Oregon Undergraduate Research Journal: McNair Special Issue
21.3 (2023) ISSN: 2160-617X (online)
ourj.uoregon.edu

Comparing Verbal Descriptions of Image Memories with Natural Language Processing

Julian Gamez*

Abstract

A goal of memory research is to understand how the brain remembers similar events. Analyzing data from human subjects, we explore how competition between memories of images influences their recall by answering the question *Does studying images from similar or differently themed categories affect the verbal content used to describe them?* The competitive condition was composed of images from a single category ("Pond 1," "Pond 2"), whereas the non-competitive condition was a set of images from different categories ("Pond 1," "Library 1"). Specifically, we aimed to quantify how verbal memories of these images varied depending on the study condition. To quantify subjects' verbal memories, we used natural language processing to map subjects' descriptions of the images onto points in a high-dimensional "text embedding" space. We performed dimensionality reduction and clustering analyses on these text embeddings and found that semantic representations of images studied in the competitive condition were similarly differentiated compared with those in the non-competitive condition. Our results suggest that verbal memories of images were influenced by the similarity of subjects' memories and that highly similar memories may push their respective representations away from one another.

1. Introduction

One of the main regions of interest involved in learning and memory is the hippocampus (Eichenbaum, 2000). Investigation of place cells in the rodent hippocampus has been able to demonstrate that perceived environmental differences can affect patterns of neural activity recorded in this brain region (Colgin et al., 2008). These patterns of activity are subject to sudden changes that come about when the rodent moves through space over time. It has also been shown that these patterns of neural activity can be affected by internal differences that reflect change due to experience (Bostock et al., 1991). The change in hippocampal neural activity tied to the rodent's movement and experience is referred to as "remapping." Remapping highlights a connection between the reported differences in rodent hippocampal activity and the study of memory interference in humans. Since many events that a person might experience are likely to share similar physical features or have been presented at a common point in time, it is appealing to posit that interference between episodic memories, particularly highly similar memories, could be resolved through this process of hippocampal remapping.

In human subjects, remapping can also be thought of in terms of the "repulsion" that has been observed when similar memories

Julian Gamez (jgamez@uoregon.edu) is graduating with a Bachelor of Science in Psychology and a minor in Entrepreneurship. Julian's research project titled "Comparing Verbal Descriptions of Image Memories with Natural Language Processing" was mentored by UO Professors Dr. Brice Kuhl of Psychology and Dr. James Murray of Biology and Mathematics, as well as assisted by graduate student Anisha Babu of the Kuhl lab. He plans to attend continue research after graduation to further explore how neuroimaging and computational methods can be used to study learning and memory.

overlapping in representational space appear to move away from each other. In one behavioral experiment, subjects studied nearly identical images of objects paired with faces, with the colors of the objects serving as the independent variable (Chanales et al., 2021). This study found that subjects' "color memories" for objects sharing a high degree of color similarity were reported as being more different from one another than the objects' actual appearances would suggest. Repulsion among highly similar color memories was most pronounced in subjects who displayed the fewest interference errors during associative recall. This finding is consistent with the notion that interference between similar memories drives remapping and reduces interference.

Previous cognitive neuroscience research in humans has explored this prediction further using fMRI experiments to test recall (Favila et al., 2016; Wanjia et al., 2021). In one experimental paradigm, subjects studied paired images of similar objects and scenes (Wanjia et al., 2021). For example, one pairing might include the object "Guitar 1" and the scene "Lighthouse 1," whereas another pairing might be "Guitar 2" and "Lighthouse 2." Subjects were tasked with learning these associations over multiple pairing trials. During the testing phase, the scenes were presented to the subjects, where the subjects' goal was to correctly identify the object associated with the scene. Learning was defined to have occurred once subjects could verbally recall these associations with high confidence. The study found that changes in certain subregions of the hippocampus shared a correspondence with learning over time—something that has similarly been found in rodent experiments examining activity in those subfields (Lee et al., 2004). More specifically, memories that shared a high degree of similarity before learning had occurred displayed a greater decorrelation amongst their hippocampal representations after memory interference had been resolved.

