Etykietowanie muzyki (music tagging) za pomocą metod sztucznej inteligencji

Jakub Robaczewski

2022L

1 Wstęp

Celem mojej pracy inżynierskiej będzie napisanie aplikacji, która umożliwi etykietowanie muzyki tagami związanymi z gatunkiem muzycznym (np. rock, pop, rap), nastrojem (smutny, wesoły), tempem oraz innymi grupami.

2 Content based filtering

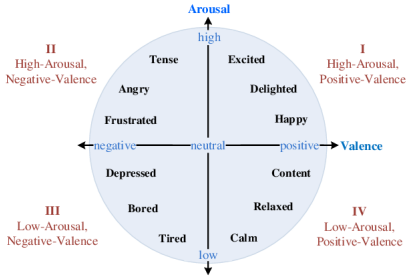
Tagowanie utworów muzycznych jest złożonym zagadnieniem, dlatego warto pochylić się nad tym jakie rodzaje klasyfikacji możemy zastosować w naszym problemie.

2.1 Gatunek

Gatunek jest najbardziej naturalnym sposobem podziału utworów, jednak największym problemem tego podziału jest to, że jest mocno subiektywny i jeden utwór może być zaklasyfikowany do wielu gatunków naraz.

2.2 Nastrój

Klasyfikowanie nastroju utworów jest jeszcze bardziej subiektywne niż gatunków. Najbardziej popularną metodą określenia tego jest wpisanie utworów w 2-wymiarową płaszczyznę, której osie oznaczają pobudzenie i wartościowość.



2.3 Charakterystyczne instrumenty

Zwykle problem tagowania instrumentów jest uproszczany do problemu klasyfikowania głównego instrumentu, ze względu na problem z zaszumieniem źródła, przez co trudne jest dokładne określenie wszystkich instrumentów.

2.4 Tempo

Oprócz tego, muzykę można klasyfikować za pomocą metryk, które można obliczyć na podstawie wycinka utworu. Jednym z nim jest tempo, które można obliczyć i następnie odpowiednio sklasyfikować jako tagi "szybko" i "wolno".

3 Algorytmy tagowania

W najnowszych zastosowaniach najskuteczniejsze są architektury uczenia nadzorowanego, ale wzrost skuteczności uczenia nienadzorowanego sprawia, że warto się zastanowić nad wykorzystaniem tych nowych sposobów.

Model	Dataset	ROC-AUC	PR-AUC
CLMR (ours)	MTAT	88.7 (**89.3**)	35.6 (**36.0**)
Musicnn [5]†	MTAT	89.0	34.9
SampleCNN [26]†	MTAT	88.6	34.4
CPC (ours)	MTAT	86.6 (88.0)	31.0 (33.0)
1D CNN [36]†	MTAT	85.6	29.6
Transformer [37]†§	MSD	**89.7**	**34.8**
Musicnn [5]†	MSD	88.0	28.7
SampleCNN [26]†	MSD	87.9	28.5
CLMR (ours)	MSD	85.7	25.0

Benchmark algorytmów tagujących[2]

3.1 Uczenie nadzorowane

Uczenie nadzorowane jest najpopularniejszym sposobem uczenia maszynowego, jego zaletą jest duża skuteczność, niestety jest to metoda wymagająca stosunkowo dużo próbek danych. Wśród architektur zasługujących na szczególną uwagę są MusiCNN (rozwinięcie idei FCN).

Model	Prepro-cessing	Input length	Front end	Back end	Training	Aggrega-tion
FCN	STFT	29.1s	2D CNN	.	song-level	.
VGG-ish / Short-chunk CNN	STFT	3.96s	2D CNN	Global max pooling	instance-level	Average
Harmonic CNN	STFT	5s	2D CNN	Global max pooling	instance-level	Average
MusiCNN	STFT	3s	2D CNN	1D CNN	instance-level	Average
Sample-level CNN	.	3s	1D CNN	1D CNN	instance-level	Average
CRNN	STFT	29.1s	2D CNN	RNN	song-level	.
Music tagging trans-former	STFT	5s-30s	2D CNN	Transformer	instance-level	Average

Wykaz algorytmów do tagowania[1]

3.1.1 Fully Convolutional Networks (FCNs)

Architektura bardzo podobne do CNN wykorzystywanych przy obróbce obrazów. Wykorzystuje 4 warstwy konwulucyjne wraz z normalizacją batchową i filtrami 3x3.

