Extending a Computational Model of the Retinal Neural Code for Natural Scenes: Skip Connections and Saliency Maps

Jenny Dewhurst (14277867)

Timo van Hattem (12172421)

Sai Siddharth Govindarajan (14242958)

Assessor: Dr. Jelle Zuidema

Advanced Neural and Cognitive Modelling (5244ANCM6Y)

Date: 26-10-2023

UNIVERSITEIT VAN AMSTERDAM

Abstract

In this study, we aimed to replicate and extend on previous work by Maheswaranathan et. al. (2023) who trained a three-layer CNN model to capture the retinal neural code in response to natural scenes and white noise. We explored how model prediction performance would change when a skip connection is added, this way incorporating a structure similar to that present in a biological human retina. The dataset consisted of intracellular recordings of retinal ganglion cells from larval tiger salamanders responding to a combination of natural scenes and white noise stimuli. Saliency analysis was conducted to extract the features to which the model gives the most importance. We successfully replicated the results from the original paper, which demonstrate the model predicted neural spike rate from both the white noise stimuli and natural scenes. However, there was not an improved task performance with the addition of a skip connection.. Saliency analysis revealed central regions were of high importance for the model, irrespective of stimuli type or model architecture. These findings suggest the presence of complex interactions behind the processing of visual stimuli beyond standard saliency interpretation.

Keywords: retina, computational models, natural scenes, white noise, saliency maps

Introduction

As the first step in the visual processing stream, the retina plays a pivotal role in shaping human visual perception (Ptito et al., 2021). Multiple layers of specialized cells within the retina capture visual information reaching the eye and transmit the encoded signals to cortical structures, such as the occipital cortex, that are responsible for higher visual processing (Masland, 2012; Ptito et al., 2021). Photoreceptors, which include rods and cones, respond to light energy within their receptive fields and convert this into electrical signals (Molday & Moritz, 2015). Bipolar cells are an intermediate cell layer in the retinal circuitry and serve to relay signals from the photoreceptors to subsequent ganglion cells (West & Cepko, 2022). Ganglion cells form the output layer of the retina and these cells ultimately transmit the signals from the eye into the cortex via the optic nerve (Xiang et al., 1996). Lastly, amacrine cells are positioned in between bipolar and ganglion cells and modulate communications between these two layers (Ptito et al., 2021).

Researchers continually strive to get a deeper understanding of the neural mechanisms of the retinal circuit and the role of different retinal neurons in shaping human visual perception. It is a challenge to capture the entirety of the retinal circuit with experimental and clinical research, using neuroimaging and electrophysiology techniques in vivo and in vitro, due to the inherent high complexity of this system (Eldret & Reh, 2021). Computational models can provide additional insights on the neural underpinnings of the retina by characterizing and predicting retinal behavior (Roberts et al., 2016). In addition, computational models can be used to test and formulate scientific hypotheses with regard to how interneurons interact to perform various computations (Roberts et al., 2016). Computational models of the retina have traditionally been

trained on artificial stimuli, such as gratings, flashed, and white noise, because these are well-understood and easily controllable (Maheswaranathan et al., 2023). To better understand how the retina generates responses in the real world, it is crucial that models are trained on natural stimuli of settings that accurately represent the outside world.

Maheswaranathan et al. (2023) recently tested different computational models in an attempt to capture the neural code of the human retina for natural scenes, while also achieving computational and mechanistic interpretability. The authors trained a simple linear-nonlinear (LN) model, generalized linear model (GLM) and deep convolutional neural network (CNN) to predict the firing rate of ganglion cells responding to images of natural scenes and white noise. Results showed that the three-layer CNN trained on natural scenes accurately predicted the firing rate of ganglion cells and outperformed the LN and GLM models. In subsequent analysis, it was shown that the CNN model trained on natural images is able to capture a variety of nonlinear retinal computations of ganglion cells, such as motion adaptation, latency encoding and predictive coding. Interestingly, it was found that these particular phenomena arose in the CNN model as a byproduct of optimization on the natural scenes dataset. The CNN trained on white noise was also able to predict firing rates of ganglion cells accurately, but it did not trigger the retinal phenomena like the CNN trained on natural scenes. Furthermore, there was a high correlation between the activation of internal units in the neural network and that of real interneurons that were recorded in a separate subject. The authors concluded that the correlations between the model interneurons and the biological interneurons was sufficiently high to provide interpretable mechanistic insights into the retinal computations. Taken together, this allowed the researchers to test various hypotheses on neural computations of the human retina and also generate new testable hypotheses using the trained models.

