

敵対的 VAE による音楽のスタイル変換

陳 毅文[†] 澤田 秀之[‡]早稲田大学先進理工学部[†] 早稲田大学理工学術院[‡]

1. はじめに

本研究では、音楽のスタイル変換に応用する敵対的 VAE (VAEGAN モデル) を提案し、これによって変換された音楽の評価、原曲との比較を行う。

2. 音楽のスタイル変換

楽曲には、クラシックやポップスなどのスタイルがある。これを他のスタイルに変換することを、スタイル変換と呼ぶ。

3. VAEGAN

本研究で構築した VAEGAN モデルを図 1 に示す。変分オートエンコーダ (VAE) のデコーダーの後ろに、敵対生成ネットワーク (GAN) を結合している^[1]。本モデルにおいて、中間の部分は VAE の Decoder であり、また GAN の Generator でもある。

4. モデル構造

本研究では、音楽の MIDI ファイルから Pitch, Instrument, Velocity の三つの音楽情報を抽出し、行列で表現する。MIDI の音高は 128 段階あるが、本研究では、よく現れる 24 (C₁) から 84 (C₆) までの 60 個の音高を取り出す。楽器に関しては、n-hot ベクトルを使用し、楽曲で使われる楽器を表すところを 1 とし、他を 0 とする。Velocity に関しては、MIDI の音符の強度に 128 段階 (0~127) あり、これを [0, 1] に正規化する。

入力は、三つの多次元行列であり、音高行列 $P \in \{0, 1\}^{P \times B \times T}$ 、楽器行列 $I \in \{0, 1\}^{I \times T}$ 、強度行列 $V \in [0, 1]^{P \times B \times T}$ としている。P は 60 で、音高の数を表す。B は 64 で、時間ステップ数を表す。T は楽曲のトラック数を表す。エンコーダとデコーダーには、ゲート付き回帰型ユニット (GRU) と全結合層を使用した。GAN の識別器^[2]には、全結合層とバッチ正規化を使用している。そして、

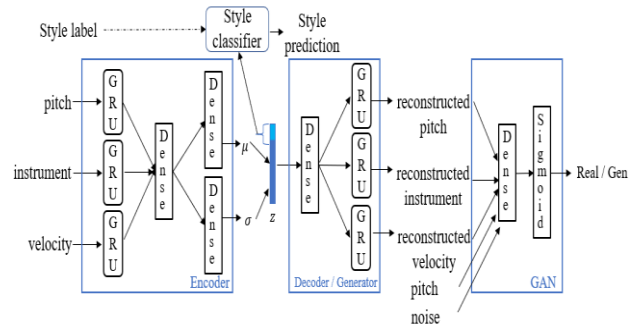


図 1 VAEGAN モデル

表 1 GAN の構造

Discriminator	
512 fully-connected,	BNorm, ReLU
256 fully-connected,	BNorm, ReLU
128 fully-connected,	BNorm, ReLU
1 fully-connected,	sigmoid

スタイル変換を実現するために、エンコーダにソフトマックス損失関数のスタイル分類器を用いる。この分類器は、潜在変数 z の上から k 次元 (本研究では二つのスタイルを相互に変換するため、 k を 2 とする) の部分を使い、エンコードしたスタイルラベル情報を残す。これにより、スタイルを S_i から S_j に変換する場合、 z^i と z^j の値を交換すれば良い。

4. 1 モデルの損失関数

まず VAE の損失関数には、再構成損失と KL 損失を設定する。再構成損失は元のデータと生成されたデータの間の差を計算する。データサンプルを x とし、潜在変数を z とすると、

$$L_{recon} = -E_{q(z|x)}[\log p(z|x)] \quad (1)$$

となる。また、

$$z \sim Enc(x) = q(z|x) \quad (2)$$

$$x \sim Dec(z) = p(x|z) \quad (3)$$

KL 損失は二つの分布間の距離を計算する。

$$L_{kl} = D_{KL}(q(z|x)p(z)) \quad (4)$$

D_{KL} は Kullback-Leibler divergence を表す。

Music Genre Transfer using VAEGAN

Y. Chen, School of Advanced Science and Engineering,
Waseda University

H. Sawada, Faculty of Science and Engineering, Waseda
University

以上より，VAE モデルの損失は，

$$L_{VAE} = L_{recon} + L_{kl} \quad (5)$$

となる．

GAN の損失関数にはクロスエントロピーを設定する．

$$L_{GAN} = \log(Dis(x)) + \log(1 - Dis(\tilde{x})) + \log(1 - Dis(x_{noise})) \quad (6)$$

