2023 Master's Thesis

Pre-controller for Safe Reinforcement Learning using Transformer with State-Action-Reward Representations

A Thesis Submitted to the Department of Computer Science and Communications Engineering,

the Graduate School of Fundamental Science and Engineering of Waseda University

in Partial Fulfillment of the Requirements for the Degree of Master of Engineering

Submission Date: January, 22, 2024

Zhiwei Shen
(5122FG09-1)
Supervisor: Prof. Shinichi Honiden
Research guidance: Research on Autonomous Agent System
CSCE, Waseda University

Contents

Chapter 1

Introduction

Reinforcement Learning (RL) is a dynamic and influential field within artificial intelligence that focuses on how agents should take actions in an environment to maximize a cumulative reward [1]. Rooted in the fundamental principles of trial-and-error learning and decision making, RL has emerged as a key technique in modern AI applications, offering a framework for solving complex problems where explicit programming is impractical.

At the heart of RL lies the exploration of the underlying Markov decision process (MDPs) [2]. The agent makes decisions or takes actions, and the environment responds to these actions by presenting a new state and a reward. The reward, a critical component of RL, serves as feedback for the actions taken, guiding the agent to learn a policy that maximizes long-term rewards.

The evolution of RL has been marked by the transition from basic models to more sophisticated systems:

Initially, RL focused on simple problems with discrete state and action spaces, using algorithms like Q-learning [3] and SARSA [4]. These models laid the groundwork for understanding how agents could learn from direct interaction with an environment.

The advent of deep learning brought about a significant leap in RL. Deep Reinforcement Learning (Deep RL) combines neural networks with RL, enabling agents to handle high-dimensional, continuous state spaces [5]. This led to breakthroughs in various domains, such as gaming (e.g., AlphaGo), robotics, and autonomous vehicles.

As RL systems began to tackle more complex tasks, the need for scalable and efficient algorithms became apparent. This led to the development of algorithms that could handle large-scale problems, involving intricate state dynamics and decision-making under uncertainty.

The introduction of sequence modeling in RL marks a crucial innovation in the field [6]. Sequence modeling, borrowed from natural language processing, offers a new perspective for handling temporal dependencies in RL tasks. It addresses the challenge of learning from sequences of states and actions, capturing the essence of decision-making in RL.

In this evolving landscape, models like StARformer represent the next step in the RL paradigm [7]. StARformer integrates the principles of sequence modeling with RL, offering a novel way to interpret and process the sequential data inherent in RL environments. This approach allows for a more nuanced understanding of the temporal relationships within RL tasks, facilitating the development of more advanced and efficient learning algorithms.

1.1 Safety Challenges in Reinforcement Learning

Reinforcement Learning (RL) presents a unique set of challenges, primarily centered around decision-making under uncertainty and the delicate balance between exploration and safety. These challenges are amplified in applications involving real-world interactions where the stakes are higher and the consequences of decisions are more significant [8].

In the realm of Reinforcement Learning (RL), the trade-off between exploration and safety is a fundamental challenge. Agents learn by exploring their environment, yet such exploration can lead to unsafe states or actions, particularly in high-risk environments where mistakes may have severe consequences. Achieving a balance between the need for exploration to learn optimal policies and the necessity to avoid dangerous situations is crucial. Additionally, RL agents, typically trained in specific environments, face significant challenges in maintaining safety and effectiveness amidst dynamic and unpredictable real-world changes. This highlights the importance of robustness to environmental changes.

Managing uncertainty and risk in RL is also pivotal. Since RL involves dealing with uncertainties in environmental dynamics and action outcomes, developing methods to quantify, manage, and mitigate associated risks is vital for ensuring safety. In many real-world applications, the high-dimensional nature of state and action spaces adds complexity to safe exploration, necessitating research into algorithms that can navigate these spaces efficiently and safely.

1.2 Reinforcement Learning to Sequence Modeling

Reinforcement Learning (RL) is commonly modeled as a Markov Decision Process (MDP), from which single-step value-estimation methods like Q-learning [9] and Temporal Difference (TD) learning [10] have been developed. These methods and their various extensions have traditionally formed the backbone of RL approaches. However, more recent developments in the field have started to view RL differently. For instance, approaches like Decision transformer [11] treat RL as a sequence modeling task. In this framework, given a sequence of recent experiences, including state-action-reward triplets, a model is trained to predict the sequence of subsequent actions. This concept of sequence modeling in RL can also be interpreted as solving RL problems by learning trajectory representations. The combination of these techniques represents a significant advancement in the field of RL, offering new perspectives and methodologies for tackling RL challenges.

1.3 Objectives of Our Study:

Our study primarily aims to explore the integration of the State-Action-Reward Transformer (StARformer) model in enhancing safety and efficiency within Reinforcement Learning (RL) tasks. The objectives are multi-faceted: firstly, to train the StARformer model to effectively assess and predict safe actions by understanding the association between current states and non-successful terminal states. Secondly, to develop and evaluate a preliminary control mechanism, termed the 'Pre-controller', which utilizes the trained StARformer for initial decision-making in various RL scenarios. Finally, our study seeks to demonstrate the practical application of the Pre-controller in real-world RL environments, focusing on its ability to filter and select actions that align with predefined safety criteria. This approach aims to seamlessly integrate safety considerations into the RL process, ensuring safer and more reliable decision-making, and potentially setting new standards in the realm of Safe RL.

1.4 Structure of the Thesis

The rest of the paper is organized as follows: Chapter 2 provides the academic background of this research. Chapter 3 explains the motivation of our research. Chapter 4 describes our proposed improvements and methods in detail. Chapter 5 presents the evaluation methodology and experimental results. Chapter 6 discusses challenges and limitations of our research. Finally, Chapter 7 summarizes the conclusions of our research and outlines future directions.

Chapter 2

Literature Review and Background

Our method relies on Sequenced RL and Safe RL. We introduce essential concepts and knowledge in the following subsections.

2.1 Sequenced RL

2.1.1 Basics of Reinforcement Learning

Reinforcement Learning (RL) is a branch of machine learning where an agent learns to make decisions by interacting with an environment to achieve a goal [1]. The fundamental principle of RL is learning through trial and error, where the agent makes sequential decisions, receives feedback in the form of rewards or penalties, and adjusts its actions accordingly. Here's a breakdown of its fundamental concepts:

In RL, an environment is typically modeled as a Markov Decision Process (MDP), characterized by a set of states, actions, and rewards. The agent, at each time step, observes its current state, selects an action, and then transitions to a new state while receiving a reward. The goal is to learn a policy - a strategy that maps states to actions - to maximize cumulative rewards over time.

RL has several key components:

1. Agent: The learner or decision-maker. 2. Environment: What the agent interacts with. 3. State: The current situation or context of the agent. 4. Action: Choices the agent can make. 5. Reward: Feedback from the environment indicating the success of an action. 6. Policy: The agent's strategy for selecting actions. 7. Value Function: Measures the expected

long-term return of states or actions, guiding the policy. 8. Model: An optional component that predicts the next state and reward.

2.1.2 Sequenced RL

Reinforcement Learning (RL) is traditionally modeled as a Markov Decision Process (MDP). Recent advancements, however, approach RL as a sequence modeling task, where models predict future actions based on a sequence of past state-action-reward triplets. This approach aligns well with offline RL and imitation learning, focusing on learning trajectory representations through supervised learning methods.

Using Transformer to Solve RL Problem

Transformers, initially prominent in natural language processing [12] [13] [14]and computer vision [15], are increasingly being explored in Reinforcement Learning (RL). Recent studies have begun to unveil their potential in this field to model interactions between a sequence of word embeddings, or more generally, unit representations or tokens. Recently, Transformers have been adopted in vision tasks with the key idea of breaking down images/videos into tokens [16] [17] [18], often outperforming convolutional networks (CNNs) in practice. Inspired by designs from both Transformers and CNNs, combining the two [19]shows further improvements. GPT [14] can be applied to RL under the sequence modeling setting.

