Competitive Retraining Reveals the Resilience and Reallocation of Functional Specialization in Deep Neural Networks

Zhengqing Miao (zhengqing.miao@uni-giessen.de)
Department of Mathematics and Computer Science, Physics, Geography
Justus Liebig University Giessen, Giessen, 35392, Germany

Katharina Dobs (katharina.dobs@uni-giessen.de)
Department of Mathematics and Computer Science, Physics, Geography
Justus Liebig University Giessen , Giessen, 35392, Germany
Center for Mind, Brain and Behavior
Universities of Marburg, Giessen, and Darmstadt, Marburg, 35032, Germany

Abstract

Cognitive neuroscientists have long documented functional specialization in the brain for tasks such as face, body or scene recognition, and recent computational studies reveal that deep neural networks (DNNs) spontaneously develop specialized populations of units for the same tasks. But are these specialized units necessary for performance, and how plastic are they? Here, we combine lesioning approaches with competitive retraining in DNNs to address these questions. In a dual-task network with localized specialized units in the final convolutional layer for face and object tasks, we ablated those units either at the onset or continuously throughout retraining. To modulate competition, we retrained the networks on a single task or both tasks simultaneously. Our findings reveal that retraining restores network performance even when these layer-specific units remain permanently disrupted, indicating they are not strictly necessary. Moreover, the extent and pattern of unit reallocation vary with retraining conditions, demonstrating substantial plasticity and suggesting that the reallocation process is an intrinsic outcome of rapid network optimization.

Keywords: Functional specificity; Task optimization; Competitive training; Lesioning; Cortical recycling

Introduction

Functional specialization is a well-established phenomenon in the human brain. For instance, distinct regions in the visual cortex respond to specific perceptual tasks, and their disruption leads to corresponding selective deficits (Kanwisher, 2010; Pitcher et al., 2009; Moscovitch et al., 1997). Understanding how and why such specialization develops is fundamental to both neuroscience and artificial intelligence. Intriguingly, deep neural networks have been shown to develop analogous specialized populations of units for tasks such as face, object, and scene recognition (Blauch et al., 2022; Dobs et al., 2022; Prince et al., 2024). Yet, it remains unclear whether these specialized units are necessary for performance and how plastic they are. In particular, do they emerge only as an efficient solution when training from scratch, or can they persist, and even reemerge, after disruption?

To address these questions, we used a dual-task network trained for face and object recognition that developed distinct specialized populations, especially pronounced in the last convolutional layer, for each task, as validated by lesioning (Dobs et al., 2022) (Fig. 1). We probed the mechanisms of this functional specialization by combining ablation of those units ("lesion init" at retraining onset or "lesion always" throughout) with varied competitive constraints. Specifically, we selectively ablated task-specific units at the onset or continuously, then retrained under single- or dual-task conditions. By evaluating performance recovery and shifts in the contributions of remaining units, we tested whether functional specialization arises intrinsically from rapid network optimization and assessed its resilience to disruption.

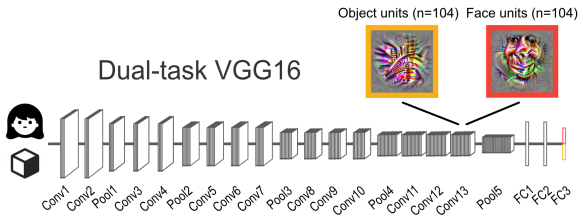

Figure 1: Schematic diagram of dual-task-trained VGG16

Methods

Datasets and Baseline Model. All experiments were conducted using a VGG16-based (Simonyan & Zisserman, 2014) model pretrained on two tasks (Dobs et al., 2022): (1) face recognition using the VGGFace2 dataset (Cao et al., 2018) (1,714 identities) and (2) object recognition using the ImageNet-2012 dataset (Deng et al., 2009) (423 categories). In this model, the top 20% of filters in the last convolutional layer, whose lesion most impaired performance on each task, were identified as face- or object-specific units (Fig. 1).

Retraining Procedure. For consistency, we used the same pretrained VGG16 checkpoint, optimizer, datasets and training strategy as in Dobs et al. (2022) during retraining. To vary competition, the network was retrained on either a single task (face or object recognition) or on both tasks simultaneously (dual-task). We employed two lesioning approaches during retraining: (1) *Lesion at Initialization*, in which units are ablated only at the beginning of retraining (allowing for potential recovery), and (2) *Lesion Always*, in which units are ablated continuously throughout retraining (Miao & Zhao, 2025).