Taken together, these findings suggest that hippocampal representations of episodic memories that are very similar to one another become differentiated as their interference with one another is minimized across learning trials. Notably, this repulsion effect can be captured by the extent to which these similar representations shift according to a reduction in interference. Our data analysis here attempts to extend this research to verbal memories by investigating whether this repulsion effect occurs for verbal descriptions of similar images in different behavioral conditions by using natural language processing techniques to quantify the degree of similarity between descriptions of related versus unrelated images.

2. Methods

Our data analysis is based upon a study of 120 participants at the University of Oregon (Figure 1). After providing their consent to participate, subjects were asked to study six different categories of scenes ("Indoor Pool," "Library," "Downtown Street," "Soccer Stadium," "Ice Skating," or "Pond"), each containing six unique images. During the study phase of the experiment, subjects were presented with images from these categories paired with an image of a random human face (Figure 1a). Subjects in the baseline condition studied scenes belonging to different categories, whereas those in the competitive condition studied images belonging to the same category. During the testing phase, subjects' memories were measured by how well they could recall the scene associated with the presented face from the study phase (Figure 1b). Subjects were then asked to use at least ten words to verbalize their memory of the studied images in as much detail as possible.

Subjects' words and phrases were recorded as the verbal memory input, and a natural language processing (NLP) algorithm called MPNET was used to convert the verbal memories into numeric values representing how images were remembered by subjects during the testing phase

(Song et al., 2020). This process allowed us to code the different descriptions that subjects provided into a set of numbers that could help us analyze the language people used to communicate their memory of an image. This tool generated a quantitative representation of subjects' verbal memories of scenes by transforming words and phrases into a 768-length vector through a process known as a "text embedding." Each column of the dataset can be thought of as a "feature" value of the text embedding, with 768 features per image (Figure 1c). The feature values of the subjects' verbal memories across both conditions were pooled together to perform most of our analysis.

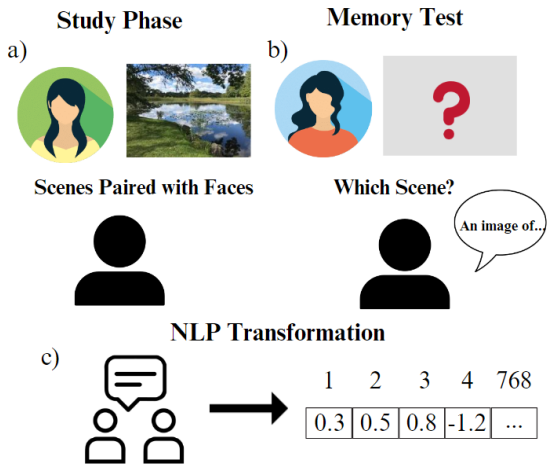

Figure 1. (a) Subjects studied images of scenes paired with a random human face. (b) Subjects' memories of scenes were tested when they were asked to remember and describe the correct scene that was associated with the face from the study phase. (c) The verbal descriptions of scenes from memory became transformed through the NLP MPNET model into numeric variables used to represent the meaning of the subjects' responses.

We normalized the dataset around a standard normal distribution with a mean of 0 and standard deviation of 1. After performing normalization on the dataset of verbal memories, we conducted principal components analysis (PCA) on the text embeddings to reduce the number of features—or dimensions—corresponding to the verbal memories of scenes. PCA is a mathematical technique that can help improve interpretability of large datasets. Instead of using all 768 features created through the NLP transformation process, PCA allowed us to "compress" our dataset into its principal components. The principal components represent linear transformations of the text embeddings into unrelated variables that can be used to communicate the maximum amount of information about the verbal memories while reducing the number of overall features used to express the dataset. We were able to reduce the number of features from the initial 768-length vector generated by the NLP algorithm to six principal components and plotted the proportion of variance explained by each component in either condition.

