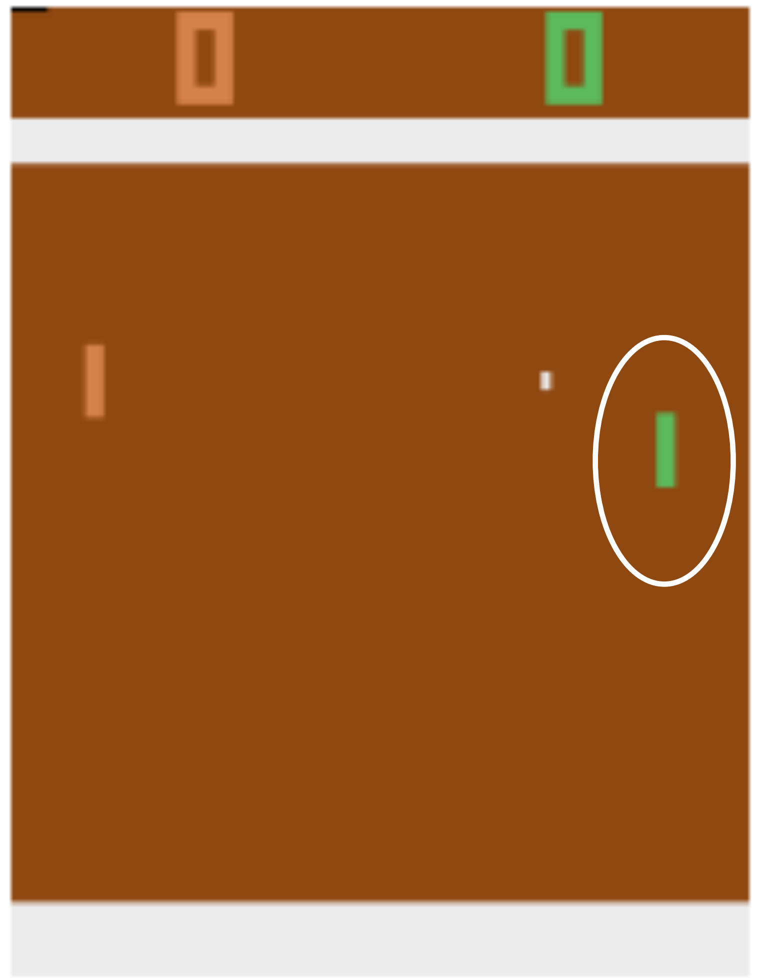
Robust Deep Reinforcement Learning through Bootstrapped Opportunistic Curriculum

