5B-01

Masked Autoencoding による触覚と固有受容感覚の経験を通した身体近傍空間の認識

野口　渉[†]　飯塚　博幸[†‡]　山本　雅人[†‡]

北海道大学　大学院情報科学研究院[†]　北海道大学　人間知・脳・AI 研究教育センター[‡]

1　はじめに

物体の操作や障害物の回避といった適応的な行動のためには，自身の身体の空間的な位置や周囲の物体配置を認識する空間認識能力が必要であるが，空間認識能力は，生得的ではなく経験を通して構築されていくものである．我々の感覚器官は空間そのものを直接観測するわけではないが，観測は感覚器官を備える身体の空間配置に依存し決定される．この感覚運動の依存関係の経験を通して，その依存関係を決定する空間について認識できると考えられている．

Laflaquière らは，ニューラルネットワークモデルに予測学習という形で感覚運動の依存関係を学習させることで，身体近傍空間の空間座標の内部表現が獲得されることを示した[1]．また，[2]は，同モデルを物体形状の認識に拡張した．一方，これらのモデルは記憶構造をもたず，局所的な観測単独では把握できない近傍空間全体の物体配置の認識，すなわち，身体近傍空間のマップを構築する能力は備わっていない．

本研究では，複数の観測を統合しつつ感覚運動の依存関係を学習するため，Transformer[3]に基づく自己注意機構を備え，Masked Autoencoding[4]学習を行うモデルを提案する．また，シミュレーションにより，提案モデルが触覚と固有受容感覚（固有感覚）の経験を通して身体近傍空間のマップ構築することを示す．

2　シミュレーション・モデル

2.1 シミュレーション環境

図１にシミュレーション環境を示す．環境平面に，３つの関節自由度をもつアームロボットが設置され，同平面上には「コ」の字形の物体が 2 つ配置されている．物体の位置と向きは可変である．アーム先端の手のひらは触覚センサーを備え，平面上の物体に触れることで物体を観測

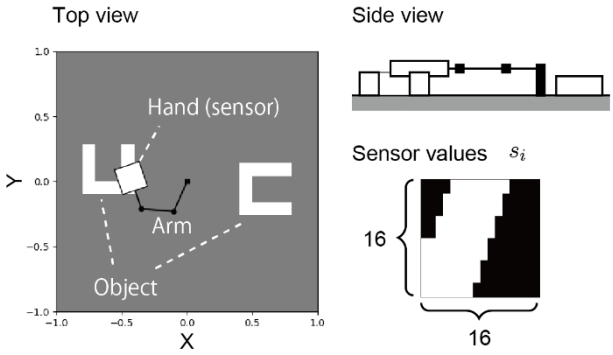

図１　シミュレーション環境

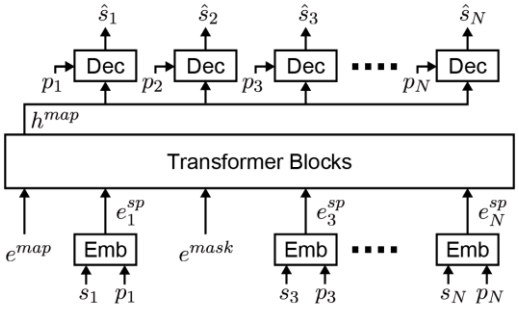

図２　提案モデルの概要

する．ロボットは様々な姿勢において，固有感覚（３関節の関節角度）$p \in \mathbb{R}^3$ と触覚 $s = f(p;o) \in \mathbb{R}^{16\times16=256}$（$f$は物体配置$o$における触覚観測を決定する関数）を得る．ここで，ロボットは，アームの空間座標や手のひらのセンサーの空間配置といった空間の事前知識を有しない．

2.2 ネットワークモデル

提案モデルを図２に示す．モデルは 1)触覚と固有感覚の埋め込みモジュール，2)複数の観測を統合する統合モジュール(Transformer Blocks)，3) 触覚予測を出力するデコードモジュールからなり，全体として触覚と固有感覚のペア群 $X = \{(s_i, p_i)\}_{i=1,\dots,N}$ を入力とし，触覚を再構築する．

統合モジュールは，自己注意機構により複数の触覚・固有感覚の埋め込みベクトルを統合する．とくに，感覚入力とは別のモデル内部に保持する入力ベクトル e^{map} を用意し，e^{map} に対応した出力を観測の統合されたベクトル h^{map} としてデコードモジュールに渡す．デコードモジュールは，h^{map} と p_i に加え，各j番目の触覚受容野に対

Recognition of peripersonal space through touch and proprioception with masked autoencoding
Wataru Noguchi[†], Hiroyuki Iizuka[†‡], and Masahito Yamamoto[†‡]
† Faculty of Information Science and Technology, Hokkaido University
‡ Center for Human Nature, Artificial Intelligence, and Neuroscience, Hokkaido University

