Knowledge Transfer in Model-Based Reinforcement Learning Agents for Efficient Multi-Task Learning

Extended Abstract

Dmytro Kuzmenko
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
kuzmenko@ukma.edu.ua

Nadiya Shvai
National University of Kyiv-Mohyla Academy
Kyiv, Ukraine
n.shvay@ukma.edu.ua
Cyclope.ai
Paris, France
nadiya.shvai@cyclope.ai

ABSTRACT

We propose an efficient knowledge transfer approach for model-based reinforcement learning, addressing the challenge of deploying large world models in resource-constrained environments. Our method distills a high-capacity multi-task agent (317M parameters) into a compact 1M parameter model, achieving state-of-the-art performance on the MT30 benchmark with a normalized score of 28.45, a substantial improvement over the original 1M parameter model's score of 18.93. This demonstrates the ability of our distillation technique to consolidate complex multi-task knowledge effectively. Additionally, we apply FP16 post-training quantization, reducing the model size by 50% while maintaining performance. Our work bridges the gap between the power of large models and practical deployment constraints, offering a scalable solution for efficient and accessible multi-task reinforcement learning in robotics and other resource-limited domains.

KEYWORDS

Model-Based Reinforcement Learning; Multi-Task Learning; Knowledge Distillation; Model Compression; Efficient RL Agents

ACM Reference Format:
Dmytro Kuzmenko and Nadiya Shvai. 2025. Knowledge Transfer in Model-Based Reinforcement Learning Agents for Efficient Multi-Task Learning: Extended Abstract. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 3 pages.

1 INTRODUCTION

Reinforcement learning (RL) has achieved significant progress in diverse domains, yet it remains challenging to efficiently train agents for multiple tasks in resource-constrained environments. Model-based RL approaches, such as TD-MPC2 [2], leverage large world models for superior performance and generalization but require substantial computational resources, limiting real-world applicability. Our work addresses this by focusing on optimizing large model-based RL agents for efficient multi-task learning through knowledge distillation and model compression. Specifically, we transfer knowledge from high-capacity TD-MPC2 models to compact 1M parameter backbones suitable for deployment in resource-limited scenarios. Our method builds on traditional teacher-student distillation [1, 3, 5, 6, 8] and incorporates quantization techniques [4], creating a training recipe for lightweight RL models. Using the MT30 benchmark [2, 7], our FP16-quantized model achieves state-of-the-art performance (28.45 normalized score), surpassing the original model (+48.5%) and outperforming models trained from scratch (+2.77%).

This work is licensed under a Creative Commons Attribution International 4.0 License.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

2 METHODS

Our approach employs a teacher-student distillation framework, adapting it to the specific challenges of model-based RL (Figure 1).

The 317M-parameter TD-MPC2 model servers as the teacher, representing the largest available checkpoint from the TD-MPC2 [2], and the respective 1M-parameter checkpoint serves as the student.

We build upon the original TD-MPC2 loss functions (consistency, reward, and value losses) by introducing an additional reward distillation loss. This new loss is computed as the mean squared error (MSE) between the rewards predicted by the teacher and student models:

$$L_{\text{distill}} = \text{MSE}(R_{\text{teacher}}(s,a), R_{\text{student}}(s,a))$$

where R_{teacher} and R_{student} are the reward predictions of the teacher and student models, respectively.

An additional reward distillation coefficient (*d_coef*) is introduced to balance the original TD-MPC2 loss with our new distillation loss:

$$total\_loss = original\_loss + d_{\text{coef}} * distillation\_loss$$

The *d_coef* acts as a hyperparameter controlling the influence of the teacher model's knowledge on the student model's learning process. We empirically find that values close to 0.5 yield the best results, with 0.4 being optimal for most training setups (Table 1).

The normalized score is used as our primary evaluation metric, consistent with [2]. Each task is scored on a scale of 1 to 1000, with the average sum divided by the number of tasks, resulting in a final range of 1-100.

We begin with a 317M parameter TD-MPC2 model checkpoint pretrained on the MT30 dataset and distill it into a 1M parameter

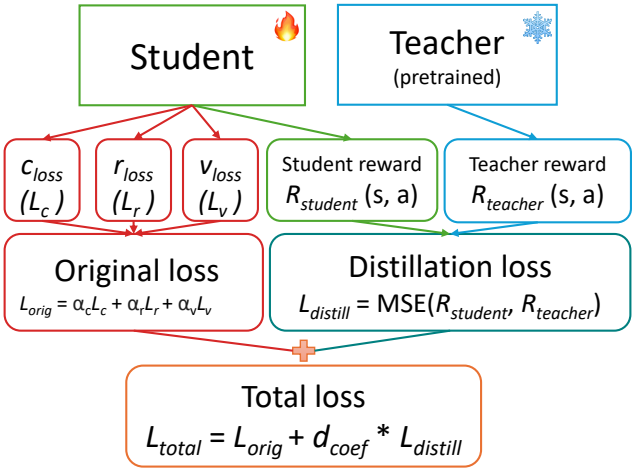

Figure 1: Our distillation approach consists of two main loss function components: the original loss from TD-MPC2 (in red), calculated as a linear combination of consistency, reward, and value losses (α denotes a scaling coefficient for each respective loss); and the distillation loss (in teal) that calculates MSE between student's (green) and teacher's (blue) rewards produced from inferring the same sample's state-action pair. The total loss is a linear combination of both losses, with the distillation loss scaled by the *d_coef*. In our scenario, the student is trainable while the teacher's weights are frozen.

student model using our proposed distillation approach, incorporating a reward distillation loss. To evaluate the effects of extended distillation, we experiment with periods ranging from 200,000 to 1,000,000 steps and varying batch sizes.

