Modeling Biological Immunity to Adversarial Examples

Edward Kim[1], Jocelyn Rego[1], Yijing Watkins[2], Garrett T. Kenyon[2]
[1]Department of Computer Science, Drexel University, PA
[2]Los Alamos National Laboratory, Los Alamos, NM
ek826@drexel.edu,jr3548@drexel.edu,twatkins@lanl.gov,gkeynon@lanl.gov

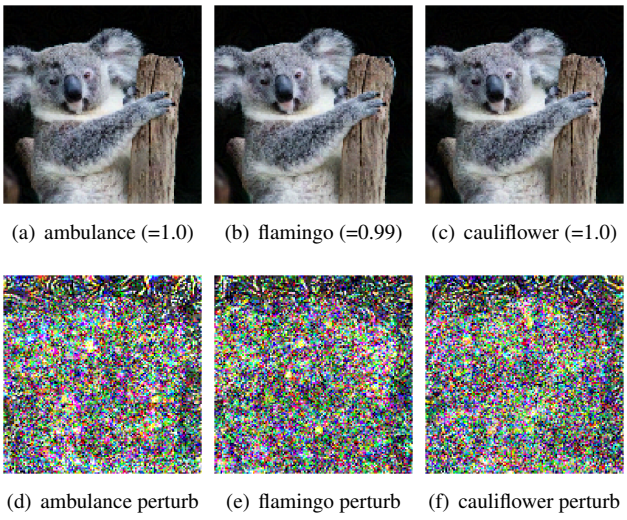
Abstract

While deep learning continues to permeate through all fields of signal processing and machine learning, a critical exploit in these frameworks exists and remains unsolved. These exploits, or adversarial examples, are a type of signal attack that can change the output class of a classifier by perturbing the stimulus signal by an imperceptible amount. The attack takes advantage of statistical irregularities within the training data, where the added perturbations can "move" the image across deep learning decision boundaries. What is even more alarming is the transferability of these attacks to different deep learning models and architectures. This means a successful attack on one model has adversarial effects on other, unrelated models.

In a general sense, adversarial attack through perturbations is not a machine learning vulnerability. Human and biological vision can also be fooled by various methods, i.e. mixing high and low frequency images together, by altering semantically related signals, or by sufficiently distorting the input signal. However, the amount and magnitude of such a distortion required to alter biological perception is at a much larger scale. In this work, we explored this gap through the lens of biology and neuroscience in order to understand the robustness exhibited in human perception. Our experiments show that by leveraging sparsity and modeling the biological mechanisms at a cellular level, we are able to mitigate the effect of adversarial alterations to the signal that have no perceptible meaning. Furthermore, we present and illustrate the effects of top-down functional processes that contribute to the inherent immunity in human perception in the context of exploiting these properties to make a more robust machine vision system.

1. Introduction

In recent years, deep learning has revolutionized nearly all machine learning fields and has been transformational to the community at large. Deep learning has shown great success in supervised learning tasks where a neural network model can been trained on a large amount of labeled training data using backpropagation and gradient descent. However, research has shown that this artificial architecture and learning mechanism can be exploited by adversarial examples [44]. Adversarial examples are a type of signal attack that can change the output class of a classifier by perturbing the stimulus signal by an imperceptible amount. The attack takes advantage of statistical irregularities within the training data, where the added perturbations can "move" the image across deep learning decision boundaries. An illustration of attacked images with their corresponding perturbation signal can be seen in Figure 1.



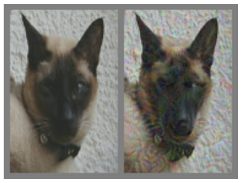
(a) ambulance (=1.0) (b) flamingo (=0.99) (c) cauliflower (=1.0)

(d) ambulance perturb (e) flamingo perturb (f) cauliflower perturb

Figure 1: Illustration of an attack using the Projected Gradient Descent method [29] to target different classes. (a)-(c) are adversarial examples altered by the noise in (d)-(f). ResNet50 [18] classifies the following images as ambulance, flamingo, and cauliflower with confidence levels of 1.0, 0.99, and 1.0, respectively.

In contrast to machine adversarial examples, there are examples that fool humans as seen in Figure 2. These ex-

(a) Hybrid Image example



(b) Time-limited example

(c) Similar features ex.

Figure 2: Examples of "attacks" on human vision. (a) Hybrid image that mixes high and low frequencies [32]. (b) Perturbed example that fools both machines and time-limited humans [14]. (c) Viral photo that illustrates similar features between dogs and food.

hibit different characteristics that rely on more semantic alterations. It is clear from this illustration that the features that the convolutional neural network (CNN) is using for classification are different than the features we humans use for object identification. Furthermore, the scale of perturbations required to alter human perception is often times orders of magnitude larger and are semantically related. In this work, we sought to answer several questions about the gap between humans and machines in perception. Specifically, what is it about biological vision that makes it so robust? What input signal does the brain actually see, i.e. what kind of processing is occurring in the retina? And then given this signal, how does the primary visual cortex process and represent this visual information? To answer these questions, we review the literature and examine the mechanisms that govern human perception.

