Tracking data flow in digital brains exposes coincidental encryption

Douglas Kirkpatrick\textsuperscript{1,3}, Victoria Cao\textsuperscript{1,3} and Clifford Bohm\textsuperscript{2,3}

\textsuperscript{1}Michigan State University, Department of Computer Science and Engineering, East Lansing, 48824, United States of America
\textsuperscript{2}Michigan State University, Department of Integrative Biology, East Lansing, 48824, United States of America
\textsuperscript{3}Michigan State University, BEACON Center for the Study of Evolution in Action, East Lansing, 48824, United States of America

cliff@msu.edu

Abstract

Brains are among the most complex evolved objects. In recent years we have seen an explosion in the development of artificial cognitive systems constructed \textit{in silico} (i.e. digital brains). In fact, we are now capable of creating digital brains whose operation is so complex that they are effectively black boxes (Castelvecchi, 2016; Gunning, 2017). Previous work (Marstaller et al., 2013; Hintze et al., 2018; Kirkpatrick and Hintze, 2019) has identified and expanded upon various information-theoretic measures that can shed light on the internal processes of digital brains. Here we introduce a new information-theoretic measure called Fragmentation ($F$) which can measure how fragmented information is in an a digital brain. To provide a example of the application of $F$ we look at the evolutionary emergence of complexity. Questions regarding the evolution of complexity have been of interest for as long as evolution has been a theory (Gregory, 1935). Nature is responsible for the development of a massive array of complex organisms, each comprised of various organs and regulatory systems that are themselves complex (McShea and Brandon, 2010). It has been observed that complexity can evolve even when complexity itself is being selected against (Beslon et al., 2021). We conclude by using $F$ to show a case of evolved complexity that results in coincidental encryption.

Fragmentation

We define fragmentation in the context of a salient feature (e.g. a state of an environment) and a system (e.g. a brain) as the size of the smallest partition of the system that can predict the feature to some minimum threshold of accuracy. Fragmentation Matrices can be generated by calculating the shared entropy of a set of features and the power set of all partitions of a system (see Fig.1).

If we use future brain states as salient features and current brain states as the system, then a fragmentation matrix can be generated that explores the causal and temporal relationships between a brains inputs, memory, and outputs. From these relationships we can generate information flow diagrams that allow us to see how information moves through a brain (see Fig.2). If we also know the connectome of the brain, this can be used to prune “impossible” connections from an information flow diagram.

Results

To provide an intuition for Fragmentation we evolve Markov Brains (Hintze et al., 2017) using the MABE digital evolution framework (Bohm et al., 2017) on a simple memory task, NBack, in which agents receive binary input values one at a time and must output values corresponding to some prior inputs. In the visualizations below we show the structure of memory (Fig.1) and information flow (Fig. 2) of the same four brains (labeled [a], [b], [c], or [d]) each of which achieves perfect performance. [a] shows a brain that implements a simple data passing process using the minimal amount of information flow and no information integration, but the other three brains ([b], [c], and [d]) demonstrate instances of different degrees of increasingly complex information flow. As configured here, the NBack task requires the the full utilization of eight memory values to maximize performance. Additionally, the brains were only provided eight memory values, and thus any perfect strategy must fully utilize all memory values. As a consequence, any resulting complexity must be a lossless compression, or encryption, of the inputs, typically executed using a bitwise XOR. However, in the case of [d] more complex encryption is clearly occurring.

Discussion

$F$ and fragmentation matrices provide new ways to view the operation of cognitive structures. Here, we share an application of $F$ that illustrates how evolution can arrive at complex solutions. Even when evolving a simple task, such as NBack, which requires no information integration and only short term memory, we observe the emergence of encryption and complex information flow. We feel that this observation is important because if complexity is common in small networks, larger networks will inevitably result in greater encryption and complexities. We hope to use these understandings, as well as other insights that $F$ may provide, to more thoroughly explore how digital brains (and perhaps, in time, biological brains) evolve to process information.

This document is a summary of a larger publication currently in development.
Figure 1: Fragmentation Matrices for the NBack Task. Four Markov brains evolved on the NBack Task that evolved perfect performance ([a],[b],[c], and [d]), chosen to represent an variety of complexities. The features (shown on the y axis) are the expected outputs on the next update and the partitions (shown on the x axis) are combinations of the brain’s 8 memory values. The amount of correlation between each feature and each partition is indicated by the value and color. Bright yellow squares indicate high correlation, orange, red and black squares represent successively less correlation. The red arrows identify the smallest partitions that predict each feature. In [a], for example, output o1 is predicted by memory value 1. A portion of each plot, containing partitions of intermediate size, is not shown to save space, indicated by the ellipses. [a] and [b] show brains where each feature is predicted by a single memory value. [c] and [d] show brains that require larger fragments to predict some of the features. Note the darker red colored cells in [d] indicating partial information.

Figure 2: Information Flow Diagrams for the NBack Task. These plots show how data flows through the same four brains shown in Fig. 1 and are labeled identically. Green, White, and Blue nodes indicate inputs (i), memory (h), and outputs (o) respectively. The numbers in the nodes indicate the information in bits that each node receives (when preceded by ‘.’) and/or delivers (when followed by ‘.’). The labels accompanying each connecting link and the link’s width both indicate the proportion of the entropy in the downstream node that can be accounted for by that link. The links going into each node represent the connections necessary to account for the total entropy in that node. In this configuration NBack agents were required to report on the outputs corresponding to t-1, t-3, t-5, t-7 and t-8 where t is the current time. Note that xor behavior is indicated in [b] by the fact that the two inputs to h2 provide no information independently, but together account from 1 bit of entropy in h2.
References

Beslon, G., Liard, V., Parsons, D. P., and Rouzaud-Cornabas, J. (2021). Of evolution, systems and complexity.

Bohm, C., CG, N., and Hintze, A. (2017). MABE (modular agent based evolver): A framework for digital evolution research. Proceedings of the European Conference of Artificial Life.

Castelvecchi, D. (2016). Can we open the black box of ai? Nature News, 538(7623):20.

Gregory, W. (1935). Reduplication in evolution. Quarterly Review of Biology - QUART REV BIOL, 10.

Gunning, D. (2017). Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2(2).

Hintze, A., Edlund, J. A., Olson, R. S., Knoester, D. B., Schossau, J., Albantakis, L., Tehrani-Saleh, A., Kvam, P., Sheneman, L., Goldsby, H., et al. (2017). Markov brains: A technical introduction. arXiv preprint arXiv:1709.05601.

Hintze, A., Kirkpatrick, D., and Adami, C. (2018). The structure of evolved representations across different substrates for artificial intelligence. In Artificial Life Conference Proceedings, pages 388–395. MIT Press.

Kirkpatrick, D. and Hintze, A. (2019). The role of ambient noise in the evolution of robust mental representations in cognitive systems. In The Proceedings of the Artificial Life Conference 2019. Cambridge, MA: MIT Press.

Marstaller, L., Hintze, A., and Adami, C. (2013). The evolution of representation in simple cognitive networks. Neural computation, 25(8):2079–2107.

McShea, D. W. and Brandon, R. N. (2010). Biology’s first law: the tendency for diversity and complexity to increase in evolutionary systems. University of Chicago Press.