情報処理学会論文誌 Vol.61 No.3 718–725 (Mar. 2020)

推薦論文

多声音楽の演奏楽器を認識するための畳み込みニューラルネットワーク

佐藤 佑樹[1,†1] 武田 敦志[1,a)]

受付日 2018年12月8日, 採録日 2019年11月29日

概要：音楽データを自動的に解析してその演奏楽器を分類できれば，それらの音楽データを自動的に分類して検索するサービスを開発することが可能となる．そこで，本論文では，畳み込みニューラルネットワークを用いて演奏音データに含まれる演奏楽器を分類する手法を提案する．近年，畳み込みニューラルネットワークを用いた演奏楽器の分類手法が研究されてきたが，従来手法には過学習を回避することが難しいという問題があった．そこで，本論文の提案手法では，複数の計算経路を持つことで高い汎化性能を発揮する畳み込みニューラルネットワーク ResNet を用いて演奏楽器の分類を行う．また，汎化性能改善手法である Signal Augmentation を導入することにより，畳み込みニューラルネットワークの分類性能の向上を目指す．本論文では，演奏楽器分類タスクのためのデータセットである IRMAS を用いた実験を行い，実験結果より提案手法が従来手法よりも高い精度で演奏楽器を分類できることを示す．提案手法は楽器分類タスクにおいて高い分類性能（F 値：0.708）を達成しており，これは現時点（2018 年 12 月）における世界最高性能（state-of-the-art）である．

キーワード：ニューラルネットワーク，深層学習，楽曲分類

Convolutional Neural Networks for Musical Instrument Recognition from Polyphonic Music

Yuki Sato[1,†1] Atsushi Takeda[1,a)]

Received: December 8, 2018, Accepted: November 29, 2019

Abstract: We will be able to develop search engines of music data if there is a useful method which classifies musical instruments played in the music data automatically. In this paper, we propose a classification method which indicates musical instruments played in audio data by using convolutional neural networks. In recent years, some researchers have studied about musical instrument classification with convolutional neural networks. The existing methods, however, have problem in avoiding overfitting. Therefore, the proposed method uses ResNets which achieve good regularization performance because ResNets contain many processing paths. In addition, in order to improve classification performance, we adapt a regularization technique "Signal Augmentation" to residual blocks in ResNets. In this paper, we show experiment results in musical instrument classification tasks of IRMAS dataset. According to the results, the proposed method achieves better performance than existing methods. Especially, F-measure of the proposed methods in IRMAS classification tasks is 0.708, which is state-of-the-art in 2018.

Keywords: neural network, deep learning, music classification

1. はじめに

近年，膨大な数のデジタルメディアが作成され，それら

[1] 東北学院大学教養学部情報科学科
Department of Information Science, Tohoku Gakuin University, Sendai, Miyagi 981–3193, Japan
[†1] 現在，テクノ・マインド株式会社
Presently with Techno Mind Corporation
[a)] takeda@cs.tohoku-gakuin.ac.jp

本論文の内容は 2018 年 1 月の支部研究会にて報告され，支部長により情報処理学会論文誌ジャーナルへの掲載が推薦された論文である．

の特徴や属性をもとに検索を行うサービスが開発されてきた．これらの検索システムの多くでは，管理者や利用者がデジタルメディアの付加情報を手動で入力し，その付加情報をもとに検索が行われている．一方，現在も膨大な量のデジタルメディアが作成されており，これらすべてのデジタルメディアに対する付加情報を手動で作成することは困難である．そのため，デジタルメディアの内容を自動的に認識し，その認識結果をもとにデジタルメディアの付加情報を自動的に作成する仕組みが必要とされている．特に音楽を演奏したときの音声データ（演奏音データ）の場合，その音楽を演奏している楽器を分類することができれば，その演奏楽器に基づいた楽曲分類や楽曲検索のサービスを構築することが可能となる．

多くの音楽は複数の楽器により演奏されているため，単純なアルゴリズムでは演奏音データから演奏している楽器を正確に分類することは困難である．そこで，畳み込みニューラルネットワークを用いて演奏音データを解析することで，演奏している楽器を自動的に分類する手法が提案されている [1]．畳み込みニューラルネットワークとは，畳込み計算を導入した多層ニューラルネットワークであり，画像認識の分野において高い分類性能を達成している [2]．一方，音声解析を行う場合，演奏音データの周波数変換を行うことにより，その演奏音データの特徴量を表現した 2 次元データを作成できる．そこで，演奏音データの周波数変換を行い，作成した 2 次元データを畳み込みニューラルネットワークで分類することにより，その演奏音データに含まれる演奏楽器を自動的に分類できる．