The objective of Maheswaranathan et al. (2023) was to create a minimal computational model that would still be able to capture the retinal neural code. The authors argued that three layers in the CNN would be sufficient to compute a set of complex retinal phenomena while maintaining mechanistic interpretability. In the current model, the first layer is interpreted as encoding the connection from photoreceptors to bipolar cells; the second layer is interpreted as encoding the connection from bipolar to amacrine cells; the third layer is interpreted as encoding the connection from amacrine cells to ganglion cells. However, arguably, the retinal circuit contains a greater complexity that can't simply be reduced to three feedforward layers (Masland et al., 2012; Guo et al., 2014; Roberts et al., 2016). It is argued here that the model of Maheswaranathan et al. (2023) is lacking a level of correspondence with the real retina that makes it a suboptimal retinal computational model. For one, direct connections between bipolar cells and ganglion cells may be observed in the real retina, which are not present in the current model (Nelson & Connaughton, 2007; Masland et al., 2012; Ptito et al., 2021). That is, in the current model by Maheswaranathan et al. (2023), all the connections go through the second layer of amacrine cells. By extending the current CNN model of Maheswaranathan et al. (2023) and adding such skip connections between the layers, we can more closely align the model to reality and potentially improve model performance.

The present study aims to first replicate findings by Maheswaranathan et al. (2023) showing that a three-layer CNN model is able to capture the retinal neural code by accurately predicting ganglion firing rates for natural images. Then, this study expands on the work of Maheswaranathan et al. (2023) by changing the architecture of the model to resemble the biological structure of the human retina more accurately. Skip connections will be added to connect layer 1 and layer 3 directly, with layer 2 serving as a relay. Lastly, interpretability

saliency analysis will be performed on the current model. The scope for each layer for different stimuli will be assessed and compared between the original and the expanded model.

Methods

Data

The retinal ganglion cell recordings were recorded from larval tiger salamanders of either sex. An array of 60 electrodes was used for the retinal ganglion cell recordings, as described in Kastner & Baccus (2011). The intracellular recordings were taken using sharp microelectrodes, as described in Manu & Baccus (2011). The datasets consisted of recordings from 5, 4, and 17 cells, from both white noise and natural scene segments.

The visual stimuli were projected using a video monitor at 30 Hz. The natural scenes images were taken from a publicly available natural images database, consisting of 5000 images of the Okavango Delta in Botswana (Tkacik et al., 2011). The authors chose this location because it is a tropical savanna and this is believed to be the habitat where the human eye evolved. The images changed every second and they were jittered in an attempt to match the statistics of fixational eye movements. The white noise stimuli consisted of black and white images, with the same frame rate, spatial size and duration as the natural scene stimuli. Figure 1. shows an example of natural image stimulus and white noise stimulus. For a detailed explanation of visual stimuli, please refer to Maheswaranathan et al. (2023).

Model architecture

The CNN consisted of two convolutional layers followed by one fully connected layer. Each convolutional layer included a linear spatiotemporal convolution and was rectified using a rectified linear unit (ReLu). Additionally, each convolutional layer consisted of eight cell types that tiled the visual input. Each cell receives input from the previous layer that was spatially limited, then the final synaptic layer receives input from across the entire spatial extent of the previous layer. The final layer contains a spatial filter, a scaling and shifting parameter for each ganglion cell, and a nonlinearity using a softplus function. For a detailed explanation of model architecture, please refer to Maheswaranathan et al. (2023).

Skip Connections

Typically, skip connections in deep neural networks (DNNs) are utilized to improve optimization and generalization. Before skip connections, DNNs would experience training and optimization failure when they had more than ten layers. However, skip connections allow outputs from earlier layers to be connected to later layers, thus alleviating these training problems and improving performance (Oyedotun et al., 2022). With the current model architecture, skip connections are not necessary for training performance, but rather to improve model alignment with the biological retinal structure. Therefore, we trained two more models using the same natural scenes and white noise datasets, but we added a skip connection between layer 1 and layer 3 into the model architecture.

Training and Testing Procedure

The model was trained to predict the firing rates of the recorded retinal ganglion cell responses to either the natural scenes or white noise stimuli. Due to limitations in computational

power, our model was only trained on the dataset with recordings from four ganglion cells. It is important to note the original paper trained three different models on all three cell recording datasets and averaged them together. To feed the responses and stimuli to the model for training and testing they were grouped into 10 ms time bins. The training set was around 54 minutes long (323,786 10 ms samples), the validation set was around 6 minutes long (35,976 10 ms samples), and the test set was around 5 minutes (5,957 10 ms samples).

The original paper already conducted a hyperparameter search and optimized the hyperparameters to the current dataset. For this reason, we did not conduct our own optimization procedure and used the hyperparameters outlined in their paper of an Adam optimizer, learning rate of 0.005, L1 regularization of 0.0001, L2 regularization of 0.001, and dropout of 0.05, trained for 40 epochs. For a detailed explanation of the training procedure and hyperparameter values, please refer to Maheswaranathan et al. (2023).

Saliency Analysis

Saliency maps are commonly used in the field of computer vision to help identify which features of the input were the most useful for the model to make its classification. This is done by using gradient calculations to assign an importance score for individual features of the input image, and it is then visualized as a heat map, showing the most important features in red/yellow (Ismail et al., 2021). We hoped the saliency maps would provide increased insight into what features of the input stimuli are most important for the different layers to predict firing rate. Potentially increasing the interpretability of this model, and giving more insight into how each of the eight cell types per layer work. To create the saliency maps, the weights of the pretrained model were loaded and a subset of the preprocessed test dataset with only 40 samples was

created. The small sample size was due to limitations in computational power. Then, the Captum toolbox in Pytorch was used to calculate the saliency maps and Matplotlib to visualize the plots (Kokhlikyan et al., 2020; Hunter, 2007).

