A Graph Neural Network Framework to Model Human Reasoning

Quan Do (qdo@bu.edu)
Graduate Program for Neuroscience and Center for Systems Neuroscience
Boston University, Boston, MA

Caroline Ahn (ahncj@bu.edu)
Graduate Program for Neuroscience, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Leah Bakst (lbakst@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Michael Pascale (mpascale@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center
Boston University, Boston, MA

Joseph T. McGuire (jtmcg@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center,
Graduate Program for Neuroscience
Boston University, Boston, MA

Michael E. Hasselmo (hasselmo@gmail.com)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center,
Graduate Program for Neuroscience
Boston University, Boston, MA

Chantal E. Stern (chantal@bu.edu)
Department of Psychological and Brain Sciences, Center for Systems Neuroscience, Cognitive Neuroimaging Center,
Graduate Program for Neuroscience
Boston University, Boston, MA

Abstract:

When confronting a new challenge in an unfamiliar and puzzling situation, humans can rapidly formulate a hypothesis based on limited interactions and come up with a solution tailored to the specific problem. However, coming up with a quick solution does not guarantee a preferable outcome. It is therefore beneficial to study the representations and neural circuits underlying human reasoning, not only to inspire the development of machines that reason flexibly and quickly like humans, but also to understand how human reasoning may be impaired in certain circumstances or by certain disorders. Here we introduced a framework to discover the representations that could lead to reasoning success and failure in humans, and explored how these representations can be realized with neural circuits. We tested and validated the framework on a human dataset (n=220) collected in our lab on a modified version of the Abstraction and Reasoning Corpus. We found that our Message Passing Graph Neural Network, when taking in graphs encoding different relationships and different levels of abstraction, can reproduce human solutions. We then mapped the space of graph representations that lead to error modes in humans to observe the link between topology and functions in reasoning.

Introduction

How can humans reason so quickly? Several lines of work suggest that humans could rely on inductive bias, or a set of assumptions other than the data, to restrict the learning space and guide the chosen hypothesis in novel circumstances (Baxter, 2000; Spelke & Kinzler, 2007; Tenenbaum et al., 2011). Inductive biases are best described and characterized as priors in probabilistic models of cognition (Griffiths et al., 2010). Priors can be represented explicitly with logical formulas and programs (Ellis et al., 2023; Lake et al., 2015), and implicitly represented in the weights and architectures of connectionist networks (Goyal & Bengio, 2022). While there are heated debates on which approach is preferable (Do & Hasselmo, 2021), graph theory might be a potential bridge between explicit representations and connectionist networks, bringing interpretability of function to connectivity and topology. Graphs were used in a probabilistic model to encode structural priors and study human learning (Kemp & Tenenbaum, 2008, 2009). On the connectionist side, Graph Neural Networks (GNNs), useful for learning representations of graphs, are gaining popularity and have achieved strong performance in algorithmic reasoning tasks such as sorting, searching, dynamic programming, pathfinding, and geometry (Cappart et al., 2023; Dudzik & Veličković, 2022; Xu et al., 2020). Furthermore, besides having relational bias (Battaglia et al., 2018), Graph Neural Networks have many variants implementing other inductive biases in their architectures (Zhang et al., 2022), adding to the expressivity of this tool. Here we introduced a novel framework with GNNs to study the representations and circuits underlying inductive bias in humans.

Methods

To demonstrate the applicability of our framework, we will leverage a novel behavioral paradigm called CogARC (Cognitive Abstraction and Reasoning Corpus), recently developed in our lab and inspired by an AI competition called ARC (Abstraction and Reasoning Corpus) (Chollet, 2019). CogARC requires participants to infer a hidden rule from 2 to 6 puzzle-solution pairs and apply the rule to a novel puzzle by drawing the solutions on an interactive interface (**Fig 1**). CogARC reasoning problems, 75 total, are varied in the types and numbers of rules that dictate the puzzles and corresponding solutions. Learned rules do not carry over across problems. Therefore, CogARC is less forgiving to random guessing or brute force methods, compared to traditional measures of human reasoning. We have a dataset containing mouse clicking patterns of 220 participants (52.27% male).