2. Background and Related Work

2.1. Mammalian Vision

Ramón y Cajal, the founder of modern neuroscience, illustrated the pathways of mammalian retina that were fundamental in the understanding of how we perceive light. To summarize at a high level, light enters the eye through the pupil and lens and is projected to the back of the eye onto the retina, a layer of tissue that lines the back of the eye and responds to light. The retina consists of neurons and photoreceptors that perform phototransduction (translates light into action potentials e.g. "neural spikes") that travel down the optic nerve, to the lateral geniculate nucleus (LGN), and then to the primary visual cortex in the occipital lobe of the brain.

Phototransduction, the concept of transforming light into neural spikes, is a critical first step in understanding how we process visual information. Unlike machine vision that reads and processes *every pixel all at once*, we instead, *sample bits and parts of the world over time*, yet are still able to recreate a holistic representation in our mind's eye. This can easily be verified given that we have a blind spot where the axons of the retina cells go back through the eye, yet we do not perceive a blind spot in our vision. We can also trace the activity and pathways of various retina cells and see that information is processed at different locations at different times. As we will describe in our methodology, our model mimics the function of the retina at the cellular level. The specific cells and their functions are describe below.

Types and Functions of Retina Cells - The first cells to interact with light are the photoreceptors, rods and cones. The rods are responsible for detecting low levels of light; whereas, the cones are capable of color vision. These photoreceptors transmit information to bipolar cells, which aggregate many photoreceptors. At this point, two major pathways emerge, the ON-center and OFF-center pathways. The ON-center bipolar cell is excited e.g. depolarized, if light is strikes the receptive field of this cell, and inhibited e.g. hyperpolarized, if light shines on the surrounding area of the receptive field. In contrast, the OFF-center bipolar cell reacts in the opposite fashion. It will become depolarized when an area of darkness strikes the center of its receptive field, and hyperpolarized when light shines on the surround. [41].

The next type of cell are horizontal cells. Horizontal cells are connected laterally to many rods, cones, and bipolar cells. Their primary role is to inhibit the activity of neighboring cells. This idea of selective suppression of nearby activity is called *lateral inhibition*. By inhibiting signals from less illuminated photoreceptors, the horizontal cells ensure that only the signal from the well lit photoreceptors reaches the ganglion cells, thus improving the contrast and definition of the visual stimulus.

Amacrine cells are also inhibitory neurons that interact with the bipolar cells and retinal ganglion cells. These cells supplement the action of horizontal cells, but also play modulatory roles, controlling the oscillations and firing frequency of retinal cells. Finally, there are the retinal ganglion cells (RGCs). As opposed to other retina cells that pass graded responses, the ganglion cells fire electrical impulses, e.g. action potentials, down a long axon that extends back towards the cortex. The firing rate (spikes per second) corresponds to the stimulus intensity within the neuron. There are many types of ganglion cells, but we focus primarily on the midget cells, which are responsible for responding to

color and contrast. These cells are connected to relatively few cones and rods (as close to a one to one ratio in the fovea) and have ON/OFF receptive fields. The midget cells relay to the parvocellular layers in the lateral geniculate nucleus.

Lateral Geniculate Nucleus (LGN) - Action potentials received from the RGCs represent a neural code that is relayed through the lateral geniculate nucleus (LGN). The LGN receives sensory input directly from the retinal ganglion cells along with many feedback connections from the primary visual cortex. These feedback connections exert both an excitatory and inhibitory influences on the LGN relay neurons [9].

Just as there are many types of retina cells, there are also many different types of LGN cells. We focus on the parvocellular neurons, as these are connected to the midget RGCs. Similar to the midget RGCs, the parvocellular neurons are sensitive to color and are more capable of discriminating fine details than other types of LGN cells. The LGN is also thought to spatially correlate signals by summing the signals received from the left and right eyes. The LGN projects this information to the primary visual cortex.

Primary Visual Cortex (V1) - Primary visual cortex, e.g. V1, is the earliest cortical visual area of the brain and has been extensively studied in the neuroscience literature. The neural representations in V1 are sparse and highly recurrent with many feedback connections. Strong evidence demonstrates that the neural code is both explicit and sparse [15] where neurons fire selectively to specific stimuli. Early research in receptive fields by Hubel and Wiesel's [21] confirmed that individual V1 neurons can primarily be described as Gabor like edge detectors.

