Multi-Task Music Representation Learning from Multi-Label Embeddings

Alexander Schindler

Center for Digital Safety and Security

Austrian Institute of Technology

Vienna, Austria

alexander.schindler@ait.ac.at

Peter Knees

Faculty of Informatics

TU Wien

Vienna, Austria

peter.knees@tuwien.ac.at

MOTIVATION

- **Learn a content-based representation** for **similarity retrieval** of (Western) **music**

- **Traditional approach:** train Machine Learning model (most recently, Deep Neural Networks) on audio spectrogram input to learn
  - Genre labels or other tags (classification)
  - Rating/listening data from users (regression)

- **Problem: Ground Truth**
  - Where to get large quantities of labeled content in high quality?
  - Vocabulary: Which task categories are captured?
  - How to incorporate similarity of labels?

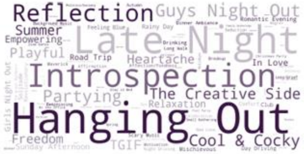
IN A NUTSHELL

- **Contribution 1: New Label Assignments for the Million Song Dataset**
  - Dataset of expert-level annotations in multiple categories; available to community
- **Contribution 2: A Novel Approach to Music Representation Learning**
  - Triplet network trained on similarity based on latent label topics
- **Contribution 3: Multi-Task Learning and Evaluation**
  - Our method improves precision up to factor 2.2 when learning across multiple tasks

- **Conclusions**
  - It makes a lot of sense and works very well
- **Future Work**
  - Applicable to digital libraries of historic and non-Western music

New Label Assignments for the Million Song Dataset (MSD)

Contribution 1



NEW MSD TAG-SET COLLECTIONS

- Million Song Dataset (MSD)
  - Currently largest music dataset
  - 1M tracks + metadata + pre-extracted features (Echonest)
- Issues
  - Harness Echonest Features (only officially provided content)
    - *Capturing the temporal domain in echonest features for improved classification effectiveness. Alexander Schindler and Andreas Rauber.*
  - Missing Ground-Truth Label Assignments
    - 2011 - Lastfm-Tags, original MSD contribution
      - User generated tags, noisy
    - *2012 - Facilitating comprehensive benchmarking experiments on the million song dataset. Alexander Schindler, Rudolf Mayer, and Andreas Rauber.*
      - *Genres, Multi-Class, custom balancing*
    - *2015 - Improving Genre Annotations for the Million Song Dataset. Hendrik Schreiber.*
      - *Aggregation of multiple label assignments, improved balancing*

NEW MSD TAG-SET COLLECTIONS

- **New Label Assignments**
  - Tag-Sets for:
    - **Genres, Styles, Moods, Themes**
  - Multi-Label assignments
  - + Expert annotated (All Music Guide)
  - + Closed vocabulary / Taxonomy
  - - Weakly labelled (per album)

	Genres	Styles	Moods	Themes
Unique Tags	21	939	286	166
Tag Combinations	688	13.589	22.577	7.322
Labelled Albums	75.339	52.304	32.148	19.375
Labelled Tracks	504.502	364.326	229.510	145.555

RANDOM ALLMUSIC BANDPAGE... FROM DUBLIN

ALLMUSIC

Search

U2

Biography by Stephen Thomas Erlewine

+ Follow Artist

Trafficking in big ideas and big sounds, a band that operated on a grander scale than any other from the '80s and attracted legions of devoted fans.

Artist Information

Active	1970s - 2010s
Formed	1976 in Dublin, Ireland
Group Members	Adam Clayton Bono Larry Mullen, Jr. The Edge

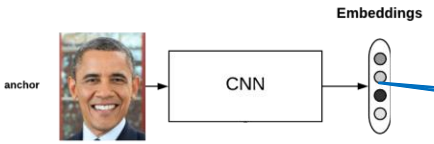
NEW MSD TAG-SET COLLECTIONS

A NOVEL APPROACH TO MUSIC REPRESENTATION LEARNING

Contribution 2

TRIPLET NETWORKS

- How do triplet networks work



Triplet Loss

- Margin optimization

$$\sum_i^N \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+$$

L_2 (anchor,positive) L_2 (anchor,negative) Margin

Triplet Loss and Online Triplet Mining in TensorFlow
https://omoindrot.github.io/triplet-loss