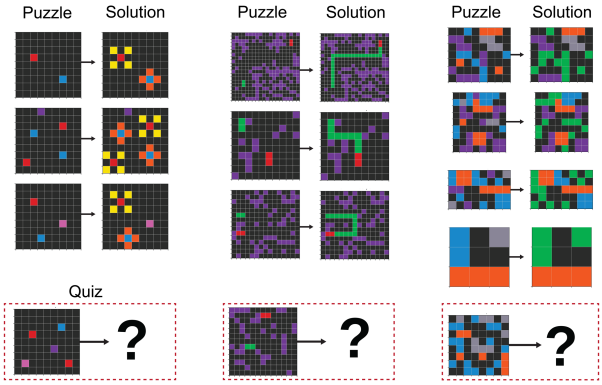

Figure 1: Sample tasks in the CogARC taskset where humans must draw their solutions after limited exposure to example puzzle-solution pairs. The tasks varied in reasoning demands.

We applied our graph framework to this dataset to reproduce human generated solutions and study how inductive bias in the representation and circuit architecture can harm or help reasoning. In our framework, a single CogARC puzzle can be converted to multiple graphs at different levels of specificity to encode different priors for problem solving. At the lowest level of abstraction, a tile in the puzzle grid can be a node. Going up the ladder of abstraction, a group of tiles, representing an object or a shape, can also be a node. The number of tiles can also be encoded in the size of the nodes. The edges in a graph can encode the relationships between adjacent tiles, nearby objects, or groups of objects. Multiple graphs can be created to test different hypotheses about the representations and relations sufficient and necessary to induce a rule (**Fig 3b**). These graphs will be forwarded to a GNN with message passing and fully connected layers (**Fig 2a**) and trained with gradient descent to output the corresponding solutions. For the message passing step, we implemented the Equivariant Graph Convolutional Layer (Satorras et al., 2021) to intentionally bias the network towards achieving translation, rotation, reflection, and permutation equivariance. After the neural network is trained on the limited number of training examples available (from 2 to 6 puzzle-solution pairs per task), the network is given a novel puzzle, and its output solution is compared to that of a human solver.

Results

EGNN exhibits few-shot learning and solves symmetry tasks when given adjacent tile graphs.

We found that when given graphs encoding adjacent tiles (**Fig 2a**), our equivariant Graph Neural Network (EGNN) can solve 9/75 tasks in the CogARC taskset. The model learned with only 2 to 6 puzzle-solution pairs (**Fig 2b,c**) and can generalize when given a novel puzzle (**Fig 2d**). The 9 tasks (only 3 are shown, **Fig 2d**) solved by EGNN all have local neighbor rules. We think that in addition to the implicit relational bias in the network architecture, the adjacent tile graphs gave the network sufficient priors to learn quickly and generalize successfully by highlighting the right level of abstraction (tile) and relation (adjacency).

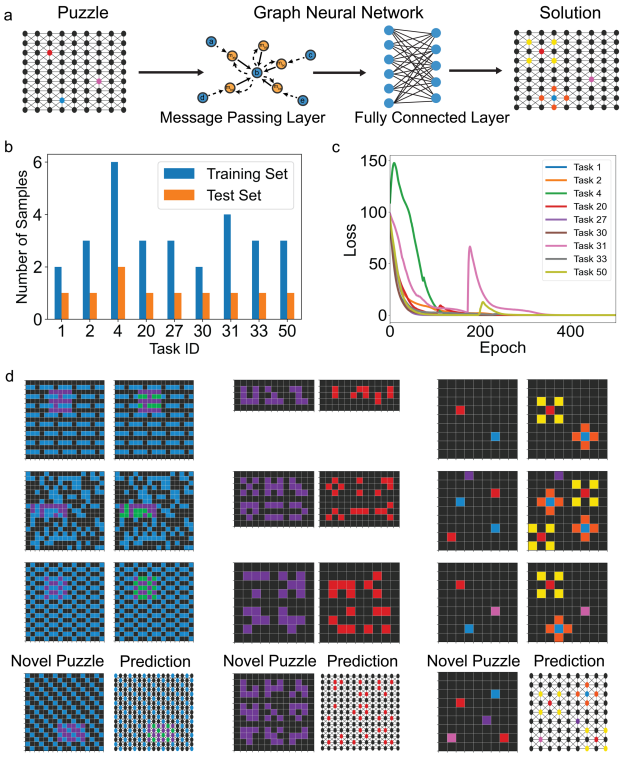

Figure 2: a) EGNN are given graphs encoding adjacent tiles. b) EGNN are trained with only 2 to 6 examples and tested on up to 2 novel puzzles. c) Learning curve showing loss over epoch for the 9 tasks EGNN can solve. d) 3 example tasks solved.