As also observed in the retina and LGN, lateral and feedback connections are extremely important to consider [39]. Lateral inhibition was discovered decades ago between neighboring columns in the visual cortex [7, 8]. Early visual neurons in V1 do not act as simple linear feature detectors as they do in artificial neural networks [37], instead, they transform retinal signals and integrate top-down and lateral inputs, which convey prediction, memory, attention, expectation, learning, and behavioral context. Such higher processing is fed back to V1 from cortical and subcortical sources [30]. Later in time, after approximately 100 ms of the presentation of a stimulus, neurons in V1 are also sensitive to the global organization of a scene [3]. These response properties stem from recurrent feedback processing from higher areas. Top-down feedback is also thought to transmit Bayesian inferences of forthcoming inputs into V1 to facilitate perception and reinforce the representation of rewarding stimuli in V1 [24]. Interestingly, there are many more feedback connections in the brain than there are feedforward.

2.2. Fundamental Biological Concepts

The process of biological perception is immeasurably complex and not something that can we hope to replicate at this time. However, there are some clear overarching, high level concepts that seem to be fundamental to vision that we can incorporate and explore in our model.

(1) The retina is transmitting signals in a reduced capacity. Photoreceptors perform convergence to the retinal ganglion cells at a factor of 150 million to 1.5 million; thus, communication relies on a form of compression [22]. In the context of neuroscience, it has been conjectured that this compression may be related to the existence of an efficient coding scheme, such as compressed sensing [4]. Indeed, one concept that we know is that natural stimuli, such as visual images, are sparse in some transform basis [43]. Thus, the retina can sample the world at sub-Nyquist frequencies and still recover the world.

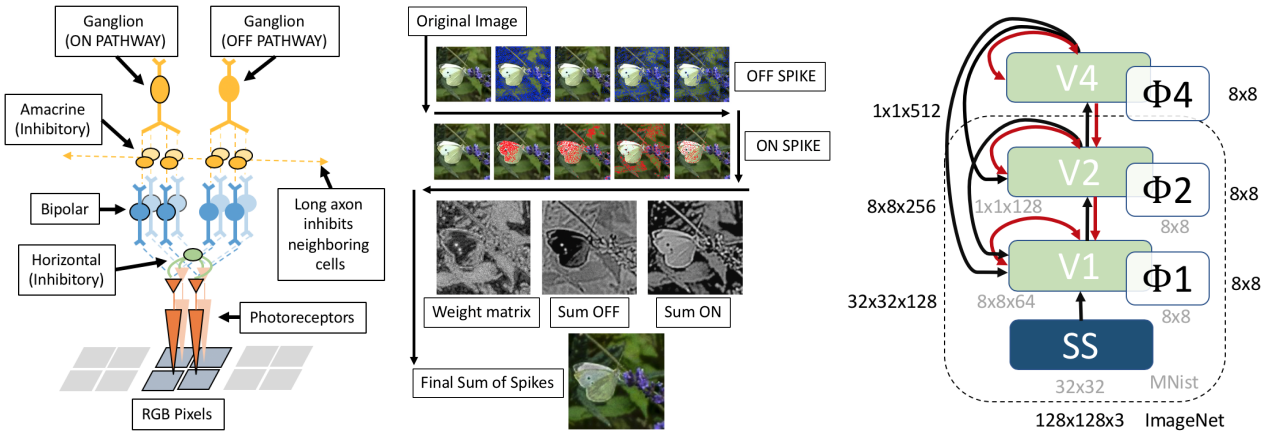
(2) The neural representation is sparse and overcomplete. Olshausen [33] has shown that sparsity is a desirable property as our natural environment can be described by a small number of structural primitives. Sparse codes have a high representational capacity in associative memory, far surpassing the number of input-output pairs that can be stored by a more dense code [5], and that biologically, sparse neural codes are more metabolically efficient and reduce the cost of code transmission [2].

Neural connections in V1 reflect the overcomplete property of the brain. For example, in a cat's V1 there are 25 times as many output fibers as there are input fibers from the LGN, and in macaque V1, the ratio is on the order of 50 to 1 [34]. These properties support a representation that is both sparse and overcomplete.

(3) Feedback is a critical component in perception. Anil Seth, neuroscientist at the University of Sussex, says that perception is a "controlled hallucination". We can see from the neural circuitry that lateral and top-down feedback connections play a significant role in vision. Evidence shows that feedback originating in higher-level areas such as V4, IT, or MT, with bigger and more complex receptive fields, can modify and shape V1 responses, accounting for contextual or extra-classical receptive field effects [10]. In summary, vision is a controlled mix of what we think we see and what we actually we see.

