Deep embeddings with Essentia models

Pablo Alonso-Jiménez   Dmitry Bogdanov   Xavier Serra
Music Technology Group, Universitat Pompeu Fabra

Essentia is an open-source **C++/Python** library for audio signal processing, developed at the MTG-UPF and licensed under **Affero GPLv3**.

Functionalities

- Audio features
  - Spectral features
  - Rhythm and tempo
  - Tonality and melody
  - Fingerprinting
- Inference with **TensorFlow** models

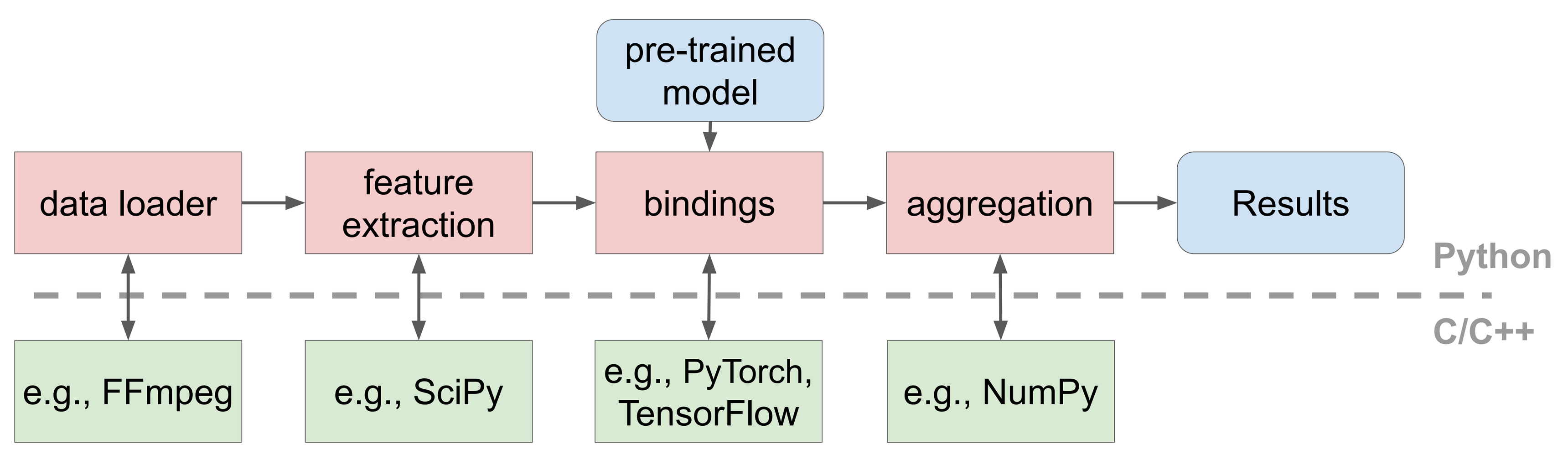
Design criteria

- **C++** with **Python** bindings
- **Large-scale** deployment
- **Real-time** processing
- **Cross-platform**
  - (Linux, MacOS, Win, iOS, Android, JS)

