1T-04

敵対的生成ネットワークによる音楽にあった画像の生成

片寄凱 † 清水郁子 †

† 東京農工大学 情報工学科

1 はじめに

近年，Youtube や SNS の普及によって個人で動画を作成し発信することが一般的となったが，高品質な動画制作には時間と技術を要する．動画制作に適用可能な技術に関連し，テキストや音声，環境音から画像を生成する先行研究は多く存在する．これらの研究では，画像と一意に対応するテキスト記述や音声データをペアとしたデータセットを用いて画像生成モデルを学習し，画像生成を行う．しかし楽曲データの場合，楽曲を構成する要素が BPM やリズム，使用されている楽器，調性など多岐に渡り，聴いた際のイメージがそれぞれ個人で異なることから，楽曲データと画像を一対一に対応づけたデータセットを作成することが困難であるため，楽曲データから画像を生成する研究はほとんど存在しない．

本手法では，楽曲から想起される感情推定に用いられる特徴量を用い，楽曲の感情カテゴリと画像データを対応づけた独自のデータセットを学習に用いることで，楽曲に合った画像を生成する手法を提案する．楽曲からの特徴抽出は，MidiBERT [1] と呼ばれるモデルを用いる．画像生成には，StackGAN [2] と呼ばれる敵対的生成ネットワーク (GAN) の一種を用いる．StackGAN は，画像とその画像を描写する複数のテキスト表現のデータセットを用いて学習し，入力されたテキストにあった写実的な画像を生成する既存の画像生成モデルである．このネットワークを用い，感情ベースで構築した独自の楽曲データセットで学習を行うことによって，楽曲の印象によって異なる画像生成を実現する．

2 提案手法

本手法では，画像に対し，その画像の印象にあう楽曲を対応付けて学習することで，入力した楽曲にあう画像を生成する手法を提案する．画像生成のネットワークには StackGAN を用いるが，StackGAN では 1 枚の画像を描写する複数のテキストから抽出した特徴を Conditioning Augumentation という手法により学習することが特徴である．本手法では，楽曲から抽出された特徴量が類似した複数の楽曲を同じ画像に対応付け，StackGAN [2] により学習する．

Generating images which match music by Generative Adversarial Networks
Gai KATAYOSE†and Ikuko SHIMIZU†
†Faculty of Engineering, Tokyo University of Agriculture and Technology,184-8588, Tokyo, Japan
s183777r@st.go.tuat.ac.jp

2.1 楽曲データからの特徴抽出

楽曲のデータは MIDI データを用い，MIDI データからの特徴抽出に MidiBERT-Piano [1] を用いる．MidiBERT では，MIDI データから想起される感情のラベルを EPOPIA [3] と呼ばれるデータセットにより事前学習し，感情分類が行えるモデルが公開されている．本手法では，感情分類タスクを解くように事前学習されたネットワークの中間層の値を取り出し，楽曲特徴量とする．この楽曲特徴量は，入力した楽曲から想起される感情に関する情報を多く含むと考えられる．感情分類の感情ラベルは，Valance（感情値）と Arousal（覚醒度）の 2 次元空間で概念化し，HVHA(high valence high arousal),HVLA(high valence low arousal),LVHA(low valence high arousal),LVLA (low valence low arousal) の 4 クラスである．

2.2 画像生成ネットワーク

画像生成には，Text to Image タスクを行うための画像生成ネットワークである StackGAN [2] を用いる．これは段階的に画像生成を行うネットワークであり，1 段階目に低解像度の画像を生成し，2 段階目に前段の生成画像を基にして高解像度の画像を生成する．

StackGAN では，1 枚の画像に対する複数の特徴記述をペアとしてネットワークの学習を行う Conditioning Augmentation を行うことで，少数の画像と特徴のペアから効率よく学習が可能である．1 枚の画像に対し，複数の楽曲の特徴を対応づける方法は次節に示す．

2.3 データセットの構築

StackGAN の学習のために，まず，EMOPIA データセット [3] の各楽曲に付与された 4 つの感情ラベルに対して，Places365 データセット [4] の異なるラベルの画像を表 1 のように対応させた．

表 1: 感情ラベルと画像ラベルの対応関係

感情ラベル	HVHA	LVHA	LVLA	HVLA
画像ラベル	desert sand	forest road	snowfield	waterfall

次に，対応するラベル内に含まれる画像と楽曲を対応づける．1 枚の画像に複数の類似した楽曲の特徴量を対応づけるために，同じラベルを持つ楽曲の特徴空間においてクラスタリングを行う．各クラスに属する楽曲数が全て N 個ずつになるように，k-means 法により楽曲の特徴空間でクラスタリングを行い，各クラスに対して，対応するラベルを持つ画像をそれぞれランダムに割り当てる．これにより，画像 1 枚に対して N 個の楽曲特徴量がペアとなったデータセットを作成することができる．

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.

