A Deep Learning Model of the Retina

Lane McIntosh and Niru Maheswaranathan
Neurosciences Graduate Program, Stanford University
Stanford, CA
{lanemc, nirum}@stanford.edu

Abstract

The retina represents the first stage of processing our visual world. In just three layers of cells, the retina transduces a raw electrical signal that varies with the number of photons that hit our eye to a binary code of action potentials that conveys information about motion, object edges, direction, and even predictions about what will happen next in the world. In order to understand how and why the retina performs particular computations, we must first know what it does. However, most models of the retina are simple, interpretable models of the retina's response to simple parametric noise, and there are no existing models that can accurately reproduce ganglion spiking during natural vision. In this paper we use convolutional neural networks to create the most accurate model to-date of retinal responses to spatiotemporally varying binary white noise. We also investigated how well convolutional neural networks in general can recover simple, sparse models on high dimensional data. Our results demonstrate that convolutional neural networks can improve the predictions of retinal responses under commonly used experimental stimuli, and provide a basis for predicting retinal responses to natural movies when such a dataset becomes available.

1. Introduction

One crucial test of our understanding of biological visual systems is being able to predict neural responses to arbitrary visual stimuli. As the first and most easily accessible stage of processing in our visual pathway, the retina is a logical first system to understand and predict. In only three layers of cells, the retina compresses the entire visual scene into the sparse responses of only a million output cells. Until the 1990s, these three layers of cells were thought to act only as a linear "prefilter" for visual images [6, 10], however retina studies in the last two decades have demonstrated that the retina performs a wide range of nonlinear computations, including object motion detection [18], adaptation to complex spatiotemporal patterns [12], encoding spatial structure as spike latency [9], and anticipation of periodic stimuli [22]. Despite the sophistication of these observations, retinal ganglion cells are still mostly described by their linear receptive field [6], and models of retinal output usually consist of this linear filter followed by a single, monotonic nonlinearity.

While simple linear-nonlinear models of the vertebrate retina provide good approximations for how the retina responds to simple noise models like spatiotemporal Gaussian white noise [1] or binary white noise [20, 21], there are no known models that can explain even half of neurons' variance during natural vision [7]. Despite this, retinal spiking during natural vision is highly stereotyped [2, 3, 5].

Here we fit convolutional neural networks to experimental data of vertebrate retinas responding to spatiotemporal binary noise as a proof of principle that convolutional neural networks can better predict retinal responses even in a regime where simpler models already perform reasonably well. This work lays the foundation for establishing a model that can accurately predict retinal responses to arbitrary stimuli, including both simple parametric noise and natural images. An accurate model of the retina is essential for understanding exactly what early computations our visual system performs, and choosing the space of convolutional neural network models allows us to compare our model architecture with state-of-the-art convolutional neural networks performing classification vision tasks.

1.1. Problem Statement

Our problem amounts to transforming grayscale natural movie clips into a single scalar number for each time point t and each neuron n. Data pairs (X, y) will consist of input $X \in R^3$, with dimensions 100 pixels x 100 pixels x 40 frames, and labels $y \in \{[0, 1]\}^n$, representing the probability p_t^n of spiking for each neuron after the 40 frames of X_t. Here the choice of 40 frames arises from the maximum duration of 400 ms for a retinal ganglion cell receptive field, and a 100 Hz frame rate stimulus.

The grayscale movie stimulus consists of 100 pixel by 100 pixel spatiotemporal binary white noise, which is commonly used for system identification since this stimulus in-

