University of Amsterdam, The Netherlands

Winter Term 2022/23

Recycling of a Convolutional Neural Network for Reading Japanese Syllables: A Conceptual Replication of Hannagan et al. (2021)

Hanna F. Widhölzl[1], Maxim Zewe[2], and Vincent L. Ott[3]

[1]Institute for Interdisciplinary Studies, University of Amsterdam, The Netherlands

[2]Institute of Physics, University of Amsterdam, The Netherlands

[3]Department of Psychology, University of Amsterdam, The Netherlands

Submitted as Mini-Project for the Course:

Foundations of Neural and Cognitive Modelling

Instructor:

Prof. Dr. Jelle Zuidema

Date of Submission:

18.12.2022

Correspondence

Hanna F. Widhölzl: hanna.widholzl@student.uva.nl

Maxim Zewe: maxim.zewe@student.uva.nl

Vincent L. Ott: vincent.ott@student.uva.nl

Recycling of a Convolutional Neural Network for Reading Japanese Letters:
A Conceptual Replication of Hannagan et al. (2021)

The article *'Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading'* by Hannagan et al. (2021) revolves around a biologically plausible deep convolutional neural network (CNN) to simulate the ventral visual pathway. The authors show how this CNN, when initially trained for object recognition, can adapt to incorporate word recognition and thus mimic the visual word form area (VWFA) in the ventral occipitotemporal (VOT) cortex. The VWFA responds selectively to orthographic features of words and shows distinctive properties, such as invariance to case, font, or size of the word. In humans, the VWFA is hypothesized to emerge through different mechanisms. Two popular proposals are the neuronal recycling hypothesis and the biased-connectivity hypothesis. The study at hand investigates the research question to what extent a minimal computational model of both hypotheses is sufficient to account for the emergence of the VWFA during reading acquisition.

Hannagan et al.´s Model Architecture and Training

Hannagan et al.´s CNN had four convolutional layers that represented the ventral visual pathway (V1, V2, V4, IT). After each convolutional layer, there was a pooling layer. The final layer was a dense layer for decoding.

The neuronal recycling hypothesis posits that existing visual recognition skills are partially reused in reading acquisition. The researchers probed this idea with their model as follows. The model was trained in two training phases, thereby imitating reading acquisition in children. That is, children learn to recognize objects before they learn how to read. As such, in the first training phase, the CNN was trained on exemplars of 1000 object classes. In the second training phase, the model was trained on exemplars of 1000 word classes interleaved with the object images from the first training phase. The goal of this interleaved training during the second training phase is to prevent catastrophic interference, so potential unlearning of the object classes. Furthermore, the word exemplars varied in size, font, and case to represent the variance in word stimuli from which children learn reading.

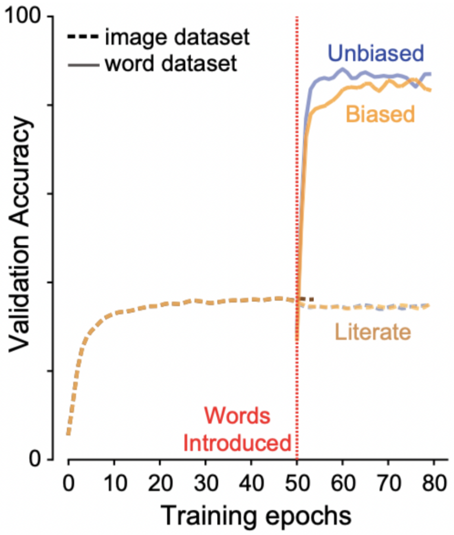
To test the biased-connectivity hypothesis, the authors include an unbiased and a biased condition: in the unbiased literate condition, the dense layer was fully connected to all existing output units whilst only around 10% of dense units were connected to the output units within the biased literate condition. This represented the biased connectivity from the VWFA to language areas compared to other subregions of the VOT cortex.

