# Task-optimized retina-inspired CNN converges to biologically-plausible functionality

**Keith T. Murray**
MIT CSAIL
Cambridge, MA
ktmurray@mit.edu

**Brabeeba Wang**
MIT CSAIL
Cambridge, MA
brabeeba@mit.edu

**Nancy Lynch**
MIT CSAIL
Cambridge, MA
lynch@csail.mit.edu

## Abstract

Convolutional neural networks (CNN) are an emerging technique in modeling neural circuits and have been shown to converge to biological functionality in cortical circuits. This functionality has not been observed in retinal circuits. We sought to observe this convergence in retinal circuits by designing a biologically inspired CNN model of a motion-detection retinal circuit and optimizing it to solve a motion-classification task. The learned weights and parameters indicated that the CNN converged to direction-sensitive ganglion and amacrine cells (both of which are biologically-plausible mechanisms) and provided evidence that task-optimization is a fair method of building retinal models. The analysis used to understand the functionality of our CNN also indicates that biologically constrained deep learning models are easier to reason about their underlying mechanisms than traditional deep learning models.

## 1 Introduction

The retina serves as the first step in visual processing for the brain in nearly all animal species [1]. While it may serve only as a first step, the retina processes visual information using complex neural circuits that have yet to be fully understood or modeled [2]. As an understanding of what these complex retinal circuits do has evolved, so too have the models used to explain the mechanisms of these circuits. Linear-nonlinear (LN) models that linearly filter and nonlinearly transform visual data [3, 4, 5] have since been replaced by deep convolutional neural networks (CNNs) that are able to more fully predict retinal cell activity through filtering and nonlinearly transforming information through many layers with weights adjusted via backpropagation algorithms [6, 7, 8]. This switch came as CNNs were also shown to accurately predict cortical activity and converge to cortical representations in a variety of perceptual tasks after task-optimization training [9, 10].

CNNs that converge to cortical representations via task-optimization training provide a possible explanation for why neural circuits in the brain are organized in a particular fashion [11, 12]. While previous CNN modeling has provided evidence that CNNs can model retinal circuits, this modeling was driven by optimization that sought to fit retinal cell data and thus cannot provide theoretical evidence as to why retinal circuits are organized in their current fashion [6, 7, 8]. We sought to find this evidence through designing biologically-constrained CNNs that are trained to optimize performance in a motion-classification task. This approach also requires developing methods to understand the trained CNN, methods which have previously been elusive [13], and could provide new algorithms for solving the motion-classification tasks.

A motion-classification task was chosen because mammalian retinas have been shown to detect and respond to moving stimuli [14]. The task involved the CNN being presented a stimulus of moving dots and the CNN outputting which direction the dots were moving fastest. The CNN would fail the task if the direction outputted was not the fastest direction. The task-stimulus had two populations of dots, a leftward and rightward direction, and the output of the CNN was fed through a linear decoder that would designate which direction was fastest. The linear decoder was trained alongside the CNN. The design of the task drew information from previous motion-detection tasks in the systems neuroscience literature where monkeys are trained to identify a synchronous motion of dots among a population of dots [15].

Our CNN model was designed to emulate the connection patterns of previous motion-detection circuits in the retina [3, 5, 2]. The architecture of the model included excitatory bipolar cells, the first stage of information processing in the retina; ganglion cells, the final stage of information processing in the retina; and inhibitory amacrine cells, cells that modulate activity between the bipolar and ganglion cells [2]. After training via the Adam optimization algorithm [16] on the motion-classification task, the CNN was systematically ablated and the individual cell types were probed to classify their functionality using a combination of deep learning and system neuroscience techniques [17, 13].

We found that our CNN model displayed functionality similar to that observed in biological retinas and performed well on our motion-classification task. The model could reliably achieve above an 80% accuracy in difficult portions of the task environment and 100% accuracy in easy portions of the task environment. For the functionality of the model, simulated ganglion and amacrine cells emerged that were direction-selective. Direction-selective ganglion cells (DSGCs) are a hallmark in motion-detecting circuits observed in mammals [14] and their emergence via task-optimization suggests that they may be optimal in motion detection. Our CNN model also provides a mechanistic model for how to robustly discriminate motion that may have implications in autonomous vehicles [18] and retinal implant technology [19].

## 2 Methods

A two-directional motion-classification task was designed that required the CNN to output the direction corresponding to a population of dots that had a greater speed. The CNN processed this task-stimulus via three layers, each layer having 8 different cell types, and a residual connection that correspond to previous motion-detection literature [3, 5, 2]. A linear decoder mapped ganglion cell activity to correspond to left-, and right-direction classes that could then be used to determine the task accuracy of the CNN. The CNN was trained using the Adam optimization algorithm [16] and ablated. The functionality of the cell types in the ablated network were classified using deep visualization [17] and neurophysiology techniques [3, 4, 20].

