Vol.2014-MUS-103 No.20
2014/5/24

Deep learningを用いた連想記憶アーキテクチャによる楽曲の記憶と生成

丹羽 孔明[1,†1,a)]　成瀬 継太郎[2]　大江 亮介[1,†1]　木下 正博[1,†1]　三田村 保[1,†1]　川上 敬[1,†1]

概要：音は聴覚の処理によって振動から認識され, また音楽は主観的な認知の処理によって時系列な音の連続から認識される. 音を構成する個々のパラメータは単純であが, 人間はこの単純なパラメータから音楽的な特徴を抽出し, 特徴の相互作用を知覚し, 感動する. 音楽に対する人間の反応は多様であり, 人間の主観に依存する音楽をコンピュータサイエンスの方法論により自動生成するシステムの実現は魅力的な研究分野である. これまでの音楽生成システムでは主に音高や音長を共にコーディングしたシンボリックな表現が用いられてきたが, 音を構成する個々のパラメータの相互作用が音楽を聴取する人間にとって重要であると仮定すると, オーディオ信号を直接学習することは, 音楽を自動生成するためのシステムにおいて効果的と期待する. 本稿は, 音楽についてより多くの情報を含むと考えられるオーディオ信号より, 音楽の特徴を学習し, また学習した特徴のシーケンスなデータを生成する音楽的特徴の連想記憶アーキテクチャを Deep learning を用いて開発する. このアーキテクチャは Restricted Boltzmann Machine (RBM) とその拡張である Conditional RBM を積層した構造とし, ある時間単位でフレームに分割したオーディオ信号を学習させる. 構築した連想記憶アーキテクチャについて, 音楽の記憶と生成の実験を行い, 音楽のオーディオ信号データを記憶, 生成した実験結果に基づいて, システムのパフォーマンスや有効性を検証する.

1. はじめに

音は聴覚の処理によって振動から認識され, また音楽は主観的な認知の処理によって時系列な音の連続から認識される. 音を構成する個々のパラメータは単純である. しかし, 単純なパラメータから音楽的な特徴を抽出し, この特徴の相互作用を知覚し, 感動する. リラックスしたり気分が高揚したりと音楽に対する人間の反応は多様である. 人間の主観に依存する音楽をコンピュータサイエンスの方法論により自動生成する知的かつ自律的なシステムの実現は魅力的な研究分野である.

コンピュータサイエンスの法論による, 音楽の自動生成は L.Hiller と L.Isaacson によって初めて行われた. このとき生成された曲は Illiac suite と呼ばれている. その後現在までに数多くの作曲システムが研究されている. D.Cope の開発した Experiments in Musical Intelligence (EMI)[1] は数ある作曲システムの中でも質の良い音楽を生成する代表的な作曲システムである. EMI はシステムが音楽を分析して獲得したルールによってあるモチーフに後続するモチーフを決定する. 近年においては, 機械学習分野で注目されている Deep learning のアプローチによる音楽生成も行われている. G.Bickerman らは Deep Belief Networks (DBNs)[2] によりコード進行からジャズの即興演奏のメロディを生成する試みを行っている [3]. N.B.Leeandowski らは高次元のシーケンスデータをモデリングするために Restricted Boltzmann Machine (RBM)[4],[5] を RNN-RBM へ拡張し, この RNN-RBM によってシンボリックな音楽の自動生成を行った [6]. これまでの音楽生成システムでは主に音高や音長を共にコーディングしたシンボリックな表現が用いられてきた. 楽譜データや MIDI データ, またはこれらに基づいた独自の記号形式などである. 音を構成する個々のパラメータの相互作用が音楽を聴取する人間にとって重要であると仮定すると, シンボリックな表現ではないオーディオ信号を直接学習することは, 機械学習の方法論に基づいた音楽を自動生成するためのアーキテクチャにおいて効果的と期待する.

本稿では音楽についてより多くの情報を含むと考えられるオーディオ信号より, 音楽の特徴を学習し, また学習した特徴のシーケンスなデータを生成する音楽的特徴の連想記憶アーキテクチャを Deep learning を用いて開発する. こ

[1] 北海道工業大学
Hokkaido Institute of Tech., 4-20, Maeda 7 jo 15 chome, Teine-ku Sapporo 006–8585, Japan
[2] 会津大学
University of Aizu
[†1] 現在，北海道科学大学
Presently with Hokkaido University of Science
[a)] r13301@hus.ac.jp

のアーキテクチャは RBM とその拡張である Conditional RBM[7] を積層した構造とし, ある時間単位でフレームに分割したオーディオ信号を学習させる. 構築したシステムについて, 音楽の記憶と生成の実験を行い, 音楽のオーディオ信号データを記憶, 生成した実験結果に基づいて, システムのパフォーマンスや有効性を検証する.