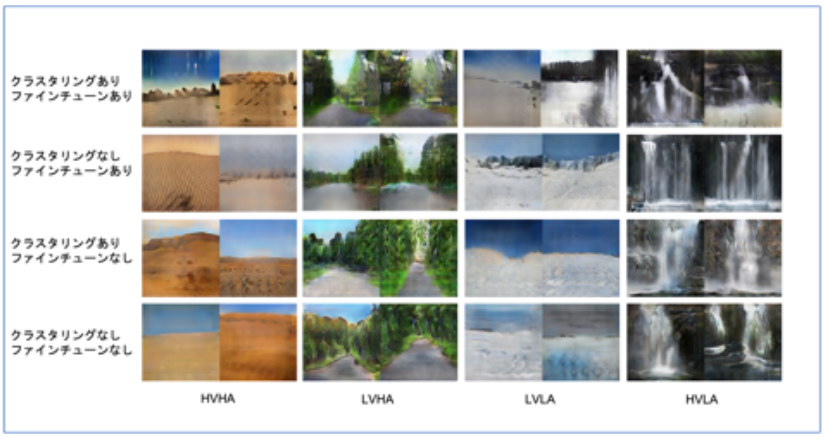

図 1: ファインチューニングあり・なし，クラスタリングあり・なしで学習したネットワークの画像生成例

3 実験

3.1 学習に用いたデータ数

実験では，$N = 5$ とし，表 2 のように，各感情ラベルごとに 1 枚の画像と 5 個の楽曲特徴量のペアを生成した．

表 2: 学習に用いるデータセットのペア数

感情ラベル	HVHA	LVHA	LVLA	HVLA
ペア数	104	134	75	89

実験では，1 段階目，2 段階目のネットワーク共にバッチサイズ 64，600 エポックの学習を行い，初期学習率 0.0002，100 エポックごとに学習率が 1/2 となる Adam オプティマイザを使用した．また，学習済みモデルを用いたファインチューニングを行う場合と，スクラッチから学習する場合の 2 つについてもそれぞれ実験を行った．ファインチューニングの学習済みモデルには，StackGAN の論文内で結果が示されている CUB データセット [5] によって学習を行なったものを用いた．以下の図 1 に，楽曲特徴量の感情ラベルごとの画像生成例を示す．

図 1 のように，それぞれ対応した風景画像が生成できることが分かる．ただし，一部の生成画像に縞模様のようなノイズが現れたり，風景画像自体を生成できていない場合も確認できた．またファインチューニングを行った場合，生成画像の多様性が著しく減少してしまう場合があった．

3.2 生成画像の評価

生成画像の品質，入力した楽曲特徴量に適した画像が生成されているかという 2 つの指標に関して，217 個のテストデータを用いて評価を行う．前者は，敵対的生成ネットワークの生成画像の評価に用いられる FID スコアで評価する．後者は，imagenet で学習済みの ResNet18 を，本実験で学習に用いた 402 枚の画像で転移学習したネットワークを用いて評価する．楽曲からの生成画像を ResNet18 に入力し，楽曲特徴量の感情ラベルごとに対応づけた画像ラベルと識別される確率 (正解率) を算出する．表 3 に結果を示す．

正解率はどれも 5 割程度以下と低い数値となった．スクラッチからの学習では，クラスタリングありの場合に正解率が高くなった．ファインチューニングあり，クラスタリングありの場合に 0.25 と正解率が低いのは，生成画像の極端な多様性の低下によるものだと考えられる．

表 3: 生成画像の FID と正解率

ファインチューニング	クラスタリング	FID	正解率
なし	なし	200.9	0.38
なし	あり	203.9	0.47
あり	なし	200.3	0.51
あり	あり	191.5	0.25

FID スコアは，ファインチューニングとクラスタリングを合わせて用いた際に最も高かった．通常，敵対的生成ネットワークの学習には大量のデータが必要であるが，本研究では画像と楽曲のペアが 402 ペアと非常に規模の小さいデータセットを用いているため，2 つを組み合わせた際に最も FID が向上したと考えられる．

4 まとめ

本稿では，楽曲から想起される感情推定に用いられる特徴量を用い，楽曲の感情カテゴリと画像データを対応づけた独自のデータセットを学習に用いることで，楽曲に合った画像を生成する手法を提案した．楽曲データから特徴抽出を行うモデルと，敵対的生成ネットワークを用いた画像生成モデルを組み合わせ，楽曲の印象によって異なる画像を生成する手法を示した．学習のために，楽曲から想起される感情をベースとしたデータセットの作成方法を提案した．今後の課題として，データセットのデータ数を増やすことが考えられる．

参考文献

[1] Yi-Hui Chou, et al. "MidiBERT-Piano: Large-scale Pre-training for Symbolic Music Understanding." arXiv:2107.05223
[2] Han Zhang, et al. "StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks" Proc. ICCV 2017
[3] Hung, Hsiao-Tzu et al. "EMOPIA : A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation", Proc. Int. Society for Music Information Retrieval Conf. 2021
[4] Zhou, Bolei et al. "Places: A 10 million Image Database for Scene Recognition", IEEE Trans. PAMI, 40(6), pp.1452–1364, 2017
[5] P. Welinder et al. "Caltech-UCSD Birds 200", California Institute of Technology 2010

Copyright ©2022 Information Processing Society of Japan.
All Rights Reserved.