

Emoji-Alchemist

VAE と BO を用いた創造的な画像生成の試み

Yuto Takagi

竹内・田地研 DL-b

June 19, 2025

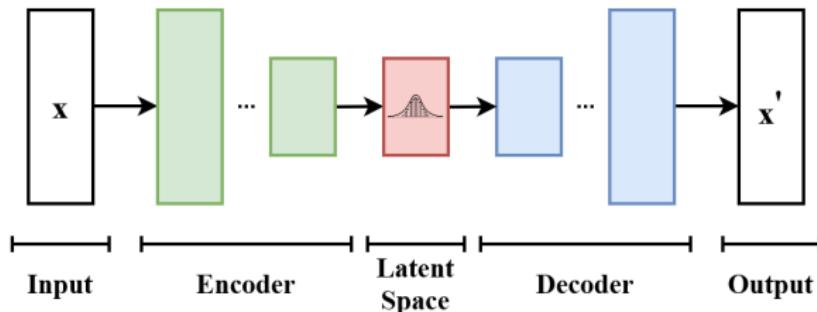
目次

- 背景と目的
- アーキテクチャ
- 実験
- 結果と結論
- Appendix

技術背景

DL-b グループのコア技術：分子探索におけるベイズ最適化 (BO)

- **VAE**: 化合物の構造的特徴を連続的な潜在空間 z に写像.
- **BO**: 高コストなブラックボックス関数（ドッキングシミュレーション等）の評価回数を最小限に抑え，最適な特性を持つ分子の潜在ベクトル z^* を発見する.



56f3: VAE と BO を組み合わせた LSBO 出典: Wikipedia, Variational autoencoder

スキルアップセミナーでやったこと

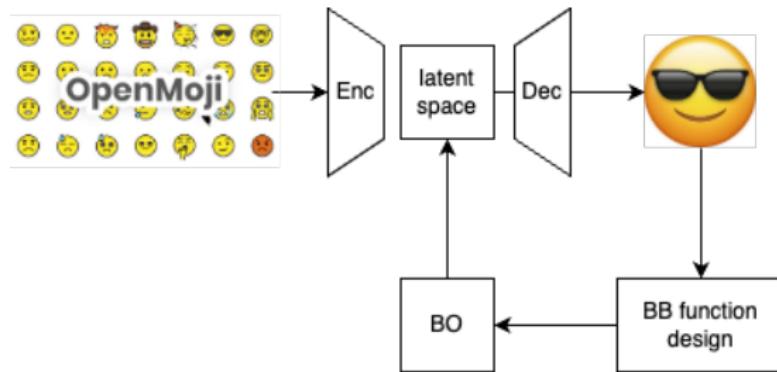
- MLP の基礎
- CNN で MNIST 手書き数字分類
- VAE で MNIST 手書き数字を潜在空間に写像
- MIL でバッグ単位で画像を分類

これらの知識を活用できるような問題設定がいいなと考えました

思いつき

・アイデア

この強力な「VAE+BO」フレームワークは、評価関数が全く異なる性質を持つ創造的タスクの領域でも同様に機能するのだろうか？



56f3: アイデア図

挑戦する「創造的タスク」の選定

なぜ「絵文字の画像生成」なのか？

- **知見の活用:**

スキルアップセミナーで扱った CNN や VAE など，画像処理の知識を直接活かせるため.

- **評価のしやすさ:**

生成されたものが「良いか悪いか」を直感的に判断しやすく，人間の感性を評価関数とする今回のテーマに最適であるため.

