

進化型画像生成システムを用いた視覚言語モデルの 音象徴的感性に関する分析

Analysis of Sound Symbolism in Visual Language Models Using Evolutionary Image Generation Systems

進藤 稜真^{†,‡}, 飯塚 博幸[‡]
Ryoma Shinto, Hiroyuki Iizuka

[†] 北海道大学 情報科学院, [‡] 人間知・脳・AI 研究教育センター
Graduate School of Information Science and Technology, Hokkaido University,
Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University
shinto.ryoma@gmail.com

概要

本研究では, Vision Language Model (VLM) が人間とどの程度類似した音象徴的感性を持つかを分析する。実験には, 人間の評価に沿って画像を進化させるシステムである CONRAD をベースに, 新たに VLM の評価も反映可能な進化型画像生成システムを構築して分析を行った。実験の結果, VLM は新たに作成した疑似単語を対象にした場合も含め, 人間と類似した音象徴的感性を示すことが確認された。

キーワード: 大規模言語モデル (LLM), 視覚言語モデル (VLM), 感覚 (Sensibility), 音象徴 (Sound Symbolism)

1. はじめに

GPT に代表される大規模言語モデル (LLM: Large Language Model) は, 大規模なテキストデータの学習により, 様々な言語処理タスクに汎用的に適用可能となった (Brown et al., 2020)。さらに, 言語に加えて画像の学習を行った視覚言語モデル (VLM: Vision Language Model) も登場し, 画像キャプションや視覚質問応答 (VQA) など, 様々なマルチモーダルタスクにおいて人間と同等のパフォーマンスを発揮している (Radford et al., 2021)。

しかし, 人間と同等の文章生成やタスク遂行が可能となっても, 人間の感性を理解し表現できるかどうかは定かではない。そこで本研究では, 「音象徴」という感性的側面に焦点を当て, VLM が人間とどの程度類似した音象徴的感性を持つかを明らかにする。

音象徴とは文字の音そのものが, ある特定のイメージを喚起する事象のことを指す。代表的な実験としては, 「bouba / kiki 実験」 (Ramachandran et al., 2001) がある。この実験では, 丸みを帯びた曲線的な図形と尖った直線的な図形に対し, ほとんどの被験者が丸みを帯びた図形を「bouba」と答え, 尖った図形を「kiki」と回答した。この現象は, 言語や年齢に関わらず共通して

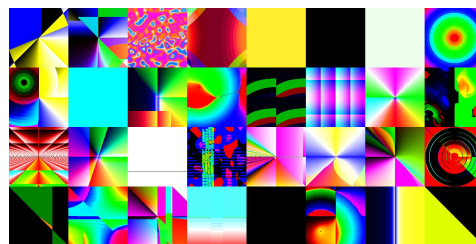


図 1: CONRAD を用いて生成した画像の例。この図は, 計 32 枚の画像を横 4 行, 縦 8 列に並べたものである。このように, 様々な図形が描かれた画像を生成できる。

みられる現象であり (Maurer et al., 2006), 文字の音が言語・聴覚的な意味を持ち, 人間の感性に影響を与えていることを示唆している。

特に感性のような定量化の難しい対象の場合, 高性能汎用 VLM はその巨大なパラメータ数や, 大規模な学習データ, 内部の詳細が非公開などの要因により, これらの観点から分析・解釈を行うことは現実的ではない。そこで本研究では, 進化の自然選択の性質を利用したアプローチを提案する。すなわち, VLM の評価結果を反映して進化した画像を分析することで, VLM の感性について明らかにする。

本研究によって得た, VLM の感性的側面やその分析手法についての新たな知見は, 実社会における VLM の制御と倫理的運用に役立つことが期待される。

2. 提案手法

本システムは「CONRAD」*1をベースに構築した進化型画像生成システムである。CONRAD はウェブ上で公開されており, サイトの訪問者による画像への 11 段階評価 (0 から 5 の 0.5 刻み) によって, 世代を重ねるごとに評価者の選択を反映させた画像へと進化するシステムである。CONRAD では, A, T, G, C の 4 種類の文字からなる計 n 文字のゲノム文字列を交叉・突然変

