



MSc. Artificial Intelligence, Master Thesis Defence, Thursday 28nd January, 2021

Contrastive Learning of Musical Representations

JANNE SPIJKERVET

Supervisor: Dr. J.A. Burgoyne

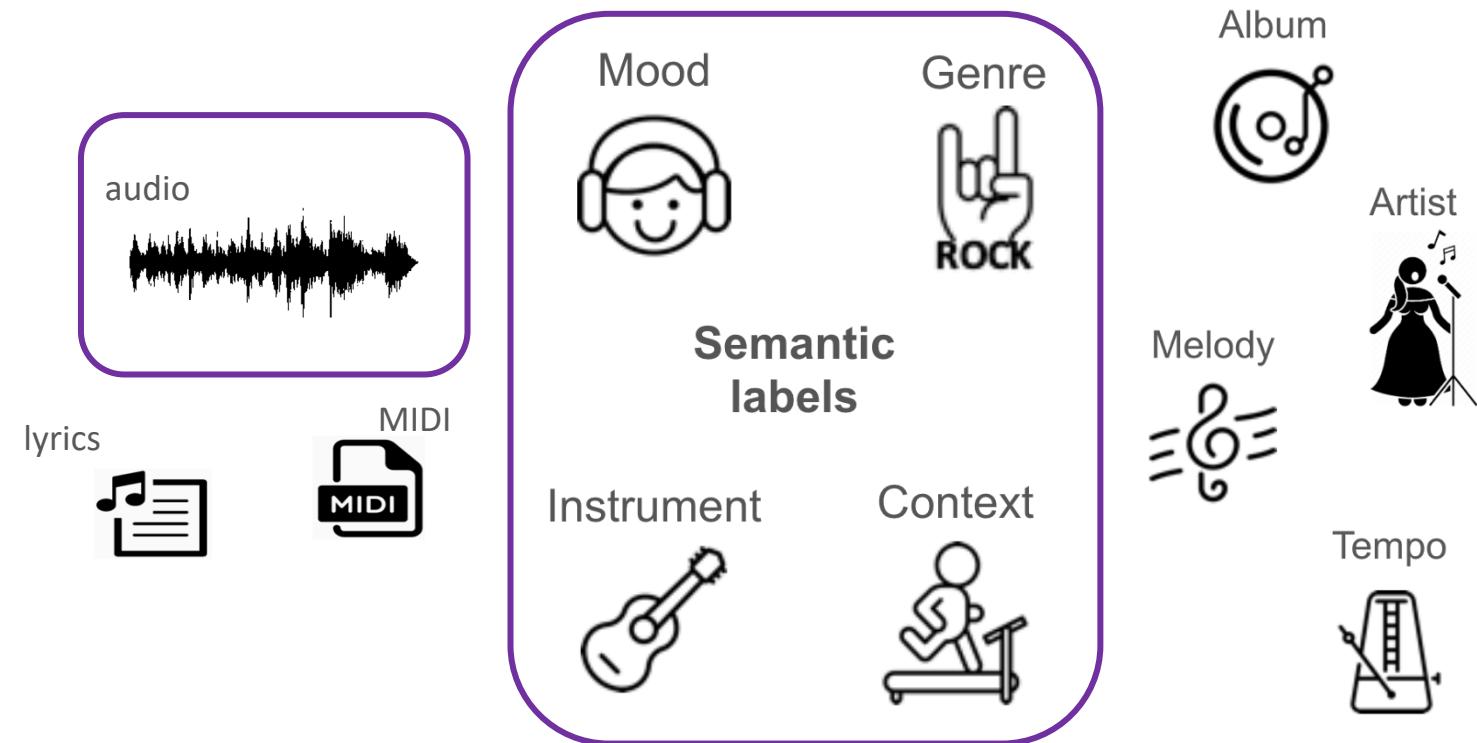
Examiner: Dr. W. Aziz

Schedule

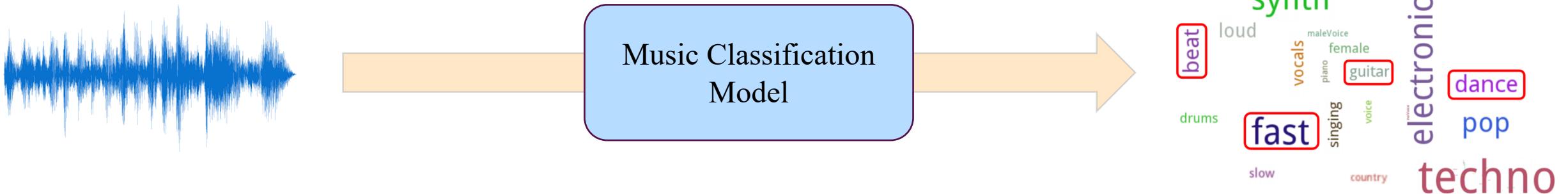
1. Introduction
2. Related Work
3. Method
4. Experimental Results
5. Interpretability
6. Discussion
7. Limitations & Future Work
8. Conclusion

Background: Music Classification

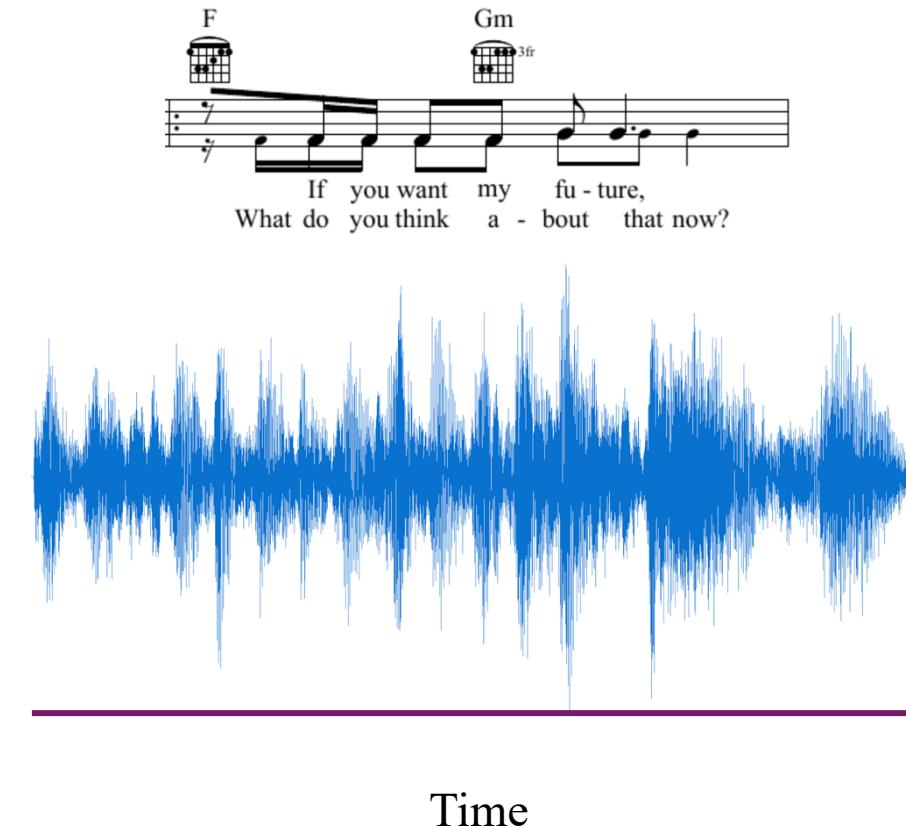
1. Music catalogs
2. Retrieval
3. Recommendation
4. Collaborative Filtering
5. Cold-start problem
6. Auto-tagging



Background: Music Classification



Background: Music Labeling



 **Wannabe**
~ Release group by Spice Girls

Overview Aliases **Tags** Details Edit

Genres

- electronic
- house
- pop
- dance-pop
- euro house
- eurodance
- europop
- pop rock
- rock
- synth-pop
- teen pop

Background: Music Labeling

- “Build larger datasets”?
- Hard to create music datasets (annotator subjectivity^[1])
- Mostly *single reference* annotations

Melody Transcription

MedleyDB [2]

108 annotations

Chord Recognition

McGill Billboard Dataset [3]

740 distinct songs

Music Classification



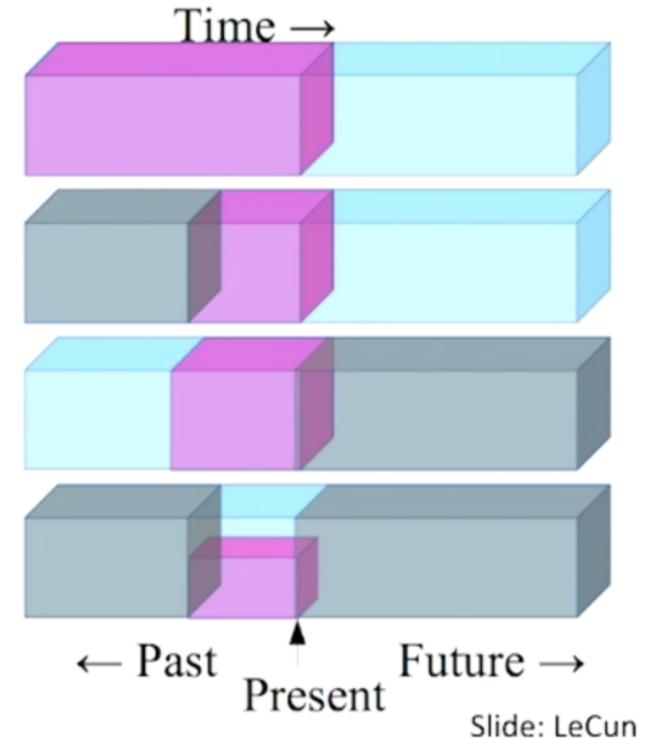
[1] Hendrik Vincent Koops, W. Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller & Anja Volk (2019) Annotator subjectivity in harmony annotations of popular music, Journal of New Music Research, 48:3, 232-252, DOI: 10.1080/09298215.2019.1613436

[2] R. Bittner, J. Salamon, M. Tierney, M. Mauch, C. Cannam and J. P. Bello, "MedleyDB: A Multitrack Dataset for Annotation-Intensive MIR Research", in 15th International Society for Music Information Retrieval Conference, Taipei, Taiwan, Oct. 2014.

[3] John Ashley Burgoyne, Jonathan Wild, and Ichiro Fujinaga, 'An Expert Ground Truth Set for Audio Chord Recognition and Music Analysis', in *Proceedings of the 12th International Society for Music Information Retrieval Conference*, ed. Anssi Klapuri and Colby Leider (Miami, FL, 2011), pp. 633–38 [1].

Self-Supervised Learning

- Predict any part of the input from any other part.
- *Predict the future from the past.*
- *Predict the future from the recent past.*
- *Predict the past from the present.*
- *Predict the top from the bottom.*
- **Predict the occluded from the visible.**
- **Pretend there is a part of the input you don't know and predict that.**