Junlin Wu · Yevgeniy Vorobeychik
Washington University in St. Louis

Background

Adversarial Deep Reinforcement Learning



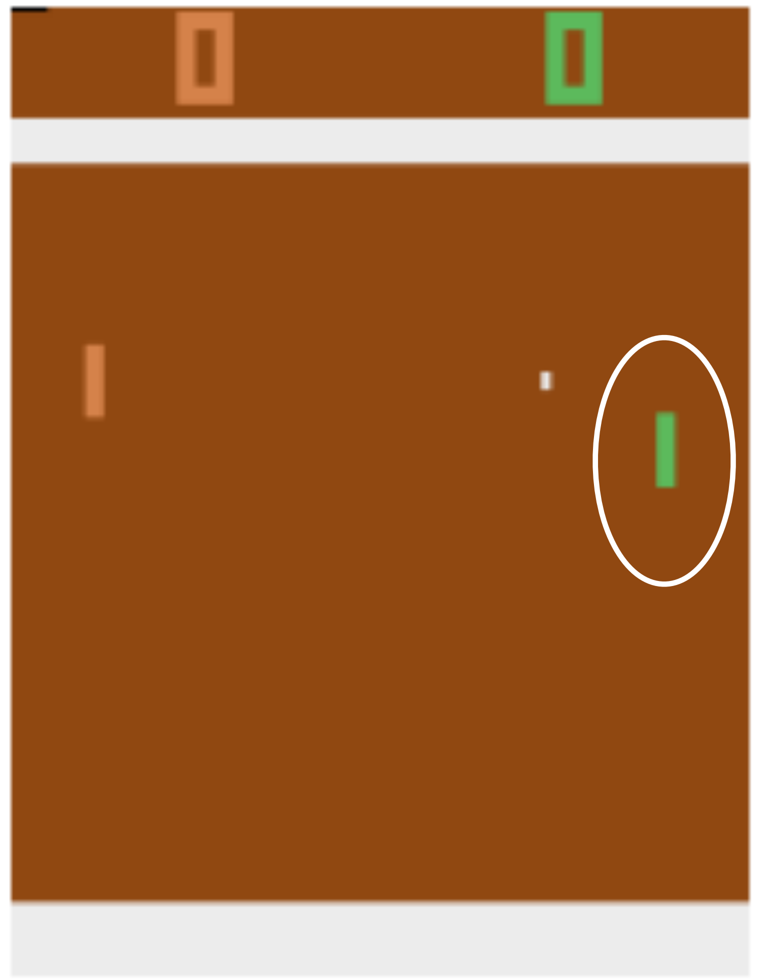
Original Image Input

NN

Action: Move up

Background

Adversarial Deep Reinforcement Learning

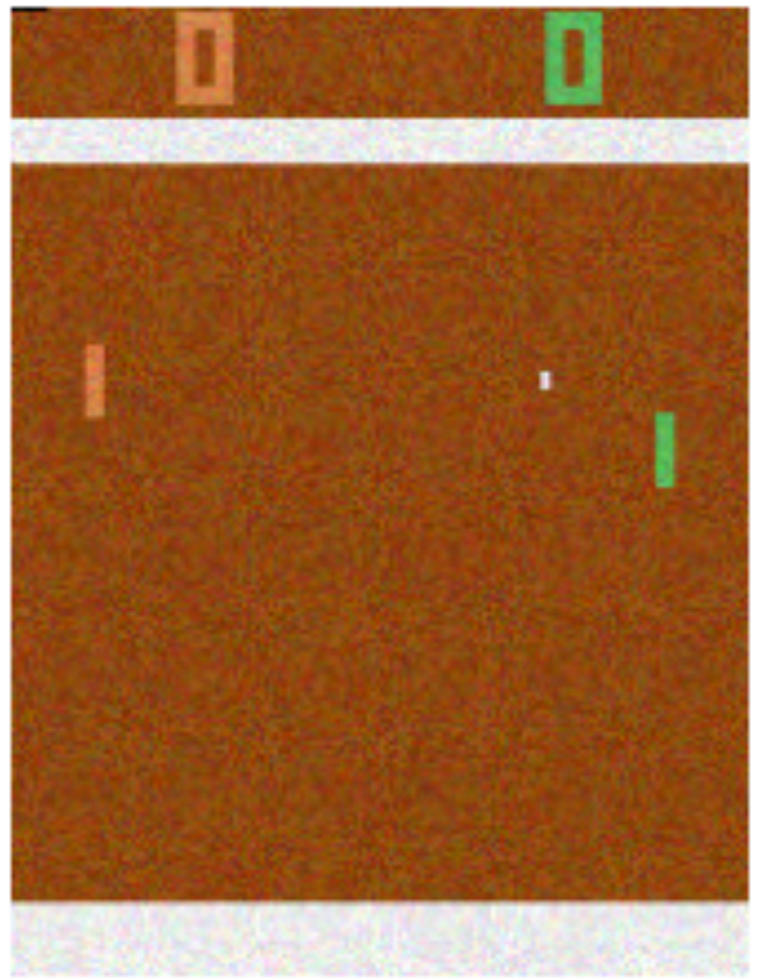


Original Image Input

NN

Action: Move up

Attacker



+ Adversarial Perturbation

NN

Action: Move Down

Background

Attacking Method

- A common attack on Deep Q-Network (DQN) aims maximize cross-entropy loss $\mathscr{L}(\mathrm{Softmax}(Q(s+\delta;\theta)),\pi(s))$ with respect to δ (adversarial perturbation), where $Q(s)$ is the vector of Q values over all actions in state s.

Background
Attacking Method

- A common attack on Deep Q-Network (DQN) aims maximize cross-entropy loss $\mathcal{L}(\mathrm{Softmax}(Q(s+\delta;\theta)),\pi(s))$ with respect to δ (adversarial perturbation), where $Q(s)$ is the vector of Q values over all actions in state s.
- A **PGD** (projected gradient descent) attack updates δ iteratively:
  $\delta_{k+1} \leftarrow \delta_k + \alpha \cdot \mathrm{sign}(\nabla_\delta \mathcal{L}(Q(x+\delta_k;\theta),\pi(s)))$
  over a fixed number of iterations with $||\delta||_\infty \leq \epsilon$.

Background

Attacking Method

- A common attack on Deep Q-Network (DQN) aims maximize cross-entropy loss $\mathscr{L}(\mathrm{Softmax}(Q(s+\delta;\theta)),\pi(s))$ with respect to δ (adversarial perturbation), where $Q(s)$ is the vector of Q values over all actions in state s.

- A **PGD** (projected gradient descent) attack updates δ iteratively:
  $\delta_{k+1} \leftarrow \delta_k + \alpha \cdot \mathrm{sign}(\nabla_\delta \mathscr{L}(Q(x+\delta_k;\theta),\pi(s)))$
  over a fixed number of iterations with $||\delta||_\infty \leq \epsilon$.

- A special class of PGD is **FGSM** (fast gradient sign method), where PGD is executed for only a single iteration and $\alpha = \epsilon$.

Prior Literature

Adversarial Deep Reinforcement Learning

- The goal is to train a robust RL agent (i.e., achieve a high reward when under adversarial attack $||\delta||_\infty \leq \epsilon$).

Prior Literature

Adversarial Deep Reinforcement Learning

- The goal is to train a robust RL agent (i.e., achieve a high reward when under adversarial attack $||\delta||_\infty \leq \epsilon$).
- The SOTA method is RADIAL [Oikarinen et al., 2021], where they leveraged interval bound propagation to increase the robustness of RL agent (e.g., robust up to 5/255 for Pong game).

Prior Literature

Adversarial Deep Reinforcement Learning

- The goal is to train a robust RL agent (i.e., achieve a high reward when under adversarial attack $||\delta||_\infty \leq \epsilon$).
- The SOTA method is RADIAL [Oikarinen et al., 2021], where they leveraged <u>interval bound propagation</u> to increase the robustness of RL agent (e.g., robust up to 5/255 for Pong game).
- Our goal is to increase the robustness of the RL agent further (robust against higher values of ϵ).

BCL Framework

Overview

- We propose ***B**ootstrapped Opportunistic Adversarial **C**urriculum **L**earning* (BCL), a novel flexible adversarial curriculum learning framework for robust reinforcement learning.

BCL Framework

Overview

- We propose ***B****ootstrapped Opportunistic Adversarial* ***C****urriculum* ***L****earning* (BCL), a novel flexible adversarial curriculum learning framework for robust reinforcement learning.

- Begins by creating a **baseline curriculum**: an increasing sequence of L attack budgets $\{\epsilon_i\}$, with $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_L$, where $\epsilon_L = \epsilon$ is our target robustness level.

BCL Framework

Overview

- We propose ***B**ootstrapped Opportunistic Adversarial **C**urriculum **L**earning* (BCL), a novel flexible adversarial curriculum learning framework for robust reinforcement learning.
- Begins by creating a **baseline curriculum**: an increasing sequence of L attack budgets $\{\epsilon_i\}$, with $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_L$, where $\epsilon_L = \epsilon$ is our target robustness level.
- In each curriculum phase, we run adversarial training (AT) **up to K times**, where each AT run is bootstrapped by the best model obtained thus far.

