

日本語の画像言語モデルの 構築

岡崎 直観

東京科学大学 情報理工学院

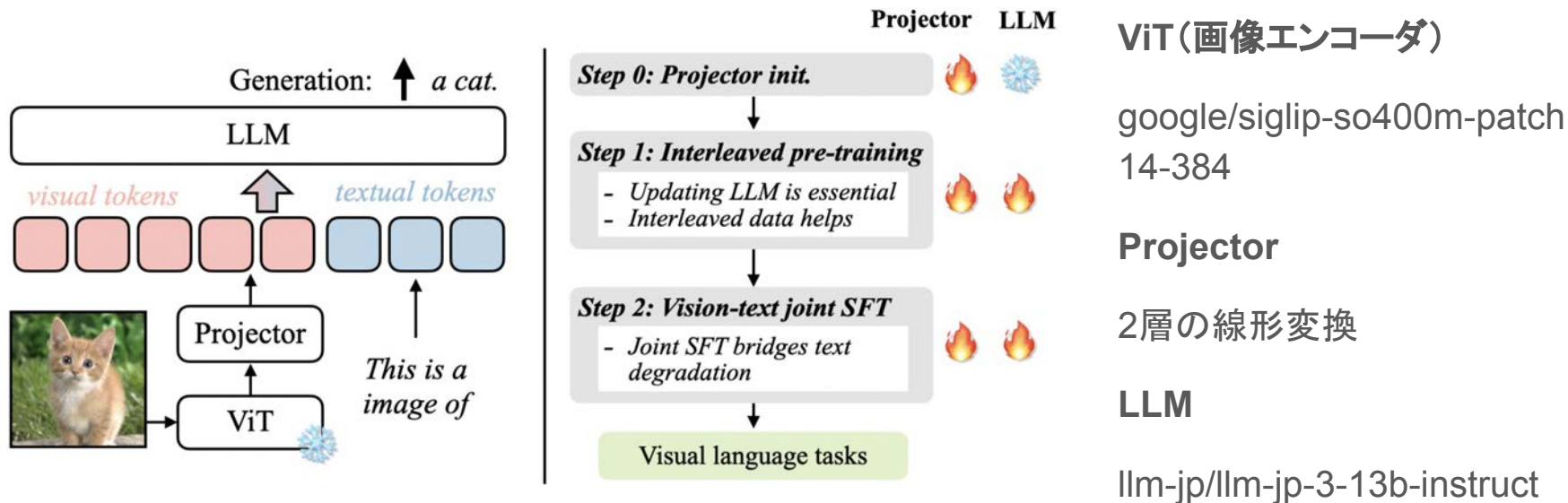
国立情報学研究所

大規模言語モデル研究開発センター

マルチモーダルWGの活動

- 画像言語モデル (VLM) の構築
 - 画像から言語への基盤モデル
 - 動画から言語への基盤モデル
 - 日本語向けCLIP
 - 日本語向けBERT
- 日本語VLMの評価基盤の構築
 - 日本語と英語のVLMタスクでの評価
 - 既存のVLMモデルのリーダーボード
- 日本語VLM構築向けのデータ整備
 - ウェブからの画像・言語対データ構築
 - 指示チューニングデータの合成
 - 日本語OCRデータの合成
- 日本語VLMの評価データの構築
 - 日本の常識・文化・習慣に関する評価セット
 - 文書画像理解の評価セット
 - 日本語OCRの評価セット

構築したVLM (LLM-jp-3 VILA 14B) の構成



日本語の画像・テキスト対データの構築

「東京科学大学」になった旧・東京工業大学、「工業」弔う墓出現
卒業生「無念さを少しは晴らしてくれたのかな」

10/3(木) 20:07 配信 52

CASTニュース
ビジネス&メディアウォッチ



2024年10月1日に東京医科歯科大学と統合し「東京科学大学」となった旧・東京工業大学（東工大）の大岡山キャンパス（東京都目黒区）内に、「工業之墓」と書かれた謎の「墓」が出現した。名称が切り替わる9月30日夜から10月1日にかけて写真が複数投稿され、Xで話題となった。

