the data challenge in music information retrieval

FRSM 2023

alexander lerch

github.com/alexanderlerch/2023-FRSM

Georgia Tech Center for Music Technology
College of Design

education

- Electrical Engineering (Technical University Berlin)
- Tonmeister (music production, University of Arts Berlin)

professional

- Associate Professor at the School of Music, Georgia Institute of Technology
- 2000-2013: CEO at zplane.development

background

- audio algorithm design (20+ years)
- commercial music software development (10+ years)
- entrepreneurship (10+ years)

www.linkedin.com/in/lerch

about

research directions

- field: music information retrieval, audio content analysis
  - audio classification
    - genre, instrument, auto-tagging, ...
  - music transcription
    - pitch, chord, performance data, ...
  - music processing
    - separation, ...
  - sound and music generation
    - controllability
- technical areas of interest
  - representation learning
  - machine learning with insufficient data
  - evaluation of generative systems

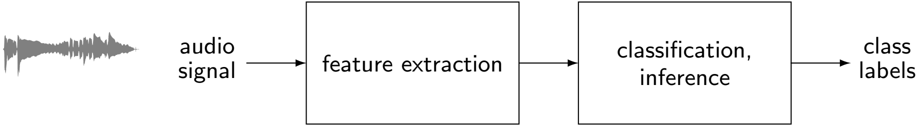
about

research directions

- field: music information retrieval, audio content analysis
  - audio classification
    - genre, instrument, auto-tagging, . . .
  - music transcription
    - pitch, chord, performance data, . . .
  - music processing
    - separation, . . .
  - sound and music generation
    - controllability
- technical areas of interest
  - representation learning
  - machine learning with insufficient data
  - evaluation of generative systems

introduction

audio classification — traditional



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

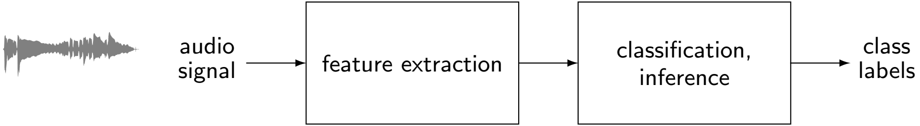
classification

- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: `http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf`.

introduction

audio classification — traditional



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

classification

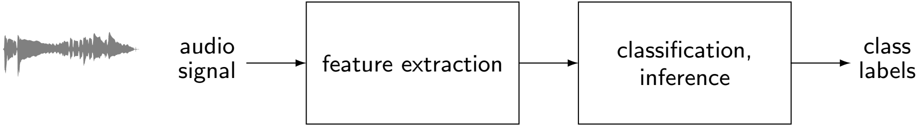
- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf.

introduction

audio classification — traditional

Georgia Tech Center for Music Technology College of Design



feature representation

- compact and non-redundant
- task-relevant
- easy to analyze
- e.g., MFCCs etc.

classification

- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

---

[1] J. J. Burred and A. Lerch, "Hierarchical Automatic Audio Signal Classification," *Journal of the Audio Engineering Society (JAES)*, vol. 52, no. 7/8, pp. 724–739, 2004. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2016/10/Burred-and-Lerch-2004-Hierarchical-Automatic-Audio-Signal-Classification.pdf.

introduction

neural network based approaches

Georgia Tech | Center for Music Technology
College of Design

- no custom-designed features anymore
- learn features from basic inputs (like spectrograms)

