深層畳み込みニューラルネットワークによる音声波形を用いた音楽アーティスト分類

情報科学科　清水 雄治　　指導教員：小林 邦和，鈴木 拓央

1　はじめに

畳み込みニューラルネットワークは, 主に画像認識に応用される順伝播型ネットワークであり [1]，これを多層化したものが深層畳み込みニューラルネットワーク (DCNN) である．近年の DCNN の開発は，生データから階層的特徴を学習する End-to-End 学習を可能にし，入力データ処理の手間を抑えている．一方で，音楽分類タスクでは音楽信号の生波形は時間-周波数表現に変換され，システムに入力されることが一般的であった．

そこで本研究では End-toEnd 学習が可能で，音楽自動タグ付けにおいて最高精度を誇る SampleCNN[2] を用いて，音楽アーティスト分類を行う．また，転移学習を行い，計算時間の削減と精度の向上を試みる．

2　提案手法

SampleCNN は音楽自動タグ付けを行うためのモデルであるが，音楽の特徴を抽出し，識別を行うという点では音楽アーティスト分類も同じであると考える．そこで，SampleCNN に以下の変更を加え，音楽アーティスト分類タスクへの適用を行う．

1. マルチレベル特徴集約 (前の 3 層を連結して，FC 層に入力)[2] の有無
2. Million Song Dataset[2] で学習させた SampleCNN を利用し，音楽アーティスト分類のデータセットで再学習を行う．
3. 活性化関数の変更
4. Basic, SE, Res-2, ReSE-2 ブロック [2] の利用
5. 入力セグメント数の変更

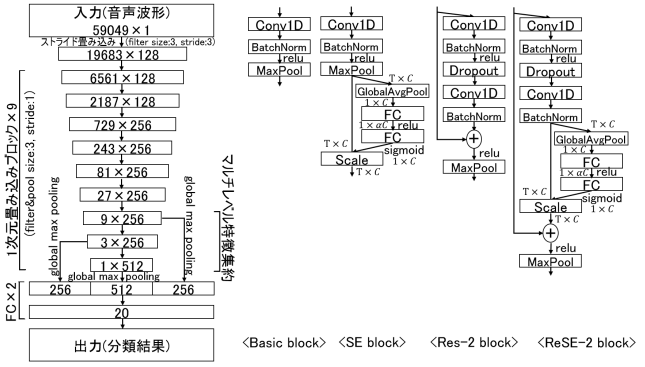

図 1　SampleCNN のネットワークと各ブロックの構造 [3]

3　計算機シミュレーション

Nesterov 運動量 0.9, ミニバッチサイズ 10 の SGD を使用して訓練を行った．初期学習率は 0.01 で，検証損失が停滞した時に学習率を 5 分の 1 に変更することを繰り返した. その後，評価を行った．

3.1　データセット

データセットは artist20[3] を利用する．これは 20 人のアーティスト，合計 1413 曲を含む (MP3,16000Hz). 収録時間に違いがあるため，入力セグメント数に応じて，0s からの切り出しを行う．入力セグメント数に満たないオーディオは取り除いた．

3.2　結果・考察

入力セグメント数以外を変更した時の精度を表 1 に，入力セグメント数変更時の精度を表 2 に示す．

表 1　入力セグメント数以外を変更した時の精度

ブロック	マルチレベル	転移学習	計算時間 (s)	活性化関数	入力セグメント数	正解率 (%)
ReSE-2				sigmoid	10	40.0
ReSE-2	○		1218	sigmoid	10	**42.7**
ReSE-2	○	○	2859	sigmoid	10	36.7
ReSE-2	○			softmax	10	39.6
Basic	○			sigmoid	10	36.1
SE	○			sigmoid	10	35.7
Res-2	○			sigmoid	10	**42.7**

表 2　入力セグメント数変更時の精度 (Res-2, マルチレベル, sigmoid)

入力セグメント数 (秒数)	訓練データ数	正解率 (%)
10 (37s)	692	42.7
20 (74s)	684	47.7
30 (111s)	670	51.7
40 (148s)	622	**53.6**
50 (185s)	538	49.9

転移学習を行ったが，計算時間の削減と精度の向上にはつながらなかった．音楽自動タグ付けと音楽アーティスト分類で抽出される特徴が似ていないことによると考えられる．ReSE-2 と Res-2 ブロックが最高精度を出したことから，SE ブロックは音楽アーティスト分類において機能していないと考えられる．Res-2 ブロックによって過学習を防ぎ，層を深くすることで精度の向上が図れた．また，入力セグメント数を増やすことでデータ数が増え，精度が向上したと考えられる．40 セグメントと比べて 50 セグメントの場合に精度が低い原因は，訓練データが急激に減少したため，データの多様性が減り，十分にアーティストの特徴を学習できなくなったためだと考えられる．

4　おわりに

本研究では End-to-End 学習が可能な SampleCNN を用いて，音楽アーティスト分類を行い，53.6 %の精度を確認した．

今後は End-to-End 学習が可能なモデルを用いた音楽アーティスト分類の精度向上を行う．そのために音楽アーティスト分類に特化したモデルの開発と，データ数の多い音楽アーティスト分類データセットの作成が必要だと考える．

参考文献

[1] 岡谷貴之: 「機械学習プロフェッショナルシリーズ　深層学習」, 講談社, pp.79-110 (2015)

[2] Taejun Kim, Jongpil Lee, Juhan Nam: “SAMPLE-LEVEL CNN ARCHITECTURES FOR MUSIC AUTO-TAGGING USING RAW WAVEFORMS”, 2018 IEEE ICASSP, pp.336-370 (2018)

[3] Daniel P. W. Ellis: “Classifying Music Audio with Timbral and Chroma Features”, Proc. Int. Conf. on Music Information Retrieval ISMIR-07 (2007)