2.3. Background in Adversarial Examples

The underlying problem of deep learning classification is the tendency of the network to learn surface regularities in the data, and not truly learn the abstract concepts of classes and objects. This makes them easily susceptible to adversarial perturbations [23]. Shwartz and Tishby [42] show that each successive layer in a deep learning model learns to throw away the data that is not used to minimize the objective loss. The network is learning how to "forget" about



(a) Retina Model (b) LGN Relay (c) Visual Cortex Model

Figure 3: Overview of our biologically inspired model of perception. The input image passes through the (a) model of the retina consisting of photoreceptors, horizontal, bipolar, amacrine, and retinal ganglion cells. The ganglion cells produce two distinct spike trains corresponding to center ON and surround OFF responses. The spike trains are (b) summed and relayed in the LGN. Finally, in (c) the rate-coded spike train is sparse coded by a hierarchical visual cortex model that uses lateral inhibition and top-down feedback. Inhibitory connections are colored red, while excitatory connections are in black. The dimensions of each layer and convolutional dictionary are shown for MNist (inside dotted line) and ImageNet (outside the dotted line).

the data. The network is not learning what an "apple" is, but rather, what features about this image can I discard in order to create a better decision boundary?

Existing defenses for adversarial examples attempt to "fix" the deep learning model by augmenting the model with adversarial examples, [29, 45] or adding stochasticity to the hidden activations [12]. Alternatively, a defense can address the input to a model by preprocessing, quantization, or compression [47, 11, 17, 19, 28].

Our work is distinct and does not exactly fit in one of these defense designations. Our contributions could be categorized as addressing both the input processing and learning model. First, we present a retina model that is akin to preprocessing the input image by sampling over space and time. Second, we describe how to train and perform inference using a sparse coding model that is driven by a rate coded input signal and modulated with lateral and top-down feedback. Our experiments and results demonstrate the effectiveness of our biologically inspired model in the context of adversarial examples; however, we would like to emphasize that the implications of this model extend to *perception in general*. The presented bio-inspired elements are not tailored for adversarial defense - *yet inherently possess this immunity*. Thus, our last contribution and hope is to motivate the community towards further research in this area.

3. Methodology

Our framework models the process of perception, starting from a digital image and ending in a visual neural representation. We create an anatomically inspired model of the retina that performs phototransduction. The spike train is aggregated into a sum of spikes that are relayed via parvocellular neurons of the LGN. This signal is then converted into an overcomplete sparse representation using a trained dictionary of generators from sparse coding. The neurons at this stage respond to input stimuli in a similar way to recorded neurons in primary visual cortex. Finally, we describe a hierarchical model that enables top-down feedback to drive the lower levels of the model towards a representation that is consistent with both the input image stimulus and top-down expectations, memories, or strong priors. For our experiments and evaluations on adversarial examples, we show that our model exhibits an inherent robustness to adversarial examples that outperforms other defenses. An illustration of the full model is shown in Figure 3.

3.1. The Model of the Retina

The model of our artificial retina consists of the major cell types, e.g. photoreceptors, bipolar, horizontal, amacrine, and ganglion cells. Our architecture extends a previously described retinal model that sought to explain synchronous, stimulus-selective oscillations between retinal ganglion cells [25, 46].

Figure 4: Retina Ganglion Cell ON (red) and OFF (blue) action potentials from ISLVRC2012_val_1003, 1-28 ms. Each image frame displays a 1 ms snapshot. The overlayed blue shows the spikes generated from the OFF pathway, and the red shows spikes generated from the ON pathway. The spikes create "waves" over the image and oscillate at specific frequencies.

Our model consists of a 128x128 array of identical local processing modules that operate over a 256x256 RGB input image. A single processing module is shown in Figure 3(a). Each module has a pixel receptive field of 2x2 patch that relays to cone photoreceptors in the outer plexiform layer. The cones are laterally inhibited by 4x horizontal cells that integrate the response over a set of cones. The cones drive the 4x biopolar cells in two pathways, the ON pathway and OFF pathway. The ON pathway is excited if the input signal is above gray (0-255, gray=128), and the OFF pathway is excited for values below 128. These cells then relay to the retinal ganglion cells with realistic stochasticity, adding an intrinsic source of noise to the model. The ganglion cells are inhibited by two sets of amacrine cells, 4x small and 4x large amacrine cells. The small amacrine cells make the ganglion cell response more transient, i.e. it enables the ganglion to respond to optimally sized small input signals, while the medium amacrine cells provide surround inhibition to the ganglion cell, providing spatial contrast. We note that the small amacrine cell is bistratified (responds to both ON and OFF signals) in order to turn off the ganglion faster and make the ganglion responses sharper.

Biologically, most of the cells in the retina can communicate through graded signals, e.g. floating point values, since minimal signal degradation occurs over short distances. However, the ganglion cells must transmit over a long optic nerve, thus requiring action potentials. The internal state of the integrate and fire cells, i.e. "membrane potential", charges up and when it exceeds a certain threshold, will activate and fire that neuron. We implement the cells of our model as leaky integrators for graded cells, or leaky integrate and fire neurons for ganglion cells with time constants consistent with biological processes.

