Evaluation of Deep Audio Representations for Semantic Sound Similarity

Recep Oguz Araz, Dmitry Bogdanov, Pablo Alonso-Jiménez, Frederic Font
Music Technology Group, Universitat Pompeu Fabra, Spain

- Audio-sharing platforms, such as Freesound, offer sound similarity functions (query-by-example) for content-based retrieval.
  - They typically use manually-engineered audio representations that:
    - Do not capture audio semantics
    - Do not leverage recent developments in deep representation learning.
- **Our goal** is to identify the best deep audio representation for semantic sound similarity and deploy it within the Freesound platform.
- GitHub repository: https://github.com/raraz15/semantic-sound-similarity

Deep representation learning models which use **audio and language** modalities together perform **significantly better** in the **semantic sound similarity** task.

Modality	Model	Pre-training Objective	Param.	Data	Dim.	MAP@N ↑		MR1 ↓
						N=15	N=150	
A	Freesound	-	-	-	846	0.09	0.03	43
A	VGGish	Classification	62M	80M	128	0.20	0.11	27
	YAMNet	Classification	4M	2M	1024	0.27	0.15	25
	FSD-SINet	Classification	5M	51K	512	0.33	0.19	18
	BEATs	Classification	90M	1.8M	768	0.37	0.20	13
A & I	OpenL3	Audiovisual correspondence	5M	296K	512	0.15	0.06	31
	CAV-MAE	Audiovisual correspondence	85M	1.8M	768	0.33	0.18	15
A & L	CLAP2022	Contrastive alignment	81M	128K	1024	0.38	0.25	20
	LAION-CLAP	Contrastive alignment	31M	2.5M	512	**0.53**	**0.37**	**8**
	CLAP2023	Contrastive alignment	31M	4.6M	1024	0.50	0.34	10
	Pengi	Question-answering	31M	3.4M	1024	0.49	0.35	12
A & I & L	AudioCLIP	Contrastive alignment	30M	1.8M	1024	0.06	0.02	56
	Wav2CLIP	Contrastive alignment	12M	200K	512	0.12	0.04	36
A & I & L & O	ImageBind	Contrastive alignment	85M	1.8M	1024	0.29	0.17	22

Model performances on the semantic sound similarity task

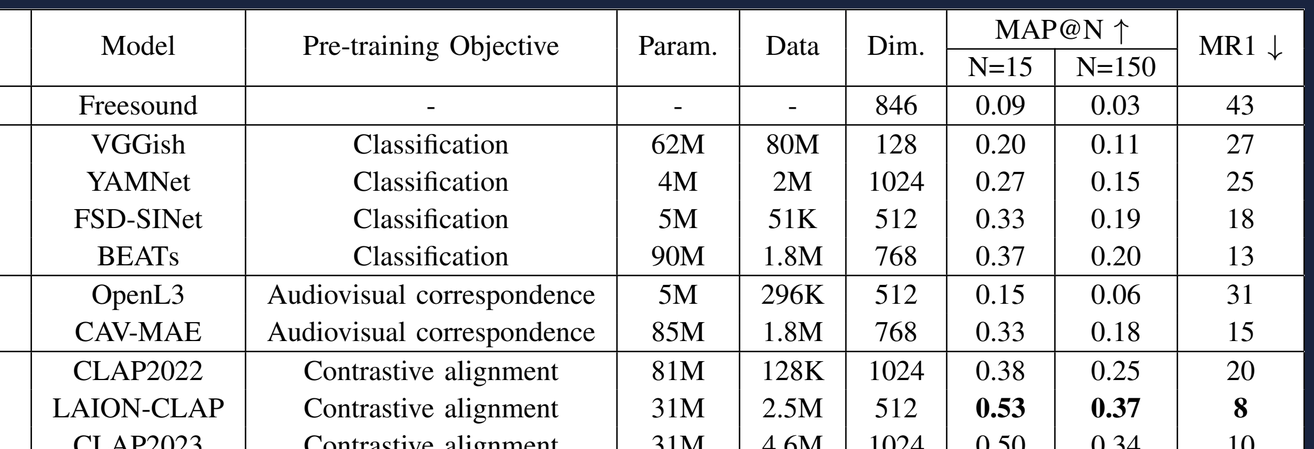
Methodology

Data: FSD50K evaluation set - 10,231 audio clips with 200 sound class labels from 7 sound families.