The core mechanism of transformers is the self-attention module, which models interactions between all pairs of input tokens to capture their relationships. Each input token is mapped into query, key, and value representations to compute self-attention, as detailed in [20]. This approach has been adapted for vision tasks by Vision Transformer (ViT) [21] [22], which processes images by dividing them into a sequence of non-overlapping local patches. These patches are then flattened and linearly mapped to token sequences for self-attention processing.

Transformer architectures, initially introduced for language processing tasks, have significantly impacted how we model interactions between sequences of word embedding or, more broadly, unit representations known as tokens. These architectures have recently made a notable transition into vision tasks. The core innovation here lies in treating images or videos as a series of tokens, a methodology that has often surpassed the performance of traditional convolutions networks (CNNs) in practical applications.

This success in vision tasks has prompted a fusion of ideas from Transformers and CNNs, leading to further enhancements in model performance.

Transformers have also demonstrated utility in processing sensory information and in one-shot imitation learning. In the realm of Reinforcement Learning (RL), Chen et al. explored the application of GPT under a sequence modeling framework. This approach in visual RL closely resembles learning from videos, where input data comprises sequences of observed images, or states.

However, applying Transformers to video data introduces challenges, particularly due to the extensive number of input tokens and the associated quadratic computational demands. Researchers have investigated several solutions to these issues, such as attention approximation techniques [23], implementing separable attention across different dimensions [24], reducing token numbers through local windows [25], generating a smaller amount of tokens adaptively, or utilizing a CNN-stem to produce a condensed set of high-level tokens [26].

StARformer

In Reinforcement Learning (RL), states, actions, and rewards across adjacent time steps often exhibit strong causal connections. Recent past states significantly influence the next action, while the immediate future state and reward are direct consequences of the current action. In Markov Decision Processes (MDP), these relationships are even more pronounced. However, a Transformer that attends to all tokens without discrimination might struggle with excess information, potentially obscuring crucial relational priors [6]. This issue becomes acute with large input sequences, both spatially and temporally, and in complex Transformer models with numerous layers [27]. Learning Markovian dependencies from scratch can be inefficient, leading to wasted computational resources.

To address these challenges, the State-Action-Reward Transformer (StARformer) for visual RL is introduced. This model explicitly captures single-step transitions, infusing a Markovian-like inductive bias and optimizing the capacity for long sequence modeling. The StARformer comprises two interleaved components: a Step Transformer and a Sequence Transformer. The Step Transformer focuses on local representations within a single timestep by self-attending to state-action-reward tokens, with image states encoded as ViT-like patches to preserve detailed spatial information. The Sequence Transformer then merges these StAR-representations with pure image state representations, derived from convolutional features, to predict actions over the entire sequence.

StARformer consists of two basic components: Step Transformer and Sequence Transformer, together with interleaving connections. Step Trans-

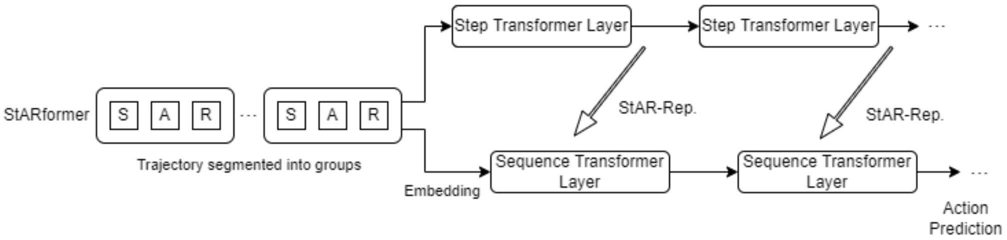


Figure 2.1: StARformer

former learns StAR-representations from strongly-connected local tokens explicitly, which are then fed into the Sequence Transformer along with pure state representations to model the whole input trajectory. At the output of the final Sequence Transformer layer is the action predictions. In the following subsections, we will introduce the two Transformer components, and their corresponding token embedding in detail.

In the StARformer framework, the input consists of a sequence of states, actions, and rewards from RL environments, akin to word encoding in NLP. Each element of this sequence is transformed into embeddings, forming the basis of our model's input representation.

Step Transformer

Grouping State-Action-Reward Segment a trajectory T into groups in order to capture strong local relationships. Each group includes a previous action a_{t-1}, reward r_{t-1}, and the current state s_t.

Patch-wise State Token Embeddings Following the Vision Transformer (ViT)approach, tokenize each state image into non-overlapping spatial patches Z s. This tokenization aims to produce detailed state embeddings, allowing the Step Transformer to understand how actions and rewards correlate with specific state regions.

Action and Reward Token Embeddings Using a linear layer to embed the action and reward tokens.

S-A-R Embeddings Combine state, action, and reward embeddings to form the input for the initial Step Transformer layer: $Z = \{z_{a_{t-1}}, z_{r_{t-1}}, z_{s_t}\}$. Across each trajectory, we have T groups of such token representations processed simultaneously by the Step Transformer with shared parameters.

Sequence Transformer

2.2 Safe RL

2.2.1 Basics of Safe Reinforcement Learning

Safe RL can be defined as the process of learning policies that maximize the expectation of the return in problems in which it is important to ensure reasonable system performance and/or respect safety constraints during the learning and/or deployment processes [28]. Reinforcement Learning (RL) presents a unique set of challenges, primarily centered around decision-making under uncertainty and the delicate balance between exploration and safety. These challenges are amplified in applications involving real-world interactions where the stakes are higher and the consequences of decisions are more significant [29] [30] [28].

In Reinforcement Learning (RL), the trade-off between exploration and safety is a fundamental challenge, as an agent's exploration for optimal policy learning can lead to unsafe states, particularly in high-stakes environments. Ensuring robustness to the unpredictable changes of real-world environments, where uncertainty in dynamics and outcomes is prevalent, is crucial for the safety and efficacy of RL agents. This necessity extends to managing risks in high-dimensional state and action spaces, making the design of safe exploration algorithms vital. Moreover, guaranteeing the safety of RL policies through rigorous verification and validation is an ongoing research focus.

The real-world deployment of RL also brings forth ethical and social considerations, especially in sensitive sectors like healthcare and autonomous driving, requiring responsible development and implementation. Additionally, making RL agent policies interpretable and explainable is key for user trust and effective system control. Lastly, ensuring the long-term safety of RL strategies, particularly in areas with lasting impacts like environmental management, poses significant challenges, underlining the importance of careful planning and strategy in RL applications.

2.2.2 Shielded learning

Shielded learning in RL introduces a safety mechanism, referred to as a 'shield,' to ensure actions comply with predefined safety specifications. In traditional RL, an agent selects actions at each timestep, receiving feedback from the environment in terms of state observations and rewards. The typical goal is to optimize accumulated rewards. The shield, derived from safety specifications and an abstraction of environmental dynamics. This shield differentiates between 'unsafe' actions, which could violate safety norms, and 'correct' ones, effectively preventing unsafe decisions by the agent.

Pre-shielding

Pre-shielding in Reinforcement Learning (RL) is a proactive strategy designed to enhance the safety and reliability of RL agents. It involves creating a safeguard mechanism that operates before the agent executes its decisions, hence the term 'pre-shielding'. This mechanism acts as a filter or a checkpoint, ensuring that the actions selected by the RL agent do not lead to undesirable or unsafe outcomes.

Pre-shielding is a safety mechanism in Reinforcement Learning (RL) that modifies the interaction loop between the learning agent and its environment to ensure safety.

In pre-shielding, the workflow at each time step t involves the following steps:

Action Set Computation by the Shield The shield takes the set of all possible actions that the agent can choose from and filters out any actions that could lead to safety violations. This process results in a set of safe actions, denoted as $\{a_1, a_2, ..., a_n\}$.

Agent's Choice of Action The agent, upon receiving this filtered list of actions, selects one of the correct actions. This selection, denoted as a_t, is guaranteed to be a safe choice because it comes from the pre-screened set provided by the shield.

Environment's Response Once the agent has selected an action, the environment executes this action, transitions to the next state s_{t+1}, and computes the associated reward r_{t+1}.