近年の畳み込みニューラルネットワークに関する研究により，単一の計算経路のみを持つ畳み込みニューラルネットワークより，複数の計算経路を持つ畳み込みニューラルネットワークの方が高い性能を発揮することが知られている [3], [4]．しかし，既存の演奏音データの楽器分類手法 [1], [5] では，単一の計算経路を持つ畳み込みニューラルネットワークを使用しており，複数の計算経路を持つニューラルネットワークについては検討されていない．そこで，本論文では，音楽を演奏している楽器を自動的に分類するための新しい畳み込みニューラルネットワークを提案する．提案手法では，複数の計算経路を有する畳み込みニューラルネットワークを設計し，このニューラルネットワークを用いて演奏音データを解析することにより，従来手法よりも高い精度で演奏楽器を分類することが可能となる．また，新しい正則化手法である Signal Augmentation を導入することにより，畳み込みニューラルネットワークの分類精度をさらに向上させる．これらの手法を導入した演奏音データの演奏楽器分類ソフトウェアを実装[*1]し，IRMAS データセット [6] を用いて実装したソフトウェアの性能を評価した．この評価により，本論文で提案する畳み込みニューラルネットワークを用いた演奏楽器分類ソフトウェアは既存手法よりも高い性能を達成すること確認した．また，新たに汎化性能改善手法 Signal Augmentation を導入することにより，畳み込みニューラルネットワークの性能が改善することを検証した．IRMAS データセットを用いた演奏楽器分類タスクにおいて，本論文の提案手法の F 値は 0.708 となった．これは，従来手法の F 値を 0.1 以上改善したものであり，現在（2018 年 12 月）の世界最高性能（state-of-the-art）である．

以降，2 章では既存の演奏音データの演奏楽器分類手法と畳み込みニューラルネットワークについて述べ，3 章では畳み込みニューラルネットワークを用いた演奏楽器分類手法と汎化性能改善手法 Signal Augmentation を提案する．また，4 章では提案手法の評価結果を示し，5 章で提案手法の考察と今後の課題を述べる．

2. 関連研究

従来より，演奏音データを解析して演奏楽器を分類することを目的とした研究がなされており，統計解析や機械学習を用いた演奏楽器分類手法が提案されている [7], [8], [9]．これらの手法では，メル周波数ケプストラム係数を用いて演奏音データの特徴量を反映したデータを作成し，混合ガウスモデルやサポートベクターマシンを用いてこれらのデータの演奏楽器を分類する．一方，近年，畳み込みニューラルネットワークに関する研究がさかんに行われており，画像認識分野における目覚ましい成果が報告されている．そこで，演奏音データに対して周波数変換を行うことで音楽の特徴量を表現した 2 次元データを作成し，この 2 次元データを畳み込みニューラルネットワークで分類することにより，従来よりも高い精度で演奏楽器を分類する手法が提案されている [1], [5]．しかし，これらの既存手法は単一の計算経路のみで構成される畳み込みニューラルネットワークを利用しているため，容易に過学習に陥るという問題がある．

畳み込みニューラルネットワークとは，畳み込み計算を導入したニューラルネットワークであり，画像認識や音声認識において高い性能を有することが知られている [2], [10]．しかし，畳み込みニューラルネットワークは表現力が非常に高いため，単一の計算経路のみで構成される畳み込みニューラルネットワークは容易に過学習状態に陥ることが分かっている [11]．一方，複数の計算経路を持つニューラルネットワークとして，浅い計算経路と深い計算経路の両方を持つ畳み込みニューラルネットワーク ResNet が提案されている [3]．ResNet は，浅い計算経路を持つため勾配消失問題を回避することができ，複数の計算経路によるアンサンブル効果により過学習問題を回避できると考えられている [12]．さらに，ResNet の計算順序の改良

[*1] https://github.com/YukiSato-ml/SignalAugML

も行われ，PreActivation ブロックを用いた ResNet [13] や SingleReLU ブロックを用いた ResNet [14] が提案されており，これらの畳み込みニューラルネットワークは画像分類において高い性能を発揮する．

本論文では，画像分類のための畳み込みニューラルネットワークである ResNet を用いて，演奏音データの演奏楽器を分類する手法を提案する．ResNet を用いることにより，従来手法よりも高い精度で演奏音データの演奏楽器を分類できると考えられる．さらに，ResNet の汎化性能を向上させるため，新しい汎化性能改善手法である Signal Augmentation を導入する．ニューラルネットワークの汎化性能を改善する手法として Dropout がある [15]．しかし，ResNet に対して Dropout を導入したとしても，ResNet の汎化性能を改善させることは難しい．一方，Signal Augmentation は複数の計算経路を有するニューラルネットワークを想定した汎化性能改善手法であり，これを用いることで ResNet の汎化性能を改善できる．

3. 畳み込みニューラルネットワークを用いた演奏楽器の分類手法

3.1 演奏楽器の分類手順

本論文では，音楽を演奏したときの音声波形データ（演奏音データ）から，その音楽を演奏している楽器を分類する手法を提案する．**図 1** に提案手法の概要を示す．提案手法では，既存手法 [1] と同様に，演奏音データに対して前処理を行い，その演奏音データの特徴量を反映した 2 次元データを作成する．そして，作成した 2 次元データを畳み込みニューラルネットワークを用いて解析することにより，その演奏音データに含まれる演奏楽器を分類する．

