［研究助成（C）］

実世界における直観に反した運動の予測と言語生成

2247010

研究代表者　お茶の水女子大学　大学院 人間文化創成科学研究科　博士後期課程　黒　田　彗　莉

［研 究 の 目 的］

過去 10 年の自然言語処理研究により，ヒトの活動や生活をサポートするような対話型ロボット（ヒトの発話を受け取り，内容に即して回答するロボット）の開発が進んだ。また，大規模言語モデル（LLM）の開発に伴い，今後の対話型ロボットはより自然に返答できるようになると期待される。しかし現在の LLM は，答えのない問いへの回答や，ヒトの会話のように曖昧な表現を含んだ返答が苦手である。また，ヒト同士の会話と，ヒトとロボットのコミュニケーションは決定的に異なる。会話においてヒトはただ相手の話を聞くだけでなく，相手の振る舞いから次の発言を予測しながら発話する。一方でロボットは，あくまでヒトからの一方的な指示に答えるに過ぎず，相手（ヒト）の言動を予測しているわけではない。ヒトからロボットへの一方的な命令という関係を超えて，今後ヒトと機械が共存していくには，ロボットが自律的に状況を判断できるようになる必要がある。さらに，判断した状況から様々な将来を予測することで，ヒトの言動を先回りしてサポートできるロボットの内部モデル構築を目指す。

たとえば，緑の円柱が青い立方体に近づいている状況があったとする。これらの物体は，この後どうなるだろうか。直観的には「緑の円柱が青い立方体にぶつかる」と想像できる。一方で，直観に反して「もし緑の円柱が止まった」とすると，「青い立方体と緑の円柱はぶつからない」と予測できる。このように直観に反した状況でも，ヒトは柔軟に推論を変更することで，その場面を画像として想像し，言語として発話できる。しかし予測研究の多くは，環境にある物体を視覚的（画像）もしくは物理的（物体の速さなど）にとらえ，直観に即した予測をするにとどまっている。また，直観に反した予測を扱う研究は，ほとんどの場合，物理シミュレーターで実世界を表し，物理情報を直接修正することで仮想環境を生成している。そのため，ヒトのように実世界からの情報をもとに柔軟に推論を修正し，起こりえない予測を生成できるモデルはまだない。

これまでの研究（例：Engelcke et al., 2020; Burgess et al., 2019）は，入力の系列情報に対する決まった（直観に即した）予測のみを扱ってきた。それに対し本研究では，直観に即した従来の予測だけでなく，直観に反する条件が与えられたときに，これまでの推論を柔軟に変更することで，状況に即したもっともらしい環境を予測できる推論モデルを構築する。与えられた条件に適合するように推論を修正して，新し

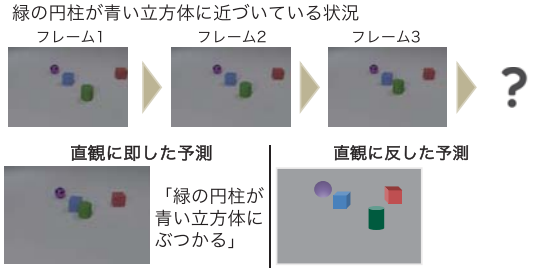

図 1　本研究の概要図

い予測を生成するようなモデルはまだ提案されていない。

また，予測推論モデルから画像を生成する手法として，シーングラフから画像を生成する研究（Johnson et al., 2018）がある。しかし，環境にある物体の速度や移動方向といった物理的な特徴や物体同士の位置関係を表したグラフ構造から画像を生成した先行研究はいまだ存在しない。

このように自身の研究に適した形で従来の研究を改善する方法を独自に考え，新たな予測モデルの構築や画像生成手法を提案することで課題を解決する。

［研究の内容，成果］

1. 提案手法

1.1 物理特性を含んだ訓練データセット作成

本研究ではデータセットとして CLEVRER（Yi et al., 2019）をメインで用いた。CLEVRER は「画像内で 3 種類の物体が動き，物理的事象は衝突のみ」というシンプルなデータセットである。また，提案モデルの入力情報は画像だけでなく，画像から得られた物理情報（位置や速度など）の 2 種を想定している。そのため，元となるデータセットから物理情報のデータセットも新たに作成する。作成の手順を図 2 に示す。