Mathematically, we can model the membrane potential, V^k, where k is a particular cell type (cone, horizontal, bipolar, amacrine, ganglion) as the following,

$$\dot{V}^k = -\frac{1}{\tau^k}\Big[V^k - b^k - L^k - \sum_{k'} W^{(k,k')} \cdot f(V^{k'}) \cdot W^{(k,k')^T}\Big] \quad (1)$$

Where τ^k are the time constants, b^k are bias currents, L^k is the light input (image pixels) $L^k = 0, k \neq 1$. The input-output relations are defined by $f(V^{k'})$, and the weight matrices $W^{(k,k')}$ are separable Gaussian functions computed with the distance between pre- and post-synaptic column locations in the weight matrix.

For the graded input-output relation function, we define $f(V^{k'})$ as a piecewise linear saturation function, or, as a step function in the case of a spiking output. The resting membrane potentials, threshold criterion, refractory periods, and chemical interactions have been meticulously tuned to be consistent with biological measurements. Exact parameterizations can be found in [25].

The ON and OFF spike trains can be seen in Figure 4, and the sum of spikes can be seen later in Figure 7 (g). Our outputs results are consistent with the literature in the functions the retina. It was shown by Atick and Redlich [1] that center-surround receptive fields of retinal ganglion cells serve to decorrelate natural visual input e.g. "whitening" of the input. Experimental evidence shows that indeed there is decorrelation in the early stages of the visual pathway [16], and that it may be a necessary preprocessing step before circuitry in V1 can achieve a sparse representation [13]. Other non-spiking retinal models, Figure 7 (f) also show that in a retina output, the mean luminance energy is attenuated, spectrum is whitened and all contours are enhanced [6]. Coincidentally, the ZCA whitening operation is often used as a data preprocessing step in deep learning and convolutional neural networks [35].

3.2. The Model of the LGN

In our model, the primary purpose of the LGN is to act as a relay from the retina to primary visual cortex. However, since the LGN also serves to spatially correlate signals by summing binocular signals, our model of the LGN also sums noisy spike trains arising from the optic nerve. The ON and OFF retinal spike trains are summed over 128 ms into a matrix with the same dimensions as the input image. We can relay this rate coded matrix to the primary cortex, or in the case of an RGB image, we can use this matrix as a weighting matrix. The final sum of spikes output is obtained by a point-wise multiplication with the weighting

matrix and original image. The process can be seen in Figure 3(b).

3.3. The Model of Primary Visual Cortex

The final component of our framework is the model of the primary visual cortex as illustrated in Figure 3(c). The algorithm governing our approach to create a plausible neural representation is based upon deep sparse coding [26, 27]. Mathematically, sparse coding is a reconstruction minimization problem which can be defined as follows. In the sparse coding model, we have some input variable $x^{(n)}$ from which we are attempting to find a latent representation $a^{(n)}$ (we refer to as "activations") such that $a^{(n)}$ is sparse, e.g. contains many zeros, and we can reconstruct the original input, $x^{(n)}$ with high fidelity. A single layer of sparse coding can be defined as,

$$\min_{\Phi} \sum_{n=1}^{\mathcal{N}} \min_{a^{(n)}} \frac{1}{2} \|x^{(n)} - \Phi a^{(n)}\|_2^2 + \lambda \|a^{(n)}\|_1 \qquad (2)$$

Where Φ is the overcomplete dictionary, and $\Phi a^{(n)} = \hat{x}^{(n)}$, or the reconstruction of $x^{(n)}$. The λ term controls the sparsity penalty, balancing the reconstruction versus sparsity term. $\mathcal{N}$ is the total training set, where n is one element of training. Φ represents a dictionary composed of small kernels that share features across the input signal.

We use the Locally Competitive Algorithm (LCA) [40] to minimize the mean-squared error in Equation 2. The LCA algorithm is a biologically informed sparse solver governed by dynamics that evolve the neuron's membrane potential when presented with some input stimulus. Activations of neurons in this model laterally inhibit units within the layer to prevent them from firing. The input potential to the state is proportional to how well the image matches the neuron's dictionary element, while the inhibitory strength is proportional to the activation and the similarity of the current neuron and competing neuron's convolutional patches, forcing the neurons to be decorrelated. The LCA model is an energy based model similar to a Hopfield network [20] where the neural dynamics can be represented by a nonlinear ordinary differential equation. Let us consider a single input signal, in our case the sum of spikes, ss. We define the internal state of a particular neuron, m, as u^m and the active coefficients as $a^m = T_\lambda(u^m)$, where T is an activation function with threshold parameter, λ.

