효율적인 평생학습 알고리즘의 모델기반 강화학습 적용에 관한 연구

이병준°, 이종민, 최윤선, 장영수, 김기응
한국과학기술원
{bjlee, jmlee, yschoi, ysjang}@ai.kaist.ac.kr, kekim@kaist.ac.kr

A Study on Application of Efficient Lifelong Learning Algorithm to Model-based Reinforcement Learning

Byung-Jun Lee, Jongmin Lee, Yunseon Choi, Youngsoo Jang, Kee-Eung Kim
KAIST

요 약

평생학습 문제는 여러 가지 다른 태스크들을 연속해서 학습하는 문제로, 범용 인공지능 에이전트의 연구에 있어 매우 중요하다. 본 논문은 지도학습 분야의 잘 알려진 평생학습 알고리즘, Efficient lifelong learning algorithm (ELLA: 효율적인 평생학습 알고리즘)을 모델기반 강화학습 분야에 적용하는 것을 목표로 한다. 제안하는 알고리즘인 MB-ELRL은 한 번에 하나의 태스크에 접근 가능하지만 태스크들의 동역학들 사이의 공유 가능한 정보를 효율적으로 학습하여 각 태스크를 독립적으로 학습하는 것에 비해 월등한 성능을 보인다.

1 서론

현실 세계에서 작동하는 자율 에이전트는 지속적으로 여러 데이터 스트림에 노출되는데, 우리는 궁극적으로는 이 에이전트가 동적 데이터 분포로부터 자율적으로 학습하고 또 필요한 정보를 기억하여 많은 종류의 태스크를 결국 모두 해결할 수 있기를 바란다. 이를 위해, 새로운 경험들로부터 지식을 점진적으로 습득하고, 이 지식을 장기간에 걸쳐 지속적으로 진화시키고, 주어진 태스크에 따라 다양한 방식으로 지식을 사용할 수 있어야 한다. 이렇듯 기존에 배운 경험을 유지하면서 새로운 지식을 수용하여 시간이 지남에 따라 지속적으로 학습할 수 있는 능력을 평생학습이라 하며, 이는 범용 인공지능 에이전트 개발의 핵심으로 오래도록 연구되어 왔다 [1, 2, 3].

평생학습 문제를 일반적인 기계학습 방법론에 기반하여 해결을 시도할 경우 일명 *파괴적 망각* 이라고 불리는, 새로운 정보에 대해 모델을 추가 학습하였을 때 기존에 학습된 지식을 치명적으로 손상하게 되는 현상에 맞닥뜨리게 된다 [4]. 일반적인 기계학습 방법론에서는 훈련 단계에서 모든 데이터를 사용할 수 있다고 가정하므로 데이터 분포가 변화하게 되면 전체 데이터에 대해 에이전트를 재훈련해야 하지만, 평생학습 문제에서는 이러한 재훈련 방식은 심각한 비효율성을 야기하여 새로운 데이터에 대한 실시간 학습을 불가능하게 만들 수 있다. 때문에 이 가정을 무시하고 새로운 데이터 입력만으로 에이전트를 지속적으로 훈련하다 보면, 기존의 지식이 새 지식으로 완전히 덮어씌워 지는 것이다.

이러한 파괴적 망각을 해결하기 위해서는 변화한 데이터 입력을 기반으로 기존 지식을 개선할 수 있어야 하고(가소성), 다른 한편으로는 기존 지식의 개선이 오래전 경험했던 태스크의 수행 능력을 크게 방해하지 않도록 해야 하며(안정성), 이 둘 사이의 균형을 맞추는 것이 바로 평생학습 문제의 쟁점이라고 할 수 있다 [5]. 이 문제를 올바르게 해결한 평생학습 에이전트는 실시간으로, 즉 일반적인 기계학습 방법론의 가정을 따라 모든 태스크 데이터에 대해 에이전트를 훈련하는 전형적인 다중-태스크 학습 알고리즘보다 매우 작은 시간 안에 학습이 가능하면서 큰 성능 손해를 보지 않을 것을 기대할 수 있다. 또한, 여러 태스크를 경험하며 태스크 사이에 공유되는 지식을 개선한 만큼 평생학습 에이전트는 각 태스크를 독립적으로 학습한 것에 비해서는 높은 성능을 가질 것이다.

