Inherent Receptive Fields in the Early Layer Enable Continual Learning Under Dynamic Environments

Minjun Kang (4401kmj@kaist.ac.kr)
Department of Brain and Cognitive Sciences
Korea Advanced Institute of Science and Technology
Daejeon, 34141, Republic of Korea

Seungdae Baek (seung7649@kaist.ac.kr)
Department of Bio and Brain Engineering
Korea Advanced Institute of Science and Technology
Daejeon, 34141, Republic of Korea

Se-Bum Paik (sbpaik@kaist.ac.kr)
Department of Brain and Cognitive Sciences
Korea Advanced Institute of Science and Technology
Daejeon, 34141, Republic of Korea

Abstract

Continuously learning new information is a fundamental ability of animals but a challenging problem for conventional deep neural networks (DNNs), which suffer from catastrophic forgetting. Unlike DNNs, whose early layers change depending on training images, the brain's early visual pathway has innate Gabor-like receptive fields that are stably maintained throughout a lifetime. Here, we demonstrate that fixing early layers of DNNs using Gabor filters, resembling the primary visual cortex (V1) cells' receptive fields, enables continual learning under dynamic environments. We first showed that networks with fixed Gabor filters maintained the previous performance even when sequentially trained on a completely different image domain, alleviating catastrophic forgetting. Moreover, representation analysis revealed that fixed Gabor filters enabled networks to have similar representations across different domains, which may enable networks to adapt better to continuous learning. Together, Gabor filters in early layers could serve as key architectures for continual learning, highlighting the functional significance of stable early visual pathways in brains.

Keywords: Deep neural network; Continual learning; Primary visual cortex; Object recognition

Introduction

Continuously learning new information is a fundamental ability for animals (Kudithipudi et al., 2022) but a challenging problem for conventional deep neural networks (DNNs). When DNNs are trained on images with different distributions, they lose previous performance, referred to as catastrophic forgetting (McCloskey & Cohen, 1989). However, the mechanisms responsible for such a functional difference between the two systems remain elusive.

An important clue may lie in the anatomical distinction between brains and DNNs. Specifically, DNNs' early layers are randomly initialized (He, Zhang, Ren, & Sun, 2015) and trained depending on the dataset (Krizhevsky, Sutskever, & Hinton, 2012). This can be an unfavorable characteristic for continual learning since subtle changes in the early layer may lead networks to forget prior information (McCloskey & Cohen, 1989). On the other hand, the brain's early visual pathway exhibits Gabor-like receptive fields (Figure 1) even before eye-opening (Gödecke & Bonhoeffer, 1996; Niell & Stryker, 2008; Paik & Ringach, 2011; Song, Jang, Kim, & Paik, 2021), which tends to remain stable throughout visual experiences (Gödecke, Kim, Bonhoeffer, & Singer, 1997; Crist, Li, & Gilbert, 2001). Thus, the early visual pathway serves as a common basis for processing various images during a lifetime.

Here, we hypothesized that the fixed early visual circuitry in the visual system facilitates robust continual learning. To test this, we applied fixed Gabor filters in the early layer of the convolutional DNN and sequentially trained them when the image domain changes. As a result, we showed that networks with fixed Gabor filters maintained previous performance and representations under dynamic environments.

Our model

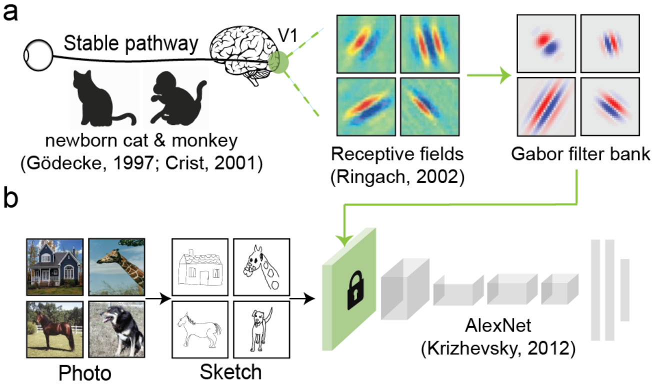

Figure 1: Our model (DNN with fixed Gabor filters). (a) Modeling V1 receptive fields as Gabor filters (b) Incorporating fixed Gabor filters in the first layer and sequentially trained networks

To explore the functional role of a stable early visual pathway, we modeled receptive fields of V1 neurons (Ringach, 2002) as Gabor filters (Figure 1a). Subsequently, we introduced them as filters of the first convolutional layer of DNN (Krizhevsky et al., 2012) to simulate the human visual systems (Figure 1b). We then trained networks using images whose domain changes over time. Particularly, we used a public dataset (Li, Yang, Song, & Hospedales, 2017) containing seven common classes with four distinct domains: Photo, Art, Cartoon, and Sketch.