Facenet: A unified embedding for face recognition and clustering. *Schroff, Florian, Dmitry Kalenichenko, and James Philbin. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.*

AUSTRIAN INSTITUTE OF TECHNOLOGY

TRIPLET SELECTION

- FaceNet - original approach
  - **Facenet: A unified embedding for face recognition and clustering**. *Schroff, Florian, Dmitry Kalenichenko, and James Philbin. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015.*

- Definition of Similarity

 is_similar 

 ~~is_similar~~ 

- Online Triplet Selection

	Person A	Person B	Person C	Person D	Person E	Person F	Person G	Person H	Person I	Person J	Person K	Person J
Person A		0	0	0	1	0	1	0	0	1	0	0

	Anchor
1	Positive Example
0	Negative Example

Positive Pairs:

<A,E>
<A,G>
<A,J>

Negative Pairs:

<A,B>
<A,C>
<A,D>
<A,F>
<A,H>
<A,I>
<A,J>

TRIPLET SELECTION

- Similarity by Artist Identity
  - **Representation learning of music using artist labels.**
    J. Park, J. Lee, J. Park, J.-W. Ha, and J. Nam, in 19th International Society for Music Information Retrieval Conference (ISMIR 2018), 2018.

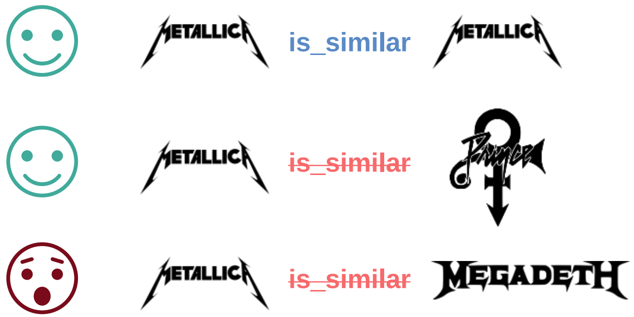
is_similar

~~is_similar~~

	Artist A	Artist B	Artist C	Artist D	Artist E	Artist F	Artist G	Artist H	Artist I	Artist J	Artist K	Artist J
Artist A		0	0	0	1	0	1	0	0	1	0	0
Artist B	0		0	1	0	0	1	0	0	0	1	0
Artist C	0	0		0	0	0	0	0	1	0	0	0
Artist D	0	1	0		1	0	0	0	0	0	1	0
Artist E	1	0	0	1		0	1	0	1	0	0	0
Artist F	0	0	0	0	0		0	0	0	0	0	0
Artist G	1	1	0	0	1	0		1	0	0	0	1
Artist H	0	0	0	0	0	0	1		0	0	0	0
Artist I	0	0	1	0	1	0	0	0		0	0	0
Artist J	1	0	0	0	0	0	0	0	0		0	0
Artist K	0	1	0	1	0	0	0	0	0	0		0
Artist L	0	0	0	0	0	0	1	0	0	0	0	

	Anchor
1	Positive Example
0	Negative Example

ISSUES



- Similarity by Artist Identity
  - **Problem**
    - Not: **Missing positive examples**
    - But: Selection of **inferior negative examples**
  - **Consequence**
    - Model focuses on **features to distinguish similar artists**
      - Smallest common denominator between similar artists
      - = Intention of original FaceNet approach (Re-Identification)
    - Model **fails to** learn features to **capture general similarity**
      - Instruments, harmonics, rhythms, modes, keys, moods, themes, etc.

MOTIVATION

- How to assess Track-Similarity from Multi-Label Tag-Sets?

