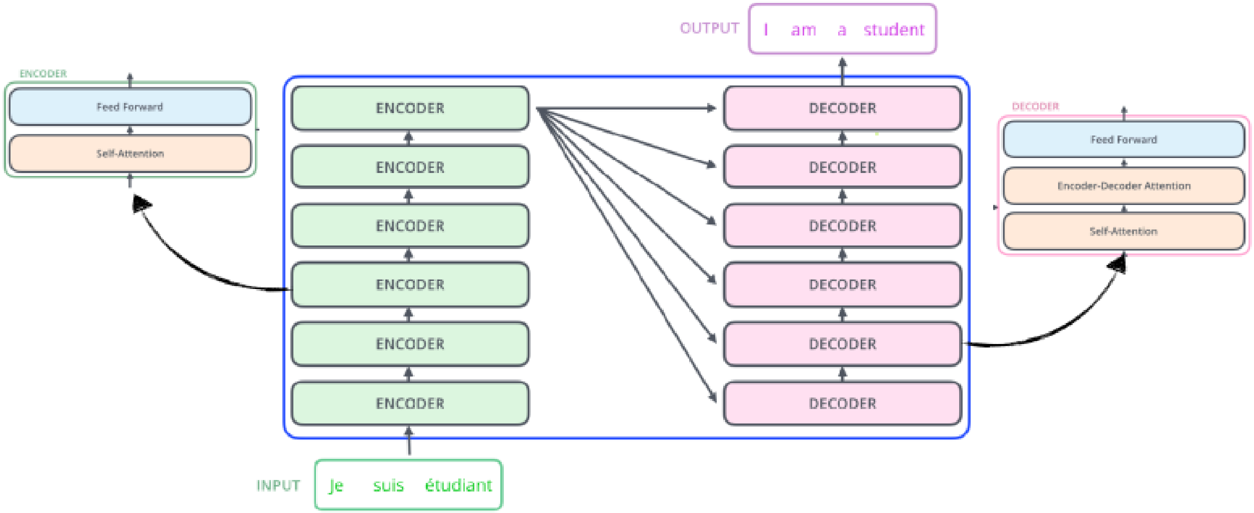
Transformers in RL: Decision Transformers vs. Traditional RL

Iman Ahmadi Amirreza Tanevardi

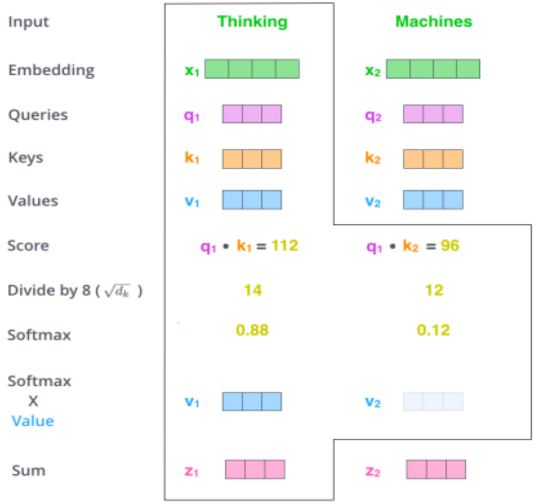
Sharif University of Technology

Transformers

Transformer Architecture



Attention as a soft-memory look up



Input
Thinking
Machines
Embedding
x_1
x_2
Queries
q_1
q_2
Keys
k_1
k_2
Values
v_1
v_2
Score
$q_1 \bullet k_1 = 112$
$q_1 \bullet k_2 = 96$
Divide by 8 ($\sqrt{d_k}$)
14
12
Softmax
0.88
0.12
Softmax
X
Value
v_1
v_2
Sum
z_1
z_2

