摘要

强化学习中，智能体通过与环境交互并利用所得经验提升策略，以完成决策任务。近年来，深度学习兴起，反向传播让强化学习策略网络中特征表示难的问题得到有效解决，深度强化学习因此被推广至棋类对弈、机器人控制、大语言模型价值对齐、游戏控制等高维复杂决策任务。然而，从头开始深度强化学习的探索成本极高、效率低而且需要手工设计复杂的奖励函数，若使用监督学习的方式冷启动则需要采集成本极高带动作标注的专家数据。为了缓解这一难题，基于视频预训练学习的策略学习范式开始被提出。互联网积累的海量不含显式动作标签的纯视觉视频蕴含丰富的视觉表征与隐含行为先验，是策略学习极具潜力的预训练资源。

本文提出一种统一的基于视频预训练的策略学习范式：上游通过视频预训练离线学习动态表征与行为先验，下游将先验知识迁移至强化学习过程提升策略。该范式具有两方面的统一性。其一是学习范式上的普适性，从上游预训练到下游策略迁移的范式可以统一适配在线和离线强化学习场景，仅在下游算法侧做相应替换。其二是任务形态上的通用性，视觉游戏控制、双手灵巧操作与室内视频导航三类差异显著的任务均可纳入相同的两阶段框架，仅需依据任务特性更换上游预训练模块与下游强化学习算法，整体结构保持一致。

在视觉游戏控制场景中，本文提出 STG 框架，对应在线强化学习场景。上游在隐空间完成视频对抗学习预训练，并附加时间距离回归辅助任务增强时序对齐能力。下游通过预训练的判别器输出构造内在奖励，驱动在线强化学习探索，全程不依赖任何外部环境奖励和动作标注的情况下在两个高维视觉任务评测基准上等上取得了较高的样本效率。

在双手灵巧操作场景中，本文提出 BiDexHD 框架，对应在线多智能体强化学习场景。上游通过预训练模型从人类双手操作视频中估计手物三维位姿轨迹，经动作重定向将人手动作映射至机器人灵巧手，自动构建去中心化部分可观测任务集。下游应用多智能体强化学习算法训练基于状态的教师策略，设计通用的三阶段复合奖励，再蒸馏为点云输入的视觉学生策略，从而实现对未见任务与物体的零样本迁移。

在室内视频导航场景中，本文提出 NOLO 框架，对应离线强化学习设定。上游以预训练光流预测模型作为隐式逆动力学先验提取器，从相邻帧恢复伪动作并构建完整的包含观测和动作的轨迹，同时学习时间一致性偏好表征。下游设计以上下文视频为条件的双向循环 Transformer 策略网络，并结合离线强化学习方法进行学习。该框架使智能体仅凭一段未见环境的上下文视频即可完成目标导航，在两大导航仿真评测基准

与真实四足机器人平台上验证了跨布局、跨房间与跨仿真器的泛化能力。

综上，本文围绕从上游视频预训练到下游策略学习的核心思路，在视觉游戏控制、双手灵巧操作与室内视频导航三类任务上分别给出 STG、BiDexHD 与 NOLO 三个不同的实例化，并相应覆盖在线与离线三种强化学习设定，从任务形态与学习范式两个维度系统验证了所提统一范式的有效性。该范式有效缓解了专家数据匮乏、稀疏奖励与跨场景泛化不足等关键挑战，推动了视频驱动策略学习在具身智能实际场景中的落地应用，对专家数据获取困难、环境交互成本高昂或需快速泛化至新场景的决策任务具有重要的参考价值与应用前景。

