Audio-based metric learning for artist disambiguation in large music catalogs

Jimena Royo-Letelier, **Romain Hennequin**, Viet-Anh Tran, Manuel Moussallam

Deep learning workshop: From theory to applications

September 5th, 2018 - Rennes

DEEZER

Large catalogs

- Music streaming/automated music recommendation became central for music consumption

- **Ever evolving**, **large music databases (catalogs)**
    - millions of artists
    - tens of millions of tracks
    - tens of thousands new tracks ingested every day

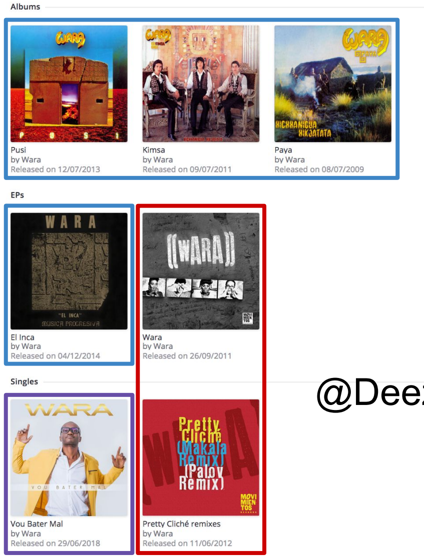
Global identifiers

Dealing with such large database needs **global identifiers** :

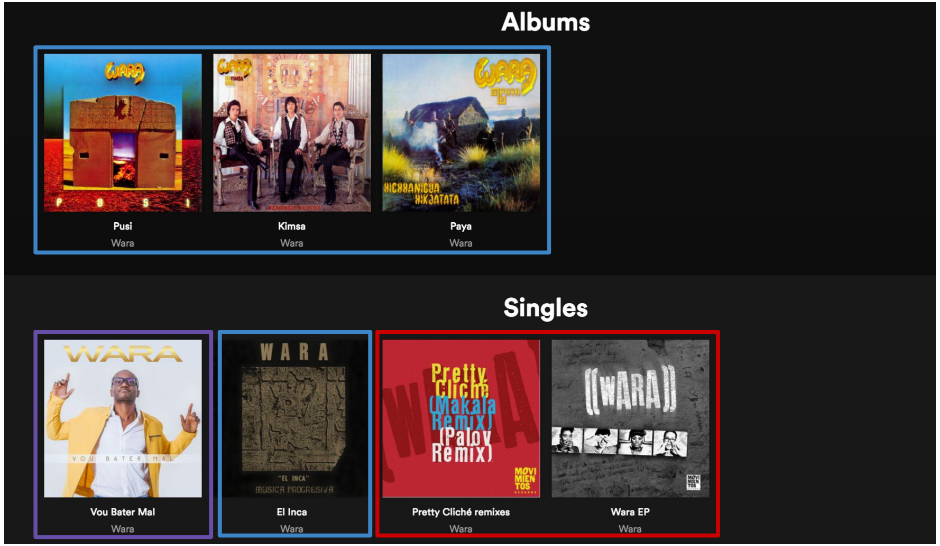
- **ISRC** describes recordings (supposedly) uniquely
- **UPC** describes albums (supposedly) uniquely
- For artists… huh?

→ Lack of an **universal** and **reliable** mean to identify music artists

And then ...



@Deezer



@some competitor *(Visited on 06/07/2018)*

Wara music group page

Artist name is used as an identifier:

→ Confusing / unpleasant catalog exploration for end users

→ May induce bad quality suggestions when notion of artist used for recommendation

Automatic artists *disambiguation*

- Given a group of recordings associated to the same artist name, identify actual artists.
- May be solved using **metadata** (release dates, titles language, record labels, etc.) but these are not always available nor reliable.

  → Rely (at least partially) on **audio** content

- A totally optimized system should use all possible sources

What is an *artist* ?

Loosely defined notion:

- Usually the main performer/band that plays a song
- Sometimes there may be several ones
- Can be sometimes the composer (mainly in classical music)
- Can be the music producer (mainly in electronic music)

=> what the provider (record label) want to put in it…

What is an *artist ?*

Can be ambiguous at the audio level, but there should be similarity :

- Singer timber
- Instrumentation
- Characteristic instrument licks
- Lyrical content
- Production
- …

Very diverse characteristics of audio are involved.

