Внаслідок тренування можна виокремити 4 кластери стратегій (рис.1)які згруповані за очікуваною винагородою та агента та їхнім рівнем кооперації. видно близько чотирьох кластерів стратегій. На рис.2 виведені результати різниці між очікуваним виграшом стратегії та результатами навченого агента.

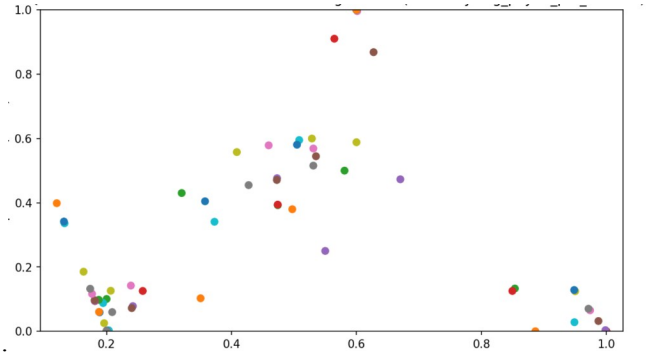

Рис.1Розподіл стратегій у просторі їхньої здатності до кооперації з агентом та зваженого результату агента.

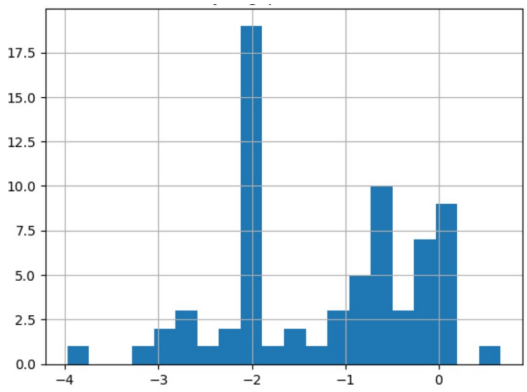

Рис.2 Різниця середнього очікуваного виграшу стратегій відносно навченого агента.

Стисле подання стану дозволяє розрізняти типи опонентів і адаптивно реагувати на їхні патерни поведніки і ефективно відтворювати їх під час фази тестування. Основними обмеженнями цього дослідження є детермінованість популяції, відсутність довготривалої пам’яті та нестаціонарність динаміки навчання. У подальших дослідженнях буде проведено порівняння з рекурентними та трансформер-моделями, введено стохастичних опоненти[4] та популяційний тренінг. У підсумку, можна стверджувати що DQN із компактним поданням показує контекстно-залежну поведінку та виявляє кластерну структуру опонентів, формуючи основу для подальших досліджень кооперації.

ДЖЕРЕЛА

1. Axelrod, Robert, and William D. Hamilton. "The evolution of cooperation." *science* 211.4489 (1981): 1390-1396.
2. Mnih, Volodymyr, et al. "Human-level control through deep reinforcement learning." *nature* 518.7540 (2015): 529-533.
3. Roderick, Melrose, James MacGlashan, and Stefanie Tellex. "Implementing the deep q-network." *arXiv preprint arXiv:1711.07478* (2017).
4. Bertrand, Quentin, et al. "Q-learners Can Provably Collude in the Iterated Prisoner's Dilemma." *arXiv preprint arXiv:2312.08484* (2023).

ЕФЕКТИВНЕ НАВЧАННЯ СТРАТЕГІЙ КЕРУВАННЯ ДЛЯ РОБОТИЗОВАНИХ МАНІПУЛЯЦІЙ ЗА ДОПОМОГОЮ ДИСТИЛЯЦІЇ ЗНАНЬ/EFFICIENT POLICY LEARNING VIA KNOWLEDGE DISTILLATION FOR ROBOTIC MANIPULATION

Севергін О.В., Кузьменко Д.О., Швай Н.О. / Severhin O., Kuzmenko D., Shvai N.
Національний університет "Києво-Могилянська Академія" / National University of Kyiv-Mohyla Academy
04655, Київ, вул. Григорія Сковороди, 2, факультет інформатики, кафедра математики
E-mail: oleksandr.severhin@ukma.edu.ua, Kuzmenko@ukma.edu.ua, n.shvay@ukma.edu.ua

The work focuses on the computational intractability of large-scale Reinforcement Learning (RL) models for robotic manipulation. While world-like models like TD-MPC2 demonstrate high performance in various manipulative tasks, their immense parameter count (e.g., 317M) hinders training and deployment on resource-constrained hardware. This research investigates Knowledge Distillation (KD) with a loss function specifically described in [1] and [2] as a primary method for model compression. This involves training a lightweight "student" model to mimic the behavior of a large, pre-trained "teacher" model. Unlike in supervised learning, distilling knowledge in RL is uniquely complex; the objective is to transfer a dynamic, reward-driven policy, not a simple input-output function.

Previous work in this area, specifically the TD-MPC-Opt study [3], established the feasibility of this approach using a static distillation coefficient. While successful in transferring capabilities to a 1M parameter student, a static coefficient presents a conceptual trade-off. A high, constant coefficient may overly bias the student towards imitation, stifling its own exploration and learning from direct environmental rewards. Conversely, a low coefficient may fail to provide sufficient guidance. This work extends the original study by hypothesizing that a dynamic distillation coefficient, which adapts over the training process, can lead to faster convergence and a more robust final policy.

