Transactions of the International Society for Music Information Retrieval

Oramas, S., Barbieri, F., Nieto, O., and Serra, X (2018). Multimodal Deep Learning for Music Genre Classification, *Transactions of the International Society for Music Information Retrieval*, V(N), pp. xx–xx, DOI: https://doi.org/xx.xxxx/xxxx.xx

ARTICLE TYPE

Multimodal Deep Learning for Music Genre Classification

Sergio Oramas[*], Francesco Barbieri[†], Oriol Nieto[‡], and Xavier Serra[*]

Abstract

Music genre labels are useful to organize songs, albums, and artists into broader groups that share similar musical characteristics. In this work, an approach to learn and combine multimodal data representations for music genre classification is proposed. Intermediate representations of deep neural networks are learned from audio tracks, text reviews, and cover art images, and further combined for classification. Experiments on single and multi-label genre classification are then carried out, evaluating the effect of the different learned representations and their combinations. Results on both experiments show how the aggregation of learned representations from different modalities improves the accuracy of the classification, suggesting that different modalities embed complementary information. In addition, the learning of a multimodal feature space increase the performance of pure audio representations, which may be specially relevant when the other modalities are available for training, but not at prediction time. Moreover, a proposed approach for dimensionality reduction of target labels yields major improvements in multi-label classification not only in terms of accuracy, but also in terms of the diversity of the predicted genres, which implies a more fine-grained categorization. Finally, a qualitative analysis of the results sheds some light on the behavior of the different modalities in the classification task.

Keywords: information retrieval, deep learning, music, multimodal, multi-label classification

1. Introduction

The advent of large music collections has posed the challenge of how to retrieve, browse, and recommend their containing items. One way to ease the access of large music collections is to keep tag annotations of all music resources (Sordo, 2012). Annotations can be added either manually or automatically. However, due to the high human effort required for manual annotations, the implementation of automatic annotation processes is more cost-effective.

Music genre labels are useful categories to organize and classify songs, albums, and artists into broader groups that share similar musical characteristics. Music genres have been widely used for music classification, from physical music stores to streaming services. Automatic music genre classification thus is a widely explored topic (Sturm, 2012; Bogdanov et al., 2016). However, almost all related work is concentrated in the classification of music items into broad genres (e.g., Pop, Rock) using handcrafted audio features and assigning a single label per item (Sturm, 2012). This is problematic for several reasons. First, there may be hundreds of more specific music genres (Pachet and Cazaly, 2000), and these may not necessarily be mutually exclusive (e.g., a song could be Pop, and at the same time have elements from Deep House and a Reggae groove). Second, handcrafted features may not fully represent the variability of the data. By contrast, representation learning approaches have demonstrated their superiority in multiple domains (Bengio et al., 2013). Third, large music collections contain different modalities of information, i.e., audio, images, and texts, and all these data are suitable to be exploited for genre classification. Several approaches dealing with different modalities have been proposed (Wu et al., 2016; Schedl et al., 2013). However, to the best of our knowledge, no multimodal approach based on deep learning architectures has been proposed for this Music Information Retrieval (MIR) task, neither for single-label nor multi-label classification.

In this work, we aim to fill this gap by proposing a system able to predict music genre labels using deep learning architectures given different data modalities. Our approach is divided into two steps: (1) A neural network is trained on the classification task for

[*]Music Tenchnology Group, Universitat Pompeu Fabra, Barcelona, Spain

[†]TALN Group, Universitat Pompeu Fabra, Barcelona, Spain

[‡]Pandora Media Inc., 94612 Oakland, USA

each modality. (2) Intermediate representations are extracted from each network and combined in a multimodal approach. Experiments on single-label and multi-label genre classification are then carried out, evaluating the effect of the learned data representations and their combination.

Audio representations are learned from time-frequency representations of the audio signal in form of audio spectrograms using Convolutional Neural Networks (CNNs). Visual representations are learned using a state-of-the-art CNN (ResNet) (He et al., 2016), initialized with pretrained parameters learned in a general image classification task (Russakovsky et al., 2015), and fine-tuned on the classification of music genre labels from the album cover images. Text representations are learned from music related texts (e.g., album reviews) using a feedforward network over a Vector Space Model (VSM) representation of texts, previously enriched with semantic information via entity linking (Oramas, 2017).

A first experiment on single-label classification is carried out from audio and images. In this experiment, in addition to the audio and visual learned representations, a multimodal feature space is learned by aligning both data representations. Results show that the fusion of audio and visual representations improves the performance of the classification over pure audio or visual approaches. In addition, the introduction of the multimodal feature space improves the quality of pure audio representations, even when no visual data are available in the prediction. Next, the performance of our learned models is compared with those of a human annotator, and a qualitative analysis of the classification results is reported. This analysis shows that audio and visual representations seem to complement each other. In addition, we study how the visual deep model focuses its attention on different regions of the input images when evaluating each genre.

These results are further expanded with an experiment on multi-label classification, which is carried out over audio, text, and images. Results from this experiment show again how the fusion of data representations learned from different modalities achieves better scores than each of them individually. In addition, we show that representation learning using deep neural networks substantially surpasses a traditional audio-based approach that employs handcrafted features. Moreover, an extensive comparison of different deep learning architectures for audio classification is provided, including the usage of a dimensionality reduction technique for labels that yields improved results. Then, a qualitative analysis of the multi-label classification experiment is finally reported.

This paper is an extended version of a previous contribution (Oramas et al., 2017a), with the main novel contributions being the addition of a single-label genre classification experiment where the differences among modalities are further explored, and a deeper qualitative analysis of the results is carried on. This paper is structured as follows. We review the related work in Section 2. In Section 3 we describe the representation learning approach from audio, images, and text with deep learning systems, and the multimodal joint space. Section 4 describes the fusion of multiple modalities into a single model and its potential benefits. Then, in Section 5 we describe the multi-label classification problem. In Section 6 the experiments on single-label classification are presented. Then, in Section 7 the experiments on multi-label classification are reported. In Section 8 we conclude our paper with a short summary of our findings.

2. Related work

Most published music genre classification approaches rely on audio sources (for an extensive review on the topic, please refer to Sturm (2012); Bogdanov et al. (2016)). Traditional techniques typically use handcrafted audio features, such as Mel Frequency Cepstral Coefficients (MFCCs) (Logan, 2000), as input to a machine learning classifier (e.g., SVM, k-NN) (Tzanetakis and Cook, 2002; Seyerlehner et al., 2010a; Gouyon et al., 2004). More recent deep learning approaches take advantage of visual representations of the audio signal in form of spectrograms. These visual representations of audio are used as input to Convolutional Neural Networks (CNNs) (Dieleman et al., 2011; Dieleman and Schrauwen, 2014; Pons et al., 2016; Choi et al., 2016a,b), following approaches similar to those used for image classification.

Text-based approaches have also been explored for this task. For instance, one of the earliest attempts on classification of music reviews is described in Hu et al. (2005), where experiments on multi-class genre classification and star rating prediction are described. Similarly Hu and Downie (2006) extend these experiments with a novel approach for predicting usages of music via agglomerative clustering, and conclude that bigram features are more informative than unigram features. Moreover, part-of-speech (POS) tags along with pattern mining techniques are applied in Downie and Hu (2006) to extract descriptive patterns for distinguishing negative from positive reviews. Additional textual evidence is leveraged in Choi et al. (2014), who consider lyrics as well as texts referring to the meaning of the song, and used for training a kNN classifier for predicting song subjects (e.g., love, war, or drugs). In Oramas et al. (2016a), album reviews are semantically enriched and classified among 13 genre classes using an SVM classifier.

There are few papers dealing with image-based music genre classification (Libeks and Turnbull, 2011). Regarding multimodal approaches found in the literature, most of them combine audio and song lyrics (Laurier et al., 2008; Neumayer and Rauber, 2007). Other modalities such as audio and video have been explored (Schindler and Rauber, 2015). In McKay and Fujinaga

(2008) cultural, symbolic, and audio features are combined for music classification.

Multi-label classification is a widely studied problem in other domains (Tsoumakas and Katakis, 2006; Jain et al., 2016). In the context of MIR, tag classification from audio (or auto-tagging) has been studied from a multi-label perspective using traditional machine learning approaches (Sordo, 2012; Wang et al., 2009; Turnbull et al., 2008; Bertin-Mahieux et al., 2008; Seyerlehner et al., 2010b), and more recently using deep learning approaches (Choi et al., 2016a; Dieleman and Schrauwen, 2014; Pons et al., 2017). However, there are few approaches for multi-label classification of music genres (Sanden and Zhang, 2011; Wang et al., 2009), and none of them is based on representation learning approaches nor multimodal data.