旧・東京工業大学の大岡山キャンパス

【画像】旧・東京工業大学に出現した「工業之墓」

<https://www.j-cast.com/2024/10/03494892.html>

日本語画像・テキストペア

“また、\n地下の木挽町広場には、「和」を意識した\nお店がたくさん入っていて、
眺めるだけでも\n楽しめます”



“もうお雛様が飾られていました\n\n”

“投稿者: 山本歯科診療所\n\n”

“2016.01.18更新\n\n”

“### ドクターズスイーツ\n\n”

“「ドクターズスイーツ」というのが\n最近流行っているようです”



主に、糖尿病で糖質の摂取が制限されている人でも食べられるお菓子などですが、\n歯科でもあります !\n\n”

<http://www.yamamoto-dental-clinic.com/blog/staff/2016/01/>

日本語版MMC4

日本語の指示チューニングデータの合成



Q: この画像には何が写っていますか？

A: 画像には、白と赤のバスが都市の通りを走っている様子が写っています。バスは比較的大きく、広告が後部に貼られているようです。また、周囲には多くの人や建物があります。

.....

llava-instruct-ja



<https://huggingface.co/datasets/ThePioneer/japanese-photos>

Q: この画像にはどんな建物が写っていますか？

A: 画像には「LaLaport」と書かれた建物が写っています。

Q: この画像の中で目立つ大きなフィギュアは何ですか？

A: 目立つ大きなフィギュアはガンダムのロボットです。

.....

japanese-photos-conv

学習の詳細 (太字は今回構築したデータ、青はVQAデータ、橙はOCRデータ)

学習段階	日本語データ	英語データ	学習時間
Step 0	日本語画像・テキストペア (558k)	LLaVA-1.5事前学習データ ³ (558k)	LLM-jp クラスタ 8xA100, 約14-15時間
Step 1	日本語画像・テキストペア (6M), 日本語版MMC4 (6M)	COYO ⁴ の一部 (6M), mmc4-core ⁵ の一部 (6M)	LLM-jp NVLINK クラスタ 8x8xA100, 約130時間
Step 2	llava-instruct-ja (156k), japanese-photos-conv (12k), ja-vg-vqa ¹ (99k), synthdog-ja ² (102k), 計369k事例	LLaVA-1.5 instruction tuningデータ ⁶ (158k), VQAv2 (53k), GQA (46k), OCR VQA (80k), TextCaps (22k), 計358k事例	LLM-jp NVLINK クラスタ 4x8xA100, 約11時間

1 <https://github.com/yahoojapan/ja-vg-vqa>

2 <https://huggingface.co/datasets/naver-clova-ix/synthdog-ja>

3 <https://huggingface.co/datasets/liuhaotian/LLaVA-Pretrain>

4 <https://github.com/kakaobrain/coyo-dataset>

5 <https://github.com/allenai/mmc4>

6 <https://huggingface.co/datasets/liuhaotian/LLaVA-Instruct-150K>

評価結果 (当時)

Heron Bench

JA-VLM-Bench-In-the-Wild

JA-VG-VQA-500

モデル

Judgeスコア (%)

ROUGE-L

Judgeスコア (5点満点)

ROUGE-L

Judgeスコア (5点満点)

[Japanese InstructBLIP Alpha \(7B\)](#)

14.0

20.8

2.42

-

-

[Japanese Stable VLM \(7B\)](#)

24.2

23.3

2.47

-

-

[Llama-3-EvoVLM-JP-v2 \(8B\)](#)

39.3

41.4

2.92

23.5

2.96

[LLaVA-CALM2-SigLIP \(7B\)](#)

43.3

47.2

3.15

17.4

3.21

[Llama-3.2-11B-Vision](#)

36.5

27.4

2.77

13.8

2.95

[InternVL2-8B](#)