2. Restricted Boltzmann Machine

2.1 Restricted Boltzmann Machine

Restricted Boltzmann Machine (RBM) は無向グラフィカルモデルによるエネルギーに基づいた生成メカニズムである. 図 1 中の左の図に示すように, 可視層の各ユニット $\boldsymbol{v} = [v_1, v_2, \ldots, v_V]$ は無向な接続を用いて隠れ層の各ユニット $\boldsymbol{h} = [h_1, h_2, \ldots, h_H]$ に接続される. V, H は $\boldsymbol{v}, \boldsymbol{h}$ それぞれの層のユニット数である. RBM の一般的なモデルにおいて, 可視ユニットと隠れユニットの状態は $[0, 1]$ のバイナリ値によって定義される. この場合に RBM のモデルは次のエネルギーを有する.

$$E(\boldsymbol{v}, \boldsymbol{h}) = -\sum_i a_i v_i - \sum_j b_i h_i - \sum_{ij} v_i h_j w_{ij} \quad (1)$$

ここで $\mathbf{W} = [w_{11}, w_{12}, \ldots, w_{VH}]$ は可視層と隠れ層間の無向接続の重み行列,$\boldsymbol{a} = [a_1, a_2, \ldots, a_V]$ と $\boldsymbol{b} = [b_1, b_2, \ldots, b_H]$ はそれらの層のバイアスである. 可視ユニット $\boldsymbol{v}$ と隠れユニット $\boldsymbol{h}$ の同時確率は, エネルギー関数によって次のように与えられている.

$$p(\boldsymbol{v}, \boldsymbol{h}) = \frac{1}{Z} \exp(-E(\boldsymbol{v}, \boldsymbol{h})) \quad (2)$$

ここで Z は $Z = \sum_{\boldsymbol{v}, \boldsymbol{h}} \exp(-E(\boldsymbol{v}, \boldsymbol{h}))$ で与えられる規格化定数である.

RBM のモデルパラメータ $\boldsymbol{a}, \boldsymbol{b}$ と $\mathbf{W}$ は対数尤度学習のための Contrastive Divergence (CD_k)[8] 学習アルゴリズムによって推定される. CD_k アルゴリズムにおけるパラメータの更新規則は次のように与えられる.

$$\Delta W = \epsilon(\langle v_i, h_j \rangle_{data} - \langle v_i, h_j \rangle_{model}) \quad (3)$$

$$\Delta a = \epsilon(\langle v_i \rangle_{data} - \langle v_i \rangle_{model}) \quad (4)$$

$$\Delta b = \epsilon(\langle h_j \rangle_{data} - \langle h_j \rangle_{model}) \quad (5)$$

ここで $\langle\rangle_{data}$ は学習データ,$\langle\rangle_{model}$ は次式の条件付き確率によって得られる各層の発火状態を用いたギブスサンプリングによって学習データを再構成したデータを表す.

$$p(h_j = 1|\boldsymbol{v}) = sigmoid(b_j + \sum_i v_i w_{ij}) \quad (6)$$

$$p(v_i = 1|\boldsymbol{h}) = sigmoid(a_i + \sum_j h_j w_{ij}) \quad (7)$$

ここで $sigmoid(x)$ は $sigmoid(x) = 1/(1 + \exp(-x))$ によって与えられる標準シグモイド関数である.

本稿では楽曲の連想記憶アーキテクチャの構築に Gaussian Bernoulli RBM (GBRBM)[9] を用いる. 学習対象のデータが楽曲のオーディオ信号であるために入力を受け取る可視ユニットが実数値を扱える必要があるためである. GBRBM は可視ユニットの状態を実数値で定義し, また隠れユニットの状態をバイナリ値で表現したモデルである. GBRBM のモデルではエネルギー関数と可視ユニットの発火状態が次のように与えられる.

$$E(\boldsymbol{v}, \boldsymbol{h}) = \sum_i \frac{(v_i - a_i)^2}{2\sigma_i^2} - \sum_j b_j h_j - \sum_{ij} \frac{v_i}{\sigma_i} h_j w_{ij} \quad (8)$$

$$p(h_j|\boldsymbol{v}) = N(h_j; b_j + \sum_i v_i w_{ij}, \sigma^2) \quad (9)$$

ここで $N(a; b; \sigma^2)$ はガウス関数, また σ は標準偏差である.

