公開

研 究 終 了 報 告 書

「与えられた指示文章に従い言語で判断を説明する AI」

研究期間:2020 年 12 月～2024 年 3 月
研 究 者: 栗田修平

1. 研究のねらい

人間からの言語指示を理解し，現実世界の複雑な課題を計画的にこなすモデルを構築する際は，いくつかの障害があることが知られている．まず，現行の自然言語処理モデルがテキストの中に含まれる「人物」や「動作」等を現実の対象と上手く対応付けて捉えられていないという問題が古くから知られている（**自然言語のグラウンディング問題**）．自然言語処理や画像処理のタスクでは，テキストのみもしくは画像のみなど限られた形態のデータのみで学習が行われることも多く，テキストと画像が紐付いた学習データセットを利用しても，実際には言語情報のみなど限られた情報を利用した場合と精度は大差ない場合も多かった（**マルチモーダルデータの問題**）．加えて，特に言語データを解釈する際は，言語データの曖昧性に対して，その文脈や日常的な知識を利用した解決がしばしば求められる．例えば「書類を持っていって，印鑑をもらってきてほしい」という指示文章は「書類に印鑑を押してもらう」という意味になりやすいが，「印鑑を受け取ってくる」という意味にも解釈できる．このような知識が欠如したモデルは，ユーザーの意図とは乖離した判断をしてしまう可能性がある（**常識的な知識の欠落に従う信頼性の問題**）．また，特に深層学習モデルについては，内部的にどのような要素に着目して判断を行ったのか容易に解釈しがたいことが指摘されている（**ブラックボックスおよび説明性の問題**）．

研究提案者は，現在の言語処理・理解モデルには物理的な実世界やそこで作業をする際に必要な常識など，外部世界の情報を扱わせることができないために，「信頼出来ない AI」になってしまっているのではないか，と考えた．本研究では，言語による直感的な指示を通して仮想世界や実世界を認識し，言語指示から複雑かつ体系的な操作を可能にする．また，このために必要な仮想世界や実世界で撮られた 3D 点群データ，動画などのモダリティとテキスト情報を結びつける手法を開発し，また，そのための基盤となるデータセットを作成する．

2. 研究成果

(1)概要

本研究では以下の二つの観点から研究を推進した．

(1) 実世界の情報と自然言語との対応付け技術については，テキストから参照された物体を画像や動画から探索する AI の基礎技術として「参照表現理解タスク」に着目し，主観視点動画からテキストで参照された物体を探すタスクや，全周カメラで撮影されたシーンをテキストにて説明するタスクなど，実世界とテキストをつなぐ各種のタスクを開発した．具体的には，RefEgo データセットおよび ArKitSceneRefer データセットなどを作成し，またベースラインモデルを作成した．(2)自然言語による人間の指示を理解し，実世界・仮想世界から情報を取得し，行動や質問応答を行う技術については，視覚と言

語を融合した実世界探索において，画像キャプションモデルを利用して言語指示と動作・視覚の結合や，3 次元の写実的なシーンからの質問応答，詳細な物体の探索などについて，タスクやモデルの研究開発と提案を行い，データセット構築を行なった．具体的には，視覚と言語のナビゲーションに加えて，ScanQA データセットおよびモデルの作成，およびその実世界への応用を進めている．

これらの成果は，視覚，言語，機械学習などの幅広いトップ国際会議に採択されている．今後もこのように実世界と言語を繋ぐ研究や，さらに近年急速に研究が進みつつある大規模言語モデルにこれらの成果を応用することが期待される．

