M1 サーベイレポート

学籍番号: 82518829 氏名: 戸倉 健登

所属: 杉浦孔明研究室

1. 関連研究

Vision-language モデル（VLM）は、ロボティクスのために広く研究されており [?,?]、[?,?] において包括的にレビューされている。多様なマルチモーダル検索手法は [?,?] にまとめられている。さらに、シーンテキストの統合は、視覚表現の曖昧性解消において著しく効果的であることが、[?,?] に記録されている。

ロボティクスのための基盤モデル最近の研究は、大規模言語モデル（LLM）や VLM [?,?,?] などの大規模基盤モデルを活用することで、操作タスクとナビゲーションタスクにわたる顕著な汎化性能を達成できることを示している。大規模な実ロボット軌跡で訓練された Vision-language-action モデル [?,?,?,?,?] は、多様な操作タスクにわたるスケーラブルな方策学習を可能にする。例えば、Gemini-Robotics [?] は、複雑な家庭タスクのために高度な視覚理解とロボット能力を統合している。しかし、これらのアプローチはシーンテキストを明示的に組み込んでおらず、テキスト情報が重要な意味的グラウンディングを提供するシーンにおいて、マルチモーダル理解が制限されている。

ロボティクスのための RAGいくつかの既存手法 [?,?,?,?,?] は、強化されたグラウンディングと計画のために、具現化エージェントに RAG を適用している。他のアプローチ [?,?] は、RAG を使用してナビゲーション指示を生成し、それにより vision-language ナビゲーションモデルを改善している。RAG ベースのメモリは、長期的推論のために提案されており [?]、他の手法は模倣方策学習のためにデモンストレーションを検索および融合するために RAG を使用している [?]。これらの進歩にもかかわらず、先行研究は検索信号としてシーンテキストを活用しておらず、これは日常環境において視覚的に類似したオブジェクトを曖昧性解消するために不可欠である。

マルチモーダル検索多数の研究 [?,?] が、自然言語指示に基づくモバイル操作タスクにおける家庭用サービスロボットの性能を評価している。いくつかの既存手法 [?,?,?,?] は、環境内のターゲットオブジェクトを識別するためにマルチモーダル検索を実行し、自然言語指示に基づいて関連画像のランク付けリストを出力する。また、マルチモーダル検索を実行してターゲットオブジェクトを見つけ、テキストベースの回答やウェイポイントを出力する手法もある [?]。この分野での急速な進歩にもかかわらず、本質的に曖昧性解消情報を含むシーンテキスト（ラベル、標識など）の活用には重要なギャップが残っている。提案された STARE は、シーンテキストを明示的に組み込む新しいマルチモーダルフレームワークを提案することで、この制限に対処し、実世界環境におけるオブジェクト検索精度の向上を可能にする。

シーンテキストシーンテキストを含む画像に関する研究は、テキスト認識 [?,?]、テキスト検出 [?,?,?]、テキスト検索 [?,?,?]、視覚的質問応答 [?,?,?]、およびテキスト合成 [?,?,?] を含む様々なタスクを網羅している。特に、シーンテキスト検索手法は、OCR で抽出されたテキストと位置情報を視覚特徴と統合して、クエリテキストとのアライメントを改善し、より高い検索精度を達成することが多い。Mishra ら [?] は、シーンテキスト検索タスクを導入し、文字を検出および分類してクエリテキストを含む画像の確率を計算する 2 段階パイプラインを提案した。Gomez ら [?] は、テキスト提案を予測し、クエリテキストとの類似性に基づいて画像をランク付けするエンドツーエンドネットワークを提案した。ViSTA [?] は、BERT [?] と線形投影をそれぞれ使用して OCR で抽出されたテキストと位置を埋め込み、マルチモーダル統合のためにトランスフォーマー内の融合トークンを介してテキストと位置を視覚特徴と融合する。これらのアプローチとは対照的に、STARE は、対応するオブジェクトとその属性との関連性を捉えたナラティブ表現にテキストと位置を統合する。さらに、STARE は、シーンテキストと参照されるオブジェクト間の複雑な関係を効率的にモデル化する補助類似度関数を導入する。

ベンチマークマルチモーダル検索とシーンテキストを含むタスクの分野で、いくつかの標準ベンチマークが提案されている。前者のうち、COCO [?] と Flickr30K [?] は、もともと画像キャプショニングベンチマークであり、マルチモーダル検索タスクに広く使用されている。REVERIE [?] と GOAT-Bench [?] は、オブジェクト目標ナビゲーションタスク用に設計されており、LTRRIE [?] は画像のランク付けによるマルチモーダル検索に焦点を当てている。シーンテキストのベンチマークは、様々な vision-language タスクにわたっている。例えば、ST-VQA [?] は、シーンテキストベースの VQA に一般的に使用され、TextCaps [?] と COCO-Text-Captioned [?] は、シーンテキストを考慮した画像キャプショニングに適している。Visual Genome [?] には、シーンテキストと詳細なアノテーションを持つ画像が含まれており、vision-language グラウンディングと推論タスクをサポートしている。最後に、RefText [?] は、参照表現理解のベンチマークであり、画像-オブジェクトペアと参照表現を提供し、その一部はコンテキストの手がかりとしてシーンテキストを活用している。マルチモーダル検索のための多くの既存ベンチマークの中で、ほとんどはシーンテキストを欠いているため、我々のタスクには適していない。シーンテキストを含むベンチマークは、通常、マルチモーダル検索用に設計されていないか、ロボットタスクに適していないか、または過度に単純な指示を含んでいる。これらの制限により、シーンテキストと視覚コンテンツにわたるきめ細かい推論を必要とする、ロボットのための実世界の指示ベースの検索タスクを評価するには不十分である。

既存手法との違いSTARE は、2 つの重要な側面で既存のマルチモーダル検索手法と異なる。第一に、ほとんどの先行手法が屋内環境での検索に焦点を当てているのに対し、STARE は広範囲の屋内および屋外環境にわたってタスクを実行し、実世界のシナリオにより適用可能である。第二に、STARE は画像からシーンテキストを抽出および利用するが、既存の手法はシー

ンテキストを明示的に組み込んでいない。次に、STARE は、シーンテキストを処理する方法において、シーンテキストを扱うまたはシーンテキストを考慮したマルチモーダル検索を実行する既存の手法と異なる。シーンテキストをそのまま使用する先行手法とは対照的に、STARE は、対応するオブジェクトとその属性との関連性を捉えたナラティブ表現にシーンテキストを統合する。また、本手法は、シーンテキストと参照されるオブジェクト間の複雑な関係を効率的にモデル化する補助類似度関数を導入する。