C2-22

MOMIJI: 日本語大規模インターリーブ視覚言語データセット

塩野 大輝[1,2], 横井 慎吾[1], 犬塚 眞太郎[1], 髙橋 翼[1], 鈴木 潤[2,3,4], 山口 祐[1]

[1] Turing 株式会社, [2] 東北大学, [3] 理化学研究所, [4] 国立情報学研究所 LLMC

Dataset Paper

背景と概要

背景

インターリーブ視覚言語データセットは、現在の LVLMs の事前学習に広く用いられる

しかし、日本語の大規模なインターリーブ視覚言語データセットは少なく、
また Interleave レイアウトの事前学習における有効性に関する証拠も一貫していない

概要

そこで、日本語大規模インターリーブ視覚言語データセット **MOMIJI** を新たに構築

さらに Interleave レイアウトと画像-テキスト間の意味的関連性を互いに独立に変化させる
10 通りの統制されたデータセット派生版を構築し、これらで LVLM を事前学習した際の
性能を比較することで、Interleave レイアウトが有効となるデータ特性を調査

MOMIJI データセット構築

MOMIJI のフィルタリングパイプライン

(1) 画像 URL を含む日本語テキストの抽出

- (1-1) WARC アーカイブのダウンロード
- (1-2) 迅速な日本語テキスト判定
- (1-3) 画像 URL プレースホルダーを含むテキスト本文の抽出 (BeautifulSoup + Trafilatura)
- (1-4) 高精度な日本語テキスト判定 (fastText)

(2) テキスト品質に基づく文書フィルタリング (Hojichar)

- (2-1) 高頻度に繰り返される要素を含むウェブ文書の除去
- (2-2) 低品質なテキストを含むウェブ文書の除去
- (2-3) 有害な表現を含む可能性のあるウェブ文書の除去

(3) 画像フィルタリング

- (3-1) 同一文書内における重複画像 URL の除去
- (3-2) 不適切な画像 URL の刈り込み
- (3-3) 他文書にまたがる重複画像 URL の除去
- (3-4) 候補画像のダウンロード (img2dataset)
- (3-5) 画像の解像度とアスペクト比によるフィルタリング
- (3-6) NSFW 画像のフィルタリング (nsfw-detector)

WARC ファイル
(2024/02 – 2025/01)

フィルタリング

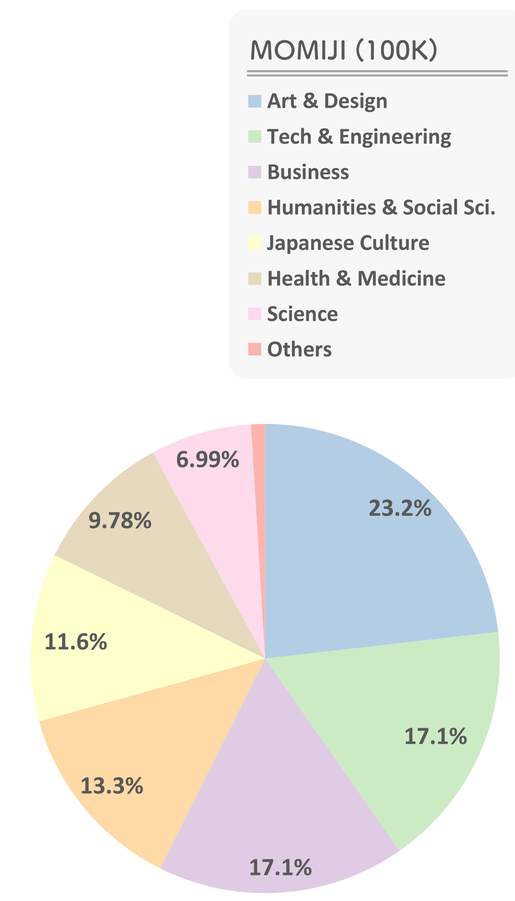
Interleave
画像-テキスト
(56M)

「七夕」の飾りや習慣とは

「七夕飾り」といえば七夕竹が有名です。天の神様への目印として江戸時代に庶民の間で七夕竹を立てたのがはじまりで、古くは7月6日の夜に立てていました。
「七夕」は機織りにちなみ手芸の上達を願うものでしたが、現代ではジャンルを問わずさまざまな願い事を短冊に書いて吊るす風習があります。