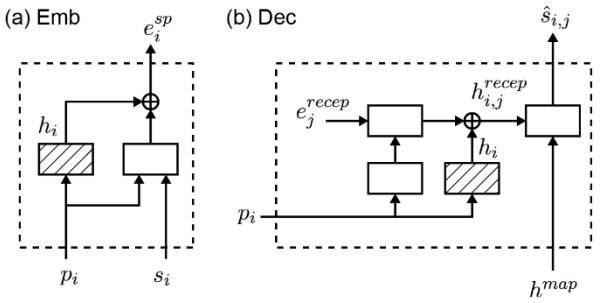

図３　(a)埋め込みモジュール（b）デコードモジュール．各ボックスは全結合層からなるサブモジュール．網かけ部は共有のサブモジュールである．

応するベクトルe_j^{recep}（e_j^{recep}はe^{map}と同じくモデル内部に保持するベクトル）を用いて，受容野ごとに出力$\hat{s}_{i,j}$を計算し，各出力をまとめることで触覚予測$\hat{s}_i$を生成する．埋め込み・デコードモジュールの詳細は図３に示す．

モデルの学習は，Mask Autoencoding[4]の形式で行う．すなわち，一部の触覚・固有感覚入力(s_i, p_i)を欠損，つまりマスクした状況で，触覚の再構築を学習する．なお，マスクされた入力の代わりにはマスクベクトルe^{mask}が入力される．マスクされていない観測を過去の観測，マスクされている観測を未来の観測と考えれば，このMask Autoencoding学習は，過去の触覚・固有感覚の観測から，未来にとる姿勢の固有感覚に対応する触覚の予測を学習することに対応する．

3　実験

3.1 実験設定

モデルの学習のために，Kパターンの物体配置で触覚・固有感覚のデータセット$D = \{X_k\}_{k=1,\dots,K}$を収集する．$X_k$は同一の物体配置の環境$o_k$で観測された触覚・固有感覚ペア群である．本実験では，同一配置での触覚・固有感覚ペア数を$N = 30$とした．ただし，物体が観測される場合が多くなるように30ペアのうち20ペアは，手の中心が物体上に置かれた姿勢のみを選択し収集した．また，物体配置のパターン数$K = 10000$とした．モデルの学習は触覚予測の予測誤差の最小化を目的とし，最適化アルゴリズムはAdamを用いた．

3.2 実験結果

予測学習後にデコードモジュール内のベクトル（図３(b)中の$h_{i,j}^{recep}$）を用いて触覚予測を可視化した．具体的には，主成分分析（PCA）を適用して$h_{i,j}^{recep}$を2次元空間にマッピングし，それぞれの点は$h_{i,j}^{recep}$と対応する予測$\hat{s}_{i,j}$により色付ける．また，身体近傍空間全体を十分観測するだけの姿勢を用意し，それぞれの固有感覚p_iに対する出力を同一の空間に可視化する．図４に可視化した結果を示す．可視化結果からは，主成分空間にシミュレーション空間と同様な物体形状が確認できる．これはベクトル$h_{i,j}^{recep}$が環境平面に対応した2次元の構造を構成しており，かつ，同ベクトル空間中に環境の物体配置を再現するように触覚がマッピングされていることを示す．つまり，学習後のネットワークが，アームの可動範囲内の身体近傍空間について触覚と紐づいた空間マップを内部に再構築したことを示す．

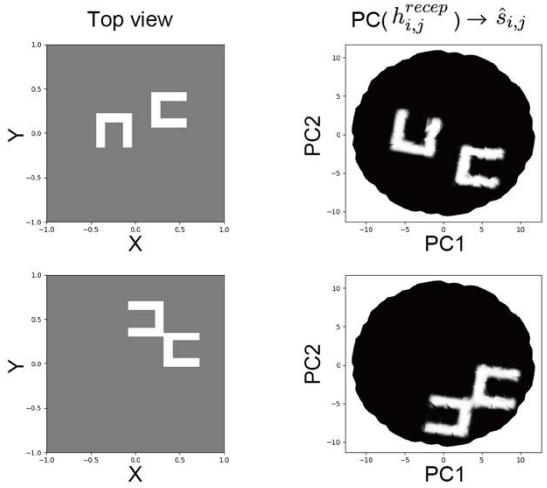

図４　学習後モデルの身体近傍空間マップ

4　おわりに

本研究では，空間の事前知識をもたないモデルが，触覚と固有感覚のMasked Autoencodingの学習により身体近傍空間のマップを構築するシミュレーションを行なった．今後は，より複雑な形状の物体に対応できるモデルの構築を試みる．

謝辞

本研究はJSPS科研費JP20K19880の助成を受けたものです．

参考文献

[1] Laflaquière, A., and Ortiz, M. G.: Unsupervised emergence of egocentric spatial structure from sensorimotor prediction, *Adv. Neural Inf. Process. Syst.* Vol.32, pp.7158-7168 (2019).

[2] 野口渉: 感覚運動予測学習による物体形状表現の獲得モデル. 情報処理学会第84回全国大会講演論文集, pp.15-16 (2022)

[3] Vaswani, A., Shazeer, N., Parmar, N., et al.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* Vol.30, pp.5998-6008 (2017).

[4] Devlin, J., Chang, M. W., Lee, K., et al.: Bert: Pre-training of deep bidirectional transformers for language understanding. *Proc. NAACL-HLT*, Vol.1, pp.4171-4186 (2019).