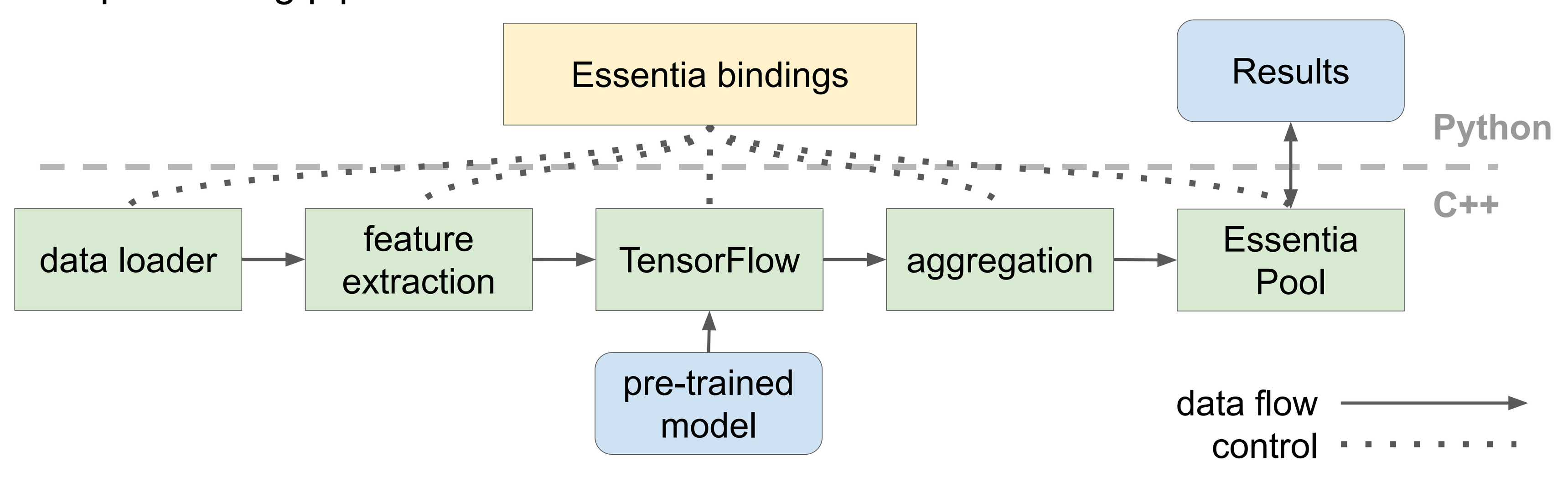
https://essentia.upf.edu/

TensorFlow integration in Essentia

Data processing pipeline found in **common MIR** projects:



Data processing pipeline in **Essentia:**



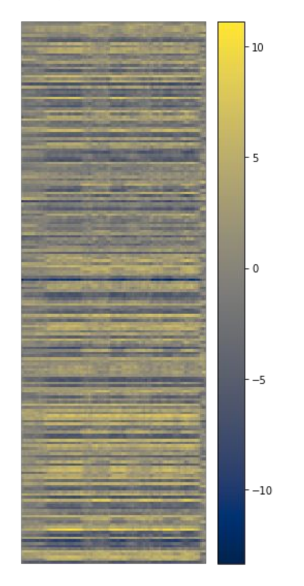
Suitable framework for **research** and **deployment** scenarios.

Pre-trained models

We have prepared various **MIR** models for several tasks. They can also be used as **embeddings extractors**. The following plots show the embeddings produced with our models for a 2 minutes rock track.

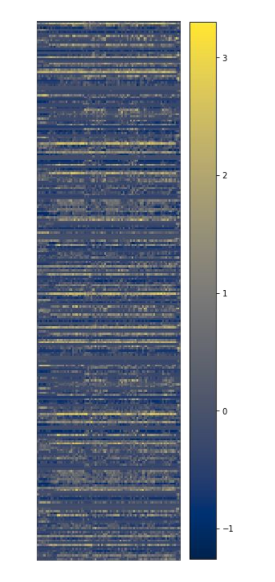
MusiCNN

music auto-tagging
787K parameters
200 embedding dimensions
220/350K training size
fully supervised



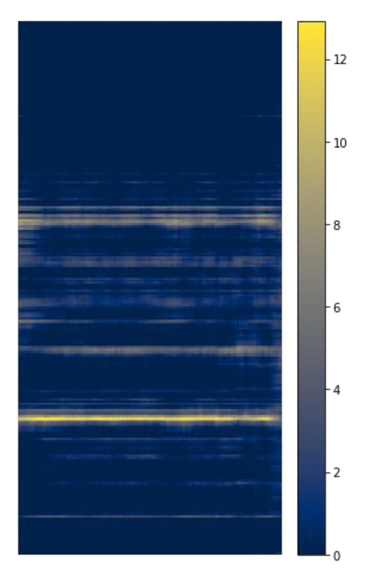
VGG-I

music auto-tagging
605K parameters
256 embedding dimensions
220/350K training size
fully supervised



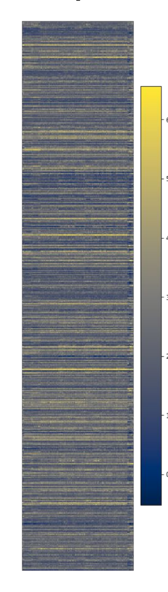
Tempo-CNN

tempo estimation
1.2M parameters
256 embedding dimensions
11K training size
fully supervised



