Towards Biologically Plausible Artificial Vision

Mason Westfall

Department of Philosophy/Philosophy-Neuroscience-Psychology Program, Washington University in St. Louis

w.mason@wustl.edu

WASHINGTON UNIVERSITY
CB 1073
ONE BROOKINGS DRIVE
ST. LOUIS, MO 63130-4899

314-935-6670

http://www.masonwestfall.com

Behavioral and Brain Sciences, forthcoming. Commentary on Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum, "The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences", *Behavioral and Brain Sciences*, 2022.

Abstract

Quilty-Dunn et al. argue that DCNNs optimized for image classification exemplify structural disanalogies to human vision. A different kind of artificial vision—found in reinforcement learning agents navigating artificial three-dimensional environments—can be expected to be more human-like. Recent work suggests that language-like representations substantially improves these agents' performance, lending some indirect support to the LoTH.

Abstract Word Count: 60
Main Text Word Count: 928
References Word Count: 315
Total Word Count: 1324

Image classifiers implemented with DCNNs have been taken by many to tell against LoT architectures. Quilty-Dunn et al. argue that this is a mistake. These image classifiers exhibit deep structural disanalogies to human vision, so, whether or not they implement LoT architectures tells us little about human vision. This is perhaps unsurprising, since biological vision is plausibly not optimized solely for image classification (Bowers et al., 2022, p. 10). Would training artificial vision under more ecologically realistic conditions produce a more realistic model of human vision? To make progress on this question, I describe some reinforcement learning (RL) agents trained to navigate artificial three-dimensional environments on the basis of how things appear from their perspective, and explain why we might expect their vision to be more human-like. Interestingly, language-like representations seem to be especially helpful to these agents. They explore more effectively, more quickly learn novel tasks, and are even facilitated in downstream image classification. These models arguably provide some indirect evidence for the LoTH about human vision, and may offer some clues as to why LoT architectures arose evolutionarily.

What *is* biological vision optimized for, and what would artificial vision that was similarly optimized be like? One answer to the first question is that biological vision is optimized for an agent's success in their environment. Success requires a number of competences that vision must contribute to simultaneously. Agents need to effectively explore, learn new behaviors, and act to achieve their goals, all while the environment changes in often surprising ways.

Recent work in RL arguably more closely approximates the optimization problem facing biological agents. Artificial RL agents can learn to do many complex tasks, across a variety of environments—most interestingly, in this context, exploring and pursuing goals in artificial three dimensional environments like Habitat (Savva et al., 2019), Matterport3D (Chang et al., 2017), Gibson Env (Xia et al., 2018), Franka Kitchen (Gupta et al., 2019), VizDoom (Kempka et al., 2016), Playroom (Tam et al., 2022) and City (Tam et al., 2022). One way of accomplishing this—especially in environments where environmental reward is sparse—is by making novelty intrinsically rewarding. These ‘curious agents’ can learn, without supervision, representations that enable them to perform navigation tasks, interact with objects, and also perform better than baseline in image recognition tasks (Du, Gan, and Isola, 2021). As the authors put it, their agents are ‘learning a task-agnostic representation for different downstream interactive tasks’ (Du, Gan, and Isola, 2021, p. 10409).

One challenge these researchers face is how to characterize novelty. Superficial differences in viewing angle or pixel distribution can easily be rated as highly novel, leading to low-level exploration that does not serve learning conducive to achieving goals. A recent innovation is to equip RL agents with ‘prior knowledge, in the form of abstractions derived from large vision-language models’ (Tam et al., 2022, p. 2). Doing so enables the state space over which novelty is defined to be characterized by abstract, semantic categories, such that novelty is defined in task-relevant ways (Mu et al., 2022). This method has been shown to substantially improve performance across a variety of tasks and environments, compared to non-linguistic ways of characterizing the state space (Schwartz et al., 2019; Tam et al., 2022; Mu et al., 2022). The improvements are especially pronounced for tasks involving relations between objects, e.g. ‘Put a OBJECT on a {bed, tray}’(Tam et al., 2022, p. 2), reminiscent of work on relations reviewed in the target article (Hafri and Firestone, 2021). As the authors note, their training on vision–language representations that encode ‘objects and relationship’ instead of on ImageNet—optimized for classification—should be expected to be more successful (Tam et al., 2022, p. 10).

