視覚的質問応答における視線情報を利用した曖昧性解消に向けて

稲積 駿 [1,2]　河野 誠也 [2]　湯口 彰重 [2,1]　川西 康友 [2,1]　吉野 幸一郎 [2,1]
[1] 奈良先端科学技術大学院大学　[2] 理化学研究所ガーディアンロボットプロジェクト
inazumi.shun.in6@is.naist.jp
{seiya.kawano,akishige.yuguchi,yasutomo.kawanishi,koichiro.yoshino}@riken.jp

概要

視覚的質問応答 (VQA: Visual Question Answering) は，画像に関する質問が与えられた時に回答を導くタスクであり，質問と画像中の情報から回答が一意に決定する状況を仮定する．しかし，VQA を人間と実世界対話を行うロボットに応用する場合，主語の省略や指示語が，視線などのマルチモーダルな補完情報（文脈情報）と共に利用される場合がある．本研究では画像中の人物が見ている対象の情報を利用した VQA データセットを構築し，視線の先の物体名を VQA の曖昧性解消に利用した結果を報告する．

1　はじめに

実世界での事物を考慮した人間とロボットの自然なインタラクションを実現することは，Vision and Langauge 研究の到達点の一つである．この際，ロボットは自身が見ている視覚情報と人間が発する言語情報を統合・理解して，適切な応答を返す能力を備える必要がある．視覚的質問応答 (VQA: Visual Question Answering) は，画像に関する質問が与えられた際に質問に対する回答を導くタスクであり，視覚・言語情報の統合に関する重要なベンチマークが公開されている [1, 2, 3]．

既存の VQA に含まれる質問は，その意図が明確であり回答が一意に決まることが多い．しかし，人間同士のインタラクションに見られるように，実際に人間が行う質問は，指示語や主語の省略によって曖昧さが生じその回答が一意に決まらない場合がある [4]．図 1 上は既存の VQA データセットに含まれる質問の例であり，図 1 下は指示語や主語の省略により曖昧性が生じる質問の例である．

本研究では，VQA における質問の曖昧性を解消するために，視線 [5] に代表されるマルチモーダルな文脈情報を用いることを想定し，視線情報付き VQA データセットおよびタスクを提案する．視線情報付き VQA タスクは画像内の人物が発するであろう曖昧な質問に対して，その人物の視線情報を考慮して回答する．例えば，図 1 の図下に示された曖昧な質問は，「傘を差している女性」が「赤いジャケットを着た男性」に注視していることを追加情報として用いれば，回答を一意に定めることができる．本研究ではさらに，視線の先の物体が扱えるように既存の VQA モデルの入力を拡張し，視線付き VQA タスクによるモデルの評価を行った．

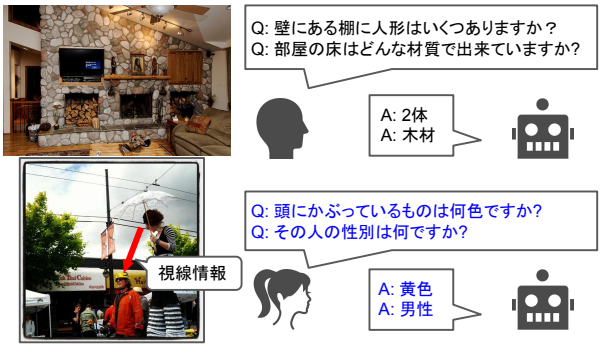

図 1　日本語 VQA データセット [3](図上) と視線情報付き VQA データセット (図下) の例．図下は，画像，質問，回答に加え視線の視点と終点が付与されている．

2　関連研究

視覚的質問応答　VQA は画像処理と自然言語処理の融合分野である Vision and Language のタスクの一つに位置する [1, 2, 3, 6]．VQA のタスク設定は，用意された回答群から回答を選択する設定 (選択型) と回答を生成する設定（生成型）の二つが存在する．本研究では，生成型の設定を想定して質問と回答を収集した．VQA では基本的に画像と質問の情報から一意に定まる質問・応答ペアが用意されるが，本研究は視線情報を持たない場合に回答が曖昧になるようなデータセットを構築した．

表 1　日本語 VQA データセット [3] と視線情報付き VQA データセットの統計情報.

