LAMDA
Learning And Mining from DatA
http://lamda.nju.edu.cn

南京大学
NANJING UNIVERSITY

Multi-Agent Imitation Learning for Driving Simulation

韩馥光

2020.2.20

Introduction

- 模仿学习在单智能体环境中可以学习驾驶策略，但是在多智能体的环境中，这些在单智能体环境中学到的策略往往表现不好，因为这样做再次引入了covariate shift
- 出现这种问题的原因是训练和测试时的观测是从不同分布中采样的

于是该文章结合GAIL和Parameter Sharing TRPO提出了PS-GAIL的算法来解决上述问题

Definition

joint observation space： $\mathcal{O} \in \mathbb{R}^{O \times M}$

joint action space： $\mathcal{A} \in \mathbb{R}^{A \times M}$ 其中M是Agent的数量

若agents是同质的，即每个agent的策略的映射是相同的，那么联合策略分布

$\pi = p(a_0, \ldots, a_M \mid o_0, \ldots, o_M)$ 可以分解为 $\prod_i p(a_i \mid o_i)$

Assumption

Homogeneous agents：每个agent的O和A相同

$$\mathcal{O}_i = \mathcal{O}_j \text{ and } \mathcal{A}_i = \mathcal{A}_j \ \forall \text{ agents } i, j$$

Independent rewards：每个agent的reward function只和本身的状态和行动有关

$$\mathcal{R}_i(s, a_1, \ldots, a_i, \ldots, a_k) = \mathcal{R}_i(s, a_i)$$

Identical reward function: 每个agent的reward function是一样的

$$\mathcal{R}_i = \mathcal{R}_j \ \forall \text{ agents } i, j$$

这些假设只是理想化的，但是可以做近似，并且结果表明可以学到实际的驾驶策略

Algorithm

Algorithm 1 PS-GAIL

Input: Expert trajectories $\tau_E \sim \pi_E$, Shared policy parameters Θ_0, Discriminator parameters ψ_0, Trust region size Δ_{KL}

for $k \leftarrow 0, 1, \ldots$ **do**

Rollout trajectories for all agents $\vec{\tau} \sim \pi_{\theta_k}$

Score $\vec{\tau}$ with critic, generating reward $\tilde{r}(s_t, a_t; \psi_k)$

Batch trajectories obtained from all the agents

Take a TRPO step to find $\pi_{\theta_{k+1}}$ maximizing Eq. (5)

Update the critic parameters ψ by maximizing Eq. (4)