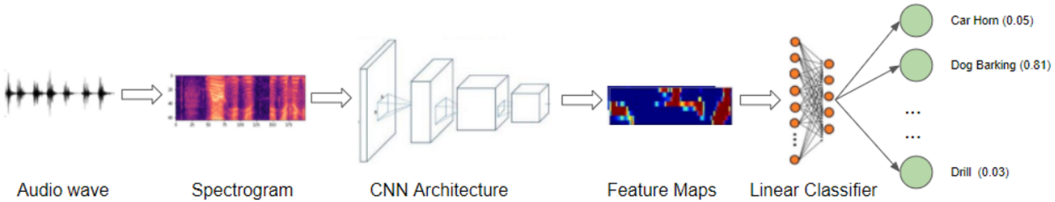
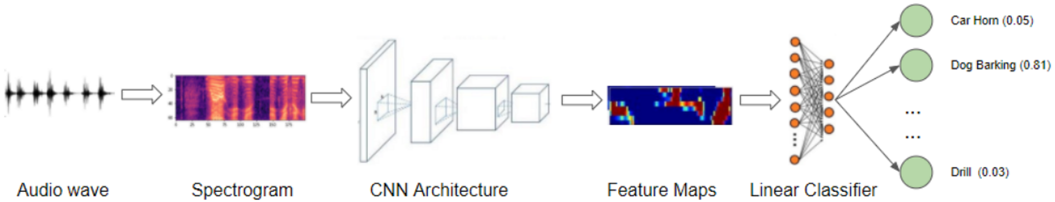


Fig.: towardsdatascience.com

- less required expert-knowledge, more complex systems
- less expert-tweaking, more rigorous experimental requirement
- much **higher data requirements**

introduction
neural network based approaches

- no custom-designed features anymore
- learn features from basic inputs (like spectrograms)



- less required expert-knowledge, more complex systems
- less expert-tweaking, more rigorous experimental requirement
- much **higher data requirements**

Fig.: towardsdatascience.com

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - subjectivity
  - noisiness
  - imbalance & bias
  - diversity & representativeness
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - subjectivity
  - noisiness
  - imbalance & bias
  - diversity & representativeness
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data

importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - subjectivity
  - noisiness
  - imbalance & bias
  - diversity & representativeness
  - amount

https://imgs.xkcd.com/comics/machine_learning.png

data
importance of data

machine learning: generic algorithm mapping an input to an output

- mapping function is learned from patterns and characteristics **from data**

⇒ model **success largely depends on training data**

- **general challenges** concerning data
  - subjectivity
  - noisiness
  - imbalance & bias
  - diversity & representativeness
  - **amount**

https://imgs.xkcd.com/comics/machine_learning.png

data
insufficient data

Georgia Tech Center for Music Technology College of Design

insufficient data in music

insufficient data in music

- **music data** itself is not scarce (although there might be copyright issues...)
- **consumer annotations** are more difficult to collect, but there are some large collections

data

insufficient data

insufficient data in music

- **music data** itself is not scarce (although there might be copyright issues...)
- **consumer annotations** are more difficult to collect, but there are some large collections
- **detailed musical annotations** are hard to come by, because
  - time consuming & tedious annotation process
  - experts needed for annotations

data

previous work on insufficient data

- literature proposes many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation[2]
  - transfer learning
  - semi- and self-supervised approaches
  - ...

[2] Y. Qin and A. Lerch, "Tuning Frequency Dependency in Music Classification," en, in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP)*, Brighton, UK: Institute of Electrical and Electronics Engineers (IEEE), 2019, pp. 401–405. DOI: 10.1109/ICASSP.2019.8683340. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2019/04/Qin-and-Lerch-2019-Tuning-Frequency-Dependency-in-Music-Classificatio.pdf.

data

previous work on insufficient data

- literature proposes many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning[2]
  - semi- and self-supervised approaches
  - ...

[2] S. Gururani, M. Sharma, and A. Lerch, "An Attention Mechanism for Music Instrument Recognition," in *ISMIR*, Delft, 2019.

data

previous work on insufficient data

- literature proposes many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning
  - semi- and self-supervised approaches[2][3]
  - ...

[2] C.-W. Wu and A. Lerch, "Automatic drum transcription using the student-teacher learning paradigm with unlabeled music data," in *ISMIR*, Suzhou, 2017.

