LEVERAGING HIERARCHICAL GRAPH ATTENTION NETWORKS FOR MENTAL REPRESENTATION INFERENCING

Muhammad Fusenig

Robert Slevc

University of Maryland, College Park

Abstract

Despite numerous advancements in reading comprehension (RC) models, many fail to adequately account for the role of memory in word sense-making and the dynamic construction of mental representations during reading. Existing computational approaches focus heavily on mechanistic processes like word identification and disambiguation, often neglecting how memory storage and retrieval shape comprehension in real-time, adaptive contexts. To address these limitations, we propose a Hierarchical Graph Attention Network (HGAN) that models memory storage, activation, and retrieval processes in RC. Our model integrates linguistic propositional arguments and experimental constraints within a neuro-symbolic architecture to emulate modular subsystems, balancing external task complexity with internal cognitive representations. Additionally, an encoder-decoder network is employed alongside a "Take the Best" algorithm to enable rapid, computationally efficient inference of text, leveraging multimodal sense data to construct interpretable mental representations. This novel framework offers a theoretically robust and computationally efficient model of memory-driven reading comprehension, bridging symbolic and connectionist paradigms to capture the evolving interplay of memory, reasoning, and contextual adaptation.

Background

Reading Comprehension: A Cognitive Challenge

Reading comprehension (RC) is among the most fundamental yet cognitively complex processes that individuals engage in (Duke and Carlisle, 2010). Transforming linguistic symbols

into meaningful concepts, retaining that information, and drawing upon it when needed has been the focus of intense theoretical and empirical study for decades (Joshi, 2000; Duke, 2010; Perfetti, 2014).

Existing RC models, such as the Componential and Active View models, have highlighted key factors influencing comprehension, including inference-making strategies, background knowledge, and individual motivation (Joshi and Aaron, 2000; Duke and Cartwright, 2021). These models provide valuable insights into the interplay of cognitive and social differences that shape RC ability. However, these models have struggled to account for how such factors influence RC in real-time, dynamic environments with variable contextual stimulus.

This challenge is further complicated by fundamental questions about how memory systems store and retrieve information and how these processes interact with external stimuli. As argued in scholarship of enactivist cognition, phenomenological sense—the ability to make meaning from experiences—emerges from a combination of prior memory activations and the integration of novel perceptual stimuli (Gallagher, 1997; Gallagher 2017). This suggests that traditional RC models that do not account for prior memory may fall short of fully capturing the fluidity and complexity of underlying cognitive processes that inform comprehension, particularly as they unfold over time and across different contexts.

The Promise of Hierarchical Graph Attention Networks

To address these challenges, Hierarchical Graph Attention Networks (HGANs) offer a novel framework for modeling the intricate cognitive processes underlying RC. HGANs enable

the representation of cognitive and contextual factors as nodes and encode their relationships as edges within a dynamic, hierarchical structure (Veličković et al., 2017; Sobolevsky 2021). Unlike static models, HGANs can provide an interpretable framework for integrating multimodal perceptual inputs from the environment, allowing the model to adapt to shifting task demands (Schapiro et al., 2012; Nastase et al. 2021).

This dynamic modeling approach situates cognition as an active, embodied process, capable of accounting for both individual differences and contextual influences. HGANs provide a mechanism to operationalize the factors associated with traditional RC theories by mapping the interactions of cognitive subsystems and their contributions to memory and comprehension. Through these structures, HGANs can capture the evolving mental representations that underpin RC, offering a more comprehensive and interpretable model of this critical cognitive process.

Theoretical Framework

Reasoning and memory are foundational cognitive processes that underpin reading comprehension and mental representation. A key component of reading comprehension is the ability to infer word meanings and causal relationships within a given text (Duke, 2010; Perfetti, 2014). This inferencing process is critical for transforming linguistic symbols into meaningful concepts and for constructing coherent mental representations of the information being read.

