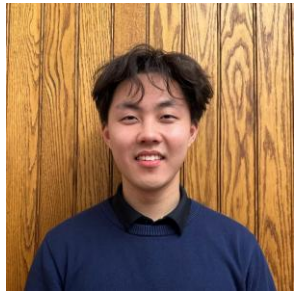
1000 Layer Networks for Self-Supervised RL: Scaling Depth Can Enable New Goal-Reaching Capabilities

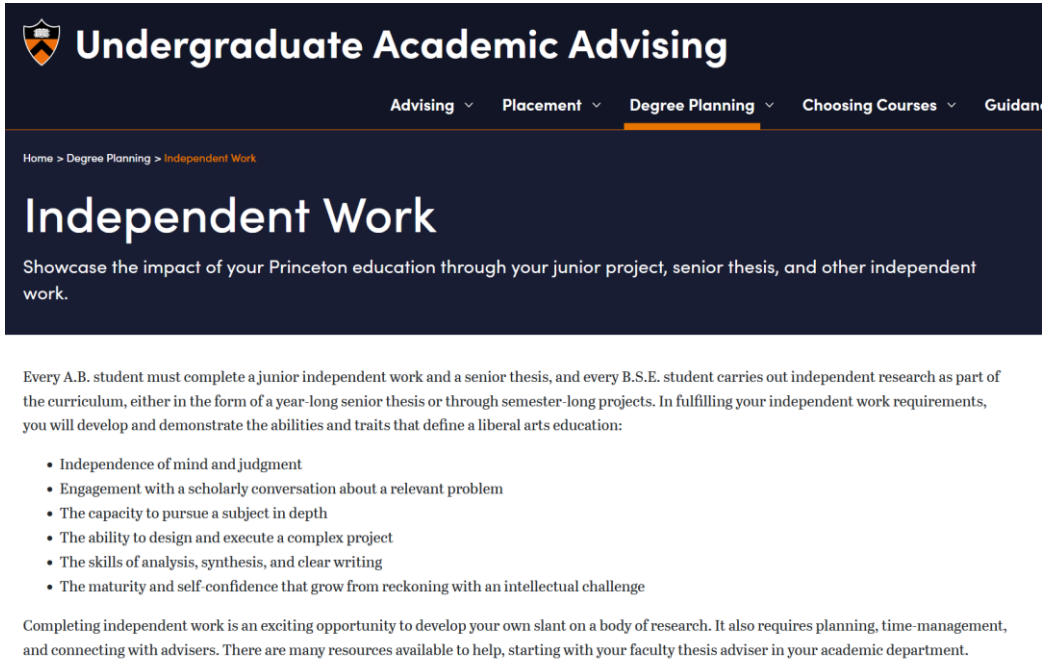
Kevin Wang, Ishaan Javali, Michał Bortkiewicz, Tomasz Trzcinski, Benjamin Eysenbach
39th Conference on Neural Information Processing Systems (NeurIPS 2025 Best Paper)

2026.03.19.
에이전트브레인스토밍 스터디
김동민

https://arxiv.org/abs/2503.14858

천재 학부생 케빈의 강화학습 정복기





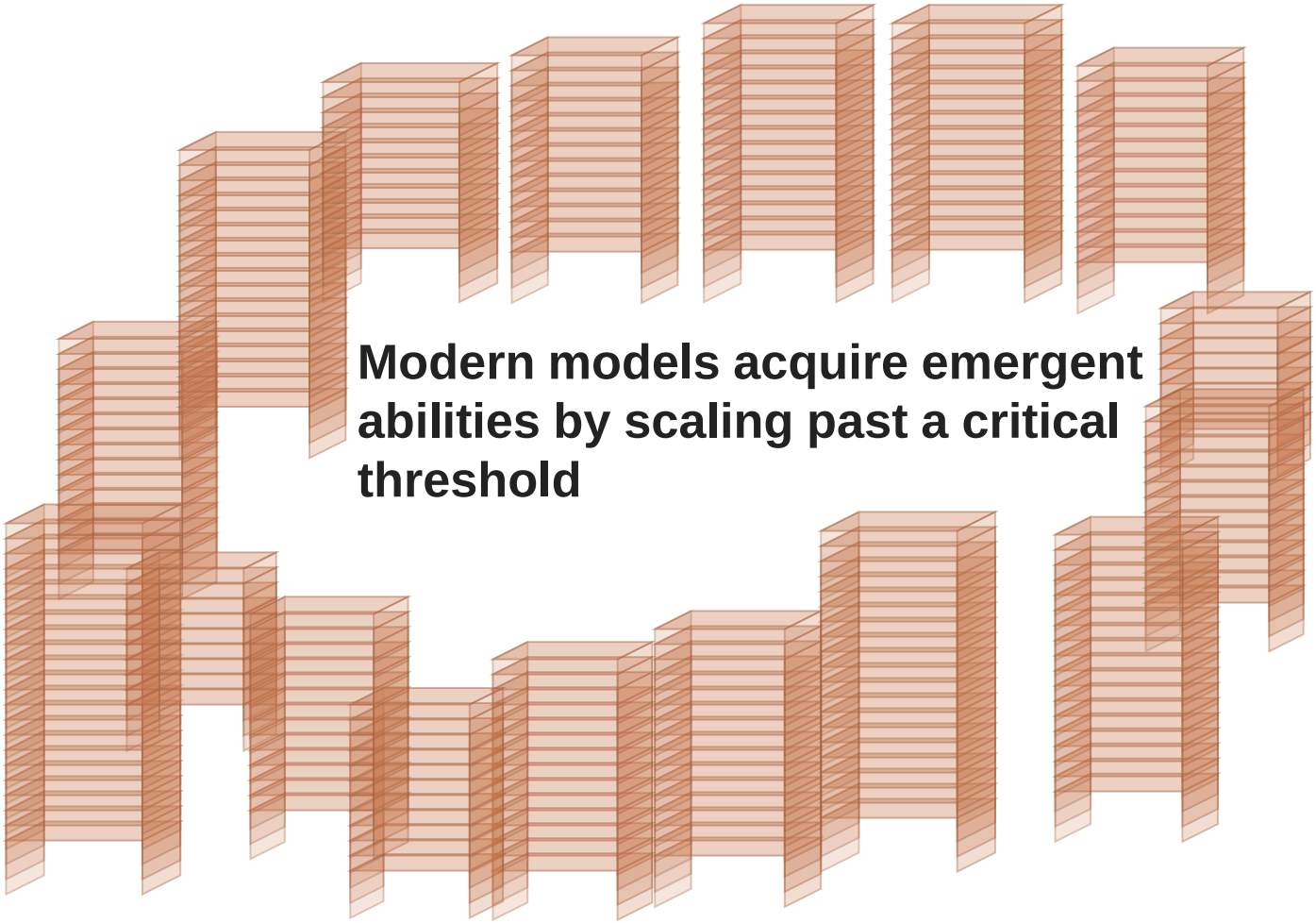
- Kevin Wang
  - 중국계 미국인
  - 2016년 알파고를 보고 인공지능에 대한 관심이 생기기 시작
  - 프린스턴대학교 22학번 (Computer Science)
  - 2025년 5월에 3년만에 조기 졸업
  - 현재 OpenAI에서 reasoning 관련 연구 수행 중
- 대학교 2학년 때 Benjamin Eysenbach의 강화학습 수업을 들으면서 강화학습을 접하게 됨
- Independent Work라는 학부생이 교수와 함께 연구를 수행하는 수업을 들으면서 처음 연구를 경험
  - Eysenbach의 오피스아워마다 찾아가서 많이 배웠다고 함
  - 처음 쓴 논문이 이 논문
  - 그리고 best paper를 받음!

