The 39th Annual Conference of the Japanese Society for Artificial Intelligence, 2025

大規模言語モデルを活用した物体間の左右の位置関係を含む画像キャプションデータセットの構築

Construction of an Image Caption Dataset Utilizing LLMs Including Left-Right Positional Relationships Between Objects

守屋 響[*1]
Hibki Moriya

大和淳司[*1]
Jyunji Yamato

[*1] 工学院大学
Kogakuin University

We constructed a new dataset for image captioning, including the left-right positional relationship between two objects appearing in an image, by utilizing LLMs for generation and cross-verification. Traditional datasets lack information about the positional relationship between objects, and were reliant on datasets where positional information was manually annotated. However, preparing a large volume of such manually annotated data is challenging. Therefore, we created a dataset by generating captions that include positional relationships from images using multiple LLMs. However, the accuracy of the generated captions depended on the models, with about 60-80% being correct. To improve the dataset's accuracy, we refined the data by re-evaluating the generated captions with LLMs to distinguish between correct and incorrect annotations. When self-evaluation was performed by the LLM that generated the captions, there was a tendency to incorrectly mark errors as correct. Nevertheless, through cross-verification between different LLMs from companies such as OpenAI, Google, and Anthropic, we were able to improve the accuracy of the generated caption dataset.

1. はじめに

画像キャプション生成は，画像の内容を説明する文章を自動的に生成する技術であり，正確なキャプションは画像の詳細な理解につながる．また，画像検索や画像に基づく質問応答などに応用できる．本研究では，画像のより詳細な理解のために，画像に登場する 2 つのオブジェクト(人物，動物，乗り物など)について位置関係，特に左右の位置関係を含むキャプションデータセットの構築を行う．

従来，モデルを調整することで元のキャプションに含まれている位置関係を抽出しようと試みられてきた．画像内のオブジェクトの位置関係を把握するために，[C Wang 2022]では Geometry Attendant Transformer (GAT)モデルが提案され，幾何学的表現を得る新たなアプローチを導入している．[Yang 2022]ではシーングラフを用いて関係情報を埋め込んだ特徴を生成する ReFormer が提案されており，Transformer を応用してペアごとの関係を記述するシーングラフの生成を可能にしている．

画像キャプション生成は一般的に人手によって作成された画像とキャプションのセットを用いて学習するが，使用される学習データセットに位置関係が含まれているのは稀である．[守屋 2024]では人手でアノテーションしたデータセットを構築している．これにより，オブジェクトの位置を属性として付与できることが確認された．しかし，人手はコストが高く容易に数を増やせないという課題がある．

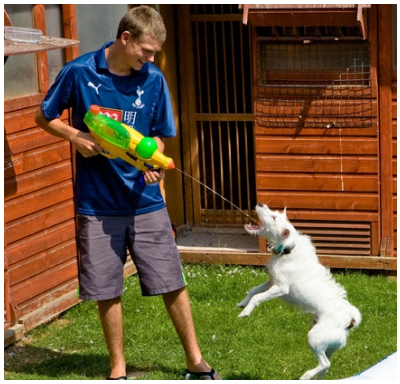
そこで，位置関係を含むキャプションの生成のために，LLM を用いてデータセットを作成した．LLM の性能向上により左右の位置関係を含む画像キャプションが生成できるようになったが，ハルシネーションと呼ばれる誤りを含んでいる可能性がある．本研究では LLM による生成と相互評価を用いるデータセット構築を提案する．複数の LLM を用いてキャプションの生成と評価を行いデータセット作成を実施する．

連絡先：守屋響，080-5874-0190，em23044@g.kogakuin.jp

2. 提案手法

2.1 左右の位置関係

左右の位置関係を含むキャプションの例を図 1 に示す．この写真では左側にいる男性が右側の犬に向かって水鉄砲を打っている様子が捉えられている．左右の情報がなければ，キャプションを読んだだけでは犬と男性の位置関係が明確ではない．左右の関係が示されることで，VQA タスクで左右の質問に対応することや，応用としては視覚障害者支援において，具体的な位置情報の提示が可能になる．



一般的なキャプション
The young man with the water gun is shooting water into the white dog mouth.
左右の位置関係を含むキャプション
The young **man on the left** with the water gun is shooting water into the white **dog on the right** mouth.