*1 <https://genetic-algorithms.com/>

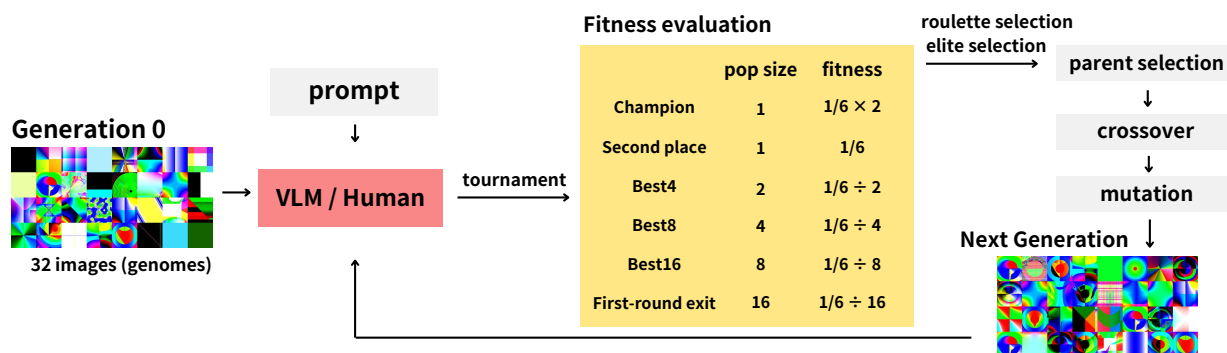


図 2: 構築したシステムの概要図。ゲノムから画像を生成した後、VLM または人間にプロンプトと画像を 2 枚ずつ提示しトーナメント戦を行う。各個体 (各画像) にはその順位に応じた適応度が割り振られ、適応度をもとに親個体を決定した後、交叉・突然変異を行う。ここまですべてを一代として、これらの処理を繰り返すうちに評価者の感性に沿った画像へと進化していく。

異させることで画像を進化させる (図 1)。ゲノム文字列から画像が生成される処理は次の通りである。まず、ゲノム文字列内の連続した 3 文字のまとまりを「コドン」とする。全部で 64 種類あるコドンにはそれぞれ数学的な関数が定義されている。最初に生成したランダムな初期値の正方形行列に対し、ゲノム内のコドンに対応する関数を適用し続け、全ての関数を適用した後の行列から画像を生成する。

この CONRAD をベースとして構築したシステムの概要図を図 2 に示す。本研究ではこのシステムを、人間だけでなく VLM が評価を行えるように次のように変更した。まず、一つの世代には m 枚の画像 (m 個のゲノム) が存在し、評価者には質問用のプロンプトと共に 2 枚ずつ画像が提示される。評価者は提示された 2 枚の画像のうち、質問プロンプトに合致する方を選択する。これを一つの世代内で優勝を決定するまでトーナメント戦を実施し、順位に応じてスコアを割り振る。その後、スコアをもとに確率的に親個体 (画像) を決定し、交叉・突然変異を経て次世代の個体を決定する。ここまですべてを一代として、これらの処理を繰り返すうちに評価者の感性に沿った画像へと進化していく。

■**適応度評価** その画像がどれだけ評価されたかを表す「適応度」の設定方法は次のとおりである。まずトーナメントの結果に基づき 32 個体を 6 つのグループに分ける。「優勝 (1 個体)」「準優勝 (1 個体)」「ベスト 4 (2 個体)」「ベスト 8 (4 個体)」「ベスト 16 (8 個体)」「1 回戦敗退 (16 個体)」である。それぞれのグループには 1 をグループ数で割ったスコア (この場合は $1/6$) を割り振り、さらにグループに割り振られたスコア ($1/6$) をグループ内の個体数で割ったスコアが各個体のスコアとなる (図 2)。また、優勝した個体は 2 倍のスコア (この場合は $2/6$) を獲得する。

■**親個体の決定** 適応度が高いほど選出される確率が高くなる「ルーレット戦略」を用いて、次世代の親個体

を決定し k 世代にわたって進化を繰り返す。また、優勝、準優勝となった 2 個体については、それぞれ交叉や突然変異処理を行わずに、次の世代のトーナメント戦において第一シード、第二シードとなるエリート戦略を取った。これにより、評価者が最後まで選び抜いた画像は変化せずに次の世代に継承されるため、ユーザーの評価を反映させる画像へと進化しやすくなる。