### 2.1 Task Design

The task design was inspired by neuroscience literature investigating the function of the middle temporal visual area (MT) of the visual cortex [15]. In the literature, MT-related tasks involved motion discrimination between a population of randomly moving dots and dots moving synchronously in a particular direction. We drew inspiration from this aspect of population discrimination among moving dots but used speed as the discriminating factor instead of motion synchronicity among the dots.

Our task consisted of randomly placed dots on a field moving either left or right (Figure 1A). All dots that moved in a particular direction would be assigned a speed. To solve the task, the model would have to indicate which population of dots, right-directional or left-directional, had the greater speed. We hypothesized that a CNN that could solve this task would have to encode motion. The mechanisms that the CNN used to encode motion could then be further studied to describe what mechanisms a retina could use to encode motion.

The data set consisted of 1000 stimuli with each stimulus being an instance of the task. Each instance of the task was a 255 by 255 dimensional video consisting of 51 frames. For each stimuli, the placement of the dots was randomly initialized with an average of 16.67 dots being in one frame at

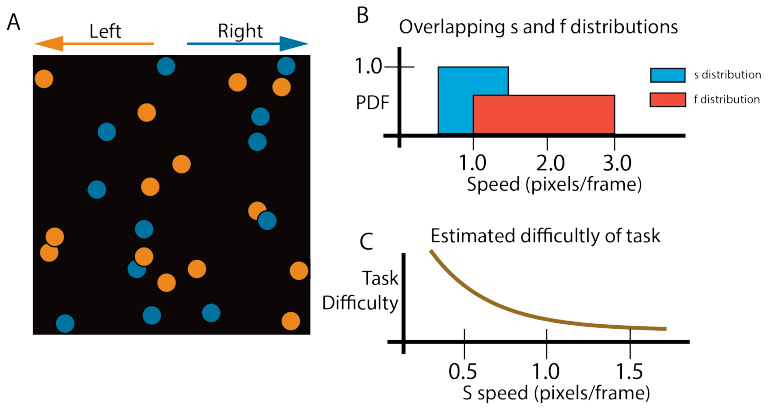


Figure 1: Illustrations of task design. (A) An example stimulus frame. Note, there are no colors in the actual stimulus. (B) The $s$ and $f$ distributions and the overlap between them. This overlap adds another layer of difficulty. (C) The estimated difficulty of the task decreases as the $S$ speed increase and creates a larger separation from $F$.

any given frame. Each dot was assigned a direction, right or left, with uniform probability. For each stimulus, the speed for the right- and left-directional was determined via the following method:

1. The slow direction ($S$) is chosen uniformly between left and right.
2. $S$ is assigned a speed by $s \sim U(0.5, 1.5)$.
3. The fast direction ($F$) is assigned a speed by $f = \alpha * s$, where $\alpha$ is the velocity multiplier.

The environment of the data set was chosen to have an $\alpha$ variable of 2 for training and testing the model (Figure 1B). The distribution of $s$ determined the difficulty of the stimulus. An $s$ near the lower end of the distribution is more difficult because the difference between $f$ and $s$ is smaller and less perceptible (Figure 1C).

### 2.2 Model Architecture and Training

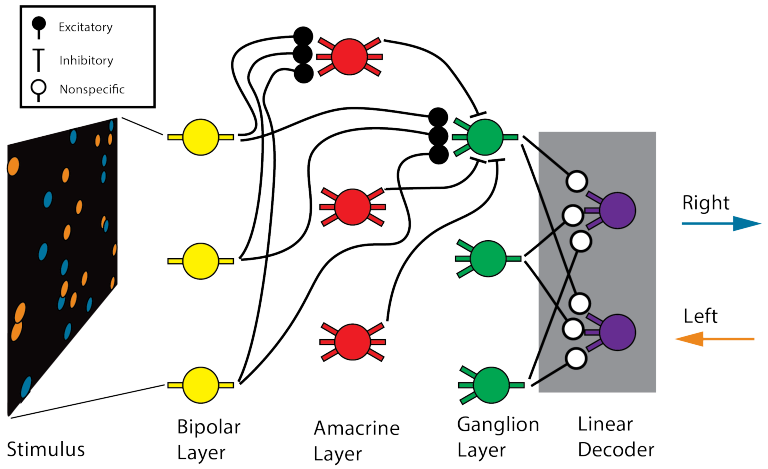


Figure 2: Diagram of the CNN and its connections. Only 3 cell types per layer are depicted here, but 8 per layer were used. Each cell type consists of a spatial convolution denoted by the connections, an internal temporal convolution, and a ReLU activation function.

