視覚言語基盤モデルに基づく画像質問応答AIによる映像解析技術 物体指向VQA

AI-Based Object-Centric VQA Capable of Answering Questions about Objects in Camera Images

三島　直 MISHIMA Nao

カメラ映像の解析は，危険検知や，顧客行動分析，業務効率改善などへの活用が期待できるが，人手による解析や用途に応じた専用AIの開発が必要であった。一方，近年，大規模な画像とテキストのデータを用いて事前学習された視覚言語基盤モデルの研究が進み，高度な映像理解が可能になった。

東芝は，視覚言語基盤モデルに基づく画像質問応答AI（VQA：Visual Question Answering）を活用し，映像中の物体ごとの質問に回答する独自の映像解析技術である物体指向VQAを開発した。物体指向VQAは，画像の内容を理解した上で様々な質問に回答できる。質問のテキストを変えれば複雑なタスクにも対応可能で，人手による開発を削減できる。また，対象とする物体の質問だけに回答するため，計算量が少ない。オフィスや，製造・物流現場，サービス業など，多くの場面で映像解析の利用が容易になる。

Camera image analysis is utilized for various applications including risk detection, customer behavior analysis, and operational efficiency improvement. Issues, however, include the need for manual analysis or the development of dedicated artificial intelligence (AI) models in accordance with each application. In line with the progress of research on a vision-language foundation model based on pre-training of large volumes of images and text data collected via networks, this model is expected to enable precise understanding of images.

Toshiba Corporation has developed a proprietary image analysis technology, called object-centric visual question answering (VQA), that provides clear answers to questions about objects in images by means of an AI developed using a visual-language foundation model. The object-centric VQA can be easily expanded to various business applications including offices, manufacturing and logistics sites, and service businesses thanks to the following features: (1) capability to appropriately answer various questions based on the understanding of content in each image, (2) applicability to complicated tasks by simply changing the questions without the need for additional manual development, and (3) minimal calculation due to only answering questions about the target object.

1. まえがき

社会の安全志向に伴い，様々な場所にカメラが設置されている。カメラ映像を分析することで，危険検知や，顧客の行動分析，業務効率改善などの様々な用途に活用できる。しかし，画像に内容の意味が記述されていないため，人手による解析が必要であった。これまで，省人化のために様々な画像認識AIが開発されてきたが，用途や分析方法ごとに専用のアプリケーションの開発が必要であった。その結果，コストや，導入までのリードタイムが掛かることから，あまり導入が進んでいなかった。

一方で近年，生成AIが大きな注目を集めている。従来のAIにはなかった高い汎用性により，様々な分野でのAI活用進展が期待されている。この生成AIの躍進を支えているのが，大規模なデータで事前学習された基盤モデル[(1)]である。特に，画像とテキストの両方の情報を用いて事前学習された視覚言語基盤モデルによって，映像を非常に高度なレベルで理解する能力をAIが獲得できることが分かってきた。画像に関する質問に回答するVQAは，基盤モデルを基に専用データセットで学習（ファインチューン）することで，人間と同等の回答能力を持つまでになった。しかし，画面内に複数の物体が映っていると，どの物体への回答かが曖昧になることがあった。

そこで東芝は，物体を検出して切り出し，物体ごとにVQAを適用する新しい映像解析技術である物体指向VQAを開発した。物体指向VQAは，画像に映った内容を理解した上でユーザーからの様々な質問に回答可能で，質問を変えるだけで様々なアプリケーションに適用できる。これにより，映像解析の省人化を強力に支援できる。

ここでは，開発した映像解析技術の概要と，評価結果について述べる。

2. 視覚言語基盤モデル

基盤モデル（ファウンデーションモデル）とは，大規模な

データで学習し，幅広い用途に適用できるAIのベースモデルである[1]。近年大きな話題となっている，対話的にテキストで課題を解決するチャットボットや，任意のテキストから高画質な画像を生成できるDiffusionモデルなどの生成AIも，基盤モデルを基にしている。特に視覚言語基盤モデルは，画像とテキストを同時に扱えるという特長があり，様々な用途で重要な役割を果たしている。視覚言語基盤モデルの用途には，画像にキャプションを自動で付ける画像キャプショニングや，テキストから画像を検索するテキスト画像検索，画像に関する質問に答える画像質問応答，フレーズに対応した画像内の要素を探し出すVisual Grounding，テキストから画像を生成するテキスト画像生成など，非常に高度な映像理解を必要とするものがある。