演奏音データは時間ごとの音波の変化を表現したものであり，音質や音色などの楽器分類に必要となる情報を直接表現しているとはいい難い．そこで，演奏音データに対して短時間フーリエ変換（short-time Fourier transform）を実行し，時間ごとの周波数スペクトルを計算する．さらに，計算した周波数スペクトルと人間の知覚尺度（メル尺度）を表現したメルフィルタバンクをかけ合わせることにより，人間が感じる音質や音色を表現したメル周波数スペクトログラムを生成する．生成したメル周波数スペクトログラムの値を，メル周波数の軸と時間軸で整理することにより，メル周波数スペクトログラムの変化を表現した 2 次元データを作成する．作成した 2 次元データは演奏音データを特徴量を反映したものとなっており，この 2 次元データを畳み込みニューラルネットワークを用いて解析することにより，演奏楽器の分類が可能となる．ただし，畳み込みニューラルネットワークは一定の大きさの 2 次元データを入力として受け付ける．しかし，この 2 次元データの時間軸の長さは演奏音データの記録時間に依存するため，作成した 2 次元データの大きさが一定になるとは限らない．そこで，作成した 2 次元データから一定の時間軸の長さ（window サイズ）を切り出すことにより，畳み込みニューラルネットワークに入力するための固定長の 2 次元データを作成する．通常，window サイズは演奏音データの記録時間よりも十分に短い値を設定するため，1 個の演奏音データから複数の 2 次元データが作成される．

図 2 に提案手法における演奏楽器の分類手順の手順を示す．提案手法では，1 個の演奏音データから作成した複数の 2 次元データをそれぞれ畳み込みニューラルネットワークへ入力し，それぞれの時間で演奏されている楽器を分類する．このとき，畳み込みニューラルネットワークの出力は N 次元ベクトル（N は分類楽器の数）であり，このベクトルの各要素値はそれぞれの楽器が演奏されているかどうかを示している．提案手法では，1 個の演奏音データから複数の出力ベクトルが得られる．そこで，これらのベクトルを時間軸方向に window サイズ 6・hop サイズ 3 の max pooling で圧縮し，圧縮した要素値の平均が閾値以上となった楽器を演奏楽器として判定する．

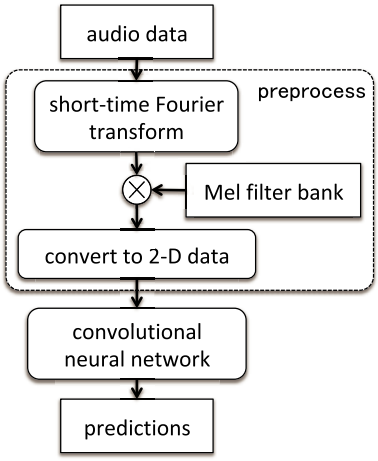

図 1 演奏楽器の分類手法の概要
Fig. 1 Outline of musical instrument classification.

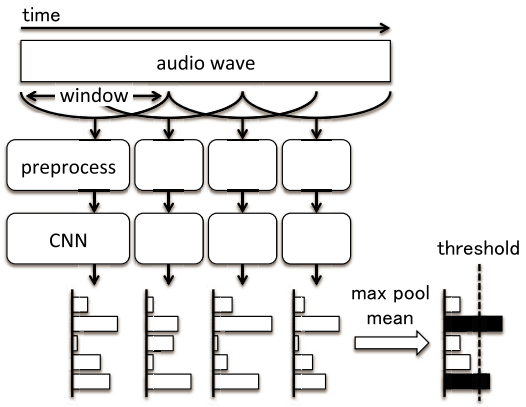

図 2 演奏楽器の分類手順
Fig. 2 Process of musical instrument classification.

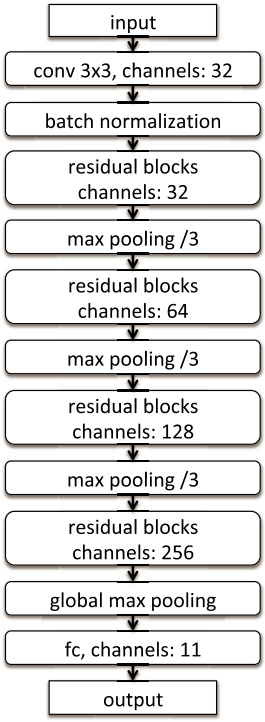

図 3 演奏楽器分類のための畳み込みニューラルネットワーク
Fig. 3 Convolutional neural network for music instrument classification.

