Text2Imageモデルを利用した楽曲歌詞に基づくサムネイル画像生成

佐々木翔一† 牛尼 剛聡††

† 九州大学大学院 芸術工学府 〒 815–8540 福岡県福岡市南区塩原 4–9–1
†† 九州大学大学院 芸術工学研究院 〒 815–8540 福岡県福岡市南区塩原 4–9–1
E-mail: †sasaki.shoichi.896@s.kyushu-u.ac.jp, ††ushiama@design.kyushu-u.ac.jp

あらまし 音楽ストリーミングサービスの配信楽曲に付与されているサムネイル画像は，視覚的に短時間で内容を把握することができるため，楽曲の選別に大きく貢献している．一方で，同一アルバム内の楽曲にすべて同じジャケット画像やアーティスト画像が割り振られる場合には，それらが個々の楽曲内容を適切に表現していない場合がある．そこで，本研究では歌詞情報と Text2Image モデルを用いて，楽曲に適したサムネイル画像を自動生成する手法を提案する．この際，歌詞全文を直接 Text2Image モデルに入力した場合，長文であるため満足な画像を得ることは難しい．そこで，本研究では歌詞の意味や感情を分析し，画像生成に適した入力テキストを再生成することで画像を獲得する．

キーワード 音楽，画像生成，Text2Image，情報抽出，テキスト感情分析

1 はじめに

近年，音楽を聞く手段として，音楽ストリーミングサービスを利用する人が増えている．日本レコード協会が行った調査によると，2012 年から 2021 年の 10 年間における音楽配信売上実績のうち，ストリーミングの占める割合が急増している．

多くの音楽ストリーミングサービスでは，個々のコンテンツに対してサムネイル画像が付与されている．サムネイル画像は，実際に映画を視聴したり音楽を聴取したりするのに比べて，視覚的に短時間で内容を把握することができる．そのため，ユーザのコンテンツに対する理解支援やユーザ体験の向上を目的とした，サムネイル画像およびストリーミング中の提示画像の最適化について盛んに研究が行われている [1] [2]．しかし，これらの手法では楽曲の特徴と画像の特徴を事前に紐付け，そこから得られた関係性をもとにサムネイル画像を生成しているため，学習過程で用いられたデータと大きく異なる特徴が入力されたときに，適切な画像を生成することが難しい．そのためこれらの手法は，生成画像の多様性が高くない．

一方，近年テキストを入力とし，その文の内容を表す画像を生成する Text2Image モデルが盛んに研究されており，「アボカドの形をした椅子」「写実的なスタイルの馬に乗った宇宙飛行士」など，教師データに含まれていない複雑な内容に対しても高水準な画像を生成できるようになった [3] [4] [5]．そこで，本研究では楽曲におけるテキストデータである歌詞に着目し，これを Text2Image モデルに入力することで，楽曲に適したサムネイル画像を自動生成することを目指す．

Text2Image では，画像と画像の内容を説明したテキストの関係を学習しているため，単一単語ではない複雑な文章を表す画像も生成することができ，生成画像の多様性という点において従来手法よりも秀でていると考えられる．

一方，Text2Image モデルに入力するテキストは「アボカドの形をした椅子」「写実的なスタイルの馬に乗った宇宙飛行士」など，比較的短く具体的な文章が想定されているため，歌詞のような長文をそのままの状態で入力したとしても，楽曲に適したサムネイル画像を出力することは難しい．そのため，より良い画像を生成するための高品質なテキストについて検討するプロンプトエンジニアリングが重要視されている．

以上より，本研究では楽曲の歌詞を分析し，Text2Image モデルによりサムネイル画像を生成するためのプロンプトを自動抽出するための手法について検討する．具体的には，楽曲内容を代表する比較的短文のテキストと，そのテキストをもとにサムネイル画像を生成した際に，画像の雰囲気を決定づけるスタイルと印象形容詞を抽出する．

まず，生成画像の内容を決定するテキストについては，歌詞全文を，生成する画像の内容を示す具体的で短い文章に変換する必要がある．そこで，本研究では楽曲歌詞における印象的フレーズに着目する．印象的フレーズはその楽曲を象徴する可能性が高いため，楽曲に適したサムネイル画像を生成する際に，画像の内容を示す短い文章として利用することができると考える．

楽曲の歌詞から，印象的フレーズを抽出することを目指した例として，山西らの研究では，楽曲におけるフレーズの繰り返しが印象に大きな影響を与えていることを，被験者実験の結果から報告している [6]．ここで言う繰り返しとは，同じ文字列の繰り返しという意味であるが，本研究ではこれに加えて，フレーズの意味的な繰り返し構造についても，印象的フレーズに影響を与えているという仮説をたてた．

楽曲には作者が一曲全体を通して伝えたいテーマがあり，この作者が表現したいテーマと意味的に最も近いフレーズが印象的フレーズになると考えられる．そして，作者が表現したいテーマは楽曲内のフレーズすべてに対して意味的な影響を与えていると考えられるため，楽曲のテーマと意味的に近い印象的フレーズは，その他のフレーズに対する類似度が大きいと考えられる．つまり，各フレーズのその他のフレーズに対する意味的な重畳性を確認すれば，印象的フレーズを推定することが可

能であると考えられる．また，このとき抽出した印象的フレーズは，作者が伝えたいテーマと意味的に近いため，楽曲にふさわしい画像を生成するためのテキストとしても利用可能である可能性がある．

