物体検出モデルの信頼値を利用したVQAモデルによるぬいぐるみ画像分類

尾関迅[1] 佐藤伶耶[2] 鈴木秀佳[2] 船田眞帆[2] 仲宗根太郎[1]
櫻井義尚[2]
[1]明治大学大学院　先端数理科学研究科
[2]明治大学　総合数理学部
{ozeki} cs233005@meiji.ac.jp
{sakurai} sakuraiy@meiji.ac.jp

概要

近年，大規模言語モデルの発展に伴い，マルチモーダルLLM(Large Language Model)の活用が進んでいる．特に画像分類においては,画像内の状況や意味を理解することで，従来の物体検出モデルでは検出困難だった画像全体を考慮した判別や新規の物体を特定する可能性が注目されている．しかしながら，マルチモーダルAIの性能は，画像とテキストのペアデータセットの収集難易度や，学習データの影響により画像分類性能に課題がある．

そこで，本研究ではこの課題を克服するために物体検出モデルとマルチモーダルLLMを組み合わせたハイブリッド画像分類手法**OConfVQAC(Object detection Confidence-guided VQA Classifier)**の提案を行う．具体的には，物体検出モデルによって算出された信頼度をVQAモデルにプロンプトとして与え物体検出モデルの検出漏れを補完し，VQAモデルを用いた画像分類の性能向上を目指す．

また，この提案手法の有効性を検証するために，テーマパークに関するSNSの投稿画像にぬいぐるみが映っているか否かの2値分類タスクに適用した．

本研究では,VQAモデルの画像分類性能の向上の可能性を検証した．

1 はじめに

近年，多くの人々がSNSを活用している．様々な種類のSNSが普及し，それぞれの用途によって個人が自分の状況や考えを投稿できる．これらのSNSプラットフォームには，個人の日常や考えが投稿されており，投稿内容で企業や政府機関は分析を行いマーケティングや政策の策定に活用している．このような企業や政府機関のSNSの情報を活用する手法を「ソーシャルリスニング」という．

ソーシャルリスニングによって，SNSの投稿情報を企業はマーケティング戦略や製品開発，ブランドイメージの向上に役立てている．

ソーシャルリスニングを活用している例としてテーマパークなどの商業施設における分析がある．テーマパークなどの商業施設において，ソーシャルリスニングは顧客の意見を抽出することや，熱心なファンである「ヘビーユーザー」を特定する目的で活用される．

尾関らの研究では，疑似ラベルを用いた手法によってSNSに投稿されたテキストからユーザーの意見を抽出することを行った[1].

また，小川，鈴木らの研究では投稿から熱心なファンを特定するヘビーユーザー特定のタスクを行っている[2].

このように顧客がSNSに投稿された文章からポジティブな意見や改善点を収集する「意見抽出」や，投稿の頻度や内容から特定の施設への関心度を評価する「ヘビーユーザー特定」といったタスクが行われているが，テーマパークにおけるソーシャルリスニングの研究ではテキストが中心であり画像を含めたものはあまりない．

画像を分析対象に含めることでテキストだけでは得ることのできなかったユーザーの情報を収集することができると考えられるが，分析の課題として大量の画像データから目的にあった画像を抽出するのが困難な点があげられる．特定の商品について，ユーザーの利用状況を分析するにはその商品が映る画像に絞ることが必要だが，SNSには様々な画像が投稿されており人手で分類するのは困難である．そこで本研究では機械学習を用いた自動の画像分類器の提案を行う．

画像内に映っている対象を判別する方法としてYOLO(You Only Look Once)やDETR(End-to-End

Object Detection with Transformers)のような物体検出モデルを活用することが考えられる．しかし，これらのモデルは特定のクラスに依存するため，未知の物体や文脈を理解する能力には限界がある．

一方で，VQA(Visual Question Answering)モデルは，画像内の状況を理解することが可能であると考えられている．例えば，「この商品はカチューシャですか？」という質問に対して，人物が頭に装着しているなどの「カチューシャらしさ」を推測して判断することが可能であると考えられる．

