Degradation and plasticity in convolutional neural networks: An investigation of internal representations

Jasmine A. Moore
Biomedical Engineering Graduate Program
University of Calgary
Calgary, AB T2N1N4
jasmine.moore@ucalgary.ca

Vibujithan Vigneshwaran
Department of Radiology
University of Calgary
Calgary, AB T2N1N4
vibujithan.vigneshwa@ucalgary.ca

Matthias Wilms
Department of Pediatrics
University of Calgary
Calgary, AB T2N1N4
matthias.wilms@ucalgary.ca

Nils D. Forkert
Department of Radiology
University of Calgary
Calgary, AB T2N1N4
nils.forkert@ucalgary.ca

Abstract

The architecture and information processing of convolutional neural networks was originally heavily inspired by the biological visual system. In this work, we make use of these similarities to create an in silico model of neurodegenerative diseases affecting the visual system. We examine layer-wise internal representations and accuracy levels of the model as it is subjected to synaptic decay and retraining to investigate if it is possible to capture a biologically realistic profile of visual cognitive decline. Therefore, we progressively decay and freeze model synapses in a highly compressed model trained for object recognition. Between each iteration of progressive model degradation, we retrain the remaining unaffected synapses on subsets of initial training data to simulate continual neuroplasticity. The results of this work show that even with high levels of synaptic decay and limited retraining data, the model is able to regain internal representations similar to that of the unaffected, healthy model. We also demonstrate that throughout a complete cycle of model degradation, the early layers of the model retain high levels of centered kernel alignment similarity, while later layers containing high-level information are much more susceptible to deviate from the healthy model.

1 Introduction

Deep learning models were originally inspired by the hierarchical organization of the brain's neural circuits, where lower-level features are progressively combined to form higher-level abstractions [1]. Convolutional neural networks (CNNs) have proven effective in computer vision tasks by using local receptive fields, which resembles the organization of visual processing in the mammalian visual cortex. The intersection of neuroscience and deep learning is a rapidly evolving field that seeks to leverage insights from the brain's structure and functioning to enhance the design and performance of deep learning algorithms. However, in recent years research has begun to explore the reverse direction of this relationship aiming to improve our understanding about the brain by using deep learning models[2, 3, 4, 5]. Not only has deep learning been used to draw parallels to the healthy brain, but also to simulate the cognitive deficits that accompany neurodegenerative diseases by imposing axonal or neuronal injury to deep learning models [6, 7]. This study uses the established similarities between CNNs and the biological visual system to simulate the progression of neurdegenerative

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

disease, for example, Alzheimer's disease (AD). In AD, one of the hallmark pathological features is the abnormal accumulation of tau protein in the form of neurofibrillary tangles. These tangles disrupt synaptic function, compromise synaptic plasticity, and contribute to cognitive decline. These disruptions contribute to the cognitive deficits observed in individuals with AD, such as impaired visual cognition, memory impairment, and executive function [8, 9, 10]. In this work, we simulate the effects of the accumulation of tau in the visual cortex by progressively decaying an increasing number of model synapses (weights) in conjunction with a biologically realistic retraining method (Fig. 1). The progressive weight decay is motivated by the synaptic degradation and impaired synaptic transmission caused by neurofibrillary tangles [11, 12]. Furthermore, we simulate neuroplasticity by retraining the model on small sets of original training data. Moreover, in this work, we examine the internal representations the model learns during continual weight decay and retraining. We investigate this by using centered kernel alignment (CKA) to investigate whether the model maintains similar representations during injury and retraining.

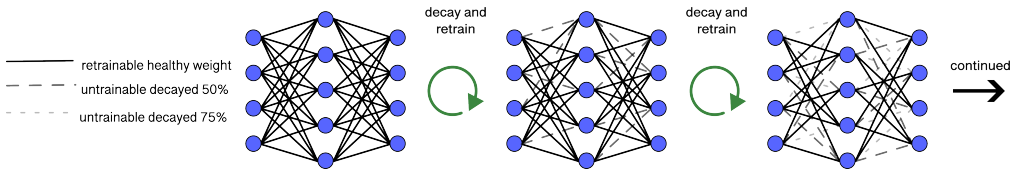

Figure 1: Pipeline of degradation to the CNN. Grey lines represent synapses that have been decayed. Only healthy, unaffected weights are subjected to retraining. All decayed weights remain frozen and will continue to decay with each iteration of model degeneration.