The CNN consisted of 3 layers where each layer modeled the bipolar, amacrine, or ganglion layer, each layer was connected to the next layer, and an extra direct connection, a *residual* connection, existed between the bipolar and ganglion layers (Figure 2). While *residual* connections have shown

to increased performance of CNNs in object-recognition tasks [21], our motivation for inserting a residual connection in our CNN came from an observed direct connection between the bipolar and ganglion layers in the retina [3, 5]. Previous CNN models of the retina have not accounted for this residual connection [6, 7, 8]. By including a residual connection, a connection restraint constraint between cell types could be enforced that had previously not existed in other retinal CNN models [6, 7, 8] but has been biologically observed [3, 5, 2]. Connection enforcement was performed through constraining weights in the amacrine and ganglion layers to be positive (negative weights were zeroed out) and taking the additive inverse from the otherwise positive output of the amacrine layer.

Each cell layer in our CNN included 8 different cell types. For example, in the bipolar layer, there were 8 different bipolar cell types where each type had access to the same information as every other type. In the amacrine and ganglion layers, each cell type had connections to every cell type in the previous layer (e.g. amacrine cell type 0 had connections to bipolar cell types 0-7). While this connection pattern may not be biologically constrained, this connection pattern is a principled pattern for allowing emergent representations through learning.

Each cell type consisted of a spatial convolution, temporal convolution, and a rectified linear unit (ReLU) activation. For the bipolar layer, spatial convolutions are analogous to information received from photoreceptors, the fundamental unit of the retina [2], and for the amacrine and ganglion layers, spatial convolutions represent the connections between the layers. The residual connection took shape in the form of an additional spatial convolution between the bipolar and ganglion layers. The temporal convolution in the bipolar layer is analogous to the feedback from horizontal cells and in the amacrine and ganglion layers represents the decay of activity. The ReLU activation was chosen because of its use in previous modeling [3, 5, 6, 7, 8].

The output from the ganglion cell layer was fed through a linear decoder that gave two outputs: each corresponding to a direction (left or right). Feeding the stimulus through the CNN model created 51 outputs from the linear decoder. The linear decoder output with the greater activation on the last frame of the stimulus would indicate the direction the model predicted of the $F$ direction. The linear decoder was used because the output could be used for training via a cross-entropy loss function [22],

$$L(y, class) = -y[class] + \log(\textstyle\sum_j \exp(y[j])),$$

where $y$ is the output from the linear decoder and $class$ indicates whether the left or right direction is the $F$ direction.

Our CNN was trained over the course of 500 epochs using the Adam optimization algorithm [16] and a dropout rate of 40% via the PyTorch libraries [22].

### 2.3 Analysis Methods

The analysis of any deep learning model has been shown to be a difficult undertaking [13], but the analysis of our model was inspired by the analysis of previous biological retinal circuits and CNN models [3, 4, 7, 8]. The trained CNN's performance was first evaluated on a variety of environments where the $s$ and $\alpha$ variables were systematically manipulated to reveal how robust the CNN is and give insights into its functionality.

In line with previous deep learning literature working to understand how deep learning models work [23], the CNN was ablated to gain an understanding about the fundamental operations performed by each cell layer. Our methods for ablating differ than previous literature in that cell types were ablated instead of randomized parameters or some other algorithm [23]. Individual cell types were ablated and the model's accuracy was tested to establish which cell types were necessary.

After ablation, each cell type's functional role was established via functional observations in performance on test sets and reasoning about their respective deep visualization. The functional observations used in analyzing cell types included measuring the response amplitude of a cell, analyzing the accuracy of the ablated model when further ablated, and reasoning about the cell type's convolutional kernels. A deep visualization has been shown to be particularly useful in classifying the functional relevancy of neural units in neural networks [17] but has not previously been utilized in classifying the functional relevancy of neural units in CNN models of neural circuits. By treating the model as a static transformer, the individual pixels in the stimulus can be treated as parameters that are tuned

to maximize the response of a particular neuron. Deep visualizations are particularly useful when gaining an intuition about how the spatial convolution and temporal convolution profiles interact.

The final step in the analysis was to merge an understanding of how all cell layers function into an understanding as to how the CNN solves the task as a whole. Our objective is to explain the performance across various environments observed in the first step and to draw conclusions about how biologically plausible the functionality of the CNN is.