[3] S. Gururani and A. Lerch, "Semi-Supervised Audio Classification with Partially Labeled Data," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, online: Institute of Electrical and Electronics Engineers (IEEE), 2021. [Online]. Available: https://arxiv.org/abs/2111.12761.

overview
overview

1. **self-supervised representation learning**
   - utilize pre-trained features to improve classification
2. **reprogramming**
   - utilize pre-trained model to improve classification
3. **data challenge revisited**

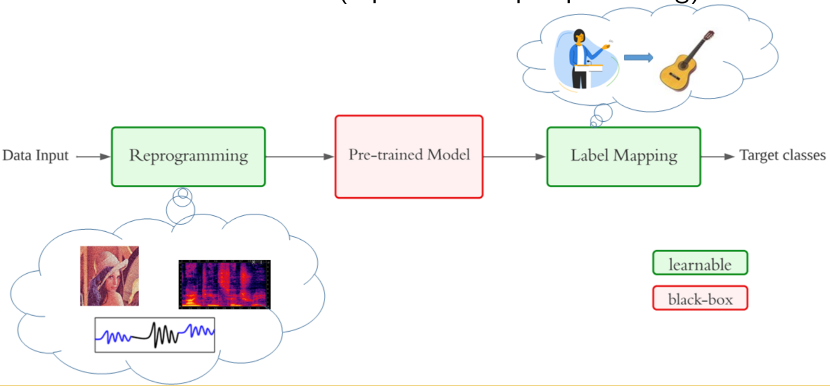
reprogramming
introduction

- **observation**
  - pre-trained deep models can be very powerful if trained with sufficient data, even for different tasks
- **idea**
  - re-using pre-trained models for a new task **without** re-training
- **goals**
  - keep number of training parameters minimal
  - utilize unmodified network trained on different task

- **observation**
  - pre-trained deep models can be very powerful if trained with sufficient data, even for different tasks
- **idea**
  - re-using pre-trained models for a new task **without** re-training
- **goals**
  - keep number of training parameters minimal
  - utilize unmodified network trained on different task

reprogramming
overview

- inspired by
  - transfer learning
  - adversarial learning
- allows for small trainable model (input and output processing)



reprogramming
experimental setup: data

- OpenMic:
  - 20 classes of musical instruments
  - 10 s audio snippets (20000)

reprogramming

experimental setup: baselines

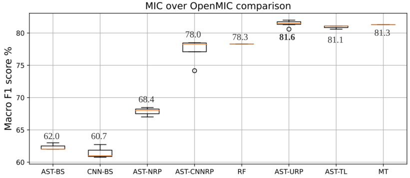
- Baseline AST:
  - good performance on audio event classification[4]
- ablation study:
  - CNN only
  - U-Net only
  - CNN + AST + FC
  - U-Net + AST + FC

[4] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech*, arXiv: 2104.01778, Brno, Czechia, Jul. 2021. [Online]. Available: http://arxiv.org/abs/2104.01778 (visited on 04/17/2022).

reprogramming

results: classification metrics

method	F1 (macro)	train. param. (M)
AST + simple output mapping	62.03	0.001
CNN	60.77	0.017
U-Net	62.73	0.017
CNN + AST + FC	78.08	0.017
U-Net + AST + FC	**81.60**	0.018



- a powerful model trained on a different task cannot easily be used directly
- proper input and output processing can significantly improve performance
- *re-programming can beat the state-of-the-art* with a fraction of trainable parameters (at least factor 10)

[5] H.-H. Chen and A. Lerch, "Music Instrument Classification Reprogrammed," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*, Bergen, Norway, 2023. [Online]. Available: https://arxiv.org/abs/2211.08379.

embeddings as teachers

introduction

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

embeddings as teachers

introduction

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

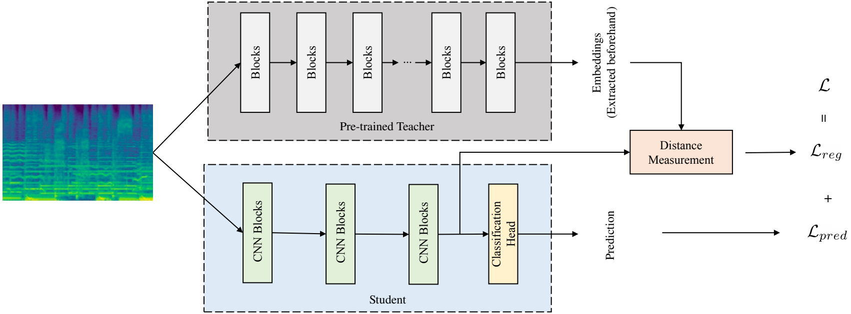
embeddings as teachers

introduction

- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings
  - *reduce model complexity*
- **general approach**:
  - combine transfer learning and knowledge distillation ideas

