A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning

Presenter: Christina Petlowany

10/11/2022

Motivation Videos

https://youtu.be/ks0Z0Is6GKU

https://youtu.be/LBmlxZFGxE8

Ross, Stéphane, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning." *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2011.

Motivation and Main Problem

- Most robot problems have some type of sequential nature
  - Force control
  - Manipulation
  - Navigation
  - Vision (sometimes)
  - And more!

- Robot Learning (RL) needs to account for temporal error accumulations
  - Especially where expert demonstrations do not cover the entire state space

Problem Setting

- Sequential problems are not **Independent and Identically Distributed**
  - The future state depends on the action input
- This is a problem in the Imitation Learning (IL) domain where expert demonstrations do not cover all possible perturbations
  - Existing approaches compound errors resulting from mistakes over time

with i.i.d -> Error $\leq \varepsilon T$

without i.i.d -> Error $\propto \varepsilon T^2$

https://supertuxkart.net/Gallery

Imitation Learning

- Implemented in cases where the reward is complex
    - Learn a reward from demonstrations ✅
    - Explicitly specify a reward ⟶ how would you design a reward function for SuperTuxKart?
        - Need to go fast overall
        - Might need to slow down for curves
        - Avoid other vehicles (but not always!)
        - Stay on the road
        - Drifting?

https://www.youtube.com/watch?v=V7CY68zH6ps

Distribution mismatch

- Training dataset != test dataset
- In this scenario, occurs when the errors accumulate in the test environment and the test environment no longer reflects the expert demonstrations
- Not always solved by adding information

De Haan, Pim, Dinesh Jayaraman, and Sergey Levine. "Causal confusion in imitation learning." *Advances in Neural Information Processing Systems* 32 (2019).

Masiha, Mohammad Saeed, et al. "Learning under distribution mismatch and model misspecification." *2021 IEEE International Symposium on Information Theory (ISIT)*. IEEE, 2021.

Related Work

- Existing supervised learning approach
  - Error ∝ εT^2
- Forward Training (Ross and Bagnell 2010)
  - Trains a policy at each time step
  - These policies are trained on the expected distribution of states for that time step
  - Very computationally expensive, must have T policies

Related Work (continued)

- SMILe (Ross and Bagnell 2010)
  - Switch between executing the trained policy and the policy of the expert
  - Can stop training at any time and remove the expert inputs

Ross, Stéphane, and Drew Bagnell. "Efficient reductions for imitation learning." *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

DAgger (Dataset Aggregation)

- Iteratively trains policies from expert demonstrations to expand the dataset

$$\begin{tikzcd} P_e \arrow[r] \arrow[r, bend right, "T_e"'] & D \arrow[r] & P_2 \arrow[r] \arrow[r, bend right, "T_2"'] & D \arrow[r] & P_3 \arrow[r] \arrow[r, bend right, "T_3"'] & D \arrow[r] & \dots \arrow[r] & P_{n+1} \end{tikzcd}$$

- Asks the experts for labeling/demonstrations when necessary based on relevant expected states from the new trained policy

DAgger (continued)

Initialize $\mathcal{D} \leftarrow \emptyset$.
Initialize $\hat{\pi}_1$ to any policy in Π.
for $i = 1$ **to** N **do**
  Let $\pi_i = \beta_i \pi^* + (1 - \beta_i)\hat{\pi}_i$.
  Sample T-step trajectories using π_i.
  Get dataset $\mathcal{D}_i = \{(s, \pi^*(s))\}$ of visited states by π_i and actions given by expert.
  Aggregate datasets: $\mathcal{D} \leftarrow \mathcal{D} \bigcup \mathcal{D}_i$.
  Train classifier $\hat{\pi}_{i+1}$ on $\mathcal{D}$.
end for
Return best $\hat{\pi}_i$ on validation.

DAgger (continued)

- Proofs for how the error is limited
- if N is $\tilde{O}(T)$: $E \le \varepsilon_n + O(1/T)$
  - ε_n is the true loss of the policy
- if N is $O(T^2 \log(1/\delta))$: with probability $(1-\delta)$, $E \le \varepsilon_n + O(1/T)$
  - ε_n is the training loss
- Proofs to show guarantee finding policy with ε surrogate loss