https://youtu.be/Z2DRWEIuAZ4?si=DSeRi46jOQ2zElMQ
https://youtu.be/25FsKN0f8gQ?si=a0rAwZOh07urLytM

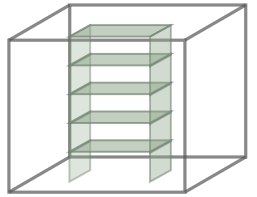
케빈의 의문점

- 강화학습에서는 왜 2개~5개 정도의 작은 레이어만 사용하지?
- 강화학습을 스케일업해보고 싶다

The Scale Era (Vision & Language)



The RL Ceiling



Standard State-Based RL: 2 to 5 Layers

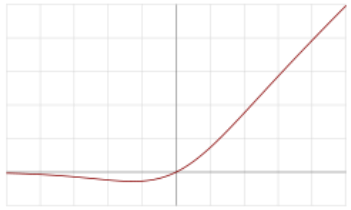
Scaling Issue in RL

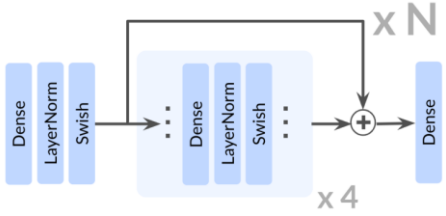
The conventional wisdom states that RL provides too few bits of feedback (sparse rewards) to train deep networks.
Can we achieve emergent phenomena by scaling RL itself?

Sparse Reward | **Q-Regression** | **Limited Batch Size**

케빈의 최초 접근

- 강화학습을 스케일업을 위해 배치사이즈를 늘려봐야겠다(큰 배치사이즈를 처리하려면 큰 네트워크가 필요하니)
- → Eyenbach에게 이런 연구를 하겠다고 얘기했고 Eyenbach는 잘 안될거 같으니 별로 관심을 안준듯 하다. 대신 컴퓨팅자원은 쓸 수 있게 해줌(H100을 제공함. 아마도 1대인듯)
- 기존 연구를 찾아보니 네트워크의 depth가 아니라 width를 늘린 연구는 있었다고 한다
- depth를 늘리려면 residual connection을 쓰는게 좋을 듯 하여 적용해봤지만 별로 효과가 없었다고 함
- 그래서 residual connection + layer normalization + swish activation을 같이 활용해봤더니 depth를 늘리는데 효과적이었음
  - 이 때 많은 삽질을 한 듯 하고 친구인 Ishaan Javali(2저자)가 위의 구조를 발견하는데 큰 공을 세운 듯 하다
  - SimBa라는 ICLR2025 논문에 제출된 구조와 유사함

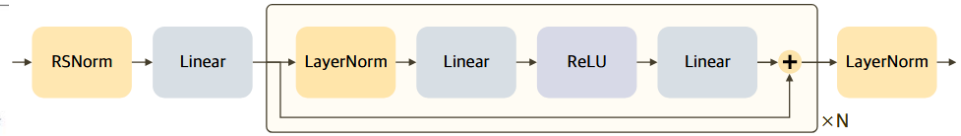




Published as a conference paper at ICLR 2025

SIMBA: SIMPLICITY BIAS FOR SCALING UP PARAMETERS IN DEEP REINFORCEMENT LEARNING

Hojoon Lee[1,2*†] Dongyoon Hwang[1,3*] Donghu Kim[1] Hyunseung Kim[1,3] Jun Jet Tai[2,4†]
Kaushik Subramanian[2] Peter R.Wurman[2] Jaegul Choo[1] Peter Stone[2,5] Takuma Seno[2]
[1]KAIST [2]Sony AI [3]KRAFTON [4]Coventry University [5]UT Austin
{joonleesky, godnpeter}@kaist.ac.kr



- 그러나 아직은 부족했다.
- 당시 Eyenbach의 연구실에는 폴란드에서 온 귀인이 방문한 상태였으니…

The 5 Building Blocks of Scalable RL

Sparse Reward

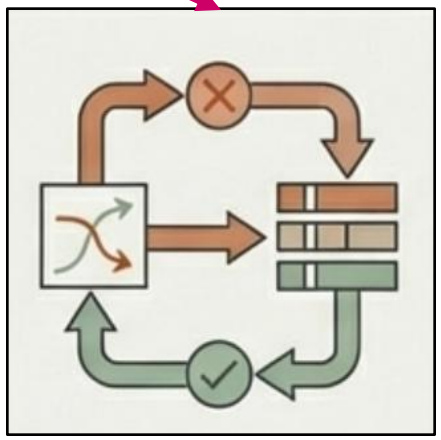
1. Self-Supervised Learning

Exploring without rewards via Contrastive RL

Benjamin Eysenbach: 케빈아 내가 만든거 써봐

Contrastive Learning as Goal-Conditioned Reinforcement Learning

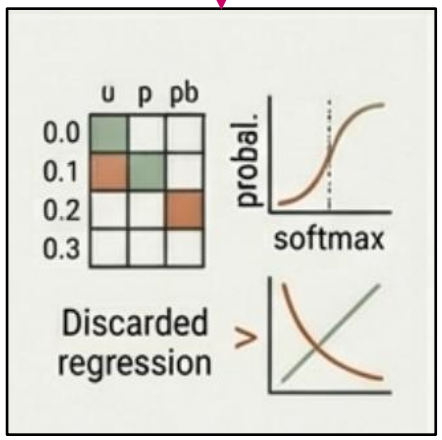
Benjamin Eysenbach[α,β] Tianjun Zhang[γ] Sergey Levine[β,γ] Ruslan Salakhutdinov[α]
[α]CMU [β]Google Research [γ]UC Berkeley