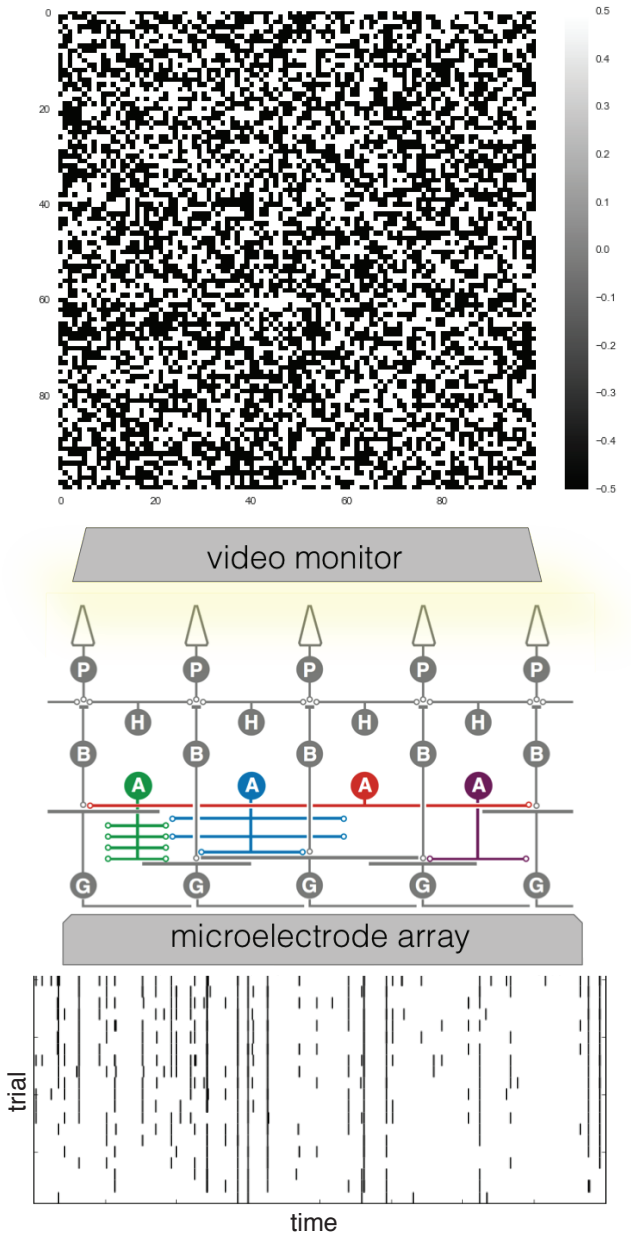

Figure 1. Top: Example binary white noise frame presented by NM to a salamander's retina in the Baccus laboratory at Stanford University. Middle: A cartoon of the experimental setup where a CRT video monitor presents the stimulus to an *ex vivo* retina situated on a 64-channel microelectrode array. Bottom: Example retinal ganglion spike rasters after presenting 20 trials of the same 80 s stimulus to the retina. Repeats were used to estimate retinal noise around empirical spike probabilities, but were not used for training the network.

cludes all spatial and temporal frequencies (limited by the spatial and temporal resolution of the experimental setup) and $2^{100\times100}$ possible patterns in a given frame (Figure 1).

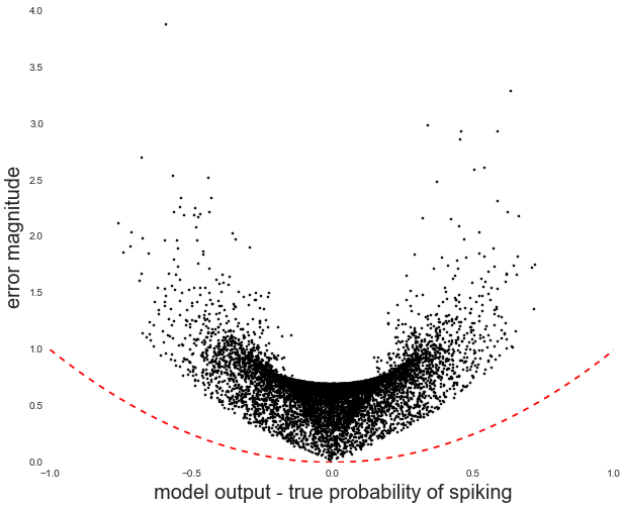

Figure 2. Loss versus error between true normalized firing rates and model's predictions for cross entropy (black) and mean squared error (red) loss functions. Since higher normalized firing rates are expected to be noisier, the cross entropy loss penalizes the same magnitude error less when the firing rate is higher.

Labels correspond to normalized number of spikes per 10ms bin, convolved with a 10 ms standard deviation Gaussian. We convolve the raw spike trains with a Gaussian smoothing window to reflect the noise inherent in any single trial of retinal ganglion cell responses, where the width of the Gaussian scales with the estimated noise of the retina. Since the value per bin represents the probability of spiking, skipping any smoothing step is equivalent to having complete certainty of spiking in bins where a spike was recorded, and zero probability of spiking in neighboring bins. Here we will use the terms probability of spiking, normalized number of spikes, and normalized firing rate interchangeably.