Hannagan et al.´s Results & Interpretation

Figure 1 shows the classification accuracies of Hannagan et al.´s CNN. After the first training phase, the model classified about 35% of the objects correctly. Once the second training phase began (epoch 50), word recognition quickly reached an accuracy of 85%. Importantly, there was only a slight drop in accuracy for object recognition, which indicates that the CNN was robust against catastrophic interference. This mimics the production of a VWFA and supports the neuronal recycling hypothesis. As the word recognition was about equal between the unbiased and the biased CNN, the results neither falsify nor confirm the biased-connectivity hypothesis. Hannagan et al. suggested that the higher accuracy for word (vs. object) recognition might have been because the word stimuli were less complex.

Figure 1

Changes in Network Performance During Training



Note. Average validation accuracy in the unbiased literate and biased literate networks on the test set, separately for images and words.

Mini-Project

Aims of the Project

Our project aims at (1) conceptually replicating, (2) testing the generalizability, and (3) extending the findings by Hannagan et al. with a simplified version of their model. We want to conceptually replicate their findings by using a different object dataset. To further test the generalizability of the reading acquisition process, we use handwritten Japanese hiragana syllables instead of French words. This allows us to test the reading acquisition process for more natural, handwritten stimuli from which children learn words (Hannagan et al. exclusively used typed words). Further, it examines if Hannagan et al.´s findings also hold true for languages with other symbols than Roman letters. We extend Hannagan et al.´s article by also training another model only on hiragana syllables first and then on a merged dataset of objecs and syllable classes. While this is not biologically plausible since children would not learn reading before internalizing object categories, this allows us to interpret ceiling effects, such as how well the network would 'be able to read' if it was not already committed to recognising object classes.

Methods

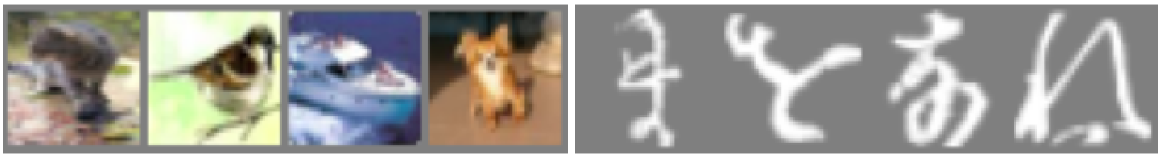
Stimuli

For the object classes, we used the CIFAR-10 (Krizhevsky & Hinton, 2009) dataset, which is built-in in PyTorch (Paszke et al., 2019). It consists of images for 10 common object classes: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. The train set has 5000 exemplars per class and the test set has 100 per class. Each image is 32x32 pixels. **Figure 2** (left side) shows four exemplars from the CIFAR-10 dataset.

For the syllables, we used the Kuzushiji-MNIST dataset (KMNIST, Clanuwat et al., 2018) from PyTorch, which consists of images for 10 classes of modern hiragana syllables. To match the size of the CIFAR-10 dataset we truncated the KMNIST dataset to 5000 exemplars per class, whereas the test set remained at 100 exemplars per class. We further resized the images from 28x28 pixels to 32x32 so that it would match CIFAR-10. Additionally, all images were coloured according to the RGB model. **Figure 2** (right side) shows four exemplars from the KMNIST dataset.

Figure 2

Four Exemplars from the CIFAR-10 Dataset (left) and Four Exemplars from the Kuzushiji-MNIST Dataset (right)



Model Architecture

The starting point for our model was the open-source CNN from Loeber (2020; 2022). Based on this, we built a simplified version of the CNN by Hannagan et al. Because our stimuli are 32x32 pixels (whereas Hannagan et al.´s stimuli were 224x224 pixels) this allowed for only three convolutional layers. After each of the convolutional layers, a pooling layer follows. Just as in the original model, the output layer of our model was preceded by a dense layer. The number of output layers in training phase 1 was 10, corresponding to the number of classes to learn. As 10 classes were added in training phase 2, 10 more output layers were added after phase 1. For more detailed information, see our GoogleColab notebook (see below).

Training

We trained all models in two training phases. The first phase always comprised 20 epochs and the second always comprised 15 epochs. Like Hannagan et al. we trained our first model (*Model 1*) only on the CIFAR-10 dataset during the first training phase. For the second training phase, we followed the interleaved training scheme used by Hannagan et al., so we merged the CIFAR-10 and KMNIST dataset and then continued to train the model. In contrast, we trained *Model 2* only on the KMNIST dataset during the first training phase. For the second training phase, we merged the CIFAR-10 and KMNIST dataset and then continued to train the model, similarly to *Model 1*.