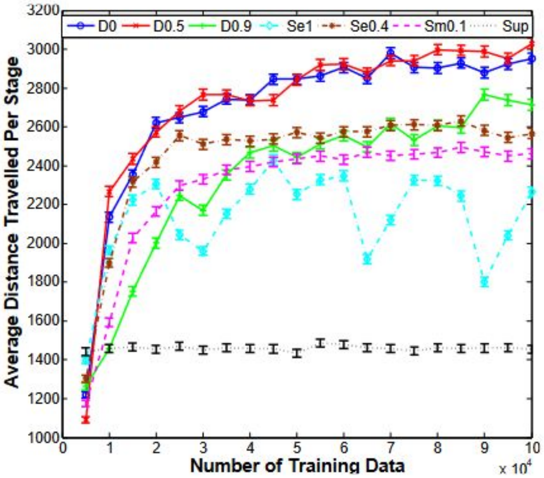
Experimental Results Videos

https://www.youtube.com/watch?v=V00npNnWzSU

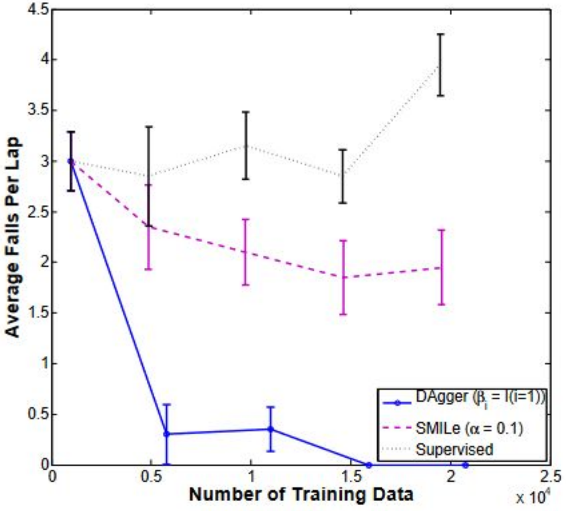
https://www.youtube.com/watch?v=anOI0xZ3kGM

Experimental Results

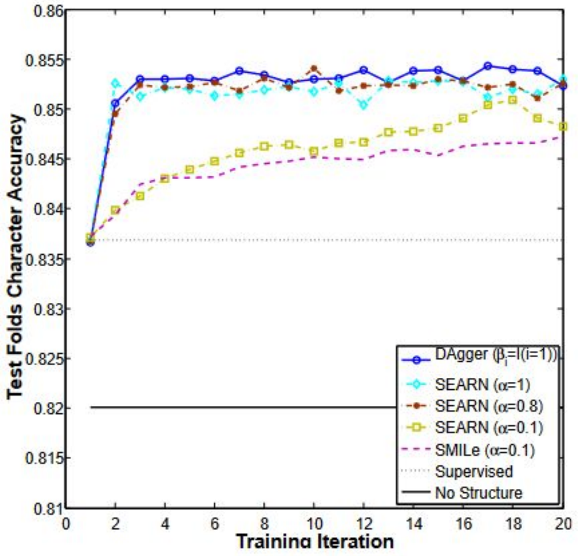
Super Mario Bros.



SuperTuxKart



Handwritten Word Recognition



Critique

- The authors mainly cite their own work which leads me to question the generability of their efforts

Proof. We here follow a similar proof to Ross and Ba (2010). Given our policy π, consider the policy $\pi_{1:t}$, which executes π in the first t-steps and then execute the expert π^*. Then

We here provide a theorem slightly more general than the one provided by Ross and Bagnell (2010) that applies to

- Very expensive, requires availability of expert

Future Work

- ❖ More sophisticated ways for generating the trajectories
- ❖ Reducing reliance on expert input
- ❖ See extended readings

Extended Readings

Brown, Daniel, et al. "Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations." *International conference on machine learning*. PMLR, 2019.

Kober, Jens, J. Andrew Bagnell, and Jan Peters. "Reinforcement learning in robotics: A survey." *The International Journal of Robotics Research* 32.11 (2013): 1238-1274.

Ho, Jonathan, and Stefano Ermon. "Generative adversarial imitation learning." *Advances in neural information processing systems* 29 (2016).

Bengio, Samy, et al. "Scheduled sampling for sequence prediction with recurrent neural networks." *Advances in neural information processing systems* 28 (2015).

Amodei, Dario, et al. "Concrete problems in AI safety." *arXiv preprint arXiv:1606.06565* (2016).

Ross, Stéphane, and Drew Bagnell. "Efficient reductions for imitation learning." *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings, 2010.

Summary

- Errors accumulate over time and IL is particularly susceptible to this
- DAgger trains policies and adds relevant demonstrations to the dataset
- DAgger shows significant performance improvements with small increases in training iterations

https://supertuxkart.net/Gallery