embeddings as teachers

method overview



- **transfer learning**
  - use embeddings from a different task for the target task
- **knowledge distillation**
  - use a teacher to train a less complex student on the same task

embeddings as teachers

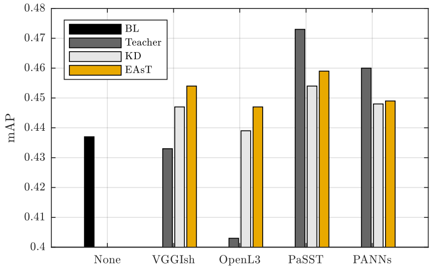
experimental setup

- task: auto-tagging
  - MagnaTagATune (MTAT) dataset:
    - 50 music tags
    - 30 s audio snippets (≈ 21000)
- systems:
  - baseline: student without teacher
  - teacher: embedding plus logistic regression
    - VGGish
    - OpenL3
    - PaSST
    - PANNs
  - KD: student trained with soft targets from teacher
  - EAsT: student regularized with teacher embeddings

embeddings as teachers

results

- student model consistently outperforms baseline
- student model consistently outperforms knowledge distillation
- student model outperforms teacher for "old" embeddings
- modern embeddings are powerful but complex



[6] Y. Ding and A. Lerch, "Audio Embeddings as Teachers for Music Classification," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Milan, Italy, 2023. DOI: 10.48550/arXiv.2306.17424. [Online]. Available: http://arxiv.org/abs/2306.17424 (visited on 07/03/2023).

data challenge revisited

insufficiency vs. representativeness

moderate improvements can be made to deal with insufficient data, but

data challenge revisited

insufficiency vs. representativeness

moderate improvements can be made to deal with insufficient data, but

is the amount of data really the main issue

data challenge revisited

insufficiency vs. representativeness

moderate improvements can be made to deal with insufficient data, but

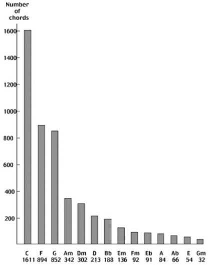
is the amount of data really the main issue

- maybe not...
  - a closer look at example music datasets for popular tasks

data challenge revisited

dataset example 1: chord detection

- Beatles dataset for chord detection
  - chord progressions from Beatles albums (181 songs)
  - chord vocabulary



- potential problems
  - stylistic homogeneity
    - timbre and instrumentation, style, release dates, audio quality...
  - chord imbalance
    - very skewed distribution, key dependence

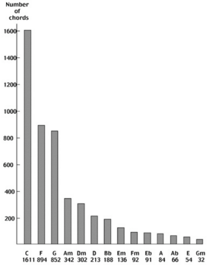
⇒ **may not generalize**

⇒ **may only recognize most common chords**

data challenge revisited

dataset example 1: chord detection

- Beatles dataset for chord detection
  - chord progressions from Beatles albums (181 songs)
  - chord vocabulary



- potential problems
  - stylistic homogeneity
    - timbre and instrumentation, style, release dates, audio quality...
  - chord imbalance
    - very skewed distribution, key dependence

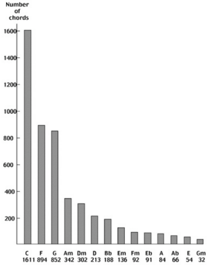
⇒ may not generalize

⇒ may only recognize most common chords

data challenge revisited

dataset example 1: chord detection

- Beatles dataset for chord detection
  - chord progressions from Beatles albums (181 songs)
  - chord vocabulary
- potential problems
  - stylistic homogeneity
    - timbre and instrumentation, style, release dates, audio quality...
  - chord imbalance
    - very skewed distribution, key dependence