We minimize the cross entropy loss (or, equivalently, the logistic loss) $L = -\frac{1}{T}\sum_{t=0}^{T}\left[y_t \log \hat{y}_t + (1-y_t)\log(1-\hat{y}_t)\right]$ with L1 regularization on the network weights to promote sparse features. The cross entropy loss corresponds to the negative log likelihood of the true spiking probabilities given the model's predictions. Here y_t is the true smoothed, normalized number of spikes in bin t, while $\hat{y}_t$ is the network's output prediction, and T is the total data duration in frames. The cross entropy loss has the nice property of penalizing large errors more strongly than mean squared error (Figure 2). Under preliminary experiments using mean squared error, the network was often content to always predict the mean firing rate.

While minimizing the cross entropy loss, we will evaluate training, validation, and test accuracy according to the Pearson correlation coefficient between the true normalized firing rates and the model's predictions.

1.2. Technical Approach

1.2.1 Visual stimulus

We generated the stimulus via Psychtoolbox in MATLAB [4, 15], which can precisely control when the stimulus computer delivers each frame down to millisecond precision. The CRT video monitor was calibrated using a photodiode to ensure the linearity of the display.

1.2.2 Electrophysiology recordings

The spiking responses of salamander retinal ganglion cells were recorded with a 64 channel multi-electrode array (Multichannel Systems) in Professor Stephen Baccus' Neurobiology laboratory at Stanford University. Since a single cell may be recorded by multiple channels, we sorted the spikes to each cell using PCA and cross-correlation methods. All experiments were performed according to the procedures approved by the Stanford University Administrative Panel on Laboratory Animal Care.

1.2.3 Modeling

All simulations and modeling were done in Python using the CS 231n convolutional neural network codebase which we modified to support analog time series prediction instead of classification, mean squared error and cross entropy loss function layers, parameterized logistic function layers, and arbitrary movie input. Although we initially invested a substantial amount of time into developing this work with Caffe [13], modifying the data layer to support movies ended up being nontrivial (in fact, the 4th google search result for "Caffe convolutional neural network movie data" is our own github repository).

Convolutional neural network architectures were inspired by retinal anatomy, where we experimented with architectures that had the same number of layers as cell types (5 layers), or had the same number of layers as cell body layers (3 layers). We also tried 2 layer convolutional neural network models and simple 1 layer models (e.g., LN models) for comparison. Figure 3 depicts an early convolutional neural network architecture that we tested.

1.3. Results

1.3.1 Performance margin

An important consideration of any modeling endeavor is to first understand both the current benchmarks to surpass (Table 1) as well as the best possible performance, which in our case is bounded by the trial-to-trial variability of the retinal responses.

The state-of-the-art performance of retinal models varies drastically on how sophisticated the input stimulus is. For one of the simplest possible stimuli, spatially uniform binary white noise, Pillow and colleagues [20, 21] have explained 90% of the variance in retinal responses, while for non-parameterized naturalistic stimuli the best model in the literature cannot explain even half of the variance.

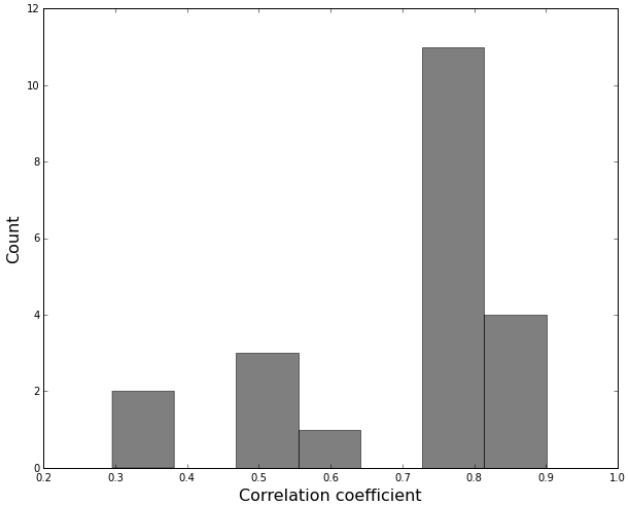

Figure 4. The correlation coefficient of a retinal ganglion cell response during a given trial to its overall normalized firing rate (averaged over many trials) ranges from as low as 30% in some cells to as high as 90% in others.

Since this is the first application of convolutional neural networks to modeling retinal responses, as a proof of principle we start from a commonly used stimulus, spatially varying binary white noise, that is relatively simple yet is capable of representing a wide range of possible spatiotemporal patterns. Since this stimulus is high contrast, retinal noise is also reduced relative to more naturalistic stimuli.

For this class of spatially varying stimuli, currently used retina models achieve around 40% explained variance on the most reliable cells (unpublished). However, the correlation of a given cell's single trial response to its overall normalized firing rate, obtained from averaging many trials, can be as high as 90% (Figure 4).