The dynamics of each node is determined by the ordinary differential equation,

$$\dot{u}^m = \frac{1}{\tau}\Big[-u^m + (\Phi^T ss) - (\Phi^T \Phi a^m - a^m)\Big] \qquad (3)$$

The $-u^m$ term is leaking the internal state, τ is the time constant, the $(\Phi^T ss)$ term is "charging up" the the state by the inner product (match) between the dictionary element and input sum of spikes signal, and the $(\Phi^T \Phi a^m - a^m)$ term (drawn as the red recurrent connection in Figure 3(c)) represents the lateral inhibition signal from the set of active neurons proportional to the inner product between dictionary elements. The $-a^m$ in this case is eliminating self interactions. In summary, neurons that match the input rate coded image charge up faster, then pass a threshold of activation. Once they pass the threshold, other neurons in that layer are suppressed proportional to how similar the dictionary elements are between competing neurons. This prevents the same image component from being redundantly represented by multiple nodes.

Upon further inspection, one can see that this is a generative model where the objective function is minimizing reconstruction error. The model is not learning decision boundaries and not throwing away information to maximize classification. In contrast, *the goal of the network is to remember everything, not forget.*

3.4. Top-down Feedback

In a recent paper from Google Brain, Elsayed et al. [14] show that adversarial examples can fool time-limited humans, but not no-limit humans. If an image is quickly presented to a human the human may make a classification mistake, but not if given unlimited time. "One possible explanation ... is that no-limit humans are fundamentally more robust to adversarial examples and achieve this robustness via top-down or lateral connections." And this, "suggests that machine learning security research should explore the significance of these top-down or lateral connections further."

In our neural network, we can stack sparse layers where each layer is attempting to reconstruct the previous layer's internal state, i.e. layer $N+1$ is reconstructing the membrane potential at layer N, u^N. The residual, or error of reconstruction of the top layer r^{N+1}, can be used as an inhibitory signal driving the mechanics of the lower layer,

$$r^{N+1} = u^N - \Phi^{N+1} a^{N+1} \qquad (4)$$

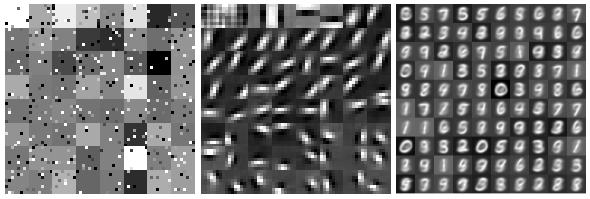
such that equation of all neurons at layer N is defined as,

$$\dot{u}^N = \frac{1}{\tau}\Big[-u^N + (\Phi^{N^T} u^{N-1}) - (\Phi^{N^T} \Phi^N a^N - a^N) - r^{N+1}\Big] \qquad (5)$$

This inhibitory connection is illustrated as the red feedback arrows from layers V4 to V2, and from layers V2 to V1 in Figure 3 (c). This connection has the effect of inhibiting neurons at a lower layer that are not consistent with high level representations.

However, as noted in biology, feedback connections are both inhibitory and excitatory. Thus, we create excitatory feedback connections from higher levels - V4 to V2, V2 to V1, and V4 to V1, as additional drivers to a sparse coding

Figure 5: The effect of top-down feedback on input images (a). (b)-(e) show the evolution of the internal representation at intervals of 1000 timesteps. Column (f) shows the converged representation at t=4000. Rows (1)-(3) illustrate top-down influences moving the image towards a learned canonical form. Rows (4),(5) show hallucination effects of different classes using strong priors from the top level of the hierarchy. The green pixels indicate addition while the red pixels indicate subtractions.



(a) Φ1 init (b) Φ1 trained (c) V2 Activity

Figure 6: (a) shows the initialization of Φ1 in the MNist model. (b) is the fully trained Φ1. (c) shows the activity triggered average of 100 neurons i.e. weighted sum of the input that activates that neuron, in layer V2.

layer. The "driver" is added through the error layer, altering equation 5 to,

$$\Phi^{N^T} u^{N-1} \rightarrow \Phi^{N^T}(u^{N-1} + \lambda \Phi^N(\Phi^{N+1} a^{N+1})) \quad (6)$$

Thus, the sparse coding layer is not only influenced by the input from the lower layer, but also guided by the activated reconstruction of the higher layer. One important caveat is that we do not turn on excitatory top-down feedback during training ($\lambda = 0$), as we want our dictionary to learn on stimuli that it actually sees, not what it thinks it sees. To illustrate the effects of top-down feedback, we train a 2-layer (V1 and V2 only) sparse coding network using the MNist dataset. Results of reconstructions at the V1 level can be seen in Figure 5 and learned dictionary in Figure 6.

As mentioned previously, our network shares properties with Hopfield memory networks, which have been used previously to model the cerebral cortex [36]. When given excitatory feedback, the model drives the input towards canonical forms of a digit learned at higher layers in the model. Given more time to "look" at the image, the model converges closer and closer towards high-level memories stored at local minimums of the network.