ほかにも，VQA モデルは物体検出モデルと比較して状況把握から施策策定まで一気通貫で可能である点で優れている．例えば「この画像に映っているぬいぐるみはどのように使用されていますか？」という質問によってユーザーが商品をどのように楽しんでいるかを把握することができる．さらに「このぬいぐるみの利用状況に基づき，どのようなマーケティング施策ができますか？」という質問によって，新商品の開発やプロモーションの方向を策定できる．

しかし，VQA モデルには画像の認識能力に課題がある．例えば「この画像にぬいぐるみはありますか？」という質問に対して実際には存在しないぬいぐるみに対してぬいぐるみが映っていると回答する課題があり，他にも，実物のぬいぐるみと 2D やイラストのキャラクターなど類似した物体の画像を見分けることができない可能性もある．

このように VQA モデルには高い状況把握能力とマーケティング施策提案までの応用性があると考えられるが，画像の分類器として利用するには課題がある．

そこで，本研究では物体検出モデルと VQA モデルを組み合わせたハイブリッドな手法の OConfVQAC によって VQA モデルの分類性能を向上させることを目指す．

また，本研究は VQA モデルの特性とその課題点を検証するために具体的な対象として「ぬいぐるみ」を選定した．ぬいぐるみは様々な種類のキャラクターやデフォルメが施されたグッズが発売され，デザインや形状が多様であり，SNS の投稿に用いられることの多い商品の一つである．このように様々な形状があり，多様な角度から撮影される商品は物体検出モデルにおいては認識するのが困難であると考えられるため，VQA モデルの優位性を検証するために適していると考えた．

さらに，ぬいぐるみはテーマパークにおいてユーザーの利用シーンやどのように楽しんでいるかを把握するのに重要な役割を果たす．例えば，ヘビーユーザー特定のタスクではぬいぐるみ商品を多く所有していることがヘビーユーザーであることを示す手掛かりにもなる．

このように本研究では「ぬいぐるみ」がマーケティング的に重要であり，VQA モデルの性能を評価する上で適切であると考えられるので対象とした．

2 関連研究

VQA(Visual Question Answering)モデルは，画像と言語を統合した研究分野であり，画像とテキスト(質問)を入力し，それに対応する回答を生成することができる．近年では，LLM の発展に伴い VQA モデルの性能も向上している．

例えば，2023 年に Liu らによって発表された LLaVA(Large Language and Vision Assistant) [3]は視覚情報と大規模言語モデルの情報を組み合わせることで画像内の情報を理解し，複雑な状況であっても回答を生成することが可能だと考えられる．しかし，これらのモデルは物体検出用に学習されているわけではないので物体が存在しないにも関わらず存在しているように回答を生成してしまうことや細かい物体の分類性能はプロンプトに依存する．

分類性能を向上させるアプローチとして，データセットを拡張することが考えられるが画像とテキストのペアのデータは収集することが困難であり多くの労力を必要とするため現実的ではない．

そこで，従来の物体検出モデルと組み合わせることで物体特定の正確性を高めるアプローチが考えられる．Jiao らの研究ではマルチモーダル大規模言語モデルの視覚的な情報の理解を向上させるために，トレーニングなしとトレーニングありの手法を比較し，その有効性を評価している．その中で，物体検出の情報を入力することでマルチモーダル大規模言語モデルの視覚的理解を向上することを示した．具体的には物体の名前や座標などをテキスト形式でモデルに入力し，それで複数のベンチマークテストにおいて性能を向上させたことが確認されている [4].

しかし，この手法では物体検出の情報を断定的な形式でモデルに入力しているため，マルチモーダル大規模言語モデルの柔軟な推論能力が十分に活用さ

れていない可能性がある．そこで，私たちは物体検出モデルによって算出された信頼度情報をマルチモーダル大規模言語モデルに与える**OConfVQAC(Object detection Confidence-guided VQA Classifier)**によってモデルの推論能力を最大化し，その性能を高めることを目指した．

3 提案手法

本研究では，物体検出モデルと VQA モデルを組み合わせたハイブリッド手法によって物体検出モデルの検出漏れを補完し, VQA モデルの分類性能を向上させる手法の **OConfVQAC(Object detection Confidence-guided VQA Classifier)**を提案する．

