Challenges and Approaches to an Information-Theoretic Framework for the Analysis of Embodied Cognitive Systems

Madhavun Candadai
Program in Cognitive Science
Indiana University
Bloomington, IN 47408
madhavun.cv@gmail.com

Eduardo J. Izquierdo
Program in Cognitive Science
Indiana University
Bloomington, IN 47408
edizquie@iu.edu

Abstract

Although information theory is well established in the study of cognitive systems, the majority of its uses rely on studying the system under open-loop conditions, where the system is presented with input by the experimenter. However, a crucial aspect of understanding cognition is understanding the flow of information in the dynamics of an organism's closed-loop interaction with its environment. We outline four key challenges that an information-theoretic framework for embodied cognitive systems faces. Such a framework must be able to quantify: (1) multivariate interactions; (2) how information in the system changes dynamically over time; (3) what specific aspects of the features of the task the information is about; and (4) information flow in a brain-body-environment system which is coupled in a closed loop. In this short review, we provide perspectives on these four challenges, explain their significance, provide examples of how they have been tackled in other work, and outline the open challenges.

1 Introduction

Although there is a vast and growing amount of data about the neuroanatomy, the neurophysiology, and the behavior of several model organisms, how cognition works is still poorly understood. Much research has been dedicated to understanding how information flows through neural systems. However, with the growing recognition of the central roles that both embodiment and situatedness play in generating and modulating neural activity, the challenge is even more difficult: to understand how behavior is grounded in the dynamics of an entire coupled brain-body-environment system. Such a challenge demands the development of an information-theoretic framework for the analysis of embodied cognitive systems.

An embodied cognitive system is a dynamic recurrent neural network that is coupled to a body, which in turn is coupled to an environment, such that together the coupled brain-body-environment dynamical system produces adaptive behavior (Figure 1A). An understanding of an embodied cognitive system entails the characterization of how the interactions between the individual components that make up the neural system, the body, and the environment give rise to adaptive behavior. Thus, the goal of an information-theoretic analysis is to characterize how information about the relevant features of the task flow through the complete brain-body-environment system in ways that are relevant for the generation of the adaptive behavior at hand (Figure 1C).

There are four main challenges that such an informational characterization of a brain-body-environment system must meet. First, it must be able to quantify multivariate interactions. Second, it

Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

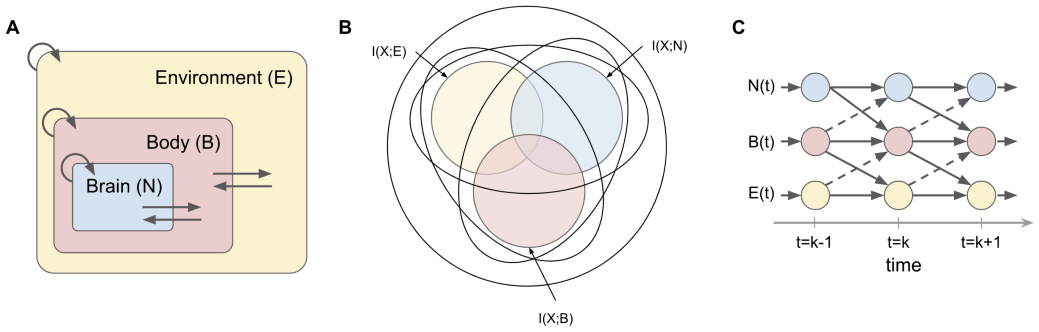

Figure 1: Information theory for brain-body-environment systems (BBE). [A] Illustration of a BBE, showing the brain interacting with the body and the body with the environment. Note information flows between each of these layers bidirectionally and thus sets up a closed-loop system. [B] Decomposition of multivariate information in the BBE system about task-relevant variable X. Each inner circle represents information each component (brain, body, environment) has about X. The overlapping regions represent redundant information across them. The non-overlapping region represents unique information and the region outside the inner circles represents synergistic information. [C] Temporal interaction in a BBE system demonstrating reason for information dynamics. Each component (brain, body, environment) is a dynamical system and when unrolled in time shows the continuous interaction between them. Each node in this graph can be modeled as a random variable to study information (decompostion) at every point in time, and hence its dynamics.

must be able to quantify how information in the system changes dynamically over time. Third, it must be able to quantify what specific aspects of the features of the task the information is about. Finally, it must be able to quantify information flow in a system that is in a closed loop of interactions with its environment.

Of these four challenges for an information-theoretic framework to analyze embodied cognitive systems, the first three have received more attention than the last. This has been in great part driven by the typical open-loop paradigm employed in experimental settings. As such, progress in this last challenge has been driven primarily through the use of synthetic data from simulation models. However, given experimental progress, increasingly model organisms are being studied in a closed-loop fashion. In this short review, we discuss each of these challenges and the approaches that have been taken to overcome them to date.

