GAIL extensions

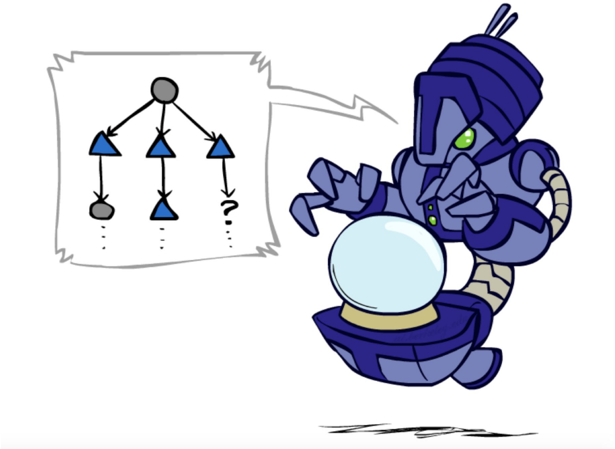
Léonard Boussioux

1. InfoGAIL
2. GAIL for BabyAI
3. Some other ideas

InfoGAIL : Interpretable Imitation Learning from Visual Demonstrations

NIPS 2017, Yunzhu Li, Jiaming Song, Stefano Ermon

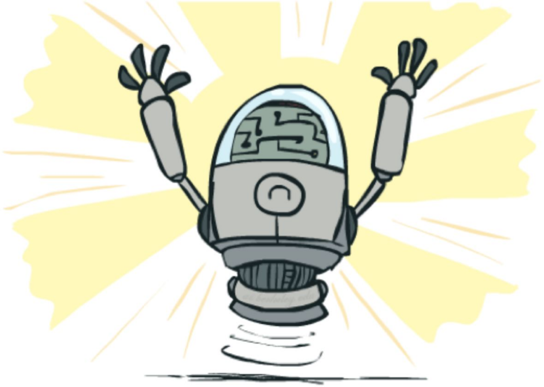
Problem : the expert policy is a mixture of expert policies.