BCL Framework
Overview

- We propose ***B**ootstrapped Opportunistic Adversarial **C**urriculum **L**earning* (BCL), a novel flexible adversarial curriculum learning framework for robust reinforcement learning.
- Begins by creating a **baseline curriculum**: an increasing sequence of L attack budgets $\{\epsilon_i\}$, with $\epsilon_1 < \epsilon_2 < \cdots < \epsilon_L$, where $\epsilon_L = \epsilon$ is our target robustness level.
- In each curriculum phase, we run adversarial training (AT) **up to K times**, where each AT run is bootstrapped by the best model obtained thus far.
- For example, based on observed performance, we could speed up the training by
  - Performing fewer than K runs for each curriculum phases;
  - Skipping forward the curriculum phases.

Adversarial Loss Function

We experiment on two types of adversarial loss functions:

Adversarial Loss Function

We experiment on two types of adversarial loss functions:

- RADIAL method:

  Use **interval bound propagation** (RADIAL-DQN [Oikarinen et al., 2021]).

Adversarial Loss Function

We experiment on two types of adversarial loss functions:

- RADIAL method:

  Use **interval bound propagation** (RADIAL-DQN [Oikarinen et al., 2021]).

- AT method:

  Use FGSM-based method and leverage the structure of Double-DQN to generate **adversarial examples** efficiently during training time.

BCL Framework

Special Cases of BCL

- AT-DQN (Adversarial Training)
- NCL-AT/RADIAL-DQN (Naive Curriculum Learning)

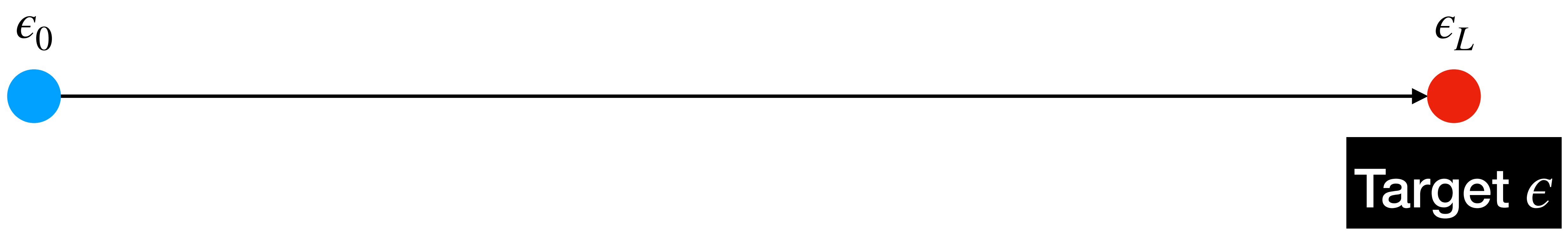
Benchmark Models

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)
- BCL-RADIAL-DQN (BCL with RADIAL approach)
- BCL-RADIAL+AT-DQN (BCL-RADIAL-DQN + BCL-C-AT-DQN)

BCL Framework

Special Cases of BCL

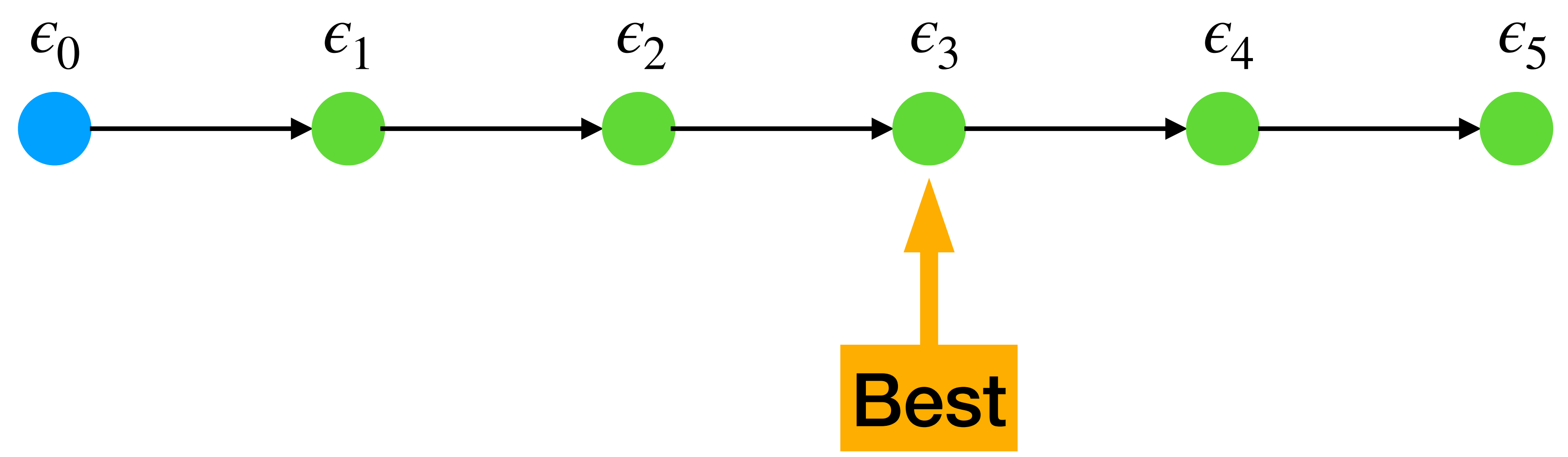
- AT-DQN (Adversarial Training)



