CONTRASTIVE LEARNING FOR CROSS-MODAL ARTIST RETRIEVAL

Andres Ferraro **Jaehun Kim** **Sergio Oramas**
Andreas Ehmann **Fabien Gouyon**
Pandora-SiriusXM, Oakland
andres.ferraro@siriusxm.com

ABSTRACT

Music retrieval and recommendation applications often rely on content features encoded as embeddings, which provide vector representations of items in a music dataset. Numerous complementary embeddings can be derived from processing items originally represented in several modalities, e.g., audio signals, user interaction data, or editorial data. However, data of any given modality might not be available for all items in any music dataset. In this work, we propose a method based on contrastive learning to combine embeddings from multiple modalities and explore the impact of the presence or absence of embeddings from diverse modalities in an artist similarity task. Experiments on two datasets suggest that our contrastive method outperforms single-modality embeddings and baseline algorithms for combining modalities, both in terms of artist retrieval accuracy and coverage. Improvements with respect to other methods are particularly significant for less popular query artists. We demonstrate our method successfully combines complementary information from diverse modalities, and is more robust to missing modality data (i.e., it better handles the retrieval of artists with different modality embeddings than the query artist's).

1. INTRODUCTION AND RELATED WORK

The MIR community has dedicated significant effort to defining and computing music similarity in the last 20 years. Music similarity can be used in multiple downstream tasks, from playlist continuation, music visualization/navigation, music categorization for organizing catalogs, or for personalized recommendations. The notion of similarity is subjective and there is no consensus on how to define and evaluate it [1]. To evaluate the performance of a music similarity algorithm, some previous works either focus on content-based aspects, such as melody or harmony. Other works measure similarity based on *cultural aspects*, such as based on the co-occurrence of items in playlists or on editorial data –this is the approach of our work.

Multiple methods have been proposed to compute music similarity based on a variety of data types related to the music, e.g., based on audio descriptors [2], document similarity [3], or graphs of musical connections [4, 5]. Some relatively recent works propose ways to produce embeddings –that can be used to compute music similarity– in a supervised or unsupervised way, by training models on large amounts of data (such as audio, text or image). Such pre-trained models, which are often released publicly, may produce feature representations –i.e. embeddings– that are effective for previously unseen tasks. Such embeddings can be computed from diverse types of modalities related to music such as audio [6–8], tags [9], album covers images [10], or biographies [4]. The multiple modalities of data that can describe a music item –such as audio, tags, or listening interactions– may contain *complementary* information. For example, the quality and scale of audio vs collaborative data has been shown to have significant influence in autotagging tasks [11]. It therefore appears beneficial to combine diverse complementary modalities to obtain a more informative representation of music items. In fact, recent research identifies the combination of diverse sources of data as specially promising for mitigating limitations and issues in music recommendation research [12].

Another aspect to take into account is that in any given music dataset, data of diverse modalities might be available for different subsets of items. Therefore, when querying with an item represented in a given modality, the maximum coverage for retrieval is limited to items for which that same modality is available, leaving out a potentially significant –and relevant– part of the dataset. For example, the availability of listening interactions or users' explicit feedback is highly dependent on item popularity. Therefore, for artists with very little listening and user feedback, it may not be possible to obtain embeddings from that modality. Embeddings from other modalities may suffer from the same issue, either because there is no data available to produce an embedding or because the quality of the available information is very low. For instance in the case of a model trained on tag annotations to produce artist embeddings, where the output embedding may not be very informative for those artists that have a single or few tag annotations. Such issues are particularly common and problematic emerging or more underground artists, for which the available information is more limited.

In order to mitigate the issue of availability of some modalities, it is important to combine and take full ad-

vantage of all information available so that when querying with an artist that has only one modality available, we can also retrieve artists for which we have a different modality information. Therefore, the focus of this work is to combine diverse modalities into a common shared space that is beneficial for 1) leveraging each modality information from the artists, and 2) allowing to operate on a single space that covers the full population of artists, ensuring that whether or not an artist is retrieved for another does not depend on the number of modalities available.