Figure 1

Example of visual stimulus used for natural scenes(A) and white noise (B) model training
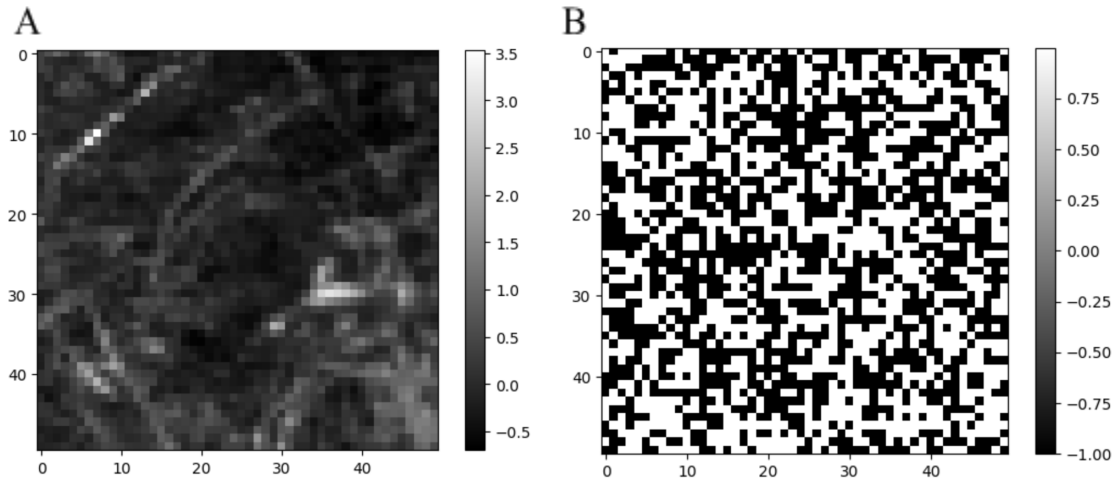
Note. The natural scenes were images of the Okavango Delta in Botswana. The images changed every second and they were jittered in an attempt to match the statistics of fixational eye movements. The white noise stimuli consisted of black and white images, with the same frame rate, spatial size and duration as the natural scene stimuli.

<div align="center">

Results

</div>

Baseline models

The model trained on natural images had a final correlation accuracy of 72.39% (Figure 2A), while the model trained on white noise had a final correlation accuracy of 80.25% (Figure 2B). The model trained on white noise is seen to start from a relatively higher initial accuracy for

both the test (~55%) and validation data (~68%) but there is a large difference between the final correlation accuracies between both of these sets.

Figure 2

Correlation accuracy curves during training of the natural scene (A) and white noise (B) baseline model
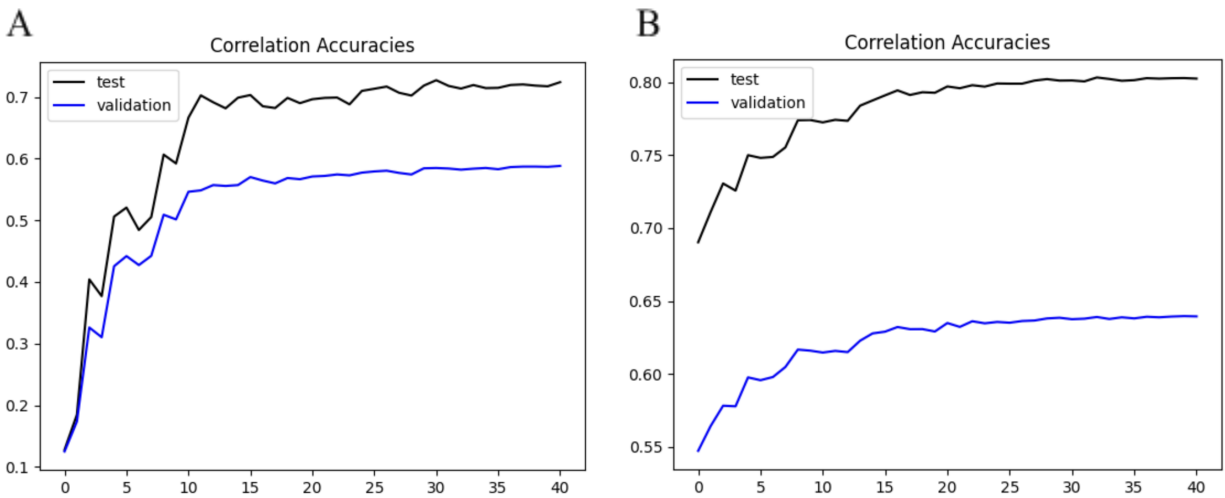

The saliency map of the whole model trained on natural scenes showed that attention was focused to (top) central regions of the image (Figure 3A). The saliency map of the whole white noise similarly highlighted a patch in the (bottom) central regions of the image (Figure 3D). The highlighted patches were of comparable size (Figure 3A & 3D).

