<https://arxiv.org/ftp/arxiv/papers/2202/2202.08444.pdf>

Domain adaptation, task adaption

Domain generalization, task generalization

Adaptation refers to its ability to learn faster without re-training from scratch and generalization refers to its ability to extrapolate beyond the learned knowledge to tackle unseen environments.

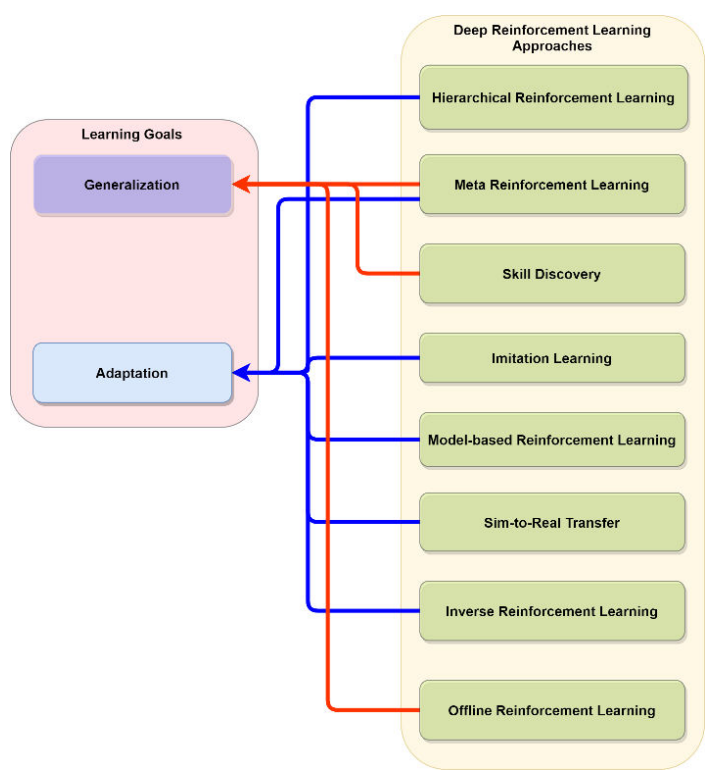
* Used as learning goals to evaluate algorithm’s learnability

Markov Decision Process (MDP):

* (S, A, p\_s, r)
* S: discrete or continuous state space
* A: discrete or continuous action space
* p\_s: transition function
* r: reward function
* P\_s and r are used to model the environment

Policy

* pi(a | s)
* Defines behavior of an agent by mapping a state s\_t in S to an action a\_t in A
* Deterministic or stochastic (preferred)



Hierarchical Reinforcement Learning

* Solve complex problems/ tasks that constitute simpler related problems/ tasks
* Transition policy gradient update for 2 level hierarchical network
  + Manager network that learns latent state- space and sets goals
  + Worker network that produces primitive actions based on goals from manager
* Multi level reasoning: method for learning hierarchy of policies
  + Low level policies: taking actions in environment
  + High level policies: planning long term decisions
  + High sample efficiency by training agent off- policy for complex tasks like robotic locomotion and object manipulation in simulation
* Priors for capturing learned knowledge by low level policies in a simple environment which can be transferred to a similar problem in a more complex environment
* Tackled problem of sparse reward using self supervision: efficient intrinsic option discovery to obtain higher rewards for the task

Meta Reinforcement Learning

* Learning to learn to generalize to unseen tasks/ domains
* Achieve faster adaptation to unseen tasks without learning from scratch by utilizing past experience
* Update rule defined in learning algorithms require re- iterative approach to tune parameters for each task
  + Meta learning provides a way to learn the rules for learning
  + Gradient based update in standard meta learning methods are expensive, degradation in generalization performance (ability to learn after given number of updates)
* Meta objective’s constraint to similar geometry as of the learner
  + Bootstrapped to assimilate information about learning dynamics into meta- objective
* Meta objective’s limited ability to generalize within given number of steps and failing to incorporate future dynamics
  + Minimizing distance to bootstrapped target using KL divergence