BCL Framework

Special Cases of BCL

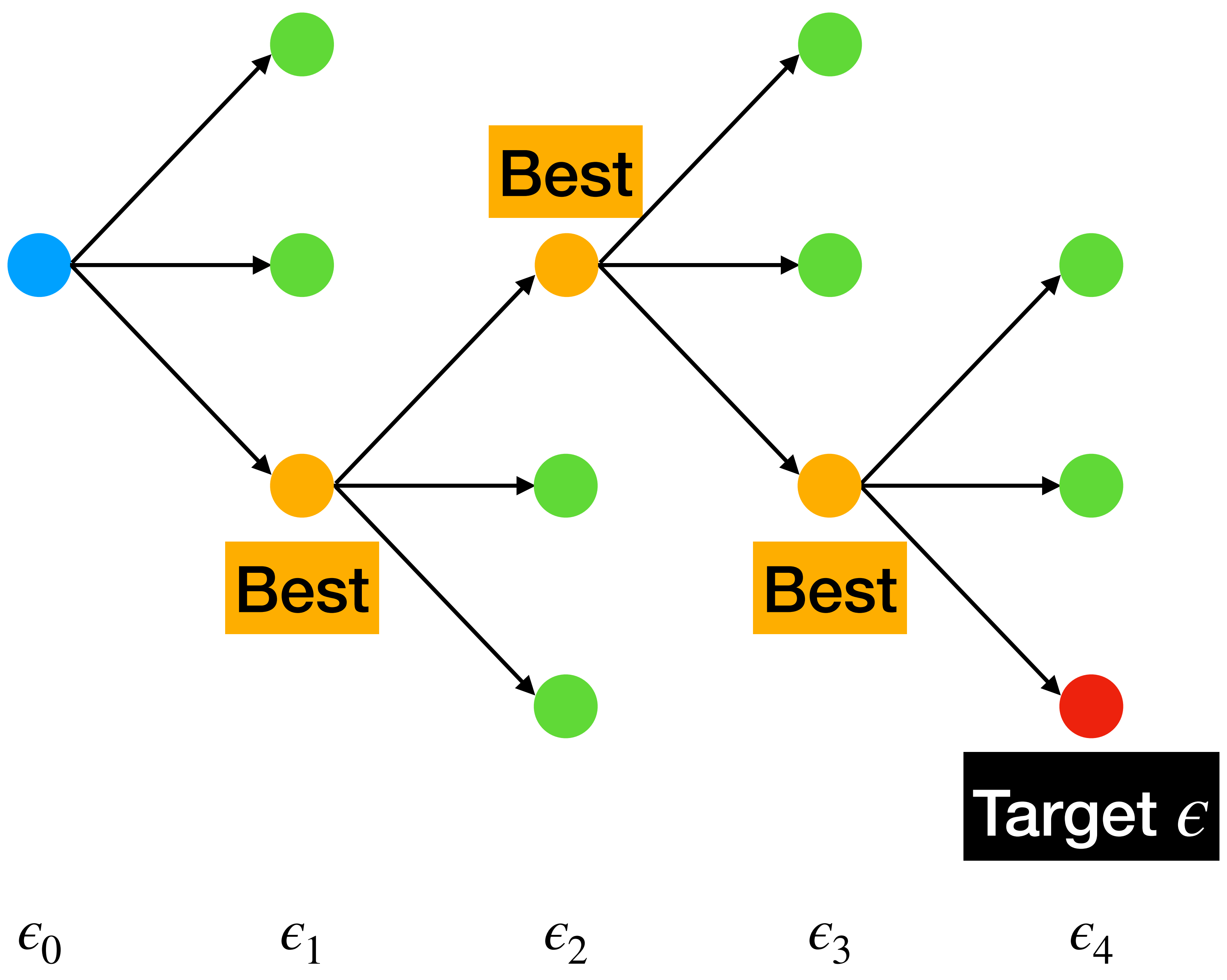
- AT-DQN (Adversarial Training)
- NCL-AT/RADIAL-DQN (Naive Curriculum Learning)



BCL Framework

Special Cases of BCL

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)

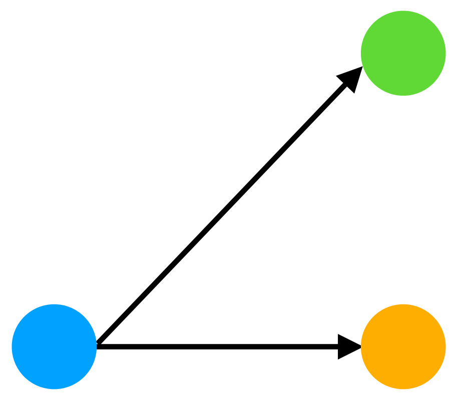


- Perform K runs for each phase
- Choose the best model among K results

BCL Framework
Special Cases of BCL

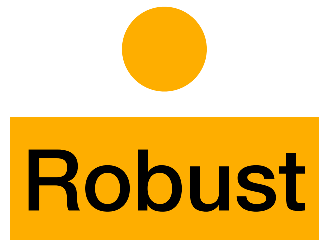
- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)

We use a threshold to decide whether a model is robust against ϵ_i



- Perform **up to** K runs for each phase

ϵ_0 ϵ_1 ϵ_2 ϵ_3 ϵ_4 ϵ_5



BCL Framework

Special Cases of BCL

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)

