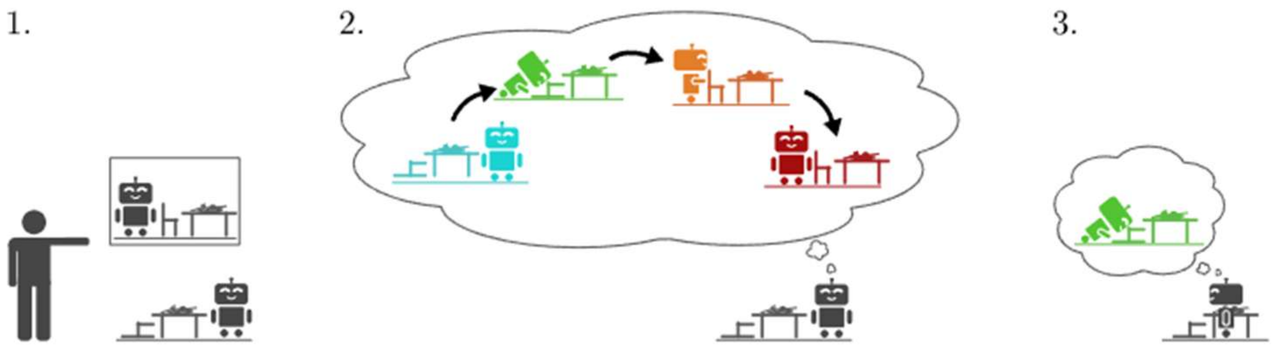
Planning with Goal-Conditioned Policies



- publish: NeurIPS 2019
- motivate: 通过feasible vector找到sub-goal

Planning with Goal-Conditioned Policies

- 定义feasible value: $V^{\pi}(\mathbf{s}, \mathbf{g}, t) = \mathbb{E}\left[\sum_{t'=t}^{T_{\max}} R(\mathbf{s}_{t'}, \mathbf{g}, t') \mid \mathbf{s}_t = \mathbf{s}, \pi \text{ is conditioned on } \mathbf{g}\right]$
- 根据feasible value定义feasible vector:

$$\vec{\mathbf{V}}(\mathbf{s}, \mathbf{g}_{1:K}, t_{1:K+1}, \mathbf{g}) = \begin{bmatrix} V(\mathbf{s}, \mathbf{g}_1, t_1) \\ V(\mathbf{g}_1, \mathbf{g}_2, t_2) \\ \vdots \\ V(\mathbf{g}_{K-1}, \mathbf{g}_K, t_K) \\ V(\mathbf{g}_K, \mathbf{g}, t_{K+1}) \end{bmatrix}$$

- Loss: $\mathcal{L}(\mathbf{g}_{1:K}) = ||\vec{\mathbf{V}}(\mathbf{s}, \mathbf{g}_{1:K}, t_{1:K+1}, \mathbf{g})||$

Planning with Goal-Conditioned Policies

- 由于 $\mathbf{g}_{1:K}$ 的维度可能很高, 因此使用VAE的方案找到latent space来降维:

$$\psi(\mathbf{z}) = \arg\max_{\mathbf{g}'} p_\theta(\mathbf{g}' \mid \mathbf{z})$$

$$\vec{\mathbf{V}}(\mathbf{s}, \mathbf{z}_{1:K}, t_{1:K+1}, \mathbf{g}) = \begin{bmatrix} V(\mathbf{s}, \psi(\mathbf{z}_1), t_1) \\ V(\psi(\mathbf{z}_1), \psi(\mathbf{z}_2), t_2) \\ \vdots \\ V(\psi(\mathbf{z}_{K-1}), \psi(\mathbf{z}_K), t_K) \\ V(\psi(\mathbf{z}_K), \mathbf{g}, t_{K+1}) \end{bmatrix}$$

- 最终, loss为:

$$\mathcal{L}_{\text{LEAP}}(\mathbf{z}_{1:K}) = ||\vec{\mathbf{V}}(\mathbf{s}, \mathbf{z}_{1:K}, t_{1:K+1}, \mathbf{g})||_p - \lambda \sum_{k=1}^{K} \log p(\mathbf{z}_k)$$

f-Policy Gradients: A General Framework for Goal Conditioned RL using f-Divergences

- publish: NeurIPS 2023
- motivate: 通过与环境的交互来识别解决任务的优化行为
- 与Planning with Goal-Conditioned Policies不同, 这篇文章的主要目的是根据goal去优化policy, 而不是找到sub-goal

f-Policy Gradients: A General Framework for Goal Conditioned RL using f-Divergences

f-divergence	$D_f(P\|\|Q)$	$f(u)$	$f'(u)$	$f'(\infty)$
FKL	$\int P(x)\log\frac{P(x)}{Q(x)}dx$	$u\log u$	$1+\log u$	Undefined
RKL	$\int Q(x)\log\frac{Q(x)}{P(x)}dx$	$-\log u$	$-\frac{1}{u}$	0
JS	$\frac{1}{2}\int P(x)\log\frac{2P(x)}{P(x)+Q(x)}+ Q(x)\log\frac{2Q(x)}{P(x)+Q(x)}dx$	$u\log u- (1+u)\log\frac{1+u}{2}$	$\log\frac{2u}{1+u}$	$\log 2$
χ^2	$\frac{1}{2}\int Q(x)(\frac{P(x)}{Q(x)}-1)^2dx$	$\frac{1}{2}(u-1)^2$	u	Undefined

