日本語 Winoground データセットの自動構築

清水博文　河原大輔
早稲田大学理工学術院
bowen1205@toki.waseda.jp　dkw@waseda.jp

概要

人間レベルの推論ができるマルチモーダルモデルの開発・評価のために、多くのマルチモーダルデータセットが構築されている。本研究では、その中でも視覚・言語の高度な理解を必要とする Winoground データセットの自動構築手法を提案し、日本語 Winoground データセットを構築する。このデータセットにより日本語の視覚・言語モデルの評価が可能になったが、自動構築による生成データセットの質に関してはさらなる改良が必要である。

1　はじめに

人間レベルの推論を行う情報処理モデルの研究が注目を浴びている中、視覚・言語の高度な理解が可能なマルチモーダルモデルの研究が進展している。視覚・言語モデルを評価するタスクの一つである Visual Reasoning は画像内の物体認識を行ったうえで、複雑な推論が必要となるタスクである。Visual Reasoning タスクに分類される Winoground [1] は、2 枚の画像と 2 つのキャプションを結びつけるタスクとなっている。2 つのキャプションで使用される単語は同じだが、並び替えることにより意味が異なっている。そのため、Winoground データセットは視覚・言語の理解能力が必要な高難易度タスクであり、多くのモデルが苦戦している。Winoground はマルチモーダルタスクとして価値のあるデータセットであるが、完全に人手によって構築されているためデータ数が少ないことが問題点としてあげられる。

本研究では、大規模言語モデルと画像生成モデルを利用した Winoground データセットの自動構築手法を提案する。実験では日本語の Winoground データセットを構築し評価を行う。その結果、日本語の視覚・言語モデルも Winoground を解くのが困難であることを確認した。しかし、自動構築したデータセットはオリジナルの Winoground に比べ、低品質なデータが多く存在していることが判明し、データセット自動構築における問題点や大規模言語モデルの弱点が浮き彫りになった。構築した日本語 Winoground データセットは公開予定である。

(a) スケートボードに乗る人を追う犬　**(b)** スケートボードに乗る犬を追う人

図 1: 日本語 Winoground の例

2　関連研究

2.1　Winoground

視覚と言語の理解を必要とするデータセットは多く存在する。画像に適したテキストを検索するタスクの Image-Text Retrieval、画像に関する質問文に回答するタスクの Visual Question Answering、画像に関する説明文を生成する Image Captioning などである。

Image-Text Retrieval タスクの一つである Winoground データセットは、2 枚の画像と 2 つのキャプションをそれぞれ結びつけるタスクとなっており、400 セットから構成される。2 つのキャプションは英語で記述され、同じ単語が使われているが、単語の順番が入れ替えられており意味が異なっている。また、キャプション内の単語の入れ替え方によって Object・Relation・Both の 3 種類のタグがつけられている。Winoground を用いたモデルの評価指標として Image Score, Text Score, Group Score の三種類が提案されている。様々な視覚・言語モデルを評価したところ、人間の正答率は 90%前後であるのに対し、モデルの正答率は最高でも 40%に満たなかったことが報告されている。そのため Winoground は、マルチモーダルモデルの画像・言語