3. Learning data representations

3.1 Audio representations

The use of CNNs and audio spectrograms has become a standard in MIR (Dieleman et al., 2011; Choi et al., 2016a). Following this principle, we have designed a convolutional architecture to predict the genre labels from the audio spectrogram of a song. Spectrogram representations are typically contained in $\mathbb{R}^{\mathcal{F}\times N}$ matrices with $\mathcal{F}$ frequency bins and N time frames. In this work we compute $\mathcal{F} = 96$ frequency bins, log-compressed constant-Q transforms (CQT) (Schörkhuber and Klapuri, 2010) for all the tracks in our dataset using `librosa` (Mcfee et al., 2015) with the following parameters: audio sampling rate at 22050 Hz, hop length of 1024 samples, Hann analysis window, and 12 bins per octave. We randomly sampled one 15-seconds long *patch* from each track, resulting in the fixed-size input to the CNN. The deep model trained with these data is defined as follows: the CQT patches are fed to a series of convolutional layers with rectified linear units (ReLU) as activations followed by max pooling layers. The output of the last convolutional layer is flattened and connected to the ouptut layer. The activations of the last hidden layer constitute the intermediate audio representation used in our multimodal approach. More details on the architectures used and the training process are detailed in Sections 6.2 and 7.3.1.

3.2 Visual representations

Deep Residual Networks (ResNets) (He et al., 2016) are a specific type of CNNs that have become one of the best architectures for several image classification tasks (Russakovsky et al., 2015; Lin et al., 2014). ResNet is a feedforward CNN with *residual learning*, which consists on bypassing two or more convolution layers (similar to previous approaches (Sermanet and LeCun, 2011)). This addresses the underfitting problem originated when using a high number of layers, thus allowing for very deep architectures. We use the original ResNet[1] architecture, where the scaling and aspect ratio augmentation are obtained from Szegedy et al. (2015), the photometric distortions from Howard (2013), and weight decay is applied to all weights and biases (i.e., not focusing on convolutional layers only). Our network is composed of 101 layers (ResNet-101), initialized with pretrained parameters learned on ImageNet. This is our starting point to fine-tune (Razavian et al., 2014; Yosinski et al., 2014) the network on the genre classification task. More details about training process are reported in Sections 6.2 and 7.3.3. The activations of the last hidden layer of the ResNet become the visual representation used in our multimodal approach.

3.3 Text representations

Given a text describing a musical item (e.g., artist biography, album review), a process of semantic enrichment is firstly applied. To semantically enrich texts, we adopt Babelfy, a state-of-the-art tool for entity linking (Moro et al., 2014). Entity linking is the task to associate, for a given textual fragment candidate (e.g., an artist name, a place), the most suitable entry in a reference Knowledge Base. Babelfy maps words from a given text to BabelNet (Navigli and Ponzetto, 2012), returning the BabelNet URI of every identified entity. In addition to Babelfy, we use ELVIS (Oramas et al., 2016b), an entity linking integration framework, which retrieves the corresponding Wikipedia[2] URL and categories given a BabelNet URI. In Wikipedia, categories are used to organize resources, and they help users to group articles of the same topic. We take all the Wikipedia categories of entities identified in each document and add them at the end of the text as new words. We apply then a VSM with tf-idf weighting (Zobel and Moffat, 1998) over the enriched texts. Note that either words or categories may be part of the vocabulary in the VSM. 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 labels (the training process of this network is described in detail in Section 7.3.2). Dropout with a factor of 0.5 is applied after the input and each one of the dense layers. The last hidden layer becomes the text representation of each musical item.

Although word embeddings (Mikolov et al., 2013) with CNNs are state-of-the-art in many text processing tasks (Kim, 2014), a traditional VSM with a feed forward network is used instead, as it has been shown to perform better when dealing with large music-related texts and high dimensional outputs (Oramas et al., 2017b).

Finally, one may argue that if text representations are available, genre information is likely to be accessible as well, thus making the task of automatic genre

[1] `https://github.com/facebook/fb.resnet.torch/`
[2] `http://wikipedia.org`

classification redundant. While this might be true in some cases, genre information provided by external sources will unlikely comply with the current taxonomy of the collection to be classified, thus a mapping of sorts will be required in such instances. Moreover, it might also be unlikely to have multiple genre labels per release, making a stronger case to further employ as much data as possible (e.g., audio, visual, text) to further refine the potential genre(s) of each item in the catalog, regardless of whichever potential genres might be externally given (which can, too, become a new different modality).

3.4 Multimodal feature space

Given data representations from two different modalities, we design a neural model that learns to embed them in a new multimodal space that better optimizes their similarity. Using a deep learning approach to learn a multimodal space has been previously used, in particular for textual and visual modalities (Srivastava and Salakhutdinov, 2012; Yan and Mikolajczyk, 2015). Our model can be described as follows: let $\mathbf{a}$ and $\mathbf{v}$ be two representation vectors of a song obtained from different data modalities (e.g., audio and video), we embed them in a shared space. Formally:

$$\mathbf{e}_a(\mathbf{a}) = W_{a2} \tanh(W_{a1}\mathbf{a})$$

$$\mathbf{e}_v(\mathbf{v}) = W_{v2} \tanh(W_{v1}\mathbf{v})$$

where W_{xn} are weight matrices of the x modality (i.e., a or v) from the n-th layer and $\tanh$ is the element-wise hyperbolic tangent function, added as a non-linear component of the network. We then iterate over each song and learn the two modality embeddings by minimizing the loss defined by the cosine distance:

$$L^+ = 1 - \cos(\mathbf{e}_a(\mathbf{a}), \mathbf{e}_v(\mathbf{v}))$$

where $\cos(\cdot,\cdot)$ is the cosine similarity between two vectors.

Moreover, for each song we select two random negative samples (Mikolov et al., 2013): $\mathbf{r}_v$ and $\mathbf{r}_a$ from each modality. We want $\mathbf{a}$ and $\mathbf{v}$ to be distant from $\mathbf{r}_v$ and $\mathbf{r}_a$, respectively. This negative sampling avoids the problematic situation where the network maps all vectors to a single point (making $L^+ = 0$, but producing an useless mapping). We define the loss for the negative samples as:

$$L_a^- = \max(0, \cos(\mathbf{e}_a(\mathbf{r}_a), \mathbf{e}_v(\mathbf{v})) - m)$$

for one modality and, analogously, for the other modality part:

$$L_v^- = \max(0, \cos(\mathbf{e}_a(\mathbf{a}), \mathbf{e}_v(\mathbf{r}_v)) - m)$$

where m, the margin, is the scalar between 0 and 1 that indicates the importance of the negative samples (i.e., if 0, the negative sample is fully considered in the loss, whereas if 1, this sampling is ignored). We found that 0.5 was the best performing margin setting[3].

To summarize, given two different modality vectors of a song, the final loss that the multimodal network minimizes is:

$$L = L^+ + L_a^- + L_v^-$$

The resulting multimodal features from the networks $\mathbf{e}_a$ and $\mathbf{e}_v$ are composed of 200 dimensions each.

4. Multimodal fusion

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 (Ngiam et al., 2011; Dorfer et al., 2016). However, experiments in Oramas et al. (2017b) reflect the contrary. They have observed, for instance, that deep networks are able to quickly find an optimal minimum 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.

Therefore, from each modality network described above, we separately obtain an internal data representation for every item after training them on the genre classification task. Concretely, the activations of the last hidden layer of each network become the feature vector for its respective modality. Given a set of feature vectors, $l2$-norm is applied on each of them for normalization. They are then concatenated into a single feature vector, which becomes the input to a simple feedforward network, where the input layer is directly connected to the output layer. For single-label classification, softmax activation is finally applied, resulting in a multinomial logistic regression model. For multi-label classification, sigmoid activation is used instead.

5. Multi-label classification

In multi-label classification, multiple target labels may be assigned to each classifiable instance. Formally: given a set of n labels $G = \{g_1, g_2, \ldots, g_n\}$, and a set of d items $I = \{i_1, i_2, \ldots, i_d\}$, 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

[3] When training the multimodal system with the loss L, we tried ten different values of m: [0.1,0.2,...,0.8,0.9].

target labels are encoded as high-dimensional sparse binary vectors (Szegedy et al., 2016). This method, which we refer as LOGISTIC, implies the assumption that the classes are statistically independent (which is not the case in music genres).