- Distribution Matching:
    - 把state和goal转换成distribution
    - 使用f-Divergence让两个policy的最终state和goal state尽量接近
    - 将f-Divergence作为loss训练policy module

f-Policy Gradients: A General Framework for Goal Conditioned RL using f-Divergences

- Loss: $J(\theta) = D_f(p_\theta(s)||p_g(s))$

- gradient:

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim p_\theta(\tau)} \Big[\Big[\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) \Big] \Big[\sum_{t=1}^{T} f'\Big(\frac{p_\theta(s_t)}{p_g(s_t)}\Big) \Big] \Big]$$

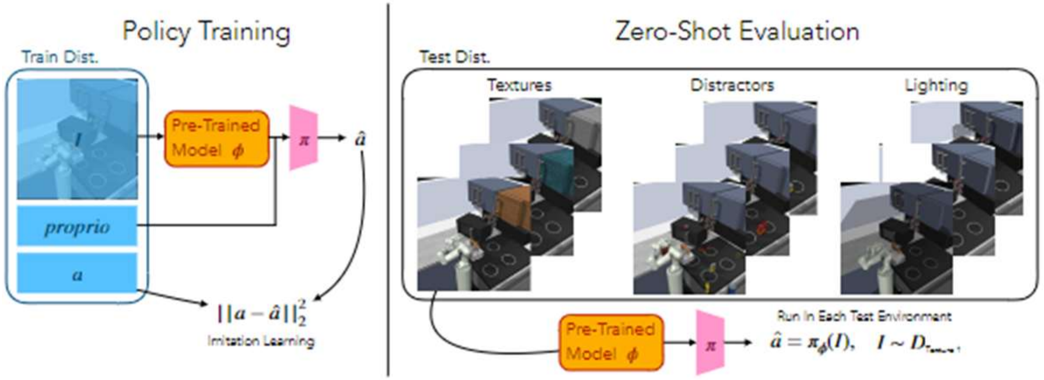
- proximal policy optimization:

$$\nabla_\theta J(\theta) = \mathbb{E}_{s_t, a_t \sim p_{\theta'}(s_t, a_t)} \Big[\min(r_\theta(s_t) F_{\theta'}(s_t), clip(r_\theta(s_t), 1-\epsilon, 1+\epsilon) F_{\theta'}(s_t)) \Big]$$

Algorithm 1 f-PG

Let, π_θ be the policy, G be the set of goals, B be a buffer
for $i = 1$ **to** num_iter **do**
  $B \leftarrow []$
  for $j = 1$ **to** num_traj_per_iter **do**
    Sample g, set $p_g(s)$
    Collect goal conditioned trajectories, $\tau : g$
    Fit $p_\theta(s)$ using KDE on τ
    Store $f'\Big(\frac{p_\theta(s)}{p_g(s)}\Big)$ for each s in τ
    $B \leftarrow B + \{\tau : g\}$
  end for
  for $j = 1$ **to** num_policy_updates **do**
    $\theta \leftarrow \theta - \alpha \nabla_\theta J(\theta)$ (Equation 6)
  end for
end for

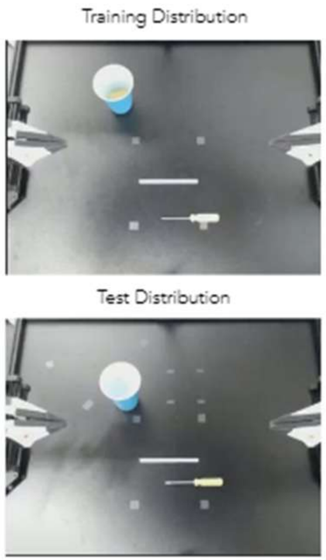
What makes Pre-trained Visual Representations Successful for Robust Manipulation



- publish: CoRL 2024
- 研究模型和数据集在各方面对泛化能力的影响

What makes Pre-trained Visual Representations Successful for Robust Manipulation

- Environment:
  - FrankaKitchen
  - Meta-World
- Distribution Shift
  - Light
  - Textural
  - Object
- Policy Train
  - Imitate Learning



What makes Pre-trained Visual Representations Successful for Robust Manipulation

- R3M 和 VIP 显出优于baseline(ImageNet)

- ImageNet学到的特征, 即使冻结也能在各种模拟控制任务中与真实状态信息竞争
- 风格化 ImageNet 更好
- 监督的存在对泛化性的影响不如其他因素

- 增强集的选择比监督的重要性更大

- 架构选择方面, ViTs 比 ResNets 略有优势