Vol.2018-MUS-121 No.22
2018/11/21

Hand-crafted特徴量に対する多入力CNNベースの楽曲タグ予測モデルの評価

穴繁 陽太[†1,a)]　千葉 一永[†1]

概要：楽曲タグ予測手法には，人間の手によって特徴量抽出プロセスが設計される hand-crafted 特徴量を用いた手法と，ニューラルネットワークを用い人間の介入を減らした end-to-end と呼ばれる手法の 2 つが存在する．後者には，生の音源波形やメル周波数スペクトログラムに対してニューラルネットワークを用いることで特徴変換を自動化できるというメリットがあるが，高精度を達成するために，ニューラルネットワークのパラメータ数が膨大になる傾向がある．そこで本研究では，end-to-end においてニューラルネットワークのみで一貫していた特徴変換の一部を音特有の特徴抽出アルゴリズムで置換し，ニューラルネットワーク自身の特徴変換を削減するという発想から，インプットデータとして hand-crafted 特徴量を使用した多入力型の CNN に基づく楽曲タグ予測モデルを提案し，評価を行う．CNN による特徴変換部の convolutional block の種類や層数，また，畳み込み層とマックスプーリング層のフィルタサイズなどが異なる計 9 種類のモデルを作成し，MagnaTagATune データセットに対して比較を行った．その結果，convolutional block に ReSE block を用いたモデルが既存研究の最大値と同等の精度を約半分のパラメータ数で達成した．

キーワード：楽曲タグ予測，多入力 convolutional neural networks，hand-crafted 特徴量

Automatic tagging using multi-input CNN for hand-crafted features

Yota Anashige[†1,a)]　Kazuhisa Chiba[†1]

Abstract: There are two methods for predicting tags of music. The first is a method using hand-crafted features that the feature extraction process is designed by the human. The second is a method called end-to-end that reduces engineering effort by the neural network. In the latter case, there is an advantage that the feature extraction process can be automated by using the neural network for the raw waveform or mel-spectrogram. However, in order to achieve high performance, the number of parameters of the neural network tends to become enormous. Therefore, in this research, from the idea of replacing part of the consistent feature transformation in the neural network at end-to-end with sound-specific feature extraction algorithms, we propose the tag prediction model based on multi-input CNN using hand-crafted features as input data and evaluate it. We create 9 models that the types and numbers of convolutional block, the filter size of the convolutional layer and max-pooling layer are different and compare it using MagnaTagATune dataset. As a result, the model using the ReSE block achieves the performance equivalent to the maximum value of previous works with about half the number of parameters.

Keywords: music auto-tagging, multi-input convolutional neural networks, hand-crafted features

1. 序論

コンピュータサイエンスにおいて，人間が持つ知覚機能を機械に付加する研究は盛んに行われている．例えば，顔

[†1] 現在，電気通信大学
Presently with The University of Electro-Communications (UEC), Chohu, Tokyo 182–8585, Japan
[a)] a1832001@edu.cc.uec.ac.jp

ⓒ 2018 Information Processing Society of Japan

認証や機械翻訳技術への応用が可能なため，機械に画像中の物体を認識させたり，言語を理解させる研究が行われている．本研究ではそのような分野の中で，音をトピックに扱う machine hearing[1] と呼ばれるテーマに着目する．機械が人間のような聴覚機能を獲得できれば，音声を文章に変換したり，雑多な音の中から人間の声だけを抽出したりと様々な技術への応用が考えられる．

Machine hearing において，人間のような知覚機能を機械に付加するために，生の音源波形を特徴量と呼ばれる音固有の表現に変換することが必要となる．様々な手法が提案されているが，それらは 2 つに大別される．

(1) 特徴量への変換プロセスが人間の手によって設計される hand-crafted 特徴量を用いた手法が存在する．Hand-crafted 特徴量をロジスティック回帰やサポートベクターマシンのような機械学習アルゴリズムのインプットデータとして使用することで，モデルを作成することが可能だ．ここでモデルとは，生の音源波形を特徴量に変換し，予測値の出力までを行う一連のアルゴリズムのことを指す．

