Semantically Meaningful Attributes from Co-Listen Embeddings for Playlist Exploration and Expansion

Ayush Patwari, Nicholas Kong, Jun Wang, Ullas Gargi
{patwaria,kongn,juwang,ullas}@google.com

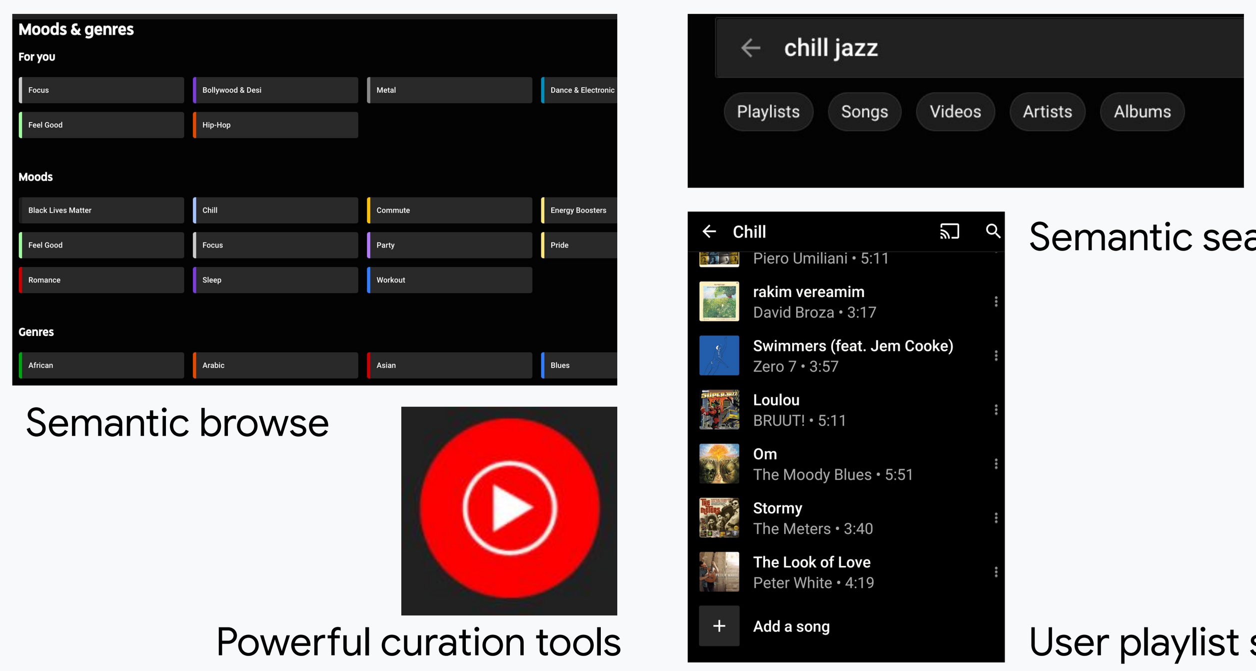
Michele Covell, Aren Jansen
{covell,arenjansen}@google.com

Music

Google

Introduction

- Deep Neural Nets can learn amazingly subtle similarities given enough training data. For example, representations of musical similarity given user co-listen behavior.
- The embedding representations generated by these networks are not immediately interpretable.
- There are practical applications in the music discovery space that require semantically meaningful annotations



