Vol.2017-MUS-116 No.11
2017/8/25

ディープラーニングによるループ音源の自動生成

細川 皓平[1,a)]　横山 想一郎[1,b)]　山下 倫央[1,c)]　川村 秀憲[1,d)]

概要：本研究では音楽や効果音のような生の音データを生成することを目的としている．ここでは，ディープラーニングの一種である，敵対的生成ネットワーク (generative adversarial network, GAN) を用いて，ループ音源と呼ばれる短い音楽データの生成を行う．実験として 2 秒のループ音源 4 つを訓練データとして学習に使用し，ランダムに生成したデータと訓練データと比較した．生成データはそれぞれの訓練データから一部を模倣したようなものとなり，かつ完全に同一のものは生成されなかった．

Generating Audio Loops using Deep Learning

Hosokawa Kouhei[1,a)]　Yokoyama Soichiro[1,b)]　Yamashita Tomohisa[1,c)]　Kawamura Hidenori[1,d)]

1. 序論

自動での音楽や効果音といった音データの生成は, 楽曲制作，BGM の生成，動画への効果音の付与など，様々な場面で求められている．このような生の音データの生成を実現するために，敵対的生成ネットワーク (Generative Adversarial Networks, GAN) を用いた手法を提案する．

音楽情報学の分野では Hiller ら [1] による「イリアック組曲」を初めとした自動作曲の研究が古くから行われており，現在もその研究は発展し続けている．しかし一般的に作曲とは楽譜を生成することを示しているため，実際の音データを生成するためには，さらに別の技術が求められる．我々の目的は音データそのものを生成することであるため，自動作曲による手法をとるためには，さらに音色などの情報も自動生成しなければならない．

一方で，Oord ら [2] は，Wavenet というモデルにより生の音データを学習，生成することを実現した．Wavenet ではディープラーニングの一種である Recurent Neural Network の構造を用いて，文章の読み上げ (text-to-speech) の分野で高い性能を示した．この研究では音楽データの生成に関する実験も行っているものの，全体的なジャンル、音量、音色、音質などについて統一性がないとしている．

生成モデルとしては近年，画像生成の分野において Goodfellow ら [3] による敵対的生成モデル (Generative Adversarial Networks, GAN) が注目されている．GAN は Generator と Discriminator と呼ばれる 2 つのネットワークにより構成されたモデルで，従来主流であった Autoencoder[4] と比べて非常に鮮明な画像の生成が実現された．Generator は乱数により生成された 100 次元ベクトル z を入力とし，画像データを出力する．一方の Discriminator は本物の画像データである訓練データと，Generator により生成されたデータを分類する．Discriminator はより正確にこの判別を行うように学習を進めていく一方で，Generator は Discriminator に本物の画像であると誤判定するように学習を進めていく．この 2 つの学習を交互に続けていくことによって，Generator は本物の画像に似ていながらも同一ではない画像を生成することが可能となる．さらに，ベクトル z は生成画像の特徴分布を示すことから，入力を操作することにより任意の特徴を持った画像を生成することができる．

しかし GAN はこのような利点がある一方で，学習が非常に不安定であるという問題がある．Radford ら [5]

[1] 北海道大学大学院情報科学研究科
Graduate School of Information Science and Technology
Hokkaido University Hokkaido, Japan

a) hosokawa@complex.ist.hokudai.ac.jp
b) yokoyama@complex.ist.hokudai.ac.jp
c) tomohisa@complex.ist.hokudai.ac.jp
d) kawamura@complex.ist.hokudai.ac.jp