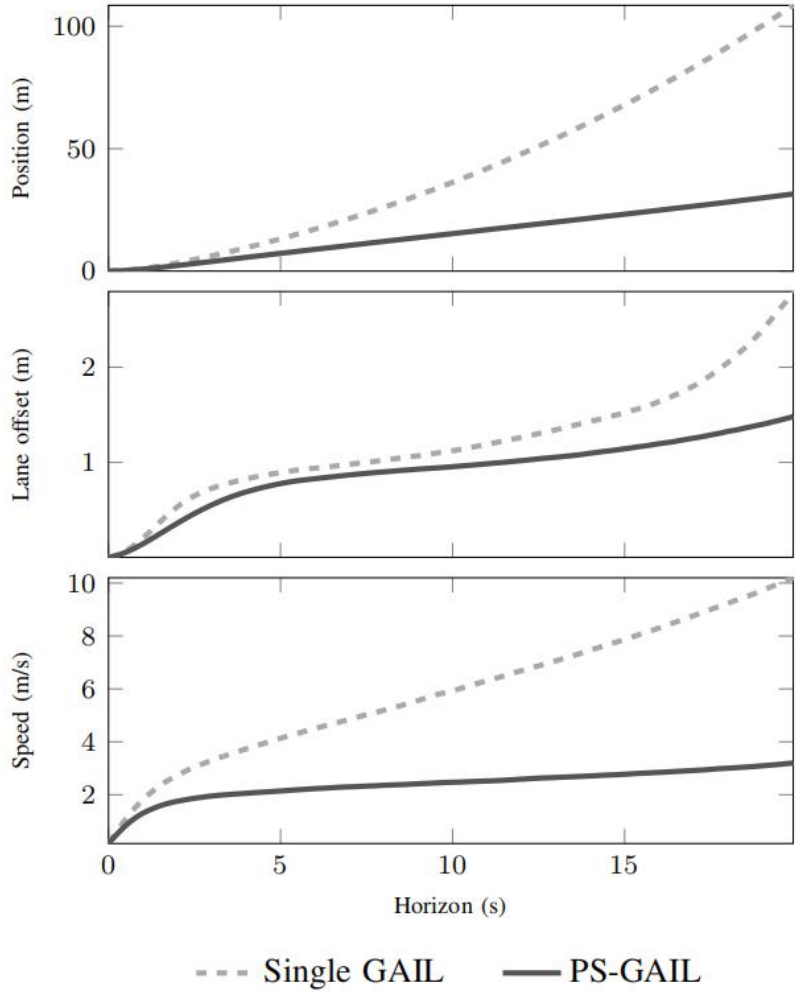
end for

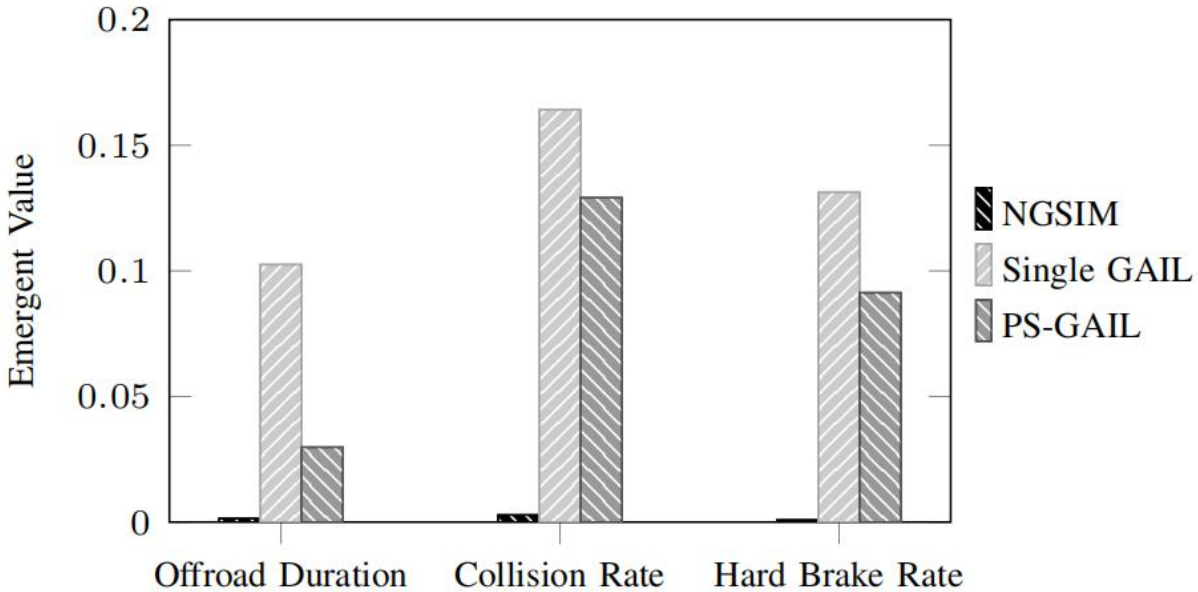
$$\min_{\theta} \max_{\psi} \mathbb{E}_{\pi_E}[D_\psi(s,a)] - \mathbb{E}_{\pi_\theta}[D_\psi(s,a)], \tag{4}$$

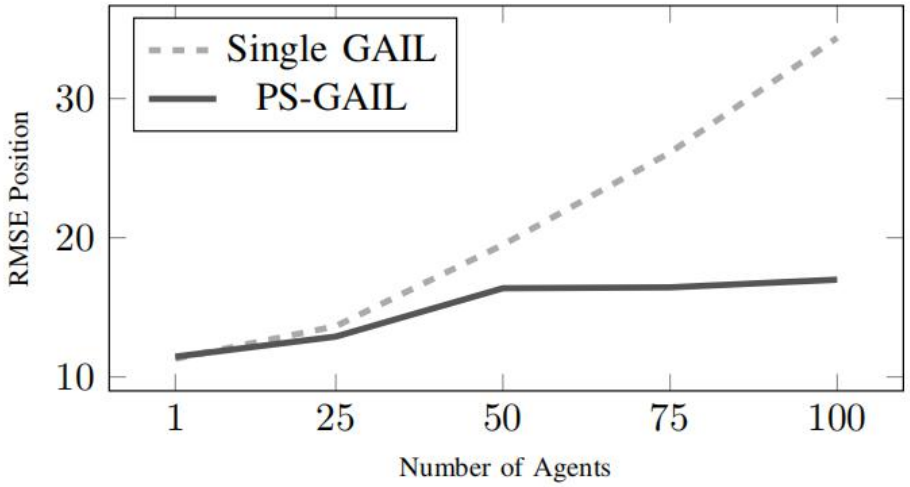
这个优化目标是对GAIL中的优化目标的的一个改进

使用Wasserstein距离而不是JS散度可以缓解梯度消失问题，并且能得到更好的policy

Experiments







Thanks!