MULTI-LABEL MUSIC GENRE CLASSIFICATION FROM AUDIO, TEXT, AND IMAGES USING DEEP FEATURES

Sergio Oramas¹, Oriol Nieto², Francesco Barbieri³, Xavier Serra¹
¹Music Technology Group, Universitat Pompeu Fabra
²Pandora Media, Inc.
³TALN Group, Universitat Pompeu Fabra
{sergio.oramas, francesco.barbieri, xavier.serra}@upf.edu, onieto@pandora.com

ABSTRACT

Music genres allow to categorize musical items that share common characteristics. Although these categories are not mutually exclusive, most related research is traditionally focused on classifying tracks into a single class. Furthermore, these categories (e.g., Pop, Rock) tend to be too broad for certain applications. In this work we aim to expand this task by categorizing musical items into multiple and fine-grained labels, using three different data modalities: audio, text, and images. To this end we present MuMu, a new dataset of more than 31k albums classified into 250 genre classes. For every album we have collected the cover image, text reviews, and audio tracks. Additionally, we propose an approach for multi-label genre classification based on the combination of feature embeddings learned with state-of-the-art deep learning methodologies. Experiments show major differences between modalities, which not only introduce new baselines for multi-label genre classification, but also suggest that combining them yields improved results.

1. INTRODUCTION

Music genres are useful labels to classify musical items into broader categories that share similar musical, regional, or temporal characteristics. Dealing with large collections of music poses numerous challenges when retrieving and classifying information [3]. Music streaming services tend to offer catalogs of tens of millions of tracks, for which tasks such as music classification are of utmost importance. Music genre classification is a widely studied problem in the Music Information Research (MIR) community [40]. However, almost all related work is concentrated in multi-class classification of music items into broad genres (e.g., Pop, Rock), assigning a single label per item. This is problematic since there may be hundreds of more specific music genres [33], and these may not be necessarily mutually exclusive (i.e., a song could be Pop, and at the same time have elements from Deep House and a Reggae groove). In this work we aim to advance the field of music classification by framing it as multi-label genre classification of fine-grained genres.

To this end, we present MuMu, a new large-scale multimodal dataset for multi-label music genre classification. MuMu contains information of roughly 31k albums classified into one or more 250 genre classes. For every album we analyze the cover image, text reviews, and audio tracks, with a total number of approximately 147k audio tracks and 447k album reviews. Furthermore, we exploit this dataset with a novel deep learning approach to learn multiple genre labels for every album using different data modalities (i.e., audio, text, and image). In addition, we combine these modalities to study how the different combinations behave.

Results show how feature learning using deep neural networks substantially surpasses traditional approaches based on handcrafted features, reducing the gap between text-based and audio-based classification [29]. Moreover, an extensive comparative of different deep learning architectures for audio classification is provided, including the usage of a dimensionality reduction approach that yields improved results. Finally, we show how the late fusion of feature vectors learned from different modalities achieves better scores than each of them individually.

2. RELATED WORK

Most published music genre classification approaches rely on audio sources [2, 40]. Traditional techniques typically use handcrafted audio features, such as Mel Frequency Cepstral Coefficient (MFCCs) [20], as input of a machine learning classifier (e.g., SVM) [39, 44]. More recent deep learning approaches take advantage of visual representations of the audio signal in form of spectrograms. These visual representations are used as input to Convolutional Neural Networks (CNNs) [5, 6, 8, 9, 34], following approaches similar to those used for image classification.

Text-based approaches have also been explored for this task. For instance, in [13, 29] album customer reviews are used as input for the classification, whereas in [4, 22] song lyrics are employed. By contrast, there are a limited number of papers dealing with image-based genre classi-
3. MULTIMODAL DATASET

To the best of our knowledge, there are no publicly available large-scale datasets that encompass audio, images, text, and multi-label annotations. Therefore, we present MuMu, a new Multimodal Music dataset with multi-label genre annotations that combines information from the Amazon Reviews dataset [23] and the Million Song Dataset (MSD) [1]. The former contains millions of album customer reviews and album metadata gathered from Amazon.com. The latter is a collection of metadata and precomputed audio features for a million songs.

