

畳み込みニューラルネットワークを用いた Hardcore techno のサブジャンル分類

小林 拓司^{1,a)} 丸山 一貴^{1,b)}

概要：楽曲には、曲名、アーティスト情報、ジャンルなどの属性が存在する。中でもジャンルには、サブジャンルというものが存在する。従来の研究では、楽曲のジャンルの分類を行っているが、サブジャンルまでの分類を行っていない。本研究では、楽曲データのメル周波数ケプストラム係数を求め、画像に変換し、畳み込みニューラルネットワークを使用して Hardcore techno の楽曲が属するサブジャンルを判定する。本研究を行うことによって、サブジャンルの属性が付いていない楽曲を分類することや、好みの近いサブジャンルを発見する手がかりにすることができる。4つのサブジャンルについて、確信度に基づく分類を行ったところ、平均 88.25%の精度であった。しかし、楽曲中に特徴が少ないサブジャンルの認識率が低い結果となった。

1. はじめに

楽曲の属性は、曲名、アーティスト名、ジャンルなどが挙げられる。中でも、楽曲のジャンルは、個人の嗜好に大きく関わるものであり、好みの楽曲を探すための重要な情報である。また、ジャンルにはサブジャンルというものが存在する。サブジャンルとは、ジャンルの中から更に細分化されたジャンルである。この情報を利用して、好みの楽曲をより正確に探索する事が可能である。しかし、従来の研究における楽曲のジャンル分類では、サブジャンルまでの分類は行っていない。

本研究では、畳み込みニューラルネットワーク (以下、「CNN」という。)を用いて Hardcore techno の楽曲が属するサブジャンルを判定するを目的とする。本研究を行うことによって、サブジャンルの属性が付いていない楽曲を分類することや、好みの近いサブジャンルを発見する手がかりにすることができる。

2. 関連研究

これまでも楽曲のジャンル分類に関する研究は多く存在している。新妻ら [1] は、RandomForest を用いて、ジャンルごとにパラメータを変えることなく、1つのチューニング済みの設定で 9 種類のジャンルの分類を行っている。

分類精度は、84%と報告している。

George ら [2] は、メル周波数ケプストラム係数 (以下、「MFCC」という。), 音色, Beats Per Minute を楽曲の特徴として扱い、10 種類のジャンルをガウス分布, ガウス混合分布モデル及び k 近傍法を用いてそれぞれで分類している。分類精度としては、61%と報告している。

Chun ら [3] は、MFCC を楽曲の特徴として使用し、LSTM を用いて分類を行っている。この研究では、最初に似た特徴を持つジャンルを 2 グループに分けてから分類し、最終的に 10 種類のジャンルに分類している。全体の分類精度としては 52.975%である。また、畳み込み層が一つである CNN での精度は 46.87%なので、高い精度が得られていると報告している。しかし、これらの研究ではサブジャンルまでの分類は行っておらず本研究と異なる。本研究では、Chun ら [3] が用いた MFCC を特徴として分類する手法を実装した。

Yandre ら [4] は、3つの音楽のデータセットに対して、Local Binary Patterns, リズムパターン, スペクトラム, リズムヒストグラムを特徴として扱い、CNN と SVM を用いてジャンルの分類を行っている。楽曲の 30 秒から 1 分 30 秒までの 60 秒を 1 つのデータとして分類を行っている。また、楽曲のジャンルの分類において、CNN を使用することの優位性を示している。本研究では、サブジャンルを扱っており、ジャンル間よりもサブジャンル間の方が類似度が高くなり分類が難しくなる。よって、本研究とは異なる。また、Yandre ら [4] が用いた手法である、楽曲の分割方法及び CNN での学習方法を応用し実装した。

¹ 明星大学情報学部情報学科
Department of Information Science, School of Information Science, Meisei University

a) 15j5056@stu.meisei-u.ac.jp

b) kazutaka.maruyama@meisei-u.ac.jp

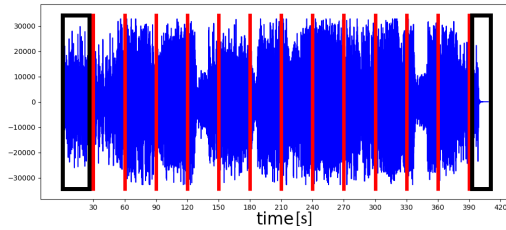


図 1 楽曲の分割

3. 提案手法

提案手法は、全体の流れとして、(1) データの抽出、(2) 学習用モデルの作成、(3) 機械学習、の 3 工程に分類される。第 3.1 節、第 3.2 節では学習させるまでの前処理について、第 3.3 節で CNN を用いた学習について述べる。