2 Methods

2.1 Models and data

The model used to simulate the 'healthy' visual system is based on the VGG-19 architecture trained on the CIFAR10 dataset [13, 14]. The initial train and test split used is 50,000/10,000 images, respectively. Previous studies have indicated that VGG-19 models exhibit a significant degree of over-parameterization, as evidenced by their ability to maintain high levels of accuracy even after undergoing significant pruning [15, 16]. This is attributed to an excessive number of weights (synapses) and neurons, leading to the learning of redundant pathways. In contrast, the brain, while also displaying some features of over-parameterization, operates under more stringent constraints such as energy consumption and physical space limitations [17, 18]. To account for potential spurious outcomes resulting from over-parameterization rather than the plasticity of the model, we performed experiments with a highly compressed version of a VGG19-like architecture. Therefore, compression of the original VGG-19 architecture was performed using structured pruning techniques based on graph dependencies as described by Li et al. and Fang et al. [19, 20]. More precisely, filters and their corresponding weights were simultaneously eliminated based on their L1 norm until the model's inference speed, measured in floating-point operations per second (FLOPS), was enhanced by a predefined margin. This regime led to a compressed model that only contained 8.54% of the original number of parameters. Therefore, all experiments were conducted using this highly optimized, more constrained version of VGG-19. This model has five convolutional blocks, each followed by a batch normalization layer. The final layers of the model consist of four max-pooling layers and a softmax activation with ten nodes corresponding to the ten classes in the dataset. Our model was pretrained on ImageNet and fine-tuned on CIFAR10 for 100 epochs with a learning rate of 0.001, using a batch size of 128, and a stochastic gradient descent optimizer with momentum 0.9.

2.2 Synaptic decay and retraining

To model tau accumulation, model synapses were decayed exponentially, as given in the following equation as $\sum_{i=1}^{N} \frac{1}{N} T\gamma^i$, where N is the number of iterations of injury that is applied to the model, T is the total number of model synapses to be subjected to atrophy, and γ is the decay constant by which weight values are multiplied by. In this study, T is the number of synapses in the convolutional and linear layers of the model, and we performed experiments with N=10 and γ=0.5. These values

were found to be representative of injury resolution while remaining computationally cost-effective. Therefore, in the first iteration of injury 10% model synapses were decayed by a factor of 0.5. In the second iteration of injury, an additional 10% of so-far unaffected model synapses were decayed by a factor of 0.5, while the initially decayed weights were decayed by another factor of 0.5. This process was repeated until all model weights given by T had been affected. After each iteration of injury, the remaining healthy weights in the model were retrained, while decayed weights remained frozen at the same value. This simulates the continual neuroplasticity of the brain, even when affected by disease. We investigated using retraining sets of data containing 5000, 1000, and 100 images randomly selected from the training set used for initial model training. In these subsets of data, the images are not balanced by class as to more biologically represent the continual learning of information processed by the biological visual cortex.

2.3 Centered kernel alignment

Centered kernel alignment (CKA) is an established method used to compare internal model representations of deep neural networks [21]. Though similarities between representations of deep neural networks have been measured in various ways, CKA is notably the most consistent and is one of the only methods that has succeeded in finding correlation in representations between models that are identical in architecture and training regimes but have different initialization seeds [22, 23, 24]. In this study, we utilize CKA to analyze if the model re-establishes similar representations of information as compared to the healthy model after being decayed and retrained. This provided insight and quantification as to whether the model is able to make use of its cognitive reserve and plasticity abilities to learn new pathways for the task of object recognition under increasing levels of injury.

