オープン LLM による翻訳を活用した日本語 CLIP の開発

杉浦一瑳 ♣,‡, 栗田修平 ♢,‡, 小田 悠介 ‡, 河原大輔 ♠,‡, 岡崎 直観 ♡,‡
♣ 京都大学, ♢ 国立情報学研究所, ♠ 早稲田大学, ♡ 東京科学大学,
‡ 国立情報学研究所 大規模言語モデル研究開発センター
sugiura.issa.q29@kyoto-u.jp {skurita, odashi}@nii.ac.jp
dkw@waseda.jp okazaki@c.titech.ac.jp

概要

CLIP は視覚言語モデルのコンポーネントとして採用される事例が増えており, 重要性が増している. しかし, オープンな日本語画像・テキスト対データセットは不足しており, モデル開発の障壁となっている. 本研究では, オープン LLM を用いた機械翻訳により, 20 億事例の日本語画像・テキスト対データセットを構築し, 日本語 CLIP を学習した. 学習済みモデルの性能を, 7 つの評価データセットを用いて評価した結果, 平均スコアが同程度のモデルサイズにおいて高水準であった一方, 日本文化に関するタスクの性能が低いことがわかった. 学習済みモデル[1)2)], 学習データセット[3)], 翻訳ツール[4)], 評価コード[5)] は公開する.

1 はじめに

CLIP [1] は, 対照学習によって画像とテキストを同じ埋め込み空間に対応付けるモデルである. CLIP は視覚言語モデル (Visual Language Model: VLM) や拡散モデルに利用されることがあり, 重要性が高まっている [2, 3, 4].

一方で, 既存の CLIP の多くは主に英語のテキストと画像が対になったデータセットで学習されており, 他の言語における性能が低いという問題がある [5, 6, 7]. 多言語の画像・テキスト対データセットを構築する試みもあるが [5, 8, 9], オープンな日本語画像・テキスト対データセットの量は依然として少ない. 実際に, 執筆時点でオープンな日本語画像・テキスト対データセットとして最大規模の LAION-5B [9] の日本語サブセットは約 1.2 億事例であり, 4 億事例のデータを用いて学習した CLIP [1] より小さい. DeepL を用いた機械翻訳により日本語画像テキスト対データセットを構築する先行研究もあるが [7], 数千万事例と小規模なものにとどまっている.

本稿では, モデルの重みが公開されている LLM (open-weight LLM, 以下オープン LLM と呼ぶ) を用いて 20 億事例の画像・テキスト対データを機械翻訳で構築し, 日本語 CLIP をフルスクラッチ学習した. 評価の結果, 平均スコアが同程度のモデルサイズの中で高水準であった一方, 日本語文化に関するタスクにおいて性能が低いことが分かった.

2 日本語の画像・テキスト対データセットの構築

日本語画像・テキスト対データセットは限られており, 十分な量のデータを確保することが課題となっている. この課題に対して, 英語で構築された大規模なデータセットを日本語に翻訳することで, 日本語データを増強できるが, その実現には高速かつ効率的に翻訳を行う必要がある. 本研究では、この課題に取り組むため, ReLAION-5B[6)] の英語サブセット[7)] を gemma-2-9b-it[8)] を用いて翻訳した. ReLAION-5B は, LAION-5B [9] から子どもの性的虐待表現物 (Child Sexual Abuse Material: CSAM) が疑われるデータを取り除いたデータセットであり, Common Crawl から画像と対応する alt 属性（テキスト）の組を収集し, CLIP でフィルタリングすることで構築された大規模画像・テキスト対データセットである. 英語, 多言語, 非言語のサブセットに分かれており, 英語サブセットは 2B 事

1) https://huggingface.co/llm-jp/llm-jp-clip-vit-base-patch16
2) https://huggingface.co/llm-jp/llm-jp-clip-vit-large-patch14
3) https://huggingface.co/llm-jp/relaion2B-en-research-safe-japanese-translation
4) https://github.com/llm-jp/text2dataset
5) https://github.com/llm-jp/clip-eval
6) https://laion.ai/blog/relaion-5b
7) https://huggingface.co/datasets/laion/relaion2B-en-research-safe
8) https://huggingface.co/google/gemma-2-9b-it

表 1　gemma による ReLAION-5B データセットの翻訳例.

