From Audio to Music Understanding

machine learning for music analysis, processing, & generation

alexander lerch

github.com/alexanderlerch/lerch-research-pr

about

about me

- **education**
  - Electrical Engineering (Technical University, Berlin)
  - Tonmeister (music production, University of Arts, Berlin)
- **professional**
  - Professor, School of Music, Georgia Tech
  - Associate Dean for Research & Creative Practice, College of Design, Georgia Tech
  - 2000-2013: CEO at zplane.development
- **experience**
  - audio algorithm design (20+ years)
  - machine learning for music (15+ years)
  - professional music software engineering & development (10+ years)
  - entrepreneurship (10+ years)
  - research administration (2+ years)

www.linkedin.com/in/lerch

introduction

vision & mission

- **vision**
  - **democratization** of
    - music making
    - music education
    - music discovery

  ⇒ through **machine understanding of music**
    - musically meaningful discovery & processing
    - musically meaningful / controllable generation
    - musically intelligent ai tutors

- **mission**
  - create new technologies transforming and enhancing how we *make, produce, perform, discover,* and *consume music*
  - advance the field of AI for audio/music through *knowledge-driven machine learning*

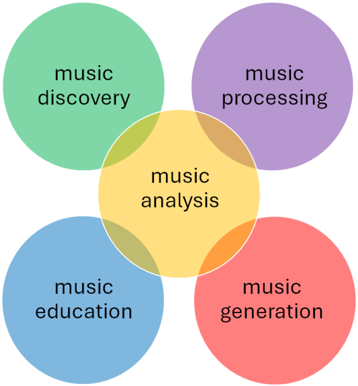
introduction

vision & mission

- **vision**
  - **democratization** of
    - music making
    - music education
    - music discovery

  ⇒ through **machine understanding of music**
    - musically meaningful discovery & processing
    - musically meaningful / controllable generation
    - musically intelligent ai tutors

- **mission**
  - create new technologies transforming and enhancing how we *make, produce, perform, discover,* and *consume music*
  - advance the field of AI for audio/music through *knowledge-driven machine learning*

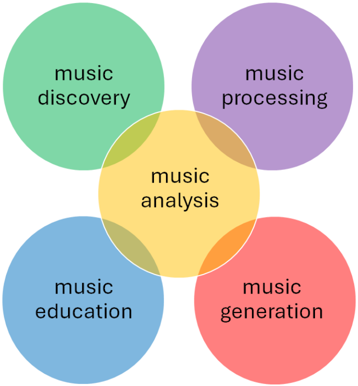


introduction
vision & mission

- **vision**
  - **democratization** of
    - music making
    - music education
    - music discovery

  ⇒ through **machine understanding of music**
    - musically meaningful discovery & processing
    - musically meaningful / controllable generation
    - musically intelligent ai tutors

- **mission**
  - create new technologies transforming and enhancing how we *make*, *produce*, *perform*, *discover*, and *consume music*
  - advance the field of AI for audio/music through *knowledge-driven machine learning*



introduction

research focus (tasks)

- **music analysis**
  - music/audio *classification*
    - genre/events [1], [2]
    - instruments [3], [4], [5]
    - tagging [5], [6], [7]
    - emotion [8]
  - music *transcription*
    - drum transcription [9]
    - chord detection [10]
    - pitch tracking [11]
  - music *performance analysis* [12], [13]
- **music processing**
  - music source separation [14], [15], [16]
- **sound and music generation**
  - evaluation [17], [18], [19], [20]

introduction

research focus (tasks)

- **music analysis**
  - music/audio *classification*
    - genre/events [1], [2]
    - instruments [3], [4], [5]
    - tagging [5], [6], [7]
    - emotion [8]
  - music *transcription*
    - drum transcription [9]
    - chord detection [10]
    - pitch tracking [11]
  - music *performance analysis* [12], [13]
- **music processing**
  - music source separation [14], [15], [16]
- **sound and music generation**
  - evaluation [17], [18], [19], [20]

introduction

research focus (tasks)

