MaskDP による事前学習のマルチドメイン拡張

○鈴木 佳三　板谷 英典（中部大学）　村瀬 卓也　佐々木 一磨（株式会社ドワンゴ）
平川 翼　山下 隆義　藤吉 弘亘（中部大学）

Masked Decision Prediction (MaskDP) はマスクされた状態と行動の軌道を復元する事前学習を行うことで強化学習の効率化を図る手法である．しかし，この手法は事前学習と追加学習のドメインが同じ必要がある．そこで MaskDP をマルチドメインへ拡張した Multi-Domain MaskDP を提案する．MuJoCo におけるロボットタスクで性能を維持したままマルチドメイン事前学習ができることを示す．

1. はじめに

自然言語処理やコンピュータビジョン分野において，基盤モデルが大きな注目を集めている．この基盤モデルは，画像や音声，文章などの多様なモダリティデータによる大規模な事前学習済みの AI モデルである．基盤モデルを特定のタスクで追加学習することにより，機械翻訳や画像認識，画像キャプション生成といった多種多様な下流タスクにおいて高いパフォーマンスを発揮することが知られている．特に Generative Pre-Trained Transformer (GPT) [1] は，高い汎用性を有していることからビジネスや教育，エンターテイメントなどの様々なアプリケーションとして利活用されている．

このような成功を背景に，深層強化学習分野では様々なタスクで高性能な制御を実現する基盤エージェントモデルが期待されている．深層強化学習は，従来の基盤モデルが対象とする教師あり学習と異なり，環境とのインタラクションから獲得する経験にもとづいたシーケンスデータを用いてエージェントの最適な制御を学習する．そのため，基盤エージェントモデルの実現には深層強化学習の性質に合わせた事前学習法の確立が必要である．

Masked Decision Prediction (MaskDP) [2] は，深層強化学習を対象とした自己教師あり事前学習法である．この手法は，経験データである状態と行動の軌道を一部マスク処理し，マスク箇所を復元する事前学習法であり，下流タスクにおける学習の効率化を実現している．しかし，MaskDP は事前学習と追加学習時のドメインが同一であることを前提とし，状態/行動次元数が異なる他ドメインへの適用が困難である．このことから，MaskDP では事前学習に複数ドメインのデータを利用できないため，ドメインを跨ぐ汎用的な特徴が獲得できないという課題が存在する．

本研究では，MaskDP をマルチドメインへ拡張した Multi-Domain MaskDP (MD-MaskDP) を提案する．MD-MaskDP では，ドメインごとの状態/行動次元数に合わせた埋め込み層とヘッドを構築することで，異なる複数ドメイン間における入出力次元数のギャップを解消する．複数ドメインの状態と行動の軌道を用いたマルチドメイン事前学習により，エージェントモデルの汎用性向上を実現する．評価実験では，物理演算エンジン MuJoCo [3] をベースとしたロボット制御タスク DeepMind Control Suite [4] を用いて，マルチドメイン事前学習の有用性を示す．

2. 関連研究

基盤モデルは高い汎用性から様々なタスクに適用可能であり,多種多様な分野で盛んに研究されている [1, 5, 6]．これら基盤モデルは大規模な事前学習により獲得される．以下では基盤モデルの事前学習について述べる．

Reed らは様々なモダリティデータを用いた大規模な教師あり学習により，高い汎用性を獲得した Gato を提案している [7]．この手法はモーダルごとに合わせた前処理を施すことで，様々なモダリティデータでの事前学習を実現している．Anthony らは，大規模な Vison-and-Language モデルを用いてロボット制御を実現した Robotic Transformer 2 (RT-2) [8] を提案している．RT-2 は，ロボットの動作をテキストトークンとして表現し，自然言語と同様にモデルの学習を可能にした．He らは，Vision Transformer (ViT) [9] に対し，マスク箇所の再構成による自己教師あり事前学習 Masked Autoencoder (MAE) を提案している [10]. MAE は, パッチ分割した入力画像の一部にマスク処理を施し，マスク箇所の再構成を行うことで，画像分類タスクにおける精度向上を実現している．Caron らは，MAE と同様に ViT を対象とした自己教師あり事前学習法である self-distillation with no labels (DINO) を提案してる [11] ．DINO は教師モデルと生徒モデルを構築し，ラベル無しデータを用いた自己蒸留による学習を行うことで，画像分類タスクの精度向上を実現している．