In the event of ambiguous input, we can force the model to hallucinate different digits by manually activating neurons at the V2 level and increasing $\lambda = 10$. Figure 5 (4)(5) illustrates the effect of activating the "5" neurons and "8" neurons, respectively. This exemplifies the idea of "controlled hallucinations" in perception. In fact, psychology literature [38] proposes that hallucinations can be understood as top-down effects on perception, mediated by inappropriate perceptual priors.

4. Experiments and Results

For our experiments on adversarial examples, we randomly sample 1,000 images from the ILSVRC2012 validation image set and run them through a pretrained ResNet50 [18] classifier. This reveals the baseline accuracy of 56.0% for top-1 and a top-5 accuracy of 79.3%. Next, we attacked the images with a state-of-the-art gradient based method, Projected Gradient Descent [29]. The resulting top-1 and top-5 accuracies drop to 9.5% and 17.7%, respectively.

For adversarial defense comparisons, we use a suite of methods available from the ART toolbox [31]. The defense methods applied to the attacked images included a spatial smoothing [47], jpeg compression [11], and total variation minimization [17]. As another comparison, we present results from a parvocellular bioinspired retinal model [19]. This model was designed to be biologically inspired to perform texture analysis and enhance details which are robust against input images luminance ranges. Quantitative results are presented in in Table 1 and qualitative results can be seen in Figure 7.

Methods that do not alter the input image significantly (like JPG compression with q=75) maintain high accuracy, but provide virtually no protection against adversarial perturbations. In contrast, our retina model enhances edges and whitens the signal, and our sparse coding model denoises the image. For ImageNet, our top-down feedback does not have significant quantitative effect, but qualitatively changes the output. We will explore these effects in future research. Overall, our model performs the best against adversarial perturbations and is closest to maintaining the accuracy between the original and attacked data.

(a) Original (b) Attacked (c) Smooth (d) JPG C. (e) TVmin (f) Parvo (g) Spikes (h) SC (i) SC+Top

Figure 7: Qualitative examples of (b) an attacked image and the output of various protection methods. (c) Spatial smooth window of 4, (d) JPEG compression quality of 50, (e) Total variation norm of 1, (f) Parvocelluar model horizontal gain of 0.001, (g) Sum of spikes from our retinal model illustrating decorrelation and edge enhancements, (h) Sum of spikes sparse coded using our cortex model, (i) Sum of spikes sparse coded with top down feedback.

Method	**T1**	**T5**	**AT1**	**AT5**
Original images	56.0	79.3	9.5	17.7
Spatial Smoothing, Xu el al. [47]				
Smooth window = 3	49.7	70.9	11.6	46.7
Smooth window = 4	44.0	66.4	21.6	55.0
Smooth window = 5	37.1	59.8	23.2	53.6
JPEG Compression, Das el al. [11]				
JPEG quality = 75	55.8	79.6	12.4	51.6
JPEG quality = 50	49.7	74.2	10.9	38.0
JPEG quality = 25	45.9	70.5	23.5	60.7
Total Variation Minimzation, Guo el al. [17]				
TV norm = 1	40.8	63.6	25.3	57.3
TV norm = 2	23.6	44.6	21.0	41.7
Parvocellular retina model, Herault el al. [19]				
Horizontal gain = 0.0001	36.2	61.2	23.8	53.2
Horizontal gain = 0.001	36.0	61.3	24.0	53.2
Horizontal gain = 0.01	35.2	60.4	23.9	52.1
Our method				
Retina Model Only	46.8	68.9	30.2	62.6
Retina & Sparse Coding (SC)	48.8	70.8	35.4	**67.7**
Retina & SC & Topdown	48.2	70.2	**35.7**	65.4

Table 1: Top 1 (T1) and top 5 (T5) classification accuracy on a subset of images from the ILSVRC 2012 validation set. AT1 is top1 and AT5 is top5 accuracy on the same set of images attacked with adversarial noise.

5. Conclusion

Given our framework, model, and results, we return to the original question posed, what is it about biological vision that makes it so robust? Some researchers theorize that media equipment (like monitors or print media) do not display the perturbations with proper fidelity and thus doesn't effect vision. Alternatively, Zhou et al. [48] suggests there are physiological limitations on human visual acuity, resolution, and sensitivity to contrast, which simply cannot match the resolving power of in silico image processing.

However, from our research, we are convinced that *humans do not even see most adversarial perturbations*. It is not the case that we see it and somehow ignore it, but instead, it probably does not even reach the cortex. The retina samples the world over space and time, exhibits inherent stochasticity within cellular responses, and converts to an optimal spike code. Even if the perturbation reaches the cortex, the sparsity and generative properties of the brain fill in and denoise a signal, all while being modulated by lateral and top-down feedback.

