Jagle: 視覚言語モデルのための 大規模日本語マルチモーダル事後学習データセットの構築

2Yin-B-01

Dataset Model

杉浦 一瑳[1,2] 笹川 慶人[3,2] 中尾 圭佑[3,2] 前田 航希[4,2] Yin Ziqi[2] Yang Zhishen[2]
栗田 修平[5,2] 小田 悠介[2] 徳久 良子[6,7] 河原 大輔[3,2] 岡崎 直観[4,2]
1 京都大学 2 NII LLMC 3 早稲田大学 4 東京科学大学 5 NII 6 愛知工業大学 7 RIKEN

概要

- **日本語として最大規模のマルチモーダル事後学習データセット　Jagle**を構築
  - **13データソース** から, **5カテゴリ**, **17サブセット**, **9.2M事例**のVQAを作成
- 2.2BのVLM学習実験の結果, **Jagleで学習したモデルは InternVL3.5-2Bを上回る日本語性能** を示した
- **Jagleを用いて学習した LLM-jp-4-VL 9B betaは, Qwen3-VL 8Bに匹敵する日本語性能** を示した

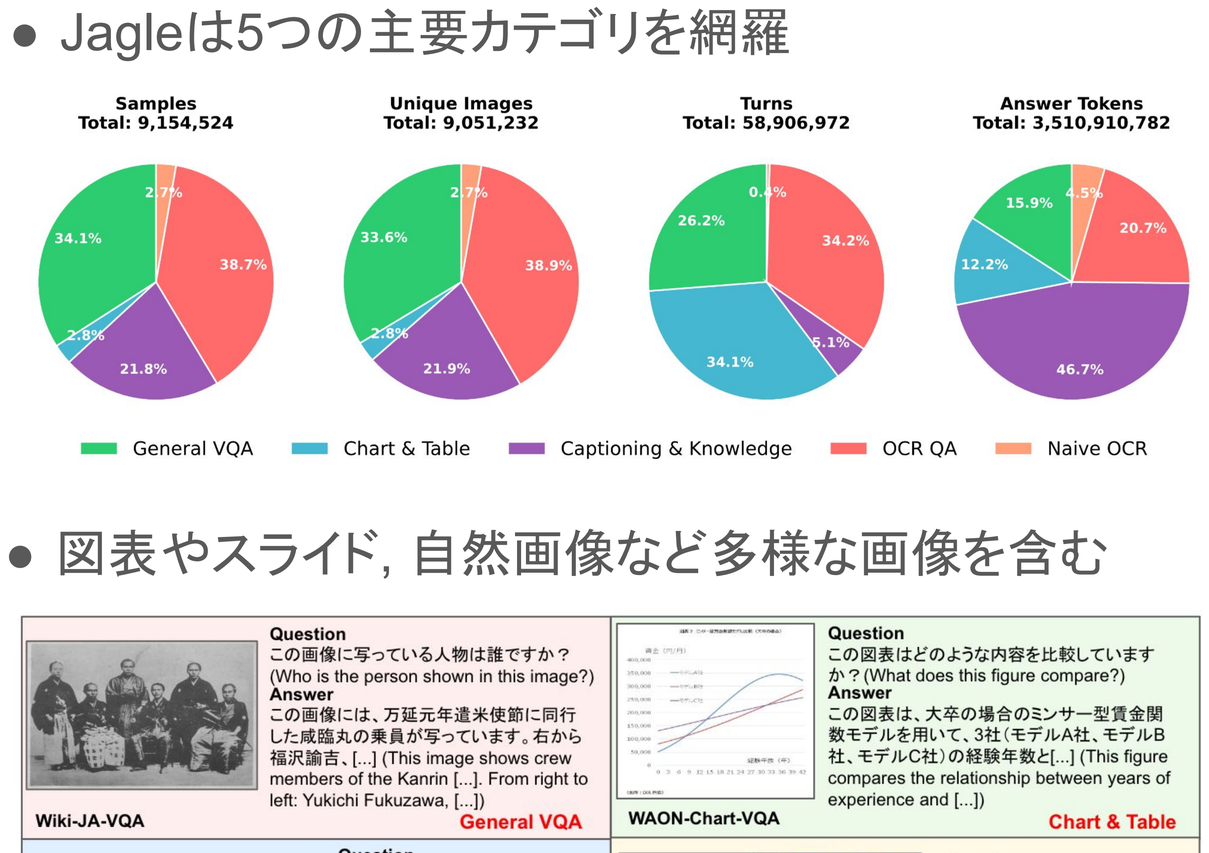
はじめに

- 既存の日本語マルチモーダル事後学習データセットは**規模・網羅性** に課題がある

Dataset	Language	Categories	Subsets	Examples
Cambrian-7B (Tong et al., 2024)	English	9	70	7.1M
FineVision (Wiedmann et al., 2025)	English	9	185	24.2M