关键词：机器学习, 强化学习, 视频预训练

Policy Learning from Video Pretraining

Zhou Bohan (Computer Application Technology)
Directed by: Prof. Lu Zongqing

ABSTRACT

In reinforcement learning, an agent improves its policy by interacting with the environment and leveraging the collected experience to accomplish decision-making tasks. In recent years, with the rise of deep learning, backpropagation has effectively resolved the longstanding difficulty of feature representation in reinforcement learning policy networks, thereby extending deep reinforcement learning to high-dimensional and complex decision-making tasks such as board games, robot control, alignment of large language models, and game playing. Nevertheless, training deep reinforcement learning from scratch suffers from prohibitive exploration costs, low efficiency, and reliance on hand-crafted reward functions, while cold-starting via supervised learning requires expert demonstrations with extremely expensive action annotations. To alleviate this dilemma, a policy learning paradigm based on video pretraining has begun to emerge. The vast amount of action-free visual videos on the Internet encode rich visual representations and implicit behavioral priors, constituting a highly promising pretraining resource for policy learning.

This thesis proposes a unified policy learning paradigm based on video pretraining. The upstream stage learns dynamical representations and behavioral priors offline through video pretraining, while the downstream stage transfers these priors to the reinforcement learning process to improve the policy. The paradigm exhibits unity in two aspects. The first is the universality at the level of the learning paradigm. The transfer from upstream pretraining to downstream reinforcement learning can be uniformly adapted to both online and offline scenarios with only corresponding replacements on the downstream side. The second is the generality at the task level. Three substantially different tasks, video games, bimanual dexterous manipulation, and indoor video navigation can all be incorporated into the same two-stage framework, with the overall structure preserved.

For video games, this thesis proposes the STG framework, which corresponds to the online reinforcement learning setting. The upstream conducts video adversarial pretraining in the latent space, augmented with a temporal distance regression auxiliary task to enhance temporal

alignment. The downstream constructs intrinsic rewards from the output of the pretrained discriminator to drive online exploration. Without relying on any external environmental rewards or action annotations, the framework achieves high sample efficiency on two high-dimensional visual task benchmarks.

For bimanual dexterous manipulation, this thesis proposes the BiDexHD framework, which corresponds to the online multi-agent reinforcement learning setting. The upstream employs pretrained models to estimate three dimensional hand–object pose trajectories from human bimanual manipulation videos, and through action retargeting maps human hand motions to robotic dexterous hands, thereby automatically constructing a decentralized partially observable task set. The downstream applies a multi-agent reinforcement learning algorithm to train state-based teacher policies, designs a generic staged compound reward composed of approaching, lifting, and target bonus, and further distills the teachers into vision-based student policies that take point clouds as input, achieving zero-shot transfer to unseen tasks and objects.

For indoor video navigation, this thesis proposes the NOLO framework, which corresponds to the offline reinforcement learning setting. The upstream adopts a pretrained optical flow prediction model as an implicit inverse dynamics prior extractor, recovering pseudo discrete actions from adjacent frames to construct complete trajectories containing both observations and actions, while simultaneously learning temporal representations with consistent ordering. The downstream designs a policy network of bidirectional recurrent Transformer conditioned on a context video, and performs conservative offline optimization via offline reinforcement learning methods. The framework enables the agent to accomplish video navigation in a previously unseen environment given only a single context video, and its cross-layout, cross-room, and cross-simulator generalization capability is verified on two navigation simulation benchmarks as well as a physical quadruped robot platform.

In summary, centered on the core principle of transitioning from upstream video pretraining to downstream policy learning, this thesis develops three distinct implementations, STG, BiDexHD, and NOLO, targeting visual games, bimanual dexterous manipulation, and indoor video navigation respectively. By correspondingly encompassing both online and offline reinforcement learning paradigms, this work systematically validates the efficacy of the proposed unified framework across both task formulation and learning paradigm dimensions. The paradigm effectively alleviates key challenges such as the scarcity of expert data, sparse rewards, and insufficient cross-domain generalization, and advances the practical application of

video-driven policy learning in real world scenarios of embodied intelligence, offering significant reference value and application prospects for decision-making tasks with difficult expert data acquisition, prohibitive environmental interaction costs, or the need for rapid generalization to novel scenarios.

KEY WORDS: Machine Learning, Reinforcement Learning, Video Pretraining