Reasoning

Traditional reasoning models suppose a multitiered memory system in which short-term and long-term memory constraints influence one's capacity for reasoning (Rugg, 2003; Chun, M. M. & Turk-Browne, 2007). In this view, mental objects are accessed and manipulated either simultaneously or in rapid succession within working memory to determine the veridicality of facts or predictions. Practically speaking, a reasoning algorithm must account for two primary components: (a) the characteristics and dimensions of the objects being reasoned over and (b) the selection algorithm that organizes and sequentially transforms mental objects. The relationship between these two components determines the ability of an information processing system (IPS) to engage in effective reasoning. As Simon (1990) observed, reasoning is dependent on "the structure of task environments and the computational capabilities of the actors (Simon 1990)." These components allow for either the construction or decomposition of objects into features and sub-features across multiple time steps. This process allows objects to be re-analyzed and defined, updating mental states, activating memory, and enabling decision-making to address underlying goals.

Memory

Memory serves as the foundation for reasoning, enabling the storage, retrieval, and manipulation of mental objects (Hayes et al., 2014; Sherman et al., 2023). Traditional theories of memory emphasize a multitiered structure, where short-term and long-term memory constraints influence the process of reasoning (Atkinson & Shiffrin, 1968; Baddeley & Hitch, 1974). These systems allow for the rapid retrieval of heuristics from long-term memory to facilitate quick

decisions, while deliberative operations in working memory enable more accurate reasoning by engaging mental objects in greater depth (Tversky & Kahneman, 1974; Squire, 2009).

In this view, specific memories are retrieved and operated over in workspaces. According to alternative views, such as those put forward by Sherman et al. (2023), memory is understood not simply as an archive of past experiences but as a dynamic interplay between previous experience and novel perceptual stimuli. These patterns are influenced by prior experiences and continuously shaped by ongoing perceptual inputs, leading to novel mental representations. Downstream reasoning is thus viewed as a set of mental operations over retrieved and generated representations. This iterative process of memory activation and updating of continually changing mental representations brings into question discrete representations of memories and mental objects.

Recent advances in cognitive neuroscience corroborate these more dynamic theories of memory storage, suggesting that activation patterns across seemingly disparate brain regions are not redundancies but instead contribute meaningfully to memory activation and retrieval processes (Sherman et al., 2023). These relationships influence both attentional mechanisms—guiding the generation of objects from the total set of perceptual stimuli—and the phenomenological experience of these objects, encompassing both unconscious and conscious appraisals (Baars et al., 2021; Sherman & Turk-Browne, 2024). These findings challenge traditional theories of multi-part memory systems, the use of heuristics, and unitary representations within localized brain regions, leading to more dynamic models and theories of memory activation.

Challenges With Current Reading Comprehension Models

While reasoning and memory are foundational to reading comprehension, existing RC models fail to capture the nuance and relationship between these two processes. A majority of models of RC do not fully account for how memory retrieval and reasoning processes adapt to contextual demands or integrate perceptual inputs to construct meaningful representations. This has resulted in a gap between models that can bridge mechanistic processes, such as word identification, with higher-order cognitive functions, such as inference and decision-making.

RC models fall into two camps: Connectionist vs. Symbolic. Connectionist models, while powerful in their focus on mechanistic processes, often struggle to incorporate higher-order cognitive factors, such as individual differences and ecological variables. Rule-based, symbolic models, on the other hand, rely on static frameworks that fail to adapt to dynamic, real-world environments where contextual factors influence cognition. These models have also been criticized for their inability to represent the full complexity of mental processes, particularly in how memory and reasoning interact to construct meaningful representations (Sprevak, 2023).

Moreover, traditional memory models often assume distinct, competing subsystems for short- and long-term storage. This segmentation has been called into question by growing evidence suggesting that brain regions associated with memory are part of a more unified network, where overlapping activation patterns contribute to memory retrieval and application (Sherman et al., 2023). Current models also struggle to account for the relational and dimensional characteristics of mental objects—features that influence reasoning accuracy and the ability to adapt to novel tasks (Kendeou & O'Brien, 2018).