$\tilde{x}$: 再構成された x

x_{noise} : 生成器がノイズから生成したもの

VAEGAN は VAE と GAN を結合したもので，識別器は VAE によって生成されたものを識別する必要があるため，一般の GAN に対して式(6)中の右辺第二項を付け加えた．

4. 2 トレーニングプロセス

ここで，VAEGAN モデルのトレーニングアルゴリズムを下記に示す．

VAEGAN アルゴリズム
各ネットワークを初期化
repeat
$X \leftarrow$ ミニバッチからランダムに取る
$Z \leftarrow Enc(X)$
$L_{kl} \leftarrow D_{KL}(q(Z X)p(Z))$
$\tilde{X} \leftarrow Dec(Z)$
$L_{recon} \leftarrow -E_{q(Z X)}[\log p(z x)]$
$noise \leftarrow$ ランダムに発生させる
$X_{noise} \leftarrow Dec(noise)$
$L_{GAN} \leftarrow \log(Dis(x)) + \log(1 - Dis(\tilde{x})) + \log(1 - Dis(x_{noise}))$

5. 実験

本研究では，クラシック曲，ポップス曲から合計 480 曲をトレーニングデータとしてモデルに学習させた．トレーニング回数を 200，バッチサイズを 64 としている．記録した損失の結果から滑らかに収束することが分かった．

変換後の曲に関しては，良好にスタイルが変換されており，主旋律が連続的で原曲が十分に解る程に保たれていることが分かった，その一方で，音符が断続的で，変換が失敗した例もあった．そこで，トレーニングデータから各スタイル 50 曲を取り出し，成功の割合とスタイル分類による変換前後のスタイル識別率を表 2，3 にまとめた．成功となる基準としては，変換後の楽曲に明らかな中断する部分がなく，音符が連続的に並んでいることとした．変換前後のスタイル識別率については，変換後の楽曲が元のスタイルであるとは認識されないことが目標なので，

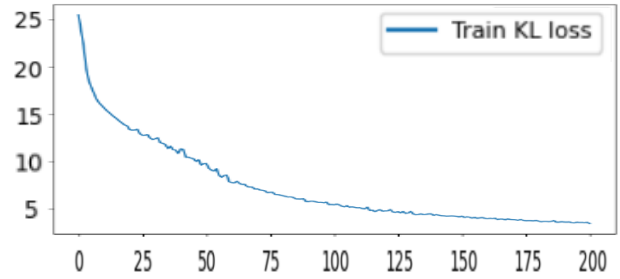


図 2 KL 損失の推移

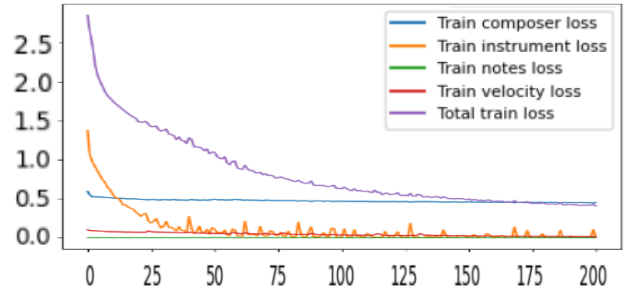


図 3 再構成損失の推移

表 2 変換後の曲の成功の割合

クラシックから ポップスへ	ポップスから クラシックへ
88%	86%

表 3 スタイル変換のパフォーマンス

テストデータ	クラシック&ポップス	
変換前	変換後	確率の差
83%	46%	37%

前後の認識率の差が大きいほど，モデルのパフォーマンスが良いと言える．

6. まとめ

本研究では，音楽のスタイル変換に VAEGAN モデルを提案した．その結果，安定して学習でき，滑らかに損失値が収束した．変換後の曲に関しては，良好にスタイルが識別されており，主旋律も保たれていることが分かった．その一方で失敗の例も見られたので，今後その原因を調べていく．

参考文献

- [1] Gino Brunner, Andres Konrad, Yuyi Wang, Roger Wattenhofer, “MIDI-VAE: modeling dynamics and instrumentation of music with applications to style transfer”, CVPR, pp.2-5, 2018.
- [2] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, Ole Winther, “Autoencoding beyond pixels using a learned similarity metric”, PMLR, pp.1-3, 2016.