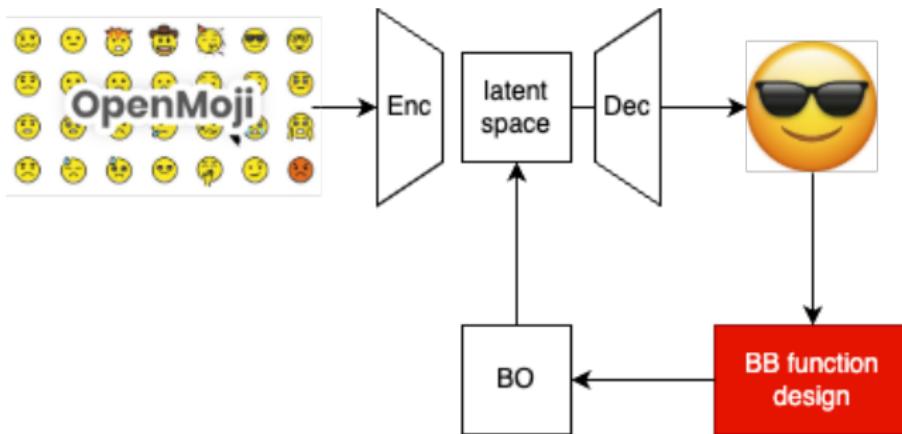
- **手頃なデータセット:**

OpenMoji という高品質なデータセットが存在し，実験の基盤として適しているため.

結論→ VAE+BO フレームワークの限界と可能性を探る最初のステップとして，「絵文字生成」は理想的な題材であると考えた.

ブラックボックス関数の設計

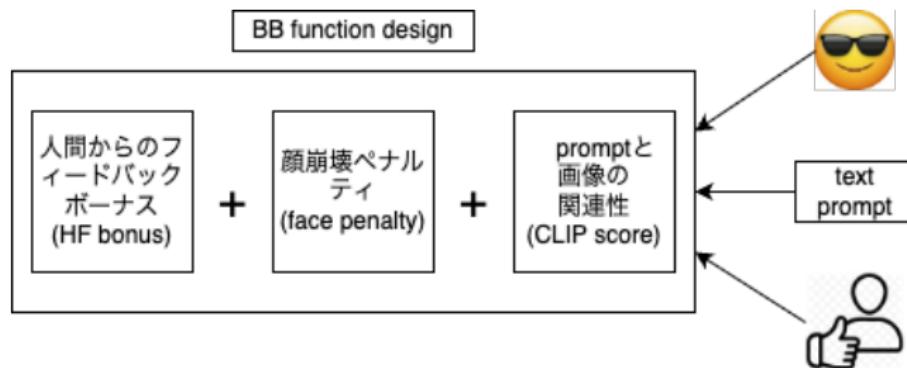
BO は評価関数に忠実 → ブラックボックス関数にユーザーの感性を取り入れる



56f3: 全体アーキテクチャ

ブラックボックス関数の設計

テキストとの類似度 + 顔検出ペナルティ + ユーザーのフィードバックボーナス
(Appendix 参照)

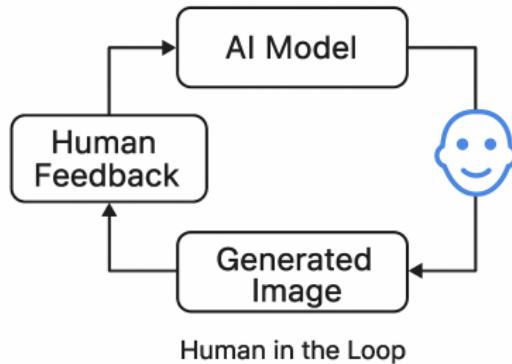


56f3: ブラックボックス関数の中身

目標: interactive な生成システム

従来手法の限界と新しい発想

- Stable Diffusionなどの従来手法では、テキストのみで画像を指定するため、感性的なニュアンス（例：「ちょっと憂いのある表情」など）を伝えるのが難しい。
 - そこで、人間の感性そのものをブラックボックス関数と見なす
 - 人間のクリックなどのフィードバックをBOにリアルタイムで取り入れることで、「人間と協調しながら進化する生成システム」を目指す。



