ニューラルネットワークによる
音楽の自動変換

陶山大輝
所属：教養学部後期課程学際科学科総合情報学コース
学籍番号：08-192021
指導教員：金子知適准教授

2020 年 12 月 29 日

概 要

目 次

第１章　はじめに

第2章　背景

2.1　ニューラルネットワーク

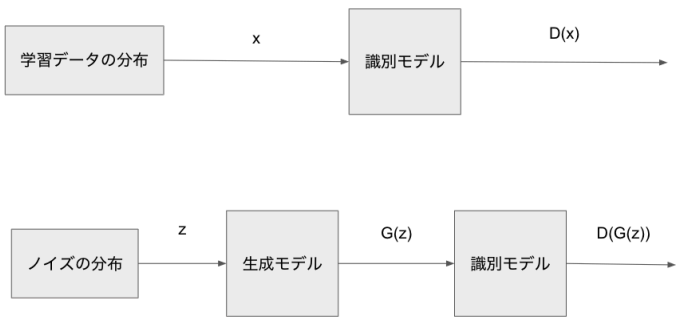
ニューラルネットワークとはニューロンとニューロン間のシナプスによる結合で形成される脳のネットワークを模した数理モデルである。入力層と出力層を持ち、シナプスの結合強度を変化させることで問題に最適なネットワークを構成することを目標とする。

2.2　ディープラーニング

ディープラーニングとは層をより深くしたニューラルネットワークを用いた機械学習の手法である。多大な計算資源を必要とするが、GPU を含む計算機の性能の向上により実用的な手法となった。

また、層を増やすと勾配の減衰による勾配消失や訓練データへの最適化による過学習などの問題が発生するが、前者の場合は活性化関数に ReLU 関数を用い後者の場合は汎化性能を測定することで避けることができる。

2.3　GAN

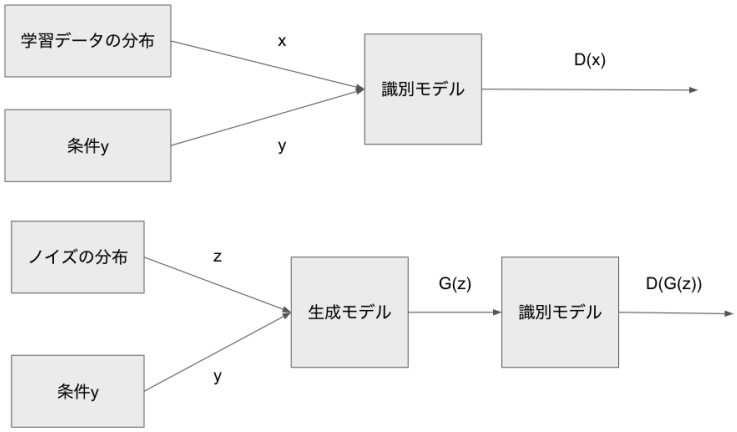


GAN(敵対的生成モデル) とは生成モデルと識別モデルが競合して学習を行うディープラーニングのモデルである。生成モデルは訓練データの分布を捉えようとし、識別モデルは訓練データである確率を推定する。

具体的には、訓練データ x の分布を p_{data}, 生成モデルの入力のノイズ z の分布を p_z, ノイズ z を元にデータを生成する関数を G, 生成モデルのデータではなく訓練データである確率を返す関数を D とした時、式 2.1 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [1]。

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}}(\boldsymbol{z})[\log(1 - D(G(\boldsymbol{z})))] \tag{2.1}$$

2.4 CGAN



CGAN(条件付き敵対的生成モデル) とは条件付きの GAN である。具体的には y という条件下で式 2.2 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [2]。

原著論文では訓練時に MNIST の数字のラベルを条件として与えることで特定の数字のラベルの画像を生成するモデルを実装している。

$$\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x} \mid \boldsymbol{y})] + \mathbb{E}_{\boldsymbol{z}\sim p_z(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z} \mid \boldsymbol{y})))] \tag{2.2}$$

2.5 pix2pix

pix2pix とはある条件下で画像間の変換を行う GAN である。先程の CGAN を元に U-Net と PatachGAN を組み合わせたネットワークになっている。U-net とは…,PatchGAN とは…。(ここは後々調べて書く)

具体的には、画像 x を条件として式 2.5 を生成モデルは最小化し識別モデルは最大化をすることを目標として学習を行う [3]。

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{x,y}[\log D(x, y)] + \mathbb{E}_{x,z}[\log(1 - D(x, G(x, z)))] \quad (2.3)$$

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{x,y,z}\left[\|y - G(x, z)\|_1\right] \quad (2.4)$$

$$\mathcal{L}_{cGAN}(G, D) + \lambda\mathcal{L}_{L1}(G) \quad (2.5)$$

第3章　実験の詳細

3.1　データセット

楽譜作成ソフトの MuseScore を利用して国際の階名表記で A0 から C8 の音を wav 形式で 52 音生成した。また、この 52 音は一般的な 88 鍵のピアノで出すことのできる全音階の音であり、その音域は人間が音程として聞き分けることのできる限界の音域として選んだ。

ここで、52 音については人間の耳で聴き分けられる程に十分に音色が異なると考えられるエレキギターとハープを選んだ。

そして、それぞれの音については四分音符を生成したのち 1 秒の長さに揃えている。これらの音は全てサンプル周波数が 44.1kHz で量子化ビットは 16 ビットである。

3.1.1　データ拡張

先程の 52 音のみではデータ数が十分ではないのでデータ拡張を行った。具体的には正規化したデータを $\alpha(0 < \alpha < 1)$ 倍した後に $[0, 1-\alpha)$ の一様乱数を加えることにした。また、今回は $\alpha = 0.01$ に固定した。

3.2　評価

第4章　提案手法

第5章　まとめ

謝辞

関連図書

[1] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[2] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets, 2014.

[3] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks, 2018.