English Caption	Japanese Caption
Iron Man Movie Poster	アイアンマン 映画ポスター
Unique 14k Gold Yellow and Blue Diamond Engagement Ring 2.64ct.	ユニークな 14 金イエローゴールドとブルーダイヤの婚約指輪 2.64ct.
”””Brent Payne ”””””Brent Payne””””” 1999 Self Released Country Nm/Nm Out Of Print Cd”””””	ブレント・ペイン ”ブレント・ペイン” 1999 年 自主制作 カントリー Nm/Nm 絶版 CD

例を含む. なお, 大規模なデータを高速に翻訳するために, `text2dataset`[9] というツールを開発した. このツールは vLLM [10] という LLM 高速推論ライブラリを利用することで, 大規模な英語データセットを日本語に高速に翻訳する.

プロンプト　LLM に翻訳を遂行させるには, 翻訳対象のデータに加え, 適切な指示文（プロンプト）を入力する必要がある [11]. 本研究では, 以下に示すプロンプトを用いた. ここで {passage} は翻訳元の文が挿入されるプレースホルダであり, このプロンプトを受け取った LLM は続けて翻訳文を出力することが期待される.

```
You are an excellent English-Japanese
translator. Please translate the
following sentence into Japanese.\n You
must output only the translation.\n
Sentence:{passage}\n Translation:
```

翻訳結果　ReLAION-5B 英語サブセット全体を, mdx 演算加速 GPU ノード上で翻訳した. NVIDIA A100 40GB を 32 個用い, 約 9 日の処理で 2,097,693,557 事例の翻訳を完了した. 本処理で得られた翻訳例を表 1 に示す. 英語のキャプションを日本語に上手く翻訳できていることが確認できる. 一方で, 先頭 1 万事例を目視で確認したところ, 翻訳結果にはいくつかの問題が見られた. まず, プロンプトで具体的に翻訳先言語を指定しているにも関わらず中国語や韓国語などに誤って翻訳された例が 1%ほど存在した. また翻訳文の最後に ”Please let me know if you have any questions.” という様な内容の文が追記されるという, 指示チューニング済み LLM 特有の現象も 0.1%ほど見受けられた. これらの問題は, より翻訳性能の高い LLM の利用や翻訳結果の事後処理によって改善できると考えられるが, 今後の課題とする.

画像は `img2dataset`[10] を用いてダウンロードした. URL のリンク切れや前処理の失敗等のためにダウンロード成功率は約 70%となり, 最終的に 1,451,957,221 事例の日本語画像・テキスト対データセットが得られた.

9）https://github.com/llm-jp/text2dataset
10）https://github.com/rom1504/img2dataset

3　CLIP の学習

前節で得られた画像・テキスト対のデータセットを用い, CLIP を学習した. ここではデフォルト設定として, llm-jp-clip-ViT-B/16 の学習設定について述べる. CLIP では画像とテキストのエンコーダのモデル選択が任意であるが, 本研究では画像エンコーダに ViT-B/16 [12], テキストエンコーダに $\text{RoBERTa}_{\text{BASE}}$ [13] を用いた. 各エンコーダの出力次元は 512 とし, どちらもフルスクラッチで学習した. トークナイザは llm-jp-tokenizer[11] を基に, CLIP 用に修正したものを用いた. テキストの最大コンテキスト長は 76 トークンとした. 画像の解像度は 224x224 とした.

最適化アルゴリズムに AdamW を採用し, ハイパーパラメータはそれぞれ $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$ とした. 学習率スケジューリングに 2000 ステップのウォームアップとコサイン減衰を採用し, 最大学習率を 5.0×10^{-4}, 最小学習率を 0.0 とした. スケジューラの終了時点は学習データの 9 エポック分に相当する 13,067,614,989 事例とし, この事例数に到達するまでデータセットを周回した. 損失関数には対照損失を用いた [1]. 損失の計算に用いたバッチサイズは 8,192 とし, バッチあたり 4 回の勾配累積で計算を行った. なお, 勾配累積の結果得られる損失はバッチサイズ 32,768 を直接用いた対照損失とは異なることに注意が必要である.