2. Signal Density

Utilizing Hindsight Experience Replay (HER) to turn every failed trajectory into a labeled learning sample

Q-Regression

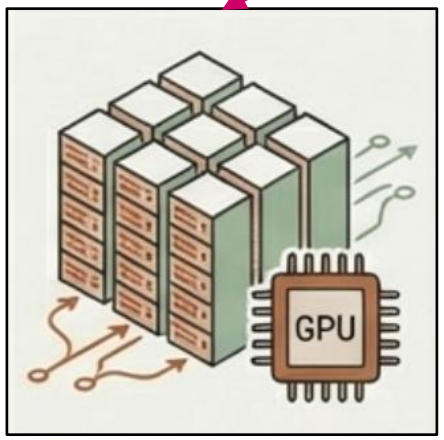


3. Classification Paradigm

Generalizing cross-entropy using an InfoNCE objective rather than regressive TD learning

$$\min_{\phi,\psi} \mathbb{E}_{\mathcal{B}}\left[-\sum_{i=1}^{|\mathcal{B}|}\log\left(\frac{e^{f_{\phi,\psi}(s_i,a_i,g_i)}}{\sum_{j=1}^{K} e^{f_{\phi,\psi}(s_i,a_i,g_j)}}\right)\right]$$

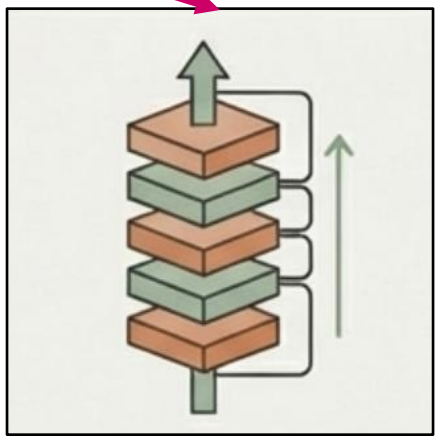
Limited Batch Size



4. Data Scale

Generating massive online data via GPU-accelerated environments (JaxGCRL)

폴란드 귀인이 만듦
Michał Bortkiewicz



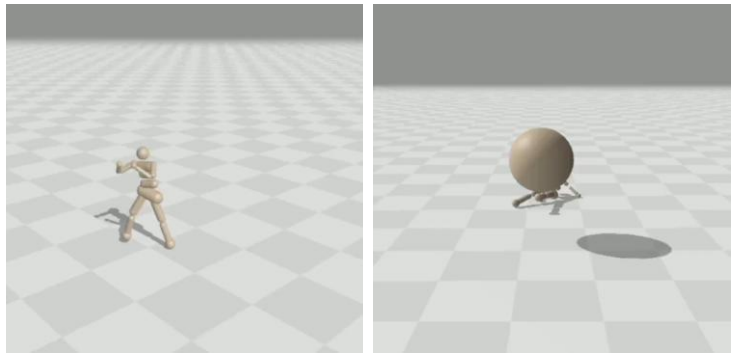
5. Stabilized Architecture

Custom residual blocks to prevent training instability at extreme depths

Empirical Results

- 레이어를 높였을 때 성능이 좋아짐을 확인함

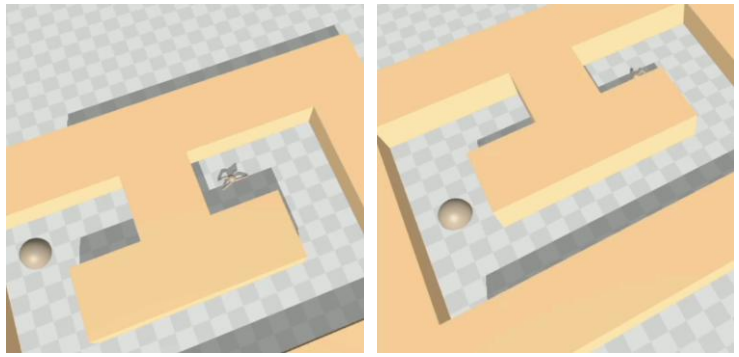
Humanoid



Conventional nets (4 layers)

Deep nets (64 layers)

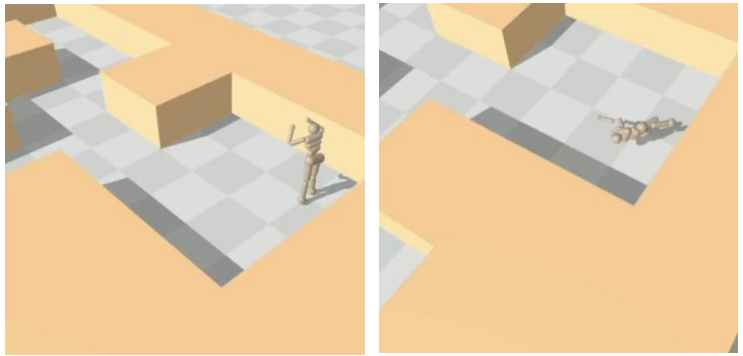
Ant U4-Maze



Conventional nets (4 layers)

Deep nets (64 layers)