1.3.2 What is enough data?

Since experimental data is difficult to obtain, an important question is how much data we need to train the convolutional neural networks on retinal responses.

As a lower bound to how much data we need, we fit one layer networks with parameterized nonlinearities to varying amounts of training data (from 1 minute to 50 minutes) and evaluated the model accuracy using the Pearson's correlation coefficient between model predictions and the true normalized firing rate on held-out data (Figure 5). This is a lower bound on the data we need because presumably as

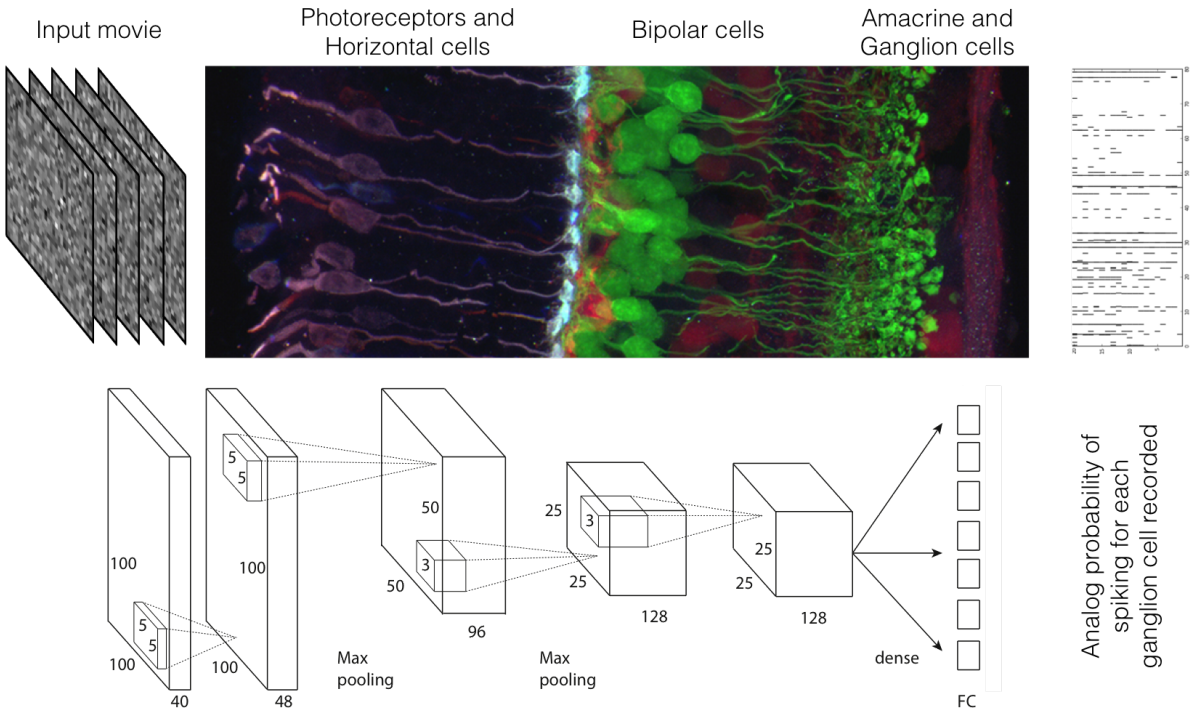

Figure 3. Top: An example 4 frame white noise movie clip is presented to a genetically labeled retina from Josh Morgan in Rachel Wong's lab at UW [17], with an example spike raster representing the binary output of the ganglion cells. Bottom: An early convolutional neural network architecture with one layer per cell type. The number of filters per layer were chosen to be greater than the number of functional subtypes per cell type. The number of filters per layer increases with each layer similarly to how the number of subtypes per cell type increases with each cell layer in the retina. In this model architecture, the input layer is 100 pixels by 100 pixels by 40 frames, followed by four convolutional layers (each with a ReLU layer) and two max pooling steps. 128 filters was chosen as a rough upper bound on the number of cell types in the retina. Output layer is fully connected with a cross entropy loss, and each scalar output represents the predicted probability of a particular retinal ganglion cell spiking.