データセット	画像数 [枚]	質問・回答セット [件]	質問 / 回答の平均文字数 [文字]	ユニークな質問 / 回答 [件]
日本語 VQA	99,208	793,664	14.942 / 4.561	358,844 / 135,743
視線情報付き VQA	14,000	26,299	14.820 / 5.319	10,311 / 7,849

注視対象推定　注視対象推定とは，画像に映る人物を選択し，その人物が注視しているオブジェクトを推定するタスクである．代表的なデータセットとして，Gazefollow がある [7]．Gazefollow は，MSCOCO [8] を含む様々な画像データセットから収集した人物を含む画像に対し，人物の視線元と視線先のアノテーションを付与したものである．視線付き VQA データセットの構築は，Gazefollow の画像と視線情報を利用した．

質問応答における曖昧性　言語に生じる曖昧さの問題は質問応答システムで活発に取り組まれており，質問の曖昧性は言い換え [9, 10] や質問生成 [11, 12] により解消するアプローチが検討されている．

VQA においても，曖昧さの議論はなされている [13, 14]．選択型の VQA では，同じ内容に対する言い換えに対応するべく，自動評価のために 10 件程度の回答が付与される．しかし，画像が不鮮明な場合や質問が曖昧である場合，全ての回答が一致するとは限らない [13]．既存研究では回答が一意になるように問い返しの質問を生成する [15] ことで選択型 VQA の問題は解消できることが示唆されている [14]．本研究ではこのようなユーザから追加情報を引き出そうとするアプローチとは異なり，画像の状況から得られる追加の文脈情報を利用することで曖昧性の解消を試みる．

3　視線情報付き視覚的質問応答

3.1　タスク設定

実験では注視対象推定のタスクを扱わず，視線の先の物体名が既に与えられていると仮定し，視線情報付き VQA のタスクを次のように定義する．

視線情報付き VQA　画像，質問，視線情報先の物体名が与えられた時，モデルは回答を生成する．

3.2　データ収集方法

視線情報付き VQA データセットは，クラウドソーシング[1)]によって構築した．Gazefollow [7] に含まれる画像に対して，人物が見ている対象に関する質問を作成し，質問に対する回答を与える．視線の先が物体を差していない場合や画像が不鮮明である場合は質問・回答作成の対象から除いている．

ワーカに対する教示を以下に示す．

- 質問の文字数は 10 文字以上とする．
- 視線の先に存在する物体名を直接質問に含めない．
- 画像内の情報のみで回答できる質問を作成する．

一つ目の項目は，多様な語彙の質問を用意する目的で設定した．二つ目の項目は，視線情報を持たない場合に曖昧となる質問を作成する目的で設定した．三つ目の項目は，画像の内容以外で曖昧になる質問をデータセットに含めない目的で設定した．

3.3　統計と分析

既存の日本語 VQA データセット [3] と今回収集した視線情報付き VQA データセットを比較する．図 2 と図 3 はそれぞれ質問と回答に関する名詞の分布であり，各データセットから質問・回答ペアを 25,000 件ランダムサンプリングした．表 1 はそれぞれの統計情報である．

質問　表 1 より，ユニークな質問の割合は視線情報付き VQA (39.2%) より日本語 VQA (45.2%) が多く，質問の文字長は同程度である．図 2 より，視線付き VQA は日本語 VQA より分布がロングテールであり，質問タイプは偏っている．とくに「what」タイプの質問が多くを占めている．これは，視線の先の物体を質問作成の対象としたことに起因する．

回答　表 1 より，ユニークな回答の割合は日本語 VQA (17.1%) より視線情報付き VQA (29.8%) が多く，回答の文字長は視線情報付き VQA が長い．これは，「人物の行動」などを問う質問が多くみられたことに起因する．結果として一意なフレーズで表現できない回答が収集された．図 3 より回答の分布に大きな差は無い．しかし，クラウドソーシング対象とした Gazefollow の画像セットが限定的であるため，視線付き VQA の回答分布の上位は，「テニス」など特定のドメインに偏っている．

