Human-in-the-Loop Generative Policy Learning from Demonstrations and Preferences

By Eiji Uchibe

The talk addressed a central challenge in RL: the difficulty of designing reward functions that reliably induce the desired behaviors. In many real robotic systems, the reward cannot be easily encoded, making conventional RL unreliable or impractical.

To address this challenge, the speaker examined the integration of learning from demonstrations and learning from preferences within a unified generative framework. Imitation learning is powerful when demonstrations are close to optimal, but it suffers from sensitivity to coverage gaps and noise. Preference-based learning, commonly used in RL from Human Feedback, is more robust to suboptimal data and provides sparse yet highly informative signals, but requires a method to generate meaningful trajectory pairs for comparison.

The proposed pipeline begins by collecting a set of expert demonstrations, forming the initial dataset. The first stage applies generative imitation learning, where two discriminators, functionally representing the policy, reward, and value networks, are trained to distinguish expert data from policy-generated samples. Through adversarial training, the reward and value functions emerge from the discriminators, and the policy is iteratively improved based on these learned structures.

Next, the learned stochastic policy generates paired trajectories, which serve as candidates for preference labeling. A prior preference is computed directly from the imitation-learned reward, but this may disagree with actual human judgment. Human evaluators then provide binary feedback indicating which trajectory is preferable. When inconsistencies arise between prior and human preferences, the reward model is updated using cross-entropy loss, ensuring alignment with human evaluation. Some human-endorsed trajectories may also be added back into the expert dataset, keeping the learning loop grounded in real behavior.

A notable architectural component is the environment discriminator, responsible for distinguishing whether a state–action–next-state tuple originates from the real system or from the learned dynamics model. This encourages the generative components to remain faithful to real-world transitions—an important aspect given that the method targets physical robotic platforms rather than simulation environments.

The speaker concluded by emphasizing the full human-in-the-loop imitation learning scheme, where imitation learning produces expert-like policies, these policies generate candidate rollouts, humans provide preference-based corrections, and the reward and value functions are continuously refined. The approach was validated on real robot experiments without simulators, highlighting its practical relevance for complex manipulation tasks such as robotic cloth folding.