Learning categories with invariances in a neural network model of prefrontal cortex

Suhas E. CHELIAN[a,1], Rajan BHATTACHARYYA[a], Randall O'REILLY[b]
[a] *HRL Laboratories, LLC,* [b] *University of Colorado, Boulder*

Abstract. Prefrontal cortex (PFC) is implicated in a number of functions including working memory and categorization. Here the Prefrontal cortex Basal Ganglia Working Memory (PBWM) model (O'Reilly and Frank, 2006) is applied to learning categories with invariances. In particular, motivated by a problem in scene recognition, objects in different locations are sequentially presented to the network for categorization. The model learns to recognize these classes without explicit programming, thus modeling human categorization along with characteristics such as generalization to novel sequences and frequency dependent effects. Future extensions to the current work including applications to other domains and modeling functionally distinct segregations of PFC and neuromodulatory systems are also described.

Keywords. Prefrontal cortex (PFC), working memory, categorization, scene recognition.

Introduction

Prefrontal cortex (PFC) is implicated in a number of functions including working memory and categorization. Working memory includes the ability to actively maintain task relevant information over delays and distracters. As early as the 1930s, ablation studies established that PFC was critical to working memory in delayed response tasks [1, 2]. These findings have since been extended with single unit (e.g., [3, 4]) and fMRI recordings (e.g., [5]). Categorization involves grouping perceptually dissimilar objects into functionally organized classes. (For current purposes, "category" and "class" will be used interchangeably.) Although posterior cortical areas are certainly involved in the recognition of individual objects, categorization is at least partly encoded by PFC (e.g., [6, 7]). For example, in the case of vision, inferotemporal cortex might represent objects such as a plum or fish, but PFC can group these into an *is-edible* class.

Perhaps the simplest task that involves both working memory and categorization is classifying sequences of objects. For example, in the case of language, if 4 character words are generated from the English letters A through Z, PLUM and FISH would belong to the *is-edible* class but other combinations (e.g., WISH, TIRE) would not. Working memory is required to differentiate between words (e.g., FISH v. WISH), while categorization is required to group individual objects (e.g., PLUM and FISH) into classes.

One recent neural network model of working memory is the Prefrontal cortex Basal Ganglia Working Memory (PBWM) model [8]. PBWM posits "stripes" within

[1] Corresponding author (sechelian@hrl.com).

PFC to hold each working memory symbol—each letter in PLUM for example—over delays and distracters. Symbols are gated into and out of PFC stripes by basal ganglia if doing so is deemed to be rewarding. Although PBWM is an extensive and sophisticated model, it has not been applied to tasks with overlapping symbols or generalization to new input sequences. For example, the two categories in the popular 1-2-AX task require rote memorization of two disjoint input sequences (1-AX and 2-BY). Categories can be learned with bio-inspired models (e.g., [9, 10]) but few claims are made with regards to detailed neuroanatomical underpinnings. Thus, this work extends the neuro-realism of PBWM to classifying sequences of symbols motivated by a problem in scene recognition.

Section 1 describes the inputs, datasets, and networks studied in this paper. Section 2 contains results, while section 3 contains the discussion and conclusion.

1. Material and Methods

Consider the task of scene recognition. This too can be seen as classifying sequences of objects. Saccades create a series of foveated image regions which must be integrated in working memory to discern the category of a scene. Each object and its relative location define a scene. Scenes may exhibit invariance over objects or positions as explained below.

1.1. Inputs and Datasets

Four scene categories are defined in Table 1. To stress working memory and categorization as opposed to visual processing, fixed scan paths (top to bottom, left to right) to 4 discrete locations are considered. Furthermore, 8 objects with 2 objects per scene are used.