본 논문에서는 지도학습 태스크들에 대해 전형적인 다중-태스크 학습 알고리즘과 거의 같은 성능을 유지하면서 실시간 학습을 가능하게 했던 기존의 평생학습 알고리즘인 ELLA [6]를 기반으로 하여 여러 순차적으로 주어지는 강화학습 태스크들을 실시간으로 학습해 나가는 모델기반 평생강화학습 알고리즘 MB-ELRL (Model-Based Efficient Lifelong Reinforcement Learning)을 제안한다.[1]

2 연구 배경

우리는 마르코프 의사결정과정(MDP; [7])으로 정의될 수 있는 순차적 의사결정 태스크들을 고려하며, 이 MDP는 일반적으로 $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, R, \gamma, d_0 \rangle$의 튜플로 형식화된다. $\mathcal{S}$는 에이전트가 직면할 수 있는 상태의 집합, $\mathcal{A}$는 에이전트가 취할 수 있는 행동의 집합, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$는 환경의 동역학을 기술하는 상태 전이 함수, $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$은 각 상태와 행동에 따라 환경이 제공하는 보상 함수, $\gamma \in [0, 1)$는 시간에 따른 보상의 중요도를 나타내는 감쇠 상수이며 $d_0 \in \Delta(\mathcal{S})$는 시작 상태 분포이다. 각 타임스텝 t에, 에이전트는 상태 $s_t \in \mathcal{S}$를 가지며 행동 $a_t \in \mathcal{A}$를 취해서 P에 따라 다음 상태 $s_{t+1} \sim p(s_{t+1}|s_t, a_t)$ 및 보상 $r_t = R(s_t, a_t)$를 받는다. 에이전트의 정책은 상태 행동 확률적 매핑으로 $\pi : \mathcal{S} \to \Delta(\mathcal{A})$으로 정의된다. 강화학습 에이전트의 목표는 감쇠된 보상합의 기댓

1) 이 논문은 2020년도 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원의 지원을 받아 수행된 연구임 (No.2020-0-00940, 안전한 강화학습 원천 기술 개발 및 자연어 처리에의 응용)

값, $R^\pi = \mathbb{E}_{d_0,P,\pi}[\sum_{t=0}^{T} \gamma^t r_t]$를 최대화하는 최적 정책 π^*를 찾는 것이다. 시작 상태에서부터 (무한할 수도 있는) 마지막 타임스텝 T까지의 상태 행동을 묶어서 하나의 궤적 $\tau = \{s_t, a_t\}_{t=0}^{T}$로 정의한다.

본 논문에서는 매 학습 에포크마다 위와 같이 정의된 MDP들: $\mathcal{M}^{(1)}, \ldots, \mathcal{M}^{(K)}$들 중 하나의 태스크가 주어지는 평생학습 상황을 가정한다. 일반성을 잃지않고, 우리는 K개의 태스크들이 전이 함수 P 및 보상 함수 R을 제외한 나머지를 공유한다고 가정한다. 즉, 각 태스크는 $\mathcal{M}^{(k)} = \langle \mathcal{S}, \mathcal{A}, P^{(k)}, R^{(k)}, \gamma, d_0 \rangle$의 형태로 정의된다. 이 가정은 수식 표현의 용이성을 위함이며, 가정을 만족하지 않는 형태로 곧바로 확장될 수 있다.

각 에포크마다 에이전트는 주어진 태스크와 상호작용하여 다수의 궤적을 얻어낼 수 있고, 이전에 한 번 학습한 적 있던 태스크가 다시 주어져 해당 태스크에 대해 중단되었던 학습을 재개할 가능성 역시 열려 있다. 하지만, 에이전트는 새 태스크가 주어지는 시점에서 어떤 태스크가 주어지는지, 그리고 어떤 확률 분포를 기반으로 태스크가 주어지는 알 수 없다. 이러한 상황에서, 에이전트는 다중-태스크 최적 정책 $\Pi : \mathcal{S} \to \Delta(\mathcal{A})^K$를 찾는 것을 목표로 한다.

