大規模視覚言語モデルの質感知覚能力の分析

松田 陵佑 [1] , 塩野 大輝 [1], Ana Brassard [2,1], 鈴木 潤 [1,2,3]

matsuda.ryosuke.t4@dc.tohoku.ac.jp ([1] 東北大学, [2] 理化学研究所, [3] 国立情報学研究所)

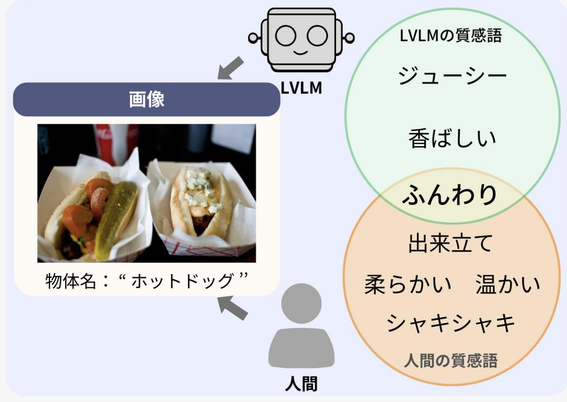
概要

- LVLMの**質感知覚能力**と， LVLM と人間との間の**質感知覚の整合性**の分析が目標.
- 質感知覚能力を評価する**分類タスク**と質感知覚の整合性を評価する**生成タスク**を設計.
- ｛画像,物体名,質感語｝ の３つの情報を結びつけた**質感データセット**を人手で作成.
- 既存の代表的なLVLMの中では， GPT-4oが高い質感知覚能力を有することを確認.さらに分類タスクの正解率が高いLVLM は，**人間との間の質感知覚の整合性**においても高いスコアを示すことを確認.

背景／動機

質感： 物性（光沢感・透明感）
状態（乾燥・凍結）
印象（美しい・醜い）

- 研究の意義：モデルと人間の間の整合性が取ることは、モデルが人間と同様の方法で知覚し、思考し、行動する際に重要.
- 目標
  1.既存の代表的なLVLMの**質感知覚能力**の分析
  2.**LVLMと人間の質感知覚の整合性**の分析



質感データセット構築

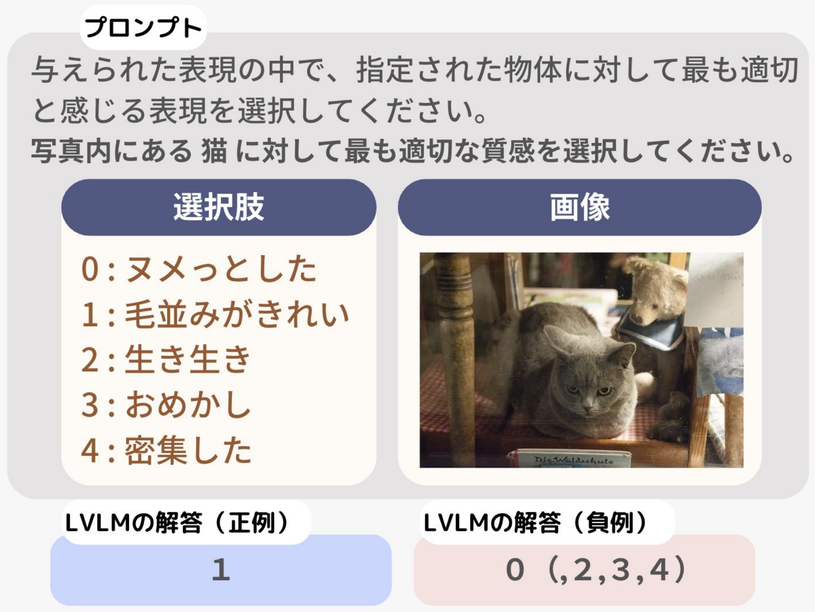
- COCOデータセットの{画像, 物体}のサンプルに，人手で質感語の情報を追加した**質感データセット**を作成.

実験

分類タスク

Q. LVLMは画像内の質感を正しく知覚できるのか？

- 画像内の物体に対し，複数の質感語から最も適切なものを選択するタスク
- 7種のLVLMおよび人手アノテーターに対して評価を行う.



結果

モデル	2択問題	5択問題
Random	50.00	20.00
Human	—	78.57
GPT-4o 2024-11-20	93.43	81.19
Llama-3.2 11BInstruct	57.31	45.07
Qwen2-VL 7BInstruct	85.37	61.79
LLaVA-OneVision 7Bov	78.21	62.09
LLaVA-NeXT 7B	64.48	33.43
Idefic2 8B chatty	66.27	39.40
LLaVA-1.5 7B	52.84	21.49

GPT-4oが最も高い正解率を示し，人間の平均正解率を上回った. LVLMの正解率は、2択・5択問題とも頑健な性能を確認.

生成タスク

Q. LVLMと人間の質感知覚の整合性の関係性は？

- 画像に含まれる物体に対してLVLMに質感語を生成させる.
- LVLMと人間の質感知覚の整合性を調べるため**平均質感語一致率**, **yes/no 人手判定スコア**，２つの評価指標で分析.

評価指標１

LVLMが生成した質感語と人間の質感語の共通部分の割合

$$\text{平均質感語一致率} = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{|S_{\text{Human}} \cap S_{\text{LVLM}}|}{|S_{\text{LVLM}}|}\right)$$

（S_{Human}:人間が書き出した質感語集合，S_{LVLM}:LVLMが生成した質感語集合 ）

評価指標２

LVLMが生成した質感語が人間にとって自然であるかの割合

$$\textit{yes/no 人手判定スコア} = \frac{1}{H}\sum_{h=1}^{H}\left(\frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_{h,i}}{w_i}\right)\right)$$

H:総アノテーター数, N:総サンプル数, w_i:LVLMが生成した質感語の数,
$y_{h,i}$:サンプル i においてアノテーター h がyesと回答した数

結果

モデル	平均質感語一致率	yes/no 人手判定スコア
GPT-4o	21.50	75.49
LLaVA-1.5 7B	11.93	57.75

GPT-4o が，yes/no判定スコアおよび平均質感語一致率の両方でLLaVA-1.5 7Bを上回る結果に.

分析と議論

- **分類タスクの正解率が高いLVLMは，生成タスクにおいても自然な質感語を出力する**傾向を確認.
- 質感データセット内の人間が回答した質感語を用いる分類タスクは，その都度人的コストをかけずにLVLMの質感知覚能力と人間知覚との整合性を評価できる有望な手法であることの可能性を確認.