LEARNING A NEURAL RESPONSE METRIC FOR RETINAL PROSTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Retinal prostheses for treating incurable blindness are designed to electrically stimulate surviving retinal neurons, causing them to send artificial visual signals to the brain. However, electrical stimulation generally cannot precisely reproduce normal patterns of neural activity in the retina. Therefore, an electrical stimulus must be selected that produces a neural response as close as possible to the desired response. This requires a technique for computing a distance between the desired response and the achievable response that is meaningful in terms of the visual signal being conveyed. Here we propose a method to learn such a metric on neural responses, directly from recorded light responses of a population of retinal ganglion cells (RGCs) in the primate retina. The learned metric produces a measure of similarity of RGC population responses that accurately reflects the similarity of the visual input. Using data from electrical stimulation experiments, we demonstrate that this metric may improve the performance of a prosthesis.

1 INTRODUCTION

An important application of neuroscience research is the development of electronic devices to replace the function of diseased or damaged neural circuits (Wilson et al., 1991; Schwartz, 2004). Artificial vision has been a particularly challenging modality due to the richness of visual information, its diverse uses in perception and behavior, and the complexity of fabricating a physical device that can interface effectively with neural circuitry (Stingl et al., 2013; Wilke et al., 2011; Jepson et al., 2014a).

The most advanced example is a retinal prosthesis: a device that replaces the function of neural circuitry in the retina lost to degenerative disease. Most of the computational work related to this application has focused on building encoding models that use the visual image to accurately predict the spiking activity of populations of retinal ganglion cells (RGCs), the output neurons of the retina that convey visual information to the brain. Leading models include linear models (Chichilnisky, 2001), probabilistic point-process models (Pillow et al., 2008) and recently proposed models employing rich nonlinearities (McIntosh et al.; Batty et al.; Shah et al., 2017).

However, an accurate encoding model, although valuable, is insufficient. Any retinal prosthesis – whether based on electrical stimulation (Sekirnjak et al., 2008) or optical stimulation (Boyden et al., 2005; Bernstein et al., 2008) – is limited in its ability to create arbitrary desired patterns of neural activity, due to inefficiencies or lack of specificity in the stimulation modality (Barrett et al., 2014; Jepson et al., 2014a). Thus, a given stimulation system can only achieve a limited vocabulary of elicited spike patterns. Hence, although a powerful and accurate encoding model might indicate that a particular spike pattern would be the natural biological response to the incident visual stimulus, the desired spike pattern might not reside within the feasible set of the stimulation device.

Previous studies (Jepson et al., 2014b) have addressed this problem by selecting the electrical stimulation which minimizes the number of unmatched spikes across cells (Hamming distance). However, the goal of a prosthesis should instead be to select electrical stimulation parameters that produce a response which is most similar to the desired pattern of activity in terms of the visual sensation it produces (Figure 1). The similarity measure ideally would capture – perhaps implicitly – the statistics of relevant visual stimuli and the signal and noise properties of individual neurons and populations of neurons (Shlens et al., 2009; Pillow et al., 2008; Field & Chichilnisky, 2007). This measure should expand upon previously developed spike train metrics that trade off spike count,

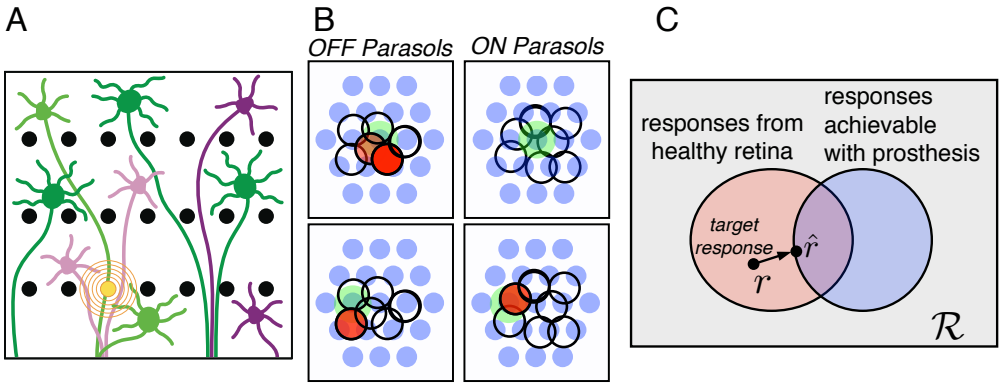

Figure 1: Setup of a retinal prosthetic system. A. Cartoon showing ON (purple) and OFF(green) RGCs overlayed on top of an electrode array (black dots). The cells firing due to passing current through the orange electrode are lightly shaded. Note that a cell can fire even if an axon of the cell passes near the stimulating electrode. Hence, single electrode stimulation leads to firing of many cells. B. *Real* recorded spiking activity in a two populations of primate retinal ganglion cells (RGCs) demonstrating the lack of specificity from electrical stimulation. The electrodes are in blue and the stimulated electrode is shaded green. C. The target firing pattern r often lies outside the set of firing patterns achievable with the prosthesis. The goal of the learned metric is to define a distance measure to identify the nearest feasible electrical stimulation $\hat{r}$. $\mathcal{R}$ denotes the set of all neural population responses.

temporal precision and neural identity, and are easy to compute (Victor & Purpura, 1996; 1997; Victor, 2005).

Here we present an efficient, learned method for selecting the electrical stimulation pattern that produces the optimal pattern of neural activity achievable within the constraints of the device (Figure 1B and C), where optimality is assessed in terms of the visual stimulus associated with the response. We first build a new spike train similarity measure that is expensive to learn but cheap to use in real time, and captures the statistics of visual images and neural responses. We fit this model to recordings of populations of RGCs in non-human primate retina, and demonstrate that it provides an intuitive, meaningful representation of the similarity between spike patterns in the RGC population. Finally, we use data from electrical stimulation experiments to demonstrate how the approach can lead to more faithful reproduction of the visual image.