45.2

33.7

2.98

11.6

3.13

[Owen2-VL-7B-Instruct](#)

54.8

45.3

3.53

16.2

3.48

LLM-jp-3 VILA 14B

57.2

52.3

3.69

16.2

3.62

GPT-4o

87.6

37.6

3.85

12.1

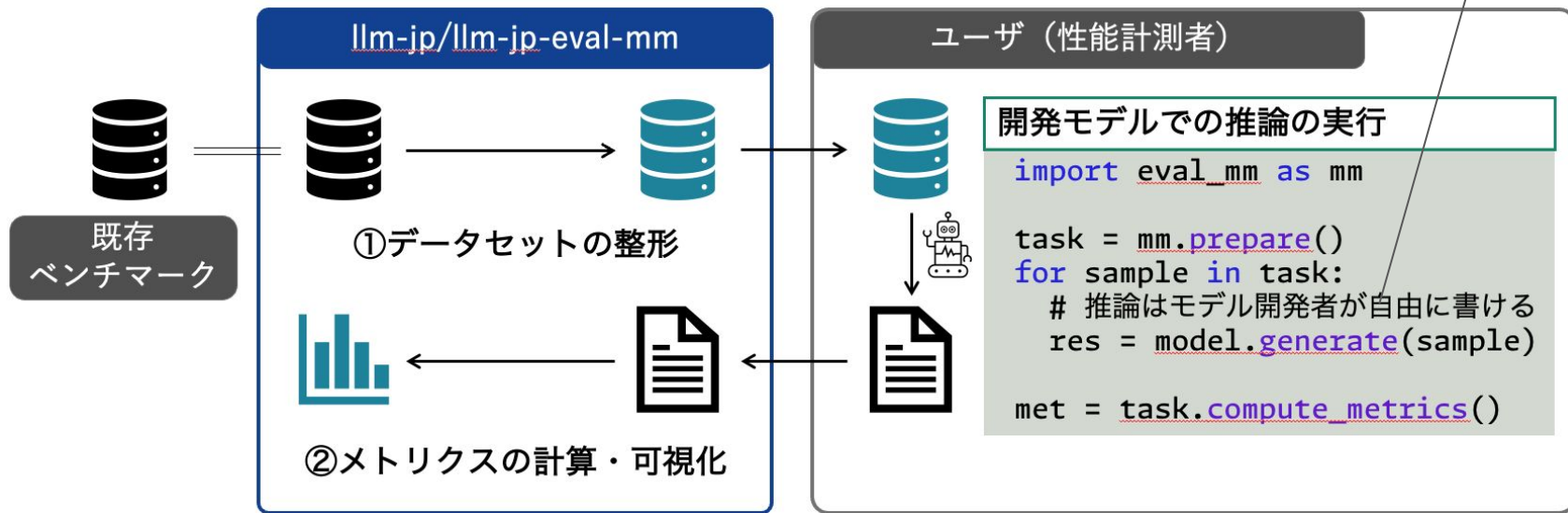
3.58

7

日本語VLMの評価基盤: llm-jp-eval-mm

日本語10タスク、英語2タスク


`eval-mm` では条件を揃えたサンプルコードを提供する



Model	CC-OCR	CVQA	Heron	JDocQA		JIC	JMMMU	JVB-ItW		LLAVA		MECHA	MMMU	MulIm-VQA		YG-VQA	
	macro f1 ↓	Acc ↓	LLM ↓	Acc ↓	LLM ↓	Acc ↓	Acc ↓	LLM ↓	Rouge ↓	LLM ↓	Rouge ↓	Acc ↓	Acc ↓	LLM ↓	Rouge ↓	LLM ↓	Rouge ↓
stabilityai/japanese-instructblip-alpha	0.3	23.6	23.5	12.4	2.1	56.6	26.4	2.3	14.2	1.3	0.1	23.3	25.6	2.4	25.0	3.0	34.1
stabilityai/japanese-stable-vlm	-	-	48.4	13.1	2.1	69.3	3.4	3.3	23.2	1.4	0.6	5.1	7.1	2.3	40.7	3.5	45.7
SakanaAI/Llama-3-EvoVLM-JP-v2	10.4	43.8	47.6	15.4	2.4	67.0	36.4	3.5	48.4	2.9	27.9	50.8	38.9	3.1	44.3	3.4	24.7
cyberagent/llava-calm2-siglip	11.4	6.9	54.1	8.2	2.0	58.5	6.1	3.7	46.3	1.9	2.9	11.3	26.7	2.8	40.6	3.6	17.7
llm-jp/llm-jp-3-vila-14b	11.6	29.6	68.0	17.3	2.5	81.3	19.0	4.1	52.4	3.4	36.0	45.6	32.7	3.5	47.0	3.9	16.2
sbintuitions/sarashina2-vision-8b	8.2	48.8	60.5	22.6	3.0	78.7	39.2	4.1	44.8	2.5	16.9	56.7	29.7	2.6	30.2	3.7	25.4
sbintuitions/sarashina2-vision-14b	23.2	56.2	60.1	23.9	3.1	80.0	43.0	4.0	44.3	2.5	15.6	64.8	33.8	2.6	35.3	3.7	25.3
MIL-UT/Asagi-14B	-	-	41.9	10.4	2.0	76.2	21.7	2.9	30.9	1.6	0.1	-	15.3	2.0	18.4	2.0	9.3
llava-hf/llava-1.5-7b-hf	14.5	40.9	43.1	14.8	2.2	44.0	29.6	3.0	40.8	2.9	34.5	38.4	34.0	2.5	35.8	3.0	14.0
llava-hf/llava-v1.6-mistral-7b-hf	20.1	26.6	30.0	14.4	2.0	58.2	25.4	2.9	28.6	3.3	30.7	34.0	35.9	2.3	25.2	3.0	11.7