To map the information from both datasets we use MusicBrainz 1. For every album in the Amazon dataset, we query MusicBrainz with the album title and artist name to find the best possible match. Matching is performed using the same methodology described in [30], following a pairwise entity resolution approach based on string similarity. Following this approach, we were able to map 60% of the Amazon dataset. For all the matched albums, we obtain the MusicBrainz recording ids of their songs. With these, we use an available mapping from MSD to MusicBrainz 2 to obtain the subset of recordings present in the MSD. From the mapped recordings, we only keep those associated with a unique album. This process yields the final set of 147,295 songs, which belong to 31,471 albums.

The song features provided by the MSD are not generally suitable for deep learning [45], so we instead use in our experiments audio previews between 15 and 30 seconds retrieved from 7digital.com. For the mapped set of albums, there are 447,583 customer reviews in the Amazon Dataset. In addition, the Amazon Dataset provides further information about each album, such as genre annotations, average rating, selling rank, similar products, cover image url, etc. We employ the provided image url to gather the cover art of all selected albums. For instance, an album annotated as Traditional Pop comes with the complete branch information Pop / Oldies / Traditional Pop. To exploit either the taxonomic and the co-occurrence information, we provide every item with the labels of all their branches. For example, an album classified as Jazz / Vocal Jazz and Pop / Vocal Pop is annotated in MuMu with the four labels: Jazz, Vocal Jazz, Pop, and Vocal Pop. There are in average 5.97 labels for each song (3.13 standard deviation).

3.1 Genre Labels

Amazon has its own hierarchical taxonomy of music genres, which is up to four levels in depth. In the first level there are 27 genres, and almost 500 genres overall. In our dataset, we keep the 250 genres that satisfy the condition of having been annotated in at least 12 albums. Every album in Amazon is annotated with one or more genres from different levels of the taxonomy. The Amazon Dataset contains complete information about the specific branch from the taxonomy used to classify each album. For instance, an album annotated as Traditional Pop comes with the complete branch information Pop / Oldies / Traditional Pop. To exploit either the taxonomic and the co-occurrence information, we provide every item with the labels of all their branches. For example, an album classified as Jazz / Vocal Jazz and Pop / Vocal Pop is annotated in MuMu with the four labels: Jazz, Vocal Jazz, Pop, and Vocal Pop. There are in average 5.97 labels for each song (3.13 standard deviation).

<p>| Table 1. Top-10 most and least represented genres |
|----------------|----------------|----------------|</p>
<table>
<thead>
<tr>
<th>Genre</th>
<th>% of albums</th>
<th>Genre</th>
<th>% of albums</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rock</td>
<td>55.29</td>
<td>Harmonica Blues</td>
<td>0.10</td>
</tr>
<tr>
<td>Rock</td>
<td>55.29</td>
<td>Harmonica Blues</td>
<td>0.10</td>
</tr>
<tr>
<td>Alternative Rock</td>
<td>27.69</td>
<td>Concertos</td>
<td>0.10</td>
</tr>
<tr>
<td>World Music</td>
<td>19.31</td>
<td>Bass</td>
<td>0.06</td>
</tr>
<tr>
<td>Jazz</td>
<td>14.73</td>
<td>European Jazz</td>
<td>0.06</td>
</tr>
<tr>
<td>Dance &amp; Electronic</td>
<td>12.23</td>
<td>Piano Blues</td>
<td>0.06</td>
</tr>
<tr>
<td>Metal</td>
<td>11.50</td>
<td>Norway</td>
<td>0.06</td>
</tr>
<tr>
<td>Indie &amp; Lo-Fi</td>
<td>10.45</td>
<td>Slide Guitar</td>
<td>0.06</td>
</tr>
<tr>
<td>R&amp;B</td>
<td>10.10</td>
<td>East Coast Blues</td>
<td>0.06</td>
</tr>
<tr>
<td>Folk</td>
<td>9.69</td>
<td>Girl Groups</td>
<td>0.06</td>
</tr>
<tr>
<td>Pop</td>
<td>84.38</td>
<td>Tributes</td>
<td>0.10</td>
</tr>
</tbody>
</table>

The labels in the dataset are highly unbalanced, following a distribution which might align well with those found in real world scenarios. In Table 1 we see the top 10 most and least represented genres and the percentage of albums annotated with each label. The unbalanced character of the genre annotations poses an interesting challenge for music classification that we also aim to exploit. Among the multiple possibilities that this dataset may offer to the MIR community, we focus our work on the multi-label classification problem, described next.