3.1 データの抽出

本研究で分類するサブジャンルは、Hardstyle, Frenchcore, Crossbreed, Oldschool の 4 つのサブジャンルを選定した。これらは、楽曲の特徴がかなり異なるサブジャンルと、似た特徴を持つサブジャンルが混在するように選定した。今回の場合、Hardstyle, Frenchcore, Crossbreed が前者に当てはまり、Frenchcore, Oldschool が後者に当てはまる。それぞれのサブジャンルの楽曲を 100 曲ずつ使用する。楽曲は、Lololyrics[5], SoundCloud[6] で選曲した。楽曲の分割には ffmpeg を使用した。

本研究では、図 1 のように、楽曲を 30 秒間隔に分割し、こうして作成されたものを 1 つのデータとして使用した。図 1 は縦軸が音圧、横軸が時間を表しており、赤線が分割の区切り目である。楽曲を分割することによって、学習データを増やすことが可能になる。また、CNN は高解像度の処理に適用してなく、一枚あたりの解像度を落とすためにも分割は必要である。図 1 の黒枠で囲まれた最初の 30 秒間と 30 秒に満たない部分（曲の最後）はデータから取り除いた。楽曲の最初と最後には無音部分があり、その楽曲の特徴を示す特徴量が乏しいため、削除した。楽曲の量子化ビット数は 16bit、サンプリング周波数は 44.1 kHz、ビットレートは 2822kbps である。

3.2 モデルの作成

librosa の MFCC を用いて、学習用の画像を作成した。作成された画像は図 2 であり、縦軸が Hz、横軸が時間である。表示されている色は、デシベルを表しており、色が白いほど音が大きいことを示している。こうして生成された画像は合計 3187 枚である。各サブジャンルごとの画像の枚数を表 1 に示す。サブジャンルごとの画像枚数の割合が 3% 以上異なる場合もあるが、Yandre ら [4] の結果から、サブジャンルごとの画像枚数の割合が 12% 以上異なってい

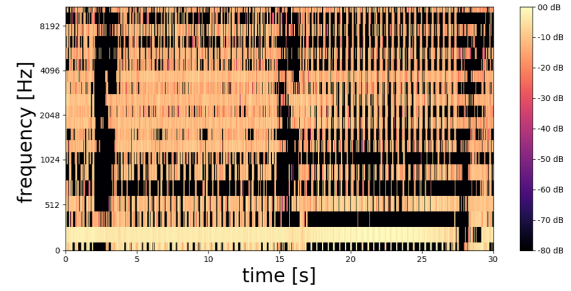


図 2 学習用の画像

表 1 作成枚数

サブジャンル	枚数
Hardstyle	747
Frenchcore	857
Crossbreed	826
Oldschool	757
合計	3187

表 2 使用した CNN の層

層	シェープ
input	(42, 102, 3)
B1_Conv1(Conv2D)	(42, 102, 64)
B1_Conv2(Conv2D)	(42, 102, 64)
B1_pool(MaxPooling2D)	(21, 51, 64)
dropout1(Dropout(0.25))	(21, 51, 64)
B2_Conv1(Conv2D)	(21, 51, 64)
B2_Conv2(Conv2D)	(21, 51, 64)
B2_pool(MaxPooling2D)	(11, 25, 64)
dropout2(Dropout(0.25))	(11, 25, 64)
flatten1(Flatten)	(17600)
Dense1	(4400)
Dense2	(1100)
Dense3	(550)
dropout3(Dropout(0.5))	(550)
Dense3	(4)

