UNSUPERVISED DEEP FEATURE LEARNING FOR MUSIC SEGMENTATION

Matthew C. McCallum
Gracenote Inc.

ABSTRACT

Music segmentation refers to the dual problem of identifying boundaries between, and labelling, distinct music segments, e.g., the chorus, verse, bridge etc. in popular music. Investigation into the performance of a range of music segmentation algorithms has shown significant differences in performance dependent on the audio features chosen at the input to such algorithms. Some approaches have proposed learning feature transformations from music segment annotation data to improve music segmentation performance. While annotated music segmentation data is a scarce resource, the amount of music data available in the industry is much greater. Here it is proposed to use unsupervised training of convolutional neural networks for audio embedding in the music segmentation problem.

1. INTRODUCTION

Music segmentation is in practice two functionally distinct but related problems. That of segment labeling, where contiguous regions of a song are labeled as the same or distinct segments, and boundary detection where the approximate time points at which changes in segment labels occur [8]. Current approaches in the literature can be broadly classed into two categories The clustering approach [4, 5] and the novelty function approach [1, 9]. In the former one of a number of clustering approaches are employed to group carefully chosen features that are similar within segments but dissimilar between segments. In the latter, a self similarity matrix (SSM) [1] or derivative thereof [9] is analyzed to detect changes in the patterns of similarity between features throughout the song. This is motivated by the observation that contiguous perceptually distinct segments in music typically contain either homogeneity in features or repetition. Sudden changes in the SSM distances may therefore indicate the beginning or end of a segment within the song. Detection of song boundaries in this way is typically performed by convolving an SSM with a "change kernel" that will emphasize changes in the SSM structure and create a novelty function. The peaks in the novelty function may then be detected as music segment boundaries via a peak picking algorithm.

2. FEATURE ENGINEERING

The literature considers a range of features from which structure representations such as SSMs may be constructed. Typically features that are representative of timbre such as Bark band coefficients [1], pitch such as harmonic pitch class profiles (HPCPs) [9], or a combination thereof [5] are considered. The decision on the specific feature employed for a given application may be motivated by utility of these features for the content of interest. A broad evaluation on the effectiveness of a range of features used in a range of algorithms is available in [7].

With the myriad of music data that exists today it is reasonable to ask whether features can be derived from such data that are not hand crafted based on knowledge of short time physical or perceptual properties (e.g., timbre or pitch), but learned based on common patterns in music. For example, a given segment in music may contain a similar tempo, rhythm, cadences, instrumentation etc., despite changes in pitch and timbre.

In recent years, several supervised approaches have been proposed that attempt to make use of large amounts of data to improve music segmentation performance [6] and [11]. In particular, McFee and Ellis [6] proposed learning a linear transformation of several aggregated features representative of timbre, pitch and repetition. The linear transformation was learned from labeled music segmentation datasets to maximize separation between classes and minimize variance within classes using ordinal linear discriminant analysis (OLDA), where each class represents a unique segment within a song. Ullrich et. al. [11] proposed to forgo explicit feature learning and map input features directly to a novelty function using convolutional neural networks (CNNs) with promising results.

3. UNSUPERVISED FEATURE LEARNING

The datasets available for addressing the music segmentation problem in a supervised manner are limited in size with respect to modern machine learning standards. The largest publicly available dataset for the task is a subset of the SALAMI dataset [10]. While this resource has proved useful in some supervised machine learning settings [6, 11], it is relatively small in comparison to many modern machine learning experiments, e.g., [2]. Due to the time or resources it takes to search a music timeline and annotate boundaries [12], it is of interest to question what can be gained from unsupervised machine learning techniques for the music segmentation problem. Recently, it was shown that by training audio embeddings on very large datasets can significantly improve audio classification performance in relatively shallow classifiers [3]. This approach employed class preserving augmentations to audio features (e.g., additive noise, time/frequency translation, source mixing etc.) and used a triplet loss function to train a CNN that could map these augmentations with complex distance relationships in the feature domain, to deep features where class relationships can be interpreted via simple distance metrics, e.g., via squared Euclidean distance.

4. DEEP FEATURES FOR MUSIC SEGMENTATION

Considering the unsupervised training methodology in [3], the significant effect of feature selection on music segmentation algorithms [7], and the difficulty of creating quality music segmentation annotations [12], it is of interest to investigate the effectiveness of employing unsupervised deep embedding methods for the music segmentation problem. Many of the distortions employed in [3] are not class preserving for the music segmentation problem. However, because distinct music segments consist of contiguous portions of a song's timeline that typically form a minorty of the length of the song, the time proximity of features in a song is an important trait that may be exploited in the unsupervised learning of deep features for music segmentation.

Here it is proposed to use triplet loss to create a contextual embedding of audio features for the purposes of music segmentation. Specifically, a CNN may be trained using gradient descent, to embed features in a space that maintains shorter Euclidean distance relationships between features that typically occur together within a short time context relative to the distances between features that rarely occur together within a similar context. A triplet loss function over a batch of anchor, positive and negative examples may be written as,

$$l(A,P,N) = \sum_{i=0}^{B} \left[|f(A_i) - f(P_i)|_2^2 - |f(A_i) - f(N_i)|_2^2 + \alpha \right]_+ , \tag{1}$$

where A, P, and N correspond to batches of anchor, positive (similar) and negative (dissimilar) examples respectively, in which individual samples are indexed by i. α corresponds to a margin that is enforced between the relative distances of anchor and positive, and anchor and negative examples.

