On the Feasibility of Cross-Task Transfer with Model-Based Reinforcement Learning

ICLR 2023

강화학습특론 Paper review 발표
10조: 강용훈 김산 허찬순
2023년 10월 24일 화요일

목차

Table of Contents

Introduction

Before this work,

- Sample Efficiency Problem!
  - Pretraining을 통해 극복할 수 있다
- RL을 위한 Pretraining framework
  - Same-task initialization + Model-free policies from an offline dataset
  - Novel instance를 이용해서 target task에 맞게 finetuning

⇒ **왜 Cross-task + Model-based + Pretraining framework는 없을까?**

Introduction

Obstacles are,

⇒ **왜 Cross-task + Model-based + Pretraining framework는 없을까?**

- **High-variance objectives**
  - Cross task에 걸친 목표가 다양해서 학습이 잘 이루어지지 못한다
- **Catastrophic forgetting**
  - 알고리즘이 학습한 World Model을 여러 task에 대한 학습을 거치는 동안 잊어버린다

Pretraining + Finetuning framework for RL이 필요하다

Introduction

In this work,

- Pretraining + Finetuning framework for RL in settings of,
  - **Cross-task** : 다양한 task에 따른 high-variance objectives 상황에서도 공통된 visual cue를 학습
  - **Model-based RL** : 최근 좋은 성능을 보여주는 RL 코드들처럼 model-based
  - **Online-RL setting** : catastrophic forgetting을 방지
  - **High dimensional input** : image처럼 input dimension이 복잡

⇒ **How and When can model-based RL be pretrained on diverse set of distinct tasks?**

Backgrounds
연구의 배경

- **Model-based algorithm**
  - Low-dimension에서는 sample efficient
  - Image input에서는 sample efficiency가 급격히 떨어져, 학습데이터가 충분하지 않은 경우 model-free 알고리즘보다도 성능이 떨어짐
- **MuZero**
  - Image input에서도 super-human performance
  - 데이터는 여전히 많이 필요.

Backgrounds

EfficientZero와 MuZero Reanalysis

- **Issues in data-limited settings**
  - 환경이 갖고 있는 불확실성
  - 환경에 대한 정보의 부족
  - Off-policy issues of multi-step value

- **EfficientZero (NeurIPS 2021)**
  - Self-Supervised Consistency Loss
  - End-to-end prediction of the Value Prefix
  - Model-Based Off-Policy Correction
- **MuZero Reanalysis (NeurIPS 2021)**
  - On-offline RL을 통합해 learning efficiency 제고

Backgrounds

환경: Arcade Learning Environment

- **연구의 목적-환경 fit**
  - High dimensional input(image) 사용 가능
  - 환경과 목적, task가 굉장히 다양한 환경
  - 인간은 공통된 visual cue를 학습에 이용
  - Model-based algorithm이 많이 보고됨

- **알려진 Model-based Algorithms**
  - MuZero
  - MuZero Reanalysis
  - EfficientZero

Proposed Method

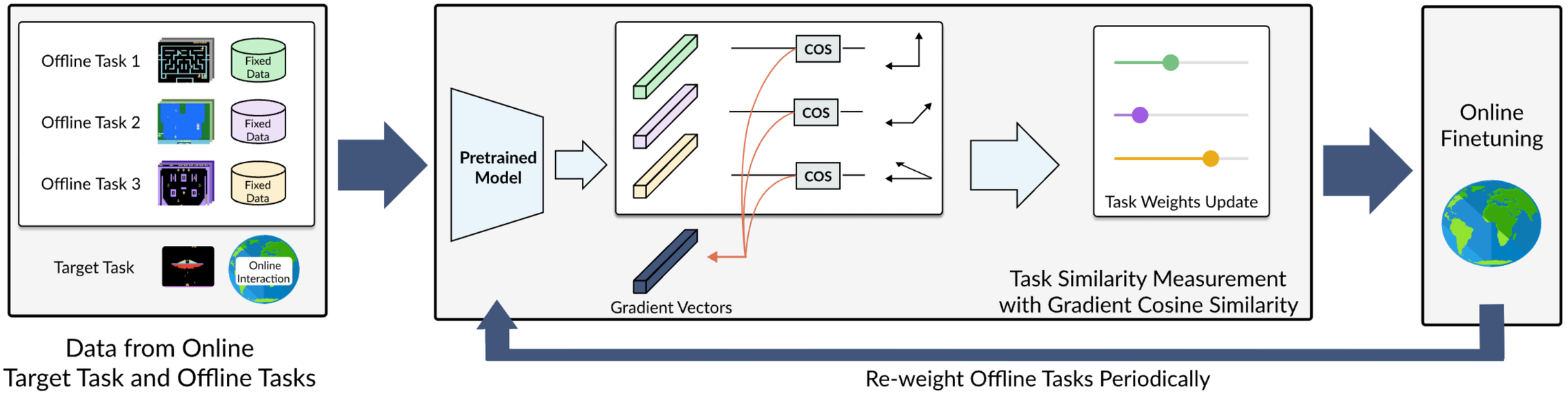
주요 아이디어

- **Model-Based Pretrain-Finetuning 구현**
  - Multi-task Pretraining → Distillation → Target-task에 대해 Finetuning
- **Catastrophic Forgetting 극복**
  - Target task에 대한 Finetuning 과정에서 사전학습 데이터도 활용
  - Target-Pretrain 간의 task 유사도를 기반으로 Gradient Reweighting

⇒ **XTRA**: model-based Cross-Task tRAnsfer Learning

Proposed Method

전체 다이어그램



Stage 1

Offline에서 다양한 task 학습
→ 이후 distillation

Transfer

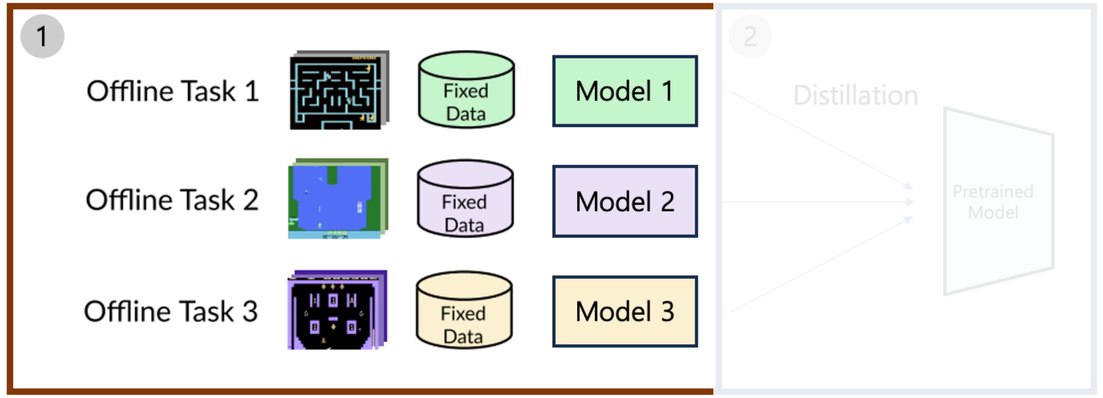
Stage 2

Online에서 model-based finetuning
(과정에서 reweighting)