The first scene category has no object or positional invariance. Changing either a constitute object or its location invalidates a scene from the category. The second scene category has object invariance but not position invariance. For example, in Table 1, if A is store, B is hotel, and Y is a surrounding gate, then AY and BY are both gated businesses. YA and YB, on the other hand, would be business surrounding a gate. The third scene category has positional invariance but no object invariance. For example, in Table 1, a hotel (B) at or above a certain ground plane (Z) would be a mountain hotel. Changing the hotel to a shop (A), however, means it is no longer a member of that scene class. The fourth scene category has both positional and object invariance. That is, both the object in the first sector and the position of the second object can vary. The first two members in this category—B(AA) in different spatial configurations (where (AA) represents an arbitrary symbol following Z)—exhibit positional invariance, while the second two members—C(AA) and D(AA)—exhibit combined position and object invariance. In addition to the four scene categories, non-target scenes were generated to provide counter-examples to the previously defined scene categories. Objects and positions are used in multiple categories to increase the difficulty of the task. For example, B is used in 3 scene categories and in the non-target class.

Table 1. Scene classes 1 through 4, which exhibit invariances between objects and locations, and examples of the non-target scenes.

Scene class	Sectors		Description
1	A	X	No invariances.
2	A	Y	Object invariant.
2	B	Y	
3	B	Z	Position invariant.
3	B		
		Z	
4	B	AA	Multiply invariant.
4	B		
	AA		

Scene class	Sectors		Description
4	C	AA	Multiply invariant (continued).
4	D		
	AA		
Non-target	C	X	Object changed wrt scene class 1
Non-target	B		Position changed wrt scene class 2
		Y	
Non-target	C		Object changed wrt scene class 3
		Z	
Non-target	D		Position and object changed wrt scene class 4
		AA	

A training set consisted of 50 random scenes generated from a probability distribution across scene categories. Testing sets were then created with 25 scenes not in the training set. Furthermore, to study the effect of relative proportions of scene categories on training and generalization, the probability of each scene class was varied in 3 ways. In the first dataset, all 4 scene categories and the non-target class were equiprobable. The second dataset assigned equal probabilities to the 4 scene categories but non-target scenes constitute half of the dataset because there are many more possible scenes in this category. The third dataset is similar to the second, but each scene class was given more examples based on the expected learning difficulty. For example, twice as many scene 4s were created than scene 3s because it involves 2 invariances instead of 1. Table 2 summarizes the class distributions in each dataset.

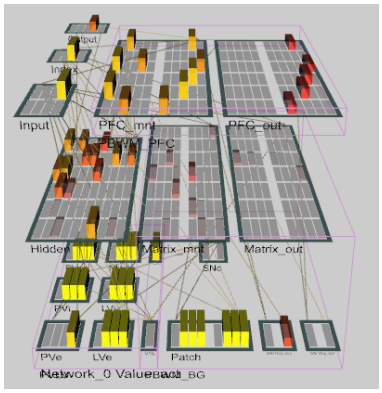
Table 2. Class distributions across datasets used to study training and generalization performance.

	Even distribution across scene classes and non-target	Even distribution across scene classes, more non-target	Distributed by expected difficulty to learn
1	20%	12.5%	4%
2	20%	12.5%	11%
3	20%	12.5%	11%
4	20%	12.5%	24%
Non-target	20%	50%	50%

1.2. Network

PBWM networks were constructed in the Emergent simulation environment [11] with 4 stripes—2 for maintenance and 2 for output. These stripes comprise the working memory component of the model. Output stripes are meant to model the immediate use of a symbol and its removal from working memory, while maintenance stripes hold symbols for longer time frames. Each stripe was 5 by 6 neurons in size. Maintenance stripes received connections from input fields and were subject to adaptation through error driven learning. Output stripes received connections from the maintenance stripes only. A hidden layer of 10 by 6 neurons received connections from all PFC stripes and the input fields. While PFC maintains a working memory of previously seen objects,

the hidden layer maps the contents of PFC, along with the current input to a motor response. This motor response corresponds to identifying a scene category, such as with a button press or vocalization. Activation in the network is shown in Figure 1a.