(2) 一方近年，ニューラルネットワークの発展により，特徴抽出プロセスへの人間の介入を減らした end-to-end と呼ばれる手法が誕生し，盛んに研究されている．end-to-end では，生の音源波形から特徴量を自動で抽出し，ニューラルネットワーク内部で一貫したモデルを作成することが可能だ．しかし，高精度なモデルを作成するために，ニューラルネットワークのパラメータ数が膨大になり，多くの学習データ，計算量を必要とする傾向がある．

そこで本研究では，hand-crafted 特徴量を用いた手法と end-to-end 手法の 2 つを組み合わせた新たな手法を提案し，人間の特徴抽出プロセスへの介入が少ない，かつ，ニューラルネットワークのパラメータ数を抑えた音知覚モデルの作成を目指す．ここで作成したモデルの評価を行うため，音楽情報検索（MIR: Music Information Retrieval）の主要なタスクの 1 つであるオートタギングを用いる．オートタギングとは，楽曲に割り振られたジャンル，ムード，楽器といったタグを高精度で予測するタスクである．本手法の要点は，生の音源波形から音色やリズムといった音特有の特徴量を手動で抽出し，ニューラルネットワークのインプットデータとして使用することで，end-to-end においてニューラルネットワークのみで一貫していた特徴抽出プロセスを 2 段階に切り分けた点にある．

以下，第 2 章に本研究で提案するハイブリット手法について説明する．第 3 章では，実験に使用したデータセットや実験手順について述べる．第 4 章では，オートタギングタスクにおける作成したモデルの数値実験結果を示す．最後に第 5 章で，結論を述べる．

2. 提案手法

本研究で使用するモデルは，hand-crafted 特徴量の抽出，convolutional block による特徴変換，Fully Connected（FC） layer による予測値の出力の 3 段階から構成される．音特有の特徴抽出アルゴリズムから得られた 13 種類の hand-crafted 特徴量各々に対し convolutional block による特徴変換を行い，得られた 13 個の新たな特徴量を global max pooling を用いて集約し，2 層の FC layer のインプットデータとして使用することで楽曲タグ予測モデルを作成する．以下に詳細を示す．

2.1 Hand-crafted 特徴量

機械学習アルゴリズムのインプットデータとして使用する hand-crafted 特徴量を抽出するための第一段階として，生の音源波形を楽曲特有の性質を反映した特徴量に変換する必要があり，それらを抽出するアルゴリズムとして様々なものが提案されている．例えば，参考文献 [2], [3] ではそれぞれ，（timbral texture，rhythmic content，pitch content）の 3 種類，（dynamics，rhythmic，spectral，harmony）の 4 種類といった様々な組み合わせの特徴量が抽出される．アルゴリズムの種類によるが，これらの特徴量は〜50 [ms] 程度の範囲から抽出されることが多い．この短い範囲は analysis window と呼ばれる．Analysis window から抽出された特徴量は，サポートベクターマシンのような機械学習アルゴリズムのインプットデータには適切ではなく，それらに適切な処理を施すことも hand-crafted 特徴量を抽出するための重要なプロセスである．参考文献 [2] では，analysis window により抽出された特徴量から 1 [s] 程度の範囲で平均値や標準偏差といった統計量を計算し，最終的なインプットデータとして使用している．

本手法では，analysis window により得られた特徴量をそのまま CNN のインプットデータとして用いることで，analysis window から得られた特徴量をより適切な特徴量に変換するプロセスを CNN 内部で行う．つまり本手法において hand-crafted 特徴量とは，analysis window から得られた特徴量のことを指す．本研究では試験的に，hand-crafted 特徴量として参考文献 [4] で定義されている Time Domain Physical, Frequency Domain Physical, Cepstral Domain Perceptual の中で，MIRtoolbox[5] で実装されている Root Mean Square（RMS），zerocross，pulseclarity，beatspectrum，flux，centroid，rolloff，flatness，entropy，skewness，kurtosis，chromagram，13 次元 Mel Frequency Cepstral Coefficients（MFCC）の計 13 種類を選択する．

