事前学習用画像データセット生成モジュール

自然画像を用いない AI の学習への挑戦

askbox

2024 年 3 月 12 日

コンテストの評価結果

予選: 2 位

順位	ユーザ名	暫定評価	最終評価
1	Petr (CZ)	1620.2100000	12.7838994
2	askbox	1557.1400000	13.9381575

本戦: 定量評価部門 3 位

順位	ユーザ名	暫定評価	最終評価
1	FYSignate1009	1088.1100000	23.1545619
2	Petr (CZ)	1136.7800000	23.7837702
3	askbox	1134.3200000	23.9095506

開発したモジュールの概要

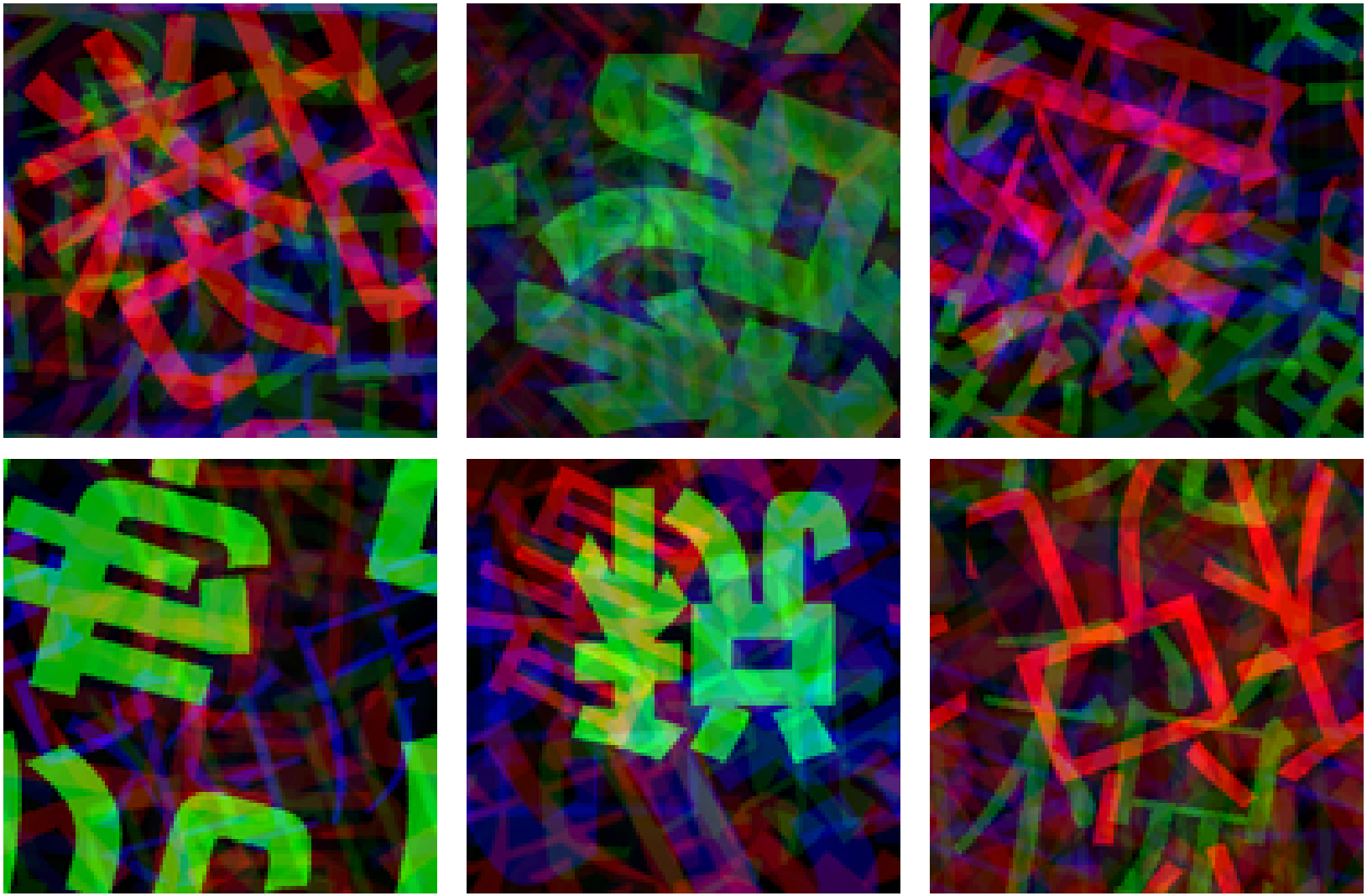
概要

既存のフラクタル画像を利用せず、独自のアイデアとして同じ漢字・文字が出現された生成画像は同じクラスと仮定する。

精度向上のポイント

漢字・文字の画像をクラスタリング処理、各クラスの文字数・各文字の大きさ・描画位置・回転角度・上下左右反転・合成の割合調整等の試行錯誤を重ねる。

2023 年「今年の漢字」[1]の生成画像サンプル



[1] トップ 6: 税, 暑, 戦, 虎, 勝, 球

モジュールのアルゴリズム

モジュールの開発

1. フェーズ 1：漢字・文字の画像のクラスタリング
2. フェーズ 2：画像生成

モジュールのアルゴリズム [フェーズ 1]

漢字・文字の画像のクラスタリング

1. 3 万以上の漢字・文字の画像[2] を VAE[3] で 100 次元に圧縮する。
2. 圧縮した 100 次元データを k-means 法、PCA[4] で 1000 クラスと各クラスが 6 文字を超えるように調整する。
3. クラスタリングしたデータを csv ファイルに書き出す。