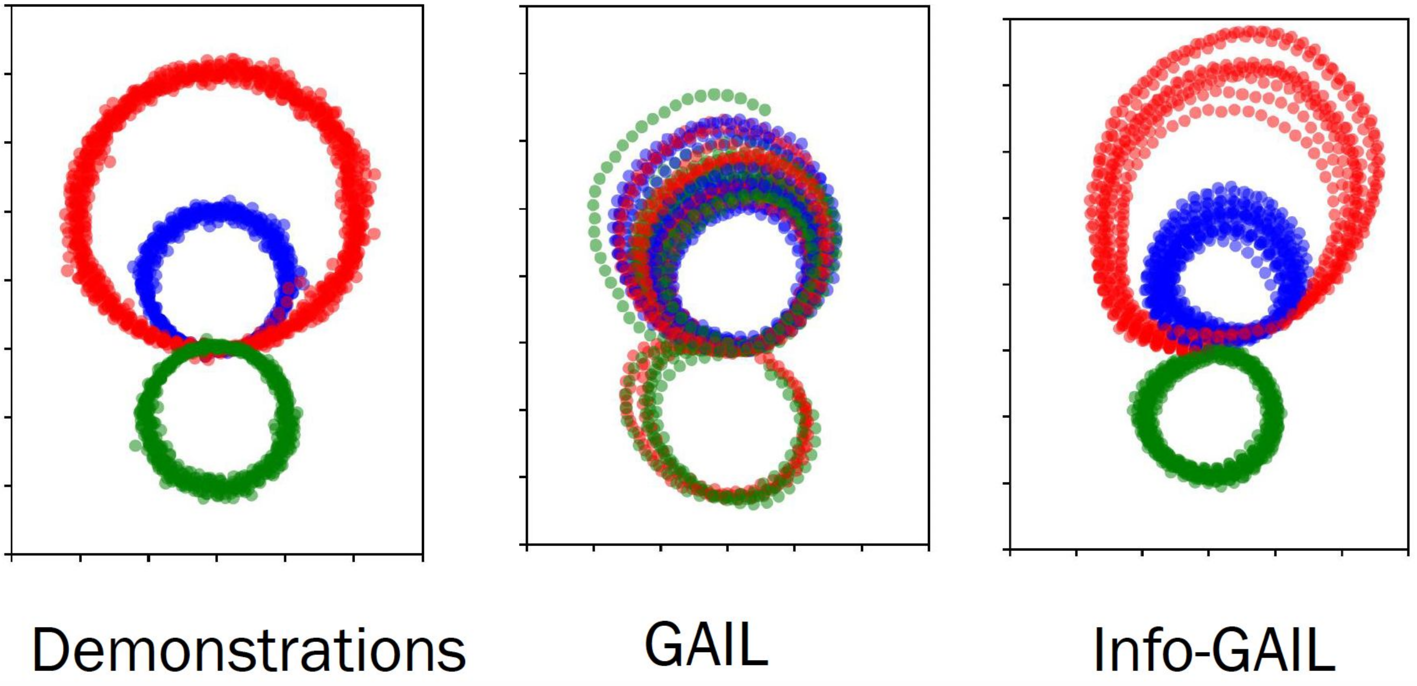


Goal : recover $\pi(a|s, c)$

where **c** is a discrete latent variable that selects a specific policy from the mixture of expert policies through $p(\pi|c)$

➔ disentangle salient latent factors of variation underlying expert demonstrations without supervision



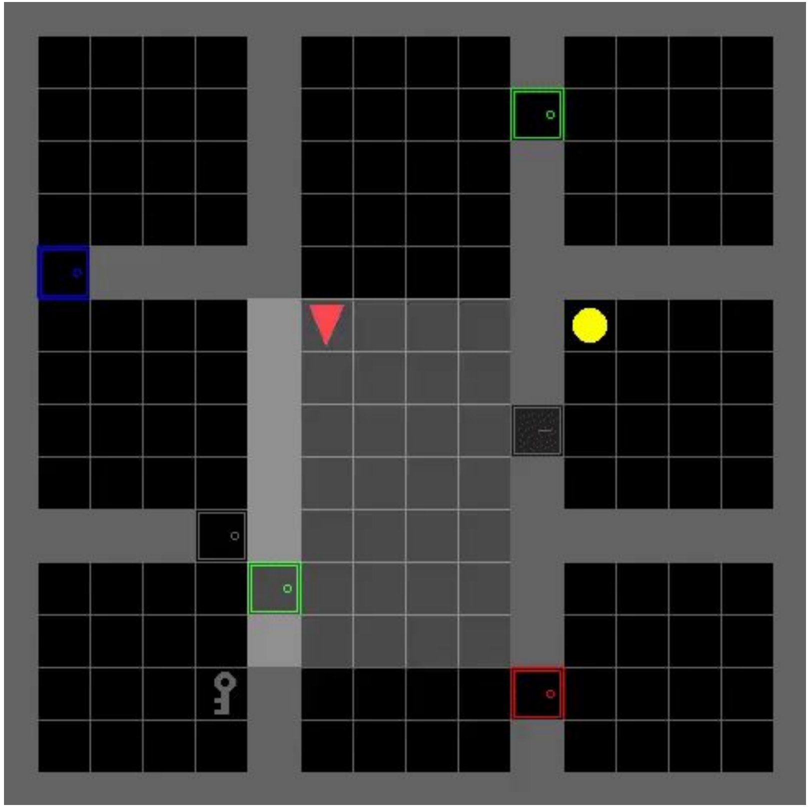


Source : InfoGAIL paper

GAIL: fails to capture the latent structure, assuming that the demonstrations are generated from a single expert ➔ tries to learn an average policy.

InfoGAIL successfully distinguishes expert behaviors and imitates each mode accordingly.