Humanoid Big Maze

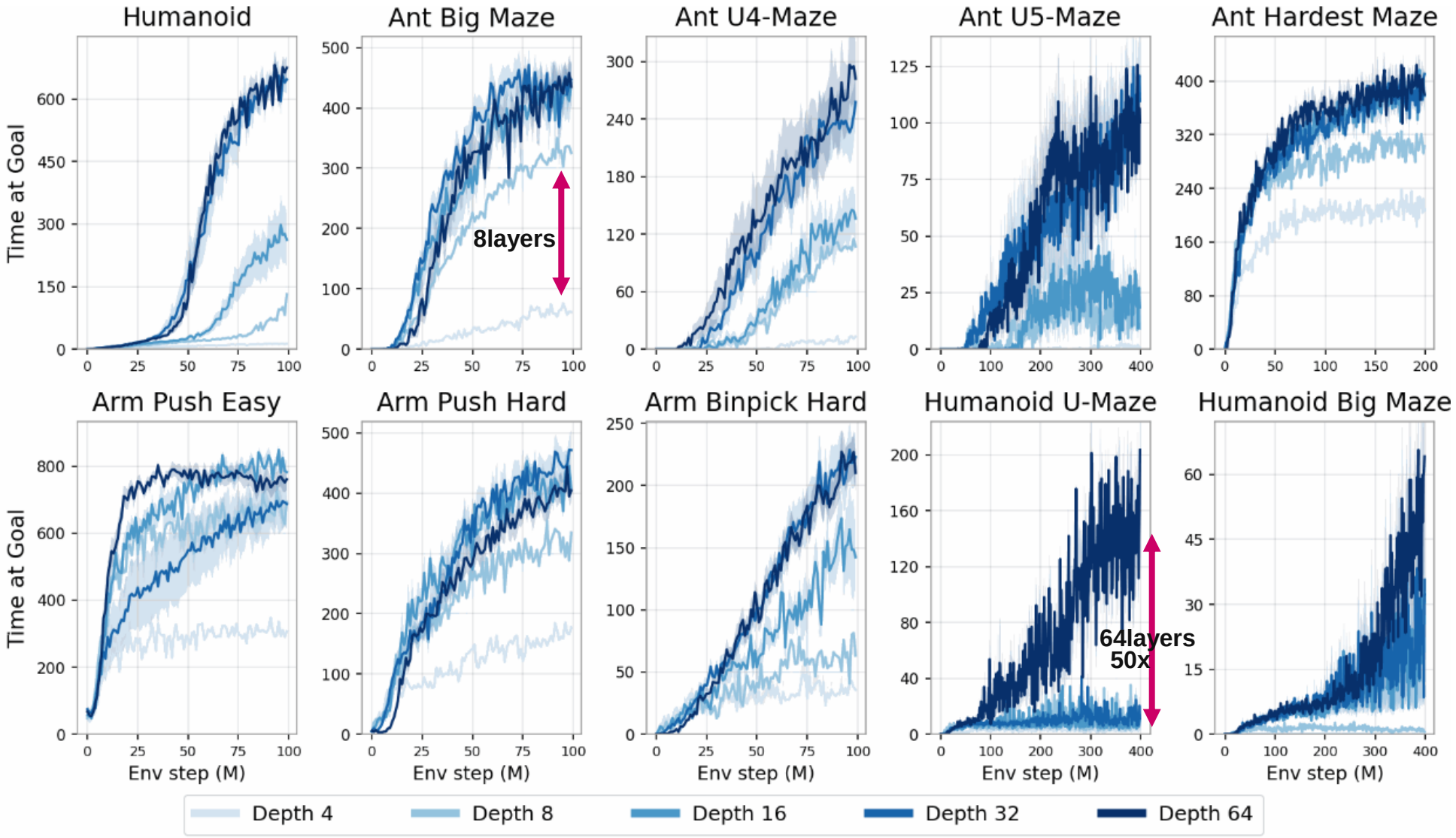


Conventional nets (4 layers)

Deep nets (64 layers)