- Tag-Relatedness measures

  - Jaccard Index $$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

  - Dice Coefficient $$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

JACCARD INDEX

Track_A = Rap

Track_B = Rap, Gangsta Rap

$|A \cap B|$ = Rap

$|A \cup B|$ = Rap, Gangsta Rap

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{Rap}}{\text{Rap, Gangsta Rap}} = 0.5$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{Rap}}{\text{Rap, }\textbf{Heavy Metal}} = 0.5$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{\text{Rap}}{\text{Rap, }\textbf{Children Music}} = 0.5$$

JACCARD INDEX

$Track_A$ = East Coast Rap

$Track_B$ = West Coast Rap

$|A \cap B|$ = []

$|A \cup B|$ = East Coast Rap, West Coast Rap

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{[]}{\text{East Coast Rap, West Coast Rap}} = 0$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{[]}{\text{Classic, Techno}} = 0$$

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{[]}{\text{Happy, Sad}} = 0$$

TAG-RELATEDNESS MEASURE

- **Goal:** Define better **Tag-Relatedness Measure**
  - Take Tag-relationships into account

LATENT SEMANTIC INDEXING (LSI)

- Classic IR approach to discover latent topics in texts (Deerwester et al., 1990)
- Based on Singular Value Decomposition (SVD)
- Topic importance ordered according to Eigenvalues
  - Truncated SVD removes less relevant topics (noise), increases robustness
- Shown effective to deal with polysemy and homonymy

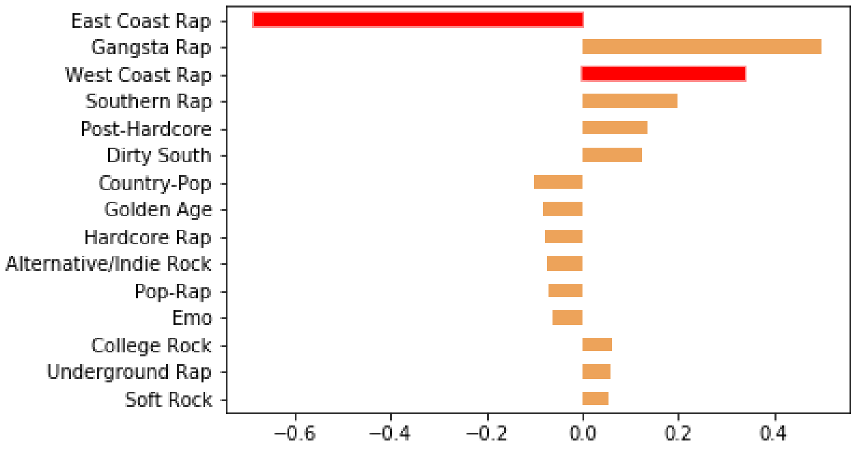
- **Regarding our approach**
  - provides a **more robust** (=less sparse) **tag-similarity function** via topics

$$\overset{X}{\begin{array}{cc} & \begin{array}{ccccc} & & (\mathbf{d}_j) & & \\ & & \downarrow & & \end{array} \\ (\mathbf{t}_i^T) \rightarrow & \begin{bmatrix} x_{1,1} & \dots & x_{1,j} & \dots & x_{1,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i,1} & \dots & x_{i,j} & \dots & x_{i,n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{m,1} & \dots & x_{m,j} & \dots & x_{m,n} \end{bmatrix} \end{array}} = (\hat{\mathbf{t}}_i^T) \rightarrow \overset{U}{\begin{bmatrix} \begin{bmatrix} \\ \mathbf{u}_1 \\ \\ \end{bmatrix} \dots \begin{bmatrix} \\ \mathbf{u}_l \\ \\ \end{bmatrix} \end{bmatrix}} \cdot \overset{\Sigma}{\begin{bmatrix} \sigma_1 & \dots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \dots & \sigma_l \end{bmatrix}} \cdot \overset{V^T}{\begin{array}{c} (\hat{\mathbf{d}}_j) \\ \downarrow \\ \begin{bmatrix} [& \mathbf{v}_1 &] \\ & \vdots & \\ [& \mathbf{v}_l &] \end{bmatrix} \end{array}}$$