2 Multivariate information

An agent embedded in an environment that is in closed-loop interaction with the environment (perception informs actions which informs perception, and so on), can be modeled as a recurrent dynamical system of recurrent dynamical components i.e. the brain, the body, and the environment. To understand the role of one of those components, one common approach is to perform an ablation study of that unit. While ablation experiments provide conclusive evidence of that unit's participation/requirement in the behavior (see note in the appendix about information ablation), they do not have any explanatory power. To this end, Partial Information Decomposition (PID) [Williams and Beer, 2010a] is an information-theoretic framework that allows decomposing the total information across multiple sources (i.e., brain, body, or environment) about a target variable (i.e., task-relevant variable) into unique information that each source has about the target, redundant information that may exist across multiple sources, as well as synergistic information that is only available from referring to multiple sources at the same time and not from just one source alone. Consider two sources of information (two random variables) $X1$ and $X2$, about the random variable Y. If $X = \{X1, X2\}$, the total mutual information $I(X;Y)$ can be written as follows

$$I(X;Y) = U(X1;Y) + U(X2;Y) + R(X;Y) + S(X;Y) \tag{1}$$

where U denotes unique information, R denotes redundant information and S denotes synergistic information from these sources about Y. Naturally, with more than two sources, redundant and syner-

gistic information will be available for all combinations of the different sources. The decomposition can be better understood when visualized as a Venn diagram as shown in Figure 1B for three sources: brain, body, and environment. Also, note that each random variable used in these descriptions is multi-dimensional.

This approach could be applied simply within each component of the BBE system. For instance, within just the brain, all neurons feeding into another neuron can be analyzed by a set of sources feeding information to the target neuron about the variable of interest. This allows us to evaluate the unique, redundant, and synergistic information that the target neuron receives from the source neurons about the variable of interest. Furthermore, when these variables are offset in time, as shown in Figure 1C, partial information decomposition has been shown to be equivalent to transfer entropy [Schreiber, 2000, Williams and Beer, 2011]. In neuroscience, PID has been utilized to study task-relevant changes in fMRI [Lizier et al., 2011, Anzellotti and Coutanche, 2018], information distribution dissociated cultures [Sherrill et al., 2021] and importantly in neural models of BBE systems [Beer and Williams, 2015].

While PID scales well for high-dimensional systems, open challenges with analyzing multivariate systems involve scaling up the number of sources of information. If one were to decompose the BBE system into a three-source PID framework as shown in Figure 1B, each source can be any dimensional random variable. However, if one were to decompose it further into the individual sub-units within each of those sources, then PID would not scale well (imagine what a 7 source Venn diagram would look like).

3 Dynamic information

All living organisms performing cognitive or adaptive behavior are doing so over time, where timing is typically a crucial component of whether the behavior is adaptive or not. As such, the tools of analysis for understanding cognitive systems must take time and timing into consideration. Counter-intuitively, the majority of ways in which information theory is used to analyze cognitive systems discard time and timing, typically gathering data over time to make a distribution. However, the framework of information theory can be easily unrolled over time.

Unrolling information over time involves modeling variables of interest as random processes rather than random variables. In other words, a variable of interest at specific temporal landmarks of a task (or even better, along each time point of the task) is modeled as random variables and information-theoretic analyses are performed at those time points by estimating probability densities at these time points (each node the temporal graph shown in Figure 1C would be a random variable). This has been applied in a few different contexts: Williams and Beer [2010b] demonstrate how information measured over time allows us to identify when a decision about catch/avoid is made in an artificial agent optimized to perform a relational categorization task; Izquierdo et al. [2015] show the dynamics of information in a computation model of *C. elegans*, demonstrating how information about the chemosensory stimulus flows through its neural circuitry over time. Based on the graph in Figure 1C, information at time t in the brain about task-relevant variable X would include two other sources of information $N(t-1)$ and $B(t-1)$. Using the PID framework discussed in the previous section, the decomposition of multivariate information of these three sources allows us to study the origin and flow of information across the BBE system components in time.

The main challenge here is that the source of information in dynamical systems can quickly get obscured in time [Amblard and Michel, 2012]. In a system where the agent and environment are continuously interacting, a signal that arises in the environment (or agent) quickly gets reflected in the agent's (or environment's) dynamics thus making it impossible to disentangle the source. Candadai and Izquierdo [2020] show this phenomenon in a simulated setting where the environment drives a dynamical neural network model or vice versa (i.e. even when the behavior does not involve mutual interaction between agent and environment). Consequently, information quantities measured across time become estimates of the generalized correlation between those two random variables and do not reveal the direction of information flow i.e. causation. Thus, it becomes crucial to study temporal fluctuation in information across components of the BBE system to track the source of information. In other words, unrolling information in time allows us to identify *when* specific milestones occur over the course of a task and *where* they are initiated.