The problem of combining embeddings from diverse modalities in a shared representation has received some attention in the last few years. In the music domain, there have been some works on combining embeddings by simple concatenation [13] or predicting one modality from another [14]. Contrastive learning techniques go beyond simple concatenation or prediction, trying to learn a shared representation between embeddings from different modalities. Some examples of research related to multimodal contrastive learning can be found in [10], where embeddings from a shared multimodal space are used as additional features for classification, or in [15, 16] where, e.g., music audio can be retrieved from natural language descriptions. In this work, we propose to apply a contrastive learning method that maps embeddings from diverse modalities to a shared embedding space, extending the advantages of multiple modalities to populations that would not be covered otherwise.

In summary, in this work we propose an approach to combine the multiple encoders of a contrastive learning method, showcasing several improvements over baselines and single-modality approaches in an artist similarity task. We show under two different contexts –using an open and an in-house dataset– that our proposed approach:

- achieves higher performance in terms of accuracy and coverage of retrieved artists (§ 3.1),
- successfully combines complementary information from diverse modalities (§ 3.2),
- is more robust to missing modality data (§ 3.3),
- particularly increases the performance for less popular query artists (§ 3.4).

2. METHODOLOGY

2.1 Single-Modality Embeddings and Contrastive Method

In this work, we use three modalities, namely: tags, user-listening interactions (i.e. collaborative filtering data, referred to as CF), and audio information. In all cases, we use pre-trained models to obtain embeddings for each of the modalities. We evaluate artist similarity performance using the embeddings from the pre-trained models directly, and compare to the performance when using the embeddings produced by our contrastive method which is trained with the same embeddings from pre-trained models.

In these experiments we apply a contrastive learning loss based on InfoNCE [17]. Specifically, we define the contrastive loss between two modalities, $\boldsymbol{\psi}_a$ and $\boldsymbol{\psi}_b$, as:

$$\mathcal{L}_{\boldsymbol{\psi}_a, \boldsymbol{\psi}_b} = \sum_{i=1}^{M} -\log \frac{\Xi(\boldsymbol{\psi}_a^i, \boldsymbol{\psi}_b^i, \tau)}{\sum_{k=1}^{2M} \mathbb{1}_{[k \neq i]} \Xi(\boldsymbol{\psi}_a^i, \boldsymbol{\zeta}^k, \tau)}$$, where M is the batch size and τ is the temperature parameter. We define $\Xi(\mathbf{a}, \mathbf{b}, \tau) = \exp(\cos(\mathbf{a}, \mathbf{b})\tau^{-1})$, based on the cosine similarity. $\boldsymbol{\zeta}^k$ is defined as $\boldsymbol{\psi}_a^k$, if $k \leq M$ and else $\boldsymbol{\psi}_b^{k-M}$. This loss function attempts to minimize the distance between the modalities of the same artist while maximizing the distance with any modality from other artists.

We use three encoders –one for each modality– that will produce three representations in our shared space for each artist. During training [1] we minimize the sum of the pairwise losses between each of the modalities as in [18]:
$$\mathcal{L}_{\text{tot}} = \mathcal{L}_{\text{Audio-Tag}} + \mathcal{L}_{\text{Audio-CF}} + \mathcal{L}_{\text{Tag-CF}}$$

Once the model is trained with the contrastive method and we want to use it for inference, for a given artist, we aggregate the output of each internal encoder by averaging all available information.

2.2 Training Data

In order to investigate the effectiveness of our contrastive method under different situations, we train our model using two independent datasets: We use a dataset based on public data to facilitate the reproducibility of some of the results. And we also use an in-house dataset that contains multimodal information for a larger set of artists.

Training our model requires *full coverage of the three modalities for all artists* –tag-based embeddings, CF embeddings, and audio embeddings. For the public dataset, we use the Million Song Dataset (MSD) [19] and its connections with other datasets to collect tags, audio and CF embeddings. We collected audio track embeddings using the public unsupervised model from [6] to extract embeddings from MSD audio previews, then we averaged all audio tracks embeddings for each artist. The tagging data was collected from the MSD500 dataset [11] and embeddings were computed using PMI factorization [13] of 500 tags. The CF embeddings were obtained using weighted matrix factorization [20] based on the Echonest Profile dataset, [2] with Gaussian process-based Bayesian hyperparameter tuning [21]. We gathered information from the three modalities for $17,478$ artists.

