Binding Actions to Objects in World Models

Ondrej Biza[1], **Robert Platt**[1], **Jan-Willem van de Meent**[1,2], **Lawson L. S. Wong**[1] **and Thomas Kipf**[3]
[1] Northeastern University, Boston, MA, USA
[2] University of Amsterdam, Netherlands
[3] Google Research, Brain Team

Northeastern University
Khoury College of Computer Sciences

World Models with Object Slots

Our goal is to learn **world models,** models that learn to compactly represent the state of the world and to predict its forward dynamics.
We work with **structured world models**, which

- represent the state of the world as the state of individual objects in a scene,
- model the dynamics of the world using a message passing neural network – a type of **graph neural network** that uses multi-layer perceptrons to send messages between objects and update their states given an executed action – and
- are trained with a self-supervised contrastive loss.

Soft and Hard Action Attention

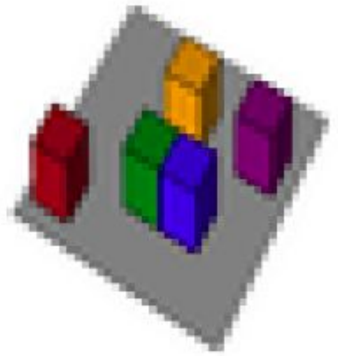
The main contribution of this paper are two attention mechanisms that learn to predict which object(s) are affected by a selected action.
In both cases, the states of K objects (z) are transformed into keys and the action is transformed into a query. The predicted attention weights are the inner product of keys and queries:

$$k = \langle j_k(z_1), j_k(z_2), ..., j_k(z_K) \rangle \quad q = j_q(a)$$
$$\alpha = \text{softmax}(k_1^T q, k_2^T q, ..., k_K^T q)$$

Hard Attention: the attention weights are used as parameters to a categorical distribution, which states that only one object slot receives information about the selected action. The model makes predictions for all possible assignments to calculate the gradient.
Soft Attention: all object slots receive the same (transformed) action multiplied by the attention weight for the corresponding slot.

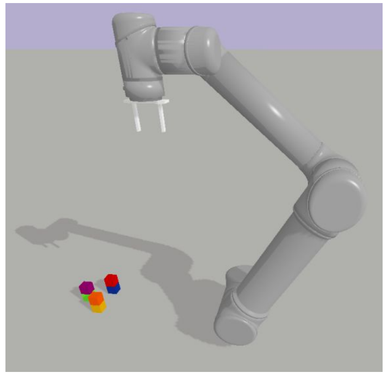
Toy Objects:

In a toy grid-world environment with five objects moving in the four cardinal direction, **hard attention correctly identifies the action-object mapping**. A baseline structured world model without attention fails to distinguish individual objects.



Robotic Manipulation:

In a realistic robotic manipulation environment, **soft attention correctly predicts which cubes are moved by the robot**. With soft attention, the robot's world model achieves high accuracy with only one graph neural network layer. A baseline requires multiple graph neural network layers to reach the same performance.



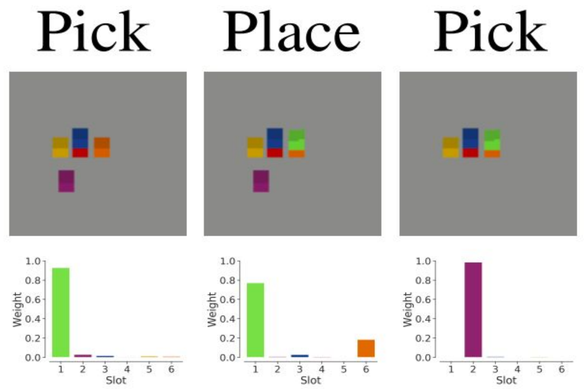


Figure: Attention weights for the robot's pick and place actions.

Atari games:

In Atari games Pong and Space Invaders, neither soft nor hard attention improves the world model. We found all slots in the structured world model to **capture nearly identical information**, defeating the point of an action-slot binding.

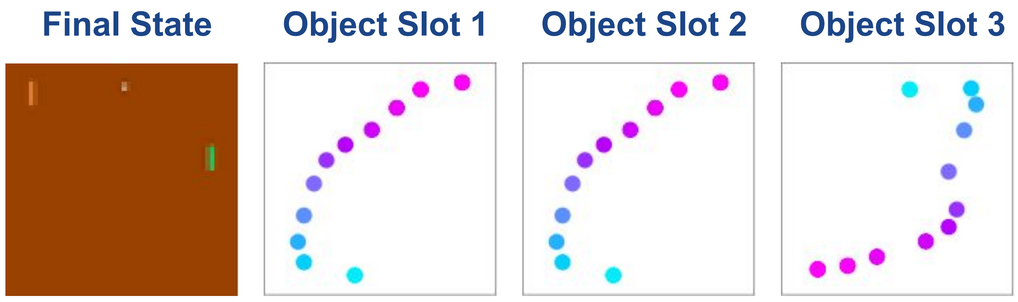


Figure: A sequence of 11 states embedded into three object slots.