3.2 畳み込みニューラルネットワーク

図 3 に提案手法で用いる畳み込みニューラルネットワークの構造を示す．提案手法では，ImageNet [16] などの画像分類タスクで用いられる畳み込みニューラルネットワーク ResNet [3] の構造を参考に，演奏楽器を分類するための畳み込みニューラルネットワークを設計した．具体的には，畳み込み計算を行う residual block とデータ縮小を行う max pooling を交互に配置した構造となっており，局所的な特徴と全体の関係の両方を計算できるようになっている．ただし，画像分類と異なり，1 個の音楽を複数の楽器で演奏している可能性があるため，演奏楽器分類タスクではマルチラベルに対応しなくてならない．そこで，提案手法の畳み込みニューラルネットワークの出力値は，softmax 関数を用いて計算した確率ではなく，sigmoid 関数を用いて正規化した値となっている．

図 3 に示す提案手法の畳み込みニューラルネットワークは PreActivation [13] の residual block を想定した構造となっている．そのため，**図 4** に示す，PreActivation の residual block や SingleReLU の residual block を適用できる．これらの residual block は高い分類性能を発揮する residual block として知られており，2018 年現在の residual block の標準となっている．さらに，提案手法では，これらの residual block に対して汎化性能を向上させる仕組みを導入する．具体的には，図 4 の「Regularize 1-4」のいずれかの場所に，汎化性能向上手法である Dropout [15] か Signal Augmentation を導入する．Dropout は多くのニューラルネットワークで利用されている汎化性能向上手法であり，ニューラルネットワークの分類性能を向上させると考えられている．しかし，ResNet などの複数の計算経路を持つニューラルネットワークの場合，Dropout を導入したことによりその汎化性能を向上したという報告は少ない．一方，Signal Augmentation は我々が新たに開発しているニューラルネットワークの汎化性能向上手法であり，ResNet に導入した場合，その分類性能の向上を期待できる．

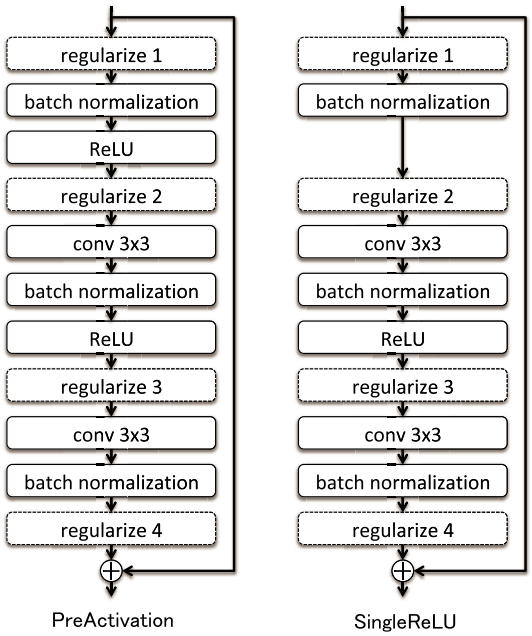

図 4 Residual Block の構造
Fig. 4 Structure of residual blocks

3.3 Signal Augmentation による汎化性能の改善

Signal Augmentation は ResNet などの複数の計算経路を持つニューラルネットワークを対象とした汎化性能向上手法である．複数の計算経路を持つニューラルネットワークでは，それぞれの経路で行われた計算結果を統合したものが最終的な出力値となる．そのため，一部の経路で誤った計算が実行されたとしても，全体としては正しい計算結果を得られることがある．そこで，一部の経路が不正確な計算を行う状態であっても，全体としては正しい計算結果を得られるようにニューラルネットワークの学習処理を実行する．これにより，訓練データでは想定されていない計算が発生し，ニューラルネットワークを順伝播する信号の種類が増加するため，ニューラルネットワークの汎化性能を改善できると考えられる．現在まで，画像分類タスクを用いた実験評価を行っており，畳み込みニューラルネットワークの分類性能を向上させることを実験を通じて確認している [17]．そこで，提案手法では，Signal Augmentation を導入することにより，畳み込みニューラルネットワークの演奏楽器分類性能の向上を目指す．

Signal Augmentation は，ニューラルネットワークの順

伝播の信号にガウスノイズをかけ合わせることで，ニューラルネットワークが過学習状態となることを防ぐ．具体的な Signal Augmentation の計算は，

$$\boldsymbol{y} = \boldsymbol{x} \cdot \mathrm{gaussian}(\sigma) \tag{1}$$

となる．ここで，$\boldsymbol{x}$ は Signal Augmentation への入力値であり，$\boldsymbol{y}$ は Signal Augmentation からの出力値である．また，gaussian は平均が 1 となるガウスノイズを発生させる関数であり，σ は発生させるガウスノイズの標準偏差である．式 (1) に示す Signal Augmentation の計算は，ニューラルネットワークの順伝播のときにのみ行う計算であり，逆誤差伝播を行うときにはこの計算を考慮しない．この計算を導入することにより，Data Augmentaion と同様に順伝播する信号の種類が増えるため，ニューラルネットワークが過学習状態に陥ることを回避できると考えられる．