提案手法は以下の 3 つのステップに分けられる．

1. 物体検出モデルによる信頼度計算

まず，画像の前処理として画像を物体検出モデルに入力できる形式に変換する．その後，物体検出モデルで推論をし，Non-Maximum Suppression を用いて後処理を行った．そして，最も高い値のぬいぐるみラベルの信頼度を取得する．

2. プロンプト作成

取得した信頼度を用いて VQA モデルに入力するプロンプトを作成する．プロンプトには様々なバリエーションを出すことができ，例えば目的のタスクを端的に述べるシンプルなプロンプトや，間違えやすい部分の注意を加えたプロンプトを作成することが可能である．

本研究では prompt1 として目的のタスクをシンプルに述べたものと，prompt2 として本タスクにおける誤検出が起きやすい，イラストや実物のキャラクターの判別の注意文を加えたプロンプトを用意した．図 1 に prompt1 と prompt2 を示す．

3. VQA モデルによる分類

最後に VQA モデルによって推論を行い，出力された回答に’yes’, ‘no’ のどちらが含まれるかを正規表現を用いて判定する．

4 実験

4.1 実験概要

物体検出モデルと VQA モデルのハイブリッドモデルである提案手法の有効性を明らかにするため，物体検出モデル, VQA モデル単体での場合と提案手法との分類性能比較を行った．また，プロンプトに

Prompt1:
"USER: <image>¥nThe confidence score of teddy bear detection by the object detection model is {confidence:.3f}. Based on this, is there a plush toy in this photo?¥nASSISTANT:"

Prompt2:
"USER: <image>¥nThe confidence score of teddy bear detection by the object detection model is {confidence:.3f}. Based on this, is there a plush toy in this photo? Please note, the plush toy must be a physical object, not an illustration or drawing.¥nASSISTANT:"

図 1 OConfVQAC のプロンプト例

与える情報の違いによる差を明らかにするため，2 タイプのプロンプトを用いた比較を行った．このように本実験では 6 つの手法による比較を行う．以下に手法を記す．

- **手法 1: YOLO**
- **手法 2: DETR**
- **手法 3: LLaVA(prompt1)**
- **手法 4: LLaVA(prompt2)**
- **手法 5: OConfVQAC (prompt1)**
- **手法 6: OConfVQAC (prompt2)**

物体検出モデルには’yolov7’ [5], ‘facebook/detr-resnet-50’ [6]の 2 種類の物体検出モデルと VQA モデルには’llava-hf/llava-1.5-7b-hf’ [3]を用いる．また，OConfVQAC では YOLO を用いて信頼度を取得し，LLaVA に入力した．

4.2 実験設定

実験設定として, YOLO と DETR の Non-Maximum Suppression の信頼度と IoU 閾値はそれぞれ,0.25, 0.45 とする．

また，使用したデータセットの収集方法，アノテーション手順は以下のようになる．

まず，「ディズニー」をキーワードとした画像付きの投稿を収集する．収集された投稿は 2024 年 2 月~2024 年 9 月のものである．その中からキーワードベースのスパムフィルターをかける．そして，月ごとに約 1250 件ずつサンプリングを行い，約

10,000 件の画像にアノテーションを実施した．アノテーションは「画像内にぬいぐるみが存在するか否か」の 2 値の分類である．ぬいぐるみの定義は「布製やフリース製の柔らかいぬいぐるみ」「抱きしめられるサイズのもの（手のひらサイズから等身大のものまで）」である．また，布製のぬいぐるみが付いているキーホルダーもぬいぐるみとして扱う．一方，プラスチック製や木製などの布製以外の人形や着ぐるみや被り物，耳カチューシャなどは対象外とした．

アノテーションは管理者やクラウドワーカーの 3 名の多数決をとり，ワーカーの判断で判定がスキップされ判定がわれてしまった場合や明らかに判定が誤っていると判断されるものが発見された場合，アノテーション業者などの管理者によって修正が加えられる．Xから収集した画像データのアノテーションの結果は以下の通りである．