Model Architecture

To address these challenges, the proposed model leverages a Hierarchical Graph Attention Network (HGAN) architecture, designed to integrate perceptual inputs, memory retrieval processes, and reasoning mechanisms into a unified, context-sensitive framework. The model has four distinct parts, (a) an encoder network responsible for attenuating to relevant objects and distilling mental representations into memory; (b) an HGAN memory network representative of underlying neurological imaging data; (c) a decoder network that processes the memory network data into operable mental objects; and (d) a reasoning module that aggregates and *reasons* over the generated mental objects utilizing a Take the Best Algorithm.

Encoder Network

The model leverages a series of encoder networks that consolidate perceptual stimulus into an integrated HGAN-based memory network. Each level in the encoder can be specified to possess certain values or can be instantiated with the outputs of a discrete, symbolic modular system. Additionally, compared to a fully encapsulated system, this structure allows complex reorganization in the training process. A fully encapsulated, modular system can be used to produce outputs that would be used to constrain a particular layer. (a) Contains an attention map that allows the model to attenuate to various objects in the environment. This attention map is refined during the backpropagation of the training phase. (b) The implementation of sub features as constrained nodes derived from relevant modular or connectionist sub systems. I.e. phonographic, orthographic, syntactic processing units, etc. (c) Distillation of relevant features, via encoding. Aggregates the influence of such features on an output node. (d) The Output node corresponds to a memory key value that (in combination with other perceptual encoder

networks) leads to particular activation of memory, which subsequently triggers the propagation of activation patterns across the memory network, forming a cascade of contextually relevant mental states and representations. This means that the encoded features serve as keys or cues that access and retrieve relevant stored information within memory that will be useful for later downstream processes.

Memory Network

The memory network architecture is informed by an enactivist interpretation of embodied cognition that simultaneously accounts for and integrates perceptual stimuli from the environment and individual information processing subsystems into memory. As such, inputs from across perceptual modalities are distilled and consolidated to operate on a shared underlying graphical representation of active brain regions and neural activity. First, neural imaging of brain regions is conducted. Second, relevant features are extracted via pretrained convolutional neural networks. Third, identified brain activation patterns are converted to graphical representations and labeled with respect to the relevant behavioral data. In this case, the sequencing of word identification. The result is a graphical representation that captures neural activation patterns associated with word identification in corresponding brain regions. Varying activations result in differentiable weighting of nodes and edges. Fourth, the aforementioned graphical representations are aggregated to create a unified HGAN. Connections between nodes with temporally or functionally correlated activation patterns are established. Fifth, an additional layer is created, synthesizing activated nodes into consolidated graphical representations, allowing for a compressed representation of underlying brain activation patterns.

Decoder Network

The decoding process begins with the transformation of graphical data from the memory network into structured representations. During decoding, the network translates the graphical data in memory into structured representations that mirror the actions or decisions made during reading comprehension tasks. These outputs capture how specific mental representations or memory activations contribute to key processes, such as inferencing, word disambiguation, or the integration of new information. Errors in prediction are corrected through backpropagation during training, ensuring that the decoder refines its outputs to closely match observed outcomes while preserving interpretability.

Similar to the encoder network, several deep learning layers integrate propositional and linguistic constraints. These constraints are preconfigured functions within respective deep learning layers that serve to guide the model during training. These constraints may include perceptual stimuli and word-sense categories. The interjection of such constraints simulate modular output that can be operated over by the neural network.

The result is symbolic and connectionist representations that self-organize to predict a set of operable mental representations. These representations, given their unique processing history, possess respective weights that will be aggregated into a Take Best Reasoning Algorithm. Node weights represent the sparsity and dimensionality of mental representation. The nodes are aggregated and used to predict the first likely satisficing optimum for a constrained reasoning task, this being word-sense making.

Take Best Algorithm

To model this process, we leverage the Take Best Algorithm, which offers greater computational efficiency and better predict human decision making compared to traditional reasoning models (Gigerenzer & Goldstein, 1996). In this model, the activation of underlying memory, combined with discrete mental processes, generates contextually relevant mental representations of external objects. Given additional stimuli—such as task demands or directions—these representations and their underlying values are used to predict next action states or satisfy a generative process leading to a "satisficing" local optimum. This optimum aggregates the set of mental representations to compute a choice or next action that maximizes a contextually relevant reward function. This reward function is variable, influenced by the characteristics, dimensions, and sequencing of the underlying mental representations. This results in a streamlined decision-making process, where the Take Best Algorithm efficiently translates mental representations from the decoder into some deliberative response that aligns with environmental and task-specific demands.