The primary task of the shield in pre-shielding is to continually modify the set of available actions for the agent at each time step. This ensures that the agent can only choose from actions that are deemed correct or safe, in accordance with a predefined safety specification. The shield acts as a proactive filter, preventing the agent from making choices that could lead to unsafe states or outcomes.

This approach is particularly beneficial in scenarios where safety is a critical concern and where wrong actions could lead to serious consequences. By integrating pre-shielding, the safety of the RL system is significantly enhanced, as it minimizes the risk of the agent engaging in harmful or undesirable behaviors. In every step, the shield just takes the provided action

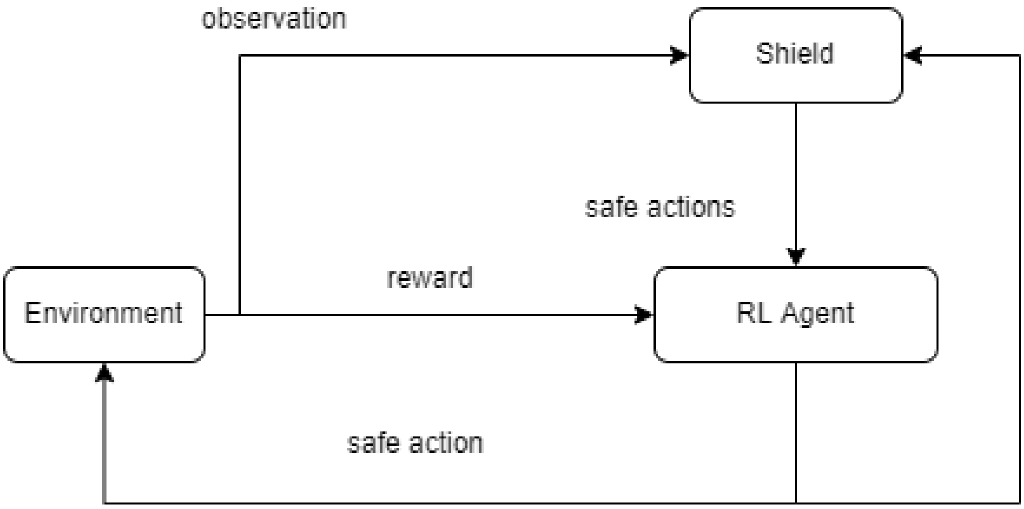


Figure 2.2: Pre-shielding

from the agent and corrects it if necessary to ensure safe operation of the system. The learning agent does not even need to know that it is shielded.

Chapter 3

Motivation

The motivation behind our research is rooted in the recognition of a critical gap in current Reinforcement Learning (RL) methodologies, particularly in the realms of safety and performance. This gap has significant implications, especially as RL systems are increasingly deployed in complex, real-world scenarios. Our study is driven by the conviction that the StARformer model has the potential to address these pivotal issues, thereby advancing the field of RL and contributing to the development of safer and more efficient AI systems.

3.1 Current Concerns

In the field of Reinforcement Learning (RL), practitioners frequently encounter two significant challenges with traditional methodologies: safety concerns and performance limitations, both of which profoundly impact the effectiveness and applicability of RL models in various scenarios. Traditional RL approaches, primarily focused on achieving optimal performance, often inadvertently compromise safety, particularly in unpredictable or complex environments. This issue becomes especially critical in scenarios where safety is paramount, such as autonomous driving, healthcare, and industrial automation. The root of the problem lies in the training and objective-setting of traditional RL models, which usually lack mechanisms to evaluate and mitigate potential risks or to foresee the long-term consequences of their actions.

Moreover, existing RL models often struggle to efficiently learn optimal policies in environments with high-dimensional state spaces or those requiring long-term strategic planning. In high-dimensional environments, these models may find it difficult to discern relevant features or patterns within the data, leading to slower learning processes or suboptimal policy formulation.

This limitation hinders the model's ability to make well-informed decisions, thus affecting its performance and applicability. Additionally, tasks requiring long-term planning pose unique challenges, as traditional RL models, especially those employing methods like Q-learning or policy gradients, tend to focus on immediate rewards and may lack the foresight needed for effective long-term strategy development.

In conclusion, while traditional RL methods can be effective in certain scenarios, they exhibit significant shortcomings in terms of safety and performance, especially in complex, high-dimensional, or environments demanding long-term planning. These limitations underscore the need for more advanced RL models like the StARformer, which integrates safety into the learning process and is adept at handling complex and long-term decision-making scenarios.

3.2 Potential of StARformer

The StARformer model marks a significant advancement in the field of Reinforcement Learning (RL) with its advanced sequence modeling capabilities, bringing about a transformative impact on decision-making, safety, and performance. At its core, the StARformer excels in interpreting complex sequences, providing a nuanced view of the environment and the decision-making process. This ability leads to more informed choices, enhancing the quality of decisions and allowing for a deeper understanding of intricate patterns and dependencies in data. This sophistication contrasts sharply with traditional models that often rely on more simplistic interpretations.

A key motivation behind the StARformer's development is the integration of safety as a fundamental component of the RL process, representing a paradigm shift from conventional models where safety might be secondary or an external addition. In the StARformer, safety is not just a consideration but a priority, with each decision evaluated for both its rewards and safety implications. This ensures that the pursuit of high rewards does not overshadow the importance of safe outcomes, aligning the model's objectives with the crucial real-world need for safety.

In addition to enhanced decision-making and a focus on safety, the StARformer excels in complex and dynamic environments. Unlike traditional RL models, which may struggle with large volumes of complex data, the StARformer's advanced architecture is adept at efficiently processing sequences of states and actions. This efficiency allows the model to develop well-informed, strategic policies that are safe and effective, even in challenging scenarios.

In essence, the StARformer stands out as a robust and versatile tool in RL. It not only offers improved decision-making abilities and prioritizes safety but also demonstrates a remarkable performance in complex settings. The design of the StARformer represents a holistic approach to RL, where safety and efficacy are seamlessly integrated into the decision-making fabric, making it a pivotal development in advancing RL technologies.

3.3 Advancing the Field of SafeRL

Our research in developing the StARformer model marks a significant advancement in the field of Safe Reinforcement Learning (SafeRL), particularly focusing on the simultaneous achievement of safety and performance in RL. This approach addresses the traditionally challenging dichotomy where optimal performance often came at the cost of safety, and vice versa. Our work with the StARformer has shown that it is indeed possible to attain high performance while strictly adhering to safety standards, a balance crucial for the effectiveness and security of RL systems, especially in scenarios where safety is non-negotiable.

A key innovation in our study is the adaptation of Transformer models, originally designed for natural language processing, for ensuring the safety of RL. These models' exceptional abilities in handling sequential data have been leveraged to enhance RL system performance and to integrate safety considerations more deeply into the learning process. This novel methodology could significantly influence future research and development in SafeRL, inspiring new solutions to the field's unique challenges.

Moreover, the potential of our research extends into real-world applications, particularly in domains where safety is paramount, such as autonomous vehicles, healthcare, and robotics. In these areas, the application of the StARformer model can greatly enhance the safety and reliability of RL systems, ensuring they make safer and more informed decisions. By contributing to the development of RL applications that are not only high-performing but also trustworthy and secure, our research has laid the groundwork for future innovations aimed at making RL systems more effective and safer in practical scenarios.

In summary, the StARformer model stands at the forefront of advancing SafeRL, bridging the gap between safety and performance with an innovative approach that holds substantial potential for a variety of real-world applications. This research not only extends the limits of what is achievable in SafeRL but also paves the way for future breakthroughs that could further enhance the safety and efficacy of RL systems in practical applications.

3.4 Rationale Behind Integration:

The integration of Transformer models into Reinforcement Learning (RL) is driven by their unique capabilities that promise to significantly enhance RL task performance, boost learning efficiency, and improve generalization. Transformers are particularly adept at parallel processing and retaining long sequences, features that are incredibly beneficial in complex RL environments where a comprehensive understanding of the sequence of states and actions is critical. This capability enables Transformers to provide a holistic view of scenarios, leading to more informed and strategic decision-making, especially in tasks requiring long-term planning.