Decision Transformer: Reinforcement Learning via Sequence Modeling

Lili Chen[*,1], Kevin Lu[*,1], Aravind Rajeswaran[2], Kimin Lee[1],
Aditya Grover[2], Michael Laskin[1], Pieter Abbeel[1], Aravind Srinivas[†,1], Igor Mordatch[†,3]
[*]equal contribution [†]equal advising
[1]UC Berkeley [2]Facebook AI Research [3]Google Brain
{lilichen, kzl}@berkeley.edu

$$s_0 \quad a_0 \quad r_0 \quad s_1 \quad a_1 \quad r_1 \quad \cdots\cdots\cdots \quad \hat{R} = \sum_{t=0}^{T-1} r_t$$

$$s_0 \quad a_0 \quad \hat{R}_0 \quad s_1 \quad a_1 \quad \hat{R}_1 \quad \cdots\cdots\cdots$$

$$\hat{R}_0 = \sum_{t=0}^{T-1} r_t \qquad \hat{R}_1 = \sum_{t=1}^{T-1} r_t$$

Methodology: Input Setup

$$\tau = (r_0, s_0, a_0, r_1, s_1, a_1, \ldots \ldots, r_T, s_T, a_T)$$

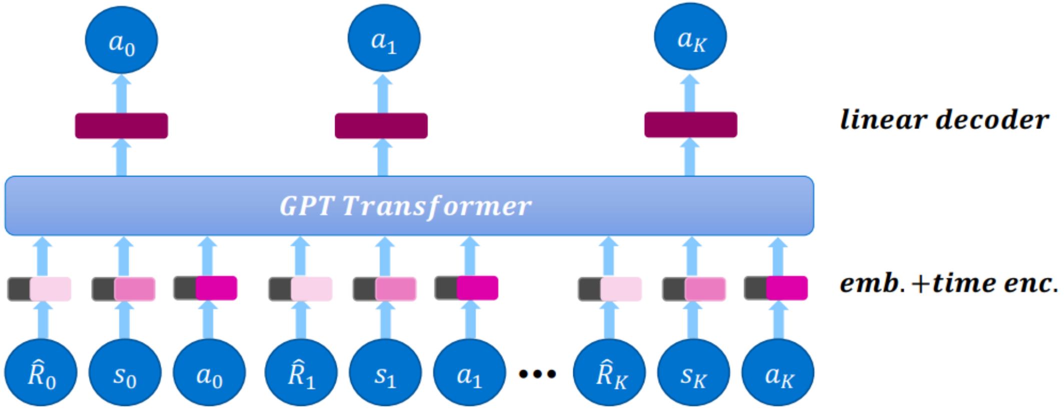
$$\downarrow$$

$$\tau = (\hat{R}_0, s_0, a_0, \hat{R}_1, s_1, a_1, \ldots \ldots, \hat{R}_T, s_T, a_T)$$

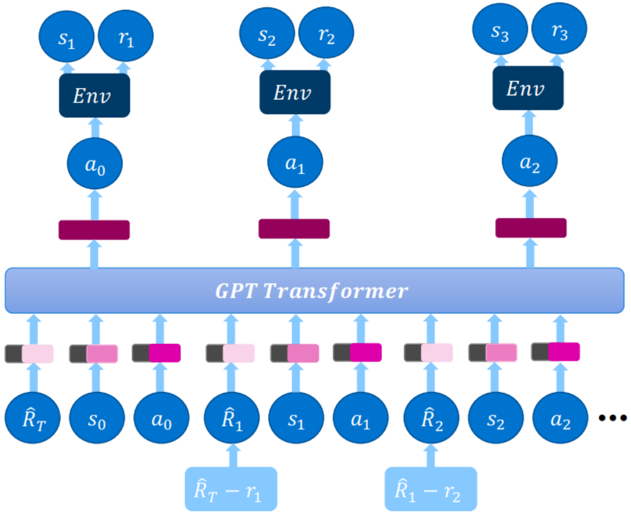
$$\hat{R}_t = \sum_{t'=t}^{T} r_{t'}$$

Rewards-to-go

Methodology: Training Pipeline



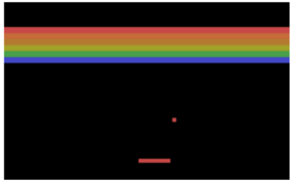
Methodology: Inference Pipeline

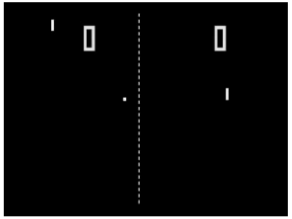


Decision Transformer

Experiments: Atari Benchmark

Baselines	Games	Challenges
• CQL [22] • REM [23] • QE-DQN [24] • BC (New)	• Breakout • Qbert • Pong (K=50) • Seaquest	• Visual Inputs • Long-term credit assignment



