Contrastive learning models based on editorial metadata from Discogs

ISMIR 2022 BENGALURU

Music Representation Learning Based on Editorial Metadata From Discogs

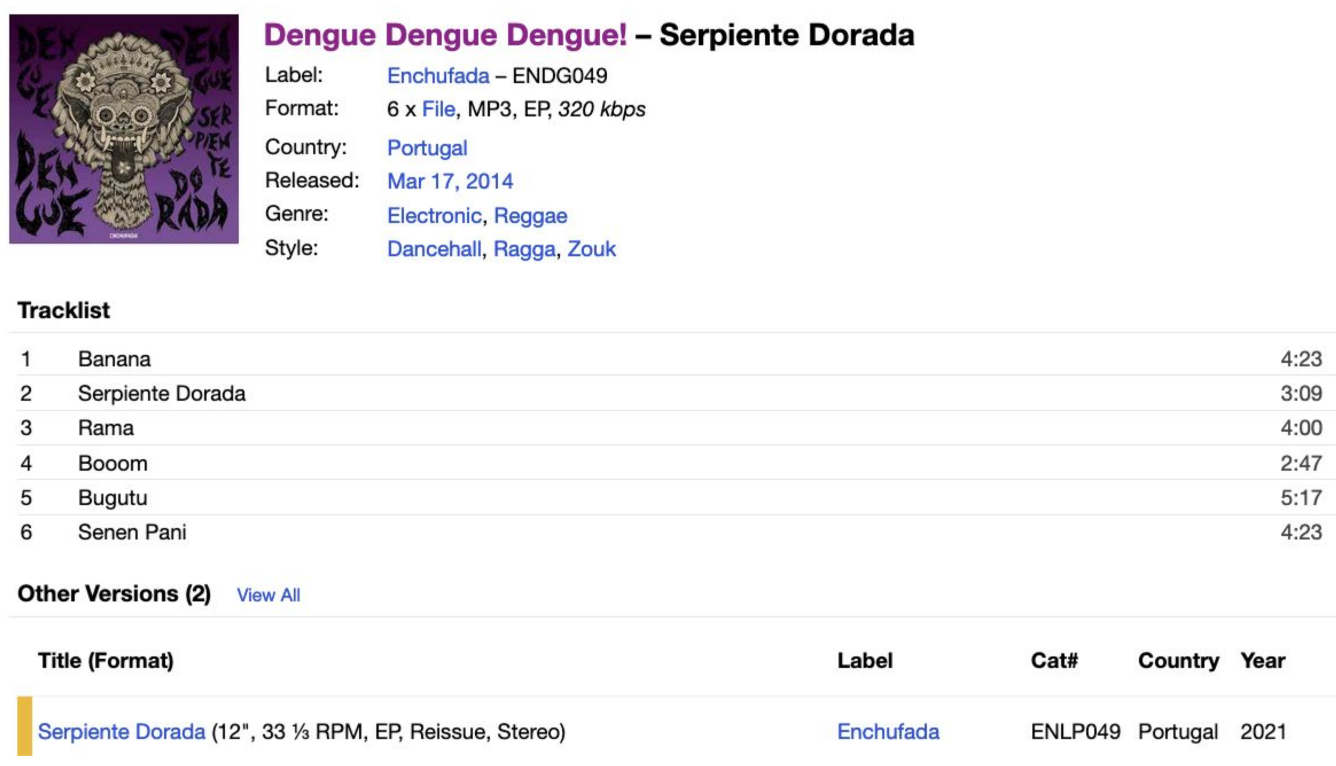
Pablo Alonso-Jiménez, Xavier Serra, Dmitry Bogdanov

Music Technology Group, Universitat Pompeu Fabra

Motivation

Descriptive tags are **difficult to obtain** and **noisy**. We need alternative ways of generating training targets for large music collections and suitable training approaches to develop **music representation** models.

Discogs metadata



Dengue Dengue Dengue! – Serpiente Dorada

Label: Enchufada – ENDG049
Format: 6 x File, MP3, EP, *320 kbps*
Country: Portugal
Released: Mar 17, 2014
Genre: Electronic, Reggae
Style: Dancehall, Ragga, Zouk