2 Metric and Similarity Learning

In this section we describe the algorithmic framework for learning pseudometrics or similarity measures in neural response space. We start by introducing notations and conventions that we use throughout the paper. We use bold face letters to denote vectors and upper case letters to denote matrices. The element wise product of two matrices A and B of the same size is denoted by $A \bullet B = \mathrm{Tr}(A^\top B)$. We denote the symmetrizing operator of a square matrix M by $\mathrm{sym}(M) = \frac{1}{2}(M + M^\top)$.

A single frame of visual stimulus, $\mathbf{s}$, is an image represented as an $n \times n$ matrix. The space of possible stimuli is $\mathcal{S} \subset \mathbb{R}^{n \times n}$. A sequence $\mathbf{s}_l, \ldots, \mathbf{s}_m$ of $m - l + 1$ frames, where $\mathbf{s}_j \in \mathcal{S}$, is denoted as $\mathbf{s}_{l:m}$. In order to simplify our notation, we define the responses of the cells to be a p dimensional vector and the space of possible responses as $\mathbf{r} \subseteq \mathbb{R}^p$. Analogously, a sequence of cell activities $\mathbf{r}_t$ for $t = l, \ldots, m$ is denoted $\mathbf{r}_{l:m}$. To simplify the presentation below, we confine the visual stimulus to be a single image and the corresponding response of each cell to be a scalar.

2.1 Empirical Loss Minimization

Given the population response space $\mathcal{R}$, we learn a function $h : \mathcal{R} \times \mathcal{R} \to \mathbb{R}$ which captures invariances in the spiking response when the same stimulus is presented multiple times. The scoring function h is viewed either as a similarity function or a pseudometric. To distinguish between the two

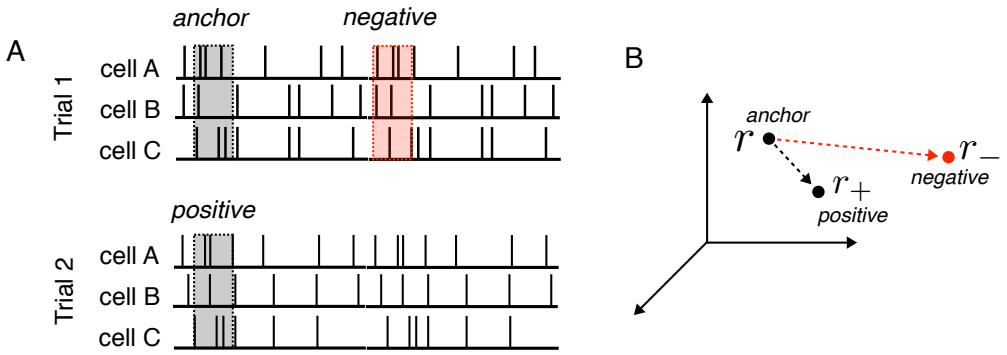

Figure 2: Method for constructing an objective for learning a neural metric using triplet-based losses. (A) Schematic of spike trains from population of neurons responding to a dynamic stimulus (not shown). The spiking response across the window is randomly selected at relative time t. The binarized representation for the population response is termed the *anchor*. In a second presentation of the same stimulus, the population response at the same relative time t is recorded and labeled a *positive*. The *negative* response is any other selected time. (B) The objective of the triplet-loss is to require that the *anchor* is closer to the *positive* response than the *negative* response with a margin. See text for details.

cases, we use $d(\cdot,\cdot)$ to denote a pseudometric. A pseudometric d needs to satisfy: $d(\mathbf{r}_1,\mathbf{r}_2) \geq 0$ and $d(\mathbf{r},\mathbf{r}) = 0$ [*positivity*], $d(\mathbf{r}_1,\mathbf{r}_2) + d(\mathbf{r}_2,\mathbf{r}_3) \geq d(\mathbf{r}_1,\mathbf{r}_3)$ [*sub-additivity*], and $d(\mathbf{r}_1,\mathbf{r}_2) = d(\mathbf{r}_2,\mathbf{r}_1)$ [*symmetry*].

During the experiments, repeats of the same sequence of visual stimuli are presented. The responses collected during the ith presentation (repeat) of visual stimulus are denoted by $(\mathbf{s}_t^i, \mathbf{r}_t^i)$. Here $\mathbf{s}_t^i$ is the stimulus history which elicits population response $\mathbf{r}_t^i$ at time t. The goal of this approach to learn a metric such that pairs of responses generated during different repeats of the same stimulus are closer, or more similar, than pairs of responses generated by different stimuli. We slice the data into triplets of the form $(\mathbf{r}, \mathbf{r}_+, \mathbf{r}_-)$ where $\mathbf{r}$ and $\mathbf{r}_+$ are responses of cells to the same stimulus while $\mathbf{r}_-$ is the response to a different visual stimuli (Figure 2A). We refer to $(\mathbf{r}, \mathbf{r}_+)$ as a positive pair and $(\mathbf{r}, \mathbf{r}_-)$ as a negative pair (Figure 2B).

Metric and similarity learning using triplets (Shalev-Shwartz et al., 2004; Weinberger & Saul, 2009) has been used extensively in several domains, such as computer vision for state of the art performance on face recognition (Schroff et al., 2015). A central theme of this work (Schroff et al., 2015; Oh Song et al., 2016) has been improving metric learning by mining hard-to-classify negatives.