Validation of Interleaved Training

To validate whether the interleaved training would prevent catastrophic interference, we trained two additional models. *Model V1* was only trained on CIFAR-10 during the first training phase. In the second training phase it was only trained on KMNIST. For *Model V2*, we presented

KMNIST and CIFAR-10 in the reverse order. If the interleaved training of Model 1 and Model 2 actually protected against catastrophic interference, then one would expect that at the end of training phase 2, Model V1 would be noticeably worse at classifying objects compared to Model 1. Similarly, Model V2 should be noticeably worse at classifying hiragana syllables compared to Model 2.

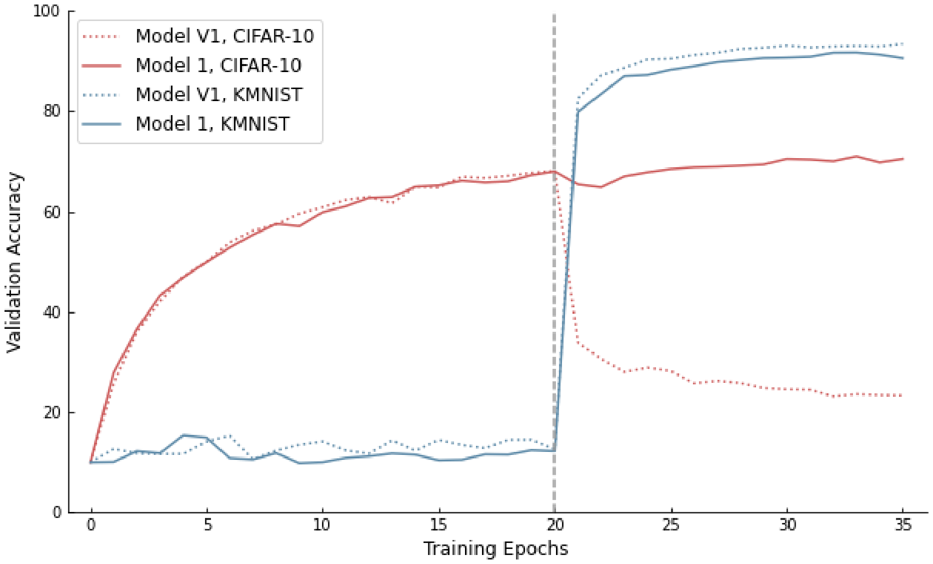
Results

Model 1

Figure 3 depicts the accuracy on the test set for each condition at each training epoch averaged over all classes in the test set for *Model 1* and *Model V1*, from which four observations are to be made. Firstly, this shows that the network successfully learned to recognise object classes in training phase 1. Secondly, during training phase 1, no notable learning for Japanese syllables occurred, indicating that learning object classes did not have learning Japanese syllables as a by-effect. Thirdly, the networks successfully learned to recognise Japanese syllables in phase 2, despite previous exposure to objects. Finally, *Model 1* retained accuracy on object recognition in phase 2 while simultaneously learning to classify Japanese syllables. The final accuracies of *Model 1* at the end of training were 68.68% for CIFAR-10 classification and 91.61% for the KMNIST test set.

Figure 3

Changes in validation accuracy performance (in %) during training for Model 1 and Model V1



