REPRODUCTION REPORT FOR WIRING UP VISION

Kanishk Gandhi
31047572
kg7g19@soton.ac.uk

Jay Hill
31074243
jdh1g19@soton.ac.uk

Anjali Nandi
29249929
an5g18@soton.ac.uk

ABSTRACT

This reproducibility report presents our methodology, implementation and results of the experiments outlined by Geiger et al. (2022). We explore each of the 3 strategies they propose for reducing the number of synaptic updates in a deep neural network to better correspond to a primate's ventral stream, which is measured by the Brain-Score metric. Our results from reproducing the first strategy (reducing the number of epochs and/or training images) differ slightly from the original. With the second strategy – compressing the weights for initialization – the results were harder to reproduce due to the paper's lack of detail. For the third strategy, we corroborate that the critical training method proposed by Geiger et al. is indeed effective in retaining relative brain predictivity scores by only training critical downsampling layers in the model. Finally, the combination of the three, despite weight compression being unreproducable, still yields promising results somewhat in line with Geiger et al. (2022).

1 INTRODUCTION

Geiger et al. (2022) highlight the problem that the training of modern deep neural networks are far removed from the development of a primate ventral systems. The total number of supervised weight updates in the models far exceed the number of synaptic weight during primate brain development. Therefore, in an attempt to perhaps steer future development of these models in a direction which is more grounded in biological inspiration, three key strategies are proposed and examined:

- Reducing the number of epochs and training images presented to the model.
- Optimizing the 'at-birth' (initialized) synaptic connectivity via model weight compression.
- Training only the critical layers of the model by freezing all other layers to imitate the heterogeneity of cortical circuits.

They also explore combining all 3 strategies to establish a way to better emulate an adult-like ventral stream. The way this similarity is evaluated is via Brain-Score (Schrimpf et al., 2018), a tool which assesses how close areas in a model's architecture line up with the V1, V2, V4, IT, and behavior areas of the ventral stream. The mean score for these 5 is taken by Geiger et al. (2022) for a single brain predictivity score (BPS) to represent the evaluated model.

The deep learning model which is experimented upon is CORnet-S (Kubilius et al., 2019), on the ImageNet1000 dataset. They establish a 'fully-trained' standard model, by training CORnet-S normally for 43 epochs which achieves a BPS of 0.42. All subsequent hypotheses which incorporate the 3 strategies are compared against this baseline score as a percentage for analysis.

2 TARGET QUESTIONS

The target questions we seek to address for demonstrating reproducibility are as follows:

1. Can 2% of the supervised updates of a fully-trained model achieve ∼80% of the model's BPS?
2. Can 54% of a fully-trained model's BPS be achieved with no training simply by improving the random distribution of 'at-birth' synaptic connectivity via weight compression?
3. By training only ∼5% of the model's synapses via critical training, can nearly 80% of a fully trained model's BPS be achieved?
4. Is 26% the best increase in relative BPS from a standard model to one using all 3 strategies as shown in Figure 4B in Geiger et al. (2022)?

3 Experimental Methodology and Implementation

This section outlines the experimental methodology employed to reproduce the results necessary to address the target questions.

To begin with, the CORnet-S model was implemented in PyTorch based on the description provided by Geiger et al. (2022). Due to certain ambiguities, the original CORnet-S paper (Kubilius et al., 2019) also had to be referred to. We noted a difference in the number of trainable parameters (synapses) in our implementation compared to Geiger et al. (2022)'s reported values (53416616 vs ∼52000000).

We obtained our 'fully-trained' model by training our CORnet-S implementation for 32 epochs on ImageNet1000, which gave a **baseline BPS of 0.41**. This deviation from the original paper (43 epochs of training, BPS of 0.42), was due to resource constraints. This model and all subsequent ones were trained using a common random seed for initialization for reproducibility purposes. A custom Brain-Score framework was developed to allow us to evaluate models locally. The code used for Brain-Score setup and measurement was adapted from `https://github.com/brain-score/candidate_models/`.

