University of Amsterdam
Faculty of Science

Music Cognition Group

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

Contrastive Learning of Musical Representations

Janne Spijkervet

Supervisor: Dr. J.A. Burgoyne

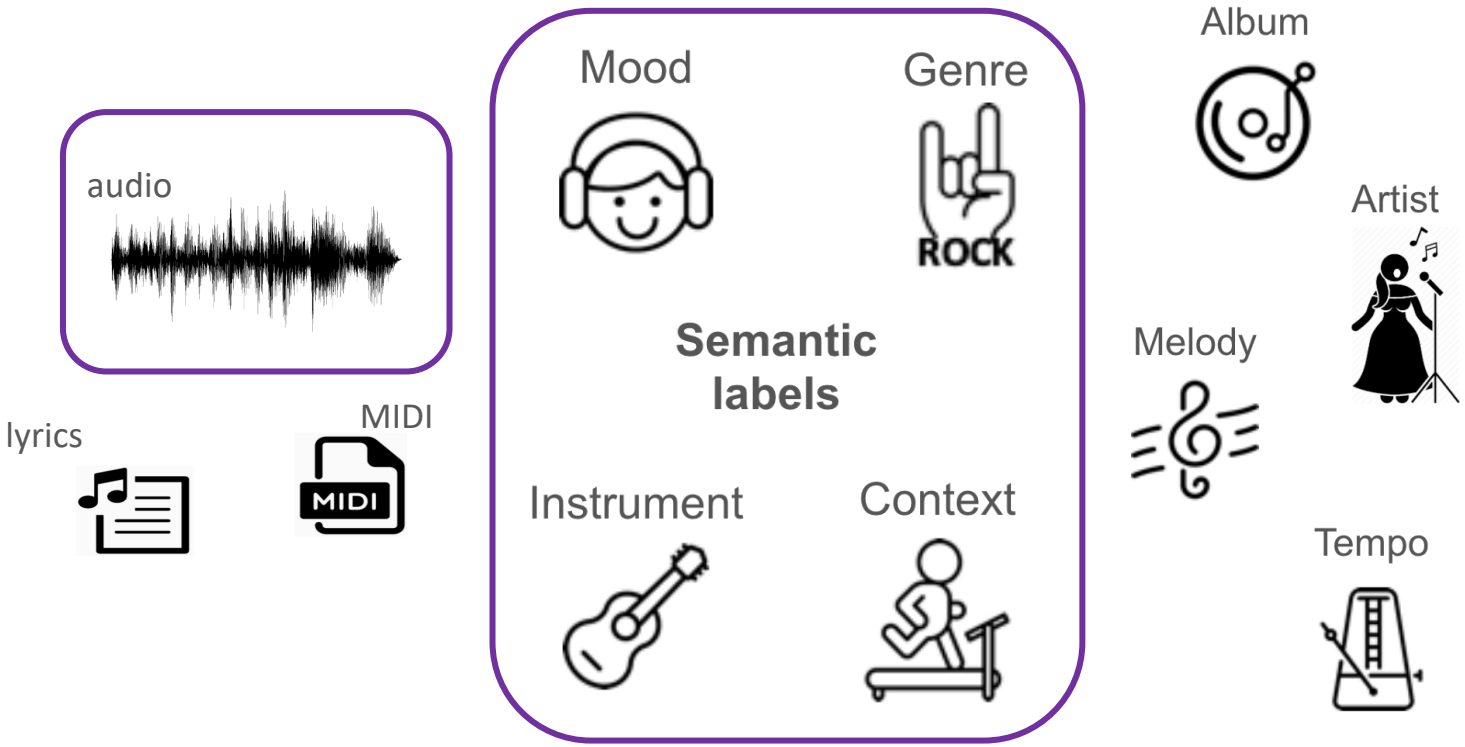
Examiner: Dr. W. Aziz

Schedule

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

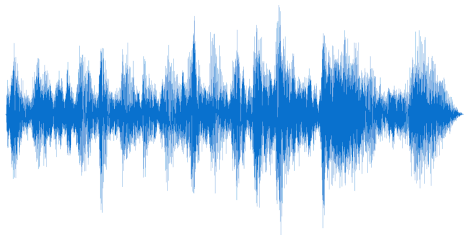
Background: Music Classification

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

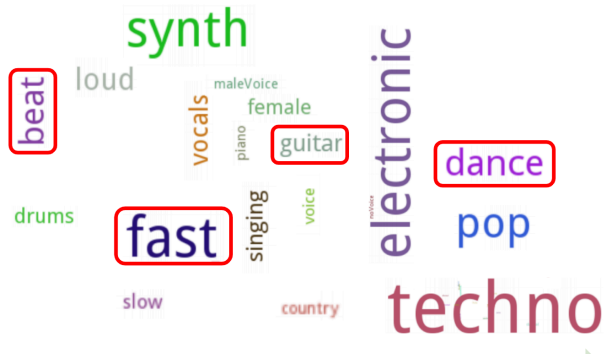


Celma, Oscar, 2010 "Music recommendation", *Music recommendation and discovery, Springer*.

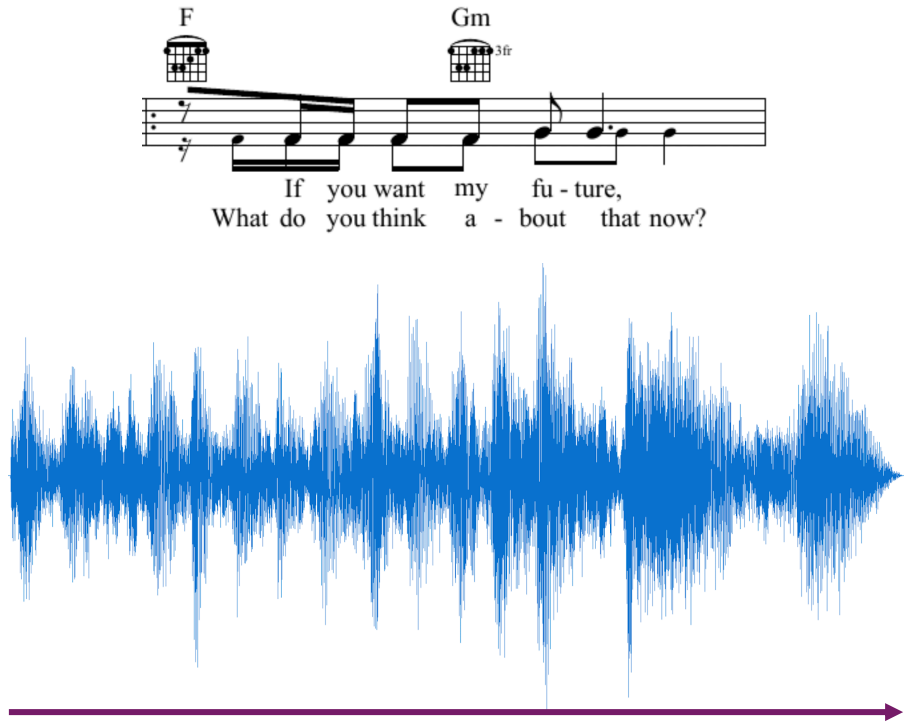
Background: Music Classification



Music Classification Model



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

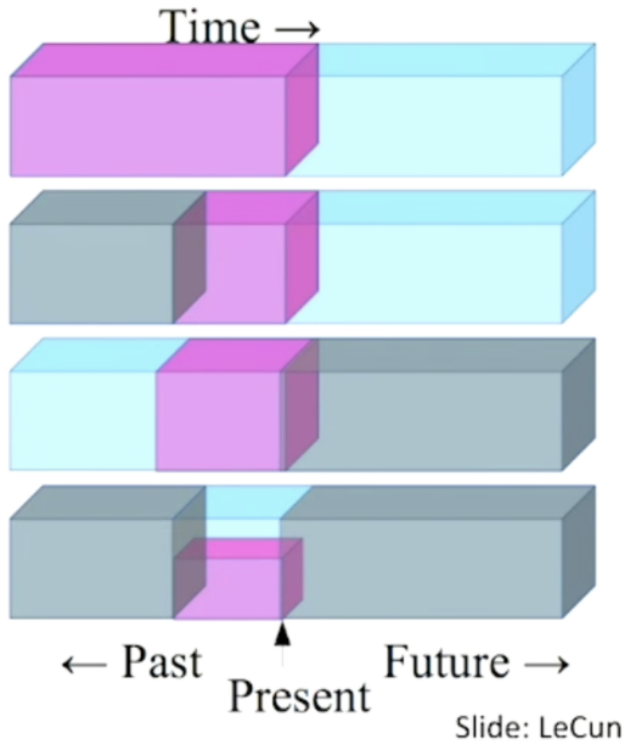
[1] Hendrik Vincent Koops, W. Bas de Haas, John Ashley Burgoyne, Jeroen Bransen, Anna Kent-Muller & Anja Volk (2019) Annotator subjectivity in harmony annotations of popular music, Journal of New Music Research, 48:3, 232-252, DOI: 10.1080/09298215.2019.1613436

[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

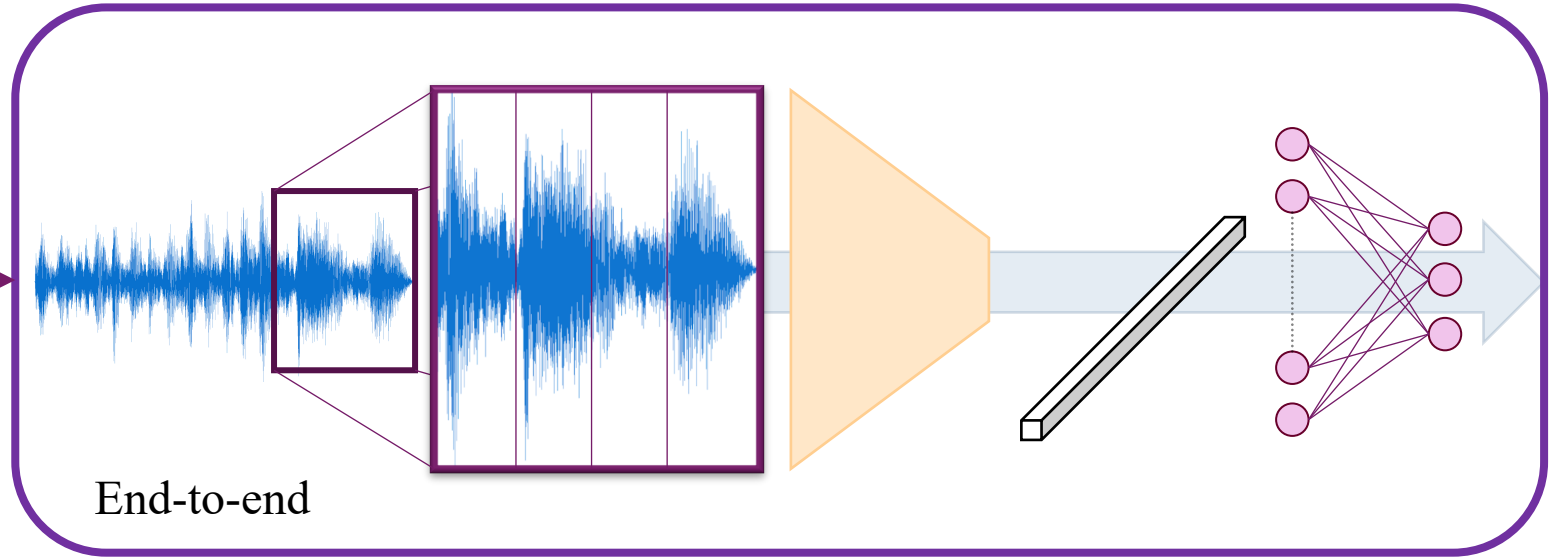
Self-Supervised Learning

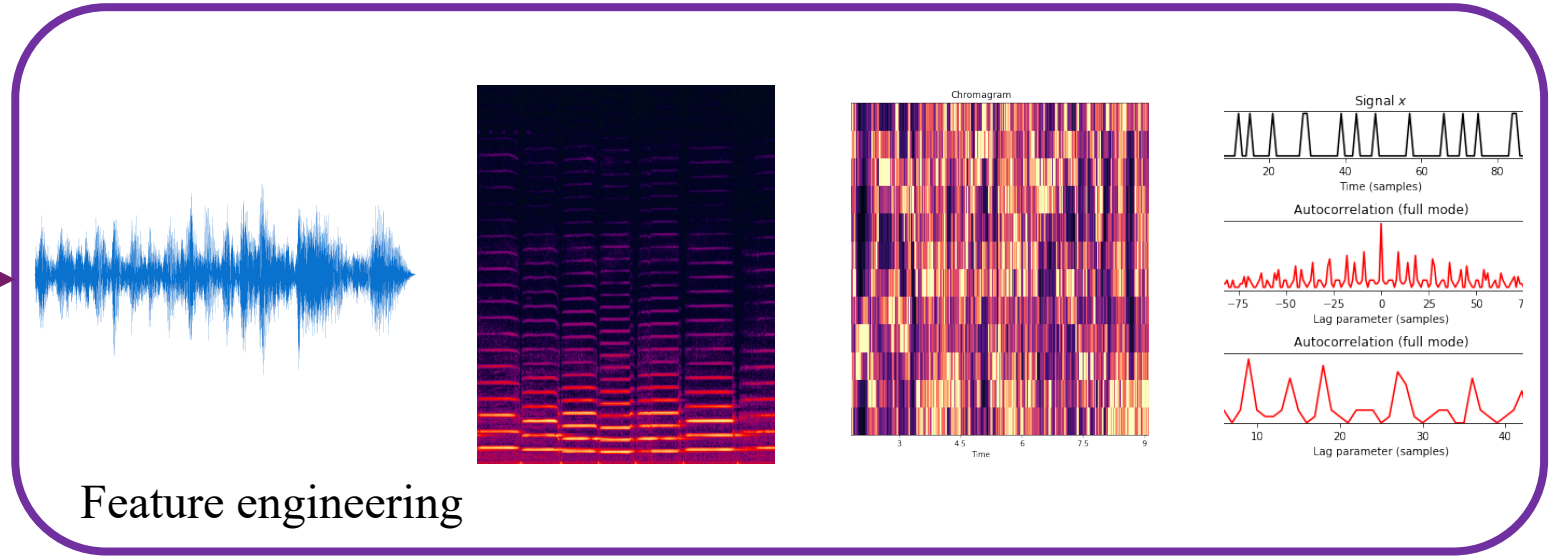
- Learn representations without ground truth.
- Pre-text tasks.
- Methodological improvement for training on smaller datasets.
- Simple and straightforward pre-training.