4. MULTI-LABEL CLASSIFICATION

In multi-label classification, multiple target labels may be assigned to each classifiable instance. More formally: given a set of labels \( L = \{l_1, l_2, \ldots, l_n\} \), and a set of \( m \) items \( I = \{i_1, i_2, \ldots, i_m\} \), we aim to model a function \( f \) able to associate a set of \( c \) labels to every item in \( I \), where \( c \in [1, n] \) varies for every item.

Deep learning approaches are well-suited for this problem, as these architectures allow to have multiple outputs in their final layer. The usual architecture for large multi-label classification using deep learning ends with a logistic regression layer with sigmoid activations evaluated with the cross-entropy loss, where target labels are encoded as high-dimensional sparse binary vectors [42]. This method, which we refer as LOGISTIC, implies the assumption that

---

1 https://musicbrainz.org/
2 http://labs.acousticbrainz.org/million-song-dataset-echonest-archive
3 https://www.upf.edu/web/mtp/mumu
the classes are statistically independent (which is not the case in music genres).

A more recent approach [7], relies on matrix factorization to reduce the dimensionality of the target labels. This method makes use of the interrelation between labels, embedding the high-dimensional sparse labels onto lower-dimensional vectors. In this case, the target of the network is a dense lower-dimensional vector which can be learned using the cosine proximity loss, as these vectors tend to be \(l_2\)-normalized. We denote this technique as COSINE, and we provide a more formal definition next.

### 4.1 Labels Factorization

Let \( M \) be the binary matrix of items \( I \) and labels \( L \) where \( m_{ij} = 1 \) if \( i \) is annotated with label \( l_j \) and \( m_{ij} = 0 \) otherwise. Using \( M \), we calculate the matrix \( X \) of Positive Pointwise Mutual Information (PPMI) for the set of labels \( L \). Given \( L_i \) as the set of items annotated with label \( l_i \), the PPMI between two labels is defined as:

\[
X(l_i, l_j) = \max(0, \log \frac{P(L_i, L_j)}{P(L_i)P(L_j)}) \tag{1}
\]

where \( P(L_i, L_j) = |L_i \cap L_j|/|I| \) and \( P(L_i) = |L_i|/|I| \).

The PPMI matrix \( X \) is then factorized using Singular Value Decomposition (SVD) such that \( X \approx U \Sigma V \), where \( U \) and \( V \) are unitary matrices, and \( \Sigma \) is a diagonal matrix of singular values. Let \( \Sigma_d \) be the diagonal matrix formed from the top \( d \) singular values, and let \( U_d \) be the matrix produced by selecting the corresponding columns from \( U \). The matrix \( C_d = U_d \cdot \sqrt{\Sigma_d} \) contains the label factors of \( d \) dimensions. Finally, we obtain the matrix of item factors \( F_d \) as \( F_d = C_d \cdot M^T \). Further information on this technique may be found in [17].

Factors present in matrices \( C_d \) and \( F_d \) are embedded in the same space. Thus, a distance metric such as cosine distance can be used to obtain distance measures between items and labels. Similar labels are grouped in the space, and at the same time, items with similar sets of labels are near each other. These properties can be exploited in the label prediction problem.

### 4.2 Evaluation Metrics

The evaluation of multi-label classification is not necessarily straightforward. Evaluation measures vary according to the output of the system. In this work we are interested in measures that deal with probabilistic outputs, instead of binary. The Receiver Operating Characteristic (ROC) curve is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. Thus, the area under the ROC curve (AUC) is often taken as an evaluation measure to compare such systems. We selected this metric to compare the performance of the different approaches as it has been widely used for genre and tag classification problems [5, 9].