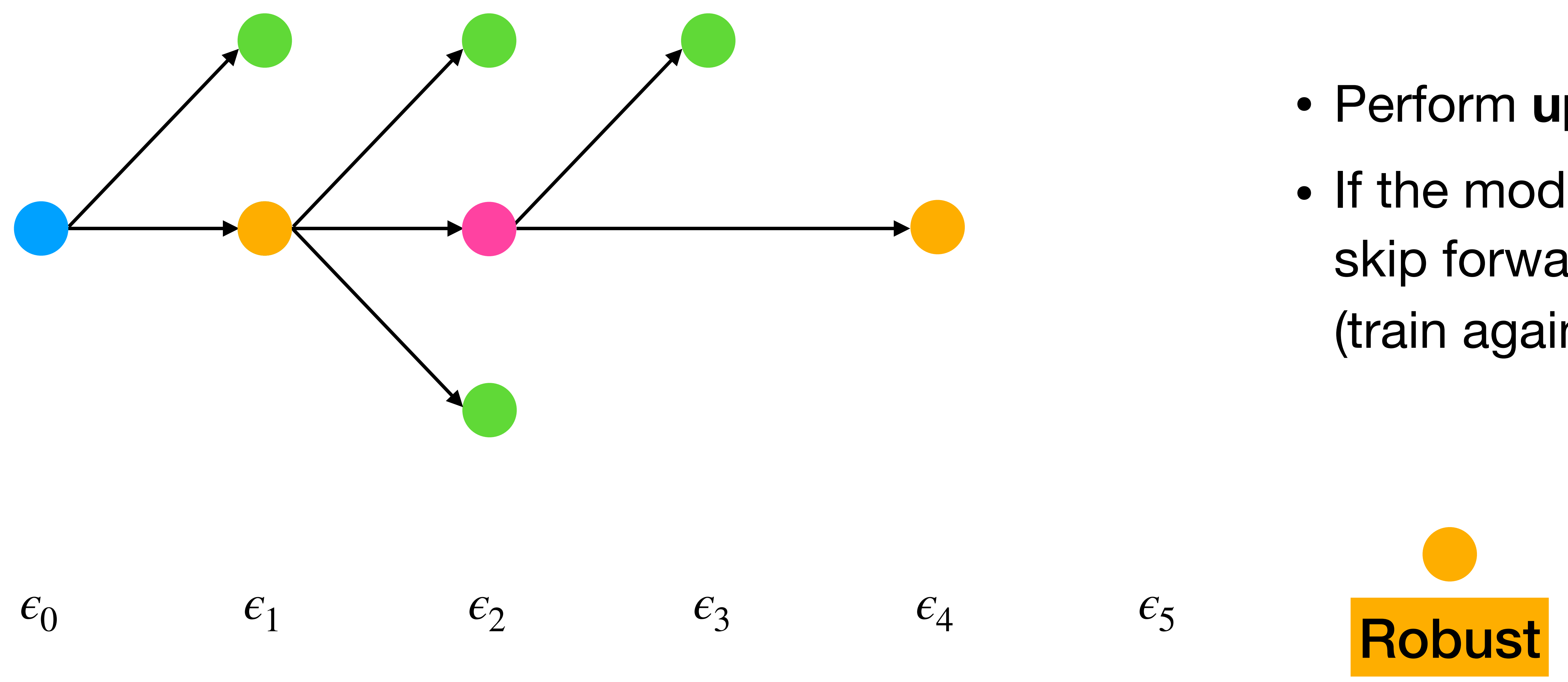
We use a threshold to decide whether a model is robust against ϵ_i

- Perform **up to** K runs for each phase

BCL Framework
Special Cases of BCL

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)

We use a threshold to decide whether a model is robust against ϵ_i



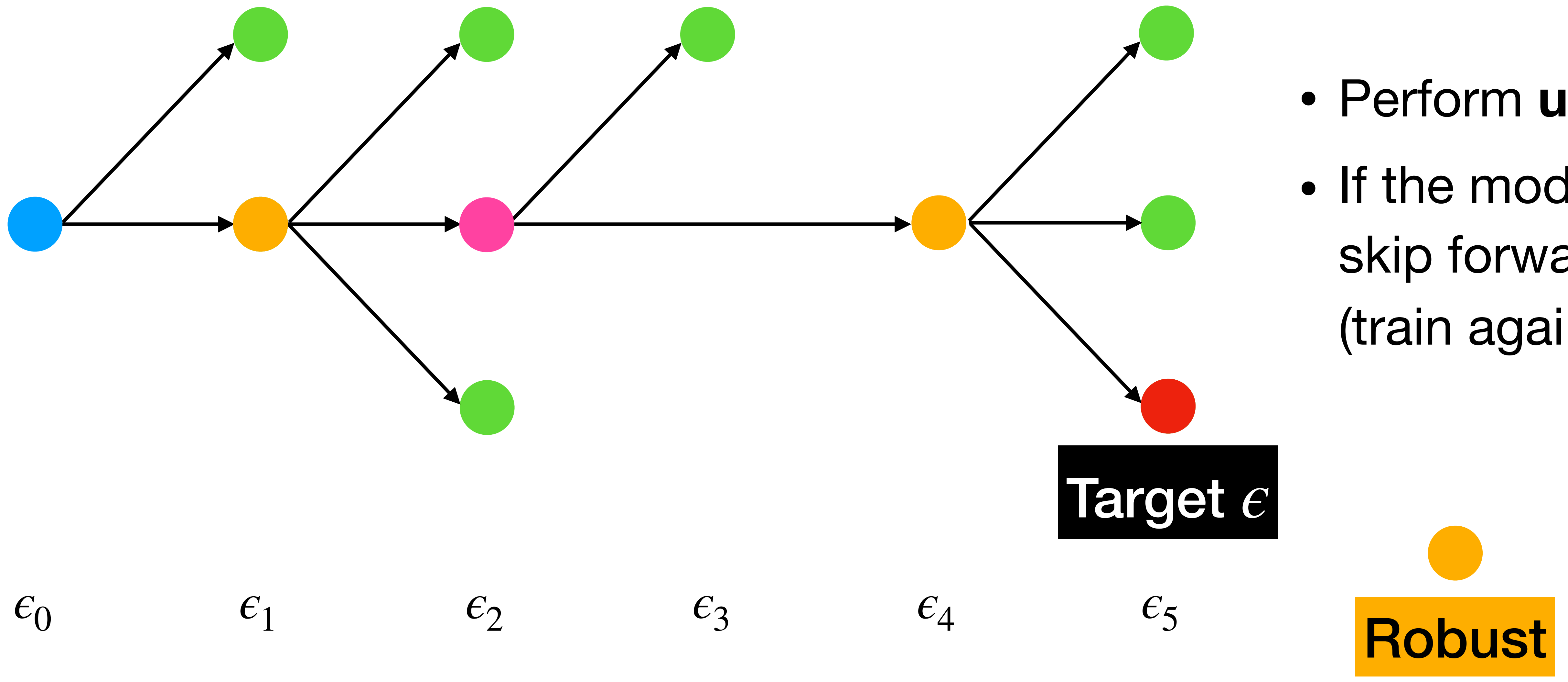
- Perform **up to** K runs for each phase
- If the model is robust against ϵ_{i+1}, skip forward the curriculum phase (train against ϵ_{i+2})