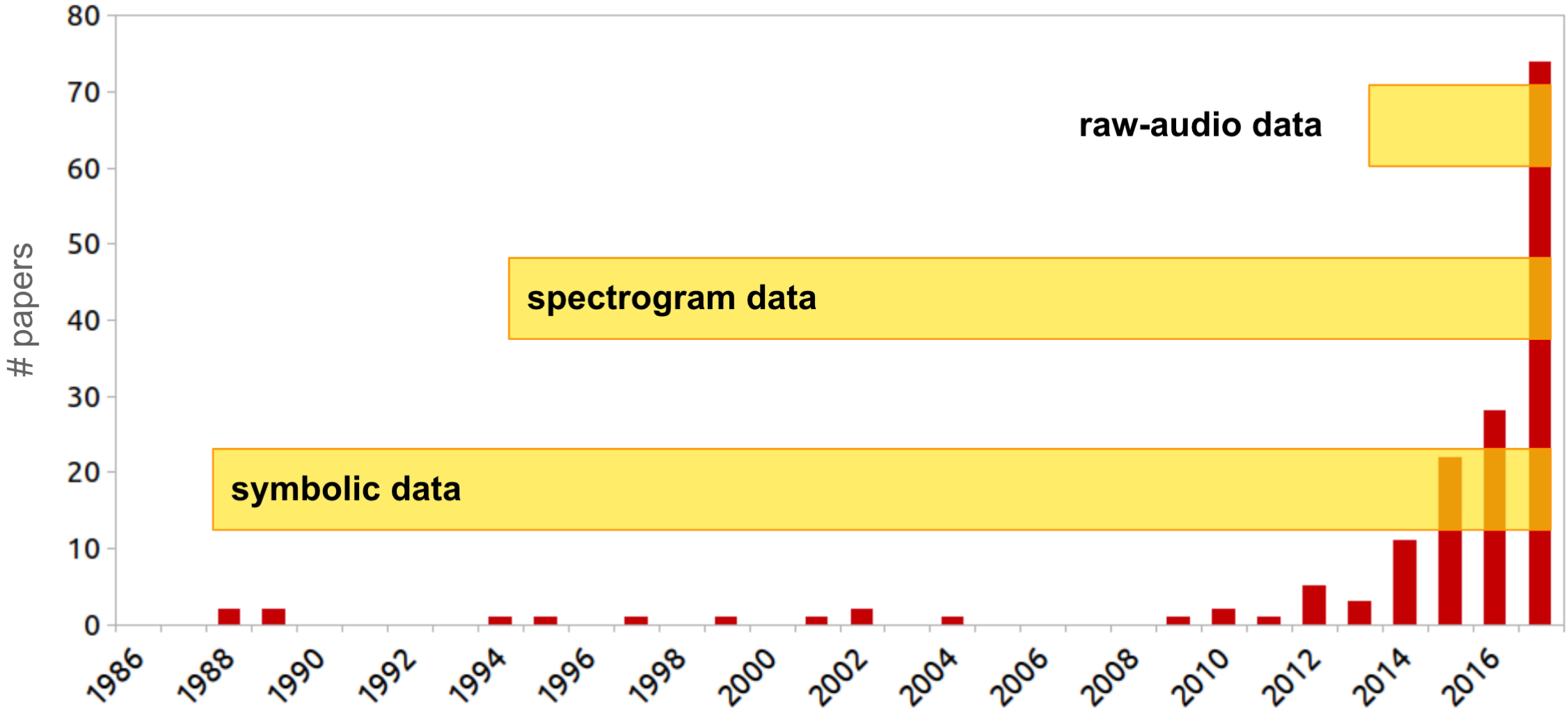
- Explore the limits of "Simple and straightforward" in MIR?

Background


WANNABE
SPICE GIRLS

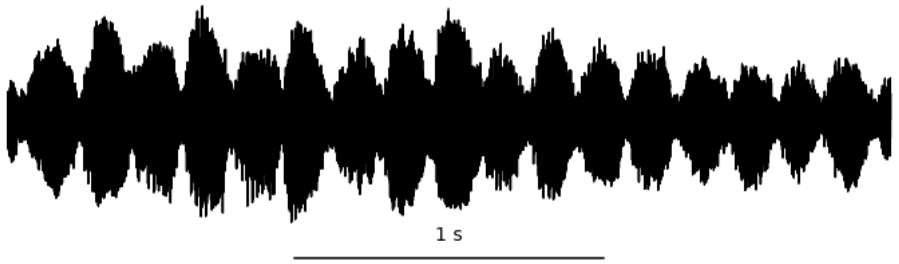

End-to-end

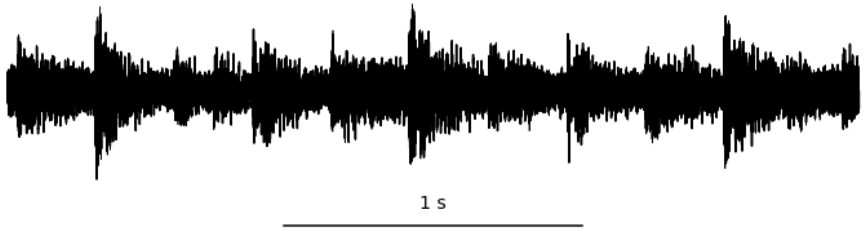

Chromagram
Signal x
Time (samples)
Autocorrelation (full mode)
Lag parameter (samples)
Autocorrelation (full mode)
Lag parameter (samples)
Feature engineering



Sander Dieleman, Jordi Pons, Jongpil Lee, "Waveform-based music processing with deep learning", *Tutorial at the International Society for Music Information Retrieval 2019*

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



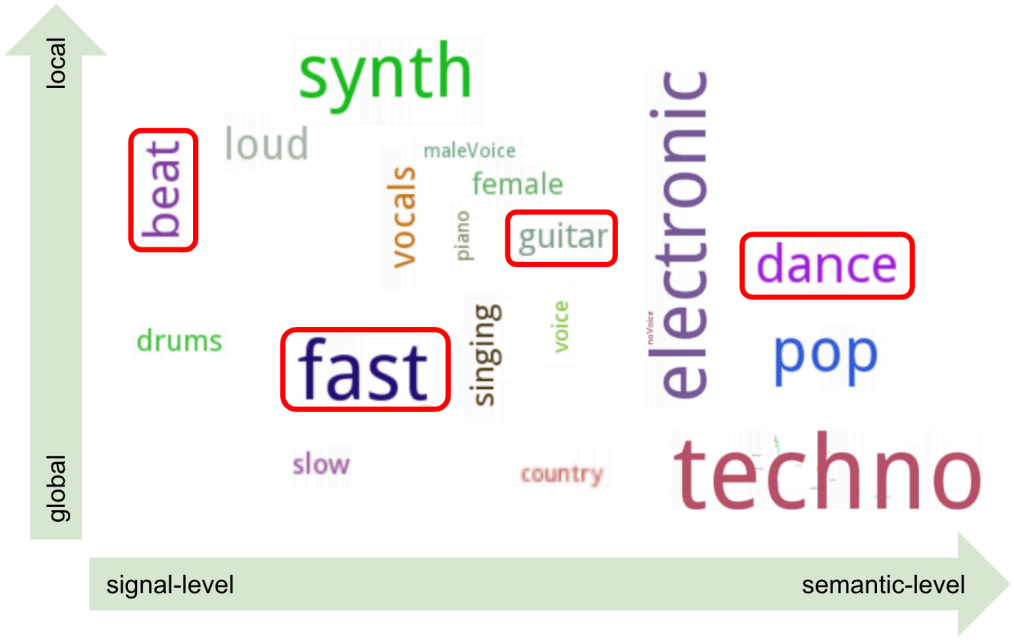


- High-dimensional
- Variable-length
- Missing hierarchical structure

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?

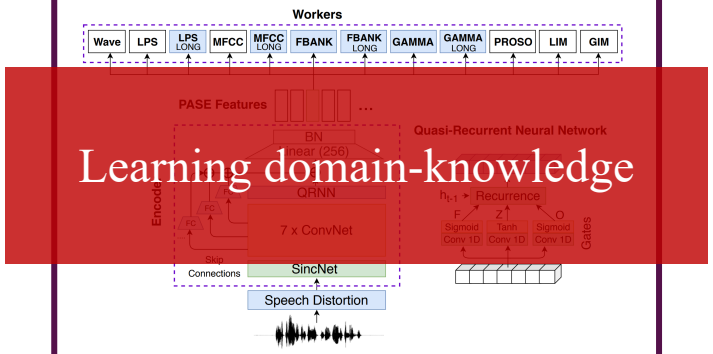
RQ3: Do these methods enable efficient classification for smaller datasets?

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

2. Related Work

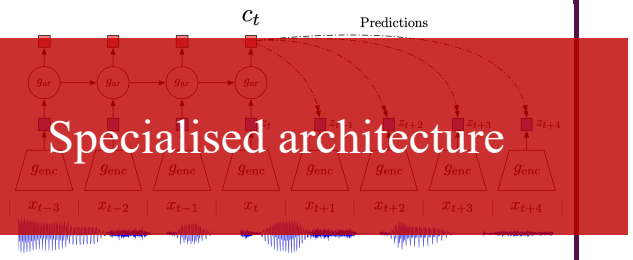
Related Work (self-supervised learning)

PASE / PASE+



Learning domain-knowledge

CPC

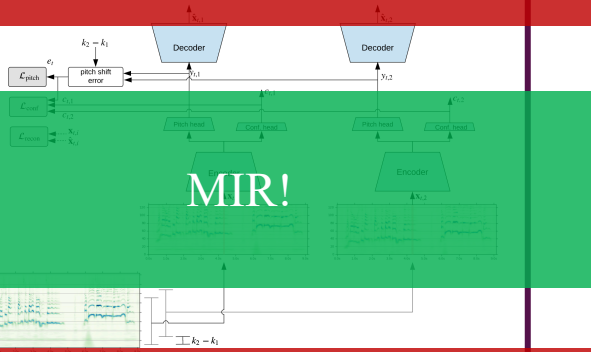


Specialised architecture

MIR!

SPICE

Too specific pre-text task

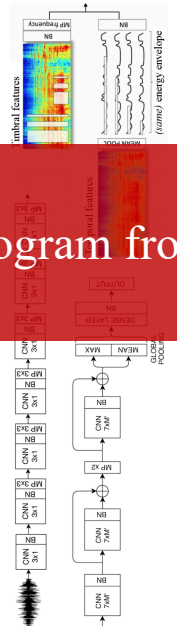


MIR!

