Evaluating Prior Knowledge of ARC Using World Models

Seungpil Lee [1], Jihwan Lee [1], Sundong Kim [1]

[1]Gwangju Institute of Science and Technology

Abstract

AI research has achieved notable success in solving specific tasks; however, progress in addressing generalized problems remains limited. One reason for this is the lack of effective methods to test universal cognitive abilities. The ARC benchmark has been introduced to precisely measure reasoning abilities. Yet, the impact of prior knowledge and the analysis of the types of prior knowledge required to solve the ARC benchmark have not been thoroughly examined, making research in this area challenging. In this study, we propose a new method using World Model algorithm to analyze the influence of prior knowledge when solving the ARC benchmark and identify the types of prior knowledge embedded in ARC.

Introduction

The difficulty in general artificial intelligence (AGI) research lies in the challenge of understanding what composes reasoning abilities and how to assess them. The Abstraction and Reasoning Corpus (ARC) benchmark was developed to precisely measure the inference capabilities of artificial intelligence. ARC is distinctive for minimizing the amount of prior knowledge and data required to solve problems, focusing solely on evaluating inference abilities. However, two key aspects have not been thoroughly researched: **1) the impact of prior knowledge** and **2) the specific types of prior knowledge needed**. As a result, there are challenges in determining whether ARC serves as appropriate dataset for evaluating only inference abilities and distinguishing the difficulty levels of each ARC problem.

In this study, we propose a novel approach using World Model algorithm to analyze prior knowledge in the ARC benchmark. World Model, a reinforcement learning algorithm that extracts prior knowledge inherent in the environment for decision-making, is employed to analyze the impact and the types of prior knowledge required when solving ARC.

Research Objectives

- **Objective 1:** Analyzing the impact of prior knowledge when solving ARC.
- **Objective 2:** Determining the types of prior knowledge required to solve ARC.

Background: ARC

The ARC benchmark is a newly developed benchmark for measuring reasoning ability. It has achieved some success in capturing the differences between humans and artificial intelligence. Humans achieve an average accuracy of 80%, while the best performing model to date achieves 30% accuracy, and transformer-based models achieve only 10% accuracy.