For the in-house dataset (hereafter, OWN) we collected tags, CF, and audio information for $38,301$ artists. This dataset is larger than MSD and includes what we believe is *higher-quality tags and CF data*, which allows us to compare the performance of our approach in a different setting. The CF information is computed from very large amounts of user-listening interactions on a streaming platform. The audio embeddings are computed using the supervised model [3] described in [6]. The tag embeddings are

---

[1] For both datasets we use Adam optimization with a learning rate of 0.0001 and temperature of 0.1. We use a fully connected layer of 256 for the CF encoder, two layers with 512 and 256 for the Audio encoder and 4 attention heads of 256 for the tag encoder. The learned space has 200 dimentions. Batch size for C_{OWN} is 2048 and for C_{MSD} is 128.

[2] Specifically, we aggregated the per-song listening counts corresponding artists such that we obtain the 'user-artist' listening matrix.

[3] i.e. a *different* model for audio embeddings than when training on MSD.

computed using PMI factorization from a total of $6,421$ different tags, which are a combination of manual and automatic annotations. Since our pre-trained models for audio and CF are at the track level, we compute artist embeddings by averaging over artist track embeddings.

In the remainder of this work, we refer to the model trained with the contrastive method with in-house data as C_{OWN} and the model trained with public data as C_{MSD}.

2.3 Evaluation Dataset

The ground truth for artist similarity is defined herein by the OLGA public dataset [22], containing artist similarity information collected from AllMusic. Our evaluations are therefore based on a *cultural* ground-truth, following [5].

We collected data from the MSD dataset for the original $17,646$ artists in OLGA. We obtained tag data from the MSD500 for $10,971$ (62%) artists, user interaction data from the Echonest Profile dataset for $15,389$ (87%) artists, and audio embeddings using MSD audio previews for 100% of the artists. [4]

We also create a subset of OLGA where *all* artists contain complete tags, user interaction, and audio information from MSD. We refer to this subset as OLGA Full Modality Coverage (FMC), which contains $9,474$ artists and it is also mapped to our internal dataset. The OLGA FMC subset is used to compare the results of multiple methods pre-trained on different and independent datasets.

2.4 Evaluation Conditions

In order to provide insights on the performance of the contrastive method, we conduct analyses under 3 different situations, varying the degree of availability of the different modalities in the evaluation data:

Raw evaluation dataset: In one condition, we compare the methods using all the artists in the OLGA dataset. In this case, we are interested to understand performance in a scenario of a real –uncontrolled– evaluation dataset, accounting for some organic imbalance of the availability of data in different modalities.

Full Modality Coverage: In another condition, we use the OLGA FMC subset where all artists contain CF, tags, and audio embeddings in both MSD and OWN datasets. In this case, we want to understand performance while factoring out the potential influence of one or another modality being only partially available in evaluation.

Systematic variation of modality coverage: We also perform multiple comparisons by grouping artists from OLGA depending on how many modalities are available. Here, we want to look at how much the contrastive method and the baselines are capable of doing cross-modality retrieval when using different modalities as input. In particular, we want to see whether or not they are capable of retrieving artists that have different modality information

available compared to the query artists. Therefore, in this part, we create 7 groups of artists –at random– of equal size with each group containing one, two, or three modalities (namely, CF, audio, tag, CF+audio, CF+tag, audio+tag, audio+CF+tag). We refer to these groups as 'Modality Groups'. It is important to highlight the artificiality of this setting. We are considering an extreme case only to evaluate cross-modality retrieval capabilities of the methods. We are not considering here the accuracy of these results since it is already evaluated in the other analyses.

2.5 Baseline multimodal approaches

For multi-modal baselines, we employ two conventional models: PCA, and Gaussian random projection [23, 24] (which we refer to as Rand). [5] For fitting these models, we consider artists who have access to all modalities. Their multimodal embeddings are concatenated and treated as a single feature vector. It yields a dimensionality of $2,063$ for the MSD dataset, and $2,528$ for the OWN dataset. We set the reduced dimensionality to 200, which is the same size as the embeddings of the contrastive model. If an artist has a missing modality in the prediction phase, we employ the global mean embedding of the missing modality. [6]

2.6 Metrics

Accuracy: We consider nDCG@200 to measure how accurate the retrieved artists are compared to the ground truth while taking into account the position in the ranking of the retrieved artists, a metric considered robust to missing relevance information [26]. [7]

Distribution: We also compute the Gini@200 index, measuring the distribution of the top 200 retrieved artists in each experimental condition across the whole set of artists. A lower value of Gini indicates that the recommendations across artists are more uniformly distributed –covering more artists retrieved– while a higher value of Gini indicates that the recommendations are focused on only the few same artists.