The saliency maps of the first and second layer of the natural scene model showed that attention was evenly distributed across the image with a tendency for increased attention to top-central regions (Figure 3B & 3C). A similar region was highlighted in the white noise model, though in this model seemingly more focalized (Figure 3E & 3F). Notably, the salience in the first and second layer of the model trained on white noise was lower around the entire image

compared to the natural scene model. For both models, there were no large, observable differences in the saliency maps of the separate layers (Figure 3B, 3C, 3E & 3F).

Figure 3

Saliency maps of the natural scene model (A-C) and white noise model (D-F) without skip connections for the various model layers
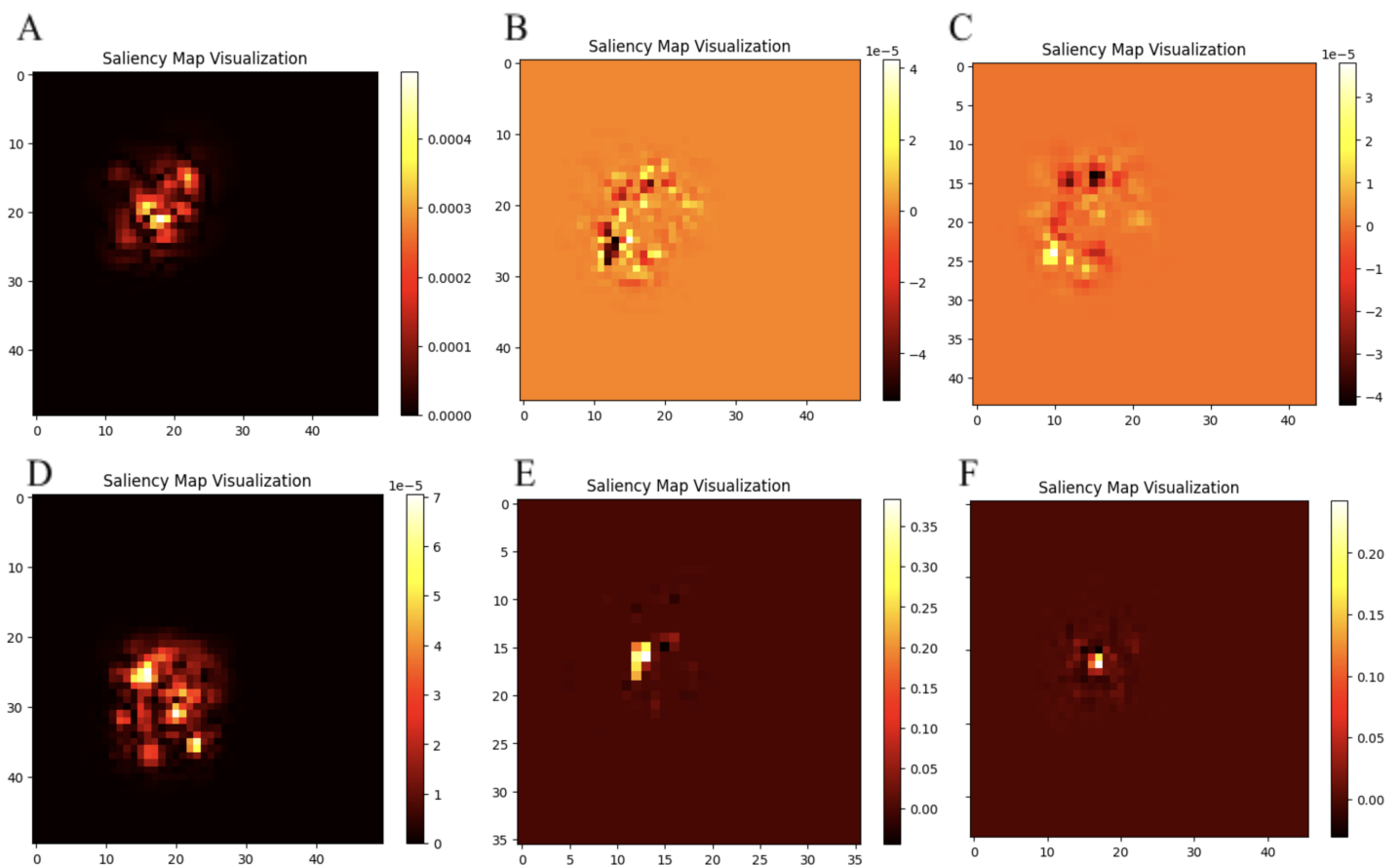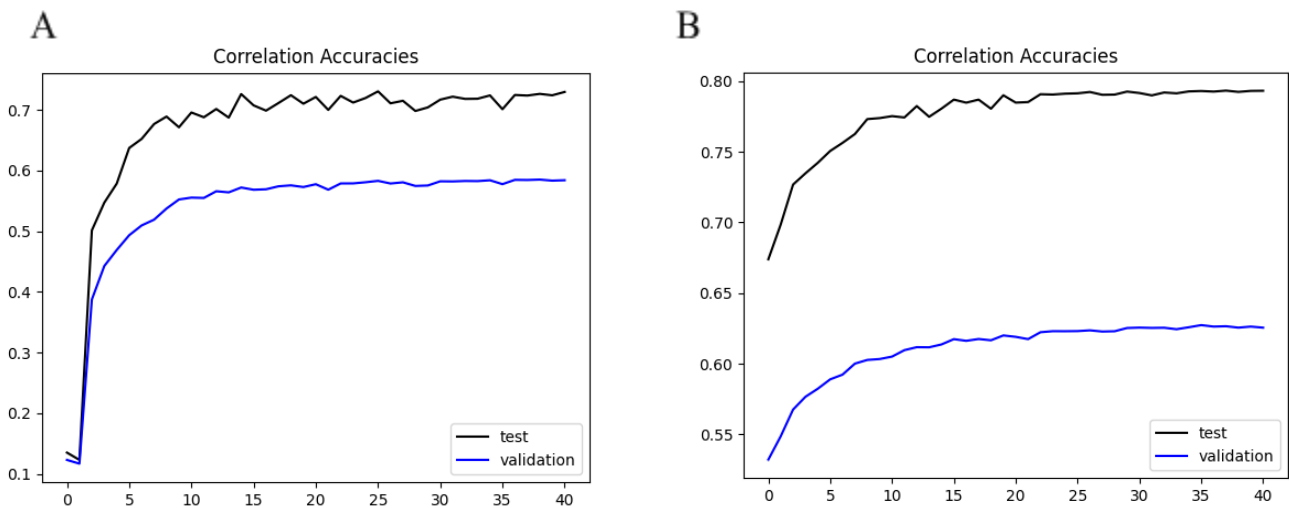

