ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Introducing Intrinsic Motivation in Elastic Decision Transformers

Leonardo Guiducci[1], Giovanna Maria Dimitri[2], Giulia Palma[1], Antonio Rizzo[1]

1- DISPOC, Universitá di Siena, Siena, Italy
Via Roma 56, 53100, Siena, Italy

2- DIISM, Universitá di Siena, Siena, Italy
Via Roma 56, 53100, Siena, Italy

Abstract. Effective decision-making is a key challenge in artificial intelligence, with Reinforcement Learning (RL) emerging as one of the main approaches. However, RL often depends on complex reward functions, which are difficult to design. Intrinsic motivation, inspired by psychological concepts like curiosity, offers an alternative by generating agent-driven rewards to foster exploration. This paper introduces intrinsic motivation into the Elastic Decision Transformer (EDT) framework for Offline RL. By using an auxiliary intrinsic loss, we enhance representation learning without altering fixed reward signals. Experiments in locomotion tasks demonstrate improved performance, underscoring the potential of intrinsic motivation to advance RL in offline settings.

1 Introduction

Decision-making in dynamic and uncertain environments is a key challenge in artificial intelligence, with applications spanning robotics, healthcare, autonomous systems, and beyond. Reinforcement Learning (RL) has emerged as a powerful approach to train agents to interact with environments and learn through experience [1]. However, much of RL's success relies on carefully designed dense reward functions, whose creation poses significant engineering challenges. An alternative approach augments sparse extrinsic rewards with dense intrinsic rewards generated by the agent itself. Intrinsic rewards, inspired by concepts such as *intrinsic motivation* [2] and *curiosity* [3], encourage exploration and learning for their own sake. This paradigm has enabled RL agents to discover novel states, acquire diverse skills, and adapt to complex environments lacking sufficient external feedback. By incorporating intrinsic rewards through modules like Intrinsic Curiosity Module (ICM) [4] and Random Network Distillation (RND) [5], RL has achieved human-level performance in exploration-intensive tasks such as complex games and robotic manipulation.

This paper investigates the integration of intrinsic motivation into Elastic Decision Transformers (EDTs) [6], a recent RL innovation that leverages Transformer [7] architectures to enhance decision-making. EDTs excel in Offline RL [8], which uses pre-collected data for training, proving effective in scenarios where real-time interactions are costly or limited. Their transformer-based design captures long-range dependencies in sequential data, achieving notable results in benchmarks like D4RL [9] and Atari Games [10]. While intrinsic motivation is widely applied in Online RL to drive exploration by modifying the reward signal, Offline RL poses unique challenges. With learning confined to a fixed dataset collected by a behaviour policy, directly adding intrinsic rewards may disrupt the training process. EDTs, which model policies based on observed trajectories, inherently avoid altering fixed reward signals. To address this, we propose an auxiliary intrinsic loss function integrated within the EDT framework. This loss function, independent of the fixed reward signals, enhances representation learning by promoting the capture of intrinsic properties such as state novelty or predictive discrepancies. Optimizing this auxiliary loss alongside the primary task enables more robust and generalized policies, improving performance without additional exploration data. Experiments on locomotion tasks demonstrate that this approach enhances performance compared to the standard EDT architecture, underscoring the potential of intrinsic motivation in transformer architecture and offline RL.

1.1 Biological Plausibility and Motivation

Elastic Decision Transformers (EDTs) excel at trajectory stitching, selecting promising segments from past experiences and dynamically adjusting the history length considered for decision-making. This allows EDTs to learn effectively from both positive and negative experiences, treating errors as valuable signals alongside rewards. Building on this principle, our approach aims to enhance biological plausibility by redefining errors (*i.e. loss*) as intrinsic rewards, not merely as state feedback but as experiences shaped by the agent's learning process. By integrating this heuristic perspective into the EDT framework, we align the learning process more closely with natural strategies, emphasizing the informative role of failures in developing robust policies. In this paper, we introduce an auxiliary intrinsic loss mechanism into the elastic decision transformer framework to enhance policy learning in offline RL. We propose two EDT variants, which integrate intrinsic motivation by leveraging RND to improve representation learning. Our approach demonstrates how intrinsic motivation can be effectively incorporated into transformer-based architectures, paving the way for better generalization and biologically plausible RL models.

2 Background

In this section, we outline the theoretical background that underlies our work and provides an essential context for the study.

