audio classification with insufficient data

alexander lerch

Georgia Tech Center for Music Technology
College of Design

introduction
about me

- **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

introduction
audio classification

- audio classification: one of the earliest and seminal tasks in Music Information Retrieval (MIR)
- includes, e.g.,
  - music/speech classification
  - genre classification
  - musical instrument recognition
  - mood recognition
  - music auto-tagging
  - artist classification
  - ...
- non-music related
  - speaker detection
  - audio event detection
  - ...

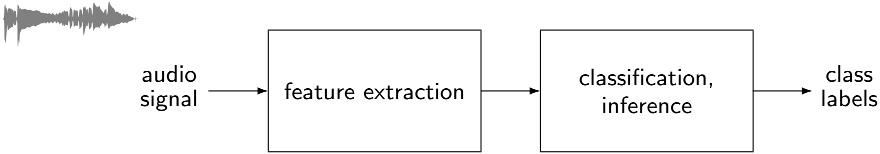
introduction

audio classification

- audio classification: one of the earliest and seminal tasks in Music Information Retrieval (MIR)
- includes, e.g.,
  - music/speech classification
  - genre classification
  - musical instrument recognition
  - mood recognition
  - music auto-tagging
  - artist classification
  - ...
- non-music related
  - speaker detection
  - audio event detection
  - ...

introduction

old work: genre classification



feature representation

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

classification

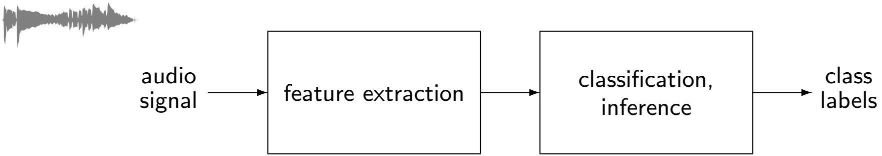
- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[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.

introduction

old work: genre classification

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.

introduction

old work: genre classification

audio signal → feature extraction → classification, inference → class labels

feature representation

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

classification

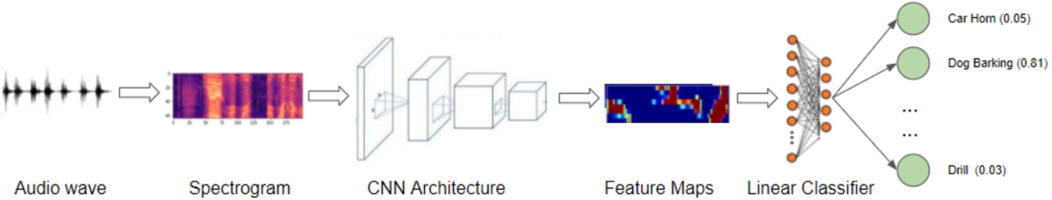
- map or convert feature to comprehensible domain
- e.g., Support Vector Machines etc.

[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.

introduction

neural network based approaches

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



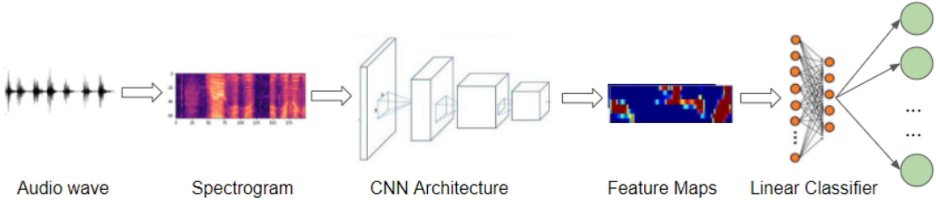
- less required expert-knowledge, more complex systems
- less expert-tweaking, more rigorous experimental requirement
- much **higher data requirements**

Fig.: towardsdatascience.com/audio-deep-learning-made-simple-sound-classification-step-by

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/audio-deep-learning-made-simple-sound-classification-step-by

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
  - noisiness
  - subjectivity
  - imbalance, bias, and diversity
  - 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
  - noisiness
  - subjectivity
  - imbalance, bias, and diversity
  - 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**