(a) 

(b)

Name	Input	Index	Output	RewT...
A				
X_C1				
A				
Y_C2				
B				
Z_C3				
C				
AA_C4				
D				
AA_NT				

Figure 1. (a) A PBWM network used to learn scene categories. (b) Input to the network representing the contents of each sector ("Input" column) and its location ("Index" column); other columns are described in the text. Here scene classes 1 through 4 are depicted followed by a non-target scene.

To represent the scenes to the network, input fields included both the object and the sector in which the object occurred. The former is meant to be the output of temporal cortex ("what") and was represented by a localist 2 by 4 input field. The latter is meant to correspond to the output of parietal cortex ("where"/ "how") and was represented by a localist 1 by 4 input field. Each object was represented with a single neuron set to be maximally active while all other neurons remained inactive. Symbol A, for example, had only the first unit on; symbol B had the second unit active and so on. Sector information was represented in a similar fashion where sector 1 had only the first neuron active, sector 2, the second neuron, and so on. Output categories were presented by a 1 by 5 output field, with one neuron per scene category, and one for non-target scenes. Again, scene category 1 had only the first neuron active, etc. With each foveation, a single object and its location was serially presented to the network with the category label and reward signal presented after the second foveation. (The reward signal tells the network when to answer, but its amplitude is dictated by the match between the prediction of the network and the ground truth label.) Inputs to the network are illustrated in Figure 1b.

During training, each network was trained until the sum of squared error (SSE) of the output field, averaged across all scenes, reached 0 for 2 consecutive epochs or 500 epochs was reached, whichever came first. An epoch contains all training scenes and the order of scenes was shuffled between epochs. Thirty random weight initializations, or batches, were run. Thus, a successful batch is one where SSE converges to 0 before 500 epochs. During testing, learning was turned off, and novel scenes were presented to determine generalization capability.

2. Results

First, PBWM networks can learn the scene categories without explicit programming as shown in Table 3. Across all datasets, approximately 70% of the networks were able to reach 0 training error within 500 epochs. Five hundred epochs was sufficient to

determine learning convergence as it was nearly 4 times as long as the average number of epochs in successful batches across all datasets (135.01). In testing, average percent correct across all successful weight initializations was 81.01%, which is far better than chance for 5 classes (20%).

Table 3. Training and generalization performance across datasets. Numbers within parenthesis are standard deviations.

	Even distribution across scene classes and non-target	Even distribution across scene classes, more non-target	Distributed by expected difficulty to learn	Average
Percentage of successful batches	56.67	73.33	73.33	67.78
Number of epochs in successful batches	157.88 (6.75)	126.59 (4.13)	120.57 (3.92)	135.01
Percent correct in successful batches	73.18 (6.75)	85.09 (4.13)	84.76 (3.92)	81.01

Looking at the generalization performance shows differences across the datasets. In the first dataset, the percentage of successful batches was 22% lower than that of the second and third datasets. Testing percent correct was also smaller than the other datasets ($t(58) = 8.36$, $p < .05$). This is because the first dataset did not have as many counter-examples to differentiate scene categories 1 through 4 from the non-target scenes. In the second dataset, the number of epochs did not change with respect to the first dataset ($t(37) = 1.24$, $p > .05$), but testing percent correct was higher ($t(37) = 6.81$, $p < .05$). Here, more counter-examples helped generalization. The second and third datasets do not differ in terms of percent correct ($t(41)=.27$, $p > .05$) or number of epochs ($t(41)=.29$, $p > .05$) as they share the same number of counter-examples. Similar training times across all datasets may reflect the limited capacity of the relatively small networks used.

3. Discussion and Conclusion

Using scene recognition as a challenge problem, PBWM networks are able of classifying sequences of symbols. However, the inputs and network structure used are generic enough to be applied to other domains that rely on working memory and categorization. Continuing in visual domain, this framework could be extended to behavior recognition where each scene is part of an action sequence (e.g., car left of gate, car at gate, car right of gate → car exit gate). Furthermore, the use of relational input encoding can introduce generality over objects or locations [12]. In addition, language involves processing chains of symbols at many scales: phonemes in words, words in sentences, sentences in paragraphs, etc. Navigation and motor sequence learning are other examples that could be modeled with this framework.