After the PCA dimensionality reduction, we then used a k-means clustering algorithm to group verbal memory representations together based on their similarity. While PCA reduces the number of dimensions of the data to display it in a more concise way through the creation of principal components, k-means clustering attempts to represent the data in the form of the tightest clusters. The assignment of each data point to a cluster is determined by how far away the point is from the center of the cluster, with the objective of maintaining a tight cluster. We decided to use six clusters for our k-means clustering of the principal components by measuring the within-cluster sum of squares error to determine the point at which an increase in the number of clusters would provide diminishing returns for our analysis. We plotted the within-cluster sum of squares error found in either condition and created graphs containing subplots of the clustering results to compare baseline and competitive verbal memory representations in the PCA space.

3. Results

PCA dimensionality reduction on our dataset revealed a decline in the amount of variation amongst the text embeddings that could be

explained by each subsequent component following the fifth component (Figure 2). We found that five principal components could explain over half of the cumulative variance in the verbal descriptions (Figure 3). This supports the idea that dimensionality reduction can be used to explore the same data with fewer variables, improving the quality of visualization.

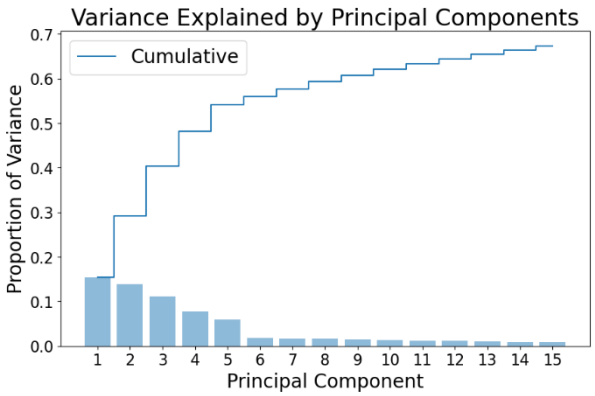

Figure 2. Proportion of variance captured by principal components constructed from the pooled set of text embeddings and plotted according to explained variance.

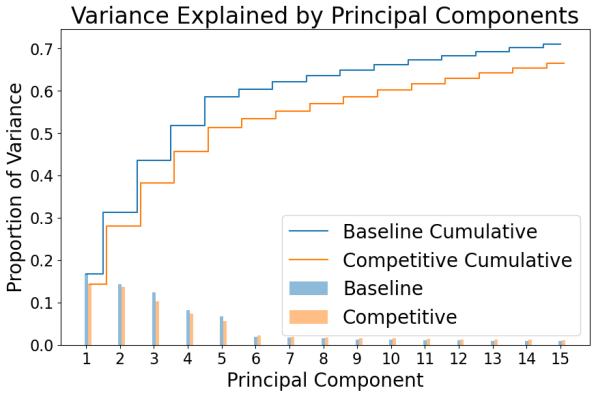

Figure 3. Proportion of variance captured by principal components constructed from data separated by condition and plotted according to explained variance.

Likewise, this result informed us that retaining more than five principal components to represent the information expressed by the text embeddings would only be marginally beneficial at capturing any additional information about the content of the verbal memories. Furthermore, to ensure that our k-means analysis would also be reflective of the most important information captured by the NLP algorithm for assessing similarity, we chose to keep six principal components to perform clustering analysis on the verbal memories delivered by subjects. By measuring the within-cluster sum of squares error and plotting it as a function of the number of clusters, we were able to observe the point at which adding more clusters into our analysis would not provide any noticeable increase to the tightness of each cluster (Figure 4; Figure 5).

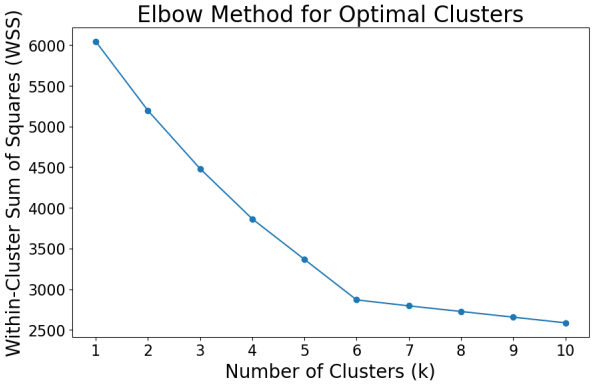

Figure 4. Elbow method plot constructed from the pooled set of principal components, useful for finding the optimal number of clusters to group principal components. The within-cluster sum of squares error decreases as the number of clusters increases. The "elbow" can be seen where the number of clusters approaches six.

