Resolving the tension between exemplar and structure learning through a mixture-of-experts model of hippocampal-mPFC interaction

Dhairyya Singh (dsin@sas.upenn.edu)
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Ashley B. Williams (ashbwill@sas.upenn.edu)
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Anna C. Schapiro (aschapir@sas.upenn.edu)
University of Pennsylvania, Philadelphia, Pennsylvania, USA

Abstract

The hippocampus has been strongly implicated in both taking snapshots of individual experiences and extracting common structure across these experiences — functions often in tension. Prior evidence suggests an anatomical division of labor: the trisynaptic pathway (TSP) employs pattern-separated representations that store episodes while the monosynaptic pathway (MSP) uses overlapping representations to support statistical learning. A fundamental mystery remains, however: how does the brain recruit the right representation at the right time? Medial prefrontal cortex (mPFC) has been proposed to exert control over hippocampal outputs. Here, we introduce a stimulus-computable mixture-of-experts system featuring MSP- and TSP-like neural network experts, along with an mPFC-inspired gating network that controls their outputs. The system performs exemplar recognition and categorization simultaneously, and learns to adaptively combine expert outputs. We found that joint training of the experts and the gating network is necessary and simple mixing is insufficient. This framework illustrates how mPFC control may harness hippocampal specialization to resolve opposing computational objectives.

Keywords: hippocampus, medial prefrontal cortex, mixture-of-experts, category learning, memory

Introduction

The hippocampus supports both episodic (Dickerson & Eichenbaum, 2010) and structure learning (Mack et al., 2018) — functions that impose opposing computational demands. C-HORSE, a biologically-grounded model of the hippocampus, provides an account of how the hippocampus could support both functions, assigning these distinct computational roles to separate pathways (Schapiro et al., 2017). The TSP is suited to building sparse pattern-separated representations that orthogonalize episodes, and the MSP to building overlapping distributed representations that extract structure. A critical puzzle remains, however — how does the brain adaptively deploy representations most useful for a given task? Prior work has proposed mPFC-hippocampal interactions as a candidate control mechanism, facilitating adaptive behavior that is responsive to task demands (Eichenbaum, 2017; Preston & Eichenbaum, 2013). We explore learning mechanisms that could allow the mPFC to exert this kind of task-specific control over hippocampal representations.

We propose a stimulus-computable Mixture-of-Experts system (MoE; Jacobs et al., 1991) composed of two experts imbued with properties of the MSP and TSP, and an mPFC-inspired output-gating network that controls their deployment. Trained end-to-end simultaneously on episodic and category learning tasks, the system (i) learns an optimal policy for engaging each expert based on task demands, and (ii) outperforms models lacking a learned gating mechanism, demonstrating the critical role of the gating mechanism and its co-adaptation with the experts during training. Together, these findings suggest how an mPFC-like gating mechanism enables the flexible use of complementary hippocampal representations, offering an account of how the brain balances specificity and generalization in memory-guided behavior.

Results

The MoE system **(Fig 1A)** comprised two single hidden layer neural network experts, each incorporating properties of hippocampal subfield circuitry. The TSP expert featured a larger hidden layer, sparse connectivity, and k-Winner-Take-All inhibition, while the MSP expert had a smaller hidden layer, full connectivity, and no inhibition. Both experts generated predictions for category and exemplar tasks. An mPFC-inspired gating network, consisting of a small hidden layer and two output units, produced task-specific mixture coefficients $\boldsymbol{\alpha_{cat}}$ and $\boldsymbol{\alpha_{exem}}$. For each input, the final prediction for each task was computed as the linear combination of expert predictions: $\alpha_{task}MSP_{pred} + (1 - \alpha_{task})TSP_{pred}$.

The system was trained end-to-end, with parameters updated with respect to the summed loss across tasks. Post-training MSP representations were more sensitive to category structure than those of the TSP, consistent with their distinct architectural biases **(Fig 1B)**.

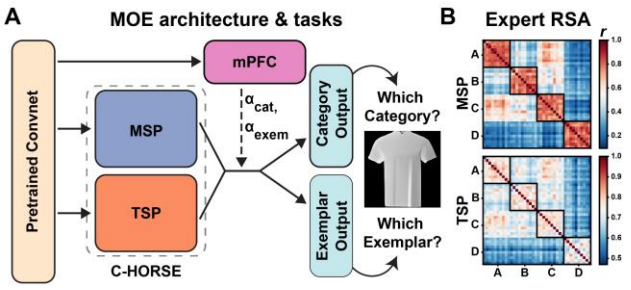

Figure 1: Mixture-of-experts system architecture, tasks, and representations. **(A)** The system included MSP and TSP experts and mPFC-like gating. Inputs were extracted from the decoder-avgpool layer of CORnet-S, a CNN pretrained on ImageNet and designed to approximate the primate ventral visual stream (Kubilius et al., 2019), and passed to both experts and the gating network. The gating network generated mixture coefficients to compute final weighted predictions from expert outputs. Training was performed for 500 epochs on a 100-image subset of Fashion-MNIST (10 per category; Xiao et al., 2017), with each image presented once per epoch. **(B)** Post- training MSP and TSP hidden-layer representations for a subset of categories.