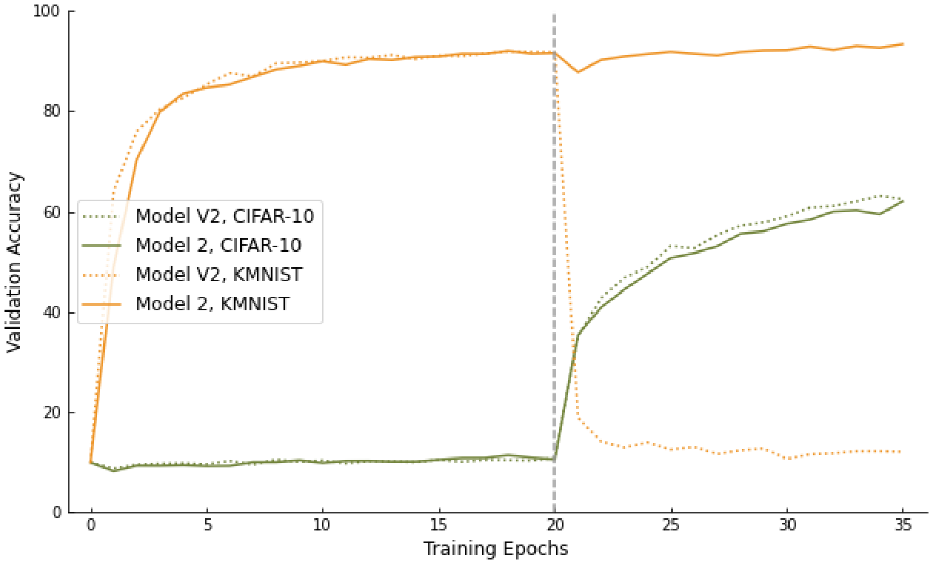
Note: the gray line indicates the start of training phase 2.

Model 2

Similarly to the results of *Model 1*, **Figure 4** demonstrates the computed average accuracy on the test set for each condition at each training epoch for *Model 2* and *Model V2*. Here, the network firstly learned to recognise hiragana syllables before exposing it to the merged data set of KMNIST and CIFAR-10. Analogous observations as for *Model 1* can be made. The network successfully learned classification of syllables in phase 1 without a by-effect for object recognition in phase 1. Additionally, phase 2 successfully taught object recognition, where *Model 2* retained syllable classification. After training, the overall accuracy of *Model 2* for CIFAR-10 classification was 61.2% and 92.69% for KMNIST classification.

Figure 4

Changes in validation accuracy performance (in %) during training for Model 2 and Model V2



Note: the gray line indicates the start of training phase 2.

Validation of Interleaved Training

Our results suggest that both *Model 1* and *2* are robust against catastrophic interference. Evidently, the interleaved training for *Model 1* protected against catastrophic interference, since the accuracy of object recognition of *Model 1* did not steeply decline once the network was exposed to linguistic input at epoch 20. Instead, *Model V1'*s object recognition accuracy did steeply decline after epoch 20, indicating that catastrophic interference occurred in phase 2 (**Figure 3**).

For *Model 2*, it is similarly evident that catastrophic interference was prevented since *Model V2* performed considerably worse from epoch 20 on hiragana syllable recognition, whereas *Model 2* retained its accuracy on the KMNIST test set (**Figure 4**).

Furthermore, while the accuracy on syllable recognition dropped to values around 10% (i.e., the average accuracy obtained by pure guessing) for *Model V2*, object recognition accuracy remained above chance levels for *Model V1* in phase 2. This indicates that catastrophic interference was less pronounced in the latter case.

Discussion

We built a CNN to conceptually reproduce and extend the findings by Hannagan and colleagues (2021), who showed how a CNN could reproduce the emergence of the VWFA during reading acquisition. To do so, we trained two CNNs on different stimuli than Hannagan et al. Both models were robust against catastrophic interference of knowledge learned in the first training phase.

Primarily, we trained our CNN firstly on images to recognize object classes and secondly on the same image data set combined with a data set of written Japanese hiragana syllables. We chose those syllables instead of words based on the Roman alphabet such as French. By doing so, this model meant to extend Hannagan et al.'s findings to languages based on other orthographic symbols. The network training resulted in high accuracies of recognising objects (68.68%) and Japanese symbols (91.61%). Hence, it tentatively suggests that the VWFA emerges similarly for a varying alphabetic system.

Our model performed better than the CNN by Hannagan et al., which had an accuracy of 35% for object recognition and 85% for word recognition. An explanation could be that our model was trained on fewer classes of objects (10 instead of 1000) and orthographic characters (10 instead of 1000). Moreover, our linguistic input consisted of syllables rather than full words (which should be more complex). Hence, the network might learn the Japanese syllables more easily than complete words. Adding to this is the fact, we did not test for invariances of size, case, or font. Had we included this, we would expect a slightly lower accuracy than was reported now as learning to recognise such variances would consume some of the network's capacity. Nevertheless, the symbols of the KMNIST dataset were handwritten. Those more natural stimuli offer greater ecological validity than using fonts, for instance Arial or Times New Roman, as done in the original publication, and thus present some natural variation in the dataset. The network was evidently able to handle this noisy input.