## 3 Results

To analyze our CNN, we first performed a **Full Model Analysis** and then proceeded to investigate the **Bipolar Functionality**, **Ganglion Functionality**, and **Amacrine Functionality**. With an intuition around each cell layer, we could hypothesize about the **Full Model Functionality**. For our most significant result, we found that the CNN emerged to have direction selective ganglion and amacrine cell types, a functionality that has been observed in biological retinas [14].

### 3.1 Full Model Analysis

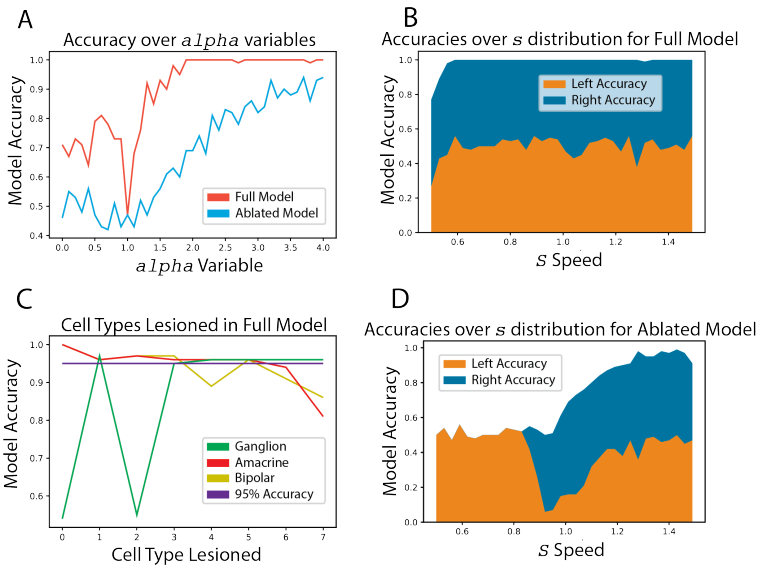


Figure 3: Analysis of the behavior and robustness of the full model. (A) Psychometric curve for the $\alpha$ parameter of full and ablated model. (B) Relative left and right accuracies for the full model over a variable $s$ parameter. The full model displays high accuracy over the full $s$ distribution. (C) Lesions of all the cell types to test their affect on the model. Any cell type that does not impact model accuracy below 95% was excluded. Ganglion cell types 0 and 2 are vitally important. (D) Relative left and right accuracies for the ablated model over a variable $s$ parameter.

The trained CNN was analyzed on its performance over different environmental conditions that were determined by the distribution of $s$ and the value of $\alpha$. Figure 3 suggests that the CNN is robust to environments that ranged over a variety of $\alpha$ parameters (Figure 3A) and nearly solved the task over the full distribution of $s$ (Figure 3B). The CNN failed to fully solve for the bottom of the distribution of $s$, indicating that the task's difficulty in low $S$ environments is high.

All cell types were lesioned (singularly ablated) to determine their significance in the functionality of the CNN (Figure 3C). The cell types in the CNN were shown to be highly redundant, except for ganglion cell types 0 and 2.

The ablation process for the CNN was performed by testing all combinations of bipolar and amacrine cell types while still maintaining ganglion cell types 0 and 2. The ablated CNN (ablated model) bipolar and amacrine combination consisted of bipolar cell type 4 and amacrine cell type 2 as they were shown to have the best performance. The performance of the ablated model was shown to be

more accurate toward the upper distribution of $s$ (Figure 3D), which must be due to the ease of the task towards that end of the distribution. The ablated models performance on various $\alpha$ environments was shown to be less that the full CNN (full model) but displayed similar trends of increasing performance in high $\alpha$ environments (Figure 3A).