Time-frequency domain

Related Work (Music Classification)

Pons et al. (Musicnn)



Spectrogram front-end

SampleCNN

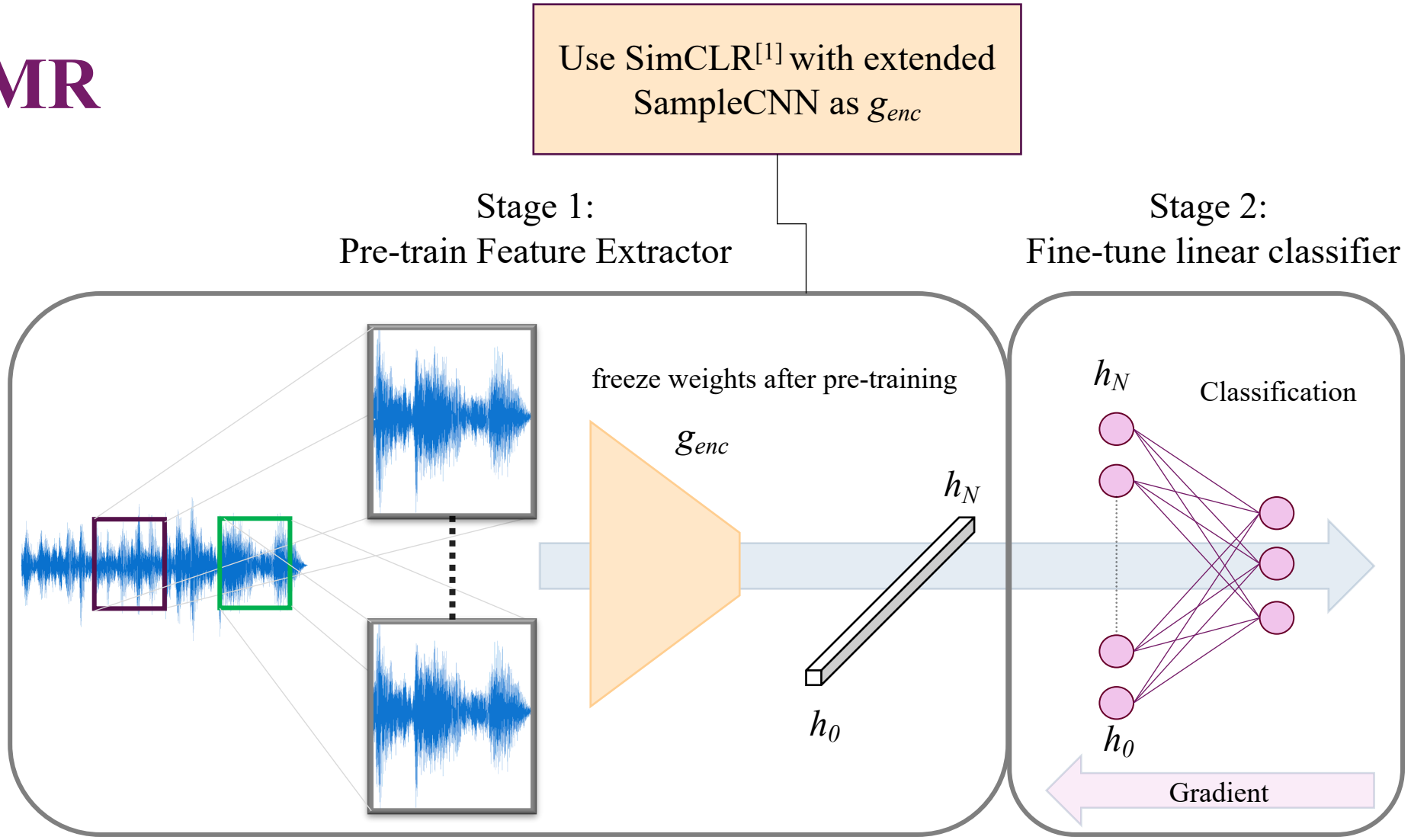
Poor performance on small datasets

Related Work

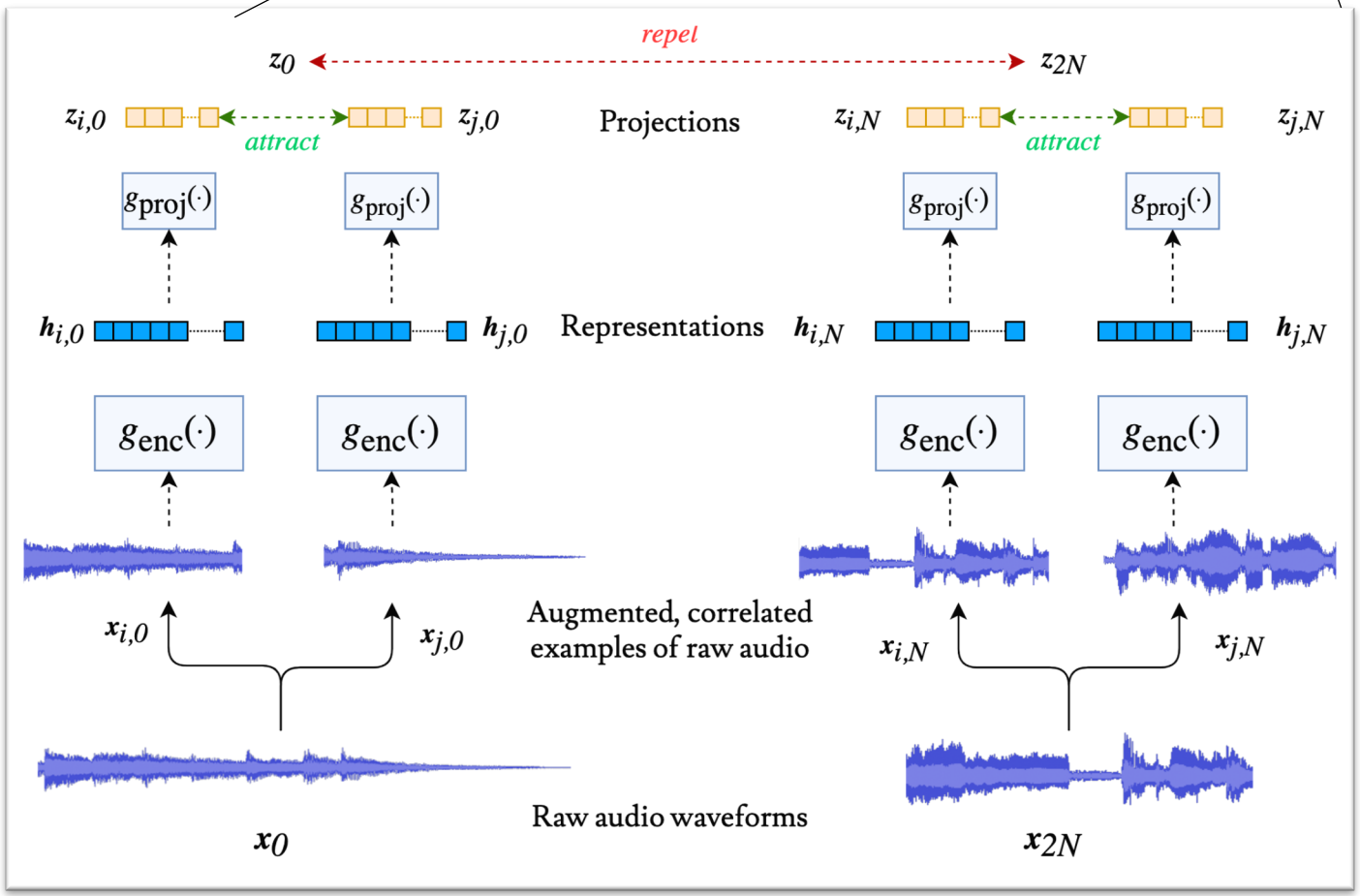
Method	No domain knowledge injection	Performs well on small datasets	Time-domain	No specialised architecture	Performs MIR task
PASE / PASE+	-	+	+	+	-
CPC	+	+	+	-	+/-
SPICE	-	+	-	+	+
Pons et al. (Musicnn)	-	-	+/-	-	+
SampleCNN	+	-	+	+	+
CLMR	+	+	+	+	+

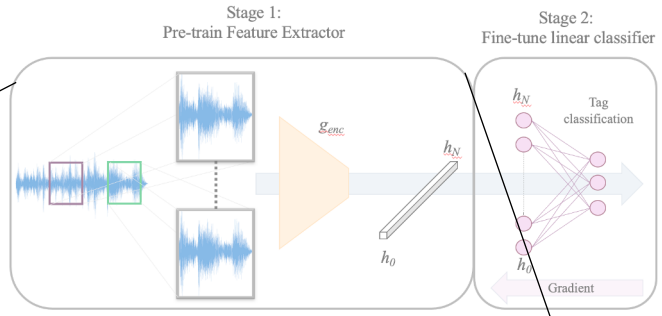
3. Method

CLMR


Use SimCLR[1] with extended SampleCNN as g_{enc}
Stage 1:
Pre-train Feature Extractor
Stage 2:
Fine-tune linear classifier
freeze weights after pre-training
g_{enc}
h_N
h_0
h_N
Classification
h_0
Gradient

SimCLR

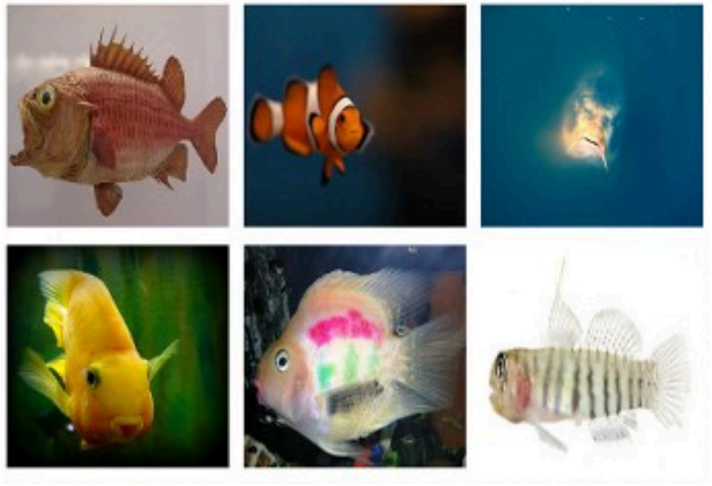


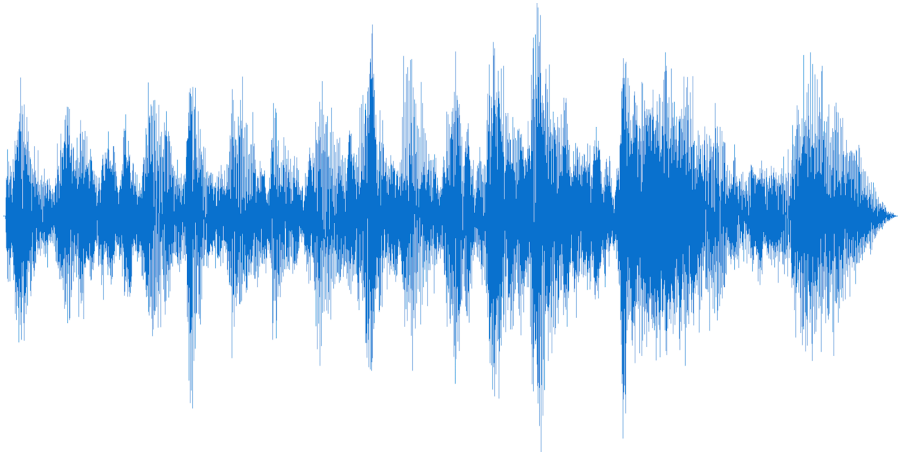


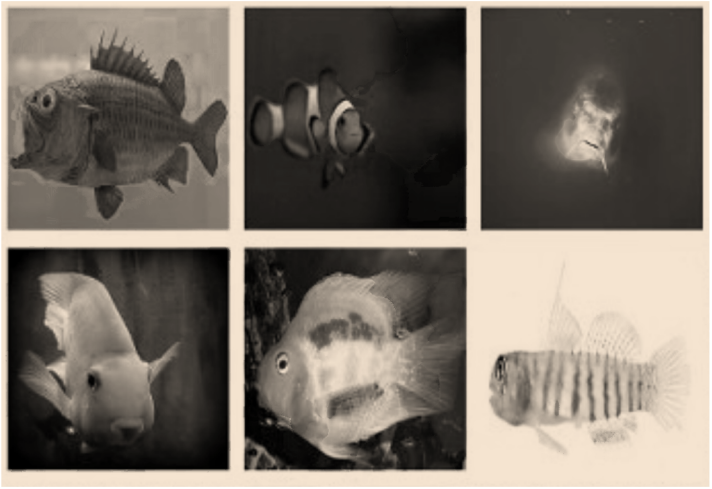
- Data augmentations.
- Extract representations with SampleCNN encoder.
- Project representations.
- Cosine similarity
- Contrastive loss

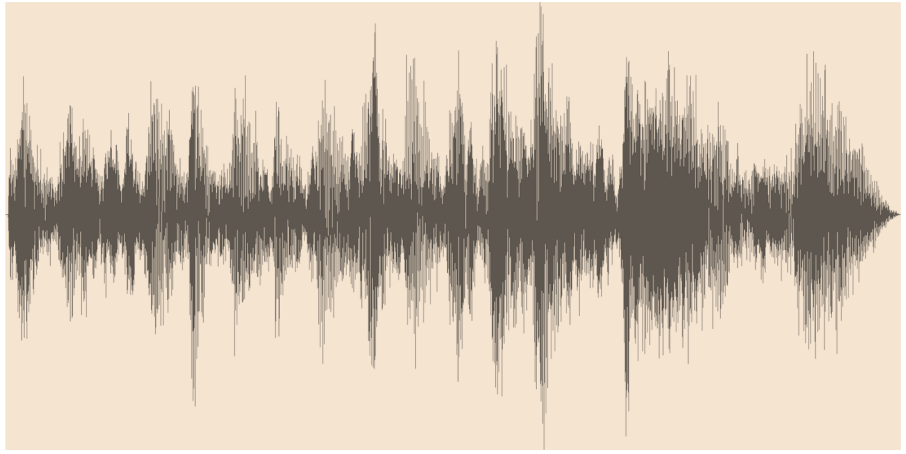
$$\ell_{i,j} = -\log \frac{\exp\left(\text{sim}\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp\left(\text{sim}\left(z_i, z_k\right)/\tau\right)}$$

