スケッチ指示を用いた深層強化学習によるロボット動作の獲得に関する研究

TR22011 本多 航也　　指導教授：藤吉 弘亘

1. はじめに

少子高齢化の拡大に伴い，介護業界では労働者不足が深刻な課題となっている．そのため，高齢者や障がい者の生活支援を目的としたロボットの実用化が期待されている．生活支援ロボットの操作を容易にするには，直感的な指示を可能とすることが重要である．これまでに，自然言語指示とスケッチ指示を用いた研究が取り組まれている．Zitkovich らは，自然言語指示を用いてロボットの行動を指示する RT-2 を提案している [1]．ロボットに移動を指示する場合，自然言語指示はどのように移動させるか指示が曖昧になりやすい．一方で，スケッチによる指示は，曲がる位置や止まる位置などを直感的かつ適切に指示できる．

そこで本研究では，スケッチ指示からのロボット動作の獲得のために，Action Q-Transformer (AQT)[2] に，スケッチ指示を考慮するための機構である Instruction Phase を導入した手法を提案する．評価実験により，提案手法はスケッチ指示を考慮するロボット動作の獲得に有効であることを示す．

2. Action Q-Transformer

AQT[2] は Transformer 構造を導入した深層強化学習手法である．AQT のモデル構造を図 1 に示す．AQT は，環境から取得した画像を CNN を用いて特徴量に変換して，位置情報を付与して Transformer Encoder に入力する．そして，Transformer Encoder からの出力を Key，Value として Transformer Decoder に入力する．Query には Action query を入力し，入力画像と Action query の関係性を算出する．Action query は行動ごとに作成した One-hot Vector を全結合層を用いてトークンにしたものである．Action query を行動数分用意することで行動ごとの Q 値を算出できる．また，AQT は入力画像と Action query との Attention を計算しているため，Attention を可視化することで行動選択に寄与する画像内の領域を解析可能となる．

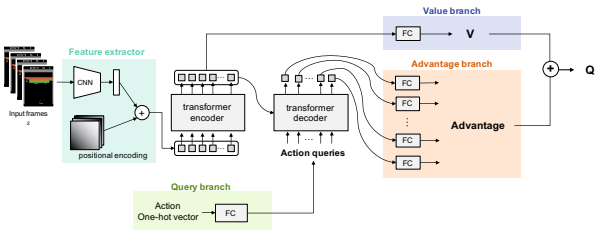


図 1: AQT のモデル構造（文献 [2] より引用）

3. 提案手法

本研究では，生活支援ロボットのナビゲーションタスクにおいてスケッチ指示を用いたロボット動作の獲得を目的とし，新たなモデル構造を提案する．

3.1 提案手法の構造

本手法は，AQT をベースにスケッチ指示を考慮できる機構を導入する．提案手法のモデル構造を図 2 に示す．提案手法は，Sketch Sequence から Instruction Feature を算出する Instruction Phase と，ロボット動作に対する Q 値を算出する Control Phase から構成されている．深層強化学習アルゴリズムとして Deep Q-Network (DQN) [3] を用いる．

Instruction Phase

Instruction Phase は事前に作成したスケッチ指示と Transformer Encoder を用いて，時系列情報を考慮しながら Instruction Feature を算出する段階である．スケッチ指示から時系列を考慮した Sketch Sequence を作成する．そして，全結合層を用いて特徴量に変換して，位置情報を付与して Transformer Encoder へ入力する．Transformer Encoder の Self-Attention を用いて Sketch Sequence 間の関係性を考慮した Instruction Feature を獲得する．

Control Phase

Control Phase はロボット動作を獲得するための段階である．提案手法は，Transformer Encoder-Decoder 構造を用いており，入力には各ステップの 128 × 128 ピクセルの画像，ロボットの位置と向き，Instruction Phase で算出した各ステップに該当する Instruction Feature を用いる．入力画像をパッチ分割したあと，位置情報を足し合わせて Transformer Encoder に入力する．Transformer Encoder の出力を Key，Value として Transformer Decoder へ入力する．Query は行動ごとに作成した One-hot Vector とロボットの位置と向き，Instruction Feature を連結して全結合層を用いて特徴量に変換した Action query を用いる．Transformer Decoder からの出力を Q 値を算出する Q-Head に入力して，最終的な Q 値を求める．

