Building on prior knowledge without building it in

A commentary on "Building machines that learn and think like people"

Steven S. Hansen[1], Andrew K. Lampinen[1], Gaurav Suri[2], and James L. McClelland[1]

1: *Stanford University*
2:San Francisco State University

sshansen@stanford.edu

lampinen@stanford.edu

rav.psych@gmail.com

mcclelland@stanford.edu

http://www.suriradlab.com/

https://web.stanford.edu/group/pdplab/

Abstract: Lake et al propose that people rely on 'startup software', 'causal models', and 'intuitive theories' built using compositional representations to learn new tasks more efficiently than some deep neural network models. We highlight the many drawbacks of a commitment to compositional representations and describe our continuing effort to explore how the ability to build on prior knowledge and to learn new tasks efficiently could arise through learning in deep neural networks.

This is a preprint version of the paper. The published version can be found in *Behavioral and Brain Sciences,* 40, E268 at https://doi.org/10.1017/S0140525X17000176

Lake et al. have laid out a perspective that builds on earlier work within the structured / explicit probabilistic cognitive modeling framework. They have identified several ways in which humans with existing domain knowledge can quickly acquire new domain knowledge and deploy their knowledge flexibly. Lake et al. also make the argument that the key to understanding these important human abilities is the use of 'startup software', 'causal models', and 'intuitive theories' that rely on a compositional knowledge representation of the kind advocated by, e.g. Fodor and Pylyshyn (1988).

We agree that humans can often acquire new domain knowledge quickly and can often generalize this knowledge to new examples and use it in flexible ways. However we believe that human knowledge acquisition and generalization can be understood without building in a commitment to domain-specific knowledge structures or compositional knowledge representation. We therefore expect that continuing our long-standing effort to understand how humans abilities can emerge without assuming special start up software will be most helpful in explicating the nature of human cognition.

The explicit compositional approach of Lake et al. is limited because it downplays the often complex interactions between the multitude of contextual variables in the task settings in which the representation is used. Avoiding a commitment to symbolic compositionality increases one's flexibility to respond to sometimes subtle influences of context and allows for the possibility of more robust learning across contexts. The recent startling improvements in computer vision (Krizhevsky, Sutskever, & Hinton, 2012), machine translation (Johnson et al., 2016), and question answering (Weston et al, 2015)

were possible, precisely because they avoided these limitations by foregoing symbolic compositionality altogether.

Although Lake et al. seek to take the computational level 'high ground' (Marr, 1982), their representational commitments also constrain the inferential procedures they rely on. Their modeling work relies on the use of combinatorially explosive search algorithms. This approach can be effective in a specific limited domain (such as Omniglot), precisely because the 'start-up software' can be hand-selected by the modeler to match the specific requirements of that specific domain. However, their approach avoids the hard question of where this startup-software came from. Appeals to evolution, whereas they may be plausible for some tasks, seem out of place in domains of recent human invention such a character-based writing systems. Also, because many naturalistic learning contexts are far more open ended, combinatorial search is not a practical algorithmic strategy. Here, the gradient-based methods of neural networks have proven far more effective (see citations above).

We believe learning research will be better off taking a domain general approach wherein the start-up software used when one encounters a task as an experienced adult human learner is the experience and prior knowledge acquired through a domain general learning process.

However, most current deep learning models don't build on prior experience. For example, the network in (Mnih et al. 2013) that learns Atari games was trained from

scratch on each new problem encountered. This is clearly not the same as human learning, which builds cumulatively on prior learning. Humans learn complex skills in a domain after previously learning simpler ones, gradually building structured knowledge as they learn. In games like Chess or Go, human learners can receive feedback not only on the outcome of an entire game – did the learner succeed or fail? – but also on individual steps in an action sequence. This sort of richer feedback can easily be incorporated into neural networks, and doing so can enhance learning (Gulcehre and Bengio, 2016).

An important direction is to explore how humans learn from a rich ensemble of multiple, partially related, tasks. The steps of a sequential task can be seen as mutually supporting sub-tasks, and a skill, such as playing chess, can be seen as a broad set of related tasks beyond selecting moves: predicting the opponent's moves, explaining positions, etc. One reason humans might be able to learn from fewer games than a neural network trained on playing chess as a single integrated task is that humans receive feedback on many of these tasks throughout learning, and this both allows more feedback from a single experience (e.g. both an emotional reward for capturing a piece and an explanation of the tactic from a teacher) and constrains the representations that can emerge (they must support all these related sub-tasks). Such constraints amount to extracting shared principles that allow for accelerated learning when encountering other tasks that use them. One example is training a recurrent network on translation tasks between multiple language pairs, which can lead to zero-shot (no training necessary) generalization, to translation between unseen language pairs (Johnson et al, 2016). Just as

neural networks can exhibit rule-like behavior without building in explicit rules, we believe that they may not require a compositional, explicitly symbolic form of reasoning to produce human-like behavior.

Indeed, recent work on meta-learning (or learning-to-learn) in deep learning models provides a base for making good on this claim (Santoro et al, 2016; Vinyals et al, 2016; Bartunov and Vetrov, 2016). The appearance of rapid learning (e.g. one-shot classification) is explained as slow, gradient-based, learning on a meta-problem (e.g. repeatedly solving one-shot classification problems drawn from a distribution). Although the meta-tasks used in these first attempts only roughly reflect the training environment that humans face (we probably don't face explicit one-shot classification problems that frequently), the same approach could be used with meta-tasks that are extremely common as a result of socio-cultural conventions, such as 'follow written instructions', 'incorporate comments from a teacher', or 'give a convincing explanation of your behavior'.

Fully addressing the challenges Lake et al. pose – rather than building in compositional knowledge structures that will ultimately prove limiting – is a long-term challenge for the science of learning. We expect meeting this challenge to take time, but that the time and effort will be well spent. We would be pleased if Lake et al. would join us in this effort. Their participation would help accelerate progress toward a fuller understanding of how advanced human cognitive abilities arise when humans are

immersed in the richly-structured learning environments that have arisen in human cultures and their educational systems.

References

Bartunov, S. & Vetrov, D. P. (2016). Fast adaptation in generative models with generative matching networks. *arXiv preprint arXiv:1612.02192*.

Fodor, J. A. & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition* 28(1-2):3-71.

Gülçehre, Ç. & Bengio, Y. (2016). Knowledge matters: Importance of prior information for optimization. *Journal of Machine Learning Research* 17(8):1-32.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z. & Hughes, M. (2016). Google's multilingual neural machine translation system: Enabling zero-shot translation. *arXiv preprint arXiv:1611.04558*.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information ProcessingSsystems* (pp. 1097-1105).

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. MIT Press

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D. & Riedmiller, M. (2013). Playing Atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602.*

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D. & Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

Vinyals, O., Blundell, C., Lillicrap, T. & Wierstra, D. (2016). Matching networks for one shot learning. In: *Advances in Neural Information Processing Systems* (pp. 3630-38).

Weston, J., Bordes, A., Chopra, S., Rush, A. M., van Merriënboer, B., Joulin, A. & Mikolov, T. (2015). Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.