図 1 左右の位置関係を含むキャプション

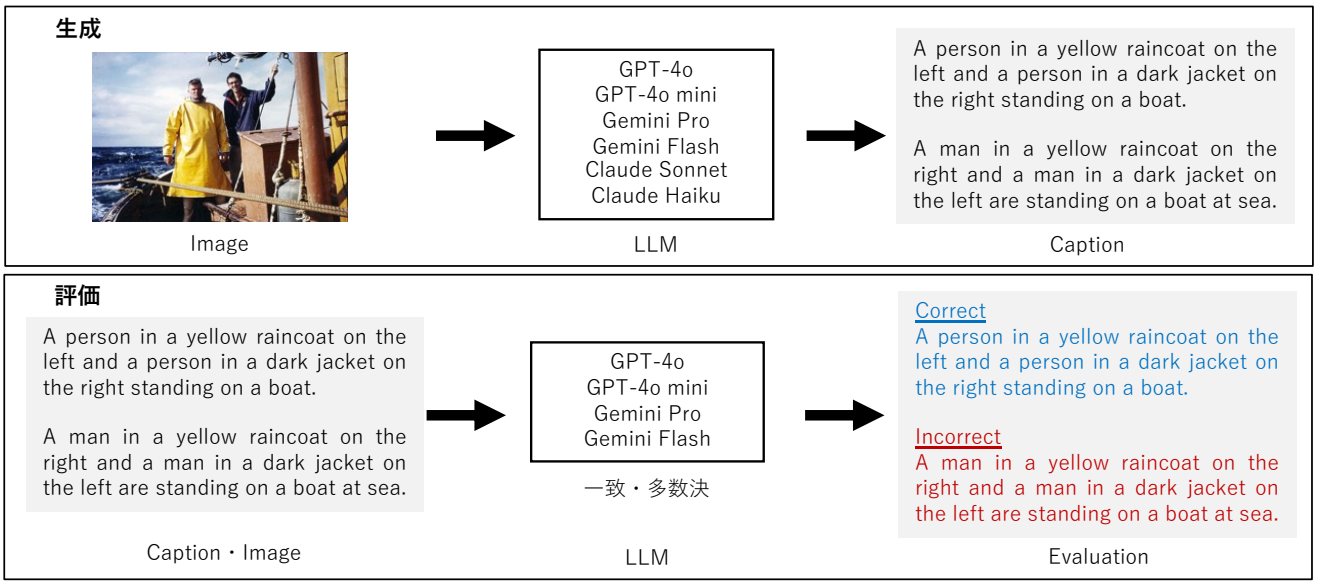


図 2　キャプション生成・相互評価の概要.

表 1　使用 LLM version

LLM　Model	version
GPT-4o	gpt-4o-2024-08-06
GPT-4o mini	gpt-4o-mini-2024-07-18
Gemini Pro	gemini-1.5-pro
Gemini Flash	gemini-1.5-flash
Claude Sonnet	claude-3-5-sonnet-20240620
Claude Haiku	claude-3-haiku-20240307

2.2 キャプション生成

ChatGPT の登場以降，LLM の利活用が急速に進展している．OpenAI の GPT をはじめとするマルチモーダル対応の LLM は，画像と言語の関係性を効果的に捉えることが可能となった．その結果，画像中の細かな文脈や位置情報を含むキャプション生成が実現されている．これにより，LLM を用いることで左右の位置情報を反映したキャプションを生成することが可能となった．しかし，LLM は計算リソースが高く小型システムには向かないことなど課題も存在する．そこで本研究ではパラメータ数が 1/1000 程度である BLIP[Li 2022]の Fine Tuning に使用するデータセット構築を，LLM を用いて実施する．図 2 に本研究の概要図を示した．

データセットに使用するキャプションの生成は多様性を確保するため，複数の LLM を用いてキャプション生成を行った．使用した各モデルを表 1 に示す．対象は GPT 系，Gemini 系，Claude 系の標準モデルおよび軽量モデルである．生成されたキャプションについては，200 枚のサンプルを用いて人手による精度評価を実施し，その結果を図に示す．なお，キャプションの内容自体は正しいものの，左右の区別が不明確な場合は誤りと定義した．評価結果から最も高い正解率を示したのは Gemini 1.5 Pro であり，最も低い正解率を示したのは Claude3 Haiku であることが確認できる．また，本サンプルにおいては生成キャプションのうち 16.3%に誤りが含まれていることが明らかになった．次に生成したキャプションの LLM による評価選別を実施した．

2.3　相互評価

本研究では，生成したキャプションの妥当性を再度 LLM を用いて評価・選別する手法を提案する．評価には表 1 に示すモデルのうち Claude 系を除く 4 モデルを用いた．その理由は，各モデルが評価結果に与える影響度と評価の妥当性に差異がみられたためである．

本手法においては，キャプションの選別手法として多数決方式と一致方式の 2 手法を検討した．

(1) 多数決方式:4 つのモデルのうち過半数が正解と判定したキャプションを使用する．

(2) 一致方式:4 つのモデルの全てが正解と判定したキャプションを使用する．

予備実験の結果，一致方式では厳格な選別が行われる結果，正解であるにもかかわらず多くのデータが除外され，データセット全体の規模が大幅に減少することが確認された．特に，多数決方式において採用されていたが一致方式では除外されたデータの中で，Claude 3.5 Sonnet に起因するものが 43.5%を占めることが明らかになった．一方で，人手による評価との一致率では，GPT 系や Gemini 系が高く，Claude 系の信頼度は他と比較して低いことが確認された．