$\Rightarrow$ model **success largely depends on training data**

- **general challenges** concerning data
  - noisiness
  - subjectivity
  - imbalance, bias, and diversity
  - 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**

$\Rightarrow$ model **success largely depends on training data**

- **general challenges** concerning data
  - noisiness
  - subjectivity
  - imbalance, bias, and diversity
  - amount

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

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

Georgia Tech | Center for Music Technology
College of Design

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 and tedious annotation process
  - experts needed for annotations

overview

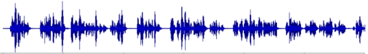
overview

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

semi-supervised audio classification
introduction

- **observation**:
  - unlabeled data is readily available
    - example: OpenMIC dataset (musical instrument classification)

Data Point



	Guitar	Drum	Bass	Violin	Piano	Flute
Fully Labeled	✓	✓	✓	✗	✓	✗
Partially Labeled	✓	✓	?	✗	?	?

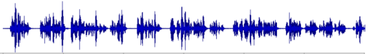
- **goal**:
  - utilize *unlabeled* data for training to improve inference

semi-supervised audio classification

introduction

- **observation**:
  - unlabeled data is readily available
    - example: OpenMIC dataset (musical instrument classification)

Data Point



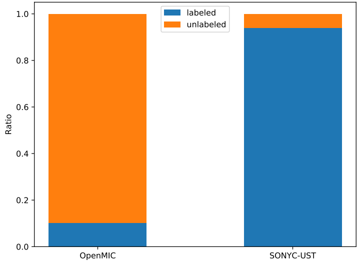
	Guitar	Drum	Bass	Violin	Piano	Flute
Fully Labeled	✓	✓	✓	✗	✓	✗
Partially Labeled	✓	✓	?	✗	?	?

- **goal**:
  - utilize *unlabeled* data for training to improve inference

semi-supervised audio classification

experimental setup: data

- OpenMic:
  - 20 classes of musical instruments
  - 10 s audio snippets (20000)
  - 90% of labels are missing
- SONYC Urban Sound Tagging:
  - 23 classes of urban noise
  - 10 s audio snippets (13538 + 4308 + 669)
  - 6% of labels are missing



semi-supervised audio classification

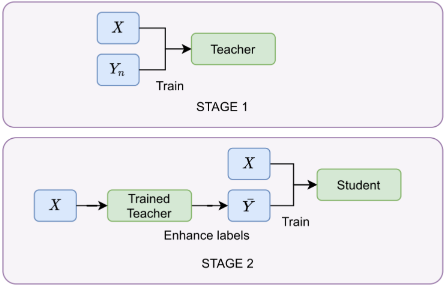
experimental setup: baselines

- Baseline 0 (B0):
  - missing labels are treated as negative labels
  - "standard approach"
- Baseline 1 (B1):
  - missing labels are masked out of the loss function

semi-supervised audio classification

method 1: label enhancing

- stage 1:
  - assume all missing labels are negative
  - train a teacher system
- stage 2:
  - predict labels with teacher
  - train student with combined training set/likely predicted labels
  - mask the loss for unlikely negatives

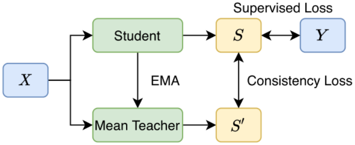


[2] E. Fonseca, S. Hershey, M. Plakal, *et al.*, "Addressing Missing Labels in Large-Scale Sound Event Recognition Using a Teacher-Student Framework With Loss Masking," *IEEE Signal Processing Letters*, vol. 27, pp. 1235–1239, 2020, Conference Name: IEEE Signal Processing Letters, ISSN: 1558-2361. DOI: 10.1109/LSP.2020.3006378.

semi-supervised audio classification

method 2: mean teacher

- teacher and student are trained simultaneously
- teacher is exponential average (EMA) of student
- consistency loss is computed from the teacher predictions
- student is updated with both consistency loss and binary cross-entropy loss



[3] P. Bachman, O. Alsharif, and D. Precup, "Learning with Pseudo-Ensembles," in *Advances in Neural Information Processing Systems*, vol. 27, Curran Associates, Inc., 2014.