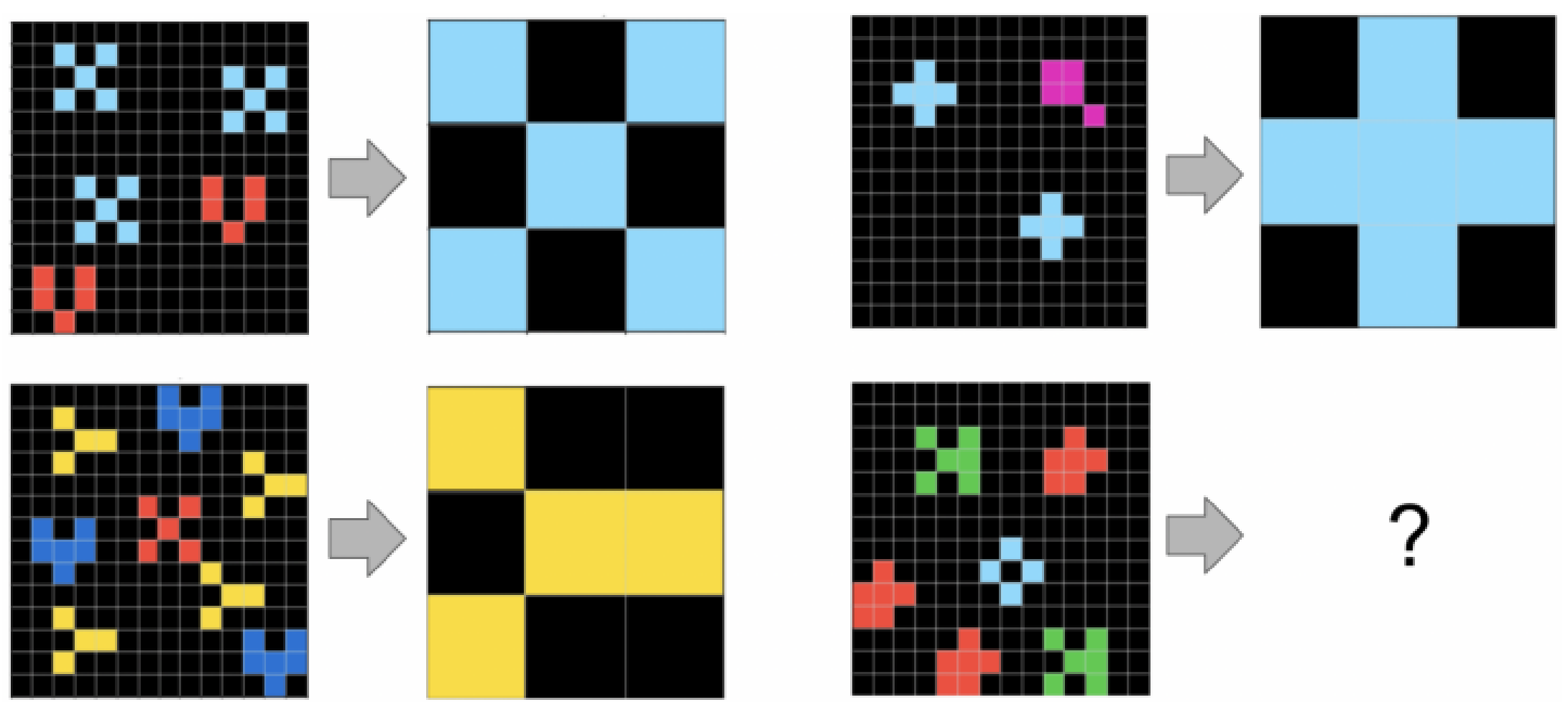

Figure 1. One example of ARC problem

Nevertheless, the current ARC benchmark has the disadvantage of not specifying what the minimum prior knowledge is required to solve each problem. This means that it does not minimize the prior knowledge given to the model before solving a problem, and thus it does not satisfy the definition of intelligence in the strict sense.

Actor-Critic

In this study, we will use Actor-Critic as an algorithm for solving real ARC problems based on the prior knowledge extracted by World Model. Actor-Critic is a reinforcement learning algorithm that learns policy and value simultaneously. It can learn faster and more stably than other reinforcement learning algorithms by separating the module that acts and evaluates the given situation.