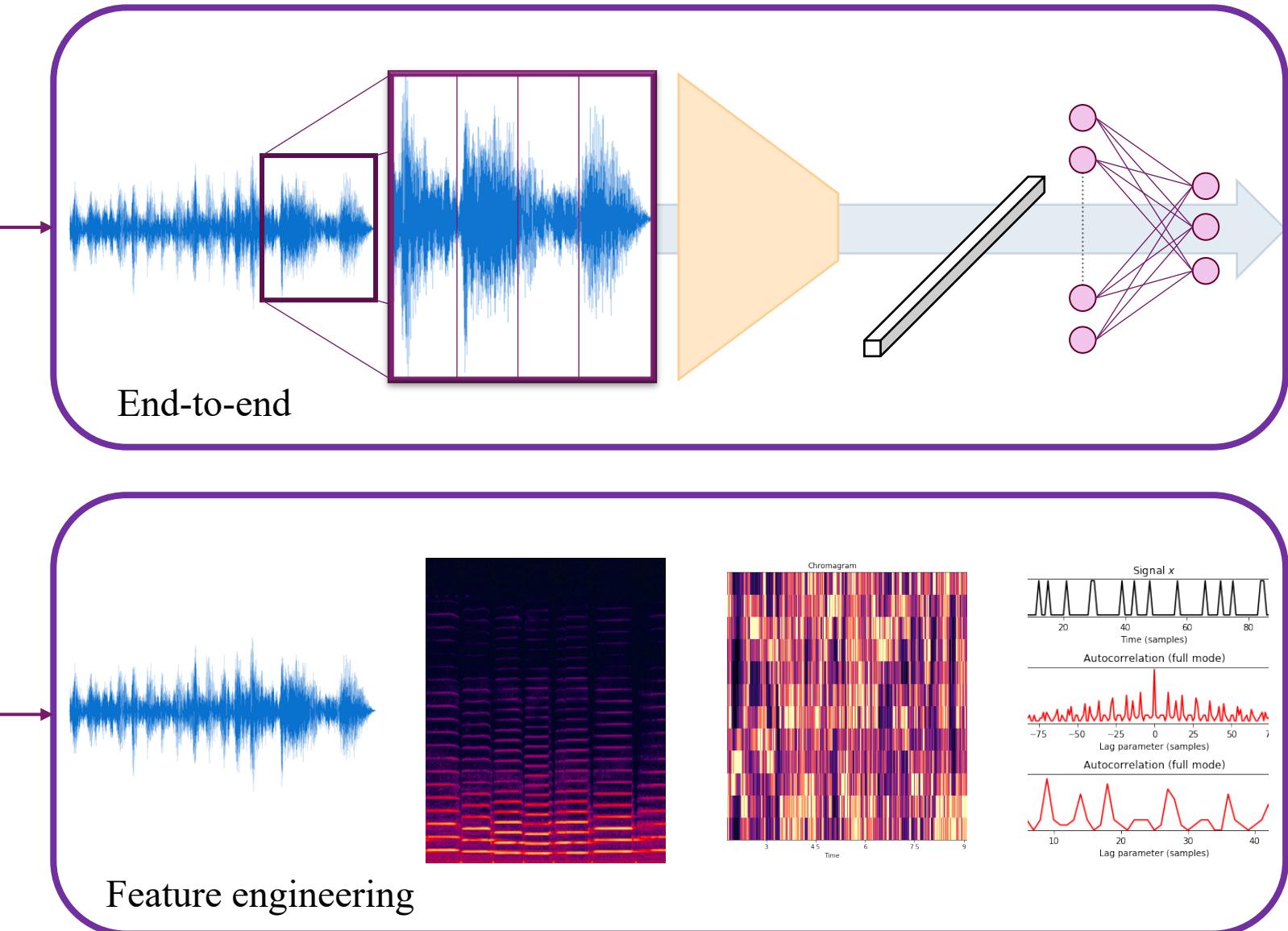
Slide: LeCun

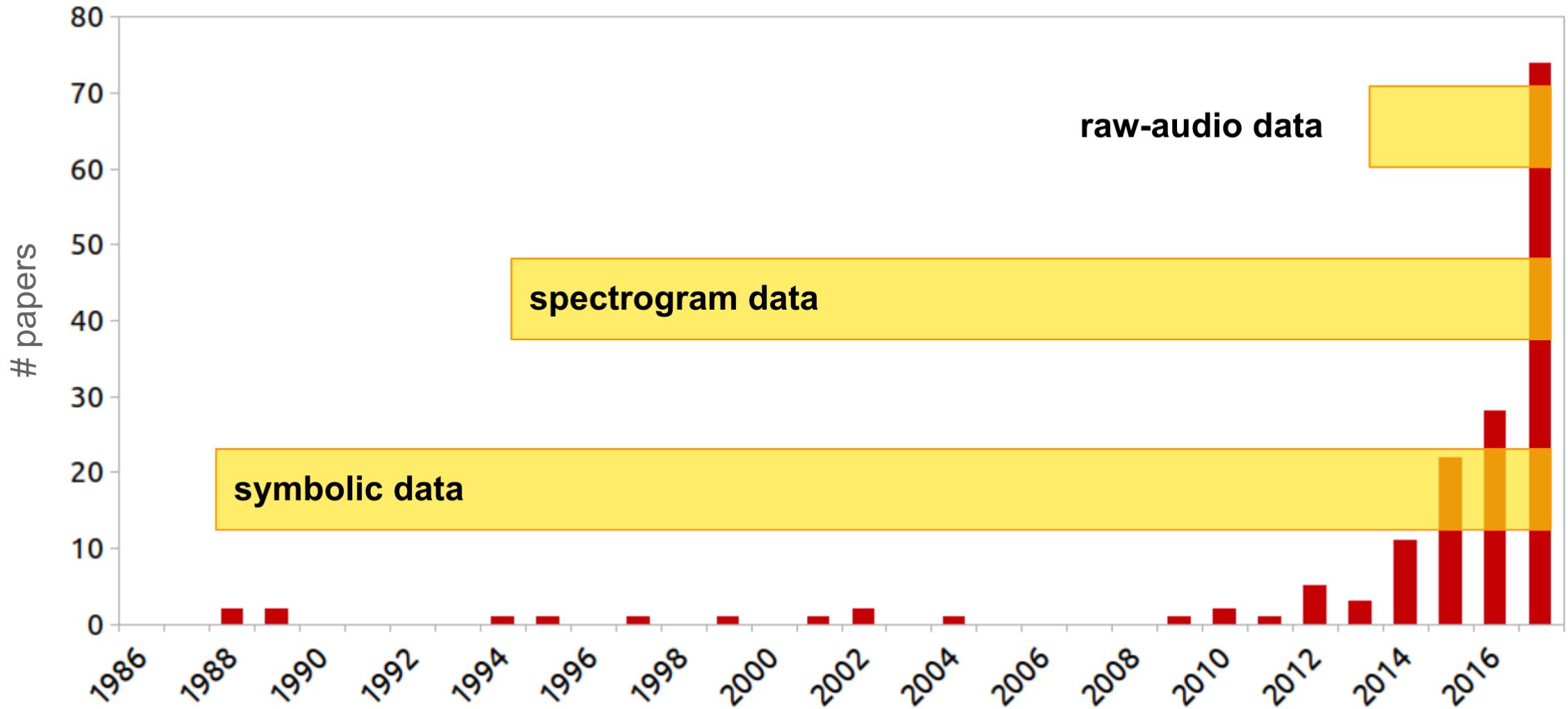
Self-Supervised Learning

- Learn representations without ground truth.
- Pre-text tasks.
- Methodological improvement for training on smaller datasets.
- Simple and straightforward pre-training.

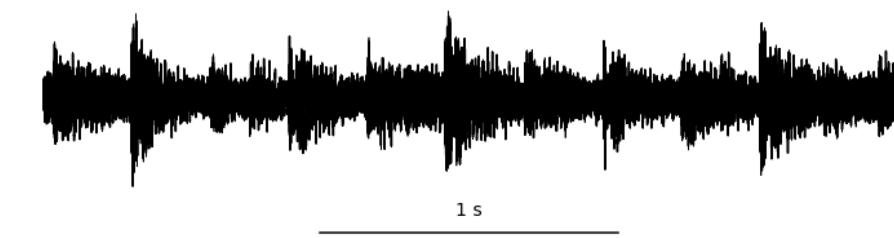
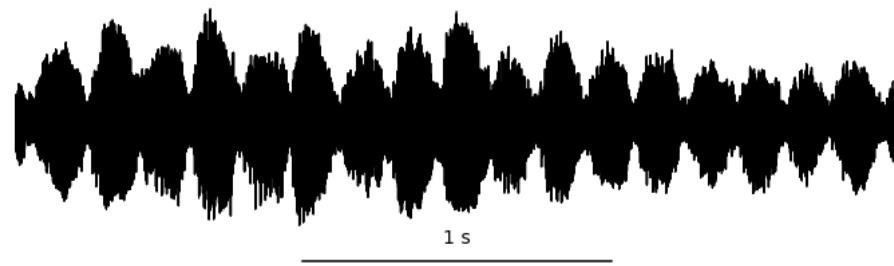
- Explore the limits of “Simple and straightforward” in MIR?

Background





Why SSL & Music Classification? (in the time-domain)

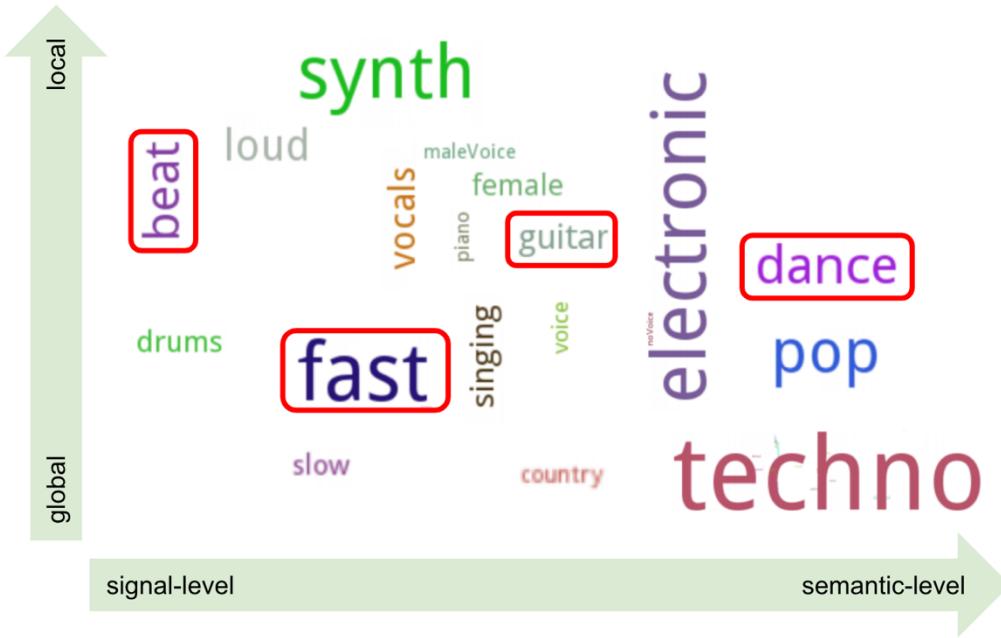


- High-dimensional
- Variable-length
- Missing hierarchical structure

But:

- No fine-tuning of pre-processing pipeline
- Learn expressive representations
- More challenging

Why SSL & Music Classification?



- Highly diverse
- Different levels of abstraction
- Multi-timbre
- Polyphonic
- Test on small and large scale
- Evaluate **versatility** of learned representations:
 - Genre
 - Instrumentation
 - Dynamics

Research Questions

RQ1: Are self-supervised learning methods effective in tagging raw audio waveforms of music?

RQ2: Do stronger data augmentations lead to more robust audio features?

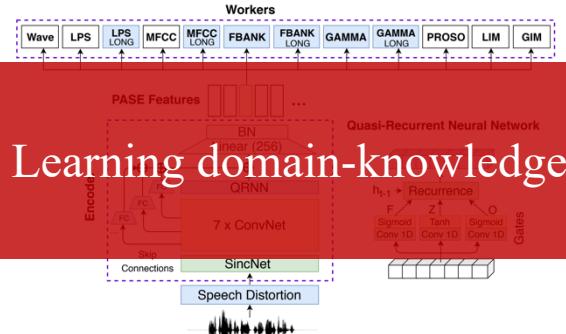
RQ3: Do these methods enable efficient classification for smaller datasets?

RQ4: Do these methods capture important, musical features, that are transferable to out-of-domain datasets?

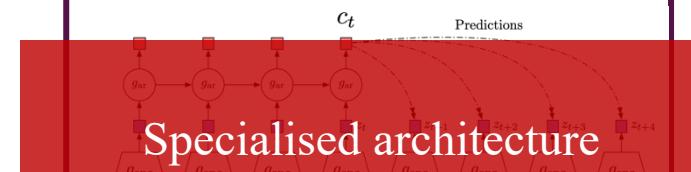
2. Related Work

Related Work (self-supervised learning)

PASE / PASE+



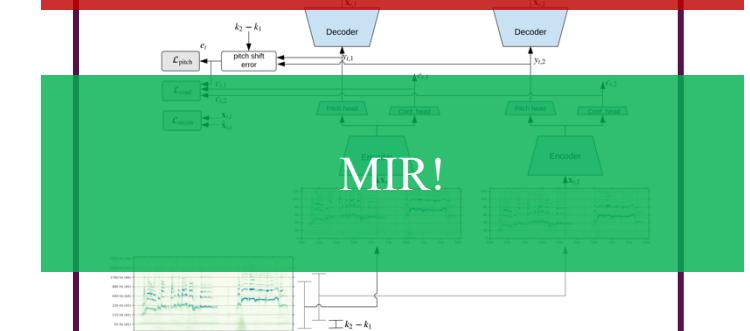
CPC



MIR!