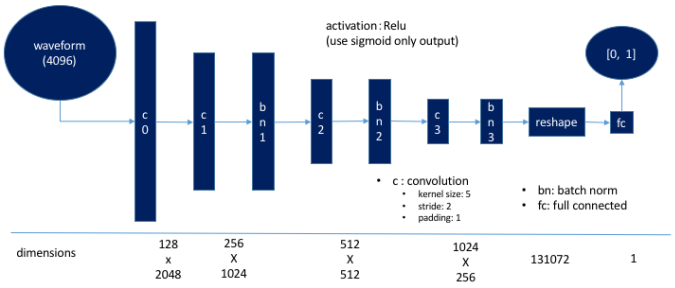

図 1 Discriminator の概略

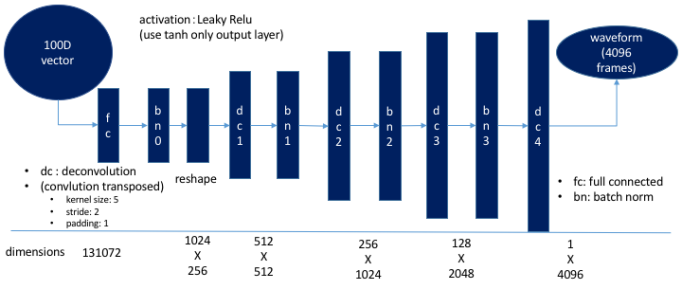

図 2 generator の概略

は Deep Convolutional Generative Adversarial networks (DCGAN) というモデルを構築し，安定した学習を行い，効果的な分布を獲得するための一例を提案している．本研究では DCGAN のモデルをベースに音データのための生成モデルを構築することにより，ループ音源の自動生成を目指す．

2. 手法

2.1 Discriminator

Discriminator のネットワーク構成は図 1 に示すとおりである．畳み込み層，バッチ正規化層，活性化関数として Rectifierd Linear Unit (Relu) を繰り返した構造となっている．畳み込み層のパラメータはカーネルサイズ 5，ストライド 2，パディング 1 となっており，出力チャネル数は入力チャネル数の半分となっている．

最終層は活性化関数にシグモイド関数を用いた全結合層となっており，出力は 0 から 1 の値をとる．

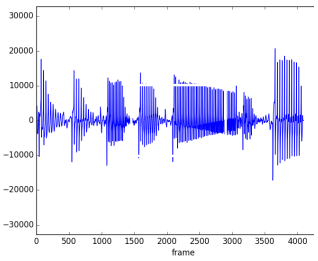

図 3 訓練データ 1

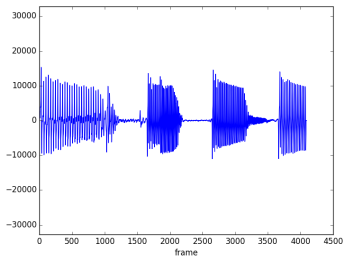

図 4 訓練データ 2

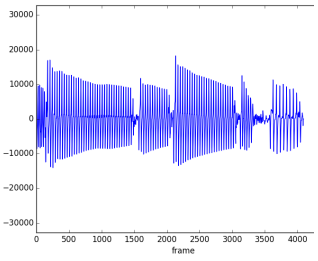

図 5 訓練データ 3

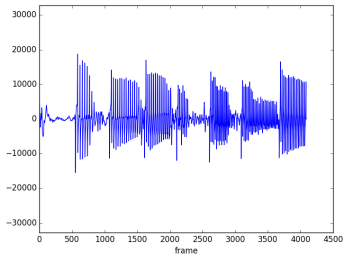

図 6 訓練データ 4

2.2 Generator

Generator のネットワーク構成は図 2 に示すとおりである．初めに，一様分布から生成した 100 次元ベクトル z を生成し，全結合層によって Discriminator の最後の畳み込み層の出力と同じ次元数に変換する．Generator の構造は基本的に Discriminator の構造と対照的になっており，逆畳み込み層 (転置畳み込み層，convolution transposed)，バッチ正規化層，活性化関数として Leaky Rectifierd Linear Unit (Leaky Relu) により構成されている．逆畳み込み層のパラメータはカーネルサイズ 5，ストライド 2，パディング 1 となっており，出力チャネル数は入力チャネル数の 2 倍となっている．最終層の出力は訓練データ同じ次元数となる．

2.3 最適化手法

最適化手法としては Adam [6] を用い，パラメータは $\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10e^{-8}$ とした．