汎用的なエージェントモデルの実現に向けた研究も取り組まれている．Lee らは，ビデオゲームタスクにおける汎用的なエージェントモデルの実現を目指し，Multi-Game Decision Transformer を提案している [12] ．この手法は，46 種のビデオゲームタスクを用いたオフライン強化学習による事前学習を行い，単一のエージェントモデルで多数のゲームタスクにおける高スコアを獲得している．Seo らは，画像ベースな強化学習における教師なし事前学習として Action-Free Pre-training from Videos (APV) を提案している [13] ．APV は，アクションフリーな動画を用いた Autoencoder による事前学習を行うことで，画像特徴と環境のダイナミクス情報の獲得している．Xiao らは, 画像ベースな強化学習エージェントの特徴抽出部に対し，MAE を用いた事前学習を提案している [14] ．画像の再構成による事前学習を行うことで，ロボット制御タスクにおいて学習効率の向上を実現している．

Liu らは深層強化学習における状態と行動のシーケンスデータに着目し，マスク処理後のシーケンスデータの再構成による事前学習 MaskDP を提案している [2]. MaskDP は, MAE のアイデアに触発されたものであり，事前学習により環境のダイナミクス情報を考慮したエージェントモデルの獲得を実現している．MaskDP では，

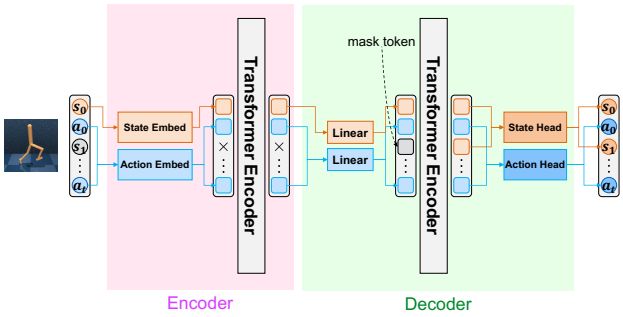

図 1　MaskDP の概略

深層強化学習特有の要素である状態と行動をそれぞれトークンとして扱い，時系列に並んだ状態と行動の関係性を学習する．それにより，追加学習においてエージェントは環境の理解が容易になり，学習効率向上に繋がる．

3. Masked Decision Prediction

本研究は基盤エージェントモデルの実現に向けた事前学習法に着目したものである．そこで，本節ではエージェントモデルの代表的な事前学習法 Masked Decision Prediction (MaskDP) [2] に触れ，問題点について述べる．

3.1 MaskDP による事前学習

MaskDP は，状態と行動から構成されるシーケンスデータの一部にマスク処理を施し，マスクされた領域を復元する深層強化学習における自己教師あり事前学習である．MaskDP の概略を図 1 に示す．

MaskDP は，マスク処理適用後の状態と行動のシーケンスデータから特徴抽出を行う Encoder と，Encoder で抽出した特徴と mask token から元の状態や行動を復元する Decoder で構成される．Encoder と Decoder は Transformer Encoder [15] により構築され，入力間の関係，つまり状態と行動の関係を表現した特徴抽出を行う．Encoder の入力はマスク処理適用後のシーケンスデータであり，これらは埋め込み層を用いて埋め込みベクトルへ変換する．ここでの埋め込む層は，状態と行動ごとに 2 つの埋め込み層 (State Embed, Action Embed) を構築する．埋め込み処理を施したのち，Transformer Encoder へ入力し特徴抽出を行う．また，Decoder の Transformer Encoder には，マスク箇所を mask token に置換した状態と行動の特徴ベクトルを入力する．ここでの mask token は復元すべき対象であり，マスク箇所全てに共通した学習可能なベクトルある．最後に Transformer Encoder の出力ベクトルを，各状態と行動ごとのヘッド (State Head, Action Head) を用いて入力値であった状態と行動に復元する．