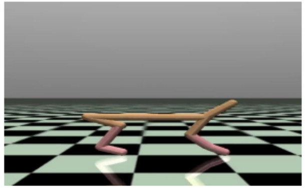


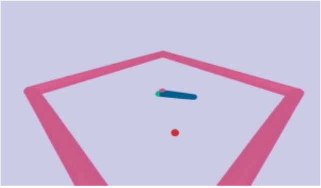
Experiments: D4RL Benchmark

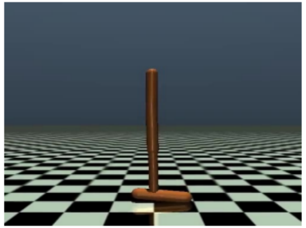
Baselines
- CQL [22]
- BEAR [25]
- BRAC [26]
- AWR [5]
- BC (New)

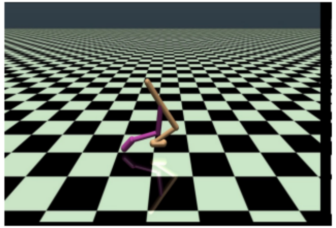
Games
- HalfCheetah
- Hopper
- Walker
- Reacher (New)

Dataset Settings
- Medium
- Medium-Replay
- Medium-Expert









Results: Atari Benchmark

Game	DT (Ours)	CQL	QR-DQN	REM	BC
Breakout	$\mathbf{267.5 \pm 97.5}$	211.1	17.1	8.9	138.9 ± 61.7
Qbert	15.4 ± 11.4	**104.2**	0.0	0.0	17.3 ± 14.7
Pong	106.1 ± 8.1	**111.9**	18.0	0.5	85.2 ± 20.0
Seaquest	$\mathbf{2.5 \pm 0.4}$	1.7	0.4	0.7	2.1 ± 0.3

Results : D4RL Benchmark

Dataset	Environment	DT (Ours)	CQL	BEAR	BRAC-v	AWR	BC
Medium-Expert	HalfCheetah	**86.8 ± 1.3**	62.4	53.4	41.9	52.7	59.9
Medium-Expert	Hopper	107.6 ± 1.8	**111.0**	96.3	0.8	27.1	79.6
Medium-Expert	Walker	**108.1 ± 0.2**	98.7	40.1	81.6	53.8	36.6
Medium-Expert	Reacher	**89.1 ± 1.3**	30.6	-	-	-	73.3
Medium	HalfCheetah	42.6 ± 0.1	44.4	41.7	**46.3**	37.4	43.1
Medium	Hopper	**67.6 ± 1.0**	58.0	52.1	31.1	35.9	63.9
Medium	Walker	74.0 ± 1.4	79.2	59.1	**81.1**	17.4	77.3
Medium	Reacher	**51.2 ± 3.4**	26.0	-	-	-	**48.9**
Medium-Replay	HalfCheetah	36.6 ± 0.8	46.2	38.6	**47.7**	40.3	4.3
Medium-Replay	Hopper	**82.7 ± 7.0**	48.6	33.7	0.6	28.4	27.6
Medium-Replay	Walker	**66.6 ± 3.0**	26.7	19.2	0.9	15.5	36.9
Medium-Replay	Reacher	**18.0 ± 2.4**	**19.0**	-	-	-	5.4
Average (Without Reacher)		**74.7**	63.9	48.2	36.9	34.3	46.4
Average (All Settings)		**69.2**	54.2	-	-	-	47.7

Q1: Does Decision Transformer perform behavior cloning on a subset of the data?