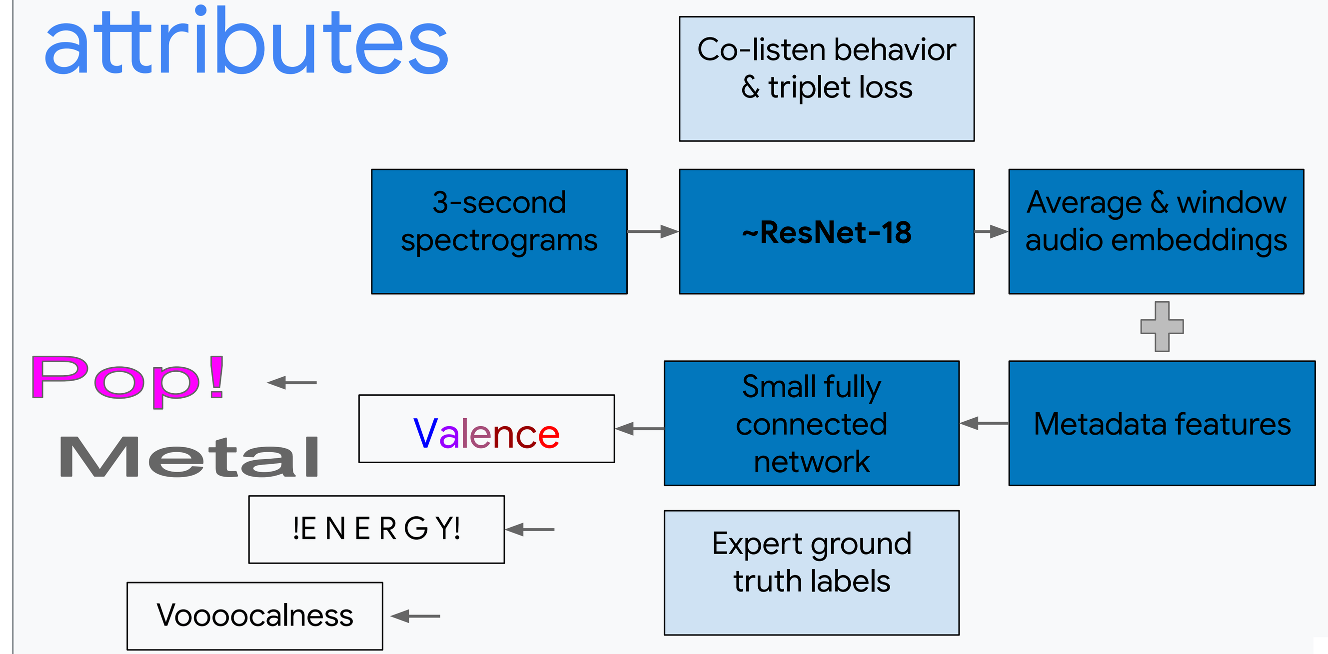
Semantic browse

Powerful curation tools

Semantic search

User playlist sequencing

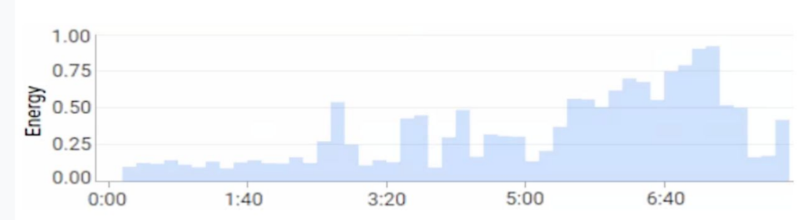
Co-Listen audio embeddings to semantic attributes



- Shallow network on top of audio-embeddings (+ other features)
- Ground-truth data from music experts
- 1k samples per genre
- ~10-20k samples for other models

Temporal Attribute-ness

- Inference on 10-second segments of audio using time-localized embeddings
- Model same as track-level
- This approach also yields temporal consistency attributes that are useful in and of themselves



Temporal Computed Energy for Stairway to Heaven.

$$\mathrm{E} = \max_{0 \le i < N-W} \frac{1}{W} \sum_{j=i}^{i+W-1} e_j \qquad (1)$$

where N is the the number of 10-second segments in a track, e_j is the raw energy estimate for the j^{th} segment, and W is the window size which also a function of N according to $W = \max\{3, \frac{N}{6}\}$.

Track level estimate from local estimates for Energy

Attribute quality

Attribute	Model type	Metric	Quality
Genres	Multi-label classifier • 16 classes	Human-expert labels	78% precision 84% recall
Valence	Regression • Output ∈ [0. 1]	Prediction < 0.33 from human-expert labels on a 4-point scale	78% accuracy
Vocalness	Binary classifier • Has vocals	Human-expert labels	97% precision 78% recall
Energy	Regression • Output ∈ [0. 1]	Prediction < 0.25 from human-expert labels on a 3-point scale	90% accuracy

Aggregating temporal attributes improved energy accuracy from 85% to 90%.