次に画像の雰囲気を決定づけるスタイルと印象形容詞を抽出するためには，テキストとそのテキストに対して人間が抱く感情の関係を分析する必要がある．Cong らは，Bi-DLSTM に Attention 機構を加えることで，楽曲歌詞に付与された印象ラベルを予測する手法を提案している [7]．一方，本研究ではサムネイル画像生成のプロンプトに適したスタイルおよび印象形容詞を予測する必要がある．そのため，抽出する情報は画像生成に用いた際に，その画像を見て感じる印象に与える影響が大きいものが望ましい．しかし，Cong らの手法では楽曲歌詞を聴取した際に感じる印象を予測することを目指しているため，予測した印象語が楽曲にふさわしいサムネイル画像を生成する上でのプロンプトに適していることは示されていない．

そこで，本研究では絵画に対して抱いた印象とその印象を抱いた理由を説明したテキストを対応付けている ArtEmis [8] というデータセットを利用する．ArtEmis データセットでは，絵画を見た際に人間が抱く印象形容詞のラベルとそのように感じた理由を絵画の表現特徴に基づき述べた文章との関連性が対応付けられている．このデータセットを用いて絵画のスタイルおよび絵画に対して抱いた印象とその印象を頂いた理由との関係性をそれぞれ BERT に学習させ楽曲歌詞のスタイルおよび印象形容詞を予測するといったアプローチをとる．

本研究では，ArtEmis データセットと BERT を用いて絵画の表現特徴と人間が絵画を見た際に抱く印象形容詞との関係を予測することができれば，楽曲歌詞の雰囲気を表す形容詞を抽出できるという仮設を立てた．これを実現するために日本語学習済み BERT を ArtEmis データセットで Fine-Tuning するというアプローチを取る．

本研究では印象的フレーズと印象形容詞を抽出した後，Text2Image モデルの一つである Stable Difusion [5] に入力し，サムネイル画像を生成する．

本論文の貢献は以下の通りである．

- Text2Image モデルを用いて楽曲のサムネイル画像を生成するためのプロンプトを，歌詞から自動抽出する．
- 楽曲歌詞の意味的な重畳性をもとに，印象的なフレーズを抽出する．
- 人間が絵画に対して抱く印象をもとに，楽曲歌詞からその楽曲にふさわしい画像のスタイルや印象を決定するための形容詞を抽出する．

本論文の構成は次の通りである．2 章では関連研究について述べる．3 章では提案手法について述べる．4 章では実験と考察について述べる．5 章では本論文のまとめを述べる．

2 関 連 研 究

2.1 サムネイル画像の自動生成

映画配信サービスや音楽ストリーミングサービスでは，個々のコンテンツに対してサムネイル画像が付与されている．サムネイル画像は，実際に映画を視聴したり音楽を聴取したりするのに比べて，視覚的に短時間で内容を把握することができる．そのため，ユーザのコンテンツに対する理解支援やクリック確率の向上を目的として，サムネイル画像の最適化について活発に研究が行われている．

たとえば，映画配信サービスの Netflix は，ユーザの視聴履歴と映像内から抽出した映像的特性に基づきユーザの好みの傾向を予測することで，各ユーザに適したシーンのフレームを自動選択し，サムネイル画像を生成する手法を提案している [9]．一方，音楽は音コンテンツであるため，それ自体に内容を表す視覚的な要素が含まれていない．そのため，Netflix のようなシーン選択を利用した手法を用いることができず，楽曲から抽出した情報をもとに適切なサムネイル画像を自動生成・選択するというアプローチを取る必要がある．

楽曲情報をもとに，その楽曲に合う適切な画像の生成・選択を目指した研究として，梅村らは画像に付与した語ラベルと楽曲との共起性を Word2Vec により計算し，曲の流れにあったスライドショーを提示する手法を提案している [2]．また，Qiu らは短時間フーリエ変換（STFT）した楽曲音源を CNN および LSTM を用いたモデルに入力することで，楽曲内容を表現することに適した特徴量を獲得し，その後 DCGAN と組み合わせることで，楽曲の雰囲気に合うサムネイル画像を自動生成する手法を提案している [1]．しかし，梅村らの手法では，あらかじめ存在する画像の中から適切なものを選択するため，対象楽曲の内容に近い画像が存在しない場合，適切なサムネイル画像を選択できない可能性がある．また，Qiu らの手法では，事前に sky，water，moutain，desert という 4 つのテーマを表す楽曲の特徴量と画像とを紐付けるような学習を行っている．そのため，これら 4 つのテーマのいずれかに関する楽曲の場合適切な画像を生成することができても，全く未知の楽曲であった場合，適切な画像を生成できない可能性がある．

このように，既存の手法では楽曲の特徴と画像の特徴を紐付け，そこから得られた関係性をもとにサムネイル画像を生成・選択しているため，事前情報として保有していない内容の入力があったときに，適切な画像を生成することが難しい．そのためこれらの手法は，一つ一つの楽曲に適したサムネイル画像を出力するという，生成画像の多様性が求められるタスクにおいて問題があると考えられる．