Study	Area	Stimulus	Model	Metric	Performance
Keat *et al*. 2001 [14]	retina	full-field Gaussian white noise	LN	Average # spikes	0.47
Lesica and Stanley 2004 [16]	LGN	Indiana Jones + noise	LNP	Correlation coeff.	0.6
David *et al*. 2004 [8]	V1	Simulated natural images	PSFT	Explained variance	0.20
Pillow *et al*. 2005 [20]	retina	full-field binary white noise	generalized IF	Explained variance	0.91
David and Gallant 2005 [7]	V1	natural images	2nd order Fourier power model	Explained variance	0.4
Carandini *et al*. 2005 [6, 23]	retina	full-field Gaussian white noise	LNP	Explained variance	0.81
Pillow *et al*. 2008 [21]	retina	full-field binary white noise	GLM	Explained variance	0.9
Ozuysal and Baccus 2012 [19]	retina	full-field Gaussian white noise	LNK	Correlation coefficient	0.88

Table 1. Summary of the recent literature in predicting early visual neuron responses. LN=linear-nonlinear, LNP=linear-nonlinear-poisson, PSFT=phase-separated Fourier model, IF=integrate-and-fire, GLM=generalized linear model, LNK=linear-nonlinear-kinetic model.

we increase the number of layers, the model complexity as measured by the number of parameters increases, and so the amount of data needed to prevent overfitting increases.

Unfortunately, we found that even after 50 minutes of

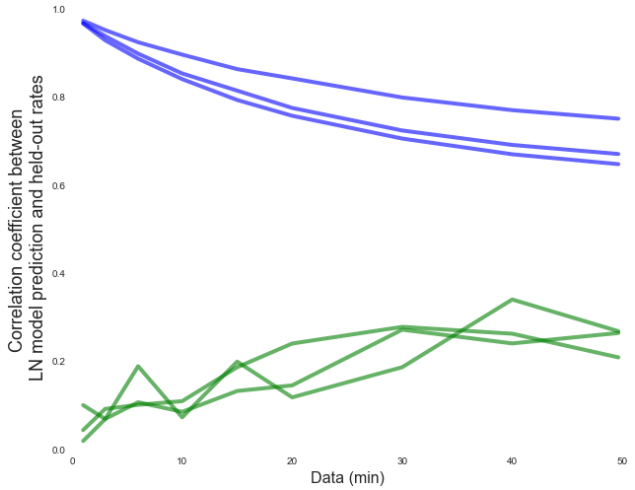

Figure 5. Correlation coefficient between the LN model prediction and held-out normalized firing rates as a function of the amount of training data used in minutes. Blue is training accuracy and green is validation accuracy, and each line represents a different cell. Training on all available data yielded an average correlation coefficient of around 0.7.

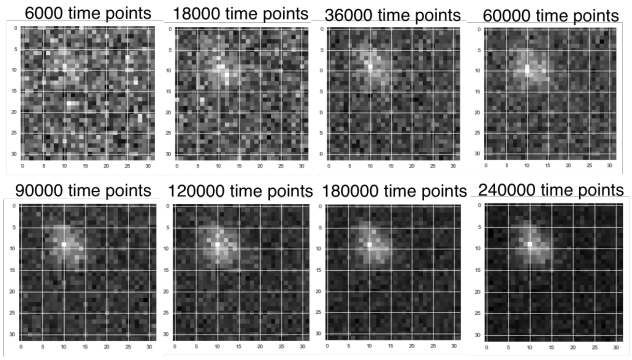

Figure 6. Linear filter estimated from reverse correlation between the binary white noise stimulus and the normalized rates. Reverse correlation required 36,000 time points (6 minutes) to recover the LN model with the same fidelity as the LN fit to a convolutional neural network after 5,000 time points (50 seconds).

training data, the validation and training accuracy did not converge despite the one layer network (e.g. LN model) learning sparse filters that looked qualitatively identical to the reverse correlation between the held-out data and the held-out true responses.

This could be related to the ability of linear mechanisms to "fool" convolutional neural networks [11], since a small amount of noise in the stimulus across its $100 \times 100 \times 40$ values could produce a large change in network's prediction.

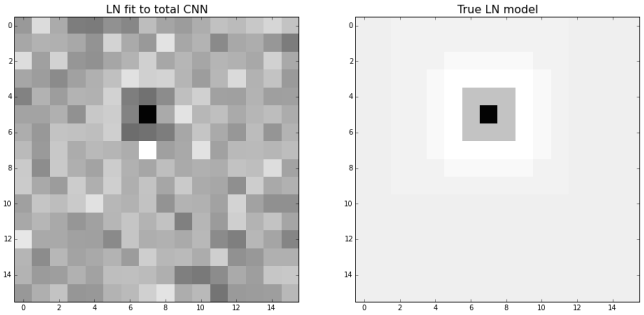

Figure 8. The LN fit to a 3 layer convolutional neural network trained on 5,000 time points (50 seconds) of simulated data generated from a sparse LN model.

