Description of the Hidden State of the World

September 25, 2023
New Bulgarian University
CSAI 2023

Dimiter Dobrev

d@dobrev.com

Institute of Mathematics and Informatics
Bulgarian Academy of Sciences

There are three ways to approach the creation of Artificial Intelligence:

1. Full Observability. The world has no internal (hidden) states.

   Image Recognition (Neural Networks)

2. Partial Observability which we reduce to Full Observability. The world has hidden states but we manage to forget about them.

   ChatGPT (Neural Networks)

3. Real Partial Observability. We describe the hidden state of the world and on that basis predict the future.

   Artificial General Intelligence

What are the weaknesses of the first two approaches?

1. The *Life Experience* should be done by a teacher.

   Our own life is full of mistakes and cannot be used for finding of the approximation function.

2. We need a huge *Life Experience*.

   When we describe life without understanding only on the basis of exceptions, then we need many examples.

3. Too much information.

   (in the second approach when the argument is the whole life)

   It is better to compress all life and represent it through the description of the current state of the world.

4. The resulting program talks like a parrot.

We are not creating a model of the world. We are only imitating the teacher.

5. We do not conduct logical reasoning.

There are two types of reasoning. The first is instant or one-step. The second is logical or multi-step. For the second type of reasoning, we need a model of the world and a description of its internal state. Exactly the multi-step transformation of the internal state is the logical reasoning.

6. We answer the question “What to do?” without understanding what is going on.

It means we have no understanding.

Two questions:

We are looking for an answer to the question:

"What to do?" or "What should be my next action?"

This is the main question, but before we ask it, we should ask the supplementary (auxiliary) question:

"What's going on?"

The first two approaches

Only the question “What to do?” is asked there.

To answer this question, you need a goal (purpose, rewards).

This is the reason why a goal must be introduced in the first two approaches.

The third approach

Here we have to answer both questions, but in this report we will only look for an answer to the question “What’s going on?”.

We assume that if we understand life and know what is going to happen, then if we choose a goal, we will know what to do to achieve that goal.

Therefore, in this report we will look for an explanation of life without being interested in the goal.

What is given and what are we looking for?

We have a finite sequence of actions and observations:

$$a_0, o_0, a_1, o_1, \dots , a_{t-1}, o_{t-1} \qquad (a_i \in \Sigma, o_i \in \Omega)$$

We call this sequence a game, the life, or lived experience.

In the first two approaches:

We are looking for an approximation of the teacher's behavior.

In the third approach:

1. We look for an explanation of the world and on this basis we predict the future.
2. Based on this, and based on some goal, we choose the action that will ensure the best future.

The first approach (*Full Observability*)

The *Life Experience* is a set of tuples:

$$Life = \{ \langle o_i, a_{i+1} \rangle \mid i < t \}$$

We can represent it as a sequence in which the order is not important.

$$o_0, a_1, o_1, a_2, \ldots , o_{t-1}, a_t, o_t$$

In this case we are looking for an approximation of some function f.

$$\forall i \; (i < t) \;\; a_{i+1} = f(o_i)$$

If f' is the approximation of f, then the next action should be:

$$a_{t+1} = f'(o_t)$$

The second approach (*Partial* to *Full Observability*)

In this case we are looking for a policy:

$$f(\,life\,) = a_t$$

$$f(\{a_0, o_0, \ldots\ , a_{t-1}, o_{t-1}\}) = a_t$$

Here the *Life Experience* is a set of tuples:

$$Life = \{\, \langle life, action \rangle \mid finite\ number\ of\ tuples\}$$

If f' is the approximation of *f*, then the next action should be:

$$a_t = f'(current\ life)$$

What is an explanation of life?

or a model of the world
or a description of life.

To us, to explain the life is to say how it will continue (to predict the future).

The model of the world

We assume that life is the result from the interaction between a world and an agent. During this interaction, the world changes its internal state from s_0 to s_t (*Partial Observability*).

$$s_0, a_0, o_0 \quad \ldots \ , s_{t-1}, a_{t-1}, o_{t-1}, s_t$$

We assume that the agent can do whatever he wants. Therefore, the explanation of life is a description of the world (model of the world).

The model of the world consists of:

S – the set of internal states.

$s \in S$ – the current state ($s = s_t$).

$g: S \times \Sigma \rightarrow \Omega \times S$ – function from state and action to observation and new state.

$$\forall i \ (0 \leq i < t) \ \ g(s_i, a_i) = \langle o_i, s_{i+1} \rangle$$

We replace the function with a relation

g: $S \times \Sigma \rightarrow \Omega \times S$ – function.

$$\forall i \ (0 \le i < t) \ \ g(s_i, a_i) = \langle o_i, s_{i+1} \rangle$$

$g \subseteq S \times \Sigma \times \Omega \times S$ – relation.

$$\forall i \ (0 \le i < t) \ \ \langle s_i, a_i, o_i, s_{i+1} \rangle \in g$$

The relation is accurate in the forward direction if:

$$\forall s_1 \ \forall a \ \exists! \, o \ \exists! \, s_2 \ \ \langle s_1, a, o, s_2 \rangle \in g$$

The relation is accurate in the backward direction if:

$$\forall s_2 \ \forall a \ \forall o \ ! \, s_1 \ \ \langle s_1, a, o, s_2 \rangle \in g$$

Example of a world model

Let the world be a game between the agent and some opponent that is part of the world.
Let the agent's observation be exactly what the opponent's action is.

Let the game be chess.

$$s_0, a_0, s'_0, o_0, s_1 \ \ldots \ , s_{t-1}, a_{t-1}, s'_{t-1}, o_{t-1}, s_t$$

s_i – the position of the chessboard when it is agent's turn.
s'_i – the position of the chessboard when the opponent is on a turn.
a_i – the agent's move.
o_i – the opponent's move.