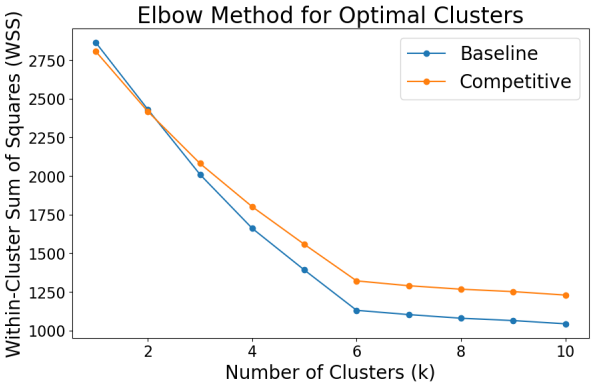

Figure 5. Elbow method plot constructed from the separated data of principal components in each condition.

Reducing the 768-length vectors of text embeddings via PCA allowed us to qualitatively visualize how subjects in each condition recalled images using these new values to represent their words and phrases. We conducted this visualization by generating subplots to compare the different principal components across both

conditions using the six predefined clusters (Figure 6). Moreover, though we had to select the number of clusters for our k-means analysis, verbal memories of each category appear distributed according to each of the six categories of the behavioral experiment. This finding reinforces our decision for the selection of six clusters, as it confirms that information about images from the same categories in the dataset aggregated together.

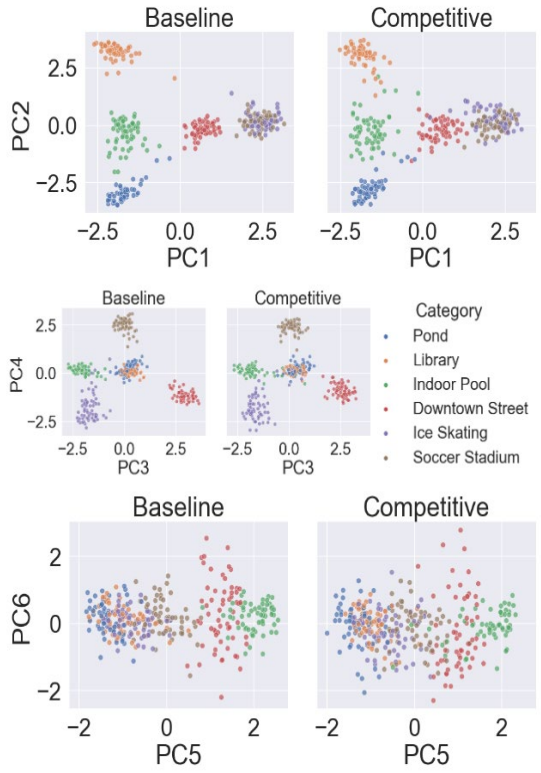

Figure 6. K-means clustering results separated by condition and plotted according to information represented along each principal component. The categories are distinguished by color to highlight the effects of the clustering on their separation.

Upon plotting the clusters with respect to each of our principal components, we identified that the distribution of verbal memory representations in the PCA space using six predefined cluster segments captured differentiation between categories in either condition. While we have not specifically performed any statistical analysis on this relationship here, Figure 6 reveals a minor qualitative difference in the tightness of clustering between conditions. This could suggest that when subjects study multiple scenes from the same category, their verbal recall tends to emphasize information that differentiates those scenes, though we cannot draw that conclusion based on these results alone.

4. Discussion

We showed using dimensionality reduction that verbal memories of similar images from the competitive condition, and of images studied under the baseline condition from unrelated categories, could be compared using text embeddings created with the NLP MPNET model. After normalizing the dataset and performing PCA and k-means clustering on text embeddings, we found that the plotted representations of verbal memories group tightly into six predefined clusters that become separated by category. This validates that the NLP algorithm was able to detect similarity between verbal recall, such that two images from the same category are found to group more closely together than two images from different categories. The clustering pattern of the principal components found in either condition suggests that scenes from different categories are differentiated by subjects' verbal memories delivered during the testing phase. The visual differences found by clustering the verbal memories address our question as to whether the content used to describe the images changes depending on the similarity of the studied images.

