

脳波デバイスを用いた大規模言語モデルとの視覚的コミュニケーション形式の提案 ”こそあど言葉”を含む指示語を用いた画像を通じた対話

Proposal for a Visual Communication Format with a Large-Scale Language Model Using an EEG Device: Dialogue Through Images with Demonstrative Words Including ”KOSOADO Words”

茶谷 瑛佑[†] 佐藤 啓宏[†]
EISUKE CHATANI[†] YOSHIHIRO SATO[†]

1. はじめに

ChatGPT や Claude3 をはじめとした大規模言語モデル (Large Language Models: LLMs) の発展により, 多くの対話型アプリケーションが開発されている. 2024 年 6 月現在, これらのプログラムは画像の内容を文章で表現することにも対応している. このことにより, 大規模言語モデルと画像を通じた人間らしいコミュニケーションをとることが可能なレベルにあると考えられる. しかし, 人間同士のコミュニケーションでは, ある程度, 曖昧な指示や表現であっても相手と意思疎通を図ることができるが, 大規模言語モデルがどのくらい曖昧で視覚的な会話を許容できる能力を有しているか不明である. 曖昧なコミュニケーションの例として, 相手に対して「これ何?」や「あれいいよね」などといった「こそあど言葉」を用いたコミュニケーションを行う場合がある. このコミュニケーションが成立する理由の一つとして, 自分と相手が同じ情報を共有していることが挙げられる. 人間同士で情報の共有を行う場合, 事前知識や視覚情報, 表情などといった言語以外の方法で共有されることが多い. しかし, コンピュータを相手にする場合, 違う手段を用いることができるという優位性がある. 例えば, 脳波デバイスを用いることでコンピュータとずっと自然にコミュニケーションを行うことができる.

そこで, 本研究は, 脳波デバイスを装着した人と大規模言語モデルを用いて曖昧な視覚的コミュニケーションを行い, その評価を行うシステムを作成することを目的とする. 具体的には, 写真とそこから抽出された情報に対して脳波デバイスを用いて入力を行い, その部分を大規模言語モデルに説明してもらうことでコミュニケーションを図る. また, 入力された部分を中心として, 写真の見える範囲を徐々に狭めていくことで, 会話を深掘りすることを図る. さらに大規模言語モデルが出力した説明文を, 面白さの観点で評価を行う方法を提案し, その結果を示す.

2. 関連研究

脳波 (Electroencephalograph: EEG) を計測してコマンドに変換するブレインコンピュータインターフェース (BCI) の

研究は古くから存在し, Vidal[1] は視覚刺激に対する脳の反応を分析する Visual Evoke Potential (VEP) という手法を開発し, EEG を使ってコンピュータや他の機器を制御することの可能性を示した. この技術を発展させ, Middendorf ら [2] は点滅刺激を注視することで誘発される Steady-State Visual Evoked Potential (SSVEP) を利用した BCI を開発し, 脳波でロボットを動かす研究の先駆けとなった.

BCI を利用した研究は現在でも多くあり, Wang ら [3] は人とロボットの共同作業において, 声や動作ができない状況で脳波デバイスを用いることで直感的なロボット制御を可能にすることを説明している. ほかに, Janapati ら [4] は BCI を利用してコンピュータのナビゲーションと制御を行うアプローチを提案し, ハンディキャップがある人でも手足の動きを想像することで操作できるようにした.

また, EEG は, 感情状態の分析にも利用される. 例えば, Tuib ら [5] は EEG 信号を用いた感情検出のために, バイナリーモスフレーム最適化 (BMFO) アルゴリズムと畳み込みニューラルネットワーク (CNN) を組み合わせたシステムを提案した. このシステムは, 従来の音声のトーンや表情の分析に基づく方法に比べて, 精度と真陽性率が向上していることが分かった.