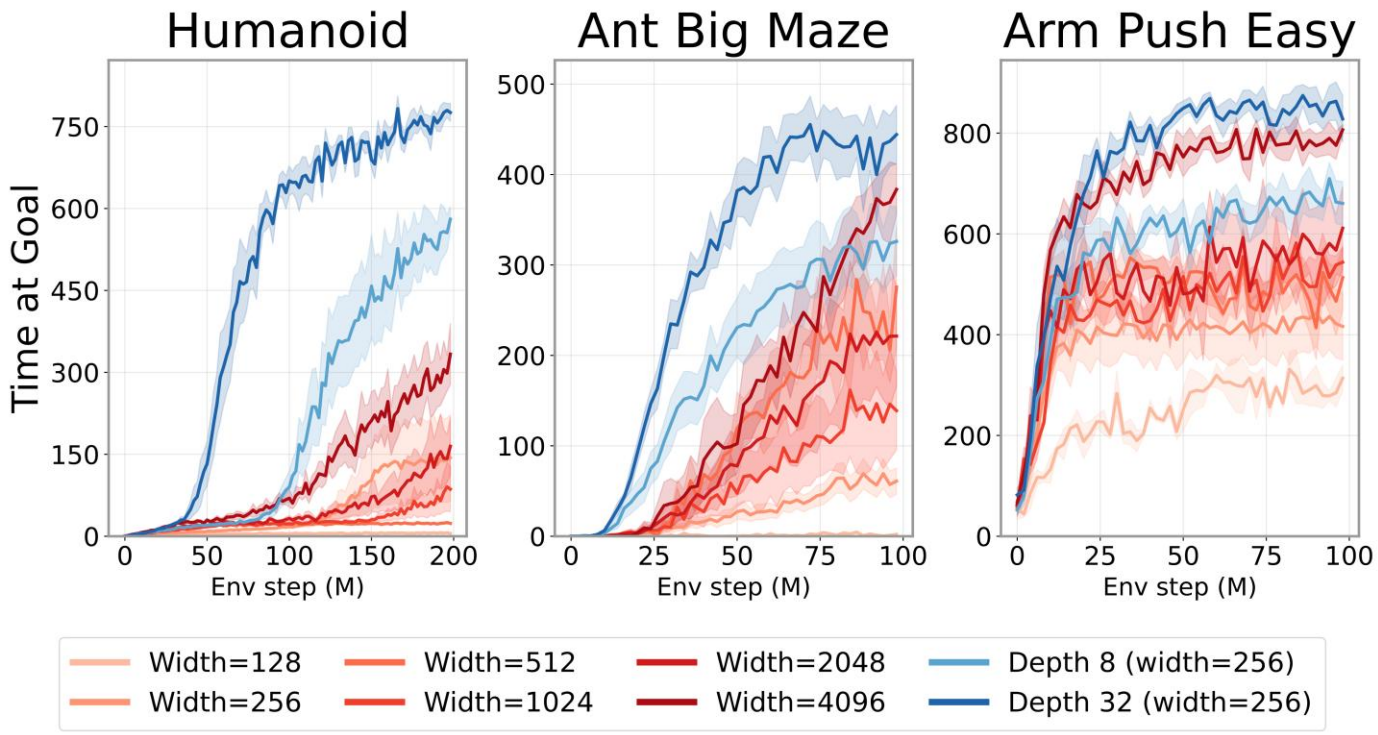
Empirical Results

- 레이어를 높였을 때 성능이 좋아짐을 확인함

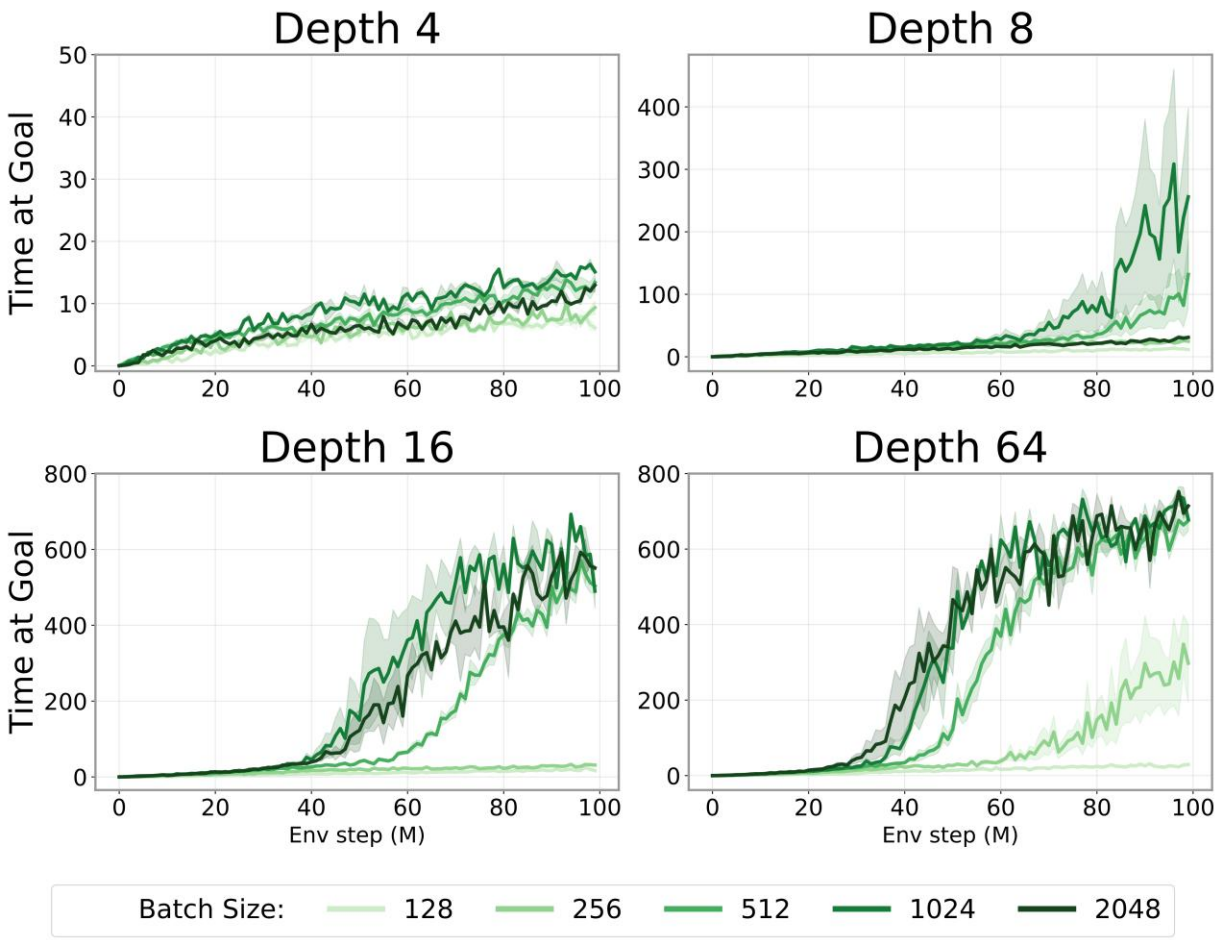


Scaling Depth Outperforms Width Scaling

- Prior RL research focused on scaling network width, often reporting negative returns for depth
- In the Humanoid environment, doubling depth to 8 (width 256) easily outperforms a massive width 4096 network