Moreover, the efficiency of Transformers in processing large volumes of data simultaneously makes the RL learning process more efficient. Unlike traditional RL models that often process data sequentially, Transformers can handle complex decision-making processes and large-scale environments more effectively, accelerating the learning process. This is especially valuable in high-dimensional state spaces or scenarios requiring rapid adaptation to new information.

A notable strength of Transformers, proven in the field of Natural Language Processing (NLP), is their remarkable ability to generalize from training data. This characteristic is immensely beneficial in RL, where models often face the challenge of generalizing from limited experiences to new, unseen situations. The Transformer's inherent generalization ability enhances the RL model's capacity to adapt and perform effectively in these novel situations, minimizing the need for extensive retraining or specific programming for each new environment.

In essence, the integration of Transformer models into RL harnesses their potential to not only improve performance in complex tasks but also to increase the learning process's efficiency and enhance model generalization. These advantages directly address some of the fundamental challenges in RL, paving the way for the development of more advanced, efficient, and adaptable RL systems.

Chapter 4

Pre-controller for Safe Reinforcement Learning using Transformer with State-Action-Reward Representations

This chapter outlines the methodology and proposal for integrating the StARformer model into Safe RL. It details the experimental setup, the technical approach, and the proposed innovations to enhance both performance and safety in RL tasks.

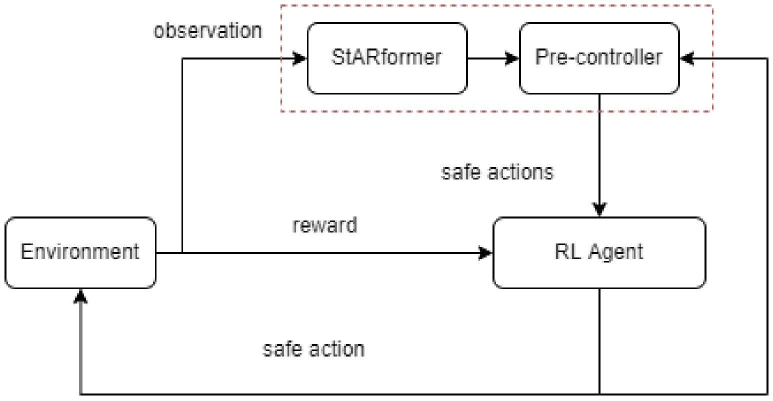


Figure 4.1: Overview

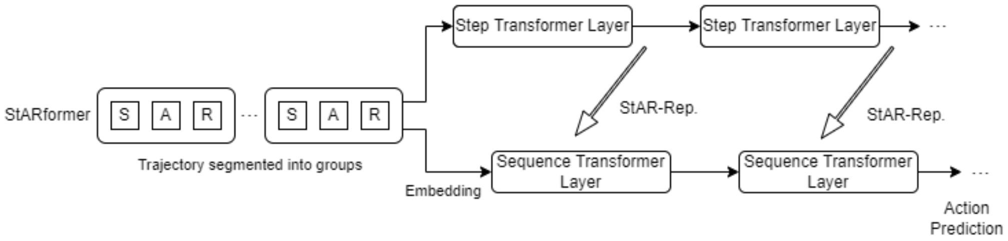


Figure 4.2: StARformer

4.1 StARformer

The StARformer represents a significant innovation in the application of Transformer models to Reinforcement Learning (RL), blending the robust capabilities of sequence modeling with the dynamic requirements of RL environments. Here's a detailed overview of the StARformer architecture:

The Transformer, originally designed for tasks in natural language processing, is renowned for its attention mechanism, which allows it to process sequences of data in parallel and capture long-range dependencies. The model's architecture eschews recurrent layers for stacked self-attention and feed-forward layers, facilitating more efficient training and better scalability.

4.1.1 Modifications for StARformer

Defining Safety: In both experimental approaches, safety is defined through the association between the current StARformer group (comprising state, action, reward, and safety or safety weight) and non-successful terminal states. This definition is rooted in the hypothesis that certain sequences of states and actions, particularly those leading to non-successful outcomes, are indicative of unsafe scenarios. By identifying and learning from these associations, the model is trained to recognize and avoid patterns that could lead to unsafe or undesirable end states.

Method 1: Integrating a Safety Factor into the Token Group of StARformer

Our approach innovates the StARformer model by incorporating a safety factor into its input, thus forming a four-element tuple (state, action, reward, safety). This safety factor is calculated from historical data to indicate the likelihood of transitioning to a non-successful terminal state from the current StAR group. Training the model with this augmented dataset enables it to

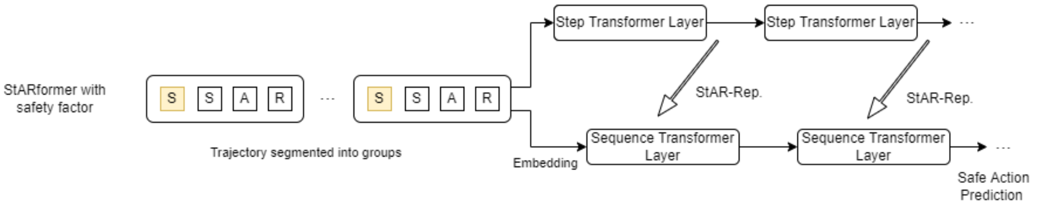

StARformer with safety factor
S
S
A
R
...
S
S
A
R
Trajectory segmented into groups
Embedding
Step Transformer Layer
Step Transformer Layer
...
StAR-Rep.
StAR-Rep.
Sequence Transformer Layer
Sequence Transformer Layer
...
Safe Action Prediction

Figure 4.3: StARformer with safety factor

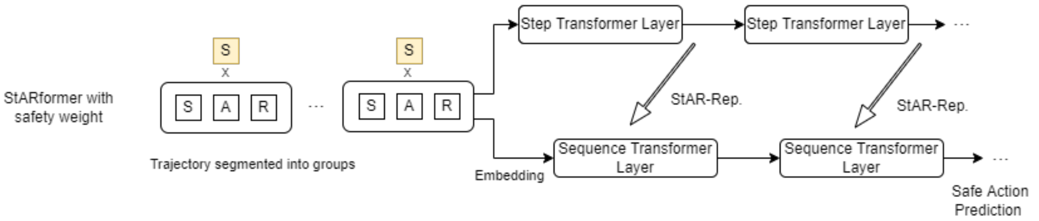

StARformer with safety weight
S
x
S
A
R
...
S
x
S
A
R
Trajectory segmented into groups
Embedding
Step Transformer Layer
Step Transformer Layer
...
StAR-Rep.
StAR-Rep.
Sequence Transformer Layer
Sequence Transformer Layer
...
Safe Action Prediction

Figure 4.4: StARformer with safety weight

discern state-action-reward combinations associated with higher safety risks. The addition of the safety element directly embeds safety considerations into the learning process, allowing the StARformer to develop policies that not only aim to maximize rewards but also uphold safety constraints.

Method 2: Incorporating Safety Weights into StARformer Tokens

We introduce a novel approach in our second method by assigning a safety weight to each token in the StARformer model. These weights, calculated based on each token's association with non-successful terminal states, adjust the reward signal. Tokens linked to historically unsafe outcomes receive higher weights, thereby diminishing the reward in such scenarios.

The StARformer model is enhanced to include safety weights in its input sequence of state, action, and reward. These weights are derived from predefined safety metrics, such as environmental risks or the inherent safety of actions. During training, the reward for each token is scaled by its safety weight, allowing the model to incorporate both traditional goals and safety considerations.

The augmented input for the StARformer model now consists of (state, action, reward, safety weight). By training on this enhanced dataset, the model learns to understand the interplay between actions, rewards, and safety, with the safety weights acting as a dynamic modifier of the reward signal.