Results and Outcomes

Validation

The HGANs will be trained and tested on existing neurological and reading comprehension datasets such as the "Alice" and "Narratives" datasets to evaluate their predictive accuracy and interpretability (Bhattasali et al., 2020; Nastase et al., 2021). Their performance will be

compared directly with traditional RC models to assess improvements in predicting comprehension outcomes. Optimization techniques, such as minimizing loss and error rates, will refine the models during training, while testing on unseen data will ensure their generalizability. By incorporating reinforcement learning and human feedback, the HGANs will be further adapted to capture cognitive processes effectively. This approach will allow the HGANs to provide more accurate and interpretable predictions of comprehension outcomes compared to traditional assessments.

Anticipated Model Outcomes

The proposed HGAN model is expected to deliver a robust, dynamic framework for representing memory storage, retrieval, and activation processes. By integrating modular subsystems and attention mechanisms, the model will generate stimulus-agnostic memory representations capable of adapting dynamically to task and environmental constraints. These representations will be discrete and interpretable, enabling the transformation of perceptual stimuli into higher-order mental constructs. By minimizing representational errors through the integration of linguistic and propositional constraints, the model ensures alignment with real-world data and experimental findings. Furthermore, its scalable architecture and integration of the Take Best Algorithm provides a computationally efficient framework for full scale cognitive models, extending beyond reading comprehension to support more general reasoning and inference-making tasks.

Such a model has significant theoretical and experimental implications within reading comprehension and beyond. By integrating connectionist and symbolic frameworks into a

unified neuro-symbolic architecture via HGANs, this model operationalizes theories of memory activation and embodied cognition while remaining dynamic and interpretable. Such research advances understanding of how overlapping neural activation patterns contribute to memory retrieval and sense-making. Furthermore, the model provides a practical framework to test hypotheses about the sequencing and weighting of activation patterns, addressing persistent challenges in traditional memory models.

Works Cited

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence (Eds.), *The psychology of learning and motivation* (Vol. 2, pp. 89–195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Baars, B. J., Geld, N., & Kozma, R. (2021). Global Workspace Theory (GWT) and Prefrontal Cortex: Recent Developments. *Frontiers in Psychology*, 12. https://doi.org/10.3389/fpsyg.2021.749868

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press. https://doi.org/10.1016/S0079-7421(08)60452-1

Bhattasali, S., Brennan, J., Luh, W.-M., Franzluebbers, B., & Hale, J. (2020). The Alice Datasets: fMRI & EEG observations of natural language comprehension. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 120–125). European Language Resources Association. https://aclanthology.org/2020.lrec-1.15

Borghi, A. M., Barca, L., Binkofski, F., Castelfranchi, C., Pezzulo, G., & Tummolini, L. (2018). Words as social tools: Language, sociality, and inner grounding in abstract concepts. *Physics of Life Reviews*. https://doi.org/10.1016/j.plrev.2018.12.001

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology, 62*, 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Chun, M. M., & Turk-Browne, N. B. (2007). Interactions between attention and memory. *Current Opinion in Neurobiology, 17*(2), 177–184. https://doi.org/10.1016/j.conb.2007.03.005

Duke, N. K., & Carlisle, J. (2010). The development of comprehension. In M. L. Kamil, P. D. Pearson, E. B. Moje, & P. P. Afflerbach (Eds.), *Handbook of reading research: Volume IV* (pp. 199–228). Routledge. https://doi.org/10.4324/9780203840412

Duke, N. K., & Cartwright, K. B. (2021). The science of reading progresses: Communicating advances beyond the simple view of reading. *Reading Research Quarterly, 56*(Suppl 1), S25–S44. https://doi.org/10.1002/rrq.411

Gallagher, S. (1997). Mutual enlightenment: Recent phenomenology in cognitive science. *Journal of Consciousness Studies, 4*(3), 195–214.