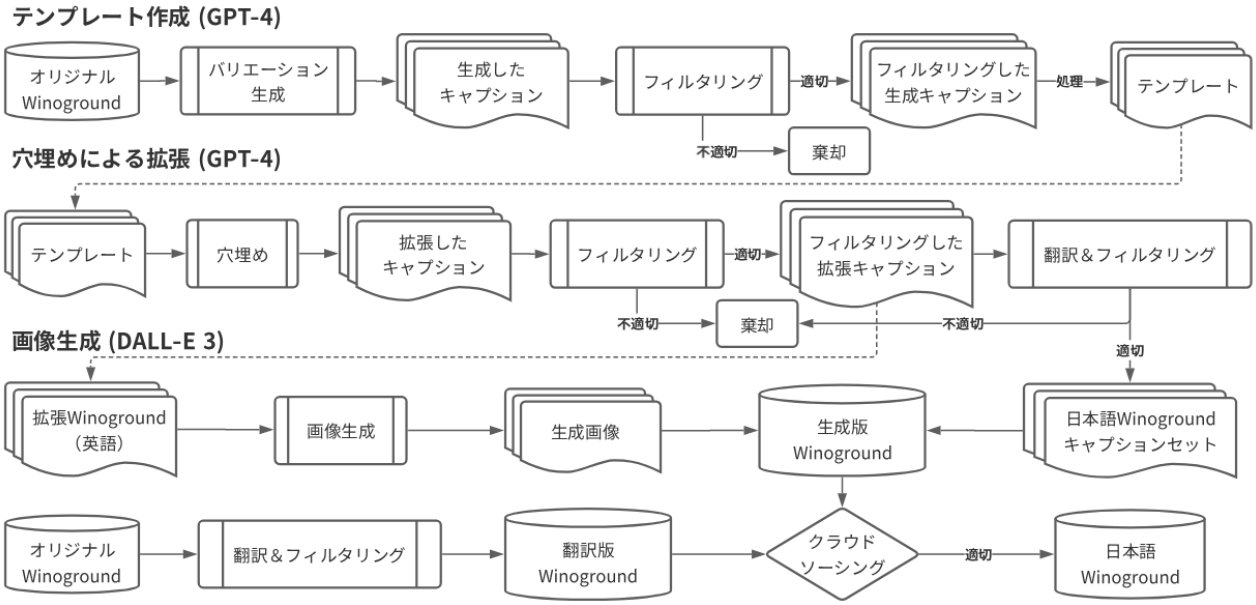


図 2: 日本語 Winoground 構築のフローチャート

理解能力が要求される高難易度のデータセットとされている。

2.2 Few-shot 学習

GPT-4 [2] のような大規模言語モデルが発展し、自然言語生成の精度が向上している。大規模言語モデルを利用し特定のタスクを学習する手法として、推論時にタスクの説明文と少量の例を与えることで特定のタスクに適応する Few-shot 学習が提案されている。ファインチューニングによる手法は大量のデータを用いて追加訓練を行い、モデルのパラメータを微調整するため、データセット収集や学習においてコストがかかる。Few-shot 学習は、必要とされるデータ量が少なく簡単に実装でき過学習を起こさない利点がある。

2.3 画像生成モデル

画像生成モデルの主な手法として GAN [3] や VAE [4] が使われていたが、拡散モデル [5] を用いた手法の研究が進展している。拡散モデルとは、データにノイズを付与されていく過程を学習することにより、ノイズデータからノイズ除去を行い画像生成する手法である。Stable Diffusion [6] や DALL-E 2 [7] で使用されており、テキストに沿った高解像度の画像の生成が可能である。また、DALL-E 3 [8] は GPT によって拡張されたプロンプトを学習することで、DALL-E 2 よりも高い精度の画像生成を可能としている。

3 構築手法

本研究ではオリジナルの英語 Winoground データセットを日本語に翻訳するとともに、新規のデータを GPT-4 および DALL-E 3 により自動生成する。Winoground の条件に適したキャプションを大規模言語モデル GPT-4 の Few-shot 学習を主に用いて生成し、2 枚の画像は画像生成モデル DALL-E 3 を用いて生成する。適切なキャプションを直接生成することは難しいため、テンプレート作成と穴埋めによる拡張の二段階で構築する。テンプレート作成ではデータセットのキャプションパターンの増強を目的とし、穴埋めによる拡張ではデータセットの数を増やす目的で行う。基本的には英語でキャプションの生成を行い、最後に日本語化する。

3.1 テンプレート作成

自動構築における一段階目はキャプションパターンの増強を目的としており、バリエーション生成、フィルタリング、テンプレート生成・拡張の 3 ステップで行う。

3.1.1 バリエーション生成

Winoground の条件に沿ったキャプションのペアを GPT-4 の 3-shot で生成する。タグごとに生成を行い、shot で使用する例はオリジナルの Winoground において同じタグが付与されているキャプションからランダムに選択する。予備実験にて 3-shot、5-shot、

表 1: バリエーション生成用プロンプト

role	content
system	You are a generative AI.
user	Generate one caption pair that have different meanings and are realistic and feasible. The two captions are using the same words. No explanation is needed. {fewshot} Generated caption pair:

表 2: 生成後フィルタリング用プロンプト

role	content
system	You are an expert in English grammar.
user	caption: {caption2} Is this given sentence grammatically correct? Just answer in 1 word, yes or no.
system	You are able to understand the meaning of English sentences.
user	caption1: {caption1} caption2: {caption2} Do the two captions have different meaning? Just answer in 1 word, yes or no.