$a_i = f(s_i)$ – the agent's move.
$o_i = g'(s'_i)$ – the opponent's move.
$s'_i = h_1(s_i, a_i)$, $s_{i+1} = h_2(s'_i, \mathrm{o}_i)$ – the rules of the game.

$$g(s_i, a_i) = \ \langle g'\big(h_1(s_i, a_i)\big), h_2\left(h_1(s_i, a_i), g'\big(h_1(s_i, a_i)\big)\right)\rangle$$

What explanation are we looking for?

We will look for an explanation that is as simple and accurate as possible.

Occam's razor principle: The simplest explanation is preferable to one that is more complex.

The more accurate the model is, the more accurately we can predict the future.

Does the world have a model?

The loosest explanation of the world:

Let S be a singleton ($S=\{s\}$) and let

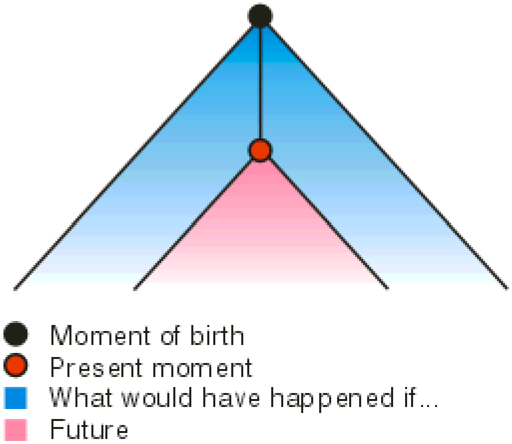
$$g = \{s\} \times \Sigma \times \Omega \times \{s\}$$

This explanation is useless because with it every observation is possible.

We need a more accurate model.

An example of an accurate model

To describe the world, we must first describe the set of its internal states *S*. The reachable states are the nodes of this tree:



For this description we can use any countable set. Let's take Σ^*.

$$S \leftrightarrow \Sigma^*, \qquad s_i \leftrightarrow a_0 a_1 \ldots a_{i-1}$$

The observations will be defined by the function g. On the life we will define the observations to match the life experience, and on the other nodes we can define the observations in an arbitrary way.

What is the result of this?

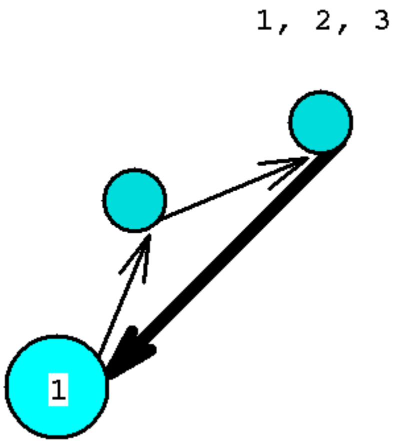
Every life has continuum many explanations, countably many computable explanations, and one computable explanation that is the simplest. All these models are accurate in both forward and backward directions.

This description of internal states says nothing specific about the world. Here all the complexity of the world is in the function g.

We want to make a model where the complexity of the world is concentrated in the description of its current state.

The first step in this description is the classification

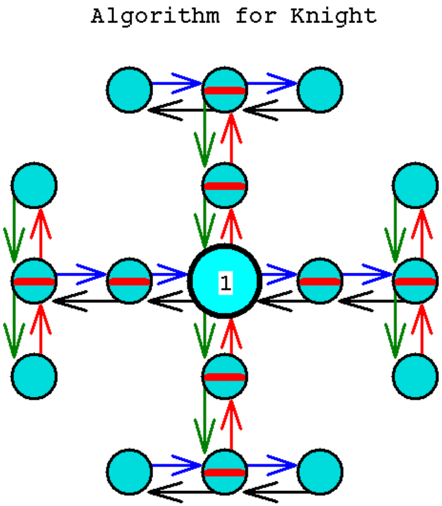
The result will be Event-Driven (ED) model. This is something like finite-state automaton in which letters are replaced by events.



A similar ED model is that of the days of the week.

Another Event-Driven model

This is the knight algorithm (the chess piece knight).



Here we have a trace here. That is, we have states in which something special happens.

What will be the interpretation of the ED model?

Let the model of the world be the only one. Let us have a function $\Sigma^* \to Q$. Let this function be coordinated with the events. We will call this function an interpretation of the ED model.

The interpretation defines a function $Q \to \mathcal{P}(S)$ which, for each state of the model q, gives us the set of states S_q the world can be in.

If $S_q \neq S$ then q gives us some information about the current state of the world.

The best case is when S_q are disjoint subsets. Then S_q are equivalence classes of some equivalence relation and Q is the quotient set of this relation.

We're not going to start with the set S, because we don't know it. Instead of that we're going to start with some quotient set. We will find this quotient set by chance.

What does description through ED models look like?

We will find *n* ED models. For each ED model we will find its current state (we will find it exactly or approximately).

The resulting description will be *n*-tuple composed of these states. For example:

$$\langle Monday, Sofia, hot, hungry \rangle$$

What are the next steps?

1. **Temporal patterns.** The ED models represent permanent patterns (which are valid all the time) but in our world we can observe a phenomenon which is a temporal pattern (valid from time to time).

2. **Algorithms.** These are also temporal patterns.

3. **Moving trace.** We suppose that in the ED model we have permanent trace. This is a state in which something special happens. If this special behavior can move to another state, then the trace is mobile.

4. **Objects and agents.** We will add this abstractions of higher level.