TOWARDS A UNIVERSAL REPRESENTATION FOR AUDIO INFORMATION RETRIEVAL AND ANALYSIS

Bjørn Sand Jensen, Rasmus Troelsgaard, Jan Larsen, and Lars Kai Hansen

DTU Compute
Technical University of Denmark
Asmussens Allé B305, 2800 Kgs. Lyngby, Denmark
{bjje,rast,janla,lkai}@dtu.dk

ABSTRACT

A fundamental and general representation of audio and music which integrates multi-modal data sources is important for both application and basic research purposes. In this paper we address this challenge by proposing a multi-modal version of the Latent Dirichlet Allocation model which provides a joint latent representation. We evaluate this representation on the Million Song Dataset by integrating three fundamentally different modalities, namely tags, lyrics, and audio features. We show how the resulting representation is aligned with common 'cognitive' variables such as tags, and provide some evidence for the common assumption that genres form an acceptable categorization when evaluating latent representations of music. We furthermore quantify the model by its predictive performance in terms of genre and style, providing benchmark results for the Million Song Dataset.

Index Terms— Audio representation, multi-modal LDA, Million Song Dataset, genre classification.

1. INTRODUCTION

Music representation and information retrieval are issues of great theoretical and practical importance. The theoretical interest relates in part to the close interplay between audio, human cognition and sociality, leading to heterogenous and highly multi-modal representations in music. The practical importance, on the other hand, is evident as current music business models suffer from the lack of efficient and user friendly navigation tools. We are interested in representations that directly support interactivity, thus representations based on latent variables that are well-aligned with cognitively (semantic) relevant variables [1]. User generated tags can be seen as such 'cognitive variables' since they represent decisions that express reflections on music content and context.

Clearly, such tags are often extremely heterogenous, high-dimensional, and idiosyncratic as they may relate to any aspect of music use and understanding.

Moving towards broadly applicable and cognitively relevant representations of music data is clearly contingent on the ability to handle multi-modality. This is reflected in current music information research that use a large variety of representations and models, ranging from support vector machine (SVM) genre classifiers [2]; custom latent variable models models for tagging [3]; similarity based methods for recommendation based on Gaussian Mixture models [4]; and latent variable models for hybrid recommendation [5]. A significant step in the direction of flexible multi-modal representations was taken in the work of Law et al. [6] based on the probabilistic framework of Latent Dirichlet Allocation (LDA) topic modeling. Their topic model representation of tags allows capturing rich cognitive semantics as users are able to tag freely without being constrained by a fixed vocabulary. However, with a strong focus on automatic tagging Law et al. refrained from developing a universal representation - symmetric with respect to all modalities. A more symmetric representation is pursued in recent work by Weston et al. [7]; however, without a formal statistical framework it offers less flexibility, e.g., in relation to handling missing features or modalities. This is often a challenge encountered in real world music applications.

In this work we pursue a multi-modal view towards a unifying representation, focusing on latent representations informed symmetrically by all modalities based on a multi-modal version of the Latent Dirichlet Allocation model. In order to quantify the approach, we evaluate the model and representation in a large-scale setting using the million song dataset (MSD) [8], and consider a number of models trained on combinations of the three basic modalities: user tags (top-down view), lyrics (meta-data view) and content based audio features (bottom-up view). First, we show that the latent representation obtained by considering the audio and lyrics modalities is well aligned—in an unsupervised manner - with 'cognitive' variables by analyzing the mutual information...
between the user generated tags and the representation itself. Secondly, with knowledge obtained in the first step, we evaluate auxiliary predictive tasks to demonstrate the predictive alignment of the latent representation with well-known human categories and metadata information. In particular we consider genre and styles provided by [9], none of which is used to learn the latent semantics themselves. This leads to benchmark results on the MSD and provides insight into the nature of generative genre and style classifiers.

Our work is related to a rich body of studies in music modeling, and multi-modal integration. In terms of non-probabilistic approaches this includes the already mentioned work of Weston et al. [7]. McFee et al. [10] showed how hypergraphs (see also [11]) can be used to combine multiple modalities with the possibilities to learn the importance of each modality for a particular task. Recently McVicar et al. [12] applied multi-way CCA to analyze emotional aspects of music based on the MSD.

In the topic modelling domain, Arenas-García et al. [13] proposed multi-modal PLSA as a way to integrate multiple descriptors of similarity such as genre and low-level audio features. Yoshii et al. [5, 14] suggested a similar approach for hybrid music recommendation integrating subject taste and timbre features. In [15], standard LDA was applied with audio words for the task of obtaining low-dimensional features (topic distributions) applied in a discriminative SVM classifier. For the particular task of genre classification et al. [16] applied the pLSA model as a generative genre classifier. Our work is a generalization and extension of these previous ideas and contributions based on the multi-modal LDA, multiple audio features, audio words and a generative classification view.