A more recent approach (Chollet, 2016), relies on matrix factorization to reduce the dimensionality of the target labels, yielding a space where learning can be made more effectively. 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 $l2$-normalized. We denote this technique as COSINE, and we provide a more formal definition next.

5.1 Labels factorization

Let M be the binary matrix of items I and labels G where $m_{ij} = 1$ if i_i is annotated with label g_j and $m_{ij} = 0$ otherwise. Using M, we calculate the matrix X of Positive Pointwise Mutual Information (PPMI) for the set of labels G. Given G_i as the set of items annotated with label g_i, the PPMI between two labels is defined as:

$$X(g_i, g_j) = max\left(0, \log \frac{P(G_i, G_j)}{P(G_i)P(G_j)}\right) \quad (1)$$

where $P(G_i, G_j) = |G_i \cap G_j|/|I|$, $P(G_i) = |G_i|/|I|$, and $|.|$ denotes the cardinality function.

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 Σ is a diagonal matrix of singular values. Let Σ_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 Levy and Goldberg (2014).

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. Both labels and items with similar sets of labels are near each other in this space. These properties can be exploited in the label prediction problem.

6. Single-label classification experiment

In this section we describe the dataset and the experimental framework for single-label genre classification from audio and images (text modality will only be used in a second set of experiments in Section 7). More specifically, we set up an experiment for track genre classification using the different data modalities: only audio, only album cover artwork, and both. Lastly, we report and discuss the results of each experiment, compare them with human performance on the task, and perform a qualitative analysis of the results.

Genre	Train	Val	Test	%
Blues	518	120	190	2.68
Country	1351	243	194	5.78
Electronic	3434	725	733	15.81
Folk	858	164	136	3.74
Jazz	1844	373	462	8.66
Latin	390	83	83	1.80
Metal	1749	512	375	8.52
New Age	158	71	38	0.86
Pop	2333	644	466	11.13
Punk	487	132	96	2.31
Rap	1932	380	381	8.71
Reggae	1249	190	266	5.51
RnB	1223	222	396	5.95
Rock	3694	709	829	16.91
World	331	123	46	1.62

Table 1: Number of instances for each genre on the train, validation and test subsets. The percentage of elements for each genre is also shown.

6.1 MSD-I dataset

The Million Song Dataset (MSD, McFee et al., 2012) is a collection of metadata and precomputed audio features for 1 million songs. Along with this dataset, a dataset with annotations of 15 top-level genres with a single label per song was released (Schreiber, 2015). In our work, we combine the CD2c version of this genre dataset [4] with a collection of album cover images gathered from `7digital.com` using the information present in the MSD/Echo Nest mapping archive.[5] The final dataset contains 30,713 tracks from the MSD and their related album cover images, each annotated with a unique genre label among 15 classes. Based on an initial analysis on the images, we identified that this set of tracks is associated to 16,753 albums, yielding an average of 1.8 songs per album. We also gathered audio previews of all tracks from `7digital.com`. To facilitate the reproducibility of this work, all metadata, splits, feature embeddings, and links to related content are released as a new dataset called the MSD-I[6].

We randomly divide the dataset into three parts: 70% for training, 15% for validation, and 15% for test, with no artist and album overlap across these sets. This is crucial to avoid possible overfitting (Flexer, 2007), as the classifier may learn to predict the artist instead of the genre. In Table 1 we report the number of instances of each genre in the three subsets, and also the percentage of genre distribution on the entire dataset.

[4] `http://www.tagtraum.com/msd_genre_datasets.html`

[5] `http://labs.acousticbrainz.org/million-song-dataset-echonest-archive`

[6] `https://www.upf.edu/web/mtg/msdi`

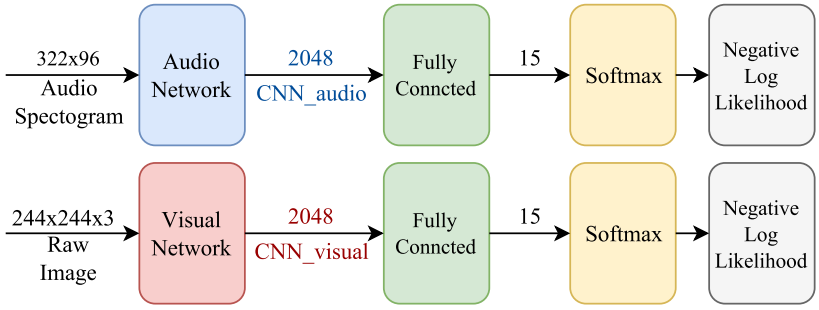


Figure 1: Scheme of two CNNs (*Top*: audio, *Bottom*: visual).

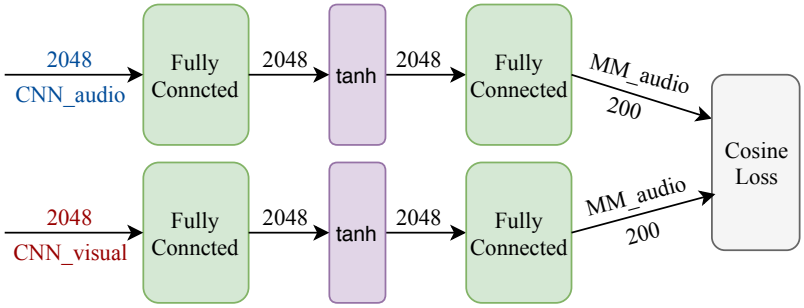


Figure 2: Scheme of the multimodal feature space network. The previously learned features from different modalities are mapped to the same space.

Rock (16.9%), Electronic (15.8%), and Pop (11.1%) are the most frequent, while Latin (1.8%), New Age (0.86%), and World (1.62%) the least represented.

6.2 Training procedure

To extract the audio features, we first train the CNN described in Section 3.1 on the genre classification task. We employ three convolutional layers, with the following number of filters, from first to last: 64, 128, and 256. Similar to van den Oord et al. (2013) the convolutions are only applied to the time axis, using a 4 frames width filter in each layer. Max pooling of 4 units across the time axis is applied after each of the first two convolutional layers, and max pooling of 2 after the third. Dropout of 0.5 is applied to all layers, as applied in Choi et al. (2016a). The flattened output of the last layer has 2048 units and the final fully connected layer has 15 units (to match the number of classes aiming to be predicted) with softmax activation. Categorical crossentropy is used as the loss function. Mini batches of 32 items are randomly sampled from the training data to compute the gradient, and Adam (Kingma and Ba, 2014) is the optimizer used to train the models, with the default suggested learning parameters. The networks are trained with a maximum of 100 epochs with early stopping. Once trained, we extract the 2048-dimensional vectors from the previous to last fully connected layer (CNN_AUDIO) for the training, validation, and test sets (see Figure 1).

The visual features are similarly extracted from the ResNet described in Section 3.2. The network is trained on the genre classification task with mini batches of 50 samples, for 90 epochs, a learning rate of 0.0001, and with Adam as optimizer. Once the network converges, we obtain the 2048-dimensional features (CNN_VISUAL) from the input to the last fully connected layer of the ResNet (see Figure 1).

Finally, we extract the multimodal features from the network described in Section 3.4. We first train the multimodal feature space, and later extract the feature vectors from the last fully connected layers (i.e., MM_VISUAL and MM_AUDIO), as shown in Figure 2. To obtain MM_AUDIO, at test time, no visual features are needed, only audio features (CNN_AUDIO). The same method is applied to the visual features, where only visual features (CNN_VISUAL) are used to obtain the MM_VISUAL features of the test set.

In all described networks, feature vectors of items from train, validation, and test sets are obtained. These feature vectors are fed to the multinomial fusion network described in Section 4, and classification results are obtained. This latter training is done with a maximum of 100 epochs with early stopping, and dropout applied after the input layer with a factor of 50%.

6.3 Results and Discussion

Table 2 shows the Precision (**P**), Recall (**R**), and F1-Scores (**F1**) for the (**Audio**), (**Visual**) and (**A + V**) approaches. Results shown are the macro average of the values obtained for every class[7]. Every experiment was run 3 times and mean and standard devi-

[7]Note that the reported F1-score is the average of the F1-score of every class, it is not calculated as the harmonic mean of the macro precision and recall values.