問題設定

目標

text = "A blue emoji made of ice" \Rightarrow



「テキストで指定したイメージ」と「ユーザーの感性」から新しい絵文字を鍛成

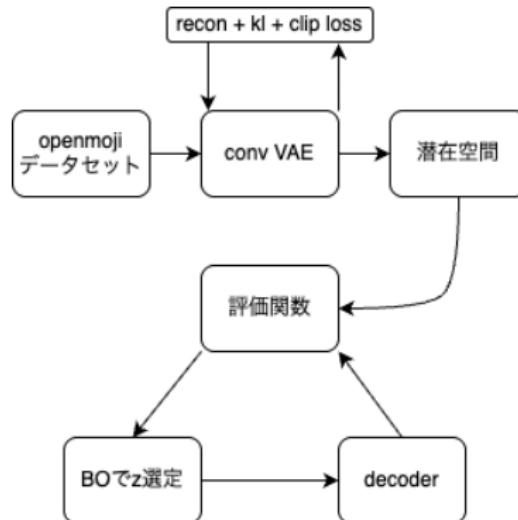
アプローチ

- **データセット:** openmoji の 4192 個の絵文字画像
- **初期入力:** ポジティブプロンプト, ネガティブプロンプト
- **フィードバック入力:** ユーザーと評価関数からのフィードバック
- **出力:** 人間の感性に合致する画像

アーキテクチャ

VAE+BO フレームワークの概要

- **VAE**: テキストプロンプトから画像を生成するための潜在空間 z を学習.
- **BO**: 潜在空間内での最適な z^* を探索し、評価関数を最大化する.



56f3: VAE と BO を組み合わせたアーキテクチャ

VAE 性能テスト

Original (top) vs VAE reconstruction (bottom)



56f3: interpolation

56f3: VAE の再構成結果

VAE性能テスト

Posterior-z decode (realistic latent sample)



z のサンプリング方法

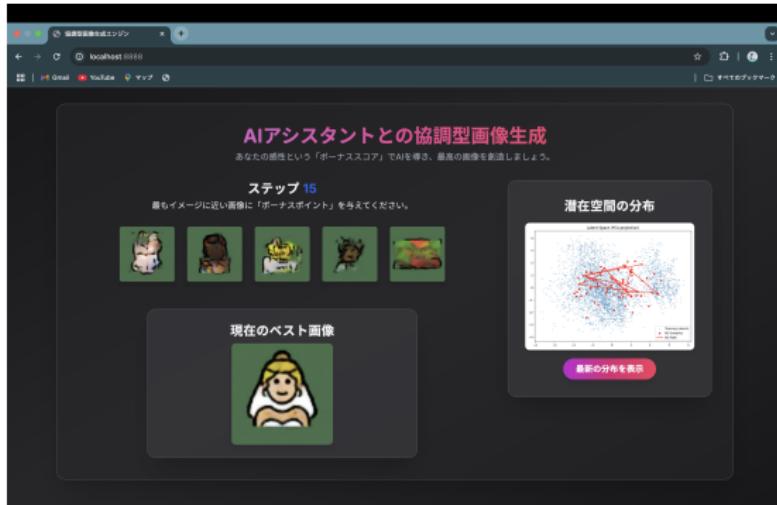
- $z = mu + eps * std$
- eps は $N(0, I)$ からサンプル

56f3: ちょっとズレた z からのデコード結果

開発成果物

BOによる探索の様子

- 画像の探索過程を示す

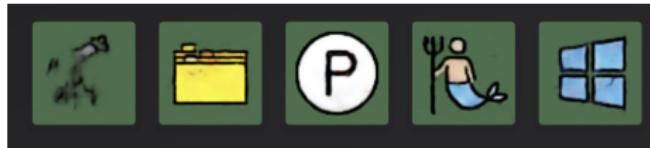


56f3: BOによる探索のデモ動画（実演予定）

結果と考察

結果：探索の失敗

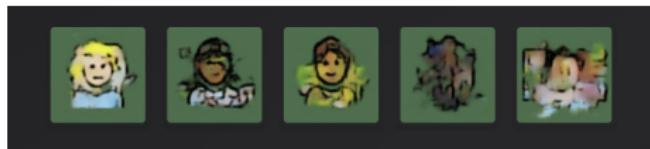
BO の最初のステップでは、整った画像が選ばれているが、5 ステップくらいで崩壊している。



