Simultaneously Learning Robust Audio Embeddings and Balanced Hash Codes for Query-by-Example

Anup Singh[1,2], Kris Demuynck[1], Vipul Arora[2]
[1]Ghent University, [2]Indian Institute of Technology-Kanpur

Introduction

What is Audio Fingerprinting?
Audio fingerprinting generates compact summary of an audio signal.

What are the applications?
Music identification, second-screen apps, broadcast monitoring, etc.

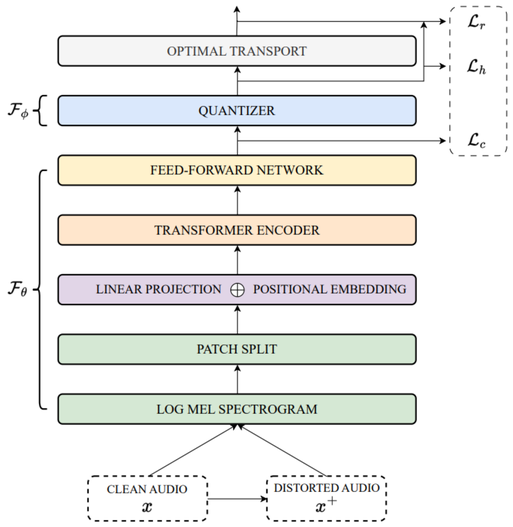
What did existing methods do?
Conventional methods rely on handcrafted audio features to design audio fingerprints. Recent methods deploy deep learning to generate compact audio fingerprints.

What makes it challenging?
✗ Robustness
✗ Efficient audio indexing
✗ Comprehensive audio search
✗ Practically deployable

What do we do?
✓ Discriminative embeddings using Transformer encoder
✓ Contrastive learning framework to achieve robustness
✓ Efficient indexing by learning balanced hash codes
✓ Adopted Optimal Transport to achieve balanced clustering
✓ Fine-grained audio search

Research Questions:
- How to learn embeddings and hash codes to make audio indexing and retrieval efficient?
- How to achieve balanced clustering?

Method



- Balanced clustering objective is combinatorial optimization problem **- Difficult to solve!**
- We adopt the optimal transport framework to obtain approximate solution.
- We also propose simple yet efficient subsequent search strategy to precisely locate the query timestamp.

Training Objectives

$$\mathcal{L}_c = -\log \frac{e^{(\mathcal{F}_\theta(x)\cdot\mathcal{F}_\theta(x^+))/\tau}}{e^{(\mathcal{F}_\theta(x)\cdot\mathcal{F}_\theta(x^+))/\tau} + \sum_{x^-} e^{(\mathcal{F}_\theta(x)\cdot\mathcal{F}_\theta(x^-))/\tau}}$$

$$\mathcal{L}_h = -\log \frac{e^{(q(x)\cdot q(x^+))/\tau}}{e^{(q(x)\cdot q(x^+))/\tau} + \sum_{x^-} e^{(q(x)\cdot q(x^-))/\tau}}$$

$$\mathcal{L}_r = -\left[\text{softmax}\left(\frac{s_{k^*}(x)}{\tau'}\right) + \text{softmax}\left(\frac{s_{k^*}(x^+)}{\tau'}\right)\right]$$

Balanced Clustering Objective

$$\max_A \mathbb{E}_x\Big[\sum_{k=1}^{2^K} A_k(x) s_k(x)\Big] \quad \text{subject to}$$

$$A_k(x) \in \{0,1\},\ \sum_{k=1}^{2^K} A_k(x) = 1\ \forall x\ ,\ \mathbb{E}_x[A_k(x)] = \frac{1}{2^K}\ \forall k$$