Automatic artists disambiguation

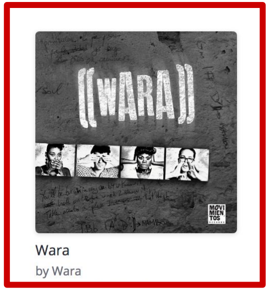
Difficult problem because:

- **Variability** across tracks/albums from a **same artist**
- Acoustics **similarities** (genre/mood/orchestration) between **different artists**

Contemporary latin music

Bolivian rock- folk band

Afro pop singer



Wara_A

Wara_B

Wara_C

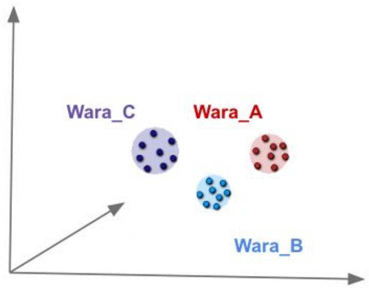
Audio clustering problem

- In Music Information Retrieval (MIR) literature:
  Usually addressed as classification of **already known artists** [Berenzweig et. al 2003], [Eghbal-Zadeh et al. 2015]... → not a real case scenario (new artists are added every day)

- Ideal system

  → **distinguish and group** recordings from the same artist

  → **unbalanced** clustering problem with **unknown number of clusters**

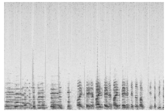


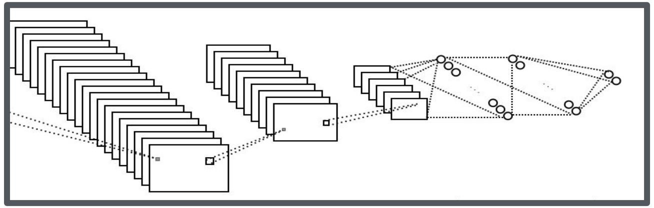
- Try to learn ***tailored representations*** from audio ***for clustering task***

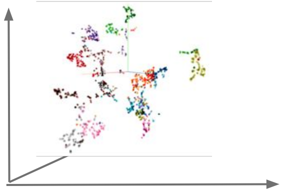
  → representations of tracks from the same (resp. different) artists must be close (resp. far) in space

Representation Learning

- Representation space (**embedding**) creation by directly learning a **parametric map** from input to representation







x = low level feature

parametric map f

(e.g.: convolutional/recurrent neural network)

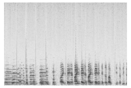
embedding $f(x)$

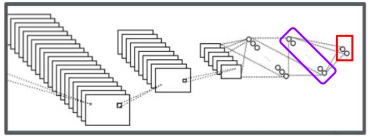
$\mathbb{R}^d$, d small

- Learn a low dimensional space that represents high-level characteristics of audio content in which **proximity** may be interpreted as **some kind of similarity**.

Representation Learning

Intermediary classifier activations





f = conv. neural network

$f(x)$ = last hidden layer activation of ***classifier*** network

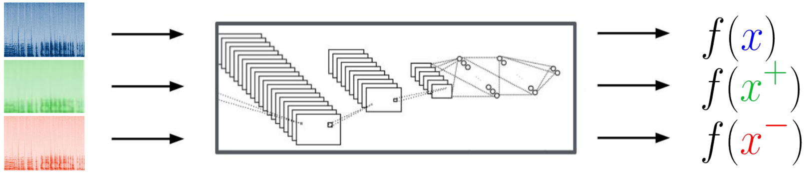
train system with cross-entropy loss

Done for artists classification in [Park et al. 2017], used as baseline.

=> provides a representation optimized for classification (linear separation) not clustering.

Representation Learning

Metric Learning



Impose metric through distance of positive/negative pairs: ***triplet loss*** (*used for face detection [Schroff et al. 2015]*)

$$\mathcal{L}(\mathcal{X}) = \Big| \|f(x) - f(x^+))\|_2^2 - \|f(x) - f(x^-))\|_2^2 + \alpha \Big|_+$$

sampled triplets: $\mathcal{X} = (x, x^+, x^-)$

positive example
same artist

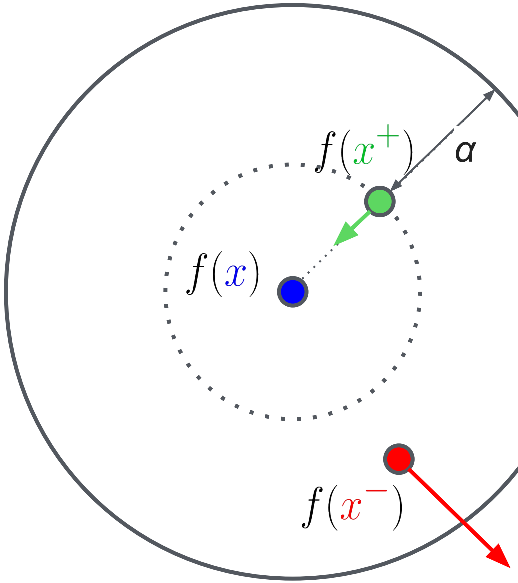
negative example
different artist

DEEZER

Representation Learning

Metric Learning

$$\mathcal{L}(\mathcal{X}) = \left| \|f(x) - f(x^+))\|_2^2 - \|f(x) - f(x^-))\|_2^2 + \alpha \right|_+$$