以上の設定を用い, さくらインターネット高火力 PHY の計算機上に構築した OpenCLIP [14] の実行環境で実験を行った. NVIDIA H100 80GB を 16 個用い, ひとつのモデルあたり 2 週間の学習時間を要した.

4　評価

学習したモデルの性能を測るために, 日本語評価データセットを用いて, 日本語および多言語対

11）https://github.com/llm-jp/llm-jp-tokenizer

表 2 各モデルのゼロショット画像分類及び画像・テキスト検索タスクの性能. **太字**は一位, 下線は二位を表す.

タスク		画像分類						検索		
モデル	# Params (M)	ImageNet	Recruit	CIFAR10	CIFAR100	Food101	Caltech101	XM3600 I→T	XM3600 T→I	平均
日本語 CLIP										
Rinna ViT-B/16 [7]	196	50.6	39.9	90.7	64.0	53.2	84.6	53.8	54.0	61.4
Rinna ViT-B/16 cloob [7]	196	54.6	41.6	88.2	60.3	57.2	80.2	53.4	53.4	61.1
LY ViT-B/16 [15]	196	52.0	**83.8**	96.3	76.7	73.9	**88.4**	**76.9**	**78.0**	**78.3**
llm-jp-clip-ViT-B/16	248	54.2	59.4	91.8	69.2	82.2	85.6	73.6	72.7	73.6
StabilityAI ViT-L/16 [16]	414	**62.4**	70.5	97.6	**84.1**	74.0	86.7	67.3	66.0	76.1
llm-jp-clip-ViT-L/14	467	59.5	62.9	96.4	77.0	**88.2**	87.8	74.1	74.1	77.5
多言語 CLIP										
SigLIP B/16-256 multi [17]	370	51.9	71.2	92.4	65.8	78.6	85.6	45.9	43.0	66.8
jina-clip-v2 [18]	865	35.8	48.1	95.1	58.3	52.0	69.4	67.3	66.4	61.6
LAION ViT-H/14 multi [9]	1193	53.0	74.5	**97.9**	78.4	74.3	85.1	75.0	72.0	76.3

応の CLIP モデルと比較をして評価した.

4.1 評価方法

ゼロショット画像分類及びゼロショット画像・テキスト検索で評価した.

ゼロショット画像分類 ゼロショット画像分類は, CLIP [1] で提案された評価手法に基づいて実施した. まず, 分類対象の画像に対応するラベルをテンプレートを利用して自然な文章形式に変換する. 例えば, “a photo of a {label}” というテンプレートの {label} の部分にラベルを挿入することでラベルを自然な文章に変換する. 次に, 画像の埋め込みとテキスト埋め込みの類似度を計算し, 最も類似度が高いラベルを画像に対する予測クラスとする. テンプレートには `japanese-clip`[12] で提供されている日本語テンプレートを使用した.

評価データセットは ImageNet-1K [19], Recruit[13], CIFAR10 [20], CIFAR100 [20], Food101 [21], Caltech101 [22] を用いた. 英語のクラスラベルは日本語に翻訳して評価した. ImageNet-1K は `japanese-clip` が提供している日本語クラスラベルを用いた. CIFAR10, CIFAR100, Food101, Caltech101 は, 英語のクラス名を DeepL を用いて著者が日本語に翻訳した.

ゼロショット画像・テキスト検索 画像検索 (Text-to-Image Retrieval) は, 与えられたテキストクエリに対して関連する画像を検索するタスクである. このタスクでは, 各テキスト埋め込みに対して, 全ての画像埋め込みとの類似度を計算し, 類似度に基づいて画像をランキングする. 本研究では, 評価指標として Recall@1(ランキング上位 1 件に正解が含まれる割合) を用いた. テキスト検索 (Image-to-Text Retrieval) は与えられた画像クエリに対して関連するテキストを検索するタスクであり, Text-to-Image Retrieval と同様の方法で評価した. 評価データセットには CrossModal-3600 (XM3600) [23] を用いた. XM3600 は 3,600 枚の画像に多言語のアノテーションが付けられたデータセットである. 今回は日本語のアノテーションを用いて評価を行った.