1)　株式会社クラウドワークス, https://crowdworks.jp/

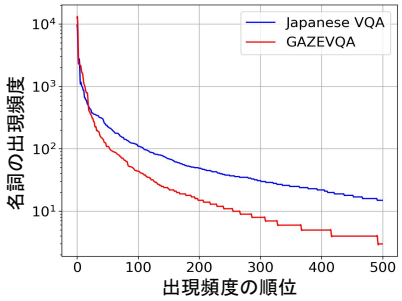

図 2　質問セットの単語数に関する分布

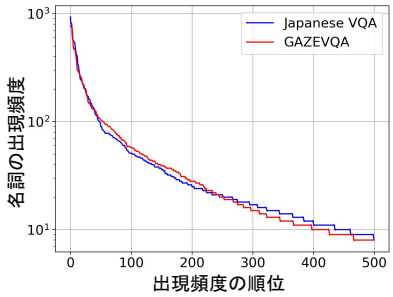

図 3　回答セットの単語数に関する分布

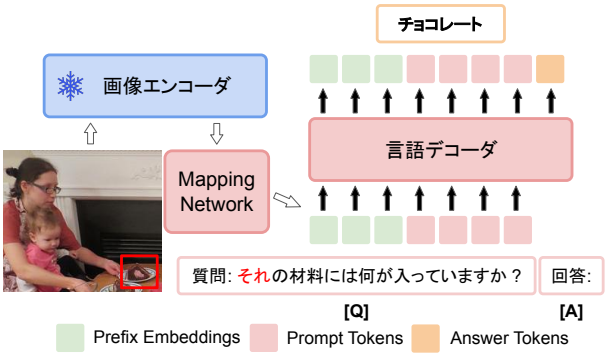

図 4　画像エンコーダと言語デコーダによるモデル．

4　評価実験

4.1　実験設定

事前学習済みの画像エンコーダ・言語デコーダモデル [16] をベースラインとして視線情報付き VQA の評価実験を行う．モデルの詳細は 4.2 節で述べる．VQA を扱う場合のモデルの性能評価，モデルの入力を変更した場合の ablation 評価を行う．

4.2　モデル

図 4 にモデルの概要を示す．モデルは CNN ベースの画像エンコーダ，エンコーダの埋め込み系列を低次元のベクトルへ圧縮する Mapping Network，自己回帰的に回答生成を行う言語デコーダで構成される．

画像エンコーダ　画像エンコーダにより，全体画像を 640 次元の画像特徴量に変換する．Shen ら [17] にならって画像エンコーダは CLIP [18] の RN×4 を使用し，Mokady ら [16] にならって画像エンコーダのパラメータを凍結する．

Mapping Network　Mapping Network を用いて，画像エンコーダから得た画像特徴量を言語デコーダへ入力可能なベクトルに圧縮する．ここで，Mapping Network は 1 層の MLP 層と 8 層の Transformer 層で構成されており，圧縮後のベクトルを Prefix Embeddings　(PEs) と呼ぶ．

言語デコーダ　初めに，VQA の質問を含むプロンプトから embedding 系列を取得する．この embedding 系列を Prompt Tokens (PTs) と呼ぶ．ここで，[Q] は質問を示し，[A] は回答生成の開始位置を示すラベルである．PEs と PTs を連結したベクトルを言語デコーダ[2] に渡す．連結したベクトルの次に出現するトークン系列を VQA の回答とみなして，学習と評価を進める．なお，今回は VQA の質問に視線の先の物体名を [objs] として prompting で与えることで視線の先の物体が認識できている状況のモデルを構築しようとした．

4.3　データセットと評価手法

データセット　日本語 VQA データセットに加え，日本語画像キャプションデータセット [19] を言語デコーダの事前学習に使用した．VQA に対するモデルの精度を評価するため，日本語 VQA データセットからテストセットを 4,000 件確保した．また，視線付き VQA データセットの訓練・開発・テストセットは 20,899 件，1,400 件，4,000 件とした．

評価手法　抽出型 QA タスクの評価指標である Exact Match スコア (EM) と F1 スコア (F1) [20] をテストデータ全体の評価に用いる．加えて，同義な回答を正しく評価するため，正解トークンと予測トークンが完全一致していないペアに対しては，BERT スコア[3](Bs) [21] によって回答フレーズの類似を考慮した評価を行う．