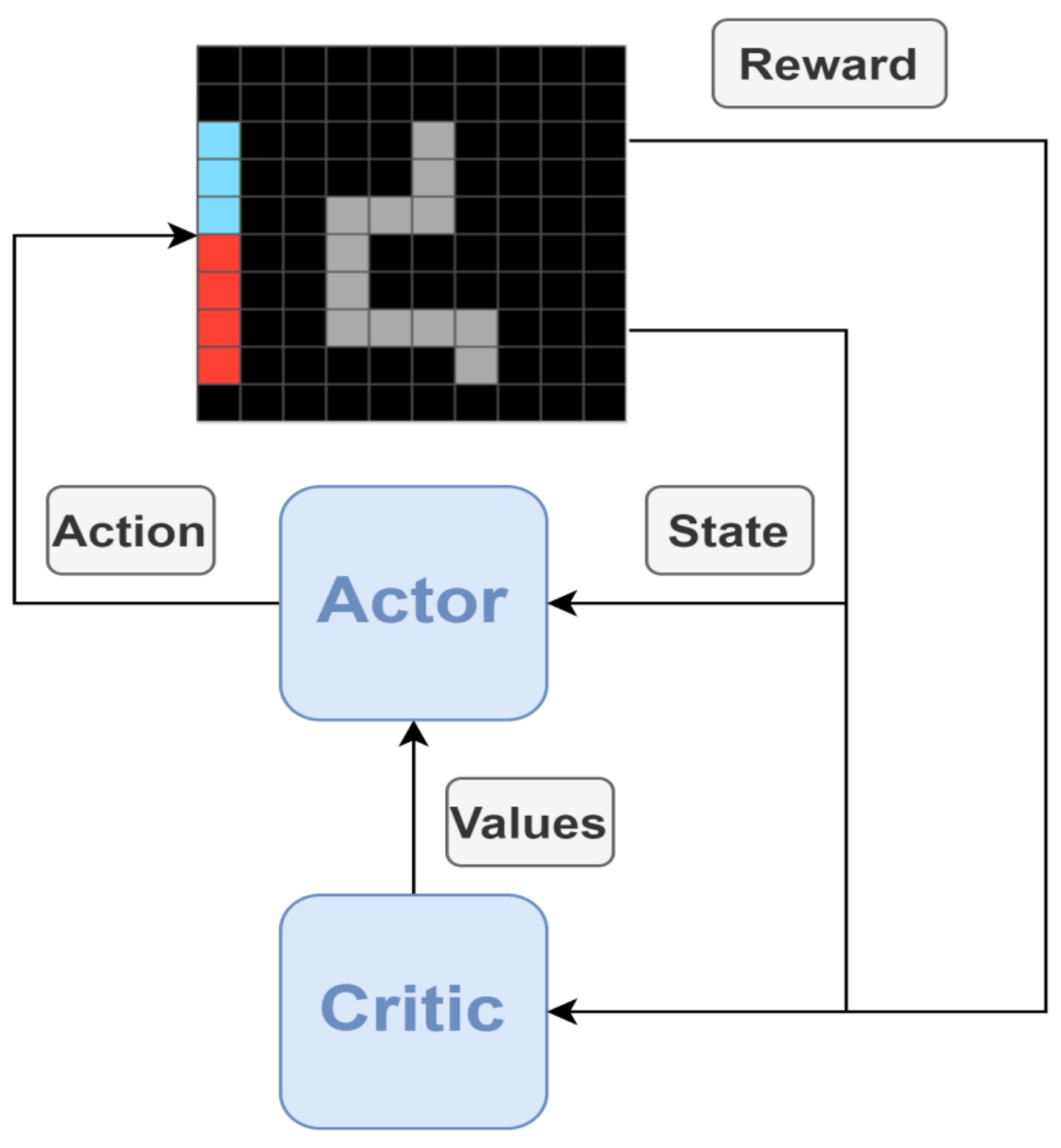

Figure 2. The way pure Actor-Critic learns

Time-Invariant DreamerV3

DreamerV3 is a reinforcement learning model based on a world model. It consists of two parts: World Model and Actor-Critic. World Model extracts prior knowledge about the domain from images, while Actor-Critic uses the extracted prior knowledge to decide the actual action.