4.2 結果

各モデルの性能を表 2 に示す. llm-jp-clip-ViT-B/16 は, 日本語 CLIP の同モデルサイズの中では, 平均スコアが LY ViT-B/16 に次ぐ性能で, 汎用性の高いモデルであることがわかる. 一方で, 日本特有の画像が多く含まれた Recruit においては, LY ViT-B/16 に対して 30 ポイント以上低いスコアとなった. この原因を探るため, LY ViT-B/16 と llm-jp-clip-ViT-B/16 の埋め込みを可視化して分析した. 分析のために, Recruit データセットのサブセットの jafood の各クラスのテキストの埋め込みと画像の埋め込みの全ての組み合わせに対して, コサイン類似度を計算した.

各モデルの画像・テキスト埋め込みの類似度行列を図 1 に示す. LY ViT-B/16 は llm-jp-clip-ViT-B/16 よりもテキスト埋め込みの正例と負例が綺麗に分離できていることがわかる. Recruit データセットは “交番”や “おでん”, “鎌倉大仏”といった日本文化特有の画像が多く含まれているが, 今回用いた翻訳元のデータセットには類似のドメインのデータが含まれていなかった可能性がある.

12) https://github.com/rinnakk/japanese-clip
13) https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset

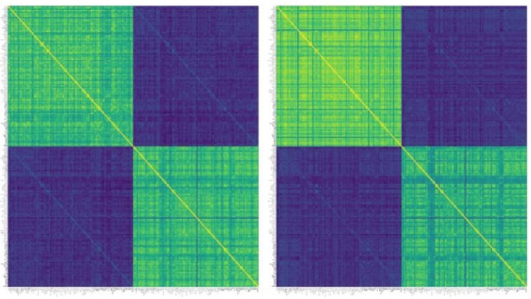

図 1 テキスト・画像埋め込みのコサイン類似度行列. 左: LY ViT-B/16. 右: llm-jp-clip-ViT-B/16. 左上のブロックはテキスト埋め込み同士, 右下のブロックは画像埋め込み同士, 右上・左下のブロックは画像埋め込みとテキスト埋め込み同士の類似度を表す. 明るい色ほど類似度が大きいことを示す.

表 3 画像エンコーダの設定の違い.

設定	ImageNet	XM3600	
		I → T	T → I
フルスクラッチ	**54.2**	**73.6**	**72.7**
継続学習	52.9	71.6	71.7
LiT	52.7	71.7	70.9

4.3 画像エンコーダのアブレーションスタディ

画像エンコーダの最良の設定を調べるため, アブレーションスタディを行った.

学習条件の効果 画像エンコーダの学習条件に関して, 以下の 3 つの条件で実験を行った. (1) フルスクラッチ学習. (2) 継続学習. (3) 学習済み画像エンコーダを固定して学習 (Locked-image Tuning; LiT [24]). 継続学習及び LiT の設定においては, 画像エンコーダモデルの重みの初期値として, LAION の CLIP-ViT-B-16[14] を使用した. テキストエンコーダは全ての設定においてフルスクラッチ学習とした. 継続学習および LiT については, ロススパイク防止のため, 学習率を 1.0×10^{-4} に下げて学習を行った.

各設定の ImageNet の性能推移を図 2 に, 最終結果を表 3 に示す. 先行研究 [24] と同様に, LiT では学習の初期段階で顕著な性能向上が見られる一方, その後の性能向上は緩やかであった. 継続学習では正解率の過程がフルスクラッチと LiT の中間的な挙動を示した. フルスクラッチでは学習初期の ImageNet の性能は低かったものの, 学習が進むにつれて大幅な性能向上が観測され, 最終的に継続学習および LiT を上回った.

モデルサイズの効果 ここでは ViT-B/16, ViT-L/14 を用いた場合の性能比較をした. 画像エンコーダ以外の設定は同一とした. 表 2 に結果を示す. 全てのタスクにおいて, ViT-L/14 が性能を上回った. モデルサイズを大きくするほど性能が向上すること [25] を再確認できた.

