An Experimental Comparison Of Multi-view Self-supervised Methods For Music Tagging

Gabriel Meseguer-Brocal, Dorian Desblancs, Romain Hennequin
Deezer Research, Paris

Abstract

- Self-supervised learning for pre-training ML models on unlabeled data is valuable, especially in music, where obtaining labelled data is challenging.
- Models are trained on pretext tasks to acquire robust features for downstream tasks. The choice of the pretext task is critical as it guides the model in shaping the feature space with meaningful constraints for information encoding.
- We investigate and compare the performance of new multi-view self-supervised methods for music tagging consistently across the training pipeline.

Keywords:
audio representations, music information retrieval, self-supervised learning

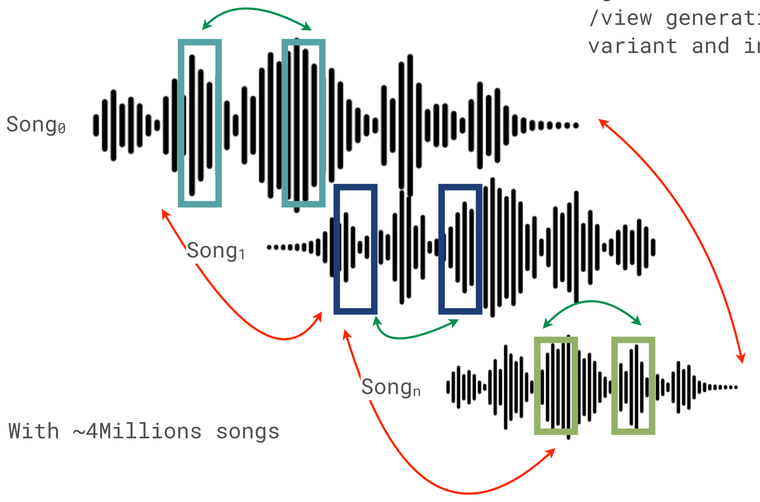
A self-supervised setup:

Data /generalisation scope /view generations: variant and invariant

Model /computational power

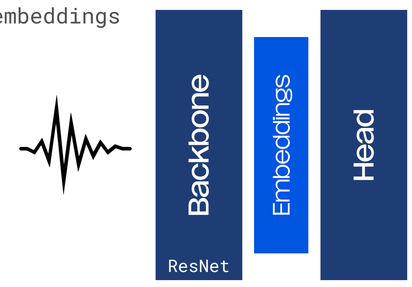
Pretext Task /actual supervised learning objective that is used to train the model

Evaluation /final downstream tasks /model validation



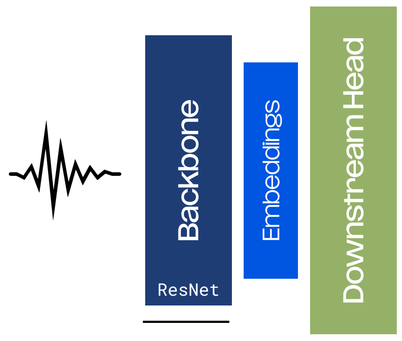
With ~4Millions songs

To compute the generic embeddings



To not tailored to the pretext method

What we aim to obtain



Frozen

Pretext task

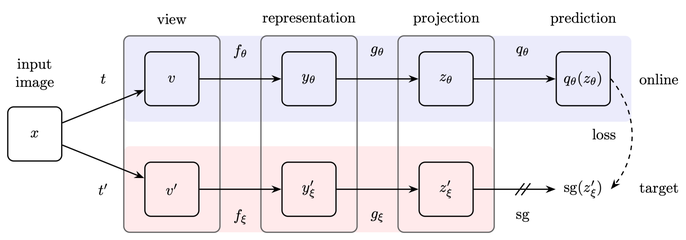
Feature similarity:

Contrastive: temperature-scaled cross-entropy loss → Negatives

$$\ell_{i,j} = -\log \frac{\exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\text{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in ICML, 2020

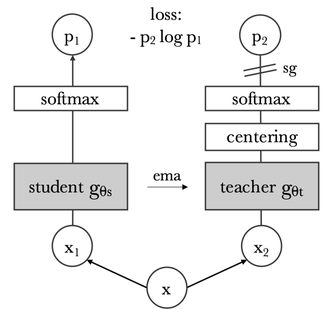
BOYL: (also a teacher-student) → EMA



Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al., "Bootstrap your own latent-a new approach to self-supervised learning," NeurIPS, 2020

Clustering:

DINO (also a teacher-student) → cross-entropy, sharping and centering.



Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jéegou, Julien Mairal, Piotr Bojanowski, and Armand Joulin, "Emerging properties in self-supervised vision transformers," CoRR, 2021.