Input	Model	P	R	F1
Audio	CNN_AUDIO	0.385 ± 0.006	0.341 ± 0.001	0.336 ± 0.002
	MM_AUDIO	0.406 ± 0.001	0.342 ± 0.003	0.334 ± 0.003
	CNN_AUDIO + MM_AUDIO	0.389 ± 0.005	0.350 ± 0.002	$\mathbf{0.346 \pm 0.002}$
Video	**CNN_VISUAL**	0.291 ± 0.016	0.260 ± 0.006	$\mathbf{0.255 \pm 0.003}$
	MM_VISUAL	0.264 ± 0.005	0.241 ± 0.002	0.239 ± 0.002
	CNN_VISUAL + MM_VISUAL	0.271 ± 0.001	0.248 ± 0.003	0.245 ± 0.003
A + V	CNN_AUDIO + CNN_VISUAL	0.485 ± 0.005	0.413 ± 0.005	0.425 ± 0.005
	MM_AUDIO + MM_VISUAL	0.467 ± 0.007	0.393 ± 0.003	0.400 ± 0.004
	ALL	0.477 ± 0.010	0.413 ± 0.002	$\mathbf{0.427 \pm 0.000}$

Table 2: Genre classification experiments in terms of macro precision, recall, and f-measure. Every experiment was run 3 times and mean and standard deviation of the results are reported.

	Human Annotator			Neural Model		
Genre	**Audio**	**Visual**	**A + V**	**Audio**	**Visual**	**A + V**
Blues	0	0.50	0.67	0.05	0.36	0.42
Country	0.40	0.60	0.31	0.37	0.21	0.40
Electronic	0.62	0.44	0.67	0.64	0.44	0.68
Folk	0	0.33	0	0.13	0.23	0.28
Jazz	0.62	0.38	0.67	0.47	0.27	0.49
Latin	0.33	0.33	0.40	0.17	0.08	0.13
Metal	0.80	0.43	0.71	0.69	0.49	0.73
New Age	0	0	0	0	0.12	0.10
Pop	0.43	0.46	0.42	0.39	0.43	0.49
Punk	0.44	0.29	0.46	0.04	0	0.30
Rap	0.74	0.29	0.88	0.73	0.39	0.73
Reggae	0.67	0	0.80	0.51	0.34	0.55
RnB	0.55	0	0.46	0.45	0.31	0.51
Rock	0.58	0.40	0.40	0.54	0.20	0.58
World	0	0.33	0	0	0	0.03
Average	**0.41**	**0.32**	**0.46**	**0.35**	**0.25**	**0.43**

Table 3: Detailed results of the genre classification task. Human annotated results on the left, and our best models on the right (CNN_AUDIO + MM_AUDIO, CNN_VISUAL, and ALL respectively).

ation of the results are reported in Table 2. The results show that the combination of audio and visual features greatly outperforms audio and visual modalities in isolation. Audio seems to be a better source of features for genre classification, as it obtains a higher performance over visual features. Furthermore, we observe that the addition of the features learned from the multimodal feature space MM_AUDIO yields better performance in the case of audio. This implies that audio features get benefited by the multimodal space, resulting in an improvement of the quality of pure audio prediction when images are only used in the training of the multimodal feature space, and not in the prediction.

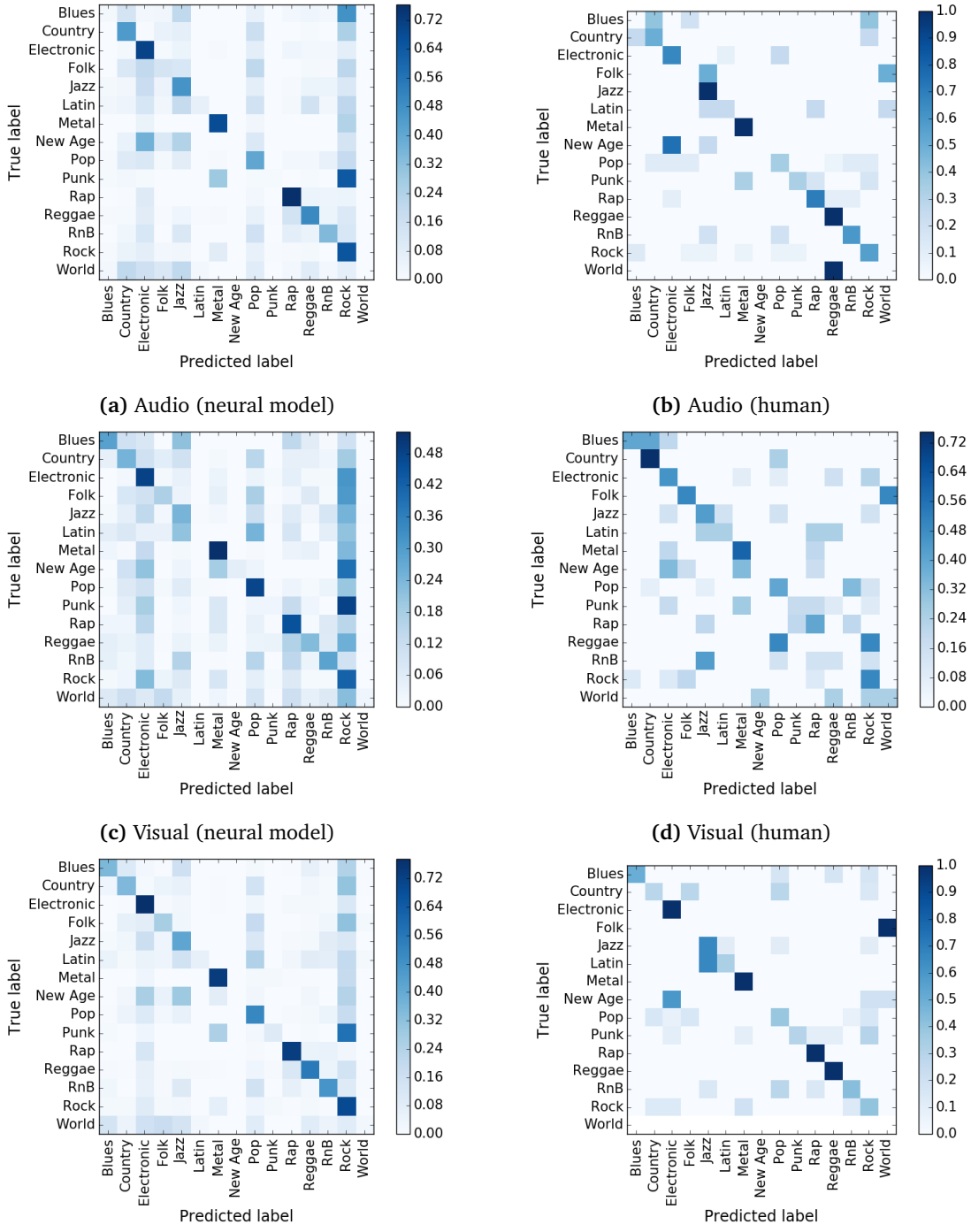
Finally, the aggregation of all feature vectors yields the highest results. It seems that every feature vector is helping to boost the performance of specific classes. Therefore, the neural network allows the aggregated features to improve the results.

We further explore the results by splitting them into the different genre classes to understand where our models perform better. In Table 3 the F1-Scores for these results are reported. The "Neural model" column displays the per class results of the best approach for each modality. The performance of the audio and visual features is correlated (Pearson correlation of 0.80), and audio features generally outperform visual features. However, visual features perform better than audio in Pop, even though this is a well populated class. Moreover, visual features clearly outperform audio in Blues and Folk. The aggregation of all features is able to combine the ability of each feature vector and obtain the best results in all classes. New Age and World obtain very low performance in all settings, being also the least represented classes in the dataset.

6.4 Human evaluation

We compare our neural network results with a human expert performing the same genre classification task.[8]

[8]The human annotator is an external music expert. Although having these genres annotated by several experts would diminish potential problems that subjectivity may arise, a single expert annotator is a fairly good indicative of human performance.



(a) Audio (neural model)

(b) Audio (human)

(c) Visual (neural model)

(d) Visual (human)

(e) A + V (neural model)

(f) A + V (human)

Figure 3: Confusion matrices of the three settings from the classification with the Neural Network models (CNN_Audio + MM_Audio, CNN_Visual and ALL) and the human annotator.

The subject annotated 300 songs of different albums and artists from the test set with their corresponding genre from the given list of 15 genres[9]. Genres of the songs were balanced following the same distribution of the test set. The content presented to the annotator was divided into 100 songs with audio tracks, 100 with cover images, and 100 with audio tracks and their corresponding cover images. The annotator can only see the album cover in the visual experiment, listen to the audio in the audio experiment, and both things in the multimodal experiment. Neither titles nor artist names were displayed.

[9]All procedures performed in this study involving human subjects were conducted in accordance with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from the participant.