The output of a multi-label classifier is a label-item matrix. Thus, it can be evaluated either from the labels or the items perspective. We can measure how accurate the classification is for every label, or how well the labels are ranked for every item. In this work, the former point of view is evaluated with the AUC measure, which is computed for every label and then averaged. We are interested in classification models that strengthen the diversity of label assignments. As the taxonomy is composed of broad genres which are over-represented in the dataset (see Table 1), and more specific subgenres (e.g., Vocal Jazz, Brit-pop), we want to measure whether the classifier is focusing only on over-represented genres, or on more fine-grained ones. To this end, catalog coverage (also known as aggregated diversity) is an evaluation measure used in the extreme multi-label classification [14] and the recommender systems [32] communities. Coverage@k measures the percentage of normalized unique labels present in the top \( k \) predictions made by an algorithm across all test items. Values of \( k = 1, 3, 5 \) are typically employed in multi-label classification.

---

### 5. ALBUM GENRE CLASSIFICATION

In this section we exploit the multimodal nature of the MuMu dataset to address the multi-label classification task. More specifically, and since each modality on this set (i.e., cover image, text reviews, and audio tracks) is associated with a music album, our task focuses on album classification.

#### 5.1 Audio-based Approach

A music album is composed by a series of audio tracks, each of which may be associated with different genres. In order to learn the album genre from a set of audio tracks we split the problem into three steps: (1) track feature vectors are learned while trying to predict the genre labels of the album from every track in a deep neural network. (2) Track vectors of each album are averaged to obtain album feature vectors. (3) Album genres are predicted from the album feature vectors in a shallow network where the input layer is directly connected to the output layer.

It is common in MIR to make use of CNNs to learn higher-level features from spectrograms. These representations are typically contained in \( \mathbb{R}^{F \times N} \) matrices with \( F \) frequency bins and \( N \) time frames. In this work we compute 96 frequency bin, log-compressed constant-Q transforms (CQT) [38] for all the tracks in our dataset using librosa [24] with the following parameters: audio sampling rate at 22050 Hz, hop length of 1024 samples, Hann analysis window, and 12 bins per octave. In addition, log-amplitude scaling is applied to the CQT spectrograms. Following a similar approach to [45], we address the variability of the length \( N \) across songs by sampling one 15-seconds long "patch" from each track, resulting in the fixed-size input to the CNN.

To learn the genre labels we design a CNN with four convolutional layers and experiment with different number of filters, filter sizes, and output configurations (see Section 6.1).
5.2 Text-based Approach

In the presented dataset, each album has a variable number of customer reviews. We use an approach similar to [13, 29] for genre classification from text, where all reviews from the same album are aggregated into a single text. The aggregated result is truncated at 1000 characters, thus balancing the amount of text per album, as more popular artists tend to have a higher number of reviews. Then we apply a Vector Space Model approach (VSM) with tf-idf weighting [47] to create a feature vector for each album. Although word embeddings [25] with CNNs are state-of-the-art in many text classification tasks [15], a traditional VSM approach is used instead, as it seems to perform better when dealing with large texts [31]. The vocabulary size is limited to 10k as it was a good balance of network complexity and accuracy.

Furthermore, a second approach is proposed based on the addition of semantic information, similarly to the method described in [29]. To semantically enrich the album texts, we adopted Babelfy, a state-of-the-art tool for entity linking [26], a task to associate, for a given textual fragment candidate, the most suitable entry in a reference KB. Babelfy maps words from a given text to Wikipedia [4]. In Wikipedia, categories are used to organize resources. We take all the Wikipedia categories of entities identified by Babelfy in each document and add them at the end of the text as new words. Then a VSM with tf-idf weighting is applied to the semantically enriched texts, where the vocabulary is also limited to 10k terms. Note that either words or categories may be part of this vocabulary.

From this representation, a feed forward network with two dense layers of 2048 neurons and a Rectified Linear Unit (ReLU) after each layer is trained to predict the genre label in both LOGISTIC and COSINE configurations.

5.3 Image-based Approach

Every album in the dataset has an associated cover art image. To perform music genre classification from these images, we use Deep Residual Networks (ResNets) [11]. They are the state-of-the-art in various image classification tasks [19], a traditional VSM approach is used instead, as it seems to perform better when dealing with large texts [31]. The vocabulary size is limited to 10k as it was a good balance of network complexity and accuracy.