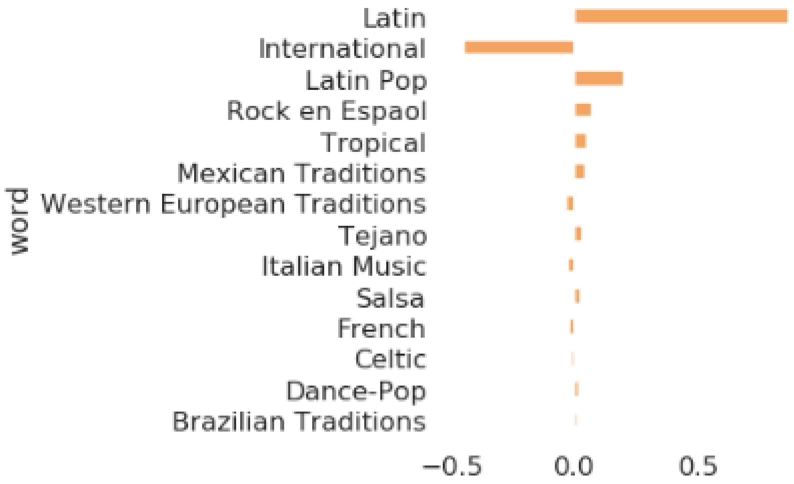
https://en.wikipedia.org/wiki/Latent_semantic_analysis

LSI TOPICS (EXAMPLES, STYLES)

- Rap

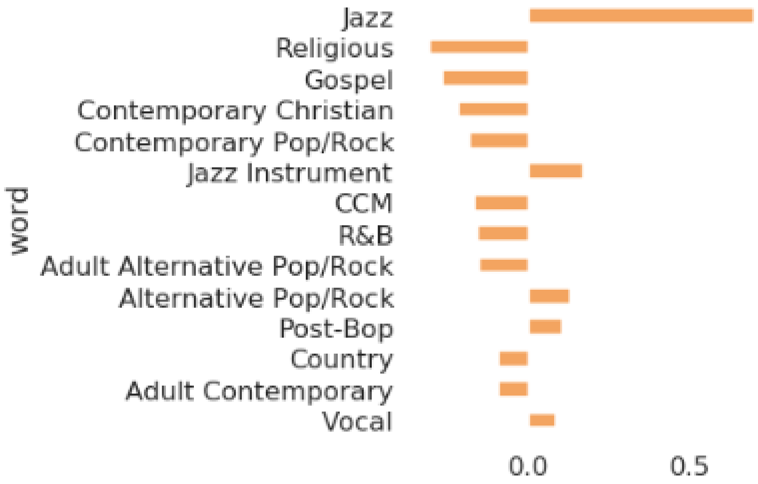


- Latin

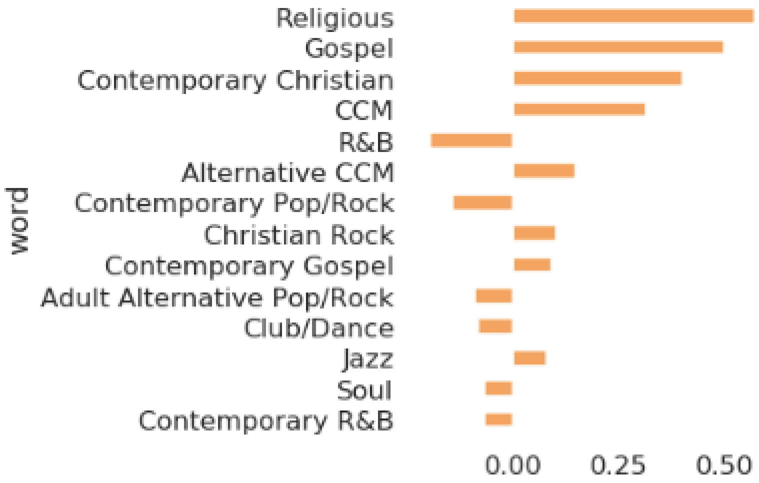


LSI TOPICS (EXAMPLES, STYLES)