3 Results

3.1 Accuracy

The performance of our model was systematically evaluated as it underwent ten iterations of synaptic decay and retraining. This iterative process aimed to investigate the impact of increasing synaptic decay on the model's accuracy and its ability to recover and adapt during the subsequent retraining process with different-sized subsets of retraining data. The initial, compressed model achieved an accuracy of 93.3% on the test dataset. After the first 10% of model synapses were decayed, the model performance was reduced by about 25% accuracy. Upon retraining, the three different training regimes revealed differences in accuracy improvements. When trained for one epoch with 5000 images, 1000 images, and 100 images, model accuracies improved to 83.0% $\pm$ 1.87%, 76.6% $\pm$ 6.06%, and 70.8% $\pm$ 21.8%, respectively. The large standard deviations in accuracy at high injury levels can be attributed to extreme synaptic sparsity that arises as most of the model is degenerated. If a salient weight gets degenerated and frozen, the sparse models may not be able to recover accuracy even with retraining. These findings seem biologically reasonable in that less training data leads to less of a regain of accuracy. However, even at high levels of injury (*i.e.,* 60% of synapses are decayed), we found that one epoch of retraining with 1000 images led to a drastic increase in performance from 15.0% $\pm$ 7.01% accuracy in its decayed state to 66.4% $\pm$ 6.12% accuracy after retraining. These trends continue throughout the full cycle of degeneration, although at injury levels of 70% and higher, the model with only 100 retraining images is unable to regain accuracy levels above chance.

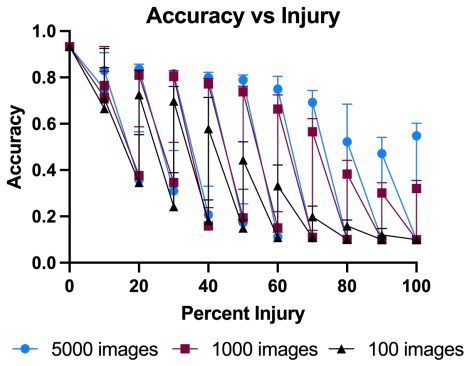

Figure 2: Model accuracy and standard deviation as an increasing number of weights are decayed, and the remaining 'healthy' weights are retrained. Accuracy curves are shown for the different subsets of retraining data. Each subset is randomly sampled from training data and is not balanced between classes. Average accuracies are reported for 10 different cycles of model degeneration.

3.2 Internal representations

While accuracy shows that the model is able to reorganize in terms of task performance, it does not provide any information if the model is regaining the same internal representations as the healthy model. Therefore, we investigated this aspect using CKA for each layer in the model as it went through the cycle of degeneration. These findings for the model retrained on the subset of 1000 images at four different levels of injury are provided in Figure 3. Even at high levels of model degeneration (*i.e.* 80% synapses have been decayed and frozen), the early layers in the model retain high representational similarity to those of the healthy model. Contrarily, the later layer representations appear to be more vulnerable to breaking down as injury is applied. When 60% of model weights have been decayed and frozen, layers 22-48 only retain CKA similarity values of 0.4 and below, as seen in the third panel of Figure 3C. However, even with only one epoch of retraining, these CKA values of later layers recover decent levels (about 0.6) of representational similarity. These patterns in results are consistent for all three retraining regimes, although more similarity to the healthy model is regained with the larger retraining subset.

