混合エキスパートによる視野予測を用いた

EmbodiedQA タスク学習と性能評価

Performance Evaluation of EmbodiedQA Task Using
Visual Field Prediction with Mixture of Experts

5119E014-4　石 晶　　指導教員　尾形 哲也　教授
SEKI Syou　　Prof. OGATA Tetsuya

概要： 本研究では，EmbodiedQAタスクの性能を向上させるため，ナビゲーションモジュールに異なるステップの画像を予測するエキスパートをゲーティングネットワークで統合し，混合エキスパートを構築した．近年，人間の指示でロボットがターゲットまでナビゲーションするため，深層学習モデルを用いて，言語指示とRGB画像から行動を生成する研究が行われている．Facebookの研究者達は，シミュレーション環境において，ロボットが言語質問を受け，一人称視点で未知環境を探索し，ターゲットまでナビゲーションするタスクとしてEmbodiedQAタスクを提案した．本研究では，EmbodiedQAタスクにおいて，ナビゲーションの性能を向上させるため，異なるステップの画像をサブタスクとして予測する，複数のエキスパートをゲーティングネットワークで統合した混合エキスパートを構築した．言語指示（例えば：ベッドはどの部屋にありますか？）を受けて行動を生成し，評価実験を行った結果，本提案手法によるナビゲーションの性能が最も高いことを確認した．

キーワード：EmbodiedQA タスク，混合エキスパート，ディープニューラルネットワーク
Keywords: EmbodiedQA Task, Mixture of Experts, Deep Neural Network,

１．初めに

現在，実世界に設置したロボットにおいて，深層学習モデルを用いて，言語指示と RGB 画像の入力からターゲットまでナビゲーションするため，直接行動を生成する研究が行われている[1][2][3]．しかし，未知環境で言語指示に従って正しくターゲットまでナビゲーションするのはまだ難しい課題である．例えば，短い言語指示で環境を探索することが困難であり，また，未知環境での成功率が低いという問題が存在している．Facebook の研究者達は，EmbodiedQA タスクを提案した．House3D シミュレーション環境でロボットは質問を受けると，一人称視点で探索し，答えを見つけ出すため，ターゲットまでナビゲーションするタスクである．深層学習モデルを用いて学習し，良い結果を得た．しかし，CG で構成したシミュレーション環境のビジョン情報は実世界に比べて少ない．実世界と近いリアルな写真で構成した複雑なシミュレーション環境で実装すると，ビジョン情報はもっと全面的に，多くの場所をカバーする必要があると思うので，単一のネットワークは最適ではないと考える．故に，本研究では，EmbodiedQA タスクをリアルな写真で構成した Matterport3D シミュレーターで実装し，性能を向上させるため，ナビゲーションモジュールの改善と評価を行う．

２．提案手法

本実験の提案手法では，ナビゲーションモジュールに異なるステップの画像を予測することをサブタスクとして学習する複数のエキスパートをゲーティングネットワークで統合し，混合エキスパートを構築する．混合エキスパートの詳細では，まず，時刻 t の画像特徴量と言語指示を LSTM ネットワークに入力し，t+n ステップの画像特徴量と t+1 の行動を予測することを１つのエキスパートとして設定する．n は１，３，５，７，９と設定し，5 つのエキスパートを用意する．そして，同じく LSTM ネットワークであるゲーティングネットワークはエキスパートと同じ入力を使い，重みを５つ出力する．５つの重みはそれぞれ先ほど用意した５つのエキスパートにかけて，足し合わせることで，どの予測ステップのエキスパートを主に使うかを決定する．モデルをまとめてみると，画像特徴量と言語指示を LSTM であるエキスパートとゲーティングネットワークに入力し，各エキスパートの出力とゲーティングネットワークの重みをかけた出力と目標との損失を用いて学習する．学習したモデルは，時刻 t の画像と言語指

示を入力として，５つのエキスパートの出力を統合した時刻 t＋１の行動を出力する．

３，実験

本実験では，Matterport3D シミュレーターを用いて，EmbodiedQA タスクを実装する[4]．Matterport3D シミュレーターでは，実世界の建物から撮影した写真を用いて再構成したシミュレーターである．合計 90 個の異なる環境がある．図 1 に示すように RGB-D 画像だけではなく，セマンティックセグメンテーション画像も提供できる．