3 본론

모델기반 강화학습의 일반적인 방법론은 전이 함수 P와 보상 함수 R을 가장 잘 근사하는 매개변수 $\boldsymbol{\theta} = (\boldsymbol{\theta}_p, \boldsymbol{\theta}_r)$를 찾고, 이를 기반으로 최적 정책을 찾는 것이다. 평생학습 문제에서 고려하는 다중-태스크 상황에서는 태스크마다 최적의 $\boldsymbol{\theta}$가 모두 다르므로, 에이전트는 태스크에 종속적인 매개변수 $\boldsymbol{\Theta} = \{\boldsymbol{\theta}^{(k)}\}_{k=1}^{K}$을 찾아야한다. ELLA [6]의 모델 설계에 기반하여 우리는 모든 태스크의 매개변수들이 같은 차원을 가진다고 가정하고, $\boldsymbol{\theta}^{(k)} = (\boldsymbol{\theta}_p^{(k)}, \boldsymbol{\theta}_r^{(k)}) \in \mathbb{R}^{d_p} \times \mathbb{R}^{d_r}$, 이가 일정 수, l 개의 잠재 매개변수에 의해 선형으로 표현된다고 가정한다. 즉, 잠재 매개변수 라이브러리 $\mathbf{L} = (\mathbf{L}_p, \mathbf{L}_r) \in \mathbb{R}^{d_p \times c_p} \times \mathbb{R}^{d_r \times c_r}$ 이 모든 태스크 사이에서 공유되고 각각의 태스크는 태스크 종속적인 웨이트 벡터인 $\mathbf{s}^{(k)} = (\mathbf{s}_p^{(k)}, \mathbf{s}_r^{(k)}) \in \mathbb{R}^{c_p} \times \mathbb{R}^{c_r}$로 표현되어 $\boldsymbol{\theta}^{(k)} = (\mathbf{L}_p \mathbf{s}_p^{(k)}, \mathbf{L}_r \mathbf{s}_r^{(k)})$의 관계식을 가진다.

만일 모든 태스크의 궤적들이 한 번에 주어졌다고 가정한다면, 우리는 동시에 모든 태스크 궤적들을 정확하게 예측하도록 학습할 것이고, 내부적으로는 여러 태스크에서 공유되는 모델의 잠재 매개변수를 잘 학습하여 전체적인 모델의 성능을 올리려고 한다. 이는 아래와 같은 목표식으로 표현된다:

$$J(\mathbf{L}) = \frac{1}{K}\sum_{k=1}^{K} \min_{\mathbf{s}^{(k)}} \left\{ \frac{1}{n_k}\sum_{i=1}^{n_k} \mathcal{L}(\mathbf{x}_{sa,i}^{(k)}, \mathbf{y}_{s',i}^{(k)}, \mathbf{y}_{r,i}^{(k)}; \mathbf{L}_p\mathbf{s}_p^{(k)}, \mathbf{L}_r\mathbf{s}_r^{(k)}) + \eta_p^{(k)}\|\mathbf{s}_p^{(k)}\|_1 + \eta_r^{(k)}\|\mathbf{s}_r^{(k)}\|_1 \right\} + \lambda_p\|\mathbf{L}_p\|_F^2 + \lambda_r\|\mathbf{L}_r\|_F^2 \quad (1)$$

위 목표식에서 $(\mathbf{x}_{sa,i}^{(k)}, \mathbf{y}_{s',i}^{(k)}, \mathbf{y}_{r,i}^{(k)})$ 는 태스크 k에 대한 i번째 경험, 즉 상태 s_i에서 행동 a_i을 실행하였을 때 보상 r_i와 다음 상태 s_i'

Algorithm 1 MB-ELRL($c_p, c_r, d_p, d_r, \lambda, \mu$)