Signal Augmentation と同様に，順伝播の信号の一部をランダムに削除し，逆誤差伝播を通常どおりに実行することにより，ResNet の汎化性能を改善する手法が提案されている [18]．しかし，従来手法ではニューラルネットワークの学習処理が不安定となるため，ニューラルネットワークの構造や分類タスクの内容によってはこの手法を適用できない場合がある．一方，Signal Augmentation ではパラメータ σ を調整することにより学習処理の安定性を制御できるため，様々なニューラルネットワークや分類タスクに適用可能である．具体的には，σ に大きい値を設定した場合はニューラルネットワークの汎化性能の改善を期待でき，σ に小さい値を設定した場合は学習処理が安定する．

4. 実験評価

4.1 データセット

演奏音データに含まれる演奏楽器を分類する精度を評価するためのデータセット IRMAS（a dataset for instrument recognition in musical audio signals）[6] を用いて，本論文の提案手法を評価した．IRMAS を用いた演奏楽器分類タスクで分類対象となる楽器の種類を**表 1** に示す．IRMAS は様々な楽曲の演奏音データを収録したデータセットであり，それぞれの演奏音データに対してその楽曲の演奏楽器の種類がラベルとして設定されている．IRMAS では，訓練データとして 3 秒の演奏音データが 6,705 個あり，それぞれの演奏音データに 1 個の楽器ラベルが正解データとして設定されている．また，同様に，テストデータとして 5〜20 秒の演奏音データが 2,874 個あり，それぞれの演奏音データに複数の楽器ラベルが正解データとして設定されている．

表 1 IRMAS の分類対象となる演奏楽器とそのデータ数
Table 1 Musical instruments and the number of data in IRMAS classification tasks.

instruments	# of train data	# of test data
Cello	388	111
Clarinet	505	62
Flute	451	163
Acoustic guitar	637	535
Electric guitar	760	942
Organ	682	361
Piano	721	995
Saxophone	626	326
Trumpet	577	167
Violin	580	211
Voice	778	1,044

4.2 実装

多層ニューラルネットワークのフレームワークである Chainer 2.0 [*2] を用いて提案手法を実装した．この実装では，22,050 Hz にダウンサンプリングした演奏音データに対し，window サイズが 2,048・hop サイズが 1,024 の短時間フーリエ変換を実行することで周波数スペクトルを計算した．また，この周波数スペクトルとフィルタ数 128 のメルフィルタバンクをかけ合わせることでメル周波数スペクトログラムを生成し，メル周波数スペクトログラムを window サイズ 1 秒・hop サイズ 0.5 秒で切り出すことで 43×128 の 2 次元データを作成した．これらのデータ処理に用いたパラメータは既存の楽器分類手法 [1] の場合と同じ値である．これにより，33,525 個の訓練用 2 次元データと 93,783 個のテスト用 2 次元データを作成した．

畳み込みニューラルネットワークは，図 4 に示す PreActivation で構成される畳み込みニューラルネットワークと SingleReLU で構成される畳み込みニューラルネットワークをそれぞれ実装した．また，図 4 の「Regularize 1-4」のいずれかの場所に，汎化性能向上手法である Dropout か Signal Augmentation を導入したニューラルネットワークも実装した．このとき，Dropout の確率は 0.2 とし，Signal Augmentation のノイズの標準偏差 σ は 0.8 とした．

畳み込みニューラルネットワークの学習では，各パラメータを MSRA [19] で初期化し，MomentumSDG（momentum=0.9, weight decay=1e-4）でこれらのパラメータを更新した．また，ミニバッチサイズを 100 とし，畳み込みニューラルネットワークの出力誤差は sigmiod cross entropy とした．IRMAS の訓練データのすべてを用いてパラメータ更新を行うまでを 1 epoch とし，650 epoch の学習を実行した．また，学習開始時の学習係数は 0.3 とし，200，350，500 epoch において学習係数を 0.1，0.01，0.001 に変更した．

提案手法では，畳み込みニューラルネットワークの出力値が閾値以上となった場合，対応する楽器が演奏されて

[*2] https://chainer.org

いるものと判断する．このとき，適切な閾値は畳み込みニューラネットワークの学習状態により異なる．そこで，畳み込みニューラルネットワークの学習を行った後，テストデータを用いたグリッドサーチにより閾値を求めた．