Images != Audio

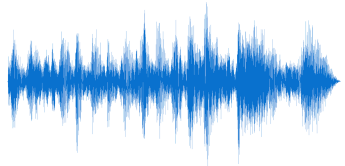




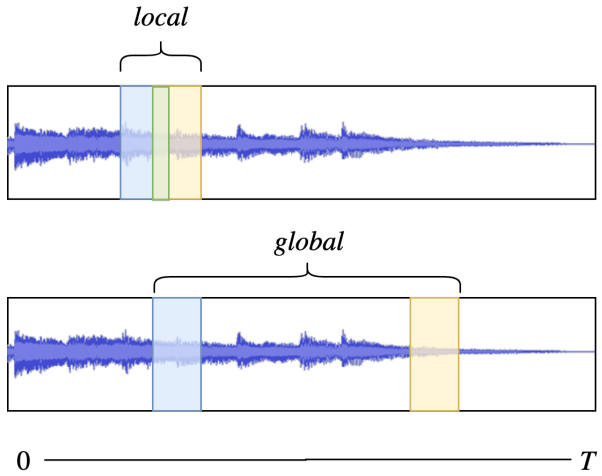




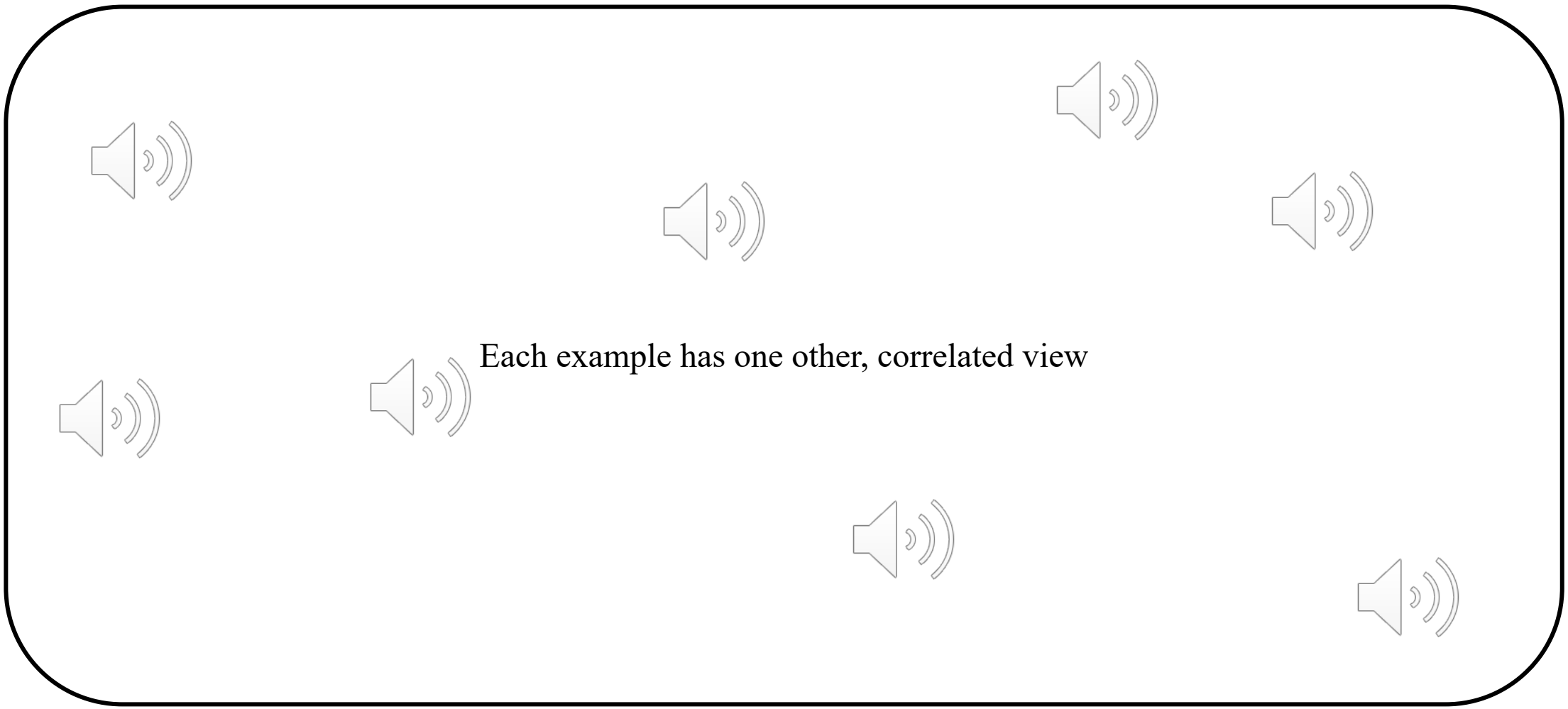
Method



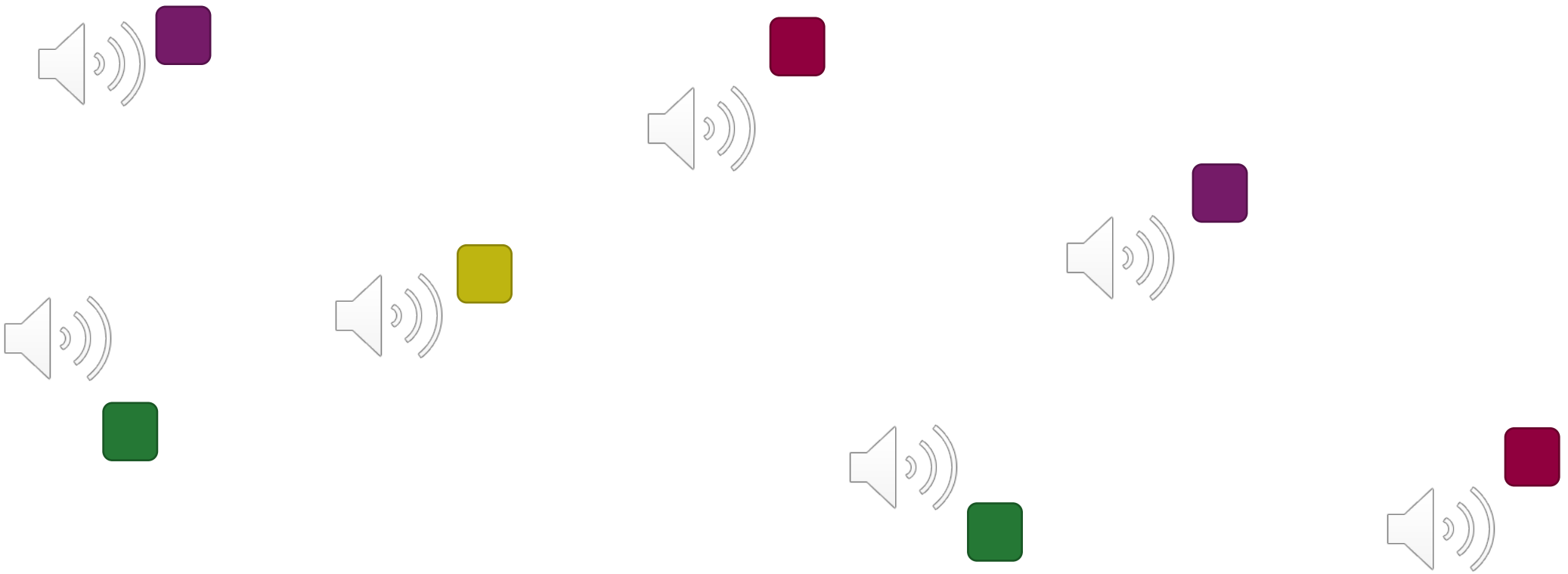
- Random Trimming — Infer local/global structure.
- Polarity Inversion — Phase-awareness
- Noise Addition — Robustness to noisy signals
- Gain Reduction — Loudness / Dynamics
- Filtering — Frequency spectrum
- Delay — Repeating signals
- Pitch Shifting — Pitch-awareness
- Reverberation — Room acoustics