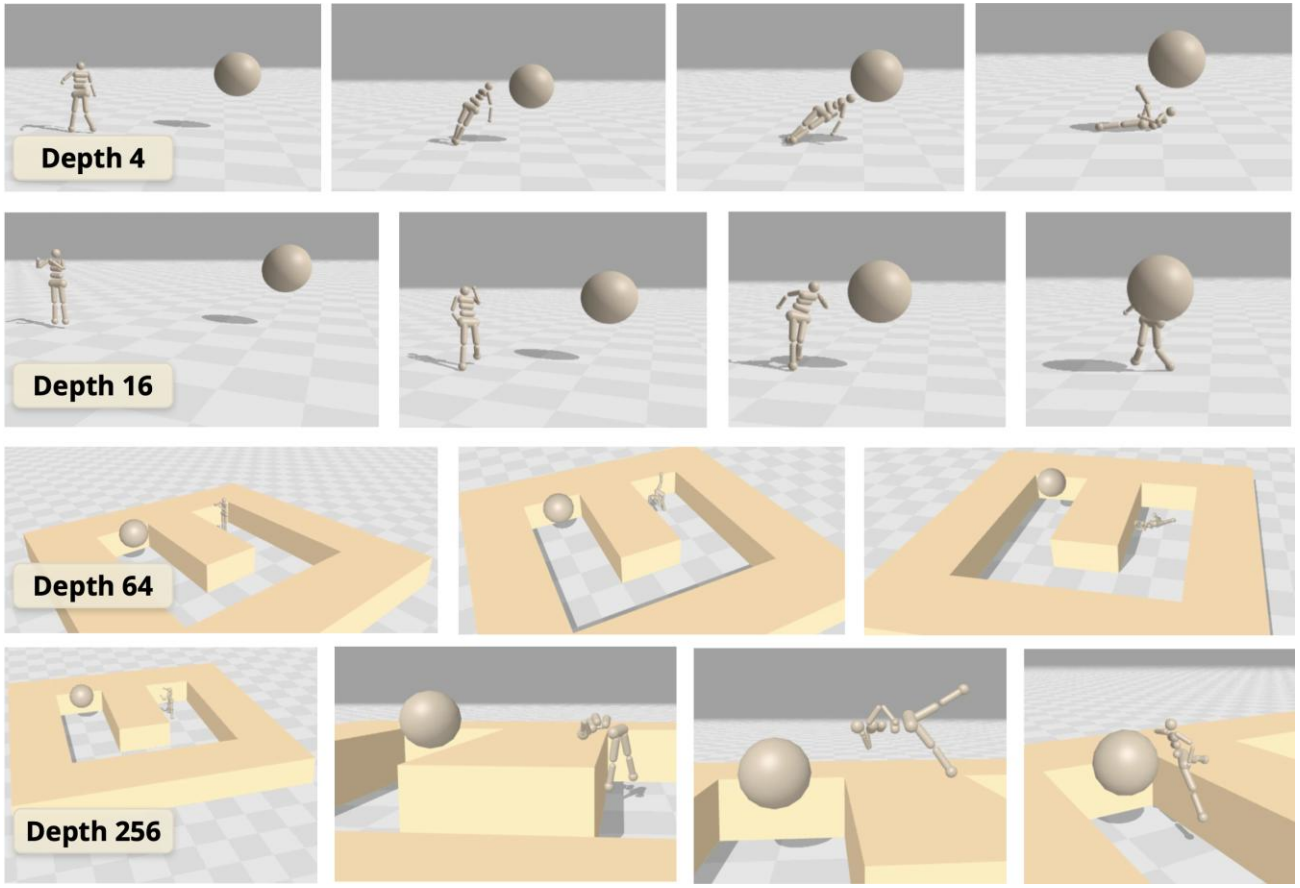


Scaling Depth Unlocks Batch Size Scaling



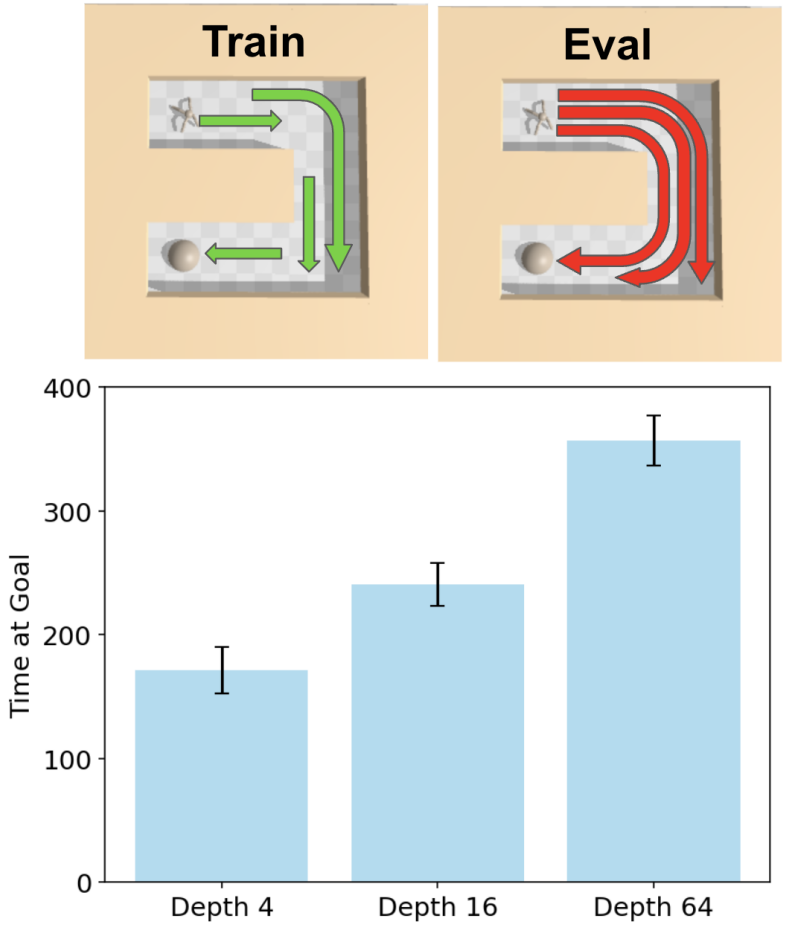
- In NLP and Vision, larger batch sizes reliably improve training
- In RL, large batches historically offer zero benefit
- Small models obscure the benefits of large batch sizes
- Once network depth scales to 64 layers, the model finally possesses the capacity to leverage large batches, unlocking an entirely new dimension of scalable self-supervised RL