Updates	Epochs	Images
100%	32	1,280,000
50%	16	1,280,000
20%	6.4	1,280,000
6.25%	2	1,280,000
2%	**2**	**409,600**
2%	**4**	**204,800**
0.2%	2	40,960

Table 1: Relative sizes in terms of supervised updates (Epochs x Trained Images) of the different CORnet-S models we trained and scored. Configurations for target question 1 in bold.

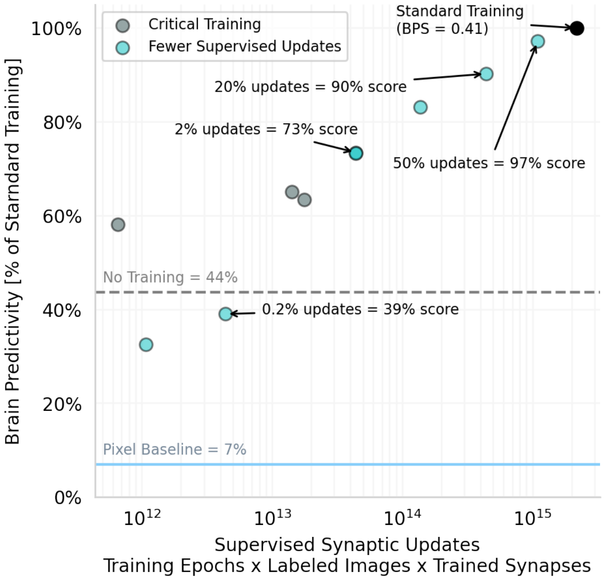

Figure 1: Reproduction of Figure 8 in Geiger et al. (2022) with our own results for key results from Figure 1A in the paper shown including 2% updates to address target question 1.

To address target question 1, as shown in table 1, we calculate the number of epochs and training images needed to achieve 20% of the 'fully-trained' model's supervised updates. We test 2 different configurations to ensure more coverage of the original paper's, as they do not specify the epoch-to-image ratio. We test other update sizes, including 20% and 0.2% which are highlighted in Figure 1 of Geiger et al. (2022) as some extra reproducibility testing, and a custom training dataset splitter was implemented to make this process efficient.

To address target question 2, our methodology followed that of section 4 of Geiger et al. (2022). CORnet-S model was randomly initialized – referred to as Kaiming Normal – 5 times (instead of 10 due to time and resource constraints), and an average BPS was found. Weight compression was partially applied to the 32-epoch fully-trained model, and 5 models initialized with weight compression were produced. To validate whether the initialization via weight compression provides an improved Brain-Score, a t-test was performed on the two sets of 5 measurements.

To address target question 3, the frozen weights indicated in figure 3B in Geiger et al. (2022) were frozen using `requires_grad = False` before training. Due to resource constraints, the dataset ImageNetMini (`https://www.kaggle.com/datasets/ifigotin/imagenetmini-1000`) was used for training. Using a smaller dataset helped to assess how well the critical training method scales. The trainable parameter totals were matched to the values shown in figure 3C of the paper. The models that we implemented include a full model (on ImageNetMini), the three Critical Training models and three Downstream Training models all shown in figure 3A.

All models were trained on the full ImageNetMini dataset for 10 epochs. BPSs were calculated for each model, and figure 3C in Geiger et al. (2022) was reproduced.