neulab/Pangea-7B-hf	18.7	48.3	57.0	16.2	2.4	85.6	37.4	3.9	33.5	3.5	25.9	57.3	43.7	3.4	40.3	4.1	54.2
mistralai/Pixtral-12B-2409	27.2	49.3	60.9	14.8	2.4	61.1	18.7	3.9	38.3	3.6	31.6	56.0	48.6	4.1	34.5	3.5	13.1
meta-llama/Llama-3.2-11B-Vision-Instruct	27.7	51.7	38.1	17.6	2.5	78.7	34.6	3.4	30.4	3.7	30.2	49.6	38.2	2.6	24.5	3.3	14.2
Efficient-Large-Model/VILA1.5-13b	20.6	43.8	46.9	14.8	2.2	58.2	33.5	3.5	42.5	3.6	35.0	46.6	37.0	3.2	40.0	3.2	13.0
OpenGVLab/InternVL2-8B	37.3	48.3	49.8	19.8	2.7	65.7	39.1	3.5	33.8	3.1	31.5	50.3	49.7	2.9	34.5	3.5	11.7
OpenGVLab/InternVL2-26B	38.5	50.3	59.7	15.3	2.6	73.9	39.0	3.1	26.7	3.8	30.6	50.9	48.2	3.3	45.4	3.6	11.6
Qwen/Qwen2.5-VL-7B-Instruct	74.8	55.7	70.3	26.5	3.6	82.7	48.2	4.3	29.6	3.9	27.1	61.2	50.0	4.1	50.4	3.7	9.0
Qwen/Qwen2.5-VL-32B-Instruct	-	-	74.8	25.2	3.8	96.0	48.8	4.3	14.7	4.0	18.9	68.4	59.1	4.6	42.2	3.8	5.2
Qwen/Qwen2.5-VL-72B-Instruct	77.1	69.5	85.5	23.9	3.9	90.4	60.6	4.4	32.0	4.0	28.7	76.2	63.0	4.8	60.9	3.9	9.8
google/gemma-3-4b-it	55.5	49.3	52.8	17.6	2.6	75.4	37.0	3.7	37.1	3.6	22.1	47.9	40.7	3.7	52.7	3.4	12.5
google/gemma-3-12b-it	64.9	58.6	72.2	20.1	3.0	85.7	47.6	4.3	35.7	4.0	22.1	62.4	48.1	4.2	59.7	3.7	12.5
google/gemma-3-27b-it	66.9	62.6	69.2	20.2	3.1	88.2	50.5	4.4	30.9	3.9	21.1	68.7	56.1	4.3	56.3	3.8	10.9
microsoft/Phi-4-multimodal-instruct	42.5	37.4	45.5	22.9	2.9	52.3	39.2	3.2	26.8	3.4	29.5	45.6	53.7	3.4	42.3	3.3	19.0
gpt-4o-2024-11-20	64.1	82.3	93.7	22.0	3.6	95.8	57.5	4.4	32.2	4.1	29.8	84.0	56.1	4.8	62.5	3.9	11.8
turing-motors/Heron-NVILA-Lite-15B	44.5	60.6	68.5	19.2	2.7	-	49.5	4.4	50.6	3.6	36.6	68.3	47.1	4.3	60.0	-	-