Skip-connection models

The model trained on natural images with skip-connections had a final correlation accuracy of 72.94% (Figure 4A), while the model trained on white noise with skip connections

had a final correlation accuracy of 79.33% (Figure 4B). The correlation accuracy of the model trained on white noise follows a similar pattern of higher initial value (~52% & ~67%) and a large difference between the final values. There does not seem to be any difference between the correlation accuracies for interneurons of the models with and without skip connections, irrespective of the dataset (Figure 6). Although, it does seem the skip connection model trained on natural scenes reaches convergence faster than the model without skip connections trained on natural scenes.

Figure 4

Correlation accuracy curves during training of the natural scene (A) and white noise (B) model with skip connections


The saliency map of the whole model trained on natural scenes with skip connections showed that attention was focused to central regions of the image (Figure 5A). For the natural scene model with skip connections, salience in the second layer a larger patch was covered

compared to the first layer (Figure 5B & 5C). Both the models retained the even distribution of attention in the first and the second layer, like the models trained without the skip connections (Figure 5B, 5C, 5E & 5F). The whole model saliency map with the skip connections does seem to have a smaller region of focus when compared to the heatmap of the models without the skip connections (Figure 5A and 5D).

Figure 5

Saliency maps of the natural scene model (A-C and white noise model (D-F) with skip connections for the various model layers.
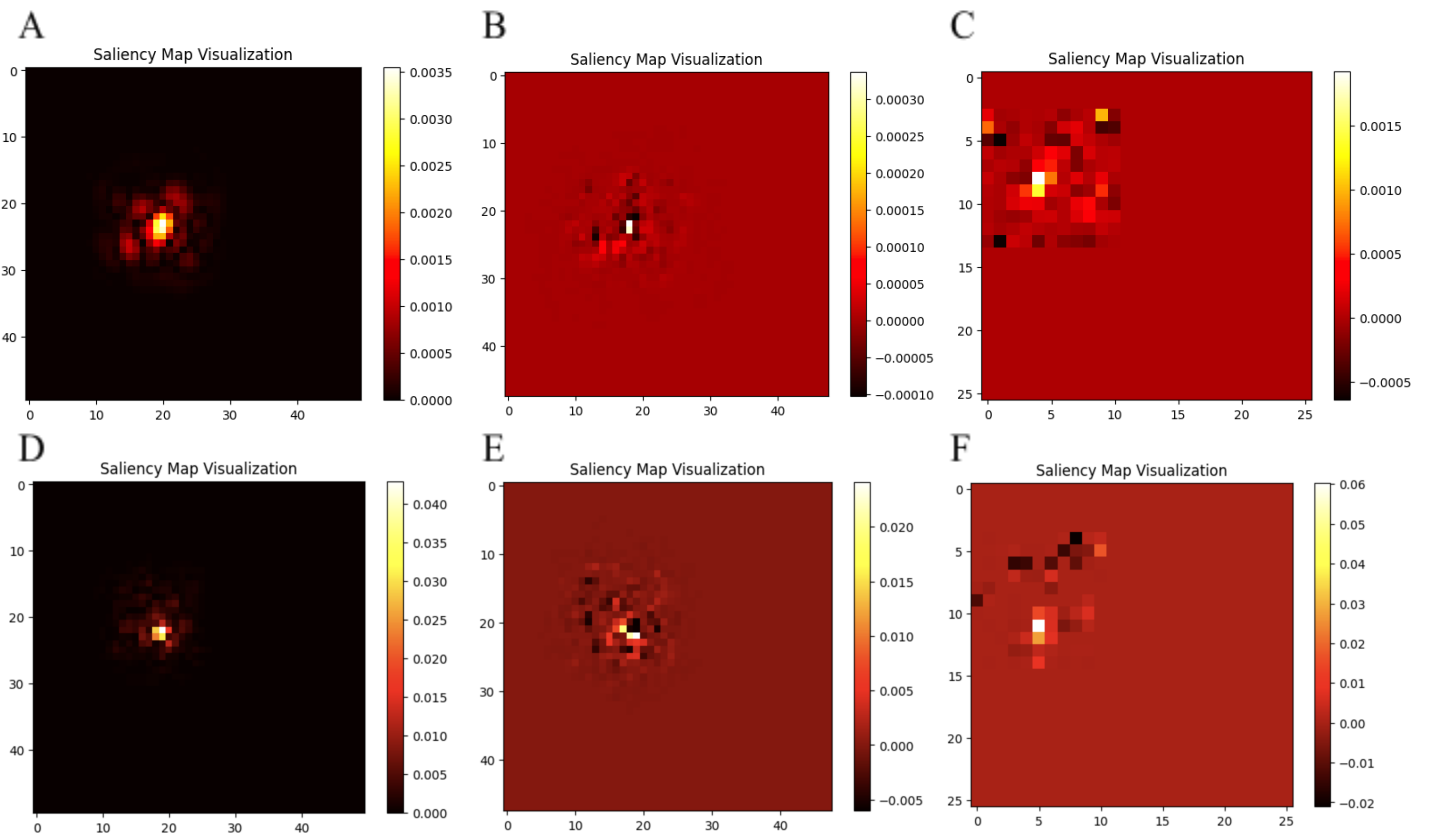