Performance and Reallocation Assessment. After retraining, we evaluated the accuracy of the network (for both single and dual tasks) and compared it to the original network's baseline performance by computing the mean difference and its standard error (SE(Δ)) across classes. Additionally, we measured the extent of reallocation of task-specific units by comparing the impact of lesioning the original face- and object-specific units after retraining to their baseline impact in the original network.

Results

Lesioning and Performance Resilience

How much does the network's performance depend on task-specific units, and how resilient is it to their disruption? Lesioning face units in the original model dramatically impaired face recognition performance, dropping accuracy from 92.54% to 30.45% (lesion impact: 59.8%). However, our lesioning experiments showed that retraining on the face task restored or surpassed baseline performance, regardless of whether the face units were ablated only at onset or continuously throughout retraining (Fig. 2a). Task performance declined only if the task was excluded from retraining (left yellow bar in Fig. 2a). Even with permanent ablation of face units, or both face and object units, the network fully recovered after retraining

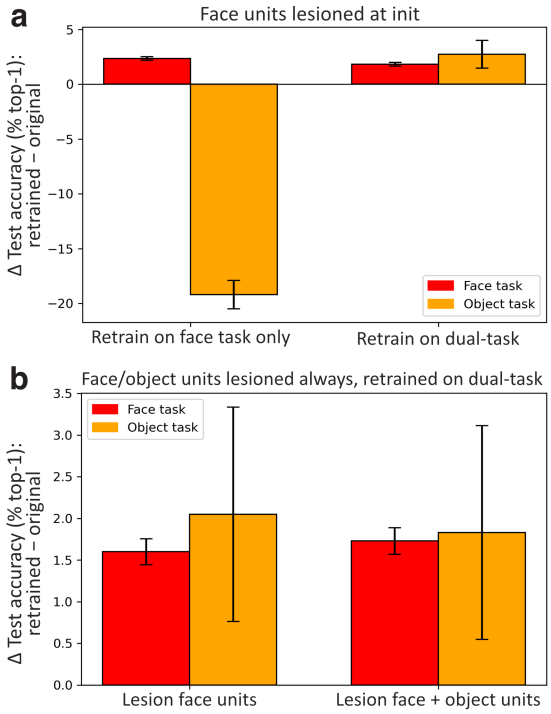

Figure 2: Test accuracy differences (% top-1) for retrained models relative to the original dual-task model. Baseline (0) corresponds to original model's accuracy; positive values indicate improved accuracy after retraining, negative values indicate a decline. **(a)** Retraining with face units ablated at initialization on the face task only (left bars) or on dual-task (right bars). **(b)** Dual-task retraining with continious ablation of face units (left bars) or of both face and object units (right bars). Error bars denote the standard error of the difference across classes.

(Fig. 2b). These results suggest that the network's task performance is not strictly dependent on the originally identified task-specific units, but can flexibly recover. This raises the critical question: does the network recycle the original task-specific units, develop new specialized units during retraining, or simply solve the task without relying on specialized units?

Lesioning and Task-Specific Unit Reallocation

To find out, we manipulated competition during retraining by training either on single or dual tasks. We then measured lesion impacts when ablating the same task-specific units relative to the original baseline (face task: 92.54%-30.45%=59.80%; object task: 52.34%-21.19%=31.15%; baseline in Fig. 3). Ablating object units after retraining produced a similar impairment as in the original model (yellow left and middle bars in Fig. 3), suggesting that object units are extensively recycled during retraining. In contrast, face units appear to be recycled more robustly under more competitive (dual-task) retraining conditions (red left and middle bars in Fig. 3). This may be because object recognition presents a more challenging optimization problem, making recycling the original object units the fastest route to recovery. In simpler (single task) retraining scenarios, networks may settle into local optima, while intense competition forces a greater reliance on task-specific units. Moreover, lesioning original object units after retraining with permanently ablated face units showed a similar impact on the object task (and vice versa for face units; right bars in Fig. 3). These results suggest that when networks recover their performance after continuous lesioning of face units (Fig. 2b), they do not recycle object units.