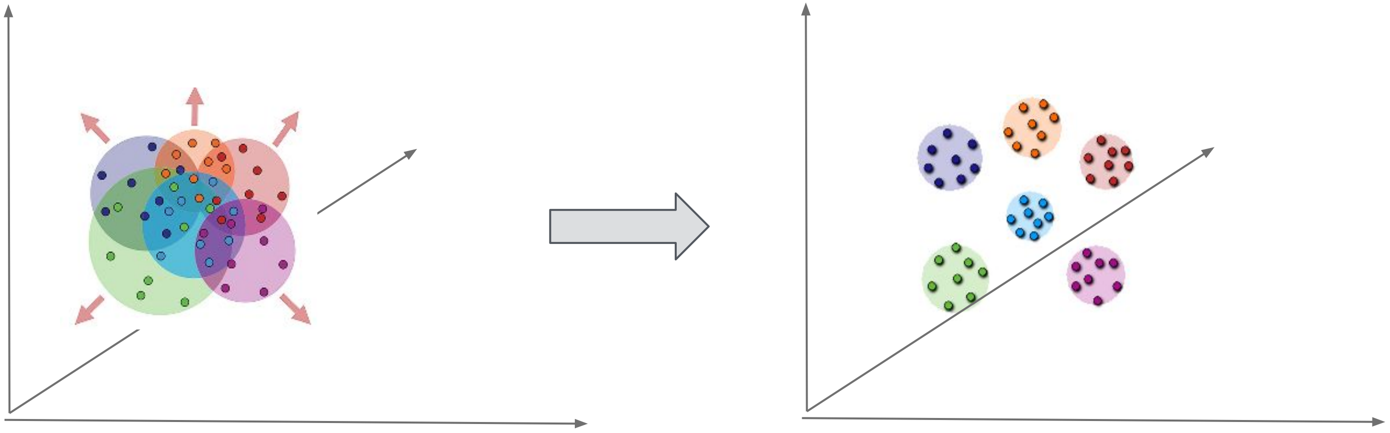


DEEZER

Representation Learning

Metric Learning

$$\mathcal{L}(\mathcal{X}) = \left| \|f(x) - f(x^{+}))\|_2^2 - \|f(x) - f(x^{-}))\|_2^2 + \alpha \right|_{+}$$



Representation Learning

Metric Learning

- Objective function designed to get a good representation for **clustering**.
- Dynamic sampling for learning: favor hard vs semi-hard triplet.
- Possibility to add **side information** (e.g. tags, usage data) to guide learning.
  - Take advantage of music hierarchical organisation for smart sampling:

    → favor positive pairs from **different albums**

    → favor negative pairs from **same genre**
- Issues: dynamic sampling may result in instabilities and mode collapse.

Evaluation

Not manually checked training dataset:

- Several thousands artists
- Used for training embedding map

Homonym artists dataset used as test dataset:

- 122 groups of 2 to 4 homonym artists.
- Clustering of albums of group of artists with the same name.

Agglomerative hierarchical clustering:

- No need of previous knowledge about number of different artists
- Cross-validation to set flat clusters threshold
- Performance evaluated with Rand index (probability that two clusterings agree on a randomly chosen pair) adjusted for chance.

Disambiguating Music Artists at Scale with Audio Metric Learning
J. Royo-Letelier, R. Hennequin, V-A. Tran, M. Moussallam
To be published at Ismir 2018

Systems Performances Evaluation

Table 1. Mean ARI performances of the metric learning and classification embedding systems on the artist clustering task (5-fold cross-validation) for *Balanced* experiment.

	25	50	100	200	400	600
CL	0.32	0.32	0.35	0.47	0.54	**0.60**
ML	0.45	0.56	0.52	0.56	**0.60**	0.58

=> Metric Learning performs better with less artists in the training set.

Disambiguating Music Artists at Scale with Audio Metric Learning
J. Royo-Letelier, R. Hennequin, V-A. Tran, M. Moussallam
To be published at Ismir 2018

Systems Performances Evaluation

Table 2. Mean ARI performances of the metric learning embedding systems on the artist clustering task (5-fold cross-validation) *unbalanced* (left) and *Side information* (right) experiments.

