- IRL
- Offline RL
    - Adapt to large untapped dataset
- Representation learning
    - Observation encoder, action decoder
- Transfer learning

Elements to read up on

- ~~Offline RL~~
- ~~Transfer learning~~
    - ~~Distillation learning~~
- ~~Inverse RL from suboptimal demonstrations~~
- Representation learning
- Curriculum learning

The proposal should be a maximum of **one page**, and should describe the idea that you plan to explore for your project and the members of your group. One proposal should be submitted per group. Your proposal should at minimum answer the following questions:

1. Which tasks or problems will you study? Where will you get your data or simulator (or real-world system)?
2. What is the main research hypothesis your project will investigate? All projects should at least attempt to evaluate novel ideas that pertain to deep RL or its applications.
3. How does the topic of your project relate to deep RL?

Unsupervised learning for RL:
https://www.youtube.com/watch?v=YqvhDPd1UEw&list=PLwRJQ4m4UJjPiJP3691u-qWwPGVKzSlNP&index=12

Curriculum Learning:
https://lilianweng.github.io/lil-log/2020/01/29/curriculum-for-reinforcement-learning.html
https://arxiv.org/abs/2006.02689
https://www.researchgate.net/publication/339873123_Curriculum_Learning_for_Reinforcement_Learning_Domains_A_Framework_and_Survey

Unsupervised curricula for meta-RL: https://arxiv.org/pdf/1912.04226.pdf
Diversity is all you need: https://arxiv.org/pdf/1802.06070.pdf
Variational Option Discovery Algorithms (VALOR): https://arxiv.org/pdf/1807.10299.pdf
IR-VIC: https://arxiv.org/pdf/1907.10580.pdf

Project Proposal

Introduction

We are particularly interested in how unsupervised agents can learn useful skills by exploring their environments. We want to study approaches for improving reward-free optional discovery in reinforcement learning. Specifically, existing variational option discovery techniques like VALOR, VIC, DIAYN, etc. can utilize curriculum learning to increase skills, and can utilize auxiliary objectives like UNREAL's pixel control or an inverse dynamics loss to encourage learning mapped skills that are effective in the given environment. We plan on using OpenAI Gym as our simulator.

In order to optimize existing variational option discovery techniques, we plan to investigate more principled curriculum learning schedules for increasing the number of discrete contexts in the distribution, as well as methods to simplify the continuous-context case in order to apply curriculum learning. We will explore various combinations of variational option discovery algorithms with curriculum learning techniques and auxiliary losses to determine which adjustments, if any, lead to significant improvements from previous results.

Research Hypothesis

We predict that more careful curriculum learning approaches will increase the information learned through unsupervised skill discovery, and that auxiliary losses associated with inverse dynamics predictions will lead to more qualitatively "natural" learned skills than in previous unregularized experiments.

Topic Relevance

Domains with unknown or underspecified rewards remain challenging for deep RL algorithms, despite the amount of information available to be captured in these experiences. A domain-agnostic unsupervised skill discovery algorithm which improves data efficiency for deep RL algorithms via curriculum learning or auxiliary loss functions would benefit few-shot learning on natural tasks, and could complement work in transfer learning to apply information gained from unsupervised simulation and observation to physical skill mastery.

---

try different curriculum generation techniques with different variational option discovery algorithms
Goal: work better in higher dimensional environments, reduce “unnaturalness”
Maybe different curriculum generation techniques will work better with different variational option discovery algorithms?
We propose that

Things To Try

1. Try reproducing VALOR, VIC, DIAYN, CARML results with/without curriculum trick on OpenAI Gym tasks (should look somewhat like these examples)
2. Experiments
    a. Tweak curriculum approach
        i. VALOR’s curriculum approach: start training with small K (where learning is easy), and gradually increase it over time as the decoder gets stronger
            1. "leads to faster and more stable convergence"
        ii. Perhaps try different approaches to choosing curriculum, ex. set up a teacher agent to choose contexts
    b. Change architecture for decoder
        i. Perhaps use a transformer instead of a bi-LSTM
    c. Introduce and weight auxiliary predictive losses to learn richer state representations
        i. Ex: CPCv2, MoCo, SimCLR, Student-Teacher variations
        ii. Hypothesis: state encodings will be more impactful for complex environments
        iii.
    d.