Attribute Embedding Generation

- Vector of the continuous logits of attribute models
- Renormalize using (regularized) square-root inverse of the pooled variance matrix
- Pooled variance matrix is estimated using a sampling of playlists treating each as a separate cluster sharing a single (pooled) variance matrix
- Post re-normalization, each playlist is a 0-mean, identity-variance distribution, allowing direct comparison between playlist distances

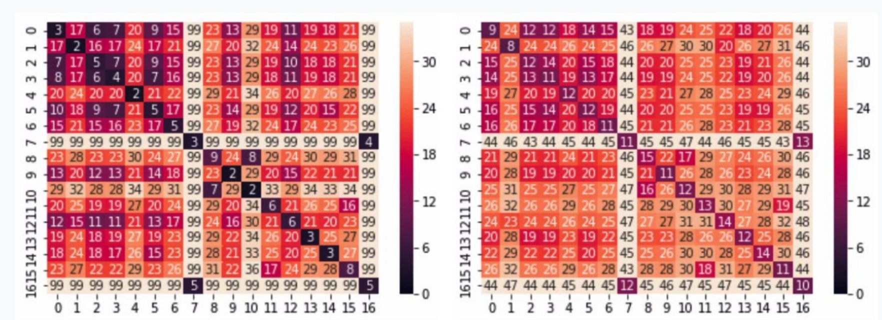
$$d_{i,k,j} = ||e_{i,k} - m_j||^2 \qquad (3)$$

where $e_{i,k}$ is the embedding-space coordinates for the i^{th} entry in the k^{th} playlist and m_j is the mean of embedding-space coordinates across all N_j entries in the j^{th} playlist: $m_j = \frac{1}{N_j} \sum_{i=0}^{N_j-1} e_{i,j}$.

Results

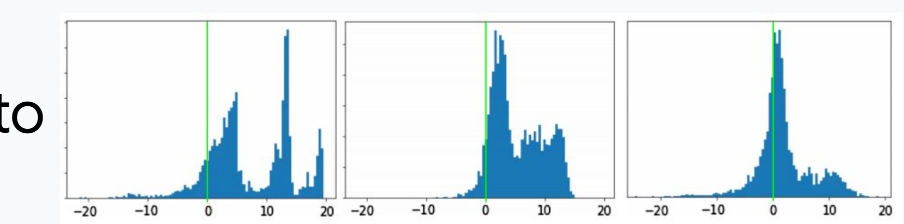
Playlist Separation

- Compared how well human-curated playlists are separated in attribute (1) and audio (2) embedding spaces
- $\Delta_{i,k} = \min_{j \neq k} d_{i,k,j} - d_{i,k,k}$: the smallest difference between each entry's distance to closest "other" mean and its own mean
- Playlists better separated in (1) than (2)



(a) using normalized attribute-embedding space (b) using normalized audio-embedding space

Average over each playlist k of di,k,j defined in Equation 3



(a) $\Delta_{i,k}^{att}$ using normalized attribute-embedding space (b) $\Delta_{i,k}^{aud}$ using normalized audio-embedding space (c) paired difference: $\Delta_{i,k}^{att} - \Delta_{i,k}^{aud}$ (matched indices)

Histograms of Δi,k using the two spaces and of their difference.

Playlist Expansion

- Generate suggestions based on attribute embedding distance for playlist expansion
- Humans rated the suggestions as either acceptable or good for thematic playlists. Suggestions found not as effective for non-thematic playlists e.g., decade-based.

Playlist	Rating			Total
	Good	Borderline	Bad	
Classical for Sleeping	36%	38%	26%	214
Classic Sunshine Soul	39%	35%	26%	101
Tranquil Spa Day	37%	63%	0%	27
Feeling Good in the 80's	22%	20%	58%	143
90's Rock Relaxation	11%	24%	65%	85

Music curator ratings on suggestions for playlist extension

Conclusions

- Temporal inference and temporal statistics on those time series perform better than inference on the temporally averaged embedding
- We demonstrate that the smaller embedding space induced by these semantic attributes separate thematic playlists better than the raw audio embedding as measured by inter- and intra-playlist distances
- Thematic playlists can also be described by recipes using a semantic attribute vocabulary and when these playlists were extended using those recipes, humans rated the suggestions as acceptable or good. Non thematic playlists such as decade-based did not.