2. DATA & REPRESENTATION

The recently published million song dataset (MSD) [8] has highlighted some of the challenges in modern music information retrieval; and made it possible to evaluate top-down and bottom-up integration of data sources on a large scale. Hence, we naturally use the MSD and associated data sets to evaluate the merits of our approach. In defining the latent semantic representation, we integrate the following modalities/data sources.

The tags, or top-down features, are human annotations from last.fm often conveying information about genre and year of release. Since users have consciously annotated the music in an open vocabulary, such tags are considered an expressed view of the users cognitive representation. The metadata level, i.e., the lyrics, is of course nonexistent for for majority of certain genres, and in other cases simply missing for individual songs which is not a problem for the proposed model. The lyrics are represented in a bag-of-words style, i.e., no information about the order in which the terms occurs is included. The content based or bottom up features are derived from the audio itself. We rely on the Echonest feature extraction¹ already available in for the MSD, namely timbre, chroma, loudness, and tempo. These are orginally derived in event related segments, but we follow previous work [17] by beat aligning all features obtaining an meaningful alignment with music related aspects.

In order to allow for practical and efficient indexing and representation, we abandon the classic representation of using for example a Gaussian mixture model for representing each song in its respective feature space. Instead we turn to the so-called audio word approach (see e.g. [18, 19, 3, 17]) where each song is represented by a vector of counts of (finite) number of audio words (feature vector). We obtain these audio words by running a randomly initiated K-means algorithm on a 5% random subset of the MSD for timbre, chroma, loudness and tempo with 1024, 1024, 32, and 32 clusters, respectively. All beat segments in a all songs are then quantized into these audio words and the resulting counts, representing the four different audio features, are concatenated to yield the audio modality.

3. MULTI-MODAL MODEL

In order to model the heterogeneous modalities outline above, we turn to the framework of topic modeling. We propose to use a multi-modal modification of the standard LDA to represent the latent representation in a symmetric way relevant to many music applications. The multi-modal LDA, mmLDA, [20] is a straight forward extension of standard LDA topic model [21], as shown in Fig. 1. The model and notation is easily understood by the way it generates a new song by the different modalities, thus the following generative process defines the model:

- For each topic $z \in [1;T]$ in each modality $m \in [1;M]$
  - Draw $\phi_z(m) \sim Dirichlet(\beta(m))$.
  - This is the parameters of the $z^{th}$ topic's distribution over vocabulary $[1;V(m)]$ of modality $m$.

- For each song $s \in [1;S]$
  - Draw $\theta_s \sim Dirichlet(\alpha)$.
  - This is the parameters of the $s^{th}$ song's distribution over topics $[1;T]$.
  - For each modality $m \in [1;M]$
    - For each word $w \in [1;N_{xm}]$
      - Draw a specific topic $z^{(m)}_w \sim Categorical(\theta_s)$
      - Draw a word $w^{(m)} \sim Categorical(\phi_z^{(m)})$

¹http://the.echonest.com
A main characteristic of mNLDA is the common topic proportions for all $M$ modalities in each song, $s$, and separate word-topic distributions $p(w_i|m,z)$ for each modality, where $z$ denotes a particular topic. Thus, each modality has its own definition of what a topic is in terms of its own vocabulary.

Model inference is performed using a collapsed Gibbs sampler [22] similar to the standard LDA. The Gibbs sampler is run for a limited number of complete sweeps through the training songs, and the model state with the highest model evidence within the last 50 iterations is regarded as the MAP estimate. From this MAP sample, point estimates of the topic-word, topic-slot and slot-topic expectations of the corresponding Dirichlet distributions can be computed based on the expectations of the corresponding Dirichlet distributions.

Evaluation of model performance on a unknown test song, $s^*$, is performed using the procedure of fold-in [23, 24] by computing the point estimate of the topic-sentence distribution, $\hat{p}(z|s)$, and the modality, $m$, specific word-topic distribution of the held out modality, $m^*$, of the training data. This is obtained by fixing the song-topic counts while updating the word-topic counts for that specific modality. This is similar to the fold-in procedure used for test songs.