However, it can be computationally demanding to mine such hard negatives. We found that a much simpler strategy of randomly sampling a common set of negatives for all the positive examples in the batch is very effective. Hence we first sample positive pairs of responses corresponding to random stimulus times and a common set of negative responses generated by stimuli distinct from any stimulus for positive responses. Hence a batch of triplets is denoted by $\mathcal{T} = \{\{\mathbf{r}^i, \mathbf{r}_+^i\}, \{\mathbf{r}_-^j\}\}$.

Given a training set of triplets $\mathcal{T}$ the goal is to find a pseudometric such that for most $(\mathbf{r}^i, \mathbf{r}_+^i, \{\mathbf{r}_-^j\}) \in \mathcal{T}$ the distance between responses of two repeats of same stimulus is smaller than their distance to any of the irrelevant response vectors,

$$d(\mathbf{r}^i, \mathbf{r}_+^i) < \min_j d(\mathbf{r}^i, \mathbf{r}_-^j) \tag{1}$$

We cast the learning task as empirical risk minimization of the form,

$$\frac{1}{|\mathcal{T}|} \sum_{(\mathbf{r}^i, \mathbf{r}_+^i, \{\mathbf{r}_-^j\}) \in \mathcal{T}} \ell(\mathbf{r}^i, \mathbf{r}_+^i, \{\mathbf{r}_-^j\}) \ ,$$

where $\ell()$ is a differential, typically convex, relaxation of the ordering constraints from (1). We use

$$\ell(\mathbf{r}^i, \mathbf{r}_+^i, \{\mathbf{r}_-^j\}) = \beta \log \left[1 + \sum_j e^{\frac{d(\mathbf{r}^i, \mathbf{r}_+^i) - d(\mathbf{r}^i, \mathbf{r}_-^j)}{\beta}} \right] \ ,$$

as the surrogate loss. We use $\beta = 10$ in this work.

In the case of similarity learning, we swap the role of the pairs and define,

$$\ell(\mathbf{r}^i, \mathbf{r}^i_+, \{\mathbf{r}^j_-\}) = \beta \log \left[1 + \sum_j e^{\frac{h(\mathbf{r}^i, \mathbf{r}^j_-) - h(\mathbf{r}^i, \mathbf{r}^i_+)}{\beta}} \right] ,$$

We implemented two parametric forms for distance and similarity functions. The first is a quadratic form where $A \succeq 0$ and

$$h_A(\mathbf{r}_1, \mathbf{r}_2) = \mathbf{r}_1^\top A \, \mathbf{r}_2 \quad \text{and} \quad d_A(\mathbf{r}_1, \mathbf{r}_2) = (\mathbf{r}_1 - \mathbf{r}_2)^\top A \, (\mathbf{r}_1 - \mathbf{r}_2) \, . \tag{2}$$

We also implemented and experimented with non-linear representations using a feed-forward convolutional network to define a mapping ϕ from the response space $\mathcal{R}$ to $\mathbb{R}^p$. Briefly, we employ a hierarchical, convolutional network topology to mirror the translation invariance expected in the receptive field organization of the retina. The convolutional network consists of 595K parameters across 7 layers and we employ batch normalization to speed up training. Please the Appendix for more details.

We defer the details of network architecture to the appendix.

$$h_\phi(\mathbf{r}_1, \mathbf{r}_2) = \phi(\mathbf{r}_1) \cdot \phi(\mathbf{r}_2) \quad \text{and} \quad d_\phi(\mathbf{r}_1, \mathbf{r}_2) = \|\phi(\mathbf{r}_1) - \phi(\mathbf{r}_2)\|^2 \, . \tag{3}$$

We learn the parameters by minimizing the loss using Adagrad (Duchi et al., 2011) for quadratic metric and Adam (Kingma & Ba, 2014) for convolutional neural network metric. For the quadratic metric, we further project A onto the space of positive semi-definite matrices space after every update. The projection is performed using singular value decomposition. Concretely, we rewrite A as, $UDU^\top$ where U is a unitary matrix and D is a diagonal matrix. We then threshold the diagonal elements of D to be non-negative.

3 Results

3.1 Experimental Setup

Spiking responses from hundreds of retinal ganglion cells (RGCs) in primate retina were recorded using a 512 electrode array system (Litke et al., 2004; Frechette et al., 2005). ON and OFF parasol RGC types were identified using visual stimulation with binary white noise and reverse correlation (Chichilnisky, 2001).

Since each analysis requires different stimulus conditions and number of cells, we leave the details of each preparation to the subsequent sections. For each analysis, spike trains were discretized at the 120 Hz frame rate of the display (bins of 8.33ms), and responses across all the cells for 1 time bin were used to generate each training example.

In the following sections, we quantitatively assess the quality of multiple learned metrics – each metric with increasing complexity – with respect to a baseline (Hamming distance). First, we assess the quality of the learned metrics with respect to traditional error analysis. Second, we assess the quality of the learned embeddings with respect to optimal decoding of the stimulus. Finally, we demonstrate the utility of the learned metric by employing the metric in a real electrical stimulation experiment.

3.2 Quantitative evaluation of learned metric space.

The quality of a neural metric is measured by the the ability of the model to determine whether a pair of firing patterns arose from the same visual stimulus or a distinct visual stimulus.

For understanding the properties of learned metric at the scale of large RGC populations, we focus our analysis on responses of a collection of 36 OFF parasol cells and 30 ON parasol cells to 600 repeats of 10s long white noise stimulus clip. The responses were paritioned into training (first 8s) and testing (last 2s) of each trial.