With training, the MSP expert specialized in the category task and the TSP expert in the exemplar task **(Fig 2A)**. The MoE system leveraged these specializations via mPFC gating, assigning higher (MSP-favoring) $\boldsymbol{\alpha_{cat}}$ values for the category task and lower (TSP-favoring) $\boldsymbol{\alpha_{exem}}$ values for the exemplar task **(Fig 2B)**. The system also engaged the TSP on the category task, consistent with prior findings (Heffernan et al., 2021; Sučević & Schapiro, 2023), allowing it to outperform MSP alone. To probe gating behavior, we examined how **α** values varied with image-level similarity. There was a positive correlation between an image's assigned $\boldsymbol{\alpha_{cat}}$ and its similarity to other category members ($\rho = .081$, $W = 1497$, $p < .001$), suggesting that MSP is favored for category-consistent inputs, while TSP is recruited for atypical ones. Conversely, $\boldsymbol{\alpha_{exem}}$ was negatively correlated with similarity to all images ($\rho = -.188$, $W = 803$, $p < .001$), indicating greater TSP involvement when fine-grained discrimination is needed.

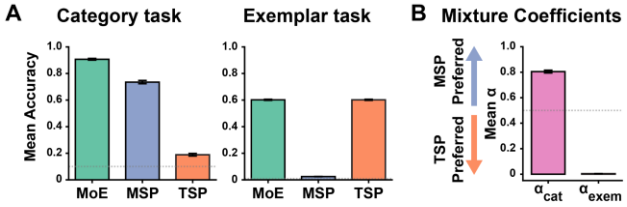

Figure 2: Task performance and learned mixture coefficients. **(A)** Performance on the category and exemplar tasks for the MoE system and individual MSP and TSP experts. Dashed line indicates chance. **(B)** Learned mPFC mixture coefficients for the category and exemplar tasks. Dashed line indicates **α** = 0.5. Error bars represent +/- SEM across 100 network initializations.

Finally, to assess the importance of adaptive gating, we compared our end-to-end system to two alternatives: (i) a system with fixed mixture coefficients (**α** = 0.5), and (ii) independently trained experts combined via equal-weight ensembling at test. Both alternative systems showed degraded category task performance **(Fig 3A)** and impaired generalization to held-out images **(Fig 3B)**, demonstrating that simple mixing is insufficient — adaptive gating must be learned alongside the experts.

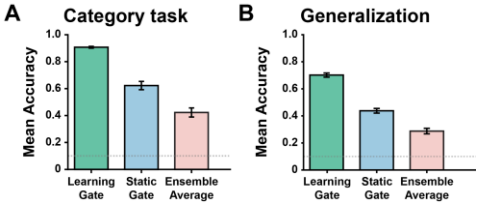

Figure 3: Performance on the **(A)** category task and **(B)** generalization across model variants with learned gating, static gating, or ensemble averaging.

Conclusion

Our model offers an account of how the mPFC could learn to flexibly coordinate complementary hippocampal pathways to support task-appropriate memory use. We demonstrate that the mPFC benefits from learning its gating concurrently with current task learning, though in the real environment it is possible that prior relevant tasks could contribute to training the gating relationship. This work provides a solution to how the brain may resolve competing demands on memory through learned control over hippocampal representations.

References

Dickerson, B. C., & Eichenbaum, H. (2010). The episodic memory system: Neurocircuitry and disorders. *Neuropsychopharmacology*, *35*(1), 86–104. https://doi.org/10.1038/npp.2009.126

Eichenbaum, H. (2017). Prefrontal–hippocampal interactions in episodic memory. *Nature Reviews Neuroscience*, *18*(9), 547–558. https://doi.org/10.1038/nrn.2017.74

Heffernan, E. M., Schlichting, M. L., & Mack, M. L. (2021). Learning exceptions to the rule in human and model via hippocampal encoding. *Scientific Reports*, *11*(1), 21429. https://doi.org/10.1038/s41598-021-00864-9

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*(1), 79–87. https://doi.org/10.1162/neco.1991.3.1.79

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *Advances in Neural Information Processing Systems*, *32*. https://papers.nips.cc/paper_files/paper/2019/hash/7813d1590d28a7dd372ad54b5d29d033-Abstract.html

Mack, M. L., Love, B. C., & Preston, A. R. (2018). Building concepts one episode at a time: The hippocampus and concept formation. *Neuroscience Letters*, *680*, 31–38. https://doi.org/10.1016/j.neulet.2017.07.061

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764–R773. https://doi.org/10.1016/j.cub.2013.05.041

Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *eLife*, *12*, e77185. https://doi.org/10.7554/eLife.77185

Xiao, H., Rasul, K., & Vollgraf, R. (2017). *Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms* (arXiv:1708.07747). arXiv. https://doi.org/10.48550/arXiv.1708.07747