The network we use is composed of 101 layers (ResNet-101), initialized with pretrained parameters learned on ImageNet. The output layer may have either a softmax or a sigmoid activation function. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

Therefore, from each modality network described above, we separately obtain an internal feature representation for every album after training them on the genre classification task. Concretely, the input to the last fully connected layer of each network becomes feature vector for its respective modality. Given a set of feature vectors, $l2$-regularization is applied on each of them. They are then concatenated into a single feature vector, which becomes the input to a simple Multi Layer Perceptron (MLP), where the input layer is directly connected to the output layer. The output layer may have either a LOGISTIC or a COSINE configuration.

6. EXPERIMENTS

We apply the architectures defined in the previous section to the MuMu dataset. The dataset is divided as follows: 80% for training, 10% for validation, and 10% for test. We first evaluate every modality in isolation in the multi-label genre classification task. Then, from each modality, a deep feature vector is obtained for the best performing configuration. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

Therefore, from each modality network described above, we separately obtain an internal feature representation for every album after training them on the genre classification task. Concretely, the input to the last fully connected layer of each network becomes feature vector for its respective modality. Given a set of feature vectors, $l2$-regularization is applied on each of them. They are then concatenated into a single feature vector, which becomes the input to a simple Multi Layer Perceptron (MLP), where the input layer is directly connected to the output layer. The output layer may have either a LOGISTIC or a COSINE configuration.

6.1 Audio Classification

We explore three network design parameters: convolution filter size, number of filters per convolutional layer, and validation sets is factorized using the approach described in Section 4.1, with a value of $k = 1, 3, 5$. The training speed per epoch and number of network hyperparameters are also reported. All source code and data splits used in our experiments are available on-line.

The matrix of album genre annotations of the training and validation sets is factorized using the approach described in Section 4.1, with a value of $d = 50$ dimensions. From the set of album factors, those annotated with a single label from the top level of the taxonomy are plotted in Figure 1 using t-SNE dimensionality reduction [21]. It can be seen how the different albums are properly clustered in the factor space according to their genre.

6.2 Multimodal approach

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.

We aim to combine all of these different types of data into a single model. There are several works claiming that learning data representations from different modalities simultaneously outperforms systems that learn them separately [10, 28]. However, recent work in multimodal learning with audio and text in the context of music recommendation [31] reflects the contrary. We have observed that deep networks are able to find an optimal minimum very fast from text data. However, the complexity of the audio signal can significantly slow down the training process. Simultaneous learning may under-explore one of the modalities, as the stronger modality may dominate quickly. Thus, learning each modality separately warrants that the variability of the input data is fully represented in each of the feature vectors.
applied with two different settings: SQUARE 3x3 filters as in [5], a filter of 4x96 that convolves only in time [45], and a musically motivated filter of 4x70, which is able to slightly convolve in the frequency domain [34]. To study the width of the convolutional layer, we try with three approaches: 5M 260s 0.862 0.12 0.33 0.48
AUDIO LOGISTIC HIGH-4x96 0.2M 140s 0.851 0.14 0.32 0.48
 AUDIO LOGISTIC LOW-4x96 5M 260s 0.862 0.12 0.33 0.48
 AUDIO LOGISTIC LOW-4x70 0.35M 200s 0.871 0.05 0.16 0.34
 AUDIO LOGISTIC HIGH-4x70 7.5M 600s 0.849 0.08 0.23 0.38
 AUDIO COSINE LOW-3x3 0.33M 400s 0.864 0.26 0.47 0.65
 AUDIO COSINE HIGH-3x3 15.5M 2200s 0.881 0.30 0.54 0.69
 AUDIO COSINE LOW-4x96 0.15M 155s 0.860 0.19 0.40 0.52
 AUDIO COSINE HIGH-4x96 4M 250s 0.884 0.35 0.59 0.75
 AUDIO COSINE LOW-4x70 0.3M 190s 0.868 0.26 0.51 0.68
 AUDIO (A) COSINE HIGH-4x70 6.5M 590s 0.888 0.35 0.60 0.74