$K \leftarrow 0$
$\mathbf{A}_p \leftarrow \mathbf{zeros}_{c_p \times d_p, c_p \times d_p}$, $\mathbf{A}_r \leftarrow \mathbf{zeros}_{c_r \times d_r, c_r \times d_r}$
$\mathbf{b}_p \leftarrow \mathbf{zeros}_{c_p \times d_p, 1}$, $\mathbf{b}_r \leftarrow \mathbf{zeros}_{c_r \times d_r, 1}$
$\mathbf{L}_p \leftarrow \mathbf{zeros}_{d_p, c_p}$, $\mathbf{L}_r \leftarrow \mathbf{zeros}_{d_r, c_r}$
while isMoreTrainingDataAvailable() **do**
  $(\mathbf{X}_{sa}^{\text{new}}, \mathbf{y}_r^{\text{new}}, \mathbf{y}_{s'}^{\text{new}}, k) \leftarrow$ getNextTrainingData()
  if isNewTask(k) **then**
    $K \leftarrow K + 1$
    $\mathbf{X}_{sa}^{(k)} \leftarrow \mathbf{X}_{sa}^{\text{new}}, \mathbf{y}_r^{(k)} \leftarrow \mathbf{y}_r^{\text{new}}, \mathbf{y}_{s'}^{(k)} \leftarrow \mathbf{y}_{s'}^{\text{new}}$
  else
    $\mathbf{A}_p \leftarrow \mathbf{A}_p - \left(\mathbf{s}_p^{(k)}\mathbf{s}_p^{(k)\top}\right) \otimes \mathbf{D}_p^{(k)}$
    $\mathbf{A}_r \leftarrow \mathbf{A}_r - \left(\mathbf{s}_r^{(k)}\mathbf{s}_r^{(k)\top}\right) \otimes \mathbf{D}_r^{(k)}$
    $\mathbf{b}_p \leftarrow \mathbf{b}_p - \text{vec}\left(\mathbf{s}_p^{(k)\top} \otimes \left(\boldsymbol{\theta}_p^{(k)\top}\mathbf{D}_p^{(k)}\right)\right)$
    $\mathbf{b}_r \leftarrow \mathbf{b}_r - \text{vec}\left(\mathbf{s}_r^{(k)\top} \otimes \left(\boldsymbol{\theta}_r^{(k)\top}\mathbf{D}_r^{(k)}\right)\right)$
    $\mathbf{X}_{sa}^{(k)} \leftarrow \left[\mathbf{X}_{sa}^{(k)}\ ;\ \mathbf{X}_{sa}^{\text{new}}\right]$
    $\mathbf{y}_r^{(k)} \leftarrow \left[\mathbf{y}_r^{(k)}\ ;\ \mathbf{y}_r^{\text{new}}\right], \mathbf{y}_{s'}^{(k)} \leftarrow \left[\mathbf{y}_{s'}^{(k)}\ ;\ \mathbf{y}_{s'}^{\text{new}}\right]$
  end if
  $(\boldsymbol{\theta}_p^{(k)}, \mathbf{D}_p^{(k)}) \leftarrow$ singleTaskTransitionLearner($\mathbf{X}_{sa}^{(k)}, \mathbf{y}_{s'}^{(k)}$)
  $(\boldsymbol{\theta}_r^{(k)}, \mathbf{D}_r^{(k)}) \leftarrow$ singleTaskRewardLearner($\mathbf{X}_{sa}^{(k)}, \mathbf{y}_r^{(k)}$)
  $(\mathbf{L}_p, \mathbf{L}_r) \leftarrow$ reinitializeAllZeroColumns($\mathbf{L}_p, \mathbf{L}_r$)
  $\mathbf{s}_p^{(k)} \leftarrow \arg\min_{\mathbf{s}_p^{(k)}} \|\boldsymbol{\theta}_p^{(k)} - \mathbf{L}_p\mathbf{s}_p^{(k)}\|_{\mathbf{D}_p^{(k)}}^2 + \eta_p^{(k)}\|\mathbf{s}_p^{(k)}\|_1$
  $\mathbf{s}_r^{(k)} \leftarrow \arg\min_{\mathbf{s}_r^{(k)}} \|\boldsymbol{\theta}_r^{(k)} - \mathbf{L}_r\mathbf{s}_r^{(k)}\|_{\mathbf{D}_r^{(k)}}^2 + \eta_r^{(k)}\|\mathbf{s}_r^{(k)}\|_1$
  $\mathbf{A}_p \leftarrow \mathbf{A}_p + \left(\mathbf{s}_p^{(k)}\mathbf{s}_p^{(k)\top}\right) \otimes \mathbf{D}_p^{(k)}$
  $\mathbf{A}_r \leftarrow \mathbf{A}_r + \left(\mathbf{s}_r^{(k)}\mathbf{s}_r^{(k)\top}\right) \otimes \mathbf{D}_r^{(k)}$
  $\mathbf{b}_p \leftarrow \mathbf{b}_p + \text{vec}\left(\mathbf{s}_p^{(k)\top} \otimes \left(\boldsymbol{\theta}_p^{(k)\top}\mathbf{D}_p^{(k)}\right)\right)$
  $\mathbf{b}_r \leftarrow \mathbf{b}_r + \text{vec}\left(\mathbf{s}_r^{(k)\top} \otimes \left(\boldsymbol{\theta}_r^{(k)\top}\mathbf{D}_r^{(k)}\right)\right)$
  $\mathbf{L}_p \leftarrow \text{mat}\left(\left(\frac{1}{K}\mathbf{A}_p + \lambda\mathbf{I}_{c_p \times d_p, c_p \times d_p}\right)^{-1}\frac{1}{K}\mathbf{b}_\mathbf{p}\right)$
  $\mathbf{L}_r \leftarrow \text{mat}\left(\left(\frac{1}{K}\mathbf{A}_r + \lambda\mathbf{I}_{c_r \times d_r, c_r \times d_r}\right)^{-1}\frac{1}{K}\mathbf{b}_\mathbf{r}\right)$
end while