（1）画像内の環境の獲得

CLEVRER に写っている物体の種類や位置情報を考慮したグラフベースの訓練データを作成するために，YOLACT を用いた。YOLACT を用いて物体検知をし，画像内の 2 次元位置情報と物体の形状・色を取得する。検知可能な種類数は物体の色 8 種類・形状 3 種類・素材 2 種類の組み合わせ 48 種類で行った。

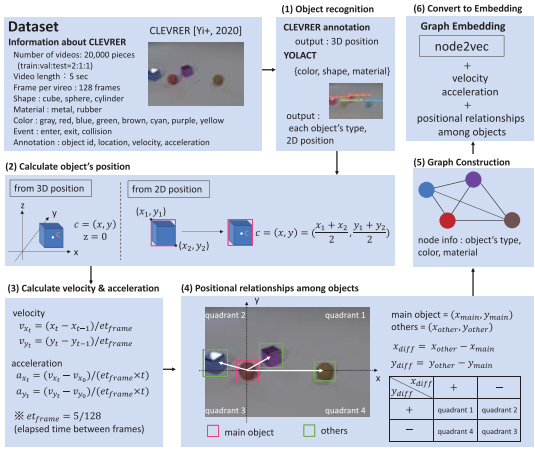

図 2　訓練データセット作成の概要図

（2）物体の座標・環境を表現したグラフの構築

獲得したバウンディングボックスの座標から物体の 2 次元位置情報を決定し，グラフを作成した。本研究ではグラフを構築した後，それらのグラフを埋め込みベクトルに変換した。グラフの埋め込みベクトル作成においては word2vec の Skip-gram から着想を得た node2vec と doc2vec の PV-DBOW から着想を得た graph2vec の 2 種類の手法を用いた。node2vec を用いた埋め込みでは 1 フレーム内に写っている物体一つずつをノードとみなし，構築したグラフのノードをもとに埋め込みベクトルを作成した。一方で graph2vec では，観測環境（CLEVRER 1 フレーム）を表したグラフ全体から埋め込みベクトルを作成した。

（3）環境内の物体の速度と加速度の算出

環境に写っている各物体についての物理特性を捉えるために，物体の速度および加速度の算出を行った。時刻t_kにおける物体の位置をx_{tk}, y_{tk}とし，次の時刻t_{k+1}における物体の位置をx_{tk+1}, y_{tk+1}とする。また時刻t_kから次の時刻t_{k+1}の経過時間は，5 秒の動画を 128 フレームに分割しているので，5/128 秒となる。加速度算出における初速v_0は，環境における各物体は停止しているところから始まるためv_0=0と設定した。

（4）物体間の位置関係の表現

環境内の物体に関して，物体同士の位置関係も重要な情報である。ヒトを例に上げると，自分を中心にして他の物体がどの位置関係にあるかを瞬時に捉えている。ここでは全ての物体同士の位置関係について，各物体が中心となったときの他物体の位置方向を算出した。

1.2　予測モデル構築

本研究では，PredNet に対して画像特徴量の変化から環境の移り変わりのタイミングを捉える研究である Variational Temporal Abstraction（VTA）の機能を組み合わせ，環境中の視覚情報と物理特性の両方の変化から環境の移り変わりを予測する予測モデルを構築した。VTA は系列情報から階層的な抽象度を見つける状態空間モデルであり，入力情報の変化点を抽出することができる。

モデルの構造は，物体の動きなどの物理特性を予測する機構と，環境の画像情報を予測する機構の二つの階層構造を並列にし，各モデルのそれぞれの機構に VTA の機構である変化点判別フラグ m を取り入れて構築した。