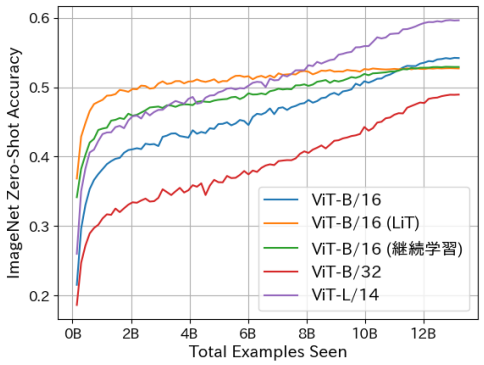

図 2 ImageNet ゼロショット画像分類の正解率の過程.

パッチサイズの効果 画像エンコーダのパッチサイズの設定の違いによる ImageNet の性能差を検証した. ViT [12] は, 画像をパッチに分割した上でパッチの系列を Transformer に入力するため, パッチサイズを小さくするほどトークン長は大きくなりメモリ使用量が増える一方, 画像をより細かく扱える. ここでは ViT-B/32 と ViT-B/16 を用いて評価した. ViT-B/32 について, 損失計算に用いるバッチサイズを 16384, バッチあたりの勾配累積を 2 回, LR を 1.0×10^{-3} とし, 他の設定は ViT-B/16 と同じとした.

ImageNet におけるゼロショット画像分類の正解率の過程の結果を図 2 に示す. ViT-B/16 は ViT-B/32 より一貫して高い性能を示していることがわかる. CLIP [1] においても, パッチサイズが小さいほど性能が高くなることが実験的に示されており, 同様の結果が得られた.

5 おわりに

本研究では, オープン LLM を用いた翻訳によって大規模な日本語画像・テキスト対データセットを構築し, 日本語 CLIP を学習した. 評価の結果, 平均スコアが同程度のモデルサイズの中で高水準であった一方, 日本文化特有のデータセットにおける性能が低かった. この原因として, 翻訳元データのバイアスの影響が考えられる. 多様性のある日本語画像・テキスト対データセットの構築や, 日本語 CLIP のさらなる性能向上については今後の課題である.

14) https://huggingface.co/laion/CLIP-ViT-B-16-laion2B-s34B-b88K/tree/main

謝辞

著者の杉浦は, 2024 年度公益財団岩垂奨学会から奨学金を受給しました.

参考文献

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In **ICML**, 2021.

[2] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In **NeurIPS**, 2023.

[3] Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. ViLA: On pretraining for visual language models. In **CVPR**, 2024.

[4] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022.

[5] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. Chinese clip: Contrastive vision-language pretraining in chinese, 2023.

[6] Federico Bianchi, Giuseppe Attanasio, Raphael Pisoni, Silvia Terragni, Gabriele Sarti, and Sri Lakshmi. Contrastive language-image pre-training for the italian language. **arXiv preprint arXiv:2108.08688**, 2021.

[7] Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. Release of pre-trained models for the Japanese language. In **LREC-COLING**, 2024.

[8] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, Chunjing XU, and Hang Xu. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. In **NeurIPS**, 2022.

[9] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. In **NeurIPS**, 2022.

[10] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In **SOSP**, 2023.

[11] Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. Multilingual machine translation with large language models: Empirical results and analysis. In **NAACL**, 2024.

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In **ICLR**, 2021.

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. **arXiv preprint arXiv:1907.11692**, 2019.

[14] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. OpenCLIP, 2021.

[15] Yokoo Shuhei, Okada Shuntaro, Zhu Peifei, Nishimura Shuhei, and Takayama Naoki. CLIP Japanese Base.

[16] Makoto Shing and Takuya Akiba. Japanese Stable CLIP ViT-L/16.

[17] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. **arXiv preprint arXiv:2303.15343**, 2023.

[18] Andreas Koukounas, Georgios Mastrapas, Bo Wang, Mohammad Kalim Akram, Sedigheh Eslami, Michael Günther, Isabelle Mohr, Saba Sturua, Scott Martens, Nan Wang, and Han Xiao. jina-clip-v2: Multilingual multimodal embeddings for text and images. **arXiv preprint arXiv:2412.08802**, 2024.

[19] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database.