これらの研究動向から, 従来は刺激に対する単純なクラス分類を行う例が多かったのに対し, 近年は感情のような曖昧なものの評価にも使われるようになっている傾向が読み取れる.

一方, 大規模言語モデルが生成した文章の評価を行う研究も盛んに行われている. 例えば, Saunders ら [6] は AI モデルによるトピック要約タスクにおけるモデルの生成, 識別, 批評の能力を体系的に測定した研究を行った. この研究では, モデルの生成能力と批判能力のギャップを改善するためのアプローチを提案している. ほかに, Peng ら [7] は LLM-Augmenter という LLM の応答文章の正確性を上げるための LLM 拡張システムの提案を行った. このシステムでは, Knowledge F1, BLEU, BLEURT などの様々な評価指標を用いて, LLM の応答文章を正確性, 流暢さ, 文脈適合性などの項目で評価した.

このように, 文脈の整合性などを確認するのに AI 技術を用いていることがわかる. しかし, 会話内容の面白さを評価している研究は少ない. その理由の一つとして, 面白さというのは主観的なものであり, 客観的に評価する基準を確立するのが難しいからだと考えられる.

[†] 京都先端科学大学 Kyoto University of Advanced Science

そこで、本研究では脳波デバイスを用いて LLM と非言語的かつ直感的なコミュニケーションを図るとともに、LLM に会話内容の面白さを評価させるシステムを考案し、面白さの基準となるものを調べる。

3. 脳波デバイスを用いた LLM との視覚的コミュニケーション

図 1 は画面上に特徴的な部分を検出する方法を表している。初めに一枚の写真に対して特徴点検出を行い特徴的な部分の座標を受け取る。その後、画面上にそれぞれの座標をもとに特徴点を配置する。特徴点検出方法は物体検知アルゴリズムや画像の特徴的な部分をクラスタリングするアルゴリズムなど様々な方法が適用可能である。

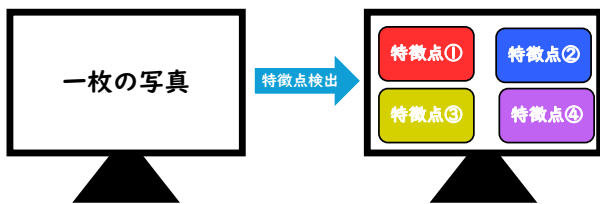


図 1 画像特徴点の検出。

また、図 2 は、LLM がチャットボットとしてユーザーが選択した画面上の特徴部分に関する内容の説明を行う図である。初めに画像の特徴的な部分を検出し、特徴点の位置に脳波の刺激マーカ―を表示する。ユーザーは特徴点 1 つを選択し、脳波デバイスを用いて入力を行う。LLM は入力された特徴点に関する内容や関連情報の説明を行う。その後、選択した特徴点を中心に画面がズームアップされ、再度特徴点の検出が行われる。この動作を繰り返すことで、ユーザーは画面上のさまざまな特徴点から細かな特徴点に話題が移動することで会話内容を深掘りすることができる。また、ユーザーが一定時間入力を行わなかった場合、元の画像に戻すことで話題を変えることができる。

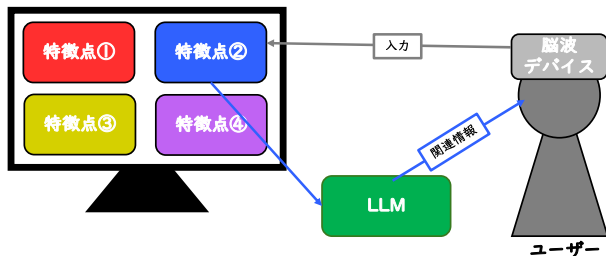


図 2 ユーザーが注目する画像特徴点に関連する情報を LLM が提供。

この形式の利点は、脳波デバイスを用いることによって「これ何」などの曖昧な指示を直感的な入力で可能にしている点にある。ユーザーは、興味に応じて画像の一部分を眺めるだけで、画像がズームアップされて、それに伴って会話内容が変わっていく。この仕組みを用いて、ユーザーの興味関心に基づいた会話へ深掘りしていくことを目指す。