4. EXPERIMENTAL RESULTS & DISCUSSION

4.1. Alignment

The first aim is to evaluate the latent representation’s alignment with a human ‘cognitive’ variable, which we previously argued could be the open vocabulary tags. We do this by including only the lower level modalities of audio and lyrics when estimating the model. Then the normalized mutual information between a single tag and the latent representations, i.e., the topics, is calculated for all the tags.

Thus for a single tag, $w_i^{(tag)}$ we can compute the mutual information between the tag and the topic distribution for a specific song, $s$ as:

$$\text{MI}(w_i, z | s) =$$

$$\frac{\text{KL}(\hat{p}(w_i^{(tag)} | z | s) \| \hat{p}(w_i^{(tag)} | s) \hat{p}(z | s))}{\text{H}(w_i^{(tag)} | s) + \text{H}(z | s)},$$

where $\text{KL}()$ denotes the Kullback-Leibler divergence. We normalize the MI to be in $[0; 1]$, i.e.,

$$\text{NMI}(w_i^{(tag)}, z | s) = \frac{\text{MI}(w_i^{(tag)}, z | s)}{\frac{1}{2} \sum \text{NMI}(w_i^{(tag)}, z | s)},$$

where $\text{H}(\cdot)$ denotes the entropy. Finally, we compute the average over all songs to arrive at the final measure of alignment for a specific tag, given by $\text{avgNMI}(w_i^{(tag)}) = \frac{1}{T} \sum_{t=1}^{T} \text{NMI}(w_i^{(tag)}, z | s)$.

Fig. 2 shows a sorted list of tags, where tags with high alignment with the latent representation have higher average NMI (avgNMI). It is notable that the combination of the audio and lyrics modality, in defining the latent representation, seems to align well with genre-like and style-like tags. On the contrary, emotional and period tags are relatively less aligned with the representation. Also note that the alignment is not simply a matter of the tag being the most popular as can be seen from Fig. 2. Less popular tags are ranked higher by avgNMI than very popular tags, suggesting that some are more specialized in terms of the latent representation than others.

The result gives merit to the idea of using genre and styles as proxy for evaluating latent representation in comparison with other open vocabulary tags, since we - from lower level features, such as audio features and lyrics - can find latent representations which align well with high-level, ’cognitive’ aspects in an unsupervised way. This is in line with many studies in music informatics on western music (see e.g. [25, 26, 27]) which indicate coherence between genre and tag categories and cognitive understanding of music structure.
In this paper, we proposed the multi-way LDA as a flexible model for analyzing and modeling multi-modal and heterogeneous music data in a large scale setting. Based on the analysis of tags and latent representation, we provided evidence for the common assumption that genre may be an acceptable proxy for cognitive categorization of (western) music. Finally, we demonstrated and analyzed the predictive performance of the generative model providing benchmark result for the Million Song Dataset, and a genre dependent performance of the generative model showing some evidence in favor of such coherence.

4.2. Prediction

Given the evidence presented for genre and style being the relatively most appropriate human categories, our second aim is to evaluate the predictive performance of the multi-modal model for genre and style, and we turn to the recently published extension of the MSD [9] for reference test/train splits and genre and style labels. In particular, we use the balanced splits defined in [9].

For the genre case, this results in 2000 labeled examples per genre and 15 genres, thus resulting in 30,000 songs. We estimate the predictive genre performance by 10-fold cross-validation. Fig. 4 shows the per-label classification accuracy (perfect classification equals 1). The total genre classification performance is illustrated in Fig. 3a. The corresponding result for style classification, based on a total of 50,000 labeled examples, is shown in Fig. 3b. Both results are generated using \( T = 128 \) topics, 2000 Gibbs sweeps and predicting using the MAP estimate from the Gibbs sampler.

We first note that the combination of all modalities performs the best and significantly better than random as seen from Fig. 3, which is encouraging, and support the multi-modal approach. It is furthermore noted that the tag modality is able to perform very well. This indicates that despite the possibly noisy user expressed view, the model is able to find structure in line with the taxonomy defined in the reference labels of [9]. More interesting is perhaps the audio and lyric modalities and the combination of the two. This shows that lyrics performs the worst for genre, possibly due to the missing data in some tracks, while the combination is significantly better. For style there is no significant difference between audio and lyrics.

Looking at the genre specific performance in Fig. 4 we find a significant difference between the modalities. It appears that the importance of the modalities is partly in line with the fundamentally different characteristics of each specific genre. For example ‘latino’ driven by very characteristic lyrics. Further insight can be obtained by considering the confusion matrices which show some systematic pattern of error in the individual modalities, whereas the combined model shows a distinct diagonal structure, highlighting the benefits of multi-modal integration.

5. CONCLUSION
6. REFERENCES