4.3 評価

データセット IRMAS と提案手法の実装を用いて実験を行い，提案手法の性能を評価した．ここでは，提案手法の楽器分類性能を評価する指標として精度（P_{micro}），再現率（R_{micro}），および，F 値（F_{micro}）を用いる．提案手法を用いてテストデータ全体の演奏音データの演奏楽器の分類を行ったとき，楽器 m（$m \in M$）について true positive だった結果の数を tp_m，false positive だった結果の数を fp_m，false negative だった結果の数を fn_m とする．このときの精度，再現率，および，F 値は，

$$P_{micro} = \frac{\sum_M tp_m}{\sum_M tp_m + \sum_M fp_m} \tag{2}$$

$$R_{micro} = \frac{\sum_M tp_m}{\sum_M tp_m + \sum_M fn_m} \tag{3}$$

$$F_{micro} = \frac{2P_{micro}P_{micro}}{P_{micro} + R_{micro}} \tag{4}$$

となる．ここでは，それぞれのニューラルネットワークについて初期化・学習・テストを 3 回実行し，これら 3 回のテストにおける tp_m，fp_m，fn_m の平均値を求め，これらの平均値から求めた P_{micro}，R_{micro}，F_{micro} をそのニューラルネットワークの評価値とする．

この評価では，PreActivation の residual block によって構成された畳み込みニューラルネットワークを用いた場合と，SingleReLU の residual block によって構成された畳み込みニューラルネットワークを用いた場合の分類性能を評価した．また，図 4 に示す「Regularize 1-4」のいずれかの場所に汎化性能を向上させるための Dropout もしくは Signal Augmentation を導入した場合を評価した．以下，これらの場所に Dropout を導入した畳み込みニューラルネットワークを「w/ Dropout 1-4」と記述し，Signal Augmentation を導入した畳み込みニューラルネットワークを「w/ SignalAug 1-4」と記述する．

図 5 に学習時のテストデータに対する分類精度の変化を示し，**図 6** に学習時の訓練データに対する誤差の変化を示す．ここで，Han らは従来手法 [1] の演奏楽器の分類手法を実装し，提案手法と同じ環境で実験した結果である．図 5 より，提案手法が従来手法よりも高い精度で演奏楽器を分類できていることを確認できる．また，Signal Augmentation を導入した畳み込みニューラルネットワークは，図 6 の誤差は高い値となっているが，図 5 では高い精度を達成している．この結果は，Signal Augmentation が過学習を防ぐことで，畳み込みニューラルネットワークの性能を向上させていることを示している．

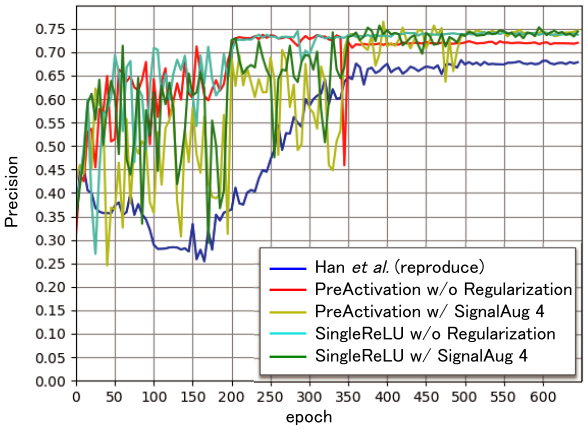

図 5 テストデータに対する精度の変化
Fig. 5 Precision values of test data.

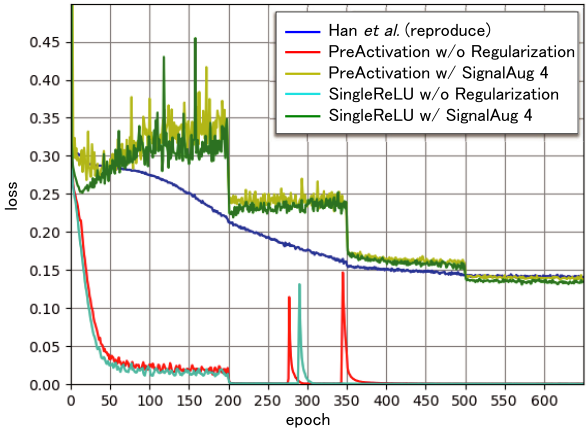

図 6 訓練データに対する誤差の変化
Fig. 6 Loss values of training data.

表 2 に提案手法を用いて IRMAS の演奏楽器分類タスクを行ったときの分類精度を示す．この結果より，従来手法と比較して，提案手法は演奏音データの演奏楽器分類タスクにおいて高い性能を達成することが確認された．特に，PreActivation の residual block によって構成された畳み込みニューラルネットワークでは，図 4 の Regularize 4 の場所に Signal Augmentation を導入することで，高い楽器分類精度（F 値：0.708）を達成した．これは，従来手法 [1] の楽器分類精度を 0.1 以上改善するものであり，現時点（2018 年 12 月）における世界最高性能である．

従来手法 [1] や従来手法 [5] は単一の計算経路で構成される畳み込みニューラルネットワークを使用しており，このニューラルネットワークは容易に過学習に陥るため，その分類性能を改善することは難しいと考えられる．特に，従来手法 [5] では，過学習状態を回避するためにニューラルネットワークのパラメータを調整しているが，その分類性能の改善には至っていない．一方，提案手法では，畳み込みニューラルネットワークとして ResNet を使用することで，従来手法よりも高い分類性能を達成した．これは，ResNet