- Jazz



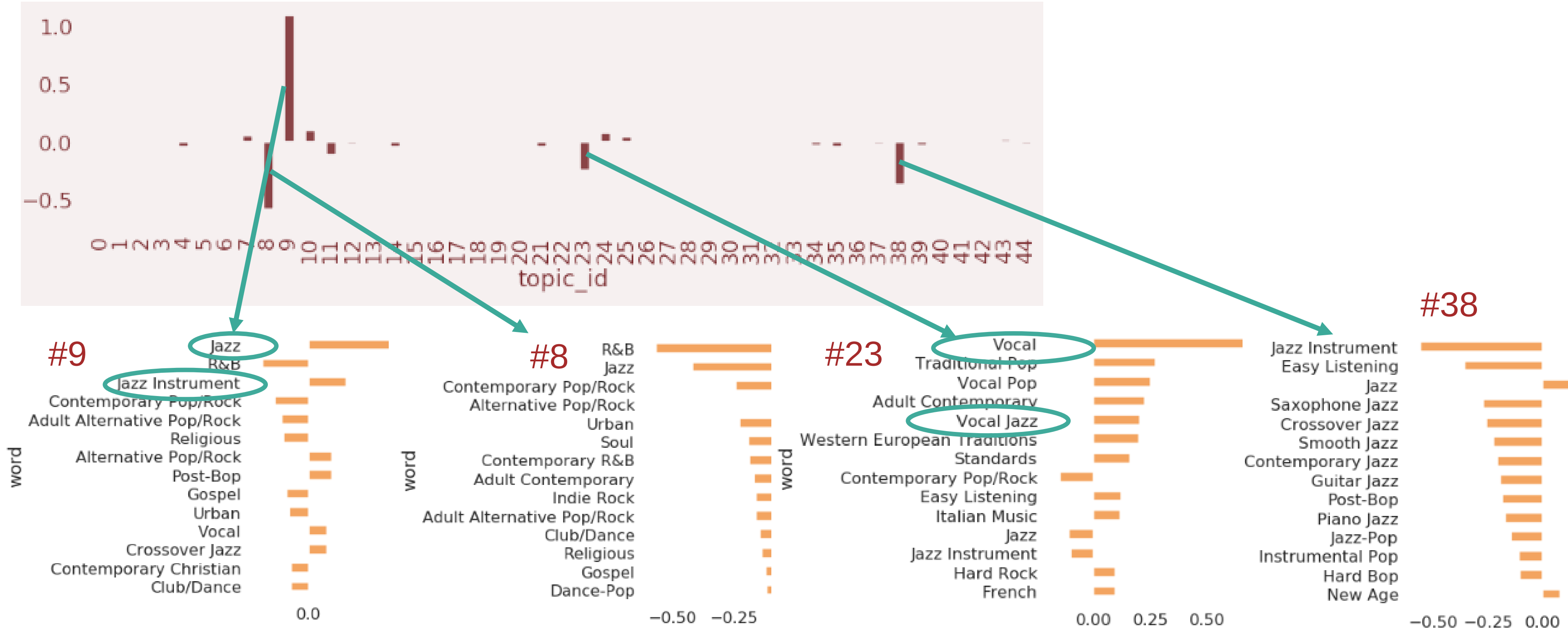
- Christian Music



LSI VECTORS (EXAMPLES, STYLES)

Miles Davis - Blue In Green (Blue Moods)

Annotated Styles: **Cool, Hard Bop, Jazz Instrument, Trumpet Jazz**



LSI-BASED ONLINE TRIPLET SELECTION

- FaceNet ➔ Binary relations
- Our approach
  - Pairwise Cosine-Distance of LSI Vectors
  - ➔ continuous similarity (range [0,1])
- Create Filter-Mask
  - Positive Examples: $\cos(LSI_1^{ts}, LSI_2^{ts})$ > *0.8*
  - Negative Examples: $\cos(LSI_1^{ts}, LSI_2^{ts})$ < *0.5*
- Select Triplets