4.4　評価結果

VQA によるモデル評価　事前学習したモデルを日本語 VQA と視線付き VQA のテストセットで評

2)　事前学習済み GPT-2 モデルを言語デコーダの初期値とする．https://huggingface.co/rinna/japanese-gpt2-medium
3)　多言語 BERT の文ベクトルを評価に用いた．https://huggingface.co/bert-base-multilingual-cased

表 2　各データセットの評価結果．

テストセット	EM	F1	Bs
日本語 VQA	37.8	52.2	86.7
視線情報付き VQA	25.0	34.5	82.7

表 3　モデルの ablation 評価．PEs は Prefix Embedding，PTs は Prompt Tokens, [objs] は視線の先の物体名を表す．

プロンプトの構成	EM	F1	Bs
PEs + PTs ([objs]+[Q]+[A])	29.7	41.5	83.8
PEs + PTs ([Q]+[A])	30.3	41.4	83.8
PEs + PTs ([A])	9.03	13.9	74.1
PTs ([Q]+[A])	19.0	26.6	81.3

価した結果を表 2 に示す．表 2 より，日本語 VQA データセットで学習したモデルは，視線情報付き VQA データセットで fine-tuning せずとも，Bs で一定の性能が保証されることが判明した．

質問または画像の欠落が結果に与える影響　視線付き VQA の訓練セットで fine-tuning したモデルの ablation 評価結果を表 3 に示す．日本語 VQA データセットによる事前学習では言語デコーダの入力を，画像情報がエンコードされた PEs，質問 ([Q]) と回答の開始位置ラベル ([A]）がエンコードされた PTs で構成した．表 3 より，言語デコーダの入力から，[Q] または PEs を除いた場合，モデルの性能は著しく低下することが判明した．

視線の先の物体名が結果に与える影響　視線付き VQA データセットに含まれる質問は意図的に視線の先の物体名 ([objs]) を欠落させているため，PTs へ [objs] を補完して fine-tuning と評価を行った．なお，[objs] は MSCOCO の物体ラベルを使用した．図 5 はモデルの入出力例であり，「それ」という指示語が含まれた曖昧な質問の先頭に，[objs] である「ケーキ」の補完を行う．表 3 より，[objs] を考慮することによる精度向上は見られなかった．

5　分析

図 6 は，質問に対する CLIP の画像エンコーダの画像特徴を Grad-CAM [22] により可視化したヒートマップである．画像右上は視線付き VQA にみられる曖昧な質問であり，画像左下・右下は曖昧さの原因となる指示語を MSCOCO の物体ラベルで補完した質問である．図 6 で示した全ての質問において，「motorcycle」の画像特徴と比べて「bicycle」の画像特徴は強調されていない．モデルの性能は CLIP の画像エンコーダに依存しているため，[objs] の補完による言語情報の補正は視線付き VQA の精度に影響を与えなかったと考える．視線が指す領域や物体の矩形などを用いて画像情報を補正することや，画像エンコーダを言語情報と画像情報を同時に扱うクロスエンコーダへ変更することで，この問題は解消できる可能性がある．

図 5　モデルの入出力例．視線の先の物体名を赤字で，生成した回答を青字で示す．

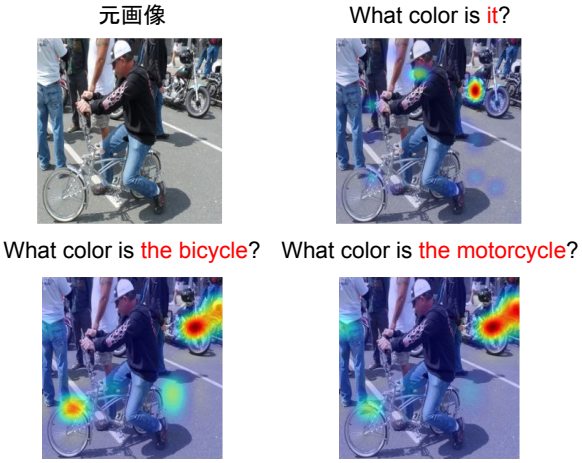

図 6　質問ごとの画像エンコーダ（CLIP の RN×4）の特徴量を Grad-CAM により可視化したヒートマップ．RN×4 の学習設定に合わせて質問を英訳した．