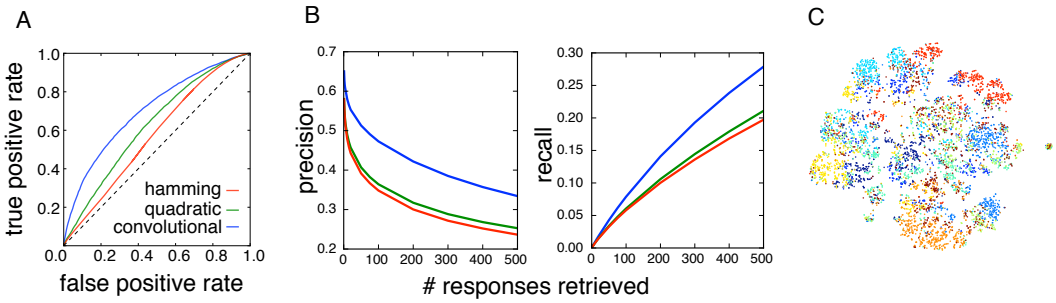

Figure 3: Quantitative analysis of learned metric. (A) ROC curve measuring false positive rate versus true positive rate across a range of thresholds for the convolutional neural network learned metric (blue, AUC: 0.7), quadratic learned metric (green, AUC:0.6), and Hamming distance (red, AUC:0.5). Note that curves closer to the upper left corner correspond to higher quality metrics. (B) Precision and recall of classification as spike pattern from same stimulus or a different stimulus across increasing number of retrieved nearest neighbors from a collection of responses consisting of all 600 repeats of 10 different stimuli. Models and colors same as (A). (C) t-SNE embedding of the collection of responses in (B) using distances estimated by the convolutional neural network. Each dot is a population response and responses with same color were generated in by the same visual stimulus.

We assessed a range of learned embedding models and baselines by employing receiver-operating characteristic (ROC) analysis. Specifically, we selected the population firing pattern, $\mathbf{r}$, at a particular offset in time in the experiment (corresponding to a visual stimulus history) and compared this firing pattern to two other types of firing patterns: (1) the firing pattern from the same group of cells at the same time during a second repeated presentation of the stimulus, $\mathbf{r}_+$; and (2) the firing pattern at a distinct, randomly selected time point, $\mathbf{r}_-$. For a given threshold, if the metric results in a correct classification of $\mathbf{r}_+$ as the same stimulus, we termed the result a *true positive*. For the same threshold, if an embedding metric incorrectly classified $\mathbf{r}_-$ as the same stimulus, we termed it a *false positive*. Note that unlike training, we do not choose a common set of negatives for testing.

Figure 3A traces out the trade-off between the false positive rate and true positive rate across a range of thresholds in an assortment of embedding models for neural population activity. Better models trace out curves that bend to the upper-left of the figure. The line of equality indicates a model that is performing at chance. A simple baseline model of a Hamming distance (red curve) performs least accurately. Quadratic metric which permit variable weight for each neuron and interaction between pairs of neurons improves the performance further (green curve). Finally, replacing a quadratic metric for a euclidean distance between embedding of responses using convolutional neural network improves the performance even further (blue curve).

The ROC analysis provides strong evidence that increasingly sophisticated embedding models learn global structure above and beyond a Hamming distance metric. We also examined how the local structure of the space is captured by the embedding metric by calculating the learned embeddings on a test dataset consisting of 600 repeats each of the 10 different visual stimuli. We randomly selected a firing pattern $\mathbf{r}$ from one presentation of the stimulus, and identified k nearest neighbors according to our metric, for increasing k. Among the k nearest neighbors, we assessed precision, i.e. what fraction of the nearest neighbors correspond to 599 other presentations of the same stimulus. A perfect learned embedding model would achieve a precision of 100% for $k \leq 599$. We also measured recall, i.e. what fraction of the remaining 599 presentations of the same stimulus are within the k nearest neighbors. A perfect learned embedding model would achieve recall of $k/5.99\%$ for $k \leq 599$. Figure 3B highlights the performance of various learned methods across increasing k. The results indicate that the precision and recall are below an optimal embedding, but the convolutional metric performs better than quadratic and hamming metrics.

Finally, we embed the 600 responses to 10 distinct stimuli using t-SNE (Maaten & Hinton, 2008) with distances estimated using the convolutional metric to visualize the separation between the responses. We see in Figure 3C, that responses corresponding to same visual stimulus (same color) cluster in the same region of embedding space reflecting the ability of the response space metric to discriminate distinct stimuli

3.3 LEARNED METRIC CAPTURES STIMULUS SIMILARITY.

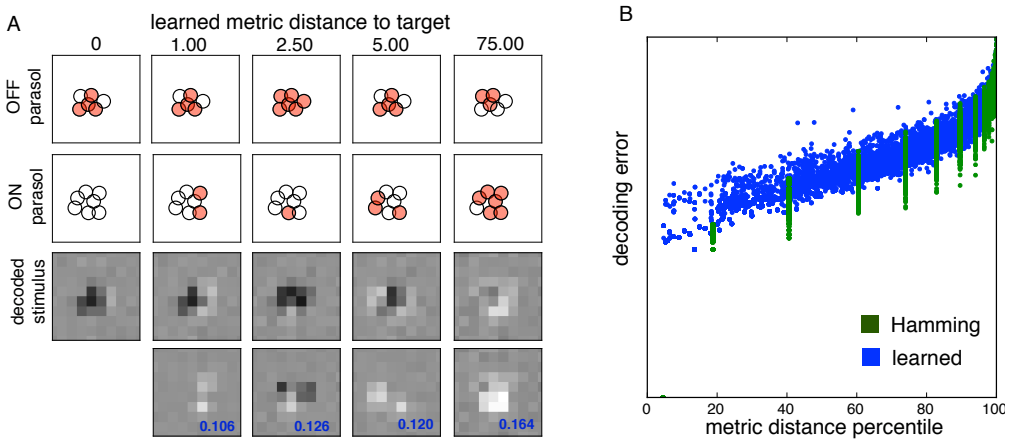

Figure 4: Decoded stimulus degrades with increasing response distance. (A) Target and retrieved responses at different distances according to a learned quadratic metric. Receptive field locations of OFF and ON parasol cells are shown (shading indicates firing). Numbers above indicate percentile of cumulative distribution across distances in learned metric space. Linearly decoded stimulus and the difference from the decoded stimulus for the target firing pattern is shown below, respectively. Numbers in the plot (blue) indicate MSE between decoded stimulus for target and that for the firing pattern above. (B) Hamming distance (green) and learned quadratic metric distance (blue) between a randomly selected pair of firing patterns plotted versus the MSE between corresponding linearly decoded stimuli.

