画像生成技術を応用した音を視覚的に認識する AI の作成

4 年 B 組　岡本　晃朋
指導教員　藤野　智美

１．要約

本研究では、音楽データを用いて楽曲のイメージに合わせたアートワークの生成を目指す。Python を用いて拡散モデルによる画像合成技術を応用し、音楽のジャンル、音の種類など、曲から連想される要素に合う画像を生成する画像生成 AI を開発する。

キーワード　AI　Python　画像生成

２．研究の背景と目的

私はパソコンを利用して音楽制作を行う DTM（デスクトップミュージック）に取り組んできた。SoundCloud 等のサービスを通じて自身が作曲した曲を他者に共有する際、アートワークと呼ばれる曲に合ったジャケット画像を用意しているが、曲のイメージに合った画像を見つけるのは大変手間がかかる。そこで、楽曲のイメージに合わせて、ジャケットを自動で作成できるシステムの構築ができないかと考え、本研究に至った。

３．研究内容

３．１　画像生成 AI の作成

画像生成システムは stable-diffusion を元に改良した物を使用する。以下に改良した内容を説明する。

３．１．１　Negative_Prompts の実装

Stable-diffusion では Prompts と呼ばれる引数から画像を導くが、反例に相当する Negative_Prompts が実装されていなかったため、自分で実装した。方法として、get_learned_conditioning(code1)に batch-size(self)と Negative_Prompts[list]の乗算で表した。

３．１．２　画像の雰囲気などの調整

生成される画像の絵柄などは学習データによるが、雰囲気などの制御ができがたいと考え、Latent Space にて生成したものを画像へ decode する回数を調整することで、質感を変えることができた。

３．２　システムの構築

システムの概要は図 1 で示す流れとなる。また、Latent Space 内の Denoising U-Net に代入する値は以下のような式で表すことができる。

$$Attention(Q,K,V) = softmax\left(\frac{QK^T}{\sqrt{D}}\right) \cdot V$$

$$Q = W_Q^{(i)}$$

$$\varphi i(Z\tau), K = W_k^{(i)} \cdot \tau_\theta(y)$$

$$V = W_V^{(i)} \cdot \tau_\theta(y)$$

生成される画像については、シード値と呼ばれる 32bit の数を設定することによっ

て同じ結果が得られる。同じシード値で複数回実施した場合、雰囲気などはある程度制御することはできるが、学習量の関係から特定の絵柄を完全に再現することは困難である。

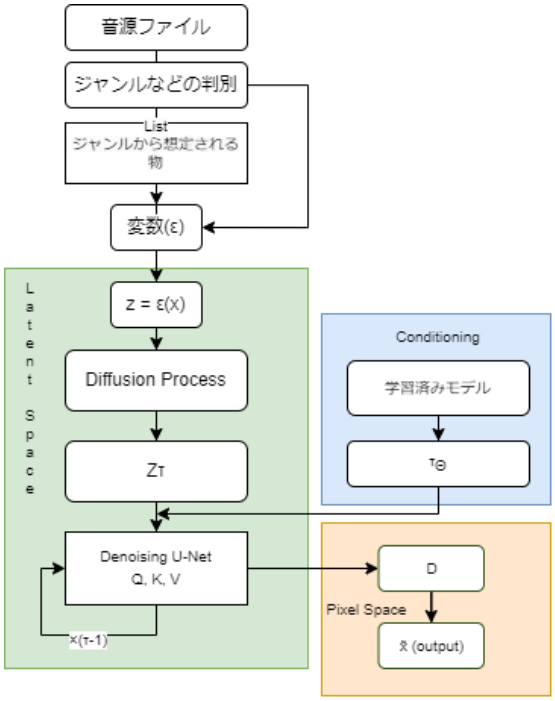

図 1 システムの概要

３．３　曲から Prompts を生成する

実際に Prompts を生成する際に、曲から音楽ジャンルや楽器の種類などを判別する仕組みを現在構築中である。方法としては、SoundCloud API より取得した音楽ファイル、ジャンルなどを学習させたモデルを利用し、利用者の作成した曲の情報から作成している。

３．３．１　データの軽量化

実際に Prompts を生成してみたが、一曲あたり大体 3 分から 4 分ほどあり、かなり時間がかかってしまうので、学習方法、インプット方法を変更することにした。方法としては、曲の中からランダムな 1 小節を切り取り、その中の要素から生成することを試みた。しかし、やはり一小節では曲の一部しか使用できないので、曲のうちの Intro, Drop の二種類をサンプルとして指定することで、より曲に近いものに仕上げることができた。簡単なニューラルネットワークの仕組みを図 2 に示す。また、活性化関数（非線形変形）にはロジスティックシグモイド関数を使用した。