Why would linguistic categories facilitate performance? One possibility is that language compress the state space in ways that facilitate successful actions. The semantic categories enshrined in natural language tend to abstract from action-irrelevant variation, and respect action-relevant variation. So, visual processing optimized relative to natural language categories is *de facto* optimized for action-relevant distinctions. The LoT architecture characteristic of object files and visual working memory seems well-suited to serving this function (though LoT plausibly is importantly different from natural languages (Green, 2020; Mandelbaum et al., 2022)). Predicating abstract properties of individual objects in a language of thought is poised to guide action, because abstract semantic categories often determine the action affordances available for some individual object, independent of nuisance variation associated with e.g. viewing angle (though viewing angle is plausibly relevant for more fine-grained control tasks (Parisi et al., 2022, p. 6)). Such abstract, task-agnostic representations are also able to transfer to new tasks or environments, in which familiar kinds take on novel relevance for action.

These recent innovations in RL arguably offer indirect support for the LoTH as applied to humans. Of course, similar performance can be achieved by distinct underlying competence, and we should not exaggerate how similar even artificial RL agents' performance actually is to humans at present. Nevertheless, language-like structures appear especially helpful for artificial agents when faced with rather more biologically plausible optimization problems than the one that faces image classifiers. Perhaps a language of thought served our ancestors similarly in an evolutionary context. Language-like structures enabled creatures to encode abstract properties in a task-agnostic way, that nevertheless facilitated downstream performance on a wide variety of tasks, as the environment changed. It's not hard to imagine why evolution might see to it that such a system stuck around.

References

Bowers, Jeffrey S et al. (2022). "Deep problems with neural network models of human vision". In: *Behavioral and Brain Sciences*, pp. 1–74.

Chang, Angel et al. (2017). "Matterport3d: Learning from rgb-d data in indoor environments". In: *arXiv preprint arXiv*:1709.06158.

Du, Yilun, Chuang Gan, and Phillip Isola (2021). "Curious representation learn- ing for embodied intelligence". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10408–10417.

Green, EJ (2020). "The perception-cognition border: A case for architectural division". In: *Philosophical Review* 129.3, pp. 323–393.

Gupta, Abhishek et al. (2019). "Relay policy learning: Solving long-horizon tasks via imitation and reinforcement learning". In: *arXiv preprint arXiv*:1910.11956.

Hafri, Alon and Chaz Firestone (2021). "The perception of relations". In: *Trends in Cognitive Sciences* 25.6, pp. 475–492.

Kempka, Michał et al. (2016). "Vizdoom: A doom-based ai research platform for visual reinforcement learning". In: *2016 IEEE conference on computational intelligence and games (CIG). IEEE,* pp. 1–8.

Mandelbaum, Eric et al. (2022). “Problems and mysteries of the many languages of thought”. In: *Cognitive Science* 46.12, e13225.

Mu, Jesse et al. (2022). “Improving intrinsic exploration with language abstractions”. In: *arXiv preprint arXiv*:2202.08938.

Parisi, Simone et al. (2022). “The unsurprising effectiveness of pre-trained vision models for control”. In: *International Conference on Machine Learning. PMLR,* pp. 17359–17371.

Savva, Manolis et al. (2019). “Habitat: A platform for embodied ai research”. In: *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9339– 9347.

Schwartz, Erez et al. (2019). “Language is power: Representing states using natural language in reinforcement learning”. In: *arXiv preprint arXiv*:1910.02789.

Tam, Allison C et al. (2022). “Semantic exploration from language abstractions and pretrained representations”. In: *arXiv preprint arXiv*:2204.05080.

Xia, Fei et al. (2018). “Gibson env: Real-world perception for embodied agents”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9068–9079.