CL A	CL B	ML 1079	ML 3023	ML 3023 genre
0.54	0.47	**0.62**	0.55	**0.64**

=> Dynamic sampling compensates for data unbalance

Disambiguating Music Artists at Scale with Audio Metric Learning
J. Royo-Letelier, R. Hennequin, V-A. Tran, M. Moussallam
To be published at Ismir 2018

DEEZER

Systems Performances Evaluation

Table 2. Mean ARI performances of the metric learning embedding systems on the artist clustering task (5-fold cross-validation) *unbalanced* (left) and *Side information* (right) experiments.

CL A	CL B	ML 1079	ML 3023	ML 3023 genre
0.54	0.47	**0.62**	0.55	**0.64**

=> Incorporation of side information provides better representation for artist discrimination

Disambiguating Music Artists at Scale with Audio Metric Learning
J. Royo-Letelier, R. Hennequin, V-A. Tran, M. Moussallam
To be published at Ismir 2018

DEEZER

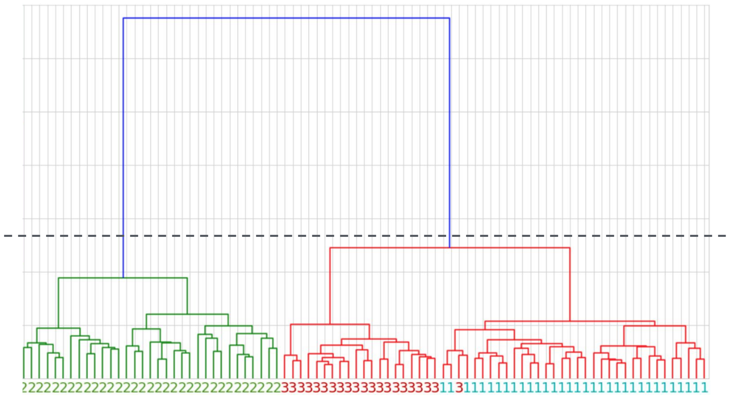
Qualitative clustering results

Eclipse_1

Eclipse_2

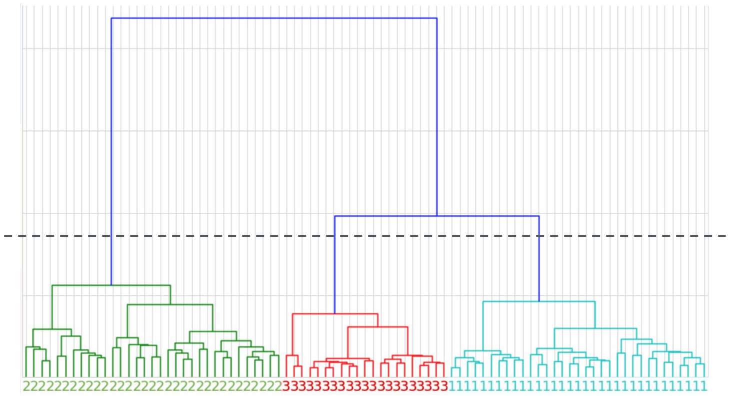
Eclipse_3

ML



Clustering threshold

ML + genre sampling



DEEZER

Take away

- Artist disambiguation from audio:
  - → useful task in a real life scenario
  - → improvement in the quality of large sized catalogs
- Addressed as representation learning + unsupervised clustering task
- Still work to do !

Next

- Leverage other information for guiding training (album covers, listening data, etc.)
- Scale evaluation
- Artist ambiguity is not only about homonymy but also about synonymy (different name for a same artist).
- Best of both representation systems: learn ***jointly*** metric learning *and* classification system
  - → regularization for ML
  - → sampling strategies for CL

Thanks !

[Berenzweig et. al 2003] A. Berenzweig, D. P. W. Ellis, and S. Lawrence. Anchor space for classification and similarity measurement of music. In International Conference on Multimedia and Expo (ICME), volume 1, pages I–29–32, 2003.

[Eghbal-Zadeh et al. 2015] Hamid Eghbal-Zadeh, Bernhard Lehner, Markus Schedl, and Gerhard Widmer. I-vectors for timbrebased music similarity and music artist classification. In ISMIR, pages 554–560, 2015.

[Park et al. 2017] Jiyoung Park, Jongpil Lee, Jangyeon Park, Jung-Woo Ha, and Juhan Nam. Representation learning of music using artist labels. CoRR, abs/1710.06648, 2017.

[Schroff et al. 2015] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In CVPR , pages 815–823. IEEE Computer Society, 2015.