GAIL for BabyAI



Goal : improve the sample efficiency of imitation learning

- partially observable environment
- requires to perform sub-tasks

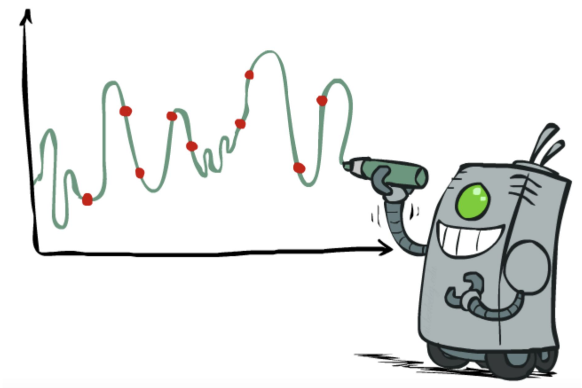
Some other ideas

End-to-End Differentiable Adversarial Imitation Learning

ICML 2017, Nir Baram, Oron Anschel, Itai Caspi, Shie Mannor

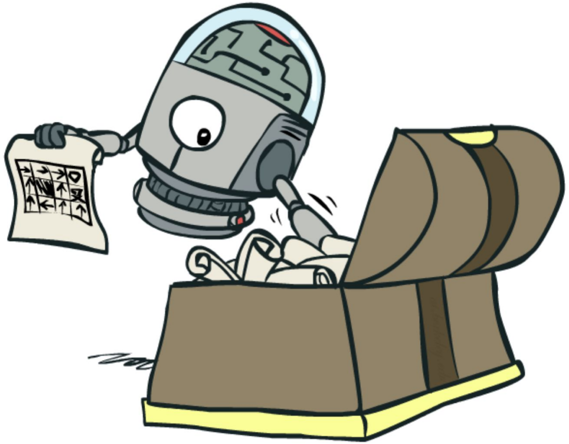
GAIL : model-free setup, generative model no longer differentiable end-to-end
➜ high-variance gradient estimation.