Training StARformer with Atari Dataset

StARformer is a replacement of DT, as training and inference procedures remain the same. StARformer can easily operate on step-wise reward without a performance drop. In contrast, it is critical to design a Return-to-go (RTG, target return) carefully in DT, which needs more trials and tuning to find the best value. We use most of the same hyper-parameters as in DT for Atari environments without extra tuning. The StARformer will undergo rigorous training with the Atari dataset, ensuring it learns to process and interpret intricate game dynamics and environmental cues. The training will emphasize the model's ability to discern safe from potentially hazardous actions, essential for its application in RL.

Predicting Safety-Focused Action Sets

Prediction: The output of the last Sequence Transformer layer is used to make action predictions

Approach: Post-training, the StARformer will be evaluated on its capability to predict actions that align with safety criteria. This involves analyzing the model's performance in various game scenarios within the Atari benchmark, focusing on its proficiency in identifying and prioritizing safe actions over high-reward but risky alternatives.

4.2 Pre-controller

Incorporating a Pre-controller within the StARformer framework is a pivotal enhancement in our methodology, aimed at significantly bolstering the model's competency in ensuring safety in Reinforcement Learning (RL) environments. This integration, abstracted from the StARformer and potentially compatible with other RL methods, emphasizes preemptive safety evaluation and decision-making refinement.

4.2.1 Role of the Pre-controller:

Safety Assessment The Pre-controller, acting as a preliminary gatekeeper, rigorously evaluates the safety of potential actions before their execution. Leveraging the StARformer's analytical capabilities, this assessment relies on a set of safety criteria meticulously tailored to the nuances of the specific RL environment. These criteria encompass not only the immediate ramifications of actions but also their potential long-term impacts on safety.

Action Filtering Central to its role, the Pre-controller carefully scrutinizes each action suggested by the RL model. By applying a stringent safety standard, it discerns whether an action is compatible with the established safety framework. Those actions falling short of safety benchmarks are either modified to align with safety norms or outright rejected. This ensures that the action repertoire presented to the environment upholds the highest safety standards.

State Evaluation Beyond action appraisal, the Pre-controller is tasked with a thorough evaluation of the current environmental state. It proactively identifies potential risks or safety hazards that could arise, influencing the choice and nature of future actions. This ongoing assessment helps in anticipating and mitigating potential safety challenges before they materialize.

4.2.2 Developing a Shield-Like Pre-controller

The conceptualization of the Pre-controller as a shield-like entity within the RL ecosystem is a novel approach. This design is akin to a dynamic safeguard, constantly adapting to the evolving landscape of the RL environment. It not only acts as a barrier against unsafe actions but also as a guide steering the RL model towards safer and more effective strategies.

Dynamic Safety Algorithm The core of the Pre-controller is a dynamic safety algorithm that continuously analyzes and learns from both the environment and the outcomes of past actions. This algorithm is embedded within the StARformer's structure, enabling it to make real-time adjustments based on the latest data and insights.

Integration with Learning Process Crucially, the Pre-controller is seamlessly integrated into the StARformer's learning process. This integration allows it to influence the model's learning trajectory, ensuring that safety considerations are ingrained in the model from the very outset of the training phase. The Pre-controller, therefore, plays a dual role of safeguarding immediate actions and shaping the model's long-term learning and decision-making patterns.

Application Across Diverse Environments Given its abstracted nature, this Pre-controller design is versatile enough to be applied across various RL environments, from gaming to more complex, real-world applications like autonomous driving or healthcare. Its adaptability and ability to work in conjunction with different RL methodologies make it a robust tool in the pursuit of safer AI systems.

Chapter 5

Evaluation

5.1 Research Question and Evaluation Metrics

Our study is guided by a primary research question: How does the integration of StARformer in Reinforcement Learning (RL) environments enhance decision-making in terms of safety and efficiency? This question directs our focus towards understanding the impact of StARformer on the quality and safety of decisions made by RL agents in various environments.

We raise two research questions to assess whether our proposal fulfills the objectives in the motivation chapter.

5.1.1 RQ1: To What Extent Does Our Method Improve Performance?

Evaluation metrics We focused on the episodic reward, which is the cumulative reward obtained in each game episode. This metric is vital for gauging our model's overall effectiveness in the Atari environment. By measuring and analyzing these rewards across episodes, we gained insights into the model's decision-making abilities, learning progression, and its adaptability to the game dynamics. The episodic reward, therefore, served as a key indicator of performance enhancement, providing a quantitative basis to evaluate the extent of improvement our method brings to the table in a complex RL setting.

5.1.2 RQ2: To What Extent Does Our Method Improve Safety Concern?

Evaluation metrics We concentrated on evaluating collision frequency as a key metric. This analysis provided crucial insights into the efficacy of the pre-controller mechanism, a core component of the StARformer, in reducing unsafe actions and states. By quantifying and examining the frequency of collisions, we were able to effectively assess the impact of the StARformer mechanism in enhancing safety, thereby demonstrating its practical value in environments where minimizing risk is essential.

5.2 Experimental Setup

In this section, we detail our experimental setup, providing an overview of the experimental environment, training dataset, model architecture, training parameters, and evaluation setup. The detail ensures the transparency and reproducibility of our experiments.

We implement offline reinforcement learning in our experiments. In the experiments, a fixed memory buffer is employed to store the history of sub-optimal trajectories.

5.2.1 DQN Replay Dataset

We utilize the DQN Replay Dataset [31] to train our model. The DQN Replay Dataset provides 200 million frames for each Atari game(60 in total) with sticky actions enabled. The frames can be represented as experience tuples (observation, action, reward, next observation), totaling 50 million instances. We select the game Breakout to conduct our experiment on. We adhere to the approach proposed in [7], selecting 1% of the experience tuples, equivalent to 500k steps.

Training Hyper-parameters We follow [7] and set the training hyper-parameters as 5.1 shows.

5.2.2 Breakout

Breakout, a classic arcade game, serves as a notable benchmark in Reinforcement Learning (RL) [32]. It challenges RL models like the StARformer with its simple yet complex gameplay. In Breakout, the player controls a paddle at the screen's bottom, aiming to break bricks at the top by bouncing a ball

Table 5.1: Hyper-parameters Setting

Hyper parameter	Value
Sequence length	10
Image size	84×84 grayscale
Frame stack	4
Frame skip	2
Layers	6
Image patch size	7
Activation function	GeLU, ReLU
Dropout	0.1
Learning rate	6×10-4
Adam betas	(0.9,0.95)
Weight decay	0.1

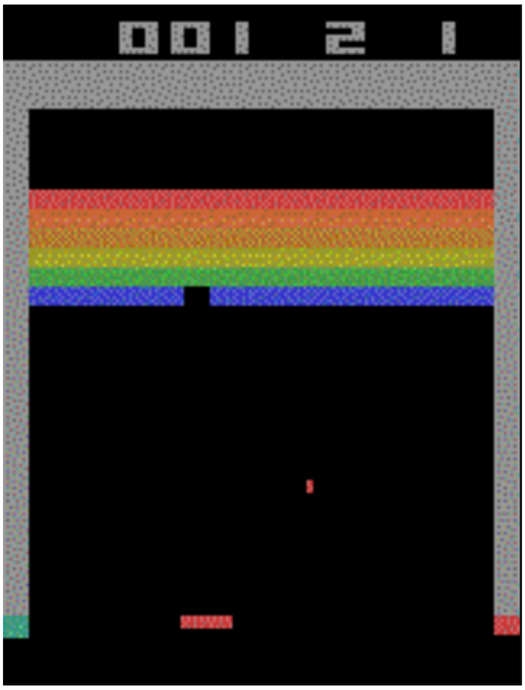
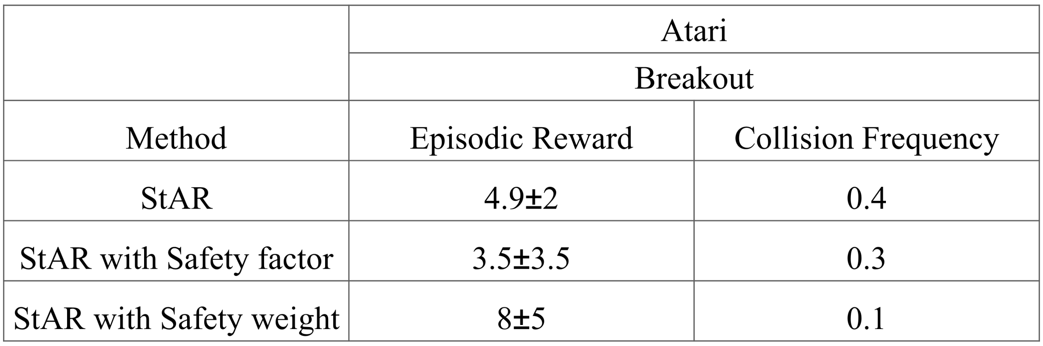