2.2 Convolutional block

参考文献 [6], [7], [8], [9] では，畳み込み層とマックス

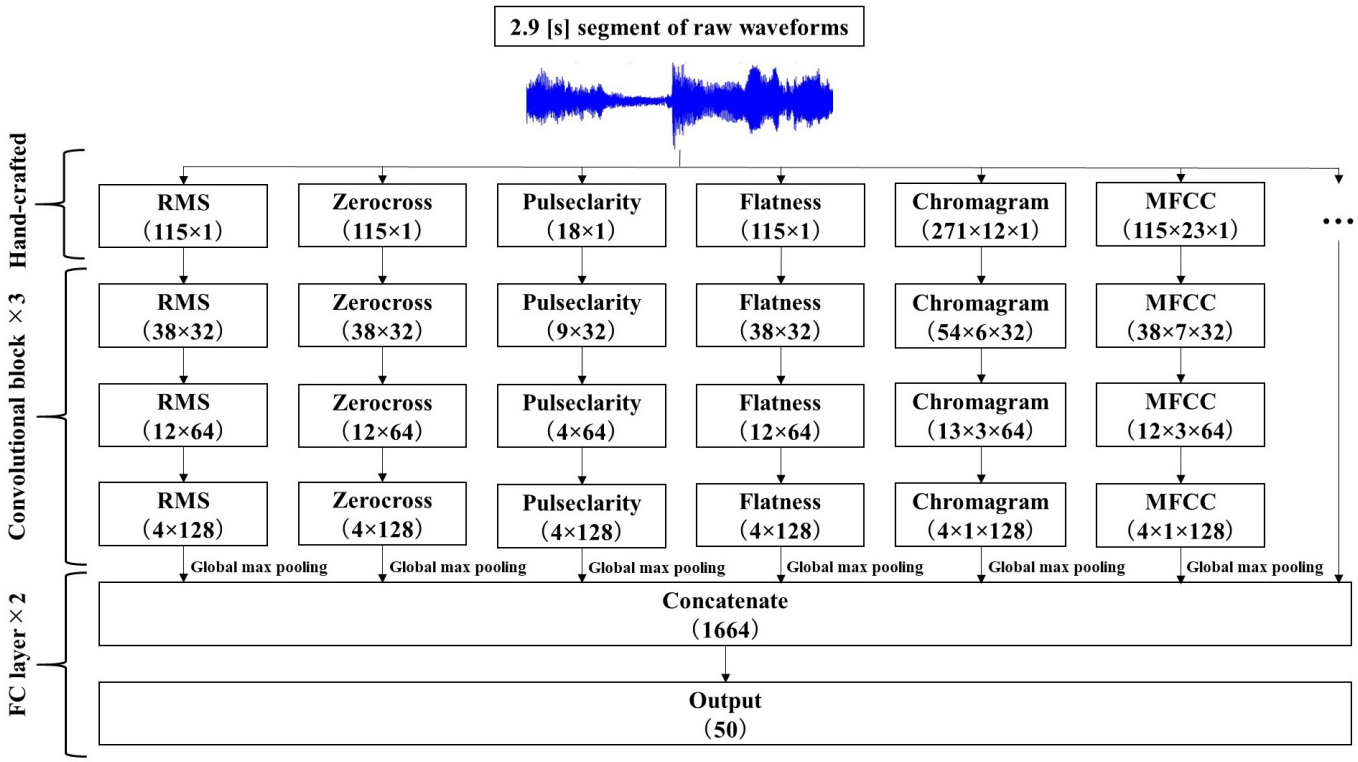


図 1 アーキテクチャ a の概略図

プーリング層からなる convolutional block が提案され，生の波形に対しては 1 次元の畳み込み層と 1 次元のマックスプーリング層が，メル周波数スペクトログラムに対しては 2 次元の畳み込み層と 2 次元のマックスプーリング層が使用される．この convolutional block を basic block と呼ぶ．また参考文献 [10] では，画像認識分野で大きな成果を挙げている *ResNets*[11] と *SENets*[12] を応用した Res，SE，ReSE blocks を提案し，精度の上昇に成功している．そこで本研究では，basic block だけでなく，参考文献 [10] で特に高い精度が得られている SE，ReSE blocks の 3 種類を hand-crafted 特徴量に対して使用したモデルを作成し，比較を行う．ここで，本研究で使用する hand-crafted 特徴量には RMS や zerocross のような 1 次元特徴量と，chromagram や MFCC のような 2 次元特徴量の 2 種類が存在する．そこで，1 次元特徴量には，生の波形と同様に 1 次元の畳み込み層と 1 次元のマックスプーリング層を使用し，2 次元特徴量には，メル周波数スペクトログラムと同様に 2 次元の畳み込み層と 2 次元のマックスプーリング層を使用する．

2.3 モデルアーキテクチャ

hand-crafted 特徴量が抽出される際，window length と hop length の値はモデルの性能に影響を与えることが想定される．また，畳み込み層とマックスプーリング層のフィルタサイズ，convolutional block の層数も同様にモデルの性能に影響を与える重要なパラメータである．しかし，それらのパラメータが異なるモデルをいくつも作成し，実験を行うことは，多くの計算時間が必要となる，また，本研究で使用するデータセットのみに過剰に適合したモデルになるといった 2 点から適切でないと考える．そこで，3 つの convolutional blocks と以下に説明するアーキテクチャ a，b，c を組み合わせた計 9 種類のモデルを作成し，比較を行う．