視覚言語基盤モデルの学習は，二つのステップから成り立つ。第1ステップでは事前学習として，比較的容易に収集できるインターネット上の画像とキャプションの大量のペアを使用し，人の教示を必要としない学習を行う。この事前学習では，大規模なモデルを豊富なデータを用いて学習することにより，非常に汎用性の高いモデルが獲得できる[2]。第2ステップでは，事前学習済みのモデルを，実際に使用する特定用途のデータでファインチューンし，活用する。

視覚言語基盤モデルの登場以降，画像質問応答の公開データセット[3]における正解率の向上が著しく，2022年には人間のパフォーマンスである81 %を超えるモデルが登場した。日々新しいモデルが提案され，各研究機関が，しのぎを削っている。

3. VQAを用いた映像解析技術

VQAは**図1**に示すとおり，画像に関する質問に答えるAIである。「はい」，「いいえ」で答えられる質問だけでなく，物の名称や色など，任意の質問に答えられる。従来の画像認識では，特定の物体を識別するためのID（識別情報）や，物体の位置を特定するための位置座標など，目的に合わせたフォーマットで情報を入力する必要があった。一方，VQAでは質問を自由なテキスト形式で記述できるため，原理的にはどのような質問でも入力できる。このため，様々な映像解析用途に適用できる可能性がある。

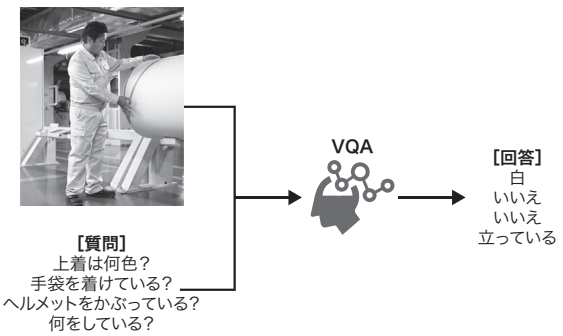

図1. VQAの概要

画像に関する任意の質問に回答できる。「はい」，「いいえ」で答える質問だけでなく，物の名称や色など，任意の質問に答えられる。

VQA outline

しかし，VQAを映像解析に適用する際には，以下のような二つの問題がある。一つ目は，画像内に二つ以上の物体が映っている場合，どの物体に対する回答かが明確でないという点である。二つ目は，質問に関係する物体が増えるほど処理量が増大するという点である。画像内に何が映っているかは事前に分からないため，画像内に映っていない物体に関する質問も処理する必要がある。これにより，処理すべき質問の数が増え，処理に要する時間やリソースが増加する。

これを解決するため，**図2**に示すとおり物体を検出して切り出し，物体ごとにVQAに入力する物体指向VQAを新たに開発した。人間と犬では問うべき質問が異なるように，物体指向VQAでは物体の種類ごとに適した質問を設定できる。画像内に物体が複数存在していても，それぞれの物体に対する回答を得られる。また，画像に映っている物体に関連する質問だけを処理するため，無駄な処理を避けられる。具体例として，人が映っている画像と犬が映っている画像，2枚の画像を考える。事前に設定する質問として，人に対して五つ，犬に対しても五つがあるとする。従来のVQAでは，人と犬に対する質問を分けられないため，人画像と犬画像に対してそれぞれ質問10個，計20回の処理が必要であった。それに対して物体指向VQAでは，切り出した人に対して質問5個，犬に対して質問5個と，計10回の処理で済むため，大幅に効率化できる。

最後に，物体指向VQAによって実現できるアプリケーション例について述べる。ここでは，保護具点検を実現する映像解析アプリケーションを考える。物体検出AIを使った従来の手法では，人物とヘルメットを検出した上で，人物矩形（くけい）内にヘルメット矩形が入っているかどうかでヘルメット着用を判定する。そのため**図3**(a)右側の人物のように，ヘルメットをかぶらずに手に持っている場合でも，安全だと誤判定することがあった。一方，物体指向VQAでは，画像に対して「ヘルメットをかぶっていますか？」というシンプルな質問をするだけで，人物とヘルメットとの関係も含めて点検内容を設定できる。このように物体指向VQAは，与える質問の内容を変えるだけで，以下のような多様なアプリケーションに対応できる。

