E-019

# Spotify が算出する音楽特徴量をニューラルネットワークの学習に使用する際の妥当性の検証

# Validation of using music features calculated by Spotify for training neural networks

瀬田 翔太郎 [1)]　石井 一夫 [1) 2)]
Shotaro Seta　Kazuo Ishii

## 1. はじめに

音楽や音声の分析にはいくつかの手法が存在するが,その中の一つに Spotify のデータベースから Spotify audio features(以下「Spotify 特徴量」と称する)を取得し,これを用いる手法がある.Spotify 特徴量とは Spotify にアップロードされた楽曲及び音声を説明する量である.この方法を採用する利点として,特徴量を抽出する手間を省くことができ,音楽的な意味を持たせた数値を取り扱うことができるという点が挙げられる.一方で,分析に利用できる特徴量が限定されるため,音楽の情報(波形の離散データ)を利用する場合に比べて情報量が少ないとみなすこともできる.

近年,機械学習を用いて音声や音楽の生成を行う場合には,学習対象の波形データをメルスペクトログラム等の画像に一度変換し,それを学習データとする手法が知られている.この手法は画像処理分野と密接な関係があり,ViT[1]や Stable Diffusion[2]等の画像処理に関係するアーキテクチャの発展とともに,黎明期に比べて高品質なデータを出力できるようになりつつある.本研究では,回帰分析において前述した 2 つの手法のどちらが優れているかを示すべく,Spotify のデータベースから得られる 30 秒のプレビューデータ(mp3 データ)と Spotify 特徴量をそれぞれ用いてニューラルネットワークによる回帰モデルを訓練した場合について,モデルの性能や訓練に必要なコスト等の面から評価する.mp3 データを用いたモデルには,波形データを入力として多クラス分類を行うことが出来る PaSST[3]を特徴抽出器として用いる.モデルの目的変数は Spotify 特徴量の一つである Popularity とし,説明変数としてその他の Spotify 特徴量を用いる.これらの変数は文献[4]中の,楽曲のトレンドを Spotify の特徴量によって説明できるという主張のもとに決定した.これは同文献中で取り扱うタスクとは異なるが,本実験で取り扱うタスクはあくまで実験的なものであり,本実験は入力データとして波形データと Spotify 特徴量のどちらが優れているかという点に注目することを留意されたい.

## 2. 準備

### 2.1 データセットの構築

Spotify が提供する Web API をコールし,135 万曲分の Spotify 特徴量とプレビューデータを取得した.広範なデータを持つデータセットを構築するために,Spotify API によって音楽ジャンルを取得した後,ジャンルごとに最大 1000 人のアーティスト情報を取得し,それぞれのアーティストがアップロードしている楽曲情報を取得した.その後,重複しているデータを削除した.また,説明変数として Spotify 特徴量から次の 10 個の数値を抽出した.

- danceability
- energy
- instrumentalness
- liveness
- loudness
- mode
- tempo
- valence

次に,学習用データとテスト用データに 8:2 の割合で分割した.学習用データに対しては図 1 に示すようにアンダーサンプリングを行った.

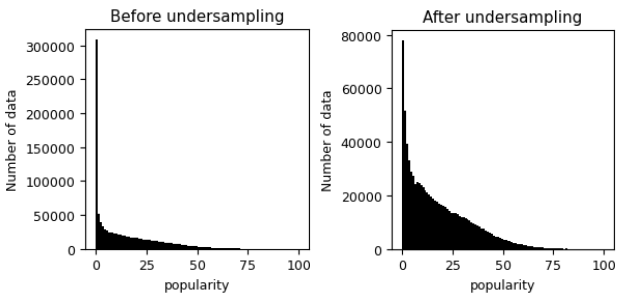


図 1　学習用データに対するアンダーサンプリング

### 2.2 モデルの訓練

mp3 を入力としたモデル(以下,単純に「PaSST」と称する)と,Spotify 特徴量を入力としたモデル(以下「Dence」と称する)を図 2 に示す.

mp3 を入力としたモデルの訓練には GPU サーバに搭載された 24 個の NVIDIA A100 を用いて DDP[9]によって訓練した.また,Spotify 特徴量を入力としたモデルは単一の NVIDIA A100 を用いて訓練した.

### 2.3 DDP(Distributed Data Parallel)[9]

DDP は複数の計算ユニットを用いてモデルの訓練を行うときに用いられる手法の一つであり,訓練のプロセスを各ユニットに分散することによって訓練の速度を向上させることが出来る.DDP では,訓練にデータセットを計算ユニットの数だけ分割しておき,それぞれのユニットに保持されているモデルで勾配計算を行った後,各ユニット間で勾配を集計し,同じ勾配値でそれぞれのモデルの更新を行うことで訓練を行う.