56f3: 1step 目

考察：なぜ分子探索と違うのか？

分子特性の評価関数は比較的なめらかである一方、画像の「創造性」や「審美性」の評価ランドスケープは極めて複雑で病的な形状を持つため、BO の代理モデルがうまく機能しない。



56f3: 5step 目

考察：BO と生成＆ランク付け，どちらが優れているか？

優劣ではなく、問題領域との相性が重要であると感じた。

特性	分子探索・材料科学	画像・テキスト生成
目的	単一の最適解の発見	多様な高品質解の発見
評価関数	高コスト・比較的滑らか	低成本・極めて複雑
探索空間	物理法則に支配	構造が不明瞭・広大
最適な戦略	ベイズ最適化 (BO)	生成＆ランク付け

結論

BO は、評価コストが高い領域での「最適化」に特化した強力なツールである。一方、創造的タスクには、まず多様な候補を生成する能力が不可欠であり、異なる思想が求められる。

まとめ

- VAE+BO フレームワークの適用限界を，画像生成タスクを通じて実践的に探求した.
- BO の「探索」と，拡散モデルの「生成」という，生成 AI における 2 つの異なる思想的アプローチを実装レベルで比較・考察した.
- この「生成＆ランク付け」という思想は，近年の **GFlowNet** にも通じる.
GFlowNet は，単に高評価なサンプルを得るだけでなく，その評価値の分布に比例した確率で多様な候補を生成するポリシーそのものを学習する，より洗練されたアプローチである.

Appendix: 評価式の全体像

BO が最大化する目的関数（スカラー評価値）

$$f(z) = \underbrace{w_{\text{pos}} s_{\text{clip},\text{pos}}(z)}_{\text{CLIP 正例}} - \underbrace{w_{\text{neg}} s_{\text{clip},\text{neg}}(z)}_{\text{CLIP 負例}} - \underbrace{w_{\text{face}} p_{\text{face}}(z)}_{\text{顔検出ペナルティ}} + \underbrace{b_{\text{human}}(z)}_{\text{Human Feedback}}$$

- z : VAE 潜在ベクトル
- 係数 $w_{\text{pos}}, w_{\text{neg}}, w_{\text{face}}$ はハイパーパラメータ
- 人間が選択したサンプルには $b_{\text{human}} = +10.0$ のボーナス

Appendix: CLIP スコアとは？

CLIP (Contrastive Language – Image Pretraining) は、画像とテキストを共通の埋め込み空間に写像し、そのコサイン類似度を数値化できるマルチモーダルモデルです。

CLIP スコアの計算

$$s_{\text{clip}} = 100 \times \cos(f_{\text{img}}, f_{\text{text}})$$

- 画像特徴 f_{img} とテキスト特徴 f_{text} のコサイン類似度に 100 を掛けた値.
- 値が大きいほど「プロンプトの意図に合致した画像」とみなせる.
- この実験では **ポジティブプロンプト**に対するスコアを加点、**ネガティブプロンプト**に対するスコアを減点項として利用.

Appendix: 各評価項目の詳細

1. CLIP ポジティブスコア

$$s_{\text{clip},\text{pos}}(z) = 100 \cos(\mathbf{f}_{\text{img}}(z), \mathbf{f}_{\text{text},\text{pos}})$$

2. CLIP ネガティブスコア（任意）

$$s_{\text{clip},\text{neg}}(z) = 100 \cos(\mathbf{f}_{\text{img}}(z), \mathbf{f}_{\text{text},\text{neg}})$$

3. 顔検出ペナルティ

$$p_{\text{face}}(z) = \begin{cases} 1 & (\text{顔が検出できない}) \\ 0 & (\text{顔が検出できる}) \end{cases}$$