10-shot を比べたところ、3-shot が生成単語の偏りが少なかったため採用する。付録 A.1 に shot 数ごとの生成文の例を示す。生成におけるプロンプトを表 1 に示す。

3.1.2 フィルタリング

GPT-4 の Winoground 生成の傾向として、一文目は意味の通るキャプションであっても、二文目は単語が入れ替わっただけの意味が通らないキャプションになっていることが多い。そのため、GPT-4 を用いて二文目が自然で意味が通る文かの確認を二回行い、さらに生成した二つのキャプションが異なる意味であるかの確認を行う。フィルタリングにおけるプロンプトを表 2 に示す。この手順により、入れ替えても同じ意味になるキャプションペア（例えば「彼は犬と猫を飼っている」「彼は猫と犬を飼っている」のようなペア）が生成されないようにする。

3.1.3 テンプレートの生成・拡張

フィルタリングによって自然な文と判断されたキャプションペアにおいて、入れ替えられた単語の部分を穴埋め部分としてテンプレートを生成する。さらに、GPT-4 を用いて、穴ではない部分の単語を別の単語に置き換えることにより、テンプレートを拡張する。

表 3: 語句の穴埋め用プロンプト

role	content
system	You are a generative AI.
user	This is a fill-in-the-blank question. {question} Answer two noun phrases that could fill each other's [blank]. Answer 10 sets of noun phrases. Generated Answers:

表 4: 拡張後フィルタリング用プロンプト

role	content
system	You are able to understand the meaning of English sentences.
user	caption: {caption2} Is this caption realistic and feasible?
system	You are a photographer.
user	caption: {caption2} Is it possible to take a photo as per the caption?
system	You are an illustrator.
user	caption: {caption2} Is it possible to illustrate as per the caption?

3.2 穴埋めによる拡張

自動構築における二段階目の穴埋めによる拡張は、データセットの量を増やすことを目的としており、語句の穴埋め、フィルタリングの 2 ステップで行う。

3.2.1 語句の穴埋め

作成したテンプレートから GPT-4 を用いてデータ拡張を行う。拡張におけるプロンプトを表 3 に示す。一つのテンプレートあたり 10 セットの語句ペアを出力するように設定する。

3.2.2 フィルタリング

3.1.2 節と同様に、自然なキャプションであるかのフィルタリングを行う。ここでは GPT-4 に簡易的なペルソナを与えることにより、画像として描写できるかどうかのフィルタリングも同時に行う。付与するペルソナは「英語話者」「写真家」「イラストレータ」の三つとする。それぞれのプロンプトを表 4 に示す。

3.3 日本語翻訳

これまでのステップで得られた、画像として描写可能で自然なキャプションについて、GPT-4 を用いて日本語に機械翻訳する。ここで、オリジナルの Winoground のキャプションの日本語翻訳も行う。

表 5: 翻訳版および生成版 Winoground を用いたモデルの評価

モデル	言語	翻訳版			生成版			翻訳版+生成版		
		Text	Image	Group	Text	Image	Group	Text	Image	Group
人間	日本語	80.42	87.30	78.31	43.59	46.67	29.74	61.72	66.67	53.65
ランダム選択	-	25.00	25.00	16.67	25.00	25.00	16.67	25.00	25.00	16.67
rinna/japanese-clip-vit-b-16	日本語	22.22	4.76	2.12	21.54	6.15	2.05	21.88	5.47	2.08
rinna/japanese-cloob-vit-b-16	日本語	23.81	7.94	5.82	21.54	6.15	3.59	22.66	7.03	4.69
openai/clip-vit-base-patch32	英語	25.40	7.41	5.82	27.69	11.28	7.69	26.56	9.38	6.77
Salesforce/blip-itm-base-coco	英語	36.51	13.76	10.58	32.31	12.82	8.21	34.38	13.28	12.68