4. 実験

本研究の実験では、画像に対して特徴点検出を用い、Unity を用いて画面上の特徴点の位置に脳波の刺激マーカ―である NeuroTag を表示し、脳波デバイス NextMind[8] で選択できるようにする。そして、選択された特徴部分について ChatGPT API が生成した文章を Python のライブラリである gTTS と pygame を用いてテキストを音声に変換し、説明を行う。また、ChatGPT を用いて、ChatGPT 自身が説明した文章の面白さの採点評価させ、結果の検討を行う。

本実験では 3 つの検出方法、

- (A) YOLOv8 による方法
- (B) Saliency Map による方法
- (C) ChatGPT による方法

を用いて、会話の面白さに違いが得られるかを調査した。

4.1 評価項目

本実験の評価項目は、(A),(B),(C) それぞれの特徴検出方法を用いて出力された文章を、ChatGPT API が面白さを基準に評価した点数である。

図 3 は ChatGPT API に説明文章の評価を行わせるためのプロンプトである。それぞれの関数のプロンプトにより点数と理由を出力させる。これらの出力結果を用いて評価を行った。

```
Score関数のプロンプト：
prompt=f“次の文章を面白さの観点で1から10の範囲で点数をつけてください。また点数だけを出力してください。:¥n¥n¥” {text}¥””

Reason関数のプロンプト：
prompt=f“次の文章を面白さの観点で評価すると{score}点です。理由を教えてください。:¥n¥n¥” {text}¥””
```

図 3 Score 関数と Reason 関数のプロンプト。

4.2 実験内容

以下が実験の流れである。

- (i) (A),(B),(C) のいずれかの特徴検出方法を用いて画面上の特徴を座標リストとして保存する。
- (ii) 座標リストをピクセル座標から Unity 座標に変換する。
- (iii) Python と Unity 間でソケット通信を行い、座標リストを Unity に送信する
- (iv) 受信した座標リストに基づいて NeuroTag を配置する。
- (v) 脳波デバイス NextMind を用いて NeuroTag に入力を行う。
- (vi) 入力された NeuroTag の座標を Python に送信し、その座標をピクセル座標に変換する。
- (vii) ChatGPT API が受け取ったピクセル座標をもとに注目した説明文を生成し、その文章の評価を行う。

この処理を繰り返し、ChatGPT と対話を行った。

図 4 は実験の様子を示している。ユーザーが脳波デバイス NextMind を用いて画面上の NeuroTag に入力を行い、ユーザーの選択に基づいて ChatGPT が説明を行う。

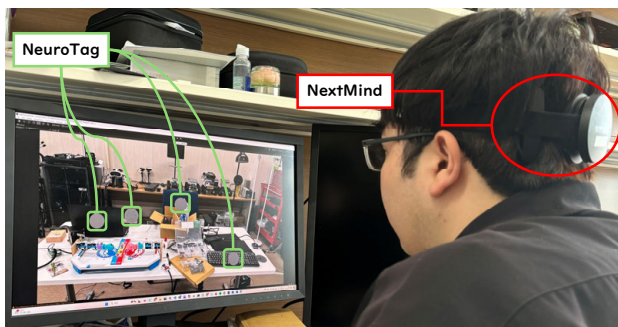


図 4 実験の様子。

また、図 5 は、Unity 上に表示される UI の一例である。画面上的様々な特徴的な部分に NeuroTag が配置されている。2 枚目ではユーザーが入力したピンク色の物体部分に画面がズームアップされ、新たな NeuroTag が配置されている。