1.3.3 Weight scales for sparse data

A common recommendation in the tuning of convolutional neural network hyperparameters to set the model's initialization weight scale to $2/N$, where N is the total number of inputs to the first layer. In our case, this translates to a weight scale of $\mathcal{O}(10e^{-7})$. Surprisingly, we found that models with small weight scale initializations consistently had predictions with much lower variance than the true responses (Figure 7). Instead, we had to select weight scales of $\mathcal{O}(1)$. This is likely due to the sparse nature of the underlying retinal ganglion cell receptive fields, since only a small number of weights are actually responsible for generating an accurate prediction.

1.3.4 Capacity for fitting small models

Given that the existing literature typically models the retina as a simple LN model, we investigated how well three and five layer convolutional neural networks could recover simple LN models after being trained on generated data. We found that these more complex convolutional neural networks could capture the LN model with very little data (Figure 8). Even though individual filters in the three and five layer networks did not resemble the original LN model used to generate this data, the overall LN fit nonetheless closely matches the true LN model after even 50 seconds of data.

1.3.5 Performance

We were able to achieve a new state-of-the-art performance of almost 80% correlation coefficient with a three layer convolutional neural network trained on retinal responses to spatially varying binary white noise. This is nearly twice the correlation coefficient achieved by simpler models, such as the LN model, that are currently used in the literature to model the retina (Figure 9).