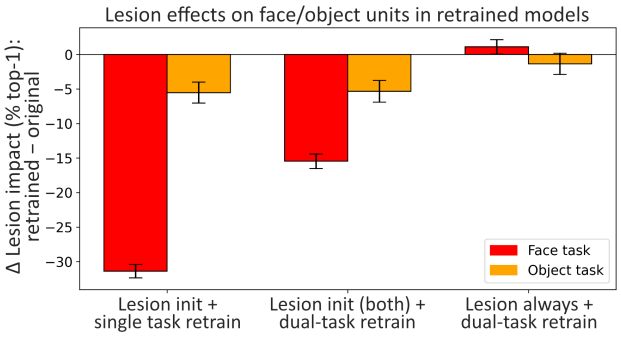

Figure 3: Lesion impact differences (% top-1) for ablating face/object units in original and retrained models. Baseline (0) corresponds to the original model's lesion impact. Negative values indicate a smaller lesioning impact in retrained models, positive values indicate an increase. **Left bars**: Lesioning face/object units at initialization, followed by retraining on the corresponding task only. **Middle bars**: Lesioning Both face and object units at initialization, retrained on dual-task. **Right bars**: Continuous ablation of face or object units during dual-task retraining, then lesion-testing the other units.

Discussion

We found that even when both face and object units are permanently disrupted, the network can recover its original performance through retraining and flexible unit reallocation. Moreover, the degree of recycling after disruption depends on task difficulty and competition during optimization, with more intensive recycling observed for challenging tasks or under strong competitive conditions. These findings suggest that, as in DNNs, functional specialization in the brain may arise as an efficient, adaptive response to task demands. However, while our models exhibit robust plasticity, this capacity may diminish in brains as development progresses (Dehaene & Cohen, 2011).

A key limitation is that we did not identify which units become newly face-selective post-retraining, so we cannot confirm whether recovery relies on comparably specialized units or a more distributed representations. Future work should lesion face-selective units across layers to test recovery limits, and examine generalization to other architectures. Such analyses will clarify the nature of recovered pathways, inform cortical recycling mechanisms, and help delineate the boundaries of plasticity in the human brain.

Acknowledgments

K.D. was supported by the ERC Starting Grant DEEPFUNC (ERC-2023-STG-101117441), the Hessian Ministry of Higher Education, Research, Science and the Arts (LOEWE Start Professorship and Excellence Program “The Adaptive Mind”), and the German Research Foundation (Collaborative Research Center SFB/TRR 135).

References

Blauch, N. M., Behrmann, M., & Plaut, D. C. (2022). A connectivity-constrained computational account of topographic organization in primate high-level visual cortex. *Proceedings of the National Academy of Sciences*, *119*(3), e2112566119. doi: 10.1073/pnas.2112566119

Cao, Q., Shen, L., Xie, W., Parkhi, O. M., & Zisserman, A. (2018). Vggface2: A dataset for recognising faces across pose and age. In *2018 13th ieee international conference on automatic face & gesture recognition (fg 2018)* (pp. 67–74).

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in cognitive sciences*, *15*(6), 254–262.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 ieee conference on computer vision and pattern recognition* (pp. 248–255).

Dobs, K., Martinez, J., Kell, A. J., & Kanwisher, N. (2022). Brain-like functional specialization emerges spontaneously in deep neural networks. *Science advances*, *8*(11), eabl8913.

Kanwisher, N. (2010). Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, *107*(25), 11163–11170.

Miao, Z., & Zhao, M. (2025). Weight-freezing: A motor imagery inspired regularization approach for eeg classification. *Biomedical Signal Processing and Control*, *100*, 107015.

Moscovitch, M., Winocur, G., & Behrmann, M. (1997). What is special about face recognition? nineteen experiments on a person with visual object agnosia and dyslexia but normal face recognition. *Journal of cognitive neuroscience*, *9*(5), 555–604.

Pitcher, D., Charles, L., Devlin, J. T., Walsh, V., & Duchaine, B. (2009). Triple dissociation of faces, bodies, and objects in extrastriate cortex. *Current Biology*, *19*(4), 319–324.

Prince, J. S., Alvarez, G. A., & Konkle, T. (2024). Contrastive learning explains the emergence and function of visual category-selective regions. *Science Advances*, *10*(39), eadl1776.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.