■**交叉** 交叉では、2 つの親個体のゲノムにおいてランダムにゲノム文字列の 2 点を決定し、その点を基に文字列を入れ替える処理を行う (二点交叉)。

■**突然変異** 突然変異では、A, T, G, C の文字列であるゲノムに対して、各文字を一定の確率で他の文字に変更する。また、一世代の個体数の十分の一にあたる個体は、初期世代の個体と同様に完全にランダムな数値を用いて新規生成される。これによって、評価者の感性を反映させつつも収束を防ぎ、探索の幅を広げることで多様な画像を生成することが可能になる。

3. 実験

本実験では「bouba/kiki 効果」をもとに、VLM と人間の音象徴的感性が類似しているかを検証する。すでに「bouba/kiki 効果」の内容が学習データとして含まれている可能性を考慮して、全く新たな疑似単語の組を作成し、それらを用いた分析も行った。

疑似単語の作成方法は次の通りである。まず (McCormick et al., 2015) に基づき、尖ったイメージを連想させる文字と丸いイメージの文字のグループに分ける。尖ったイメージの母音は「i/e」、子音は無声閉鎖音の「t/k/p」である。また、丸いイメージの母音は「u/o」、子音は共鳴音の「m/n/l」である。それぞれのグループにおいて、これらの文字から構成される 4 から 6 文字の単語を総当たりで作成し、既存の単語と重複するものを除外した後、最終的に尖ったイメージを持つ疑似単語として「tikike」、丸いイメージを持つ疑似単語と

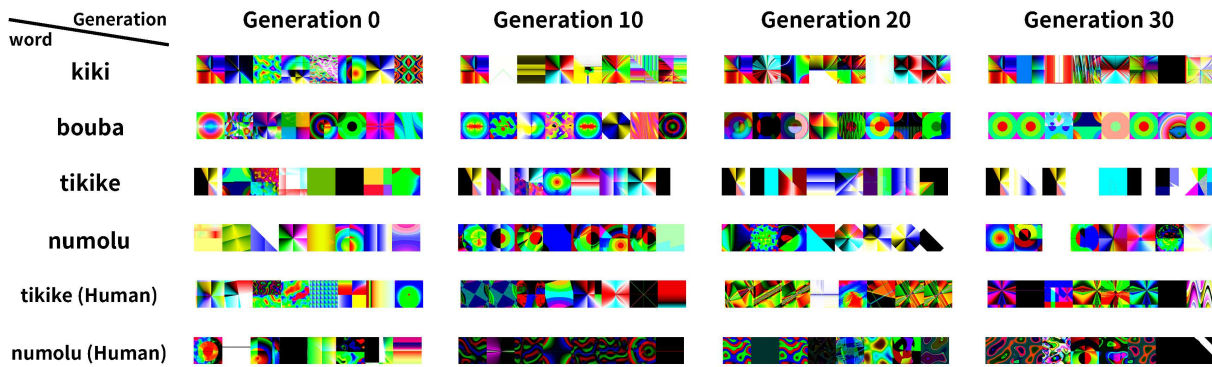


図3: 単語別の各世代上位8枚の画像. それぞれ左から, Champion(1枚), Second place(1枚), Best4(2枚), Best8(2枚) という順番に並べている. ここでは, 第0世代, 第10世代, 第20世代, 第30世代の推移を示す. それぞれの単語から喚起されるイメージに沿って, 画像が進化する様子を観察できる.

して「numolu」を選定した.

以上の疑似単語の組 (tikike/numolu) と kiki/bouba を, 以下のプロンプトの「word1, word2」と置き換えて評価者に提示する.

プロンプト

入力した画像は2つの画像が左右に結合されています.
質問
左側の画像と右側の画像において, どちらが「word1」でどちらが「word2」ですか?
図の形状に注目し, 「word1」だと思える方を教えてください.
左の画像を「word1」とするならば「1」を, 右の画像を「word1」とするならば「2」を出力してください.
int 型の整数1文字以外は何も出力しないでください.
出力:

画像生成システムのパラメータは, 世代数を30, 1世代の個体数を32, 突然変異率を0.1に設定した.

以上の実験設計で, VLMと人間を対象にした以下の2つの実験を行った.

■実験1 - VLM モデルは gemini-1.0-pro-vision-001(Gemini-Team et al., 2023) を用いた. 出力の一貫性を保つため, temperature パラメータは0に設定した.