Table 2. Results for Multi-label Music Genre Classification of Albums

<table>
<thead>
<tr>
<th>Modality</th>
<th>Target</th>
<th>Settings</th>
<th>Params</th>
<th>Time</th>
<th>AUC</th>
<th>C@1</th>
<th>C@3</th>
<th>C@5</th>
</tr>
</thead>
<tbody>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-3x3</td>
<td>0.01M</td>
<td>1s</td>
<td>0.792</td>
<td>0.04</td>
<td>0.14</td>
<td>0.22</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-4x96</td>
<td>0.5M</td>
<td>390s</td>
<td>0.849</td>
<td>0.14</td>
<td>0.34</td>
<td>0.48</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-3x3</td>
<td>16.5M</td>
<td>2280s</td>
<td>0.640</td>
<td>0.20</td>
<td>0.43</td>
<td>0.69</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-4x96</td>
<td>0.2M</td>
<td>140s</td>
<td>0.851</td>
<td>0.14</td>
<td>0.32</td>
<td>0.48</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-4x96</td>
<td>5M</td>
<td>260s</td>
<td>0.862</td>
<td>0.12</td>
<td>0.33</td>
<td>0.48</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-4x70</td>
<td>0.35M</td>
<td>200s</td>
<td>0.871</td>
<td>0.05</td>
<td>0.16</td>
<td>0.34</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-4x70</td>
<td>7.5M</td>
<td>600s</td>
<td>0.849</td>
<td>0.08</td>
<td>0.23</td>
<td>0.38</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-3x3</td>
<td>0.33M</td>
<td>400s</td>
<td>0.864</td>
<td>0.26</td>
<td>0.47</td>
<td>0.65</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-3x3</td>
<td>15.5M</td>
<td>2200s</td>
<td>0.881</td>
<td>0.30</td>
<td>0.54</td>
<td>0.69</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-4x96</td>
<td>0.15M</td>
<td>155s</td>
<td>0.860</td>
<td>0.19</td>
<td>0.40</td>
<td>0.52</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-4x96</td>
<td>4M</td>
<td>250s</td>
<td>0.884</td>
<td>0.35</td>
<td>0.59</td>
<td>0.75</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>LOW-4x70</td>
<td>0.3M</td>
<td>190s</td>
<td>0.868</td>
<td>0.26</td>
<td>0.51</td>
<td>0.68</td>
</tr>
<tr>
<td>LOW</td>
<td>TIMBRE MLP</td>
<td>HIGH-4x70</td>
<td>6.5M</td>
<td>590s</td>
<td>0.888</td>
<td>0.35</td>
<td>0.60</td>
<td>0.74</td>
</tr>
</tbody>
</table>

Figure 1. t-SNE of album factors.

and target layer. For the filter size we compare three approaches: square 3x3 filters as in [5], a filter of 4x96 that convolves only in time [45], and a musically motivated filter of 4x70, which is able to slightly convolve in the frequency domain [34]. To study the width of the convolutional layers we try with two different settings: HIGH with 256, 512, 1024, and 1024 in each layer respectively, and LOW with 64, 128, 128, 64 filters. Max-pooling is applied after each convolutional layer. Finally, we use the two different network targets defined in Section 4, LOGISTIC and COSINE. We empirically observed that dropout regularization only helps in the HIGH plus COSINE configurations. Therefore we applied dropout with a factor of 0.5 to these configurations, and no dropout to the others.

Apart from these configurations, a baseline approach is added. This approach consists in a traditional audio-based approach for genre classification based on the audio descriptors present in the MSD [1]. More specifically, for each song we aggregate four different statistics of the 12 timbre coefficient matrices: mean, max, variance, and L2-norm. The obtained 48 dimensional feature vectors are fed into a feed forward network as the one described in Section 5.4 with LOGISTIC output. This approach is denoted as TIMBRE-MLP.

The results show that CNNs applied over audio spectrograms clearly outperform traditional approaches based on handcrafted features. We observe that the TIMBRE-MLP approach achieves 0.792 of AUC, contrasting with the 0.888 from the best CNN approach. We note that the LOGISTIC configuration obtains better results when using a lower number of filters per convolution (LOW). Configurations with fewer filters have less parameters to optimize, and their training processes are faster. On the other hand, in COSINE configurations we observe that the use of a higher number of filters tends to achieve better performance. It seems that the fine-grained regression of the factors benefits from wider convolutions. Moreover, we observe that 3x3 square filter settings have lower performance, need more time to train, and have a higher number of parameters to optimize. By contrast, networks using time convolutions only (4x96) have a lower number of parameters, are faster to train, and achieve comparable performance. Furthermore, networks that slightly convolve across the frequency bins (4x70) achieve better results with only a slightly higher number of parameters and training time. Finally, we observe that the COSINE regression approach achieves better AUC scores in most configurations, and also their results are more diverse in terms of catalog coverage.

