Memory representation and retrieval in neuroscience and AI

Surya Narayanan Hari
Department of Biology and Biological Engineering
California Institute of Technology
shari@caltech.edu

Matt Thomson
Department of Biology and Biological Engineering
California Institute of Technology
mthomson@caltech.edu

Abstract

Image generation and retrieval are important emerging areas of generative AI but are also useful for world models to do visual understanding, navigation and reasoning. Unlike language which has a linear structure, visual data is non-linear and hierarchical with spatial and semantic connections between its primitives. A major question between ML and neuroscience is how to efficiently store and retrieve image primitives to generate and reason over a world of rich hierarchical visual data. Theories from neuroscience including the hippocampal memory index theory, top-down analysis, and bottom-up synthesis loops using image grammars proposed by Mumford provide a framework for AI might efficiently solve this problem. Here, we propose an image component database with image components extracted using segment-anything (SAM), along with image grammars to enable spatial and semantic retrieval. We show that our database has value by training models on foundational tasks that beat the performance of large vision models like GPT-4o. Together, we lay the foundation to train vision foundation model that is able to reason over visual data without having to simplify it into text.

Hippocampal Index Theory

While the parallels between AI systems and systems of neuroscience have been established at the neuron, circuit, and algorithm level Mumford [2020]. Trends in AI are leaning towards systems that are more agentic, which also bears roots in theoretical neuroscience Minsky [1986]. To make systems that are more agentic, we propose a system combining smaller models Hari et al. [2024] that have access to a large database of visual components extracted from previously viewed scenes. This database has evidence in the hippocampal index theory Koch and Davis [1994].

The brain solves a variety of retrieval problems, from retrieving data, as in the case of arithmetic to retrieving whole circuits, as in the case of a reflex action involving multiple muscle groups. In AI systems too, retrieval is important for problems ranging from Question Answering Lewis et al. [2021], image search Radford et al. [2021], Girdhar et al. [2023] and generation since the tokens LLMs generate are retrieving over a fixed dictionary via attention score optimization.

In the hippocampal index theory, objects are represented in the brain as representations that allow for partial recall and semantic traversal. Building a system using AI requires defining image primitives, understanding spatial relations between them and encoding semantic associations between them. Previous literature has built on defining visual primitives and parametrizing the spatial and semantic

Preprint. Under review.

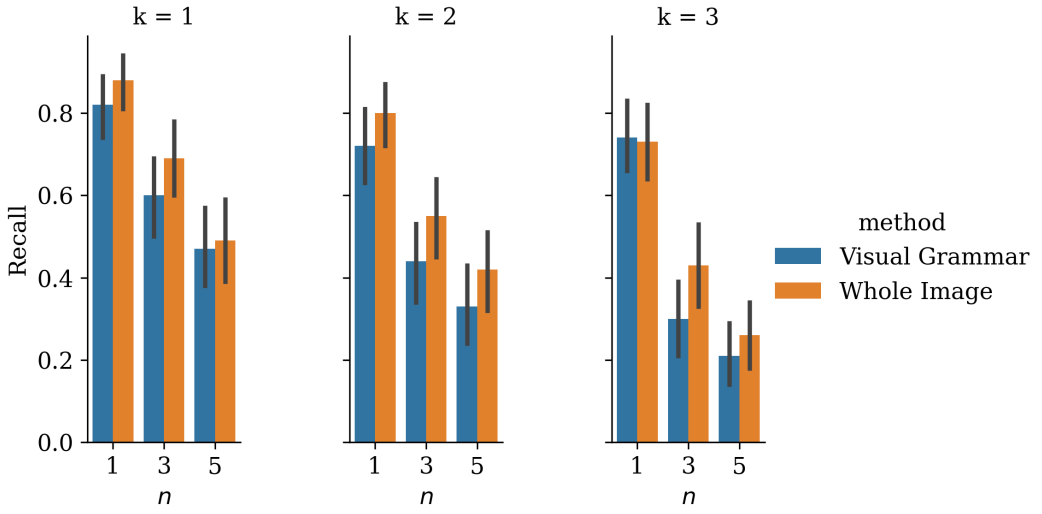

Figure 1: Recall performance of a query with k components over the top n images returned by the recall system. Our results are shown in blue, those using Imagebind embeddings are shown in orange

relationships between them Zhu and Mumford [2006], Mumford and Desolneux [2018], Koch and Davis [1994].