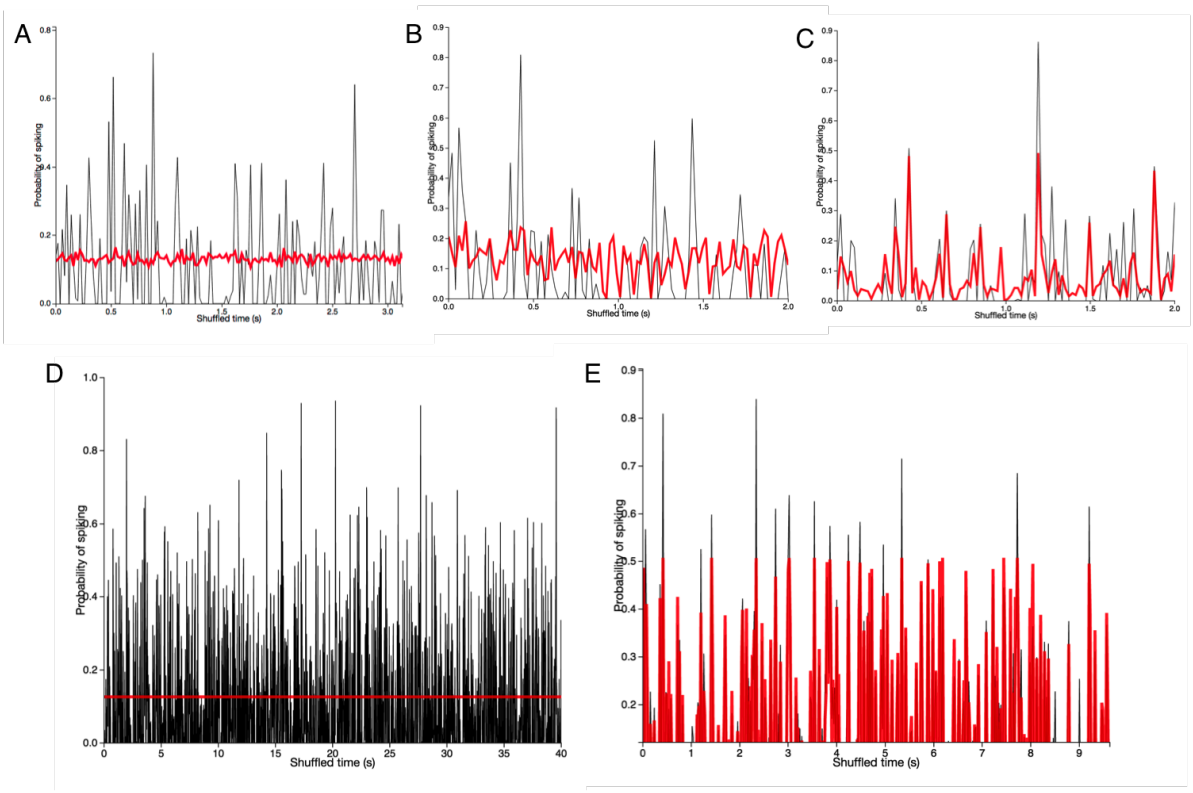

Figure 7. A, B, D: Examples of model predictions (red) versus true probability of spiking from simulated data (black) at different temporal scales when the the convolutional neural network is initialized with a weight scale smaller than $\mathcal{O}(1)$. C, E: Convolutional neural network predictions (red) versus true probability of spiking (black) with weight scale $\mathcal{O}(1)$. The (black) underlying data was generated from a LN model with sparse filters. The simulated data was convolved with a 10 ms Gaussian smoothing window to reproduce the level of noise experienced in the real data.

1.4. Conclusion

This work provides a strong proof-of-principle that convolutional neural networks can and do provide more accurate models of retinal responses compared to simpler, single stage models. Moreover, convolutional neural networks still retain some of the main advantages of simpler models since the overall linear and nonlinear properties of these more complicated networks can still be evaluated by fitting an LN model to the convolutional neural network. While these results apply to spatially varying binary white noise, these models can be easily trained on arbitrary stimuli once more datasets are collected.

This convolutional neural network model of the retina will be useful not only for researchers studying the retina, but can also be used for high-level vision scientists and computer scientists interested in having a virtual model of the retina to ask how the output of early stages of biological vision is used and transformed by the brain to yield the ability to perform complicated tasks like recognizing faces and quickly detecting objects where humans still retain state-of-the-art performance.

2. Acknowledgements

Special thanks to Ben Poole for helpful discussions and advice.

References

[1] S. A. Baccus and M. Meister. Fast and slow contrast adaptation in retinal circuitry. *Neuron*, 36(5):909–919, 2002.

[2] M. J. Berry and M. Meister. Refractoriness and neural precision. *The Journal of Neuroscience*, 18(6):2200–2211, 1998.

[3] M. J. Berry, D. K. Warland, and M. Meister. The structure and precision of retinal spike trains. *Proceedings of the National Academy of Sciences*, 94(10):5411–5416, 1997.

[4] D. H. Brainard. The psychophysics toolbox. *Spatial vision*, 10:433–436, 1997.

[5] D. A. Butts, C. Weng, J. Jin, C.-I. Yeh, N. A. Lesica, J.-M. Alonso, and G. B. Stanley. Temporal precision in the neural code and the timescales of natural vision. *Nature*, 449(7158):92–95, 2007.

[6] M. Carandini, J. B. Demb, V. Mante, D. J. Tolhurst, Y. Dan, B. A. Olshausen, J. L. Gallant, and N. C. Rust. Do we know what the early visual system does? *The Journal of Neuroscience*, 25(46):10577–10597, 2005.