XTRA: model-based Cross-Task tRAnsfer Learning

Method Details

Stage1.1: Offline Multi-Task Pretraining



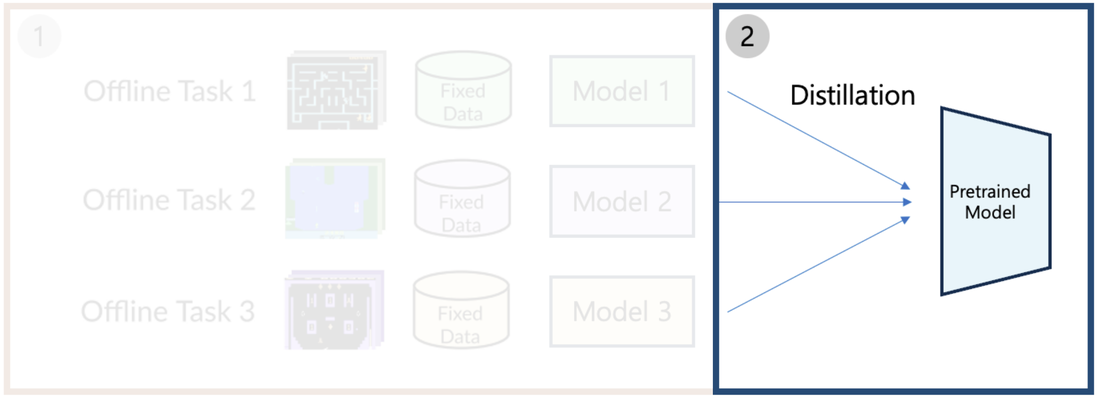
- **다양한 Offline Task에 대해 사전학습 진행**
  - **목표** : Unseen Task에 대한 Good Initialization (General model을 찾는게 아님!)
  - **어려움** :
    - 다양한 Task를 하나의 모델에 학습시켜야함
    - 다양한 Task에 걸친 데이터셋을 offline으로 학습하니 extrapolation error 커짐

⇒ **student-teacher 세팅(Distillation) 사용**

- 이 세팅이 pretraining task의 갯수를 늘릴 수 있는 유의미한 방법임을 경험적으로 확인

Method Details

Stage1.2: Distillation



1. Train individual EfficientZero teacher on each dataset of diverse tasks
   (이때 모델의 Prediction (Policy, Value) 값도 함께 저장)

2. Teacher model prediction에서 p, v,
   Environment reward에서 u를 샘플링해서 → (p, v, u) 확보

3. Teacher model의 (p, v, u)로 student model의 quantity target (π, z, u) 학습

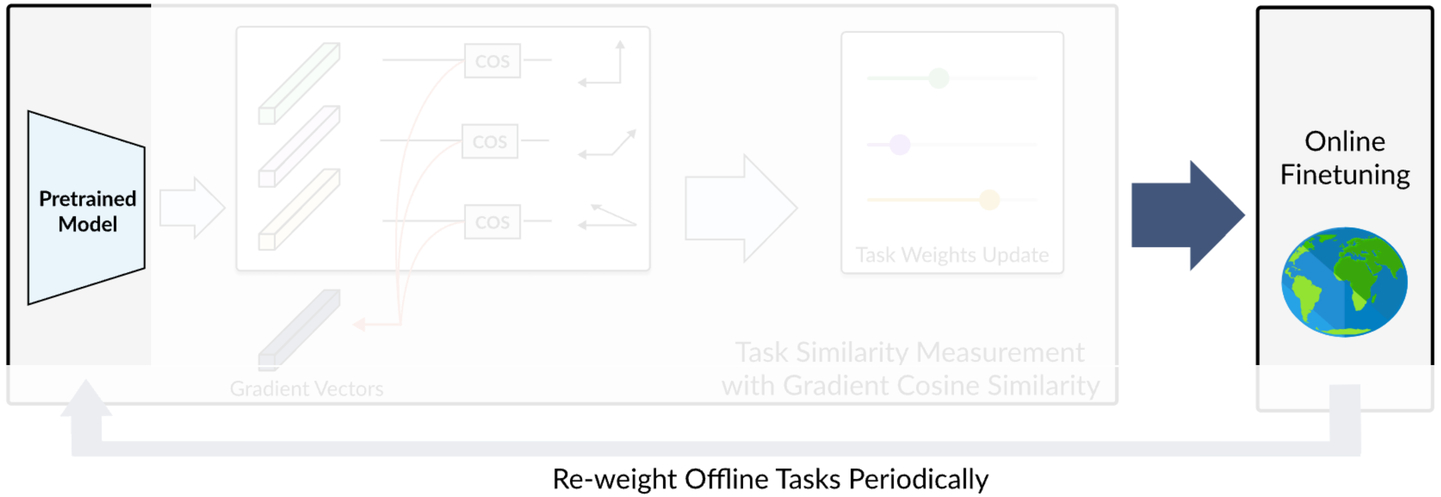
4. Multi-tasking single agent인 student 모델을 얻음

- MuZero Reanalysis에서는 teacher model을 사용하지 않아서 scalability가 떨어졌었음

⇒ 이제 pretrained model이 online interaction을 통해 finetuning 될 준비 완료!