Image Component Database

We built a large database of 3 million image components extracted from 30,000 images. 10,000 images were taken from Flickr hlky [2024] and the remaining were frames of videos taken by the authors. To obtain object primitives from these images, we used SAM Kirillov et al. [2023] and treated all object masks as invidual objects. We overlaid the mask onto the image to isolate each object and used GPT-4o to label the object. The image components returned by SAM are paired with centroids that allow us to build spatial graphs of connectivity between the components or their corresponding labels. Objects that were unclear and noisy labels were culled using a verification step by passing the object paired with its label to GPT-4o, along with the prompt "Does this label accurately describe this image", and only components that were returned in the affirmative were retained. The labels of verified objects were shortened to one-word labels to enable grouping. Components were given a unique ID using a random number generator and embedded using Imagebind Girdhar et al. [2023], to create a database enabling the retrieval of the image component, its label, short label, image embedding, position and metadata (source image), making the system suitable for vector database operations.

Recall

We use AND grammars, originally proposed by Song-Chun Zhu and David Mumford Zhu and Mumford [2006] to retrieve images from the database. A query q can composed of multiple objects joined with an `AND` operator of the form $\texttt{q}_1$ `AND` $\texttt{q}_2$ $\cdots$ `AND` $\texttt{q}_m$, (for example "`dog AND ball`)". We represent each image as a bag of components $C = \{c_1, c_2, \cdots, c_k\}$. A similarity score s is produced for each image as

$$s = \prod_{i=1}^{|q|} \max_k \; q_i \cdot c_k$$

where $\cdot$ represents the dot product and q is the number of terms in the query. The query terms are embedded using Imagebind to project them in the same space as the component vectors. Images are returned to the user in descending order of their similarity scores.

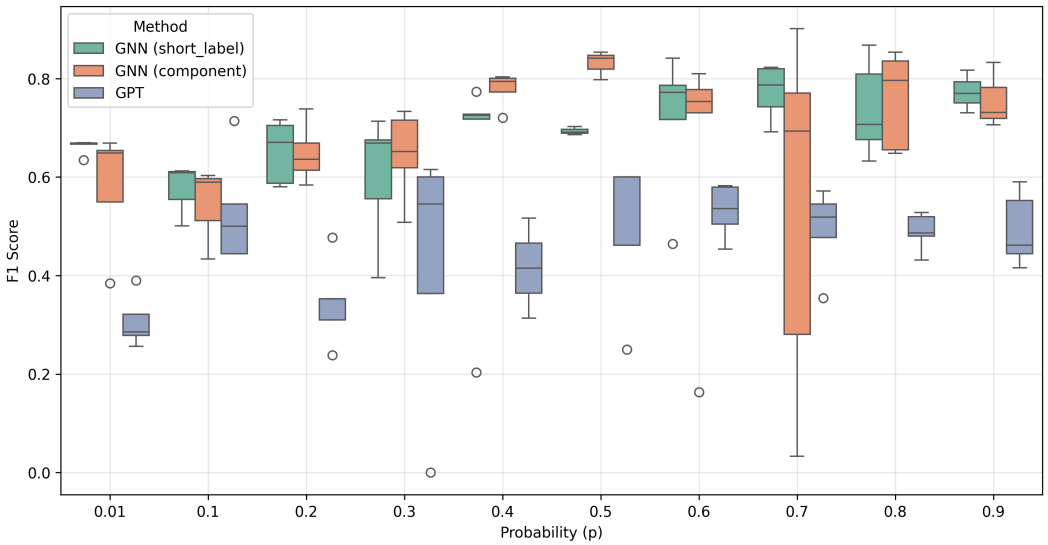

Figure 2: Comparison of models trained to identify whether a graph is of an image whose components are shuffled or not. GPT-4o (Blue) is compared against our GNNs trained with nodes as embeddings of image components (orange) and embeddings of short labels of objects (teal). x axis shows fraction of nodes shuffled.