CPC aux loss variant:

- Given a context frame, assign weights to all other frames according to some function of time and rollout ID
    - Ex: Gaussian centered at 5 seconds in the future or MoG with fwd/bwd weights
    - Different rollout ID → possibly weighted by same function if assuming rollouts are sufficiently similar, or else can be uniform
- Sample frames according to weight distribution on current rollout and other rollout, learning supervised classification of states based on same rollout or different rollout

Hyperparams:

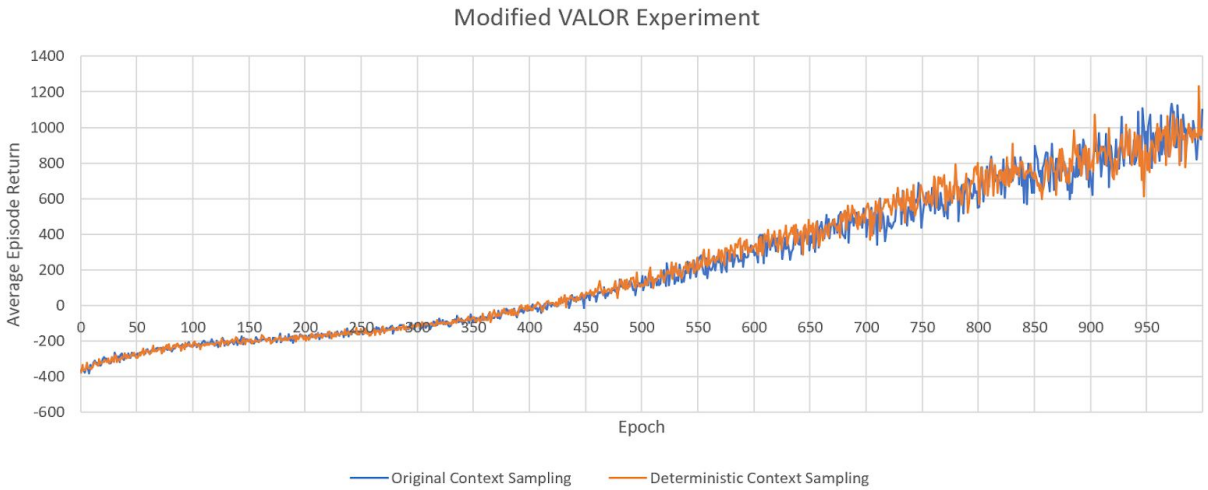
- Double-sided or single-sided
- Sample weight distribution (can parametrize, e.g. mean(s) of Gaussian or MoG)
- Weight on

To Do List

- ✓ Clone Steven Ho's implementations (maybe just VALOR for now)
- Each of us make sure we can run it / results are expected as from the paper
- Try to tweak something small from there and generate a graph of it for milestone report
    - Possible changes: decoder architecture, actor/critic
    - Write about it
- Write rest of milestone report
- Develop a plan for more proposed tweaks

Milestone Report

Experiments Conducted



Our initial experiment modified the way contexts are sampled in VALOR to be deterministic. The plot demonstrates that this modification did not cause significant differences in the episode return, and differences in stability could be attributed to noise. These differences make sense, given the small number of contexts currently utilized, but provide a helpful basis for future experiments that choose contexts more intelligently.

Changes

We have not significantly changed our problem statement. However, since writing our proposal, we have further developed our research direction and have a clearer idea of our hypothesis and the experiments we have conducted and will conduct.

In writing our initial proposal, we thought there could be room for improvement in selecting the way that the curriculum over contexts is imposed during training. VALOR, which is one of the works that is inspiring this project, proposes a rather simple curriculum - when the expected log probabilities of a context given a trajectory is high enough, the algorithm samples from a wider swath of contexts (i.e. including more difficult ones). In the same vein as CARML, we believe it would be valuable to experiment how applying different curriculum strategies may increase skills, notably in categories including teacher-student interaction.

We've also decided to primarily focus on such architectural experiments, rather than significantly trying to improve qualitatively "natural" learned skills. Therefore, we will broaden our idea of adding auxiliary losses to an analysis of multiple combinations of state representations and standard RL algorithms. A more compact, feature-rich state space should help to prevent against overparameterization and allow our algorithm to more easily maximize the variational lower bound proposed in VALOR.