Looking at Table 3 we see how the human outperforms the best neural models in the three experiments. However, the distances among scores between the annotator and the model are small, especially in the multimodal experiment. This implies that deep learning models are not too far away from human performance when classifying music by genre. Furthermore, we observe a strong correlation between the annotator and our model in the audio experiment (Pearson correlation 0.87), whereas there is no correlation in the visual experiment (0.24). This observation suggests that our audio model may be using similar features than those employed by humans, while our visual model is learning differently. Although intuitively the human performance should not depend on the number of instances per class in the training set, we observe that classes where the human and the model fail are those with a lower number of instances. This may suggest that some of these classes are difficult for audio-based classification regardless of the number of instances.

(a) Heavy Metal albums

(b) New Age albums misclassified as Heavy Metal

Figure 4: Heavy Metal and New Age album covers

6.5 Qualitative analysis

6.5.1 Error analysis

To better understand the role of each modality in the classification, we analyzed the confusion matrices (see Figure 3) of the neural model approaches and the human annotator present in Table 3. We observe again that audio features perform poorly on less populated classes (e.g., Blues, Latin, New Age, Punk and World), whereas visual features are able to achieve better results on Blues and New Age. This might be one of the reasons the two modalities complement each other well.

We observe that World music albums are highly misclassified in all the approaches. Apart from the reduced number of instances this class has, World is a too broad genre that may encompass very different types of music, making the classification harder or almost impossible from a human perspective. In addition, many albums are incorrectly classified as Rock, which is more evident in the visual approach, something that does not happen to the human annotator. The same problem appears when dealing with audio features, but the effect appears diminished. Rock is one of the most populated classes in our dataset, implying a high degree of musical variation. In all modalities, there are also an important number of albums incorrectly classified as Electronic, Jazz or Pop.

Moreover, it is worth noting that New Age albums are sometimes incorrectly classified as Heavy Metal. In Figures 4a and 4b we observe how the classifier may be identifying horns as a visual characteristic of Metal albums. In some instances, there are clear visual similarities on the cover images of these genres that, by contrast, do not exist in the audio signal.

In general, audio features seem to be more fine grained for the classification, but we need more instances in all classes to properly feed the classifier. We observe that the Audio + Visual approach produces fewer errors in general, with Rock being the most misclassified class.

6.5.2 Visual heatmaps

Recently Zhou et al. (2016) proposed an approach useful to visualize the areas of an image where a CNN focuses its attention to drive the label-prediction process. By performing global average pooling on the convolutional feature maps obtained after the chain of layers of a CNN, they are able to build a heatmap, referred to as Class Activation Mapping: this heatmap highlights the portions of the input image that have mostly influenced the image classification process. The approach consists in providing a heatmap for each class, which is very useful for recognition of objects in images. Since Resnet includes a GAP layer we just forward images of the test set and extract the weights of the GAP layer.

Using this technique we tried to properly study the misclassification problems observed in the previous section. We observed that the attention of the network is often focused on faces for Rap, Blues, Reggae, R&B, Latin, and World genres. For Jazz, the network seems to focus more on instruments, typographies, and clothes; for Rock and Electronic on backgrounds; for Country on faces, hats, and jeans; and for Folk on typographies. We observed that the network is also focusing on aging aspects of faces, associating for instance old black men with Blues. We also observed that the network tends to identify covers with nude parts of the body as Pop. In Figure 5 we present some examples of these observations. We provide all the images of the test set mapped with the attention heat-map [10] to bet-

[10] https://fvancesco.github.io/saliency/saliency.html

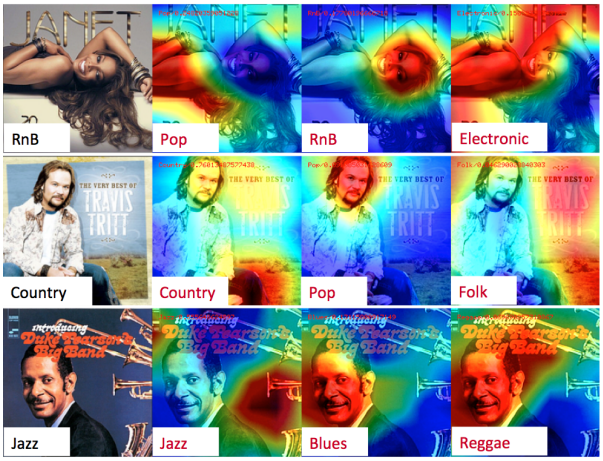


Figure 5: Examples of heatmaps for different genre classes. The genres on the left column are the ground truth ones.

ter explore where the network focuses during the predictions. Finally, thanks to this technique we corroborated the assumption presented in the previous subsection about the relation between cover arts with horns and Metal genre, as shown in Figure 6.

Figure 6: Heatmap for Metal genre class of a Metal (top) and a New Age (bottom) album with horns.

7. Multi-label classification experiment

In this section we describe the dataset and the experimental framework for multi-label genre classification from audio, text, and images. More specifically, and since each modality used (i.e., cover image, text reviews, and audio tracks) is associated with a music album, our task focuses this time on album classification, instead of track classification. Lastly, we report and discuss the results of each experiment and present a qualitative analysis of the results.

7.1 MuMu dataset

To the best of our knowledge, there are no publicly available large-scale datasets that encompass audio, images, text, and multi-label genre annotations. Therefore, we present *MuMu*, a new Multimodal Music dataset with multi-label genre annotations that combines information from the Amazon Reviews dataset (McAuley et al., 2015) and the MSD. The former contains millions of album customer reviews and album metadata gathered from Amazon.com.

To map the information from both datasets we use MusicBrainz[11], an open encyclopedia of music metadata. 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 Oramas et al. (2015), following a pair-wise 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[12] 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. We also use in these experiments audio previews retrieved from `7digital.com` (see Section 6.1). 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. The mapping between the three datasets (Amazon, MusicBrainz, and MSD), genre annotations, data splits, text reviews, and links to images are released as the *MuMu* dataset[13].

7.1.1 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 both 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).

[11] `http://musicbrainz.org`
[12] `http://labs.acousticbrainz.org/million-song-dataset-echonest-archive`
[13] `https://www.upf.edu/web/mtg/mumu`

Genre	% of albums	Genre	% of albums
Pop	84.38	Tributes	0.10
Rock	55.29	Harmonica Blues	0.10
Alternative Rock	27.69	Concertos	0.10
World Music	19.31	Bass	0.06
Jazz	14.73	European Jazz	0.06
Dance & Electronic	12.23	Piano Blues	0.06
Metal	11.50	Norway	0.06
Indie & Lo-Fi	10.45	Slide Guitar	0.06
R&B	10.10	East Coast Blues	0.06
Folk	9.69	Girl Groups	0.06

Table 4: Top-10 most and least represented genres.

The labels in the dataset are highly unbalanced, following a distribution that might align well with those found in real world scenarios. In Table 4 we see the top 10 most and least represented genres and the percentage of albums annotated with each label. The unbalanced nature of the genre annotations poses an interesting challenge for music classification that we also aim to exploit.

7.2 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, by plotting the true positive rate (TPR) against the false positive rate (FPR). 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 (Choi et al., 2016a; Dieleman and Schrauwen, 2014).

The output of a multi-label classifier is a label-item matrix. This matrix contains the probabilities of each class for every item when using the LOGISTIC configuration, and the cosine similarity between items and labels latent factors for the COSINE configuration. This matrix 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 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 that are over-represented in the dataset (see Table 4) and more specific subgenres (e.g., Vocal Jazz, Britpop), we want to measure whether the classifier is focusing only on over-represented genres, or on more fine-grained ones. We assume that an ideal classifier would exploit better the taxonomic depth. To measure this, we use aggregated diversity (Adomavicius and Kwon, 2012), also known as catalog coverage. ADiv@N measures the percentage of normalized unique labels present in the top K predictions across all test items. Values of $k = 1, 3, 5$ are typically employed in multi-label classification (Jain et al., 2016) .

CNN layer	Filter	Max pooling
1	3x3	(4,2)
2	3x3	(4,2)
3	3x3	(4,1)
4	1x1	(4,5)
1	4x96	(4,1)
2	4x1	(4,1)
3	4x1	(4,1)
4	1x1	-
1	4x70	(4,4)
2	4x6	(4,1)
3	4x1	(4,1)
4	1x1	-

Table 5: Filter and max pooling sizes applied to the different layers of the three audio CNN approaches used for multi-label classification.

