Harnessing High-Level Song Descriptors towards Natural Language-Based Music Recommendation

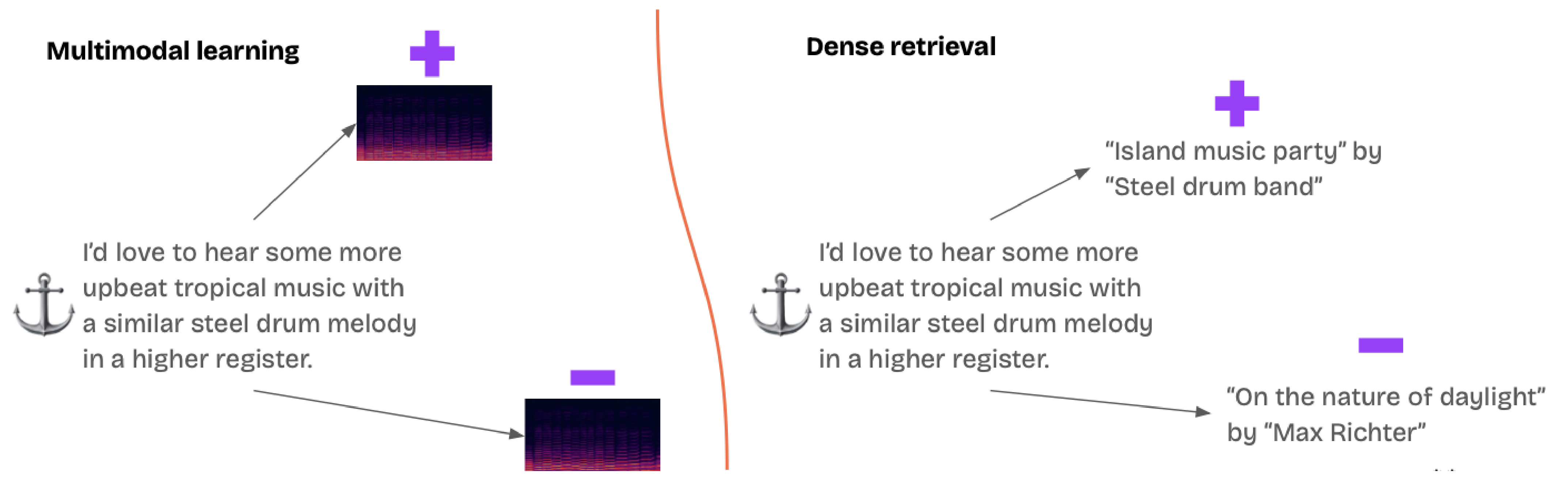
Elena V. Epure, Gabriel Meseguer Brocal, Darius Afchar, Romain Hennequin

1. Research Context

NL interfaces for search and recommendations are highly in demand

- integration of both search and recommendation
- ideally conversational: the system engages in a conversation
- promising results with LLMs, but still many challenges

Existing approaches to music focus on learning to embed user input / song description together with music representation (from audio [1] or from embedded metadata [2]) using contrastive learning.



Multimodal learning

I'd love to hear some more upbeat tropical music with a similar steel drum melody in a higher register.

Dense retrieval

"Island music party" by "Steel drum band"

I'd love to hear some more upbeat tropical music with a similar steel drum melody in a higher register.

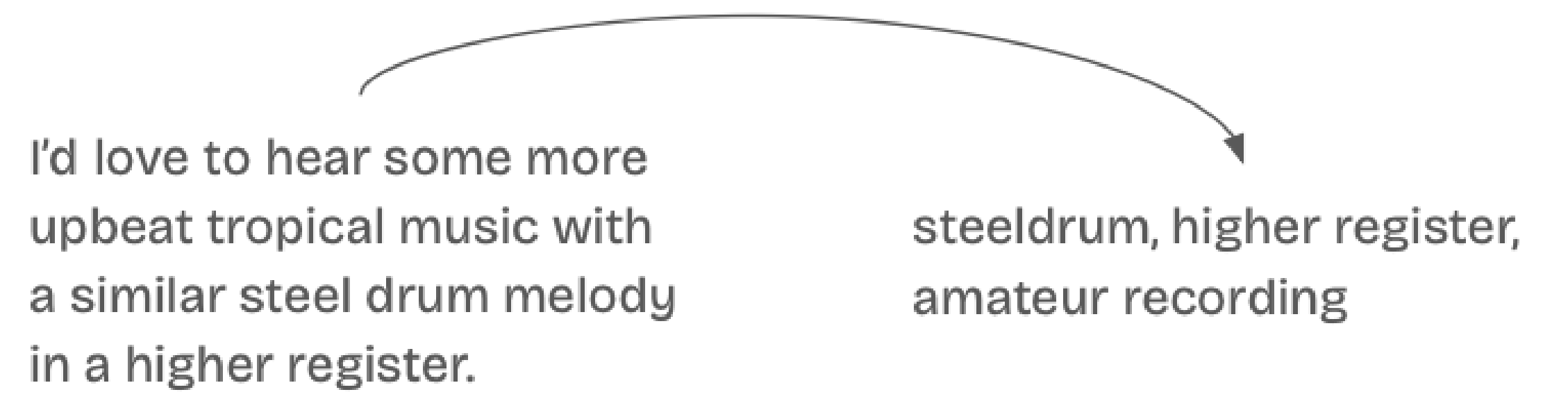
"On the nature of daylight" by "Max Richter"