Based on the theory that inference occurs when memories compete for recall, we would have expected to see evidence of more diffuse clustering between the similar memories of each category in the competitive condition compared to the unrelated images seen in the baseline condition. However, our prediction that this differentiation would be more exaggerated when memories for two images compete to be remembered cannot be supported by our current

analysis. The subtle differences captured in Figures 3 and 5, rather, demonstrate the capacity for NLP to tease apart nuanced ways in which similar images become remembered differently as a result of interference. The qualitative results displayed in Figure 6 convey that these small variabilities between conditions can be represented graphically to the extent of the principal components' ability to capture variance from within either condition.

While the scope of our data analysis is not able to support conclusions about the involvement of specific subregions of the hippocampus that might be involved in learning to differentiate these scenes, it could be the case that memories remap according to prediction errors calculated through dissociations in the specified subregions (Dimsdale-Zucker et al., 2018; Keinath et al., 2020). Our findings support existing research that aims to assess how verbal memories taken at a recall during behavioral experiments compare to neuroimaging data from the same subjects while being scanned in an MRI machine.

Future research efforts might explore NLP text embeddings using representational similarity analysis to correlate representations in the brain at specific timepoints during learning (Kriegeskorte et al., 2008). NLP methods could help make inferences about the types of information that correspond to hippocampal remapping of specific memories. Performing analysis on text embeddings recovered before and after learning might reveal quantitative differences about how people remember images over time as interference is minimized. Our analysis of NLP text embeddings here is an important first step towards that goal by offering some justification for the use of these methods in the context of verbal memory analysis.

References

Chanales, A. J. H., Tremblay-McGaw, A. G., Drascher, M. L., & Kuhl, B. A. (2021). Adaptive Repulsion of Long-Term Memory Representations Is Triggered by Event Similarity. *Psychological Science*, *32*(5), 705–720. https://doi.org/10.1177/0956797620972490

Colgin, L. L., Moser, E. I., & Moser, M.-B. (2008). Understanding memory through hippocampal remapping. *Trends in Neurosciences*, *31*(9), 469–477. https://doi.org/10.1016/j.tins.2008.06.008

Dimsdale-Zucker, H. R., Ritchey, M., Ekstrom, A. D., Yonelinas, A. P., & Ranganath, C. (2018). CA1 and CA3 differentially support spontaneous retrieval of episodic contexts within human hippocampal subfields. *Nature Communications*, *9*(1), Article 1. https://doi.org/10.1038/s41467-017-02752-1

Eichenbaum, H. (2000). A cortical–hippocampal system for declarative memory. *Nature Reviews Neuroscience*, *1*(1), Article 1. https://doi.org/10.1038/35036213

Favila, S. E., Chanales, A. J. H., & Kuhl, B. A. (2016). Experience-dependent hippocampal pattern differentiation prevents interference during subsequent learning. *Nature Communications*, *7*(1), Article 1. https://doi.org/10.1038/ncomms11066

Keinath, A. T., Nieto-Posadas, A., Robinson, J. C., & Brandon, M. P. (2020). DG–CA3 circuitry mediates hippocampal representations of latent information. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-16825-1

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Lee, I., Rao, G., & Knierim, J. J. (2004). A Double Dissociation between Hippocampal Subfields: Differential Time Course of CA3 and CA1 Place Cells for Processing Changed Environments. *Neuron*, *42*(5), 803–815. https://doi.org/10.1016/j.neuron.2004.05.010

Song, K., Tan, X., Qin, T., Lu, J., & Liu, T.-Y.

(2020). *MPNet: Masked and Permuted Pre-training for Language Understanding* (arXiv:2004.09297). arXiv. http://arxiv.org/abs/2004.09297

Wanjia, G., Favila, S. E., Kim, G., Molitor, R. J., & Kuhl, B. A. (2021). Abrupt hippocampal remapping signals resolution of memory interference. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-25126-0