BCL Framework

Special Cases of BCL

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)

We use a threshold to decide whether a model is robust against ϵ_i



- Perform **up to** K runs for each phase
- If the model is robust against ϵ_{i+1}, skip forward the curriculum phase (train against ϵ_{i+2})

BCL Framework

Special Cases of BCL

- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)
- BCL-RADIAL-DQN (BCL with RADIAL approach)

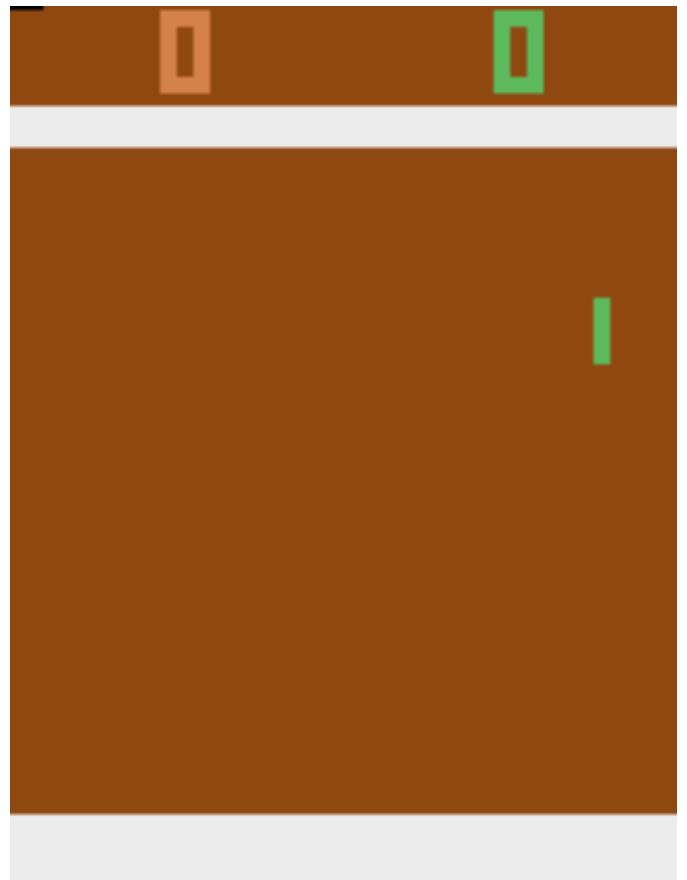
BCL Framework

Special Cases of BCL

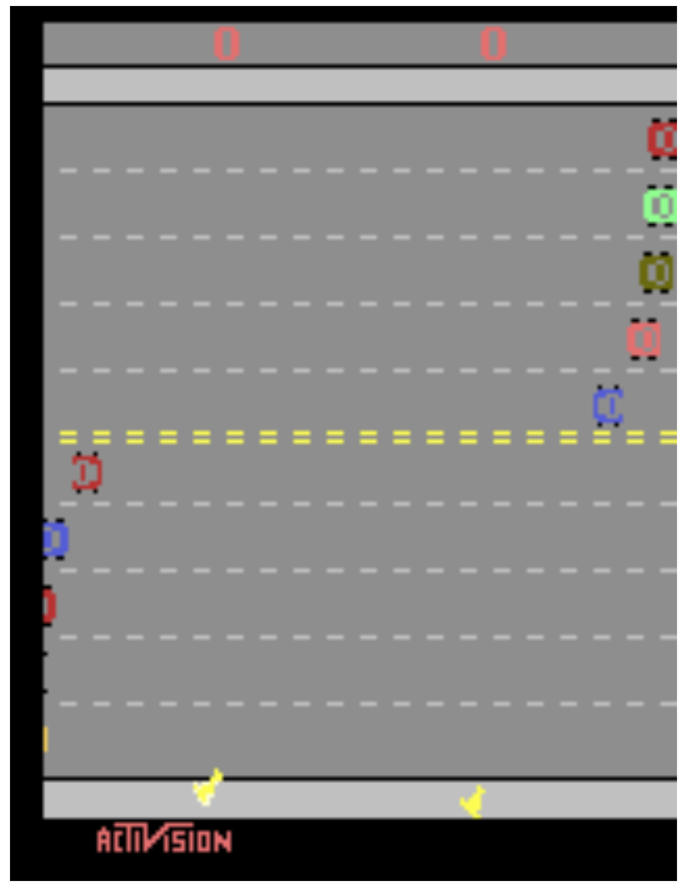
- BCL-C-AT-DQN (Conservatively Bootstrapped Curriculum Learning)
- BCL-MOS-AT-DQN (Maximum Opportunistic Skipping)
- BCL-RADIAL-DQN (BCL with RADIAL approach)
- BCL-RADIAL+AT-DQN (BCL-RADIAL-DQN + BCL-C-AT-DQN)