We train the metric only based on whether a pair of responses are generated by same stimulus or not. However, a useful embedding model would implicitly capture further visual stimulus dependencies. We test this by decoding the stimulus linearly from the neural response.

To answer this question, we focus on a spatially localized and overlapping population of 13 RGCs consisting of 6 ON parasol cells and 7 OFF parasol cells (Figure 1B). We focus on a smaller population of cells in this (and subsequent) section as we can explicitly list out all the possible 2^{13} response patterns. Training data was accrued from RGC responses to 5 repeated presentations of a 5 minute long white noise sequence. The first 4 minutes of each presentation was employed for training; the last minute was employed for testing.

We examined the similarity between the decoded stimulus and the target stimulus, for responses that, according to our learned quadratic metric, were increasingly distant from the target. Figure 4A bottom left shows the spatial profile of the linear decoding 20ms prior to the target response. We next calculate the distance of this target firing pattern to all 2^{13} firing patterns and rank order them based on the learned metric. Figure 4A, top rows, shows firing patterns at the 1%, 2.5%, 5% and 75% percentiles. Below these firing patterns are the associated linearly decoded stimuli, and the errors with respect to the target firing pattern. As we choose patterns farther from the target in terms of our metric, the distance between the decoded stimulus for the chosen firing pattern and target firing pattern systematically increases.

We quantify this observation in Figure 4B by randomly selecting pairs of responses from the test data and calculating the optimal linearly decoded stimulus associated with them (see Methods). We then plot the RMS distance between the linearly decoded stimuli against the normalized metric distance between the responses. The decoding error systematically increases as the metric distance between the corresponding responses increases, for both the learned metric (blue) as well the Hamming distance (green). However, the distances generated by Hamming distance are discrete and therefore provide a less granular representation of the decoding errors associated with the stimuli.

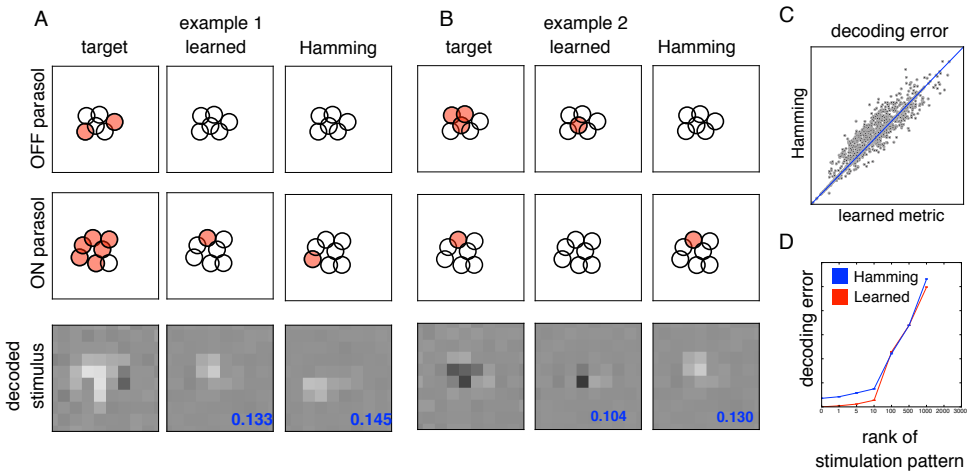

Figure 5: Learned quadratic response metric gives better perception than using a Hamming metric. (A) First column: target firing pattern in ON and OFF parasol cells (shading indicates firing) and decoded stimulus. Middle column: nearest feasible, elicited firing pattern as judged by learned metric and associated decoded stimulus. The scaled mean squared error from target response is shown in the decoded figure. Right column: nearest feasible, elicited firing pattern as judged by Hamming metric and associated decoded stimulus. (B) Same as (A) but for a second target firing pattern. (C) MSE of decoded stimulus for nearest firing pattern for learned metric versus Hamming metric for all observed firing patterns. Note that $33\%, 49\%, 18\%$ of patterns reside above, on and below the line of equality, (D) Rank order of stimulation patterns selected by the learned metric (red) and Hamming metric (blue) versus MSE of decoded stimulus for the first 2000 closest patterns

3.4 Learned response metric could improve performance of retinal prosthesis.

Using recorded experimental data, we now show how response metrics could improve retinal prostheses by choosing optimal electrical stimulation patterns. For a given target response, we use the learned quadratic metric to select the best electrical stimulation pattern, and evaluate the effectiveness of this approach by linearly decoding the stimulus from the elicited responses.

Calibration of RGC responses to electrical stimulation patterns was performed by repeating a given electrical stimulation pattern 25 times, at each of 40 current levels, in a random order. Due to the limited duration of experiments, we focused on stimulation patterns in which only one electrode was active. The data was spike sorted and spiking probability was computed for each cell by averaging across trials for each electrical stimulation pattern (Mena et al., 2016). The probability of firing for each cell in response to increasing current amplitude through each electrode was approximated with a sigmoid function.