$\Rightarrow$ **may not generalize**

$\Rightarrow$ **may only recognize most common chords**

data challenge revisited

dataset example 2: music genre classification

- GTZAN dataset for genre classification
  - 10 classes
  - 1000 30 s snippets
- problems with labeling
  - what are the 10 most relevant music genres, why 10?
  - how are genres categorized? examples:
    - baroque, christmas songs, fugue, indian art music, symphony, fusion
  - single-label vs. multi-label

⇒ **mismatch between dataset labels and 'real' task**

Disco
Country
Hip Hop
Rock
Blues
Reggae
Pop
Metal
Classical
Jazz

data challenge revisited

dataset example 2: music genre classification

- GTZAN dataset for genre classification
  - 10 classes
  - 1000 30 s snippets
- problems with labeling
  - what are the 10 most relevant music genres, why 10?
  - how are genres categorized? examples:
    - baroque, christmas songs, fugue, indian art music, symphony, fusion
  - single-label vs. multi-label

⇒ **mismatch between dataset labels and 'real' task**

Disco
Country
Hip Hop
Rock
Blues
Reggae
Pop
Metal
Classical
Jazz

data challenge revisited

dataset example 2: music genre classification

- GTZAN dataset for genre classification
  - 10 classes
  - 1000 30 s snippets
- problems with labeling
  - what are the 10 most relevant music genres, why 10?
  - how are genres categorized? examples:
    - baroque, christmas songs, fugue, indian art music, symphony, fusion
  - single-label vs. multi-label

⇒ **mismatch between dataset labels and 'real' task**

Disco
Country
Hip Hop
Rock
Blues
Reggae
Pop
Metal
Classical
Jazz

data challenge revisited

dataset example 3: source separation

- MUSDB dataset for source separation
  - 150 songs
  - 4 stems: vocals, drums, bass, other

- problems
  - stem selection does not reflect many real world scenarios
  - dataset size cannot reflect a diverse set of popular music

⇒ **mismatch between dataset and 'real' task**

data challenge revisited

dataset example 3: source separation

Georgia Tech | Center for Music Technology
College of Design

- MUSDB dataset for source separation
  - 150 songs
  - 4 stems: vocals, drums, bass, other
- problems
  - stem selection does not reflect many real world scenarios
  - dataset size cannot reflect a diverse set of popular music

⇒ **mismatch between dataset and 'real' task**

data challenge revisited
dataset example 3: source separation

- MUSDB dataset for source separation
  - 150 songs
  - 4 stems: vocals, drums, bass, other

- problems
  - stem selection does not reflect many real world scenarios
  - dataset size cannot reflect a diverse set of popular music

⇒ **mismatch between dataset and 'real' task**

data challenge revisited

dataset examples: summary

- false homogeneity/**non-representativeness impacts generalization**
  - system cannot learn what is hasn't seen or what seems irrelevant
- imbalance can lead to **unwanted bias**
  - *training*: system wrongly favors certain categories
  - *testing*: results may imply good performance yet cannot be generalized
- mismatch between dataset labels and real task may **feign good performance**
  - misleading results
  - architectural bias

conclusion
data challenge

- we presented **2 recent approaches**
  - a novel *self-supervised regularization loss*
  - *reprogramming* for audio classification
- all approaches perform **at or above the state-of-the-art** with different trade-offs between
  - *training complexity*
  - *inference complexity*
  - *classification accuracy*
- **but:** maybe we tried to solve the wrong challenges

thank you!

Georgia Tech Center for Music Technology
College of Design

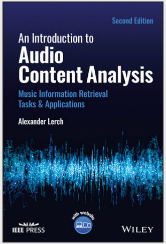
links

alexander lerch: www.linkedin.com/in/lerch

mail: alexander.lerch@gatech.edu

book: www.AudioContentAnalysis.org

music informatics group: musicinformatics.gatech.edu



github.com/alexanderlerch