3. 実験

3.1 データセット

データセットとして，エレキベースのループ音源 40 個を用意した．エレキベースのループ音源は比較的単純で，単音のフレーズが多いことから選んだ．しかし，GAN の学習は訓練データの数が増えれば増えるほど難しくなっていくことから，このうち 4 個のみを使用し，今後徐々に増やしていくこととした．今回使用した訓練データを図示したのが図 3，図 4，図 5，図 6 である．元のデータはサンプリングレートが 44.1 kHz，ビット深度が 16 bit であったが，学習を簡単にするためにサンプリングレートを 2,048 kHz とし，$(-1, 1)$ の 32 ビットの浮動小数点型で表した．

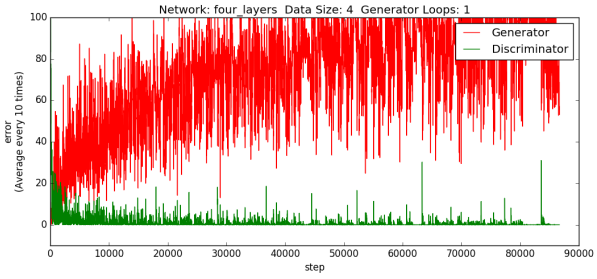

図 7 学習の過程をプロットした図．上が Generator，下が Discriminator の誤差を示している．

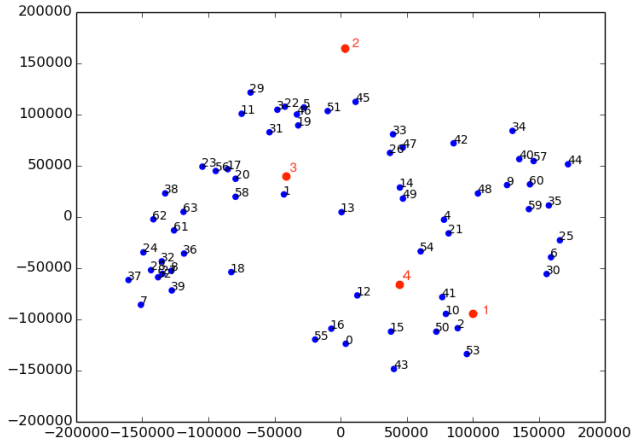

図 8 全データに関して，PCA を使って平面にプロットした図

3.2 訓練

学習の過程を確認するために，訓練を始める前にあらかじめ 100 個のベクトル z を生成しておき，100 エポックごとに検証を行った．学習に必要なエポック数はデータ数やパラメータの設定によって大きく変わるため，生成データに聴覚上の変化が無くなるまで行った．最終的には 50000 エポックの学習を行った．

4. 結果

図 8 は 4 つの訓練データと 100 個の生成データを主成分分析をしてプロットしたものである．この画像のうち，訓練データ 1 に最も近い生成データ 2 を比較していく．訓練データ 1(図 10) と生成データ 2(図 9) の画像を比較すると，生成データ 2 は右の丸の部分など，訓練データ 1 に非常によく似ているように見える．一方で図 11 の丸の部分に関しても類似していることが見受けられる．

これらから，この生成データは訓練データの一部を模倣しつつも，全く同一のものは生成していないということがわかる．これは他の生成データに関しても同様であった．また，聴覚上においてもそれらが認識できる他，低音域が強調されており，エレキベースの音を鮮明に聞くことができた．音質に関しても訓練データと遜色のないものであった．一方で，図 10 中の左の丸で囲まれた部分に関しては，音が重なりあって聞こえるように聞こえ，不自然に感じるものであった．それぞれの図を比較しても，他の部分と比較して他の訓練データの要素同士が比較的近いことから，模倣する区間を適切に学習する必要があると言える．

