

1707.02747 - Robust Imitation of Diverse Behaviors

- **Yunqiu Xu**
- Other reference:
 - 最前沿：机器人学习Robot Learning之模仿学习Imitation Learning的发展
 - DeepMind发表物理智能最新研究：如何在仿真环境中生成灵活行为
 - Robust Imitation of Diverse Behaviors
- I think this is similar to DART: broaden the demonstration via noise injection → make learning more robust

1. Introduction

- Challenge for imitation learning:
 - Supervised learning, VAE (behavior cloning):
 - Can model diverse behaviors without dropping modes
 - Not robust, hard to handle agent trajectory diverges from the demonstrations
 - Need large training datasets for non-trivial tasks
 - Generative adversarial imitation learning
 - Can learn more robust policies with fewer demonstrations
 - More difficult to train: oscillating / model collapse
- Our work
 - Combine SL and GAIL
 - SL: new VAE for supervised imitation → learn semantic policy embeddings
 - GAIL
 - More robust than supervised learning
 - Avoid model collapse
- The model learns, from a moderate number of demonstration trajectories
 - A semantically well structured embedding of behaviors
 - A corresponding multi-task controller that allows to robustly execute diverse

behaviors from this embedding space

- An encoder that can map new trajectories into the embedding space and hence allows for one-shot imitation.

2. A Generative Modeling Approach to Imitating Diverse Behaviors

- Behavior cloning with VAE

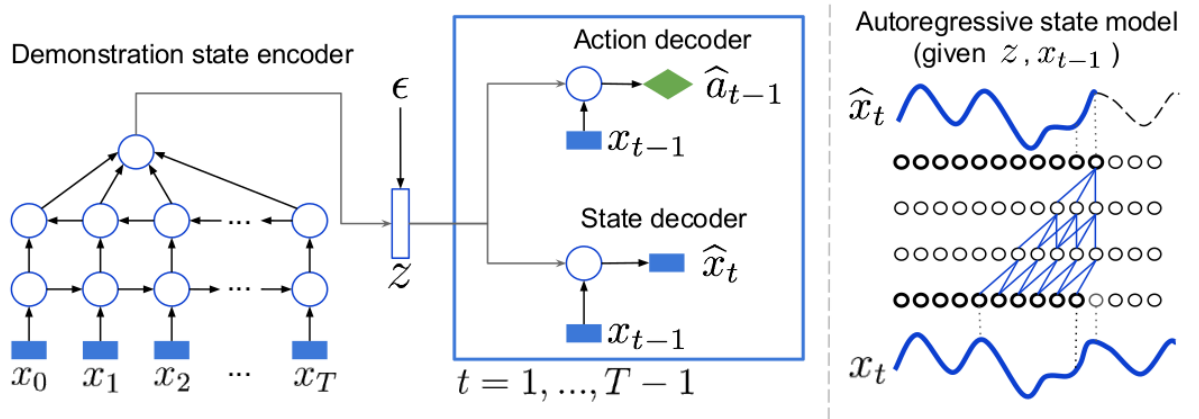


Figure 1: Schematic of the encoder-decoder architecture. **LEFT:** Bidirectional LSTM on demonstration states, followed by action and state decoders at each time step. **RIGHT:** State decoder model within a *single* time step, that is autoregressive over the state dimensions.

- Try to minimize

$$\mathcal{L}(\alpha, w, \phi; \tau_i) = -\mathbb{E}_{q_\phi(z|x_{1:T_i}^i)} \left[\sum_{t=1}^{T_i} \log \pi_\alpha(a_t^i | x_t^i, z) + \log p_w(x_{t+1}^i | x_t^i, z) \right] + D_{KL}(q_\phi(z|x_{1:T_i}^i) || p(z))$$

- Diverse generative adversarial imitation learning

- Enable GAIL to produce diverse solutions
- Discriminator:

$$\max_{\psi} \mathbb{E}_{\tau_i \sim \pi_E} \left\{ \mathbb{E}_{q(z|x_{1:T_i}^i)} \left[\frac{1}{T_i} \sum_{t=1}^{T_i} \log D_\psi(x_t^i, a_t^i | z) + \mathbb{E}_{\pi_\theta} [\log(1 - D_\psi(x, a | z))] \right] \right\}. \quad (4)$$

- Value function

$$\min_G \max_D V(G, D) = \int_y p(y) \int_z q(z|y) \left[\log D(y|z) + \int_{\hat{y}} G(\hat{y}|z) \log(1 - D(\hat{y}|z)) d\hat{y} \right] dy dz.$$

- Pseudo code

Algorithm 1 Diverse generative adversarial imitation learning.