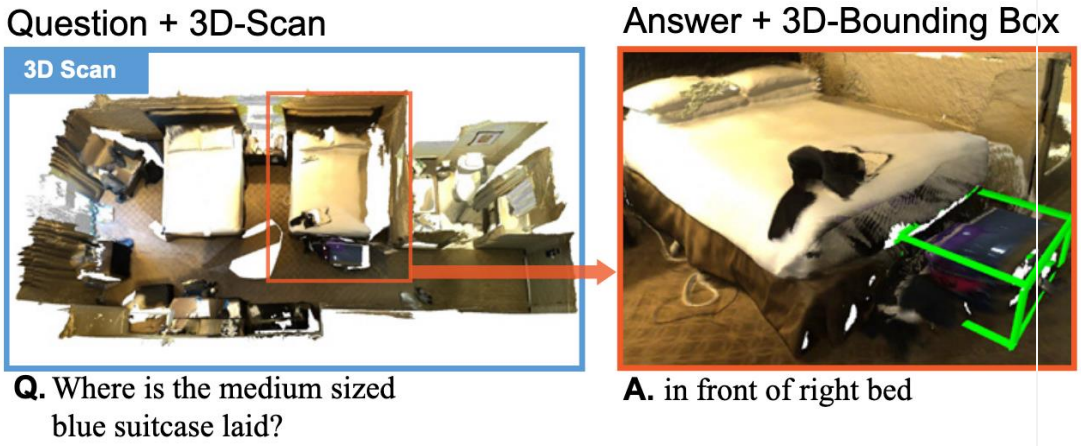
(2)詳細

(1) 実世界の情報と自然言語との対応付け技術について，動画や360画像のような多様な実世界の環境情報とテキスト情報を統合するために，参照表現理解や環境情報のテキスト説明をテーマとして，研究を進めた．実世界を移動する人やロボットなどを念頭に，テキストから実世界の物体などを結びつける，参照表現理解タスクに注力した．まず，一人称視点動画から特定の物体をテキストに基づいて探索するための基盤整備や，ScanQA の実世界での応用について研究を進めた．既存の研究では，主として画像からテキストで指示された物体を探索していたが，実世界を移動するロボットなどに応用するには対象物体を含む適切な画像を入力として選択する必要があった．実世界環境に近い写実的な仮想環境上で，与えられた質問に応じて環境内部をナビゲーションし，質問の正解物体を探して質問応答を行う研究を進めた．特に，大規模な一人称視点動画セットである Ego4D からテキスト参照表現に基づいて特定の物体を探索する参照表現理解手法と物体追跡手法を組み合わせた手法およびデータセットを提案した．提案した手法を使うことで，追跡物体を動画のフレーム外に出てしまったなどの理由で見失っても，テキスト情報を利用することで同じ物体を特定し再び追跡することができるようになる．この一人称視点動画におけるテキストからの物体追跡に関する研究(RefEgo) は，画像系のトップ国際会議 ICCV2023 にて筆頭で論文が採録され発表を行った．

図：RefEgo より，“the large white bowl with broccoli inside that is used to load the pan of broccol”に対応する物体の追跡タスク例（緑色）．途中のフレームでは物体が画面

外へと逸失しているが，モデルは継続して物体を追跡する必要がある．

(2) 自然言語による人間の指示を理解し，実世界・仮想世界から情報を取得し，行動や質問応答を行う技術について，視覚と言語によるナビゲーションなどを中心に研究を進めた（Generative Language Grounded Policy）．さらに，実世界上での具体的なタスクに落とし込むための研究を進めた．主要な成果として，3 次元の写実的な環境から必要な情報を取得し，ユーザーの質問に応答するモデルの研究のため，3 次元上での質問応答大規模なデータセットである ScanQA データセット，および点群を入力に取る質問応答モデルを提案した．これは 3D の室内環境を理解し，“Where is the blue suitcase laid?” のような質問に対して室内環境から物体を探索するための大規模なデータセットを作成したである．国際電気通信基礎技術研究所 (ATR)等の研究者と共同で行われた．この研究は画像系のトップ国際会議である CVPR2022 で口頭発表を行った (ScanQA)．また，自律移動ロボットへの応用を進めている．



図：ScanQA より，三次元世界でテキストから物体を特定するタスクの例．

これらの成果に加えて，指導学生を中心とした研究グループにより，三次元空間での小物物体のグラウンディングを行うための ArKitSceneRefer データセットや，360 度画像上での注目する物体に関するキャプション生成に関する研究にて作成した QuIC-360◦ データセットなどの研究成果が得られた．これらの研究は，いずれも特定の物体を動画や 3 次元空間上などでテキストから探しだす，テキストで指示された内容を詳述するなど，今後の研究展開に必要な基盤成果であり，自然言語処理系のトップ国際会議 EMNLP にて 2 本の論文が findings 採録された．さらに，当初の予想を上回る成果として，屋内の 3D シーンに限らず，街レベルの 3D スキャンに対し，建物などをテキストから特定する研究である CifyRefer データセットを作成した．この成果は機械学習系のトップ会議である NeurIPS dataset and benchmark トラックに共著で採録され発表を行った．加えて，複数の手法で作られた屋内シーンでの物体のグラウンディングを行う Cross3DVG