### 3.2 Bipolar Functionality

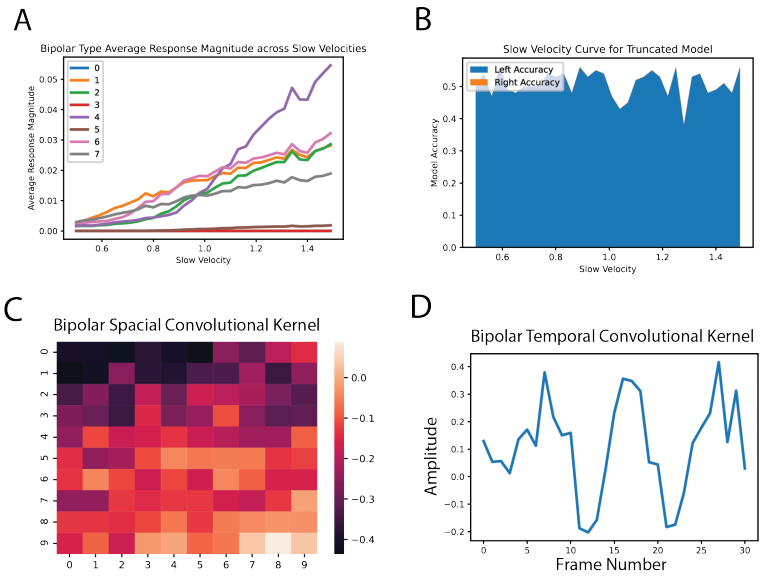


Figure 4: Analysis of the functionality of the bipolar cell for the ablated model. (A) Response amplitudes for all bipolar cells in the full model over the distribution of $s$. Bipolar cell type 4 was selected in the ablating process. (B) Relative left and right accuracies for the ablated model with the bipolar cell lesioned. The model biased towards left directions. (C) Spacial convolutional kernel for the ablated bipolar cell type. No center-surround profile observed. (D) Temporal convolutional kernel for the ablated bipolar cell type.

The functionality of the bipolar cell types was variable according to what part of the distribution of $s$ the task was in. The ablated bipolar cell type, bipolar cell type 4, was shown to have the lowest activation in the lower ranges of $s$ but increased dramatically in high ranges of $s$ (Figure 4A), indicating that bipolar cell types have activation thresholds in which they begin to respond. The ablated cell types activation magnitude in the higher ranges of $s$ corresponds to the ablated models performance (Figure 3D).

A bias in the full model was shown by the accuracy of the model when the bipolar cell type was lesioned (Figure 4B). This bias was towards the left direction and may be due to a bias in the linear decoder because no activity propagates through the model when the bipolar cell type is lesioned.

The convolutional kernels of the ablated bipolar cell type were shown to be relatively smooth, but not necessarily biologically plausible. Multiple experimental studies have shown center-surround effects in the bipolar cell [2, 1], however this effect was not observed in our CNN (Figure 4C). The temporal kernel of the ablated bipolar cell type appeared biologically plausible (Figure 4D) due to its directional preference towards the rightward direction, but no metrics were used to validate this notion of biological plausibility.

### 3.3 Ganglion Functionality

As observed in (Figure 3C), ganglion cell types 0 and 2 showed distinct activation and functionality during the task. Across the distribution of $s$, ganglion cell types 0 and 2 in the full model were the most active and showed increasing activity as $S$ speed increased (Figure 5A). When lesioned in the ablated model, the direction solved was dependent on the ganglion cell type lesioned (Figure 5B-C), indicating that these ganglion cell types are direction selective.

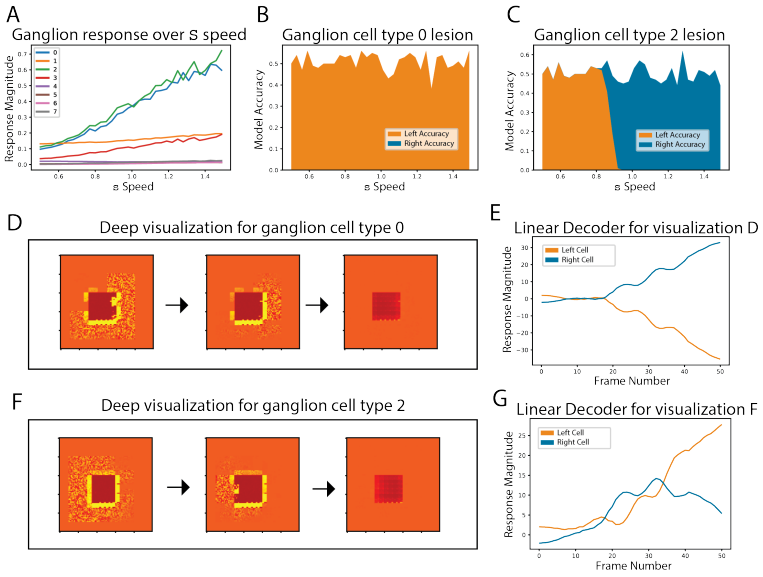


Figure 5: Analysis of the functionality for the ganglion cells for the ablated model. (A) Response amplitudes for all ganglion cell types in the full model over the distribution of $s$. Ganglion cell types 0 and 2 show the largest response which corresponds to their importance in the lesion analysis (3C). (B) Relative left and right accuracies for the ablated model with the ganglion cell type 0 lesioned. The model biased towards left directions indicating that ganglion cell type 0 solves for right directions. (C) Relative left and right accuracies for the ablated model with ganglion cell type 2 lesioned. The model biased towards right directions after bipolar activation, indicating the ganglion cell type 2 solve for left directions. (D) Deep visualization for ganglion cell type 2. This deep visualization displays activity that fades toward the left direction, indicating direction selectivity. (E) Activity of the linear decoder when the deep visualization (D) was input. This activity confirms that ganglion cell type 2 is direction selective for the left direction. (F) Deep visualization for ganglion cell type 0. The deep visualization displays activity that fades toward the right direction, indicating direction selectivity. (G) Activity of the linear decoder when the deep visualization (F) was input. This activity confirms that ganglion cell type 0 is direction selective for the right direction.