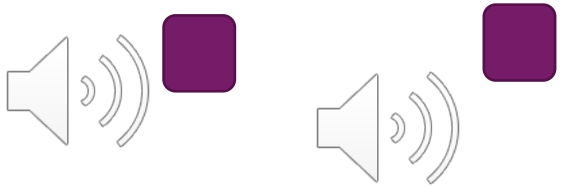
Examples

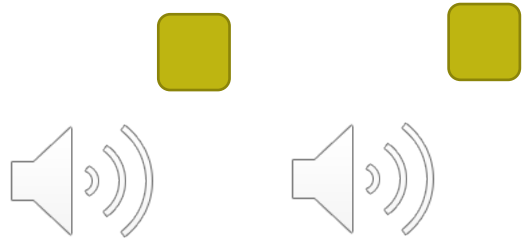


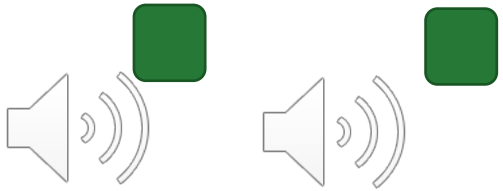
Examples

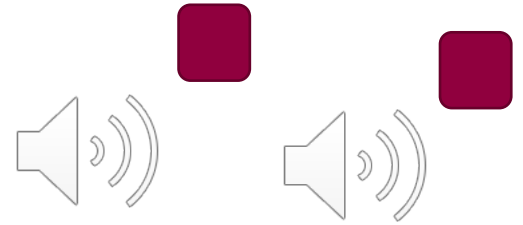


Examples







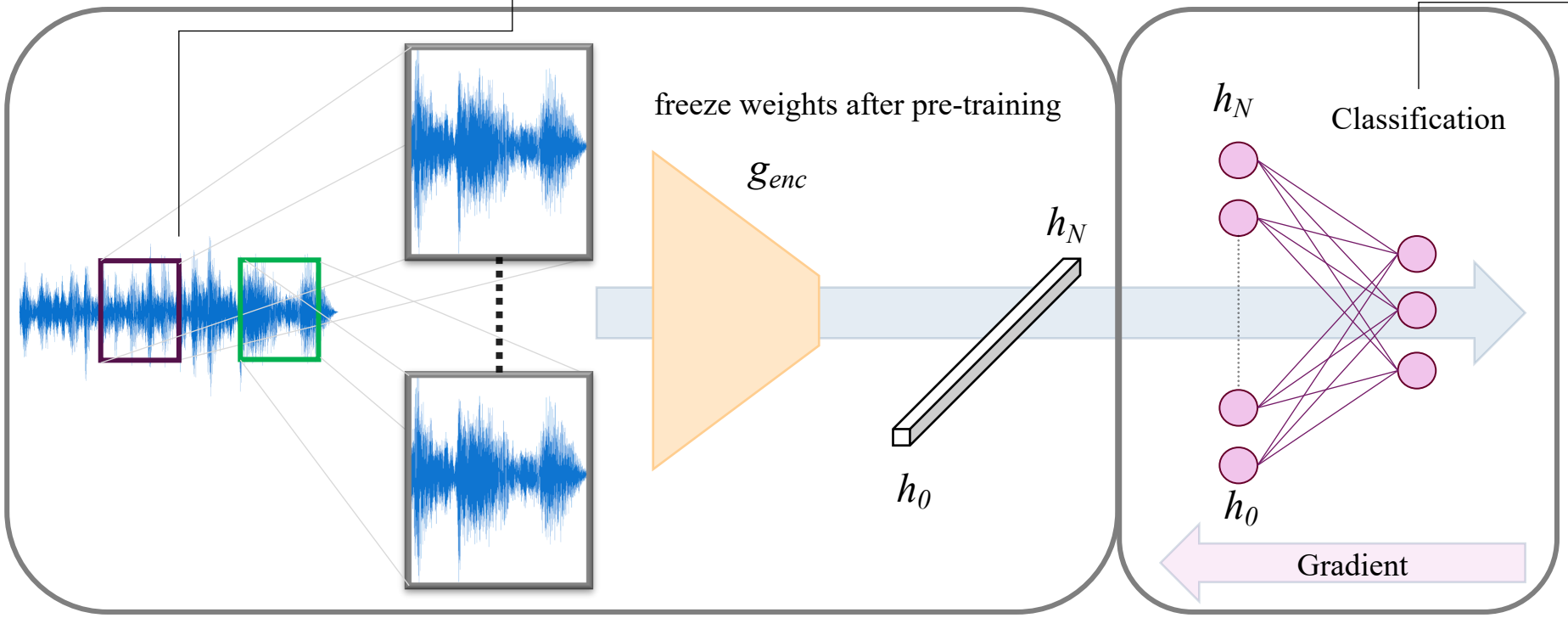


Compare to fully supervised, end-to-end baseline

CLMR

Test on two datasets:
MagnaTagATune
Million Song Dataset

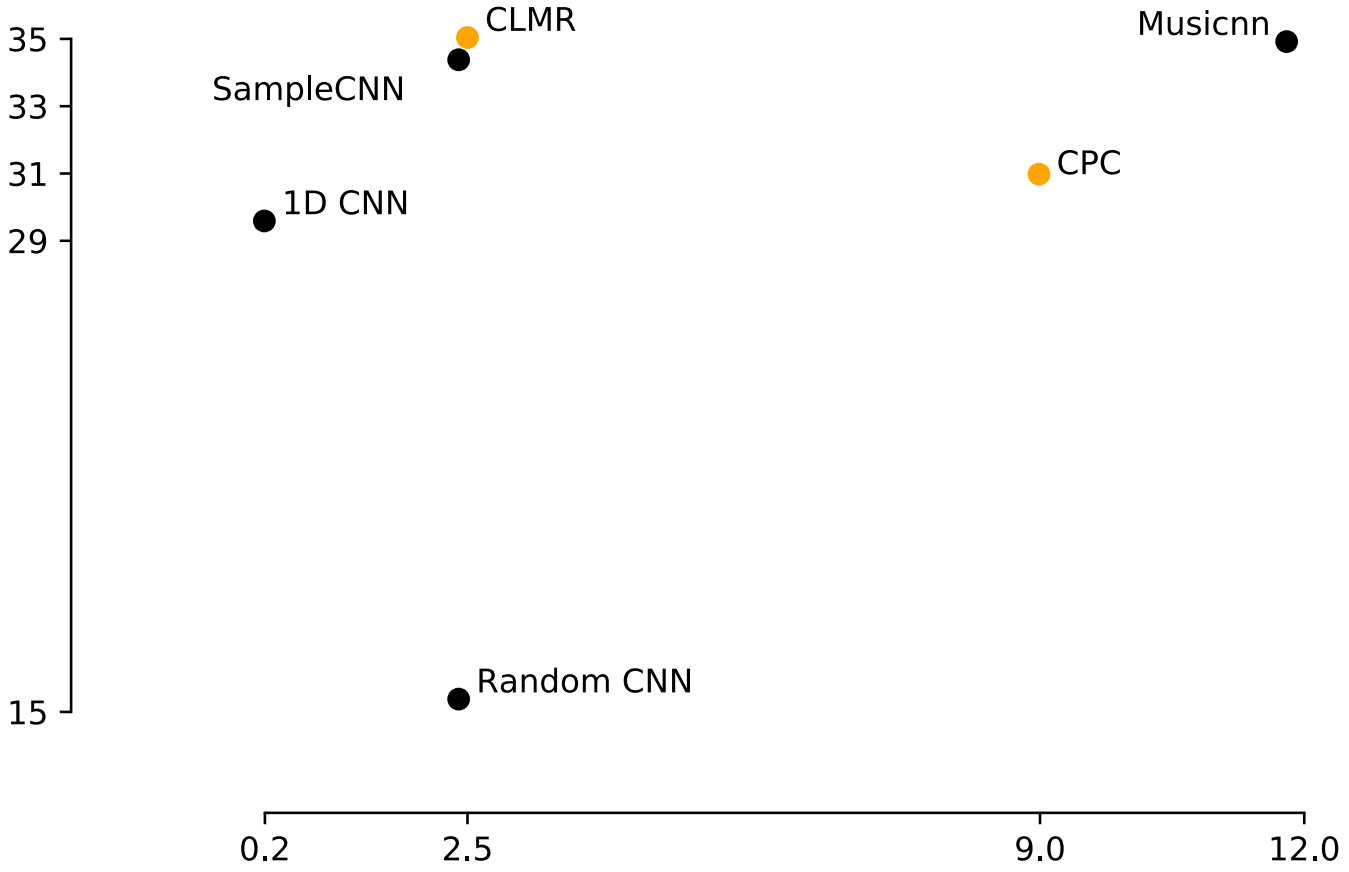
Music Classification



4. Experiments

Experiment 1: Self-Supervised Music Tagging

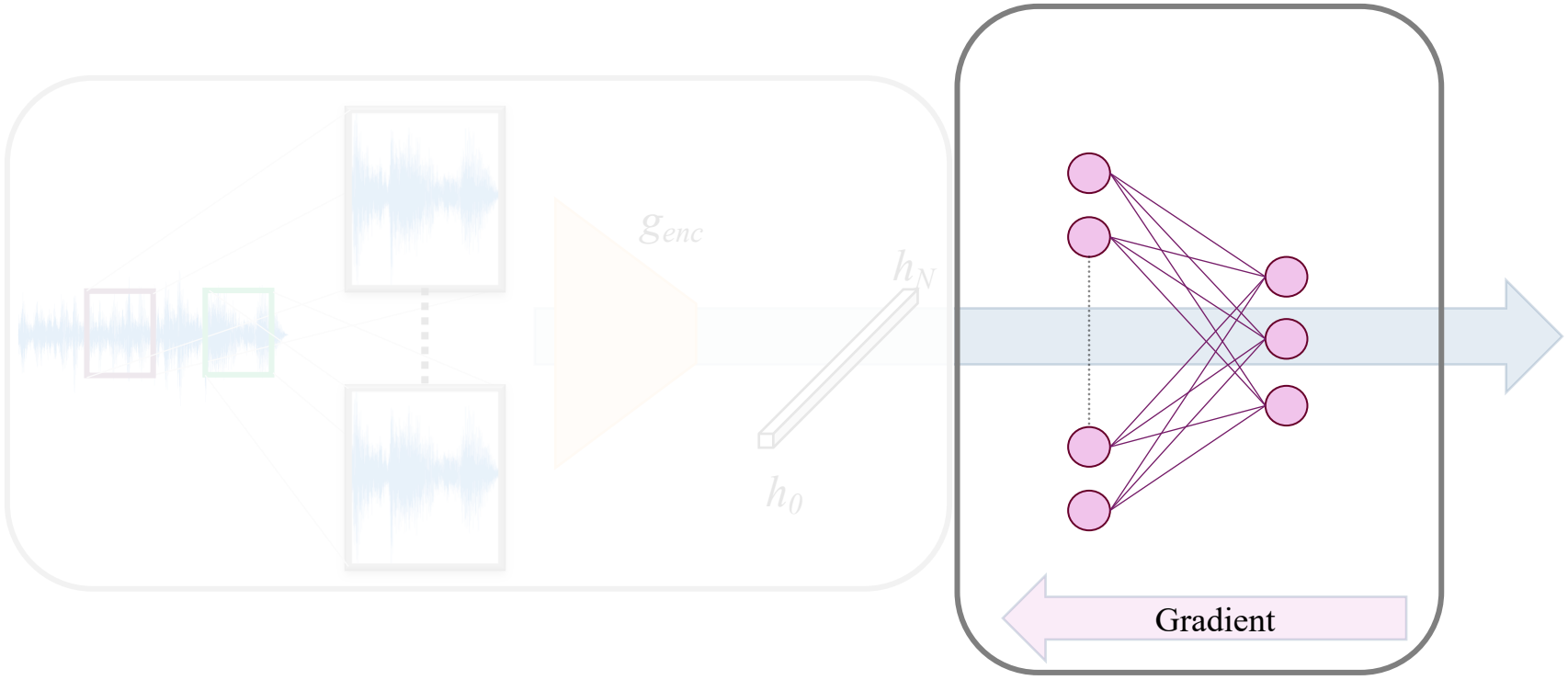
- MagnaTagATune Dataset
- 25.863 unique songs
- Top-50 tags



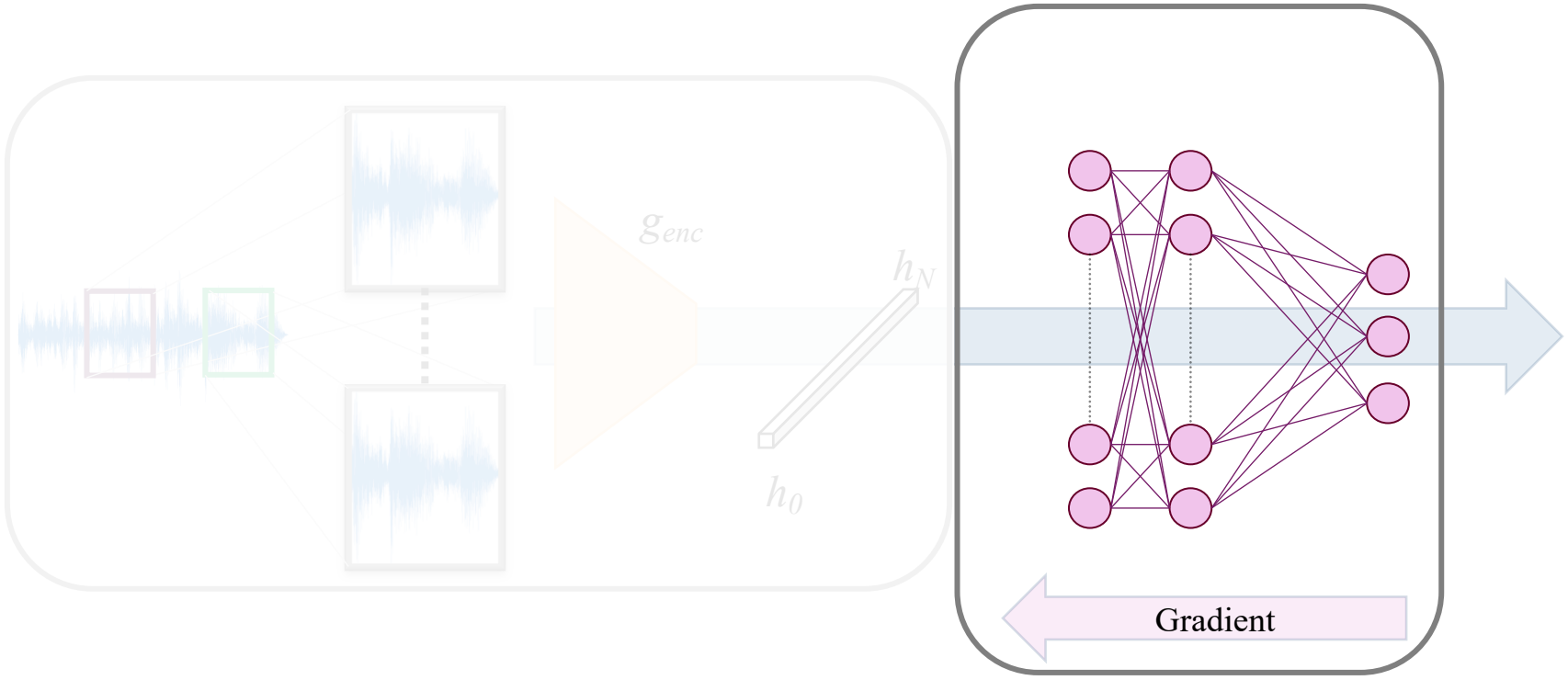
Experiment 1: Self-Supervised Music Tagging

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

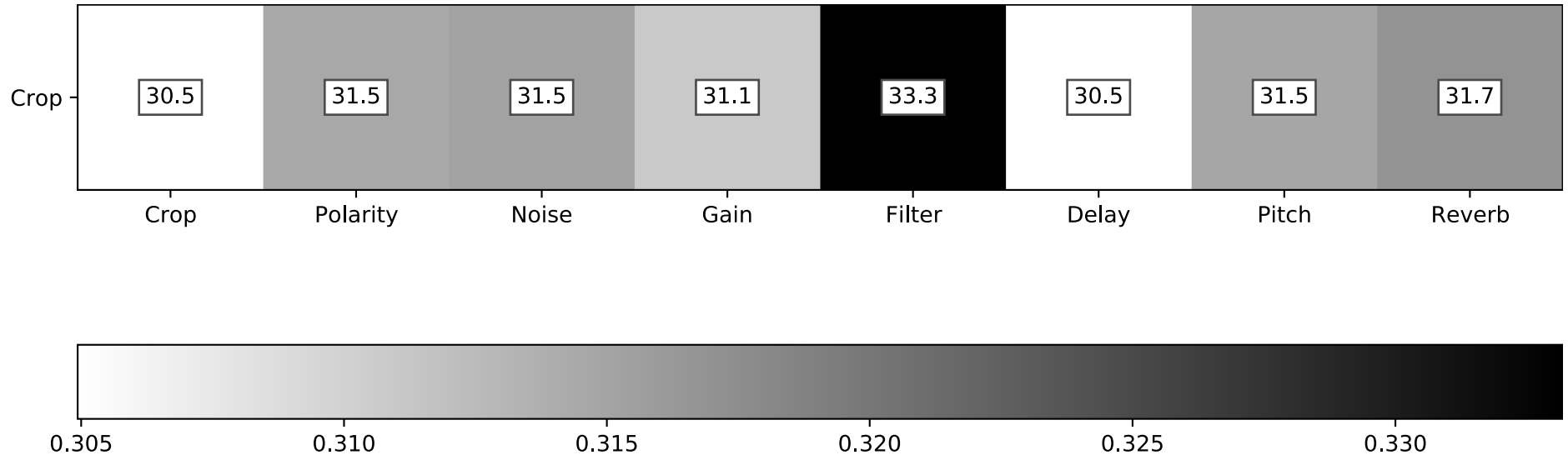
Model	Dataset	ROC-AUC$_{TAG}$	PR-AUC$_{TAG}$
CLMR (ours)	MTAT	88.49 (**89.25**)	**35.37** (**35.89**)
Pons et al.†	MTAT	89.05	34.92
SampleCNN†	MTAT	88.56	34.38
CPC (ours)	MTAT	86.60 (87.99)	30.98 (33.04)
1D CNN†	MTAT	85.58	29.59