Since the RGC response to each current stimulation is probabilistic, we evaluate each stimulation pattern by the *expected* distance between the elicited responses and the target firing pattern. For a quadratic response metric this can be easily computed in closed form. Given a response metric, we rank different stimulation patterns based on the expected distance to the target firing pattern. In Figure 5A and B (first columns) we show example target response patterns and the corresponding linearly decoded visual stimulus. We then analyze the best stimulation pattern determined by the learned quadratic metric, and by the Hamming distance. The responses sampled from the response distributions for the selected stimulation patterns are shown in Figures 5A and B (second and third columns each). We find that the linearly decoded stimuli are closer to the target when the stimulation was chosen via the learned response metric compared to the Hamming distance.

To quantify this behavior, we calculated the mean squared error between the decoded stimuli when the stimulation was chosen using learned metric and the Hamming distance (Figure 5C). We find that the learned metric and Hamming metric identify the same stimulation pattern and hence achieve the same error for 49% for the target responses observed. However, on 33% of the target responses,

the learned metric achieves lower mean squared error then the Hamming distance; conversely, the learned metric has larger MSE then Hamming distance on 18% of the target responses.

The above analysis demonstrates the benefit of using the learned metric over Hamming distance to choose the best stimulation pattern. However, the collection of available electrical stimulation patterns might change over time due to hardware or biophysical constraints. To assess the improvement under such cases, we next ask how well the learned metric performs relative to Hamming distance if we choose the kth best current pattern using each metric. (Figure 5D). Increasing rank order (k) in the learned metric leads to higher MSE in terms of a decoded stimulus. Importantly, the learned metric achieves systematically lower MSE than the Hamming distance across the nearest $k \leq 10$ stimulation patterns. These results indicate that the learned metric systematically selects better electrical stimulation patterns for eliciting reasonably close firing patterns.

4 Discussion

The learned metric approach has two major potential implications for visual neuroscience. It provides a novel method to find "symbols" in the neural code of the retina that are similar in the sense that they indicate the presence of similar stimuli Ganmor et al. (2015). We speculate that our method may be more robust to training with small data sets than methods based on mutual information or covariance, a potentially significant advantage, particularly for high-dimensional responses (many simultaneously-recorded cells, time-resolved). Second, as shown above, the metric has application to retinal prosthesis technology, in which imperfect hardware constraints demand that the set of neural responses that can be generated with a device be used to effectively transmit useful visual information. For this application, a metric on responses that reflects visual stimulus similarity could be extremely useful.

Additional techniques may also be helpful in extending our method to data involving many cells, temporal responses, and additional response structure. For example, using recurrent neural networks (Lipton et al., 2015) to embed responses may help compute distances between spiking patterns consisting of multiple time bins, perhaps of unequal length. Boosting (Freund & Schapire, 1999) may help combine multiple efficiently learned metrics for a smaller, spatially localized groups of cells. Other metrics may be developed to capture invariances learned by commonly used encoding models (Chichilnisky, 2001; Pillow et al., 2008). Also, triplet mining techniques (i.e., choosing hard negatives), a commonly used trick in computer vision, may improve efficiency (Schroff et al., 2015; Oh Song et al., 2016). Novel metrics could also be learned with additional structure in population responses, such as the highly structured correlated activity in RGCs Mastronarde (1983); Greschner et al. (2011). This noise correlation structure may be learnable using *negative* examples that destroy the noise correlations in data, but preserve light response properties, by taking responses of different cells from different repeats of the stimulus.

Note that the convolutional metric outperforms the quadratic metric at both global (ROC curves) and local (precision recall curves) scales. However, using current retinal prosthesis technologies, we might be able to resolve information only up to a particular scale. We think that for current retinal prostheses, capturing global structure is probably of greatest importance, because state-of-the-art technology has a relatively coarse vocabulary for stimulating RGCs (Humayun et al., 2012; Zrenner et al., 2011) (see also Figure 1). Specifically, the "nearest" elicited firing pattern is "far" in terms of the corresponding visual stimulus (Figure 5) . In terms of the proposed learned metric, the nearest feasible firing pattern achievable by electrical stimulation in our experiments is at the 10th percentile of all possible firing patterns. In this context, the average closest stimulation pattern, expressed as a percentile of the learned metric distances, provides a valuable benchmark to measure the performance of a prosthesis and how that performance is affected by advances in the underlying hardware and software.

References

John Martin Barrett, Rolando Berlinguer-Palmini, and Patrick Degenaar. Optogenetic approaches to retinal prosthesis. *Visual neuroscience*, 31(4-5):345–354, 2014.

Eleanor Batty, Josh Merel, Nora Brackbill, Alexander Heitman, Alexander Sher, Alan Litke, E.J. Chichilnisky, and Liam Paninski. Multilayer recurrent network models of primate retinal ganglion

cell responses.

Jacob G Bernstein, Xue Han, Michael A Henninger, Emily Y Ko, Xiaofeng Qian, Giovanni Talei Franzesi, Jackie P McConnell, Patrick Stern, Robert Desimone, and Edward S Boyden. Prosthetic systems for therapeutic optical activation and silencing of genetically targeted neurons. In *Biomedical Optics (BiOS) 2008*, pp. 68540H–68540H. International Society for Optics and Photonics, 2008.

Edward S Boyden, Feng Zhang, Ernst Bamberg, Georg Nagel, and Karl Deisseroth. Millisecond-timescale, genetically targeted optical control of neural activity. *Nature neuroscience*, 8(9): 1263–1268, 2005.

EJ Chichilnisky. A simple white noise analysis of neuronal light responses. *Network: Computation in Neural Systems*, 12(2):199–213, 2001.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.

Greg D Field and EJ Chichilnisky. Information processing in the primate retina: circuitry and coding. *Annu. Rev. Neurosci.*, 30:1–30, 2007.