SPICE

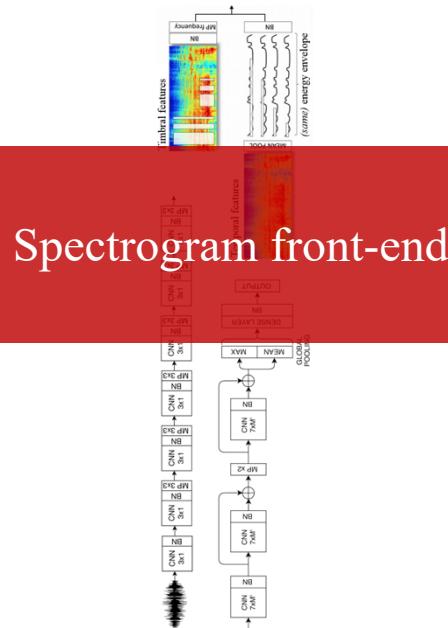
Too specific pre-text task



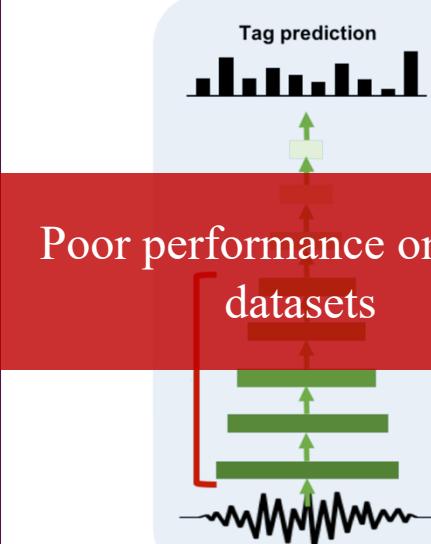
Time-frequency domain

Related Work (Music Classification)

Pons et al.
(Musicnn)



SampleCNN



Poor performance on small datasets

Related Work

Method	No domain knowledge injection	Performs well on small datasets	Time-domain	No specialised architecture	Performs MIR task
PASE / PASE+	-	+	+	+	-
CPC	+	+	+	-	+/-
SPICE	-	+	-	+	+
Pons et al. (Musicnn)	-	-	+/-	-	+
SampleCNN	+	-	+	+	+
CLMR	+	+	+	+	+

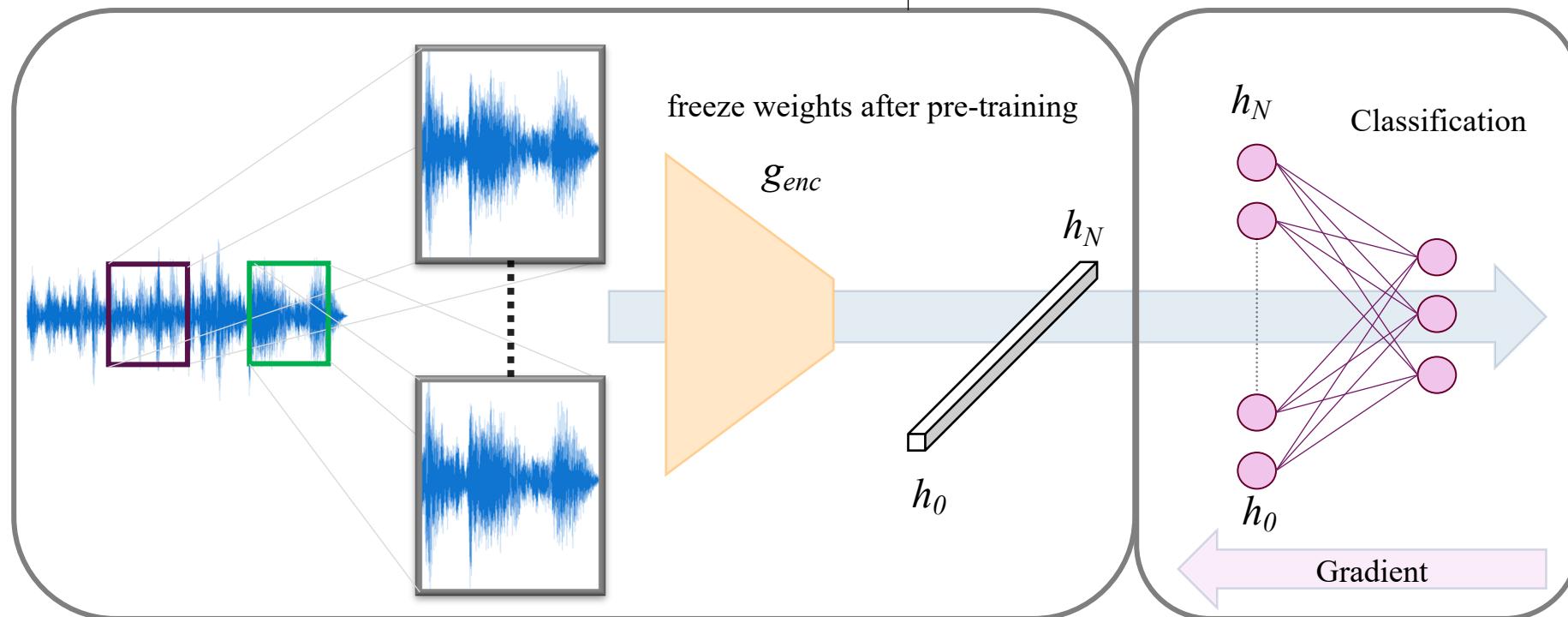
3. Method

CLMR

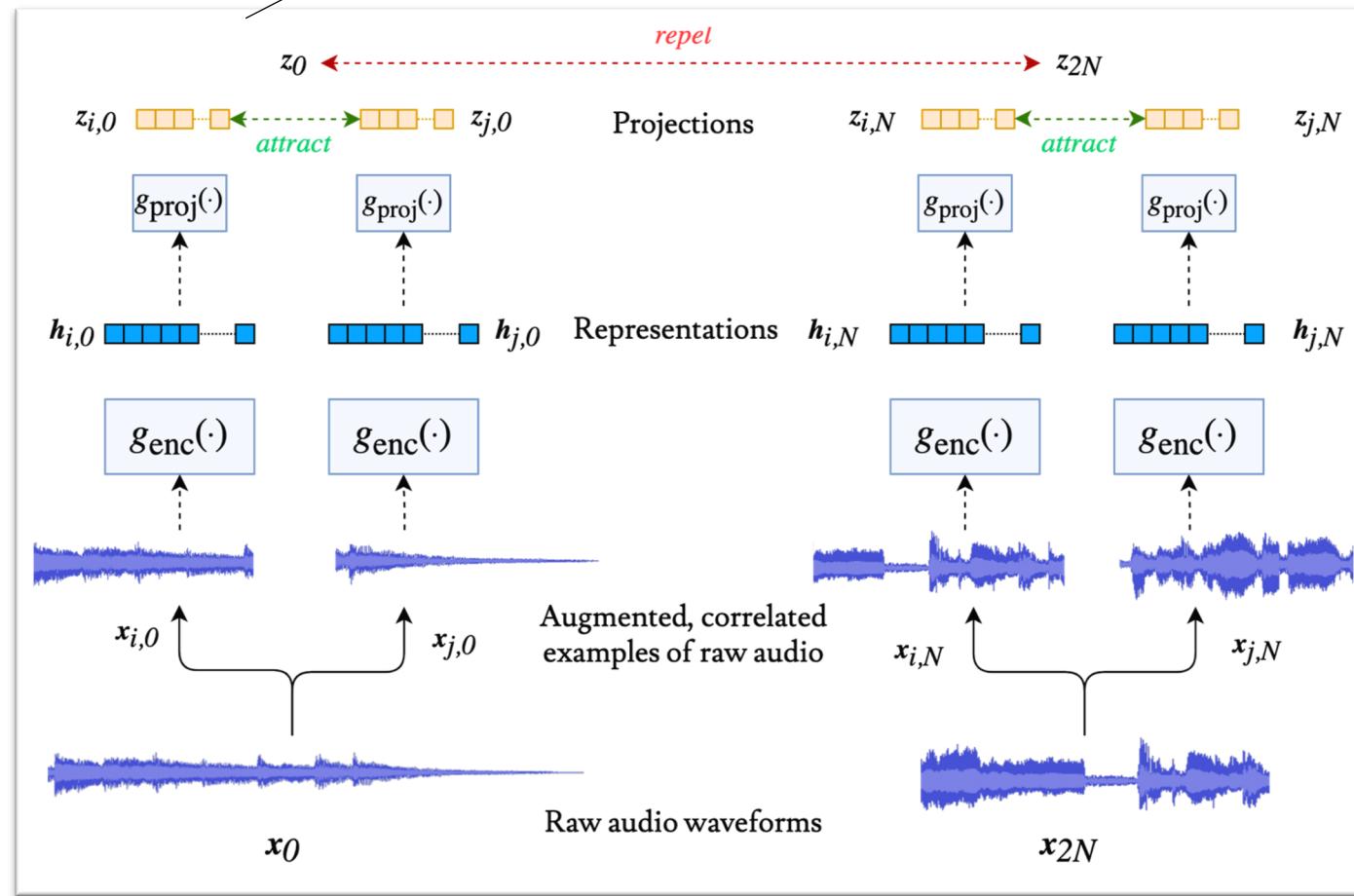
Use SimCLR^[1] with extended
SampleCNN as g_{enc}

Stage 1:
Pre-train Feature Extractor

Stage 2:
Fine-tune linear classifier



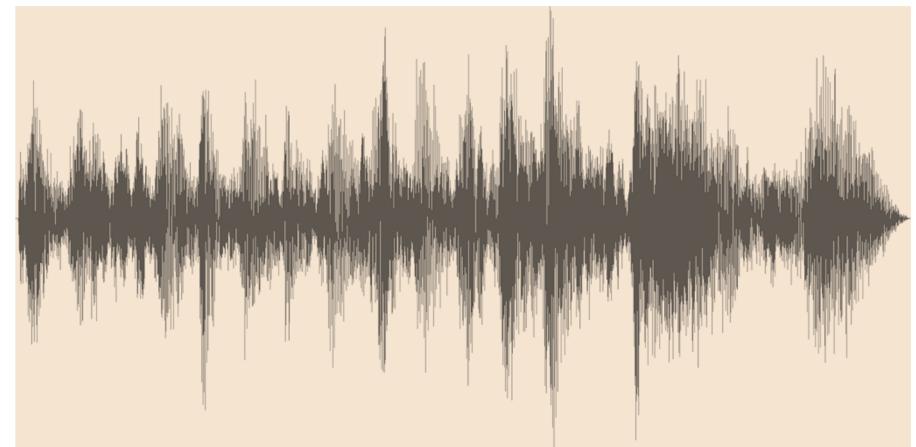
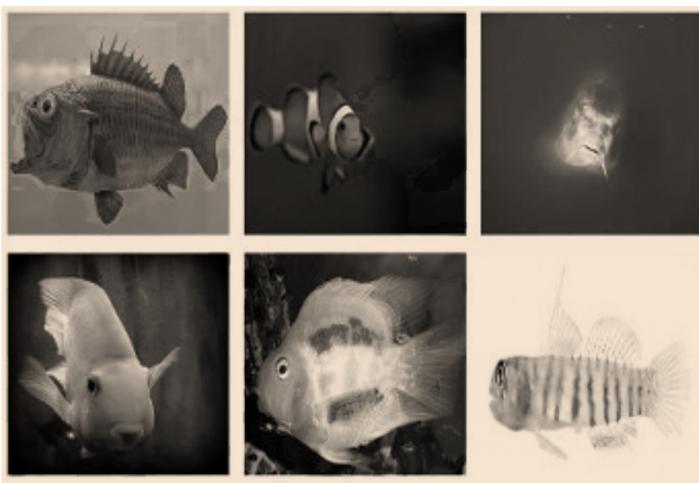
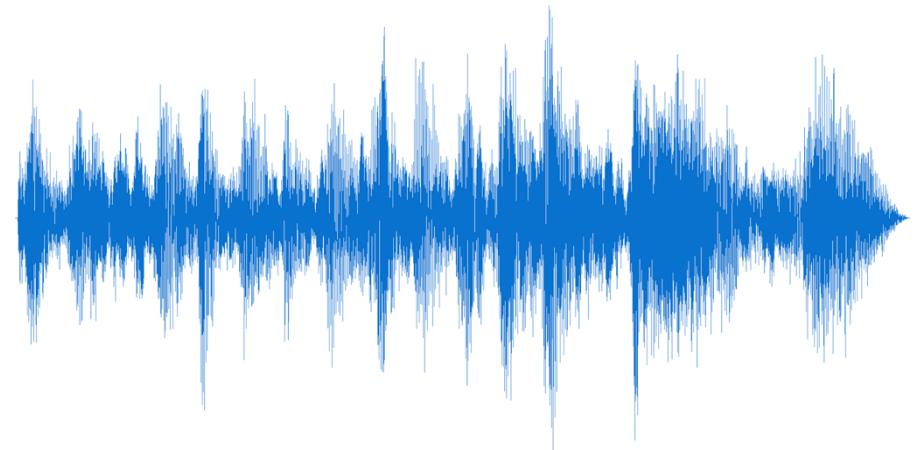
SimCLR