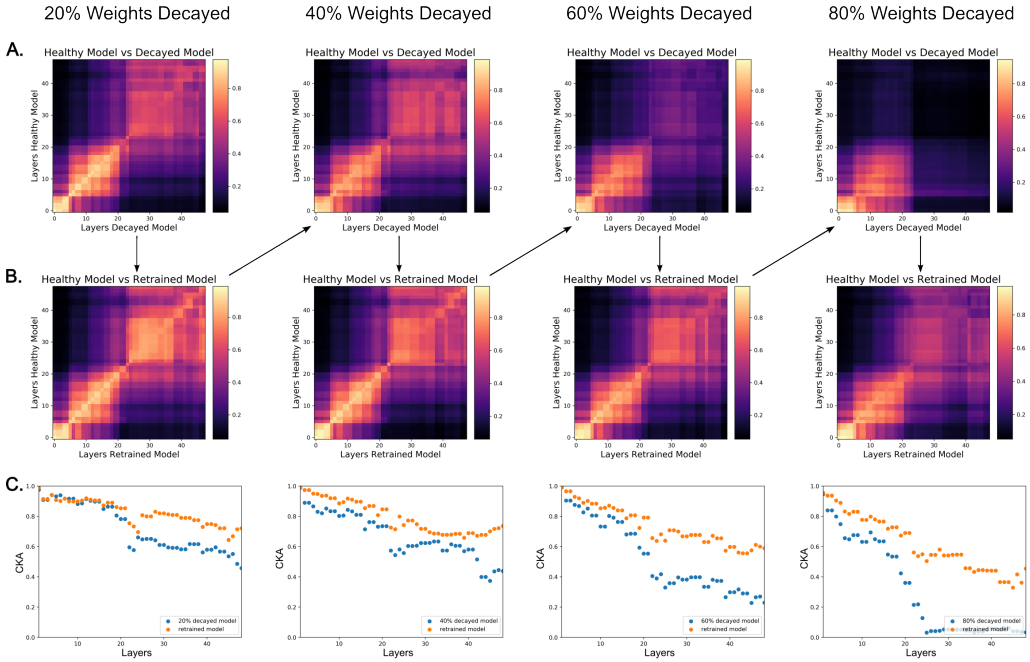

Figure 3: CKA matrices providing a layer-wise comparison of representations between the baseline model and the model with the retraining subset of 1000 images at increasing levels of synaptic decay and retraining. Row (A) shows CKA analysis between the healthy model and the decayed model. Row (B) shows CKA analysis between the healthy model and the retrained model. Row (C) displays the diagonal (1:1 layer comparison) CKA values for each injury level. Blue points denote the decayed model's layers, and orange points denote the layers after retraining.

4 Discussion

The results of this work show that, while weight decay led to initial reductions in accuracy, the model exhibited a degree of resilience and adaptability during retraining, even with limited amounts of unbalanced retraining data. These initial results raised the question if retraining results in the formation of new pathways for achieving the objective function, even as an increasing number of parameters are decayed and frozen. Alternatively, does retraining involve learning how to effectively transmit information through these compromised synapses? Monitoring the indices of the highest magnitude weights within various layers demonstrated that both scenarios appeared to be occurring. More precisely, we found that the indices of maximal values weights changed, and therefore a new pathway for information flow was utilized. However, we also observed occasionally that the index of maximal weights did not change, but the values of the weights themselves did. Thus, it can be inferred

that the model is learning new weight distributions to accomplish object recognition. Interestingly, even while the internal structure of the model changes, the same representations are recovered each time the model undergoes retraining. Therefore, internal representations seem not to rely as much on model structure but on the training task. Additionally, we found that the similarity of earlier layers remains largely unaffected even at higher levels of injury. This can be explained by locality in earlier layers when compared to more compressed global information in deeper layers, where representational similarity is essentially a function of model accuracy. It has been shown that the accumulation of tau leads to degradation and eventual atrophy of synapses, which can build for years before clinically presenting in cognitive decline [25]. It is assumed that the cognitive reserve and plasticity in the brain allow for compensation of these decaying pathways [26]. The results of this work show that CNNs can behave in similar ways by finding new pathways and scaling information to retain object recognition performance. Thus, we present a biologically inspired in silico model equipped with a realistic retraining step to represent continual learning as well as an increasing load of decay that precedes full atrophy to represent tau accumulation and eventual atrophy.

References

[1] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29(9):2352–2449, Sept 2017.

[2] Daniel L.K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, 111(23):8619–8624, Jun 2014.

[3] Daniel L.K. Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365, Mar 2016.

[4] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, Jan 2021.

[5] Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, Nov 2019.

[6] Anup Tuladhar, Jasmine A. Moore, Zahinoor Ismail, and Nils D. Forkert. Modeling Neurodegeneration in silico With Deep Learning. *Frontiers in Neuroinformatics*, 15, Nov 2021.

