Interactive Vision: Lessons from Attentional Reinforcement Learning

Introduction

Philosopher Patricia Smith Churchland posed the questions in 1994:

> What simplifications in the learning problem can be achieved by abandoning Pure Vision's rich-replica assumption? How much mileage can we get out of the reinforcement learning paradigm if we embrace the assumption that the perceptual representations are semi-world representations consisting of, let us say, goal-relevant properties? How might that work? (Churchland, Ramachandran and Sejnowski 45)

The conventional wisdom on visual perception has largely been founded on the theory of pure vision. The theory of pure vision entails the rich-replica assumption, that the visual system creates a perfect replica of the visual world independently of other sensory modalities, such as motor processing (Churchland, Ramachandran and Sejnowski 23). In contrast to pure vision, Churchland proposed a theory of interactive vision, which states that the motor system contributes significantly to what is being literally seen. Drawing upon experiments on saccadic eye movements, Churchland describes that visual attention enables and supports the visual semi-world hypothesis, that what we see is a partially elaborated representation of the visual world.

Churchland suggests that reinforcement learning, a learning paradigm by which desirable actions are reinforced by rewards based on the idea of operant conditioning, is limited by the rich-replica assumption due to the credit assignment problem, the difficulty of assigning the relevance of the many perceived visual features to the received reward upon taking an action. With the recent advancement of the attention mechanism in the field of deep learning, integrated into the reinforcement learning framework to solve visual tasks, what can attentional reinforcement learning now teach us about visual perception?

What is the rich-replica assumption and its implications?

The rich-replica assumption is implied by the theory of pure vision, which characterises what we literally see as "a fully elaborated representation of the visual scene" (Churchland, Ramachandran and Sejnowski 24). Pure vision also entails a dependency relation, that higher levels in the visual processing "hierarchy" depend on lower levels, but not vice versa.

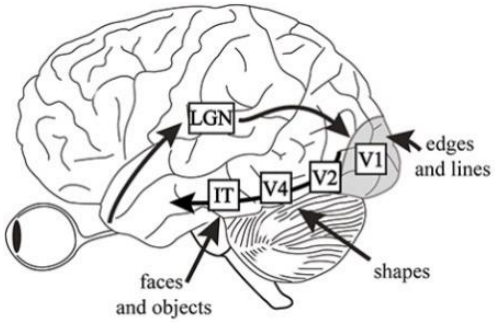

Figure 1: The classical view of hierarchical feed-forward processing *(Herzog and Clarke)*.

The dependency relation, described as the "classical view of hierarchical feed-forward processing" (Medathati, Neumann and Masson) in Figure 1, can be illustrated through the propagation of signals from the initial retinal stages, to the lateral geniculate nucleus (LGN), to the later visual cortical processing stages (e.g. V1, V2) (Churchland, Ramachandran and Sejnowski 24). The dependency relation implicates that vision operates independently of other sensory modalities, such as learning and motor execution. However, Churchland argues that what we see at any given moment is only the "visual semi-world", a partially elaborated representation of the visual scene, based on what is immediately

relevant to us, because visual perception evolved from the organism's need to excel in survival. By attending only to immediately relevant segments of the visual scene, the theory of interactive vision states that we construct only a partially elaborated representation of the visual world in our visual perception.

What is visual attention and how does it improve reinforcement learning?

Visual attention is the ability to focus on important parts of the visual world without distraction from irrelevant details (Tang and Ha). Visual attention enables people to condense broad visual information to be used in decision making. The attention mechanism is a recent breakthrough in the application of deep learning to many fields, including image processing.

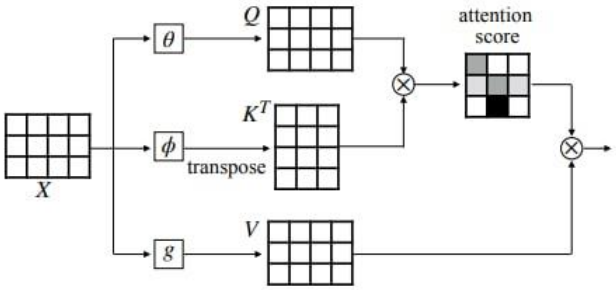

Figure 2: The attention mechanism *(Singh)*.

As illustrated on Figure 2, the attention mechanism produces an attention score that weights each segment of the visual input by its importance according to the agent's received reward, or the "utility in the predictive game" as described by Churchland, to construct a useful representation of the visual input. By weighting each visual input segment with an attention score, segments with a lower score have a smaller influence, to the extent of negligibility, over the action chosen by the agent. As a result, the attention mechanism contributes to a partially elaborated representation of the visual scene, simplifying the credit assignment problem as the agent can now determine which segment of the visual input is relevant to the utility received. In contrast, the construction of a perfect representation of the visual world within an agent, which is independent of the reward received, tends to be a disadvantage because the construction consumes additional time, space and energy (Churchland, Ramachandran and Sejnowski 47).

