E-023

# Residual CNNを用いた楽曲コード進行推定法
# Music chord progression estimation method using Residual CNN

中山翔太 †　荒井秀一 †
Shota Nakayama　Shuichi Arai

## 1. まえがき

音楽情報検索の課題の一つにコード進行推定があり，近年盛んに研究されている．特に最近はディープラーニングを用いたモデルが数多く提案されている．しかし提案されているどのモデルもネットワークの層が浅く，ディープラーニング本来の性能を引き出していないと考える．そこで本稿では，Residual Network(Res Net) に着目し，深いネットワークの学習を可能にする．ResNet は隠れ層を深くするほど精度が上がると言われている手法であり，Res Net と Convoulutional Neural Network(CNN) を組み合わせたコード進行推定モデルを提案することで，更なる精度向上を目指す．

## 2. 先行研究

コード進行は，楽曲における雰囲気を決定づける重要な要素である．楽曲信号からコード進行を正確に推定できれば，楽曲の雰囲気や類似性に基づいた楽曲検索や楽曲推薦，またカバー曲の検索等への応用が可能になる．従来は HMM(Hidden Markov Model) を用いた推定法が主流であったが，近年は画像処理などの分野で用いられている深層学習をコード進行推定に応用した手法が盛んに提案されている．深層学習を用いたコード進行推定法として，F. Koreniowski ら [1] はコード認識の自動化を目的に，DNN(Deep Neural Network) と HMM によるコード進行推定法を提案しているが，この手法は HMM のパラメタ推定の際に大量の学習データを必要とするという問題点がある．また，X. Zhou ら [2] は，高度なコード進行推定を目指し，ボトルネック構造の DNN によるコード進行推定を行った．そして先行研究として挙げられる全ての手法は，深層学習を用いているものの隠れ層が多いもので 3 層ほどであり，ディープラーニング本来の性能を引き出していないと考える．また S.Nakayama ら [3] は DNN を Resnet 化し，コード進行推定に用いていたが，CNN を Resnet 化したネットワークモデルは提案されていない．そこで本稿では，Residual Network[4] に着目し，深いネットワークの学習を可能にさせ，更なる認識精度の向上を図れるのではないかと考えた．

## 3.Residual Network

ResNet のアイデアはとてもシンプルで，隠れ層への入力と隠れ層からの出力を足し合わせることで，勾配法による最適化が容易になるというものである．例えば，図 1 左のようなネットワークを考える．ここで”Weight Layer”は通常の隠れ層であり，この例では，入力 $x$ を 2 層の隠れ層に通した出力 $H(x)$ を勾配法で最適化するというのが一般的な手法である．これに対し，ResNet では図 1 右のように最適化したい $H(x)$ と入力 $x$ との残差関数 $F(x) := H(x) - x$ をネットワークの出力として学習させる．すなわち，本来の出力 $H(x)$ は $H(x) = F(x); x$ と書けるので，図のように Shortcut Connection をネットワークに追加すればよい．この Shortcut Connection で挟まれた区間を残差ブロックと呼び，この残差ブロックを多段に積み上げて，深いネットワークを構成する．

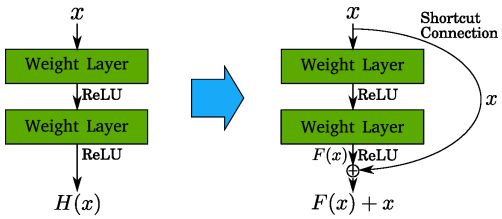


図 1: 一般的な隠れ層から残差ブロックへの変更

これにより，仮に深い層での最適な関数が恒等写像のような関数であった場合，$H(x) := F(x) + x \approx x$ となるよう，零写像に近い $F(x)$ を学習で見つけ出すことになる．これは，小さな勾配を基に重みを更新し，恒等関数のような $H(x)$ を見つけ出すより非常に簡単であり，勾配消失の影響が少ないため，深いネットワークにおける学習が可能になる．

## 4. コード進行推定モデル

本稿では ResNet を組み込んだ CNN と CRF( Conditinal Random Fields) を用いたコード進行推定モデルを提案する．モデル構造を図 2 に示す．ここで，図 2 の”conv”は畳み込み層を表し，$(3 \times 3)$ はフィルタサイズ，Pad はパディング数を表す．各畳み込み層に繋がる入力と出力の矢印に添えられた数字は，それぞれ入力，出力の特徴マップの数を表す．”Batch Normalization”はバッチ正規化層を表す．Max Pooling は Max プーリングの層を表し，Average Pooling は Average プーリングの層を表し，各プーリング層内に記載された括弧内の値は，プーリングの際のフィルタサイズを表す．”Block”は残差ブロックを $n$ 個重ねていることを表す．この推定モデルは，CNN で特徴を抽出し，”Full Connection (全結合層)”で 25 個のラベルに分類し，その結果を CRF で音楽的文脈も加味して最終的なラベル系列として出力する．