We compute the confidence interval using the bootstrap method [27] on the evaluation artist population. We report them in Figure 1 at 95% confidence level.

Expected Contrastive Loss: We propose an additional metric that we named Expected Contrastive Loss (ECL). We use this measure to analyze to what extent an artist is coherent with respect to their multimodal representations. From how we defined the loss in Section 2.1, a high loss value implies that the artist is relatively difficult to be distinguished from other artists. Once the training is reasonably progressed, we employ ECL to quantify how "coherent" the artist is with respect to their internal representations obtained from the different modalities, which is defined as: $ECL(i, u, v) = d_{ii}^{uv} - \mathbb{E}_{j\setminus i}[d_{ij}^{uv}]$,

---

[4] Note that we don't control for artist separation between MSD, OWN and OLGA. But even if some artists may be present in both train and test sets, the artist similarity information from OLGA is *only* used for evaluation, and is never used during the training of the single-modality embeddings nor the contrastive models on either MSD or OWN.

[5] For both algorithms, we employ the standard implementation provided from `scikit-learn` [25].

[6] This does not happen in FMC

[7] We focus on nDCG@200 in this work, as we experimentally observed high correlation with other retrieval metrics such as precision, recall, and R-Precision.

where i and j denote artist index, while u and v refer to the modality index. d_{ij}^{uv} means the cosine distance between artist i from modality u and artists j from modality v. Taking expectation over all the possible modality pairs leads to the final coherency measure for artist i:
$$ECL(i) = \mathbb{E}_{u,v \setminus u}[ECL(i, u, v)].$$

Clustering: We further analyze the multimodal embedding space of the contrastive model, by investigating how well the artist embeddings are clustered. The contrastive method essentially can be seen as a "supervised" clustering task, where we minimize the distance among "positive points" (i.e., multimodal embeddings from an artist) and maximize the distance between those to the "negative points" (i.e., embeddings belonging to the other artists). It implies that an artist will get a higher training loss when the embeddings are dispersed and overlapped with the embedding cluster of other artists, while the opposite cases will get lower values. The model will fit the multimodal embedding space such that the artist embeddings poorly clustered initially have more concentrated and distant clusters. While the contrastive learning implements this naturally by its loss function, there are other well-known measures for the validation of the clustering methods, such as *intra-cluster distance* (CD_{intra}) indicating how an artist embeddings are well clustered together, and *inter-cluster distance* (CD_{inter}) indicating how an artist-specific embedding cluster is far and distinct from others'. [8]

3. RESULTS

3.1 Performance comparison of contrastive method

We now look at the performance of the contrastive method when some modality information is missing in the evaluation dataset (using the raw OLGA dataset) and when all modalities are available for each artist (FMC subset). We also compare the performance of the contrastive method to the baseline methods and to single-modality embeddings.

3.1.1 Performance with incomplete modality information

Focusing on the different combinations of input modalities to the contrastive method, we can see in Table 1 that the highest nDCG result is obtained when combining all modalities as input. We therefore focus only on this model for the remainder of the work.

Figure 1a shows the results for all artists in OLGA. We can see that when using features from MSD, the contrastive method outperforms the baselines and the original embeddings in all the metrics. The contrastive method always gives a better Gini compared to the other methods –which means that the distribution of retrieved artists is more uniform– while outperforming the other models in nDCG. [9]

---

[8] we compute CD_{intra} as the mean cosine distance between multimodal embeddings of an artist to their centroid in the multimodal space of learned contrastive model. CD_{inter} is computed as the mean distance between the centroids of target artist and of all the other artists.

[9] OWN-OGLA is omitted since we observe a similar behaviour.

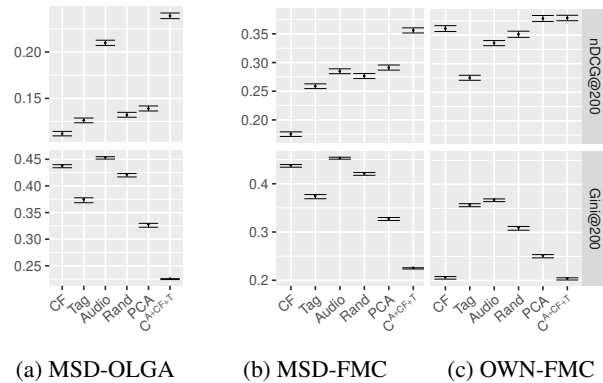(a) MSD-OLGA   (b) MSD-FMC   (c) OWN-FMC

Figure 1: Performance comparison between contrastive and other methods. Training with MSD (a and b) or with OWN (c), Evaluation on OLGA (a) or on FMC (b and c).