7.3 Training procedure

The dataset is divided as follows: 80% for training, 10% for validation, and 10% for test. Following the same artist filter used in Section 6.1, all sets contain albums from different artists to avoid overfitting. The matrix of album genre annotations of the training and validation sets is factorized using the approach described in Section 5.1, with a value of $d = 50$ dimensions.

7.3.1 Audio

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, as in the network described in Section 4.

To learn the track genre labels we design a CNN as the one described in Section 3.1, with four convolutional layers. We experiment with different number of filters, filter sizes, and output configurations. For the filter size we compare three approaches: square 3x3 filters as in Choi et al. (2016a), a filter of 4x96 that convolves only in time (van den Oord et al., 2013), and a musically motivated filter of 4x70, which is able to slightly convolve in the frequency axis (Pons et al., 2016). To study the width of the convolutional layers we try two different settings: HIGH with 256, 512, 1024, and 1024 filters in each layer respectively, and LOW with 64, 128, 128, 64 filters. Max pooling is ap-

plied after each convolutional layer (see Table 5 for further details about convolutional filter sizes and max pooling layers). Finally, we use the two different network targets defined in Section 5, 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 (Bertin-Mahieux et al., 2011). 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 4 with LOGISTIC output. This approach is denoted as TIMBRE-MLP.

All these networks are trained with a maximum of 100 epochs and early stopping, using mini batches of 32 items, randomly sampled from the training data to compute the gradient, and Adam is the optimizer used to train the models, with the default suggested learning parameters unless otherwise specified.

7.3.2 Text

In the presented dataset, each album has a variable number of customer reviews. We use an approach similar to the one described in Oramas et al. (2016a) for genre classification from text, where all reviews from the same album are aggregated into a single text. The aggregated result is truncated at approximately 1500 characters (incomplete sentences are removed from the end of the truncated text), thus balancing the amount of text per album, as more popular artists tend to have a higher number of reviews. As reviews are chronologically ordered in the dataset, older reviews are favored in this process. After truncation, we apply the semantic enrichment and Vector Space Model approaches described in Section 3.3. The vocabulary size of the VSM is limited to 10k as it yields a good balance of network complexity and accuracy.

For text classification, we obtain two feature vectors as described in Section 3.3: one built from the texts (VSM), and another built from the semantically enriched texts (VSM+SEM). Both feature vectors are trained in the multi-label genre classification task using the two output configurations LOGISTIC and COSINE. This network is also trained with mini batches of 32 items, and Adam as optimizer.

7.3.3 Images

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) described in Section 3.2 with LOGISTIC output. The network is trained on the genre classification task with mini batches of 50 samples for 90 epochs, a learning rate of 0.0001, and with Adam as optimizer.

7.4 Results and Discussion

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 approach in terms of AUC (**A**, **V**, and **I**). Finally, the three modality vectors are combined in a multimodal network as the one described in Section 4. All results are reported in Table 6 and are discussed next. Performance of the classification is reported in terms of AUC score and ADiv@N with $N = 1, 3, 5$. The training speed per epoch and number of network hyperparameters are also reported.

The results on audio classification show that CNNs applied over audio spectrograms clearly outperform our baseline approach 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 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 better in terms of aggregated diversity.

Results on text classification show that the semantic enrichment of texts clearly yields better results in terms of AUC and diversity. Furthermore, we observe that the COSINE configuration slightly outperforms LOGISTIC in terms of AUC, and greatly in terms of aggregated diversity. The text-based results are overall slightly superior to the audio-based ones.

Results show that genre classification from images underperforms in terms of AUC and aggregated diversity compared to the other modalities. Due to the use of an already pre-trained network with a LOGISTIC output (ImageNet, Russakovsky et al., 2015) as initialization of the network, it is not straightforward to apply the COSINE configuration. Therefore, we only report results for the LOGISTIC configuration.

From the best performing approaches in terms of AUC of each modality (i.e., AU-

Modality	Target	Settings	Params	Time	AUC	ADiv@1	ADiv@3	ADiv@5
AUDIO	LOGISTIC	TIMBRE-MLP	0.01M	1s	0.792	0.04	0.14	0.22
AUDIO	LOGISTIC	LOW-3X3	0.5M	390s	0.859	0.14	0.34	0.54
AUDIO	LOGISTIC	HIGH-3X3	16.5M	2280s	0.840	0.20	0.43	0.69
AUDIO	LOGISTIC	LOW-4X96	0.2M	140s	0.851	0.14	0.32	0.48
AUDIO	LOGISTIC	HIGH-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	135s	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**
TEXT	LOGISTIC	VSM	25M	11s	0.905	0.08	0.20	0.37
TEXT	LOGISTIC	VSM+SEM	25M	11s	0.916	0.10	0.25	0.44
TEXT	COSINE	VSM	25M	11s	0.901	0.53	0.44	0.90
TEXT (T)	**COSINE**	**VSM+SEM**	**25M**	**11s**	**0.917**	**0.42**	**0.70**	**0.85**
IMAGE (I)	LOGISTIC	RESNET	1.7M	4009s	0.743	0.06	0.15	0.27
A + T	LOGISTIC	MLP	1.5M	2s	0.923	0.10	0.40	0.64
A + I	LOGISTIC	MLP	1.5M	2s	0.900	0.10	0.38	0.66
T + I	LOGISTIC	MLP	1.5M	2s	0.921	0.10	0.37	0.63
A + T + I	**LOGISTIC**	**MLP**	**2M**	**2s**	**0.936**	0.11	0.39	0.66
A + T	COSINE	MLP	0.3M	2s	0.930	**0.43**	**0.74**	**0.86**
A + I	COSINE	MLP	0.3M	2s	0.896	0.32	0.57	0.76
T + I	COSINE	MLP	0.3M	2s	0.919	0.43	0.74	0.85
A + T + I	COSINE	MLP	0.4M	2s	0.931	0.42	0.72	0.86

Table 6: Results for Multi-label Music Genre Classification of Albums. Number of network hyperparameters, epoch training time, AUC-ROC, and aggregated diversity at $N = 1, 3, 5$ for different settings and modalities.

DIO/COSINE/HIGH-4X70, TEXT/COSINE/VSM+SEM and IMAGE/LOGISTIC/RESNET), an internal feature representation is obtained as described in Section 3. Then, these three feature vectors are aggregated in all possible combinations, and genre labels are predicted using the feedforward network described in Section 4. Both output configurations LOGISTIC and COSINE are used in the learning phase, and dropout of 0.7 is applied in the COSINE configuration (we empirically determined that this dropout factor yields better results).

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 achieve better 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 aggregated diversity, thanks to the spatial properties of the factorized space.

7.5 Qualitative Analysis

From the set of album factors obtained from the factorization of the training set (see Section 5.1), those annotated with a single label from the top level of the taxonomy are plotted in Figure 7 using t-SNE dimensionality reduction (Maaten and Hinton, 2008). It can be seen how the different albums are properly clustered in the factorized space according to their genre.

In addition, we studied the list of Top-3 genres predicted for every album in the test set for the best LOGISTIC and COSINE audio-based approaches in terms of AUC (LOGISTIC/LOW-4X70 AND COSINE/HIGH-4X70). In Table 7 we see these predictions for the first 20 albums in the test set. We clearly observe in these results the higher diversity of the predictions of the COSINE approach. A listening test on tracks of the predicted albums suggests that the predictions of the COSINE approach are more fine-grained that those provided by the LOGISTIC approach, as we observed that COSINE results accurately include labels from deeper levels of the taxonomy.

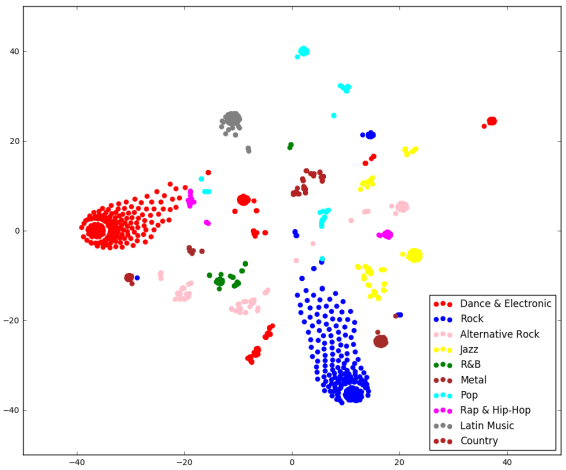


Figure 7: t-SNE of album factors.