Changing graph priors allows EGNN to solve a complex task and make errors like humans.

To test our theory that the graph priors are essential for rule induction, we kept the same circuit architecture, but changed the graph representations and implemented our model in a more complex task (**Fig 3a**). We found that GNNs can be trained to solve and fail at tasks like humans, when taking in different graph representations encoding different priors. In our example task (**Fig 3a**), solving the puzzle requires counting the number of connected tiles with the same color, and changing the color of the connected tiles to green if the count is less than three. The most common error by humans (n=16/155) involves recoloring all the blue and gray tiles green (first column, bottom left of **Fig 3c**). In this case, we can create an input graph where nodes of the same colors, either gray or blue, are connected (first column, top left of **Fig 3c**). Nodes in this graph represent the connected tiles of the same color. All the nodes are of equal size. Given this prior, the trained GNN will mistakenly recolor blue and gray to green, irrespective of number of connected tiles like humans (first column, bottom left of **Fig 3c**). Another common error by humans (n=16/155) is turning all blue connected tiles to green (but not the gray tiles), so again we can create a graph where a few of the blue nodes are connected and all the nodes are equal in size. This prior will push the neural network to the error solution that humans made (**Fig 3c**, second column). The correct solution requires counting the number of connected tiles, so we can drop all the edges from the graph and add the count information by changing the node's size correspondingly. This allows GNN to solve the task (**Fig 3c**, third column). Thus far, we have only considered the graph representations of the novel puzzle, but the graphs that GNN saw during training also determine how it made errors. We explored this space next.

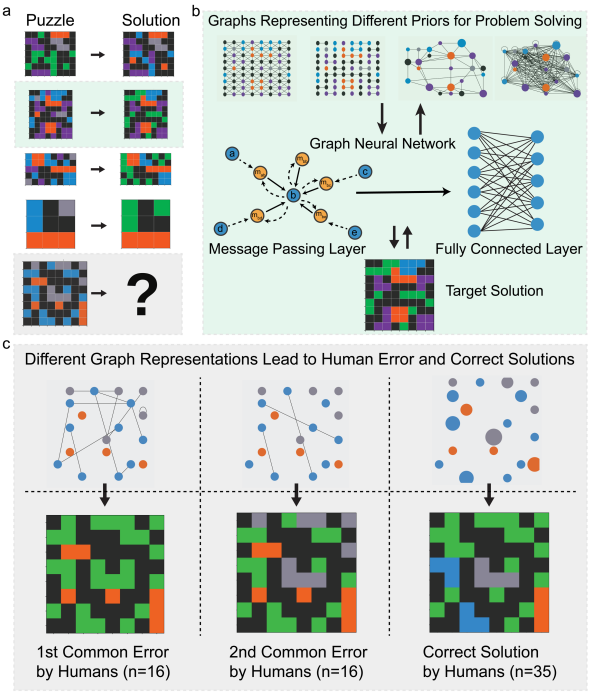

Figure 3: a) Puzzle-solution pairs for a task with complex rule. b) The GNN can be trained with different graph priors. c) The trained network given different test graphs will produce error and correct solutions like humans. Common errors and the correct solution by humans can be visualized as graphs.

Exploring the landscape of priors that could lead to human errors.

There are many graph priors that can lead to the same human-generated errors. Here we mapped the landscape of graph priors with EGNN to explore the relationship between graph topology and functions. Specifically, from the CogARC puzzle-solution pairs (training set), we generated training graphs in which nodes are connected tiles of the same color, and the connectivity of the graph is controlled by a parameter that varies from 0 (not connected) to 0.1 (densely connected). We did the same for the novel puzzle (test set). We trained EGNN on the training graphs, had the network output solutions given different novel puzzle graphs, and observed the likelihood with which different combinations would lead to human solutions.

