transforms. The best recall of 61% is achieved with NMF of an MFCC representation.

B. Influence of NMF Hyperparameters

The purpose of this second experiment is to examine the influence of NMF hyperparameters on the recall of the active retrieval procedure. The regularization of NMF is of key importance in the methods' capability to separate distinct sounds. In particular, the sparsity parameter encourages separation into a sparse dictionary of sounds, i.e., every frame of the source data can be expressed as a sum of only a few components. Additionally, the number of components itself is a hyperparameter that will affect the results significantly.

Fig. 2 shows the result of comparison of recall depending on the number of NMF components. There is a clear negative

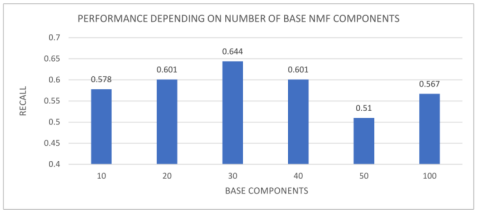

Fig. 2. Results of retrieval depending on number of NMF components.

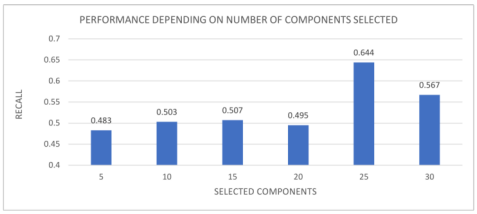

Fig. 3. Results for different number of selected features.

trend past 20 components, suggesting that number of components is sufficient to express relevant information about sounds within a single file.

We have also tested changes to L1 ratio and alpha parameter of regularization. Overall, the default parameters of NMF supplied in Librosa appear to give good results. The changes from parameter tuning offer a small gain over the default parameterization, and the use of regularization compared to lack thereof improves the recall by a margin of 1% at best. The use of L1 regularization, despite some intuitive basis for it, only worsens the results.

C. Influence of feature selection

In this experiment, we evaluate the influence of feature selection on the final result. The experiments are performed with the best parameters chosen from previous tests: 30 NMF components, $\alpha = 1$, and no L1 regularization. The results are presented in Fig. 3.

There is a visible positive effect of active feature selection on the results when the number of features selected is slightlly lower than the number of base components.

V. Conclusions and Future Work

We have defined a practically motivated sound retrieval task and presented a dataset and a simple baseline approach for its evaluation. Our task concerns retrieval of sounds of subjective interest within a single audio file based on user interaction in the form of simple yes-no answers. The presented solution uses Nonnegative Matrix Factorization to identify a base of audio components and feature selection to focus on components specific to the sound of interest. Simple nearest neighbor is then used to find the potential answers to the user's query.

The baseline demonstrates performance of 64% recall, which leaves significant room for improvement. The key area of future work is the potential use of feature learning, in particular deep network representations trained on sufficiently large unannotated data. It is also likely that the feature selection approach can improved beyond simple greedy selection based on the distance criterion.

VI. Acknowledgements

This research was carried out in cooperation with Scalac sp. z o.o. and Vixa Games game development team as part of the "Elaboration and implementation of audio modules (audio tracks indexing and analysis) and video modules (visualization of processed data) involving unique sound and graphic design use for the multimedia applications requirements" project co-funded by the European Union.

References

[1] Rainer Typke, Frans Wiering, and Remco Veltkamp. A survey of music information retrieval systems. pages 153–160, 01 2005.

[2] Douglas Turnbull, Luke Barrington, David Torres, and Gert Lanckriet. Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476, 2008.

[3] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recogn.*, 44(3):572–587, March 2011.

[4] Allen Huang and Raymond Wu. Deep learning for music. *arXiv preprint arXiv:1606.04930*, 2016.

[5] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(7):1757–1772, 2012.

[6] Wei Wang, Vincent W Zheng, Han Yu, and Chunyan Miao. A survey of zero-shot learning: Settings, methods, and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 10(2):1–37, 2019.

[7] Zhao Shuyang, Toni Heittola, and Tuomas Virtanen. Active learning for sound event classification by clustering unlabeled data. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 751–755. IEEE, 2017.

[8] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. Zero-shot learning for audio-based music classification and tagging. *arXiv preprint arXiv:1907.02670*, 2019.

[9] Yifan Fu, Xingquan Zhu, and Bin Li. A survey on instance selection for active learning. *Knowledge and information systems*, 35(2):249–283, 2013.

[10] A. Holub, P. Perona, and M. C. Burl. Entropy-based active learning for object recognition. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

[11] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[12] Jeong Choi, Jongpil Lee, Jiyoung Park, and Juhan Nam. Zero-shot learning for audio-based music classification and tagging. *arXiv preprint arXiv:1907.02670*, 2019.

[13] Vibha Tiwari. Mfcc and its applications in speaker recognition. *International journal on emerging technologies*, 1(1):19–22, 2010.

[14] Brian McFee, Colin Raffel, Dawen Liang, Daniel PW Ellis, Matt McVicar, Eric Battenberg, and Oriol Nieto. librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8, 2015.

[15] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.