P7-21

大規模視覚言語モデルに関する指示追従能力の検証

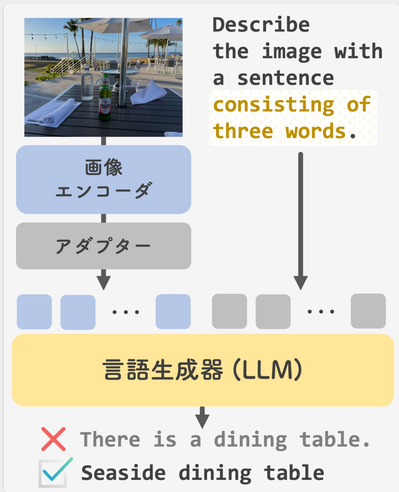
塩野 大輝[1], 宮脇 峻平[1], 田中 涼太[1,2], 鈴木 潤[1,3]

[1] 東北大学, [2] NTT人間情報研究所, [3] 理化学研究所

Paper Home

背景

- LVLMは組み込まれる前のLLMが有していた指示追従能力を示さず、タスク指示に従わない事例が**定性的**に観測される [Fu+, '23]
- 著者は既存の Visual Instruction Tuning データセットには出力形式に関する指示が含まれていない場合が多いことを観察



概要

- 追加学習後の LVLM の指示追従能力の低下を**初めて定量的**に示した
- **出力形式に関する指示を含む追加学習データセットを作成・使用**し、追加学習したモデルの指示追従能力を調査したところ、**LVLM の指示追従能力の低下に大きな影響を与えているのは、出力形式に関する指示の有無である可能性が高い**

提案法

(Visual) Instruction Tuning データセットの作成

指示追従能力評価用データセットの作成

Examples of evaluation datasets

If a movie review is **positive**,
you need to output "{label_0}".
If a movie review is **negative**,
you need to output "{label_1}".

Movie review: lovely and poignant.

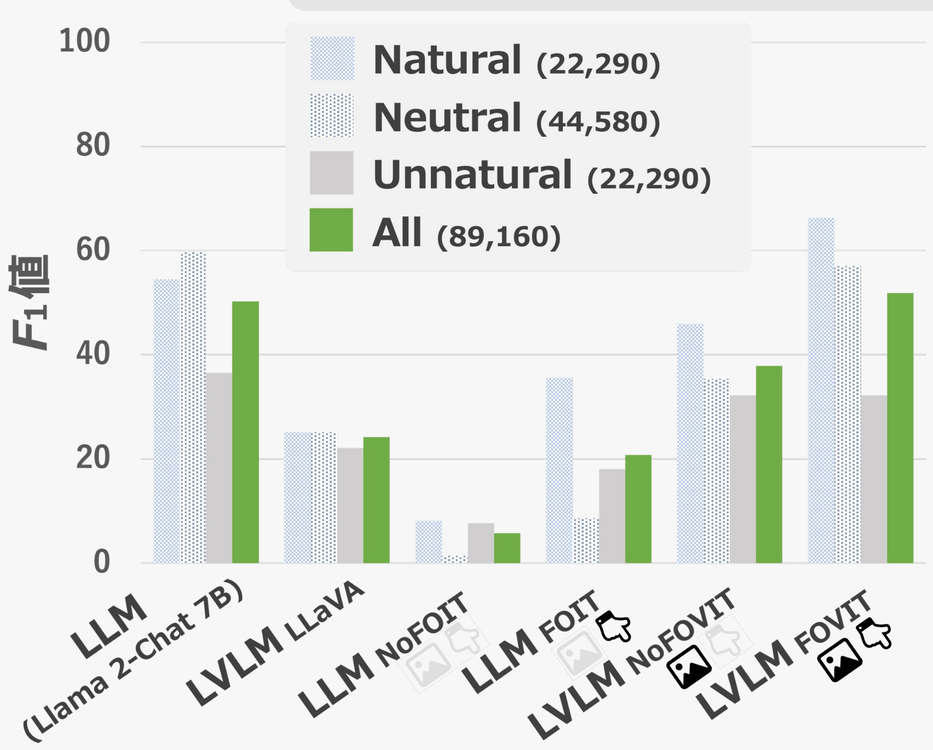
Answer:

Label System by Contextual Consistency		If positive.. label_0	If negative.. label_1
Natural	high	positive	negative
Neutral	↕	foo	bar
Unnatural	low	negative	positive

- Liら [Li+, '23] に倣い、9つの二値分類データセット (SST-2, FP, EMOTION, SNLI, SICK, RTE, QQP, MRPC, SUBJ) それぞれに対して verbalizer manipulation を実施して評価データセットを構築
- ラベルの意味表現と学習時の文脈知識の整合性別に３つのラベル体系を定義

実験

LVLMの構成要素:
- (LLM) Llama 2-Chat 7B
- (画像エンコーダ) CLIP ViT-Large/14
- (アダプター) 1層の線形層



※ All は、Natural, Neutral, Unnatural の F1値のマクロ平均を示す

✔ LVLM の指示追従能力の低下を定量的に確認

- モデルに、本来の意味とは異なるラベルで解答するよう指示した場合 (Unnatural) において、全ての追加学習済み LVLM が LLM (Llama 2-Chat 7B) を下回った

✔ 追加学習データ中の出力形式に関する指示の有無が影響

- All において、LVLM NoFOVIT よりも LVLM FOVIT の方がF1値が高い
- All において、LLM NoFOIT よりも LLM FOIT の方がF1値が高い
- 出力形式を明示的に与えることにより視覚モダリティによらず、ベースLLMが有する指示追従能力の低下を抑制できることを示唆