레이어 깊이에 따른 질적 향상



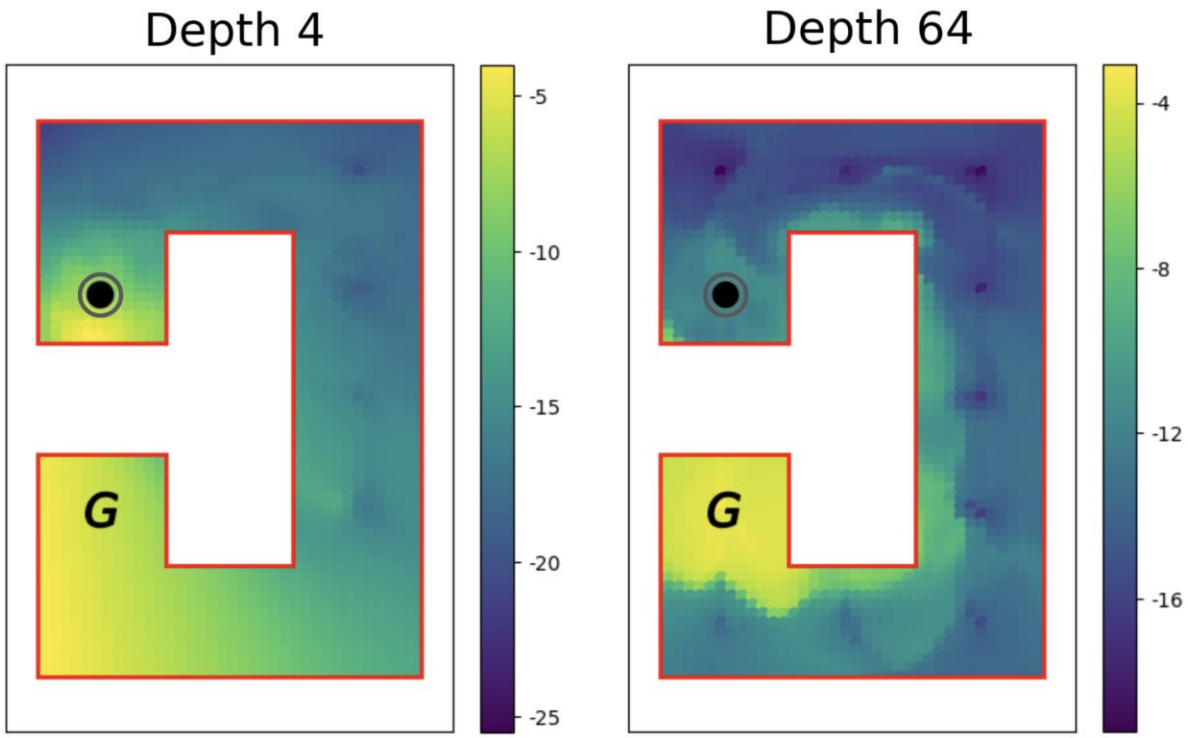
Increasing depth results in new capabilities:
Row 1: A humanoid agent trained with network depth 4 collapses and throws itself towards the goal, as opposed to in
Row 2, where the depth 16 agent gains the ability to walk upright.
Row 3: At depth 64, the humanoid agent in U-Maze struggles to reach the goal and falls.
Row 4: An impressively novel policy emerges at depth 256, as the agent exhibits an acrobatic strategy of compressing its body to vault over the maze wall.
그런데 논문소개 웹페이지에도
이 현상을 보여주는 동영상이 없음
(혹시 재현이 잘 안되는지?)

레이어 깊이에 따른 질적 향상



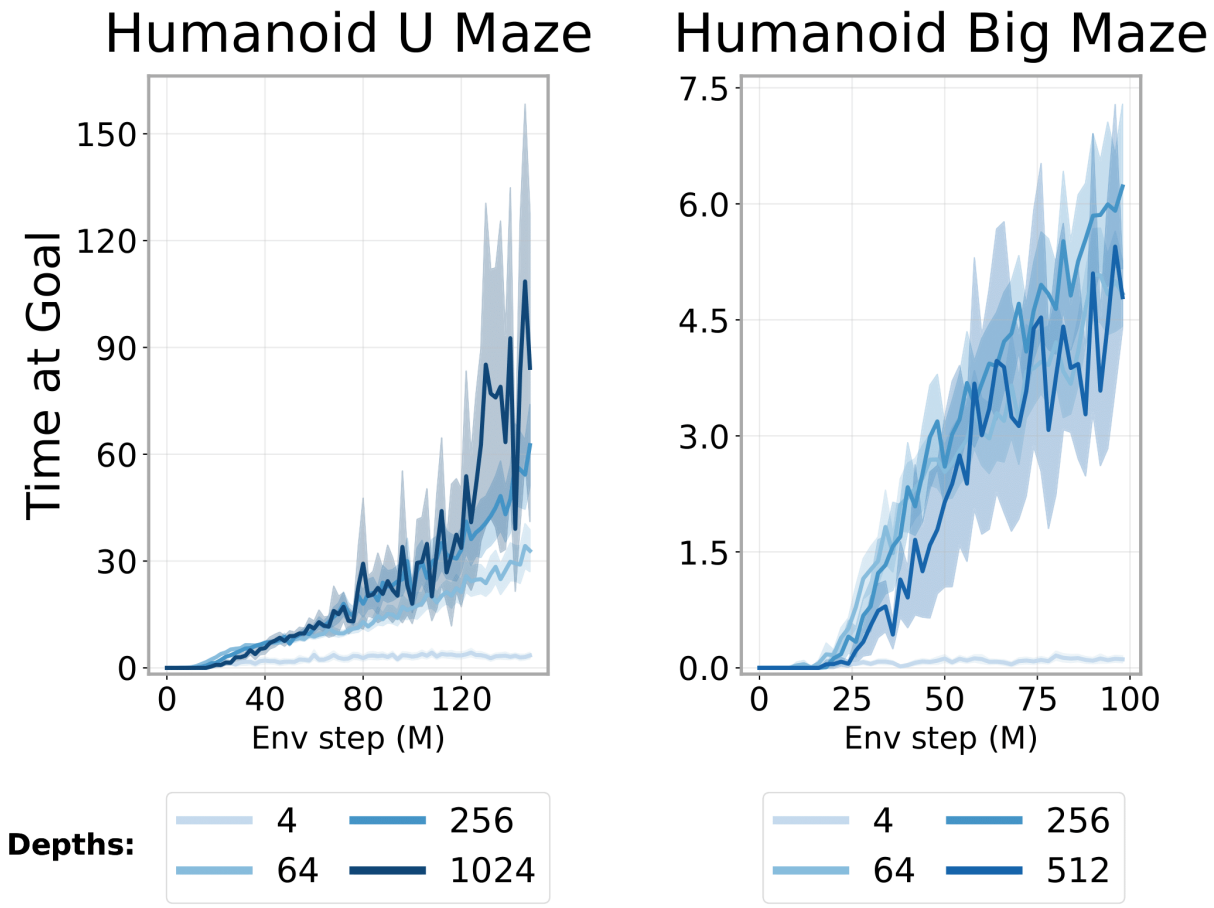
Deep networks exhibit improved generalization.
(Top left) We modify the training setup of the Ant U-Maze environment such that start-goal pairs are separated by ≤3 units. This design guarantees that no evaluation pairs (top right) were encountered during training, testing the ability for combinatorial generalization via "stitching."
(Bottom) Generalization ability improves as network depth grows from 4 to 16 to 64 layers.