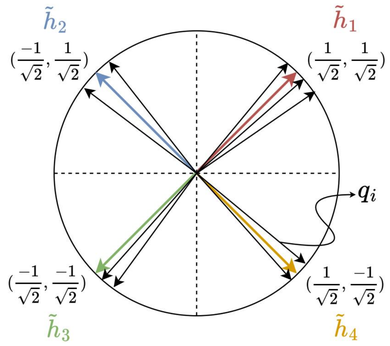


Illustration of balanced clustering in two-dimensional space

Experiments

- Database: Free Music Archive (FMA)
- Evaluation: Identify as a match if the located timestamp is within +-50 ms.
- Baselines: NAFP[1] and Audfprint[2]
- We compare our indexing performance gain against the Locality Sensitive Hashing (LSH[3])

Findings

- Our **fingerprints are robust and discriminative** and thus achieve superior performance.
- Our system delivers **reliable performance with short query lengths** at high distortion levels.
- We achieve **more than two-folds search speedup** than LSH due to balanced hash codes.
- We achieve **higher efficiency as well efficacy** compared to baselines.

Results

Length		Noise 0	5	10	15	20	Noise + Reverb 0	5	10	15	20	Reverb 0.2	0.4	0.5	0.7	0.8	speedup
1	NAFP + LSH	50.7	69.7	73.7	76.0	76.9	20.8	43.9	55.5	58.5	58.9	62.5	61.5	58.9	49.3	45.0	1x
	FE + LSH	**66.6**	**82.6**	87.6	90.0	90.6	**44.8**	**62.6**	73.8	**79.4**	**82.0**	83.8	83.8	78.4	**69.4**	64.4	1x
	FE + HT (Ours)	64.9	81.7	**87.9**	**90.1**	**90.9**	41.9	62.0	**74.6**	79.2	81.2	**84.6**	**82.3**	**80.0**	68.7	**64.5**	**2.4x**
2	NAFP + LSH	71.0	83.4	85.5	87.6	87.5	37.8	65.6	73.5	75.2	76.7	78.6	76.4	75.7	68.0	59.5	1x
	FE + LSH	**80.4**	88.2	91.6	93.2	94.8	**63.4**	78.2	84.6	86.0	85.4	90.4	86.4	85.0	74.8	67.8	1x
	FE+HT(Ours)	80.2	**89.4**	**93.1**	**94.1**	**94.9**	61.4	**80.8**	**85.0**	**87.5**	**88.3**	**92.4**	**87.7**	**86.4**	**77.3**	**70.7**	**2.4x**
3	NAFP + LSH	77.7	84.8	88.9	89.2	89.1	50.1	72.6	80.0	79.5	80.1	83.6	79.9	77.9	70.2	63.8	1x
	FE + LSH	83.2	88.4	92.6	94.4	95.2	**71.6**	82.6	85.6	86.2	87.4	91.8	87.8	85.6	74.4	68.4	1x
	FE+HT(Ours)	**84.7**	**90.8**	**95.5**	**96.3**	**97.1**	70.7	**84.1**	**88.6**	**89.0**	**90.1**	**93.8**	**90.3**	**87.9**	**77.9**	**71.6**	**2.3x**
5	NAFP + LSH	82.6	89.2	90.2	90.5	91.2	60.2	79.1	83.4	82.8	83.1	85.6	83.4	79.3	74.1	65.7	1x
	FE + LSH	85.6	90.0	92.8	94.2	95.8	80.0	87.0	87.1	87.6	87.2	93.8	88.8	86.6	75.0	69.0	1x
	FE+HT(Ours)	**88.0**	**93.4**	**95.3**	**96.2**	**97.4**	**80.6**	**88.6**	**90.8**	**91.3**	**91.5**	**96.4**	**91.1**	**88.9**	**78.7**	**71.2**	**2.4x**

References

1. S. Chang, D. Lee, J. Park, H. Lim, K. Lee, K. Ko, and Y. Han, "Neural audio fingerprint for high-specific audio retrieval based on contrastive learning," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 3025–3029.
2. https://github.com/dpwe/audfprint
3. Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni, "Locality-sensitive hashing scheme based on p-stable distributions," in Proceedings of the twentieth annual symposium on Computational geometry, 2004, pp. 253–262.

Paper Link

Contact: anup.singh@ugent.be