Method Details

Stage2.0: Online Finetuning



- Online Fine-Tuning
  - Target-Task와의 Interaction을 통해 데이터 수집 및 Distillation된 모델 Finetuning
- 그러나 Target-Task 만으로 Finetuning → Catastrophic Forgetting 문제 발생
  - ∵ Pretrained ↔ Target 사이에서 task의 차이가 크면 perturbation이 심해지고 결국 finetuning 과정에서 pretrain 되었던 inductive priors를 제거

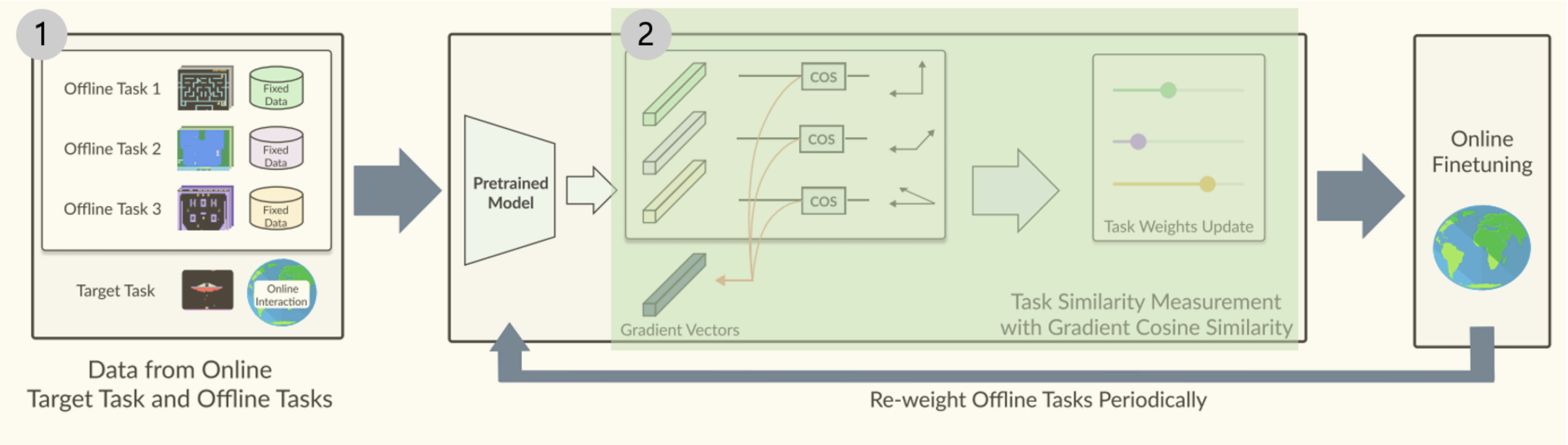
- **이러한 문제를 막기 위해, Concurrent Cross-Task Learning를 제안함**

Method Details

Stage2.0: Online Finetuning

- **Concurrent Cross-Task Learning**

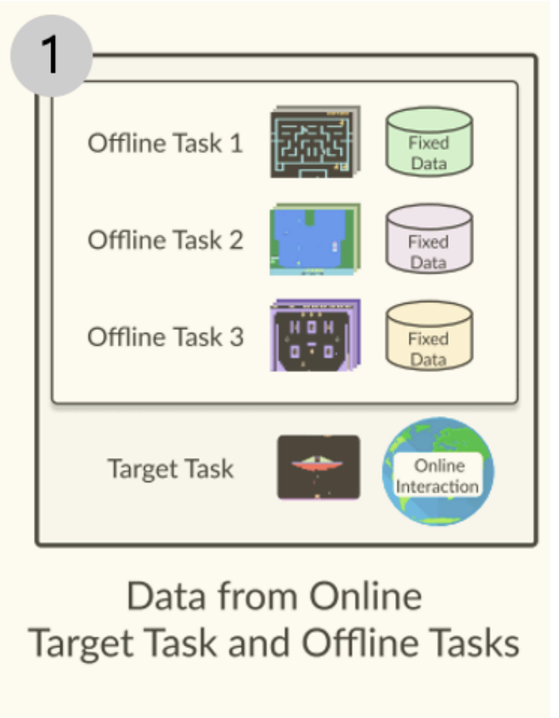
1. Cross-Task Learning
2. Reweighting Gradientby η



Method Details

Stage2.1: Cross-Task Learning

- **Cross Task Learning**
  - Finetuning 학습을 진행할때
    - Finetuning과정에서 얻은 Online Interaction 데이터와
    - Pretraining에 사용된 Offline 데이터를
  - 함께 사용



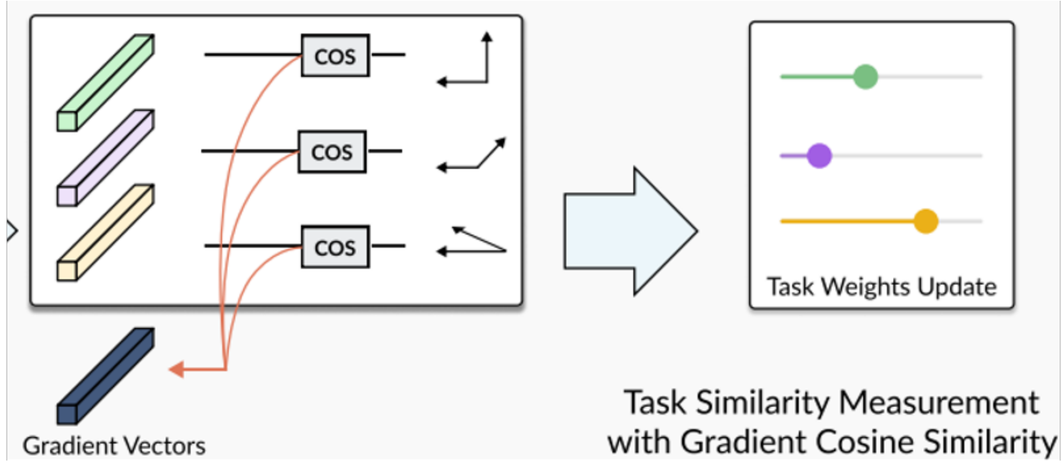
Target-Task

Offline-Task

$$\mathcal{L}_t^{\text{adapt}}(\theta) = \mathcal{L}_t^{\text{ez}}(\mathcal{M}) + \sum_{i=1}^{m} \eta^i \mathcal{L}_t^{\text{ez}}(\tilde{\mathcal{M}}^i)$$