2.2 Text2Image とプロンプトエンジニアリング

近年，テキストを入力とし，その文の内容を表すような画像を生成する Text2Image モデルが盛んに研究されている．DALL-E 2，Imagen，Stable Difusion などでは，入力された文に対して，高水準で内容を表す画像を生成できることが報告されている [3] [4] [5]．また，これらの手法では，Qiu らの手法のように sky，water といった特定のクラスと画像とをセットとして学習が行われているのではなく，画像と画像の内容を説明したテキストとの関係を学習している．そのため，単一単語ではない複雑な文章を表す画像も生成することができ，生成画

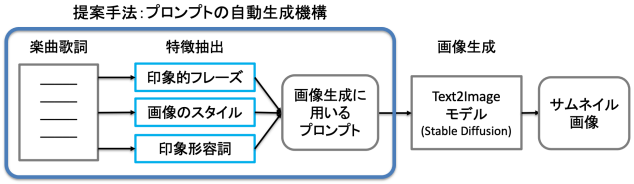


図 1　提案手法の概要

像の多様性という点において従来手法よりも秀でていると考えられる．一方で，入力するテキストの記述方法によっては目的の画像とは全く異なるものが出力されるといった課題点もあり，説明文を Text2Image モデルの理解しやすい形に調整する必要がある．そのため，理想的な画像を生成するための高品質なテキストについて検討するプロンプトエンジニアリングが重要視されており盛んに研究が行われている [10].

2.3　楽曲における印象的フレーズ

楽曲の歌詞から，印象的フレーズを抽出することを目指した例として，山西らの研究では，楽曲におけるフレーズの繰り返しが印象に大きな影響を与えていることを，被験者実験の結果から報告している [6]．ここで言う繰り返しとは，同じ文字列の繰り返しという意味であるが，本研究ではこれに加えて，フレーズの意味的な繰り返し構造についても，印象的フレーズに影響を与えているという仮説をたてた．楽曲には作者が一曲全体を通して伝えたいテーマがあり，この作者が表現したいテーマと意味的に最も近いフレーズが印象的フレーズになると考えられる．そして，作者が表現したいテーマは楽曲内のフレーズすべてに対して意味的な影響を与えていると考えられるため，楽曲のテーマと意味的に近い印象的フレーズは，その他のフレーズに対する類似度が大きいと考えられる．つまり，各フレーズのその他のフレーズに対する意味的な重畳性を確認すれば，印象的フレーズを推定することが可能であると考えられる．また，このとき抽出した印象的フレーズは，作者が伝えたいテーマと意味的に近いため，楽曲にふさわしい画像を生成するためのテキストとしても利用可能である可能性がある．

3　提 案 手 法

3.1　提案手法概要

本研究では，楽曲歌詞から印象的フレーズと画像のスタイルおよび雰囲気を表す印象形容詞を抽出し，そのテキストを利用してサムネイル画像の自動生成を行う手法について検討する．印象的フレーズの抽出は，楽曲歌詞の各フレーズを Doc2Vec を用いて分散表現に変換し，それらの意味的な重畳性を用いて行う．スタイルおよび印象形容詞の抽出には，ArtEmis データセットを用いてファインチューニングした BERT を利用する．その後，印象的フレーズと雰囲気を表す形容詞をそれぞれ英語に翻訳し，連続文とした後 Stable Difusion に入力しサムネイル画像を獲得する．提案手法の流れを図 1 に示す．

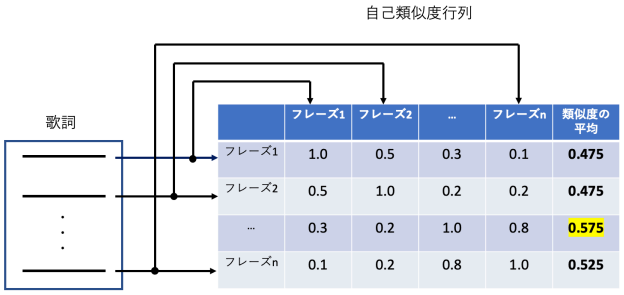


図 2　自己類似度を用いた印象的フレーズの決定法

3.2　印象的フレーズの抽出

印象的フレーズの抽出は，楽曲の歌詞における繰り返し構造をもとに実施する．具体的には，各フレーズ間での自己類似度を用いてフレーズの重要度を算出することによりこれを実現する．

フレーズ間での自己類似度を用いた印象的フレーズの決定法を図 2 に示す．楽曲の歌詞は，時間的な順序関係を有するフレーズの列と考えることができる．まず，対象とするフレーズ p に対して，同じ楽曲に含まれるすべてのフレーズとの類似度を求める．その後，求めたすべての類似度の平均を取ることで，フレーズ p の重要度を予測する．この計算を，対象とする楽曲に含まれるフレーズすべてを対象として行った後，求めた重要度が最も大きいものを，その楽曲の印象的フレーズとする．

我々は，印象的フレーズについて，文字列の繰り返し構造だけでなく，意味的な繰り返し構造が影響を与えているという仮説を立てた．そこで，提案手法における類似度の計算には，文字列の類似度だけでなく，意味的な類似度についても検証を行い，両者を比較する．

フレーズの意味的な類似度を計算するためには，まず各フレーズの文字列をそのフレーズが持つ意味を表現できる形式に変換し，その後類似度を計算する必要がある．フレーズの形式変換には，テキストの分散表現を利用する．フレーズの分散表現を求めるためには，フレーズを直接 Doc2Vec で変換する方法と，フレーズを構成する各トークンを Word2Vec で変換し，それらの平均をフレーズの分散表現とする方法の 2 つを採用する．分散表現の類似度の計算には，cos 類似度を利用する．