Our methodology is centered on a composite loss function that governs the student model's training. This loss function has two primary components. The first is the student's original learning loss, which includes standard RL objectives such as reward loss, value loss, and consistency loss (a key feature of TD-MPC ensuring stable predictions). This component allows the student model to learn directly from its own environmental interactions. The second component is distillation loss. For this, we specifically target the teacher's learned value function. We chose this target because the value function encapsulates the teacher's long-term, high-level understanding of task success, providing a much denser and more stable training signal than distilling a volatile, high-variance policy directly. The loss is calculated as the mean squared error between the reward predictions of the teacher and the student for identical states and actions.

The entire process is balanced by the distillation coefficient, a critical hyperparameter that determines the weighting between self-learning and teacher imitation at any given point in training. The higher the coefficient the bigger effect teacher imitation has on the student model.

The central investigation of this paper is the behavior of this distillation coefficient. We moved beyond the static baseline to experiment with several adaptive schedules around 50,000 training steps. These included: Linear Decay (a high initial coefficient gradually decreasing to the baseline of 0.4, prioritizing guidance early); Cosine Decay (a smoother, non-linear version of decay); Increase (testing the inverse hypothesis of starting with self-learning and adding guidance later); and a Four Phase stepped schedule to test for sensitivity to abrupt changes.

We conducted our experiments on the MT30 dataset, distilling knowledge into a 1M parameter student model. The results demonstrate that dynamic coefficient strategies significantly outperform the static baseline.

Analysis of Results: The dynamic schedules showed a clear performance hierarchy. Both Linear Decay and Cosine Decay achieved the highest normalized score (8.64), surpassing the static Constant (0.4) coefficient's score of 8.39. This finding strongly supports our hypothesis: "front-loading" the teacher's guidance is the most effective strategy. At the same time same dynamic distillation coefficient showed worse results when they were starting with 1 and ending with 0 (lower that the baseline of 0.4) that means that not only "front-loading", but also a quite a strong "support" closer to end of training and KD is important.

The success of the decay schedules suggests that the student model benefits most from heavy imitation during the initial, random phases of its policy. As the student's own policy improves, the relative

importance of this guidance is reduced, allowing the student to fine-tune its behavior based on direct environmental feedback.

Conversely, strategies that delayed or provided erratic guidance underperformed. The "Increase" schedule, which delayed guidance, yielded a poor score of 8.09. The "Four Phase" schedule (8.32) and another "Decrease" variant (8.07) also failed to match the baseline. The failure of the "Increase" schedule is particularly informative, as it implies that withholding guidance initially is detrimental. The student fails to form a competent base policy on its own and cannot effectively leverage the teacher's complex knowledge when it is finally introduced. The results show that not all dynamic strategies are beneficial, but a clear pattern emerges: the most effective methods apply a high coefficient at the beginning of training and gradually decrease it.

Subsequent work will proceed in two main directions. First, we will validate these findings on more complex and diverse domains, specifically the ManiSkill 3 dataset. This benchmark includes tasks with complex object interactions and deformable objects, which will rigorously test the generalization of our dynamic distillation method in both single-task and multi-task regimes. We aim to verify if these decay schedules remain optimal or if greater task complexity necessitates a different approach.

Second, we will investigate a highly promising architectural direction: distilling knowledge into Mixture of Experts (MoE) models using a routing architecture [4]. Rather than compressing knowledge into a single dense student, we can distill it into a set of smaller, specialized "experts." This architecture is a natural fit for multi-task robotics, where one expert might learn "picking" motions and another "placing." We will explore using the teacher's knowledge to train not only the experts but also the routing mechanism that selects the appropriate expert for a given task. This could achieve higher performance and efficiency than a monolithic student.

Список джерел

1. Sheng R., Tao H., Wuyue Y. Tailored knowledge distillation with automated loss function learning. PLOS One. 2025. URL: https://doi.org/10.1371/journal.pone.0325599.
2. Romero A., Ballas N., Kahou S. E., Chassang A., Gatta C., Bengio Y. FitNets: Hints for Thin Deep Nets. arXiv preprint arXiv:1412.6550. 2014. URL: https://arxiv.org/abs/1412.6550.
3. Kuzmenko D., Shvai N. TD-MPC-Opt: Distilling Model-Based Multi-Task Reinforcement Learning Agents. arXiv preprint arXiv:2507.01823. 2025. URL: https://arxiv.org/abs/2507.01823.
4. Kuzmenko D., Shvai N. MoIRA: Modular Instruction Routing Architecture for Multi-Task Robotics. arXiv preprint arXiv:2507.01843. 2025. URL: https://arxiv.org/abs/2507.01843.

КООРДИНАЦІЯ МУЛЬТИАГЕНТНИХ LLM-СИСТЕМ / COORDINATION OF MULTI-AGENT LLM-BASED SYSTEMS

Діхтяр І.Ю., Нагірна А.М. / Dikhtiar I.Yu., Nahirna A.M.
Національний університет “Києво-Могилянська Академія” / National University of Kyiv-Mohyla Academy
04655, Київ, вул. Григорія Сковороди, 2, факультет інформатики, кафедра інформатики
E-mail: i.dikhtiar@ukma.edu.ua, a.nahirna@ukma.edu.ua

Abstract. This research focuses on analyzing the current state and challenges in the coordination of multi-agent systems based on Large Language Models (LLMs). We review the evolution from single autonomous agents (like ReAct) to complex multi-agent architectures. The analysis of current approaches reveals a key limitation: a reliance on deterministic, predefined coordination rules. This paper identifies the lack of adaptive, learned coordination policies as a research gap. We conclude by highlighting Reinforcement Learning (RL) as a promising, though challenging, direction for developing dynamic meta-agents capable of overcoming these limitations.