損失関数は，入力であるマスク処理前のシーケンスデータと復元されたシーケンスデータ間における平均二乗誤差を採用する．なお，強化学習では長期的な状態遷移を考慮することが重要であるため，マスクされていない領域も含めたシーケンス全体で損失計算を行う．

3.2 MaskDP の問題点

MaskDP は環境のダイナミクス情報を考慮することができ，下流タスクで高い性能を達成している．しかし，事前学習と追加学習時のドメインが同一であることを前提としているため，状態や行動数の異なる他ドメインへの適用が困難である．これは Transformer Encoder への入力である埋め込みベクトルを，特定のドメインに対する状態と行動の埋め込み層のみを用いて算出しているためである．このことから，事前学習済みモデルを適用する下流タスクは事前学習時のドメインが同一でなければならないという制約が存在する．この問題を解決することで，ドメインを跨ぐ環境のダイナミクス情報の獲得，および事前学習済みモデルの利活用における応用範囲の拡張に繋がる．

4. Multi-Domain MaskDP

本研究では基盤エージェントモデルの実現に向け，3.2 節で述べた問題に着目する．そこで，MaskDP を複数ドメインの経験による学習を可能に拡張した Multi-Domain MaskDP (MD-MaskDP) を提案する．本手法は，異なるドメイン間における入出力次元数のギャップを解決し，エージェントモデルの汎用性向上を図る．

4.1 異なる複数ドメインの経験を用いた事前学習

MD-MaskDP の概略を図 2 に示す．MD-MaskDP では Encoder 部の状態と行動の埋め込み層 (State Embed, Action Embed) をドメインごとに独立して構築する．ドメインごとに適した状態と行動の埋め込み処理を行うことで，Transformer Encoder へ入力する特徴ベクトルの次元数を統一し，ドメイン間における入力次元数のギャップを解決する．Decoder 部も同様に，特徴ベクトルから状態と行動を復元するヘッド (State Head, Action Head) をドメインごとに構築する．ドメインに適した再構成により，ドメインごとの状態と行動に合わせた復元を実現する．また Encoder と Decoder 部における Transformer Encoder は各ドメイン間で共通して用いることで，複数ドメインに共通した特徴抽出を図る．これらにより，MD-MaskDP では異なる複数ドメインの状態と行動の軌道を用いたマルチドメイン事前学習を可能としている．

MD-MaskDP の事前学習では，環境のダイナミクス情報をドメインを跨いで汎用的に学習することが目的である．そこで事前学習には，内部報酬のみ用いて学習した強化学習エージェントによる状態と行動のシーケンスデータを用いる．内部報酬はエージェントの内発的な動機付けや目標にもとづく報酬であり，エージェントの探索行動の促進に重要な要素である．そのため，本事前学習データはタスクに依存しないドメイン固有のシーケンスデータである．

4.2 Behavior Cloning による追加学習

本研究では，下流タスクへの適応に模倣学習の一種である Behavior Cloning による追加学習を行う．追加学習時のモデル構造を図 2 に示す．追加学習時のモデルは，状態を埋め込みベクトルに変換する埋め込み層 (State Embed) と Transformer Encoder，行動を出力するための Actor Head で構築する．ここで，State Embed と Transformer Encoder は事前学習済みモデルの Encoder 部から流用する．追加学習として，モデルが出力した行動と，教師データである行動の平均二乗誤差による損失に基づいてモデルパラメータ更新を行う．