	OLGA		FMC	
	nDCG@200	Gini	nDCG@200	Gini
C^{A+CF+T}	**0.2387**	0.2264	**0.3560**	0.1666
C^{A+T}	0.2282	0.2035	0.3407	0.1559
C^{A+CF}	0.1381	0.3425	0.2319	0.1873
C^{CF+T}	0.1781	0.3467	0.3082	0.1917
C^{A}	0.2338	**0.1857**	0.3471	**0.1353**
C^{T}	0.1232	0.4939	0.2554	0.1745
C^{CF}	0.1381	0.3425	0.2319	0.1873

Table 1: Evaluation of the contrastive method trained with MSD data using all combinations of modalities for OLGA dataset and FMC subset.

3.1.2 Performance with complete modality information

When we look at the results with Full Modality Coverage (Figures 1b and 1c), the contrastive method outperforms the baselines and the pre-trained models in all the metrics both when trained with MSD data or with OWN data.

When looking at baseline performance between OLGA and FMC (Figures 1a and 1b), we can see that in the latter, baselines are relatively close to the best single-modality embeddings, but in the former (i.e. with incomplete modality information) their performance drops significantly lower than the best single-modality embeddings. This is something we do not observe with the contrastive method, which suggests that the baseline models are more limited in the capabilities of retrieving artists that miss some of the modalities from the query artist, while our contrastive method may be more robust to missing modality information. We investigate this further in Section 3.3.

### 3.2 Combining complementary modality information

If we focus only on the single-modality approaches, and MSD pre-training, Audio gives the best single-modality performance in both OLGA and FMC (Figure 1a and 1b). On the other hand, when pre-trained with OWN, CF is slightly better than Audio and Tag (Figure 1c). These results suggest that performance is highly dependent on the quality of the data used to pre-train the single-modality embeddings. Results from Figure 1b and 1c also sug-

	Audio	CF	Tag	Rand	PCA	C_{MSD}	C_{OWN}
Entropy	0.76	0.79	0.79	0.73	0.96	**1.86**	1.59

Table 2: Entropy of each model for Modality Groups. Higher values indicate better distributed retrieved artists.
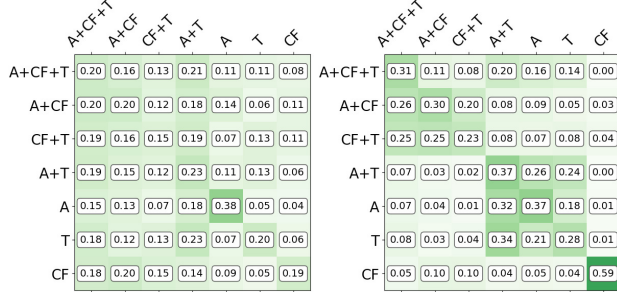
(a) Contrastive - MSD training (b) Contrastive - OWN training
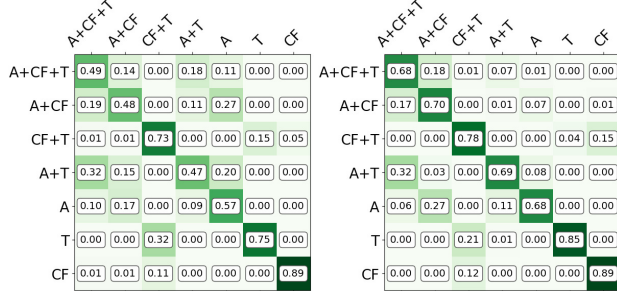
(c) PCA - MSD training (d) PCA - OWN training

Figure 2: Analysis of modality-group dependency ratio when restricting the information available for each group to one, two, or three modalities. Rows indicate the groups used to make the queries and the columns are the groups of retrieved artists. Darker green indicates a higher concentration of the retrieved artists in that cell. The color scale is normalized across all figures. Groups of artists are randomized, so an ideal situation is a homogeneous color in the full matrix.

gest that, whichever single-modality embedding is best, our contrastive method is able to successfully build on top of it and still gain in performance by combining complementary information from other embeddings.