To enhance scalability for resource-constrained deployments, we apply FP16 post-training quantization.

For the MT30 benchmark, we utilize the full available dataset of 690,000 episodes (345,690,000 transitions). We conduct our experiments on a desktop PC with a single RTX 3060 GPU (12GB VRAM).

3 RESULTS

Short-term distillation results showed a significant improvement over the baseline. The distilled model with *d_coef* = 0.4 achieved a normalized score of 17.85, compared to the baseline score of 14.04, representing a 27.1% improvement. Extended distillation yielded similar results. The distilled model (*d_coef* = 0.45) achieved a normalized score of 28.12, compared to the model trained from scratch with a score of 27.36, representing a 2.77% improvement. Quantizing the model with FP16 further improves the performance to 28.45.

Distillation with a batch size of 256 offers an optimal balance of performance, efficiency, and hardware compatibility (Table 2).

Comparing distillation from different teacher model sizes revealed the value of using a larger, more knowledgeable teacher model. Distillation from the 48M parameter teacher resulted in a score of 13.61, while distillation from the 317M parameter teacher achieved 17.85, representing a 31.2% relative improvement.

Table 1: Impact of *d_coef* on student's performance: 200K distillation steps with 317M parameter teacher and batch size of 256.

Distillation coefficient	*Normalized score*
0.05	13.61
0.25	14.49
0.4	**17.85**
0.55	16.08
0.6	14.83
0.9	13.79

Table 2: Results of different setups of 1M-parameter model distillation vs training from scratch on MT30 benchmark.

Method	*Batch size*	*Training steps, #*	*Score*
distill	128	200K	17.37
from scratch	256	200K	14.04
distill	256	200K	**17.85**
from scratch	1024	200K	**18.7**
distill	1024	200K	18.11
from scratch	1024	337k	**26.94**
distill	1024	337K	25.44
from scratch	256	1M	27.36
distill	256	1M	**28.12**

Our experiments reveal that incorporating next-state latent distillation alongside reward distillation presents a significant challenge due to the dimensional mismatch between teacher (1376-dim) and student (128-dim) models. We attempted to apply linear projection and PCA as ways to bridge such mismatch. However, that introduced substantial information loss and led to poor generalization. Linear projection yielded a normalized score of 7.69, while PCA marginally improved to 8.78, but both underperformed compared to reward-only distillation (17.85).

In contrast, reward distillation directly aligns with task-specific performance, simplifying the learning process and enabling better outcomes, particularly in reward-oriented tasks like *pendulum-swingup* or *cup-catch*.

4 CONCLUSION

Our work demonstrates the potential of combining knowledge distillation and quantization to develop efficient, deployable multi-task RL agents, significantly reducing model size while maintaining performance. However, key limitations remain, including the need for real-world deployment on hardware platforms, the reliance on high-capacity teacher models, and the narrow focus on the MT30 benchmark. Despite these challenges, our approach lays a foundation for scaling RL models to resource-constrained environments, paving the way for more accessible and practical multi-task RL systems.

REFERENCES

[1] Wojciech M Czarnecki, Razvan Pascanu, Simon Osindero, Siddhant Jayakumar, Grzegorz Swirszcz, and Max Jaderberg. 2019. Distilling policy distillation. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1331–1340.

[2] Nicklas Hansen, Hao Su, and Xiaolong Wang. 2024. TD-MPC2: Scalable, Robust World Models for Continuous Control. *arXiv preprint arXiv:2310.16828* (2024).

[3] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

[4] Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. 2018. Mixed precision training. *arXiv preprint arXiv:1710.03740* (2018).

[5] Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. 2015. Actor-mimic: Deep multitask and transfer reinforcement learning. *arXiv preprint arXiv:1511.06342* (2015).

[6] Andrei A Rusu, Sergio Gomez Colmenarejo, Caglar Gulcehre, Guillaume Desjardins, James Kirkpatrick, Razvan Pascanu, Volodymyr Mnih, Koray Kavukcuoglu, and Raia Hadsell. 2015. Policy distillation. *arXiv preprint arXiv:1511.06295* (2015).

[7] Yuval Tassa, Yotam Doron, Alistair Muldal, Tom Erez, Yazhe Li, Diego de Las Casas, David Budden, Abbas Abdolmaleki, Josh Merel, Andrew Lefrancq, et al. 2020. Dm_control: Software and tasks for continuous control. *arXiv preprint arXiv:2006.12983* (2020).

[8] Yee Whye Teh, Victor Bapst, Wojciech M Czarnecki, John Quan, James Kirkpatrick, Raia Hadsell, Nicolas Heess, and Razvan Pascanu. 2017. Distral: Robust multitask reinforcement learning. In *Advances in Neural Information Processing Systems*. 4496–4506.