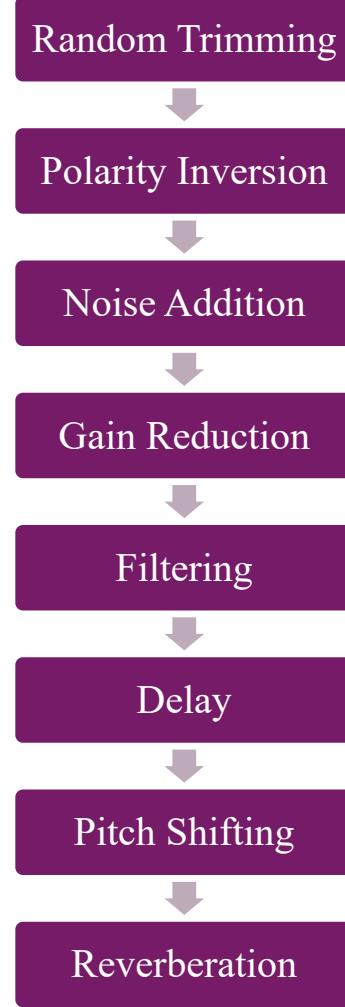
- Data augmentations.
- Extract representations with SampleCNN encoder.
- Project representations.
- Cosine similarity
- Contrastive loss

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(z_i, z_j) / \tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(z_i, z_k) / \tau)}$$

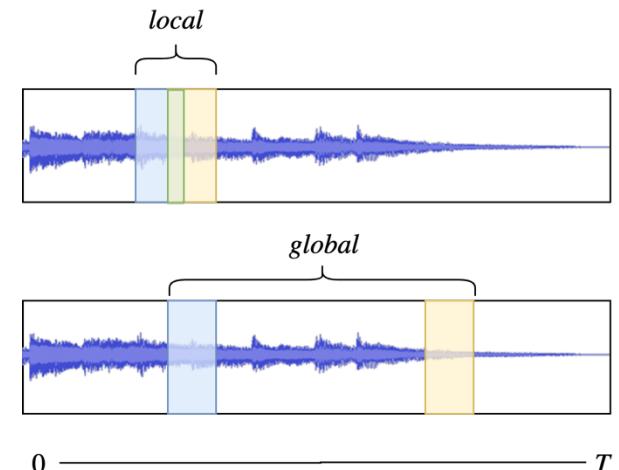
Images != Audio



Method



- Infer local/global structure.
- Phase-awareness
- Robustness to noisy signals
- Loudness / Dynamics
- Frequency spectrum
- Repeating signals
- Pitch-awareness
- Room acoustics

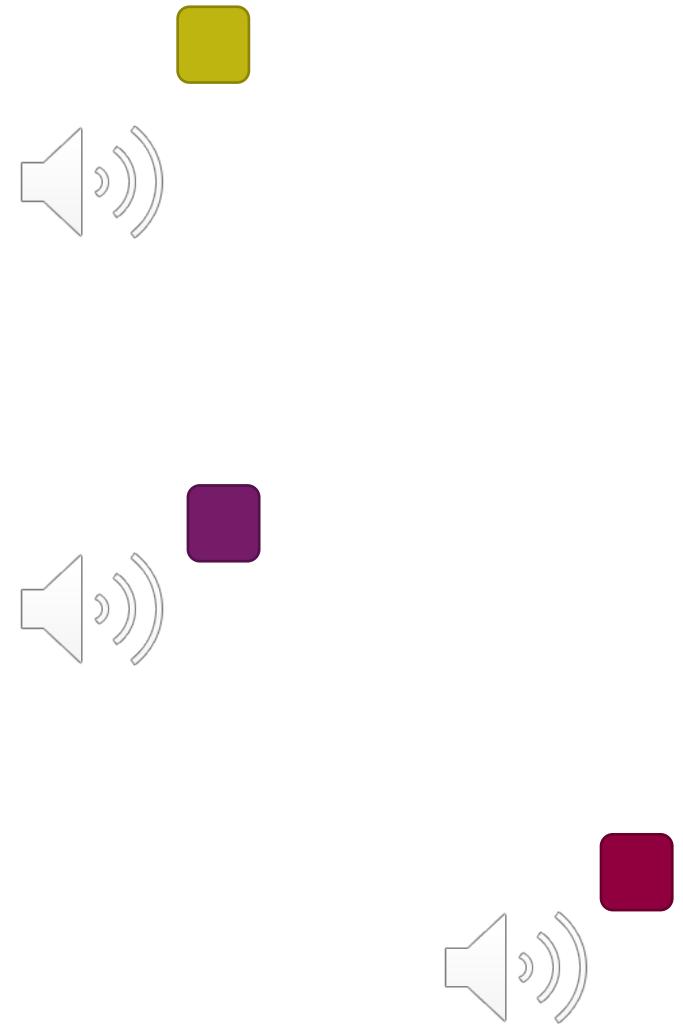
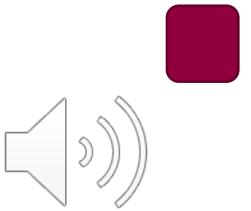


Examples

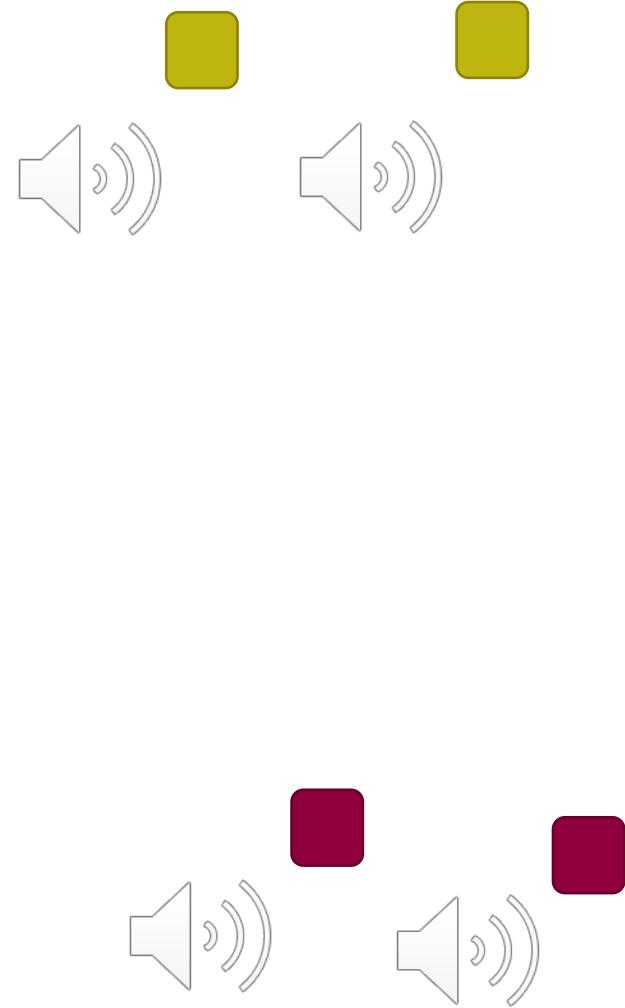
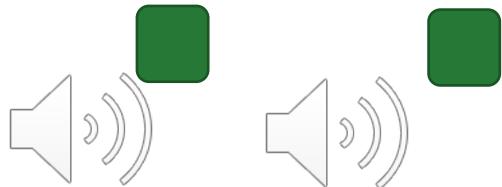
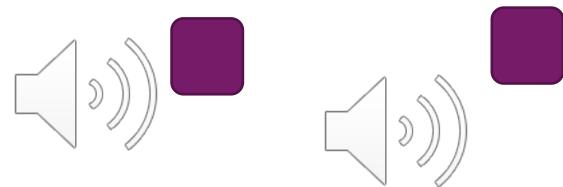


