Full World Model

Hadelin de Ponteves

February 10, 2024

Abstract

This guide delves into the Full World Model within artificial intelligence, focusing on the mathematical formulations that underpin the optimization of its core components: the Vision Model, Memory Model, and Controller. By elaborating on the optimization objectives with detailed mathematics, we aim to provide a clear understanding of how these models are trained to achieve efficient and effective decision-making in complex environments.

Contents

1 Introduction

The Full World Model in reinforcement learning decomposes the agent's understanding and interaction with its environment into three interconnected components: the Vision Model (V), the Memory Model (M), and the Controller Model (C). This document explores the detailed mathematical formulations that underpin these models.

2 Vision Model (V)

The Vision Model processes raw, high-dimensional observations, such as RGB frames from the environment, into a more manageable, lower-dimensional representation.

2.1 Formulation

Given an observation $s_t \in \mathbb{R}^n$ at time step t, where n is the dimensionality of the observation space, the Vision Model employs a convolutional neural network (CNN) to encode s_t into a compact representation $v_t \in \mathbb{R}^m$:

$$v_t = V(s_t; \theta_V), \tag{1}$$

where θ_V are the parameters of the Vision Model. The CNN effectively captures spatial hierarchies in the input states.

3 Memory Model (M)

The Memory Model captures temporal dependencies and dynamics, allowing the agent to make informed decisions based on past observations.

3.1 Formulation

The Memory Model updates its hidden state h_t based on the current encoded observation v_t and the previous hidden state h_{t-1}:

$$h_t = \text{LSTM}(v_t, h_{t-1}; \Theta_M), \tag{2}$$

where LSTM denotes a Long Short-Term Memory unit with parameters Θ_M, and $h_t, h_{t-1} \in \mathbb{R}^p$. The LSTM is capable of maintaining information across time steps for sequences of data.

4 Controller Model (C)

The Controller Model directly maps the current state representation and memory to an action.

4.1 Formulation

The action $a_t \in \mathbb{R}^k$, where k is the dimensionality of the action space, is determined by:

$$a_t = \tanh(W_c[v_t; h_t] + b_c), \tag{3}$$

where $W_c \in \mathbb{R}^{k \times (m+p)}$ and $b_c \in \mathbb{R}^k$ are the weights and biases of the Controller Model. The tanh function ensures that the actions are within a bounded range.

5 Integration and Training

The models are trained to maximize cumulative rewards. The parameters of V and M can be trained using gradients derived from the loss between predicted and actual future observations, while C is trained using reinforcement learning techniques to maximize expected returns.

5.1 Objective

The objective is to maximize the expected cumulative reward:

$$\max_{\Theta} \mathbb{E}\left[\sum_{t=0}^{T} \gamma^t R(s_t, a_t)\right], \tag{4}$$

where $R(s_t, a_t)$ is the reward function, $\gamma \in (0, 1]$ is the discount factor, and Θ represents the parameters of all three models.

6 Conclusion

The Full World Model offers a structured approach to decomposing complex environments into manageable components, facilitating the learning of rich representations and behaviors. Through its distinct yet integrated models, it achieves a comprehensive understanding and interaction with the environment, exemplifying a significant advancement in reinforcement learning.