	Track A	Track B	Track C	Track D	Track E	Track F	Track G	Track H	Track I	Track J	Track K	Track J
Track A		0	1	1	1	0	1	0	0	1	0	0
Track B	0		0	1	1	1	1	0	1		1	0
Track C	1	0		1	0	1	1	0	1	1	1	0
Track D	1	1	1		1		0	1	0	0	1	0
Track E	1	1	0	1		0	1	1	1	0	0	
Track F	0	1	1		0		1	1		0	0	0
Track G	1	1	1	0	1	1		1	0	1	0	1
Track H	0	0	0	1	1	1	1		0	0	0	0
Track I	0	1	1	0	1		0	0		0	1	0
Track J	1	1	1	0	0	0	1	0	0		0	0
Track K	0	1	1	1	0	0	0	0	1	0		0
Track L	0	0	0	0		0	1	0	0	0	0	

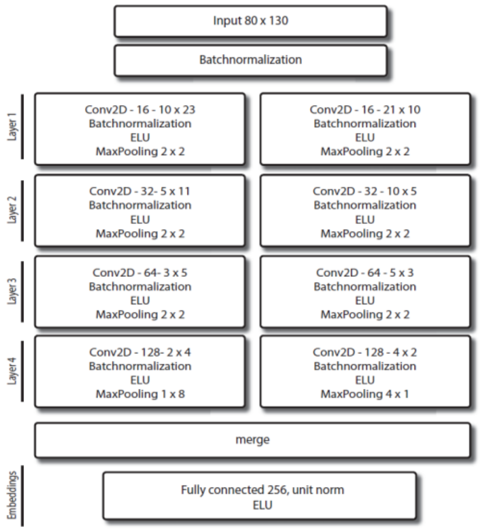
	Anchor
1	Positive Example
0	Negative Example

MULTI-TASK LEARNING AND EVALUATION

Contribution 3

PARALLEL DEEP NEURAL NETWORK

- Parallel CNN Filter Stacks
  - Timbre
    - Pooling X-axis
  - Rhythm
    - Pooling Y-axis
- Rectangular Filter shapes
- Works well on small datasets



- **Parallel convolutional neural networks for music genre and mood classification.** Lidy, Thomas, and Alexander Schindler. MIREX2016 (2016).
- **CQT-based convolutional neural networks for audio scene classification.** Thomas Lidy and Alexander Schindler. In *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2016 Workshop (DCASE2016)*, 2016.
- **Comparing shallow versus deep neural network architectures for automatic music genre classification.** Alexander Schindler, Thomas Lidy, and Andreas Rauber. In *Proceedings of the 9th Forum Media Technology (FMT2016)*, St. Poelten, Austria, 2016.
- **Multi-Temporal Resolution Convolutional Neural Networks for Acoustic Scene Classification.** Alexander Schindler, Thomas Lidy and Andreas Rauber. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE2017), November 2017.

EXPERIMENTAL SETUP

- **Same model for all experiments**
  - Controlled random processes (kernel intitializers, dropout, shuffle, etc.)
  - Batch-size 800 tracks
  - 100 epochs
- **Identical splits for all experiments** (train, val, test)
  - Grouped-Shuffle Split
    - Group-by Artist-ID (intrinsic Album-Filter to avoid „Album effect“)
- **Early stopping / save best model**
  - Patience 20 epochs
- **Evaluation Metric: Precision @100**
  - Euclidean Distance
- **Intersected Dataset**
  - Labels for all 4 Tag-sets available

	Genres	Styles	Moods	Themes
Unique Tags	21	833	285	166
Tag Combinations	449	7.446	14.300	7.298
Labelled Albums	19.107	19.107	19.107	19.107
Labelled Tracks	**143.587**	**143.587**	**143.587**	**143.587**

RESULTS

- **Task: Similar Artist/Album Retrieval**
  - Invariance to Artist/Album effects
  - Artist-Filter on Pairwise Cosine-Similarity Matrix

	Tag-Set	LSI Topics	Prec. Genres	Prec. Styles	Prec. Moods	Prec. Themes	Prec. Artists	Prec. Album
Single-Task	genres	10						
	genres	3						
	moods	160						x 1.83
	moods	20						
	styles	200						
	styles	20						
	themes	60						
	themes	20						
2-Tasks								x 1.91
								x 2.05
3-Tasks								x 2.18
4-T.								x 2.20