Figure 5.1: Breakout

off the paddle. The game's difficulty escalates as the ball's speed increases with each brick broken. The primary RL task involves learning to maneuver the paddle to keep the ball in play and break all the bricks. Losing the ball results in a lost life, and the game ends when all lives are depleted. The RL model's state includes the positions of the paddle, ball, and bricks, with some models considering the ball's velocity. The action space is discrete: moving the paddle left or right, or keeping it stationary. Rewards are earned for breaking bricks, with potential penalties for losing the ball. This scenario offers a comprehensive testbed for evaluating RL models' efficiency, adaptability, and learning prowess in a dynamic environment.

5.2.3 Evaluation Setup

The evaluation of our method poses significant challenges due to the complex nature of the Arcade Learning Environment [32], characterized primarily by its high-dimensional visual inputs. Additionally, the intricacies of credit assignment are compounded by the inherent delay observed between the execution of actions and the subsequent realization of rewards. In our experimental approach, we draw upon the DQN Replay Dataset, specifically analyzing a subset that constitutes 1 percent of the total samples. This subset represents



	Atari	
	Breakout	
Method	Episodic Reward	Collision Frequency
StAR	4.9±2	0.4
StAR with Safety factor	3.5±3.5	0.3
StAR with Safety weight	8±5	0.1

Figure 5.2: Evaluation of episodic returns and collision frequency in StARformer (StAR) , StARformer with Safety factor and StARformer with Safety weight in Atari. Seq. = 10, Life = 1.

a significant volume of data, amounting to 500 thousand transitions out of the 50 million observed by an online DQN agent during its training phase.

We leverage the Arcade Learning Environment [32] as our evaluation platform, which provides versatile interfaces facilitating game interactions and visualization. For each model derived from distinct epochs, we conduct evaluations 10 times within the Arcade Learning Environment, specifically using the Breakout game. The environment undergoes a reset when the agent exhausts all 5 lives.

To evaluate performance, the rewards are first accumulated and then averaged across the evaluation runs. Additionally, we record the number of steps taken by the agent before the game is over. This step-count metric serves as an indicator of the model's capability to play the game while minimizing unsafe actions.

In light of the absence of a definitive game-over signal in the Arcade Learning Environment. The game concludes when the agent either exhausts all 5 lives or successfully completes the game. To address this ambiguity, we choose to reset the environment upon the loss of 1 life, allowing for a more precise evaluation of safety when assessing the models.

5.3 Analysis of Model Performance

Our approach innovates on the StARformer model by integrating safety factors into its learning process through two distinct methods. The first method involves expanding the token group within the original StARformer structure from a triplet of (state, action, reward) to a quadruplet that includes safety, thus forming (state, action, reward, safety) for training and learning. The second method involves multiplying a safety weight with the reward component of each triplet.

Experimental results indicate that the method of multiplying the safety weight with the reward in each triplet yields the best episodic reward outcomes and the lowest collision frequency. On the other hand, incorporating the safety factor as an element in the token group results in the least favorable episodic rewards but achieves a moderate collision frequency. The original StARformer method ranks in the middle in terms of episodic rewards but exhibits the highest collision frequency. This suggests that directly integrating safety considerations into the reward signal is more effective at enhancing safe behavior in the model compared to simply including safety as an additional learning factor.

5.3.1 Answer to RQ1

Result shows that the original model StARformer achieved an episodic reward of 4.9±2; StARformer with Safety Factor incorporates the safety factor resulted in a lower episodic reward of 3.5±3.5; StARformer with Safety Weight yields the highest episodic reward of 8±5. Our method, which innovates on the StARformer model by integrating safety considerations, shows notable improvements in performance as measured by episodic rewards. Specifically, the approach of multiplying the safety weight with the reward in each state-action-reward (S-A-R) triplet has demonstrated the most effective results. This method yielded the best episodic reward outcomes, indicating a significant enhancement in the model's decision-making abilities and its adaptability to the game dynamics.

5.3.2 Answer to RQ2

Result shows that the original model StARformer exhibited a collision frequency of 0.4; StARformer with Safety Factor slightly improved safety, with a collision frequency reduced to 0.3; StARformer with Safety Weight demonstrated the most significant improvement in safety, reducing the collision frequency to 0.1.

Regarding safety, our method has shown considerable success in reducing collision frequency, a critical safety metric. The method that proved most effective in enhancing safety was again the integration of safety weights with the reward component. This approach significantly reduced unsafe actions and states, as evidenced by the lowest observed collision frequency. This reduction highlights the efficacy of the pre-controller mechanism in the

StARformer, underscoring its practical value in environments where risk minimization is paramount. Conversely, the method involving the addition of a safety factor as an element within the token group also achieved moderate success in reducing collision frequency, though it was less effective than directly influencing the reward signal.

5.3.3 Key Findings

Performance Enhancement Through Safety Integration Our innovative approach in modifying the StARformer model, by integrating safety factors directly into the learning process, abstracting out a pre-controller architecture, resulted in substantial performance improvements. This was particularly evident in the method where the safety weight was multiplied with the reward component, leading to the highest episodic rewards.

Moderate Success with Safety Factor Inclusion Including safety as an additional element within the token group (state, action, reward, safety) also positively impacted safety, demonstrated by a moderate reduction in collision frequency. However, this approach was less effective in improving performance compared to the safety weight multiplication method.

Comparison with Original StARformer Method When compared to the original StARformer method, both of our proposed approaches showed improvements in safety. The original method, while delivering moderate episodic rewards, exhibited the highest collision frequency, indicating a lesser focus on safety.

Balancing Safety and Performance Goals The key to enhancing safety without compromising on performance lies in effectively integrating safety considerations into the reward mechanism. This integration ensures a more nuanced and dynamic approach to achieving the dual objectives of maximizing rewards and adhering to safety constraints.

In summary, our research demonstrates that embedding safety factors into the StARformer model's learning process leads to significant improvements in both performance and safety, offering a robust framework for developing safer and more effective reinforcement learning systems.

Chapter 6

Discussion

6.1 Interpretation of Findings

Our comprehensive analysis has delved deeply into the StARformer model, a groundbreaking advancement in Reinforcement Learning (RL) that uniquely intertwines safety with performance. The essence of our study has been to unravel and understand the delicate balance the StARformer maintains between these two pivotal aspects within the complex landscape of RL.

A crucial aspect of our investigation centered on the model's resource utilization and scalability. Although the StARformer demonstrates efficiency in learning and synthesizing the pre-controller, it faces challenges regarding computational resource demands, particularly when scaled up for more intricate tasks and complex environments. Nevertheless, the model impressively sustains consistent performance across diverse scenarios, indicating its vast potential for real-world applications.

Central to the StARformer model is the interplay between safety and performance. Our detailed examination of safety-specific metrics, such as the rate of unsafe actions and compliance with established safety rules, shed light on the model's prioritization of safety within its decision-making processes. Concurrently, we assessed traditional performance metrics, including reward accumulation and task completion rates, to gauge the impact of the model's safety orientation on overall performance. This bifocal analysis brought into focus the trade-offs navigated by the StARformer, often opting to forgo immediate rewards to bolster safety, thereby epitomizing the intricate balancing act it successfully achieves in the realm of RL.