3.3 Robustness to missing modality data

In this subsection, we further analyze how the contrastive method would be able to retrieve artists depending on the available information for the query artists and the candidates for retrieval. In Figure 2, we can see how artists are retrieved from each of the Modality Groups when only considering the top 5 results for each query artist. Typically we see that with the contrastive method, the same group used for query comprises between 15-38% of the retrieved artists. We see however an exception for the CF group which obtains a larger portion of the retrieved artists (59%) when using OWN data to train the models.

When we do a similar comparison for the PCA baseline method, we see in Figure 2 that there are higher percent-
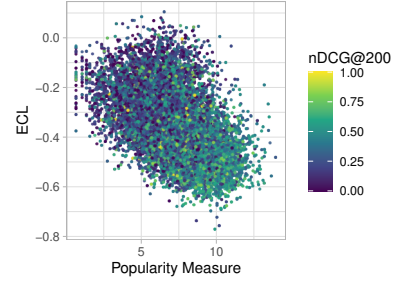

Figure 3: Scatter plot of artists based on the popularity proxy measure and the ECL. Each point represents an artist, where the color brightness represents the per-artist retrieval performance (nDCG@200). It is computed on the FMC subset with MSD data.
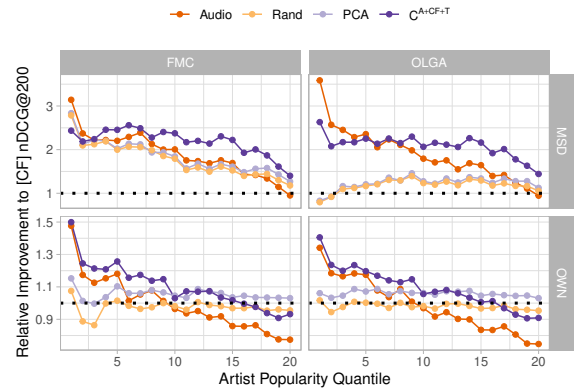

Figure 4: Relative retrieval improvement against CF modality. The x axis represents the grouped popularity quantile in 20 levels, meaning the first group includes artists whose popularity is under 5% percentile, while the top 5% popular artists belongs to the last group. The y axis is proportional improvement of nDCG@200 compared to the CF embedding model. The dotted horizontal line indicates the retrieval performance of CF modality. FMC and OLGA are evaluation datasets. MSD and OWN are training conditions.

ages in the diagonal of the matrix. This indicates that most of the retrieved artists are concentrated in the same modality group used to make the query. Therefore, these results highlight the difficulty for the PCA baseline method to retrieve artists beyond the query artist's modality.

In Table 2 we compare the entropy of each model for the Modality Groups. A higher entropy indicates that retrieved artists are better distributed across the different modality groups, i.e. that retrieval is less biased by the query modality –or more robust to partial modality data in the query. We can see that the contrastive model is more robust to missing modality data than the single-modality embeddings and the baseline approaches to combining modalities. This is true when trained with MSD or with OWN.

3.4 Effect of Popularity

Artist popularity may be a deterministic factor in artist retrieval, both for training and evaluation. Intuitively, we likely have more data about popular artists, which implies more multimodal data is available for training. At the same time, the scale of evaluation metric themselves can be inflated as more popular artists would have more ground truths (annotated as 'similar artists'). To confirm this, we compute a proxy measure for the artist popularity (POP) as POP(artist) $= \log(\#\text{listen} + 1)$,[10] and then further compare it to other training and evaluation measures.

Firstly, we compare POP with ECL and the retrieval performance. Figure 3 shows that there is correlation among POP, ECL, and nDCG. In particular, ECL has a negative correlation with nDCG. This is a desirable outcome as a model that minimizes the contrastive loss recommends "similar" artists even though such a model is not being explicitly shown artist-relatedness ground truth during training. Meanwhile, POP also correlates with nDCG, which demonstrates the confounding effect of popularity to the task itself.

Further, we investigate how multimodal models interact with artists with different popularities. One of the benefits of employing multiple modalities is the potential mitigation of the information void for "cold-start" artists from their music audio data. For MIR applications, audio is likely accessible even when some of the other modalities are not readily available. For instance, the CF modality is not available before artists' songs are consumed by the listeners. To confirm whether the audio and further multimodal embedding models would benefit less popular artists via multimodality, we divided the artists in 20 groups by popularity quantiles. For each group, we further compute the relative improvement of retrieval performance (nDCG) compared to the CF single modality model.