4. Human Feedback ボーナス

ユーザーが「良い」とクリックした画像のみ

$$b_{\text{human}}(z_{\text{selected}}) = +10.0$$

Appendix: 評価アルゴリズムのフロー

- ① **生成:** VAE から z をサンプリングし画像を生成
 - ② **自動評価:**
 - ① CLIP で $s_{clip, pos}, s_{clip, neg}$ を計算
 - ② OpenCV で顔検出 $\rightarrow p_{face}$
 - ③ **人間からのフィードバック:** GUI 上でユーザーが気に入った画像をクリック
 - ④ **スコア統合:** 前スライドの式で $f(z)$ を算出
 - ⑤ **BO 更新:** GP で $p(f | \mathcal{D})$ を推定し, qUCB により次の z 候補を提案
- 自動化と人間の感性を統合したインタラクティブ最適化ループ

Appendix: 「生成＆ランク付け」アーキテクチャを試してみた

戦略転換

「未知の最適点を探す」という BO の思想を放棄し、代わりに「高品質な候補を大量生産し、その中から選別する」という、現代的な生成 AI の思想へ転換する。

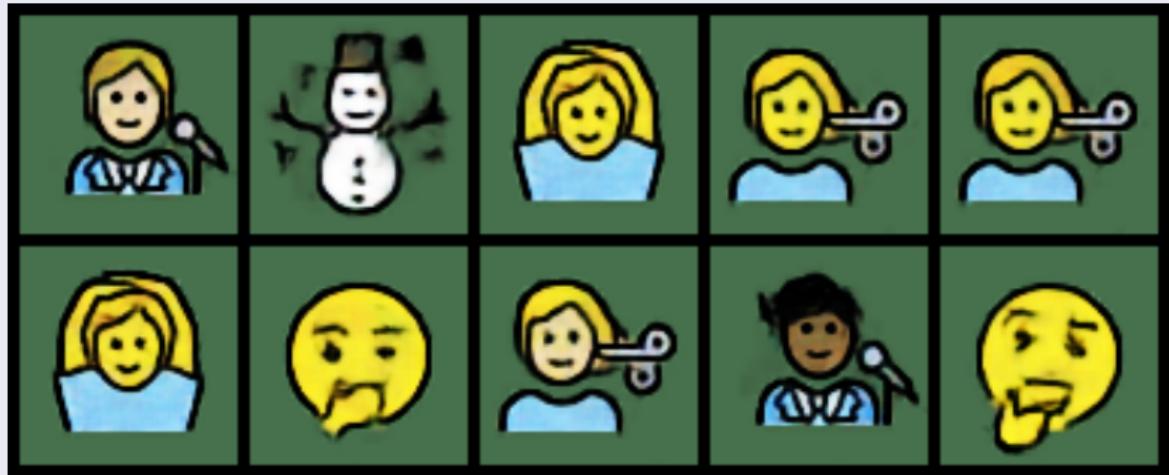
- ① **候補生成 (Generate)**: 潜在 z を教師に訓練した**拡散モデル**を用いて高品質な潜在ベクトルを生成し、デコーダーで候補 z を 1000 個生成。
- ② **評価・選別 (Rank & Select)**: CLIP スコアや各種ペナルティから成る評価関数で全候補をランク付けし、上位を選出。

アーキテクチャの意義

これは、高品質な生成を担う**生成モデル（拡散モデル）**と、評価を担う**評価モデル（CLIP）**の役割を明確に分離するという点で、Stable Diffusion と同じ

生成結果（拡散モデル）

生成例：「A blue emoji made of ice, freezing cold」



56f3: 「生成＆ランク付け」アーキテクチャによる最終結果。画像の崩壊なく、安定して多様な候補が得られた。しかし、既存の絵文字とほぼ同じであり、新規性のある絵文字を生成するには、もっと大量の VAE の学習データが必要であるとわかった。