[2] 64x64pixel のグレースケール画像
[3] VAE(Variational Auto-Encoder): 変分オートエンコーダー
[4] PCA(Principal Component Analysis): 主成分分析

モジュールのアルゴリズム [フェーズ 1]

クラスタリングした 1000 クラスのサンプル

クラス	c0	c1	c2	c3	c4	c5
0	語	餾	脂	諧	諂	諂
1	屬	肩	署	暑	署	署
2	坡	玻	陂	敍	敍	被
3	璜	煌	僥	煌	瑛	磺
…	…	…	…	…	…	…
995	呦	唎	吻	吻	唎	唎
996	涵	涸	涸	湄	涵	滙
997	浦	埔	浦	消	埔	浦
998	雷	雷	罍	畐	皇	皇
999	圳	垉	垉	垉	垉	池

モジュールのアルゴリズム [フェーズ 2]

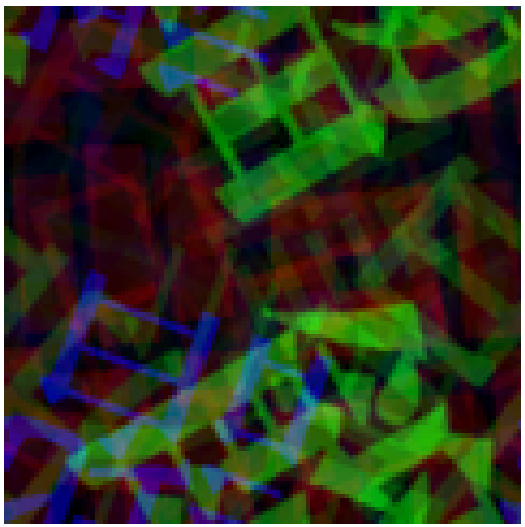
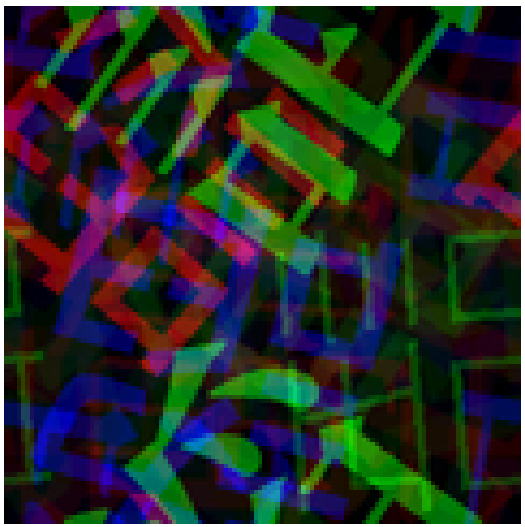
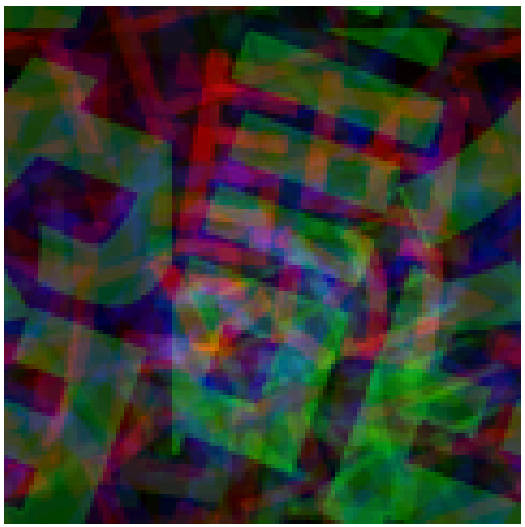
画像生成

1. 事前にクラスタリングした 1000 クラスの csv ファイルを読み込む。
2. 各クラスに出現する漢字・文字を抽出する。
3. 抽出した漢字・文字の組み合わせの数を増やすため、RGB 各チャンネルに文字の画像を生成する前に下記の項目をランダムに選択する。
   - フォントの種類[5] ・文字の大きさ
   - 描画位置・回転角度・上下左右反転
4. 生成した画像リストの混合比もランダムに選択する。
5. 生成した RGB 画像ファイルを書き出す。

[5] Ubuntu 20.04.6 LTS で使用可能なフォント

モジュールのアルゴリズム [フェーズ 2]

クラス 0[6] による生成画像サンプル

[6] クラス 0: 語, 餾, 脂, 諧, 諂, 諂

モジュールのアルゴリズム [フェーズ 2]

利用したフォント一覧[7]

```
NotoSansCJK-Black.ttc NotoSansCJK-Bold.ttc
NotoSansCJK-DemiLight.ttc NotoSansCJK-Light.ttc
NotoSansCJK-Medium.ttc NotoSansCJK-Regular.ttc
NotoSansCJK-Thin.ttc NotoSerifCJK-Black.ttc
NotoSerifCJK-Bold.ttc NotoSerifCJK-ExtraLight.ttc
NotoSerifCJK-Light.ttc NotoSerifCJK-Medium.ttc
NotoSerifCJK-Regular.ttc NotoSerifCJK-SemiBold.ttc
TakaoGothic.ttf TakaoMincho.ttf TakaoPGothic.ttf
TakaoPMincho.ttf ipaexg.ttf ipaexm.ttf
ipag.ttf ipagp.ttf ipam.ttf ipamp.ttf
```

[7] Ubuntu 20.04.6 LTS で使用可能なフォント

考察

3 位入賞の要因として、多様な漢字・文字のパターンを活かし、生成した画像の混合比をランダムに調整し、画像処理の組み合わせの数を増やしたことで、画像合成後に多様な拡張画像が生成された。