Finally, to address question 4, we combine the 3 strategies by initializing a CORnet-S model's weights via weight compression, and trained it via Critical Training of the downsampling layers (suggested by Geiger et al. (2022) to be the best). Then, for 2 epochs on 100 training images, we recorded the BPS and compared the differences between a standard model trained for 2 epochs and 100 training images. The implementation repository can be found at `https://github.com/COMP6258-Reproducibility-Challenge/Wiring-Up-Vision-Reproduction`

4 Results & Analysis

The results from our experiments addressing **target question 1** can be seen in figure 1. We ran two models with 2% of the supervised updates of the fully-trained model, and both achieved a BPS of 73% of the fully trained model, which is relatively close to Geiger et al. (2022)'s result of 76%. However, we see in figure 1 that this slight decrease in BPS is present for all the models that we evaluated. The lower the amount of updates, the larger the amount of disparity; 0.2% was reported as 50% relative BPS in the paper, but we found it to be 39%. We also observed a discrepancy between the paper's claim (in Section 3 of Geiger et al. (2022)) of achieving a 97% relative BPS with only 20% of supervised updates. Our findings were that 50% of the updates were necessary to achieve the 97% relative BPS, also shown in Figure 1.

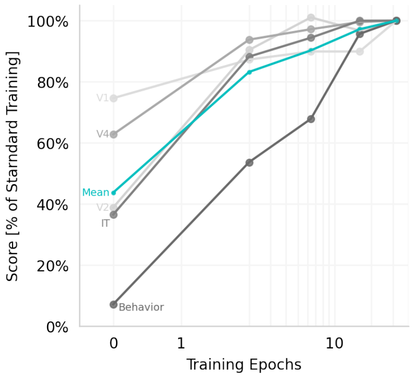

Figure 2: Individual brain predictivity scores over epochs, a reproduction of figure 1B in Geiger et al. (2022).

Based on our analysis, the discrepancy can potentially be attributed to our frame of reference being a model trained for only 32 epochs instead of 43 epochs. As you can see in figure 1A of Geiger et al. (2022), there is more variance in the relative BPS of the models with fewer synaptic updates, and more uniformity for higher synaptic updates. Having only 32 epochs puts our results in this area of more variance, hence resulting is greater disparities. Therefore it may be feasible to reproduce Geiger et al. (2022)'s results by using a model fully-trained to 43 epochs. Despite these differences, we were able to successfully reproduce the general trend, and as shown in fig. 2, the trend for the individual Brain-Score metrics can also be successfully reproduced.

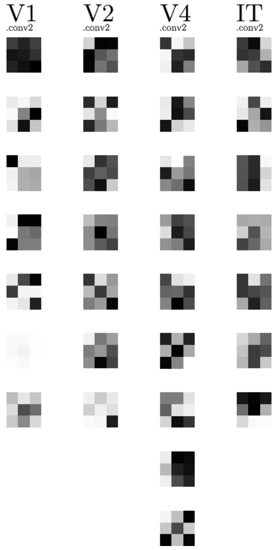

Figure 3: Visualizations of weight compression parameters reproducing Figure 2B in Geiger et al. (2022) in an attempt to answer target question 2.

Target question 2 was unfortunately unable to be answered due to a lack of clarity provided by Geiger et al. (2022) for implementing weight compression. The instructions surrounding the weight compression of the first layer, despite being seemingly the most import layer for compression, were followed but to no avail. Details provided in the main body and appendix were not sufficient to reproduce the results of compression by k-means clustering. Clustering was easily performed on the weights, but the method of sampling these clusters for initialization was explained poorly. We attempted to produce a method which somewhat matched their description, but this appears to be a poor reimplementation, as shown in fig. 3 with its distinctions. The kernels from cluster centers show patterns, like the equivalent diagram in the paper, yet all quite different. Attempts at initializing the weights using this method resulted in a significant *decrease* in at-birth BSP when compared to random (Kaiming) initialization: 1% significance for V4 and IT. 10% for V2. V1 and Behavior had no significant decrease when calculating p value from a t-test (8 d.f). This is opposite to Geiger et al. (2022)'s result, but is likely due to our implementation built on assumptions being poor.

Regarding **target question 3**, the paper claims that nearly 80% of a fully trained model's BPS can be achieved by training only approximately 5% of the model's synapses via Critical Training. As shown in fig. 4, critical training achieves over 80%, albeit on a smaller training dataset (ImageNetMini). We also corroborate that Critical Training is more effective than Downstream Training. The plot shows an initial increase in BPS followed by a decrease, which can be attributed to the smaller dataset causing more variance. However it can be argued that the Critical Training method still scales, and in terms of reproducibility, the strategy is generalizable.