Method Details

Stage2.2: Reweighting Gradient

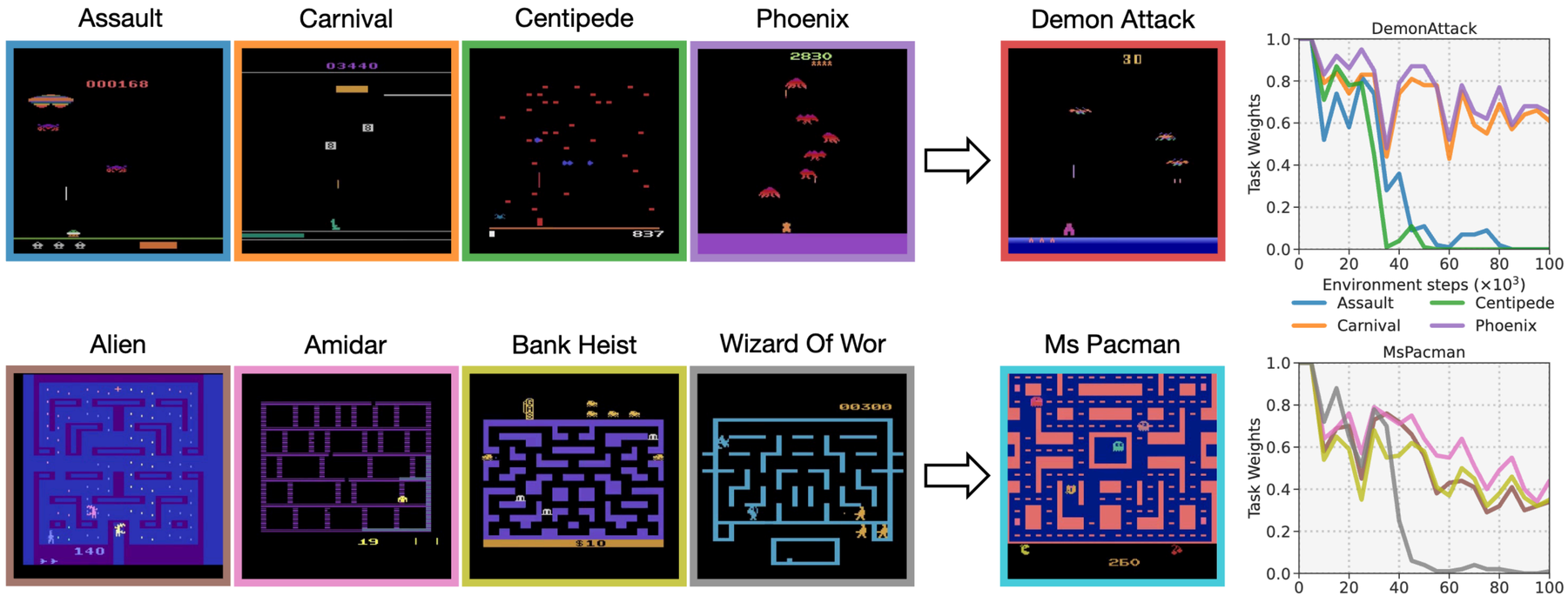


- **Reweighting Gradient by η**
  - Offline Task의 데이터가 Target-Task 학습속도를 떨어뜨려서 Sample efficiency를 저해할 수 있음
  - Offline-Task와 Target-Task와의 Gradient Similarity로 Gradient를 Weighting!

$$\mathrm{Sim}(\tilde{\mathcal{M}}^i, \mathcal{M}) = \frac{\tilde{\mathcal{G}}_n^i \cdot \mathcal{G}_n}{\|\tilde{\mathcal{G}}_n^i\| \|\mathcal{G}_n\|}$$

Method Details

Stage2: Visualization of Concurrent Cross-Task Learning



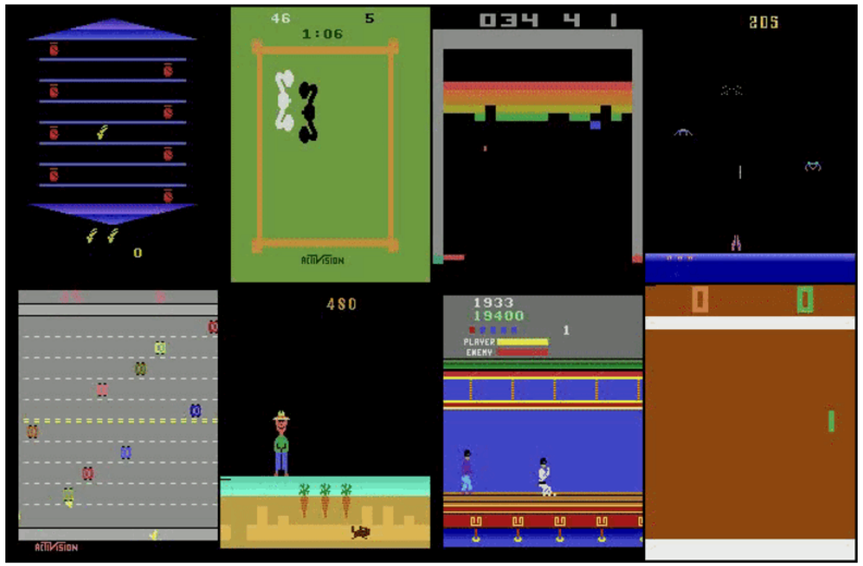
(a) Cross-Task Transfer from 4 offline games (left) to 1 target game (right). *(b)* Task weights.

- 학습을 진행하면서 비슷한 pretraining task에 집중 ↑

Experiments

Settings

- **Dataset**
  - Atari100k (14개의 게임 사용)
- **Base Architecture**
  - EfficientZero
- **Baselines**
  - EfficientZero (NeurIPS 2021) : Model-Based RL
  - CURL (ICML 2020) : Model-Free RL



Experiments

Similar Pretraining Result

- 비슷한 Task로 Pretraining하면 **XTRA가 EfficientZero 보다 성능이 개선됨**
- Shooter와 Maze에서 평균 1.36배, 1.23배 성능 향상을 보임

Table 1. **Atari100k benchmark results (*similar* pretraining tasks).** Methods are evaluated at 100k environment steps. For each game, XTRA is first pretrained on all 4 other games from the same category. Our main result is highlighted. We also include three ablations that remove *(i)* cross-task optimization in finetuning (only online RL), *(ii)* the pretraining stage (random initialization), and *(iii)* task re-weighting (constant weights of 1). We also include zero-shot performance of our method for target tasks in comparison to behavioral cloning. Mean of 5 seeds and 32 evaluation episodes.