It is worth noting that the saliency maps had variations in the scales of saliency across different levels of the model architecture, in both the models trained with and without skip connections. The range of saliency was lowest in the initial layers (Figure 3E, 5B & 5E) which increased as the processing advanced to the final layer (Figure 3F, 3D, 5C, 5A, 5F & 5D). Interestingly, there was discrepancy in this pattern of scale of saliency observed in the natural scene model without skip connections. In this model, the saliency map for the whole model (Figure 3A) exhibited a lower scale compared to the saliency maps of the first and the second layers (Figure 3B & 3C).

Figure 6

Comparison of correlation accuracies of models trained on natural scene images and white noise, with and without skip connections.
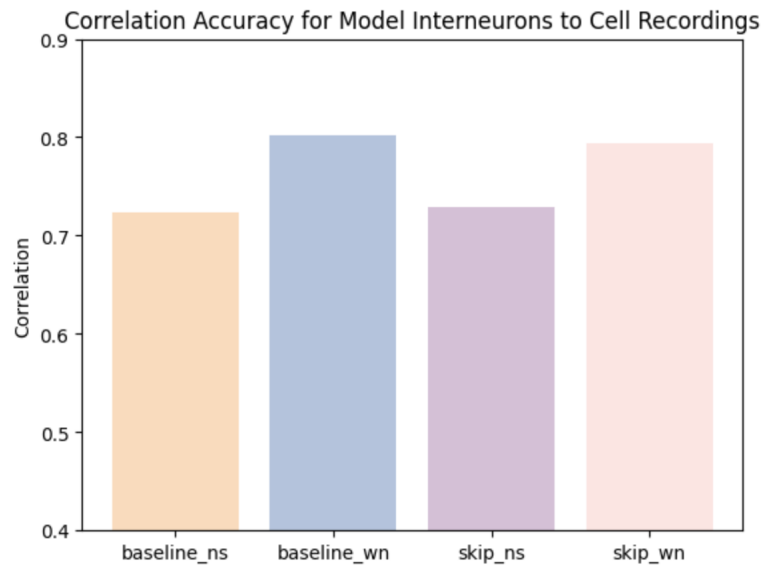
Note. Baseline_ns = baseline model natural scenes without skip connections; Baseline_wn = baseline model white noise without skip connections; Skip_ns = model natural scenes with skip connections; Skip_wn = model white noise with skip connections.

Discussion

The goal of this study was to replicate findings of a previous study that utilized a three-layer CNN model to capture the retinal neural code with high computational and mechanistic interpretability. In addition, this study aimed to expand on previous work by changing the model architecture to resemble the biological structure of the human retina more accurately. Results showed that the models trained on natural scene images performed less well than the models trained on white noise, irrespective of presence of skip connections. Adding skip connections did not significantly improve the performance of the models in predicting ganglion cell firing rates. Furthermore, saliency analysis did not show differences between model layers in attended regions of the images. Taken together, our analysis could successfully replicate the findings of the study by Maheswaranathan et al. which confirms the robustness of their findings, even if we used a smaller training set. Their results were also averaged across multiple models, while we only used a single model average which explains the discrepancy of our lower accuracy rates compared to theirs.