There are also challenges in the temporal analysis of information quantities on the experimental side; accurate measurements of variables of interest at specific landmarks during tasks over multiple trials (to build a distribution) is challenging. However, with increasingly advanced technology becoming available, such experiments are starting to become viable thus further increasing the scope for application of information theory [Yawo et al., 2013, Wang et al., 2011].

4 Specific information

Mutual information measures how much information X has about Y. In most cognitive systems, breaking down the analysis in this way is simply not sufficient to understand the operation of the nervous system. Typically, we would like to also know what information specifically X has about Y. For example, let's imagine that we are studying a decision-making task involving temperature in the environment and we observe that two different neurons have the same amount of mutual information about temperature. Ideally, we would like to further probe the system to learn whether the information that both neurons have about temperature is the same or overlaps, or whether each neuron has information about a different range of temperatures that are relevant to the task. Although specific information is not often used to study cognitive systems, information theory does allow us to unroll information about the specific features of the task-relevant variable.

The main challenge that opens up when we consider specific information in embodied and dynamic cognitive systems is that it can reveal complex, distributed, dynamic ways of ways in which the nervous system can contain information about the task at hand. For instance, population coding could be regarded as distributed where some neurons store information about a certain slice of the overall distribution of the variable of interest whereas other neurons could encode information about other slices. Measuring information about *specific* values of the variable of interest would enable acquiring a more granular explanation of the operation of a BBE system (X in Figure 1B, could refer to only a specific value of that task relevant variable). This involves, as one might expect, building a distribution of the BBE variables in response to several trials of a particular value (or a set of values) of the task-relevant variable, and then simply measuring the same information quantities such as mutual information. Ma et al. [2017] and Steinmetz et al. [2019] are examples of studies where spatial distribution of neural encoding was studied but all data required to perform a specific information study was available.

5 Embodied information

An agent embedded in an environment that is in closed-loop interaction with the environment (perception informs actions which informs perception, and so on), can be modeled as a recurrent dynamical system of recurrent dynamical components i.e. the brain, the body and the environment. To understand the role of one of those components, namely the brain, in such a setting requires answering questions along the lines of: Where in the system does information about X originate? When does it originate? Where does it exist at a given point in time? In order to answer these questions using the tools of information theory, unrolling information over time, over multiple variables, and over specific aspects of task-relevant variables, we must gather data from the cognitive system by opening up the loop of interaction, providing the system of interest with certain stimuli at certain times and tracking the flow of information over time. If the task-relevant variable of interest is independent of the embodied cognitive system (for example, an agent that has to decide to catch an object based on its size, where the task-relevant variable is the size of the object), then studying the flow of information in the system while breaking open the loop can be an appropriate approximation to the operation of the circuit in closed-loop interaction. However, when the task-relevant variable depends on the behavior of the system (for example, an agent moving in a temperature gradient, where the temperature perceived by the agent is the task-relevant variable), then studying the information flow of the system while it is decoupled from its environment is no longer as informative of the operation of the original embodied system. Unfortunately, there is little to no work in using information theory in the latter condition. As such, these remains one of the most unexplored challenges in the information-theoretic analysis of embodied cognitive systems.

6 Conclusion

In summary, this short review outlines four major challenges that an informational characterization of an embodied cognitive system must address. Information theory with its Partial Information Decomposition extension provides the framework to address these challenges. Of the four challenges discussed, the first three have received more attention but are still underused in the context of understanding embodied cognitive systems. The fourth challenge has been less explored and remains most unaddressed. In addition, there is plenty of ongoing research that helps expand the field of information theory through the use of better estimators, better ways of studying polyadic interactions, and tools that allow for easier ways to study these complex systems, as well as better ways of scaling the calculations computationally (we provide references to some software tools in the appendix).

References

Paul L Williams and Randall D Beer. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*, 2010a.

Thomas Schreiber. Measuring information transfer. *Physical review letters*, 85(2):461, 2000.

Paul L Williams and Randall D Beer. Generalized measures of information transfer. *arXiv preprint arXiv:1102.1507*, 2011.

Joseph T Lizier, Jakob Heinzle, Annette Horstmann, John-Dylan Haynes, and Mikhail Prokopenko. Multivariate information-theoretic measures reveal directed information structure and task relevant changes in fmri connectivity. *Journal of computational neuroscience*, 30(1):85–107, 2011.