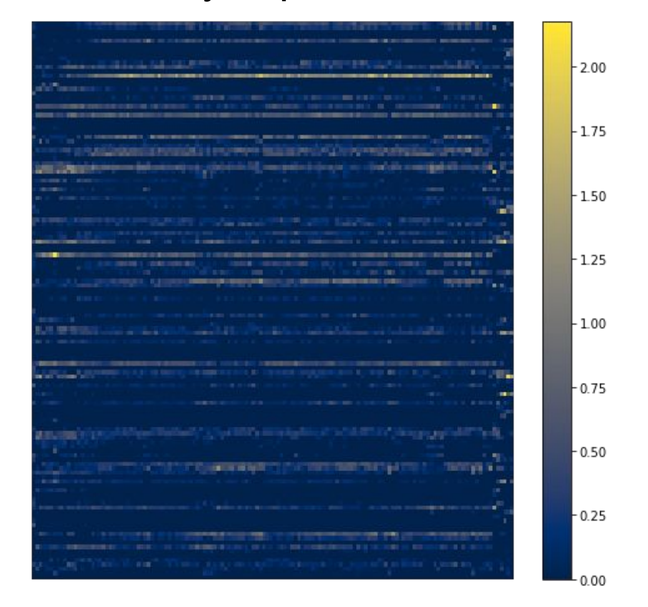
OpenL3

feature extractor
4.7M parameters
512 embedding dimensions
296K training size
self supervised



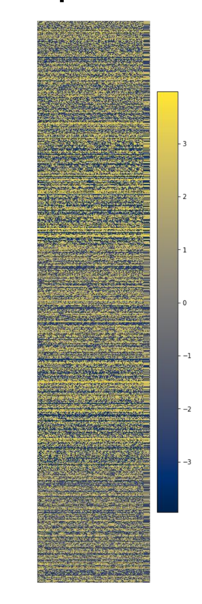
VGGish

feature extractor
62M parameters
128 embedding dimensions
70M training size
fully supervised



Spleeter

source separation
49M parameters
1280 embedding dimensions
unknown training size
fully supervised



MusiCNN and **VGG-I** are trained on two versions of **MSD-Last.fm** targeting the top 50 and the top 200 tags (*T200 models*), resulting in training sizes of 220K and 350K.

These models are available online at https://essentia.upf.edu/models/

Extractring embeddings

Essentia has dedicated algorithms to perform **inference** with each model.

With the `output` parameter we can select the **layer** of the network to retrieve. It is defaulted to the main task of the network (*e.g., music tag indices, bpm bins, separated audio*) but it can be set to point to any layer of interest.

On the **music auto-tagging** models we retrieved the penultimate layer as embeddings, on **Tempo-CNN** the logits of the last layer, on **Spleeter** we used the concatenation of the bottleneck layers of each stem as embeddings, and on the **feature extractor** models we used directly the output proposed by the authors.

```
audio = MonoLoader(filename='your_song.mp3', sampleRate=16000)()

musicnn_embs = TensorflowPredictMusiCNN(graphFilename='msd-musicnn-1.pb',
                                        output='model/dense/BiasAdd')(audio)
```

```
audio = MonoLoader(filename='your_song.mp3', sampleRate=11025)()

tempocnn_embs = TensorflowPredictTempoCNN(graphFilename='deepsquare-k16-3.pb',
                                          output='1x1/Relu0_reshape')(audio)
```

More examples at: https://essentia.upf.edu/machine_learning.html

Downstream tasks

We compared the capabilities of the pre-trained models as feature extractors in 16 **downstream** tasks.

genre recognition

dortmund	gtzan	rosamerica
alternative, blues, electronic, folk-country, funksoulrnb, jazz, pop, raphiphop, rock	blues, classic, country, disco, hiphop, jazz, metal, pop, reggae, rock	classic, dance, hip hop, jazz, pop, rhythm and blues, rock, speech
1820 excerpts	1000 excerpts	400 full tracks

mood detection

acoustic	aggressive	electronic	happy	party	relaxed	sad
acoustic, non-acoustic	aggressive, non-aggressive	electronic, non-electronic	happy, non-happy	party, non-party	relaxed, non-relaxed	sad, non-sad
321 full tracks	280 full tracks/excerpts	332 full tracks/excerpts	302 excerpts	349 full excerpts	446 full tracks/excerpts	230 full tracks/excerpts

miscellaneous audio tasks

voice/instrumental	tonal/atonal	gender	danceability	fs-loop-ds	urbansound8k
voice, instrumental	tonal, atonal	female, male	danceable, non-danceable	bass, fx, melody, percussion, other	air conditioner, car horn, children playing, dog bark, drilling, engine idling, gunshot, jackhammer, siren, street music
1000 excerpts	345 full tracks	3311 full tracks	306 full tracks	2104 excerpts	8732 excerpts

The **pre-trained** models are compared by training a **multilayer perceptron** for each task on top of the proposed embeddings.