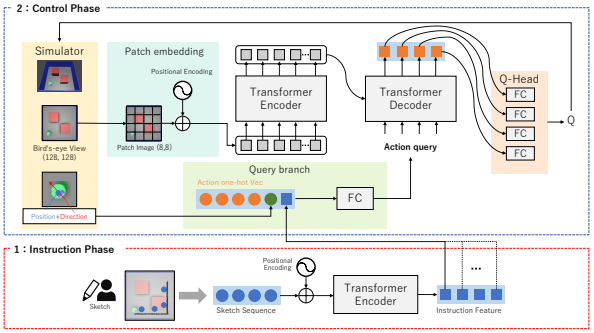


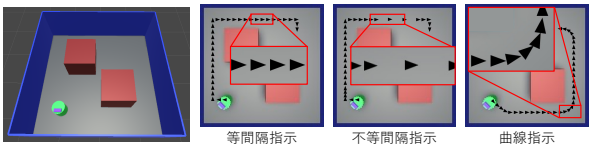
図 2: 提案手法のモデル構造

4. 評価実験

提案手法の有効性を確認するため，生活支援ロボットのナビゲーションタスクを用いて Displacement Error による定量的評価，ロボット動作と Attention の可視化による定性的評価を行う．

4.1 実験概要

本研究では，物理シミュレータである Unity3D を用いて作成した生活支援ロボットのナビゲーションタスクを対象とする．学習で使用する実験環境とスケッチ指示の例を図 3 に示す．移動可能なロボットをスケッチ指示の通りに移動させることを目標とする．エージェントの行動はその場で停止する Noop，前方へ進む Forward，右回転の Right，左回転の Left の 4 種類である．報酬は，各ステップのスケッチ指示の位置を waypoint として，waypoint とロボットの位置のユークリッド距離の差を負の報酬とする．



(a) 実験環境　　(b) スケッチ指示の例

図 3: 実験環境とスケッチ指示の例

本研究の学習で用いるスケッチ指示は，等間隔のスケッチ指示を 4 種類，不等間隔のスケッチ指示を 2 種類，曲線のスケッチ指示を 4 種類，計 10 種類のスケッチ指示である．等間隔指示，不等間隔指示，曲線指示それぞれのスケッチ指示の例を図 3(b)に示す．

提案手法の比較手法として，スケッチ情報を用いた CNN ベースのモデルを用いる．スケッチ情報を用いた CNN ベースのモデルは，提案手法と同じく Instruction Phase と Control Phase で構成されている．Instruction Phase は，提案手法と同様の構造を用いる．Control Phase は画像から特徴量を算出する Image-Feature extractor，ロボットの

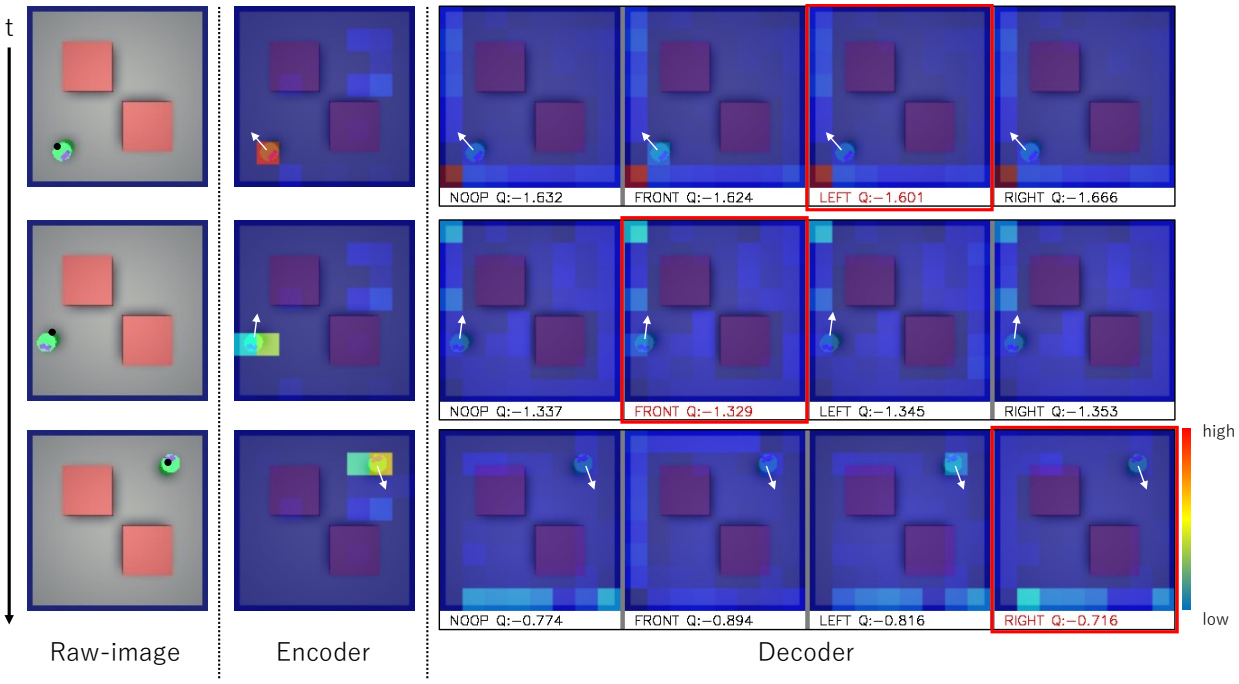


図 4: ロボット動作と Attention の可視化

位置と向きから特徴量を算出する Position-Head，Image-Feature extractor の出力と Position-Head の出力，各ステップに該当する Instruction Feature から Q 値を算出する Q-Head で構成されている．