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

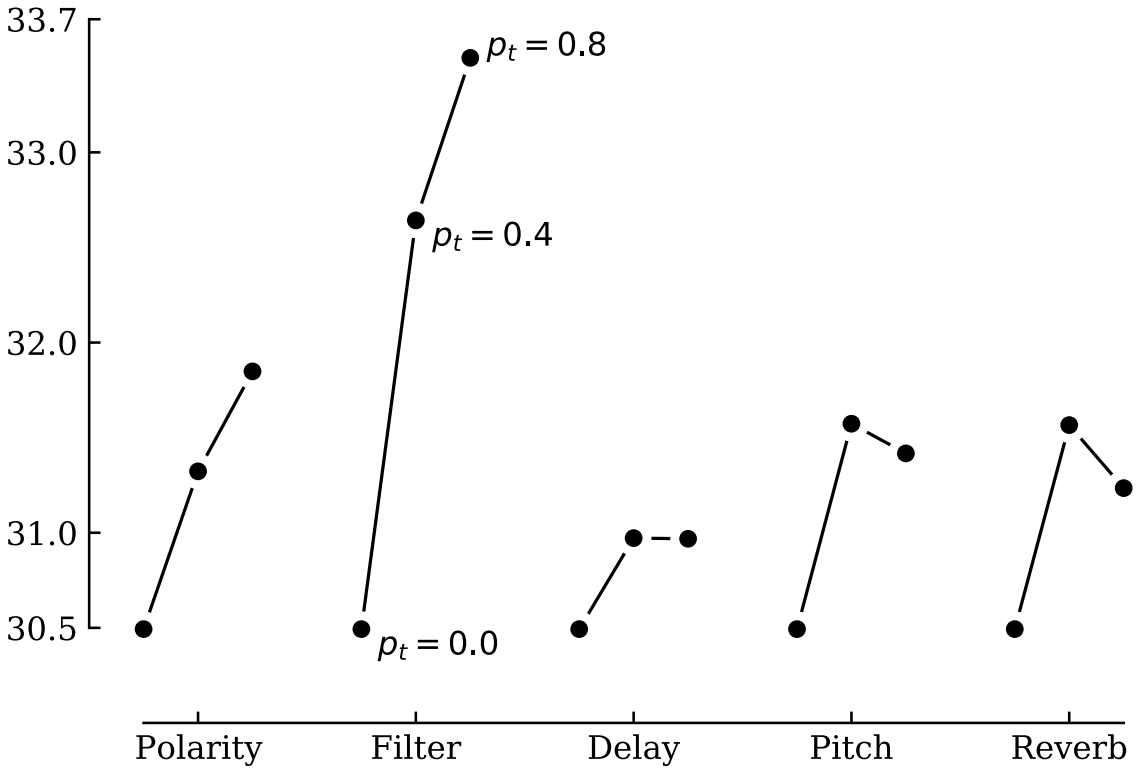


p=1

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

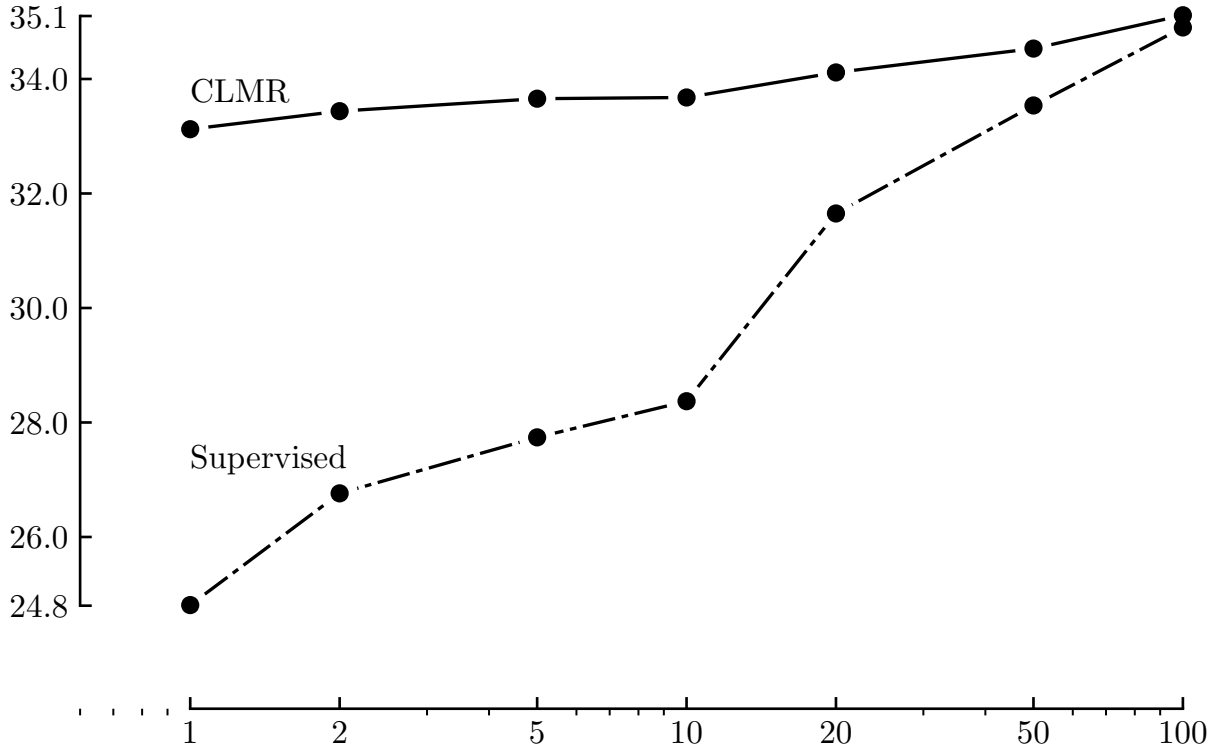
- Pre-train on all available data.
- Freeze feature extractor weights.
- Fine-tune linear classifier on 1% of the labeled data (train split).
- Test on complete test split.

Experiment 3: Efficient Classification

MagnaTagATune Dataset:

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

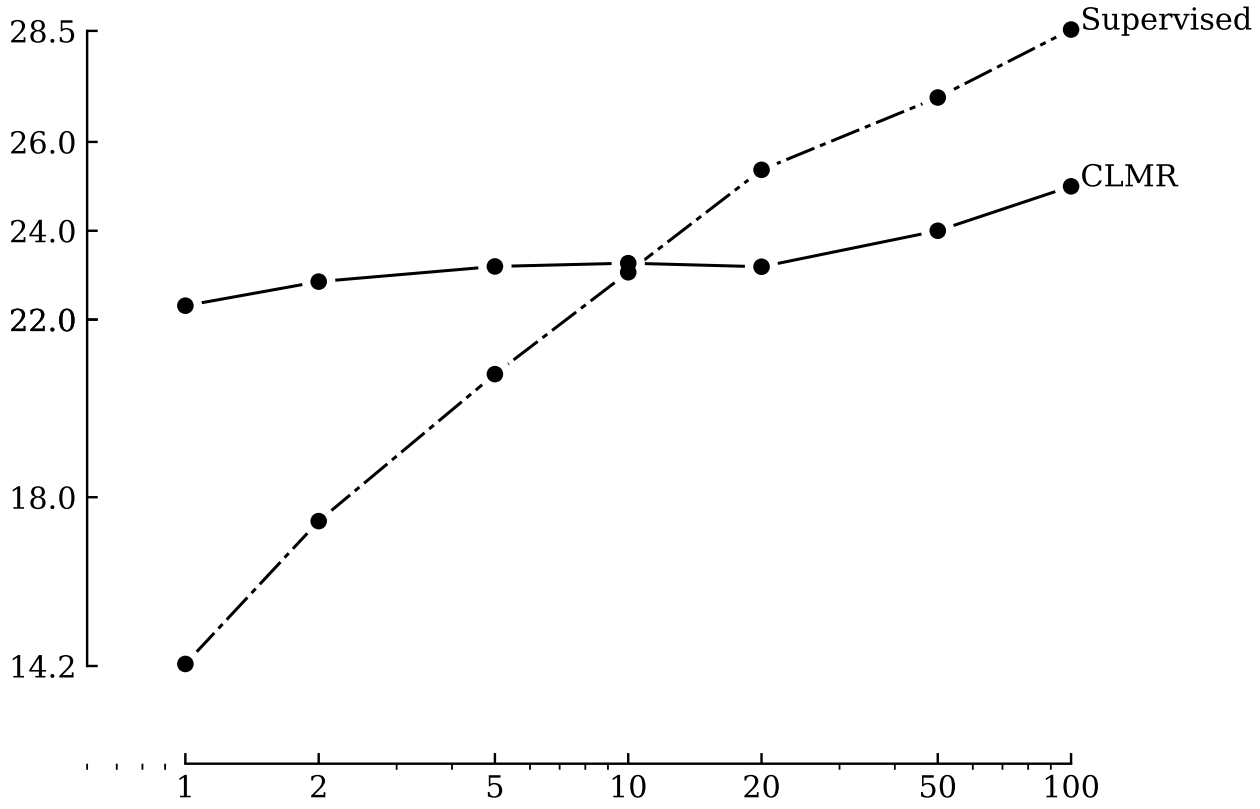
(1% = 259 songs)



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

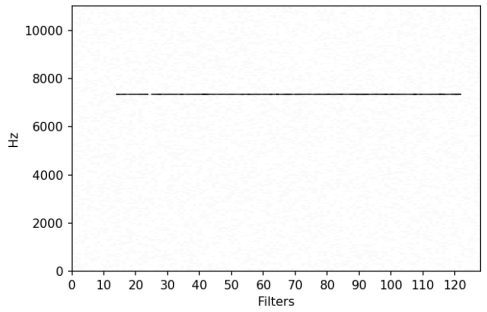
Epochs	ROC-AUC$_{TAG}$	PR-AUC$_{TAG}$	ROC-AUC$_{CLIP}$	PR-AUC$_{CLIP}$
10 000	88.47 (89.25)	35.37 (35.89)	93.16 (93.48)	69.32 (70.03)
3 000	88.49 (88.94)	35.11 (35.46)	93.07 (93.27)	69.20 (69.74)
1 000	88.31 (88.64)	34.40 (34.86)	92.89 (93.08)	68.59 (69.15)

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

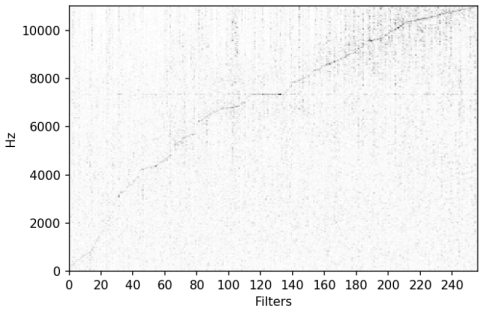
5. Interpretability

Interpretability: Filter Visualisation

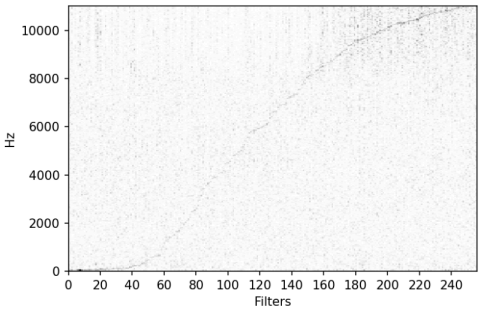
- Gradient ascent on a random waveform of 729 samples.
- Magnitude spectrum.
- Vertical line is a magnitude spectrum of a single filter.
- Sorted by frequency magnitude.



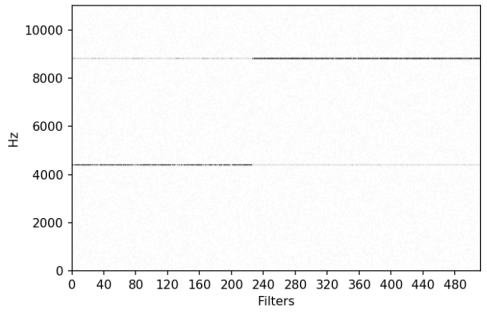
(a) $\text{CLMR}^{(1)}_{\text{MTAT}}$



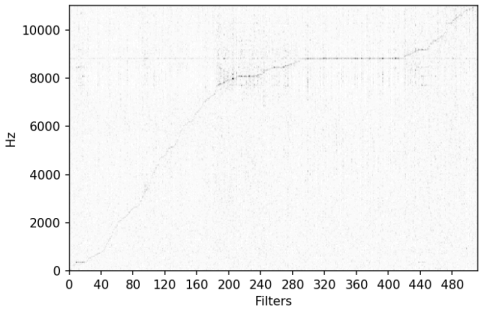
(b) $\text{CLMR}^{(4)}_{\text{MTAT}}$



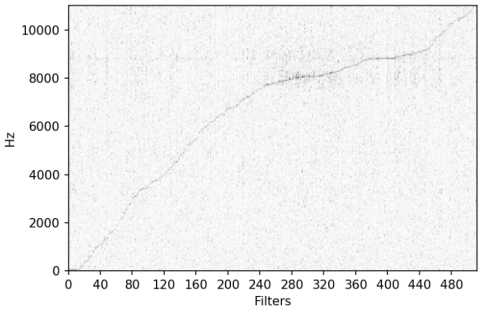
(c) $\text{CLMR}^{(6)}_{\text{MTAT}}$



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



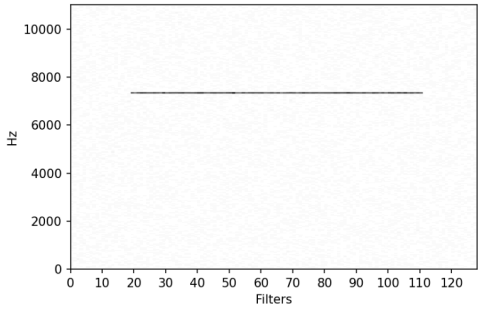
(e) $\text{CPC}^{(4)}_{\text{MTAT}}$



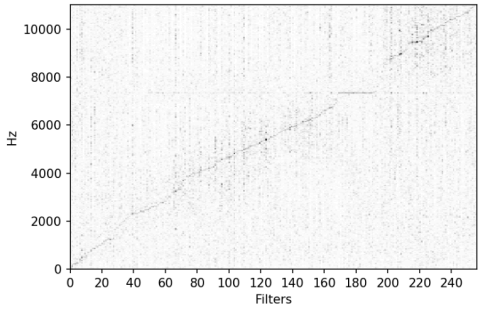
(f) $\text{CPC}^{(6)}_{\text{MTAT}}$

Interpretability: Filter Visualisation

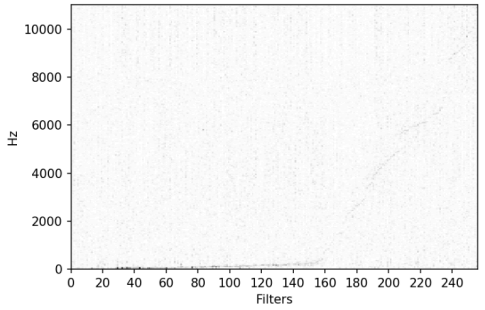
- Sensitive to a single band of frequencies around 7500 Hz.
- Higher layers: spread first linearly, then non-linearly across the full range.