ES Frechette, A Sher, MI Grivich, D Petrusca, AM Litke, and EJ Chichilnisky. Fidelity of the ensemble code for visual motion in primate retina. *Journal of neurophysiology*, 94(1):119–135, 2005.

Yoav Freund and Robert Schapire. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

Elad Ganmor, Ronen Segev, and Elad Schneidman. A thesaurus for a neural population code. *Elife*, 4:e06134, 2015.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010.

Martin Greschner, Jonathon Shlens, Constantina Bakolitsa, Greg D Field, Jeffrey L Gauthier, Lauren H Jepson, Alexander Sher, Alan M Litke, and EJ Chichilnisky. Correlated firing among major ganglion cell types in primate retina. *The Journal of Physiology*, 589(1):75–86, 2011.

Mark S Humayun, Jessy D Dorn, Lyndon Da Cruz, Gislin Dagnelie, José-Alain Sahel, Paulo E Stanga, Artur V Cideciyan, Jacque L Duncan, Dean Eliott, Eugene Filley, et al. Interim results from the international trial of second sight's visual prosthesis. *Ophthalmology*, 119(4):779–788, 2012.

Lauren H Jepson, Paweł Hottowy, Keith Mathieson, Deborah E Gunning, Władysław Dąbrowski, Alan M Litke, and EJ Chichilnisky. Spatially patterned electrical stimulation to enhance resolution of retinal prostheses. *Journal of Neuroscience*, 34(14):4871–4881, 2014a.

Lauren H Jepson, Pawel Hottowy, Geoffrey A Weiner, Władysław Dabrowski, Alan M Litke, and EJ Chichilnisky. High-fidelity reproduction of spatiotemporal visual signals for retinal prosthesis. *Neuron*, 83(1):87–92, 2014b.

Eric R Kandel, James H Schwartz, Thomas M Jessell, Steven A Siegelbaum, A James Hudspeth, et al. *Principles of neural science*, volume 4. McGraw-hill New York, 2000.

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

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Zachary C Lipton, John Berkowitz, and Charles Elkan. A critical review of recurrent neural networks for sequence learning. *arXiv preprint arXiv:1506.00019*, 2015.

AM Litke, N Bezayiff, EJ Chichilnisky, W Cunningham, W Dabrowski, AA Grillo, M Grivich, P Grybos, P Hottowy, S Kachiguine, et al. What does the eye tell the brain?: Development of a system for the large-scale recording of retinal output activity. *IEEE Transactions on Nuclear Science*, 51(4):1434–1440, 2004.

Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008.

DAVID N Mastronarde. Correlated firing of cat retinal ganglion cells. i. spontaneously active inputs to x-and y-cells. *Journal of Neurophysiology*, 49(2):303–324, 1983.

Lane T McIntosh, Niru Maheswaranathan, Aran Nayebi, Surya Ganguli, and Stephen A Baccus. Deep learning models of the retinal response to natural scenes.

Gonzalo E Mena, Lauren E Grosberg, Pawel Hottowy, Alan Litke, John Cunningham, EJ Chichilnisky, and Liam Paninski. Electrical stimulus artifact cancellation and neural spike detection on large multi-electrode arrays. *bioRxiv*, 2016. doi: 10.1101/089912. URL `http://biorxiv.org/content/early/2016/12/02/089912`.

Hyun Oh Song, Yu Xiang, Stefanie Jegelka, and Silvio Savarese. Deep metric learning via lifted structured feature embedding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4004–4012, 2016.

Jonathan W Pillow, Jonathon Shlens, Liam Paninski, Alexander Sher, Alan M Litke, EJ Chichilnisky, and Eero P Simoncelli. Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, 454(7207):995–999, 2008.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

Andrew B Schwartz. Cortical neural prosthetics. *Annu. Rev. Neurosci.*, 27:487–507, 2004.

Chris Sekirnjak, Pawel Hottowy, Alexander Sher, Wladyslaw Dabrowski, Alan M Litke, and EJ Chichilnisky. High-resolution electrical stimulation of primate retina for epiretinal implant design. *Journal of Neuroscience*, 28(17):4446–4456, 2008.

N Shah, N Brackbill, C Rhoades, A Tikidji-Hamburyan, G Goetz, A Sher, A Litke, L Paninski, E P Simoncelli, and EJ Chichilnisky. Model-based inference of nonlinear subunits underlying responses of primate retinal ganglion cells. In *Computational and Systems Neuroscience (CoSyNe)*, Salt Lake City, Utah, 23-26 Feb 2017.

S. Shalev-Shwartz, A. Ng, and Y. Singer. Online and batch learning of pseudo-metrics. In *Proc. of 21st Intl. Conf. on Machine Learning*, 2004.

Jonathon Shlens, Greg D. Field, Jeffrey L. Gauthier, Martin Greschner, Alexander Sher, Alan M. Litke, and E. J. Chichilnisky. The structure of large-scale synchronized firing in primate retina. *Journal of Neuroscience*, 29(15):5022–5031, 2009. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.5187-08.2009. URL `http://www.jneurosci.org/content/29/15/5022`.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Katarina Stingl, Karl Ulrich Bartz-Schmidt, Dorothea Besch, Angelika Braun, Anna Bruckmann, Florian Gekeler, Udo Greppmaier, Stephanie Hipp, Gernot Hörtdörfer, Christoph Kernstock, et al. Artificial vision with wirelessly powered subretinal electronic implant alpha-ims. In *Proc. R. Soc. B*, volume 280, pp. 20130077. The Royal Society, 2013.

Jonathan D Victor. Spike train metrics. *Current opinion in neurobiology*, 15(5):585–592, 2005.