## 5. 入力処理

HCQT(Harmonic Constant Q Transform) の窓関数はハニング窓を使用し，各パラメータはサンプリングレートが 22050 Hz，フレーム周期が 1024 sample，最低周波数は 32.700 Hz，最高周波数は 2093.005 Hz で

†東京都市大学大学院　工学研究科

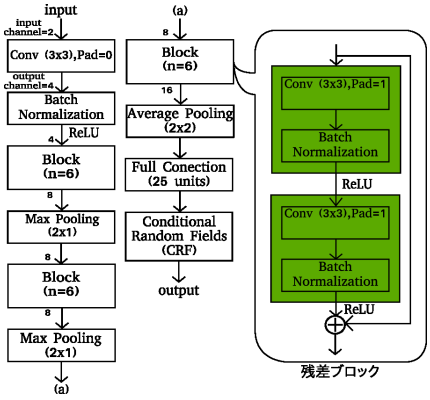


図 2: 提案モデルの構造

あり，これら 2 つの周波数の差は 6 オクターブ分ある．さらに，1 オクターブあたり対数周波数 bin の数を 24 個とし，6 オクターブの間で対数周波数 bin の数が 144 個になるよう設定する．

## 6. 実験条件

### 6.1. 使用データセット

使用したデータセットは isophonics[‡] が公開しているコード進行のアノテーションである．使用したデータセットの内訳を表 1 に示す．

表 1: 使用データセット内訳

| アーティスト | 楽曲数 |
|---|---|
| Beatles | 180 |
| Queen | 20 |
| Zweieck | 18 |
| Total | 218 |

### 6.2. 推定ラベルの種類

本稿では，12 個のピッチクラスと 2 種類 (major,minor) のコードの種類からなる 24 種類のコードと，無音，単音またはパーカッションのみの区間を表す”No Chord”の計 25 種類のラベルの推定を目的とする．

### 6.3. 評価基準

評価基準として WCSR(Weighted Chord Symbol Recall) を使用した先行研究が数多くあるため，本稿では評価基準として WCSR を用いる．また本稿では，交差検証を用いて評価を行う．その際の分割数は 6 である．

$$WCSR = \frac{\text{推定ラベルが正解した時間の総和}}{\text{楽曲の長さの総和}} \quad (1)$$

## 7. 実験結果

本提案手法の推定精度を他のアルゴリズムの推定精度と比較した．比較に用いるモデルは，(1)unit 数 256，隠れ層 3 層の DNN でコードを推定した後，HMM により音楽的文脈を考慮した推定をおこなう DNN-HMM[1] と，(2)8 層の CNN でコードを推定し，その後段に CRF を置くことで音楽的文脈を考慮した推定をおこなう ConvNet[5]，の 2 つである．この 2 つの手法と提案手法を WCSR を用いて評価した結果を表 2 に示す．

表 2: 他のアルゴリズムとの精度比較

| Algorithm | WCSR [%] |
|---|---|
| DNN-HMM | 76.0 |
| ConvNet | 77.6 |
| Proposed | 77.8 |

### 7.1. 検討及び結論

DNN-HMM と比較すると，本提案手法の推定精度は 1.8% 高い．この結果から，DNN-HMM よりも深い隠れ層によって構成される本提案手法を用いることで，より複雑な特徴を学習できたと考える．

ConvNet と比較すると，本提案手法の推定精度との差は 0.2% であり，大きな差はなかった．この理由としては，出力特徴マップの数の違いが考えられる．CNN は各畳み込み層で抽出した特徴を出力特徴マップとして出力する．ConvNet の特徴マップの要素の総数は本提案手法の約 2 倍の数であり，その分だけ多くの特徴を抽出していることが要因だと考える．しかしこれは，ResNet で層を深くすることで，出力特徴マップの要素を半分に削減しても，同等の精度で推定できることを意味し，ResNet の有用性を示す結果であると言える．

これらの結果から，コード進行推定において ResNet が有効であることが分かった．

## 参考文献

[1] Filip Korzeniowski and Gerhard Widmer. On the futility of learning complex frame-level language models for chord recognition. Vol. abs/1702.00178, , 2017.

[2] Zhou X. and Lerch A. Chord detection using deep learning. *In Proceedings of the 16th ISMIR Conference*, pp. 52–58, 2015.

[3] Shota Nakayama and Arai Shuichi. Residual dnn-crf model for audio chord recognition. *International Conference on Intelligent System and Image Processing (ICISIP)*, pp. 92–98, 2017.

[4] He Kaiming et al. Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition.*, pp. 770–778, 2016.

[5] Filip Korzeniowski and Gerhard Widmer. A fully convolutional deep auditory model for musical chord recognition. *Machine Learning for Signal Processing (MLSP)*, pp. 1–6, 2016.

[‡]http://isophonics.net/datasets