Experiments

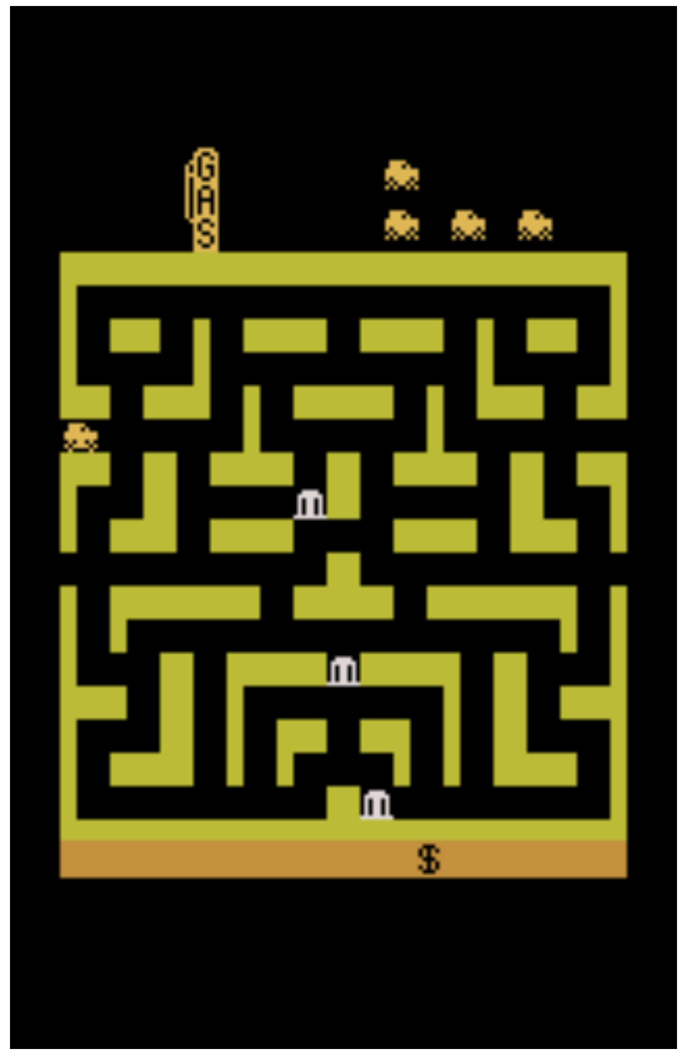
- We evaluate the proposed approach using four Atari-2600 games from the OpenAI Gym with discrete action space:



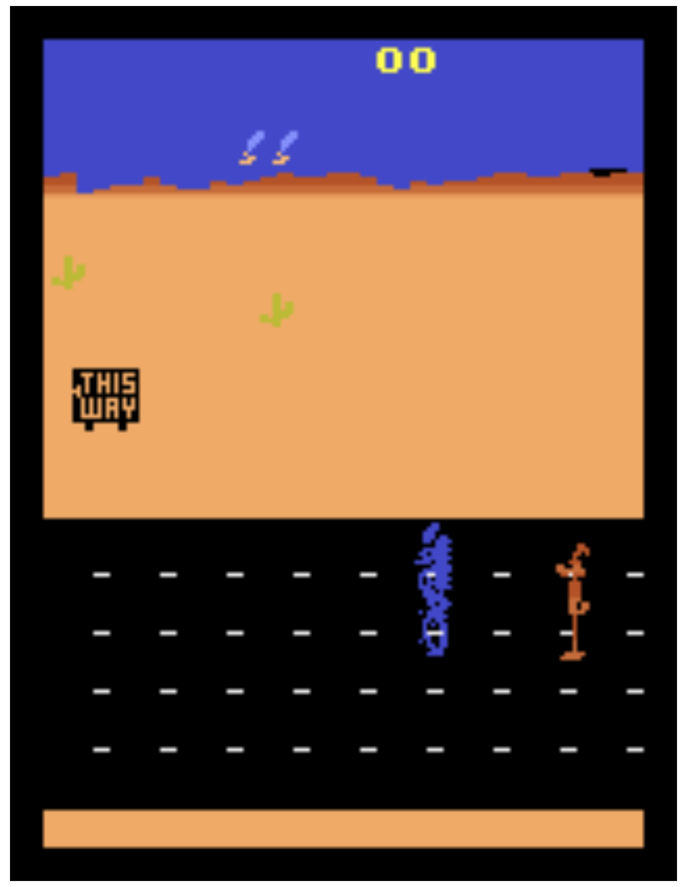
Pong



Freeway



BankHeist



RoadRunner

Experiments

Benchmark Models

- DQN (Vanilla)
- SA-DQN (Convex) [Zhang et al., 2020]
- RADIAL-DQN [Oikarinen et al., 2021]

- AT-DQN (standard adversarial training)
- NCL-AT-DQN (naive curriculum learning with adversarial examples)
- NCL-RADIAL-DQN (naive curriculum learning with RADIAL method)

Experiments

Results — Pong

- Our BCL models trained with adversarial examples (BCL-C/MOS-AT-DQN) significantly outperforms all benchmark models for higher values of ϵ.