레이어 깊이에 따른 질적 향상



Deep networks learn better representations.
In the U4-Maze, the start and goal positions are indicated by the ⊙ and G symbols respectively, and the visualized Q values are computed via the L_2 distance in the learned representation space, i.e., $Q(s,a,g) = \|\varphi(s,a) - \psi(g)\|_2$. The shallow depth-4 network (left) appears to naively rely on Euclidean proximity, exhibited by the high Q values of the semicircular gradient near the start position, despite the maze wall.
In the depth-64 heatmap (right), the highest Q values cluster at the goal, gradually tapering along the maze's interior boundary. These results highlight how increasing depth is important for learning value functions in goal-conditioned settings, which are characterized by long horizons and sparse rewards.

How Far Can We Scale Depth?



Exploring the limits of scaling.
We push the boundaries of network depth scaling up to 1024 layers. In Humanoid Big Maze, performance plateaus. However, on Humanoid U-Maze, we observe continued performance improvements with network depths of 256 and 1024 layers. This may be because the maneuver being learned (flipping over the wall) is exceptionally complex (see Figure above), and requires greater network depth to learn. Note that for the 1024-layer training runs, we observed the actor loss exploding at the onset of training, so we maintained the actor depth at 512 while using 1024-layer networks only for the two critic encoders.

후일담

- 논문은 솔직히 잘 못썼다고 느꼈음
  - 논문이라기보다는 실험레포트 느낌이 있음
  - “해보니까 되더라”(LLM scale-up 논문들에서 느껴지는 것과 유사한…)
  - 이론적인 내용은 아예 없음
  - 그럼에도 best paper
- 논문의 편집이 덜 된 느낌 (가령 Figure의 번호가 순차적이지 않음. Figure 3 다음에 Figure 5, 6이 나오고 그 다음에 Figure 4가 나옴. 본문에 없는 Figure 12설명을 한참하는데 어펜딕스에 있음)
  - 논문마감일 맞추다가 정리가 덜 된 느낌
  - 그럼에도 best paper
- 제목에 1000개 레이어를 강조하는데 논문 내에서는 특정 실험에서 1000개까지 해봤다 정도임
  - 제목에서 느껴지는 인상은 많은 실험에서 1000개 레이어를 활용해서 뭔가를 한 것 같은데 그 정도는 아니었음. 과장이 좀 있어 보임
  - 그럼에도 best paper
- 그러나 논문이 다루고 있는 문제가 기존 강화학습의 한계를 정면으로 다룬다는 점에서 임팩트가 크다고 여겨지는 듯 함
  - 그럼에도 불구하고 이 논문이 best paper인 것은 솔직히 의문
  - 그러나 CRL을 활용한 접근법에는 관심이 가며, 무엇보다 이제는 JAX를 써봐야겠다고 다짐

후일담

- 의외의 소득(?)
  - SAC, TD3 등의 네트워크 깊이는 4레이어면 충분하다

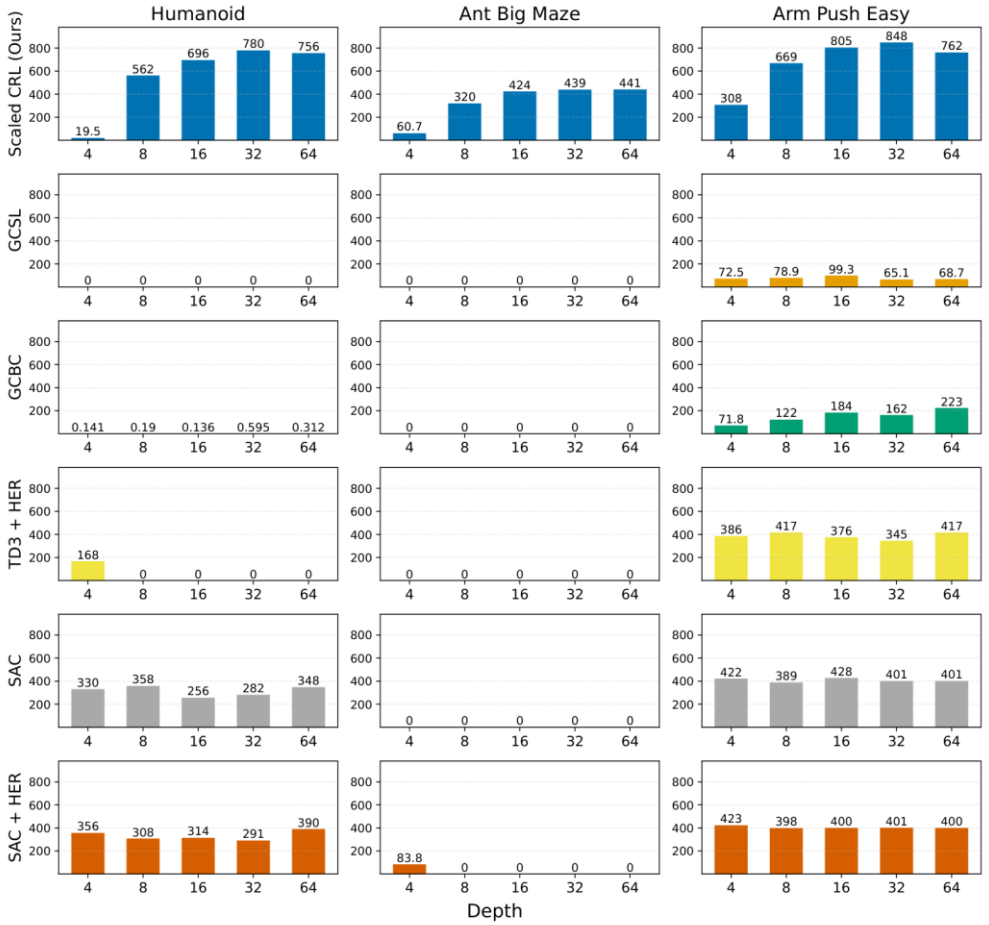


Figure 13: Depth scaling yields limited gains for SAC, SAC+HER, TD3+HER, GCSL, and GCBC.