以上の条件で，楽曲歌詞に図 2 で示した提案手法を適用し，各フレーズで求めた cos 類似度の平均値が最も大きいフレーズを印象的フレーズと決定する．

なお，単純な平均では，各フレーズに対して全体的に意味的な類似性を示すフレーズが印象的フレーズとして選択される．そのため，強い意味的な繰り返しが行われているフレーズを考慮した抽出が行えない．そこで，各フレーズに対する cos 類似度が，一定のしきい値を超えた回数をカウントし，そのカウント数も印象的フレーズの選択に利用する．具体的に比較して用いる印象的フレーズの決定方法は，(1) 各フレーズに対する cos 類似度の算術平均，(2) 一定のしきい値を超えた回数，(3) しきい値を超えた回数を重みとする重みつき平均，の 3 つである．

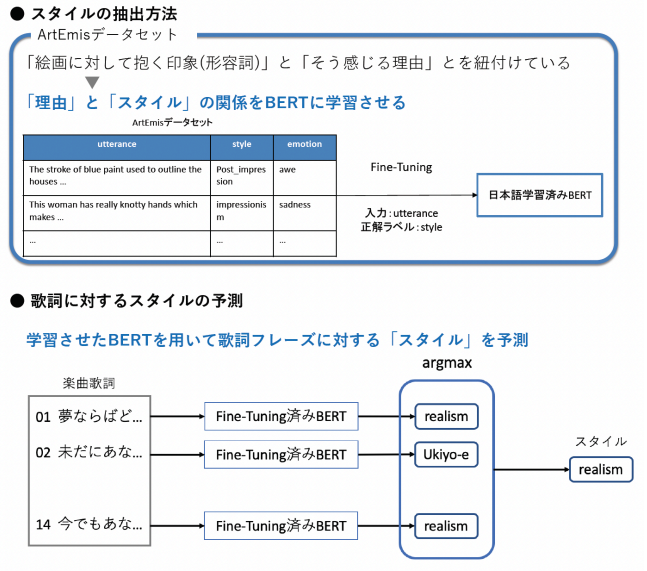


図 3　スタイルの抽出法

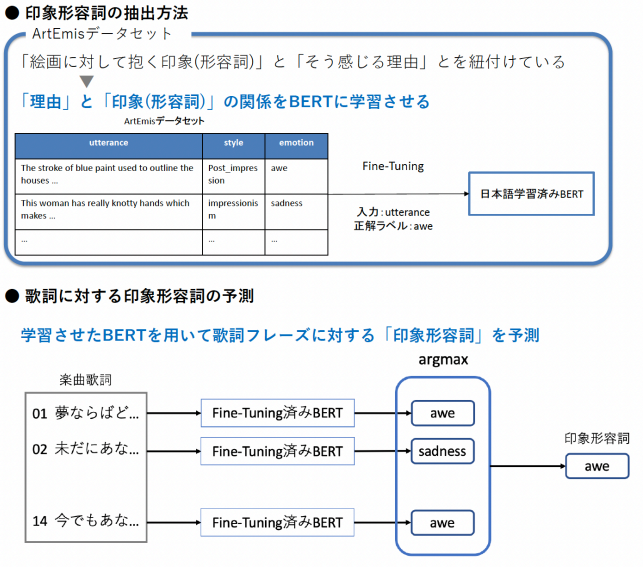


図 4　印象形容詞の抽出法

3.3　スタイルおよび印象形容詞の抽出

サムネイル画像のスタイルおよび雰囲気を表す印象形容詞の抽出は，ArtEmis を用いてファインチューニングした BERT を用いて，楽曲歌詞に最も近いラベルを予測するという流れで行う．

まず，乾らによって提案された日本語学習済み BERT モデル [1]を，ArtEmis データセットによりファインチューニングする．ArtEmis データセットには被験者実験により得られた，絵画に対する 9 つの印象ラベル（Amusement，Awe，Contentment，Excitement，Anger，Disgust，Fear，Sadness，Something Else）とその印象ラベルを付与した理由の説明文が収録されている．これらの印象ラベルは，Jana Machajdik が画像の感情分類に関する研究で用いた 4 つの Positive な感情ラベル，4 つの Negative な感情ラベル，それ以外の合計 9 つで構成されている [11]．

このデータセットを用いて，説明文を入力とし，その説明文に対するスタイルを予測する分類問題および印象ラベルを予測する分類問題のタスクを BERT にそれぞれ学習させる．この際，感情ラベルが「Something else」となっているデータについては除外する．また，入力する説明文は Google によって提供される googletrans を用いて日本語に翻訳する．

次に，ファインチューニングを行った BERT に対して楽曲歌詞の各フレーズを入力しスタイルおよび印象形容詞を予測させる．その後，1 曲の全フレーズを通じて予測された回数が最も多かったラベルをその楽曲を代表するスタイルおよび印象形容詞とする．

提案手法の処理の流れを図 3，図 4 に示す．

3.4　画像生成

3.2 および 3.3 で述べた手法により歌詞から抽出した印象的フレーズと予測した印象形容詞を，Text2Image モデルに入力することで，サムネイル画像を生成する．画像生成を行う Text2Image モデルは，Robin Rombach らによって提案された Stable Difusion [5] を使用する．