a: window length と hop length に MIRtoolbox のデフォルト値を使用したアーキテクチャである．hand-crafted 特徴量ごとに window length と hop length の値は異なるため，抽出される特徴量のサイズは異なる．そこで，全ての特徴量において層数は 3 に固定し，convolutional block の最終層から得られる特徴量の次元数が（4 × フィルタ数）となるように，マックスプーリング層のフィルタサイズを 1 から 5 の範囲で調節する．

b: 抽出される全ての hand-crafted 特徴量のサイズが 3 の倍数である 243 になるように window length，hop length を調節したアーキテクチャである．マックスプーリング層のフィルタ数を 3，層数を 5 とすることで，convolutional block の最終層から得られる特徴量の次元数が（1 × フィルタ数）となるように調節する．

c: 抽出される全ての hand-crafted 特徴量のサイズが 2 の倍数である 128 になるように window length，hop length を調節したアーキテクチャである．マックスプーリング層のフィルタ数を 2，層数を 7 とすることで，convolutional block の最終層から得られる特徴量

表 1　MTAT における 9 種類のモデルの結果

	アーキテクチャ		
Block	a	b	c
Basic	0.9076	0.9084	0.9068
SE	0.9083	0.9091	0.9093
ReSE	0.9110	0.9104	0.9096

の次元数が（1 × フィルタ数）となるように調節する．ここでアーキテクチャ b，c において，window length と hop length の比が MIRtoolbox のデフォルト値と同じになるように調節を行い，全てのアーキテクチャにおいて，畳み込み層のフィルタサイズはマックスプーリング層のフィルタサイズと同じ値を使用する．また 13 種類の特徴量の中で，pulseclarity は window length，hop length のデフォルト値が 5 [s]，0.5 [s] と大きな値が設定されている．そこで例外として，アーキテクチャ a では window length，hop length の値を 1 [s]，0.1 [s] とし，アーキテクチャ b，c では抽出される特徴量のサイズが 27，32 となるように調節，かつ，層数を 3，5 と減らすことで対応する．最後にアーキテクチャ a，b，c の順にニューラルネットワークのパラメータ数，計算コストが少ないことを明記しておく．アーキテクチャ a の概略図を図 1 に示す．