One limitation of this work is that the inputs and network structure used are relatively simple. Longer sequences, sequences with repeated symbols, or sequences with variable lengths would increase task difficulty. Similarly, more symbols, symbols with distributed representations, and the use of distracter symbols would stress network performance. The network structure could also be extended to include a functional segregation of PFC such as “what” inputs feeding into ventral regions, “where”/ “how”

inputs feeding into dorsal regions, and more anterior regions of PFC representing more abstract concepts or longer term working memory [13]. In the present work, "what" and "where" information was fused directly due to the simple nature of the input. Neuromodulatory systems, although connected with nearly every region of the brain, interact tremendously with PFC and anterior cingulate cortex (ACC) [14]. These interactions change working memory and categorization through short and long term dynamics respectively, and remain a ripe area for research.

Acknowledgement

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of the Interior (DOI) contract number D10PC20021. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. The views and conclusions contained hereon are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DOI, or the U.S. Government.

References

[1] C.F. Jacobsen, Functions of frontal association area in primates, *Archives of Neurology and Psychiatry* **33** (1935), 558-569.
[2] C.F. Jacobsen, Studies of cerebral function in primates. I. The functions of the frontal associations areas in monkeys, *Comparative Psychology Monographs* **13** (1936), 3-60.
[3] J.M. Fuster and G.E. Alexander, Neuron activity related to short term memory, *Science* **173** (1971), 652-654.
[4] K. Kubota and H. Niki, Prefrontal cortical unit activity and delayed alternation performance in monkeys, *Journal of Neurophysiology* **34** (1971), 337-347.
[5] J. D. Cohen, S. D. Forman, T.S. Braver, B.J. Casey, D. Servan-Schreiber, and D.C. Noll, Activation of the prefrontal cortex in a nonspatial working memory task with functional MRI, *Human Brain Mapping* **1** (1993), 293-304.
[6] R. Vogels, Categorization of complex visual images by rhesus monkeys, *European Journal of Neuroscience* **11** (1999), 1223-1238.
[7] D.J. Freedman, M. Risenhuber, T. Poggio, and E.K. Miller, A comparison of primate prefrontal and inferior temporal cortex during visual categorization, *Journal of Neuroscience* **23** (2003), 5235-5246.
[8] R.C. O'Reilly and M.J. Frank, Making working memory work: a computational model of learning in the frontal cortex and basal ganglia, *Neural Computation* **18** (2006), 283-328.
[9] G.A. Carpenter, S. Martens, and O.J. Ogas, Self-organizing information fusion and hierarchical knowledge discovery: a new framework using ARTMAP neural networks, *Neural Networks* **18** (2005), 287-295.
[10] B.J. Rhodes, Taxonomic knowledge structure discovery from imagery-based data using the neural associative incremental learning (NAIL) algorithm, *Information Fusion* **8** (2007), 295-315.
[11] B. Aisa, B. Mingus, and R.C. O'Reilly, The emergent neural modeling system, *Neural Networks* **21** (2008), 1146-1152.
[12] R.C. O'Reilly and R.S. Busby, Generalizable relational binding from coarse-coded distributed representations. In *Advances in Neural Information Processing Systems (NIPS) 14*, T.G. Dietterich, S. Becker and Z. Ghahramani (Eds.), Cambridge, MA; MIT Press (2002).
[13] R.C. O'Reilly, The what and how of prefrontal cortical organization, *Trends in Neurosciences* **33** (2010), 355-361.
[14] J.L. Krichmar, The neuromodulatory system - a framework for survival and adaptive behavior in a challenging world, *Adaptive Behavior* **16** (2008), 385-399.