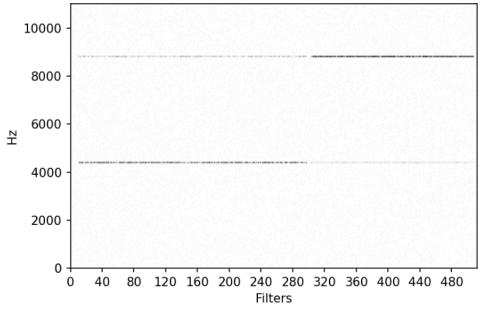
(a) $\text{CLMR}^{(1)}_{\text{Billboard}}$



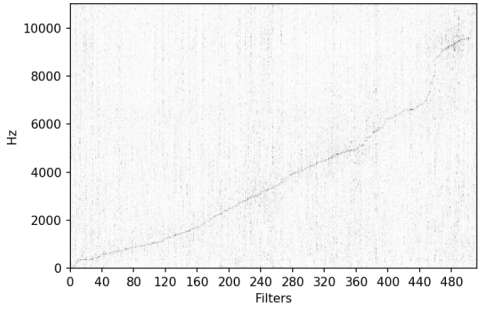
(b) $\text{CLMR}^{(4)}_{\text{Billboard}}$



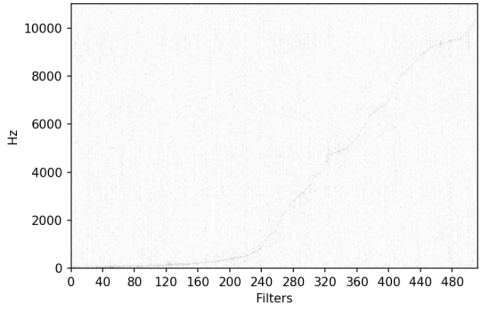
(c) $\text{CLMR}^{(6)}_{\text{Billboard}}$



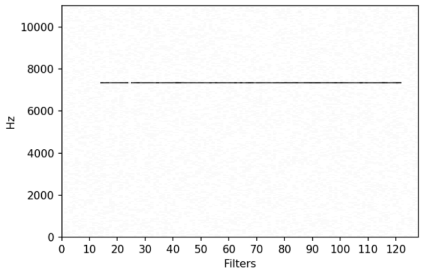
(d) $\text{CPC}^{(1)}_{\text{Billboard}}$



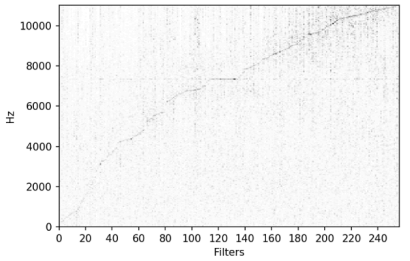
(e) $\text{CPC}^{(4)}_{\text{Billboard}}$



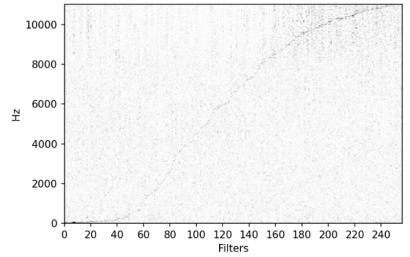
(f) $\text{CPC}^{(6)}_{\text{Billboard}}$



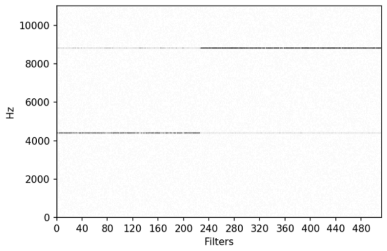
(a) $\text{CLMR}^{(1)}_{\text{MTAT}}$



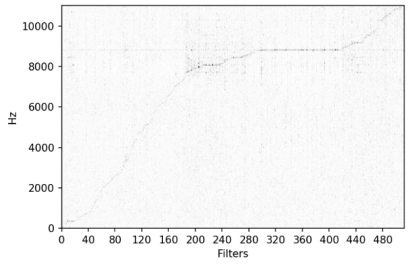
(b) $\text{CLMR}^{(4)}_{\text{MTAT}}$



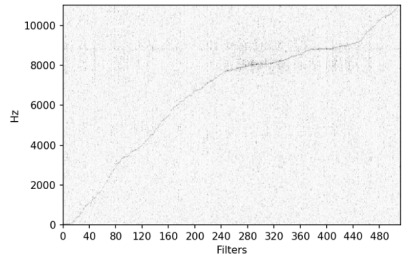
(c) $\text{CLMR}^{(6)}_{\text{MTAT}}$



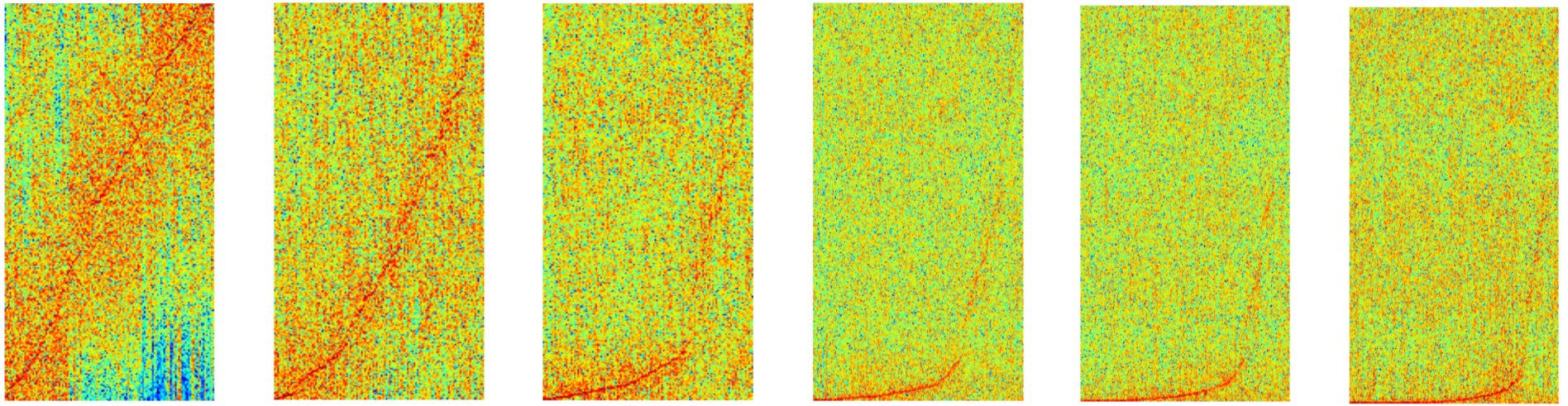
(d) $\text{CPC}^{(1)}_{\text{MTAT}}$



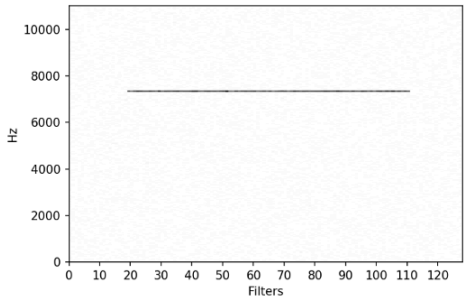
(e) $\text{CPC}^{(4)}_{\text{MTAT}}$



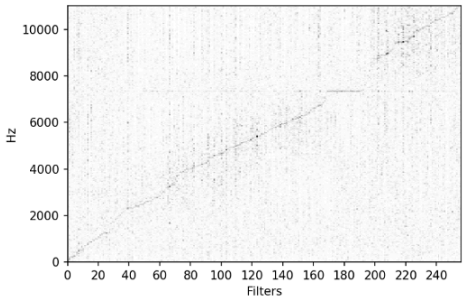
(f) $\text{CPC}^{(6)}_{\text{MTAT}}$



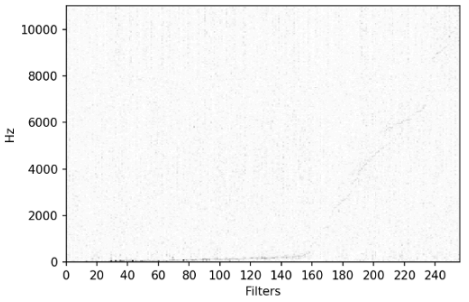
Layer 1 Layer 2 Layer 3 Layer 4 Layer 5 Layer 6



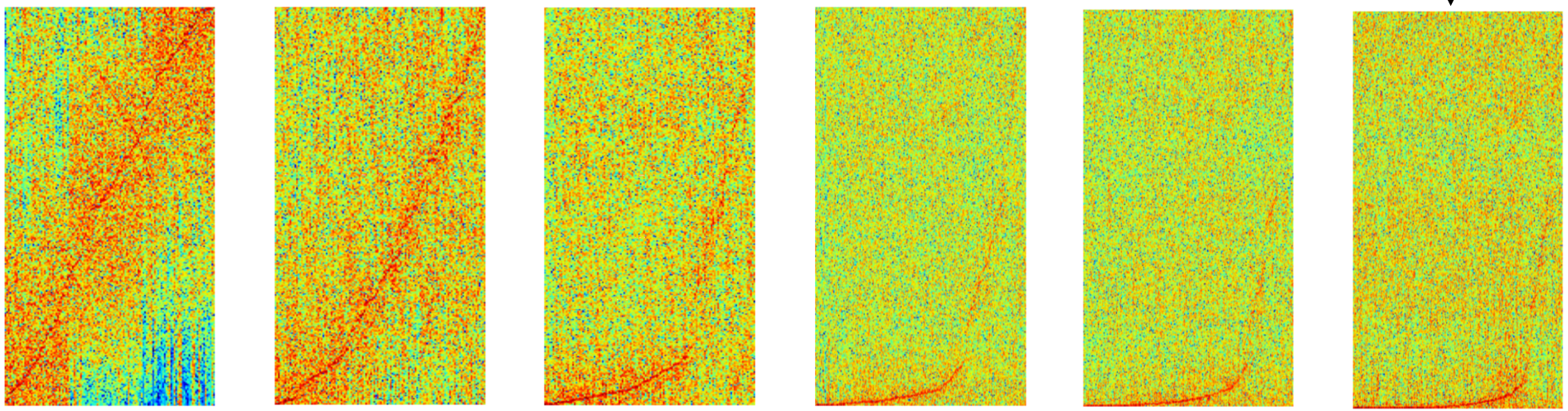
(a) $\text{CLMR}^{(1)}_{\text{Billboard}}$



