Emergent human-like path preferences and implicit subgoal selection in transformers learning graph traversal

Yuxuan Li (liyuxuan@stanford.edu), **James McClelland** (jlmcc@stanford.edu)
Department of Psychology, Stanford University, Stanford, CA, 94305, USA

Abstract

Cognitive scientists have proposed normative and heuristic principles that describe human subgoal choices and their partitioning of problems into smaller ones. Here we study the processes through which these choices or partitions arise. Building on the graph-based tasks from prior work, we train neural networks on shortest-path traversal to test whether human-like task decomposition emerges over learning. We find that a simple transformer develops a preference for paths containing nodes that occur frequently on the shortest paths in the graph, consistent with human subgoal preferences. This preference is strongest early in model learning, a phenomenon that might also be observed in human learners. We also find evidence of implicit subgoal selection in the models. These results lay the ground for using neural networks to study how humans learn to decompose tasks and select subgoals by integrating over relevant experiences.

Keywords: task decomposition; subgoal discovery; learning; neural networks

Background

Researchers have long sought to understand human task decomposition and subgoal choices. Recent work advanced this understanding by identifying common "hub" states people choose as subgoals on graphs of connected states, then mathematically characterizing the considerations that might guide such subgoal selection (Correa, Ho, Callaway, Daw, & Griffiths, 2023; Solway et al., 2014; Tomov, Yagati, Kumar, Yang, & Gershman, 2020). For example, Correa et al. (2023) collected human subgoal choices from diverse graphs and systematically compared them with prior accounts, finding that human choices were highly consistent with a simple graph property: betweenness centrality (BC). That is, people tend to choose as subgoals those states that most frequently appear on the shortest paths in the graph.

Normative or heuristic principles closely capture human subgoal choices as an end-product, but it is unclear what processes implement such choices or whether subgoal preferences can be learned. Here, we explore the idea that centrality-based subgoal choices may be learned through traversal experiences within graph-like environments. By optimizing towards more efficient (i.e., shorter) traversals, we become sensitive to nodes with high BC and come to choose paths with such nodes over other paths.

We explore these questions using neural networks, analyzing the behavior of models trained to find shortest paths. We use simple two-layer transformers (Vaswani et al., 2017), following earlier work showing that transformers can acquire sensitivity to structures in the training data and even develop the ability to decompose tasks (Li & McClelland, 2023; Manning, Clark, Hewitt, Khandelwal, & Levy, 2020).

Method

Task and dataset. We built on the 30 8-node graphs used in the human experiment in Correa et al. (2023) and generated 10 isomorphic graphs for each unique graph. The dataset comprises all start-goal pairs in these graphs whose shortest path(s) contain at least one intermediate node. We use 75% of the pairs during training and hold out the rest for evaluation. Each time a start-goal pair with multiple shortest paths appears during training, we uniformly sample a candidate path to supervise the model. We also created a dataset using 50 randomly generated 15-node graphs to examine whether the same results scale. Our reported results are based on 8-node graphs, but we found similar results using 15-node graphs.

Model. We train both a standard autoregressive decoder-only model to generate shortest paths one node at a time and a masked model that generates all intermediate nodes in parallel (Fig 1B). For our main results, the input to both models includes a learnable graph token for each isomorph of each graph, a start node, and a goal node. After an input embedding layer, both transformers include two layers of single-headed self-attention and feed-forward sublayers, with future-masked attention in the autoregressive models and all-to-all attention in the masked models. Autoregressive models are trained using teacher-forcing and evaluated with top1 rollout. Masked models are trained and evaluated using path completion where intermediate nodes on the target path are masked out. All results are aggregated across four model seeds.

Results and Discussion

Our models successfully learn the shortest path task and generalize to held-out paths with 80% or greater sequence-level accuracy (Fig 1C). The key behavior of interest is which path the models choose on start-goal pairs with more than one possible shortest path. For each isomorph of each graph, we supply its learned graph token and evaluate on held-out start-goal pairs with at least two intermediate nodes.

Model path preference. Models show preferences for paths with high centrality scores (Fig 1D1). The model-predicted paths more often have the highest average BC across intermediate nodes compared to randomly-chosen shortest paths. Because paths with high average BC often coincide with paths with high average degree centrality (DC), the models also show a DC preference. We use path regression similar to Correa et al. (2023) to compare how the quan-