2.2 Conditional RBM

Conditional RBM (CRBM) は可視層の過去の状態を考慮することで時系列な状態を学習し, その時系列データを生成可能とした RBM を拡張したモデルの一つである. 図 1 の右の図は CRBM の各層の接続を表す. RBM と同様に可視層の各ユニットは $\boldsymbol{v}$ は無向の接続によって隠れ層の各ユニット $\boldsymbol{h}$ へ接続される. 図中の層 $\boldsymbol{u} = [v_{t-1}, v_{t-2}, \ldots, v_{t-N}]$ はある時刻 t の時点における可視ユニットの過去の状態を表す履歴ベクトル,N はその履歴数である. この履歴ベクトルは有向な接続によって可視層のユニットと隠れ層のユニットに接続される. CRBM のモデルは次式で与えられるエネルギーを有する.

$$E(\boldsymbol{v}, \boldsymbol{h}|\boldsymbol{u}) = \frac{1}{2}\sum_i (v_{i,t} - \hat{a}_{i,t})^2 - \sum_{ij} w_{ij} v_{i,t} h_{j,t} \sum_j \hat{b}_{j,t} h_{j,t} \quad (10)$$

ここで $\mathbf{W}$ は可視層と隠れ層間の無向接続の重み行列,t は現在の時刻である. $\hat{\boldsymbol{a}} = [\hat{a}_1, \hat{a}_2, \ldots, \hat{a}_i]$ と $\hat{\boldsymbol{b}} = [\hat{b}_1, \hat{b}_2, \ldots, \hat{b}_i]$ は各層のバイアスであり, 履歴ベクトル $\boldsymbol{u}$ を用いて次のように求められる.

$$\hat{a}_{i,t} = a_i + \sum_k A_{ki} u_k \quad (11)$$

$$\hat{b}_{j,t} = b_j + \sum_k B_{kj} u_k \quad (12)$$

ここで $\mathbf{A}$ は履歴ベクトル $\boldsymbol{u}$ から可視ユニット $\boldsymbol{v}$ への有向接続の重み行列, $\mathbf{B}$ は履歴ベクトル $\boldsymbol{u}$ から隠れユニット $\boldsymbol{h}$ への有向接続の重み行列を表す.

CRBM のモデルパラメータは RBM と同様に CD_k アルゴリズムににによって推定される. 各層のユニットのある時刻での発火状態は式 (6) と (7) を用いて求められる. ここで式 (6) 中の変数 a_i は $\hat{a}_{i,t}$ に, また式 (7) 中の変数 b_j は $\hat{b}_{j,t}$ にそれぞれ変更される. 各パラメータの更新規則は式 (3),(4),(5) および次の式が用いられる.

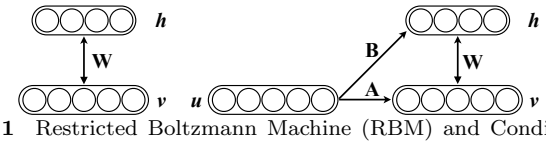

図 1 Restricted Boltzmann Machine (RBM) and Conditional RBM (CRBM). left image illustrates RBM model, and reight shows CRBM model.

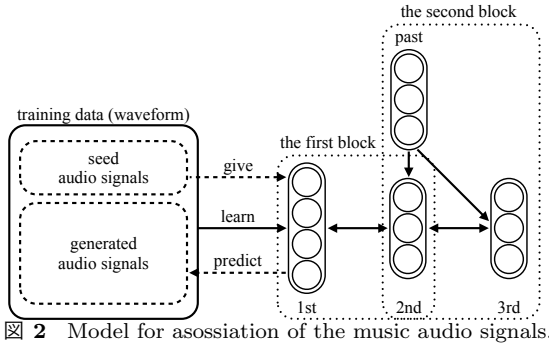

図 2 Model for asossiation of the music audio signals.

表 1 The number of units for each layer in model

1st layer	2nd layer	past stats	3rd layer
16,000	200	1,400	200

$$\Delta A = \epsilon(\langle v_i, u_k\rangle_{data} - \langle v_i, u_k\rangle_{model}) \quad (13)$$

$$\Delta B = \epsilon(\langle h_i, u_k\rangle_{data} - \langle h_i, u_k\rangle_{model}) \quad (14)$$

3. 音楽的な特徴の連想記憶アーキテクチャ

3.1 連想記憶アーキテクチャの構成

連想記憶アーキテクチャは GBRBM と CRBM を積層し, ある時間単位でフレームに分割したオーディオ信号を学習する (図 2). 連想記憶アーキテクチャの第一ブロックは GBRBM により各フレーム単位で音楽の特徴を学習する. オーディオ信号はこの第一ブロックにおいて特徴ベクトルへ変換され, また特徴ベクトルからオーディオ信号へ再合成される. 第二ブロックでは第一ブロックで学習された音楽の特徴ベクトルのシーケンスデータを学習する. 学習完了後の第二ブロックにおいて, 新しい特徴ベクトルのシーケンスデータを生成し, これを第一ブロックにおいてオーディオ信号へと戻す. 各層のユニット数は表 1 に示すユニット数で設定した.

