COMP 596 - Essay Assignment:
“Common Sense” from Vision

Hong, Joon Hwan
260832806
Winter 2021

Introduction

While the definition of common sense is not universal, for the context of discussion, common sense in vision is defined to be the understanding of object perception and permanence that humans exhibit with working memory. When discussing animal and human behaviour in response to a visual system/test, assumptions are made of information beyond the given visual information. In other words, people are able to perceive objects as objects with properties, not as a collection of shapes and colours that is categorized. Additionally, people are able to exhibit an understanding of object permanence and spatial occupancy; if an object is placed inside a container like a box, they are able to perceive that the object still spatially exists, just not directly visible. Human use of vision is more than just the raw visual input to determine and understand its environment, it is able to have a higher-order cognitive understanding: to perceive and understand individual objects in a given space. In other words, visual input contributes to human cognition, and thus understanding of the environment.

Object Perception and Exploitation

The behaviour of perceiving objects as tools is a basic example of animals displaying their capability in perceiving visual inputs as objects with useful properties. Numerous bird species are able to discover and perceive an unfamiliar object and use it as a tool for a given task. An example of this is the Aesop’s fable paradigm test (commonly known as “the crow and the pitcher”), where the birds have to drop objects into a tube containing water and floating food to eventually reach the food with their beaks. To accomplish the task, behaviourally the birds have to raise the water level; they have to understand and perceive that the tube must contain a liquid object rather than a solid, and that the dropped objects would sink and not float (Jelbert, Taylor, Cheke, Clayton, & Gray, 2014). Thus, to succeed the task, the animal must recognise new objects, exploit the fact that they can be used as tools/picked up, and dropped to eventually displace a liquid.

AI: Current Capabilities, Issues, and Hypothetical Implementation

For the purpose of the discussion, reinforcement learning models are focused as they present an AI system which can learn via trial and error maximizing the expected reward over iterations, emulating animal behaviour and methods. Then for today's AI capabilities, two questions arise: can an AI emulate such behaviour of object perception and utilization; can an AI agent actually perceive objects from visual input? For the first question, there is a literature on a deep reinforcement learning agent that is capable of performing such tasks in virtual three-dimensional environments (Mirowski, et al., 2017). However, it can not be claimed that the agent understands and perceives objects. The agent accomplishes tasks by exploiting correlations at the pixel level in visual inputs (Mirowski, et al., 2017). The agent does not generate an internalized understanding of the environment; a general consequence of such learning methods is that when presented with an unfamiliar object it would most likely display characteristics of poor transfer learning.

Consequently, it can be said that modern RL agents lack effective out-of-distribution generalization (Finn, Yu, Fu, Abbeel, & Levine, 2017) if its visual system and comprehension is based on a pixel-based correlational method. Thus, to an extent, modern AI is capable of emulating the behaviour of object perception and utilization in a virtual environment, however the means which the agent accomplishes this is drastically different than the perceptual and cognitive understanding that humans and animals exhibit through their visual system. The second question – which asks if current AI can learn to perceive objects – perhaps can be implemented in the following means: by facilitating reinforcement learning in an increased/ higher level of abstraction beyond pixel-based understanding.

To give modern AI such capabilities at a computational level, the AI should attempt learn perceptual groupings (Herzog, 2018) of different pixel inputs and implement a means to reward "exploratory behaviour". If the system can be designed to attempt to learn, generalize, and group pixel-patterns into object paradigms of certain properties/definitions (such as physical state, length, etc.) and reward policies/actions which result in interacting with pixels corresponding to learned perceptual groupings, it would emulate animal probing behaviour of unfamiliar objects and result in the AI developing a means to categorize pixel groups: similar to perceiving an object.

Algorithmically, it would essentially function as a means of object recognition, except that all object classes would be derived internally from pixel-to-pixel information to form arbitrary perceptual groupings based on its experience without predefined labels beyond general feature definitions such as lines and edges. On the assumption that perceptual groupings can be learned by the AI, physical properties of objects within the groupings can be learned. Physical state of objects – liquid, solid, etc. – can potentially be derived from the temporal difference in individual pixels (and their next temporal positions) belonging to a label based on their change of position in relation to relative motion of the agent. The agent needs the capability of interacting with two or more objects of distinct perceptual groupings and be able to observe pixel interactions. Consequently, the AI model should consist of two subnetworks: one which receives internal information such as motion of the agent's (assuming human-like) hands, joint angles, and motion (causation); the other to receive external information of the visual environment to generate perceptual groupings (consequence). Successful information gain of perceptual groupings from the second subnetwork from policies taken by the first subnetwork – interacting with the world with its self-motions – would be rewarded. While if conflicting information is generated from the subnetworks, such that the properties of an object from the second subnetwork conflicts with the observed changes in pixels from the first subnetwork's actions, reward would decrease as a penalty.

Concluding Remarks

In terms of environmental needs to train such a model, a three-dimensional simulation with reasonably accurate physics would be necessary. The AI would be human-like agent that possesses proprioception. In the virtual environment, various objects which the agent can interact would be available with varying physical properties. One issue to consider is during the AI's natal stages, where all stimulus and input are novel. To combat such issue, a method to temporally reduce the reward value of a continued interaction with a supposed perceptual grouping by means of momentum – using a running average of reward – or other means of inhibiting reward from extended interaction should be implemented.

While current AI vision systems can mimic complex motions in virtual simulations, they face difficulty in emulating object perception; the means which current intelligent systems achieve such functions is by exploiting pixel-by-pixel correlations (Mirowski, et al., 2017), a method far different from how biological systems form cognitive understanding from visual input. An AI agent in a virtual three-dimensional environment could exhibit behaviour similar to perceiving and exploiting objects, emulating humans/tool-using animals by implementing the computational and algorithmic needs mentioned previously.

References

Finn, C., Yu, T., Fu, J., Abbeel, P., & Levine, S. (2017). Generalizing Skills with Semi-Supervised Reinforcement Learning. *International Conference on Learning Representation (ICLR)*.

Herzog, M. H. (2018). Perceptual Grouping. *Current Biology*, R679-R694.

Jelbert, S. A., Taylor, A. H., Cheke, L. G., Clayton, N. S., & Gray, R. D. (2014). Using the Aesop's Fable Paradigm to Investigate Causal Understanding of Water Displacement by New Caledonian Crows. *PLoS One*, e92895.

Mirowski, P., Pascanu, R., Viola, F., Soyer, H., Ballard, A. J., Banino, A., . . . Hadsell, R. (2017). Learning to Navigate in Complex Environments. *International Conference on Learning Representations (ICLR)*.