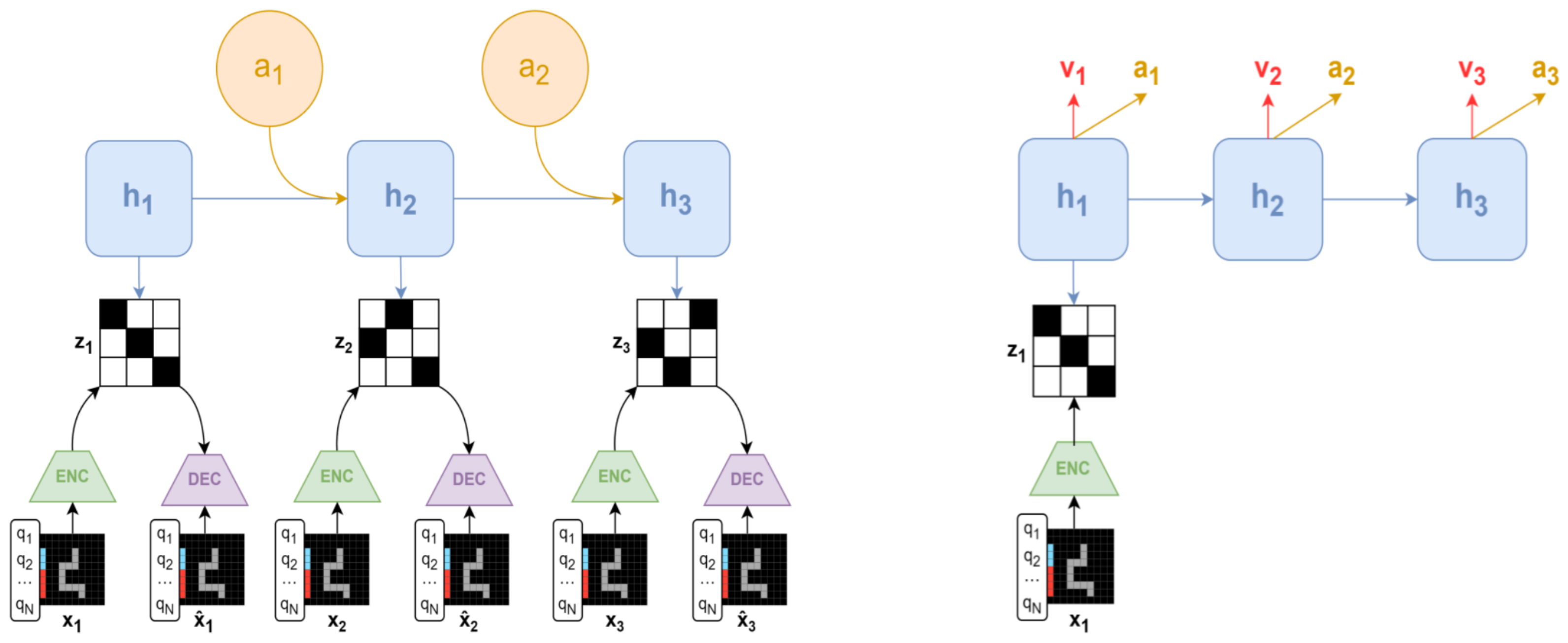

Figure 3. Learning procedure of DreamerV3: Left for World Model and Right for Actor-Critic

DreamerV3 has solved the Minecraft task, which was not solvable by traditional artificial intelligence methods, due to the large amount of information (gravity, various objects, and interactions) and the rapid changes. Since DreamerV3 has shown strong performance in extracting prior knowledge embedded in the environment, we adopt DreamerV3 for analysis among several World Model implementations.