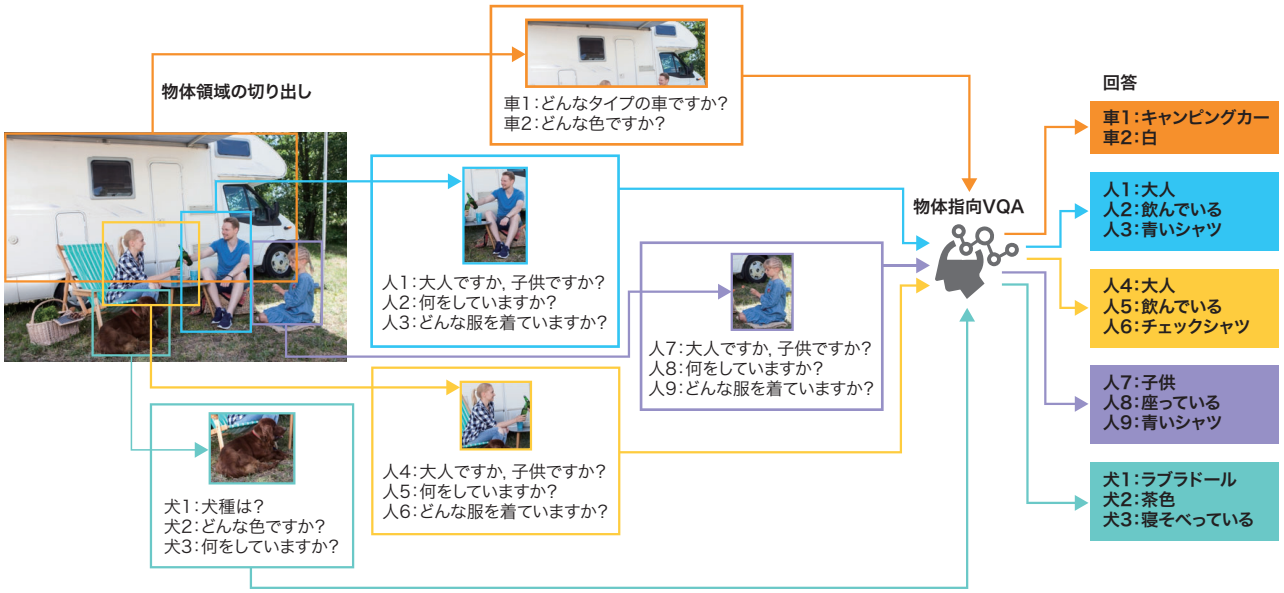

図2. 物体指向VQAによる処理の特長

映像から物体を検出して切り出し，物体ごとにVQAに入力することで，無駄な処理を避けながら，高精度に判定できる。

Features of processes for object-centric VQA

図3. 物体指向VQAによる判定結果の例

3種類の応用例で，テキストによる質問に対し，正しく判定して回答できた。

Results of identification of target objects obtained by object-centric VQA

(1) オフィス・工場

(a) 歩行中のスマートフォン利用の点検：「歩きスマホしていますか?」

(b) ポケットに手を入れたままでの歩行の点検：「ポケットハンドしていますか?」

(c) PC（パソコン）開き持ちの点検：「ノートPCを開いたまま持っていますか?」

(2) 工場

(a) 保護具点検：「ヘルメットをかぶっていますか?」

(b) 脚立安全点検：「脚立に乗っていますか?」,「ヘルメットをかぶっていますか?」

(c) 倒れている人検知：「人が倒れていますか?」

(d) ぬれた階段検知：「階段はぬれていますか?」

(3) 道路

(a) 路面状況判定：「路面状況はどうなっていますか?」

4. 開発した映像解析技術の評価

物体指向VQAの有効性を明らかにするため，様々な応用例に対して，実際に物体指向VQAで処理して評価した。ここで用いた物体指向VQAは，視覚言語基盤モデルの一つであるBEiT-3[4]をVQAv2データセット[3]でファインチューンした。図3(a)は，保護具点検の例である。「ヘルメットをか

ぶっていますか？」という質問を処理することで，ヘルメットの有無を正しく判定できた。特に右側の女性はヘルメットをかぶらずに手で持っているため，従来技術ではうまく判定できない例である。図3(b)は脚立安全点検の例である。一般的に脚立使用時はヘルメットを装着する必要がある。そこで，「脚立に乗っていますか？」，「ヘルメットをかぶっていますか？」という質問を設定した。左の人はヘルメットをかぶって脚立に乗っている（安全）が，中央の人はヘルメットをかぶらずに脚立に乗っている（不安全）。また右の人はヘルメットをかぶっていないが脚立には乗っていない（安全）。物体指向VQAは，全ての場合を正しく判定できた。図3(c)はテレビ映像やSNS（Social Networking Service）画像の分析に活用できる食事分析の例である。「何を食べたり飲んだりしていますか？」という質問を設定したところ，「りんご」と飲食物をある程度の分解能で分析できることが分かった。以上のように，物体指向VQAが，様々なアプリケーションに応用可能なことが示された。