3. 実験

3.1 データセット

本研究では，MagnaTagATune（MTAT）[13] データセットを用いる．MTAT は長さ 29 [s] 程度のファイル 25863 曲からなり，各曲に対してジャンル，楽器，ムードといった 188 個のタグが割り振られているデータセットである．参考文献 [6], [7], [8], [9], [10] と同じように，トレーニング，バリデーション，テストデータは 12:1:3 に分割し，最もよく使用される 50 個のタグを使用する．

3.2 評価

MTAT は各曲に割り振られているタグの総数が 50 個でそれぞれ異なる．そこで評価指標として，不均衡なデータセットに対して頑健性を持つ Area Under an receiver operating characteristic Curve（AUC）を用いる．50 個のタグ各々に対し AUC を計算し，50 個の値の平均値を最終的な評価指標として使用する．また本研究では，各曲を 10 個のセグメントに分割し，モデルのインプットデータとして使用する．そこで，10 個のセグメントから得られる予測値の平均値を使用することで各曲ごとに評価を行う．

3.3 各種パラメータの詳細

FC layer の output 層の活性化関数として，sigmoid 関数を用いる．また損失関数には，binary crossentropy を用いる．全てのモデルは，nesterov momentum が 0.9 の SGD を用い，バッチサイズ 23 でトレーニングを行う．初期学習率は 0.01 に設定し，20 エポックが経過するごとに 1/5 に減らす．FC layer の最終層の手前でドロップアウト率が 0.5 のドロップアウト層を使用する．全ての convolutional block において，活性化関数として ReLU 関数を使用し，バッチ正則化，weight decay が 0.001 の L2 正則化を適用する．Hand-crafted 特徴量はインプットデータとして使用する前，スケールを整えるために正規化を行う．

表 2　MTAT における既存研究との比較

モデル	AUC	パラメータ数
Temporal pooling and multiscale[14]	0.8610	
Transfer learning[15]	0.8800	
2D CNN + mel-spec[6]	0.8815	
Deep bag-of-features[16]	0.8880	
Timble CNN[17]	0.8930	191k
2D CNN + mel-spec[8]	0.8940	22M
Bag of multi-scaled features[18]	0.8980	
PersistentCNN[19]	0.9013	
Multi-level + multi-scale[9]	0.9017	
SampleCNN-ReSE + multi-level[10]	0.9113	27M
a-ReSE	0.9110	13M

4. 結果

4.1 MTAT での作成したモデルの比較

表 1 に，MTAT における 9 種類のモデルの AUC を示す．3 つのアーキテクチャ全てで ReSE，SE，basic の順に高い精度が得られた．また，basic の中で b，SE の中で c，ReSE の中で a において最も高い AUC が得られた．それぞれの block とアーキテクチャの組み合わせには相性が存在することが分かる．さらに，層数が 3 と少なく，パラメータ数の最も少ないアーキテクチャ a に ReSE block を用いたモデルが最も高い 0.9110 の AUC を達成した．

4.2 MTAT での既存研究との比較

表 2 に，本研究で最大性能が得られた a-ReSE モデルと既存研究との比較を示す．a-ReSE モデルは，既存研究で最も高い AUC を示すモデルと同程度の値を約半分のパラメータ数で達成した．ReSE は basic と比べ，高い精度が得られる反面パラメータ数が多く，層数が深くなるにつれモデル全体のパラメータ数が膨大になる傾向がある．本手法で作成したアーキテクチャ a は，3 層と少ない ReSE block で十分に高性能な特徴量を生成することができた．その結果，精度を保ったままパラメータ数を抑えることに成功したといえる．

5. 結論

本研究では，楽曲タグ予測において大別していた hand-cradted 特徴量を用いた手法とニューラルネットワークを用いることで人間の介入を減らした end-to-end 手法の 2 つから着想を得た，hand-crafted 特徴量に対する多入力 CNN に

ⓒ 2018 Information Processing Society of Japan

基づく楽曲タグ予測モデルを提案した．MagnaTagATune データセットに対する評価実験の結果，CNN による特徴変換部に ReSE block を 3 層用いたモデルが最大精度を示し，約半分のパラメータ数で既存研究の最大値と同等の精度を達成した．これらの結果から，ニューラルネットワークで一貫していた特徴変換プロセスを音特有の特徴抽出アルゴリズムで置換することで，ニューラルネットワーク自身のパラメータ数を削減し，高精度を達成できる可能性を示した．