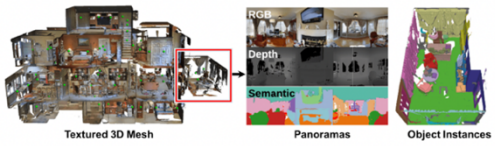

図 1　Matterport3D シミュレーター概要図

本実験で使用するモデルでは図２に示す．混合エキスパートの有効性を検証するため，サブタスクを使わない LSTM のみのモデル，一つのエキスパート（t+1 ステップの画像を予測する）及び混合エキスパートを同じ学習データで学習し，成功率と接近距離を比較する．

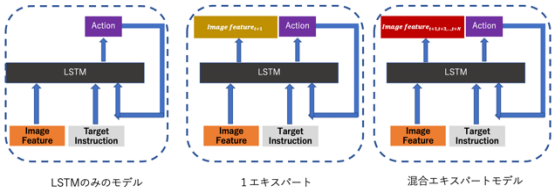

図２　本実験で使用したモデル概要図

本実験は EQA_MP3D データセットを使用した[3]．Matterport3D シミュレーター環境で，3 種類の質問と 83 の環境シーンを組み合わせて，合計 11796 エピソードを生成した．表１に示すように，訓練データ，バリエーションデータとテストデータを分割して，実験を行う．言語指示は３種類である，このオブジェクトはどこの部屋にありますか？このオブジェクトの色は何ですか？この部屋のこのオブジェクトの色は何ですか？オブジェクトの種類は 25 である．部屋の種類は 18 である．

表１：本実験で使用したデータセット

	環境シーン	エピソード
訓練データ	57	9024
バリエーションデータ	24	2747
テストデータ	2	25

評価基準は二つがある．成功率：ナビゲーションが終了する際に，最後の１フレームにターゲットがあると成功として設定する．高い方は性能が良い．接近距離：ナビゲーションが終了する際にターゲットまでの距離である．低い方は性能が良いことを示す．

４，結果と考察

異なるモデルの成功率及び接近距離の比較結果は図 3 に示す．左の方は成功率である．混合エキスパートの成功率が一番高いことが分かった．また，右に方は接近距離をバイオリンプロットにした結果である．低い方は性能が良いので，混合エキスパートの方も性能が良いことを示した．一つのエキスパートを使うモデルは LSTM のみのモデルより，良い結果を示したので，画像を予測することをサブタスクとして学習することはナビゲーションの性能に良い影響を与えることが分かった．また，一つのエキスパートを使うことより，混合エキスパートは異なるエキスパートを切り替えることで，性能を向上させることが分かった．

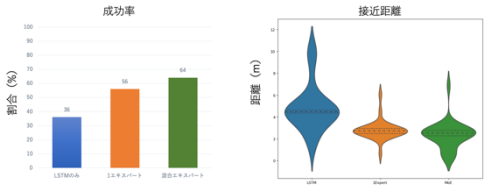

図３　３種類のモデルの成功率と接近距離の比較

５，まとめ

本実験では，EmbodiedQA タスクの性能を向上させるため，ナビゲーションモジュールに異なるステップの画像を予測する５つのエキスパートをゲーティングネットワークで統合し，混合エキスパートを構築した．Matterport3D シミュレーター上で評価実験を行った結果は混合エキスパートの性能が最も高いことを確認した．

６，未来の展望

未来の展望では，シミュレーション環境ではなく，実機で実装する．また，人間とのインタラクションを増やす．例えば，人間と交流しながら，タスクを完成する．

７，参考文献

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, Anton van den Hengel,Vision-and-Language Navigation: Interpreting visually-grounded navigation instructions in real environments. CVPR 2018

[2] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, Dhruv Batra. Embodied Question Answering. CVPR 2018

[3] Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, Dhruv Batra. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. CVPR 2019

[4] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Nießner, Manolis Savva, Shuran Song, Andy Zeng, Yinda Zhang. Matterport3D: Learning from RGB-D Data in Indoor Environments. International Conference on 3D Vision (3DV) 2017