判定精度の定量評価には，保護具点検の公開データセットConstructionSafetyV2[5]，及びPPEsV8[6]を用いた。保護具点検では，教示されている保護具の種類に合わせて「ヘルメットを付けていますか？」という質問を設定した。また比較対象として，物体検出AIのYOLOv5[7]について，保護具を検出するように学習したモデルを用意した。そして，誤検出の少なさを示す適合率と検出漏れの少なさを示す再現率を指標として，判定精度を比較した（**表1**）。物体指向VQAは，PPEsV8の適合率を除いてYOLOv5を上回る判定精度であることが確認できた。YOLOv5は，再現率が適合率に比べて低いことが分かった。これは，保護具が小さく，検出しにくいことが原因である。PPEsV8には，ヘルメット以外にゴーグル，手袋，靴といった小さい物体が多く含まれることが，再現率が低い原因として考えられる。それに対して，物体指向VQAは，人物矩形を元に保護具の装着有無を判定するため，YOLOv5に比べると再現率が大幅に高いことが分かった。これは，物体指向VQAが人物の領域と保護具の関連性をより強く捉えているためであると考えられる。また，YOLOv5はそれぞれのデータセットに対して，数千枚の学習用画像を用いて再学習する必要があるが，物体指向VQAは再学習が不要であるため，学習のコストや時間が掛からない。このように，物体指向VQAはYOLOv5と比較して，より高性能な映像解析をより低コストで実現可能であることが示された。

3種類の応用例の判定評価と，定量評価の結果，物体指向VQAの有効性を確認した。

表1．保護具点検の判定精度の定量評価結果

Comparison of results of protective equipment identification accuracy evaluations of YOLOv5 object detection AI and object-centric VQA using open datasets

項　目	再学習	ConstructionSafetyV2		PPEsV8	
		適合率 (%)	再現率 (%)	適合率 (%)	再現率 (%)
YOLOv5	必要	95.7	91.2	99.8	47.4
物体指向VQA	不要	97.8	96.0	98.8	89.9

＊適合率，再現率ともに，数値が大きいほど判定精度が高い

5. あとがき

当社は，視覚言語基盤モデルに基づくVQAを使い，質問を変えるだけで様々なアプリケーションに適用可能な物体指向VQAを開発した。物体指向VQAは，3種類の応用例と二つの公開データセットで有効性を確認済みであり，オフィスや，製造・物流，飲食やサービスなど，カメラを使用して映像解析する様々な場面に応用できる。

今後は，様々な用途での実証実験を進めるとともに，計算処理の高速化を図り，早期の実用化を目指して研究開発を進めていく。

文　献

(1) Bommasani, R. et al. On the Opportunities and Risks of Foundation Models. 2021, arXiv:2108.07258v3, 212p. <https://arxiv.org/pdf/2108.07258.pdf>, (accessed 2024-02-26).
(2) Radford, A. et al. "Learning transferable visual models from natural language supervision." Proceedings of the 38th International Conference on Machine Learning (PMLR). 2021, PMLR 139, p.8748–8763.
(3) Goyal, Y. et al. "Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering." Proceedings of CVPR 2017. Honolulu, HI, 2017-07, IEEE. 2017, arXiv:1612.00837v3, p.1-11. <https://arxiv.org/pdf/1612.00837.pdf>, (accessed 2024-02-26).
(4) Wang, W. et al. "Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks." Proceedings of 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR 2023). Vancouver, BC, 2023-06, IEEE. 2022, arXiv:2208.10442v2, p.1–18. <https://arxiv.org/pdf/2208.10442.pdf>, (accessed 2024-02-26).
(5) roboflow. "construction safety". <https://universe.roboflow.com/roboflow-100/construction-safety-gsnvb>, (accessed 2024-02-26).
(6) roboflow. "PPEs Computer Vision Project". <https://universe.roboflow.com/personal-protective-equipment/ppes-kaxsi>, (accessed 2024-02-26).
(7) ultratics. "yolov5". <https://github.com/ultralytics/yolov5>, (accessed 2024-02-26).

三島　　直　MISHIMA Nao
研究開発センター 知能化システム研究所
コラボレイティブAIラボラトリー
Collaborative AI Lab.