Category	Game	BC (finetuned)	Efficient Zero	Efficient Zero-L	XTRA (Ours)	Ablations (XTRA)			Zero-Shot	
						w.o. cross-task	w.o. pretraining	w.o. task weights	BC	XTRA (Ours)
Shooter	Assault	838.4	1027.1	1041.6	**1294.6**	1246.4	1257.5	1164.2	0.0	92.8
	Carnival	1952.4	3022.1	2784.3	**3860.9**	3544.4	2370.0	3071.6	93.75	719.3
	Centipede	1814.1	3322.7	2750.7	5681.4	3833.2	**6322.7**	5484.1	162.2	1206.8
	Demon Attack	825.5	11523.0	4691.0	14140.9	6381.5	9486.8	**51045.9**	73.8	113.6
	Phoenix	427.6	10954.9	3071.0	14579.8	10797.3	9010.6	**22873.9**	0.0	8073.4
	Mean Improvement	0.42	1.00	0.69	1.36	1.02	1.11	**2.06**	0.02	0.29
	Median Improvement	0.55	1.00	0.83	1.28	1.15	0.82	**1.65**	0.01	0.24
Maze	Alien	152.9	695.0	641.5	**954.8**	722.8	703.6	633.6	108.1	294.1
	Amidar	25.5	109.7	84.2	90.2	**121.8**	70.8	69.7	0.0	5.2
	Bank Heist	178.8	246.1	244.5	**304.9**	280.1	225.1	261.4	0.0	7.3
	Ms Pacman	550.0	1281.4	1172.8	**1459.7**	1011.1	1122.6	809.2	147.6	448.9
	Wizard Of Wor	163.8	1033.1	928.8	985.0	**1246.1**	654.4	263.5	100.0	9.4
	Mean Improvement	0.35	1.00	0.90	**1.11**	1.06	0.82	0.70	0.07	0.17
	Median Improvement	0.23	1.00	0.92	**1.14**	1.11	0.88	0.64	0.10	0.05
Overall	Mean Improvement	0.39	1.00	0.79	1.23	1.04	0.96	**1.38**	0.05	0.23
	Median Improvement	0.33	1.00	0.91	**1.25**	1.12	0.85	1.04	0.02	0.16

Experiments

Diverse Pretraining Result

Table 2. **Atari100k benchmark results (diverse pretraining tasks).** XTRA results use the *same* set of pretrained model parameters obtained by offline pretraining on 8 diverse games. Mean of 5 seeds each with 32 evaluation episodes. Our result is highlighted. All other results are adopted from EfficientZero (Ye et al., 2021). We also report human-normalized mean and median scores.

Game	XTRA (Ours)	EfficientZero	Random	Human	SimPLe	OTRainbow	DrQ	SPR	MuZero	CURL
Assault	**1742.2**	1263.1	222.4	742.0	527.2	351.9	452.4	571.0	500.1	600.6
BattleZone	14631.3	13871.2	2360.0	37187.5	5184.4	4060.6	12954.0	**16651.0**	7687.5	14870.0
Hero	**10631.8**	9315.9	1027.0	30826.4	2656.6	6458.8	3736.3	7019.2	3095.0	6279.3
Krull	**7735.8**	5663.3	1598.0	2665.5	4539.9	3277.9	4018.1	3688.9	4890.8	4229.6
Seaquest	749.5	1100.2	68.4	42054.7	683.3	286.9	301.2	583.1	208.0	384.5
Normed Mean	**1.87**	1.29	0.00	1.00	0.70	0.41	0.62	0.65	0.77	0.75
Normed Median	0.35	0.33	0.00	1.00	0.08	0.18	0.30	**0.41**	0.15	0.36

- Pretraining에서 다양한 게임을 학습하는 상황 (Carnival, Centipede, Phoenix 등 카테고리가 다른 게임들)
- **Diverse Pretraining에서 XTRA가 EfficientZero와 기존 Baseline보다 좋은 성능**을 보임

Experiments

Diverse Pretraining Result

Table 7. **XTRA ablation (number of tasks in pretraining & cross-task finetuning) for tasks that share *diverse* game mechanics**. Results for EfficientZero are adopted from EfficientZero (Ye et al., 2021). All other results are based on the average of 5 runs.

Game	XTRA	Ablations (XTRA)			EfficientZero
	8 Games	4 Games	2 Games	0 Games	0 Games
Assault	**1742.2**	1676.7	1463.8	1255.9	1263.1
BattleZone	**14631.3**	9581.3	9550.0	10125.0	13871.2
Hero	**10631.8**	9654.9	8506.5	6815.1	9315.9
Krull	**7735.8**	7375.6	7348.9	5590.6	5663.3
Seaquest	749.5	656.4	627.5	770.8	**1100.2**
Normed Mean	**1.87**	1.74	1.65	1.23	1.29
Normed Median	**0.35**	0.29	0.25	0.22	0.33

- Pretraining task의 종류에 대한 Ablation study
  - **XTRA setting에서 다양한 Task를 쓰는 것이 성능 향상에 도움이 됨**