- **music analysis**
  - music/audio *classification*
    - genre/events [1], [2]
    - instruments [3], [4], [5]
    - tagging [5], [6], [7]
    - emotion [8]
  - music *transcription*
    - drum transcription [9]
    - chord detection [10]
    - pitch tracking [11]
  - music *performance analysis* [12], [13]
- **music processing**
  - music source separation [14], [15], [16]
- **sound and music generation**
  - evaluation [17], [18], [19], [20]

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* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *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* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *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* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *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* of annotations
  - *noisiness* (bad quality, bad annotations, . . .)
  - *imbalance & bias* (distribution is skewed, biased)
  - *diversity & representativeness*
  - **amount**

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

data
insufficient data

- **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
insufficient data

- **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

insufficient data

- **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

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning [21]
  - semi- and self-supervised approaches [22][3]
  - ...

data

previous work on insufficient data

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning [21]
  - semi- and self-supervised approaches [22][3]
  - ...

data

previous work on insufficient data

- there are many ways of **dealing with insufficient data**
  - data synthesis
  - data augmentation
  - transfer learning [21]
  - semi- and self-supervised approaches [22][3]
  - ...

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

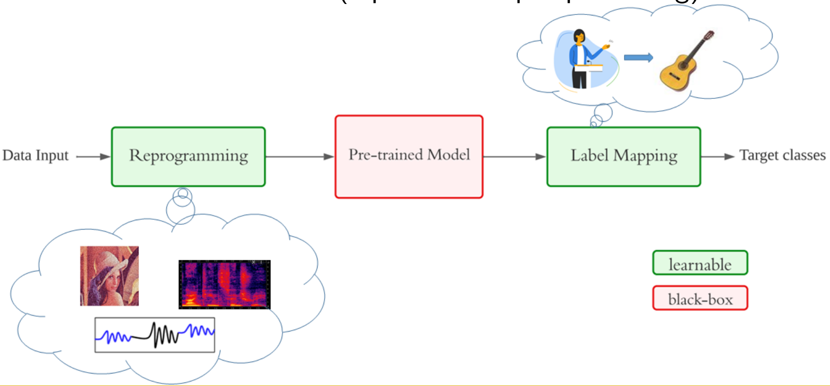
reprogramming
introduction

- **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

reprogramming

overview

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



reprogramming

experimental setup: baselines

- baseline AST:
  - good performance on audio event classification [23]

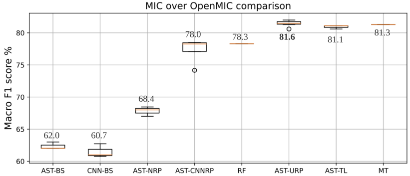
- data
  - OpenMic (instrument classification):
    - 20 classes of musical instruments
    - 10 s audio snippets (20000)

- ablation study:
  - CNN only
  - U-Net only
  - CNN + AST + FC
  - U-Net + AST + FC

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 [4]
- proper input and output processing can significantly improve performance
- *re-programming can beat the state-of-the-art* at a fraction of trainable parameters (at least factor 10)

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

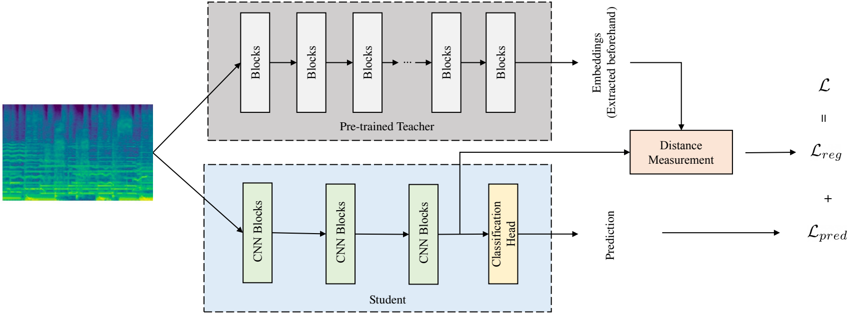
- **question**:
  - how can we provide extra training information without additional data labels
- **idea**:
  - use proven pre-trained embeddings (e.g., VGGish, OpenL3, . . .)
- **goals**:
  - *impart knowledge* of pre-trained deep models
  - *improve model generalization* by utilizing pre-trained embeddings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