2.1 Elastic Decision Transformer

This study examines a decision-making agent within the framework of Markov Decision Processes (MDPs). The agent interacts with the environment in discrete steps: it observes the state o_t, selects an action a_t, and receives a corresponding reward r_t. The goal is to learn an optimal policy distribution $P^*_\theta(a_t \mid o_{\leq t}, a_{<t}, r_{<t})$ that maximizes the cumulative reward $R_t = \sum_{k>t} r_k$ over time, based on the agent's interactions with the environment. Elastic decision transformers [6] are designed to address the challenges of offline reinforcement learning [8], where agents are trained, without real-time interaction with the environment, on fixed datasets $\mathcal{D} = \{(o_t, a_t, r_t, o_{t+1})\}$ collected by a behaviour policy $\pi_b(a_t \mid o_t)$. In offline RL, the static nature of datasets introduces issues such as distributional shift, requiring robust methods to generalize effectively from limited data. EDTs overcome the limitation seen in Decision Transformers (DT) [11] by leveraging the sequential modelling capabilities of transformer architectures to process trajectories as sequences $\tau = \{(o_t, a_t, r_t)\}_{t=1}^{T}$, capturing long-range dependencies between states, actions, and rewards. This enables EDTs to perform trajectory stitching, combining suboptimal trajectory segments into coherent policies, making them particularly effective for offline RL tasks. Their flexibility and performance in domains like D4RL benchmarks highlight their utility in reinforcement learning applications.

2.2 Curiosity-Driven Learning

Curiosity-driven learning introduces intrinsic motivation as a mechanism for driving agent behavior independently of externally defined rewards. In this paradigm, the agent generates intrinsic rewards r_t^{int} based on measures such as state novelty, prediction error, or uncertainty, which are formalized as $r_t^{\text{int}} = f(o_t, a_t)$, where f is an intrinsic reward function [4]. This encourages exploration and the discovery of diverse skills, even in environments with sparse or no external rewards. It draws inspiration from developmental psychology, where curiosity fosters learning and adaptation in humans. In reinforcement learning, curiosity has been applied successfully to tackle sparse-reward environments and improve policy robustness in complex tasks [12]. In this work, we adapt these principles to Offline RL by introducing an auxiliary intrinsic loss $\mathcal{L}_{\text{int}}$, which enhances the agent's representation learning. Unlike in online RL, where intrinsic rewards directly influence agent actions, our approach incorporates $\mathcal{L}_{\text{int}}$ to guide the learning process without modifying the fixed reward signals in the dataset.

3 Methods

In this section, we describe the methodological framework developed to integrate intrinsic motivation into the EDT for offline reinforcement learning. Our approach is motivated by the need to enhance representation learning and policy robustness without altering the fixed reward signals inherent to offline datasets. We achieve this by introducing an auxiliary intrinsic loss $\mathcal{L}_{\text{int}}$ that complements the primary learning objective of EDTs, encouraging the agent to learn richer representations by leveraging intrinsic motivation signals.

3.1 Intrinsic Auxiliary Loss

We adapt the principles of curiosity-driven learning to the offline RL paradigm by introducing an intrinsic loss $\mathcal{L}_{\text{int}}$ to complement the EDT standard loss. Inspired

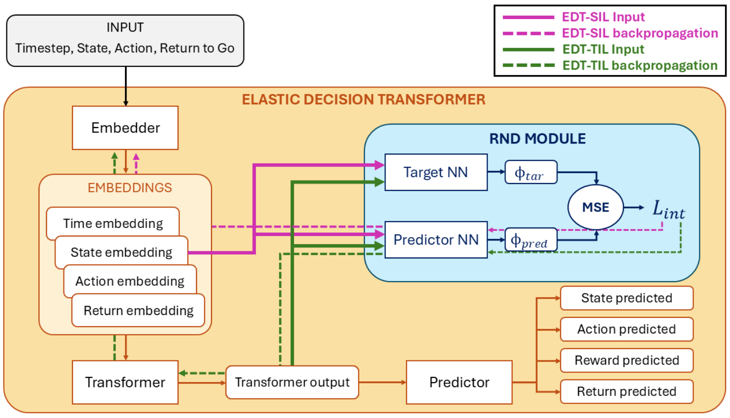

Fig. 1: Figure representing the proposed architecture with the two EDT variants. We highlight the blocks subject to backpropagation in each of the two variants. The target network of the RND has its weights frozen and is never updated, as shown in [5].

by the random network distillation module proposed in [5], which computes novelty using the next state as input, we extend this concept to two EDT-based variants (see Fig. 1). In the first variant, *EDT-SIL* (State Intrinsic Loss), the RND module operates on the embedded state, encouraging the agent to explore under-represented areas of the state space. In the second variant, *EDT-TIL* (Transformer Intrinsic Loss), we modify the classic EDT architecture to use the transformer's output as input to the RND module, aligning novelty computation with the model internal representation. For both variants, we use the outputs of the target and predictor networks, internal to the RND module, to calculate $\mathcal{L}_{\text{int}}$ as the mean squared error between the features of the two networks. Then, without scaling factors, we sum the $\mathcal{L}_{\text{int}}$ to the standard EDT loss $\mathcal{L}_{\text{EDT}}$. We embed the RND module directly within the EDT architecture. By doing so, the intrinsic loss $\mathcal{L}_{\text{int}}$ not only updates the weights of the RND module but also those of the preceding components: the embedder in the EDT-SIL case, and the embedder and transformer in the EDT-TIL case. This mechanism allows us to seamlessly integrate the concept of intrinsic motivation into the offline RL paradigm, fostering richer representation learning and encouraging exploration-driven strategies, even in the absence of environmental interaction.