4　実験と考察

4.1　印象的フレーズの抽出

4.1.1　データセットの作成

楽曲歌詞に対する聴取者の印象を評価するための基準となるデータセットを被験者実験により作成した．

対象楽曲は，歌詞検索サービスの歌ネットにおける歴代人気曲ランキング，上位 25 曲（2022 年 7 月 31 日アクセス時点）とし，クラウドソーシングにより 30 名〜40 名程度の被験者に印象的フレーズを評価させた（楽曲により回答者数が異なる）．実験では，被験者に 1 曲ずつ楽曲の歌詞を提示し，その楽曲において印象的であると感じるフレーズを，1 行単位で任意数選択させた．その後，回答を集計し，被験者により選択された回数により各フレーズに対して印象的度合いのランク付けを行った．

4.1.2　印象的フレーズの抽出

4.1.1 で述べた楽曲 25 曲に対して，提案手法を適用し，文字列の繰り返し構造を用いた手法（編集距離を用いた手法），意味的な繰り返し構造を用いた手法（Word2Vec を用いた手法，Doc2Vec を用いた手法）で，それぞれ印象的フレーズを求めた．なお，Word2Vec[2]および Doc2Vec[3]は日本語学習済みモデルを利用した．また，意味的な繰り返し構造を用いた手法では，印象的フレーズを求める際に，cos 類似度の算術平均，cos 類似度が一定のしきい値を超えた回数，cos 類似度が一定のしきい値を超えた回数を重みとする重みつき平均の 3 つを用いた．cos 類似度のしきい値は 0.9 とした．

その後，提案手法により予測された印象的フレーズと，被験

1：https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking

2：http://www.cl.ecei.tohoku.ac.jp/ m-suzuki/jawiki$_v ector$/

3：https://yag-ays.github.io/project/pretrained$_d oc2vec_w ikipedia$/

表 1　予測した印象的フレーズと正解データとの一致度（編集距離）

1 位正解率	5 位以内正解率	平均順位	分散
0.32	0.68	4.16	13.31

表 2　予測した印象的フレーズと正解データとの一致度（Word2Vec）

フレーズの決定方法	1 位正解率	5 位以内正解率	平均順位	分散
算術平均	0.08	0.32	8.68	24.48
0.9 を超えた回数	0.28	0.72	4.12	14.61
重みつき平均	0.24	0.76	3.96	13.71

表 3　予測した印象的フレーズと正解データとの一致度（Doc2Vec）

フレーズの決定方法	1 位正解率	5 位以内正解率	平均順位	分散
算術平均	0.04	0.16	9.36	18.66
0.9 を超えた回数	0.36	0.84	3.24	10.27
重みつき平均	0.20	0.72	4.48	13.18

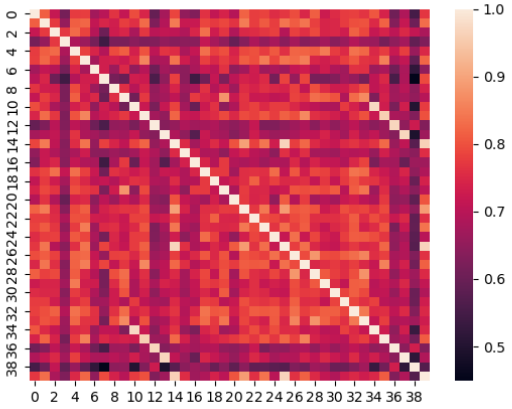


図 5　「Lemon ／米津玄師」に Doc2Vec を適用した際の自己類似度行列

者実験により定めた印象的フレーズとの一致度を求めた．

提案手法により予測した印象的フレーズと，被験者実験により定めた印象的フレーズとの一致度を，表 1, 表 2, 表 3 に示す．

表の各列は実験に用いた 25 曲に対して予測した印象的フレーズがそれぞれ，回答数 1 位のフレーズと一致した割合，回答数が 5 位以内のフレーズと一致した割合，予測したフレーズに付与された正解ランクの平均，予測したフレーズに付与された正解ランクの分散である．

また，最もスコアが高かった，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決定した際の，抽出結果を表 4 に示す．表 4 には，例として歌ネットにおける歴代人気曲ランキング上位 10 曲を示す．

まず，ベースラインとなる文字列の繰り返し構造を用いた手法では，全体の 3 割程度，回答数 1 位のフレーズを予測できている．そのため，楽曲歌詞における文字列の繰り返し構造は，印象的フレーズに大きく影響を与えていることが示唆された．しかし，文字列の繰り返し部分以外には印象的フレーズを抽出することができないため，予測を大きく外している楽曲もあった．

次に，意味的な繰り返し構造を用いた手法では，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数により印象的フレーズを決定した際に，最も高いスコアを示した．特に，5 位以内正解率および予測したフレーズに付与された正解ランクの平均が，それぞれ 0.68 から 0.84，4.16 から 3.24 に向上している．つまり，印象度合い 1 位のフレーズを予測できない場合でも，予測を大きく外すことが少なくなっている．これは，cos 類似度のしきい値を 0.9 としたため，ある程度の文字列の類似性も考慮することができ，その上で，意味的に強く繰り返されている部分が抽出できているからではないかと考えられる．