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



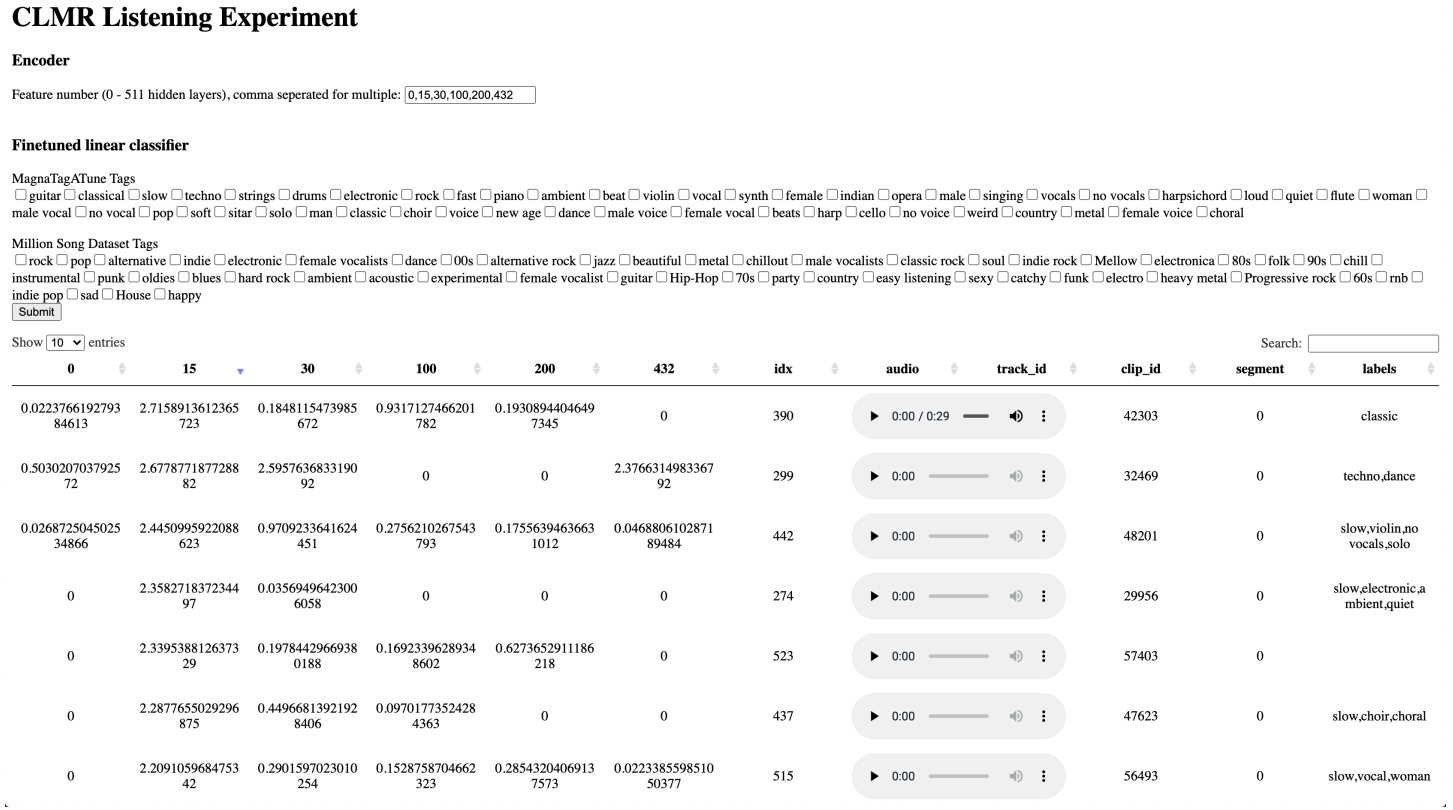
(c) $\text{CLMR}^{(6)}_{\text{Billboard}}$



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 seperated for multiple: 0,15,30,100,200,432

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

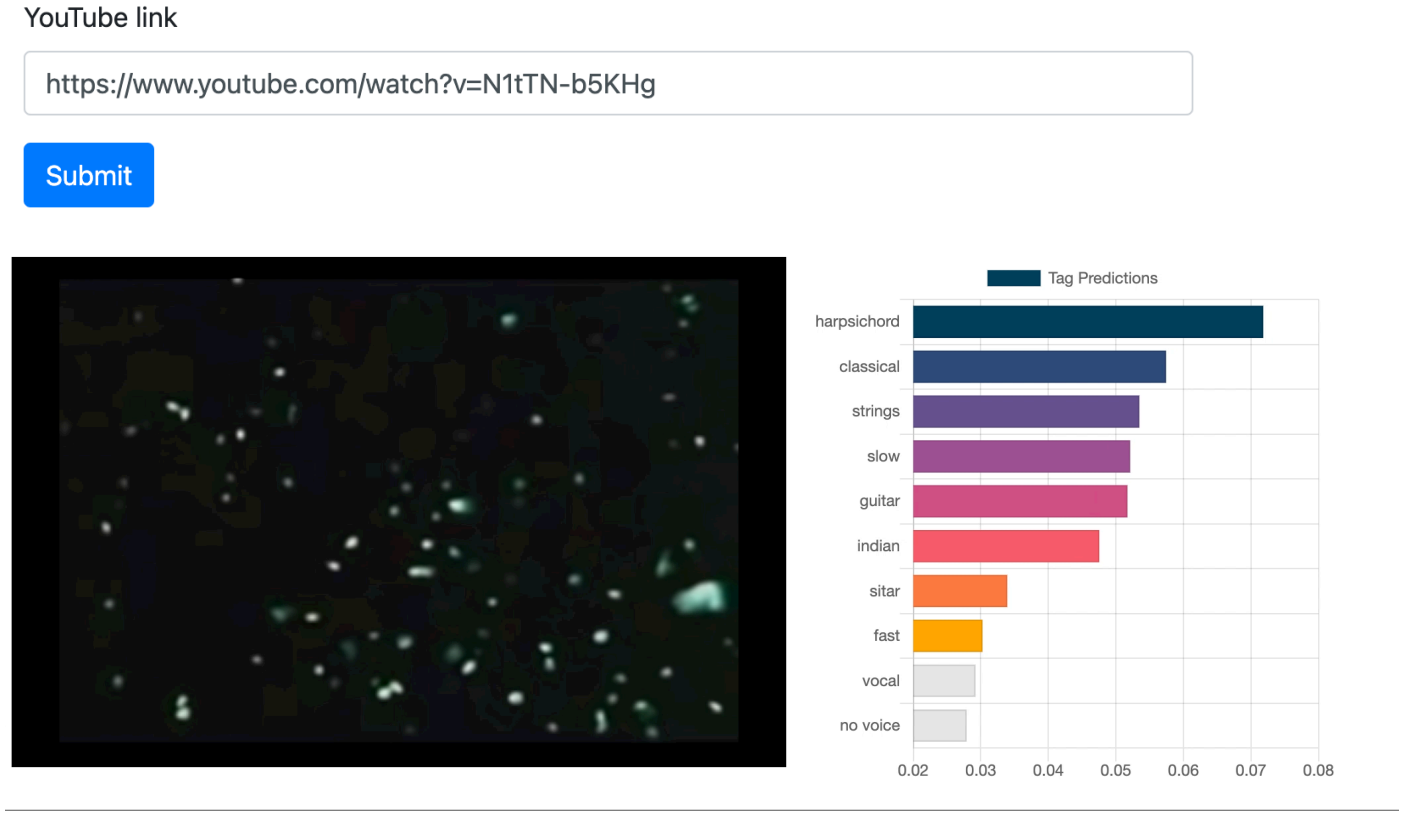
Submit

Show 10 entries Search:

0	15	30	100	200	432	idx	audio	track_id	clip_id	segment	labels
0.022376619279384613	2.7158913612365723	0.1848115473985672	0.9317127466201782	0.19308944046497345	0	390	0:00 / 0:29		42303	0	classic
0.503020703792572	2.677877187728882	2.595763683319092	0	0	2.376631498336792	299	0:00		32469	0	techno,dance
0.026872504502534866	2.445099592208862 3	0.9709233641624451	0.2756210267543793	0.17556394636631012	0.046880610287189484	442	0:00		48201	0	slow,violin,no vocals,solo
0	2.358271837234497	0.035694964230060058	0	0	0	274	0:00		29956	0	slow,electronic,ambient,quiet
0	2.339538812637329	0.19784429669380188	0.16923396289348602	0.6273652911186218	0	523	0:00		57403	0	
0	2.2877655029296875	0.44966813921928406	0.09701773524284363	0	0	437	0:00		47623	0	slow,choir,choral
0	2.209105968475342	0.290159702301025 4	0.1528758704662323	0.28543204069137573	0.022338559851050377	515	0:00		56493	0	slow,vocal,woman

Interpretability: Listening Interface

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



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

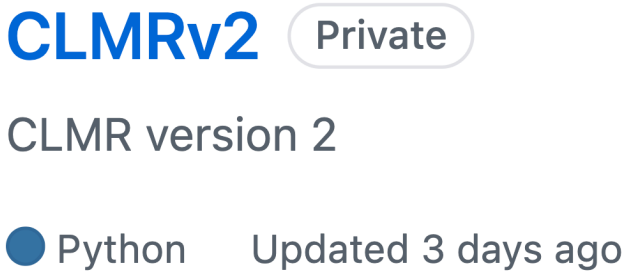
RQ4: Do these methods capture important, musical features, that are transferable to out-of-domain datasets?

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

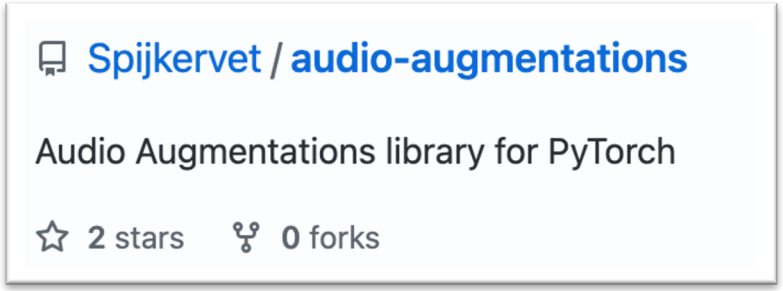
- Self-supervision mainly constrained to mini-batch size and long training.
  - Momentum Contrast, BYOL.
  - Performance on larger audio datasets.
- Self-supervision models for other MIR tasks.
- Extend to time-frequency domain.
- Specialised fine-tune model.



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



Thank you

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

UNIVERSITY OF AMSTERDAM
Faculty of Science

Contrastive Loss

Maximised

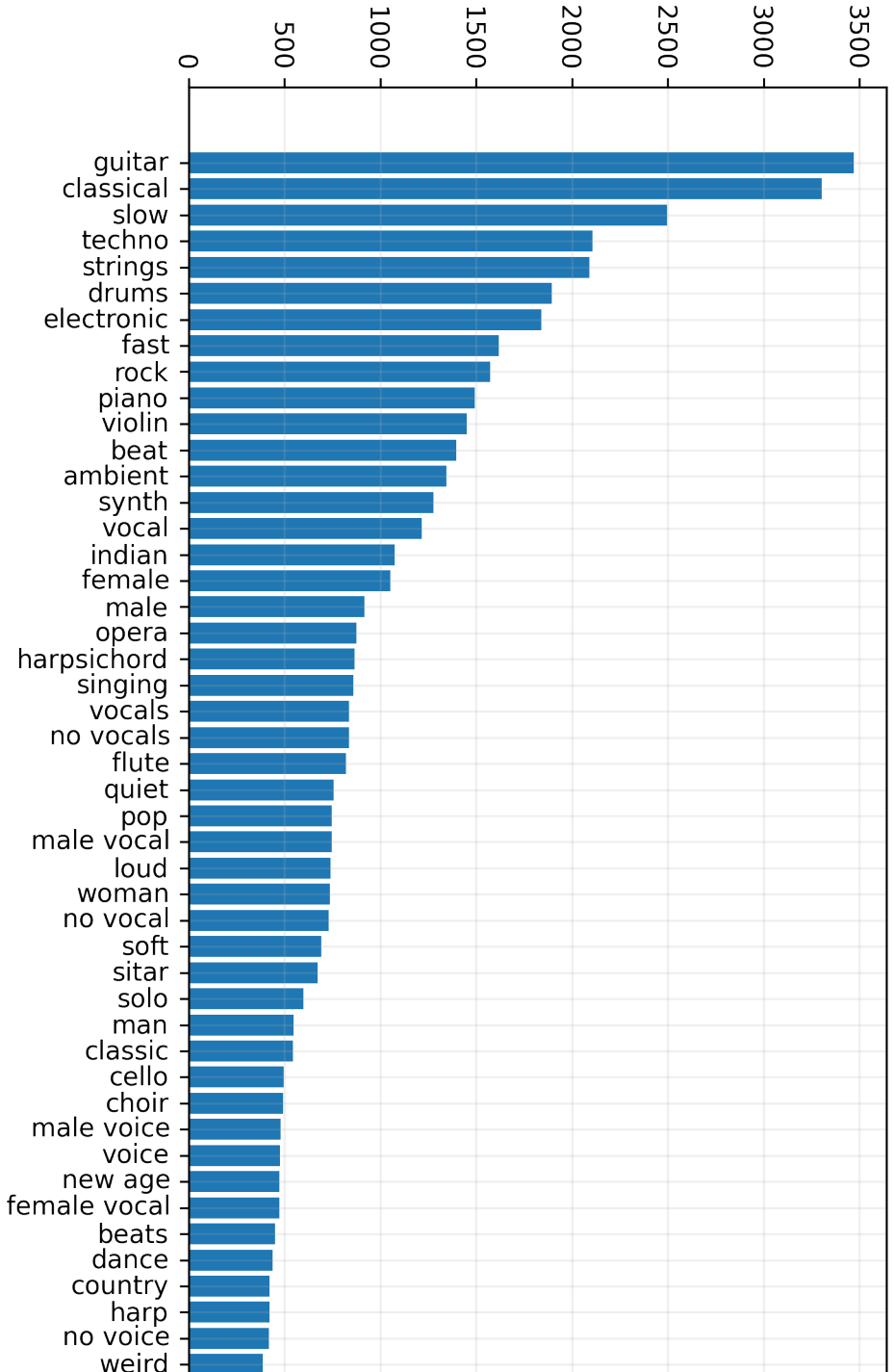
$$\mathcal{L}_{i,j} = -\log \frac{\exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_j)/\tau)}{\sum_{k=1}^{2n} \mathbf{1}_{[k \neq i]} \exp(\text{sim}(\mathbf{z}_i, \mathbf{z}_k)/\tau)}$$

Minimised

No higher than 1

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

MagnaTagATune



Million Song Dataset