Figure 4 suggests that the original audio embedding achieves better performance for the less popular artists in all training and evaluation conditions. The contrastive model shows improvements for the majority of the groups compared to the audio, while it may have smaller or no improvement over audio in the least popular group for the MSD dataset. In the OWN dataset, a similar trend is observed where the contrastive model shows a small decline for the most popular groups compared to the original CF embeddings. The two baseline models indicate relatively flat results except in the case of the MSD-FMC subset, which implies that their prediction may be more reliant on the CF modality. For the MSD-FMC subset, both baselines follow similar trends to the audio and contrastive model.

3.5 Multimodal Embedding Space Analysis

We conduct a correlation study of multiple measures where, for each artist, we compute clustering measures and other key indicators such as contrastive loss ECL, retrieval performance (nDCG@200), and finally the popularity measure.
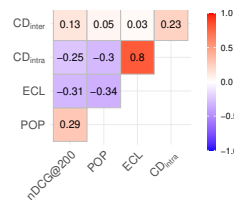

Figure 5: Correlation (Kendall's τ) among variables of interest. Each cell indicates value of τ between two associated variables. POP denotes the popularity measure.

In this way, we expect to obtain a better understanding of what contrastive learning achieves in terms of clustering of embeddings, and how they are connected to retrieval performance and popularity.

The result of the correlation study can be found in Figure 5. We see that the ECL is highly correlated to CD_{intra}, while almost independent to CD_{inter}. Notably, in terms of magnitude, all other measures (ECL, CD_{intra}, and POP) are relatively more correlated to nDCG compared to CD_{inter}, and also correlated to each other.[11]

These relations suggest that our contrastive learning method aims at producing an artist embedding space where the diverse modalities of an artist occupy a coherent region, but not necessarily a region that is unique to the artist. CD_{inter} shows lower correlation with most of the other measures, which confirms its relatively small connection to the contrastive learning and the artist retrieval downstream task. We hypothesize that this is because the maximization of CD_{inter} is constrained by the artist similarity inherent in the multimodal information and ultimately preserved. This is desirable if the ultimate goal is a representation that can measure artist similarity.

4. CONCLUSION AND FUTURE WORK

In this work, we propose a method based on contrastive learning to combine multiple artist modalities into a single representation. In an artist similarity task, we show our method yields clear improvements over other methods in terms of retrieval accuracy and coverage, and successfully combines complementary information from diverse modalities. In particular, we investigate retrieval bias towards the query's modality. Although our method exhibits a slight bias towards retrieving artists with similar modality to the query, we show it handles cross-modal retrieval better than other methods. Future work may be dedicated to further mitigate this bias. Additionally, we show that our method is particularly beneficial for less popular artists.

Our method appears to generate an artist representation space with high local coherence for intra-artist modalities, but at the cost of inter-artist separation. Depending on the final application, this is a property that could perhaps be managed by iterating on the contrastive learning method, for instance, by adapting the loss function or by adapting the size of the training sample batch as suggested in [28].

---

[10] #listen denotes the total listening count of the artist, computed from the MSD-Echonest Profile dataset.

[11] We focus on the magnitude, as the goal of this study is to investigate the degree to which some of the key indicators are associated with clustering quality measures in absolute manner

5. ACKNOWLEDGEMENT

6. REFERENCES

[1] D. P. Ellis, B. Whitman, A. Berenzweig, and S. Lawrence, "The quest for ground truth in musical artist similarity," in *ISMIR*, 2002.

[2] T. Pohle, D. Schnitzer, M. Schedl, P. Knees, and G. Widmer, "On rhythm and general music similarity." in *ISMIR*, 2009, pp. 525–530.

[3] M. Schedl, D. Hauger, and J. Urbano, "Harvesting microblogs for contextual music similarity estimation: a co-occurrence-based framework," *Multimedia Systems*, vol. 20, pp. 693–705, 2014.

[4] S. Oramas, M. Sordo, L. Espinosa-Anke, and X. Serra, "A semantic-based approach for artist similarity," in *ISMIR*, 2015.