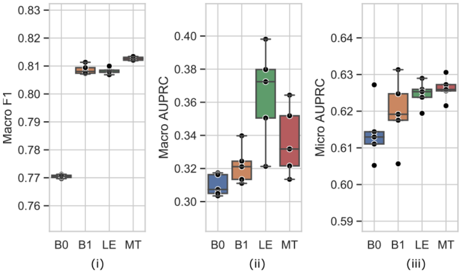
semi-supervised audio classification

results: classification

- general observations
  - B0 always worse performance
  - B1 much better but can be outperformed

(i) OpenMic:
- Mean Teacher outperforms Label Enhancing

(iii) SONYC Urban Sound Tagging:
- comparable performance of Mean Teacher and Label Enhancing

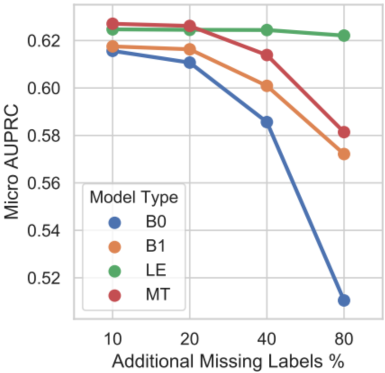


[4] S. Gururani and A. Lerch, "Semi-Supervised Audio Classification with Partially Labeled Data," in *Proceedings of the IEEE International Symposium on Multimedia (ISM)*, online: Institute of Electrical and Electronics Engineers (IEEE), 2021.

semi-supervised audio classification

results: data dependency

- removing labels from SONYC Urban Sound Tagging
  - baselines deteriorate much faster



[5] 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.

self-supervised representation learning

introduction

- **question**:
  - how can we provide extra training information without additional data labels (related approaches: transfer learning, multi-task learning)
- **idea**:
  - use proven pre-trained features (e.g., VGGish, OpenL3)
- **goals**:
  - *impart knowledge* of pre-trained deep models (VGGish, L3)
  - *improve model generalization* by utilizing pre-trained features
  - use pre-trained features *only during training*

self-supervised representation learning

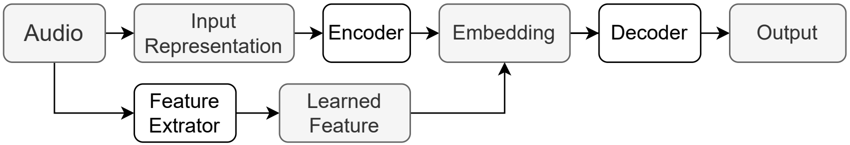
introduction

- **question**:
  - how can we provide extra training information without additional data labels (related approaches: transfer learning, multi-task learning)
- **idea**:
  - use proven pre-trained features (e.g., VGGish, OpenL3)
- **goals**:
  - *impart knowledge* of pre-trained deep models (VGGish, L3)
  - *improve model generalization* by utilizing pre-trained features
  - use pre-trained features *only during training*

self-supervised representation learning

method overview

Georgia Tech | Center for Music Technology
College of Design



- **method 1: "Con-Reg"**
  - make embedding space more similar to embedding space of features
- **method 2: "Dis-Reg"**
  - force distances between pairs of embedding vectors to be similar to feature distances

self-supervised representation learning

experimental setup: baselines

- standard **transfer** learning
  1. extract features with pre-trained network
  2. train classifier for new task with feature input

- **concat**enation:
  - concatenate the pre-trained features and the learned embeddings
  - classifier has the combined information (trained and pre-trained)

self-supervised representation learning

experimental setup: data

- DCASE 17:
  - 17 audio event classes
  - 10 s audio snippets ($\approx$ 53000)
- MagnaTagATune (MTAT):
  - 50 music tags
  - 30 s audio snippets ($\approx$ 21000)

self-supervised representation learning

results: classification metrics

	Methods	DCASE 17 (F1)				MTAT (PR-AUC)			
		None	VGGish	OpenL3	Combined	None	VGGish	OpenL3	Combined
BL	Won et al.	0.547	-	-	-	0.465	-	-	-
	transfer	-	0.496	0.477	0.501	-	0.454	0.454	0.456
	concat	-	0.529	0.492	0.495	-	0.457	0.464	0.458
Prop.	Con-Reg	-	**0.568**	**0.557**	**0.576**	-	**0.471**	0.466	**0.469**
	Dis-Reg	-	0.548	0.543	0.563	-	0.464	**0.468**	0.463