[20] Alex Krizhevsky. Learning multiple layers of features from tiny images, 2009.

[21] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101 – mining discriminative components with random forests. In **ECCV**, 2014.

[22] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022.

[23] Ashish Thapliyal, Jordi Pont-Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A Massively Multilingual Multimodal Evaluation Dataset. In **EMNLP**, 2022.

[24] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. LiT: Zero-shot transfer with locked-image text tuning. In **CVPR**, 2022.

[25] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In **CVPR**, 2023.

表 4　Recruit データセットのサブタスクごとの性能.

モデル	# Params	Recruit				
		jafacility20	jafood101	jaflower30	jalandmark10	Overall
Rinna ViT-B/16 [7]	196M	63.0	28.4	56.5	60.3	39.9
Rinna ViT-B/16 cloob [7]	196M	61.5	27.3	63.5	69.4	41.6
LY ViT-B/16 [15]	196M	**82.0**	**83.8**	**90.5**	**91.8**	**83.8**
llm-jp-clip-ViT-B/16	248M	72.4	52.7	67.0	82.2	59.4
StabilityAI ViT-L/16 [16]	414M	70.8	65.1	89.0	78.6	70.5
llm-jp-clip-ViT-L/14	467M	75.3	55.8	73.5	84.7	62.9
SigLIP B/16-256 multi [17]	370M	64.9	70.7	88.5	68.0	71.2
jina-clip-v2 [18]	865M	80.0	47.1	44.0	48.5	48.1
LAION ViT-H/14 multi [9]	1.2B	80.5	69.1	85.4	89.1	74.5

表 5　ゼロショット画像分類の評価に用いたプロンプトの例.

英語テンプレート	日本語テンプレート
a photo of the {label}	{label} の写真
a sketch of a {label}	{label} のスケッチ
a photo of the cool {label}	かっこいい {label} の写真

A　ゼロショット画像分類の評価に用いたテンプレートの例

japanese-clip[15] にある 37 個のテンプレートのうちの一部を表 5 に示す. {label} にクラスラベルが挿入される.

B　評価データセットの詳細

評価データセットのデータ数, クラス数, ラベルの言語を表 6 に示す. また, 以下に評価データセットの一部の詳細を示す.

Recruit Dataset[16]　7.65K の画像 url とクラスのデータセットであり, 日本固有の概念や事物に関する 4 つの画像分類タスクから構成される. 1.) jafood101: 101 種類の日本料理と食材. 2.) jaflower30: 日本の花 30 種類. 3.) jafacility20：日本の施設 20 種類. 4.) jalandmark10: 日本のランドマーク 10 種類. 画像 url の一部はリンク切れで最終的に 7654 個中 7586 個の画像が取得できた.

XM3600　XM3600 は 1 つの画像に対して複数のキャプションが付けられている. 今回は 1 つ目のキャプションを用いた.

C　Recruit データセットのサブタスクごとの性能

Recruit データセットのサブタスクごとの各モデルの性能を表 4 に示す.

D　比較対象のモデルの詳細

表 7 に比較に用いたモデルが学習に用いたデータセットを示す.

表 6　評価データセットの詳細.

データセット	事例数	クラス数	言語
ImageNet-1K	50,000	1,000	En
Recruit	7,654	161	Ja
CiFAR-10	10,000	10	En
CIFAR-100	10,000	100	En
Food101	25,250	101	En
Caltech101	8,677	101	En
XM3600	3,600	–	En, Ja, etc

表 7　日本語・多言語 CLIP モデルの学習データセット.

モデル	データセット
OpenAI ViT-B/16 [1]	WIT
Rinna ViT-B/16 [7]	CC12M
Rinna ViT-B/16 cloob [7]	CC12M
LY ViT-B/16 [15]	CC12M, YFCC100, CommonCrawl
StabilityAI ViT-L/16 [16]	CC12M, MS-COCO
LAION ViT-H/14 multi [9]	LAION-5B
jina-clip-v2 [18]	DFN, CommonPool
SigLIP B/16-256 multi [17]	WebLI

15）https://github.com/rinnakk/japanese-clip
16）https://huggingface.co/datasets/recruit-jp/japanese-image-classification-evaluation-dataset