Figure 1. t-SNE of album factors.
ever, it is remarkable that the texts have important information gain values. How-
different genres. We observed that genre labels present in
are overall slightly superior to the audio-based ones.
greatly in terms of catalog coverage. The text-based results
**LOGISTIC** for Rock, **COSINE** for Pop, and **dope** for Rap albums. Place names have
also important weights, as **Jamaica** for Reggae, **Nashville** for Country, or **Chicago** for Blues.

### 6.3 Image Classification

Results show that genre classification from images has lower performance in terms of AUC and catalog coverage compared to the other modalities. Due to the use of an already pre-trained network with a logistic output (ImageNet [35]) as initialization of the network, it is not straightforward to apply the **COSINE** configuration. Therefore, we only report results for the **LOGISTIC** configuration.

In Figure 2 a set of cover images of five of the most frequent genres in the dataset is shown using t-SNE over the obtained image feature vectors. In the left top corner the ResNet recognizes women faces on the foreground, which seems to be common in Country albums (red). The jazz albums (green) on the right are all clustered together probably thanks to the uniform type of clothing worn by the people of their covers. Therefore, the visual style of the cover seems to be informative when recognizing the album genre. For instance, many classical music albums include an instrument in the cover, and Dance & Electronics covers are often abstract images with bright colors, rarely including human faces.

### 6.4 Multimodal Classification

From the best performing approaches in terms of AUC of each modality (i.e., **AUDIO /cosine** / **HIGH-4X70**, **TEXT /cosine** / **VSM+SEM and IMAGE /LOGISTIC / **RESNET**), a feature vector is obtained as described in Section 5.4. Then, these three feature vectors are aggregated in all possible combinations, and genre labels are predicted using the MLP network described in Section 5.4. Both output configurations **LOGISTIC** and **COSINE** are used in the learning phase, and dropout of 0.7 is applied in the **COSINE** configuration.

Results suggest that the combination of modalities outperforms single modality approaches. As image features are learned using a **LOGISTIC** configuration, they seem to improve multimodal approaches with **LOGISTIC** configuration only. Multimodal approaches that include text features tend to improve the results. Nevertheless, the best approaches are those that exploit the three modalities of MuMu. **COSINE** approaches have similar AUC than **LOGISTIC** approaches but a much better catalog coverage, thanks to the spatial properties of the factor space.

### 7. CONCLUSIONS

An approach for multi-label music genre classification using deep learning architectures has been proposed. The approach was applied to audio, text, image data, and their combination. For its assessment, MuMu, a new multimodal music dataset with over 31k albums and 135k songs has been gathered. We showed how representation learning approaches for audio classification outperform traditional handcrafted feature based approaches. Moreover, we compared the effect of different design parameters of CNNs in audio classification. Text-based approaches seem to outperform other modalities, and benefit from the semantic enrichment of texts via entity linking. While the image-based classification yielded the lowest performance, it helped to improve the results when combined with other modalities. Multimodal approaches appear to outperform single modality approaches, and the aggregation of the three modalities achieved the best results. Furthermore, the dimensionality reduction of target labels led to better results, not only in terms of accuracy, but also in terms of catalog coverage.

This paper is an initial attempt to study the multi-label classification problem of music genres from different perspectives and using different data modalities. In addition, the release of the MuMu dataset opens up a number of unexplored research possibilities. In the near future we aim to modify the ResNet to be able to learn latent factors from images as we did in other modalities and apply the same multimodal approach to other MIR tasks.
8. ACKNOWLEDGMENTS

This work was partially funded by the Spanish Ministry of Economy and Competitiveness under the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The Tesla K40 used for this research was donated by the NVIDIA Corporation.

9. REFERENCES