➜ Model-based GAIL, fully differentiable

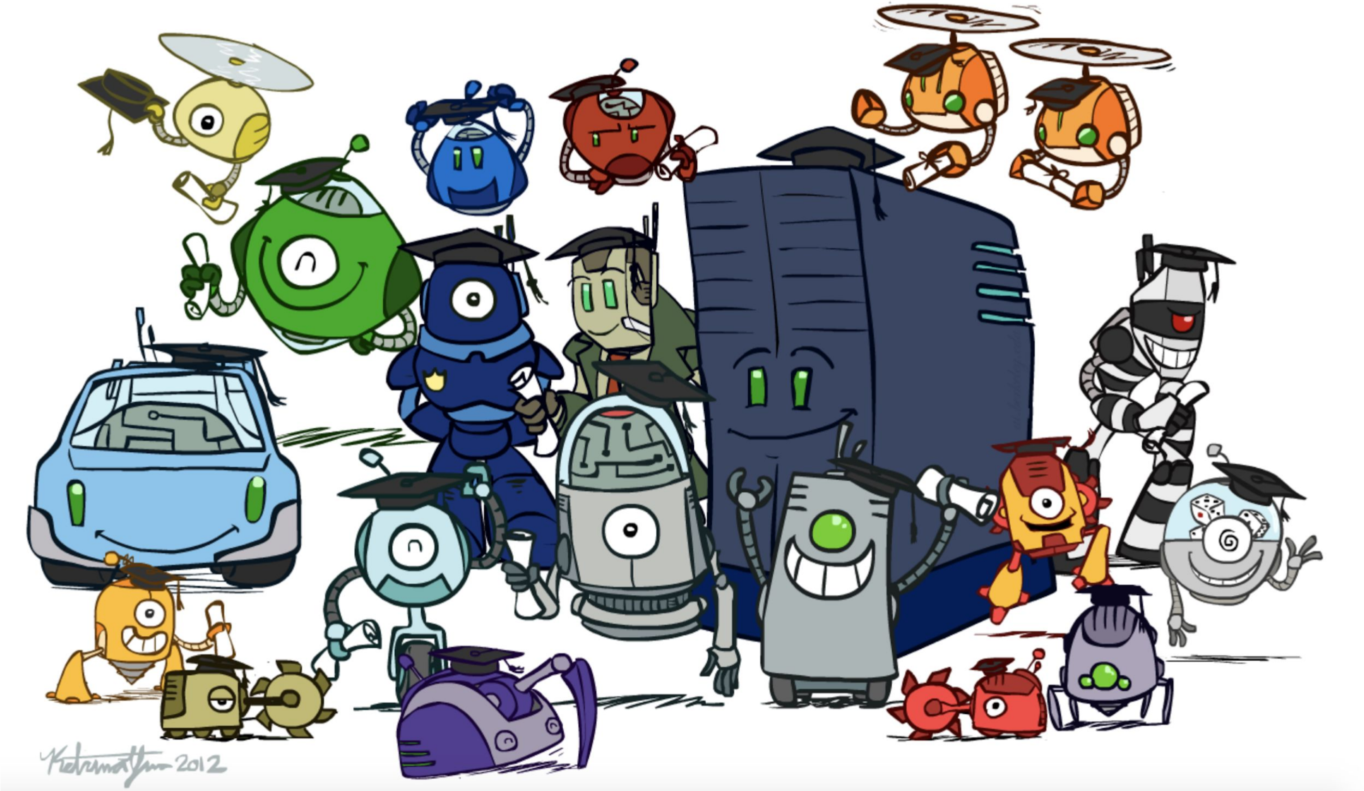


Generative Adversarial Self-Imitation Learning (ICLR 2019 reject)

➜ encourage the agent to imitate past good trajectories



Thank you for your attention!



References

InfoGAIL : Interpretable Imitation Learning from Visual Demonstrations

NIPS 2017, Yunzhu Li, Jiaming Song, Stefano Ermon
https://arxiv.org/pdf/1703.08840.pdf

End-to-End Differentiable Adversarial Imitation Learning

ICML 2017, Nir Baram, Oron Anschel, Itai Caspi, Shie Mannor
http://proceedings.mlr.press/v70/baram17a/baram17a.pdf

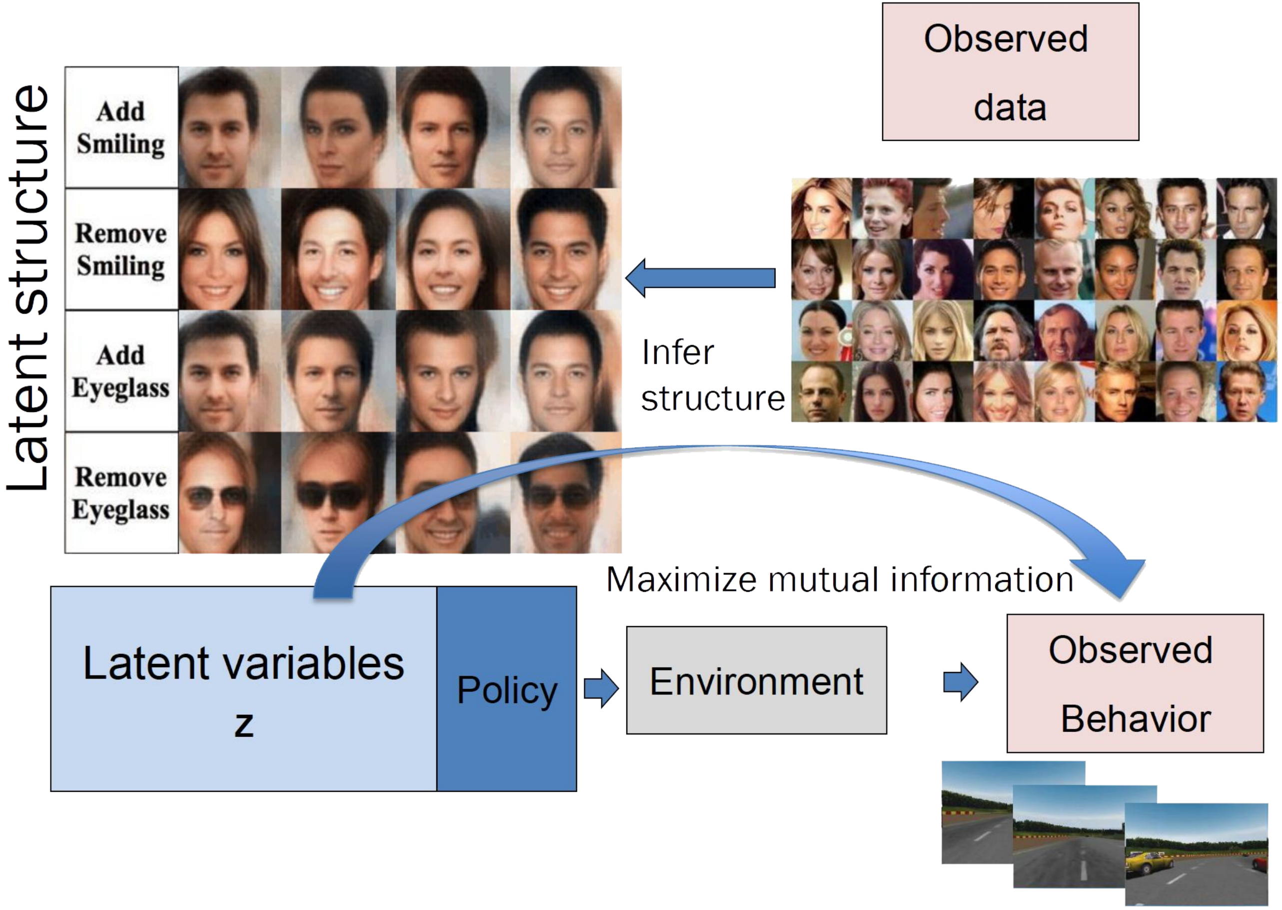
Generative Adversarial Self-Imitation Learning (ICLR 2019 reject)
Junhyuk Oh, Yijie Guo, Satinder Singh, Honglak Lee
https://openreview.net/forum?id=HJeABnCqKQ