Our second model followed a similar paradigm as described above, but we initially trained the CNN solely on Japanese characters rather than objects, whilst in the second training phase, we again used the merged data set of hiragana letters and object images. Validation accuracy after training for object recognition was 61.2% and for letter classification 92.69%. Whilst this is not biologically plausible since children would not encounter written words/letters before learning image classifications, we wanted to investigate how well the CNN would perform by only learning Japanese symbols without being pre-trained on object classes to help interpret possible ceiling effects of accuracy performance of letter recognition. Since the network was uncommitted before learning reading in this case, more neurons were available for word recognition. In the brain, this would mean that the whole VOT cortex was available to form a VWFA rather than specific areas already being specialized to object or face recognition. Hence, we expected a higher accuracy for classifying the characters than if the CNN first got trained on objects (accuracy of *Model 1*). Yet, we did not find a considerably higher accuracy. This could be due to the low number of classes that our CNN was required to learn. The input of objects and Japanese characters might not have been sufficient in satiating the network and thus, a similarly high accuracy for word recognition is reported independent of whether the network was already trained on objects. We can then conclude that the prior learning of object classes did not interfere with letter recognition since final accuracies of both *Model 1* and *2* align with each other.

As an additional comparison between the two models, the catastrophic interference for *Model V1* (**Figure 3**) appeared to be less detrimental than in *Model V2* (**Figure 4**). This suggests that object recognition is more robust to catastrophic interference than syllable recognition. One explanation could be that the Japanese symbols have some physical features that are relevant to more general object recognition, such as horizontal edges and curves (see **Figure 2**). As a result, training *Model V1* on KMNIST rather than CIFAR-10 may conserve some of the object recognition skills it had developed in phase 1. In contrast, the CIFAR-10 images may contain more features than the hiragana syllables, such as color, and thus demand more neuronal capacity to be encoded, leading to the strong unlearning of syllables observed in *Model V2*. As such, this would suggest that catastrophic interference is less detrimental when learning a class of stimuli that is simpler than what previously had been learned.

A limitation of our project is that we did not validate our CNN's architecture by having it learn recognition of French words given that our network does not perfectly correspond with the one by Hannagan et al., such as that we only utilized three instead of four convolutional layers. Hence, our findings of *Model 1* cannot confidently be interpreted as a replication of the original findings. In addition, the scale of the network is smaller, which broadens the gap with the original publication.

Notwithstanding these shortcomings, our project shows two things. Firstly, it provides direct evidence for the importance of interleaved learning to prevent catastrophic interference in a convolutional neural network. Secondly, it demonstrates that a CNN can learn to recognise both objects and hiragana syllables while initially only being exposed to one type of stimuli. As such, it extends on the work by Hannagan et al. to other orthographic symbols and thus provides a model for how object recognition and reading acquisition may co-develop.

Data and Code Availability

The data and code for our analyses can be found here: https://colab.research.google.com/drive/12Xg7_TJiY9wXDFoDOZAkbgL7156emgzO?usp=sharing

References

Clanuwat, T., Bober-Irizar, M., Kitamoto, A., Lamb, A., Yamamoto, K., & Ha, D. (2018). Deep Learning for Classical Japanese Literature. ArXiv, abs/1812.01718.

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, *118*(46), e2104779118.

Krizhevsky, A., & Hinton, G. (2009). Learning multiple layers of features from tiny images.

Loeber, P. (2020). *PyTorch Tutorial 14—Convolutional Neural Network (CNN)*. https://www.youtube.com/watch?v=pDdP0TFzsoQ

Loeber, P. (2022). *Patrickloeber/pytorchTutorial* [Python]. github.com/patrickloeber/... (Original work published 2019)

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Advances in Neural Information Processing Systems 32 (pp. 8024–8035). Curran Associates, Inc.