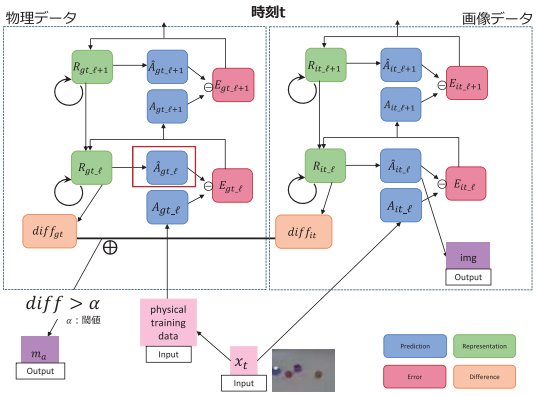

図 3　提案する予測モデルの概要図

1.3　言語データのレンプレート作成

変化点予測モデルを用いて予測した埋め込みベクトルから言語を生成するために，言語情報を新たに学習する必要がある。そのため言語データセットとして，物理特性を表すグラフ表現の埋め込みベクトルとそのときの状態を説明した文章のペアデータを作成した。グラフ表現の埋め込みベクトルは，CLEVRER のアノテーションデータから作成した。またペアとなる文章は，3（衝突前・衝突・衝突後）×3（文章の種類）の 9 種類のテンプレートに当てはまるように作成した。テンプレートの詳細を以下に示す。衝突した 2 つの物体は A・B と表し，A と B はそれぞれ「“灰，赤，青，緑，茶，水，紫，黄”色の“球，円柱，立方体”」の情報をもつ。また衝突のデータだけでなく，物体が衝突の前に近づくと衝突後に離れるときをデータセットとして作成した。近づくときは衝突の 5 フレーム前，離れるときは衝突の 5 フレーム後を対象とした。作成したペアデータの例は図 4 に示す。

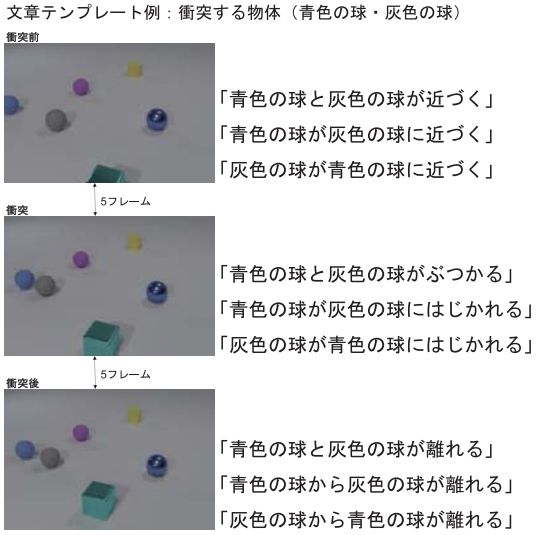

図 4　言語データのテンプレート作成の例

1.4　言語生成モデル

言語生成モデルは，Transformer の Decoder のみを用いた。図 5 に Decoder モデルを示す。従来の Transformer は Encoder-Decoder モデルで構築されているが，本研究では提案モデルの変化点予測モデルにおけるグラフの埋め込みベクトルの予測結果を Encoder における出力結果とみなし，この結果を Encoder から Decoder への入力とした。Decoder の学習は図 4 で作成したペアデータを用いた。学習データは 219303 個，テストデータは 10965 個とした。

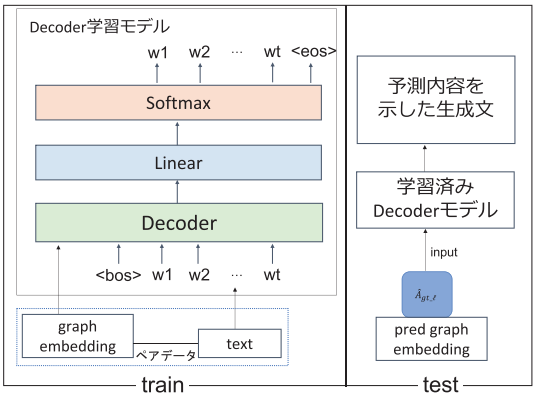

図５　言語モデル

2. 実験

実験では，予測して物理情報の変化点を言語情報として生成し，予測内容について解釈可能にすることを目的とする。

2.1 結果と考察

例１では，緑色の球と赤色の円柱がぶつかる状況に対して，赤色の円柱が反対方向に動いたときの動きを検証した。言語生成モデルを用いたときは「緑色の円柱が赤色の円柱からはなれる」という文章になり，予測画像は図６に示すものになった。これは想定する正解文に近い文章が生成できたといえる。一方で，予測画像については物体の物理特性を表すグラフ構造から画像を作成したが，ノイズがかかり，また動きが大きく変わった画像を生成することはむずかしいことがわかった。