Gallagher, S. (2017). Enactivist interventions: Rethinking the mind. Oxford University Press.

Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review, 103*(4), 650–669. https://doi.org/10.1037/0033-295X.103.4.650

Hao, S., Gu, Y., Ma, H., Hong, J. J., Wang, Z., Wang, D. Z., & Hu, Z. (2023). Reasoning with language model is planning with world model. *arXiv*. https://doi.org/10.48550/arXiv.2305.14992

Hayes, B. K., Heit, E., & Rotello, C. M. (2014). Memory, reasoning, and categorization: Parallels and common mechanisms. *Frontiers in Psychology, 5*, 529. https://doi.org/10.3389/fpsyg.2014.00529

Joshi, R. M., & Aaron, P. G. (2000). The component model of reading: Simple view of reading made a little more complex. *Reading Psychology, 21*(2), 85–97. https://doi.org/10.1080/02702710050084428

Kendeou, P., & O'Brien, E. J. (2018). Reading comprehension theories: A view from the top down. In M. F. Schober, D. N. Rapp, & M. A. Britt (Eds.), *The Routledge handbook of discourse processes* (2nd ed., pp. 7–21). Routledge/Taylor & Francis Group. https://doi.org/10.4324/9781315687384-2

Li, M., & Chen, Z. (2020). Hierarchical graph attention network for visual relationship detection. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 13883–13892. https://doi.org/10.1109/CVPR42600.2020.01390

Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T. Y., & Tegmark, M. (2024). KAN: Kolmogorov-Arnold Networks. *arXiv.* https://doi.org/10.48550/arXiv.2404.19756

Nastase, S. A., Connolly, A. C., Oosterhof, N. N., Halchenko, Y. O., Guntupalli, J. S., Visconti di Oleggio Castello, M., Gors, J., Gobbini, M. I., & Haxby, J. V. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex, 27*(8), 4277–4291. https://doi.org/10.1093/cercor/bhx138

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., Nguyen, M., Chang, C. H. C., Baldassano, C., Lositsky, O., Simony, E., Chow, M. A., Leong, Y. C., Brooks, P. P., Micciche, E., … Hasson, U. (2021). The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific Data, 8*(250). https://doi.org/10.1038/s41597-021-01087-0

Perfetti, C., & Stafura, J. (2014). Word knowledge in a theory of reading comprehension. Scientific Studies of Reading, 18(1), 22–37. https://doi.org/10.1080/10888438.2013.827687

Rugg, M. D., & Yonelinas, A. P. (2003). Human recognition memory: A cognitive neuroscience perspective. *Trends in Cognitive Sciences, 7*(7), 313–319. https://doi.org/10.1016/S1364-6613(03)00131-1

Schapiro, A. C., Kustner, L. V., & Turk-Browne, N. B. (2012). Shaping of object representations in the human medial temporal lobe based on temporal regularities. Current Biology, 22(17), 1622–1627. https://doi.org/10.1016/j.cub.2012.06.056

Sherman, B. E., Turk-Browne, N. B., & Goldfarb, E. V. (2023). Multiple memory subsystems: Reconsidering memory in the mind and brain. Perspectives on Psychological Science, 19(1), 103–125. https://doi.org/10.1177/17456916231179146

Sherman, B. E., & Turk-Browne, N. B. (2024). Attention and memory. Oxford Handbook of Human Memory (M. J. Kahana & A. D. Wagner, Eds.), Oxford University Press. Preprint.

Simon, H. A. (1990). Invariants of human behavior. *Annual Review of Psychology, 41,* 1-19.