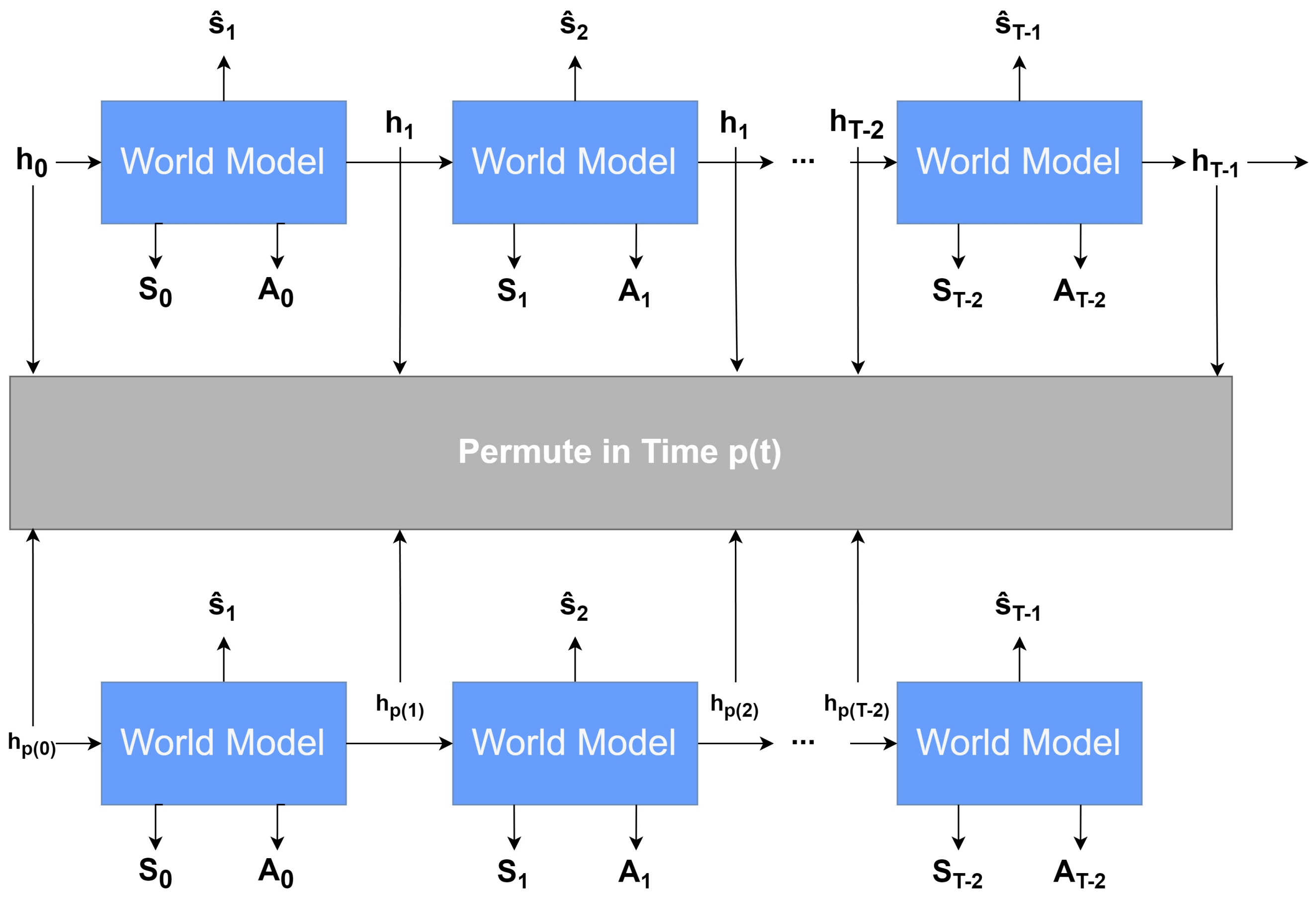

Figure 4. The way permuting time to add time-invariance on DreamerV3

In addition to this, we propose to add time invariance on DreamerV3 for smooth comparison of extracted features. In general, it is known that the feature vectors extracted by the model have a large covariance and are difficult to analyze. Christopher Reale et al. showed that adding time invariance to World Model makes feature vectors linearly, which is easy to analyze. This is because by adding a loss function with shuffled time, model is able to extract the unique features of environment that do not change over time.

Experiment Proposal

The study aims to analyze the impact of prior knowledge in solving ARC problems and to identify the types of prior knowledge included in ARC problems. To this end, two experiments are proposed.

First, the performance of DreamerV3 (World Model + Actor-Critic) and a pure Actor-Critic model in solving ARC problems will be compared to see how much the prior knowledge helps problem solving. It is expected that the difference in accuracy and solution time between the two models will be greater if solutions are strongly affected by prior knowledge.

Second, the features extracted by DreamerV3 will be clustered to see which problems share similar solutions and how the entire problem can be divided into several types. Since the feature vectors extracted represent the prior knowledge of the task, it is expected that this process will identify the types of prior knowledge required to solve the ARC benchmark.

The results of these experiments are expected to provide insights into the role of prior knowledge in reasoning ability and to help develop more effective AI models.

Conclusions

- **By analyzing the amount of prior knowledge** of ARC benchmark, we can confirm how well it meets its goal of using only prior knowledge.
- **By determining the types of prior knowledge** needed for each ARC problem, we provide a new criterion for subdividing ARC problems.

KSC 2023 Poster

iamseungpil@gm.gist.ac.kr