階層型強化学習MLSHにおける枝刈りによるサブポリシ数調整
A Sub-policy Pruning Method for Meta Learning Shared Hierarchies

4Rin1-10

洪 青[2,1]、谷村 勇輔[1,2]、○中田 秀基[1,2] (1. 産業技術総合研究所、2. 筑波大学)

概要

背景
類似してはいるが異なる複数のタスク列が対象
強化学習は報酬が疎なため学習に時間がかかる
→ 以前のタスクに関する知識を流用したい

既存研究：MLSH[1]
階層型強化学習-サブポリシをマスタポリシで選択
サブポリシという形でタスクに関する知識を共有
→ サブポリシの数を設定する必要

研究の目的
MLSHのサブポリシ数を自動的に定める

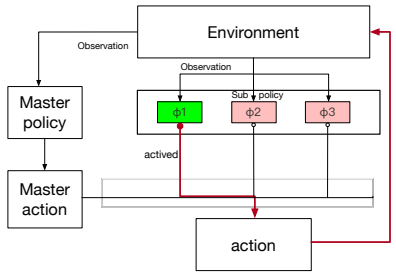
手法の概要
十分多い数を初期値として枝刈りしていく

成果
いくつかの枝刈り指標を提案
枝刈りの効果を確認

Meta Learning Shared Hierarchies

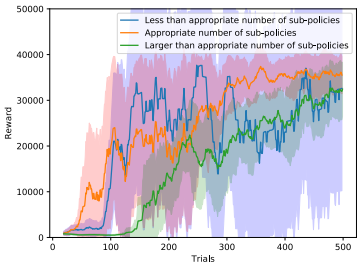
サブポリシ：
特定のサブゴールに対応したアクションを提案
タスク間で共有される
タスクをまたがって引き継がれる知識として機能

マスタポリシ：
サブゴールの提案の中から行うアクションを選択
タスクが切り替わると初期化される



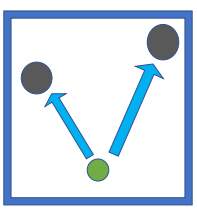
理想：
サブポリシ数
＝ サブゴール数

しかし一般には問題の
サブゴール数は不明



タスク列の例 – 2D Moving bandit

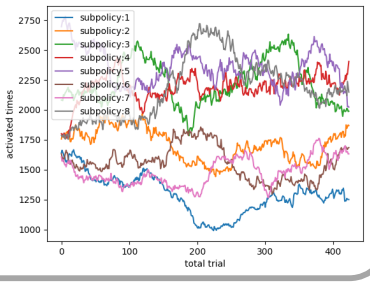
・２次元空間にいくつかの「点」が置かれる
・「点」のうちの一つが正解で、それに十分近寄ると報酬が得られる
・タスクごとに点の位置、正解となる点がランダムに 切り替わる



提案手法

基本的なアイディア
・十分な数のサブポリシで開始して「余分な」サブポリシを削除していく
・サブゴールの数に対してサブポリシの数が多いと一部のサブポリシは使われなくなるはず
→ 使われていないサブポリシを「余分」と判断して削除する

課題
・いつから ー
ある程度学習が進まないと判断できない
・どれを ー
最大値の 50%、60%で足切り

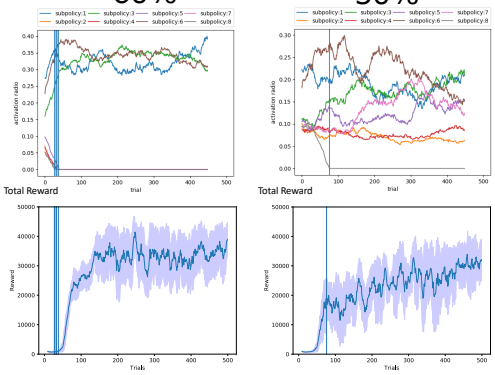


結果

2サブゴール、8サブポリシで実験
- 2サブゴールまで減らせれば正解

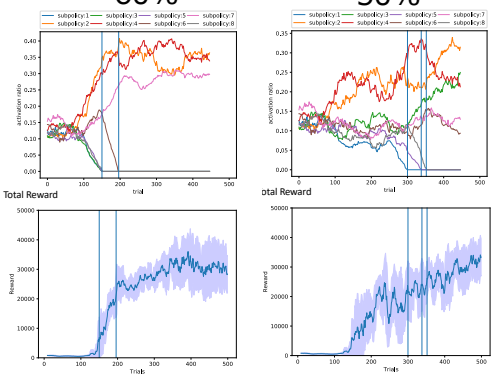
50タスク実行したところで枝刈り開始

60% 50%



150タスク実行したところで枝刈り開始

60% 50%



議論
・保守的手法（50%）では十分に枝刈りできない
・このケースでは50タスク実行時点で十分に「余分な」ゴールが判断できている
・手法の一般性の検証については今後の課題
- 他のタスクでの検証が必要

[1] Meta Learning Shared Hierarchies, Kevin Frans, Jonathan Ho, Xi Chen, Pieter Abbeel, John Schulman, ICLR '18.

この成果の一部は、国立研究開発法人新エネルギー・産業技術総合開発機構 （ＮＥＤＯ）の委託業務の結果得られたものです。
本研究はJSPS科研費 JP16K00116の助成を受けたものです。