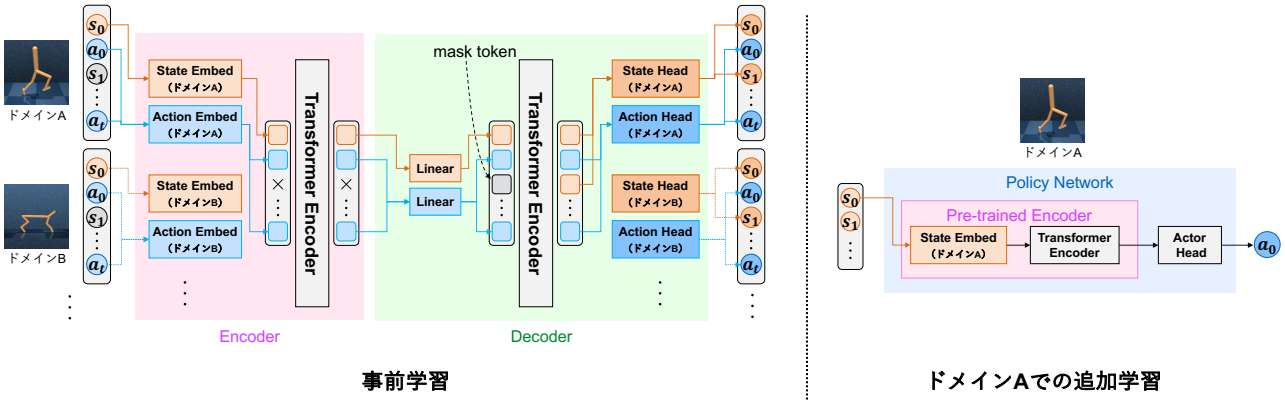

図 2　MD-MaskDP の概略

表 1　ドメインの状態次元数及び行動次元数

ドメイン	状態次元数	行動次元数
Walker	24	6
Cheetah	17	6
Quadruped	78	12

追加学習には，外部報酬を用いて学習した強化学習エージェントによる状態と行動のシーケンスデータを用いる．本追加学習データは，タスク固有の外部報酬にもとづくデータであるため，タスクに依存したシーケンスデータである．

5.　評価実験

MD-MaskDP の有用性を確認するため，ロボット制御タスクを用いた評価実験を行う．追加学習時の各タスクにおける報酬推移を比較することで，MD-MaskDP の有効性を示す．

本実験では DeepMind Control Suite [4] を用いて評価を行う．DeepMind Control Suite は，物理演算エンジンである MuJoCo [3] をベースとする強化学習環境のフレームワークであり，様々な強化学習向けの制御タスクが用意されている．使用するドメインは Walker, Cheetah, Quadruped の 3 つである．各ドメインの状態次元数及び行動次元数を表 1にまとめる．Walker ドメインは，2 次元動作による人型の 2 足歩行ロボットである．Cheetah ドメインは，2 次元動作によるチーターのような形状を持つ 2 足ロボットである．Quadruped ドメインは，3 次元動作による 4 足ロボットである．Walker ドメインが取り組む stand タスクの報酬は，ドメインの胴体の高さと直立具合によって決定され，胴体がより高い位置で垂直に近い状態であるときに大きな報酬が得られる．Walker ドメインと Quadruped ドメインが取り組む walk タスクの報酬は，stand タスクの報酬に加え，ドメインの並進速度によって決定される．walk タスクでは，より安定した姿勢を維持しながら，指定の並進速度以上で移動している場合に多くの報酬が得られる．Walker ドメインと Cheetah ドメインが取り組む run タスクの報酬は，walk タスクと同様のルールで決定されるが，指定される並進速度が walk タスクより高い値になる．そのため run タスクでは，walk タスク以上に速く移動することが求められる．

5.1　実験条件

比較手法として，我々の手法であるマルチドメイン事前学習によるモデル (MD-MaskDP)，下流タスクと同一ドメインでの事前学習によるモデル (MaskDP)，事前学習なしのモデル (scratch) の 3 つを用いる．つまり，MD-MaskDP は Walker，Cheetah，Quadruped の 3 ドメインによる事前学習済みモデルであり，MaskDP は追加学習時のドメインと同様な 1 ドメインによる事前学習済みモデルである．