The complex connectivity between the ganglion layer and the amacrine and bipolar cell types made utilizing deep visualizations [17] necessary to determine the response patterns of the ganglion cell types. A parameterized tensor initialized with random weights was fed through the ablated model and underwent optimization via the Adam algorithm [16] to maximize activity of each ganglion cell in the ablated model. The resulting deep visualizations of the ganglion cell types revealed a direction selective functionality (Figure 5E-G) and a center-surround organization (Figure 5D-F). These results implicate the ganglion cells as direction selective and the crucial mechanism by which the CNN solves the task.

### 3.4 Amacrine Functionality

The functionality of the amacrine cell type in the ablated model was revealed to be linked to the direction selectivity observed in the ganglion cell types. The ablated amacrine cell type, amacrine cell type 2, was shown to be the amacrine cell type that displayed the largest magnitude of response over the distribution of $s$ (Figure 6A). This response of the ablated amacrine cell type displays the same pattern in relation to the non-ablated amacrine cell types as the ablated bipolar cell type display in relation to the non-ablated bipolar cell types (Figure 4A). This suggests that bipolar and amacrine cell types are matched according to their response magnitudes during different ranges of $s$.

The lesion results of the ablated amacrine cell type revealed that it was critical in solving for left direction $F$ speeds (Figure 6B). This is supported by the spacial convolutional kernel of the ablated amacrine cell type which displays an increasing magnitude of suppression to stimuli that move in the right direction (Figure 6C). When the $F$ direction is left, stimuli that move in the right direction

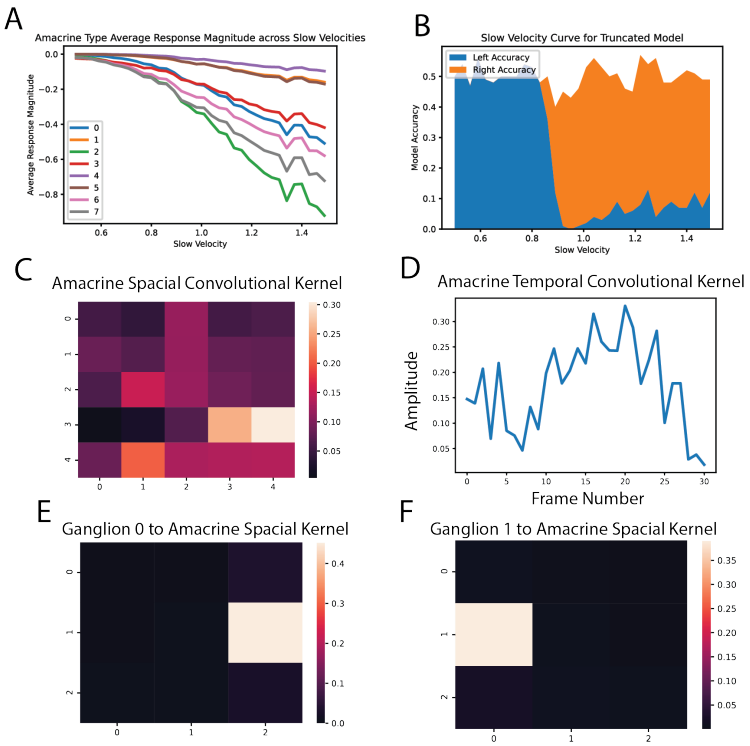


Figure 6: Analysis of the functionality of the amacrine cell for the ablated model. (A) Response amplitudes for all amacrine cells in the full model over the distribution of $s$. Amacrine cell type 2 was selected in the ablation process. (B) Relative left and right accuracies for the ablated model with the amacrine cell type lesioned. The model biased towards the right direction after bipolar activation indicating that the amacrine cell is right direction selective. (C) Spacial convolutional kernel for the ablated amacrine cell type. There is an increase in weights toward the right direction confirming the selectivity apparent in (B). (D) Temporal convolutional kernel for the ablated amacrine cell type. (E,F) Spacial convolutional kernel between ganglion cell type 0, ganglion cell type 2, respectively, and the ablated amacrine cell type. The strong weight on the left of ganglion cell type 2 and the increase in activity of (D) indicates that the ablated amacrine cell type suppressed activity of rightward moving stimuli.

would increase the inhibitory effects that the amacrine cell type would have on the right DSGC type (Figure 6D-F). Without the ablated amacrine cell type, right directions cannot be suppressed to allow for the ablated model to solve the right direction. This functionality is visible in Figure 3D as the left-direction accuracy increases as $s$ increases since a faster $S$ would provide more activity to the amacrine cell type that could intern inhibit the right DSGC type.