[5] F. Korzeniowski, S. Oramas, and F. Gouyon, "Artist similarity for everyone: A graph neural network approach," *Transactions of the International Society for Music Information Retrieval*, vol. 5, no. 1, 2022.

[6] M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, and A. F. Ehmann, "Supervised and unsupervised learning of audio representations for music understanding," in *ISMIR*, 2022.

[7] P. Alonso-Jiménez, D. Bogdanov, J. Pons, and X. Serra, "Tensorflow audio models in Essentia," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 266–270.

[8] P. Alonso-Jiménez, X. Serra, and D. Bogdanov, "Music representation learning based on editorial metadata from Discogs," in *ISMIR*, 2022.

[9] S. Dieleman, P. Brakel, and B. Schrauwen, "Audio-based music classification with a pretrained convolutional network," in *ISMIR*, 2011, pp. 669–674.

[10] S. Oramas, F. Barbieri, O. Nieto Caballero, and X. Serra, "Multimodal deep learning for music genre classification," *Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.*, 2018.

[11] M. Won, S. Oramas, O. Nieto, F. Gouyon, and X. Serra, "Multimodal metric learning for tag-based music retrieval," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 591–595.

[12] A. Ferraro, "Music cold-start and long-tail recommendation: Bias in deep representations," in *Proceedings of the 13th ACM Conference on Recommender Systems*, 2019, p. 586–590. [Online]. Available: https://doi.org/10.1145/3298689.3347052

[13] S. Oramas, O. Nieto, F. Barbieri, and X. Serra, "Multi-label music genre classification from audio, text, and images using deep features," in *ISMIR*, 2017.

[14] A. Van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," *Advances in neural information processing systems*, vol. 26, 2013.

[15] Q. Huang, A. Jansen, J. Lee, R. Ganti, J. Y. Li, and D. P. Ellis, "Mulan: A joint embedding of music audio and natural language," in *ISMIR*, 2022.

[16] I. Manco, E. Benetos, E. Quinton, and G. Fazekas, "Contrastive audio-language learning for music," in *ISMIR*, 2022.

[17] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[18] A. Ferraro, X. Favory, K. Drossos, Y. Kim, and D. Bogdanov, "Enriched music representations with multiple cross-modal contrastive learning," *IEEE Signal Processing Letters*, vol. 28, pp. 733–737, 2021.

[19] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, "The million song dataset," in *ISMIR*, 2011.

[20] Y. Hu, Y. Koren, and C. Volinsky, "Collaborative filtering for implicit feedback datasets," in *Proceedings of the 8th IEEE International Conference on Data Mining (ICDM 2008)*, 2008, pp. 263–272.

[21] T. Head, MechCoder, G. Louppe, I. Shcherbatyi, fcharras, Z. Vinícius, cmmalone, C. Schröder, nel215, N. Campos, T. Young, S. Cereda, T. Fan, rene rex, K. K. Shi, J. Schwabedal, carlosdanielcsantos, Hvass-Labs, M. Pak, SoManyUsernamesTaken, F. Callaway, L. Estève, L. Besson, M. Cherti, K. Pfannschmidt, F. Linzberger, C. Cauet, A. Gut, A. Mueller, and A. Fabisch, "scikit-optimize/scikit-optimize: v0.5.2," Mar. 2018. [Online]. Available: https://doi.org/10.5281/zenodo.1207017

[22] F. Korzeniowski, S. Oramas, and F. Gouyon, "Artist similarity with graph neural networks," in *ISMIR*, 2021.

[23] E. Bingham and H. Mannila, "Random projection in dimensionality reduction: applications to image and text data," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, 2001, pp. 245–250.

[24] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984. [Online]. Available: https://doi.org/10.1090/conm/026/737400

[25] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[26] D. Valcarce, A. Bellogín, J. Parapar, and P. Castells, "Assessing ranking metrics in top-n recommendation," *Information Retrieval Journal*, vol. 23, pp. 411–448, 2020.

[27] B. Efron and R. Tibshirani, *An Introduction to the Bootstrap*. Springer, 1993. [Online]. Available: https://doi.org/10.1007/978-1-4899-4541-9

[28] C. Chen, J. Zhang, Y. Xu, L. Chen, J. Duan, Y. Chen, S. Tran, B. Zeng, and T. Chilimbi, "Why do we need large batchsizes in contrastive learning? a gradient-bias perspective," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 33 860–33 875.