6　おわりに

本研究では，人間と円滑なインタラクションを行う対話ロボットの実現を志向し，視線情報で補完される曖昧な質問を含む VQA データセットをクラウドソーシングで構築した．視線の先の物体名をプロンプトへ補完した上で視線付き VQA の評価を行ったが，回答精度の向上は見られなかった．この問題へ対処すること，および注視対象推定から回答生成までを end-to-end で処理するシステムを開発することは今後の課題としたい．

謝辞

本研究は JSPS 科研費 JP22H04873 の助成を受けた．

参考文献

[1] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: Visual Question Answering. In **ICCV**, pp. 2425–2433, 2015.

[2] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. **International Journal of Computer Vision**, Vol. 123, No. 1, pp. 32–73, 2017.

[3] Nobuyuki Shimizu, Na Rong, and Takashi Miyazaki. Visual question answering dataset for bilingual image understanding: A study of cross-lingual transfer using attention maps. In **COLING**, pp. 1918–1928, 2018.

[4] Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. Visual coreference resolution in visual dialog using neural module networks. In **ECCV**, pp. 160–178, 2018.

[5] Nathan J. Emery. The eyes have it: the neuroethology, function and evolution of social gaze. **Neuroscience & Biobehavioral Reviews**, Vol. 24, No. 6, pp. 581–604, 2000.

[6] Danna Gurari, Qing Li, Chi Lin, Yinan Zhao, Anhong Guo, Abigale Stangl, and Jeffrey P. Bigham. VizWiz-Priv: A dataset for recognizing the presence and purpose of private visual information in images taken by blind people. In **CVPR**, pp. 939–948, 2019.

[7] Adria Recasens*, Aditya Khosla*, Carl Vondrick, and Antonio Torralba. Where are they looking? In **NIPS**, Vol. 1, pp. 199–207, 2015.

[8] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In **ECCV**, pp. 740–755, 2014.

[9] Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. AmbigQA: Answering ambiguous open-domain questions. In **EMNLP**, pp. 5783–5797, 2020.

[10] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. Question rewriting for conversational question answering. In **WSDM**, pp. 355–363, 2021.

[11] Vaibhav Kumar and Alan W Black. ClarQ: A large-scale and diverse dataset for clarification question generation. In **ACL**, pp. 7296–7301, 2020.

[12] Yuya Nakano, Seiya Kawano, Koichiro Yoshino, Katsuhito Sudoh, and Satoshi Nakamura. Pseudo ambiguous and clarifying questions based on sentence structures toward clarifying question answering system. In **DialDoc**, pp. 31–40, 2022.

[13] Danna Gurari and Kristen Grauman. Crowdverge: Predicting if people will agree on the answer to a visual question. In **CHI**, pp. 3511–3522, 2017.

[14] Nilavra Bhattacharya, Qing Li, and Danna Gurari. Why does a visual question have different answers? In **ICCV**, pp. 4271–4280, 2019.

[15] Yining Li, Chen Huang, Xiaoou Tang, and Chen Change Loy. Learning to disambiguate by asking discriminative questions. In **ICCV**, pp. 3419–3428, 2017.

[16] Ron Mokady, Amir Hertz, and Amit H Bermano. ClipCap: CLIP prefix for image captioning. arXiv:2111.09734, 2021.

[17] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How Much Can CLIP benefit vision-and-language tasks? In **ICLR**, 2022.

[18] Radford, Alec and Kim, Jong Wook and Hallacy, Chris and Ramesh, Aditya and Goh, Gabriel and Agarwal, Sandhini and Sastry, Girish and Askell, Amanda and Mishkin, Pamela and Clark, Jack and Krueger, Gretchen and Sutskever, Ilya. Learning transferable visual models from natural language supervision. In **ICML**, Vol. 139, pp. 8748–8763, 2021.

[19] Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In **ACL**, pp. 417–421, 2017.

[20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In **EMNLP**, pp. 2383–2392, 2016.

[21] Tianyi Zhang* and Varsha Kishore* and Felix Wu* and Kilian Q. Weinberger and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In **ICLR**, 2020.

[22] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In **ICCV**, pp. 618–626, 2017.