一方で，予測を大きく外す楽曲もいくつか存在したため，該当する楽曲の正解データを確認したところ，文字列の繰り返しでもなく意味的な繰り返しでもない部分が印象的フレーズとして選択されていた．たとえば，楽曲においてアーティストが感情を込めて強く歌う部分や，タイトル名と関連のある部分，楽曲の音声的な構造で特異かつ目立つ部分などが選択されていた．このようなフレーズについては，意味的な繰り返し構造を用いた手法では抽出できないため，正解することが難しいと考えられる．

また，意味的な繰り返し構造を用いた手法において算術平均を用いて印象的フレーズを決定した際には，各スコアが大きく低下する結果となった．これは，各フレーズに対して全体的に意味的な類似性を示すフレーズが選択されているため，強い意味的な繰り返しが行われているフレーズを抽出できていないことが要因であると考えられる．例として，図 5 に，「Lemon ／米津玄師」の歌詞に対して Doc2Vec を用いた手法を適用した際の cos 類似度の自己類似度行列を示す．被験者により，選定された印象的フレーズは 14，25，39 番（「今でもあなたは私の光」）であり，cos 類似度のしきい値を用いた場合では類似度が 0.9 を超えた回数をカウントしているため，正解フレーズを抽出できている．しかし，算術平均を用いた場合では cos 類似度の平均をとるため，自己類似度が全体的に高いフレーズである 4 番（「戻らない幸せがあることを」）が選択されてしまっている．このように算術平均を用いた手法では，意味的に重要でないフレーズの影響も多分に受けてしまうため，結果的に目的とする印象的フレーズを抽出できない傾向にあると考えられる．

4.2　スタイルおよび印象形容詞の抽出

3.3 で述べた手法に基づき，BERT モデルのファインチューニングとスタイルおよび印象形容詞の予測を行った．

まず，BERT モデルのファインチューニングに用いた ArtEmis データセットは表 5 のような構成になっており，utterance を入力とし emotion を正解ラベルとする分類問題，utterance を入力とし art style を正解ラベルとする分類問題を学習させた．この際，emotion が something else となっているデータは除外した．学習の際は，something else を除いた全データ 401682 件を，学習データ 6 割，検証データ 2 割，テストデータ 3 割にランダムに分割し実験を行った．

表 4　正解データにおける回答数 1 位のフレーズと提案手法により選択されたフレーズ

楽曲名／アーティスト名	被験者アンケート回答数 1 位のフレーズ	提案手法によって選択されたフレーズ	選択されたフレーズの正解データ内での順位
Lemon ／米津玄師	今でもあなたは私の光	今でもあなたは私の光	1
クリスマスソング／ back number	君が好きだ	君が好きだ	1
キセキ／ GReeeeN	せめて言わせて 「幸せです」と	君に巡り会えた それって奇跡	2
花束／ back number	君と抱き合って手を繋いでキスをして	僕は何回だって何十回だって	2
前前前世（movie ver.）／ RADWIMPS	君の前前前世から僕は 君を探しはじめたよ	君の前前前世から僕は 君を探しはじめたよ	1
Pretender ／ Official 髭男 dism	君の運命のヒトは僕じゃない	グッバイ	2
LOSER ／米津玄師	アイムアルーザー どうせだったら遠吠えだっていいだろう	アイムアルーザー どうせだったら遠吠えだっていいだろう	1
紅蓮華／ LiSA	紅蓮の華よ咲き誇れ！ 運命を照らして	強くなれる理由を知った 僕を連れて進め	4
ハッピーエンド／ back number	あなたを好きなままで消えてゆく	なんてね 嘘だよ 元気でいてね	9
RPG ／ SEKAI NO OWARI	空は青く澄み渡り 海を目指して歩く	空は青く澄み渡り 海を目指して歩く	1

次に，ファインチューニングした BERT モデルを用いて，楽曲に対応するスタイルおよび印象形容詞をそれぞれ予測した．歌ネットにおける歴代人気曲ランキング上位 10 曲に対する予測結果を表 6，表 7 に示す．

4.3 画像生成

提案手法により抽出した印象的フレーズと印象形容詞を，Stable Difusion モデルに入力し，サムネイル画像を生成した．

生成の手順は，印象的フレーズを googletrans に入力し，出力された英文と予測したスタイルおよび形容詞とを連続文とし Stable Difusion モデルに入力することで，サムネイル画像を獲得した．生成したサムネイル画像の例を図 6，図 7，図 8 に示す．また，生成に用いたプロンプトを表 8 に示す．

4.3.1 画像生成に「印象的フレーズ」のみを用いた場合

画像生成に印象的フレーズのみを用いた場合は，プロンプトが具体的な心象風景と一致しやすいものに関しては妥当な画像を生成できたものの，プロンプトが心象風景に繋がりにくいものに関しては楽曲の雰囲気と合致しない画像が生成された．たとえば，図 7 の「RPG ／ SEKAI NO OWARI」のサムネイル画像は比較的楽曲の雰囲気およびフレーズの内容を表すことができていると思われる．これは，生成に用いた「空は青く澄み渡り 海を目指して歩く」というフレーズが，具体的な心象風景としてイメージしやすいため，結果として画像化しやすくなっているからであると考えられる．一方，図 6「Lemon ／米津玄師」や図 8「白日／ King Gnu」は，印象的フレーズが「今でもあなたはわたしの光」「真っ新に生まれ変わって」など，具体的な心象風景に繋がりにくいため，楽曲の雰囲気と合致しない画像が生成されたと考えられる．