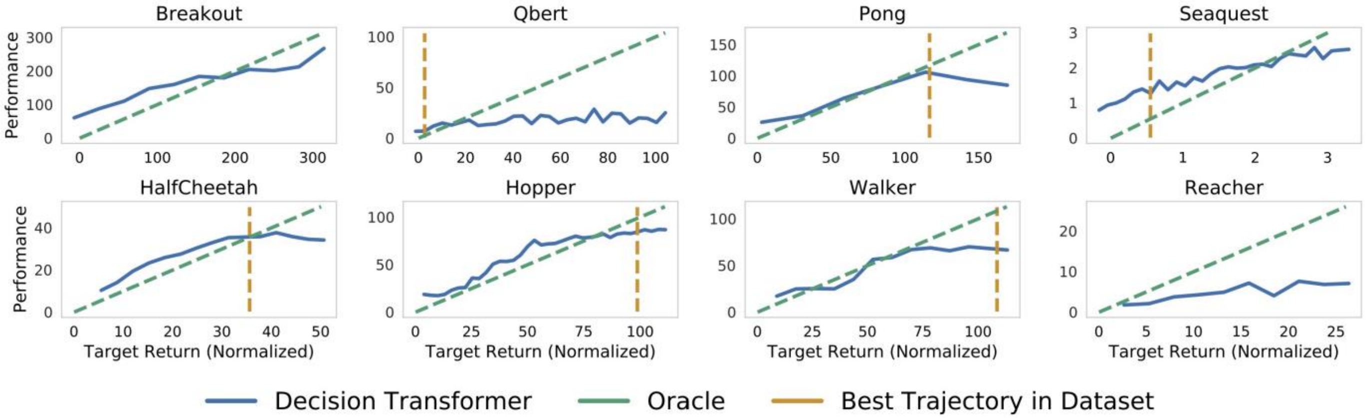
Large Dataset

Dataset	Environment	DT (Ours)	10%BC	25%BC	40%BC	100%BC	CQL
Medium	HalfCheetah	42.6 ± 0.1	42.9	43.0	43.1	43.1	**44.4**
Medium	Hopper	$\mathbf{67.6 \pm 1.0}$	65.9	65.2	65.3	63.9	58.0
Medium	Walker	74.0 ± 1.4	78.8	**80.9**	78.8	77.3	79.2
Medium	Reacher	51.2 ± 3.4	51.0	48.9	58.2	**58.4**	26.0
Medium-Replay	HalfCheetah	36.6 ± 0.8	40.8	40.9	41.1	4.3	**46.2**
Medium-Replay	Hopper	$\mathbf{82.7 \pm 7.0}$	70.6	58.6	31.0	27.6	48.6
Medium-Replay	Walker	66.6 ± 3.0	**70.4**	67.8	67.2	36.9	26.7
Medium-Replay	Reacher	18.0 ± 2.4	**33.1**	16.2	10.7	5.4	19.0
Average		56.1	**56.7**	52.7	49.4	39.5	43.5

Small Dataset

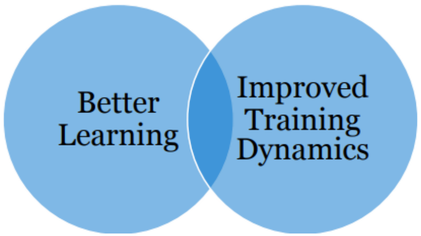
Game	DT (Ours)	10%BC	25%BC	40%BC	100%BC
Breakout	$\mathbf{267.5 \pm 97.5}$	28.5 ± 8.2	73.5 ± 6.4	108.2 ± 67.5	138.9 ± 61.7
Qbert	15.4 ± 11.4	6.6 ± 1.7	16.0 ± 13.8	11.8 ± 5.8	$\mathbf{17.3 \pm 14.7}$
Pong	$\mathbf{106.1 \pm 8.1}$	2.5 ± 0.2	13.3 ± 2.7	72.7 ± 13.3	85.2 ± 20.0
Seaquest	$\mathbf{2.5 \pm 0.4}$	1.1 ± 0.2	1.1 ± 0.2	1.6 ± 0.4	2.1 ± 0.3

Q2: How well does Decision Transformer model the distribution of returns?

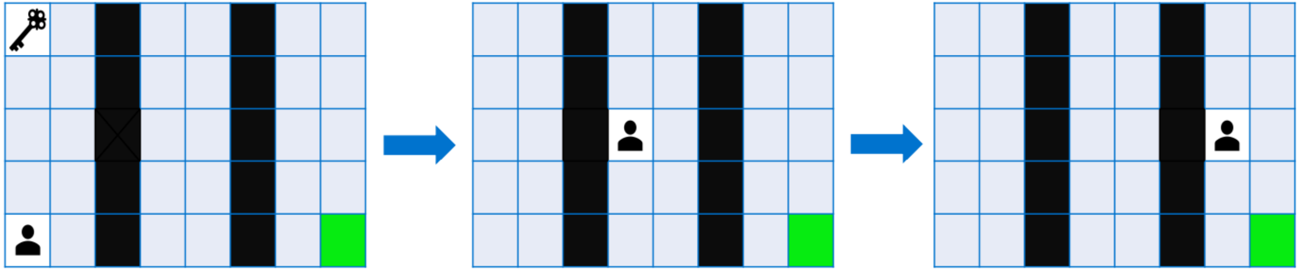


Q3: What is the benefit of using a longer context length?