2. Research Objective

There is still a high semantic gap to bridge between the high-level user input and low-level music information such as audio or embedded metadata (e.g. song title, artist name).

What if we rely on song descriptors instead?

New objective: given **user input**, retrieve / recommend songs with **matching descriptor set**?



I'd love to hear some more upbeat tropical music with a similar steel drum melody in a higher register.

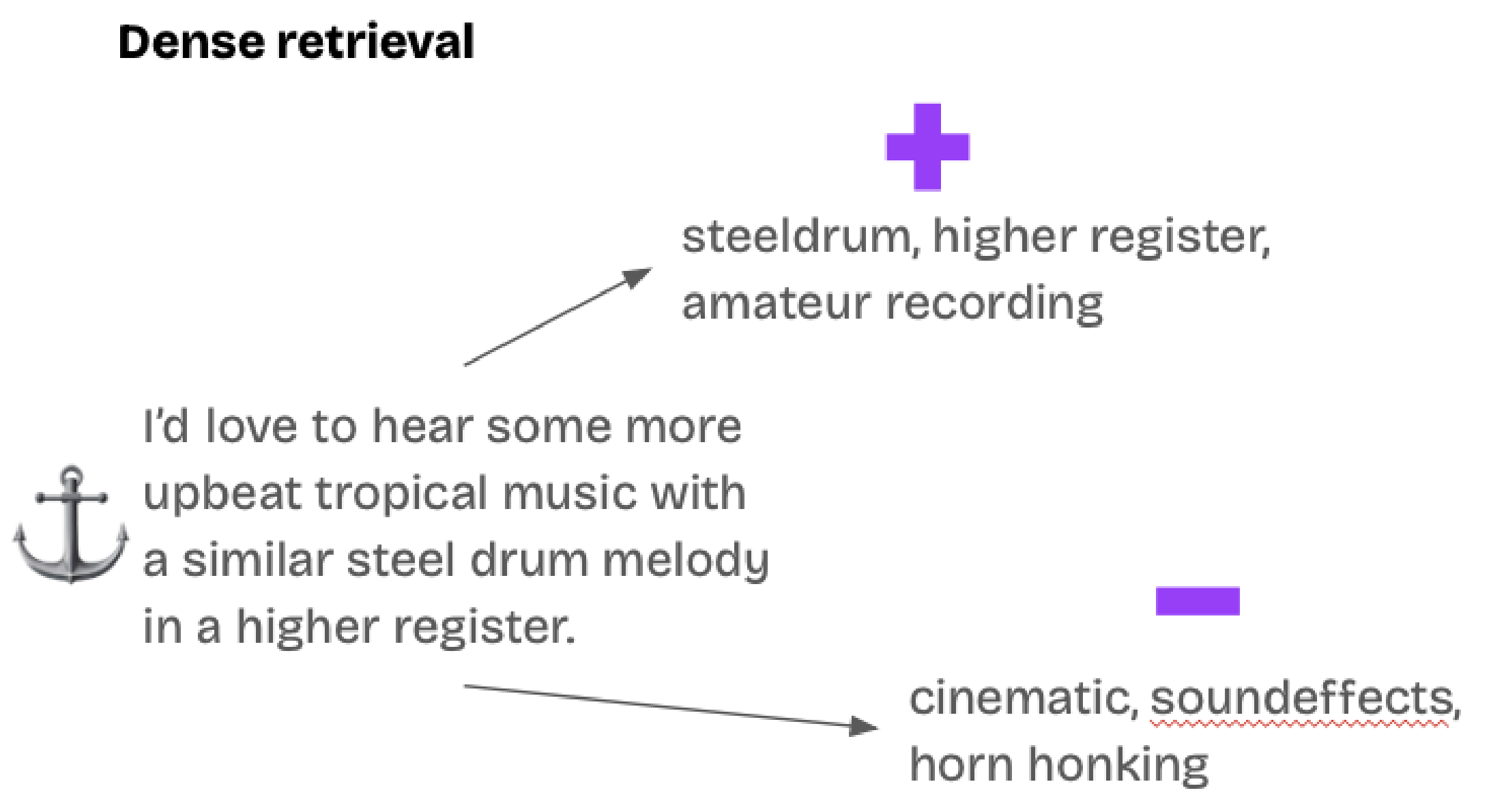
steeldrum, higher register, amateur recording