ても影響がないと推測できるため、画像枚数の正規化は行わないこととした。

3.3 機械学習

本研究では、Keras(Ver.2.2.2) と TensorFlow(Ver.1.11.0) を用いて CNN で学習する。CNN の層は、Benjamin ら [7], Luiz ら [8] の手法を参考にし、本研究に適合するように作成した。使用した層を表 2 に示す。使用した活性化関数は、Dense2 は softmax, それ以外は relu である。入力画像のサイズは、 42×102 ピクセル、batchsize は 32 で行った。epoch 数の選定は、epoch 数 200 回で学習した中で最も精度が高かった 50 回を採用した。epoch 数 200 回までの精度と誤差関数のグラフを、図 3 に示す。図 3 から、acc が上がるにつれ、loss が低下しているの、学習が正常に行えたことが確認できる。

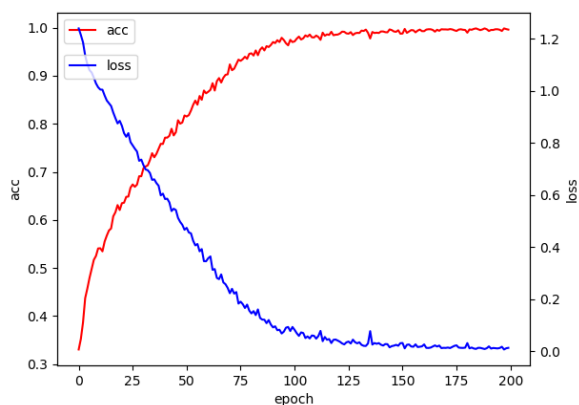


図 3 学習精度と誤差関数

表 3 使用環境

OS	Windows7
CPU	Intel CPU Core i7 i7-2600 @ 3.4GHz
RAM	DDR3-1600MHz 16GB
GPU	GeForce GTX 1050 Ti

表 4 分類結果

分類方法	精度
最大の確信度	88.25%
多数決法	84.25%
確信度の合算	88.75%

表 5 最大の確信度による分類結果

	(1)	(2)	(3)	(4)
(1)Hardstyle	100	0	0	0
(2)Frenchcore	1	96	0	4
(3)Crossbreed	6	5	86	3
(4)Oldschool	4	23	5	71

表 6 多数決法による分類結果

	(1)	(2)	(3)	(4)
(1)Hardstyle	100	0	0	0
(2)Frenchcore	1	98	0	1
(3)Crossbreed	3	7	85	5
(4)Oldschool	8	28	10	54

4. 評価実験

実験に使用したコンピュータの構成を表 3 に示す。各サブジャンル 100 曲を用いてテストを行った。本研究では、最大の確信度での評価、多数決法での評価、確信度の合算での評価、以上 3 つの方法で精度の算出を行った。それぞれの精度を表 4 に示す。また、分類方法別のサブジャンルごとの分類の結果を表 5、表 6、表 7 に示す。縦軸が実際のジャンルで、横軸が分類した結果である。

5. 考察

表 4 から、どの分類方法でも 84% を超える結果となった。分類結果の報告、新妻ら [1]、George ら [2]、Chun ら [3] の

表 7 確信度の合算による分類結果

	(1)	(2)	(3)	(4)
(1)Hardstyle	99	0	0	1
(2)Frenchcore	1	97	0	2
(3)Crossbreed	4	5	89	2
(4)Oldschool	3	25	2	70

分類結果がそれぞれ、84%、61%、46.87%であり、それらの結果を上回っているのが正確に分類されていると言える。

表 5、表 6、表 7 から、Hardstyle の分類は、どの精度の算出方法でも正確に行われていることがわかる。また、Frenchcore も同じことがいえる。この 2 つのサブジャンルが正確に分類されていたのは、楽曲の全体にそのサブジャンルの特徴となる部分が存在していたからであると考えられる。

Crossbreed に関しては、前 2 つのサブジャンルより精度は低いが正確に分類されていた。これは、前の 2 つのサブジャンルより音量が低い部分が存在し、その部分が間違っ

て分類されたため、精度が下がっていると考えられる。Oldschool に関しては、他のサブジャンルに比べると、どの算出方法でも精度は低く、Frenchcore として分類されているケースが多くなっている。これは、いずれのサブジャンルとも、シンプルな曲調を持つ楽曲が多く誤分類されていたと考えられる。逆に、Frenchcore が Oldschool と分類されていないのは、先に述べた通り、楽曲の全体的にそのサブジャンルの特徴となる部分が存在していたからである。