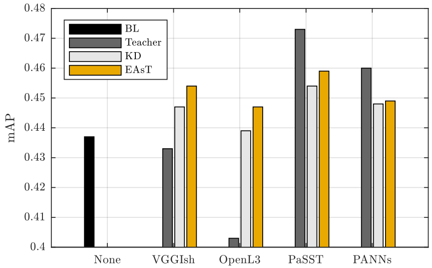
- task: auto-tagging
  - MagnaTagATune (MTAT) dataset:
    - 50 music tags
    - 30 s audio snippets ($\approx$ 21000)

- 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 [5]
- student model consistently outperforms knowledge distillation
- student model outperforms teacher for "old" embeddings
- modern embeddings are powerful but complex



conclusion

summary

- music analysis offers a **wide range of tasks** that require **specialized solutions** and **domain knowledge**
- **training data** availability remains an open problem for many music tasks
- **knowledge transfer** methodologies allow for compensating limited (and potentially biased) training data

conclusion

future work

1. **knowledge transfer**
   - transfer knowledge between tasks/modalities
   - inject expert knowledge
2. **representation learning**
   - interpretability of embedding spaces
   - understanding of learned information
3. **machine learning with insufficient data**
   - approaches reducing the risk of overfitting
   - lsynthesis and augmentation techniques
4. **evaluation of generative systems**
   - extendable framework for objective evaluation metrics
   - detection of generated content

thank you!

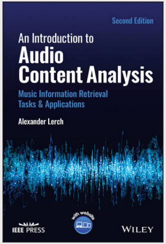
links

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

references I

references

[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.

[2] Y.-N. Hung, C.-H. H. Yang, P.-Y. Chen, and A. Lerch, "Low-Resource Music Genre Classification with Cross-Modal Neural Model Reprogramming," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: Institute of Electrical and Electronics Engineers (IEEE), 2023. DOI: 10.1109/ICASSP49357.2023.10096568. [Online]. Available: https://arxiv.org/abs/2211.01317.

[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.

[4] 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.

[5] 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. Accessed: Jul. 3, 2023. [Online]. Available: http://arxiv.org/abs/2306.17424.

[6] Y. Ding and A. Lerch, "Embedding Compression for Teacher-to-Student Knowledge Transfer," in *Proceedings of the International Conference on Acoustics Speech and Signal Processing (ICASSP) - Satellite Workshop Deep Neural Network Model Compression*, Seoul, Korea: Institute of Electrical and Electronics Engineers (IEEE), Feb. 2024. DOI: 10.48550/arXiv.2402.06761. Accessed: Feb. 27, 2024. [Online]. Available: http://arxiv.org/abs/2402.06761.

[7] T. A. Ma and A. Lerch, "Music auto-tagging in the long tail: A few-shot approach," in *Proceedings of the AES Convention*, New York, Sep. 2024. DOI: 10.48550/arXiv.2409.07730. Accessed: Sep. 13, 2024. [Online]. Available: http://arxiv.org/abs/2409.07730.

references II

references

[8] K. N. Watcharasupat, Y. Ding, T. A. Ma, P. Seshadri, and A. Lerch, "Uncertainty Estimation in the Real World: A Study on Music Emotion Recognition," in *Proceedings of the European Conference on Information Retrieval (ECIR)*, Lucca, Italy: arXiv, 2025. DOI: 10.48550/arXiv.2501.11570. Accessed: Jan. 30, 2025. [Online]. Available: http://arxiv.org/abs/2501.11570.

[9] C.-W. Wu et al., "A Review of Automatic Drum Transcription," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1457–1483, 2018, ISSN: 2329-9290. DOI: 10.1109/TASLP.2018.2830113. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2018/05/Wu-et-al.-2018-A-review-of-automatic-drum-transcription.pdf.

[10] X. Zhou and A. Lerch, "Chord Detection Using Deep Learning," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Malaga: ISMIR, 2015. DOI: 10.5281/zenodo.1416968. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2015/10/Zhou_Lerch_2015_Chord-Detection-Using-Deep-Learning.pdf.