表 2 IRMAS の楽器分類タスクにおける各手法の評価値
Table 2 Evaluation results in IRMAS musical instrument classification tasks.

method	threshold	P micro	R micro	F micro
Han ら [1]		0.655	0.557	0.602
Pons ら [5]		0.650	0.538	0.589
PreActivation: 4 residual blocks				
w/o Regularization	0.39	0.719	0.686	0.702
w/ SignalAug 1	0.92	0.346	0.626	0.445
w/ SignalAug 2	0.38	0.703	0.686	0.695
w/ SignalAug 3	0.48	0.737	0.650	0.691
w/ SignalAug 4	0.50	0.728	0.689	**0.708**
w/ Dropout 1	0.38	0.704	0.665	0.684
w/ Dropout 2	0.39	0.704	0.676	0.690
w/ Dropout 3	0.47	0.724	0.657	0.689
w/ Dropout 4	0.29	0.586	0.688	0.633
PreActivation: 8 residual blocks				
w/o Regularization	0.37	0.708	0.671	0.689
w/ SignalAug 4	0.48	0.702	0.688	0.695
PreActivation: 16 residual blocks				
w/o Regularization	0.39	0.706	0.689	0.697
w/ SignalAug 4	0.56	0.722	0.665	0.692
SingleReLU: 4 residual blocks				
w/o Regularization	0.36	0.720	0.692	0.706
w/ SignalAug 1	0.92	0.346	0.626	0.445
w/ SignalAug 2	0.44	0.724	0.660	0.691
w/ SignalAug 3	0.51	**0.742**	0.650	0.693
w/ SignalAug 4	0.46	0.718	**0.697**	0.707
w/ Dropout 1	0.38	0.724	0.672	0.697
w/ Dropout 2	0.39	0.725	0.668	0.696
w/ Dropout 3	0.40	0.719	0.680	0.699
w/ Dropout 4	0.39	0.661	0.644	0.653

が複数の計算経路で構成されたニューラルネットワークであり，単一の計算経路で構成されたニューラルネットワークよりも高い汎化性能を有するためと考えられる．さらに，新たな汎化性能改善手法である Signal Augmentation を適用することにより，提案手法の分類性能を改善した．

一方，汎化性能改善手法として Dropout を導入したとしても，楽器分類精度の改善は見られなかった．現在まで，residual block に Dropout を導入することでニューラルネットワークの性能が改善したという報告は少ない．複数の計算経路を含む畳み込みニューラルネットワークの場合，Dropout による分類性能の改善は限定的であると考えられる．この実験においても，residual block で構成されたニューラルネットワークの性能を Dropout を用いて改善する方法を見つけることはできなかった．また，提案手法の residual block の数を増ふやしたとしても，楽器分類精度の改善は見られなかった．一般的に，ResNet は residual block の数を増やすことで表現力が向上するため，分類性能が改善すると考えられている．しかし，提案手法ではメル周波数スペクトログラムを畳み込みニューラルネットワークの入力値としており，この入力値が十分に特徴量を表現しているため，畳み込みニューラルネットワークには多くの表現力を必要としなかったと考えられる．

5. おわりに

本論文では，膨大な量の演奏音データを分類するため，演奏音データからその音楽を演奏している楽器を自動的に分類する手法を提案した．今までにも，畳み込みニューラルネットワークを用いて演奏楽器を分類する手法が提案されてきた．しかし，従来手法では単一の計算経路のみを持つ畳み込みニューラルネットワークを用いているため，過学習を回避することが難しいという問題がある．そこで，提案手法では，複数の計算経路を持つ畳み込みニューラルネットワークである ResNet を用いて演奏楽器の分類を行った．また，汎化性能改善手法である Signal Augmentation を導入することにより，畳み込みニューラルネットワークの分類性能を向上させた．演奏楽器分類タスクのためのデータセットである IRMAS を用いた実験を行い，実験結果より提案手法が従来手法よりも高い分類性能を発揮することを確認した．提案手法の楽器分類性能の F 値は 0.708 であり，これは現時点（2018 年 12 月）における世界最高性能である．

今後の課題としては，演奏音データの前処理を高度化することがあげられる．本論文の提案手法では，演奏音データをメル尺度に変換することで，畳み込みニューラルネットワークへの入力値を作成している．しかし，メル尺度は人間の知覚を基準とした尺度であるため，楽器分類のための尺度として適切であるとは限らない．たとえば，多層ニューラルネットワークを用いて演奏音データの前処理を行うことにより，楽器分類に適した特徴量を表現したデータを作成できると考えられる．

参考文献

[1] Han, Y., Kim, J., Lee, K., Han, Y., Kim, J. and Lee, K.: Deep Convolutional Neural Networks for Predominant Instrument Recognition in Polyphonic Music, *IEEE/ACM Trans. Audio, Speech and Language Processing*, Vol.25, No.1, pp.208–221 (2017).