Results

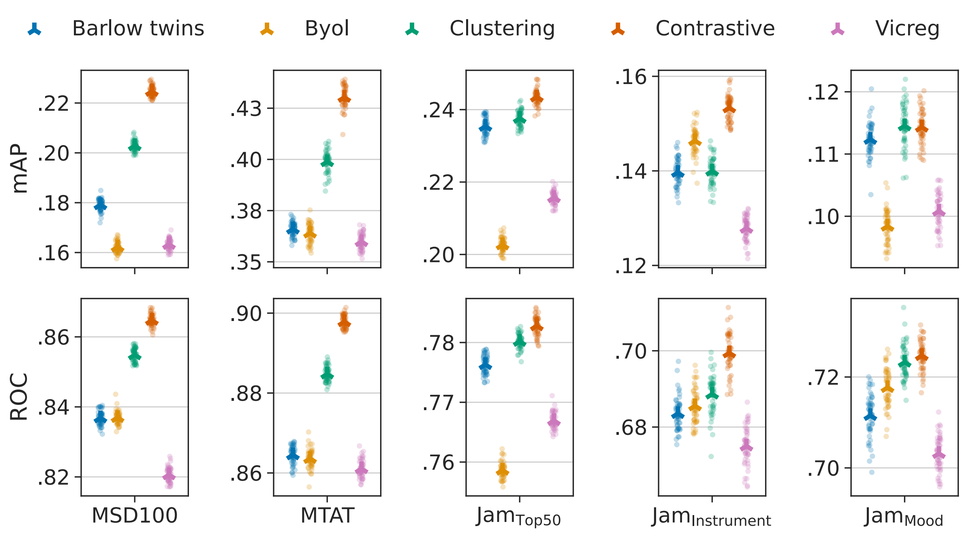
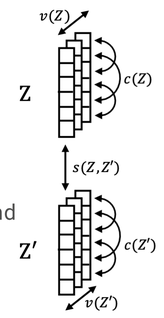


Fig.1 - Downstream results: We apply transfer learning to each task by training an MLP classifier on top of the embeddings generated by the frozen pretext model. We utilize bootstrapping. Each dot represents the metric of a resampled batch. The marker indicates the mean of each result.

Feature statistics:

VICReg: → balance between variance, invariance and covariance



v : maintain variance
c : bring covariance to zero
s : minimize distance

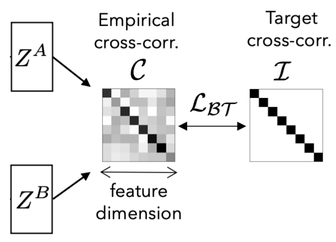
T : distribution of transformations
t,t' : random transformations

f_θ, f'_θ : encoders
h_ϕ, h'_ϕ: expanders

I : batch of images
X, X' : batches of views
Y, Y' : batches of representations
Z, Z' : batches of embeddings

Adrien Bardes, Jean Ponce, and Yann Lecun, "VICReg: Variance-invariance-covariance regularization for self-supervised learning," in ICLR, 2022

Barlow-Twins: diagonalisation and independence of each embedding dimension → Cross-correlation



Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny, "Barlow twins: Self-supervised learning via redundancy reduction," in ICML, 2021

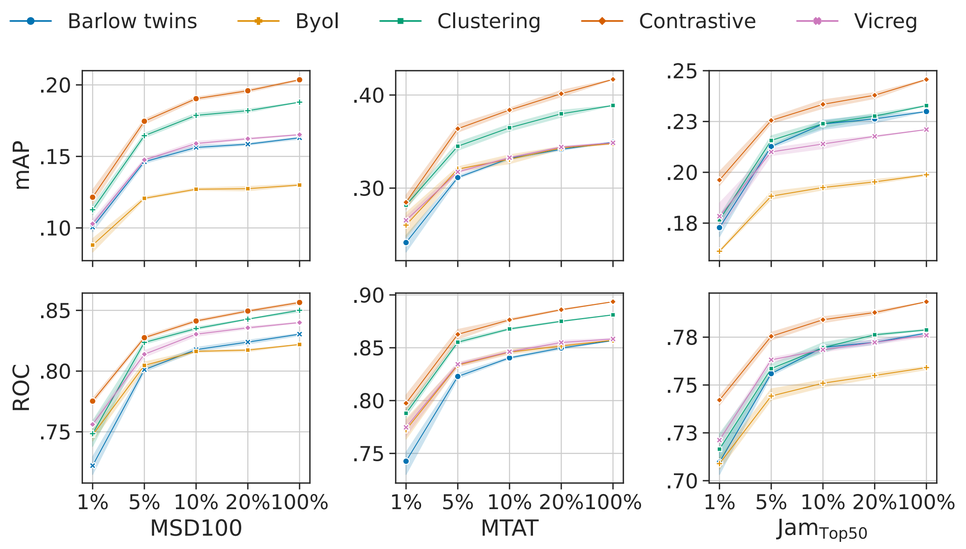


Fig.2 – Limited data results. Each dataset's train set is randomly sampled four times at four different percentages. We report the mean test set metrics obtained for each approach and include the results using the full train set to showcase each model's performance compared to these.

Takeaway

- The study compared the performance of various self-supervised methods for music tagging using a simple ResNet model trained on a diverse catalogue of millions of tracks.
- Choosing a relevant pretext task is crucial for effective self-supervised learning in the music domain.
- Contrastive learning consistently outperforms other methods in downstream tagging tasks, with minimal hyperparameter tuning required.
- Clustering shows promise but requires further investigation to address issues such as hyperparameter tuning and collapse.
- All methods yield consistent results with limited data.

Future work

- Further exploration and refinement of the clustering method are needed to improve performance and address the collapse issue.
- Investigating strategies to enhance the performance of other self-supervised pretext tasks in the music domain.
- Explore strategies to enhance the performance of self-supervised methods beyond contrastive learning in music applications.

49th IEEE International Conference on Acoustics, Speech, & Signal Processing (ICASSP 2024)

research@deezer.com
https://github.com/deezer/multi-view-ssl-benchmark