The models we trained on white noise outperformed those trained on natural scenes in predicting the ganglion cell firing rates. The model's performance seems to be influenced by the nature of the input stimuli, with white noise being more conducive for accurate predictions, compared to natural stimuli. This could be attributed to the simpler patterns in the white noise dataset which has uniform statistical properties across space and time which might in turn make it easier for the model's architecture to learn from and predict the ganglion cell responses (McIntosh et al., 2016). This could also explain the higher initial accuracy seen in the models trained on white noise as there is lesser complexity, making the stimuli less unpredictable

compared to natural scenes. The large difference between the correlation accuracies on the test and validation data within the model trained on white noise images could be attributed to the inability of the model's complexity to generalize the patterns it had learnt from the test data to the validation data. The model could have also similarly overfitted to specific characteristics of the natural scene images, which is known to happen when using smaller datasets for complex images, leading to reduced performance (Dawson et al., 2023).

An interpretability saliency analysis was performed to assess which features of the input stimuli were most important to the model for predicting spike rate. For all models and layers, it was found that the central regions were of highest importance to make accurate predictions of the firing rates. That is, regardless of model architecture and whether the model was trained on white noise or natural images, the saliency maps revealed most attention was given to one specific region. . There are a number of ways to interpret these results, firstly the saliency maps might not really be informative, as natural scenes were shown as movies and not as 2D images. The central region importance observed could be due to this incongruence between standard saliency analysis and the nature of the presented stimuli. Additionally, there seems to be a bias towards the detection of motion in the computational processes of the visual cortex which drive attentional distribution (Appleby & Manookin, 2020). The dynamic nature of the stimuli could have influenced the model to respond similarly to temporal patterns of these scenes which might not be evident from these static saliency maps. Though this confirms the robustness of the model's ability to capture the biological relevance of the visual processing systems, the model could also be concentrating on the central regions as a simplified learning strategy, one which it seems to employ constantly, owing to the model's simplified architecture. Furthermore, it is possible the dataset utilized of recordings from only four ganglion cells was not large enough,

and saliency maps produced from a model trained on the full dataset could be more informative. Therefore, the results of this analysis should be interpreted with caution.

Regardless, the model seems to increasingly get more focused and allocates specific attention to certain regions as it advances to the final layer. Along with the saliency scales that seem to increase consistently, the model seems to reflect the ability of the visual system to attend to specific characteristics of a dynamic stimuli for the optimal processing of information in higher cortical layers. Furthermore, the discrepancy of the scale variance pattern being absent in the model trained on natural scenes without the skip connection could indicate the role of skip connections in the distribution of attention within the model when attending to complex visual stimuli. This needs to be investigated further to delineate the specific factors contributing to this unique pattern.

However, it is important to mention that our study's main limitation arises from the use of a relatively small dataset, consisting of only 4 cells, compared to the previous work, where they used 3 separate datasets with recordings of 4, 5 and 17 cells (a total of 26). These datasets were used to train separate models which were all averaged to get their results. Our smaller computational availability could only allow the analysis of 4 cells that would explain the discrepancies in our results. Mainly, the lack of difference between the accuracies of the models trained with or without skip connections using both white noise and natural scenes could be attributed to the use of such a small dataset. Moreover, the salamander has been used in previous research investigating the retinal ganglion cells (Das et al., 2019) due to the relatively close resemblance to the human retina and the easy accessibility for intracellular recordings. Though this allowed both the authors of the previous work and us to bridge the gap between computational models and real neural responses, we need to continuously emphasize on the

species-specific variations (including the congruity of skip connections-type architecture) when interpreting these findings within the context of human visual processing. Considering these restraints, we recommend future research that combines the exploration of changing the model architecture to improve computational sophistication (such as adding more layers) with the ability of the model to replicate how our visual systems work (adding skip connections and recurrency in the model).

Conclusions

We set out to investigate the improvement of the computational and mechanistic interpretability of a model of a CNN trained to predict the responses of ganglion cells to visual stimuli. We analyzed the saliency maps of the model's prediction to see which parts of the input were the most important for the model's classification and added skip connections to make it resonate closer to the biological organization of visual information processing. Our analysis could replicate the findings of the previous work, though less equivalently owing to the use of data from only four cells compared to twenty seven, as used by the authors of the original paper. Our accuracies of prediction of the neural responses remained higher even though the addition of skip connection did not increase it as we predicted. Contrary to the original study, models trained on white noise stimuli performed better than the models trained on natural scenes, irrespective of the use of skip connections. Based on this, we conclude that the models' performance is significantly influenced by the nature of the stimuli. The analysis of the saliency maps of the models revealed a focus on the central region of the stimuli which we believe could be due to the model's simple architecture, use of relatively lesser data, biological similarity based on detection

of motion, or a complex interplay of all of these factors. The discrepancy between the saliency maps between the different layers of the model suggests that the model's prediction of the retinal ganglion cells' responses are more complex than initially thought, involving dynamical processes similar to that of visual processing.