Game	DT (Ours)	DT with no context ($K = 1$)
Breakout	**267.5 ± 97.5**	73.9 ± 10
Qbert	**15.1 ± 11.4**	13.6 ± 11.3
Pong	**106.1 ± 8.1**	2.5 ± 0.2
Seaquest	**2.5 ± 0.4**	0.6 ± 0.1



Q4: Does Decision Transformer perform effective long-term credit assignment?

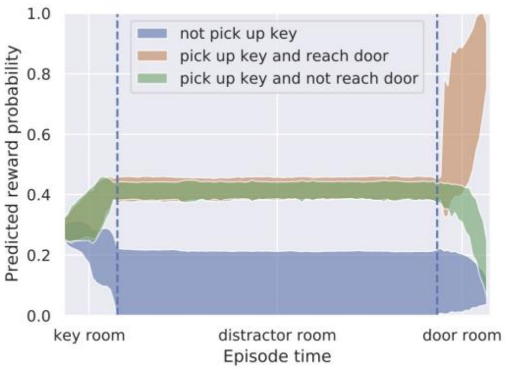


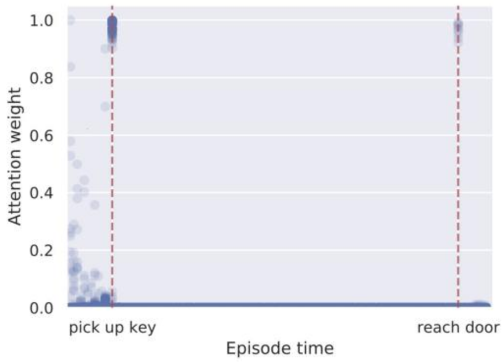
Phase 1 *Phase 2* *Phase 3*

Dataset	DT (Ours)	CQL	BC	%BC	Random
1K Random Trajectories	**71.8%**	13.1%	1.4%	69.9%	3.1%
10K Random Trajectories	94.6%	13.3%	1.6%	**95.1%**	3.1%

Q5: Can transformers be accurate critics in sparse reward settings?

Experiment Details: → Predict reward+action tokens → No initial returns-to-go





Q6: Does Decision Transformer perform well in sparse reward settings?

Experiment Details:

No rewards within trajectory

Final Cumulative reward at the end

Dataset	Environment	Delayed (Sparse) DT (Ours)	Delayed (Sparse) CQL	Agnostic BC	Agnostic %BC	Original (Dense) DT (Ours)	Original (Dense) CQL
Medium-Expert	Hopper	**107.3 ± 3.5**	9.0	59.9	102.6	107.6	111.0
Medium	Hopper	60.7 ± 4.5	5.2	63.9	**65.9**	67.6	58.0
Medium-Replay	Hopper	**78.5 ± 3.7**	2.0	27.6	70.6	82.7	48.6

Extra Observations

- No regularization or value pessimism needed
- Implicit representation of the value function
- Decision Transformer can benefit sample-efficient online regimes
- Can act as a strong model for behaviour generation

Conclusion

- Effective model-free supervised offline RL algorithm using sequence modelling.
- No reliance on any of the traditional RL concepts.
- Solves credit assignment and distribution shift problems seen in other RL algorithms.
- Match or surpass offline model-based RL state-of-the-art methods.

Limitations

Dependency on Context Length

Computational Time

Prior Knowledge on rewards

Loss of theoretical guarantees

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., & Mordatch, I. (2021). ***Decision Transformer: Reinforcement Learning via Sequence Modeling*** **(arXiv:2106.01345). arXiv. https://arxiv.org/abs/2106.01345**

Choudhury, S. (2023, Fall), ***CS 6756: Advanced Reinforcement Learning*****, Cornell University**

Any Questions?

Thank you for your time!