The models are compared in two ways. Using **5-fold cross-validation (5F)** and evaluating in the **MTG-Jamanendo dataset (JA)** for which we collected annotations following the taxonomies of all our tasks.

The table below also contains a column for the **accuracy drop (AD),** the difference between both metrics as a proxy for the **generalization** capabilities on each case.

The best embeddings for each task are shaded light/medium grey for each metric. The results are expressed in terms of **class-weighted accuracies**.

	MusiCNN			MusiCNN-T200			VGG-I			VGG-I-T200			VGGish			OpenL3			Spleeter			Tempo-CNN		
Task	5F	JD	AD	5F	JD	AD	5F	JD	AD	5F	JD	AD	5F	JD	AD	5F	JD	AD	5F	JD	AD	5F	JD	AD
dortmund	61	46	15	46	41	5	54	12	42	26	22	4	50	48	2	38	21	17	35	24	11	16	17	-1
gtzan	86	54	32	79	47	32	83	53	30	46	26	20	84	62	22	58	14	44	57	32	25	26	15	11
rosamerica	94	58	36	90	60	30	93	59	34	66	32	34	93	59	34	84	24	60	70	33	37	46	33	13
voice/instrum.	98	83	15	93	82	11	97	79	18	78	71	7	98	87	11	89	54	35	76	65	11	58	55	3
tonal/atonal	87	60	27	91	61	30	92	61	31	78	55	23	93	64	29	89	51	38	89	60	29	70	59	11
gender	87	82	5	79	76	3	84	80	4	70	65	5	83	79	4	55	53	2	55	62	-7	51	53	-2
danceability	98	66	32	94	70	24	94	68	26	71	62	9	94	70	24	90	58	32	90	62	28	66	59	7
acoustic	96	70	26	93	74	19	93	73	20	83	64	19	93	74	19	89	55	34	89	62	27	75	61	14
aggressive	97	72	25	97	76	21	99	67	32	82	70	12	99	67	32	91	52	39	93	58	35	69	59	10
electronic	93	78	15	88	77	11	88	76	12	74	70	4	94	81	13	77	57	20	77	63	14	64	55	9
happy	86	57	29	77	55	22	89	62	27	69	58	11	86	60	26	76	51	25	70	55	15	68	57	11
party	92	77	15	92	75	17	64	68	-4	84	73	11	90	75	15	77	57	20	87	66	21	73	63	10
relaxed	89	71	18	86	67	19	91	71	20	79	65	14	90	71	19	81	53	28	80	61	19	72	60	12
sad	87	67	20	88	65	23	86	68	18	83	62	21	89	65	24	85	55	30	83	60	23	84	62	22
fs-loop-ds	56	-	-	49	-	-	53	-	-	38	-	-	53	-	-	53	-	-	46	-	-	24	-	-
urbansound8k	81	-	-	40	-	-	82	-	-	35	-	-	89	-	-	77	-	-	70	-	-	10	-	-

- MusiCNN tends to be more successful in **5-fold cross-validation**
- VGG-like models tend to suffer less **accuracy drop** (*better generalization*)
- In the MusiCNN model, **more tags** and **data** tend to be beneficial for generalization
- Models not trained for classification (*OpenL3, Spleeter, Tempo-CNN*) are not so powerful

Uses in MIR

Our main goal is to provide fast **C++ inference** for state-of-the-art deep learning models in **Essentia** suitable for **deployment** in diverse MIR applications.

We host a collection of **models** for specific use-cases (*auto-tagging, tempo estimation, source separation, music classification by genre, mood, and instrumentation*).

Some of these models produce **embeddings** suitable for **transfer learning**.

upf. Universitat Pompeu Fabra Barcelona
MTG Music Technology Group

ESSENTIA