RESULTS

- Task: **Similar Genre** Retrieval
- Task: **Similar Style** Retrieval
- Task: **Similar Mood** Retrieval
- Task: **Similar Theme** Retrieval

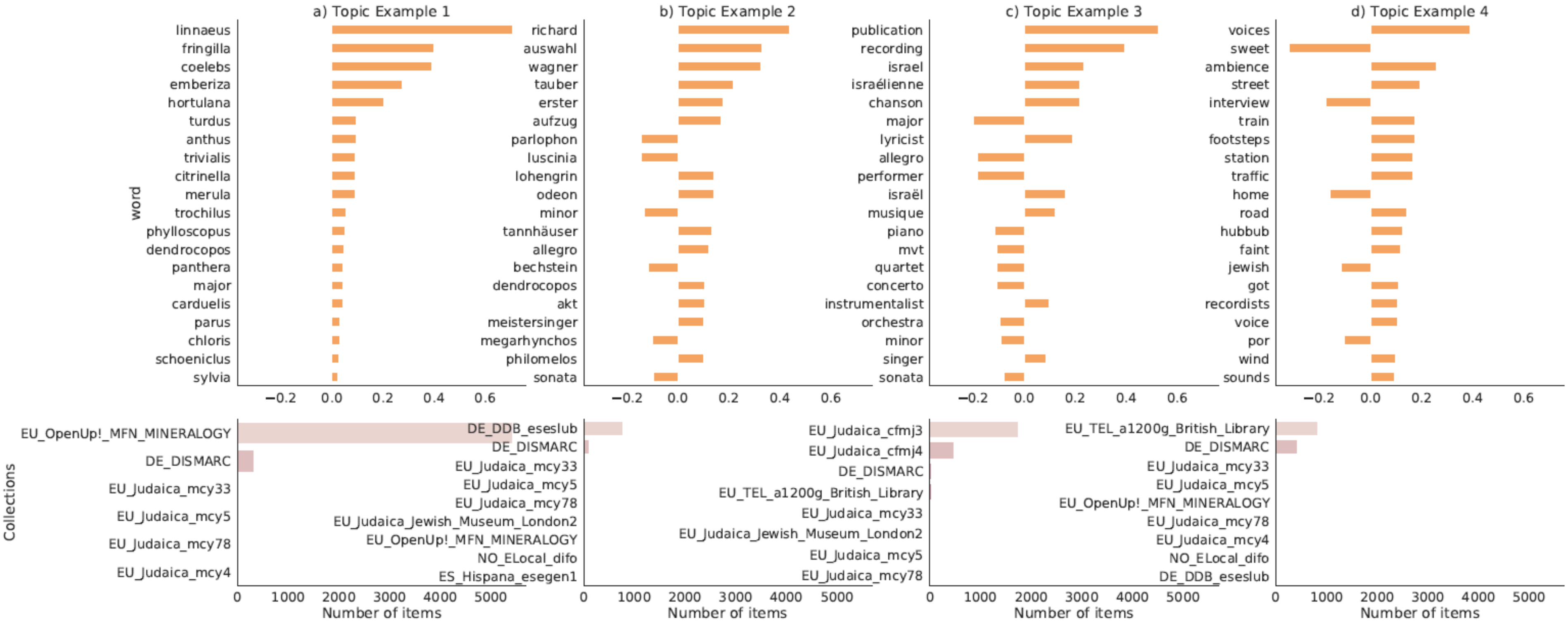
- Evaluate influence of different Tag-sets on the specific tasks

	Tag-Set	LSI Topics	Prec. Genres	Prec. Styles	Prec. Moods	Prec. Themes	Prec. Artists	Prec. Album
Single-Task	genres	10	0.3951	0.0091	0.0060	0.0076		
	genres	3	**0.3971**	0.0082	0.0055	0.0070		
2-Tasks								
3-Tasks								
4-T.								