事前学習時の学習条件はバッチサイズを 384，学習率を $1e-4$ とする．事前学習データの収集は Proto-RL [16] による内部報酬のみで学習した強化学習エージェントを用い，1 ドメイン当たり $10,000,000$ step 分の状態および行動のシーケンスデータを収集した．また，1 ドメイン当たりの学習回数を揃えるため，イテレーション数を MD-MaskDP では $1,200,000$，MaskDP では $400,000$ とした．一方で，追加学習時の学習条件はバッチサイズを 384，学習率を $1e-4$ とする．追加学習データの収集は Twin Delayed DDPG (TD3) [17] による外部報酬で学習した強化学習エージェントを用い，$2,000,000$ step 分の状態および行動のシーケンスデータを収集した．

5.2　追加学習時の報酬推移による比較

各比較手法における追加学習時の報酬推移を図 3 に示す．Walker ドメインでは stand, walk, run タスクのいずれにおいても，MD-MaskDP が学習初期から scratch より高い報酬を獲得し，MaskDP と同等であることが分かる．これらの結果は，Walker ドメインとは状態および行動次元数の異なる Cheetah ドメインと Quadruped ドメインにおいても同様に確認できる．また Quadruped ドメインは Walker や Cheetah と比較し，状態および行動ともに次元数が大きいドメインであるため，Quadruped の walk は難易度の高いタスクである．MD-MaskDP はこのタスクにおいても十分な性能を獲得できていること確認できる．

ここで MaskDP は事前学習と追加学習が同一ドメインであるため，追加学習に最も効果的である事前学習が行われた手法である．また 3 ドメインにおける MaskDP はドメインごとに独立したモデルであるのに対し，MD-MaskDP は 3 ドメインで事前学習した単一なモデルである．つまり，MD-MaskDP は各ドメインに特化した事

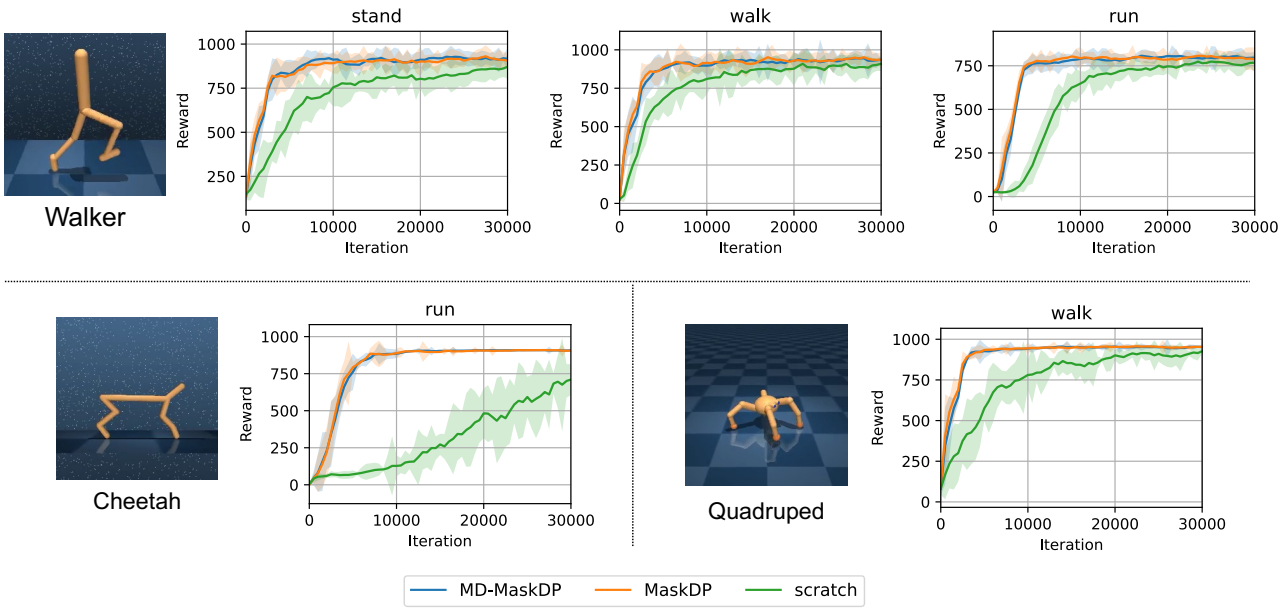

図 3　3 ドメインにおける追加学習時の報酬推移：実線は 6 つのシード値における平均値，色付き領域は 95% 信頼区間を示す．

前学習と同様の性能を達成し，3 ドメインで高い性能を獲得可能な汎用的なモデルであると考えられる．これらの結果から，異なる複数ドメインのシーケンスデータを用いるマルチドメイン事前学習が有効であると示した．