Sprevak, M. (2023). Philosophical issues in computational cognitive sciences. In R. Sun (Ed.), *The Cambridge handbook of computational cognitive sciences* (pp. 1201–1227). Cambridge University Press. https://doi.org/10.1017/9781108755610.043

Squire, L. R. (2009). Memory and brain systems: 1969–2009. *The Journal of Neuroscience*, 29(41), 12711–12716. https://doi.org/10.1523/JNEUROSCI.3575-09.2009

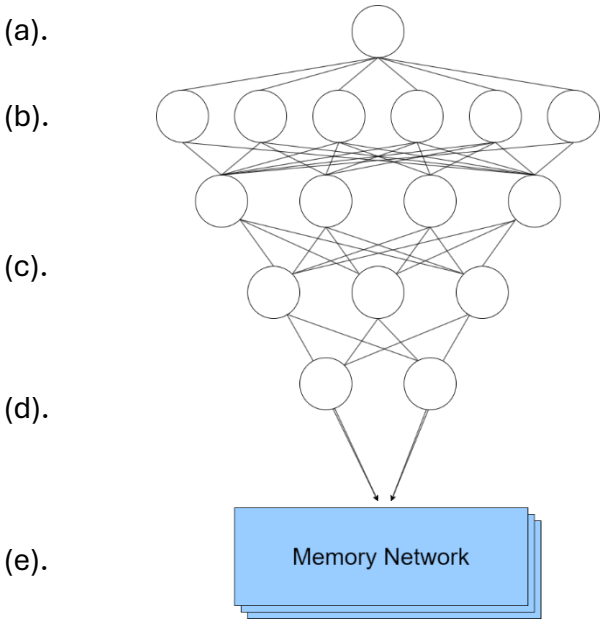
Sobolevsky, S. (2021). Hierarchical Graph Neural Networks. *Center For Urban Science+Progress, New York University*. Brooklyn, NY, USA.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Liò, P., & Bengio, Y. (2017). Graph attention networks. arXiv. https://doi.org/10.48550/arXiv.1710.10903