	Pong			
Method/Metric	Nominal	30-step PGD/RI-FGSM Attack		
ϵ	0	10/255	20/255	25/255
DQN (Vanilla)	21.0	-21.0	-21.0	-21.0
SA-DQN (Convex)	21.0	-21.0	-21.0	-21.0
RADIAL-DQN	21.0	-21.0	-21.0	-21.0
AT-DQN	21.0	18.0	-0.8	-19.4
NCL-AT-DQN	21.0	20.4	-21.0	-21.0
NCL-RADIAL-DQN	21.0	-20.6	-21.0	-21.0
BCL-C-AT-DQN	21.0	21.0	21.0	21.0
BCL-MOS-AT-DQN	21.0	21.0	20.9	20.9
BCL-RADIAL-DQN	21.0	21.0	-20.9	-21.0

Experiments

Results — BankHeist

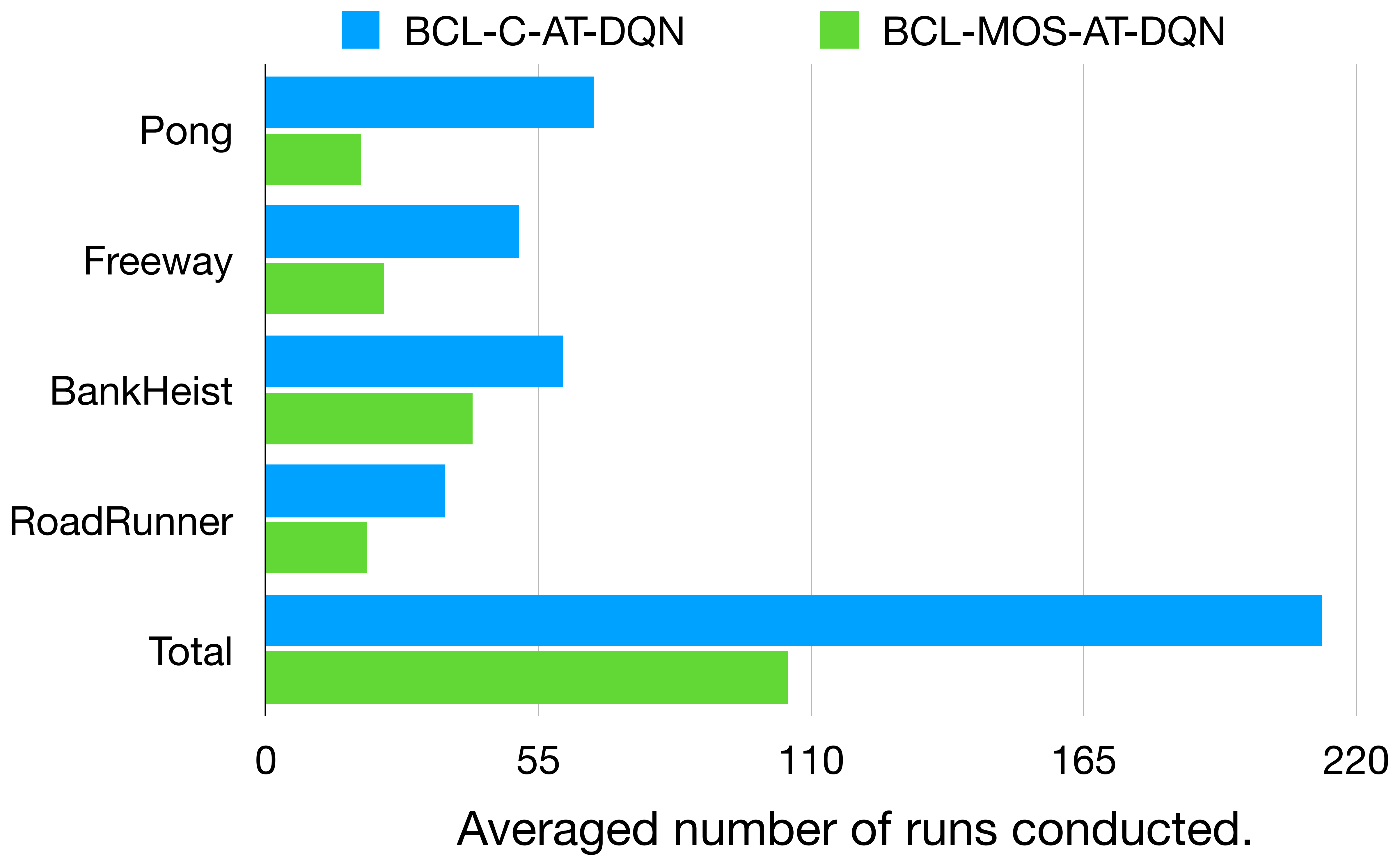
- Our BCL models outperform all benchmarks.
- BCL-RADIAL+AT-DQN yields the most significant result.

		BankHeist		
Method/Metric	Nominal	30-step PGD/RI-FGSM Attack		
ϵ	0	5/255	10/255	15/255
DQN (Vanilla)	1325.5	0.0	0.0	0.0
SA-DQN (Convex)	1237.5	1126.0	63.0	16.0
RADIAL-DQN	1349.5	581.5	0.0	0.0
AT-DQN	1271.0	129.0	5.5	0.0
NCL-AT-DQN	1311.0	245.0	1.0	0.0
NCL-RADIAL-DQN	1272.0	1168.0	59.5	9.0
BCL-C-AT-DQN	1285.5	1143.5	988.5	250.5
BCL-MOS-AT-DQN	1307.5	1095.5	664.0	586.5
BCL-RADIAL-DQN	1225.5	1225.5	1223.5	228.5
BCL-RADIAL+AT-DQN	1215.0	1093.0	1010.5	961.5

Maximum Opportunistic Skipping

BCL-C-AT-DQN vs BCL-MOS-AT-DQN

- BCL-MOS-AT-DQN significantly reduces training time (in terms of the number of training phases) and the performance is as good as BCL-C-AT-DQN.



Conclusion

In summary, we make the following contributions:

- A novel flexible **adversarial curriculum learning framework for reinforcement learning** (BCL), in which bootstrapping each phase from multiple executions of previous phase plays a key role.
- A novel opportunistic adaptive generation variant that **opportunistically skips forward** in the curriculum.
- An approach that composes interval bound propagation and FGSM-based adversarial input generation as a part of adaptive curriculum generation.
- An extensive experimental evaluation using OpenAI Gym **Atari games (DQN-style)** and **Procgen (PPO-style, Appendix)** that demonstrates significant improvement in robustness due to the proposed BCL framework.

Thank you!

Please check our poster at Hall E #915.