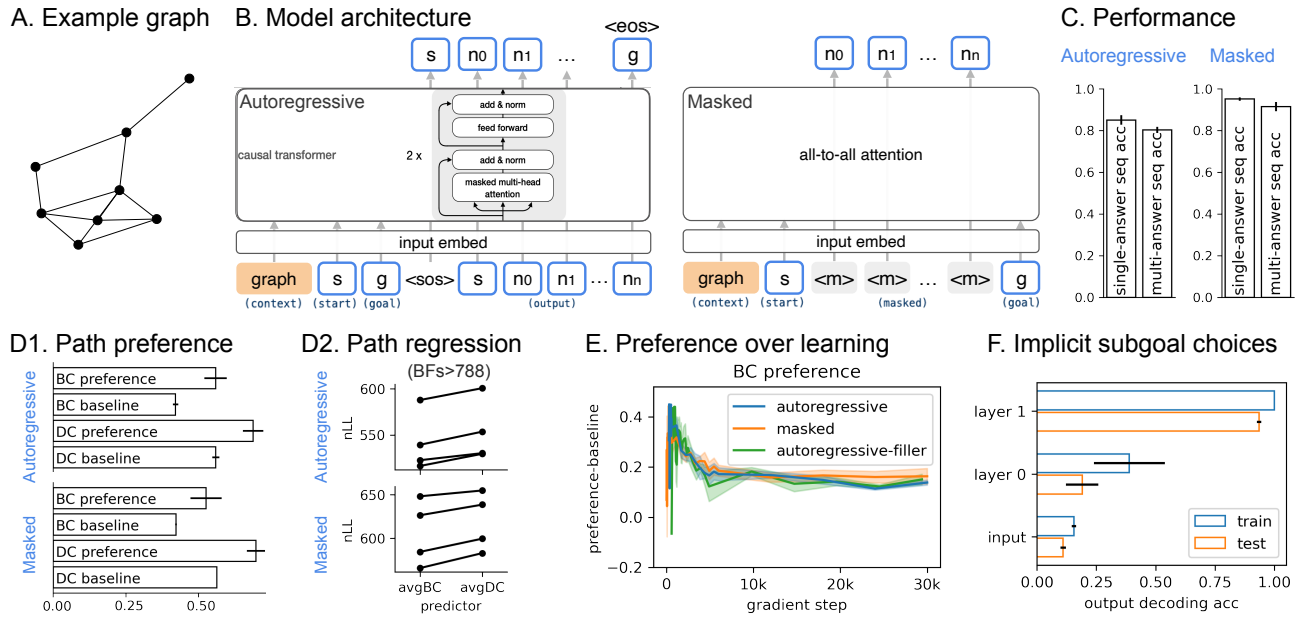

Figure 1: **A**. Example graph. **B**. Model architecture. **C**. Sequence-level accuracy on held-out paths. **D**. D1. Preference scores: how often the model-predicted path has the highest average BC/DC score; BC=betweenness centrality; DC=degree centrality. Baseline: preference score under randomly chosen shortest path. D2. Path regression using normalized log BC/DC values to predict model path choices; BF=Bayes Factor. **E**. BC preference over learning (baseline adjusted). Filler models are trained with additional filler paths to balance state visitation. **F**. Accuracy for decoding output nodes in each layer of the masked models. Error bars in C, D1, F indicate standard deviation across model runs. Error shades in E indicate the standard error of the mean.

titative BC and DC scores track model path preferences, and confirmed that BC is a better predictor (Fig 1D2). Although the effects are small, they are statistically reliable (Bayes Factors for regression pairs across seeds > 788). To disentangle BC with state frequency, we also generated a dataset with additional filler paths to increase visitation to less-visited states in each graph. We found similar path preferences in models trained on this alternative dataset.

Centrality preference over learning. We next examined how model path preference changed (relative to baseline) as models learn the task over 30k gradient steps (Fig 1E). The learning trajectory shows an initial surge in reliance on centrality followed by a gradual decrease, suggesting that models acquire sensitivity to these frequently-encountered nodes on shortest paths early and learn paths through these nodes first.

Implicit subgoal selection. We have compared model path choices to human subgoal choices as, unlike humans, we cannot explicitly probe model subgoal choices. However, we tested if the models implicitly choose subgoals on the way toward specification of the full path. We trained multinomial logistic regression classifiers to decode the output nodes on the predicted path from the token representations in each layer of the masked models (node frequencies balanced with downsampling). The decoders can successfully predict some output nodes after the first layer (Fig 1F, layer 0). We further tested whether node distance to start/goal and node BC contributed to decoder success in predicting output nodes at layer 0, on samples held out from decoder training. Both node BC, $\chi^2(1) = 16.89$, $p<0.001$, and node distance, $\chi^2(1) = 15.56$, $p<0.001$, were significant predictors, with no significant interaction. The fitted coefficients suggest that the decoders more likely predicted nodes of higher BC and in the middle of the path correctly, consistent with the view that models implicitly choose high BC nodes as subgoals in the early layer.

Generalizing to new graphs. The models discussed so far generalize to unseen paths in known graphs but cannot be tested on novel graphs. We experimented with replacing the graph token with edge tokens to signal graph connectivity information to the model, where each edge token is the summed embedding of two nodes. Two-layer edge-token models are somewhat successful at finding shortest paths in completely novel graphs (sequence-level accuracy: autoregressive≈80%, masked≈60%). These models also strongly benefit from more layers to achieve higher accuracy due to the use of in-context search over edge tokens. In addition, edge token models appear to rely more on degree centrality, especially on smaller graphs, potentially learning to count degree centrality from edge tokens. We continue to explore the different solutions learned by these different models.

Conclusion

Transformers trained to find shortest paths show path preferences consistent with human subgoal preferences and show signs that early model layers implicitly select subgoals. Model learning dynamics suggest hypotheses about human learning.

Acknowledgments

We thank members of the PDP lab for helpful feedback throughout this project.

References

Correa, C. G., Ho, M. K., Callaway, F., Daw, N. D., & Griffiths, T. L. (2023). Humans decompose tasks by trading off utility and computational cost. *PLoS Computational Biology*, *19*(6), e1011087.

Li, Y., & McClelland, J. (2023). Representations and computations in transformers that support generalization on structured tasks. *Transactions on Machine Learning Research*.

Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.

Solway, A., Diuk, C., Córdova, N., Yee, D., Barto, A. G., Niv, Y., & Botvinick, M. M. (2014). Optimal behavioral hierarchy. *PLoS Computational Biology*, *10*(8), e1003779.

Tomov, M. S., Yagati, S., Kumar, A., Yang, W., & Gershman, S. J. (2020). Discovery of hierarchical representations for efficient planning. *PLoS Computational Biology*, *16*(4), e1007594.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.