を作成し，3 次元視覚情報処理に関する会議である 3DV に共著で採録され発表を行った．

3. 今後の展開

ScanQA や RefEgo に代表されるテキストと実世界をつなげる基盤が整備されつつあることや，近年の大規模言語モデルの隆盛により，言語処理を実世界にて応用する機会は着実に広がっているものと考える．今後は，さきがけ研究で作成された手法やデータセット基盤を，大規模言語モデルやマルチモーダル基盤モデルを用いた学習・推論と組み合わせることで，実世界の状況を理解して複雑な推論を行うモデルや，そのロボットへの応用などに展開できるものと考えている．具体的には，ScanQA を実ロボットを用いた質問応答実験に応用することや，同様に RefEgo にて得られた手法・データセットを，物体の状態や動作などを理解する基盤モデルへと拡張していく方針が考えられる．より長期的には，先述のように大規模言語モデルによる推論や計画立案・外部ツールの実行などと，さきがけ研究による言語と実世界の対応付けを組み合わせることで，実世界で自律的に実験操作などを遂行するエージェントやロボットなど，幅広い学術および社会応用を模索していく方針でいる．

4. 自己評価

研究成果にて記したように本研究では以下の二つの観点から研究を推進した．

(1) 実世界の情報と自然言語との対応付け技術については，テキストから参照された物体を画像や動画から探索する AI の基礎技術として「参照表現理解タスク」に着目し，主観視点動画からテキストで参照された物体を探すタスクや，全周カメラで撮影されたシーンをテキストにて説明するタスクなど，実世界とテキストをつなぐ各種のタスクを開発した．(2)自然言語による人間の指示を理解し，実世界・仮想世界から情報を取得し，行動や質問応答を行う技術については，視覚と言語を融合した実世界探索において，画像キャプションモデルを利用して言語指示と動作・視覚の結合や，3 次元の写実的なシーンからの質問応答，詳細な物体の探索などについて，タスクやモデルの研究開発と提案を行い，データセット構築を行なった．これらテキストと実世界をつなげる基盤を作成するという研究目的はおおむね達成されたものと考える．

これらの成果は，多数の優れた論文として，視覚・言語処理から深層学習にいたる AI 分野のトップ国際会議等に採択され，国際的にも注目されている．さらに，タスクとモデルの提案，データセットの公開などを通じて，今後の先端的な言語・視覚 AI の研究に不可欠な研究情報基盤の構築に貢献している点に国際的な貢献がある．これらは，従来の言語理解手法では扱うことができない物理世界の情報をテキストと対応付けて取扱うことを可能にする新しい「信頼される AI の基盤技術」の開発を促進・支援するものであり，学術・産業・社会的な波及効果が高い．また，今後の波及に記したように，自然言語処理，画像処理，ロボティクスなどの幅広い分野の AI 技術基盤となることが期待されると考える．

5. 主な研究成果リスト

(1)代表的な論文(原著論文)発表

研究期間累積件数: 12 件

1. Shuhei Kurita and Kyunghyun Cho, Generative Language-Grounded Policy in Vision-and-Language Navigation with Bayes' Rule, Ninth International Conference on Learning Representations (ICLR2021), Online (May 2021).
3 次元仮想世界である Matterport 3D simulator 上で, 指示にしたがったナビゲーション動作を行うために, マルチモーダル言語モデルである言語キャプション生成モデルを応用したモデルを提案した.
2. Daichi Azuma(*), Taiki Miyanishi(*), Shuhei Kurita(*) and Motoaki Kawanabe, ScanQA: 3D Question Answering for Spatial Scene Understanding, The 2022 Conference on Computer Vision and Pattern Recognition (CVPR2022). (*): Equally contributed.
実世界から作成された写実的な 3D 点群データセット ScanNet 上での大規模な質問応答データセットの作成と, ベースラインモデルの提案を行った.
3. Shuhei Kurita, Naoki Katsura and Eri Onami, RefEgo: Referring Expression Comprehension Dataset from First-Person Perception of Ego4D, The 2023 International Conference on Computer Vision (ICCV2023).
参照表現理解タスクを拡張し, 一人称視点動画上でテキストから物体を特定し, 追跡し続けるためのデータセットを作成し, またベースラインモデルを提案した.

(2) 特許出願

該当なし.

(3) その他の成果(主要な学会発表, 受賞, 著作物, プレスリリース等)

- スポンサー賞(Money Forward 賞). 言語処理学会第 30 回年次大会.
- スポンサー賞(PKSHA Technology 賞). 言語処理学会第 30 回年次大会.
- 若手奨励賞. 言語処理学会第 30 回年次大会(筆頭著者学生に対する受賞).
- 若手奨励賞. 言語処理学会第 29 回年次大会(筆頭著者学生に対する受賞).
- 委員特別賞. 言語処理学会第 27 回年次大会.