We also studied the information gain of words in the different genres from the best text-based classification approach. We observed that genre labels present inside the texts have high information gain values. It

Amazon ID	LOGISTIC	COSINE
B00002SWJF	Pop,Dance & Electronic,Rock	Dance & Electronic,Dance Pop,Electronica
B00006FX4G	Pop,Rock,Alternative Rock	Rock,Alternative Rock,Pop
B000000PLF	Pop,Jazz,Rock	Jazz,Bebop,Modern Postbebop
B00005YQOV	Pop,Jazz,Bebop	Jazz,Cool Jazz,Bebop
B0000026BS	Jazz,Pop,Bebop	Jazz,Bebop,Cool Jazz
B0000006PK	Pop,Jazz,Bebop	Jazz,Bebop,Cool Jazz
B00005O6NI	Pop,Rock,World Music	Blues,Traditional Blues,Acoustic Blues
B000BPYKLY	Pop,Jazz,R&B	Smooth Jazz,Soul-Jazz & Boogaloo,Jazz
B000007U2R	Pop,Dance & Electronic,Dance Pop	Dance Pop,Dance & Electronic,Electronica
B002LSPVJO	Rock,Pop,Alternative Rock	Rock,Alternative Rock,Metal
B000007WE5	Pop,Rock,Country	Rock,Pop,Singer-Songwriters
B001IUC4AA	Dance & Electronic,Pop,Dance Pop	Dance & Electronic,Dance Pop,Electronica
B000SFJW2O	Pop,Rock,World Music	Pop,Rock,World Music
B000002NE8	Pop,Rock,Dance & Electronic	Dance & Electronic,Dance Pop,Electronica
B000002GUJ	Rock,Pop,Alternative Rock	Alternative Rock,Indie & Lo-Fi,Indie Rock
B00004T0QB	Pop,Rock,Alternative Rock	Rock,Alternative Rock,Pop
B0000520XS	Pop,Rock,Folk	Singer-Songwriters,Contemporary Folk,Folk
B000EQ47W2	Pop,Rock,Alternative Rock	Metal,Pop Metal,Rock
B00000258F	Pop,Rock,Jazz	Smooth Jazz,Soul-Jazz & Boogaloo,Jazz
B000003748	Pop,Rock,Alternative Rock	Alternative Rock,Rock,American Alternative

Table 7: Top-3 genre predictions in albums from the test set for LOGISTIC and COSINE audio-based approaches.

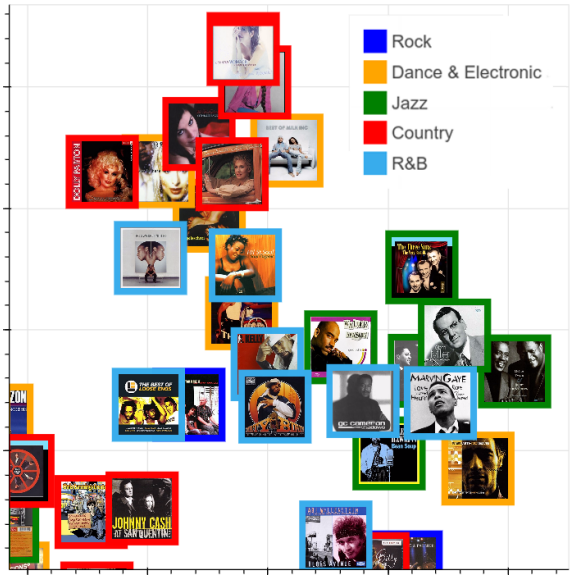


Figure 8: t-SNE visualization of randomly selected image vectors from five of the most frequent genres.

is also remarkable that *band* is a very informative word for Rock, *song* for Pop, and *dope*, *rhymes*, and *beats* are discriminative features for Rap albums. Location names have also important weights, as *Jamaica* for Reggae, *Nashville* for Country, or *Chicago* for Blues[14].

In Figure 8 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. We observe how album feature vectors of the same genre cluster well in the space. In the left top corner the ResNet recognizes women faces on the foreground, which seems to be common in Country albums (red). Also the R&B genre appears to be generally well clustered, since black men that the network sucessfully recognizes tend to appear on the cover. The jazz albums (green) on the right are all clustered together, perhaps thanks to the uniform type of clothing worn by the people of their covers, or the black and white images. Therefore, similarly to the qualitative analysis presented in Section 6.5, we observed that the visual style of the cover seems to be informative when recognizing the album genre.

[14]The complete list of words is available on-line at `https://www.upf.edu/en/web/mtg/mumu`

8. Conclusions

In this work we have proposed a representation learning approach for the classification of music genres from different data modalities, i.e., audio, text, and images. The proposed approach has been applied to a traditional classification scenario with a small number of mutually exclusive classes. It has also been applied to a multi-label classification scenario with hundreds of non mutually exclusive classes. In addition, we have proposed an approach based on the learning of a multimodal feature space and a dimensionality reduction of target labels using PPMI.

Results show in both scenarios that the combination of learned data representations from different modalities yields better results than any of the modalities in isolation. In addition, a qualitative analysis of the results have shed some light on the behavior of the different modalities. Moreover, we have compared our neural model with a human annotator, reveling correlations and showing that our deep learning approach is not far from human performance.

In our single-label experiment we clearly observed how visual features perform better in some classes where audio features fail, thus complementing each other. In addition, we have shown that the learned multimodal feature space seems to improve the performance of audio features. This space increases accuracy, even when the visual part is not present in the prediction phase. This is a promising result, not only for genre classification, but also for other applications such as music recommendation, especially when data

from different modalities are not always available for every item. However, more experimentation is needed to confirm this finding.

In our multi-label experiment we provide evidence of 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. Furthermore, the dimensionality reduction of target labels led to better results, not only in terms of AUC, but also in terms of aggregated diversity.

To carry out the experiments, we have collected and released two novel multimodal datasets for music genre classification. First, *MSD-I*, a dataset with over 30k audio tracks and their corresponding album cover artwork and genre annotation. Second, *MuMu*, a new multimodal music dataset with over 31k albums, 147k audio tracks, and 450k album reviews.

To conclude, this work has deeply explored the classification problem of music genres from different perspectives and using different data modalities, introducing novel ideas to approach this problem coming from other domains. In addition, we envision that the proposed multimodal deep learning approach may be easily applied to other MIR tasks (e.g., music recommendation, audio scene classification, machine listening, cover song identification). Moreover, the release of the gathered datasets opens up a number of potentially unexplored research possibilities.

9. Reproducibility

Both datasets used in the experiments are released as MSD-I[15] and *MuMu*[16]. The released data includes mappings between data sources, genre annotations, splits, texts, and links to images. Audio and image files are not released due to copyright issues. The source code to reproduce the audio, text, and multimodal experiments[17] and the visual experiments[18] is also available.

Acknowledgements

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).

[15] https://doi.org/10.5281/zenodo.1240484

[16] https://doi.org/10.5281/zenodo.831188

[17] https://github.com/sergiooramas/tartarus

[18] https://github.com/fvancesco/music_resnet_classification

References

Adomavicius, G. and Kwon, Y. (2012). Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911.

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Bertin-Mahieux, T., Eck, D., Maillet, F., and Lamere, P. (2008). Autotagger: A model for predicting social tags from acoustic features on large music databases. *Journal of New Music Research*, 37(2):115–135.

Bertin-Mahieux, T., Ellis, D. P. W., Whitman, B., and Lamere, P. (2011). The Million Song Dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR)*.

Bogdanov, D., Porter, A., Herrera, P., and Serra, X. (2016). Cross-collection evaluation for music classification tasks. In *Proceedings of the 17th International Conference on Music Information Retrieval (ISMIR)*.

Choi, K., Fazekas, G., and Sandler, M. (2016a). Automatic tagging using deep convolutional neural networks. *Proceedings of the 17th International Conference on Music Information Retrieval (ISMIR)*, pages 805–811.

Choi, K., Fazekas, G., Sandler, M., and Cho, K. (2016b). Convolutional Recurrent Neural Networks for Music Classification. *arXiv preprint arXiv:1609.04243*.

Choi, K., Lee, J. H., and Downie, J. S. (2014). What is this song about anyway?: Automatic classification of subject using user interpretations and lyrics. *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries*, pages 453–454.

Chollet, F. (2016). Information-theoretical label embeddings for large-scale image classification. *arXiv preprint arXiv:1607.05691*.

Dieleman, S., Brakel, P., and Schrauwen, B. (2011). Audio-based music classification with a pretrained convolutional network. In *Proceedings of the 12th International Society for Music Information Retrieval Conference (ISMIR)*, pages 669–674.