CONCLUSION

- LSI-based representation learning works well, if
  - Diversity in Corpus is high enough
    - Otherwise density in cosine-similarity space is centered at 1
    - Similarity cannot be assessed satisfactory
  - Diversity in provided Tag-Set is high
    - Especially for Moods and Styles
    - Much higher in Free-text
  - Can be extended to project any semantic information from one corpus onto another
    - Free-Text Metadata (prepared for publishing)

FREE-TEXT METADATA FROM EUROPEANA



CONCLUSION

- **LSI-based representation learning** works well, if
  - Diversity in Corpus is high enough
    - Otherwise density in cosine-similarity space is centered at 1
    - Similarity cannot be assessed satisfactory
  - Diversity in provided Tag-Set is high
    - Especially for Moods and Styles
    - Much higher in Free-text
  - Can be extended to project any semantic information from one corpus onto another
    - Free-Text Metadata (prepared for publishing)
    - Album reviews (ongoing)
    - Lyrics (Future work)
    - Salient Visual Concepts (Future Work)

CONCLUSION

- **MSD Ground-Truth Assignments**
  - Proven effective in learning music representation
    - Music Tagging (Ongoing)
    - Transfer Learning (Ongoing)

LARGE SCALE TRANSFER LEARNING USING 4 TAG-SETS

AUSTRIAN INSTITUTE OF TECHNOLOGY

```
Metallica - ...And Justice For All

Moods/Themes:
Y-TRUE: Aggressive, Angry, Bitter, Bleak, Cathartic, Cerebral, Confrontational, Crunchy, Dramatic, Earnest, Epic, Fierce, Fi
ery, Gloomy, Gritty, Harsh, Hostile, Intense, Malevolent, Maverick, Menacing, Nihilistic, Ominous, Rambunctious, Rebellious,
Revolutionary, Searching, Suffocating, Tense/Anxious, Theatrical, Thuggish, Uncompromising, Victory, Visceral, Volatile

Y-PRED: Aggressive, Angry, Bleak, Confrontational, Harsh, Hostile, Intense, Malevolent, Menacing, Nihilistic, Ominous, Visce
ral

Genres/Styles:
Y-TRUE: Hard Rock, Heavy Metal, Pop/Rock, Speed/Thrash Metal

Y-PRED: Heavy Metal, Pop/Rock
```

```
Green Day - When I Come Around (Album Version)

Moods/Themes:
Y-TRUE: Boisterous, Brash, Cool & Cocky, Cynical/Sarcastic, Drinking, Energetic, Exuberant, Freewheeling, Fun, Guys Night Ou
t, Hanging Out, Humorous, Irreverent, Paranoid, Playful, Poignant, Quirky, Raucous, Rebellious, Rollicking, Rousing, Rowdy,
TGIF, Wry

Y-PRED: Cynical/Sarcastic, Energetic, Fun, Hanging Out, Irreverent, Playful, Quirky, Rambunctious, Rousing

Genres/Styles:
Y-TRUE: Alternative Pop/Rock, Alternative/Indie Rock, Pop/Rock, Post-Grunge, Punk Revival, Punk-Pop

Y-PRED: Alternative Pop/Rock, Alternative/Indie Rock, Pop/Rock, Punk Revival, Punk-Pop
```

```
Rihanna - Don't Stop The Music

Moods/Themes:
Y-TRUE: Amiable/Good-Natured, Boisterous, Brash, Carefree, Celebratory, Confident, Exuberant, Freedom, Fun, Girls Night Out,
Happy, Innocent, Joyous, Partying, Playful, Sex, Sexy, Summer, Summery, Sweet, TGIF, Warm

Y-PRED: Carefree, Celebratory, Club, Energetic, Exuberant, Fun, Partying, Playful, Stylish

Genres/Styles:
Y-TRUE: Contemporary R&B, Dance-Pop, Pop, Pop/Rock

Y-PRED: Club/Dance, Dance-Pop, Electronic, Pop, Pop/Rock
```

AUSTRIAN INSTITUTE OF TECHNOLOGY
TOMORROW TODAY

THANK YOU!

Alexander Schindler, 04.09.2019