6.2 Comparison with Existing Models

Our comparative analysis reveals that the StARformer distinguishes itself significantly from existing Reinforcement Learning (RL) models and safety mechanisms. A key differentiation is its approach to safety; unlike many RL models where safety is an afterthought or an external constraint, the StARformer integrates safety considerations directly into the learning process. This integration not only ensures holistic adherence to safety standards but also empowers the model to dynamically adapt to various safety criteria. The StARformer's proactive safety measures, such as predictive risk assessments and mitigation strategies, mark a stark contrast to other models that primarily focus on reward optimization.

In terms of architectural advancements, the StARformer diverges from conventional RL models that typically rely on deep Q-networks or policy gradient methods. Instead, it capitalizes on the Transformer's ability to process sequential data, thereby more effectively capturing temporal dependencies and contextual nuances crucial in RL tasks. This architectural innovation contributes to its enhanced performance in diverse and complex environments.

When evaluating efficiency in learning, the StARformer demonstrates superior speed and resource efficiency in learning optimal policies compared to traditional models. Its proficiency in parallel processing and handling long sequences contributes to faster convergence and more efficient learning, particularly evident in environments that demand long-term strategic planning and rapid adaptation to changing scenarios.

Furthermore, the StARformer's integrated safety approach stands out when compared with existing safety mechanisms in RL. Traditional safety approaches often view safety as a constraint applied either externally or as a post-hoc correction. In contrast, the StARformer's design allows for a more consistent and integrated approach to safety, enhancing its adaptability to different safety standards and its ability to dynamically adjust decision-making processes based on safety assessments. This proactive approach to safety, combined with its predictive capabilities, positions the StARformer as a unique and advanced model in the realm of SafeRL, offering new possibilities in terms of efficiency, adaptability, and, most importantly, safety in RL applications.

6.3 Practical Implications

Our findings reveal that the StARformer model holds significant practical implications, particularly in real-world scenarios where safety is crucial. Its ability to process complex sequences and predict outcomes makes it versatile and applicable across various domains, especially those requiring decision-making under uncertainty.

In the realm of autonomous vehicles, the StARformer enhances decision-making capabilities, especially in unpredictable traffic conditions, by predicting and mitigating potential road hazards, thus improving safety in dynamic environments. This ability also extends to efficient route planning, where it can optimize routes not just for speed but also for safety.

In robotics, the StARformer ensures safe human-robot interactions, crucial in industries like manufacturing and services. By predicting human actions, it adjusts robot behavior for safer interactions. Additionally, it can develop adaptive control systems for robots in changing conditions, maintaining consistent performance and adhering to safety protocols.

In healthcare, the StARformer's potential is particularly impact. Robots equipped with this model in patient care can make safer and more informed decisions, vital in sensitive tasks like surgery or elderly care. Its proficiency in analyzing complex patient data sequences aids in creating personalized treatment plans, enhancing the efficacy of treatments while foreseeing and mitigating potential risks.

The model's applications extend to financial trading, where it can make safer investment decisions in uncertain market conditions by effectively balancing risks and rewards. In energy management, the StARformer optimizes the distribution and consumption of resources, foreseeing potential issues and adapting to demand fluctuations, thereby ensuring efficient and safe energy management.

In summary, the StARformer's versatility and proficiency in ensuring safety while making informed decisions under uncertainty open up a wide range of applications. Its integration into diverse domains promises not only enhanced performance but also a significant elevation in safety standards, making it a pivotal tool in advancing current technologies.

6.4 Challenges and Limitations

During the implementation of the StARformer model in Reinforcement Learning (RL) environments, various challenges and limitations were encountered, providing key insights into the model's current capabilities and highlighting potential areas for future development.

One primary challenge was the collection and availability of high-quality, diverse datasets that accurately reflect complex real-world scenarios. The lack of sufficient or relevant data at times hindered the model's learning effectiveness. Additionally, inherent biases in training data posed significant challenges, as they could lead to skewed learning outcomes. Ensuring the data was representative and unbiased was crucial yet often difficult.

The preprocessing of raw RL data to fit the Transformer architecture's requirements was another complex task. This process was essential for the model to capture the nuances of different environments, but it sometimes limited its ability to fully comprehend certain aspects of these environments.

The Transformer architecture is known for its high demand for computational resources, and this was evident in the resource intensity faced when scaling the model for more complex tasks or larger datasets. Furthermore, the complexity of the model and the size of the datasets often resulted in lengthy training times, which posed a challenge for rapid development and iteration.

Regarding scalability, while the StARformer showed proficiency in certain environments, extending its capabilities to a wide variety of RL tasks and environments was challenging. This included adapting the model to different types of state spaces and action dynamics. As the complexity of tasks increased, the model's scalability was at times limited, requiring a careful balance between model complexity and computational feasibility.

The model's ability to generalize across different tasks and not just perform well in its training environments was a significant challenge. Additionally, customizing the model for specific tasks, particularly those with unique safety requirements or reward structures, required substantial effort and posed challenges in terms of model adaptability.

In summary, the implementation of the StARformer model faced several challenges in data handling, computational resources, scalability, and adaptability. These challenges underscore the need for further research and development to enhance the model's effectiveness and broaden its applicability across various RL scenarios.

6.5 Future Research Directions

Reflecting on our research, the StARformer model emerges as a groundbreaking development in Safe Reinforcement Learning (SafeRL), setting new standards in adaptability, generalization, and integrated safety. However, our journey has also been marked by challenges related to data limitations, computational demands, and scalability. These hurdles have not only provided valuable insights but have also illuminated paths for future research and development in the field.

The potential implications of our work in SafeRL extend well beyond this study, opening doors to explore StARformer's integration in more varied and intricate environments. Future research can build upon our foundational methodology, refining the model and extending its application to scenarios where balancing safety and performance is crucial.

Advancing the model's adaptability to complex environments is a key area for future exploration. This includes developing enhanced learning algorithms to better equip the StARformer for dynamic and multifaceted settings and expanding its cross-domain adaptability to bolster its capacity to generalize across various domains. Enhancing the model's proficiency in handling high-dimensional data will also be instrumental in broadening its real-world applicability.

In terms of safety, future iterations could focus on advancing risk assessment algorithms, enabling the StARformer to more accurately predict and mitigate potential hazards. Incorporating real-time safety monitoring features would allow for immediate action adjustments in response to changing environmental conditions, further bolstering its safety capabilities.

Integrating the StARformer with other RL methodologies presents another exciting avenue. Hybrid models could combine the strengths of various approaches, leading to more robust systems. Collaborative learning systems, especially in multi-agent settings, could facilitate shared learning dynamics, enhancing overall system intelligence.

Addressing the computational challenges faced by the StARformer is another critical area for future research. Efforts to optimize computational efficiency are necessary to make the model more practical and scalable. Employing distributed and parallel computing techniques could mitigate the high computational demands, particularly for large-scale applications, making the StARformer more accessible for widespread use.

In conclusion, our exploration of the StARformer in SafeRL not only contributes significantly to the current landscape but also lays a solid foundation for future advancements. By addressing these challenges and building on the model's strengths, there is immense potential to further revolutionize SafeRL, making it more adaptable, efficient, and safe for a wide range of applications.

6.6 Contribution to the Field

Our research marks a substantial contribution to the field of Safe Reinforcement Learning (RL), with the development and implementation of the StARformer model. This advancement redefines the integration and prioritization of safety within RL systems, setting a new standard in the field.

A significant achievement of our work is the integrated safety approach adopted by the StARformer. Unlike traditional RL methods where safety is often secondary or externally imposed, the StARformer inherently embeds safety within its core architecture. This ensures that safety considerations are an intrinsic part of the policy learning and execution process, elevating the importance of safe decision-making within the RL framework.

Another crucial aspect of our contribution lies in balancing safety with performance. We tackle one of the most pressing challenges in Safe RL – achieving high task performance without compromising safety. The StARformer model demonstrates that it is possible to maintain high performance levels while ensuring safety, a critical advancement for applying RL in real-world scenarios.