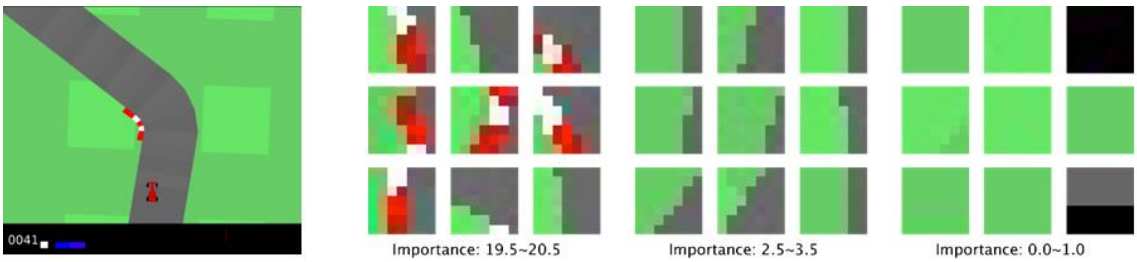

Figure 3: The car racing simulation (left) and attention score for various visual input segments (right) *(Tang, Nguyen and Ha)*

As shown in Figure 3, after attentional reinforcement learning agents were trained on a car racing simulation, visual input segments containing red-white markers that indicate a sharp turn are assigned a high attention score by the agent (Tang, Nguyen and Ha 8). When the background of the simulation was modified by the experimenters, the agent remains performant. By limiting sensory information through an attentional bottleneck, the agent is shown to learn to ignore information that is non-critical to the task. By encouraging the agent to attend to only a small fraction of its visual input, as Churchland reasoned, the agent learns to focus its attention to visual targets that benefits it in the survival game, as

an attended target of the visual scene is more probable to be causally related to the utility received (Churchland, Ramachandran and Sejnowski 44).

What does attentional reinforcement learning tell us about visual perception?

The attentional reinforcement learning agent achieves competitive results with a significantly less complex processing (Tang, Nguyen and Ha 9), showing that downsizing of the visual input using attention simplifies the credit assignment learning problem. The agent learned to attend to hints in the visual scene that are task-critical and is therefore able to generalise and remain performant in situations where task-irrelevant elements are changed. However, such an improved generalisation capability is attributed to the agent attending to the right thing rather than logical reasoning of the agent (Tang, Nguyen and Ha 9). In the simulation, upon adding a parallel lane next to the true lane, the agent happened to attend to the parallel lane and drove there instead, whereas a human driver would have reasoned that it is actually the opposite lane and not drive onto the parallel lane. The attentional agent remains unable to reason about questions of counterfactual nature without first having received prior information about the imagined reality, thus lacking the ability to perform causal inference. Without first experiencing driving on the parallel lane and receiving a utility penalty, the agent is unable to realise that driving on the parallel lane is a bad choice.

Therefore, although the attention mechanism has proven its effectiveness on visual reinforcement learning tasks, highlighting the appeal of the visual semi-world hypothesis, it also suggests that there is more to be realised about visual processing, such as possibly the top-down influence of logical reasoning or causal inference on vision. The integration of causal inference into the framework of reinforcement learning as such opens possibilities of counterfactual decision making of the visual agent, birthing the new field of causal reinforcement learning, holding promise to reveal much more about the nature of visual perception.

Conclusion

Since 1994, the theory of interactive vision has given rise to ideas that advanced the state of existing methods and our understanding of vision, such as in the integration of attention in reinforcement learning. However, attentional reinforcement learning reveals that there remains more to be realised about visual perception, as such agents have yet to learn to choose actions based on logical reasoning and causal inference. Nonetheless, the theory of interactive vision, much like how pure vision provided the groundwork for interactive vision as described by Churchland, has in turn provided the intellectual infrastructure required to further advance our understanding of visual perception, perhaps now, excitingly, in the direction of causal reinforcement learning.

References

Churchland, Patricia S., V. S. Ramachandran and Terrence J. Sejnowski. "A Critique of Pure Vision." Koch, Christoff and Joel L. Davis. *Large-Scale Neuronal Theories of the Brain*. The MIT Press, 1994. 23-48.

Herzog, Michael H. and Aaron M. Clarke. "Why vision is not both hierarchical and feedforward." *Frontiers in Computational Neuroscience* (2014): 2.

Medathati, N. V. Kartheek, et al. "Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision." *Computer Vision and Image Understanding* (2016): 3.

Singh, Praphul. *Oracle Artificial Intelligence and Data Science Blog - Multi-Head Self-Attention in NLP*. 13 May 2020. 21 11 2020.

Tang, Yujin and David Ha. *Google Artificial Intelligence Blog - Using Selective Attention in Reinforcement Learning Agents*. 18 June 2020. 21 11 2020.

Tang, Yujin, Duong Nguyen and David Ha. "Neuroevolution of Self-Interpretable Agents." *Genetic and Evolutionary Computation Conference* (2020).