Results

Fixed Gabor filters enable networks to maintain previous performance

We first applied sequentially training using Photo and Sketch domains, which show the highest difference in low-level statistics such as spatial frequency spectrum. We showed that fixed Gabor filters enable networks to maintain previous performance under a domain change (Figure 2). For conventional DNNs, when networks were sequentially trained on a different domain (Sketch), the DNNs entirely lost the first domain's performance (Photo) to the chance level, demonstrating catastrophic forgetting (Figure 2b, top; one sample t-test, NS: P=0.07). On the contrary, DNNs equipped with fixed Gabor filters robustly maintained their previous performance (Figure 2b, bottom; one sample t-test, ***: P<0.001). The accuracy loss of the Photo images was significantly lower in DNN with fixed Gabor filters than in conventional DNNs (Figure 2d; paired t-test, ***: P<0.001).

Then, we applied the same sequential training using all domain pairs to confirm the above is not a dataset-specific effect, applicable only to Photo and Sketch images. We trained networks using 12 unique pairs among four domains and measured the accuracy loss for each pair. As a result, networks with fixed Gabor filters better maintained the initial perfor-

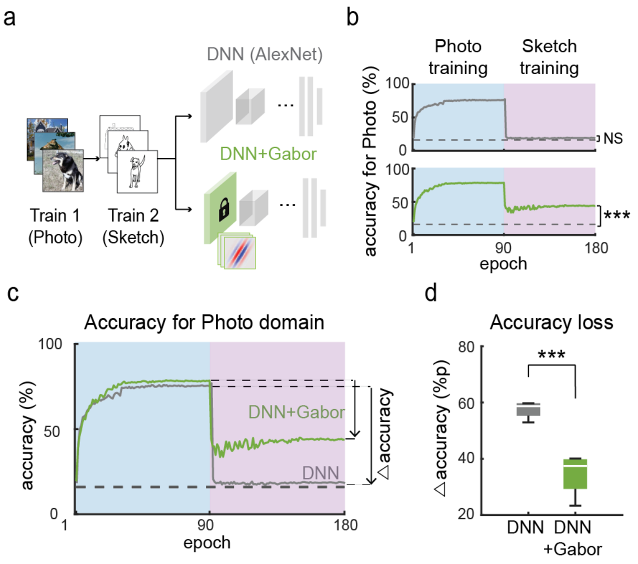

Figure 2: DNN+Gabor maintains previous performance (a) Sequential training of DNN and DNN+Gabor (b) Accuracy for Photo images of both networks (c) Different accuracy maintenance of DNN and DNN+Gabor (d) Smaller accuracy drop in DNN+Gabor

mance for most domain pairs (paired t-test, $P<0.05$ for nine pairs).

Fixed Gabor filters enable networks to have similar representations across different domains

To explain how our model could maintain the performance, we analyzed representations of Photo and Sketch images (Figure 3). We found that fixed Gabor filters enable networks to have similar representations across different domains. Initially, we measured the activations of the first fully connected layer (fc6), which is known to encode categorical information (Bao, She, McGill, & Tsao, 2020), of DNN and DNN with fixed Gabor filters by feeding Photo and Sketch images. Then, we conducted dimension reduction using t-SNE (Figure 3a). We hypothesized that the same classes across different domains (e.g., Photo dogs and Sketch dogs) would have similar representations clustered together in the latent space of DNNs with fixed Gabor filters but not of DNNs (Figure 3b). We expected this clustering would help DNNs with fixed Gabor filters to maintain their previous parameters and performance in continual learning.

As a result, we noticed that the same classes, represented as identical colors, tend to cluster better in the latent space of DNN with fixed Gabor filters than of DNN (Figure 3d). We then evaluated this clustering tendency using the silhouette index (Kaufman & Rousseeuw, 2009), in which higher values indicate better clustering (Figure 3c). We showed that DNNs with fixed Gabor filters demonstrated a higher silhouette index than DNNs, confirming the better clustering of classes (Figure 3e; paired t-test, ***: $P<0.001$). These results imply that the same classes have similar representations in our model, which can eventually lead to better adaptation in continuous learning scenarios.