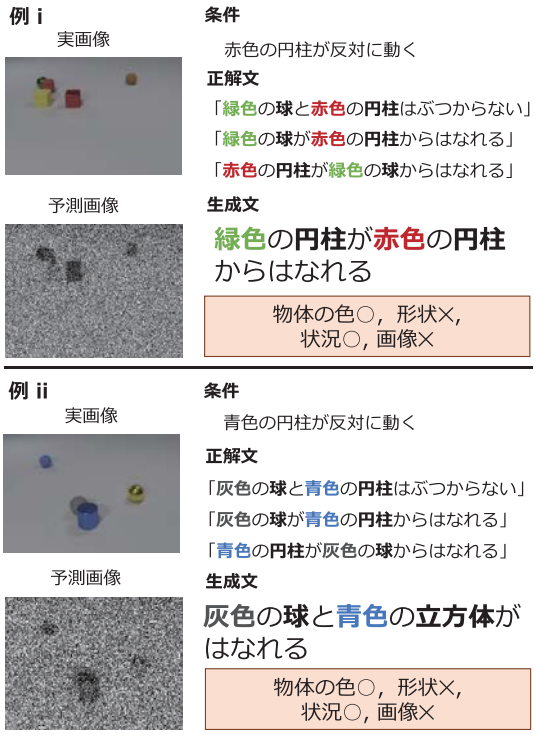

図６　実験結果

例２では，灰色の球と青色の円柱がぶつかる状況に対して，青色の円柱が反対に動いたときの動きを検証した。言語生成モデルを用いたときは「灰色の球と青色の立方体がはなれる」という文章になり，予測画像は図６に示すものになった。例１と同様に，生成文は正解文に近い文章が生成できたが，予測画像については想定していた画像とは異なるものが生成された。

2.2 文章生成の評価

次に各モデルで生成した文章を自動評価指標を用いて精度を算出した。生成した文を自動評価指標で評価した結果を表１に示す。自動評価指標には，言語生成タスクで用いられているBLEU，キャプション生成タスクで用いられるMETEOR，CIDEr を使用した。生成文１文に対し正解データは３文用意したため，各スコアの平均を生成文のスコアとした。予測推論内容から生成した生成文と正解文のスコアは全体を通して約 75 程度であり，観測環境の状況を高い精度で生成することができたといえる。

表１　文章生成の評価結果

モデル	BLEU@2	BLEU@3	BLEU@4	METEOR	CIDEr
PredNet ベース	79.3	75.1	72.3	70.4	73.1

3. まとめと今後の課題

本研究では観測した環境を視覚的・物理的に予測できるモデルを構築し，状況を説明できるモデルを提案した。また環境の物理的な特徴を予測した結果を言語として生成し，モデルによる予測内容を解釈可能にした。現実には起こり得ない状況についての文章は高い精度で生成できたが，そのときの場面を鮮明に描写することには課題が残った。

また実験に使用したデータセットは，我々人

間が目にする実環境よりも簡単なデータになっているため，予測モデルとしても言語生成としてもまだ改善の余地はある。今後の課題としてより実世界に近いデータセットを利用した実験を考えている。

［成果の発表，論文など］

1．Eri Kuroda & Ichiro Kobayashi: Predictive Inference Models for Real-world Physical Environments, Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 29, No. 3, pp. 456-468 (2025).

2．Eri Kuroda, & Ichiro Kobayashi. Verbal Representation of Object Collision Prediction Based on Physical CommonSense Knowledge. 2025 17th International Conference on Machine Learning and Computing (ICMLC2025), Long Paper, Oral, Guangzhou, China, Feb 14th-17th, 2025.

3．Eri Kuroda, Yuki Taya & Ichiro Kobayashi. Verbal Description Focusing on Physical Properties of Real-World Environments. 2024 Joint 13th International Conference on Soft Computing and Intelligent Systems and 25th International Symposium on Advanced Intelligent Systems (SCIS&ISIS2024), Long Paper, Oral, Hyogo, Japan, Nov 9th-12th, 2024.