Dieleman, S. and Schrauwen, B. (2014). End-to-end learning for music audio. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 6964–6968. IEEE.

Dorfer, M., Arzt, A., and Widmer, G. (2016). Towards score following in sheet music images. *Proceedings of the 17th International Society of Music Information Retrieval Conference (ISMIR)*.

Downie, J. S. and Hu, X. (2006). Review mining for music digital libraries: phase II. *Proceedings of the 6th ACM/IEEE-CS Joint Conference on Digital Libraries*, page 196.

Flexer, A. (2007). A Closer Look on Artist Filters for Musical Genre Classification. In *Proceedings of the 8th International Society for Music Information Retrieval Conference (ISMIR)*.

Gouyon, F., Dixon, S., Pampalk, E., and Widmer, G. (2004). Evaluating rhythmic descriptors for musical genre classification. In *Proceedings of the AES 25th International Conference*, pages 196–204.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778.

Howard, A. G. (2013). Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*.

Hu, X. and Downie, J. (2006). Stylistics in customer reviews of cultural objects. *SIGIR Forum*, pages 49–51.

Hu, X., Downie, J., West, K., and Ehmann, A. (2005). Mining Music Reviews: Promising Preliminary Results. In *Proceedings of the 6th International Society of Music Information Retrieval Conference (ISMIR)*.

Jain, H., Prabhu, Y., and Varma, M. (2016). Extreme Multi-label Loss Functions for Recommendation, Tagging, Ranking & Other Missing Label Applications. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 935–944. ACM.

Kim, Y. (2014). Convolutional Neural Networks for Sentence Classification. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 1746–1751.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Laurier, C., Grivolla, J., and Herrera, P. (2008). Multimodal music mood classification using audio and lyrics. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on*, pages 688–693. IEEE.

Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems*, pages 2177–2185.

Libeks, J. and Turnbull, D. (2011). You can judge an artist by an album cover: Using images for music annotation. *IEEE MultiMedia*, 18(4):30–37.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, pages 740–755.

Logan, B. (2000). Mel Frequency Cepstral Coefficients for Music Modeling. In *Proceedings of the 1st International Society of Music Information Retrieval Conference (ISMIR)*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

McAuley, J., Targett, C., Shi, Q., and Van Den Hengel, A. (2015). Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 43–52.

McFee, B., Bertin-Mahieux, T., Ellis, D. P., and Lanckriet, G. R. (2012). The million song dataset challenge. *Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion*, page 909.

Mcfee, B., Raffel, C., Liang, D., Ellis, D. P. W., Mcvicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference*, (Scipy):1–7.

McKay, C. and Fujinaga, I. (2008). Combining features extracted from audio, symbolic and cultural sources. In *Proceedings of the 9th International Society of Music Information Retrieval Conference*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation: a Unified Approach. *Transactions of the Association for Computational Linguistics (TACL)*, 2:231–244.

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193:217–250.

Neumayer, R. and Rauber, A. (2007). Integration of text and audio features for genre classification in music information retrieval. In *European Conference on Information Retrieval*, pages 724–727.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., and Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 689–696.

Oramas, S. (2017). Semantic enrichment for similarity and classification. In *Knowledge Extraction and Representation Learning for Music Recommendation and Classification*, chapter 6, pages 75–88. PhD Dissertation, Universitat Pompeu Fabra.

Oramas, S., Espinosa-Anke, L., Lawlor, A., and Serra, X. (2016a). Exploring Customer Reviews for Music Genre Classification and Evolutionary Studies. In *Proceedings of the 17th International Society for Music Information Retrieval Conference (ISMIR)*.

Oramas, S., Espinosa-Anke, L., Sordo, M., Saggion, H., and Serra, X. (2016b). ELMD: An Automatically Generated Entity Linking Gold Standard Dataset in the Music Domain. In *In Proceedings*

of the 10th International Conference on Language Resources and Evaluation, LREC 2016.

Oramas, S., Gómez, F., Gómez, E., and Mora, J. (2015). FlaBase: Towards the Creation of a Flamenco Music Knowledge Base. In *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Oramas, S., Nieto, O., Barbieri, F., and Serra, X. (2017a). Multi-label music genre classification from audio, text, and images using deep features. *Proceedings of the 18th International Society of Music Information Retrieval Conference ISMIR 2017*.

Oramas, S., Nieto, O., Sordo, M., and Serra, X. (2017b). A deep multimodal approach for cold-start music recommendation. *2nd Workshop on Deep Learning for Recommender Systems, co-located with RecSys 2017*.

Pachet, F. and Cazaly, D. (2000). A taxonomy of musical genres. In *Content-Based Multimedia Information Access-Volume 2*, pages 1238–1245.

Pons, J., Lidy, T., and Serra, X. (2016). Experimenting with musically motivated convolutional neural networks. In *Content-Based Multimedia Indexing (CBMI), 2016 14th International Workshop on*, pages 1–6. IEEE.

Pons, J., Nieto, O., Prockup, M., Schmidt, E. M., Ehmann, A. F., and Serra, X. (2017). End-to-end learning for music audio tagging at scale. *arXiv preprint arXiv:1711.02520*.

Razavian, A. S., Azizpour, H., Sullivan, J., and Carlsson, S. (2014). Cnn features off-the-shelf: an astounding baseline for recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 512–519. IEEE.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252.

Sanden, C. and Zhang, J. Z. (2011). Enhancing Multi-label Music Genre Classification Through Ensemble Techniques. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '11, pages 705–714.

Schedl, M., Orio, N., Liem, C., and Peeters, G. (2013). A professionally annotated and enriched multimodal data set on popular music. In *Proceedings of the 4th ACM Multimedia Systems Conference*, pages 78–83. ACM.

Schindler, A. and Rauber, A. (2015). An audio-visual approach to music genre classification through affective color features. In *European Conference on Information Retrieval*, pages 61–67.

Schörkhuber, C. and Klapuri, A. (2010). Constant-q transform toolbox for music processing. In *7th Sound and Music Computing Conference, Barcelona, Spain*, pages 3–64.

Schreiber, H. (2015). Improving genre annotations for the million song dataset. *Proceedings of the 16th International Society for Music Information Retrieval Conference (ISMIR)*.

Sermanet, P. and LeCun, Y. (2011). Traffic sign recognition with multi-scale convolutional networks. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 2809–2813. IEEE.

Seyerlehner, K., Schedl, M., Pohle, T., and Knees, P. (2010a). Using block-level features for genre classification, tag classification and music similarity estimation. *Submission to Audio Music Similarity and Retrieval Task of MIREX*, 2010.

Seyerlehner, K., Widmer, G., Schedl, M., and Knees, P. (2010b). Automatic music tag classification based on block-level. *Proceedings of Sound and Music Computing 2010*.

Sordo, M. (2012). Semantic annotation of music collections: A computational approach. In *PhD Dissertation, Universitat Pompeu Fabra*.

Srivastava, N. and Salakhutdinov, R. R. (2012). Multimodal learning with deep boltzmann machines. In *Advances in neural information processing systems*, pages 2222–2230.

Sturm, B. L. (2012). A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimedia Retrieval*, pages 29–66.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2818–2826.

Tsoumakas, G. and Katakis, I. (2006). Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3).

Turnbull, D., Barrington, L., Torres, D., and Lanckriet, G. (2008). Semantic annotation and retrieval of music and sound effects. *IEEE Transactions on Audio, Speech, and Language Processing*, 16(2):467–476.

Tzanetakis, G. and Cook, P. (2002). Musical genre classification of audio signals. *IEEE Transactions on Speech and Audio Processing*, 10(5):293–302.

van den Oord, A., Dieleman, S., and Schrauwen, B. (2013). Deep content-based music recommendation. In *Advances in neural information processing systems*, pages 2643–2651.

Wang, F., Wang, X., Shao, B., Li, T., and Ogihara, M. (2009). Tag Integrated Multi-Label Music Style Classification with Hypergraph. In *Proceedings of*

the 10th International Society of Music Information Retrieval Conference (ISMIR).

Wu, X., Qiao, Y., Wang, X., and Tang, X. (2016). Bridging Music and Image via Cross-Modal Ranking Analysis. *IEEE Transactions on Multimedia*, 18(7):1305–1318.

Yan, F. and Mikolajczyk, K. (2015). Deep correlation for matching images and text. In *Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on*, pages 3441–3450. IEEE.

Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. (2014). How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., and Torralba, A. (2016). Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2921–2929.

Zobel, J. and Moffat, A. (1998). Exploring the similarity space. *ACM SIGIR Forum*, 32(1):18–34.