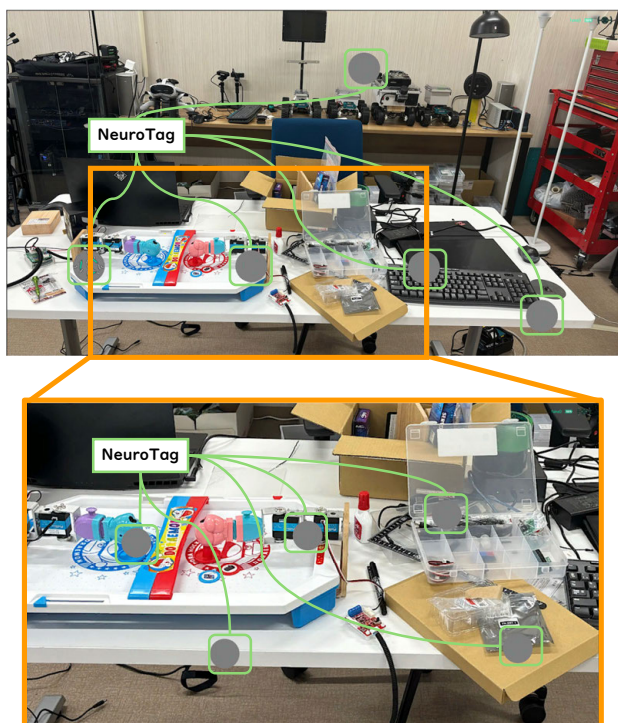


図 5 NeuroTag の配置の例とズームアップの様子。

次に、実験に用いた 3 種類の特徴検出方法を詳しく説明する。

(A) YOLOv8 を用いた特徴点の検出方法

この方法は物体検出アルゴリズム YOLOv8 を使用して画像内の特徴点を検出した。YOLOv8 は YOLO (You Only Look Once) シリーズのアルゴリズムのひとつで、物体検出タスクにおいて高速で高精度かつ画像解像度が低くても物体検出を安定に実現することができるという特徴を持つ。

以下の手順で特徴点の座標を取得した。

- (1) **モデルの準備:** 学習済み YOLOv8 モデルを読み込む。
- (2) **画像の入力:** 対象となる画像をモデルに入力する。
- (3) **物体検出:** 画像内の物体を検出し、そのバウンディングボックスを出力する。
- (4) **特徴点の出力:** 検出されたバウンディングボックスの中心

点を特徴点の座標として出力する。

この手法のメリットは、YOLOv8 によって画像内の主要な物体とその座標を特定し、座標を出力することができることである。

(B) Saliency Map を用いた特徴点の検出方法

この方法は畳み込みニューラルネットワークモデル VGG16 と Grad-CAM(Gradient-weighted Class Activation Mapping) を使用して画像内の特徴点を検出した。Grad-CAM とは特定のクラスに対する重要な領域を可視化する手法である。

以下の手順で特徴点の座標を取得する。

- (1) **モデルの準備:** 学習済み VGG16 モデルを読み込む。
- (2) **画像の入力:** 対象となる画像をモデルに入力する。
- (3) **特徴マップの抽出:** Grad-CAM を使用してモデルの最後の畳み込み層から特徴マップを抽出する。
- (4) **サリエンシーマップの生成:** 抽出された特徴マップに基づいたサリエンシーマップを生成する。
- (5) **クラスタリング:** サリエンシーマップ上の重要な領域を K-means クラスタリングを用いて複数のクラスタに分ける。
- (6) **特徴点の出力:** それぞれの分けられたクラスタの中心を特徴点の座標として出力する。

この手法のメリットは、物体認識に基づく方法のように明示化された事前知識を用いることなく、画像内の視覚的に注目すべき領域を特定し座標として出力することができることである。