謝辞 GPU サーバを御提供頂いた電気通信大学 人工知能先端研究センター，研究室内の計算機環境構築に御尽力頂いた澤原 雅隆氏に感謝する．

参考文献

[1] Richard F. Lyon, "Machine hearing: An emerging field," IEEE signal processing magazine, vol. 27, no. 5, pp. 131-139, 2010.
[2] G. Tzanetakis and P. cook, "Musical genre classification of audio signals," IEEE transactions on speech and audio processing, vol. 10, no. 5, pp. 293-302, 2002.
[3] Babu Kaji Baniya, Joonwhoan Lee and Ze-Nian Li, "Audio feature reduction and analysis for automatic music genre classification," IEEE international conference on Systems, Man, and Cybernetics (SMC), 2014.
[4] Francesc Alías, Joan Claudi Socoró Carrié and Xavier Sevillano, "A review of physical and perceptual feature extraction techniques for speech, music and environmental sounds," Applied Sciences, vol. 6, no. 5, pp. 143, 2016.
[5] Olivier Lartillot and Petri Toiviainen, "A matlab toolbox for musical feature extraction from audio," International conference on digital audio effects, 2007.
[6] Sander Dieleman and Benjamin Schrauwen, "End-to-end learning for music audio," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2014.
[7] Jongpil Lee, Jiyoung Park, Keunhyoung Luke Kim and Juhan Nam, "Sample-level deep convolutional neural networks for music auto-tagging using raw waveforms," Sound and Music Computing Conference (SMC), 2017.
[8] Keunwoo Choi, George Fazekas and Mark Sandler, "Automatic tagging using deep convolutional neural networks," International Society of Music Information Retrieval (ISMIR), 2016.
[9] Jongpil Lee and Juhan Nam, "Multi-level and multi-scale feature aggregation using pre-trained convolutional neural networks for music auto-tagging," IEEE signal processing letters, vol. 24, no. 8, pp. 1208-1212, 2017.
[10] Taejun Kim, Jongpil Lee and Juhan Nam, "Sample-level CNN architectures for music auto-tagging using raw waveforms," IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2018.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren and Jian Sun, "Deep residual learning for image recognition," IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[12] Jie Hu, Li Shen and Gang Sun, "Squeeze-and-excitation networks," arXiv preprint arXiv: 1709.01507, 2017.
[13] Edith Law, Kris West, Michael Mandel, Mert Bay and J. Stephen Downie, "Evaluation of algorithms using games: the case of music annotation," International Society of Music Information Retrieval (ISMIR), 2009.
[14] Philippe Hamel, Simon Lemieux, Yoshua Bengio and Douglas Eck, "Temporal pooling and multiscale learning for automatic annotation and ranking of music audio," International Society of Music Information Retrieval (ISMIR), 2011.
[15] Aäron van den Oord, Sander Dieleman and Benjamin Schrauwen, "Transfer learning by supervised pre-training for audio-based music classification," International Society of Music Information Retrieval (ISMIR), 2014.
[16] Juhan Nam, Jorge Herrera and Kyogu Lee, "A deep bag-of-features model for music auto-tagging," arXiv preprint arXiv: 1508.04999, 2015.
[17] Jordi Pons, Olga Slizovskaia, Rong Gong and Emilia Gómez and Xavier Serra, "Timble analysis of music audio signals with convolutional neural networks," arXiv preprint arXiv: 1703.06697, 2017.
[18] Sander Dieleman and Benjamin Schrauwen, "Multiscale approachs to music audio feature learning," International Society of Music Information Retrieval (ISMIR), 2013.
[19] Jen-Yu Liu, Shyh-Kang and Yi-Hsuan Yang, "Applying topological persistence in convolutional neural network for music audio signals," arXiv preprint arXiv: 1608.07373, 2016.

ⓒ 2018 Information Processing Society of Japan