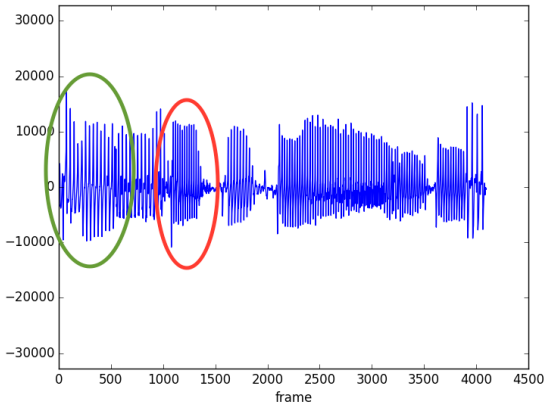

図 9 図 8 の生成データ 2

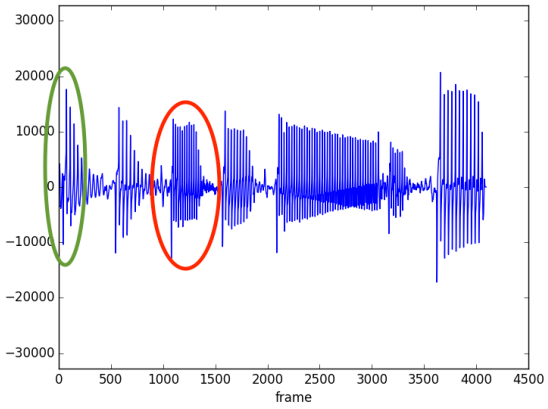

図 10 図 8 の訓練データ 1

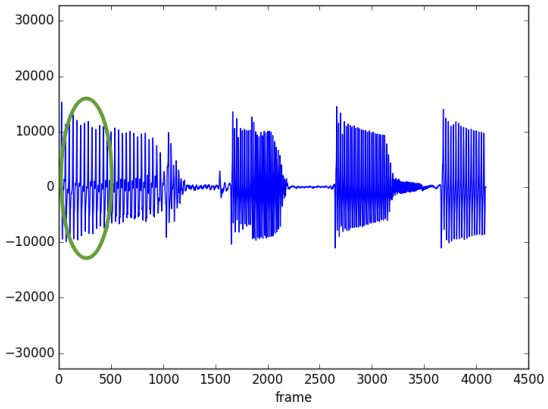

図 11 図 8 の訓練データ 2

5. 結論

本研究では，生の音データの自動生成に関する GAN の有用性を示した．生成データは訓練データに似ていながら

も，完全には同一ではないものを生成できた．また，それらは聴覚上でも訓練データと遜色ない程度に鮮明であることが確認することができた．

一方で，そもそもの訓練データの数や音質が低過ぎたことから，音色の違いを確認することができず，生成データの特徴を操作するという目的に関してはまだ達成できていない．これまでは学習の簡単化のために少ないデータ数，低い音質で行ったが，より多いデータ数，高い音質での学習，生成を実現する必要がある．これらを実現するためには，‘mode collapse’と呼ばれる問題を解決する必要がある．

‘mode collapse’は Generator の学習が進まなくなってしまう問題であり，GAN の大きな課題である．そのため，Salimans ら [7] や，Metz ら [8] などによって，これを避けるための研究が多く進められている．

今後の課題としては，これらの手法を導入していくことにより学習の安定化を実現することにより，より多くのデータセット、より高い音質での学習を行う．そして，任意の特徴を反映させたループ音源の生成を実現させる．

参考文献

[1] L. Hiller and L. M. Isaacson, *Illiac suite, for string quartet.* New Music Edition, 1957, vol. 30, no. 3.
[2] A. van den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. Senior, and K. Kavukcuoglu, “WaveNet: A Generative Model for Raw Audio,” sep 2016. [Online]. Available: http://arxiv.org/abs/1609.03499
[3] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Nets,” *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, 2014. [Online]. Available: http://papers.nips.cc/paper/5423-generative-adversarial-nets.pdf
[4] G. E. Hinton and R. R. Salakhutdinov, “Reducing the Dimensionality of Data with Neural Network,” vol. 313, no. July, pp. 504–507, 2006.
[5] A. Radford, L. Metz, and S. Chintala, “Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks,” pp. 1–16, nov 2015. [Online]. Available: http://arxiv.org/abs/1511.06434
[6] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” dec 2014. [Online]. Available: http://arxiv.org/abs/1412.6980
[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, “Improved Techniques for Training GANs,” jun 2016. [Online]. Available: http://arxiv.org/abs/1606.03498
[8] L. Metz, B. Poole, D. Pfau, and J. Sohl-Dickstein, “Unrolled Generative Adversarial Networks,” nov 2016. [Online]. Available: http://arxiv.org/abs/1611.02163