Offline Reinforcement Learning as One Big Sequence Modeling Problem

Michael Janner Qiyang Li Sergey Levine

University of California, Berkeley

Summary

We investigate how much of the usual machinery of RL algorithms can be replaced with the tools that have found widespread use in large-scale language modeling.

Reframing RL as a (self-)supervised prediction problem

- simplifies design decisions: no model ensembles, behavior policy constraints
- allows for use of high-capacity models trained with stable algorithms

Standard (offline) RL

Training

while not converged **do**
  Sample transition $\{\mathbf{s}_t, \mathbf{a}_t, r_t, \mathbf{s}_{t+1}\}$
  Update Q-function to minimize (regularized) Bellman error:
    $\min_\theta (r_t + \gamma V(\mathbf{s}_{t+1}) - Q_\theta(\mathbf{s}_t, \mathbf{a}_t))^2 + \text{regularization}(\theta)$
  Improve policy based on updated Q-function:
    $\max_\phi \mathbb{E}_{a \sim \pi_\phi(\cdot|s_t)} [Q_\theta(s_t, a)]$
end

Action selection

$\mathbf{a} \sim \pi_\phi(\cdot \mid \mathbf{s})$

Transformer trajectory optimization

Training

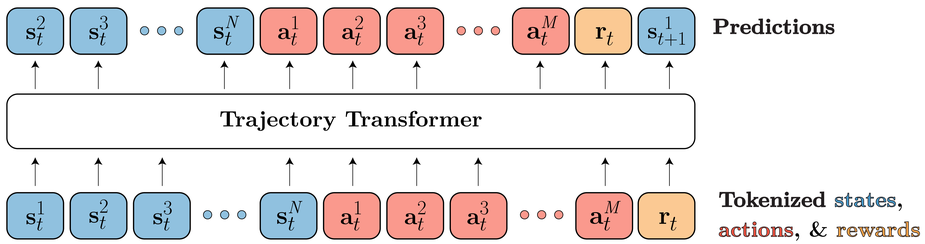
while not converged **do**
  Sample trajectory $\tau = \{\mathbf{s}_1, \mathbf{a}_1, r_1, \ldots, \mathbf{s}_T, \mathbf{a}_T, r_T\}$
  Maximize likelihood assigned to τ by transformer:
    $\max_\theta P_\theta(\tau)$
end

Action selection

Imitation: maximize $P_\theta(\tau \mid \mathbf{s}_0 = \mathbf{s})$
Goal-conditioned RL: maximize $P_\theta(\tau \mid \mathbf{s}_0 = \mathbf{s}, \mathbf{s}_T = \mathbf{g})$
Offline RL: maximize $\sum r_t$ using beam search

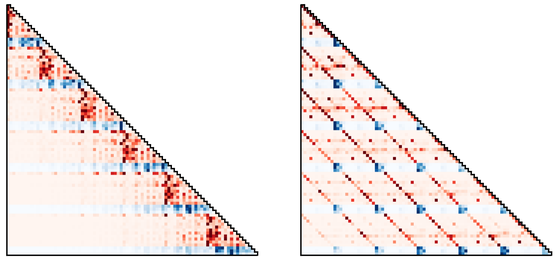
Trajectories as Generic Sequential Data

A single model serves as an action prior, dynamics model, and value estimator.



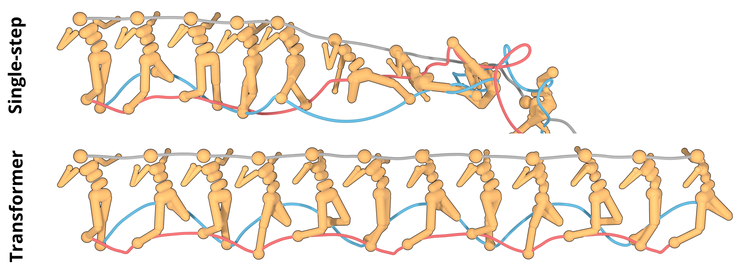
The transformer is trained on trajectories using the standard teacher-forcing objective:

$$\mathcal{L}(\tau) = \sum_{t=1}^{T} \Big(\sum_{i=1}^{N} \log P_\theta(\mathbf{s}_t^i \mid \mathbf{s}_t^{<i}, \tau_{<t}) + \sum_{j=1}^{M} \log P_\theta(\mathbf{a}_t^j \mid \mathbf{a}_t^{<j}, \mathbf{s}_t, \tau_{<t}) + \log P_\theta(r_t \mid \mathbf{a}_t, \mathbf{s}_t, \tau_{<t}) \Big)$$



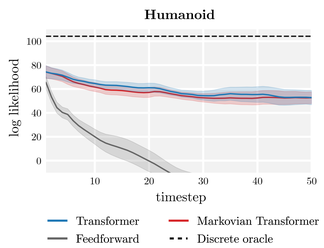
Attention patterns of the Trajectory Transformer, showing (left) a discovered Markovian stratetgy and (right) a visually striated approach with action smoothing.

Transformers as Dynamics Models



Standard predictive dynamics models have excellent single-step error, but poor long-horizon accuracy due to compounding errors

Trajectory Transformers are more reliable long-horizon predictors than single-step models, even in continuous Markovian domains



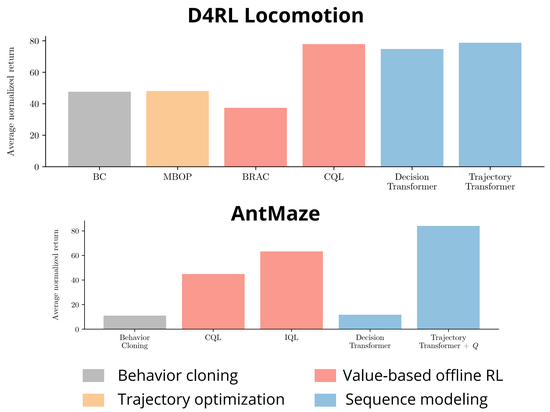
Model accuracy versus rollout length

Beam Search as a Trajectory Optimizer

Imitation learning Decoding a Trajectory Transformer with unmodified beam search gives rise to a model-based imitative method that optimizes for entire predicted trajectories to match those of an expert policy.

Goal-conditioned RL Conditioning trajectories on a future desired state alongside previously-encountered states yields a goal-reaching method.

Offline RL Replacing log-probabilities from the sequence model with reward predictions yields a model-based planning method, surprisingly effective despite lacking the details usually required to make planning with learned models effective.

Combining with Q-functions Using an IQL [Kostrikov 2021] Q-function as a planning heuristic dramatically improves performance in long-horizon, sparse-reward tasks compared to model-free and return-conditioning approaches.

Goal-reaching in the four-rooms environment Start Goal