INPUT: Demonstration trajectories $\{\tau_i\}_i$ and VAE encoder q .

repeat

for $j \in \{1, \dots, n\}$ **do**

 Sample trajectory τ_j from the demonstration set and sample $z_j \sim q(\cdot | x_{1:T_j}^j)$.

 Run policy $\pi_\theta(\cdot | z_j)$ to obtain the trajectory $\hat{\tau}_j$.

end for

 Update policy parameters via TRPO with rewards $r_t^j(x_t^j, a_t^j | z_j) = -\log(1 - D_\psi(x_t^j, a_t^j | z_j))$.

 Update discriminator parameters from ψ_i to ψ_{i+1} with gradient:

$$\nabla_\psi \left\{ \frac{1}{n} \sum_{j=1}^n \left[\frac{1}{T_j} \sum_{t=1}^{T_j} \log D_\psi(x_t^j, a_t^j | z_j) \right] + \left[\frac{1}{\widehat{T}_j} \sum_{t=1}^{\widehat{T}_j} \log(1 - D_\psi(\hat{x}_t^j, \hat{a}_t^j | z_j)) \right] \right\}$$

until Max iteration or time reached.

3. Experiments

- Robotic arm reaching

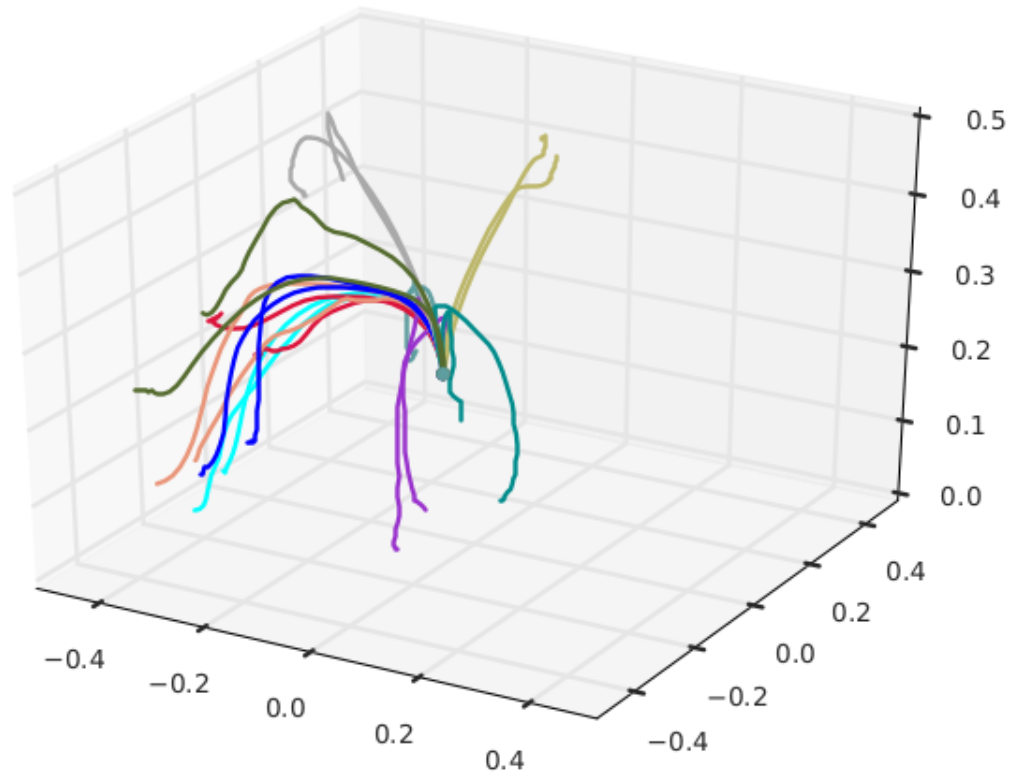


Figure 2: Trajectories for the Jaco arm's end-effector on test set demonstrations. The trajectories produced by the VAE policy and corresponding demonstration are plotted with the same color, illustrating that the policy can imitate well.

