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

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

DEEZER

# NL interfaces for search and recommendations highly in demand

→ integrating both search and recommendation

→ (ideally) conversational

→ promising results with LLMs, but still many challenges

Existing approaches for music

Learn to embed user input / song description together with music representation (from audio or from embedded metadata) using contrastive learning

Multimodal learning
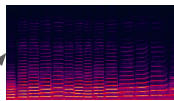
Dense retrieval

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

I'd love to hear some more upbeat tropical music with a similar steel drum melody in a higher register.
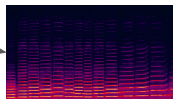
"Island music party" by "Steel drum band"

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

DEEZER

Still a high semantic
gap to bridge

What about using
song descriptors instead?

Research objective

Given user input / song description retrieve / recommend songs with a 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

Island Cruise Steel Band Presents...

Island Cruise Steel Band
20 tracks | 1 h 23 minutes | 09/08/2011 | 67 fans

Tropical Steel Drum Essentials

Island Steel Drum Players
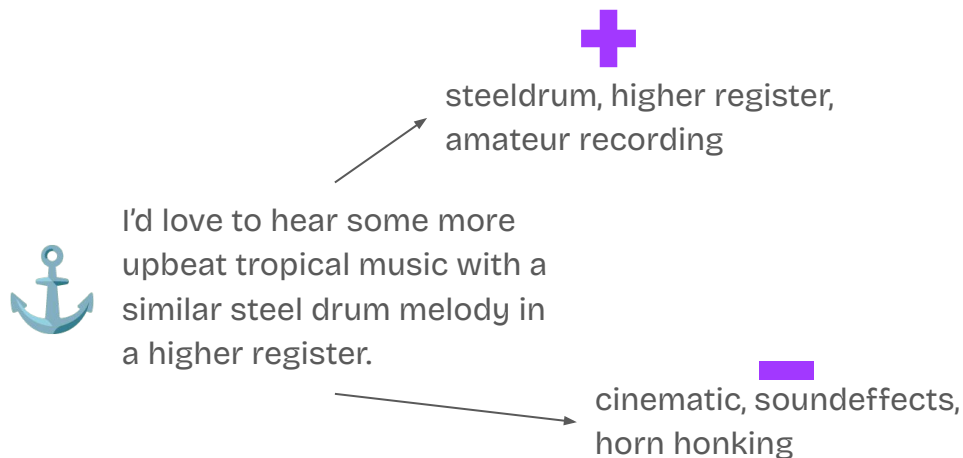30 tracks | 1 h 49 minutes | 01/07/2010 | 38 fans

DEEZER

Baselines

Given user input / song description retrieve / recommend songs with a matching descriptor set

sentence transformers

(dual encoders fine-tuned on sentence similarity)

text encoders from multimodal models

PLMs

sparse vectors

	Tf-Idf	CLAP$_{text}$	TTMR$_{text}$	BERT	all-MiniLM	msmarco-BERT
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
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
MC	$\mathbf{89.4 \pm 0.4}$	36.5 ± 1.1	19.9 ± 0.2	24.3 ± 0.9	59.9 ± 0.9	66.1 ± 0.2
MC$_{reco}$	$\mathbf{77.7 \pm 0.5}$	27.6 ± 0.4	17.9 ± 1.2	16.3 ± 0.1	48.3 ± 0.4	50.7 ± 1.0