翻訳したキャプションが Winoground の条件に適しているか、さらにキャプションの意味が通じるかについて GPT-4 で確認する。

3.4 画像生成

3.2 節で生成した英語のキャプションに基づいて画像を DALL-E 3 で生成する。3.2.2 節で行った画像描写可能性のチェックに基づいて、プロンプトに画像のスタイル（イラストまたは実写）を追加する。生成した画像が日本語キャプションと合致しているかを Yahoo!クラウドソーシングで確認する。画像とキャプションの合致度を 3, 2, 1, 0 の 4 段階の中から選択するタスクを 5 人に実施し、平均スコアが 2.0 を超えるデータを採用することで、日本語 Winoground データセットが完成する。

4 実験

4.1 データセット

日本語 Winoground は 384 セットからなる。そのうちオリジナル Winoground の翻訳が 189 セット (「翻訳版」と呼ぶ)、3 節の手法で生成したものが 195 セット (「生成版」と呼ぶ) である。キャプションは英語と日本語が含まれているが、オリジナルにあるタグ付けは実施していない。

4.2 実験設定

日本語 Winoground を用いて、日本語と英語の視覚・言語モデルの評価を行う。日本語モデルは CLIP モデルの rinna/japanese-clip-vit-b-16 と CLOOB モデルの rinna/japanese-cloob-vit-b-16 [9] を使用し、英語モデルは CLIP モデルの openai/clip-vit-base-patch32 [10] と BLIP モデルの Salesforce/blip-itm-base-coco [11] を使用する。モデルは画像・キャプションペアのスコアを計算し、スコアを比較することで評価を行う。さらに、Yahoo!クラウドソーシングを用いて、人間による評価も行った。

評価指標は Winoground で提案されている Text Score, Image Score, Group Score の三種類を用いる。これらの詳細は付録 A.3 に示す。

4.3 実験結果

結果を表 5 に示す。英語モデルにおいては Text Score がランダム選択を超えたが、Image Score および Group Score は大きく下回った。一方、日本語モデルではどの指標もランダム選択を超えることができない結果となった。人手評価においては、翻訳版における精度は 80%前後あるのに対し、拡張版においては 50%に満たない精度となった。

4.4 考察

人手評価の結果から、生成版の Winoground は翻訳版よりも低品質なデータが多く存在していることがわかる。データセットを分析すると、二つのキャプションの意味がほぼ同じになってしまっているものが多く見つかった。付録 A.4 の (c,d) に生成された低品質データの例を示す。3.1.2 節のフィルタリングで同じ意味の文ペアの除外ができていないことに原因があるため、GPT-4 は細かな意味理解が苦手であることが示唆される。

5 おわりに

本研究では、日本語の Winoground データセットを大規模言語モデルと画像生成モデルを用いて構築し、日本語と英語の視覚・言語モデルを評価した。モデルの評価には使用できるものの、構築したデータセットには低品質なデータが多く含まれていることが判明した。高品質なデータを大量に生成するためには、自動構築の手順や言語生成におけるプロンプトの改良が必要である。

謝辞

本研究は SB Intuitions 株式会社と早稲田大学の共同研究により実施した。

参考文献

[1] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In **CVPR**, 2022.

[2] OpenAI. Gpt-4 technical report. **ArXiv**, Vol. abs/2303.08774, , 2023.

[3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, **Advances in Neural Information Processing Systems**, Vol. 27. Curran Associates, Inc., 2014.

[4] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In **2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings**, 2014.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. **arXiv preprint arxiv:2006.11239**, 2020.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

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

[8] James Betker, Gabriel Goh, Li Jing, † TimBrooks, Jianfeng Wang, Linjie Li, † LongOuyang, † JuntangZhuang, † JoyceLee, † YufeiGuo, † WesamManassra, † PrafullaDhariwal, † CaseyChu, † YunxinJiao, Aditya Ramesh. Improving image generation with better captions.