### 3.5 Full Model Functionality

The results of the analysis on the ablated model indicate that the ablated model solves the task in the following functionality:

1. The bipolar cell thresholds stimulus of speeds. After the threshold, activity is propagated to the amacrine and ganglion layers.

2. The amacrine cell activates after receiving high enough magnitudes of activity from the bipolar cell to suppress stimuli moving in the right direction.

3. Activity in the DSGCs accumulates until the end of the task where the ganglion cell type with the greatest activity is decoded by the linear decoder and that corresponding direction is chosen for the $F$ direction.

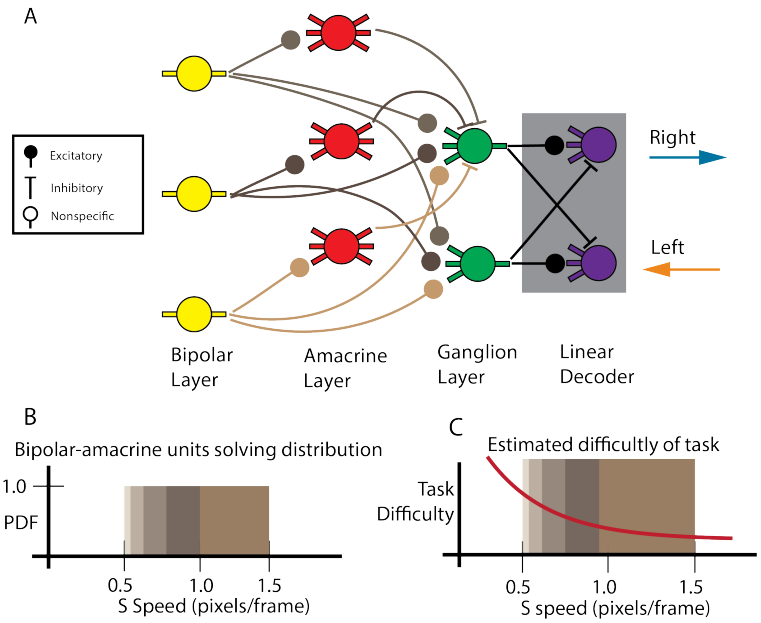


Figure 7: Explanation of the proposed organization and functionality of the full CNN. (A) Diagram of the proposed organization of the full CNN. While the diagram only displays 3 bipolar-amacrine units, there is evidence for 5 in total in the CNN. (B) Plot of the proposed functionality of the CNN on the distribution of $s$. Each bipolar-amacrine unit solves a portion of the distribution according to the activation-threshold of the bipolar cell. (C) Plot of how this proposed functionality of the CNN is due to the difficulty of the task. As the difficulty increases, the area of the distribution covered by a bipolar-amacrine unit decreases.

This interpretation of the CNN lends itself to be organized in bipolar-amacrine units (Figure 7A). Each bipolar-amacrine unit responded to a subsection of the $s$ distribution (Figure 7B). The area that each of these units responded to modulated by the difficulty of the task (Figure 7C). As the difficulty decreases, these bipolar-amacrine solved a wider range of the distribution of $s$. The activity of the units that solved lower distributions of $s$ could then be dominated by increased activity of units that solved higher distributions (Figure 4A and Figure 6A).

## 4 Discussion

These results indicate that the CNN converged to a biologically plausible organization of the retina (Figure 5D-F). This is surprising given that task-optimized deep learning models have only been shown to converge to biological representations and functionality in sensory cortical neural circuits [9, 10]. While the retina is a sensory neural circuit, the neurons and motifs observed have been much different than those in the cortex [2]. These results, however, do not indicate that backpropagation algorithms (e.g. Adam [16]), are the mechanisms neural circuits use to organize. The relationship backpropagation algorithms have with the brain remains inconclusive [24]; thus, these results indicate only that backpropagation for task-optimization is a sufficient mechanism to organize a CNN in the manner observed in the retina.

An objection to this conclusion may come from the lack of a center-surround receptive field in the ablated bipolar cell type (Figure 4C). Center-surround receptive fields have been observed in other bipolar layers of previous CNN models of the retina [6, 7, 8]. However, those models utilized natural stimuli collected from typical environments of the animal from which their retinal ganglion cells were used for fitting. It is possible that our model could converge to center-surround receptive fields if the stimulus used was less artificial. This presents a possible future direction for the model.