Jonathan D Victor and Keith P Purpura. Nature and precision of temporal coding in visual cortex: a metric-space analysis. *Journal of neurophysiology*, 76(2):1310–1326, 1996.

Jonathan D Victor and Keith P Purpura. Metric-space analysis of spike trains: theory, algorithms and application. *Network: computation in neural systems*, 8(2):127–164, 1997.

Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.

Robert Wilke, Veit-Peter Gabel, Helmut Sachs, Karl-Ulrich Bartz Schmidt, Florian Gekeler, Dorothea Besch, Peter Szurman, Alfred Stett, Barbara Wilhelm, Tobias Peters, et al. Spatial resolution and perception of patterns mediated by a subretinal 16-electrode array in patients blinded by hereditary retinal dystrophies. *Investigative ophthalmology & visual science*, 52(8):5995–6003, 2011.

Blake S Wilson, Charles C Finley, et al. Better speech recognition with cochlear implants. *Nature*, 352(6332):236, 1991.

Eberhart Zrenner, Karl Ulrich Bartz-Schmidt, Heval Benav, Dorothea Besch, Anna Bruckmann, Veit-Peter Gabel, Florian Gekeler, Udo Greppmaier, Alex Harscher, Steffen Kibbel, et al. Subretinal electronic chips allow blind patients to read letters and combine them to words. *Proceedings of the Royal Society of London B: Biological Sciences*, 278(1711):1489–1497, 2011.

A APPENDIX

A.1 CONVOLUTIONAL METRIC

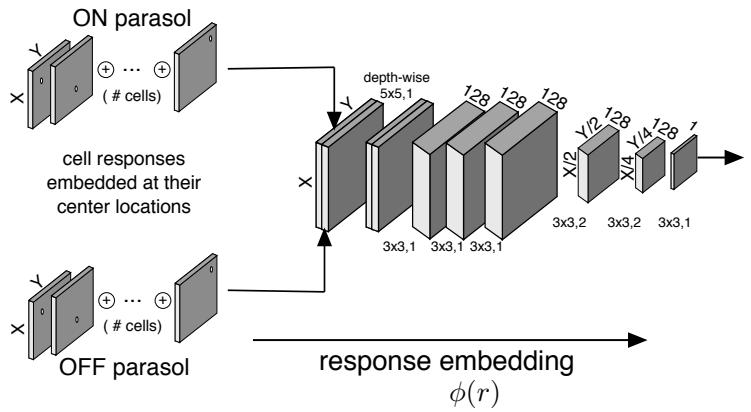

Figure 6: Convolutional network used for response embedding. The convolutional filtering is specified between each layer. For example $3x3, 1$ means convolution of a $3x3$ filter with stride 1.

We build a hierarchical, convolutional network to mirror the translation invariance expected in the receptive field organization of the retina. The goal of this network is to flexibly capture population activity of ON and OFF cells but employ minimal knowledge about cell receptive fields. The reason for this approach is to build a model that may be amenable to a retinal prosthetic in which the characterization of individual retinal ganglion cells is limited (Jepson et al., 2014a;b).

In particular, the network employs knowledge of the receptive field locations and firing rates of individual cells but the network is independent of the number of cells in the retina. The latter point is achieved by embedding the responses of neurons in a 2 channel pathway corresponding to the ON and OFF pathways known form biology (Kandel et al., 2000).

The network receives as input the spiking activity of ON and OFF parasols and embeds these spike patterns as one-hot vectors placed at the spatial locations of each cell's receptive field. The resulting pattern of activations is summed across all cells in the ON and OFF populations, respectively, and passed through several convolutional layers of a network. Successive layers shrink the spatial activation size of the representation, while increasing the number of filter channels (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014). The final embedding response vector has $1/16$th number of pixels in stimulus and represents the flattened representation of the last layer of the network.

Let the RGC population response be $\mathbf{r} \in \{0, 1\}^{\text{cells}}$.

- Represent responses with +1/-1 with $\tilde{r} = 2 * (r - 0.5)$.
- Compute the scale for each cell as a function of the mean firing rate: $s_i = a_0\mu_i^3 + a_1\mu_i^2 + a_2\mu_i^3 + a_3$.
- Map each cell to its center location on a grid with spatial dimensions same as those of visual stimulus. Let M_i be grid embedding on cell i. So, M_i has zero for all positions except center of cell.
- Perform a separable 5x5 convolution (stride = 1) of each cell's M_i to get RF estimate of cell ($\tilde{M}_i$).
- Add the activation of cells of the same type to get the total activation for a given cell type. Hence, activation map for each cell type $A_i = \sum_i \tilde{r}_i s_i \tilde{M}_i$. Subsequent layers receive input as a two layered activation map corresponding to ON and OFF parasol cells.
- The convolutional layers further combine information accross multiple cells, of different types. The details of different layers is shown in figure 6.

A.1.1 HYPERPARAMETERS

Operation	Kernel size	Stride	Feature maps	Padding	Nonlinearity
Network – $X \times Y \times 2$ input					
Separable	5	1	1	SAME	ReLU
Convolution	3	1	128	SAME	ReLU
Convolution	3	1	128	SAME	ReLU
Convolution	3	1	128	SAME	ReLU
Convolution	3	2	128	SAME	ReLU
Convolution	3	2	128	SAME	ReLU
Convolution	3	1	1	SAME	ReLU
Total number of parameters	595229				
Normalization	Batch normalization after every convolution				
Optimizer	Adam (Kingma & Ba, 2014) ($\alpha = 0.01$, $\beta_1 = 0.9$, $\beta_2 = 0.999$)				
Parameter updates	20,000				
Batch size	100				
Weight initialization	Xavier initialization (Glorot & Bengio, 2010)				

Table 1: Details of convolutional network for embedding neural responses.