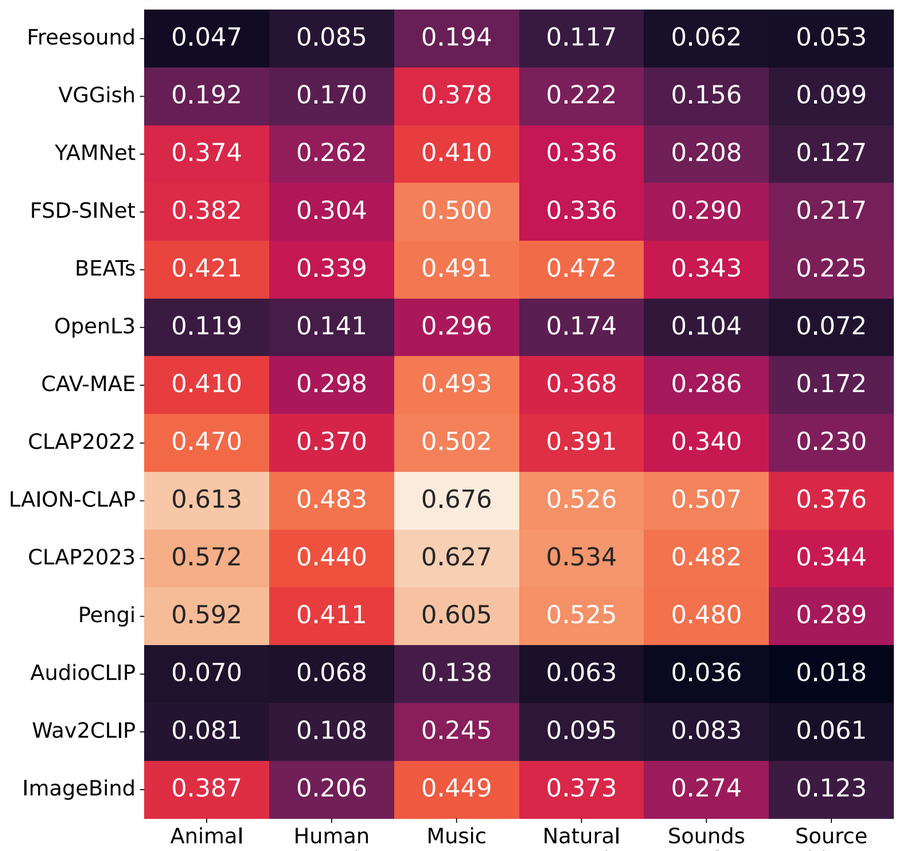
Task: Evaluate the audio representations of Freesound and 13 neural networks. For each representation, optimize:
- Embedding processing parameters,
- Similarity search functions.

Evaluation:
- **Objective :** MAP@15, MAP@150, and MR1. On,
  - Class-wise (e.g., birds, bells, motor)
  - Family-wise
  - Macro-averaged
- **Subjective:** Using our web interface, available online.

Results



Macro-averaged MAP@15 scores of QbE systems that use FSD-SINet VGG42-tlpf



Family-wise MAP@15 scores

Conclusion

Learning paradigm
- Input modalities are crucial for retrieval performance
  - Audio & language > audio > audio & image
- LAION-CLAP works the best across all families.
- Models that outperform others in the sound event classification task underperform in the semantic sound similarity task.

Embedding processing
- Dimensionality can be greatly reduced by as much as 90%, while increasing performance slightly.

Similarity search
- Maximum Cosine Similarity Search (MCSS) works the best for all.

Universitat Pompeu Fabra Barcelona
MTG Music Technology Group
freesound

Recep Oguz Araz
PhD Student
recepoguz.araz@upf.edu
https://github.com/raraz15

Scan the QR code to access the project's GitHub repository! with the **code, full paper**, and **additional material**.