Experiments

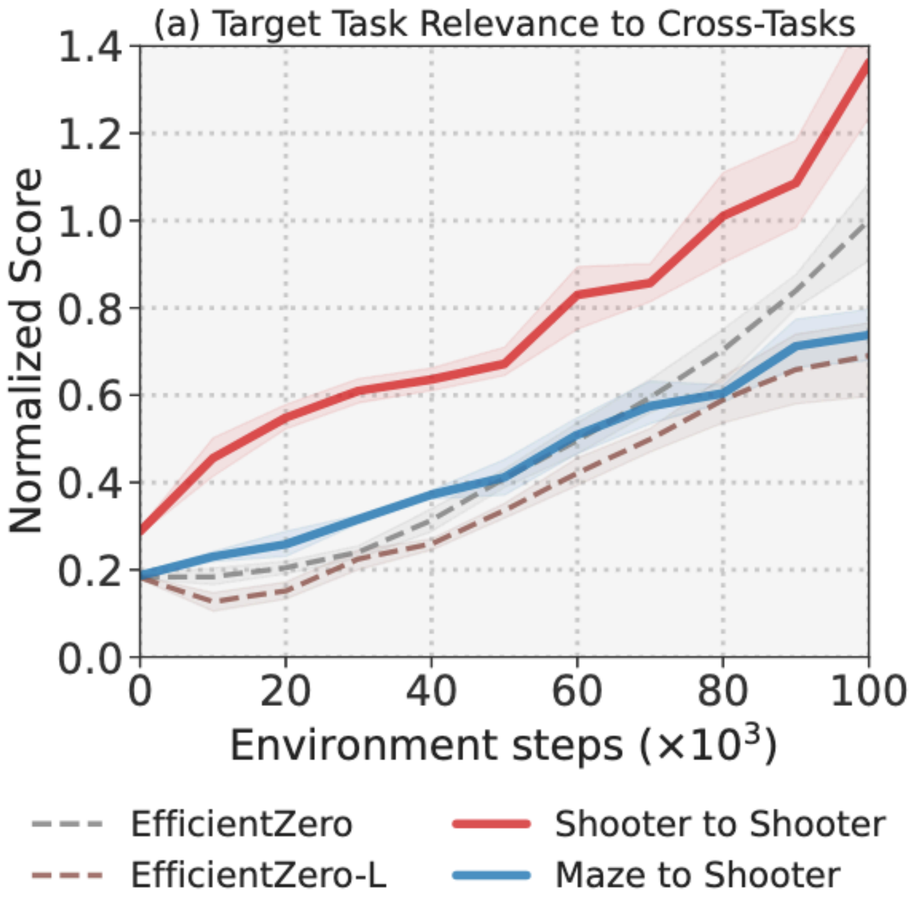
Ablation Studies

Category	Game	Ablations (XTRA)		
		w.o. cross-task	w.o. pretraining	w.o. task weights
Shooter	Assault	1246.4	1257.5	1164.2
	Carnival	3544.4	2370.0	3071.6
	Centipede	3833.2	**6322.7**	5484.1
	Demon Attack	6381.5	9486.8	**51045.9**
	Phoenix	10797.3	9010.6	**22873.9**
	Mean Improvement	1.02	1.11	**2.06**
	Median Improvement	1.15	0.82	**1.65**
Maze	Alien	722.8	703.6	633.6
	Amidar	**121.8**	70.8	69.7
	Bank Heist	280.1	225.1	261.4
	Ms Pacman	1011.1	1122.6	809.2
	Wizard Of Wor	**1246.1**	654.4	263.5
	Mean Improvement	1.06	0.82	0.70
	Median Improvement	1.11	0.88	0.64
Overall	Mean Improvement	1.04	0.96	**1.38**
	Median Improvement	1.12	0.85	1.04

- **Task reweight를 제외시켰을 경우**
  - Shooter에서는 큰 성능 향상
  - Maze에서 큰 성능하락
- 게임의 목표가 비슷하다면 같은 카테고리에서는 Pretraining gradient에 대한 Reweighting이 없어도 pretrain이 도움이 됨
- 그러나 게임의 목표가 완전히 다른 경우 같은 카테고리 더라도 Reweighting을 통해 집중할 게임을 정해주지 않으면 오히려 성능이 떨어짐

Experiments

How do the individual components of framework influence its success?
Task Relevance에 관해,

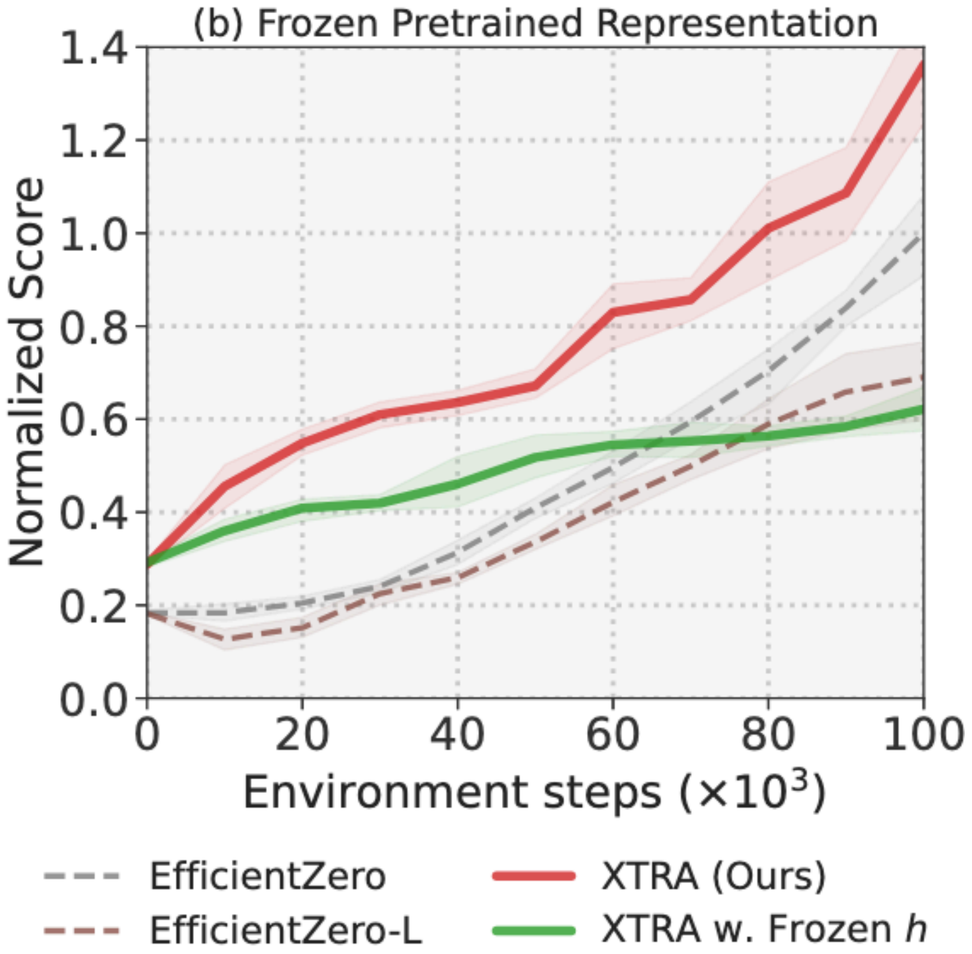


- **Task 관련성이 높은 경우**
  - 비슷한 Task를 Pretraining하여 Finetuning 하면 초기 훈련에 큰 이점을 줌
- **Task 관련성이 낮은 경우**
  - Finetuning이 일반 학습과 비슷하게 이루어짐

