mir_ref

Music Information Retrieval **Representation Evaluation Framework**

Christos Plachouras, Pablo Alonso-Jiménez, Dmitry Bogdanov
Music Technology Group, Universitat Pompeu Fabra, Spain

Code & Results

What is it?

`mir_ref` is an **open-source library** for **evaluating audio representations** (embeddings or others) on a variety of music-related downstream tasks and datasets.

It provides **ready-to-use** tasks, datasets, deformations, embedding models, and downstream models for **config-based**, **no-code** experiment orchestration. Components are **modular**, so it's easy to add custom embedding models, datasets, metrics etc. Audio-specific results **analysis and visualization tools** are also provided.

What is it for?

- Easily **reproducible**, **holistic** evaluation experiments
- Local aid for **embedding model development**
- **Benchmarking**
- To answer questions like:
  - How large should the downstream model be?
  - How densely should I sample embeddings?
  - How robust is my model to pitch shifting?
  - Can my model distinguish pitch classes?

How do I use it?

Clone/Fork `mir_ref`, install requirements, and run

```
$ python run.py -c my_config
```

to run all experiments in the config file. Individual components can be run with `deform`, `extract`, `train`, and `evaluate`. Sharing the config file allows anyone to reproduce your experiment.

Please give us feedback and tell us use-cases!

Why is it needed?

Representation evaluation in MIR is
- **fragmented**
- **tedious** to set up (gathering/handling data, complexity)
- **narrow-scoped** (robustness? efficiency? explainability?)

Downstream implementation details in embedding model papers

		model		optimization			output
	code	type	layer(s)	HPO	initial lr	wd	aggr.
EffNet-Discogs		MLP	512		$1e^{-3}$	$1e^{-5}$	pred.
MusiCNN	✓	SVM	NA		NA	NA	pred.
OpenL3		MLP	512-128	✓	$1e^{\{-5,..,-3\}}$	$1e^{\{-5,..,-3\}}$	pred.
NeuralFP		LC	NA		?	?	?
CLMR	✓	LC	NA		$3e^{-4}$	$1e^{-6}$	repr.
MERT	✓	MLP	512	✓	$1e^{\{-4,..,-2\}}$	?	repr.
COALA	✓	MLP	256		$1e^{-3}$	$1e^{-4}$	repr.
JukeMIR	✓	LC/MLP	NA/512	✓	$1e^{\{-5,..,-3\}}$	$1e^{\{-3,..,0\}}$	repr.
MuLaP	✓	MLP	512		$1e^{-3}$	$1e^{-2}$	pred.

	MTG J. genre	MTG J. instr.	MTG J. mood	MTG J. top50	GTZAN genre	MTAT tagging	MSD tagging	FMA genre	FMA identity	EMO emotion	GS key	NSynth instr.	Nsynth pitch	Vocalset singer	Vocalset tech.
EffNet-Discogs	✓	✓	✓	✓		✓		✓							
MusiCNN					✓										
NeuralFP					✓				✓						
CLMR						✓	✓								
MERT	✓	✓	✓	✓	✓	✓				✓	✓	✓	✓	✓	✓
COALA					✓							✓			
JukeMIR					✓	✓				✓	✓				
MuLaP	✓	✓	✓	✓		✓		✓		✓		✓			

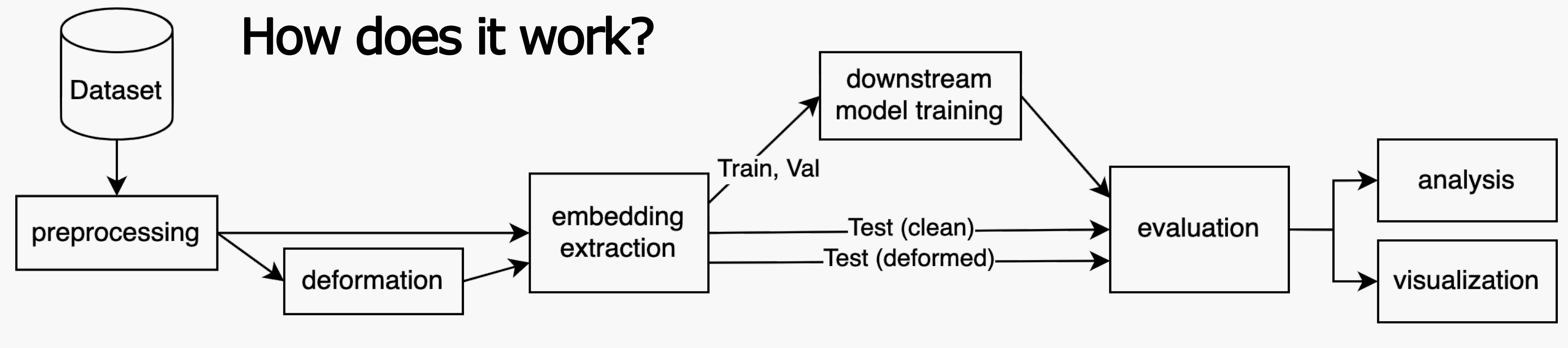
Example evaluation

We conducted an evaluation of 7 embedding models, 6 datasets and tasks, 4 deformations, and 5 downstream model configurations, and found:

- Most models struggle significantly with white noise and gain reduction, but do better with mp3 compression.
- The downstream setup often impacts performance significantly; some information is not linearly separable.
- Most models can't distinguish pitch classes.

(Scan QR for full results, or github.com/chrispla/mir_ref)

How does it work?


Dataset
preprocessing
deformation
embedding extraction
Train, Val
downstream model training
Test (clean)
Test (deformed)
evaluation
analysis
visualization