- 2D Walker

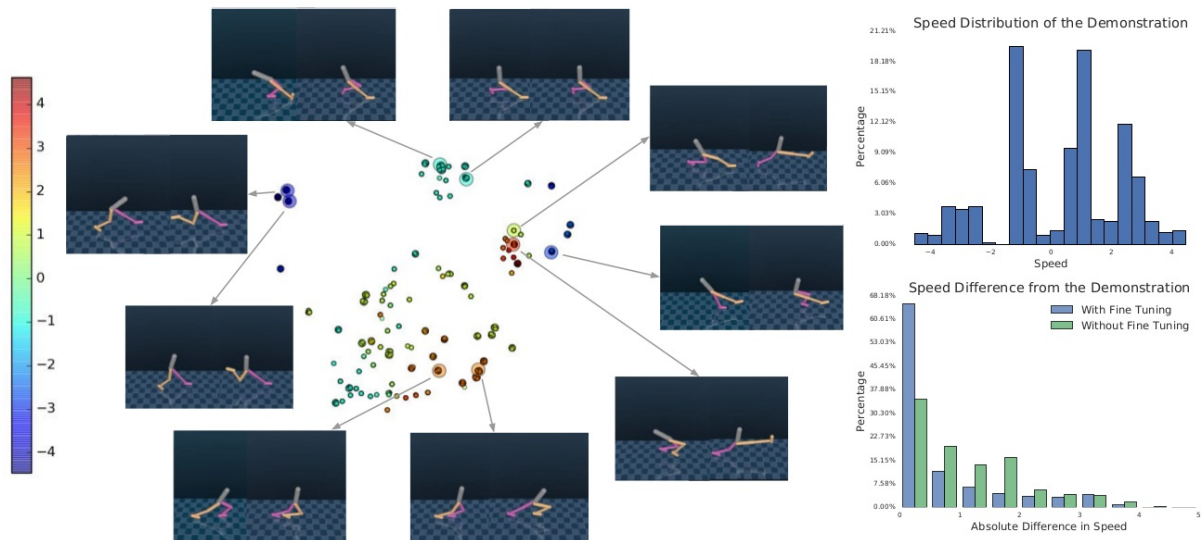


Figure 4: **LEFT:** t-SNE plot of the embedding vectors of the training trajectories; marker color indicates average speed. The plot reveals a clear clustering according to speed. Insets show pairs of frames from selected example trajectories. Trajectories nearby in the plot tend to correspond to similar movement styles even when differing in speed (e.g. see pair of trajectories on the right hand side of plot). **RIGHT, TOP:** Distribution of walker speeds for the demonstration trajectories. **RIGHT, BOTTOM:** Difference in speed between the demonstration and imitation trajectories. Measured against the demonstration trajectories, we observe that the fine-tuned controllers tend to have less difference in speed compared to controllers without fine-tuning.

- Complex humanoid

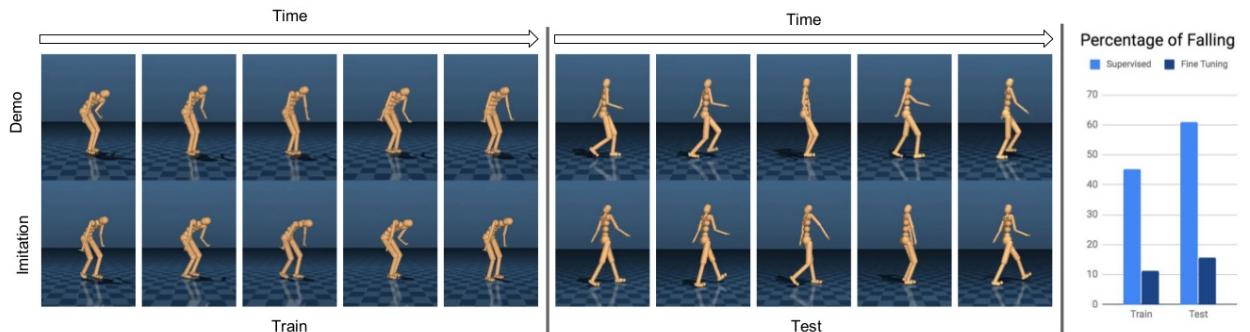


Figure 5: **Left:** examples of the demonstration trajectories in the CMU humanoid domain. The top row shows demonstrations from both the training and test set. The bottom row shows the corresponding imitation. **Right:** Percentage of falling down before the end of the episode with and without fine tuning.

- The evaluation is based on the diversity of policies, not game score
- GAIL policies are more robust than those of VAE policies

4. Conclusion

- Combine the strength of some generative models: VAE + GAIL

- Note that VAE is a method for supervised imitation (BC)