4.3.2 画像生成に「印象的フレーズ+スタイル」を用いた場合

画像生成に印象的フレーズとスタイルを用いた場合は，生成画像の妥当性が大きく改善した．印象的フレーズのみで画像を生成した際に，サムネイル画像として扱いづらい画像となった図 6「Lemon ／米津玄師」や図 8「白日／ King Gnu」も，スタイルを指定することによりサムネイル画像として扱いやすくなった．これは,「Realism」や「Synthetic Cubism」といった絵画のスタイルを指定することで,「人間が鑑賞する対象である」という解釈が行われたためであると思われる.「白日／ King Gnu」の印象的フレーズは「真っ新に生まれ変わって」であり，このフレーズだけでは具体的な心象風景はイメージしづらいが，スタイルと合わせて明示することで人間が鑑賞する対象であることが情報として加わり，図形模様の画像を生成することができた．

4.3.3 画像生成に「印象的フレーズ+印象形容詞」を用いた場合

画像生成に印象的フレーズと印象形容詞を用いた結果，印象形容詞の有無は生成する画像の内容を大きく変化させるような影響を与えなかった．一方で，画像の雰囲気を変化させる効果があると思われる．図 6〜図 8 より，印象形容詞の有無による生成画像の違いを確認すると，画像のメインとなる内容は印象的フレーズにより決定され，印象形容詞はその雰囲気を決定づけていると考えられる．

4.3.4 画像生成に「印象的フレーズ+スタイル+印象形容詞」を用いた場合

画像生成に印象的フレーズ，スタイル，印象形容詞を用いた結果,「印象的フレーズ+スタイル」によって生成された画像を,「印象形容詞」の持つ雰囲気に近づけた画像が生成された．図 6〜図 8 のいずれの場合も,「印象的フレーズ+スタイル」で生成された画像と「印象的フレーズ+スタイル+印象形容詞」で生成された画像の内容は似通っているが,「Awe impression」「Fear impression」といったそれぞれが持つ印象形容詞を反映させた雰囲気に変換させる効果があることがわかる．

5 まとめ

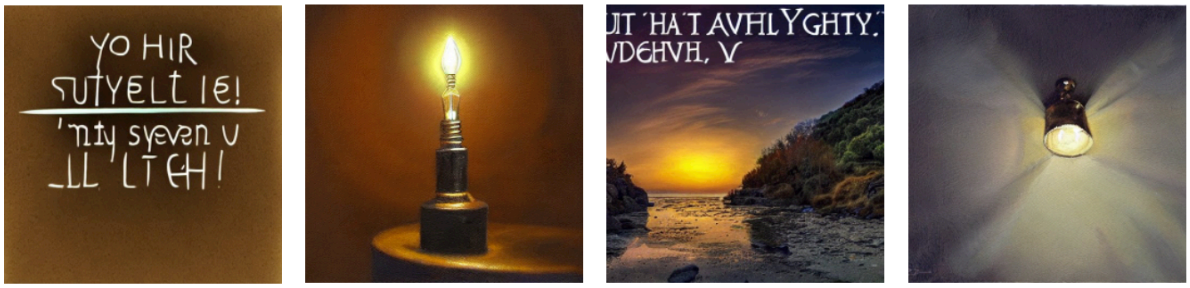
本研究では楽曲歌詞の意味的な重畳性に基づいた印象的フレーズの抽出方法と ArtEmis データセットと BERT を利用したスタイルおよび印象形容詞の予測方法を提案した．その後，抽出した印象的フレーズおよび印象形容詞を Text2Image モデルに入力することで，サムネイル画像の生成を行った．

実験の結果，印象的フレーズの抽出については，Doc2Vec を用いて他のフレーズに対する cos 類似度が 0.9 を超えた回数によりフレーズを決定した際に，文字列の繰り返し構造に着目した手法（ベースライン）よりも，高精度で抽出が行えることがわかった．一方，アーティストが強く歌う部分や，音声的な構造で目立つ部分など，文字列の繰り返しでもなく意味的な繰り返しでもない部分が印象的フレーズとして選択されている楽曲は，抽出を行うことが難しかった．今後は，楽曲においてアーティストが感情を込めて強く歌う部分や，タイトル名と関連のある部分，楽曲の音声的な構造で特異かつ目立つ部分などに基づく印象的フレーズも正しく抽出が行えるように，分散表現取得モデルの再選定，楽曲の音声を考慮したフレーズ抽出方法の提案を行っていく予定である．

スタイルおよび印象形容詞の抽出については，ArtEmis デー

表 5　ArtEmis データセットの構成

art style	painting	emotion	utterance	repetition
Post Impressionism	vincent-van-gogh	something else	”She seems very happy in the picture, and you want to know what what is behind the smile.”	10
Post Impressionism	vincent-van-gogh	sadness	This woman has really knotty hands which makes her look like she has arthritis.	10
Post Impressionism	vincent-van-gogh	something else	”When looking at this woman I am filled with curiosity about what she is thinking about...”	10
Expressionism	wassily-kandinsky	something else	”The way the image is presented, with large chunks of paint used to depict each of the subjects...	7
Impressionism	konstantin-korovin	awe	the stroke of blue paint used to outline the houses in the painting	7
Impressionism	konstantin-korovin	sadness	Some of the wooden panels are falling from the buildings	7