(C) ChatGPT を用いた特徴点の検出方法

この方法は OpenAI の GPT-4-vision を使用して画像内の特徴点を検出した。

以下の手順で特徴点の座標を取得する。

- (1) **画像のエンコード:** 対象となる画像を Base64 形式にエンコードする。
- (2) **プロンプトの作成:** ChatGPT に対して、画像内の面白そうな位置の座標を比率で表現し、左を 0、右を 1、上を 0、下を 1 として指示する。
- (3) **特徴点の抽出:** 指定されたプロンプトに基づいて画像を解析し、面白そうな位置の座標を出力する。
- (4) **座標の変換:** 出力した座標を画像のピクセル座標に変換し、出力する。

この手法のメリットは自然言語を用いて画像内の注目すべき領域の候補を選定し、他の手法とは異なり意図的な出力を期待しないアプローチで特徴点を特定できることである。

4.3 実験結果

表 1 は、11 枚の画像に対して、それぞれ (A)(B)(C) の方法を用いてコミュニケーションを行い、得られた会話の面白さを 1 から 10 の点数で評価した結果を示しており、各画像に対する 20 回の試行の平均をとったものである。

この表を見ると、(B)Saliency Map による方法が最も高く 3.06 点、次に (C)ChatGPT による方法が 2.87 点、そして最も低かったのが (A)YOLOv8 による方法で 2.76 点だった。

表 1 (A)(B)(C) の特徴検出方法を用いて得られた出力の点数

画像 [枚目]	YOLOv8[点]	Saliency Map[点]	ChatGPT[点]
1	2.88	3.21	3.00
2	2.82	2.87	2.80
3	2.75	3.27	2.95
4	2.92	3.38	2.80
5	2.70	2.98	2.55
6	2.52	3.08	2.95
7	2.94	2.74	2.95
8	2.08	2.78	2.85
9	2.75	2.87	2.75
10	2.94	3.17	2.95
11	3.03	3.29	3.00
平均	2.76	3.06	2.87

5. 考察

実験結果では、YOLOv8 を用いた特徴検出に比べて Saliency Map を用いた特徴検出のほうが若干高い点数を取ることができた。これは、YOLOv8 が強力な物体検出アルゴリズムであり解像度にあまり影響されることなくトレーニング済みの物体を検出できてしまうために、2 回目以降も同じ部分が検出されてしまい、会話のバリエーションが広がらなかったためである。さらに、ラベル付けされた既知の物体に対してタグが付けられることになるため、その物体自身の説明のみを繰り返すことが多く、その結果として会話の深掘りが思うように進まなかった。

一方、Saliency Map を用いた特徴検出は YOLOv8 に比べて画像上のさまざまな部分に特徴点を検出された。中には人間が特徴的だと思わないようなところが検出されたため、会話の内容が転々とする事となった。ChatGPT を用いた特徴検出は、期待していたよりも点数が低かった。また、実験を 20 回試行する間に 30 回程度のエラーが発生するという問題が発生した。LLM の発展途中の過渡的な問題だと考えられるが、現時点では、Saliency Map を用いた特徴検出方法の方がコミュニケーションを円滑に進めることができた。

このことから、偶然や曖昧性が会話の広がりを引き出すきっかけになっていると推測され、面白さにつながる重要な要素であると考えられる。

5.1 評価のためのプロンプトの改良

全体的に平均点が 3 点前後という結果から、ChatGPT が評価する点数は間違っている可能性があるのではないかと考えた。そこで、会話中に説明を求めるプロンプトに「面白く説明して」と追加して再度実験を行ったところ、図 6 に示すような出力が生成された。

視覚的な情報だけでなく、ChatGPT 自身が「面白い」というプロンプトの指示からバリエーションの広がる説明を付け加えていることがわかる。点数も 7 点と高得点を獲得することができていた。このことから、ChatGPT による面白さの評価の指標はある程度正確であることがわかり、また、ただ説明を繰り返すような会話では高得点を獲得することは難しく、返答方法を具体的に指示するプロンプトの影響が大きいこともわかった。