[2] Krizhevsky, A., Sutskever, I. and Hinton, G.E.: Imagenet classification with deep convolutional neural networks, *Advances in Neural Information Processing Systems*, pp.1097–1105 (2012).

[3] He, K., Zhang, X., Ren, S. and Sun, J.: Deep residual learning for image recognition, *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.770–778 (2016).

[4] Chollet, F.: Xception: Deep Learning with Depthwise Separable Convolutions, *Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), pp.1800–1807 (2017).

[5] Pons, J., Slizovskaia, O., Gong, R., Gómez, E. and Serra, X.: Timbre analysis of music audio signals with convolutional neural networks, *Proc. 25th European Signal Pro-*

cessing Conference (*EUSIPCO*), pp.2744–2748 (2017).
[6] Bosch, J.J., Janer, J., Fuhrmann, F. and Herrera, P.: A Comparison of Sound Segregation Techniques for Predominant Instrument Recognition in Musical Audio Signals, *Proc. 13th International Society for Music Information Retrieval Conference* (*ISMIR*), pp.559–564 (2012).
[7] Essid, S., Richard, G. and David, B.: Musical instrument recognition on solo performances, *Proc. 12th European Signal Processing Conference* (*EUSIPCO*), pp.1289–1292 (2004).
[8] Heittola, T., Klapuri, A. and Virtanen, T.: Musical instrument recognition in polyphonic audio using source-filter model for sound separation, *Proc. 10th International Society for Music Information Retrieval Conference* (*ISMIR*), pp.327–332 (2009).
[9] Duan, Z., Pardo, B. and Daudet, L.: A novel cepstral representation for timbre modeling of sound sources in polyphonic mixtures, *Proc. 2014 IEEE International Conference on Acoustics, Speech and Signal Processing* (*ICASSP*), pp.7495–7499, IEEE (2014).
[10] Van Den Oord, A., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., Kalchbrenner, N., Senior, A.W. and Kavukcuoglu, K.: WaveNet: A Generative Model for Raw Audio, arXiv preprint arXiv:1609.03499 (2016).
[11] Zhang, C., Bengio, S., Hardt, M., Recht, B. and Vinyals, O.: Understanding deep learning requires rethinking generalization, *Proc. 5th International Conference on Learning Representations* (2017).
[12] Veit, A., Wilber, M.J. and Belongie, S.: Residual networks behave like ensembles of relatively shallow networks, *Advances in Neural Information Processing Systems*, pp.550–558 (2016).
[13] He, K., Zhang, X., Ren, S. and Sun, J.: Identity mappings in deep residual networks, *Proc. 14th European Conference on Computer Vision* (*ECCV*), Springer, pp.630–645 (2016).
[14] Han, D., Kim, J. and Kim, J.: Deep pyramidal residual networks, arXiv preprint arXiv:1610.02915 (2016).
[15] Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I. and Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting, *Journal of Machine Learning Research*, Vol.15, No.1, pp.1929–1958 (2014).
[16] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge, *International Journal of Computer Vision*, Vol.115, No.3, pp.211–252 (2015).
[17] 野田遼太郎，今井信太郎，武田敦志：CNN による画像認識精度向上のための付加ノイズの検討，第 17 回情報科学技術フォーラム（FIT2018）講演論文集 (2018).
[18] Gastaldi, X.: Shake-Shake regularization of 3-branch residual networks (2017).
[19] He, K., Zhang, X., Ren, S. and Sun, J.: Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, *Proc. 2015 IEEE International Conference on Computer Vision* (*ICCV*), pp.1026–1034 (2015).

推薦文

音楽データの検索において，データ中に含まれる楽器の情報は，より細かいサービスを構成するために重要であるが，その自動取得は困難である．本論文では，深層学習技術である Residual Neural Network を用い，音楽の波形データからその音楽を演奏している楽器を高精度に分類する手法を提案している．また，データセットを用いた実験評価により，提案手法が世界最高の分類精度（2018 年 1 月現在）を達成していることを示している．本論文で提案されている手法の独創性，論文自体の完成度からも，推薦に値すると判断できるため，本論文を支部からの推薦論文として推薦する．

（情報処理学会東北支部支部長　菅沼 拓夫）

佐藤 佑樹

1996 年生．2018 年東北学院大学教養学部情報科学科卒業．現在，テクノ・マインド株式会社において検査業務の AI 化の概念実証，およびニューラルネットワークを使用した商材の開発に従事．ニューラルネットワークを用いたデータ解析手法に興味を持つ．

武田 敦志 （正会員）

1977 年生．2000 年東北大学工学部電子工学科卒業．2005 年同大学大学院博士後期課程修了．2005 年東北文化学園大学助手, 同大学講師を経て, 2010 年東北学院大学講師，2011 年同大学准教授，現在に至る．自律分散システム，ニューラルネットワークの研究開発に従事．博士（情報科学）．2007 年 FIT 論文賞，2017 年船井ベストペーパー賞各受賞．本会，電子情報通信学会各会員．