Each example has one other, correlated view

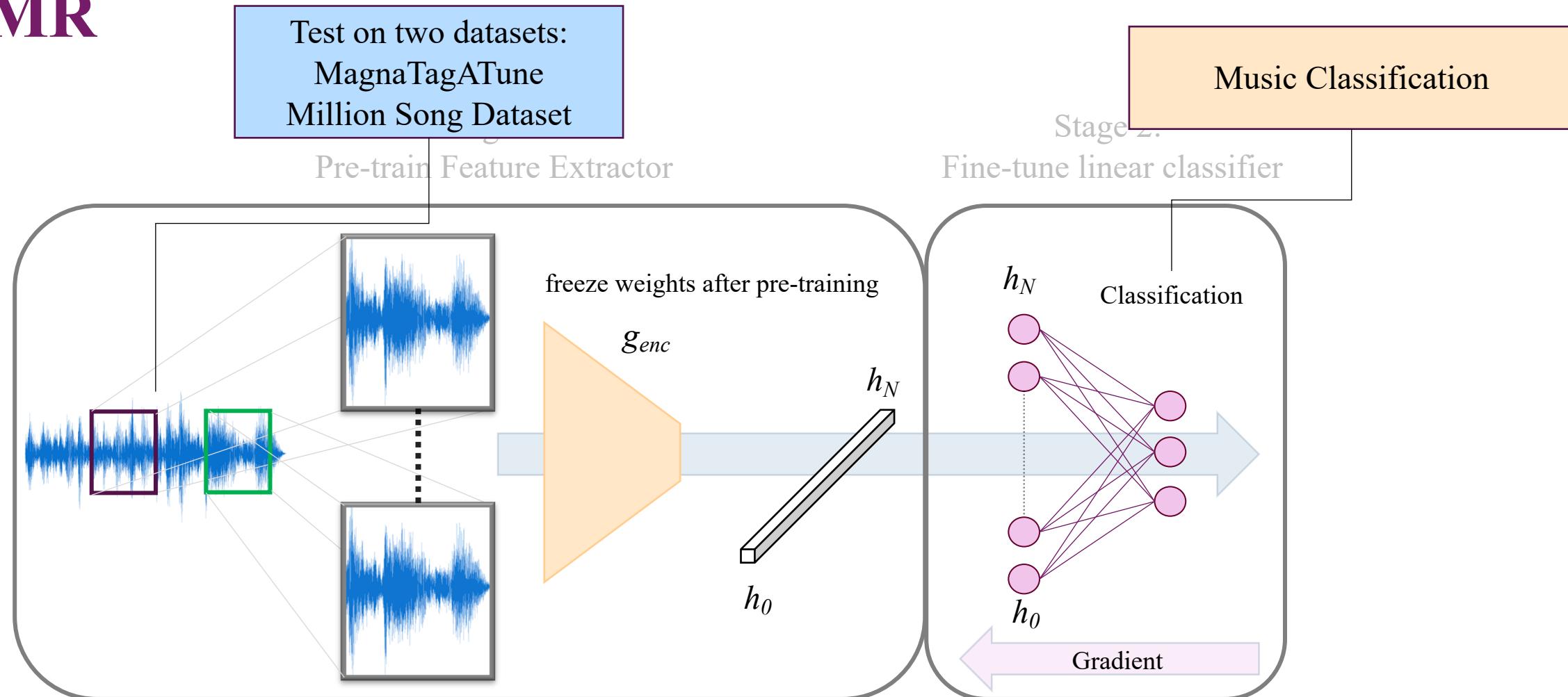
Examples



Examples



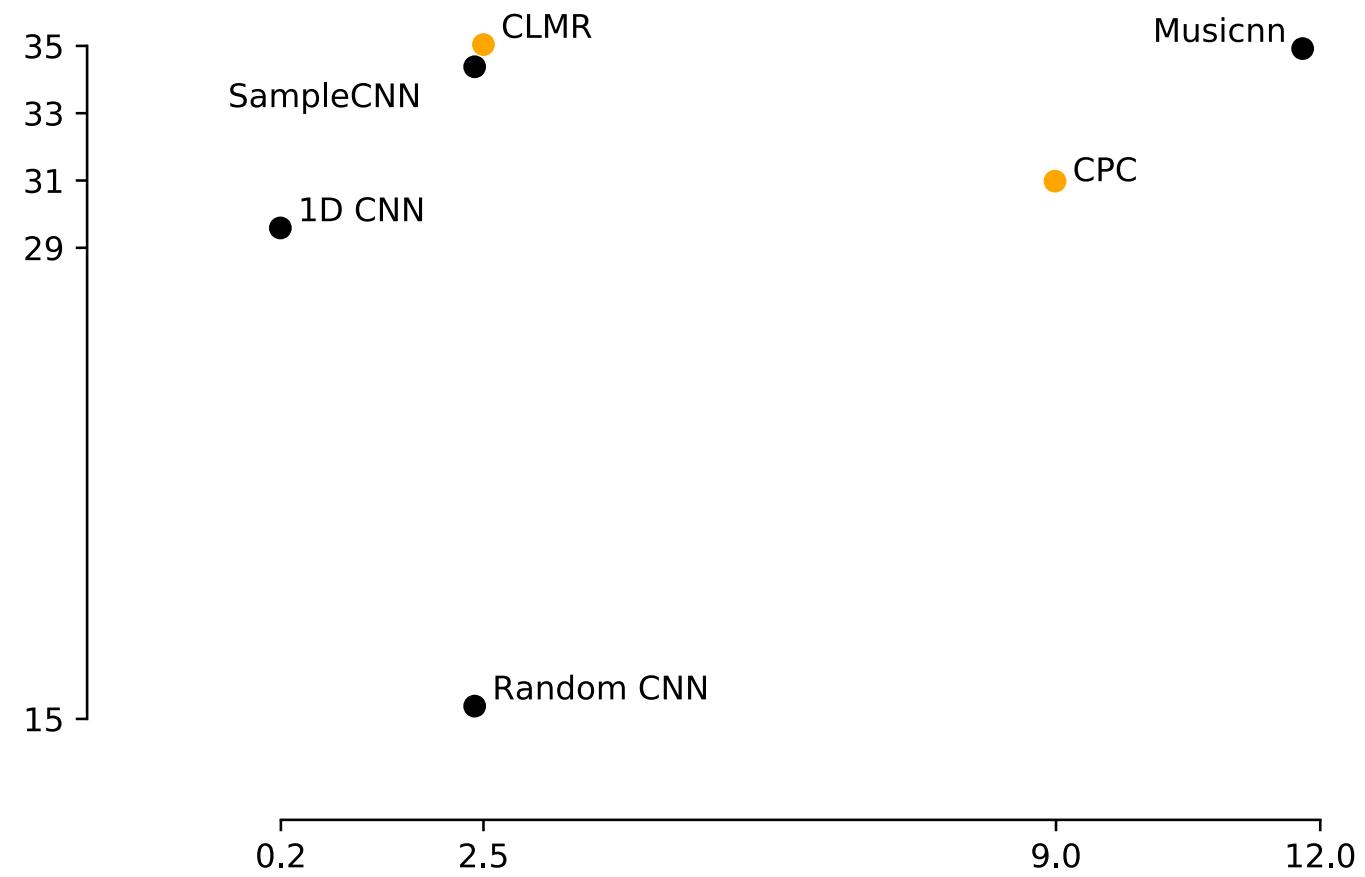
CLMR



4. Experiments

Experiment 1: Self-Supervised Music Tagging

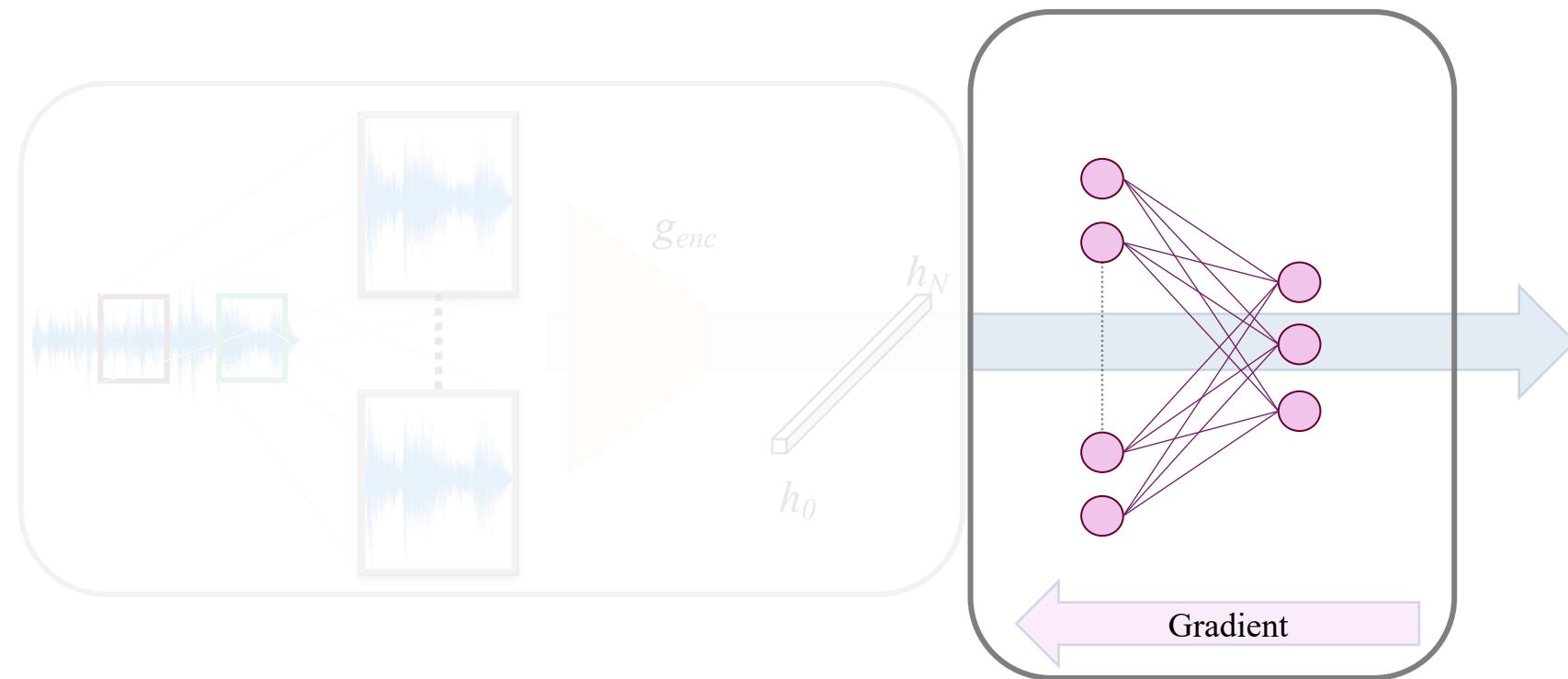
- MagnaTagATune Dataset
- 25.863 unique songs
- Top-50 tags



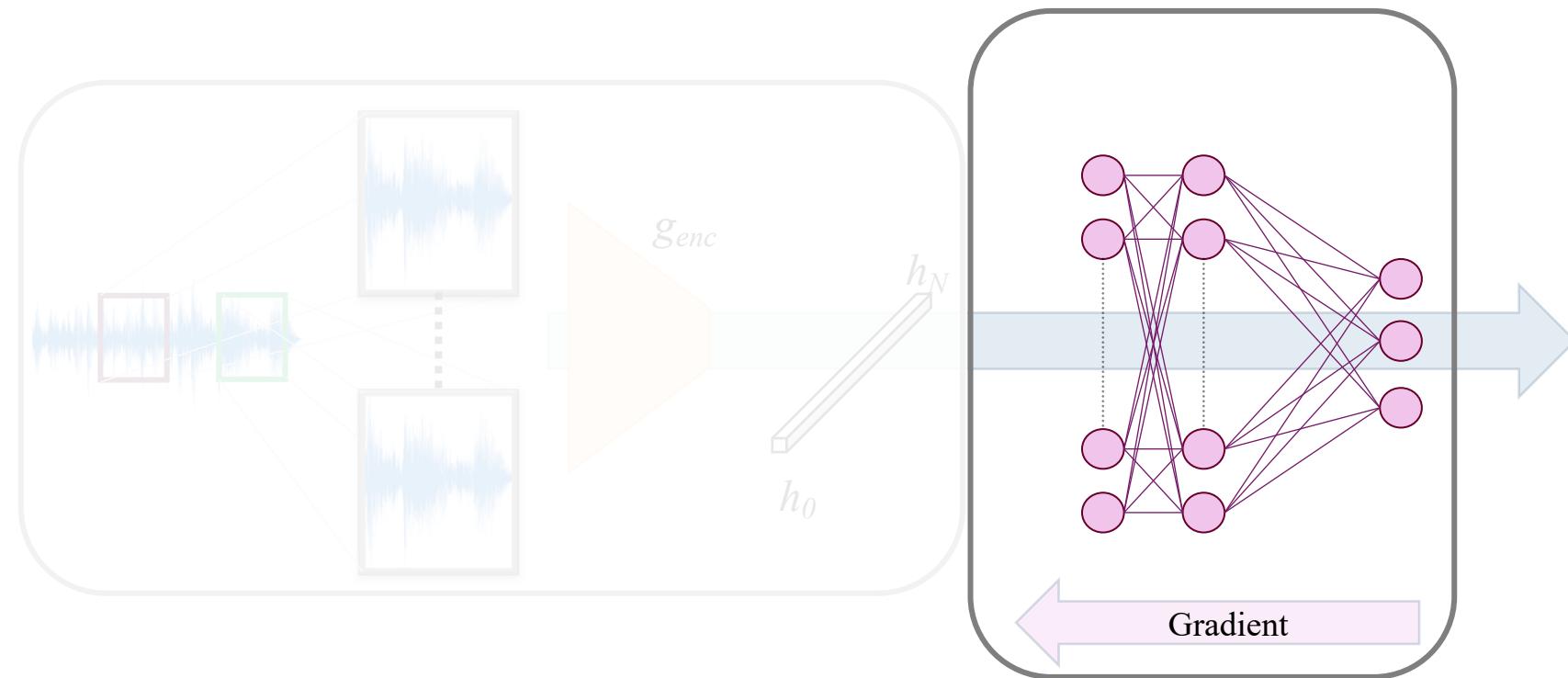
Experiment 1: Self-Supervised Music Tagging

Model	Dataset	ROC-AUC _{TAG}	PR-AUC _{TAG}
CLMR (ours)	MTAT	88.49 (89.25)	35.37 (35.89)
Pons et al. [†]	MTAT	89.05	34.92
SampleCNN [†]	MTAT	88.56	34.38
CPC (ours)	MTAT	86.60 (87.99)	30.98 (33.04)

Experiment 1: Self-Supervised Music Tagging



Experiment 1: Self-Supervised Music Tagging



Experiment 1: Self-Supervised Music Tagging

Model	Dataset	ROC-AUC _{TAG}	PR-AUC _{TAG}
CLMR (ours)	MTAT	88.49 (89.25)	35.37 (35.89)
Pons et al. [†]	MTAT	89.05	34.92
SampleCNN [†]	MTAT	88.56	34.38
CPC (ours)	MTAT	86.60 (87.99)	30.98 (33.04)

Experiment 1: Self-Supervised Music Tagging

Model	Dataset	ROC-AUC _{TAG}	PR-AUC _{TAG}
CLMR (ours)	MTAT	88.49 (89.25)	35.37 (35.89)
Pons et al. [†]	MTAT	89.05	34.92
SampleCNN [†]	MTAT	88.56	34.38
CPC (ours)	MTAT	86.60 (87.99)	30.98 (33.04)

Experiment 1: Self-Supervised Music Tagging

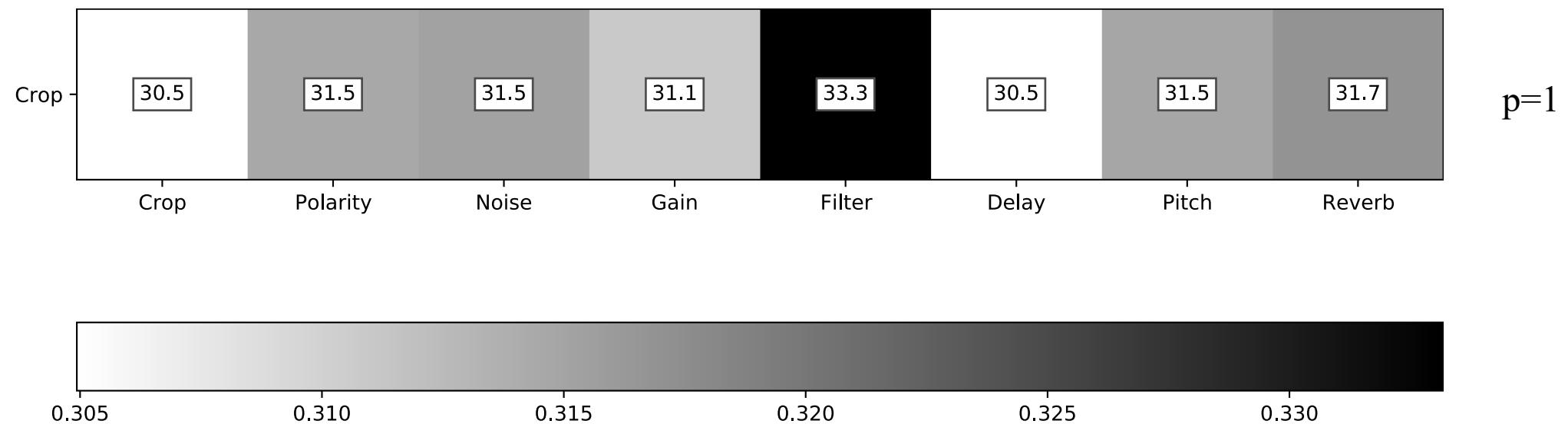
Model	Dataset	ROC-AUC _{TAG}	PR-AUC _{TAG}
CLMR (ours)	MTAT	88.49 (89.25)	35.37 (35.89)
Pons et al. [†]	MTAT	89.05	34.92
SampleCNN [†]	MTAT	88.56	34.38
CPC (ours)	MTAT	86.60 (87.99)	30.98 (33.04)
1D CNN [†]	MTAT	85.58	29.59

