SUPERVISED AND UNSUPERVISED LEARNING OF AUDIO REPRESENTATIONS FOR MUSIC UNDERSTANDING

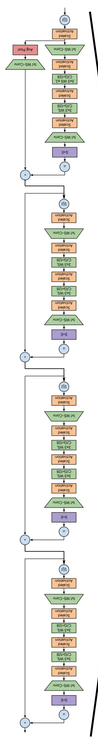
M. C. McCallum, F. Korzeniowski, S. Oramas, F. Gouyon, A. F. Ehmann

Sirius XM, USA

OBJECTIVES

- Provide a **broad set of baselines** for music understanding tasks
- **Compare** the effectiveness of **supervised and unsupervised learning** objectives **at scale**
- Investigate the **impact of training dataset content** and batch size for training unsupervised models
- Release a model to **enable and accelerate downstream research** in audio and / or multimodal understanding for music.

MODEL



Short Fast Normalizer Free Net (SF-NFNet-F0)

Stage	Slow path	Fast path
spectrogram	-	-
data layer	stride 4, 1	stride 1, 1
stem1	$1 \times 3, 16$ stride 2, 2	$3 \times 3, 2$ stride 2, 2
stem2	$1 \times 3, 32$ stride 1, 1	$3 \times 3, 4$ stride 1, 1
stem3	$1 \times 3, 64$ stride 1, 1	$3 \times 3, 8$ stride 1, 1
stem4	$3 \times 3, 128$ stride 2, 2	$3 \times 3, 16$ stride 2, 2
block1	$\begin{bmatrix} 1 \times 1, 128 \\ 1 \times 1, 128 \\ 1 \times 3, 128 \\ 1 \times 1, 256 \end{bmatrix} \times 1$	$\begin{bmatrix} 1 \times 1, 16 \\ 3 \times 1, 16 \\ 1 \times 3, 16 \\ 1 \times 1, 32 \end{bmatrix} \times 1$
block2	$\begin{bmatrix} 1 \times 1, 256 \\ 1 \times 1, 256 \\ 1 \times 3, 256 \\ 1 \times 1, 512 \end{bmatrix} \times 2$	$\begin{bmatrix} 1 \times 1, 32 \\ 3 \times 1, 32 \\ 1 \times 3, 32 \\ 1 \times 1, 64 \end{bmatrix} \times 2$
block3	$\begin{bmatrix} 1 \times 1, 768 \\ 3 \times 1, 768 \\ 1 \times 3, 768 \\ 1 \times 1, 1536 \end{bmatrix} \times 6$	$\begin{bmatrix} 1 \times 1, 96 \\ 3 \times 1, 96 \\ 1 \times 3, 96 \\ 1 \times 1, 192 \end{bmatrix} \times 6$
block4	$\begin{bmatrix} 1 \times 1, 768 \\ 3 \times 1, 768 \\ 1 \times 3, 768 \\ 1 \times 1, 1536 \end{bmatrix} \times 3$	$\begin{bmatrix} 1 \times 1, 96 \\ 3 \times 1, 96 \\ 1 \times 3, 96 \\ 1 \times 1, 192 \end{bmatrix} \times 3$
Global average pool & concatenate		

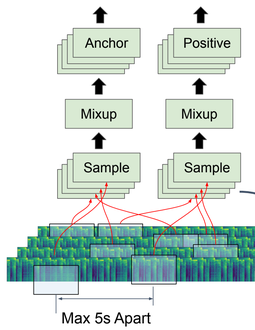
63M Params

PRETRAINING METHODOLOGY

Datasets

- **Musicset**
  - 1.8M music tracks
  - 117k hours
  - 500 tags
- **Audioset**
  - 1.7M 10s snippets
  - 4.8k hours
  - 527 tags

Sampling and Augmentation

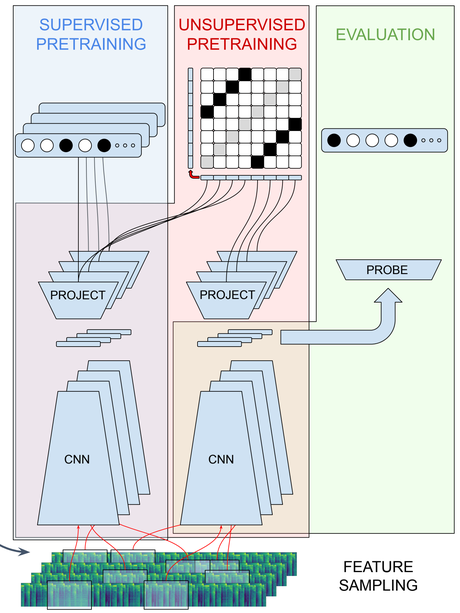


Sup

Loss: Binary CE
Batch Size: 512
Train Time: 30 hrs
Compute: 6xV100

ULarge/USmall

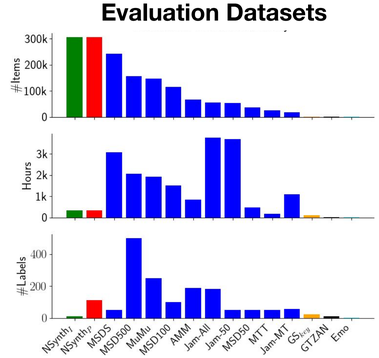
Loss: SimCLR
Batch Size: 1920/256 (pairs)
Training Time: 80hrs / 30hrs
Compute: 16xA100 / 8xV100