3.1.2 VGG-ish / Short-chunk CNNs

Bardzo podobna architektura do FCN, ale operująca na fragmentach utworów, a nie na całych utworach.

3.1.3 Harmonic CNNs

Architektura posiadająca warstwy konwulucyjne podobne do Short-chunk CNNs, ale używająca inne wejścia. Poza tym wykorzystuje trenowalne maski, które dają modelu więcej elastyczności.

3.1.4 MusiCNN

Rozwiązanie, które zamienia filtry 3x3 wykorzystywane w FCN na zbiór filtrów ręcznie wykonanych, by przechwytywać konkretne charakterystyki.

3.1.5 Music tagging transformer

Architektura typu transformer wykorzystująca atencję, na początku wyłapuje akustykę utworu, a na końcu podsumowuje sekwencję. Do wyłapywania akustyki architektura używa CNN, a podsumowuje je za pomocą klasycznego transformera.

3.2 Uczenie nienadzorowane

3.2.1 CLMR[2]

Najlepsza skuteczność wśród algorytmów uczenia nienadzorowanego osiąga model CLMR, który wykorzystuje tagi tylko w pre-treningu. Uzyskuje to dzięki skomplikowanemu procesowi argumentacji danych (redukcja Gaina do [-6, 0], dodanie opóźnionego pogłosu, filtry częstotliwościowe etc.). Sam proces działa na zasadzie autokodera, który umiejscawia dane w przestrzeni i oblicza prawdopodobieństwo wystąpienia taga jako odległość jednej danej od drugiej.

4 Zbiory danych

Do trenowania i późniejszej walidacji modeli wykorzystam 2 najpopularniejsze zbiory danych zawierające utwory:

- Million Song Dataset[7] - największy publicznie dostępny zbiór utworów muzycznych (280 GB) posegregowanych według artystów i tagów.
- MagnaTagATune[8] - zbiór ponad 25 tyś 29-sekundowych wycinków z utworów różnych gatunków oznaczonych przez 188 tagów.

5 Ewaluacja

Ocena działania modeli jest najbardziej kluczowym elementem algorytmu uczenia maszynowego, dlatego warto podzielić predykcje modelu na kilka kategorii, w przypadku tagowaniu oceniamy obecność tagu jako pozytywny, a jego brak jako negatywny:

- True positives (TP) - Wynik poprawnie przewidziany jako pozytywny
- False positives (FP) - Wynik przewidziany jako pozytywny, w rzeczywistości negatywny.
- False negatives (FN) - Wynik przewidziany jako negatywny, w rzeczywistości pozytywny.
- True negatives (TN) - Wynik poprawnie przewidziany jako negatywny.

6 Aplikacja

Aplikacja do prezentacji rezultatów pracy będzie napisana w języku Python i udostępniona jako aplikacja webowa. Za jej pomocą użytkownik będzie mógł wgrać swoje utwory i przygotować playlistę wykorzystując określone tagi. Prototyp aplikacji jest dostępny w repozytorium GitHub.[9]

Literatura

[1] https://arxiv.org/ftp/arxiv/papers/2111/2111.11636.pdf

[2] https://arxiv.org/pdf/2103.09410.pdf

[3] https://dida.do/blog/beat-tracking-with-deep-neural-networks

[4] http://infolab.stanford.edu/ũllman/mmds/ch11.pdf

[5] https://developers.google.com/machine-learning/crash-course/embeddings/obtaining-embeddings

[6] https://tlkh.github.io/text-emotion-classification/

[7] http://millionsongdataset.com/

[8] https://mirg.city.ac.uk/codeapps/the-magnatagatune-dataset

[9] https://github.com/Robak132/MIR