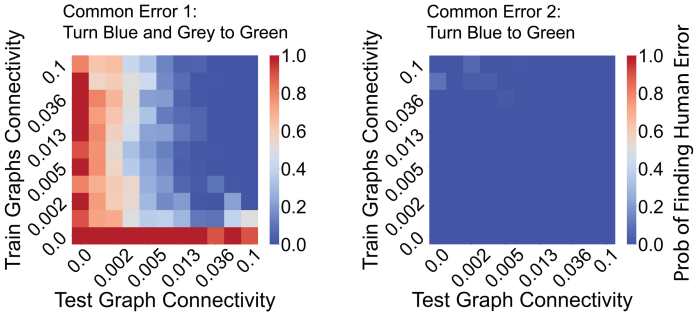

Figure 4. Connectivity of train versus test graphs and the probability of finding human common errors. Connectivity is a parameter that varies from 0 to 0.1 (arbitrary unit).

We found that while common error 2 is harder to get to, our model can arrive at both errors with densely connected graph priors during training, and by not considering connectivity when confronting the novel puzzle. It would be interesting to test if this observation holds true for humans.

Acknowledgements

We thank Jingxuan Guo for valuable discussions and advice. This work is supported by the Office of Naval Research ONR MURI N00014-19-1-2571, ONR MURI N00014-16-1-2832, and Kilachand Fund Award.

References

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). Relational inductive biases, deep learning, and graph networks. *arXiv:1806.01261 [Cs, Stat]*. http://arxiv.org/abs/1806.01261

Baxter, J. (2000). A Model of Inductive Bias Learning. *Journal of Artificial Intelligence Research*, *12*, 149–198. https://doi.org/10.1613/jair.731

Cappart, Q., Chételat, D., Khalil, E. B., Lodi, A., Morris, C., & Veličković, P. (2023). Combinatorial Optimization and Reasoning with Graph Neural Networks. *Journal of Machine Learning Research*, *24*(130), 1–61.

Chollet, F. (2019). On the Measure of Intelligence. *arXiv:1911.01547 [Cs]*. http://arxiv.org/abs/1911.01547

Do, Q., & Hasselmo, M. E. (2021). Neural circuits and symbolic processing. *Neurobiology of Learning and Memory*, *186*, 107552. https://doi.org/10.1016/j.nlm.2021.107552

Dudzik, A. J., & Veličković, P. (2022). Graph Neural Networks are Dynamic Programmers. *Advances in Neural Information Processing Systems*, *35*, 20635–20647.

Ellis, K., Wong, L., Nye, M., Sablé-Meyer, M., Cary, L., Anaya Pozo, L., Hewitt, L., Solar-Lezama, A., & Tenenbaum, J. B. (2023). DreamCoder: Growing generalizable, interpretable knowledge with wake–sleep Bayesian program learning. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *381*(2251), 20220050. https://doi.org/10.1098/rsta.2022.0050

Goyal, A., & Bengio, Y. (2022). Inductive biases for deep learning of higher-level cognition. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *478*(2266), 20210068. https://doi.org/10.1098/rspa.2021.0068

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364. https://doi.org/10.1016/j.tics.2010.05.004

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692. https://doi.org/10.1073/pnas.0802631105

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58. https://doi.org/10.1037/a0014282

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338. https://doi.org/10.1126/science.aab3050

Satorras, V. G., Hoogeboom, E., & Welling, M. (2021). *E(n) Equivariant Graph Neural Networks*. https://doi.org/10.48550/ARXIV.2102.09844

Spelke, E. S., & Kinzler, K. D. (2007). Core knowledge. *Developmental Science*, *10*(1), 89–96. https://doi.org/10.1111/j.1467-7687.2007.00569.x

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to Grow a Mind: Statistics, Structure, and Abstraction. *Science*, *331*(6022), 1279–1285. https://doi.org/10.1126/science.1192788

Xu, K., Li, J., Zhang, M., Du, S. S., Kawarabayashi, K., & Jegelka, S. (2020). *What Can Neural Networks Reason About?* (arXiv:1905.13211). arXiv. https://doi.org/10.48550/arXiv.1905.13211

Zhang, Y., Wang, N., Yu, J., Yongchareon, S., & Nguyen, M. (2022). A Short Survey on Inductive Biased Graph Neural Networks. *2022 International Conference on Service Science (ICSS)*, 64–71. https://doi.org/10.1109/ICSS55994.2022.00019