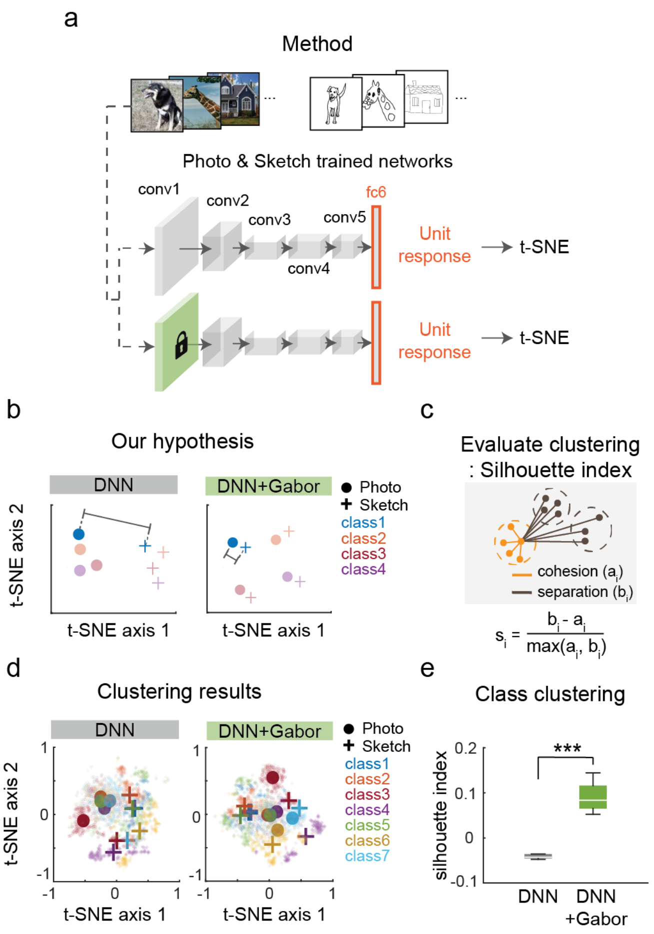

Figure 3: Representations of different domains and classes (a) Dimension reduction using t-SNE of activations of fc6 layer (b) Our hypothesis that classes will be clustered in DNN+Gabor (c) Evaluating clustering with Silhouette index (d) Clustering results for classes across different domains (e) DNN+Gabor shows better clustering of same classes across different domains

Conclusion

In summary, we showed that fixed Gabor filters enable robust continual learning by maintaining the performance of an initial domain after training a new domain. We also showed that this continuous learning ability may be caused by generating similar representations across various image domains in our model. These results highlight the importance of fixed early layers in continuous learning scenarios and underscore the functional significance of hard-wired early visual pathways in brains.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korean government (MSIT) (No. NRF-2022R1A2C3008991 to S.-B.P.) and by the Singularity Professor Research Project of KAIST (to S.-B.P.).

References

Bao, P., She, L., McGill, M., & Tsao, D. Y. (2020). A map of object space in primate inferotemporal cortex. *Nature*, *583*(7814), 103–108.

Crist, R. E., Li, W., & Gilbert, C. D. (2001). Learning to see: experience and attention in primary visual cortex. *Nature neuroscience*, *4*(5), 519–525.

Gödecke, I., & Bonhoeffer, T. (1996). Development of identical orientation maps for two eyes without common visual experience. *Nature*, *379*(6562), 251–254. doi: 10.1038/379251a0

Gödecke, I., Kim, D. S., Bonhoeffer, T., & Singer, W. (1997). Development of orientation preference maps in area 18 of kitten visual cortex. *European Journal of Neuroscience*, *9*(8), 1754–1762. doi: 10.1111/j.1460-9568.1997.tb01533.x

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the ieee international conference on computer vision* (pp. 1026–1034).

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

Kudithipudi, D., Aguilar-Simon, M., Babb, J., Bazhenov, M., Blackiston, D., Bongard, J., … Siegelmann, H. (2022). Biological underpinnings for lifelong learning machines. *Nature Machine Intelligence*, *4*(3), 196–210. doi: 10.1038/s42256-022-00452-0

Li, D., Yang, Y., Song, Y. Z., & Hospedales, T. M. (2017). Deeper, Broader and Artier Domain Generalization. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 5543–5551. doi: 10.1109/ICCV.2017.591

McCloskey, M., & Cohen, N. J. (1989, 1). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. *Psychology of Learning and Motivation - Advances in Research and Theory*, *24*(C), 109–165. doi: 10.1016/S0079-7421(08)60536-8

Niell, C. M., & Stryker, M. P. (2008). Highly selective receptive fields in mouse visual cortex. *Journal of Neuroscience*, *28*(30), 7520–7536. doi: 10.1523/JNEUROSCI.0623-08.2008

Paik, S.-B., & Ringach, D. L. (2011). Retinal origin of orientation maps in visual cortex. *Nature neuroscience*, *14*(7), 919–925.

Ringach, D. L. (2002). Spatial structure and symmetry of simple-cell receptive fields in macaque primary visual cortex. *Journal of Neurophysiology*, *88*(1), 455–463. doi: 10.1152/jn.2002.88.1.455

Song, M., Jang, J., Kim, G., & Paik, S.-B. (2021). Projection of orthogonal tiling from the retina to the visual cortex. *Cell Reports*, *34*(1).