Department of Computer Science

Hit Song Prediction: Leveraging Low- and High-level Audio Features

Eva Zangerle*, Ramona Huber*, Michael Vötter*, Yi-Hsuan Yang**

* Universität Innsbruck, Austria

** Academia Sinica, Taiwan

Goals and Take-Aways

Q1: How can we predict hit songs based on acoustic features extracted from the song's audio in a deep learning scenario?
Q2: Which role do individual features (groups of features) and the release year of a song play in this task?

- In a Wide-and-Deep neural network, we combine low- and high-level features in a regression task.
- Combining high- and low-level features improves results.
- Release year, mood and vocals are important features.
- Dataset is made available on Zenodo to foster further research on hit song prediction[1].

Data and Experiments

Million Song Dataset ∩ Billboard Hot 100

- Songs with release year information
- Essentia for feature extraction from audio

- Song is considered a hit if it is featured once in Hot 100
- Undersampling for balanced data; 6k hits and non-hits

Regression Experiments

- Predict highest rank in charts for test songs
- Five-fold cross validation
- MSE as loss function

Prediction Approach

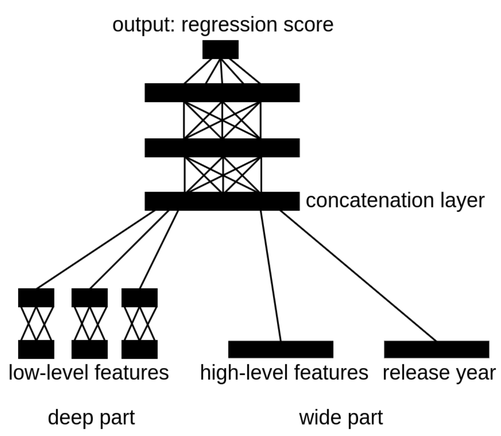
Acoustic Descriptors

40 basic low-level features (e.g., MFCCs, dissonance) (LL)
11 rhythm features (e.g., beats per minute or onset-rate)
13 tonal features (e.g., key or harmonic pitch class profiles)

Feature Categorization

Category	Features
Mood	acoustic, aggressive, electronic, happy, party, relaxed, sad; Hu and Downie's 5 clusters of mood
Genre	blues, classic, country, disco, hip-hop, jazz, metal, pop, reggae, rock
Voice	voice, instrumental, female voice, male voice
Rythm/ beat	bpm, beats count, bpm histogram, beats loudness, beats loudness band ratio, onset rate, danceability
Chords	strength, change rate, number rate, key, scale, harmonic pitch class profile, tuning strength and frequency

Regression via Wide-and-Deep Neural Network



Results

Highest Rank Prediction Task on Feature Sets

Features low-level	Features high-level	RMSE	MAE	Acc.
---	Year, voice, mood, genre	57.11	48.50	72.08%
LL, rhythm, chords	---	60.82	52.09	66.94%
LL, rhythm, chords	Year, voice, mood, genre	**55.45**	**43.84**	**75.04%**
LL, rhythm, chords	Year, genre	55.93	45.80	73.84%
LL, rhythm, chords	Year, mood	57.12	45.66	73.55%
LL, rhythm, chords	Year, voice	56.63	46.04	72.04%
LL, rhythm, chords	Genre	64.14	52.84	65.11%
LL, rhythm, chords	Mood	61.77	52.82	67.92%
LL, rhythm, chords	Voice	61.18	52.50	68.00%
LL, rhythm, chords	Year	57.51	46.35	72.29%
LL, rhythm, chords	Year, mood, voice	56.22	45.53	74.46%
LL, rhythm, chords	Year, mood, genre	57.35	45.38	73.63%
LL, rhythm, chords	Year, genre, voice	56.06	45.66	73.60%

Main Findings

- Combining low- and high-level features improves prediction performance.
- Wide-and-Deep networks are well-suited for this task.
- Release year information is important for temporally contextualizing a song to reflect musical trends and fashion.
- Vocal, mood and genre information also contribute to the performance, in line with previous research.

https://dbis.uibk.ac.at
eva.zangerle@uibk.ac.at
[1] https://doi.org/10.5281/zenodo.3258042