$$h = a(\mu) = \frac{1}{1 + \exp(\mu)}$$

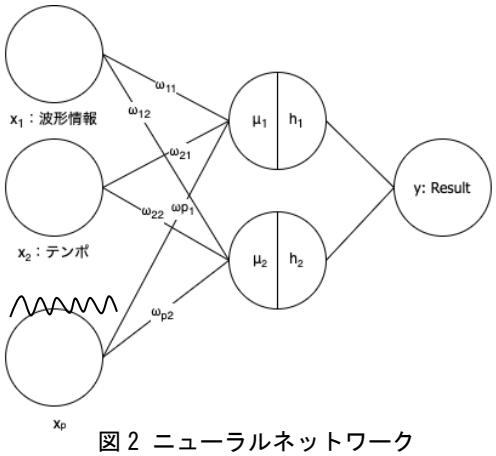

図 2 ニューラルネットワーク

３．３．２　交差エントロピーの使用

交差エントロピーとは、分類問題を解く際に用いられる目的関数である。

$$L = -\sum_{n=1}^{n}\sum_{k=1}^{k} t_{n,k} \log y_{n,k}$$

ある入力xについて、所属するクラスの正解がワンホットベクトル

$$t = [t_1, t_2, t_3, \quad , t_k]^T$$

となる時、K個のニューラルネットワークのうちのk番目にいく確率は、

$$y_k = p(y = k|x)$$

となり、これらのことから交差エントロピーは以下のように定義できる。

$$-\sum_{k=1}^{k} t_k \ log y_k$$

これを 3.3.1 にて作成したものの中間層に追加する。

３.３.３　交差エントロピーの最適化

現在のニューラルネットワークに訓練を行う。損失関数の値が最小のとき、重さが適切なので、勾配降下方を利用して最適な値を見つける。勾配降下法は学習率をηとした時、下のように表せる。

$$w \leftarrow w - \eta \frac{\partial \mathrm{L}}{\partial \mathrm{w}} | w$$

勾配は$\frac{\partial L}{\partial w}$で表せられ、$L$の値を小さくしたいので、傾きの逆方向に$w$を変化させる。こうすることで、$L$を最小にする$w$に徐々に近づけていくことができる(図 3)。

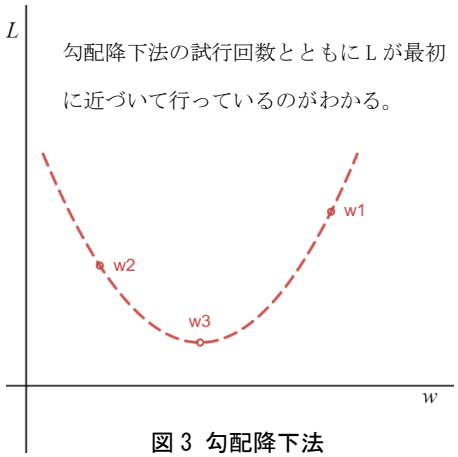

図 3 勾配降下法

４.結果

前述したシステムの実装の結果、作成できたアートワークを図 4 に示す。今回の研究により、曲から Prompts を生成し、複数の雰囲気の画像を生成することができた。

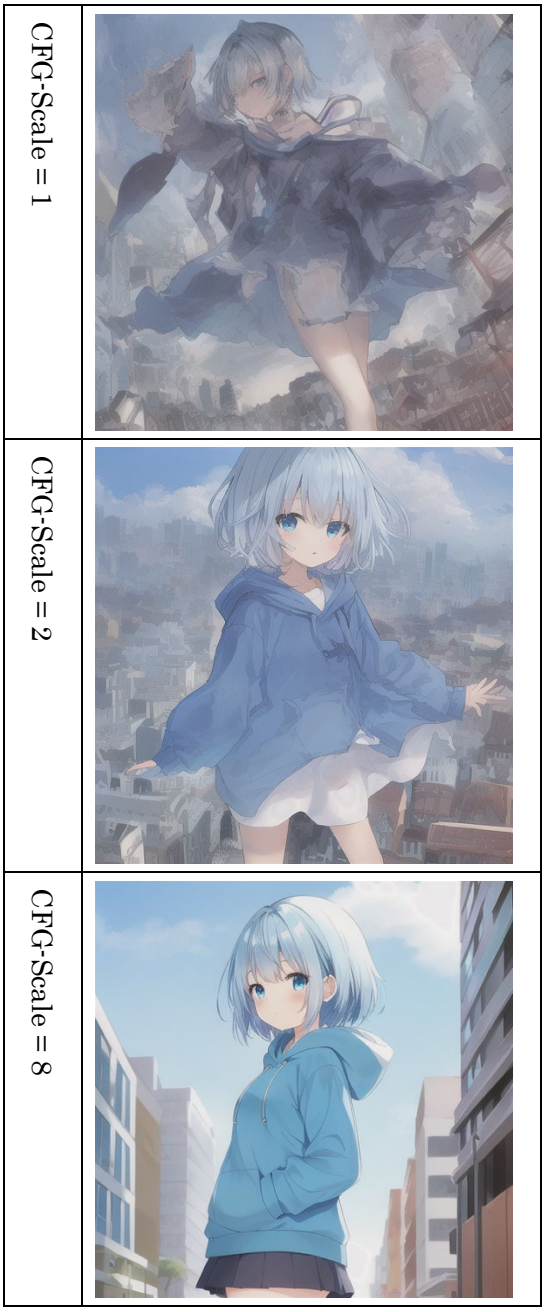

図 4 作成したアートワーク

５.今後の展望

現在は Python をメインにローカル環境でしか実行できないか、サーバー上にホスティングしてどのような機種でも簡単に本サービスを利用できるようにしたい。

また、現在は学習量の関係で十分な結果が得られていないため、より精度を向上させることで、アートワークを見ただけでど

のような雰囲気の曲なのか、ある程度判別できるようにしたい。

さらに、Spleeter にて音源分離を行った後に使われている楽器などの情報を得て、そこから想像されうるPromptsを作成する方法や、他の情報源として歌詞データなどを利用することも考えている。

今後、音楽だけではなく日常の音などにも利用できれば、聴覚に困難さを感じる方々に対する標識の作成などにも役立てるかもしれない。継続して今後の発展について研究したい。

６．参考文献

High-Resolution Image Synthesis with Latent Diffusion Models
https://ommer-lab.com/research/latent-diffusion-models/

７．謝辞

今回の研究を行うにあたり、顧問の藤野先生、物理班の諸先輩方には多大なご指導を賜りました。この場を借りて深く御礼申し上げます。