印象的フレーズのみ　印象的フレーズ＋スタイル　印象的フレーズ＋印象形容詞　印象的フレーズ＋スタイル＋印象形容詞

図 6　生成した画像例（Lemon ／米津玄師）

印象的フレーズのみ　印象的フレーズ＋スタイル　印象的フレーズ＋印象形容詞　印象的フレーズ＋スタイル＋印象形容詞

図 7　生成した画像例（RPG ／ SEKAI NO OWARI）

印象的フレーズのみ　印象的フレーズ＋スタイル　印象的フレーズ＋印象形容詞　印象的フレーズ＋スタイル＋印象形容詞

図 8　生成した画像例（白日／ King Gnu）

表 6　提案手法により予測されたスタイル（歴代人気曲上位 10）

楽曲名／アーティスト名	提案手法によって予測されたスタイル
Lemon ／米津玄師	Realism
クリスマスソング／ back number	Rococo
キセキ／ GReeeeN	Romanticism
花束／ back number	Rococo
前前前世（movie ver.）／ RADWIMPS	Pop Art
Pretender ／ Official 髭男 dism	Pointillism
LOSER ／米津玄師	Art Nouveau Modern
紅蓮華／ LiSA	Impressionism
ハッピーエンド／ back number	Pop Art
RPG ／ SEKAI NO OWARI	Synthetic Cubism

表 7　提案手法により予測された印象形容詞（歴代人気曲上位 10）

楽曲名／アーティスト名	提案手法によって予測された印象形容詞
Lemon ／米津玄師	awe
クリスマスソング／ back number	contentment
キセキ／ GReeeeN	awe
花束／ back number	contentment
前前前世（movie ver.）／ RADWIMPS	contentment
Pretender ／ Official 髭男 dism	awe
LOSER ／米津玄師	excitement
紅蓮華／ LiSA	excitement
ハッピーエンド／ back number	amusement
RPG ／ SEKAI NO OWARI	fear

タセットを用いて BERT をファインチューニングし，それをもとに印象的フレーズに対応したラベルの予測を行った．実験の結果，スタイルを指定することで生成画像を鑑賞対象として妥当なものに変換する効果があること，また，印象形容詞を指定

表 8　画像生成に用いたプロンプト

楽曲名／アーティスト名	印象的フレーズのみ	印象的フレーズ+スタイル	印象的フレーズ+印象形容詞	印象的フレーズ+スタイル+印象形容詞
Lemon 米津玄師	You are still my light	You are still my light, Realism painting style	You are still my light, Awe impression	You are still my light, Realism painting style, Awe impression
RPG SEKAI NO OWARI	The sky is clear blue and I walk towards the sea	The sky is clear blue and I walk towards the sea, Expressionism painting style	The sky is clear blue and I walk towards the sea, Fear impression	The sky is clear blue and I walk towards the sea, Expressionism painting style, Fear impression
白日 King Gnu	Completely reborn	Completely reborn, Synthetic Cubism painting style	Completely reborn, Fear impression	Completely reborn, Synthetic Cubism painting style, Fear impression

することで画像の雰囲気を変化させる効果があることが示唆された．今後は，予測されたスタイルおよび形容詞の妥当性について被験者実験を行い検証する予定である．

謝　　辞

本研究は JSPS 科研費 19H04219 の助成を受けたものです．

文　　献

[1] Yue Qiu, Hirokatsu Kataoka: *Image generation associated with music data*, CVPR workshop (2018).
[2] 梅村允康, 保利武志, 土屋駿貴, 嵯峨山茂樹: *Word2Vec* を用いて歌詞と写真を対応づけたスライドショー生成システム, 情報処理学会第 81 回全国大会 (2019).
[3] Aditya Ramesh et al.: *Hierarchical Text-Conditional Image Generation with CLIP Latents*, arXiv:2204.06125 (2022).
[4] Chitwan Saharia et al.: *Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding*, arXiv:2205.11487 (2022).
[5] Robin Rombach et al.: *High-Resolution Image Synthesis with Latent Diffusion Models*, arXiv:2112.10752 (2022).
[6] 山西良典, 鍵田里沙子, 西原陽子, 福本淳一: 共起語の特異性と繰り返しに着目した歌詞からの印象的フレーズ抽出, 日本感性工学会論文誌, Vol.14, No.1(特集号), pp.29-35(2015).
[7] Cong Jin et al.: *Attention-Based Bi-DLSTM for Sentiment Analysis of Beijing Opera Lyrics*, Wireless Communications Mobile Computing Volume 2022 (2022).
[8] Achlioptas et al.: *ArtEmis: Affective Language for Visual Art*, CoRR, abs/2101.07396 (2022).
[9] AVA: The Art and Science of Image Discovery at Netflix, https://netflixtechblog.com/ava-the-art-and-science-of-image-discovery-at-netflix-a442f163af6 (2022.08.08).
[10] Takeshi Kojima et al.: *Large Language Models are Zero-Shot Reasoners*, arXiv:2205.11916 (2022).
[11] Jana Machajdik, Allan Hanbury: *Affective image classification using features inspired by psychology and art theory*, MM'10: Proceedings of the 18th ACM international conference on Multimedia, pp.83-92 (2010).