1E5-3

深層学習を用いた音の生成モデル

SNDGAN: The generative model of sounds using Generative Adversarial Networks

土井樹 *1 Itsuki Doi　小島大樹 *1 Hiroki Kojima　池上高志 *1 Takashi Ikegami

*1東京大学総合文化研究科
Graduate School of Arts and Sciences, University of Tokyo

Deep neural networks have been showing remarkable ability especially for images, not only in detection, but also in generation. However, no deep neural network systems for generating sound textiles have been proposed. Here we propose new sound generation system, SNDGAN, which can generate sound by learning input sound datasets. We accomplished this by converting sounds into image by invertible constant-Q transformation, which not only enable us to use powerful deep neural network system for image generation but also biologically plausible. We show some examples from this framework, and evaluate the capability of the system.

1. はじめに

機械学習と大量のデータによって生まれる「思考のパターン」が、我々人類とはどの程度異なった/同じものになるのか。この問題は、人工知能について考える際の、ひとつの重要な問いとなる。Alpha GO[Silver 2016] による対局は、人と機械学習では異なる思考パターンを示す、そのようなひとつの例と考えることが出来る。深層学習は、近年は特に我々の視覚野の認識能力と深層学習を様々な観点から比較する試みが増えている [Yamins 2016] 一方で、その生成モデルと脳の生成能力（記憶の呼び出し、想像）を比較するというアプローチの例は少ない。

生成モデルとしての深層学習は、近年非常に興味深い結果を出している。特に Goodfellow らが 2014 年に提案した Generative Adversarial Network(GAN) [Goodfellow 2014] により、それまで生成モデルとして知られていた Variational Autoencoder に比べ遥かに明瞭な画像を生成することが可能となり、その後も Deep convolutional generative adversarial network (DCGAN) [Radford 2015] 、Variational Autoencoder with GAN(VAEGAN) [Larsen 2015] など、様々なバリエーションが提案されている。

これまで、音声・音楽の分野においては音声認識についての研究は数多くなされている [Hinton 2012]。しかし視覚の研究と比較して、我々の聴覚野との比較また生成に関する研究は未だ数少ない。特に、音楽の構造ではなく音色そのものを生成する試みは殆ど存在していない。音楽において深層学習の研究が少ない理由のひとつとして、画像に比べ (タグ付けが行われたという意味で）良質なデータセットが多くないことが考えられる。特に音色 (sound textile) は抽象的に表現せざるをえない場合も多く、サウンドに対するタグ付けが画像に比べ難しい。

本研究では、音に対する我々の脳と深層学習を比較し音の知覚モデルを提案することを目標とし、その初めとしてタグを必要としない unsupervised learning の枠組みを用いる GAN の特徴を利用することで、音色を生成することが可能かどうかを検討した。これはサウンド生成において重要な問題を解くことに相当する。

連絡先: 土井樹: idoi@sacral.c.u-tokyo.ac.jp

2. Model

音をニューラルネットワークの入力する場合、i). 音源の各時刻での値をそのまま時系列データとして入力する。ii). 音をフーリエ変換などで周波数成分に変換して入力する。という 2 つの方法が考えられる。本研究では後者の方法を用い、ニューラルネットに DCGAN を用いた。

2.1 入力

DCGAN は画像生成のためのニューラルネットワークであり、これを用いて音を学習、生成させるためには、まず音を画像に変換する必要がある。この変換は、1) 逆変換可能であること（生成した画像から音に戻す必要が有るため)、2) 音の特徴が画像の特徴として表現されること、の 2 点を満たすことが求められる。

逆変換可能な画像への変換としては、フーリエ変換によるスペクトログラムへの変換が考えられる。しかし、フーリエ変換を用いた場合、周波数方向の軸は線形となるが、音の表現としては、周波数方向は対数スケールの表示が自然であることが知られている。例えば、オクターブが一つ上がるごとに周波数の値は 2 倍となることがその一例である。実際に、人の聴覚器官である蝸牛での coding も対数スケールになっている [Irino 1993]。従って、音の特徴を画像の特徴として表現するには適切でないと考えられる。