日本の常識・文化・習慣に関する評価セット: MECHA-Ja


- 日本の日常生活文化を背景とするVLMベンチマーク
 - 603枚の画像と1821件の質問応答ペアで構成
 - 日本文化に根ざした知識・推論を必要とする QAタスク
 - 画像中の事物だけでなく、日本の生活文化を理解するために不可欠な無形知識の有無を測定可能
- 構築方法
 - 日本の日常・風景・文化・習慣に関する写真を収集
 - 質問とその回答となる選択肢(4択)を人手で作成
 - factoid: 事物の名称や単語のみを答える質問
 - non-factoid: 理由や方法を説明させる質問
 - 日本の日常や文化を問う質問になるように指示した



問題: 写真の下半分に写っている設備について、白いラベルには通常、何が書かれていますか。

選択肢: A. 使用者へのメッセージ B. 使用期限
C. 使用目的 D. 使用者の氏名


解説: 各下駄箱には靴の取り違えを防ぐために通常、使用者の氏名を書いたラベルがつけられています。



問題: 写真に写っている料理に、味噌を使った調味料をかけて食べることで有名な都道府県はどこですか。

選択肢: A. 群馬県 B. 長野県 C. 愛知県 D. 大阪府

解説: 写真に写っている料理はトンカツです。味噌を使った調味料をトンカツにかけて食べる料理は味噌カツと呼ばれます。この料理が有名な都道府県は愛知県です。同県は他にも味噌煮込みうどんなど、味噌を使った名物料理があります。



問題: 写真に写っている生き物が冒頭で登場する、日本のおとぎ話の主人公は誰ですか。

選択肢: A. 桃太郎 B. 金太郎
C. 浦島太郎 D. 一寸法師

解説: 写真に写っている生き物は亀です。亀が冒頭で登場する、日本のおとぎ話は浦島太郎です。このおとぎ話の主人公は、亀を助けて竜宮城に連れていかれる浦島太郎です。

現在進行中の取り組み

- 新しいVLMの構築

- ViT (画像エンコーダ): SigLIP2 (google/siglip2-so400m-patch16-512)
- LLM (言語デコーダ): LLM-jp-4-8B (開発中)
- 訓練データの拡充・改良
 - WAON (ウェブやWikipediaから抽出した画像言語対データ)
 - FineVision (英語のVLM訓練データ)
 - 指示チューニングデータ: 日本語OCRデータ、PDFからの文書画像理解・質問応答データ、図表からの文書画像理解・質問応答データなど

- 評価データの拡充・改良

- 日本の常識・文化・習慣に関する評価セット: 日本らしさに特化した評価事例の拡充
- 日本語の文書画像理解のための質問応答データセット
- 日本語の自然画像OCRデータセット(帳票・レシート、手書き文字、看板など)