를 관측했다는 데이터를 나타낸다. 유한 MDP에서 $\mathbf{x}_{sa,i}$ 및 $\mathbf{y}_{s',i}$는 각각 $|\mathcal{S} \times \mathcal{A}|$ 크기 및 $|\mathcal{S}|$ 크기의 one-hot 벡터로, $\mathbf{y}_{r,i}$는 실수로 표현한다. $\mathcal{L} = \mathcal{L}_p + \mathcal{L}_r$은 환경 모델 학습 시에 사용하는 손실 함수이며, 본 논문에서는 유한 MDP에 대한 일반적인 선택을 따라 $\boldsymbol{\theta}_p$에 대한 교차 엔트로피 손실 함수 $\mathcal{L}_p$와 θ_r에 대한 L_2 손실 함수 $\mathcal{L}_r$의 합으로 사용하였다. 모델의 일반화 능력을 위해 추가된 정규화 항들을 포함한 등식 (1)를 최소화하게 하면 우리는 최적 모델 매개변수 $\mathbf{L}$ 및 모든 k에 대한 $\mathbf{s}^{(k)}$들을 얻을 수 있다.

하지만 해당 목적식은 모든 태스크의 데이터를 동시에 필요로 하기 때문에 매우 비효율적일 수 있으며, 평생학습 문제에서 고려하고 있는 실시간 학습 방식에 적합하지 않다. 따라서 등식 (1)의 계산 비효율을 해결하기 위해 ELLA [6]에서와 마찬가지로 $\boldsymbol{\theta}^{(k)*} = \arg\min_{(\boldsymbol{\theta}_p, \boldsymbol{\theta}_r)} \frac{1}{n_k}\sum_{i=1}^{n_k} \mathcal{L}(\mathbf{x}_{sa,i}^{(k)}, \mathbf{y}_{s',i}^{(k)}, \mathbf{y}_{r,i}^{(k)}; \boldsymbol{\theta}_p, \boldsymbol{\theta}_r)$ 근방에서 $\frac{1}{n_k}\sum_{i=1}^{n_k} \mathcal{L}(\mathbf{x}_{sa,i}^{(k)}, \mathbf{y}_{s',i}^{(k)}, \mathbf{y}_{r,i}^{(k)}; \mathbf{L}_p\mathbf{s}_p^{(k)}, \mathbf{L}_r\mathbf{s}_r^{(k)})$의 2차 테일러 전개를 취하여 목적식을 근사한다. 2차 테일러 전개의 결과를 식 (1)에 대입하면 다음과 같은 목적식을 얻는다:

$$g(\mathbf{L}) = \frac{1}{K}\sum_{k=1}^{K} \min_{\mathbf{s}^{(k)}} \left\{ \frac{1}{n_k}\|\boldsymbol{\theta}_p^{(k)} - \mathbf{L}_p\mathbf{s}_p^{(k)}\|_{\mathbf{D}_p^{(k)}}^2 + \frac{1}{n_k}\|\boldsymbol{\theta}_r^{(k)} - \mathbf{L}_r\mathbf{s}_r^{(k)}\|_{\mathbf{D}_r^{(k)}}^2 + \eta_p^{(k)}\|\mathbf{s}_p^{(k)}\|_1 + \eta_r^{(k)}\|\mathbf{s}_r^{(k)}\|_1 \right\} + \lambda_p\|\mathbf{L}_p\|_F^2 + \lambda_r\|\mathbf{L}_r\|_F^2 \quad (2)$$

where

$$\mathbf{D}_p^{(k)} = \frac{1}{2}\nabla^2_{\boldsymbol{\theta}_p,\boldsymbol{\theta}_p}\frac{1}{n_t}\sum_{i=1}^{n_k}\mathcal{L}_p(\mathbf{x}_{sa,i}^{(k)},\mathbf{y}_{s',i}^{(k)};\boldsymbol{\theta}_p)\Big|_{\boldsymbol{\theta}_p=\boldsymbol{\theta}_p^{(k)*}}$$

$$\mathbf{D}_r^{(k)} = \frac{1}{2}\nabla^2_{\boldsymbol{\theta}_r,\boldsymbol{\theta}_r}\frac{1}{n_t}\sum_{i=1}^{n_k}\mathcal{L}_r(\mathbf{x}_{sa,i}^{(k)},\mathbf{y}_{r,i}^{(k)};\boldsymbol{\theta}_r)\Big|_{\boldsymbol{\theta}_r=\boldsymbol{\theta}_r^{(k)*}}$$

여기서 $\|\mathbf{v}\|^2_{\mathbf{D}} = \mathbf{v}^\top\mathbf{D}\mathbf{v}$. 식 (2)는 각 태스크마다 데이터셋 전체에 대한 내부 합이 제거되어, 테스크가 추가될 때마다 모든 데이터셋에 대해 재계산할 필요가 없어 식 (1)에 비해 훨씬 효율적이다. 이와 함께, 새로운 태스크 k 가 추가될 때마다 $\{\mathbf{s}^{(j)}\}_{j=1}^{k-1}$ 전체를 재계산하는 대신, 고정된 $\{\mathbf{s}^{(j)}\}_{j=1}^{k-1}$에 대하여 $\mathbf{L}$ 및 $\mathbf{s}^{(k)}$만을 점진적으로 업데이트하여 효율을 더욱 극대화 한다. 사용한 최종 알고리즘의 의사 코드는 Algorithm 1에 나타내었다.

4 실험

제안하는 알고리즘을 유한 상태 및 행동 공간을 갖는 두가지 문제에 대하여 비교하였다. 첫번째로 Chain 도메인 [8]은 5개의 상태와 2개의 행동을 가지며, 다양한 미끄러질 확률 (0.1 ∼ 0.2) 을 갖는 다중-태스크에 대하여 성능을 평가하였다. 두 번째 실험은 RandomMDP 도메인에서 이루어졌으며, Chain 도메인과 동일하게 5개의 상태와 2개의 행동을 가진다. 각 상태 및 행동에 대한 전이 함수들은 디리클레 분포를 따르며 이 때 1번 행동을 할 경우 자기 자신으로의 전이가 평균적으로 높은 확률을 가지도록, 2번 행동을 할 경우 다음 상태로의 전이가 평균적으로 높은 확률을 가지는 디리클레 분포를 사용하였다. 각 상태 및 행동에 대한 보상 함수들은 표준 정규분포를 따른다. 이와 같은 도메인은 도메인마다 공유되는 정보가 뚜렷하여 다중-태스크 학습론의 실험 도메인으로 적합하다.