また、3 つの分類法の中で多数決法を用いた分類結果が最も精度が低かった。原因として考えられるのは、Oldschool は楽曲の中にそのサブジャンルと判定することができる部分が、他のサブジャンルより少ないことが挙げられる。

これらの誤分類を軽減するためには以下の手法が考えられる。

- 楽曲の音量が低い部分をデータから取り除く
- 分割統治法を用いる
- CNN の調整

音量が低い部分には、サブジャンルを特定できる特徴が存在しない楽曲が多い。したがって、音量が低い部分を取り除くことによって、楽曲の学習時にノイズとなる情報を除去することができる。

分割統治法とは、1 つの問題を段階に分けて、最終的に全ての問題を解決する手法であり、Chun ら [3] もこれを利用している。例えば本研究で対象とした楽曲については、最初に、4 つのサブジャンルに対して、曲調が単純なサブジャンルと複雑なサブジャンルの 2 種類に分けるとする。次に、分けた 2 種類の中でサブジャンルを分類する。最後にそれぞれの結果を統合し、最終的なサブジャンルを決定する。このような手順で分類を行えば、少ないクラスの中から最終的なサブジャンルを決定するので、ノイズとなる情報が少なくなり、正確に分類されるようになると考えら

れる。これは、分類させるサブジャンル数が増えたときにも有効な手法だと考える。

特徴量の抽出、画像への変換するときのパラメータの調整をすることによって、精度が向上すると考える。

CNN を調整することによっても精度の向上が考えられる。例えば、層の数の調整や dropout 率などの最適化が挙げられる。

今回は、4 つのサブジャンルの分類しか行っていない。Hardcore techno において、サブジャンルはこれ以外にも存在するので、更にサブジャンルを増やして分類したときに、分類する数が増え、似たサブジャンルも増加するので誤分類する可能性が高くなる。それを解消するためにも、epoch 数や層の数など CNN の調整をする必要があると考える。

6. まとめ

本論文では、楽曲のサブジャンルを判定するために、楽曲を MFCC を用いて画像に変換し、CNN を用いて学習させて分類する手法を提案した。本研究の手法で分類した結果では、最大の確信度での評価で、平均 88.25% の精度が得られたのでこの手法の有効性を確認した。しかし、一部の楽曲に対しての分類精度が低く、改善の余地がみられる。今後の課題としては、第 5 章で述べた 3 つの手法

- 楽曲の音量が低い部分をデータから取り除く
- 分割統治法を用いる
- CNN の調整

を用いて分類の精度の向上を目指す。また、本研究では、サブジャンルを 4 つしか扱っていない。Hardcore techno において、サブジャンル数はこれ以外に存在するので、更にサブジャンルを増やし、分類をしていく予定である。

参考文献

- [1] 新妻雅弘, 斎藤博昭, “Random Forest を用いた音楽ジャンル分類”, 情報処理学会論文誌, Vol. 50, No. 12, pp.2910–2922, 2009.
- [2] George Tzanetakis, Perry Cook, “Musical Genre Classification of Audio Signals”, IEEE Transactions on Speech and Audio Processing, Vol.10, No.5, pp.293–302, July, 2002.
- [3] Chun Pui Tang, Ka Long Chui, Ying Kin Yu, Zhiliang Zeng, Kin Hong Wong, “Music Genre classification using a hierarchical Long Short Term Memory (LSTM) model”, ICMR18, pp.11–14, June, 2018.
- [4] Yandre M.G. Costa, Luiz S. Oliveira, Carlos N. Silla, “An evaluation of Convolutional Neural Networks for music classification using spectrograms”, Applied Soft Computing, Vol.52, pp.28–38, March, 2017.
- [5] atomicoz, “Lololyrics”, <<https://www.lololyrics.com>> (参照 2018/12/20).
- [6] Alexander Ljung, Eric Wahlforss, “SoundCloud”, <<https://soundcloud.com>> (参照 2018/12/20).
- [7] Benjamin Murauer, Günther Specht, “Detecting Music Genre Using Extreme Gradient Boosting”, WWW 2018,

pp.23–27, April, 2018.

- [8] Luiz G. Hafemann, Luiz S. Oliveira, Paulo Cavalin, “Forest Species Recognition using DeepConvolutional Neural Networks”, ICPR, pp.1103–1107, 2014.