Experiment 1: Self-Supervised Music Tagging

- Million Song Dataset
- 241.904 unique songs
- Top-50 tags

Model	Dataset	ROC-AUC _{TAG}	PR-AUC _{TAG}
Pons et al. [†]	MSD	87.41	28.53
SampleCNN [†]	MSD	88.42	-
CLMR (ours)	MSD	85.66	24.98

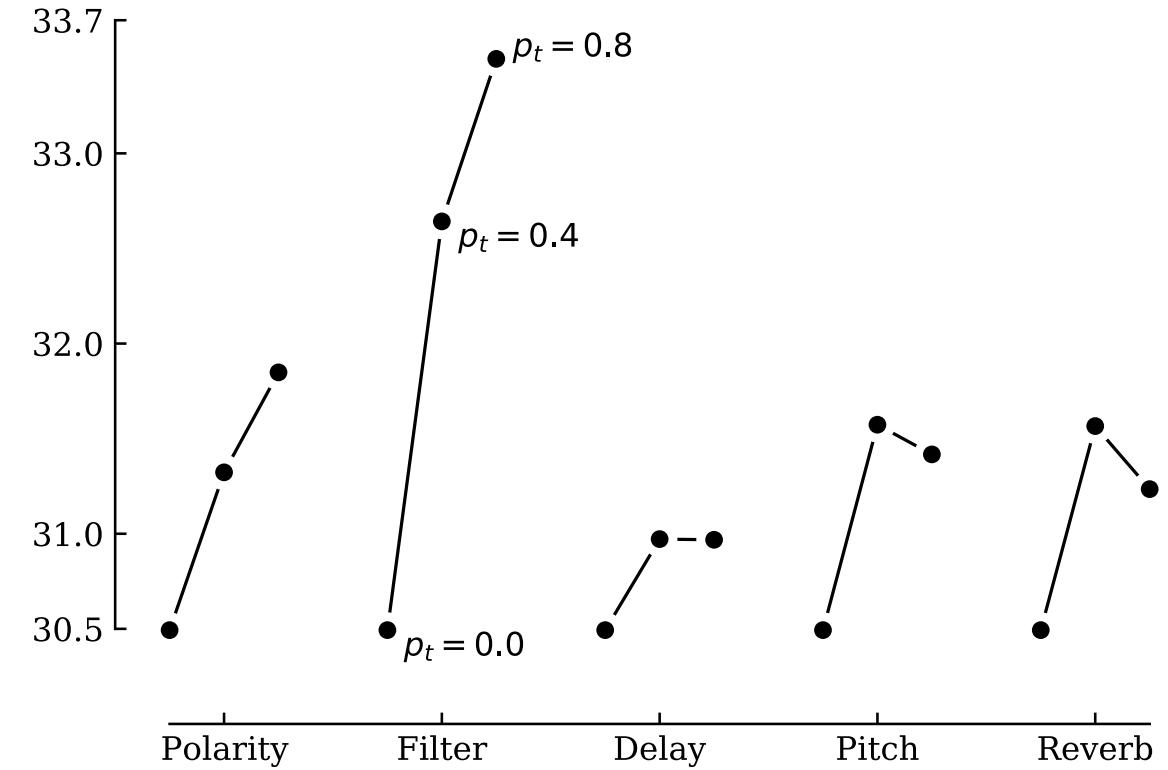
Experiment 2: Stochastic Data Augmentations



Augment 1 sample, not the other.

Experiment 2: Stochastic Data Augmentations

- Strong augmentations improve downstream task performance.
- Filtering augmentation is most effective.
- Pitch and Reverb augmentations can be too strong.



Experiment 3: Efficient Classification

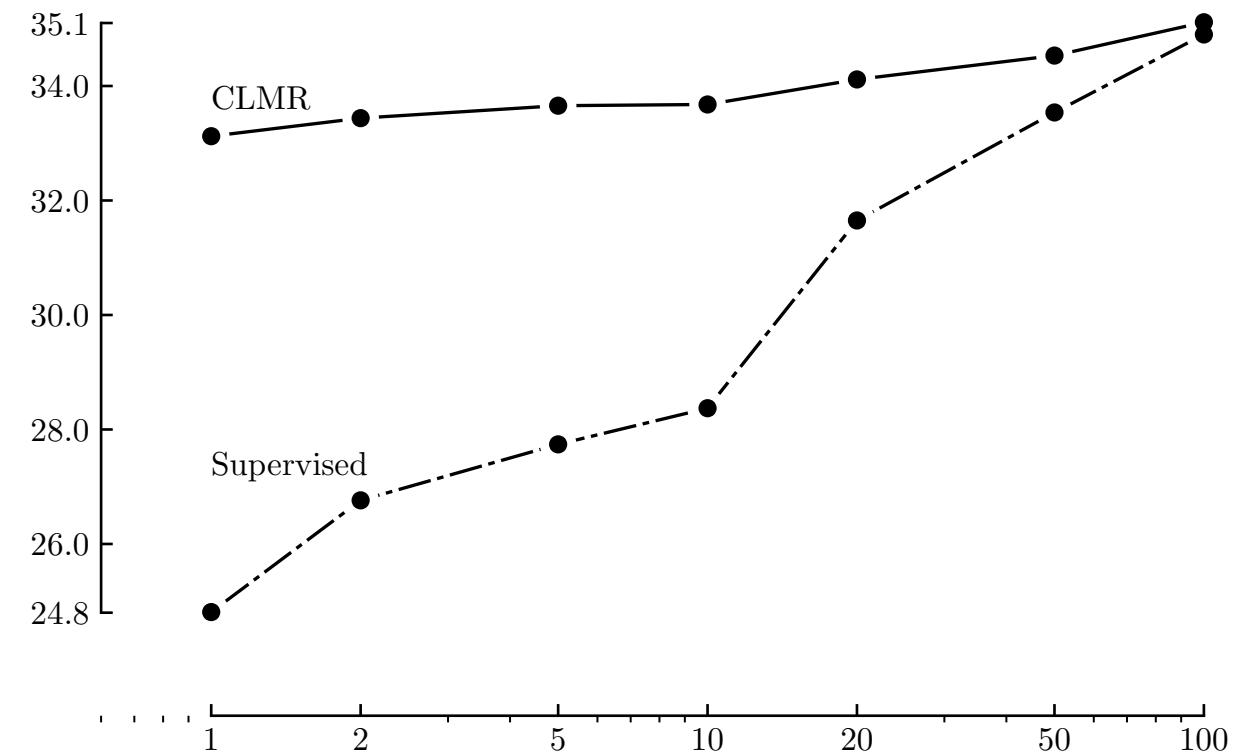
- Pre-train on all available data.
- Freeze feature extractor weights.
- Fine-tune linear classifier on 1% of the labeled data (train split).
- Test on complete test split.

Experiment 3: Efficient Classification

MagnaTagATune Dataset:

- CLMR performs better with less labeled data.
- Even with 1% of labeled data, it performs competitively.

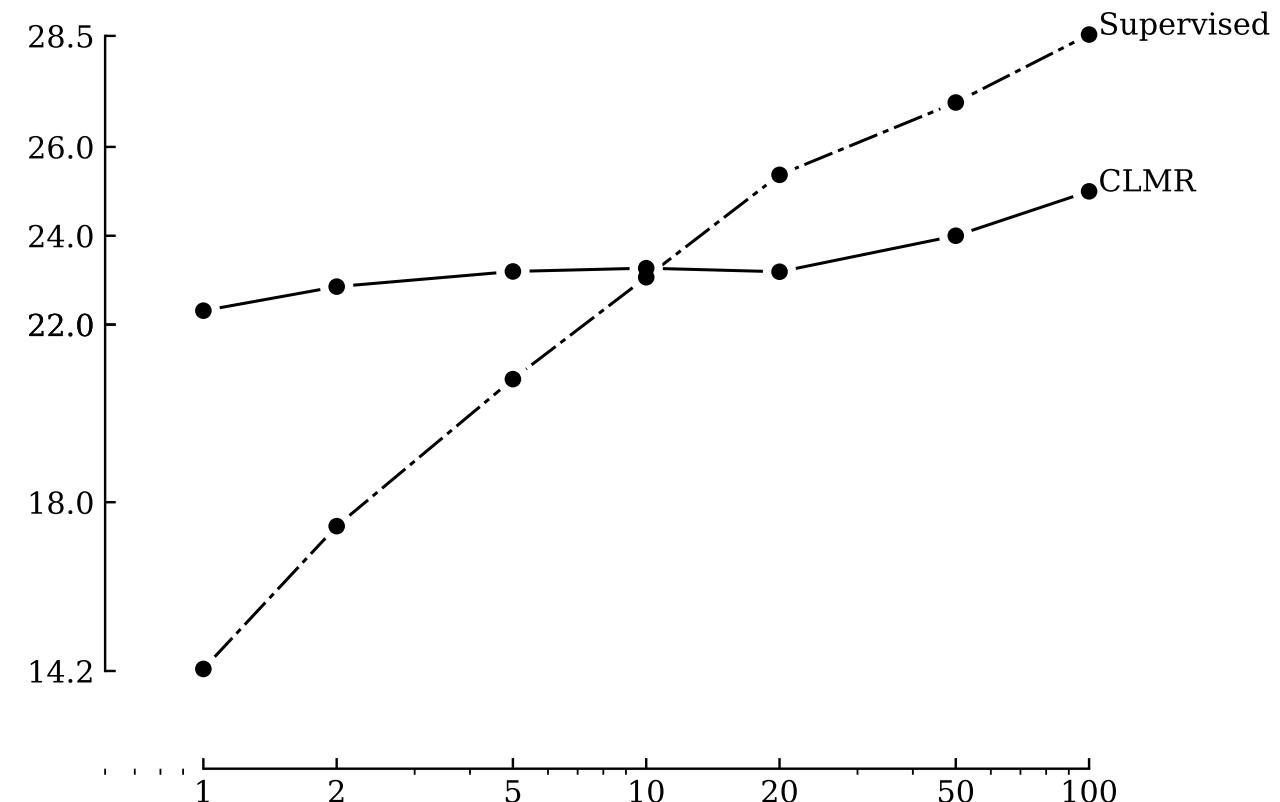
(1% = 259 songs)



Experiment 3: Efficient Classification

Million Song Dataset:

- CLMR performs better with fewer labeled datapoints.
- Supervised end-to-end model exceeds CLMR > 10% labeled data.
- ($10\% = 24.190$ songs)



Experiment 4: Transfer Learning

- Features are transferable from out-of-domain datasets.
- Reasonable performance.
- CPC performs better on smaller datasets (GTZAN / Billboard)

Model	Train Dataset	Eval. Dataset	ROC-AUC	PR-AUC
CLMR	MSD	MTAT	86.57	32.04
CPC	FMA	MTAT	86.34 (87.79)	30.71 (32.47)
CLMR	FMA	MTAT	86.22 (86.63)	30.58 (31.22)
CPC	Billboard	MTAT	85.78 (86.25)	29.68 (30.15)
CPC	GTZAN	MTAT	83.44 (86.06)	26.88 (29.72)
CLMR	Billboard	MTAT	82.73 (84.22)	26.86 (27.82)
CLMR	GTZAN	MTAT	81.88 (85.43)	26.18 (29.49)

Experiment 4: Transfer Learning

Model	Train Dataset	Eval. Dataset	ROC-AUC	PR-AUC
CLMR	MSD	MTAT	86.57	32.04
CPC	FMA	MTAT	86.34 (87.79)	30.71 (32.47)
CLMR	FMA	MTAT	86.22 (86.63)	30.58 (31.22)
CPC	Billboard	MTAT	85.78 (86.25)	29.68 (30.15)
CPC	GTZAN	MTAT	83.44 (86.06)	26.88 (29.72)
CLMR	Billboard	MTAT	82.73 (84.22)	26.86 (27.82)
CLMR	GTZAN	MTAT	81.88 (85.43)	26.18 (29.49)

Experiment 5: Mini-Batch Sizes

- Training benefits from larger batch-sizes, but:
- Task may be too hard to:
 - Infer positive pair from 2.6 second long audio fragments.
 - In a pool of 912 negative samples
- May require longer training

Mini-batch Size	ROC-AUC _{TAG}	PR-AUC _{TAG}	ROC-AUC _{CLIP}	PR-AUC _{CLIP}
456	88.13	34.87	92.96	68.90
96	88.49	35.11	93.07	69.20
48	87.91	34.56	92.88	68.75

Experiment 6: Training Duration

Epochs	ROC-AUC _{TAG}	PR-AUC _{TAG}	ROC-AUC _{CLIP}	PR-AUC _{CLIP}
10 000	88.47 (89.25)	35.37 (35.89)	93.16 (93.48)	69.32 (70.03)
3 000	88.49 (88.94)	35.11 (35.46)	93.07 (93.27)	69.20 (69.74)
1 000	88.31 (88.64)	34.40 (34.86)	92.89 (93.08)	68.59 (69.15)

- Longer training yields better results.
- Especially when adding extra hidden layer.
- In line with findings from SimCLR.