Skill Discovery

* Skill: latent conditioned policy that can be trained to perform useful tasks in a sparse/ unknown reward environment
* Variational inference based option discovery method for training agent to discover and learn skills through environment interaction without needing to maximizing cumulative reward
* Introduced a novel hierarchical RL algorithm which is capable of learning in a continuous, low level, latent space
  + Model capable of predicting output of learned skills
  + Crucial role in solving more complex tasks at a higher level
* Task generation is a major challenge for achieving multi task solving ability by an agent
* Priors help in deciding which skill is more important to explore while performing a particular action

Imitation Learning

* Agent to learn by observing an expert demonstrating the required task
* Off- policy Actor- critic algorithm that focuses on learning to imitate the agent's past good experiences (replay buffer)
* Effective way to solve autonomous driving by learning from human driving demonstrations
  + Allows agent to learn the preferences and goals of humans in a safe manner
* Distributional shift is a fundamental problem in imitation learning
  + Gives rise to causal misidentification effect that leads to setting wrong correlation between actions and their causes
  + Algorithm that learns mapping from causal graphs to the policies and then utilizes knowledge of experts to select correct policy for target tasks
* Providing high confidence bound on agent’s performance with respect to expert’s demonstration is a less- explored area
  + Most use Bayesian Inverse RL to measure reward uncertainty and error over policy generalization
  + Suffers from complex computation of MDPs leading to hindrance against safety and efficiency of model in unknown MDPs or high dimensional problems
  + Providing preferences over goals
* Generative Intrinsic Reward driven by Imitation Learning: better than expert performance from one expert demonstration
  + VAE to generate diverse future states and corresponding action latent variables
* Confidence aware imitation learning if expert demonstrations are insufficient (learned policies are suboptimal)
  + Learns policy and confidence value for every state action pair in expert’s demonstration

Model based RL

* Model free RL has shown tremendous success for solving tasks in simulated environment but they are highly sample inefficient
* Model based is sample efficient by performing policy optimization against learned dynamics of environment (can be used in real world)
* SimPLE algorithm
* Online RL typically requires iterative collection of experiences during training
  + Deployment efficiency: records number of increments in the data collection policy during training
  + Behavior Regularized Model Ensemble: aims to learn an ensemble of environment dynamics models

Sim2Real Transfer

* Develop a technique that allows agent to adapt to an unseen domain
  + Train in sim and test in real
  + Domain randomization
  + Complexities in real domain such as contact dynamics, soft bodies, and hidden information in general prevent optimal transfer
* Algorithm that can learn disentangled representation of the environment’s generative factors and utilize this information to learn a robust source policy that can be transferred to target domain (real world)
* Most modern RL involves robotics manipulation on rigid objects but deformable objects is an important yet underexplored area
  + Large configuration space change involved in manipulating deformable objects

Inverse RL

* Understanding objectives and rewards of agent by observing its behavior
  + Infers reward function from rollouts of expert policy -> policy improvement and generalization
  + Knowledge of reward is a primary goal -> apprenticeship is commonly used to acquire such policy learning from expert
* Use apprenticeship learning to utilize prior knowledge and experiences from actions of an expert to derive a probability distribution over space of reward functions
* Fails to adapt to locally consistent constraints
  + Divide complex task into several subtasks with corresponding set of local constraints

Offline RL

* Policy learning from pre- collected dataset of trajectories/ experiences for tasks where agent is not allowed to interact with environment
  + Overcome practical limitation of online RL such as expensive interaction
* Poor off policy evaluation causes inaccurate Q value estimation
  + Distribution shift: causes evaluation error and propagates to evaluation step of current policy
  + Iterative error: error between Q estimate’s errors and tend to overestimate each step, reusing data causes amplification of such error

Discussion

* Distributional shifts in data between training and testing domain is one of the primary causes behind shortcomings in RL research
  + Most focus on task adaption/ domain adaptation but important to focus on generalization
* Large, diverse datasets have allowed application of Offline RL to achieve generalization
  + Hard to scale because of longer training time and mostly applied to simulation or controlled environments
  + Need real world datasets and incorporate contextual information about environment and other agents
* Meta RL requires less training time to achieve faster adaptability/ generalizability
  + Only in controlled environments
  + Measure their reliability before deployed in open world environments
  + Skill discovery algorithms have also demonstrated ability to perform generalization
* Autonomous vehicles are an example of open world environment
  + Current RL research fails to incorporate notion of safety for undesirable situations