∵ Reweight가 작용하여 모든 Pretraining Gradient를 낮추기 때문에 성능 향상이 적음

Experiments

How do the individual components of the framework influence its success?
Pretrained Representation에 관해,

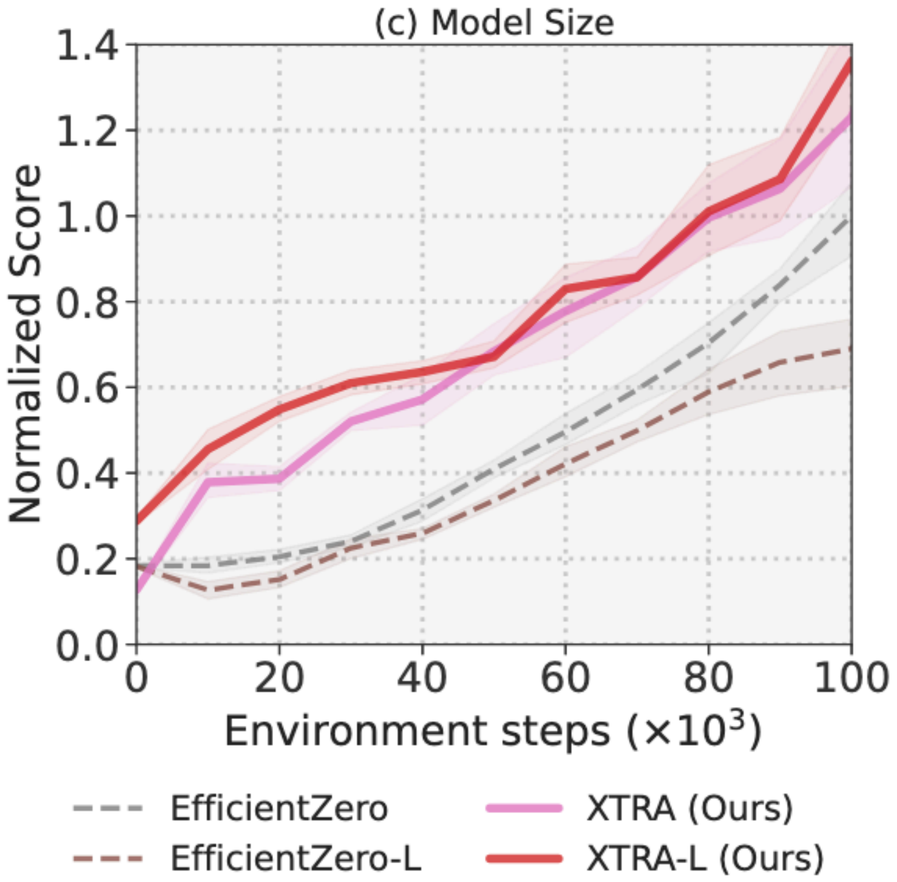


- 고정된 Representation은
  - Fine-Tuning 시, 초기에는 성능 향상
  - 최종적인 성능은 저해

⇒ **고정된 Representation이 모델이 Target-Task에 수렴하는 것을 방해**

Experiments

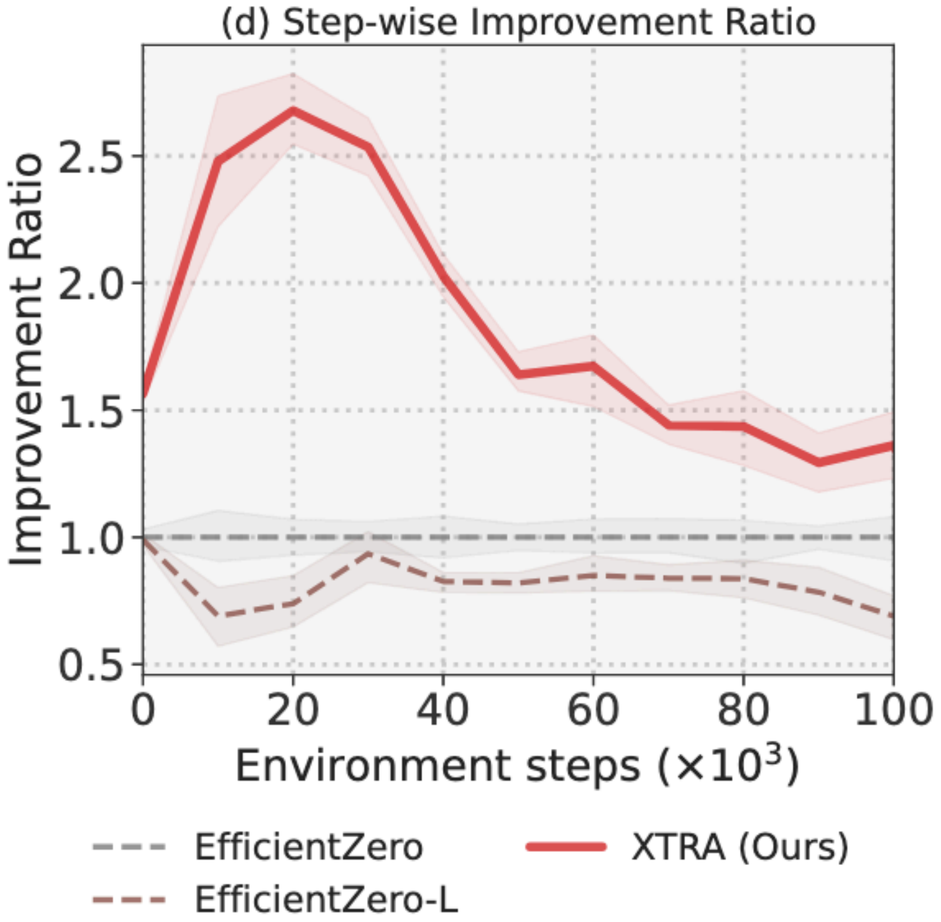
How do the individual components of the framework influence its success?
Model Scale에 관해,