With our limitations in mind, we propose further investigation into the development of biologically plausible models which are also computationally more robust. We hope that the importance of the nature of the stimuli and the preservation of complexity is highlighted in the future of computational models that aim to capture the neural mechanisms of the retina.

References

Appleby, T. R., & Manookin, M. B. (2020). Selectivity to approaching motion in retinal inputs to

the dorsal visual pathway. *ELife*, *9*, e51144. https://doi.org/10.7554/eLife.51144

Das, Gautham & Vance, Philip & Kerr, Dermot & Coleman, Sonya & McGinnity, Thomas &

Liu, Jian. (2018). Computational Modelling of Salamander Retinal Ganglion Cells using

Machine Learning Approaches. Neurocomputing. 325. 10.1016/j.neucom.2018.10.004

Dawson, H. L., Dubrule, O., & John, C. M. (2023). Impact of dataset size and convolutional

neural network architecture on transfer learning for carbonate rock classification.

Computers & Geosciences, *171*, 105284. https://doi.org/10.1016/j.cageo.2022.105284

Eldred, K. C., & Reh, T. A. (2021). Human retinal model systems: Strengths, weaknesses, and

future directions. Developmental biology, 480, 114–122.

https://doi.org/10.1016/j.ydbio.2021.09.001

Guo, T., Tsai, D., Bai, S., Morley, J. W., Suaning, G. J., Lovell, N. H., & Dokos, S. (2014).

Understanding the retina: a review of computational models of the retina from the single

cell to the network level. Critical reviews in biomedical engineering, 42(5), 419–436.

https://doi.org/10.1615/critrevbiomedeng.2014011732

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science &*

Engineering, 9(3), 90–95. https://doi.org/10.1109/mcse.2007.55

Ismail, A., Feizi, S., & Bravo, H. (2021). Improving deep learning interpretability by saliency

guided training. *35th Conference on Neural Information Processing Systems.*

Kastner, D.B., and Baccus, S.A. (2011). Coordinated dynamic encoding in

the retina using opposing forms of plasticity. Nat. Neurosci. 14, 1317–1322.

https://doi.org/10.1038/nn.2906.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J. Melnikov, A.,

Kliushkina, A., Araya, C., Yan, S. & Reblitz-Richardson, O. (2020). Captum: A unified

and generic model interpretability library for PyTorch. arXiv:2009.07896 [cs.LG].

https://doi.org/10.48550/arXiv.2009.07896

Manu, M., and Baccus, S.A. (2011). Disinhibitory gating of retinal output by transmission from

an amacrine cell. Proc. Natl. Acad. Sci. USA 108, 18447–18452.

https://doi.org/10.1073/pnas.1107994108.

Masland R. H. (2012). The neuronal organization of the retina. Neuron, 76(2), 266–280.

https://doi.org/10.1016/j.neuron.2012.10.002

McIntosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S., & Baccus, S. A. (2016). Deep

Learning Models of the Retinal Response to Natural Scenes. *Advances in Neural

Information Processing Systems*, *29*, 1369–1377.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5515384/

Molday, R. S., & Moritz, O. L. (2015). Photoreceptors at a glance. Journal of cell science,

128(22), 4039–4045. https://doi.org/10.1242/jcs.175687

Nelson, R., & Connaughton, V. (2007). Bipolar Cell Pathways in the Vertebrate Retina. In H.

Kolb (Eds.) et. al., Webvision: The Organization of the Retina and Visual System.

University of Utah Health Sciences Center.

Oyedotun, O. K., Ismaeil, K. A., & Aouada, D. (2023). "Why Is Everyone Training Very Deep

Neural Network With Skip Connections?," in IEEE Transactions on Neural Networks and

Learning Systems, vol. 34, no. 9, pp. 5961-5975, Sept. 2023, doi: 10.1109/TNNLS.2021.3131813.

Ptito, M., Bleau, M., & Bouskila, J. (2021). The Retina: A Window into the Brain. Cells, 10(12), 3269. https://doi.org/10.3390/cells10123269

Roberts, P. A., Gaffney, E. A., Luthert, P. J., Foss, A. J. E., & Byrne, H. M. (2016). Mathematical and computational models of the retina in health, development and disease. Progress in retinal and eye research, 53, 48–69. https://doi.org/10.1016/j.preteyeres.2016.04.001

Tkacik, G., Garrigan, P., Ratliff, C., Milcinski, G., Klein, J.M., Seyfarth, L.H., Sterling, P., Brainard, D.H., and Balasubramanian, V. (2011). Natural im- ages from the birthplace of the human eye. PLoS One 6, e20409. https://doi.org/10.1371/journal.pone.0020409.

West, E. R., & Cepko, C. L. (2022). Development and diversification of bipolar interneurons in the mammalian retina. Developmental biology, 481, 30–42. https://doi.org/10.1016/j.ydbio.2021.09.005

Xiang, M., Zhou, H., & Nathans, J. (1996). Molecular biology of retinal ganglion cells. Proceedings of the National Academy of Sciences of the United States of America, 93(2), 596–601. https://doi.org/10.1073/pnas.93.2.596