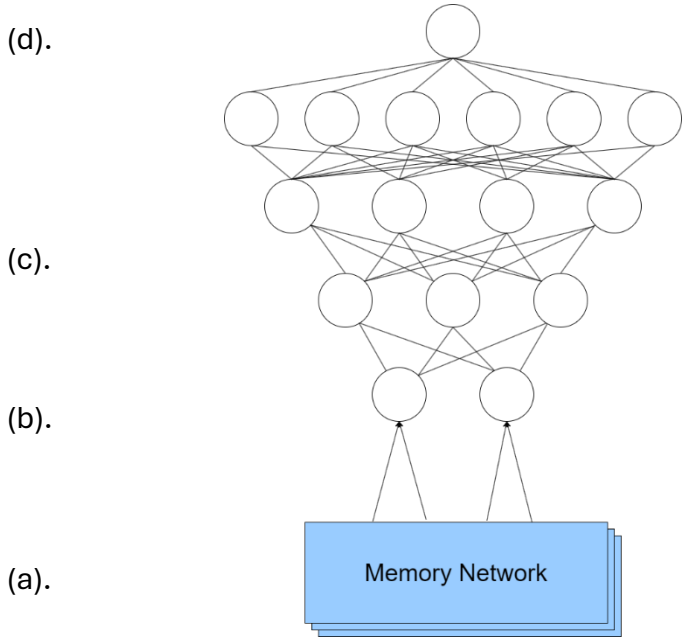
Appendix

Figure 1. Encoder



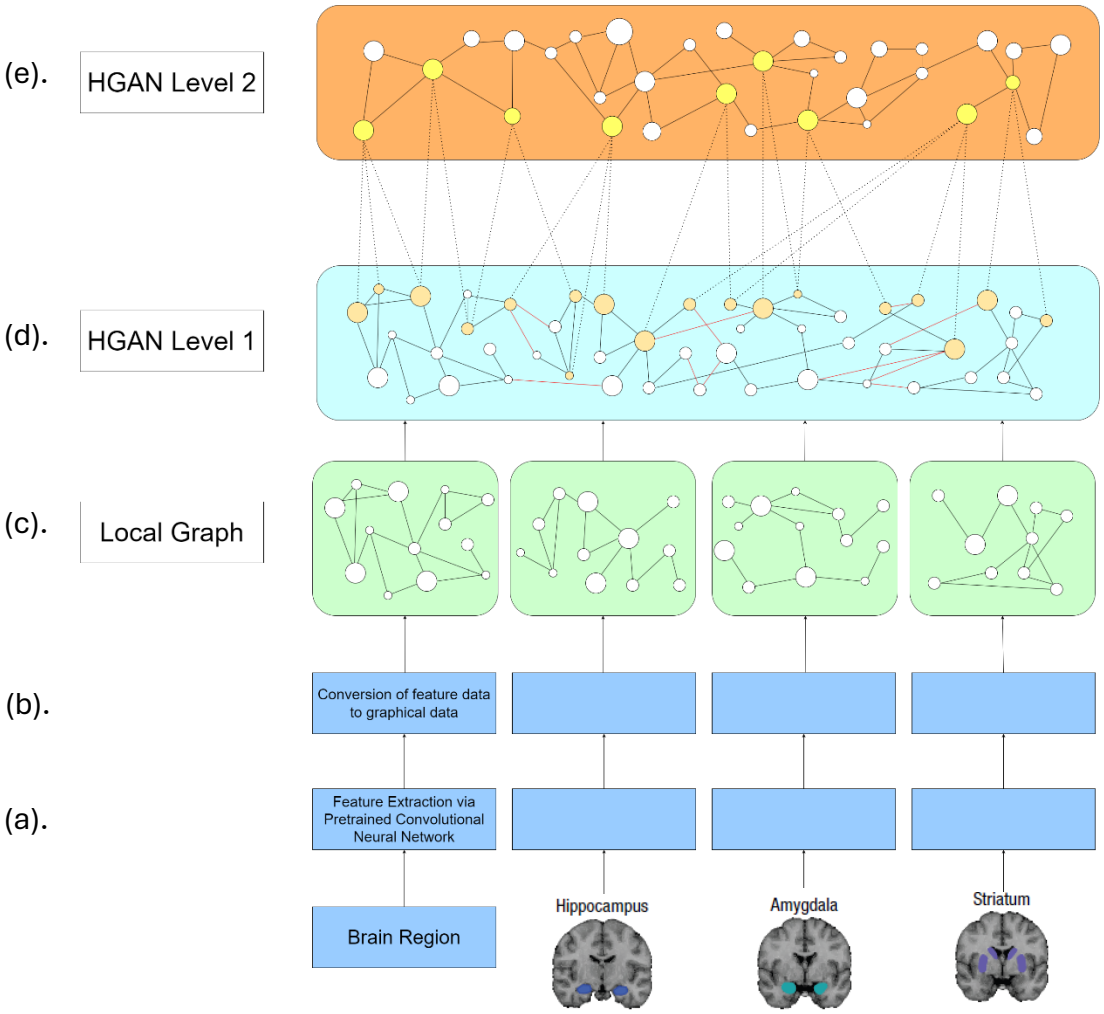
Description: Each level in the decoder can be specified to possess certain values, or can be instantiated with the outputs of a discrete, symbolic modular system. Additionally, compared to a fully encapsulated system, this structure allows complex reorganization in the training process. A fully encapsulated, modular system can be derived from the training paradigm. This trained, modular system can then be applied to out of distribution tasks. (a) The processing of a singular observed object into the encoder network. (b) The process of feature extraction. (c) The distillation of relevant features. (d) The most relevant features are reduced to a set of memory activation keys that will later act upon the memory network. (e) The encoder output instantiates some change in the memory network, given its specific key value.

Figure 2. Decoder Network



Description: The decoder network operates similarly to the encoder network, however, the decoder disambiguates the changes in the activation patterns within the memory network as a result of the encoding key value. (a) As a result of the encoder key value, the specific activation functions within the memory network are updated. (b) Most relevant activation patterns are attenuated to. (c) Information is processed up to a set of discrete nodes and learned weights. (d) The discrete nodes and learned weights in *(c)*, in addition to a task demand, are aggregated in a Task Best Algorithm to select for the satisficing option.

Figure 3. Memory Network



Description: (a) Localized brain activation data is processed through a pretrained convolutional neural network, which identifies relevant features and activation patterns. (b) The strength and boundaries of activation patterns are recorded. (c) Activation pattern data is transformed into localized graph networks. (d) Local graph networks are synthesized into an HGAN. (e) The original HGAN is abstracted, such that each node represents a pattern of activated lower level nodes.