■実験2 - 人間 新しく作成した疑似単語の「tikike, numolu」は, (McCormick et al., 2015) に基づき作成されたものの, 実際に人間にどのような形状の印象をあたえるかは定かではない. そこで, tikike/numolu の組を用いて人間を対象とした実験を行った. 被験者は男性(22歳), 女性(23歳) 各一名である.

次に, 以上の実験1,2で生成された画像の形状特性を評価する分析を行った. 各画像が尖った形状(直線的な図形)であるか, 丸みを帯びた形状(曲線的な図形)であるかを次のように評価する. まず, 画像をグレースケールに変換し, Canny エッジ検出アルゴリズムを用いて画像内のエッジを検出する. 次に, ハフ変換

を適用することで, 画像内の直線と曲線をそれぞれ検出し, 各世代ごとにそれらの本数を平均する. 平均直線数を横軸, 平均曲線数を縦軸とするグラフを作成し, 原点を通る傾き1の直線を基準として, どちらの領域により多くの点がプロットされるかを観察した.

なお, この分析においては各世代の上位8個体(画像)のみを対象にする. その理由は, 交叉や突然変異によって形状が大きく変化した画像も同世代に含まれるため, トーナメントを最低2回勝ち上がった画像(ベスト8以上)を分析対象とすることが, VLM の選択傾向に対する適切な評価となるためである.

■Result まず, 実験1および実験2にて生成された画像を図3に例示する. 図3では, 各世代上位8枚の画像を左から順に Champion(1枚), Second place(1枚), Best4(2枚), Best8(2枚) と横に並べ, 10世代ごとに示したものである. なお, 各単語においてそれぞれ5回実験を行ったが, 紙面の関係上1回目の結果のみを例示するに留める. 図3からは, 「kiki, tikike」を用いた場合は, 直線的で尖った図形へ変化し, 「bouba, numolu」の場合は曲線的な図形に変化していく様子が見られる. これは, 人間の被験者に対して行った実験(tikike(Human), numolu(Human))からも同様の傾向を読み取ることができる.

次に, これらの結果を定量的に評価したグラフを図4に示す. このグラフは, 縦軸に検出した平均曲線数, 横軸に検出した平均直線数を設定して各世代の値をプロットしたものであり, プロットの偏りから全体を通して画像の図の形状が尖っている(直線的)か, 丸みを帯びている(曲線的)かを確認できる. 図4では人間とVLMの場合のいずれにおいても, 「kiki, tikike」を用いた場合には, 基準となる直線より下側に多数の点がプロットされる傾向を見ることができる. また, 「bouba, numolu」を用いた場合には, 基準線より上側にプロットされる点が多く見られた.