以上の結果から，本評価実験では評価精度が低く，かつ影響度が高いと考えられる Claude 系モデルを外し，4 モデルによる多数決方式および一致方式の比較を行った．

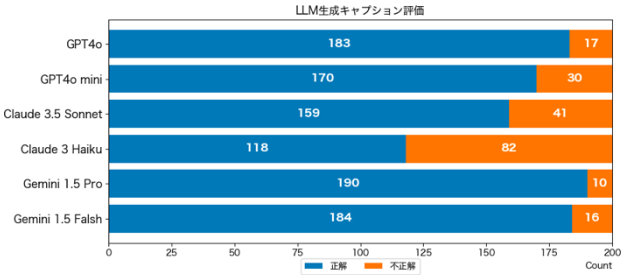


図 3　LLM 生成キャプション評価結果

2.4 評価結果

図 3 に示している 1200 件のキャプションに対して，LLM を用いた評価を実施した．その結果を，表 2(一致方式の場合)および表 3(多数決方式の場合)に示す．

表の列は正解を T，不正解を F とし，左側が実際の正誤，右側が評価結果となっている．

一致方式では，不正解と判定されたキャプションの割合は 4.5%であった．一方，多数決方式では 8.1%が不正解と判定された．なお，元データにおける誤り率は 16.3%であったことから，いずれの方式においてもデータ品質の向上が確認された．
また，それぞれの方式を比較すると，正解として判定されたキャプションの数は，一致方式で 393 件，多数決方式で 789 件となった．多数決方式では不正解の割合がやや増加するが，正解であるが不正解として除外されるキャプションの数が大幅に減少していることが確認できる．

表 2　一致方式

一致方式	T→T	T→F	F→T	F→F
GPT-4o	58	125	2	15
GPT-4o mini	64	106	5	25
Gemini Pro	75	84	6	35
Gemini Flash	30	88	4	78
Claude Sonnet	89	101	1	9
Claude Haiku	59	125	0	16

表 3　多数決方式

多数決方式	T→T	T→F	F→T	F→F
GPT-4o	133	50	5	12
GPT-4o mini	120	50	14	16
Gemini Pro	129	30	23	18
Gemini Flash	67	51	11	71
Claude Sonnet	141	49	6	4
Claude Haiku	135	49	5	11

3.　考察

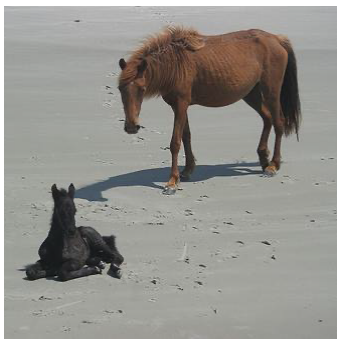
本手法では，LLMを用いた選別処理により，選別前のデータと比較して誤りが含まれる件数を大幅に低下できることが確認された．表 4 に Claude 3 Haiku（不正解：82）に対し各 LLM で誤りを正解とした件数を示した．これより，評価に用いるモデルが単一である場合は誤りの検出が不十分であることがわかる．複数の LLM による相互評価を実施することで，各モデルの弱点を補完し，誤りの除去効果が向上することが確認できる（表 2，表 3 参照）．ただし，複数のモデルで選別を行った結果，総データ数は約半分まで減少するというトレードオフの関係となった．

	Pro	4o	Flash	4o mini
Claude Haiku	25	9	23	39

表 4　Haiku に対する評価

また，評価に用いる LLM の系列について検討する．実験では，各系列の標準モデル（GPT-4o，Gemini Pro）と軽量モデル（GPT-4o mini，Gemini Flash）の 2 種類を用いたが，系列内のモデル同士の評価は，他系列に対する評価と比較して評価が甘くなる傾向が見受けられた．これより，異なる系列のモデルを組み合わせた相互評価は有効性があると考えられる．

次に評価選別によって取り除くことできなかったキャプションについて分析する．図 4 は左右が逆になっているが全ての LLM においてこのキャプションを正解としていた．図 5 では右側の人がかぶっている帽子が青色であるが，帽子を誤ってジャケットとして記述している．このように誤って評価した原因として，例えば両者とも「馬」という同種のオブジェクトであることや「青い帽子」と「青いジャケット」という似た属性を持っている点で誤評価している．また，図 6 では両者は似た属性を持っており区別できるポイントとしては姿勢や，持っているものが考えられる．実際には区別できないが，キャプション単体では区別ができている点で正解としている可能性が考えられる．



A large brown horse is on the left, and a smaller dark-colored foal is on the right, resting on the beach.