6.　おわりに

従来の MaskDP では，入出力次元数が異なるようなドメインを跨ぐ事前学習が困難であった．これは，エージェントモデルの応用範囲を限定することに繋がり，エージェントモデルの汎用性低下を招く問題である．そこで本研究では，基盤エージェントモデルの実現に向け，MaskDP をマルチドメインへ拡張した事前学習法 MD-MaskDP を提案した．ドメインごとに適した状態と行動の埋め込み層とヘッドを構築することで，ドメイン間の入出力次元数のギャップを解決し，マルチドメイン事前学習を実現した．DeepMind Control Suite のロボット制御タスクを用いた評価実験から，3 ドメインで事前学習した単一モデルの MD-MaskDP が，1 ドメイン特化の MaskDP と同等の性能を維持しながら，複数タスクでの学習効率向上を確認した．これらの結果により，異なる複数ドメインの状態と行動の軌道を利用した MD-MaskDP によるマルチドメイン事前学習の有用性を示した．

今後は MD-MaskDP での事前学習時に用いるドメインの増加によるエージェントモデルの汎用性向上，および事前学習データに含まない未知ドメインへの適応実験について取り組む予定である．

参　考　文　献

[1] T. Brown, *et al.*: “Language models are few-shot learners,” NeurIPS, vol. 33, pp. 1877–1901, 2020.

[2] F. Liu, *et al.*: “Masked Autoencoding for Scalable and Generalizable Decision Making,” NeurIPS, vol. 35, pp. 12608–12618, 2022.

[3] E. Todorov, *et al.*: “MuJoCo: A physics engine for model-based control,” IROS, pp. 5026–5033, 2012.

[4] S. Tunyasuvunakool, *et al.*: “dm_control: Software and tasks for continuous control,” Software Impacts, vol. 6, pp. 100022, 2020.

[5] R. Anil, *et al.*: “Gemini: A Family of Highly Capable Multimodal Models,” *arXiv:2312.11805*, 2023.

[6] A. Kirillov, *et al.*: “Segment Anything,” ICCV, pp. 3992–4003, 2023.

[7] S. Reed, *et al.*: “A generalist agent,” *arXiv : 2205.06175*, 2022.

[8] A. Anthony, *et al.*: “RT-2: Vision-Language-Action Models Transfer Web Knowledge to Robotic Control,” *arXiv : 2307.15818*, 2023.

[9] A. Dosovitskiy, *et al.*: “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” ICLR, 2021.

[10] K. He, *et al.*: “Masked Autoencoders Are Scalable Vision Learners,” CVPR, pp. 16000–16009, 2022.

[11] M. Caron, *et al.*: “Emerging properties in self-supervised vision transformers,” ICCV, pp. 9650–9660, 2021.

[12] K. Lee, *et al.*: “Multi-game decision transformers,” NeurIPS, vol. 35, pp. 27921–2793, 2022.

[13] S. Seo, *et al.*: “Reinforcement Learning with Action-Free Pre-Training from Videos,” PMLR, pp. 19561–1957, 2022.

[14] T. Xiao, *et al.*: “Masked visual pre-training for motor control,” *arXiv : 2203.06173*, 2022.

[15] A. Vaswani, *et al.*: “Attention is All You Need,” NeurIPS, vol. 30, 2017.

[16] D. Yarats, *et al.*: “Reinforcement learning with prototypical representations,” ICML, pp. 11920–11931, 2021.

[17] S. Fujimoto, *et al.*: “Addressing function approximation error in actor-critic methods,” ICML, pp. 1587–1596, 2020.