4 Experiments

The application of the two proposed variants of EDT to benchmark scenarios is presented in this section. We compare the performance of our variants with the performance of the standard EDT

4.1 Dataset and metrics

We evaluated our approach on locomotion tasks from MuJoCo environments (D4RL benchmark [9]), using medium-replay datasets for Hopper, Walker2d, HalfCheetah, and Ant. These datasets, collected during the training of a behavior policy transitioning from suboptimal to near-optimal performance, offer a broader and noisier data distribution, making learning more challenging than standard medium datasets. We conducted experiments using the architecture outlined by Wu et al. [6]. Performance was evaluated following the scoring methodology specified in the same work, utilizing human-normalized scores (HNS) [13]. Specifically, HNS is calculated as $HNS = \frac{score-score\_random}{score\_human-score\_random}$, ensuring a consistent scale across games.

4.2 Results

In our experiments, we trained both the basic EDT and our EDT variants using five seeds for each training environment. For each environment, we then tested all five trained models for three evaluation rounds of 100 episodes each so as to ensure robustness in the results. The results of our experimentation are shown in Table 1. The mean and standard deviation of the HNS scores (4.1) obtained during the evaluation phase are reported for each environment.

Model	**Hopper**	**Walker**	**Cheetah**	**Ant**
EDT	81.56±9.96	62.25±5.21	37.32±2.46	**85.51±5.06**
EDT-SIL	**84.67±4.80**	57.21±8.54	37.64±2.44	84.02±3.72
EDT-TIL	81.72±9.27	**65.06±3.81**	**38.60±1.28**	83.72±4.13

Table 1: Mean and standard deviation of human-normalized scores (HNS) for EDT and its variants (EDT-SIL, EDT-TIL) across four tasks. Best results per task are highlighted.

From the results we can appreciate that, in all environments except *Ant*, using the intrinsic loss mechanism leads to better results. In particular, using the EDT-SIL model results in a strong increase in performance, but only in the first test environment. On the other hand, with the EDT-TIL model we get a performance increase in all the first three environments. Note that the evaluation phase occurs in real time, using new interactions with the environment rather than relying on the fixed training dataset. This justifies the addition of intrinsic loss, as it encourages exploration-like behaviour, improving the model's capacity to generalize during evaluation. Inspired by biological learning, where curiosity adapts to new stimuli, our intrinsic loss enables dynamic refinement of representations, bridging the gap between offline training and real-time decision-making.

5 Conclusions

This study introduces intrinsic motivation into the Elastic Decision Transformer (EDT) framework, demonstrating its potential to enhance performance in offline reinforcement learning tasks. By integrating an auxiliary intrinsic loss function, we bridge the gap between agent-driven exploration strategies and the constraints of fixed reward signals in offline settings. Experimental results from locomotion tasks in benchmark datasets highlight the efficacy of this approach, showcasing improvements in policy learning. Although this is preliminary research, it lays a promising foundation for future in-depth investigations into the use of intrinsic motivation within transformer architectures and the offline RL paradigm. Future research should investigate the impact of the auxiliary intrinsic loss on model learning dynamics and its applicability to other domains. Expanding the benchmark in this way would enable a more comprehensive statistical analysis of the differences between the tested models. Exploring this direction further could unlock more robust and adaptive methods for generalizing in complex environments. Moreover, integrating biological plausibility into these models may provide insights into natural learning mechanisms, enabling the design of RL frameworks that align closely with cognitive principles. This work not only emphasizes the value of intrinsic motivation in offline settings but also offers a biologically inspired pathway to bridge the gap between static training datasets and dynamic real-world evaluation scenarios.

References

[1] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theories and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[2] Pierre-Yves Oudeyer and Frederic Kaplan. What is intrinsic motivation? a typology of computational approaches. *Frontiers in neurorobotics*, 1:6, 02 2007.

[3] Paul J. Silvia. 157 curiosity and motivation. In *The Oxford Handbook of Human Motivation*. Oxford University Press, 02 2012.

[4] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. *CoRR*, abs/1705.05363, 2017.

[5] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. Exploration by random network distillation. *CoRR*, abs/1810.12894, 2018.

[6] Yueh-Hua Wu, Xiaolong Wang, and Masashi Hamaya. Elastic decision transformer. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017.

[8] Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *CoRR*, abs/2005.01643, 2020.

[9] Justin Fu, Aviral Kumar, Ofir Nachum, George Tucker, and Sergey Levine. D4RL: datasets for deep data-driven reinforcement learning. *CoRR*, abs/2004.07219, 2020.

[10] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. The arcade learning environment: An evaluation platform for general agents. *CoRR*, abs/1207.4708, 2012.

[11] Lili Chen, Kevin Lu, Aravind Rajeswaran, and Kimin Lee et al. Decision transformer: Reinforcement learning via sequence modeling. *CoRR*, abs/2106.01345, 2021.

[12] Yuri Burda, Harri Edwards, and Deepak Pathak et al. Large-scale study of curiosity-driven learning. *CoRR*, abs/1808.04355, 2018.

[13] Volodymyr Mnih, Koray Kavukcuoglu, Silver, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.