EVALUATION

- **7 Distinct Audio Collections**
- **15 Datasets / Annotations**
- Embeddings **global-average pooled** along track length.
- Probes consist of **MLPs**
- Probe **hyperparameters optimized**, respecting **same restrictions as previous** audio representation work

Evaluation Datasets

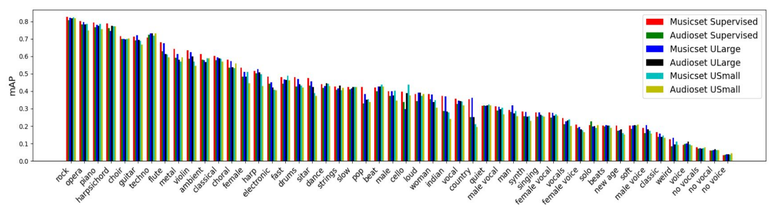


RESULTS

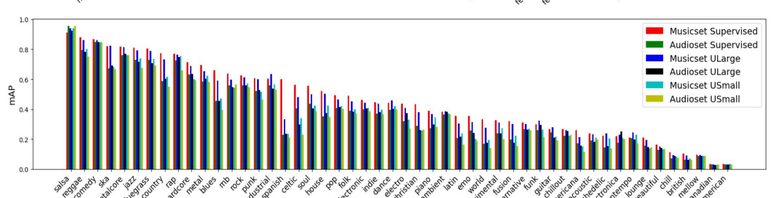
Model	MSDS mAP	MSDS ROC	MSD50 mAP	MSD50 ROC	MSD100 mAP	MSD100 ROC	MSD500 mAP	MSD500 ROC	MuMu mAP	MuMu ROC	AMM mAP	AMM ROC	Jam-MT mAP	Jam-MT ROC
Musicset-Sup	**0.363**	**0.903**	**0.459**	**0.913**	**0.346**	**0.906**	**0.169**	**0.898**	**0.257**	**0.908**	**0.180**	**0.791**	**0.161**	**0.786**
Audioset-Sup	0.308	0.880	0.375	0.883	0.278	0.877	0.128	0.874	0.191	0.867	0.156	0.760	0.137	0.749
Musicset-ULarge	0.351	**0.900**	0.438	**0.908**	0.321	**0.897**	0.152	**0.891**	0.235	0.893	**0.174**	**0.784**	**0.158**	**0.781**
Audioset-ULarge	0.311	0.885	0.377	0.886	0.276	0.878	0.121	0.873	0.162	0.855	0.156	0.763	0.142	0.765
Musicset-USmall	0.319	0.888	0.384	0.892	0.283	0.881	0.129	0.878	0.190	0.871	0.155	0.762	0.138	0.757
Audioset-USmall	0.286	0.876	0.353	0.878	0.251	0.870	0.110	0.868	0.152	0.850	0.151	0.753	0.136	0.753
SOTA	0.348 [15]	**0.897** [15]	0.386 [14]	**0.921** [14]	0.185 [22]	- -	- -	- -	0.888* [42]		0.163 [37]	0.773 [37]	**0.161**† [49]	**0.781**† [49]

Model	MTT mAP	MTT ROC	GTZAN Acc	NSynth$_P$ Acc	NSynth$_I$ Acc	Emo$_V$ r^2	Emo$_A$ r^2	GS$_{Key}$ W. Acc	Jam-50 mAP	Jam-50 ROC	Jam-All mAP	Jam-All ROC
Musicset-Sup	**0.413**	**0.917**	**0.835**	0.793	0.731	0.566	**0.726**	0.286	**0.321**	**0.843**	**0.162**	**0.839**
Audioset-Sup	0.386	0.904	0.748	0.819	0.676	0.341	0.545	0.210	0.284	0.822	0.135	0.813
Musicset-ULarge	**0.404**	**0.914**	0.735	**0.892**	0.740	0.577	0.700	0.667	**0.317**	**0.839**	**0.159**	**0.833**
Audioset-ULarge	0.391	0.906	0.672	0.805	0.721	0.438	0.624	0.287	0.285	0.826	0.131	0.816
Musicset-USmall	0.389	0.905	0.686	0.824	0.714	0.389	0.668	0.508	0.292	0.828	0.138	0.817
Audioset-USmall	0.375	0.897	0.648	0.777	0.698	0.386	0.609	0.197	0.268	0.817	0.127	0.809
Jukebox [23]	**0.414**	**0.915**	0.797	-	-	**0.617**	**0.721**	0.667	-	-	-	-
Prev. SF-NFNet-F0 [2]	0.395	-	-	0.880	**0.782**	-	-	-	-	-	-	-
SOTA Excl. [2, 23]	0.384 [37]	**0.92** [12]	0.821 [11]	- -	0.741 [43]	0.556 [44]	0.704 [45]	**0.796*** [46]	0.298 [47]	0.832 [47]	-	-

Magnatagatune



MSD50



KEY TAKE-AWAYS

- **Supervised models achieve SotA** on all multilabel tagging tasks
- **Unsupervised models generalize better** to novel tasks like pitch and key
- Music understanding models **perform better when pretrained on purely music data**

Musicset-ULarge Model Available Here:
https://github.com/PandoraMedia/music-audio-representations