3. Proposed Method

Learn to embed user input / song descriptions with music descriptor sets

How? Generative Pseudo-labeling [3]:

1. **Hard negative mining** with a music fine-tuned cross-encoder
2. **Pseudo-labeling** of training dataset: $\delta_s = g(r, \mathrm{concat}(T_s)) - g(r, \mathrm{concat}(T_s^-))$
3. **Fine-tuning a bi-encoder** to mimic the score margin δ_s



Dense retrieval

steeldrum, higher register, amateur recording

I'd love to hear some more upbeat tropical music with a similar steel drum melody in a higher register.

cinematic, soundeffects, horn honking

4. Experiments

No large dataset linking natural language user preferences (r) with high-level song descriptors (T_s) exist in music recommendation.

Re-purpose a dataset created for music captioning, LP-MusicCaps (MSD, MTT, MC) [4].

- Split paragraphs in sentences
- Sample up to 3 variations of high-level descriptors from the original T_s: both from overlapping and non-overlapping ones.
- Rephrase song descriptions as requests for recommendations with Llama-3-8B-Instruct (LP-MusicCaps MC test split)

test	#Requests	#Descriptors	Shared Words
MTT	4462	188	0.15
MSD	34631	1054	0.23
MC	2357	6930	0.41
MC$_{reco}$	2357	6930	0.34

6. Qualitative Analysis

Song Description / Request

This pop song features a captivating teen female vocal delivering melodic singing over an acoustic guitar and simple drum track that evoke a melancholic, emotional vibe.

Music Descriptors (Ground-truth)
acoustic guitar, emotional, teen female vocal, melodic singing, simple drum track, pop music

Top 5 Predictions by `Ours`

1. acoustic guitar, emotional, teen female vocal, melodic singing, simple drum track, pop music
2. acoustic guitar
3. acoustic drums
4. bass drum
5. pop, acoustic rhythm guitar, quiet playback, resonant, heartfelt, noisy, emotional, passionate female vocal

5. Results

`Tf-idf` is a strong encoder on datasets where there is a large exact term overlap between the song description r and the high-level descriptors T_s.

`BERT` and `MPNET` achieve poor results, likely because of insufficient context to derive meaningful embeddings for short text (words, phrases).

`msmarco-BERT` is the best dense retrieval model from the `sentence-transformers`.

Fine-tuning `BERT` (or variants) on text-audio similarity leads to better results(`CLAP`$_{text}$ and `TTMR`$_{text}$), yet still unsatisfactory overall.

Our approach achieves significantly higher scores than all the other dense retrievers and by design, it should generalise.

Baselines' scores on the rephrased MC test set are lower compared to those on the original MC dataset, but this might be due to the information loss.

	`Tf-Idf`	`CLAP`$_{text}$	`TTMR`$_{text}$	`BERT`	`all-MiniLM`	`msmarco-BERT`	`Ours`
MTT	57.7 ± 0.8	13.5 ± 0.3	7.8 ± 0.6	4.8 ± 0.4	33.3 ± 0.6	32.1 ± 0.2	**62.8 ± 0.5**
MSD	30.6 ± 2.3	3.4 ± 0.1	5.1 ± 0.1	4.5 ± 0.0	19.5 ± 0.2	20.7 ± 0.1	**47.9 ± 0.3**
MC	**89.4 ± 0.4**	36.5 ± 1.1	19.9 ± 0.2	24.3 ± 0.9	59.9 ± 0.9	66.1 ± 0.2	84.8 ± 0.2
MC$_{reco}$	**77.7 ± 0.5**	27.6 ± 0.4	17.9 ± 1.2	16.3 ± 0.1	48.3 ± 0.4	50.7 ± 1.0	70.1 ± 0.4

Recall@10 (mean ± std) of the all baselines on the LP-MusicCaps test splits.

7. Resources

Deezer Research website **https://research.deezer.com/** · Code for experiments **https://github.com/deezer/nlp4musa_melscribe**

7. References

[1] Oramas et al. 2024. Talking to Your Recs: Multimodal Embeddings For Recommendation and Retrieval.MuRS.
[2] Chaganty et al. 2023. Beyond single items: Exploring user pref- erences in item sets with the conversational playlist curation dataset. SIGIR.
[3] Wang et al. 2022. GPL: Generative pseudo labeling for unsupervised domain adaptation of dense retrieval. NAACL.
[4] Doh et al. 2023 Lp-musiccaps: Llm-based pseudo music captioning. ISMIR.