学習条件は学習率を 0.0001，割引率を 0.99，replaybuffer のサイズを 250,000 とし，学習の終了条件は 1,000,000 エピソードに達した場合，エピソードの終了条件は 100 ステップが経過した場合のみとする．

4.2 定量的評価

本研究では評価指標に，Average Displacement Error (ADE) 及び， Final Displacement Error (FDE) の 2 つの評価指標を用いて評価を行う．CNN ベースモデルと提案手法を用いた評価結果を表 1 に示す．表 1 はスケッチ指示ごとの ADE および FDE の平均と全スケッチ指示の平均である．表 1 より，すべてのスケッチ指示において提案手法が CNN ベースモデルより ADE および FDE が小さい．このことから，Instruction Feature を用いた Action query を Decoder の Query に入力することで，直感的に作成したスケッチ指示にも対応することができると考えられる．また，提案手法のスケッチ指示ごとの ADE および FDE のばらつきが小さいことが確認できる．このことから，時系列を考慮した Sketch Sequence を Instruction Phase で解析し，Decoder に入力することで汎化性能が向上して，どのスケッチ指示であっても安定した精度を獲得することが可能である．

表 1: Displacement Error を用いた評価結果

比較手法	CNN モデル		提案手法	
評価指標	ADE	FDE	ADE	FDE
等間隔指示	0.41	0.25	**0.27**	**0.19**
不等間隔指示	0.40	0.24	**0.29**	**0.22**
曲線指示	0.92	3.10	**0.25**	**0.25**
平均	0.61	1.39	**0.26**	**0.22**

4.3 定性的評価

提案手法で獲得した図 3(b)の等間隔指示におけるロボット動作と Attention を可視化して，定性的評価を行う．提案手法で獲得したロボット動作と Attention の可視化例を図 4 に示す．Raw-image の黒い点はスケッチ指示から生成した waypoint，Decoder の各図下の文字は行動と Q 値を表しており，赤文字の行動が選択された行動，画像中の白矢印がロボットの向きである．

ロボット動作の可視化

図 4 の Raw-image から，スケッチ指示に追従するようにロボット動作を獲得していることが確認できる．このことから，AQT に Instruction Phase を導入することでスケッチ指示を考慮したロボット動作の獲得が可能であると言える．

Attention の可視化

図 4 の Encoder の Attention から，ロボットの領域に注目していることがわかる．このことから，Encoder は入力画像から適切なロボット動作の獲得に重要となる領域に注目していると考えられる．図 4 の Decoder の Attention から，選択した行動が強い Attention を獲得していることが確認できる．また，選択された Action query が「FRONT」の場合（2 行目の赤枠）はロボットの前方が,「RIGHT」の場合（3 行目の赤枠）はロボットの右側の領域に注目していることがわかる．このことから，Decoder の Query に Action query を用いることで行動選択に寄与する領域を注目できると言える．

5. おわりに

本研究では，スケッチの情報を考慮したロボット動作の獲得のために，Transformer を導入した深層強化学習モデルにスケッチ指示を考慮するための機構である Instruction Phase を導入した手法を提案した．定量的評価より，人が直感的に作成したスケッチ指示を用いる場合でも安定した精度を獲得できることを確認した．また，定性的評価からスケッチ指示を考慮するロボット動作の獲得に有効であることを示した．今後は，さらなる精度向上と大規模な環境を用いての実験に取り組む予定である．

参考文献

[1] B. Zitkovich, *et al.*, “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control”, PMLR, 2023.

[2] H. Itaya, *et al.*, “Action Q-Transformer: Visual Explanation in Deep Reinforcement Learning with Encoder-Decoder Model using Action Query” , arXiv:2306.13879, 2023.

[3] V. Mnih, *et al.*“Human-level control through deep reinforcement learning”, Nature, 2015.

研究業績

[1] 本多航也 等, “Mask-attention 機構を導入した PPO による物体把持動作の視覚的説明”, 日本ロボット学会学術講演会, 2022.