[9] シーン誠, 趙天雨, 沢田慶. 日本語における言語画像事前学習モデルの構築と公開. In **The 25th Meeting on Image Recognition and Understanding**, 2022.

[10] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[11] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

A　付録

A.1　shot 数による生成キャプションの違い

shot 数を増やすと同じ単語が多く生成されてしまうので、本研究では N=3 を採用した。表 6 に shot 数ごとの生成キャプションの例を示す。なお、生成キャプションはキャプションペアの一文目のみを表示している。

表 6: shot 数と生成キャプション

N-shot	生成キャプション
3	a doctor operating on a patient a dog sits on a man Running shoes on the beach the glass is half full of water She can't bear the children
5	firefighters rush to save the forest a cat sitting on a mat Running through the park a man eating a crab light upon the leaves
10	a cat sitting on a laptop The cat is under the table with a hat the light is under the bridge the cat is chasing the mouse a light is hanging over the painting

A.2　ステップごとのデータ生成・除外数

バリエーション生成における文セット生成数のみを指定し、日本語キャプションの翻訳およびフィルタリングまでを自動で行っている。

- テンプレート作成
  - バリエーション生成：500 セット (指定)
  - フィルタリング：159 セット
  - 穴埋め問題作成：291 セット
- 穴埋めによる拡張
  - 穴埋め：2,910 セット
  - フィルタリング：775 セット
- 日本語化
  - 翻訳&フィルタリング：306 セット

A.3　評価指標

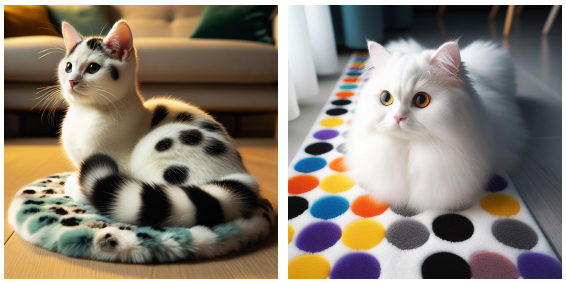
2 枚の画像を I_0, I_1、 2 つのキャプションを C_0, C_1 とする。なお、同じインデックスの画像とキャプションは対応している。モデルは画像・キャプションペアのスコア $s(C \cdot I)$ を 4 通り計算し、それぞれスコアを比較することで評価を行う。Text Score は与えられた画像に対して、正しいキャプションを選択できるかの指標 (式 (1)) であり、Image Score は与えられたキャプションに対して正しい画像を選択できるかの指標 (式 (2)) である。Group Score は 2 枚の画像と 2 つのキャプションを完璧にマッチングできるかどうかの指標 (式 (3)) であり、Text Score と Image Score によって決定される。

$$f(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_1, I_0) \\ & \text{and } s(C_1, I_1) > s(C_0, I_1), \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

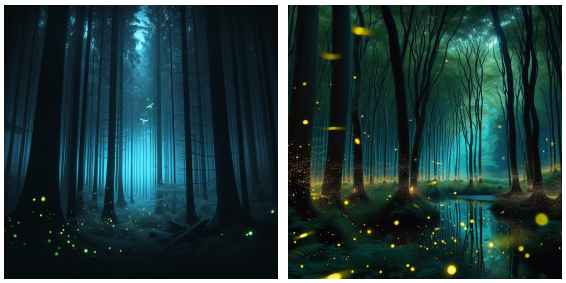
$$g(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } s(C_0, I_0) > s(C_0, I_1) \\ & \text{and } s(C_1, I_1) > s(C_1, I_0), \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

$$h(C_0, I_0, C_1, I_1) = \begin{cases} 1 & \text{if } f(C_0, I_0, C_1, I_1) \\ & \text{and } g(C_0, I_0, C_1, I_1), \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

A.4　生成版 Winoground の具体例



(a) 模様のある白い猫がふわふわのマットに座っている　**(b)** ふわふわの白い猫が模様のあるマットに座っている



(c) 夜の森のホタル　**(d)** ホタルの森の夜

図 3: 高品質データ（a,b）と低品質データ (c,d)