一方、周波数方向を対数スケールであらわす方法として、フーリエ変換をそのまま対数スケールで表示する、もしくは constant-Q 変換と呼ばれる手法で変換するといったことが可能であるが、いずれもそのまま逆変換することができず、ここでは用いることができない。

そこで本研究では、この 2 つの要件をみたすものとして、invertible constant-Q transform [Holighaus 2013] を用いることとした。この変換を用いると、入力した音が、横軸に時間、縦軸に周波数成分として表現され、周波数方向は対数スケールとなる。さらにこの変換の場合、非定常 Gabor 変換を用いることで、完全な逆変換が可能となっている。本研究ではこの invertible constant-Q transform を用いて、音を画像に変換し、入力データとして使用した。

また、データセットととしてロックミュージック (データセット A) とピアノ曲 (データセット B) を用いた。

2.2 ニューラルネットワーク: DCGAN

DCGAN は、Generative Adversarial Network(GAN) を画像生成に特化させるかたちで拡張した生成モデルである。GAN では学習データセットの分布を事前に与えず、分布の形状自体を Discriminator と Generator と呼ばれる学習機に学習させることで、学習データセットと見分けがつかないようなデータを生成する Generator を獲得していく。Generator には一様分布などからサンプルされた乱数 z が入力され、これを種として x を生成する。Generator の学習には Discriminator が用いられる。これは入力が学習データセット由来か、Generator の生成したデータかを判別する classifier となっている。DCGAN では、Discriminator は通常の Convolutional neural network (CNN)、Generator は z からスタートする逆方向の CNN を用いる。

本研究では Radford らのモデルを元に、より大きな画像を扱えるようにするために入力と出力の画像データを 64x64 から 96x96 に変更し、計算コストを抑えるためにいくつかの変更を行ったものを使用した。

2.3 出力

DCGAN の出力は、invertible constant-Q transform のスペクトログラムとなっている。この変換は、完全に逆変換可能であるため、出力をそのまま逆変換することで、音に変換される (図 1)。

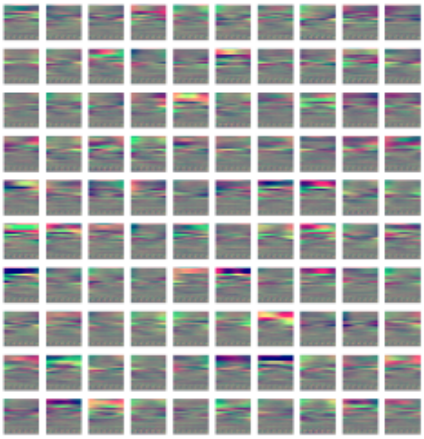

図 1: ニューラルネットワークから出力された画像のパターン例 (100 バリエーション)。x 軸が時間、y 軸が周波数を表す。

3. 解析

Generative Adversarial Network は単一の目的関数を最適化する問題となっていないため学習曲線を書くことが困難である。ここでは、96x96 のスペクトログラムを各 epoch 毎に 500 パターン出力させ、非線形次元削減法のひとつである t-SNE [Van 2008] によって二次元に次元削減し可視化した。

4. 結果

データセット A、B ともに 0epoch ではホワイトノイズ様の音が生成された。その後データセット A では、徐々にオリジナルデータに近づいていき最終 epoch では、入力データに含まれているボーカルやギターの音と極めて近い音が出力されるようになった。t-SNE による可視化においてもホワイトノイズから徐々に入力データに近づいていき 120 epoch 以降は入力データに極めて近い場所にデータ点が存在することが確認できた (図 2)。

