深層強化学習における MaskDP を用いたマルチドメイン事前学習

ER20045 鈴木 佳三　　指導教授：藤吉 弘亘

1. はじめに

自然言語処理やコンピュータビジョン分野では，多種多様なタスクに適応可能な基盤モデルが注目されている．基盤モデルは，膨大なデータを用いた大規模な事前学習後，特定のタスクにファインチューニングすることで様々な下流タスクに適応できる．同様に，深層強化学習においても様々なタスクを解くことができる基盤エージェントモデルの実現が期待されている．その 1 つのアプローチとして，深層強化学習における自己教師あり学習法である Masked Decision Prediction (MaskDP) [1] が提案されている．MaskDP では事前学習と下流タスクが同一のドメインである必要がある．そのため，適応可能な下流タスクが限定される．そこで本研究では，異なる複数ドメインの経験を用いた Multi-Domain MaskDP を提案する．様々なドメインの経験を用いた学習により各ドメインに共通する特徴を獲得し，下流タスクにおける学習効率の向上及び基盤エージェントモデルの実現を図る．

2. Masked Decision Prediction (MaskDP)

MaskDP は，深層強化学習における自己教師あり学習法である．MaskDP のネットワーク構造を図 1 に示す．MaskDP は，状態と行動のシーケンスデータに対して Masked Autoencorder (MAE) [3] のようにマスク箇所の再構成を行う事前学習手法である．Encoder では一部の領域をマスクしたシーケンスデータから特徴抽出を行う．Decoder では Encoder で抽出した特徴とマスクトークンから元の状態及び行動を再構成する．

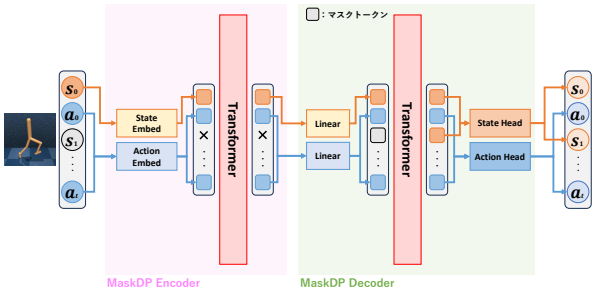

図 1：MaskDP の概略

3. 提案手法

異なるドメインのシーケンスデータから共通した特徴を獲得できる，自己教師あり事前学習手法 Multi-Domain MaskDP (MD-MaskDP) を提案する．MD-MaskDP のネットワーク構造を図 2 に示す．MaskDP による事前学習に異なるドメインのシーケンスデータを用いた場合，ドメイン間における入出力次元数の違いが問題となる．そこで，Encoder 部の埋め込み層をドメインごとに独立して構築する．入力データに適した埋め込み層を用いることで，Transformer Block へ入力する特徴の次元数を統一する．また，Decoder 部も同様に，状態及び行動を再構成する Head をドメインごとに構築し，ドメインに適した再構成を行う．これにより，ドメイン間の入力及び出力次元数の違いによる問題を解決する．

4. 評価実験

事前学習なしである scratch，MaskDP，MD-MaskDP におけるファインチューニング時の報酬推移を比較する．

4.1 実験条件

強化学習環境には DeepMind Control Suite [2] の Walker, Quadruped, Cheetah を使用する．事前学習の対象ドメインとして，MaskDP では Walker，MD-MaskDP では Walker, Quadruped, Cheetah を用いる．学習条件はバッチサイズを 384，イテレーション数を 400,000，学習率を 1e-4 とする．事前学習データセットは，内部報酬に基づきエージェントが収集した対象ドメインの経験データを使用

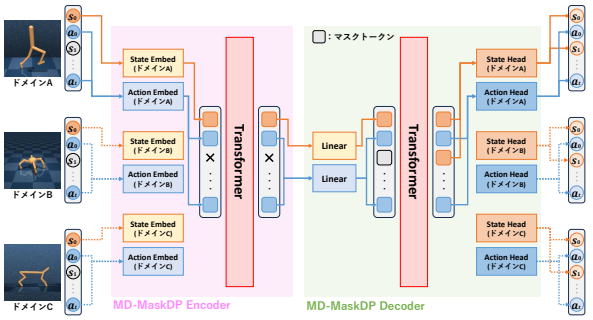

図 2：Multi-Domain MaskDP の概略

する．本データはタスクに依存しないドメイン固有の経験である．下流タスクとして，Walker の stand, walk, run タスクを用いたオフライン強化学習によるファインチューニングを行う．学習条件はバッチサイズを 384，イテレーション数を stand タスクでは 100,000，walk タスクでは 200,000，run タスクでは 1,000,000，学習率を 1e-4 とする．ファインチューニングデータセットは，外部報酬に基づきエージェントが収集した Walker における対象タスクでの経験データを使用する．本データはタスクに依存した経験データである．

4.2 報酬推移による比較

ファインチューニング時の報酬推移を図 3 に示す．図 3 より，stand 及び walk タスクにおいて，MaskDP と MD-MaskDP は同等の報酬を獲得しており，学習初期で scratch より高い報酬を獲得している．このことから，提案手法によりドメイン間で共通した特徴を捉えることができ，Walker のみを対象とした事前学習と同等の効果を得られたと考えられる．一方，run タスクではいずれのエージェントも学習初期から中期にかけて同等の報酬を獲得している．これは，MD-MaskDP 及び MaskDP による事前学習が不足しており，stand 及び walk タスクより難易度の高い run タスクでの事前学習の効果は低いといえる．また，学習終盤では MD-MaskDP は MaskDP より高い報酬を獲得したが，scratch より報酬が低下している．この現象の原因は現時点では明らかになっておらず，今後の追加実験で原因調査を行う予定である．

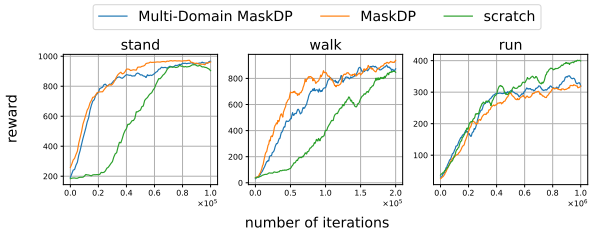

図 3：Walker における対象タスクでの報酬推移

5. おわりに

本研究では，基盤エージェントモデルの実現を目的とし，複数ドメインの経験を用いた自己教師あり学習法である MD-MaskDP を提案した．評価実験では，提案手法による下流タスクでの学習効率向上を確認できた．今後は，run タスクでの問題について調査を行う．また，ドメイン数の増加及び他ドメインでの評価実験に取り組む．

参考文献

[1] F. Liu, *et al.*, “Masked Autoencoding for Scalable and Generalizable Decision Making”, NeurIPS, 2022.

[2] Y. Tassa, *et al.*, “dm_control: Software and tasks for continuous control”, Software Impacts, 2020.

[3] H. Kaiming, *et al.*, “Masked Autoencoders Are Scalable Vision Learners”, CVPR, 2022.