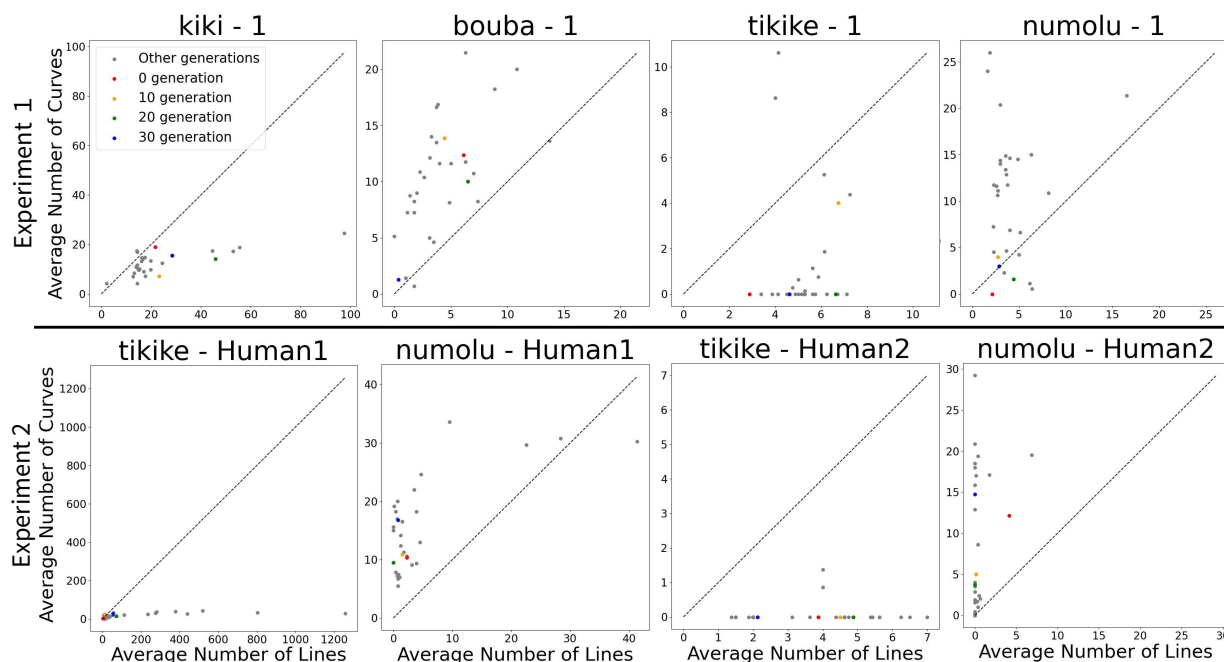


図4: 実験1と2の結果。縦軸は平均曲線数, 横軸は平均直線数を表し, プロットされた点は各世代の値を表している。上段はVLMが評価を行った実験1の各単語の結果, 下段は人間の被験者二人が評価を行った実験2の結果である。傾き1の対角線を基準として, 点の偏りを視覚化することで, 各単語に関連する音象徴と図形の形状の関係について, 人間とVLMが持つ傾向を理解することができる。グラフの第0, 10, 20, 30世代の点には異なる色を付与している。

4. 考察

まず実験2の結果からは, 人間の被験者に対して疑似単語「tikike/numolu」を提示したところ, 両者とも「tikike」は直線的, 「numolu」は曲線的な画像へと変化する傾向が見られた。この結果は, これらの疑似単語が人間に対して「kiki」「bouba」と同様の音象徴的感觉を与えていることを示唆する。

その上で, 実験1では, VLMも同様に「kiki, tikike」を提示すると直線的な尖った画像へ, 「bouba, numolu」を提示すると曲線的な丸みを帯びた画像へ変化する傾向が見られた。この結果は, VLMと人間の音象徴的感觉について一定の類似を示唆している。

音象徴は, 発音時の口の形状や音の高さなど, 身体感覚や音韻の特徴が要因の一つとして考えられている (Ramachandran et al., 2001)。それにもかかわらず, 身体を持たないVLMが, 新たに作成した疑似単語でも人間と一定の類似を示したことは興味深い。一つの仮説としては, 本研究で用いたVLMは単語単位ではなく, サブワード単位で単語ベクトルを構築している。そのため, VLMが文字あるいは音節単位で音象徴に関する知識を学習したことで, 本実験のような結果が得られた可能性が考えられる。例えば, 「母音「u/o」や共鳴音「m/n/l」といった文字が丸みを帯びた図形のイメージと結びつけられる」という知識を既に学習しており, その知識をもとに単語を構成する文字に注目することで, 画像がより曲線的である方を選択できたのではな

いかと考えられる。よって, 今後の課題として, VLMが学習によって獲得した単語ベクトルから形状に関する意味を抽出し, 実験で用いた単語がどのような形状特徴をベクトルとして表現しているかを分析する。また, 使用したモデルや実験で用いた単語が限定的であり, さらに被験者の数も少ないため, 得られた結果がどの程度一般性を持つのかについては慎重な議論が必要である。さらに, 本システムのパラメータ設定が結果に与える影響についても, より詳細な検討が求められる。

謝辞

本研究はJSPS科研費JP23K25163の助成を受けたものです。

文献

- Brown et al., (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33, 1877-1901.
- Radford et al., (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748-8763.
- Ramachandran et al., (2001). Synaesthesia—a window into perception, thought and language. *Journal of consciousness studies*, 8(12), 3-34.
- Maurer et al., (2006). The shape of boubas: Sound-shape correspondences in toddlers and adults. *Developmental science*, 9(3), 316-322.
- McCormick et al., (2015). Sound to Meaning Mappings in the Bouba-Kiki Effect. *CogSci*, 1565-1570.
- Gemini-Team et al., (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, <https://doi.org/10.48550/arXiv.2312.11805>.