3.2 連想記憶アーキテクチャの学習

オーディオ信号をフレームへ分割したシーケンスなデータを訓練データとし,CD_k アルゴリズムによって連想記憶アーキテクチャの各ブロックのモデルパラメータを推定した. 学習の際の各学習パラメータの設定は表 2 の通りに設定した. 訓練データは Vivaldi 作曲の「和声と創意への試み」より「春」「夏」「秋」「冬」, 各曲の第 1〜第 3 楽章から冒頭 15 秒間を切り出し, これを繋げた 180 秒のオーディオ信号を用いた. この訓練用のオーディオ信号は表 3 の設定を用いてリサンプリングし, また各フレームへの分割を行った.

表 2 Learning parameters for each blocks in model

rearning rate	momentum	weight decay	learning epochs
0.0001	0.9	0.001	6000

表 3 Parameters for resampling and frame spliting

sampling rate	quantization bit rate	channels
16kHz	16bit	256kbps
window size	window function	overlap
16,00	rectangular window	0

4. 実験結果

まず, 訓練データの冒頭 30 秒間を学習を完了した連想記憶アーキテクチャに与え,180 秒間のオーディオ信号を生成した. 生成されたオーディオ信号は開始から約 7 秒間程度はノイズの様であったが, その後の区間において訓練データと同様の音を聞き取ることのできる信号が得られた (図 3).

次に, 未学習のオーディオ信号を与えて 360 秒間のオーディオ信号を生成した. このとき与えたオーディオ信号は Vivaldi 作曲の「和声と創意への試み」より協奏曲第 9 番の第 1 楽章から冒頭の 20 秒間である. このとき連想記憶アーキテクチャから生成されたオーディオ信号は開始から約 90 秒区間ではノイズのようであったが, 徐々に訓練データと同様の音を聞き取れるようになり, およそ 145 秒から 205 秒の区間においてノイズが減少し, 訓練データと同様の音をはっきりと聞き取れる (図 4).

未学習のオーディオ信号を与えた際について, 学習したオーディオ信号を生成する他に, 強いノイズの中に時折, 訓練データの音楽の断片が含まれたオーディオ信号を生成するケースが確認された. 連想記憶アーキテクチャの第二ブロックにおいて, 新しい音楽の特徴のシーケンスデータの生成が, 初期値として与えているオーディオ信号より得られる特徴に強く依存していることが考えられる.

これらの生成されたオーディオ信号には多くのノイズが含まれている. しかしながら, 生成されたオーディオ信号はノイズを含むにもかかわらず人間のテスターとっては元の曲と認識できるオーディオ信号が得られている. これにより, 構築した連想記憶アーキテクチャは RBM と CRBM によりオーディオ信号より音楽の特徴を学習, 記憶し, また記憶したオーディオ信号の再生が可能であることが確認された.

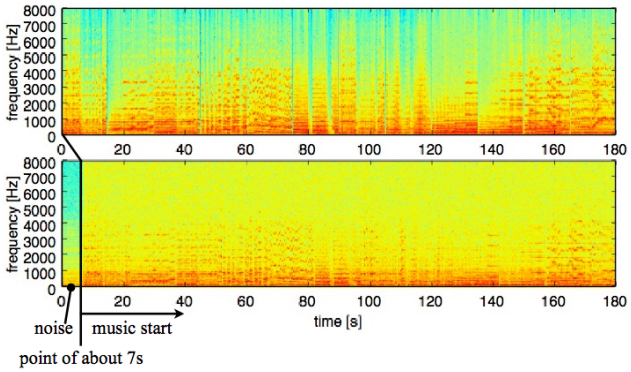

図 3 Power spectrogram of the learning audio signals and generated audio signals. upper image is spectrogram of the audio signals for learning, and lower image is spectrogram of the audio signals by model generation.