In addition to these contributions, our research enhances the understanding of sequential decision-making in RL. The Transformer-based architecture of the StARformer offers novel insights into managing temporal dynamics, showcasing the importance of understanding and leveraging sequential data for effective decision-making. This approach is particularly impactful in environments where decisions have long-term implications.

Furthermore, the StARformer's capacity for long-term planning and sophisticated risk assessment deepens our understanding of these critical aspects in Safe RL. This is especially pertinent in scenarios with far-reaching consequences, where short-sighted decisions can lead to significant risks or failures. Our work with the StARformer thus not only advances the technical aspects of Safe RL but also contributes to a more holistic and nuanced approach to decision-making, risk assessment, and planning in complex environments.

Chapter 7

Conclusion

Our research has led to pivotal findings regarding the application of the synthesis of pre-controller using StARformer in Reinforcement Learning (RL) tasks. We successfully integrated the StARformer into safe RL environment, demonstrating its effectiveness in enhancing performance of ensuring safety. A significant highlight is the model's role in improving safety, a critical aspect often overlooked in traditional RL approaches. Our experiments and analyses revealed key achievements of the StARformer, such as its proficiency in strategic decision-making and its capability to reduce unsafe actions significantly, thereby enhancing overall safety in diverse RL scenarios.

The integration of the StARformer model into the synthesis of pre-controller in RL signifies an advancement in the field, particularly in the realm of Safe RL. This research contributes to the evolution of sequence modeling approaches within RL, offering a novel perspective on how safety can be inherently embedded into the learning process.

The implications of our findings for future research are profound. Our work lays the groundwork for further exploration into the integration of safety considerations in RL models. It opens new avenues for applying the StARformer in more diverse and complex environments, pushing the boundaries of what is currently achievable in Safe RL. Future research can build upon our methodology to explore further the intersection of sequence modeling and safety in RL, potentially leading to even more sophisticated and robust RL systems.

Reflecting on the research process, we acknowledge the challenges we faced, including data limitations, computational resource constraints, and the scalability of the model. These challenges provided valuable lessons and insights, shaping the course of our research. Acknowledging these limitations highlights areas for improvement in future work. We also discuss the practical challenges in implementing advanced models like StARformer in real-world scenarios, underscoring the need for continued innovation and adaptation.

As RL continues to permeate various aspects of technology and daily life, the development of models that prioritize safety in decision-making processes becomes increasingly important. Our work with the StARformer model represents a step forward in this direction, contributing to the creation of more reliable, efficient, and safe AI systems for the future.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction.* MIT press, 2018.

[2] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming.* John Wiley & Sons, 2014.

[3] B. Jang, M. Kim, G. Harerimana, and J. W. Kim, "Q-learning algorithms: A comprehensive classification and applications," *IEEE access*, vol. 7, pp. 133 653–133 667, 2019.

[4] D. Zhao, H. Wang, K. Shao, and Y. Zhu, "Deep reinforcement learning with experience replay based on sarsa," in *2016 IEEE symposium series on computational intelligence (SSCI).* IEEE, 2016, pp. 1–6.

[5] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Processing Magazine*, vol. 34, no. 6, pp. 26–38, 2017.

[6] M. Janner, Q. Li, and S. Levine, "Offline reinforcement learning as one big sequence modeling problem," *Advances in neural information processing systems*, vol. 34, pp. 1273–1286, 2021.

[7] J. Shang, X. Li, K. Kahatapitiya, Y.-C. Lee, and M. S. Ryoo, "Starformer: Transformer with state-action-reward representations for robot learning," *IEEE transactions on pattern analysis and machine intelligence*, 2022.

[8] J. Garcıa and F. Fernández, "A comprehensive survey on safe reinforcement learning," *Journal of Machine Learning Research*, vol. 16, no. 1, pp. 1437–1480, 2015.

[9] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

[10] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, pp. 9–44, 1988.

[11] L. Chen, K. Lu, A. Rajeswaran, K. Lee, A. Grover, M. Laskin, P. Abbeel, A. Srinivas, and I. Mordatch, "Decision transformer: Reinforcement learning via sequence modeling," *Advances in neural information processing systems*, vol. 34, pp. 15 084–15 097, 2021.

[12] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[13] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever *et al.*, "Improving language understanding by generative pre-training," 2018.

[14] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[15] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, "Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition," *Advances in Neural Information Processing Systems*, vol. 34, pp. 11 960–11 973, 2021.

[16] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "Vivit: A video vision transformer," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 6836–6846.

[17] M. Chen, A. Radford, R. Child, J. Wu, H. Jun, D. Luan, and I. Sutskever, "Generative pretraining from pixels," in *International conference on machine learning.* PMLR, 2020, pp. 1691–1703.

[18] K. Han, A. Xiao, E. Wu, J. Guo, C. Xu, and Y. Wang, "Transformer in transformer," *Advances in Neural Information Processing Systems*, vol. 34, pp. 15 908–15 919, 2021.

[19] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "Coatnet: Marrying convolution and attention for all data sizes," *Advances in neural information processing systems*, vol. 34, pp. 3965–3977, 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[21] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[22] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision.* Springer, 2020, pp. 213–229.

[23] N. Kitaev, L. Kaiser, and A. Levskaya, "Reformer: The efficient transformer," in *International Conference on Learning Representations*, 2019.

[24] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *ICML*, vol. 2, no. 3, 2021, p. 4.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[26] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," *Advances in neural information processing systems*, vol. 34, pp. 30 392–30 400, 2021.

[27] H. Touvron, M. Cord, A. Sablayrolles, G. Synnaeve, and H. Jégou, "Going deeper with image transformers," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 32–42.

[28] J. Garcia and F. Fernández, "Safe exploration of state and action spaces in reinforcement learning," *Journal of Artificial Intelligence Research*, vol. 45, pp. 515–564, 2012.

[29] O. Mihatsch and R. Neuneier, "Risk-sensitive reinforcement learning," *Machine learning*, vol. 49, pp. 267–290, 2002.

[30] A. Hans, D. Schneegaß, A. M. Schäfer, and S. Udluft, "Safe exploration for reinforcement learning." in *ESANN*, 2008, pp. 143–148.

[31] R. Agarwal, D. Schuurmans, and M. Norouzi, "An optimistic perspective on offline reinforcement learning," in *International Conference on Machine Learning.* PMLR, 2020, pp. 104–114.

[32] M. G. Bellemare, Y. Naddaf, J. Veness, and M. Bowling, "The arcade learning environment: An evaluation platform for general agents," *Journal of Artificial Intelligence Research*, vol. 47, pp. 253–279, 2013.

Acknowledgement

This academic journey, filled with challenges and enriching experiences, would not have been possible without the support and contributions of many. To everyone who played a part in this significant phase of my life, I extend my deepest gratitude.

At the forefront of my gratitude are Professor Shinichi Honiden and Associate Professor Kenji Tei from Waseda University. Their unparalleled guidance, unwavering support, and profound academic insights have been instrumental in shaping both my research and personal growth. The honor and privilege of working under their tutelage cannot be overstated.

I owe a special acknowledgment to my senior colleagues, Jialong Li and Jinyu Cai. Their wisdom-rich mentorship has illuminated my research path, offering encouragement and insightful discussions that have significantly deepened my passion and understanding of our field.

The support from Tianchen Wang deserves a particular mention. His assistance was critical in the successful realization of my experiment. My heartfelt appreciation also goes to all members of our lab. Your camaraderie, support, and shared knowledge have been the pillars supporting my journey through this Master's program.

To my dear family and friends in both China and Japan, old and new alike, your unwavering love, support, and encouragement have been my fortress. The belief you've shown in me, whether through mental support, physical visits, or shared adventures, has been a source of boundless joy and inspiration.

I am indebted to everyone mentioned, as well as to those who have supported me in countless other ways, named or unnamed. Your contributions, encouragement, and faith in my abilities have been vital to the success of this research.

To all, I extend my best wishes in your future endeavors, just as you have done for mine.