この部分には、一見すると普段目にする事務用品の山がありますが、よく見ると小さな秘密基地のような雰囲気を感じさせる空間です。そこには壁に囲まれておりながら、四角い窓のような開口部があり、これがあることで監視塔の覗き穴のようです。その隙間からは、外の世界を伺いながら、こっそりと業務を遂行する謎の人物がそのような気配を醸し出しています。オフィスの片隅で生まれたこの小さな領域は、日常とは切り離された独自の世界を持ち、まるで大人の秘密基地のような役割を果たしているかもしれません。
Score: 7
Reason: この文章の評価が7点に止まった理由は以下の通りです。
1. **創造性と独創性**: オフィスの事務用品の山を小さな秘密基地に見立てる発想は面白く、創造性があります。しかし、他の題材も使用してさらに非現実的な要素を追加することで、より魅力的にする余地があります。
2. **共感性と魅力**: 「大人の秘密基地」という表現は多くの人が共感できるものですが、具体的なエピソードや具体像に欠けるため、読者がより具体的に想像し共感を深めるための要素が不足しています。
3. **展開と緊張感**: 文章全体は詩的で雰囲気をよく伝えています。大きなストーリー展開やサブライズ要素が少ないため、読み進めるうちに一部の読者が興味を失う可能性があります。
4. **情景描写**: 情景描写はある程度詳細ですが、さらなる具体性や独特な表現を加えることで、より一層読者を引き込むことができます。例えば、秘密基地の具体的なインテリアや使用するアイテムを描くことでリアリティを持たせることができます。

図 6 “面白く説明して”と説明プロンプトを変更した際の出力。

6. まとめ

本研究では、大規模言語モデルと脳波デバイスを用いて曖昧で非言語かつ直感的なコミュニケーションを行う仕組みを提案した。また、その説明の面白さの評価方法の提案も行った。実験として、画像上の特徴点に対して脳波デバイス NextMind で入力を行い、ChatGPT に説明文を生成させ、その結果得られた会話の面白さの評価を行った。実験結果として、明確に物体に特徴点があるものよりも、曖昧な場所に特徴点がある方が、LLM の面白い説明を引き出す事ができることが分かった。

また、本実験では 1080p のサイズの画像を用いて実験を行ったが、さらに高解像度の画像を使用することで、画像をズームしていくたびに新たな特徴が検出され、会話内容が変わっていく可能性がある。さらに、動画にもこのシステムを実装できれば、ユーザーのしている画面に対してリアルタイムで流れるような会話を展開することができると考えられる。

参考文献

- [1] J. J. Vidal, "Toward Direct Brain-Computer Communication," Annual Review of Biophysics and Bioengineering, vol. 2, no. 1, pp. 157-180, 1973.
- [2] M. Middendorf, et al., "Brain-computer interfaces based on the steady-state visual-evoked response," IEEE Trans. on Rehabilitation Engineering, vol. 8, no. 2, pp. 211-214, 2000.
- [3] L. Wang, et al., "Function block-based human-robot collaborative assembly driven by brainwaves," CIRP Annals, vol. 70, no. 1, pp. 5-8, 2021.
- [4] R. Janapati, et al., "Computer Navigation and Control Using BCI," in 2023 Inter. Conf. on Advanced & Global Engineering Challenges (AGEC), pp. 112-117, 2023.
- [5] T. A. Tuib, et al., "Convolutional neural network with binary moth flame optimization for emotion detection in electroencephalogram," IAES Inter. Journal of Artificial Intelligence (IJ-AI), vol. 13, no. 1, pp. 1172-1178, 2024.
- [6] W. Saunders, et al., "Self-critiquing models for assisting human evaluators," arXiv:2206.05802, 2022.
- [7] B. Peng, et al., "Check your facts and try again: Improving large language models with external knowledge and automated feedback," arXiv:2302.12813, 2023.
- [8] <https://github.com/Snapchat/NextMind>