- **모델의 크기**
  - 모델이 커지면 약간의 성능 향상은 일어난다
  - 그러나 시간복잡성이 증가하기 때문에
    - 모델의 크기를 늘리는 것은 효율적이지 않다

Experiments

How do the individual components of the framework influence its success?
Relative Improvement vs. Environment Steps 비교



- **데이터가 적은 상황**
  - XTRA는 경쟁 모델보다 더 좋은 성능 향상을 보인다

⇒ **극단적으로 데이터가 적으면 XTRA가 훨씬 유리하다**

- 물리적 제약으로 환경 상호작용이 어려운 곳에서 효용이 기대됨

Conclusion

요약

- **본 논문에서는 다른 작업의 추가 보조 데이터를 사용하여**
  - 학습된 세계 모델의 확장 가능한 사전 훈련 및 미세 조정 기능을 갖춘
  - Sample Efficient Online RL을 위한 프레임워크인 XTRA를 제안함
- **XTRA가**
  - 대부분의 작업에서 Sample Efficiency를 크게 향상시킴에 따라
  - EfficientZero의 Avg, Median 성능을 각각 23%, 25% 향상시킴

Conclusion

Research Question에 대한 대답

Cross-task + Model-based + Pretraining framework

- Student-Teacher Distillation ⇒ multi-task 문제 극복
- Offline Pretrain + Online Finetuning ⇒ Model-based Transfer Learning 구현

⇒ How and When can model-based RL be pretrained on multi-task setting?

- Student-Teacher Distillation Framework가 Sample efficiency를 높이는 효과 확인
- Task 관련성이 높을 때 Offline Pretraining이 유의미한 효과를 보였음
  - Offline dataset의 Task 관련성을 알 수 없을 때는 Representation의 Cosine Similarity Reweighting을 적용

Review Comments

Pros and Cons

- **Pros**
  - Pretraining을 RL에 적용하여 데이터 효율성을 높임
  - Distillation을 통해서 multi-task RL agent의 pretraining 과정에서 발생하는 문제들을 우회할수 있음을 제시
- **Cons**
  - 비교 데이터셋이 Atari100k 밖에 없음
  - 교사모델 학습 + 학생모델로 distillation 학생모델의 파인튜닝까지 하는 과정에서 컴퓨팅 파워가 많이 필요

프로젝트 활용

Multi-task RL with Contrastive representation learning

- **프로젝트 계획**
  - Multi-task Single-agent RL을 만들기 위해
    - Continual Learning 과
    - Contrastive Representation Learning 방법을 적용
- **리뷰한 논문과의 연관성**
  - Pretraining과 우리의 Continual Learning 기반 모델 학습은 비슷한 성격을 가지고 있음
  - Pretraining이 Task Transfer 시 도움이 된다는 것은, 이전의 Task가 다음 Task에게 도움을 줄 수 있음
  - Task간의 representation alignment가 pretrain-finetuning 과정에 영향을 준다는 것을 확인
  - 논문은 Model-Based 였지만, 우리는 Model-Free에서도 시도해보고자 함

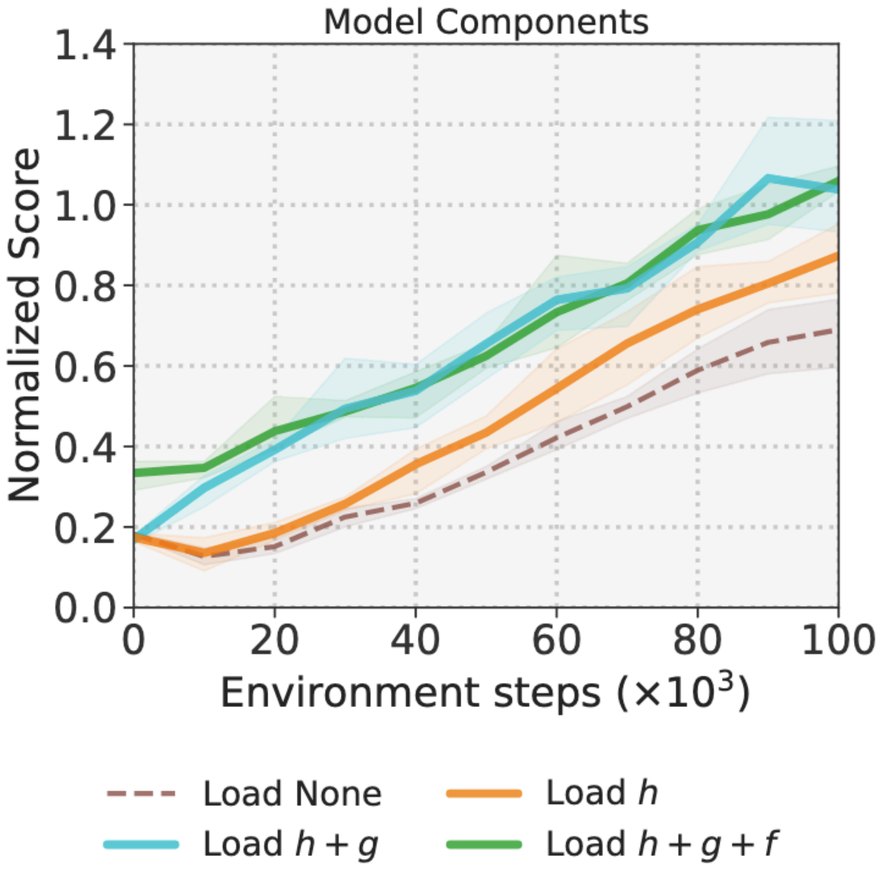
Thanks

On the Feasibility of Cross-Task Transfer with Model-Based Reinforcement Learning 리뷰

강화학습특론 Paper review 발표
10조: 강용훈 김산 허찬순
2023년 10월 24일 화요일

Experiments

How do the individual components of framework influence its success?
Task Relevance에 관해,



- Representation Network h
- Dynamics Network g
- Prediction Network f

- h, g, f 중 어떤 것이 큰 효과를 주는지 파악하기 위해,
  사용하지 않는 parameter는 초기화하고 Fine-Tuning 진행
- Representation Network만 사용하더라도
  사용하지 않은 것보다 성능 향상이 있음
- 그러나 prediction network는 큰 도움을 주지 않음을 확인
- **좋은 Dynamic network를 학습 하는 것이**
  적은 샘플 상황에서 큰 도움이 됨