[7] Jasmine A. Moore, Anup Tuladhar, Zahinoor Ismail, Pauline Mouches, Matthias Wilms, and Nils D. Forkert. Dementia in Convolutional Neural Networks: Using Deep Learning Models to Simulate Neurodegeneration of the Visual System. *Neuroinformatics*, Sept 2022.

[8] Casey Cook, Silvia S. Kang, Yari Carlomagno, Wen-Lang Lin, Mei Yue, Aishe Kurti, Mitsuru Shinohara, Karen Jansen-West, Emilie Perkerson, Monica Castanedes-Casey, Linda Rousseau, Virginia Phillips, Guojun Bu, Dennis W. Dickson, Leonard Petrucelli, and John D. Fryer. Tau deposition drives neuropathological, inflammatory and behavioral abnormalities independently of neuronal loss in a novel mouse model. *Human Molecular Genetics*, 24(21):6198–6212, Nov 2015.

[9] Albin John and P. Hemachandra Reddy. Synaptic basis of Alzheimer's disease: Focus on synaptic amyloid beta, P-tau and mitochondria, Jan 2021.

[10] Wiesje Pelkmans, Rik Ossenkoppele, Ellen Dicks, Olof Strandberg, Frederik Barkhof, Betty M. Tijms, Joana B. Pereira, and Oskar Hansson. Tau-related grey matter network breakdown across the Alzheimer's disease continuum. *Alzheimer's Research & Therapy*, 13(1):138, Dec 2021.

[11] Alejandra D. Alonso, Leah S. Cohen, Christopher Corbo, Viktoriya Morozova, Abdeslem ElIdrissi, Greg Phillips, and Frida E. Kleiman. Hyperphosphorylation of Tau Associates With Changes in Its Function Beyond Microtubule Stability. *Frontiers in Cellular Neuroscience*, 12, Oct 2018.

[12] Tania F Gendron and Leonard Petrucelli. The role of tau in neurodegeneration. *Molecular Neurodegeneration*, 4(1):13, 2009.

[13] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. 2009.

[14] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, Dec 2015.

[15] Jonathan Frankle and Michael Carbin. The Lottery Ticket Hypothesis: Finding Sparse, Trainable Neural Networks. *International Conference on Learning Representations*, Mar 2018.

[16] Babajide O. Ayinde, Tamer Inanc, and Jacek M. Zurada. Redundant feature pruning for accelerated inference in deep neural networks. *Neural Networks*, 118:148–158, Oct 2019.

[17] David A. Drachman. Do we have brain to spare? *Neurology*, 64(12):2004–2005, Jun 2005.

[18] Beatriz E.P. Mizusaki and Cian O'Donnell. Neural circuit function redundancy in brain disorders. *Current Opinion in Neurobiology*, 70:74–80, Oct 2021.

[19] Hao Li, Asim Kadav, Igor Durdanovic, Hanan Samet, and Hans Peter Graf. Pruning Filters for Efficient ConvNets. *Conference on Computer Vision and Pattern Recognition*, Aug 2016.

[20] Gongfan Fang, Xinyin Ma, Mingli Song, Michael Bi Mi, and Xinchao Wang. DepGraph: Towards Any Structural Pruning. *arXiv*, 2023.

[21] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pages 3519–3529. PMLR, 2019.

[22] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, Nov 2008.

[23] Maithra Raghu, Justin Gilmer, Jason Yosinski, and Jascha Sohl-Dickstein. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6076—-6085. Curran Associates, Inc., 2017.

[24] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation. *Advances in Neural Information Processing Systems 31*, Jun 2018.

[25] H. Moreno, G. Morfini, L. Buitrago, G. Ujlaki, S. Choi, E. Yu, J.E. Moreira, J. Avila, S.T. Brady, H. Pant, M. Sugimori, and R.R. Llinás. Tau pathology-mediated presynaptic dysfunction. *Neuroscience*, 325:30–38, Jun 2016.

[26] Margaret M Esiri and Steven A Chance. Cognitive reserve, cortical plasticity and resistance to Alzheimer's disease. *Alzheimer's Research & Therapy*, 4(2):7, 2012.