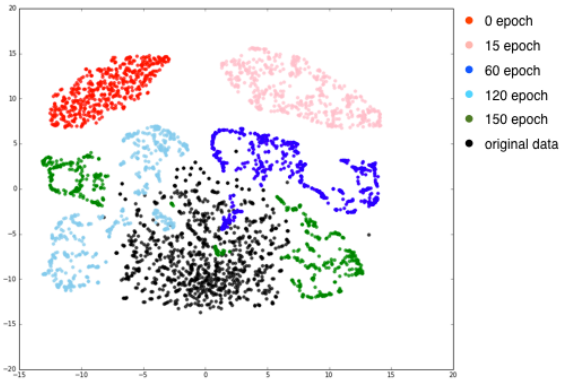

図 2: データセット A から生成された出力を t-SNE によって二次元に圧縮したもの。各点が 1 つのスペクトラム画像に対応し、色は epoch 毎に異なる。

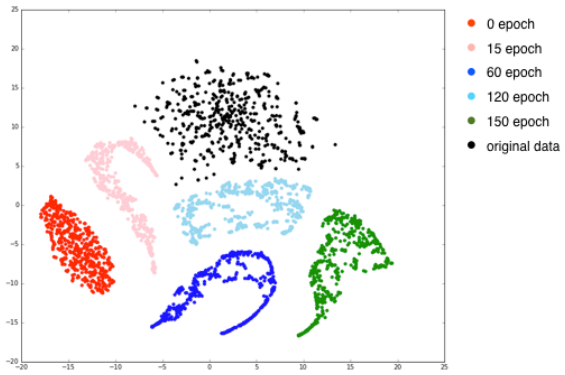

図 3: データセット B から生成された出力を t-SNE によって二次元に圧縮したもの。色などは図 1 に同じ。

一方、データセット B では、初期の epoch では入力データに含まれるピアノの音色に近づいていくものの、epoch を重ねるとノイズ様の音とピアノに近い音色を行き来し、最終 epoch ではどういった乱数 z から音を生成してもノイズを生成するようになった。t-SNE による可視化においても epoch 120 では入力データに近い音を生成しているが、最終 epoch では入力データから離れた場所にプロットが存在することが確認できた (図 3)。

5. 議論

本研究により、DCGAN の枠組みを用いることで音でも入力データと非常によく似た、しかし入力とは異なった出力を得

られることが明らかとなった。しかし、いかなる場合でも良い出力が得られわけではなく、特にデータセット B の結果やその他の実験により、今回のパラメータ設定ではアコースティック楽器が多く含まれる音源を入力とした場合は、epoch を重ねることが必ずしも良い結果を得るとは限らないことが明らかとなった。このデータセット A と B の学習結果の差は、パラメータ設定に加え画像に変換された入力データの特徴を DCGAN 内の convolutional neural net が捉えることが出来るか否かに依存していると考えられる。

会場では、より幅広いデータセットを用い Z 空間上にどのように音色或いは音楽の構造が配置されうるのかを解析した結果について、またサウンドデータをスペクトログラムに変更せず、直接ニューラルネットワークに入力した場合についても実験を行い比較検討することで、音に対する人の知覚以外の方法が深層学習によって可能であるか、可能な場合どのような知覚モデルを立てることが出来るかについて議論したい。

参考文献

[Silver 2016] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... and Dieleman, S. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484-489.

[Yamins 2016] Yamins, D. L., and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience, 19(3), 356-365.

[Goodfellow 2014] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... and Bengio, Y. (2014). Generative adversarial nets. In Advances in Neural Information Processing Systems (pp. 2672-2680).

[Radford 2015] Radford, A., Metz, L., and Chintala, S. (2015). Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

[Hinton 2012] Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A. R., Jaitly, N., ... and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. Signal Processing Magazine, IEEE, 29(6), 82-97.

[Larsen 2015] Larsen, A. B. L., Sønderby, S. K., and Winther, O. (2015). Autoencoding beyond pixels using a learned similarity metric. arXiv preprint arXiv:1512.09300.

[Holighaus 2013] Holighaus, Nicki, et al. "A framework for invertible, real-time constant-Q transforms." Audio, Speech, and Language Processing, IEEE Transactions on 21.4 (2013): 775-785.

[Irino 1993] Irino, Toshio, and Hideki Kawahara. "Signal reconstruction from modified auditory wavelet transform." IEEE Transactions on Signal Processing 41.12 (1993): 3549-3554.

[Van 2008] Van der Maaten, L., and Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(2579-2605), 85.