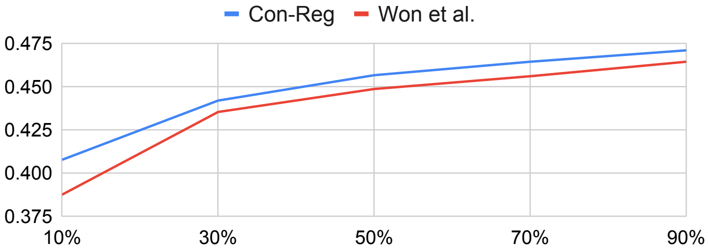
- two baselines *cannot outperform* the trained system without additional features
- *combining VGGish and L3 generally improves* on the individual feature results
- *approach improves embedding space* by using pre-trained features during training

[6] Y.-N. Hung and A. Lerch, "Feature-informed Embedding Space Regularization for Audio Classification," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022. DOI: 10.48550/arXiv.2206.04850.

self-supervised representation learning

results: data dependency

- Con-Reg outperforms non-regularized system in all cases
- larger improvement for lower amounts of data



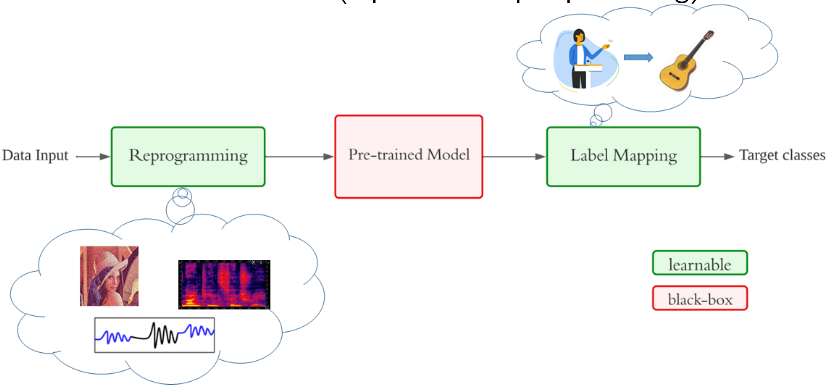
[7] Y.-N. Hung and A. Lerch, "Feature-informed Embedding Space Regularization for Audio Classification," in *Proceedings of the European Signal Processing Conference (EUSIPCO)*, Belgrade, Serbia, 2022. DOI: 10.48550/arXiv.2206.04850.

- **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:
  - state of the art performance on audio event classification[8]

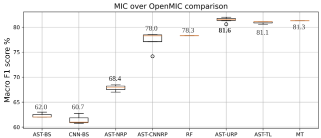
- ablation study:
  - CNN only
  - U-Net only
  - CNN + AST + FC
  - U-Net + AST + FC

[8] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Proceedings of Interspeech*, arXiv: 2104.01778, Brno, Czechia, Jul. 2021.

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)

[9] H.-H. Chen and A. Lerch, "Music Instrument Classification Reprogrammed," in *Proceedings of the International Conference on Multimedia Modeling (MMM)*, Bergen, Norway, 2023.

conclusion

learning with insufficient data

- literature presents many ways of **dealing with insufficient data**
  - data augmentation
  - data synthesis
  - transfer learning
  - semi- and self-supervised approaches
  - ...
- we presented **3 recent approaches**
  - state-of-the-art *semi-supervised learning*
  - 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*

thank you!

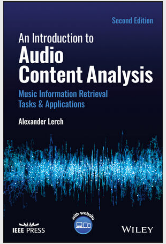
Georgia Tech | Center for Music Technology
College of Design

contact

Alexander Lerch:
alexander.lerch@gatech.edu

www.AudioContentAnalysis.org
www.alexanderlerch.com

Music Informatics Group
musicinformatics.gatech.edu



github.com/alexanderlerch