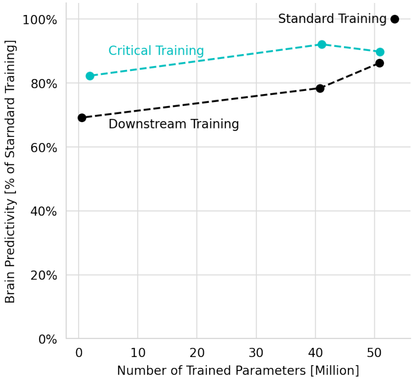

Figure 4: Reproduction of figure 3C in Geiger et al. (2022) verifying target question 3.

Due to the inability to reproduce the weight compression successfully, a model which incorporates it and Critical Training was not possible to replicate from Geiger et al. (2022). However we used our naïve version to try to answer **target question 4**, and interestingly we find that the increase in BPS – for 2 epochs and 100 training images by using our naïve weight compression and Critical Training – was still 10%. While this is not as high as the reported 26%, we have perhaps shown how effective their weight compression method is if done right, and it is a shame Geiger et al. (2022) do not provide the necessary details to achieve this.

5 Conclusion

Our results successfully address target questions 1 and 3, partially address question 4, but fail to address question 2. For both questions 1 and 3 however, compromises were made in the pursuit of reproducibility due to lack of resources, i.e. a 32 epoch fully trained model and testing Critical Training on ImageNetMini. The full model was trained for 85 hours on an NVDIA RTX 2080TI with 12GB VRAM, compared to the 60 hours for Geiger et al. (2022), likely due to their ability to use double the batch size we could. For question 2, as stated, the descriptions provided for reproducing weight compression successfully - a complicated task in itself - was insufficient for reproduction, which impacted question 4 too.

Another issue which impacted the reproducibility was that the specific Brain-Score measurement used was not specified. There is a public benchmark which can be used locally (which we did) but also a private benchmark which requires submission of your model. If Geiger et al. (2022) indeed used the private benchmark, this would certainly contribute to disparities in the BPSs.

On the topic of BPS, as stated it was assumed that this was calculated as the mean of the 5 individual scores for V1, V2, V4, IT and Behaviour, however it was not clarified if standard errors were included in the calculation to weight each component's contribution to the mean as would be statistically appropriate. Again this distinction could also impact the validity of the reproduced results. Lastly, some of the results reported by Geiger et al. (2022) took liberties with rounding, introducing inaccuracies. One of many examples being the abstract stated that a model with 5% of fully trained model parameters achieved nearly 80%, but in reality, it was only 2.6%. This meant we had to ensure the results we were reproducing actually reflected the true findings of the 'Wiring up Vision' paper.

To conclude, through attempting to reproduce the core experiments presented by Geiger et al. (2022), we can corroborate that 2 of their 3 proposed strategies demonstrate promising potential for enhancing the resemblance of machine learning models to neural synaptic behavior (by reducing the number of supervised updates) , with the third (weight compression) needing more clarity in the paper to demonstrate the same.

References

Franziska Geiger, Martin Schrimpf, Tiago Marques, and James J. DiCarlo. Wiring up vision: Minimizing supervised synaptic updates needed to produce a primate ventral stream. In *International Conference on Learning Representations*, 2022. URL `https://openreview.net/forum?id=g1SzIRLQXMM`.

Jonas Kubilius, Martin Schrimpf, Ha Hong, Najib Majaj, Rishi Rajalingham, Elias Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel Yamins, and James Dicarlo. Brain-like object recognition with high-performing shallow recurrent anns. 09 2019.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint*, 2018. URL `https://www.biorxiv.org/content/10.1101/407007v2`.