図 4　オブジェクトの左右の位置関係を逆に記述している

A construction worker in a gray outfit is on the left, and a person in a blue jacket is on the right, standing in front of a building entrance.

図 5　認識ミス：青いジャケットではなく帽子をかぶっている

A man in a neon green vest is on the left, and a man in a hard hat is on the right, digging in the ground.

図 6　どちらのオブジェクトにも含まれている属性であるため区別できない

不正解を誤って正解としているキャプションの多くは属性の一部が正しい場合が多く確認された．例えば，色は合っているがオブジェクト名（T シャツをズボンと間違える）が間違っている場合である．また，プロンプトにはオブジェクトの区別がつくようにキャプションが記述されている必要があると記載しているが，図 6 のように共通する属性を持っているオブジェクトについて，区別ができないキャプションを誤りと判定できていないことが多くある．これらの判定ミスを減らすために，実験では共通のプロンプトを使用していたがモデルごとの最適化やモデルの性能による評価の重み付けなどが考えられる．

4. まとめ

本研究では，画像内に存在する 2 つのオブジェクト間の左右の位置関係を明示的に記述するキャプション生成を目指し，LLM を活用したデータセット構築およびその相互評価による選別手法を検討した．LLM 同士による相互評価によって誤りキャプションを検出し，一致方式・多数決方式により選別を実施すると，データセットの品質向上につながることが明らかとなった．しかし，本手法ではデータ量と品質はトレードオフになり，精度向上を目指すとデータ量が大幅に減少した．正解キャプションを不正解と判定しないよう，プロンプトの最適化を実施する必要性が考えられる．一方で，LLM によるキャプション生成は人手作業と比較して大幅にコストが低いことから，データ量が確保できるまで複数回キャプションを生成することで，精度を担保し，データ量を担保するアプローチが考えられる．

参考文献

[C Wang 2022] Chi Wang, Yulin Shen, Luping Ji: “Geometry Attention Transformer with position-aware LSTMs for image captioning”, Expert Systems with Applications, 2022.

[Yang 2022] Xuewen Yang, Yingru Liu, Xin Wang: “ReFormer : The Relational Transformer for Image Captioning”, Proceedings of the 30th ACM International Conference on Multimedia, 2022.

[Li 2022]Junnan Li, Dongxu Li, Caiming Xiong, Steven Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation", 2022

[守屋 2024] 守屋響, 大和淳司 “登場する物体の左右の位置関係を含む画像キャプション生成”, JSAI, 2024.

付録（プロンプト）

生成プロンプト

```
Please generate a caption in English for an image
   containing two objects, indicating which object
   is on the left and which is on the right.
For example:'A girl in a white shirt is on the
   left, and a girl in a black and white shirt is
   on the right are playing game.'
Do not use the same attributes so that objects can
   be distinguished from each other.
Ensure that the left and right are correct,
   objects are distinguishable, and there are no
   recognition errors.
Do not group multiple representations of an object
   into one. For example:' Two girls are playing
   a game.'
Generate a caption that does not have these
   issues. Only generate the caption.
```

評価プロンプト

```
    You are an expert evaluator tasked with
 determining whether a given caption accurately
  and specifically describes the left and right
    sides of an image. Your evaluation should
    consider the following detailed criteria:

1. Presence of Left/Right Information
 a. Confirm that the caption explicitly mentions
both "left" and "right" sides of the image.
 b. If the caption includes only one side or
omits positional details altogether, note this
as an error.

2. Specificity and Clarity of Descriptions
 a. Check whether the caption includes
sufficiently detailed attributes (e.g., color,
size, shape, posture) that distinguish objects
on the left side from those on the right side.
 b. A vague description (e.g., "a dog on the
left and a dog on the right") that lacks
differentiating details should be considered
insufficient.
 c. Conversely, a detailed caption like "A white
dog is sitting on the left, and a black dog is
standing on the right" meets the criteria.

3. Consistency with the Actual Image
 a. Evaluate if the objects and their described
attributes (including positions) match what is
visible in the image.
 b. Identify any misidentifications or
discrepancies in object positions or attributes.

4. Final Evaluation:
 Using the reasoning above, decide if the
caption meets the following:
  • It accurately conveys what is on the left
side of the image.
  • It accurately conveys what is on the right
side of the image.
  • It provides sufficient specific details for
the observer to distinguish between the objects
on each side.
  • Minor grammatical errors do not count as
mistakes unless they impact clarity.
 If all criteria are met, the caption is
correct.

Evaluation Format:
 Caption: {caption}
 Final Answer:
  • Output only "TRUE" if the caption fully
satisfies the criteria.
  • Output only "FALSE" if there are any
shortcomings.
```