Tracklist

1	Banana	4:23
2	Serpiente Dorada	3:09
3	Rama	4:00
4	Booom	2:47
5	Bugutu	5:17
6	Senen Pani	4:23

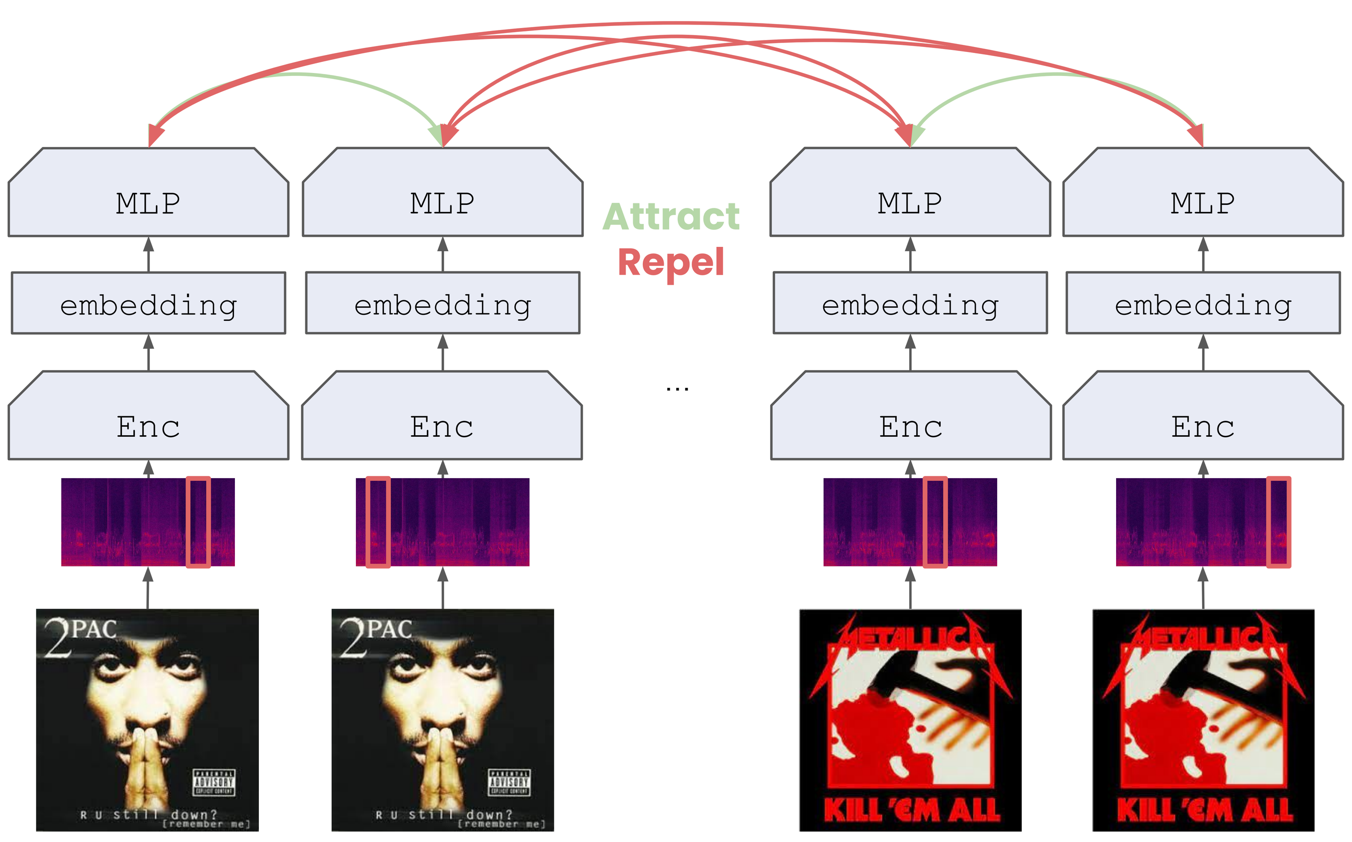
Other Versions (2) View All

Title (Format)	Label	Cat#	Country	Year
Serpiente Dorada (12", 33 ⅓ RPM, EP, Reissue, Stereo)	Enchufada	ENLP049	Portugal	2021

Discogs is an **extensive** community-maintained database of music metadata released under **CC0 license**.
We matched 3.3M tracks tracks to Discogs metadata:

- ❖ 2M releases (e.g., albums)
- ❖ 257K artists
- ❖ 142K record labels
- ❖ 400 style tags

Contrastive learning pre-training



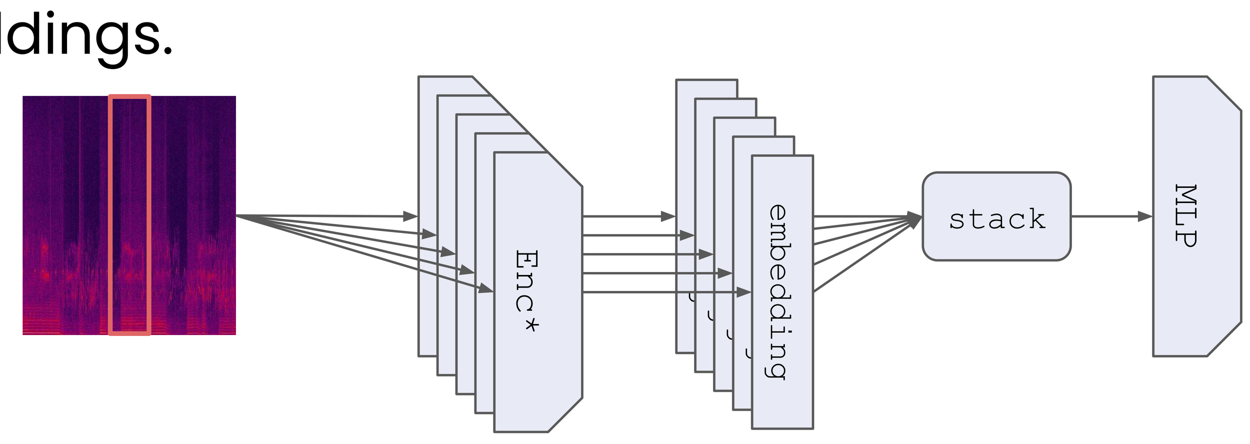
- ❖ We target editorial metadata associations similar to previous works on metric learning [1].
- ❖ We use a contrastive approach approach based on **COLA** [2] with an **EfficientNet** architecture [3].
- ❖ The considered models are:
  - **Track associations:** attract two fragments from the same song.
  - **Release associations:** attract two songs from the same release.
  - **Artist associations:** attract two songs from the same artist.
  - **Label associations:** attract two songs from the same label.
  - **Multi-task:** Jointly learn the track and artist objectives.