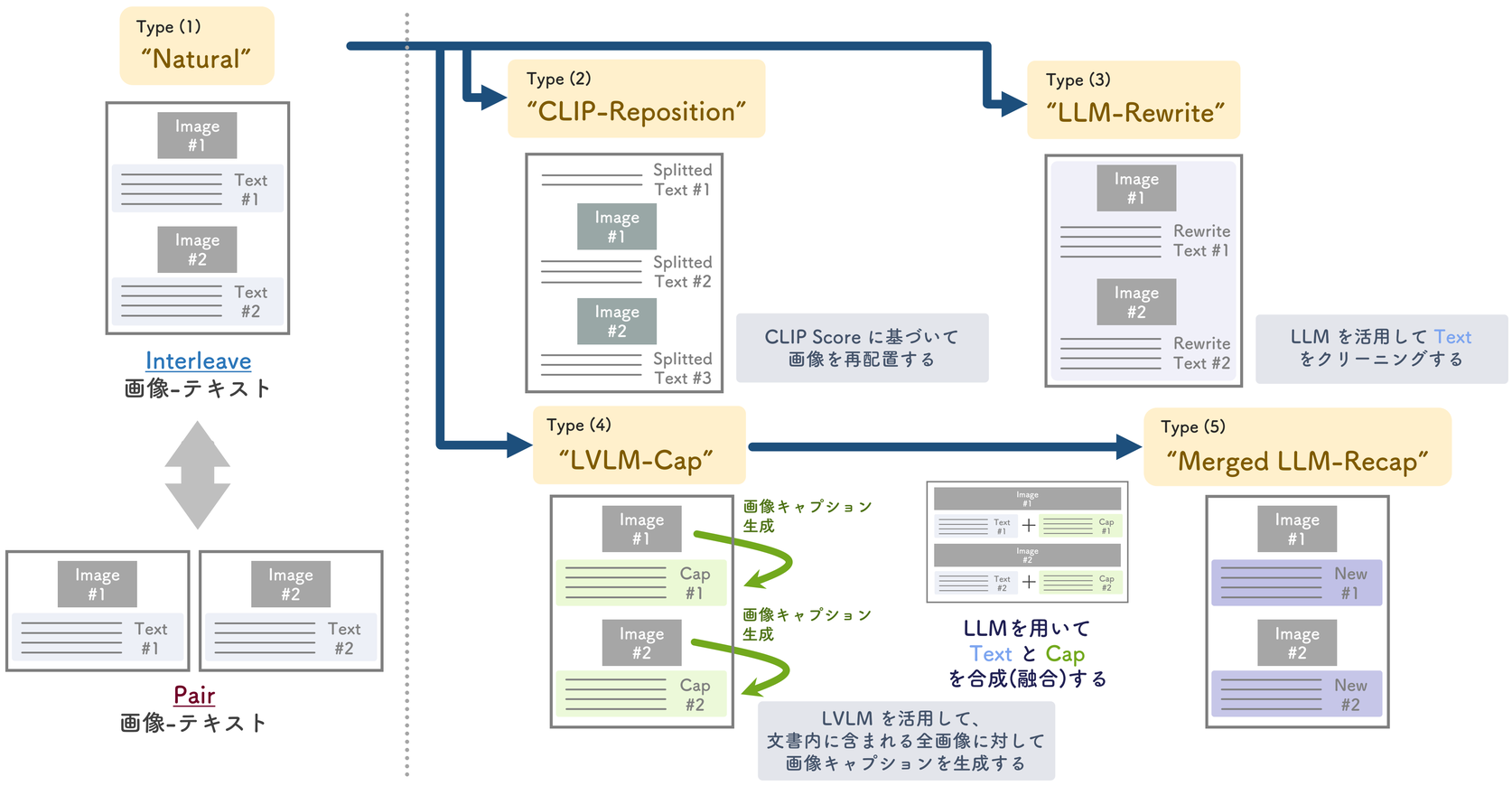
…

「七夕」の行事食はそうめんです。「七夕」にそうめんを食べると一年間無病息災でいられると言われていて、7月7日は「そうめんの日」にも制定されています。



実験: Interleave レイアウトが有効となるデータ特性の調査

MOMIJI を活用し、10 通りの学習データ派生版を作成



Stage 1) プロジェクタアライメント

- 学習データ: LLaVA-Pretrain (558k), llm-jp-japanese-image-text-pairs (6M subset)
- Trainable: {プロジェクタ}, Frozen: {LLM, 視覚エンコーダ}

Stage 2) 事前学習

ここを変える！

- 学習データ: MOMIJI 由来のデータ派生版 (1M)
- Trainable: {LLM, プロジェクタ}, Frozen: {視覚エンコーダ}

評価結果

ラベル	レイアウト	後処理タイプ	平均※ スコア (%)
(a)	Interleave	Natural	36.5
(b)	Interleave	CLIP-Reposition	37.9
(c)	Interleave	LLM-Rewrite	40.4
(d)	Interleave	LVLM-Cap	45.2
(e)	Interleave	Merged LLM-Recap	42.6
(f)	Pair	Natural	37.1
(g)	Pair	CLIP-Reposition	36.8
(h)	Pair	LLM-Rewrite	36.1
(i)	Pair	LVLM-Cap	38.4
(j)	Pair	Merged LLM-Recap	38.5

※ JMMMU, Heron-Bench, JA-VLM-Bench-In-the-Wild, JA-Multi-Image-VQA, JA-VG-VQA500, JDocQA の平均スコア

- レイアウト比較 (Interleave vs. Pair)
  - Pair よりも Interleave を好む傾向
- 後処理タイプ別の比較
  - 画像の再配置や LLM によるテキストの洗練だけでは有意な利点は観察できない
  - 画像内容と密に関係するテキストを注入することが効果的

言語処理学会第32回年次大会 (NLP2026)