To benchmark the recall using our component database, we used Imagebind to retrieve images by embedding the query q and retrieving the n most similar images sorted by descending cosine similarity. To measure recall, we used GPT-4o to check whether the recalled image actually contained the queried terms. In Figure 1, we compare our system (orange) against using Imagebind embeddings (blue), we measure recall over 100 queries with k components each and measure recall against the top n image retrieved. For $k = 1$, the 100 queries used were the 100 most common single word labels of objects in our dataset; for $k = 2$, the 100 queries were the 100 most common co-occuring pairs (triplets for $k = 3$) of objects.

Shuffling

The database of image components that we develop allows a foundational model to learn spatial and semantic relations between objects in images. To demonstrate the utility of such a dataset, we define a foundational model task of identifying whether an image has its objects shuffled. Half the images in our dataset were shuffled [1]. Amongst the images where components were shuffled, a fraction p of its components were taken and their locations permuted, while keeping the connectivity the same - i.e. if an image was composed of components $c_1, c_2, \cdots c_n$ at locations $l_1, l_2, \cdots l_n$ with edge matrix E, then after shuffling, the mapping would be of the form $\{(\mathtt{c}_2, \mathtt{l}_1), (\mathtt{c}_n, \mathtt{l}_2), \cdots, (\mathtt{c}_2, \mathtt{l}_n)\}$ and still have edge matrix E.

We benchmarked a Graph Neural Network (GNN) trained with 4 layers of GAT modules and 2 MLP layers (< 5M params total) in this shuffling task against an off the shelf visual reasoning model (We used GPT-4o in this work). The vision model we benchedmarked against was shown the image as a picture of a graph where the nodes were blue circles placed at the centroids of the image components they represented and a one-word description of the component was overlaid on top of the node. The model was also passed a prompt that asked it to identify whether the graph that was passed in was from an image whose components were shuffled or not.

As the fraction of nodes shuffled increased, our trained model was able to identify whether the graph was shuffled with greater accuracy, as shown in Figure 2. Error bars shown are over 5 random folds shuffling different sets of nodes. To create the adjacency matrix of the graph used both for the

[1] images taken from Flickr only were used in this experiment to avoid distributional leak between frames of a video

GNN and GPT, we used a kNN graph, drawing the edges between a component and its k nearest components (results in Figure 2 used $k = 5$).

References

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All, May 2023. URL http://arxiv.org/abs/2305.05665. arXiv:2305.05665 [cs].

Surya Narayanan Hari, Rex Liu, and Matt Thomson. Herd: Using multiple, smaller LLMs to match the performances of proprietary, large LLMs via an intelligent composer, September 2024. URL http://arxiv.org/abs/2310.19902. arXiv:2310.19902 [cs].

hlky. Flickr. https://huggingface.co/datasets/bigdata-pw/Flickr, 2024.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

Christof Koch and Joel L. Davis, editors. *Large-scale neuronal theories of the brain*. Computational neuroscience. MIT Press, Cambridge, Mass, 1994. ISBN 978-0-262-11183-6.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks, April 2021. URL http://arxiv.org/abs/2005.11401. arXiv:2005.11401 [cs].

Marvin Minsky. *The society of mind*. Simon and Schuster, New York, 1986. ISBN 978-0-671-60740-1.

David Mumford. The Convergence of AI code and Cortical Functioning – a Commentary, October 2020. URL http://arxiv.org/abs/2010.09101. arXiv:2010.09101 [cs].

David Mumford and Agnès. Desolneux. *Pattern theory: the stochastic analysis of real-world signals*. CRC Press Taylor & Francis Group, Boca Raton, second edition edition, 2018. ISBN 978-1-138-05396-0. OCLC: 1039146455.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision, February 2021. URL http://arxiv.org/abs/2103.00020. arXiv:2103.00020 [cs].

Song-Chun Zhu and David Mumford. A Stochastic Grammar of Images. *Foundations and Trends® in Computer Graphics and Vision*, 2(4):259–362, 2006. ISSN 1572-2740, 1572-2759. doi: 10.1561/0600000018. URL http://www.nowpublishers.com/article/Details/CGV-018.