Stefano Anzellotti and Marc N Coutanche. Beyond functional connectivity: investigating networks of multivariate representations. *Trends in cognitive sciences*, 22(3):258–269, 2018.

Samantha P Sherrill, Nicholas M Timme, John M Beggs, and Ehren L Newman. Partial information decomposition reveals that synergistic neural integration is greater downstream of recurrent information flow in organotypic cortical cultures. *PLoS computational biology*, 17(7):e1009196, 2021.

Randall D Beer and Paul L Williams. Information processing and dynamics in minimally cognitive agents. *Cognitive science*, 39(1):1–38, 2015.

Paul L Williams and Randall D Beer. Information dynamics of evolved agents. In *International Conference on Simulation of Adaptive Behavior*, pages 38–49. Springer, 2010b.

Eduardo J Izquierdo, Paul L Williams, and Randall D Beer. Information flow through a model of the c. elegans klinotaxis circuit. *PloS one*, 10(10):e0140397, 2015.

Pierre-Olivier Amblard and Olivier JJ Michel. The relation between granger causality and directed information theory: A review. *Entropy*, 15(1):113–143, 2012.

Madhavun Candadai and Eduardo J Izquierdo. Sources of predictive information in dynamical neural networks. *Scientific reports*, 10(1):1–12, 2020.

Hiromu Yawo, Toshifumi Asano, Seiichiro Sakai, and Toru Ishizuka. Optogenetic manipulation of neural and non-neural functions. *Development, growth & differentiation*, 55(4):474–490, 2013.

Kaiyu Wang, Yafeng Liu, Yiding Li, Yanmeng Guo, Peipei Song, Xiaohui Zhang, Shaoqun Zeng, and Zuoren Wang. Precise spatiotemporal control of optogenetic activation using an acousto-optic device. *PLoS One*, 6(12):e28468, 2011.

Chaolin Ma, Xuan Ma, Jing Fan, and Jiping He. Neurons in primary motor cortex encode hand orientation in a reach-to-grasp task. *Neuroscience Bulletin*, 33(4):383–395, 2017.

Nicholas A Steinmetz, Peter Zatka-Haas, Matteo Carandini, and Kenneth D Harris. Distributed coding of choice, action and engagement across the mouse brain. *Nature*, 576(7786):266–273, 2019.

Madhavun Candadai and Eduardo J Izquierdo. infotheory: A c++/python package for multivariate information theoretic analysis. *arXiv preprint arXiv:1907.02339*, 2019.

Ryan G James, Christopher J Ellison, and James P Crutchfield. dit: a python package for discrete information theory. *Journal of Open Source Software*, 3(25):738, 2018. doi: 10.21105/joss.00738.

Patricia Wollstadt, Joseph T. Lizier, Raul Vicente, Conor Finn, Mario Martınez-Zarzuela, Pedro Mediano, Leonardo Novelli, and Michael Wibral. Idtxl: The information dynamics toolkit xl: a python package for the efficient analysis of multivariate information dynamics in networks. *Journal of Open Source Software*, 4(34):1081, 2019. doi: 10.21105/joss.01081.

Joseph T Lizier. Jidt: An information-theoretic toolkit for studying the dynamics of complex systems. *Frontiers in Robotics and AI*, 1:11, 2014. doi: 10.3389/frobt.2014.00011.

Michael Lindner, Raul Vicente, Viola Priesemann, and Michael Wibral. Trentool: A matlab open source toolbox to analyse information flow in time series data with transfer entropy. *BMC neuroscience*, 12(1):1–22, 2011.

A Appendix

A.1 Information ablation

Information in natural systems is often encoded in the distribution of a signal rather that just its presence. To this end, to better mimic absence of information from a particular source, rather then completely remove any signal coming from that source we could manipulate the source signal distribution. A natural way to achieve this between two neurons would be to replace the signal from source neuron to target neuron to be equal to the source neuron's activity's expected value rather than 0. This helps remove confounds that physical lesions can bring, such as in situations where a source neuron while providing no information could still act as a constant clamp-like input to a target neuron to drive the behavior - physical lesions alone would not be able to detect such behavior.

A.2 Software packages for information theoretic analyses

1. infotheory - C++/Python; PID measures as well as transfer entropy estimation for upto 3 sources [Candadai and Izquierdo, 2019]
2. dit - Python; PID for discrete variables [James et al., 2018]
3. IDTxL - Python; PID and transfer entropy for upto 2 sources [Wollstadt et al., 2019]
4. JDIT - JAVA; primarily designed for measuring transfer entropy but also includes other measures such as entropies and mutual information [Lizier, 2014]
5. TRENTOOL - MATLAB toolbox which primarily caters to analog neural data such as MEG [Lindner et al., 2011]