Experiment 6: Training Duration

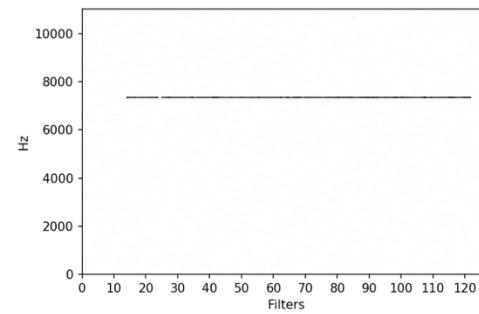
Epochs	ROC-AUC _{TAG}	PR-AUC _{TAG}	ROC-AUC _{CLIP}	PR-AUC _{CLIP}
10 000	88.47 (89.25)	35.37 (35.89)	93.16 (93.48)	69.32 (70.03)
3 000	88.49 (88.94)	35.11 (35.46)	93.07 (93.27)	69.20 (69.74)
1 000	88.31 (88.64)	34.40 (34.86)	92.89 (93.08)	68.59 (69.15)

- Longer training yields better results.
- Especially when adding extra hidden layer.
- In line with findings from SimCLR.

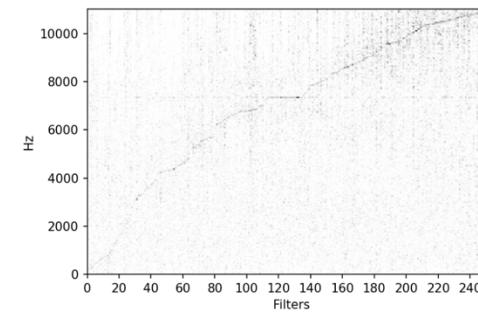
5. Interpretability

Interpretability: Filter Visualisation

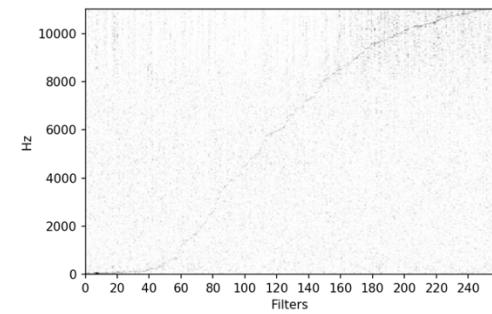
- Gradient ascent on a random waveform of 729 samples.
- Magnitude spectrum.
- Vertical line is a magnitude spectrum of a single filter.
- Sorted by frequency magnitude.



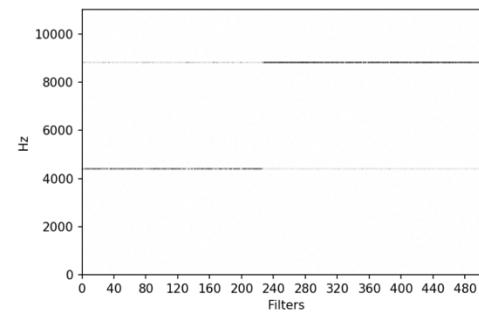
(a) $\text{CLMR}_{\text{MTAT}}^{(1)}$



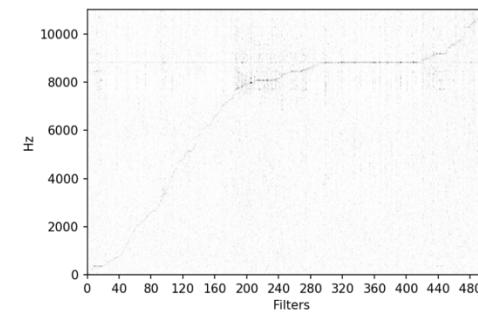
(b) $\text{CLMR}_{\text{MTAT}}^{(4)}$



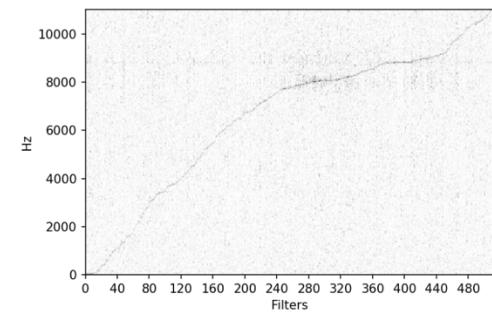
(c) $\text{CLMR}_{\text{MTAT}}^{(6)}$



(d) $\text{CPC}_{\text{MTAT}}^{(1)}$



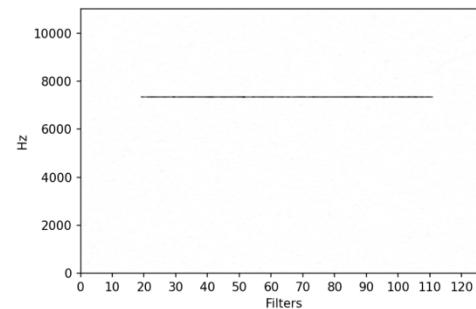
(e) $\text{CPC}_{\text{MTAT}}^{(4)}$



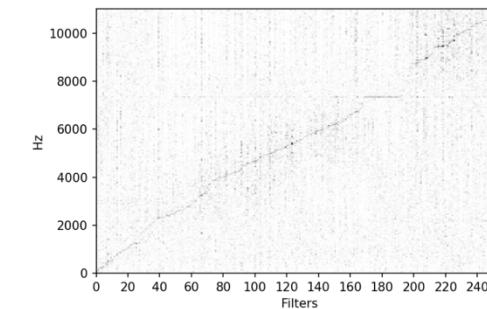
(f) $\text{CPC}_{\text{MTAT}}^{(6)}$

Interpretability: Filter Visualisation

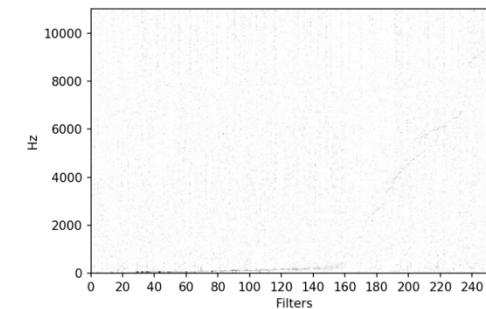
- Sensitive to a single band of frequencies around 7500 Hz.
- Higher layers: spread first linearly, then non-linearly across the full range.