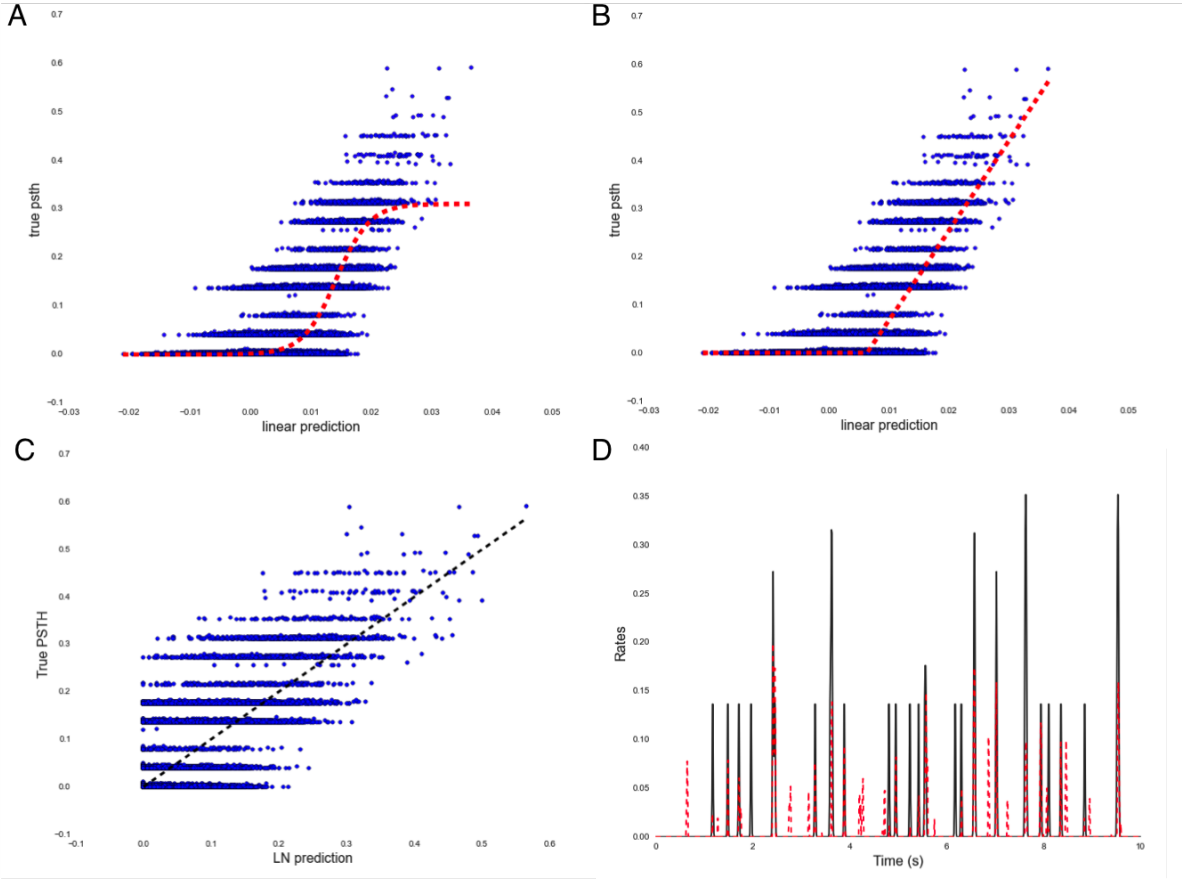

Figure 9. Performance of existing models. A: The best sigmoid nonlinearity for an LN model fit to retinal responses to binary white noise. B: The best rectifying linear nonlinearity, parameterized by its slope and bias, for an LN model fit to retinal responses to binary white noise. C: True firing rates versus the LN model predictions. D: 10 seconds of true firing rate (black) and the LN model's predictions (red).

[7] S. V. David and J. L. Gallant. Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, 16(2-3):239–260, 2005.

[8] S. V. David, W. E. Vinje, and J. L. Gallant. Natural stimulus statistics alter the receptive field structure of v1 neurons. *The Journal of Neuroscience*, 24(31):6991–7006, 2004.

[9] T. Gollisch and M. Meister. Rapid neural coding in the retina with relative spike latencies. *Science*, 319(5866):1108–1111, 2008.

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

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[12] T. Hosoya, S. A. Baccus, and M. Meister. Dynamic predictive coding by the retina. *Nature*, 436(7047):71–77, 2005.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[14] J. Keat, P. Reinagel, R. C. Reid, and M. Meister. Predicting every spike: a model for the responses of visual neurons. *Neuron*, 30(3):803–817, 2001.

[15] M. Kleiner, D. Brainard, D. Pelli, A. Ingling, R. Murray, and C. Broussard. Whats new in psychtoolbox-3. *Perception*, 36(14):1, 2007.

[16] N. A. Lesica and G. B. Stanley. Encoding of natural scene movies by tonic and burst spikes in the lateral geniculate nucleus. *The Journal of Neuroscience*, 24(47):10731–10740, 2004.

[17] J. Morgan and R. Wong. Visual section of the mouse retina. http://wonglab.biostr.washington.edu/gallery.html.

[18] B. P. Ölveczky, S. A. Baccus, and M. Meister. Segregation of object and background motion in the retina. *Nature*, 423(6938):401–408, 2003.

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

[20] J. W. Pillow, L. Paninski, V. J. Uzzell, E. P. Simoncelli, and E. Chichilnisky. Prediction and decoding of retinal ganglion cell responses with a probabilistic spiking model. *The Journal of Neuroscience*, 25(47):11003–11013, 2005.

[21] J. W. Pillow, J. Shlens, L. Paninski, A. Sher, A. M. Litke, E. Chichilnisky, and E. P. Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

[22] G. Schwartz, R. Harris, D. Shrom, and M. J. Berry. Detection and prediction of periodic patterns by the retina. *Nature neuroscience*, 10(5):552–554, 2007.

[23] K. A. Zaghloul, K. Boahen, and J. B. Demb. Contrast adaptation in subthreshold and spiking responses of mammalian y-type retinal ganglion cells. *The Journal of neuroscience*, 25(4):860–868, 2005.