表 1　画像数とクラスの関係

	ぬいぐるみあり	ぬいぐるみなし
画像枚数	948	9051

上記のモデル，実験設定，データを用いてぬいぐるみの存在する画像の分類を行った．

5 結果・考察

実験結果を表 2 に示す．

また，本実験により導き出された結論は以下の通りである．

- **物体検出モデルでは YOLO が適しているが検出漏れに課題がある**

物体検出モデルで YOLO と DETR を比較すると YOLO の F1-score が 0.69 となり DETR と比べて精度が高い結果となった．また，物体検出モデルは VQA モデルと比較すると recall が低い値になった．このことから検出漏れに課題があると言える．

- **VQA モデルは取りこぼしが少なく誤検出が多いがプロンプトで調整可能**

VQA モデルでは LLaVA(prompt1)で recall が 0.99 と高い値になったが F1-score は 0.35 と低い値になった．このことから，ぬいぐるみ画像の取りこぼしが少ないが多くの画像をぬいぐるみありと判定してしまっているため，精度が低くなっていると考えられる．誤検出の例として，Xに投稿されたキャラクターのイラストなどをぬいぐるみとして判定していることが考えられる．そこで LLaVA(prompt2)において，ぬいぐるみはイラストではなく実態のある物体だという注意を加えることで F1-score が 0.43 と prompt1 と比較して向上させることができた．

- **OConfVQAC によって分類精度向上が可能**

すべての手法の中で提案手法の OConfVQAC の F1-score が高くなった．このことから信頼度を用いることで VQA モデルの分類性能を向上させることができたと示された．また，prompt1 よりも prompt2 の精度が高くなったことから本手法ではプロンプトの構成次第でさらに精度が高まると期待できる．

6 おわりに

本研究ではテーマパークに関する投稿の画像にぬいぐるみが含まれているか否かを判別するタスクにおいて，物体検出モデルと VQA モデルのハイブリッド分類器の手法として OConfVQAC の提案を行い，その有効性を示した．

今後の課題として，本研究では物体検出モデルのパラメーターや信頼度，IoU の閾値を任意に設定したが，本タスクに最適な閾値の調整が必要であると考えた．また，VQA モデルは物体検出モデルと比較して推論時間が長くなる傾向があるため，低コストの分類方法についても検討をすべきである．

表 2　実験結果

	物体検出モデル		VQA モデル		提案手法	
	YOLO	DETR	LLaVA(prompt1)	LLaVA(prompt2)	OConfVQAC (prompt1)	OConfVQAC (prompt2)
Accuracy	0.93	0.92	0.65	0.75	0.93	0.94
Precision	0.62	0.54	0.21	0.28	0.61	0.67
Recall	0.78	0.73	0.99	0.99	0.82	0.77
F1-score	0.69	0.62	0.35	0.43	0.70	0.72

謝辞

本研究の遂行にあたり，多大なるご支援をいただいた明治大学櫻井研究室の皆様に深く感謝申し上げます．また、本研究のデータ収集およびアノテーションにご協力いただいたクラウドワーカーの皆様にも心より御礼申し上げます．

本研究は JSPS 科研費 20K11960 の助成を受けたものです．

引用文献

[1] Y. S. Y. T. Jin Ozeki, “Opinion Classifier Transfer Learning from Review Data,” IEEE Symposium Series on Computational Intelligence, 2023.

[2] A. Ogawa, M. Suzuki and Y. Sakurai, "Identification of Disney Heavy Consumers on," IEEE Region 10 Conference 2024, 2024.

[3] C. L. Y. L. Y. J. L. Haotian Liu, “Improved Baselines with Visual Instruction Tuning,” Computer Vision and Pattern Recognition, 2023.

[4] D. C. Y. H. Y. L. Y. S. Qirui Jiao, “From Training-Free to Adaptive: Empirical Insights into MLLMs' Understanding of Detection Information,” arXiv, 2024.

[5] A. B. WongKinYiu, “YOLOv7: Trainable Bag-Of-Freebies Sets New State-Of-The-Art for Real-Time Object Detectors,” https://github.com/WongKinYiu/yolov7, 2022.

[6] F. M. G. S. N. U. A. K. S. Z. Nicolas Carion, "End-to-End Object Detection with Transformers," https://huggingface.co/facebook/detr-resnet-50, 2020.