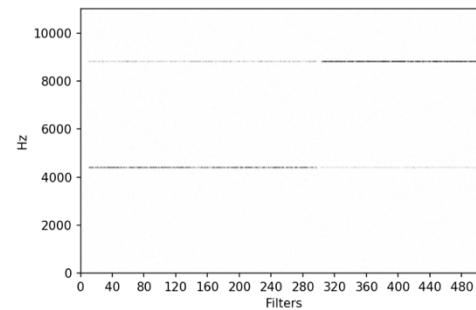
(a) CLMR⁽¹⁾_{Billboard}



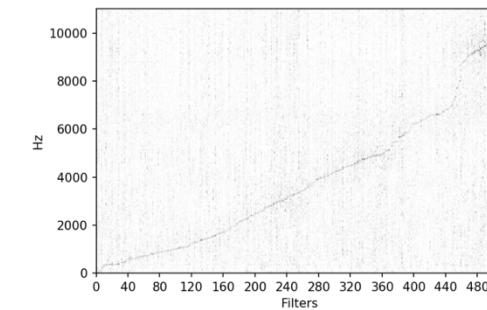
(b) CLMR⁽⁴⁾_{Billboard}



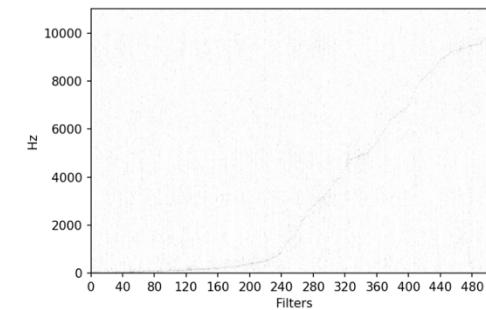
(c) CLMR⁽⁶⁾_{Billboard}



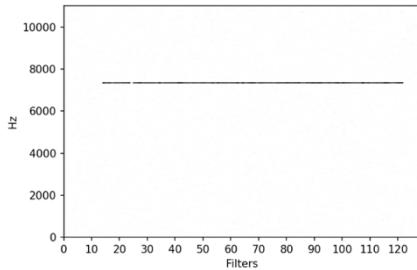
(d) CPC⁽¹⁾_{Billboard}



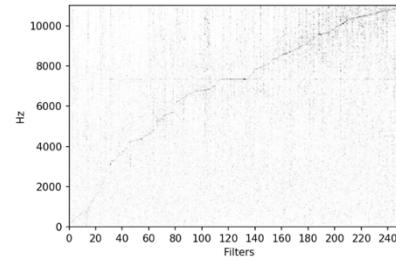
(e) CPC⁽⁴⁾_{Billboard}



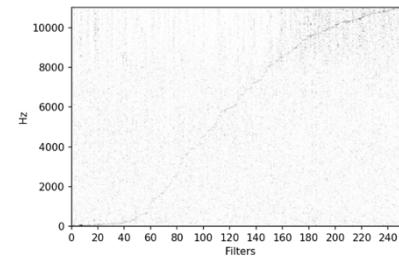
(f) CPC⁽⁶⁾_{Billboard}



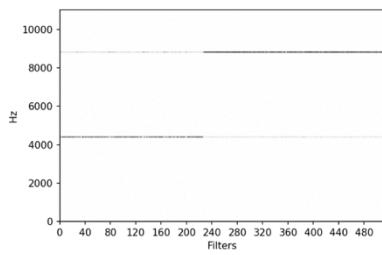
(a) CLMR⁽¹⁾_{MTAT}



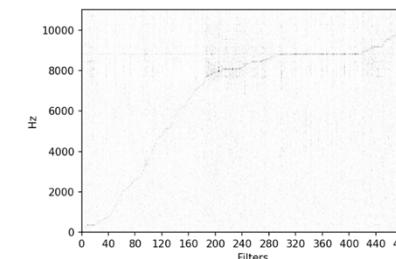
(b) CLMR⁽⁴⁾_{MTAT}



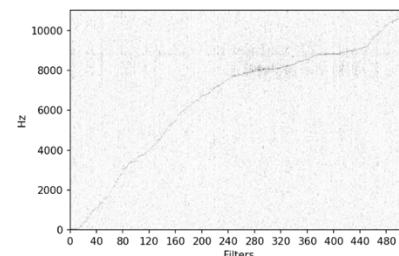
(c) CLMR⁽⁶⁾_{MTAT}



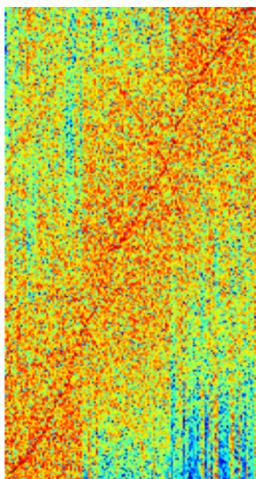
(d) CPC⁽¹⁾_{MTAT}



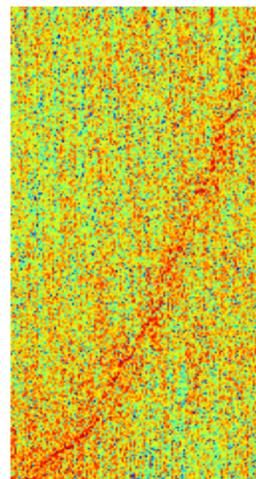
(e) CPC⁽⁴⁾_{MTAT}



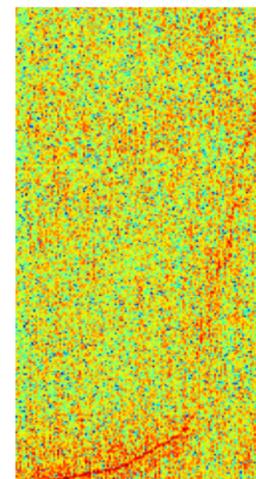
(f) CPC⁽⁶⁾_{MTAT}



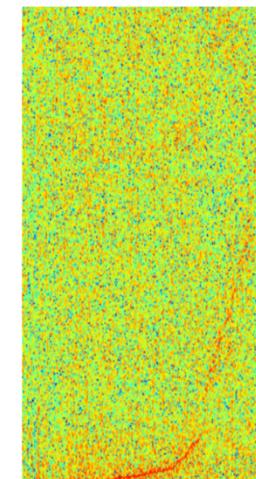
Layer 1



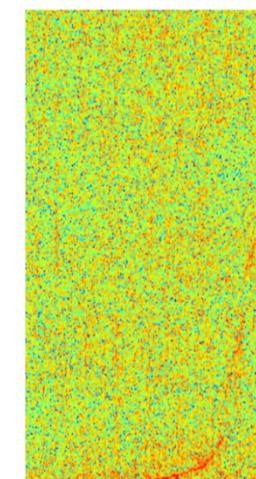
Layer 2



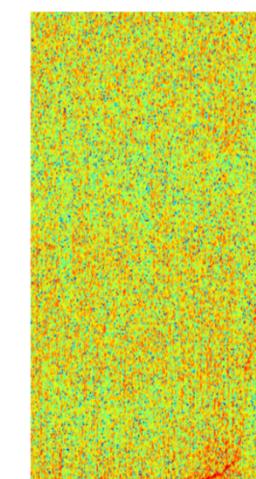
Layer 3



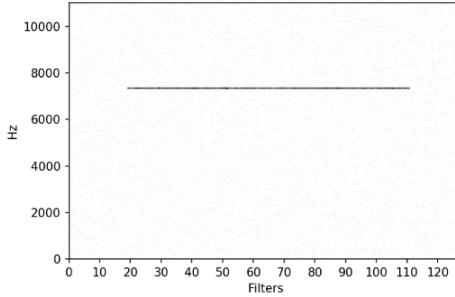
Layer 4



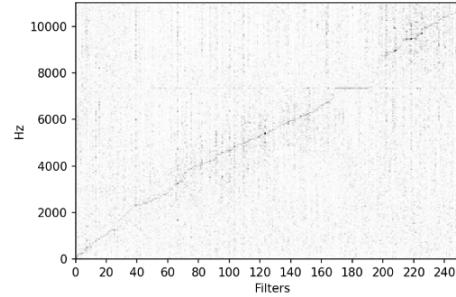
Layer 5



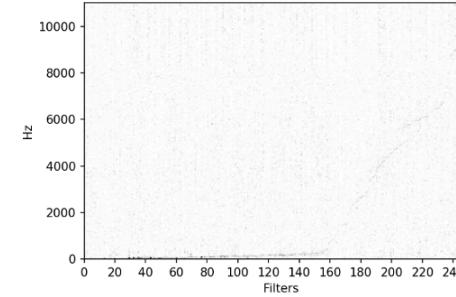
Layer 6



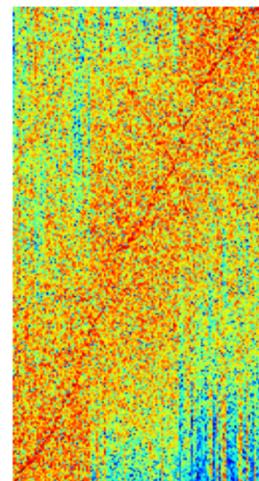
(a) CLMR⁽¹⁾_{Billboard}



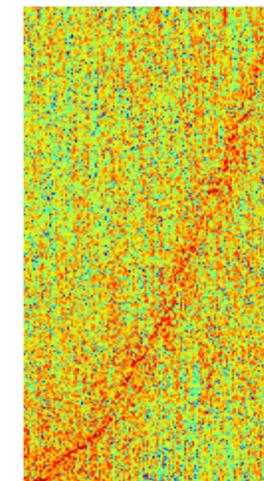
(b) CLMR⁽⁴⁾_{Billboard}



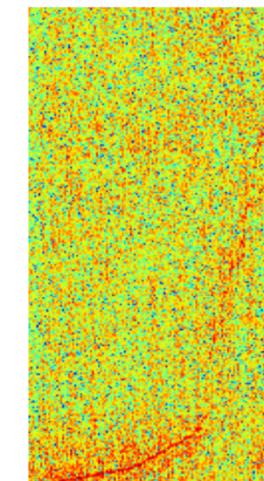
(c) CLMR⁽⁶⁾_{Billboard}



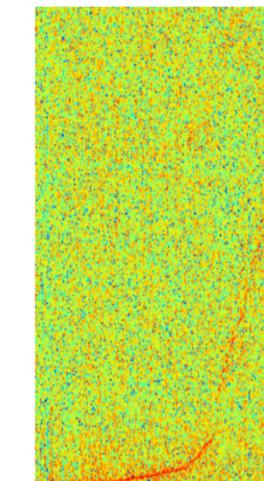
Layer 1



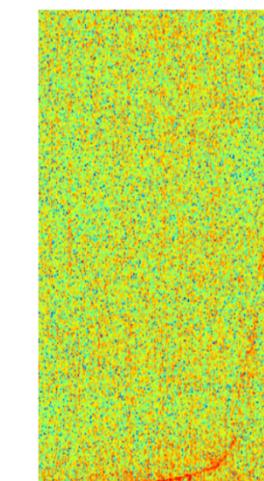
Layer 2



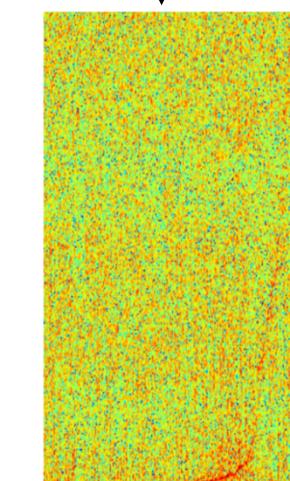
Layer 3



Layer 4



Layer 5



Layer 6



Interpretability: Listening Experiment

- Extract activation of each clip from last, self-supervised trained encoder layer.
- Sort, for each filter, by activation value.
- Compare music clips from the same filter, but different datasets.
- Qualitatively test out-of-domain generalisability.
- Ability to group specific timbre, pitch and loudness features.

CLMR Listening Experiment

Encoder

Feature number (0 - 511 hidden layers), comma separated for multiple:

Finetuned linear classifier

MagnaTagATune Tags

guitar classical slow techno strings drums electronic rock fast piano ambient beat violin vocal synth female indian opera male singing vocals no vocals harpsichord loud quiet flute woman male vocal no vocal pop soft sitar solo man classic choir voice new age dance male voice female vocal beats harp cello no voice weird country metal female voice choral

Million Song Dataset Tags

rock pop alternative indie electronic female vocalists dance 00s alternative rock jazz beautiful metal chillout male vocalists classic rock soul indie rock Mellow electronica 80s folk 90s chill instrumental punk oldies blues hard rock ambient acoustic experimental female vocalist guitar Hip-Hop 70s party country easy listening sexy catchy funk electro heavy metal Progressive rock 60s rnb indie pop sad House happy

Show entries

0	15	30	100	200	432	idx	audio	track_id	clip_id	segment	labels
0.0223766192793 84613	2.7158913612365 723	0.1848115473985 672	0.9317127466201 782	0.1930894404649 7345	0	390	<input type="button" value="▶ 0:00 / 0:29"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	42303	0	classic	
0.5030207037925 72	2.6778771877288 82	2.5957636833190 92	0	0	2.3766314983367 92	299	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	32469	0	techno,dance	
0.0268725045025 34866	2.4450995922088 623	0.9709233641624 451	0.2756210267543 793	0.1755639463663 1012	0.0468806102871 89484	442	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	48201	0	slow,violin,no vocals,solo	
0	2.3582718372344 97	0.0356949642300 6058	0	0	0	274	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	29956	0	slow,electronic,a mbient,quiet	
0	2.3395388126373 29	0.19784429696938 0188	0.1692339628934 8602	0.6273652911186 218	0	523	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	57403	0		
0	2.2877655029296 875	0.4496681392192 8406	0.0970177352428 4363	0	0	437	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	47623	0	slow,choir,choral	
0	2.2091059684753 42	0.2901597023010 254	0.1528758704662 323	0.2854320406913 7573	0.0223385598510 50377	515	<input type="button" value="▶ 0:00"/> <input type="button" value=" "/> <input type="button" value="⋮"/>	56493	0	slow,vocal,woman	

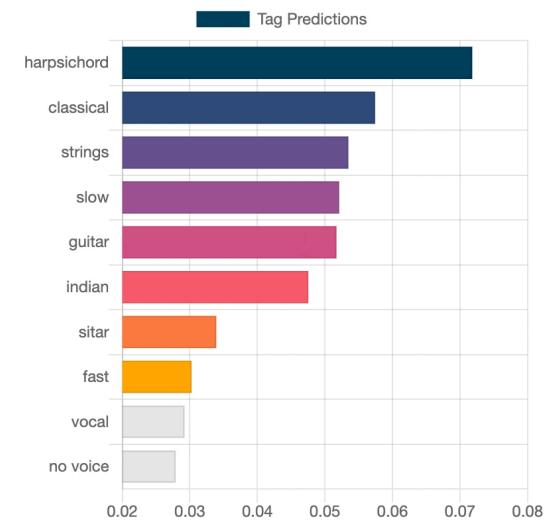
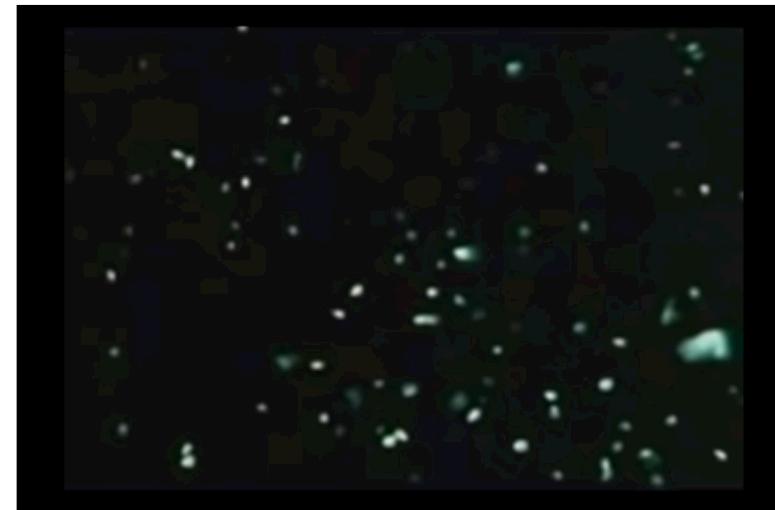
Interpretability: Listening Interface

- Listen to out-of-domain classifications
- Qualitatively analyse what the network is doing.

YouTube link

<https://www.youtube.com/watch?v=N1tTN-b5KHg>

Submit



6. Discussion

Discussion

RQ1: Are self-supervised learning methods effective in tagging raw audio waveforms of music?

Yes, they perform better on one dataset compared to fully supervised models, but perform less, however still competitively, on a larger dataset.

Discussion

RQ2: Do stronger data augmentations lead to more robust audio features?

Yes, stronger and chained data augmentations lead to a better downstream music classification performance.

We noticed large performance improvements using:

- Polarity inversion
- Filtering
- Pitch shifting
- Reverberation

Discussion

RQ3: Do these methods enable efficient classification for smaller datasets?

Yes, CLMR performs better on both datasets when training on less data.

On the MagnaTagATune dataset, the performance is still competitive using only 1% of the data (259 songs).

On the Million Song Dataset, fully supervised end-to-end models surpass CLMR with >10% of the labeled data.

Discussion

RQ4: Do these methods capture important, musical features, that are transferable to out-of-domain datasets?

Yes, we find that both CLMR and CPC get a reasonable performance on the evaluation dataset, despite being trained on out-of-domain data.

When pre-training on larger, out-of-domain corpora of music, we find that those models perform better.

Limitations & Future Work

- Self-supervision mainly constrained to mini-batch size and long training.
 - Momentum Contrast, BYOL.
 - Performance on larger audio datasets.
- Self-supervision models for other MIR tasks.
- Extend to time-frequency domain.
- Specialised fine-tune model.

CLMRv2 Private

CLMR version 2

 Python Updated 3 days ago

Conclusion

- No pre-processing and no ground-truth required to learn effective, musical representations of raw audio waveforms in a complex MIR task.
- Simple and straightforward.
- Efficient classification.
- *Linear* classification exceeds baselines for challenging MIR task.
- Out-of-domain transferability of representations.
- Foster reproducibility and self-supervised learning in MIR and beyond

Reproducibility & Self-Supervised Learning in MIR

 [Spijkervet / CLMR](#)

 MIT License

 8 stars  1 fork

 [Spijkervet / audio-augmentations](#)

Audio Augmentations library for PyTorch

 2 stars  0 forks

 [contrastive-predictive-coding](#)

PyTorch implementation of Representation Learning with Contrastive Predictive Coding by Van den Oord et al. (2018)

 Python  14  4

 [Spijkervet/SimCLR](#)

PyTorch implementation of SimCLR: A Simple Framework for Contrastive Learning of Visual Representations by T. Chen et al.

 Python  274  55

Thank you

- Ashley Burgoyne
- Wilker Aziz
- Jordan B.L. Smith
- Sander Dieleman
- Bas Veeling
- Sindy Löwe

Contrastive Loss

Maximised

$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2n} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

No higher than 1

Minimised

Sums over pairs:
Restricted from minimising the denominator or maximising the numerator, without doing the other as well.