BabyAI: First Steps Towards Grounded Language Learning With a Human In the Loop

Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, Yoshua Bengio, ICLR 2019

https://arxiv.org/abs/1810.08272

Appendix



Observed data
Latent structure
Add Smiling
Remove Smiling
Add Eyeglass
Remove Eyeglass
Infer structure
Maximize mutual information
Latent variables z
Policy
Environment
Observed Behavior

$$\min_{\pi, Q} \max_{D} \mathbb{E}_{\pi}[\log D(s,a)] + \mathbb{E}_{\pi_E}[\log(1 - D(s,a))] - \lambda_1 L_I(\pi, Q) - \lambda_2 H(\pi)$$

$$\begin{aligned} L_I(\pi, Q) &= \mathbb{E}_{c \sim p(c), a \sim \pi(\cdot|s,c)}[\log Q(c|\tau)] + H(c) \\ &\leq I(c; \tau) \end{aligned}$$

approximation of the true posterior p(clt)

Figure 2. **(a)** Block-diagram of the model-free approach: given a state s, the policy outputs μ which is fed to a stochastic sampling unit. An action a is sampled, and together with s are presented to the discriminator network. In the backward phase, the error message δ_a is *blocked* at the stochastic sampling unit. From there, a high-variance gradient estimation is used (δ_{HV}). Meanwhile, the error message δ_s is flushed. **(b)** Discarding δ_s can be disastrous as shown in the following example. Assume some $\{s, a\}$ pairs produced by the expert and G. Let $s = (x_1, x_2)$, and $a \in \mathbb{R}$. **(c)** Assuming the expert data lies in the upper half-space ($x_1 > 0$) and the policy emits trajectories in the lower half-space ($x_1 < 0$). Perfect discrimination can be achieved by applying the following rule: $sign(1 \cdot x_1 + 0 \cdot x_2 + 0 \cdot a)$. Differentiating w.r.t the three inputs give: $\frac{\partial D}{\partial x_1} = 1, \frac{\partial D}{\partial x_2} = 0, \frac{\partial D}{\partial a} = 0$. Discarding the partial derivatives w.r.t. x_1, x_2 (the state), will result in zero information gradients.