[11] A. Lerch, *An Introduction to Audio Content Analysis: Music Information Retrieval Tasks and Applications*, en-us, 2nd ed. Hoboken, N.J: Wiley-IEEE Press, 2023, ISBN: 978-1-119-89094-2. Accessed: Nov. 4, 2022. [Online]. Available: https://ieeexplore.ieee.org/servlet/opac?bknumber=9965970.

[12] A. Lerch, *Software-Based Extraction of Objective Parameters from Music Performances*. München: GRIN Verlag, 2009, ISBN: 978-3-640-29496-1. [Online]. Available: 10.14279/depositonce-2025.

[13] K. A. Pati, S. Gururani, and A. Lerch, "Assessment of Student Music Performances Using Deep Neural Networks," en, *Applied Sciences*, vol. 8, no. 4, p. 507, 2018. DOI: 10.3390/app8040507. Accessed: Mar. 27, 2018. [Online]. Available: http://www.mdpi.com/2076-3417/8/4/507/pdf.

references III

references

[14] Y.-N. Hung and A. Lerch, "Multi-Task Learning for Instrument Activation Aware Music Source Separation," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Montreal: International Society for Music Information Retrieval (ISMIR), 2020. [Online]. Available: https://musicinformatics.gatech.edu/wp-content_nondefault/uploads/2020/08/Hung-and-Lerch-2020-Multi-Task-Learning-for-Instrument-Activation-Awar.pdf.

[15] K. N. Watcharasupat et al., "A Generalized Bandsplit Neural Network for Cinematic Audio Source Separation," *IEEE Open Journal of Signal Processing*, vol. 5, pp. 73–81, 2024, ISSN: 2644-1322. DOI: 10.1109/OJSP.2023.3339428. Accessed: Jan. 2, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10342812.

[16] K. N. Watcharasupat and A. Lerch, "A Stem-Agnostic Single-Decoder System for Music Source Separation Beyond Four Stems," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, San Francisco, Jun. 2024. DOI: 10.48550/arXiv.2406.18747. Accessed: Aug. 8, 2024. [Online]. Available: http://arxiv.org/abs/2406.18747.

[17] L.-C. Yang and A. Lerch, "On the Evaluation of Generative Models in Music," en, *Neural Computing and Applications*, no. 32, pp. 4773–4784, 2020, ISSN: 1433-3058. DOI: 10.1007/s00521-018-3849-7. Accessed: Nov. 4, 2018. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2018/11/postprint.pdf.

[18] A. Pati and A. Lerch, "Is Disentanglement Enough? On Latent Representations for Controllable Music Generation," en, in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Online, 2021, p. 8. [Online]. Available: https://arxiv.org/abs/2108.01450.

[19] K. N. Watcharasupat, J. Lee, and A. Lerch, "Latte: Cross-framework Python Package for Evaluation of Latent-based Generative Models," en, *Software Impacts*, p. 100 222, 2022, ISSN: 26659638. DOI: 10.1016/j.simpa.2022.100222. Accessed: Jan. 13, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2665963822000033.

references IV

references

[20] A. Vinay and A. Lerch, "Evaluating Generative Audio Systems and their Metrics," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, arXiv:2209.00130 [cs, eess], Bangalore, IN, Aug. 2022. DOI: 10.48550/arXiv.2209.00130. Accessed: Sep. 3, 2022. [Online]. Available: http://arxiv.org/abs/2209.00130.

[21] S. Gururani, M. Sharma, and A. Lerch, "An Attention Mechanism for Music Instrument Recognition," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Delft: International Society for Music Information Retrieval (ISMIR), 2019. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2019/07/Gururani-et-al.-2019-An-Attention-Mechanism-for-Music-Instrument-Recogn.pdf.

[22] C.-W. Wu and A. Lerch, "Automatic Drum Transcription using the Student-Teacher Learning Paradigm with Unlabeled Music Data," in *Proceedings of the International Society for Music Information Retrieval Conference (ISMIR)*, Suzhou: International Society for Music Information Retrieval (ISMIR), 2017. DOI: 10.5281/zenodo.1415904. [Online]. Available: http://www.musicinformatics.gatech.edu/wp-content_nondefault/uploads/2017/07/Wu_Lerch_2017_Automatic-drum-transcription-using-the-student-teacher-learning-paradigm-with.pdf.

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