그림 1은 30개의 태스크에 대해서 제안한 MB-ELRL과 각 태스크마다 독립적으로 학습을 진행한 Independent MLE의 정규화 성능의 평균값을 나타낸 그래프이다. 그림 1에서의 가로축은 학습에 사용한 각 태스크의 데이터의 개수를 나타내며, 학습에 사용한 각 태스크의 데이터는 균등한 행동 정책을 통해 수집하였다. 수집된 데이터를 이용해 각 태스크의 전이, 보상 함수를 추정하고 추정된 모델에 대한 최적 행동을 계산하여 정규화 성능을 비교하였다. 정규화 성능은 각 태스크에서 에이전트의 최적 정책의 성능을 1, 균등 행동 정책의 성능을 0으로 하였을 때 가지는 상대적 성능치를 뜻한다. Chain 도메인 실험의 결과를 보면, 데이터 수의 증가에 따라 제안한 MB-ELRL의 성능이 Ind.MLE 보다 항상 좋은 성능을 보인다는 것을 알 수 있다. 또한 RandomMDP의 결과에서도, 제안하는 방법론 MB-ELRL이 Ind.MLE에 비해 현저하게 높은 정규화 성능을 보여준다. 특히, 각 태스크마다 50개의 학습 데이터양만으로도 모든 태스크에 대해 MB-ELRL은 평균적으로 높은 성능을 보이지만, Ind.MLE는 매우 낮은 성능을 보인다. 이는 태스크 간 정보 공유를 통해 MB-ELRL이 효율적으로 학습할 수 있었음을 의미한다.

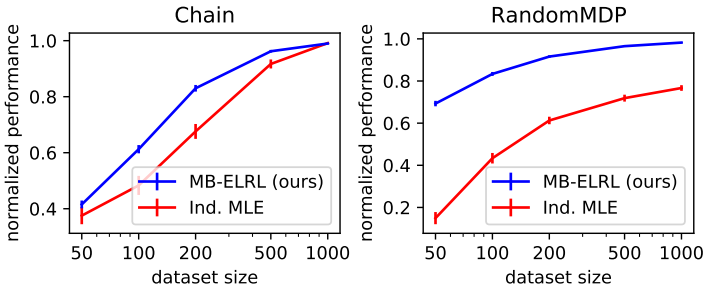

그림 1: 100번 반복 실험 결과. 에러바는 2 × (표준 오차)임.

5 결론

본 논문에서는 효율적인 평생학습 알고리즘의 모델기반 강화학습 적용을 위한 MB-ELRL 알고리즘을 제안한다. 제안하는 방법론은 모델기반 강화학습에서 다양한 태스크들 사이의 공유 가능한 정보를 효율적으로 학습하며, 실험 결과에서는 각 태스크를 독립적으로 학습하는 것에 비해 성능이 향상됨을 보여준다. 향후 연구 방향으로는 로봇 제어 시뮬레이션 환경과 같은 보다 복잡한 문제들에 대하여 제안하는 방법론의 확장 및 추가 실험 등을 고려할 수 있다.

참고 문헌

[1] S. Thrun and T. M. Mitchell, "Lifelong robot learning," *Robotics and autonomous systems*, vol. 15, no. 1-2, 1995.

[2] D. Hassabis, D. Kumaran, C. Summerfield, and M. Botvinick, "Neuroscience-inspired artificial intelligence," *Neuron*, vol. 95, no. 2, 2017.

[3] G. I. Parisi, R. Kemker, J. L. Part, C. Kanan, and S. Wermter, "Continual lifelong learning with neural networks: A review," *Neural Networks*, vol. 113, 2019.

[4] M. McCloskey and N. J. Cohen, "Catastrophic interference in connectionist networks: The sequential learning problem," in *Psychology of learning and motivation*, vol. 24, Elsevier, 1989.

[5] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, 2015.

[6] P. Ruvolo and E. Eaton, "ELLA: An efficient lifelong learning algorithm," vol. 28 of *Proceedings of Machine Learning Research*, (Atlanta, Georgia, USA), PMLR, 17–19 Jun 2013.

[7] R. S. Sutton, A. G. Barto, *et al.*, *Introduction to reinforcement learning*, vol. 135. MIT press Cambridge, 1998.

[8] R. Dearden, N. Friedman, and S. Russell, "Bayesian q-learning," in *AAAI/IAAI*, 1998.