DEJIMA (Katsube et al., 2025)	Japanese	2	2	3.9M
LLM-jp-3 VILA (Sasagawa et al., 2025b)	Japanese	3	4	0.4M
Jagle (Ours)	Japanese	**5**	**17**	**9.2M**

Jagleの構築

- 13データソースをもとに9.2M事例のVQAを構築



Jagleの統計・事例分析

- Jagleは5つの主要カテゴリを網羅

- 図表やスライド, 自然画像など多様な画像を含む

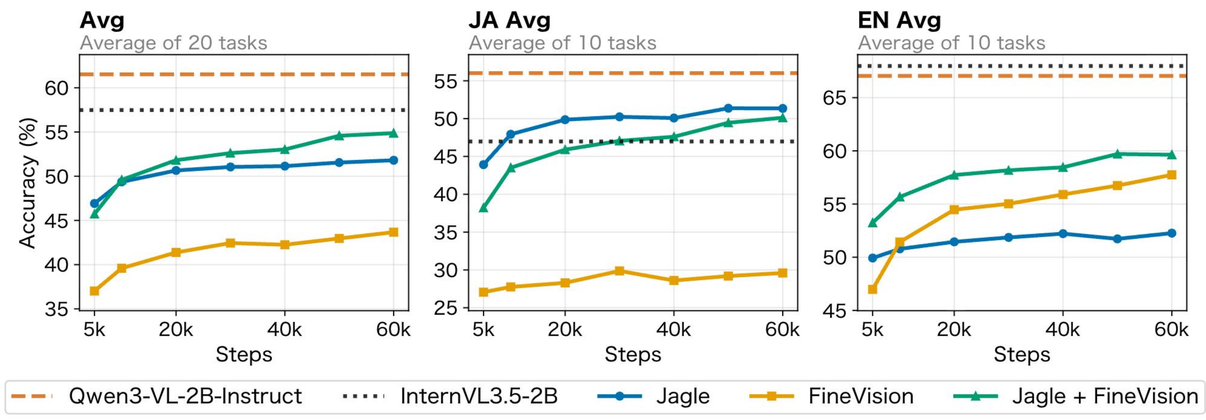
Jagleの評価

2.2BのVLMを以下の3つの設定で学習し性能を評価

- (1) Jagle, (2) FineVision, (3) Jagle + FineVision

VLMのアーキテクチャ:

- LLM: Qwen3 1.7B, ViT: SigLIP2 So400M



- **Jagleは日本語平均タスク性能を大幅に向上し InternVL3.5 2Bを上回った**
- **Jagle + FineVisionはFineVision単体より英語性能が伸びた**

LLM-jp-4-VL 9B betaの開発

- LLM-jp-4-instruct 8Bをベースに FineVision + Jagle で学習された**LLM-jp-4-VL 9B betaは, 日本語タスクで Qwen3-VL 8Bに匹敵**
- **性能は最後まで伸び続けており，学習量を増やすことでさらに性能が上がる可能性** がある