Downstream evaluation

We consider several music classification datasets.

Dataset	# tracks	Classes	Type
MTG-Jamendo Genre	55,215 ft	87	multi-label
MTG-Jamendo Instrument	25,135 ft	40	multi-label
MTG-Jamendo Moods	18,486 ft	56	multi-label
MTG-Jamendo Top 50 tags	54,380 ft	50	multi-label
MagnaTagATune	25,860 exc	50	multi-label
FMA small	8,000 exc	8	single label

We evaluate the pre-trained models as **frozen embedding extractors** by training MLP classifiers.
We also considered training classifiers on **stacks** of embeddings to **assess the complementarity** of the embeddings.



The classifiers are evaluated with the **ROC-AUC** and **PR-AUC** metrics. Additionally we report the performance of SOTA model from the literature and embeddings from random weights, a model trained on style tags, and the VGGish model [4].

	Genre		Instrument		Mood		Top50		MTAT		FMA
	ROC	PR	ROC	PR	ROC	PR	ROC	PR	ROC	PR	Acc.
Lileonardo	-	-	-	-	**77.5**	**15.1**	-	-	-	-	-
Harmoic CNN	-	-	-	-	-	-	83.2	29.8	*91.3	*45.9	-
MusiCNN	-	-	-	-	-	-	-	-	90.7	38.4	-
MuLaP	85.9	-	76.8	-	76.1	-	82.8	-	*89.3	*40.2	**61.1**
CALM	-	-	-	-	-	-	-	-	**91.5**	**41.4**	-
Random weights	50.7	3.1	49.9	6.4	50.4	3.4	48.3	6.5	50.0	5.3	12.5
Style tags	87.7	19.9	77.6	19.8	75.6	13.6	83.1	29.7	90.2	37.4	59.1
VGGish	86.3	17.2	**77.8**	**20.2**	76.3	14.1	83.2	28.2	90.2	37.2	53.0
Track associations	86.3	18.0	69.9	16.7	74.0	12.8	82.9	29.4	89.7	36.4	58.9
Release associations	86.9	18.9	71.9	17.2	72.8	11.7	83.2	29.8	90.3	37.1	60.9
Artist associations	**87.7**	**20.3**	69.7	16.9	76.3	14.3	**83.6**	**30.6**	90.7	38.0	59.1
Label associations	87.0	19.4	75.0	18.2	74.8	12.8	83.1	29.9	88.7	34.2	59.5
Stack	86.9	19.4	74.7	18.8	74.3	13.0	83.4	30.0	90.8	38.6	59.8
Multi-task	87.2	19.9	70.5	17.2	76.1	14.4	83.5	30.3	90.8	37.8	60.0

Conclusions

- ❖ **Artist associations** produce the **best embeddings**.
- ❖ The features are **complementary** and stacking them is beneficial in some cases.
- ❖ Some metadata-based embeddings are **superior to** models obtained from **classification**.

Proposed models are publicly available:
https://essentia.upf.edu/models.html

Contact: pablo.alonso@upf.edu @pablo__alonso

[1] Park, J. Lee, J.W. Ha, and J. Nam. "Representation learning of music using artist label.," *in Proceedings of the 19th International Society for Music Information Retrieval Conference (ISMIR)*. 2018.
[2] Saeed, Aaqib, Grangier, David, and Zeghidour, Neil. "Contrastive learning of general-purpose audio representations." *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2021.
[3] Tan, Mingxing, and Quoc Le. "Efficientnet: Rethinking model scaling for convolutional neural networks." *International Conference on Machine Learning (ICML)*. 2019.
[4] Hershey, Shawn, et al. "CNN architectures for large-scale audio classification." *International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2017.

This research was carried out under the project Musical AI - PID2019-111403GB-I00/AEI/10.13039/501100011033, funded by the Spanish Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación.

upf. MTG Music Technology Group

ESSENTIA MUSICAL AI_

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA, INNOVACIÓN Y UNIVERSIDADES

AGENCIA ESTATAL DE INVESTIGACIÓN