Biological retinas have been observed to adapt quickly to the variation in its environment [2, 1]. Our model does not adapt to the $S$ speed it is in, but instead has different mechanisms for different environments. This may be due to the lack of dynamical mechanisms in deep learning models and

by including dynamical mechanisms, our CNN may be more adaptable [20]. A future direction could be to incorporate dynamical aspects into the CNN via recurrent layers and analyze what their contribution is to solving various environments.

While it has previously been difficult to probe the functionality of deep learning models [13], the analysis of our CNN was fairly simple. This may be due to the biologically inspired constraints applied during training and the biologically inspired organization. By designing deep learning models using more biologically inspired constraints, deep learning models could become more explainable as more techniques from computational and systems neuroscience could be used during analysis. A future direction for our work could be to develop a more systematic methodology for analyzing the functionality and using insights into the functionality to create an algorithmic understanding of our CNN.

The final understanding of the model (Section 3.5) also provides an understanding of the mechanisms of motion-detection circuits in the retina. These circuits have been shown to be reliable and robust to noise [14], a much needed quality in autonomous vehicles [18]. Through an understanding of how our CNN works, it may be possible to engineer new computer vision techniques that are capable of operating in many environments and malicious attacks. This understanding may also inform how to build better retinal implants as this model is biologically plausible and electrically implementable [19]. An increased understanding of the retina through modeling has not only deep theoretical implications about the organizations of neurons, but engineering applications to create more explainable and robust technologies.

## References


[1] Tom Baden, Thomas Euler, and Philipp Berens. Understanding the retinal basis of vision across species. *Nature Reviews Neuroscience*, 21(1):5–20, 2020.

[2] Tim Gollisch and Markus Meister. Eye smarter than scientists believed: neural computations in circuits of the retina. *Neuron*, 65(2):150–164, 2010.

[3] Bence P Ölveczky, Stephen A Baccus, and Markus Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003.

[4] Toshihiko Hosoya, Stephen A Baccus, and Markus Meister. Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77, 2005.

[5] Stephen A Baccus, Bence P Ölveczky, Mihai Manu, and Markus Meister. A retinal circuit that computes object motion. *Journal of Neuroscience*, 28(27):6807–6817, 2008.

[6] Lane T McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A Baccus. Deep learning models of the retinal response to natural scenes. *Advances in neural information processing systems*, 29:1369, 2016.

[7] Niru Maheswaranathan, Lane T. McIntosh, Hidenori Tanaka, Satchel Grant, David B. Kastner, Josh B. Melander, Aran Nayebi, Luke Brezovec, Julia Wang, Surya Ganguli, and Stephen A. Baccus. The dynamic neural code of the retina for natural scenes. *bioRxiv*, 2019. doi: 10.1101/340943. URL `https://www.biorxiv.org/content/early/2019/12/17/340943`.

[8] Hidenori Tanaka, Aran Nayebi, Niru Maheswaranathan, Lane McIntosh, Stephen A Baccus, and Surya Ganguli. From deep learning to mechanistic understanding in neuroscience: the structure of retinal prediction. *arXiv preprint arXiv:1912.06207*, 2019.

[9] Daniel LK 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*, 111(23):8619–8624, 2014.

[10] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

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

[12] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, pages 1–13, 2020.

[13] David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current opinion in neurobiology*, 55:55–64, 2019.

[14] Wei Wei. Neural mechanisms of motion processing in the mammalian retina. *Annual review of vision science*, 4:165–192, 2018.

[15] William T Newsome and Edmond B Pare. A selective impairment of motion perception following lesions of the middle temporal visual area (mt). *Journal of Neuroscience*, 8(6):2201–2211, 1988.

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[17] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*, 2015.

[18] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1625–1634, 2018.

[19] Alice T Chuang, Curtis E Margo, and Paul B Greenberg. Retinal implants: a systematic review. *British Journal of Ophthalmology*, 98(7):852–856, 2014.

[20] Yusuf Ozuysal and Stephen A Baccus. Linking the computational structure of variance adaptation to biophysical mechanisms. *Neuron*, 73(5):1002–1015, 2012.

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[22] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019. URL `http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf`.

[23] Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the state of neural network pruning? *arXiv preprint arXiv:2003.03033*, 2020.

[24] Timothy P Lillicrap, Adam Santoro, Luke Marris, Colin J Akerman, and Geoffrey Hinton. Backpropagation and the brain. *Nature Reviews Neuroscience*, 21(6):335–346, 2020.