1)　公立諏訪東京理科大学　Suwa University of Science
2)　久留米大学　Kurume University

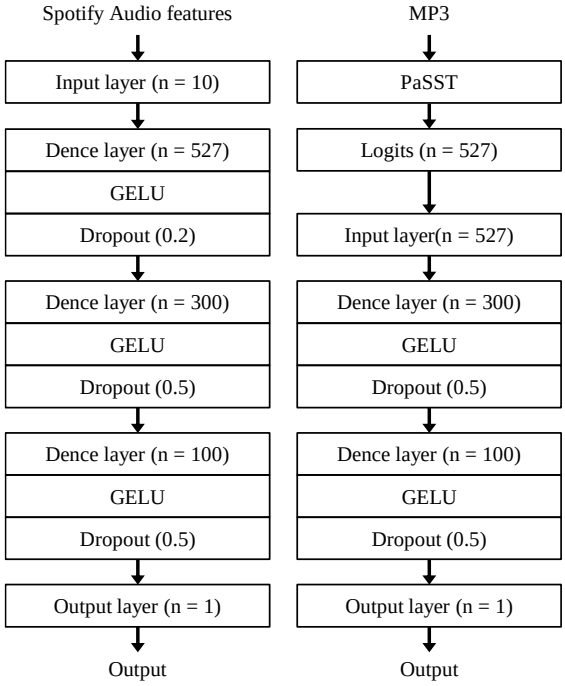


図 2　本実験で扱うモデル

## 3. 実験

### 3.1 ハイパーパラメータ

PaSST はついてはバッチサイズ 8,学習率を 1e-5 とした.バッチサイズについては GPU メモリの都合上これ以上増やすことが不可能であり,学習率のみチューニングを行った.また,Dence についてはパラメータチューニングによってバッチサイズ 512,学習率を 1e-3 とした.

### 3.2 結果

全てのデータを用いてモデルを訓練した結果を表 1 に示す.なお,目的変数 Popularity は 0.5 ~ -0.5 にスケーリングし,Dence に用いる説明変数に対しては標準化を行った.

表 1　全てのデータを用いた場合のモデル性能

| model | RMSE | MAE | hours |
|---|---|---|---|
| PaSST | **0.134** | **0.107** | 2.350 |
| Dence | 0.148 | 0.119 | **0.003** |

hours の項目は 1epoch の訓練に要した時間を示す.

また,音楽ジャンルごとにデータを分割した場合でそれぞれモデルを訓練した結果を一部抜粋したものを表 2 に示す.

表 2 より,データセットを少数に分割した場合においても PaSST のモデル性能が高いことがわかる.一方で,特定の音楽ジャンルでは,Dence が PaSST を上回る性能であることがわかる.

## 4. 考察

本実験の結果では,データセットを分割しない場合,両者のモデルの性能の差はさほど見られなかった.このことから,広範なデータを取り扱うタスクの場合,実行時間が短いことによるパラメータチューニングの容易さ等の観点から,

表 2　音楽ジャンルごとのモデル性能

| genre | PaSST | | Dence | |
|---|---|---|---|---|
| | rank | MAE | rank | MAE |
| all data | 44 | **0.107** | 29 | 0.119 |
| grindcore | 1 | **0.034** | 5 | 0.085 |
| honky-tonk | 2 | 0.040 | 1 | **0.031** |
| ambient | 23 | **0.071** | 23 | 0.113 |
| club | 24 | **0.072** | 28 | 0.118 |
| edm | 46 | **0.112** | 43 | 0.153 |
| hip-hop | 47 | **0.116** | 46 | 0.167 |

rank は音楽ジャンルについての性能の順位を示す.
完全なデータは次のリンク先に記載している.
https://docs.google.com/spreadsheets/d/1thk840TrmcSA7aINceXxA3gl98rGV72-T04yeH8Sinc/edit?usp=sharing

Spotify 特徴量を使用する学習は十分実用的であると考えられる.一方,目的変数を Spotify 特徴量で説明できないタスクの場合や,高いモデル性能が必要なケースでは,PaSST のように波形情報から特徴を抽出できるモデルを用いる手法が望ましいと考えられる.

## 5. おわりに

本実験では特定のタスクや広範なデータかつ Spotify 特徴量が目的変数を説明できる場合,Spotify 特徴量を用いる学習方法が実用的であることがわかった.一方,PaSST[3]のような画像処理を活用するモデル構造は画像処理分野の発展の影響を受けるため,今後の発展次第で本稿で示した結果を凌駕する性能を持ったモデルが発表される可能性があり,将来的には Spotify 特徴量を扱う利点はその簡便さだけになるだろう.

回帰タスクでは文献[5]で提案されるオーバーサンプリングによる Data Augmentation を行う手法があるが,これを回帰タスクの音声信号に適用する方法を著者調べでは確認できなかった.これについては研究の余地があるため,今後の発展に期待したい.

### 参考文献

[1] Dosovitskiy, Alexey, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.
[2] Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.
[3] Koutini, Khaled, et al. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.
[4] Middlebrook, Kai, and Kian Sheik. Song hit prediction: Predicting billboard hits using spotify data. *arXiv preprint arXiv:1908.08609*, 2019.
[5] Branco, Paula, Luís Torgo, and Rita P. Ribeiro. SMOGN: a pre-processing approach for imbalanced regression. First international workshop on learning with imbalanced domains: Theory and applications. PMLR, 2017.
[6] Nijkamp, Rutger. Prediction of product success: explaining song popularity by audio features from Spotify data. BS thesis. University of Twente, 2018.
[7] Vaswani, Ashish, et al. Attention is all you need. Advances in neural information processing systems 30, 2017.
[8] Akiba, Takuya, et al. Optuna: A next-generation hyperparameter optimization framework. Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining, 2019.
[9] Li, Shen, et al. Pytorch distributed: Experiences on accelerating data parallel training. *arXiv preprint arXiv:2006.15704*, 2020.