DEEZER

Our proposal

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

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

DEEZER

Our proposal

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

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.
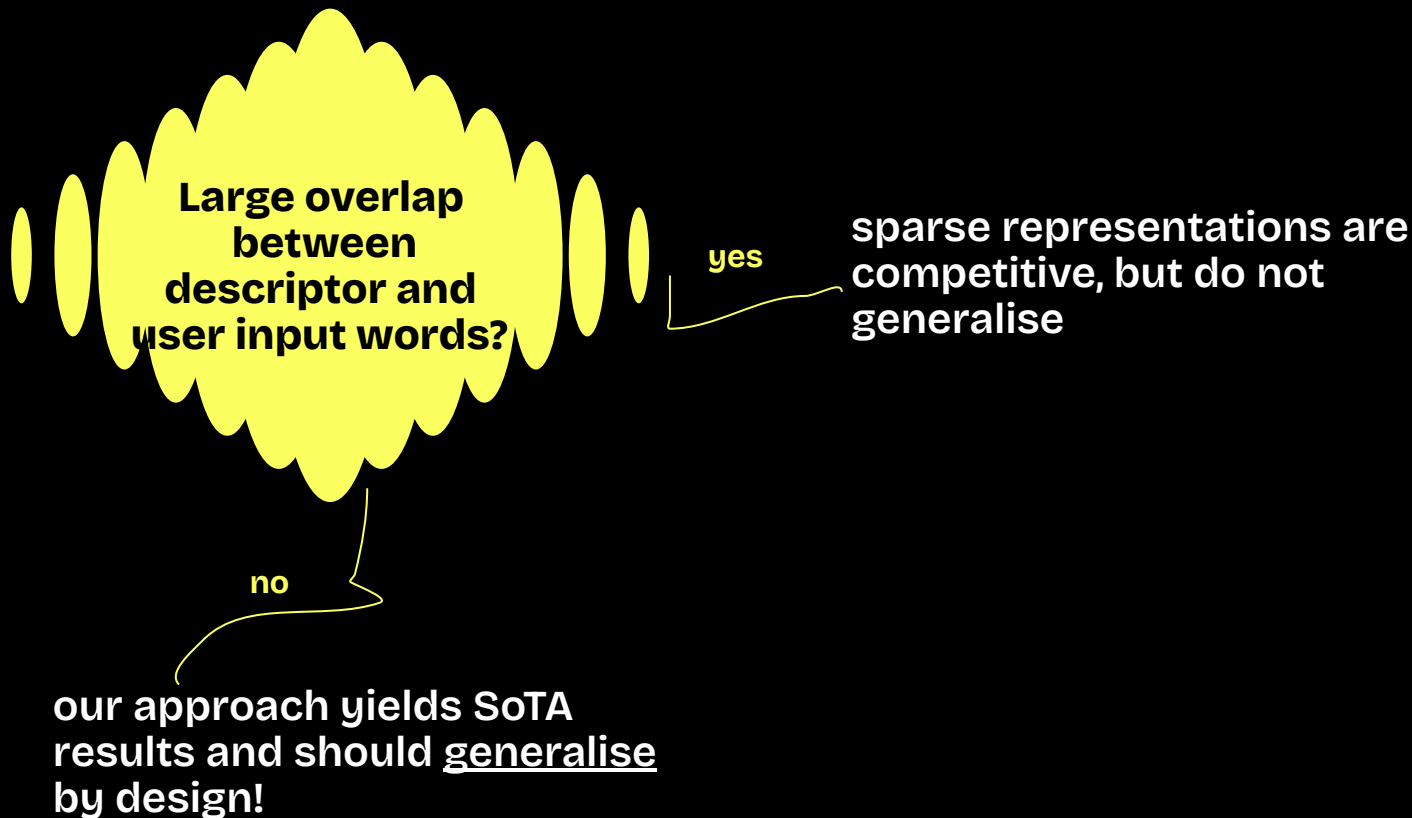
—

cinematic, soundeffects, horn honking

Generative Pseudo-labeling Method (GPL)

→ Hard negative mining with a music fine-tuned cross-encoder

→ Pseudo-labeling of training dataset (anchor, positive, negative)

→ Fine-tuning a bi-encoder to mimic the score margin between the positive and negative

DEEZER

Results

Large overlap between descriptor and user input words?

yes — sparse representations are competitive, but do not generalise

no — our approach yields SoTA results and should <u>generalise</u> by design!

DEEZER

Thank you

DEEZER