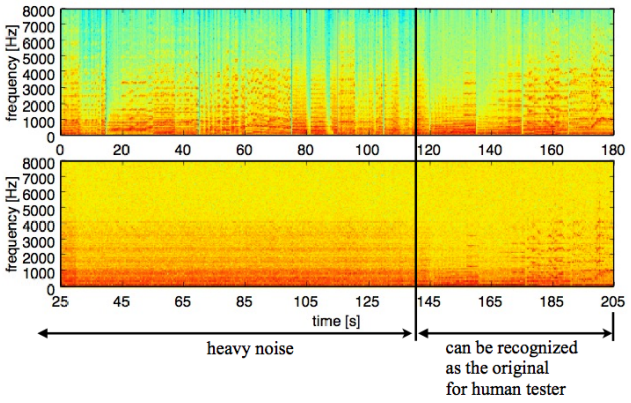

図 4 Power spectrogram of the generated audio signals using unlearned data. upper image is spectrogram of the audio signals for learning, and lower image is spectrogram of the audio signals by model generation.

4.1 おわりに

本稿では音楽についてより多くの情報を含むと考えられるオーディオ信号より, 音楽の特徴を学習し, また学習した特徴のシーケンスなデータを生成する音楽的特徴の連想記憶アーキテクチャを Deep learning によって構築する. このアーキテクチャは RBM とその拡張である Conditional RBM を積層した構造とし, ある時間単位でフレームに分割したオーディオ信号を訓練データとして学習させる. この連想記憶アーキテクチャについて, オーディオ信号を記憶し再生できることを確認した. しかしながら学習した音楽の特徴を反映した新しい音楽の生成は確認されていない.

本稿で訓練に用いた音楽のデータは 4 曲 (12 楽章分) かつそれぞれから冒頭 15 秒を切り出したのみである. これは音楽を自動生成するアーキテクチャのための訓練データとしては小規模な量である. このため, 音の時間的な変化や反復といった音楽の 1 つの楽曲に含まれる特徴を学習するには不十分であった可能性がある. また音楽をシンボリックに表現したデータでは高度に情報が抽象化されているが, 生のオーディオ信号はそうではない. Deep learning では層が深くなるにつれ, 下位の層で得られる特徴を統合したより高次の特徴を学習することが知られている. しかしながら, 本稿での連想記憶アーキテクチャは RBM が 1 ブロック,CRBM が 1 ブロックの 3 層構造である. このため十分に高次な特徴を学習できていない可能性が考えられる. Deep learning によりオーディオ信号中からよく特徴を学習するためにも, 現在より多くの層を大規模なオーディオ信号のデータセットにより学習することが求められる.

今後, 学習した音楽に基づいた新しい音楽のオーディオ信号を構築した連想記憶アーキテクチャにより生成するために, より多くの音楽のオーディオ信号をこのアーキテクチャへ学習させる. また, 連想記憶アーキテクチャの各レイヤーとそのユニット数などのネットワークのモデルパラメータについて, オーディオ信号ベースの音楽の自動生成を行うために適切なパラメータを検討する. これらに加え, 生成されたオーディオ信号に含まれるノイズを除去する方法についても重要な検討課題の一つである.

参考文献

[1] Cope, D.: Computers and Music Style, Computer Music and Digital Audio Series (1991).
[2] Bengio, Y.: Learning deep architectures for AI, Foundations and trends in Machine Learning, 2.1, 1-127 (2009).
[3] Bickerman, G., Bosley, S., Swire, P., and Keller, R. M.: Learning to Create Jazz Melodies Using Deep Belief Nets, First Inteirnational Conference on Computational Creativity (2010).
[4] Smolensky , P .: Information processing in dynamical systems: Foundations of harmony theory, Parallel Distributed Processing: Explorations in the Microstructure of Cognition, vol 1, pp.194-281, MIT Press (1986).
[5] Hinton, G. E.: A Practical Guide to Training Restricted Boltzmann Machines, Tech. Rep. Department of Computer Science, University of Toronto (2010).
[6] Lewandowski, N. B., Bengio, Y., and Vincent, P.: Modeling temporal dependencies in high-dimensional sequences: Application to polyphonic music generation and transcription, 29th International Conference on Machine Learning (ICML 2012) (2012).
[7] Taylor, G. W., and Hinton, G. E.: Factored conditional restricted Boltzmann machines for modeling motion style, Proceedings of the 26th annual international conference on machine learning, pp.1025-1032 (2009).
[8] Hinton, G. E.: Training products of experts by minimizing contrastive divergence, Neural computation, vol.14-8, pp.1771-1800 (2002).
[9] Cho, K. H., Ilin, A., Raiko, T.: Improved Learning of Gaussian-Bernoulli Restricted Boltzmann Machines, Artificial Neural Networks and Machine Learning-ICANN 2011, pp.10-17 (2011).