The positive and negative examples may be sampled from the CQT representation of a single song in an unsupervised fashion. Specifically, it is proposed to choose an arbitrary CQT window index and select a 2D CQT representation centered on this index as an anchor example (e.g., all CQT analysis windows within +/- 256 windows or +/- 2 beats for beat synchronized CQT representations, of an arbitrary index). A single positive example is then selected by uniformly sampling an index upon which to center a similar contiguous segment at a range of +/- δ_p from the anchor index. A single negative example is then selected by uniformly sampling the center index of a third contiguous from all indexes within the entire timeline of the song, excluding indeces at a distance of $< \delta_n$ from the selected anchor index. This sampling paradigm may be performed many times on each of N tracks to form a single batch of training data.

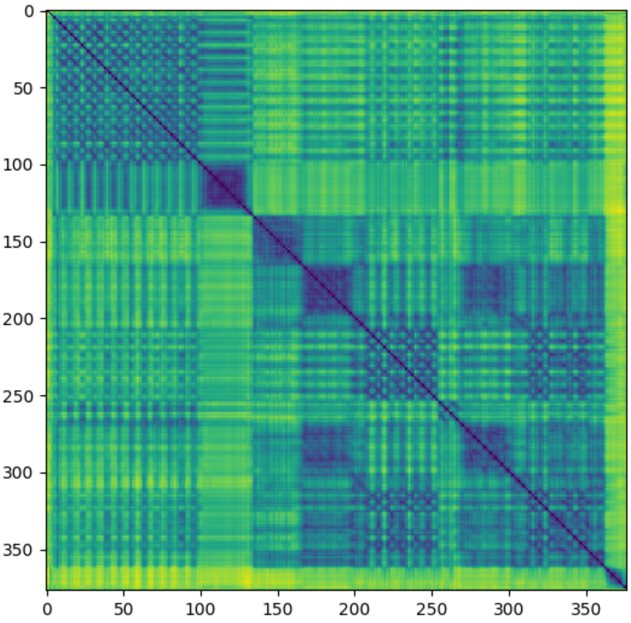

Figure 1. Self similarity matrix of Beatles - Birthday computed on Euclidean distance between deep embeddings of beat synchronized CQT features

Trimmed Metric	F-Measure	Precision	Recall
Deep Features	0.6624	0.6631	0.6909
CQT Features	0.4548	0.5128	0.4261

Table 1. Music segmentation boundary detection performance with checkerboard kernel method on CQT and deep features. Evaluated on the Beatles TUT dataset.

Once trained, the CNN may be used to analyze the CQT representation of a given song. Typically this analysis portrays distinct regions of similarity and dissimilarity throughout a song with clear jumps in the feature embedding. This is evident in the SSM depicted in Fig. 1, which unlike SSM representations usually encountered in music segmentation algorithms, consists almost entirely of block structures with few path structures.

The deep embedded representation of a song may be analyzed with one of a number of previously proposed song segmentation algorithms. Perhaps the simplest of which is the checkerboard kernel method of Foote [1]. Several features were investigated for this music segmentation method in [7], with the CQT being the most effective. It can be seen in Table 1 that when employing the same method on the CQT representation and the deep embedded representation of the Beatles TUT dataset, a significant improvement is observed with respect to the trimmed 3-second tolerance segmentation boundary hit rate metrics of [8]

5. REFERENCES

[1] Jonathan Foote. Automatic audio segmentation using a measure of audio novelty. In *Multimedia and Expo (ICME). IEEE International Conference on*, volume 1, pages 452–455. IEEE, 2000.

[2] Jort F Gemmeke, Daniel PW 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 *Acoustics, Speech and Signal Processing (ICASSP), IEEE International Conference on*, pages 776–780. IEEE, 2017.

[3] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. *arXiv preprint arXiv:1711.02209*, 2017.

[4] Mark Levy and Mark Sandler. Structural segmentation of musical audio by constrained clustering. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):318–326, 2008.

[5] Brian McFee and Dan Ellis. Analyzing song structure with spectral clustering. In *ISMIR*, pages 405–410, 2014.

[6] Brian McFee and Daniel PW Ellis. Learning to segment songs with ordinal linear discriminant analysis. *Self*, 275:330, 2014.

[7] Oriol Nieto and Juan Pablo Bello. Systematic exploration of computational music structure research. In *ISMIR*, pages 547–553, 2016.

[8] Colin Raffel, Brian McFee, Eric J Humphrey, Justin Salamon, Oriol Nieto, Dawen Liang, Daniel PW Ellis, and C Colin Raffel. mir_eval: A transparent implementation of common mir metrics. In *ISMIR*, 2014.

[9] Joan Serra, Meinard Müller, Peter Grosche, and Josep Ll Arcos. Unsupervised music structure annotation by time series structure features and segment similarity. *IEEE Transactions on Multimedia*, 16(5):1229–1240, 2014.

[10] Jordan Bennett Louis Smith, John Ashley Burgoyne, Ichiro Fujinaga, David De Roure, and J Stephen Downie. Design and creation of a large-scale database of structural annotations. In *ISMIR*, volume 11, pages 555–560, 2011.

[11] Karen Ullrich, Jan Schlüter, and Thomas Grill. Boundary detection in music structure analysis using convolutional neural networks. In *ISMIR*, pages 417–422, 2014.

[12] Cheng-i Wang, Gautham J Mysore, and Shlomo Dubnov. Re-visiting the music segmentation problem with crowdsourcing. In *ISMIR*, pages 738–744, 2017.