We also conclude that *feedback is fundamental to human perception.* Lateral inhibition starts in the early stages of the retina, and top-down feedback exists everywhere in the LGN and cortex, yet is often ignored in machine learning models. Our work begins to scratch the surface on the effects of inhibitory and excitatory feedback in a hierarchical model, but more research is warranted.

In summary, we created a biologically inspired model that begins with a digital image that is processed by an artificial retina model. The retina creates a noisy spike train that is rate coded and relayed to a plausible model of the primary visual cortex that incorporates lateral and top-down feedback. Through this encoding process, we demonstrate that the resulting output is inherently immune to adversarial examples and we explain this result both biologically and mathematically. Ultimately, the goal of this work is not to be a comprehensive, new defense against adversarial attack, but rather, a work that emphasizes the benefit and need for more research on biologically inspired models of vision.

6. Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1954364.

References

[1] Joseph J Atick and A Norman Redlich. What does the retina know about natural scenes? *Neural computation*, 4(2):196–210, 1992.

[2] Roland Baddeley. Visual-perception-an efficient code in v1. *Nature*, 381(6583):560–561, 1996.

[3] Lauren Barghout. *Vision: Global Perceptual Context Changes Local Contrast Processing Updated to include computer vision techniques*. Scholars' Press, 2014.

[4] Victor J Barranca, Gregor Kovačič, Douglas Zhou, and David Cai. Sparsity and compressed coding in sensory systems. *PLoS computational biology*, 10(8):e1003793, 2014.

[5] Eric B Baum, John Moody, and Frank Wilczek. Internal representations for associative memory. *Biological Cybernetics*, 59(4-5):217–228, 1988.

[6] Alexandre Benoit, Alice Caplier, Barthélémy Durette, and Jeanny Hérault. Using human visual system modeling for bio-inspired low level image processing. *Computer vision and Image understanding*, 114(7):758–773, 2010.

[7] Colin Blakemore, Roger HS Carpenter, and Mark A Georgeson. Lateral inhibition between orientation detectors in the human visual system. *Nature*, 228(5266):37–39, 1970.

[8] Colin Blakemore and Elisabeth A Tobin. Lateral inhibition between orientation detectors in the cat's visual cortex. *Experimental Brain Research*, 15(4):439–440, 1972.

[9] Javier Cudeiro and Adam M Sillito. Looking back: corticothalamic feedback and early visual processing. *Trends in neurosciences*, 29(6):298–306, 2006.

[10] István Czigler and István Winkler. *Unconscious Memory Representations in Perception: Processes and mechanisms in the brain*, volume 78. John Benjamins Publishing, 2010.

[11] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017.

[12] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

[13] Eric McVoy Dodds, Jesse Alexander Livezey, and Michael Robert DeWeese. Spatial whitening in the retina may be necessary for v1 to learn a sparse representation of natural scenes. *BioRxiv*, page 776799, 2019.

[14] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, pages 3910–3920, 2018.

[15] Peter Foldiak. Sparse coding in the primate cortex. *The handbook of brain theory and neural networks*, 2003.

[16] Katrin Franke, Philipp Berens, Timm Schubert, Matthias Bethge, Thomas Euler, and Tom Baden. Inhibition decorrelates visual feature representations in the inner retina. *Nature*, 542(7642):439, 2017.

[17] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens Van Der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017.

[18] 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.

[19] Jeanny Hérault. *Vision: Images, signals and neural networks: Models of neural processing in visual perception*. World Scientific, 2010.

[20] John J Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the national academy of sciences*, 81(10):3088–3092, 1984.

[21] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[22] Guy Isely, Christopher Hillar, and Fritz Sommer. Deciphering subsampled data: adaptive compressive sampling as a principle of brain communication. In *Advances in neural information processing systems*, pages 910–918, 2010.

[23] Jason Jo and Yoshua Bengio. Measuring the tendency of cnns to learn surface statistical regularities. *arXiv preprint arXiv:1711.11561*, 2017.

[24] Hulusi Kafaligonul, Bruno G Breitmeyer, and Haluk Öğmen. Feedforward and feedback processes in vision. *Frontiers in psychology*, 6, 2015.

[25] Garrett T Kenyon, Bartlett Moore, Janelle Jeffs, Kate S Denning, Greg J Stephens, Bryan J Travis, John S George, James Theiler, and David W Marshak. A model of high-frequency oscillatory potentials in retinal ganglion cells. *Visual neuroscience*, 20(5):465–480, 2003.

[26] Edward Kim, Darryl Hannan, and Garrett Kenyon. Deep sparse coding for invariant multimodal halle berry neurons. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1111–1120, 2018.

[27] Edward Kim, Edgar Lawson, Keith Sullivan, and Garrett T Kenyon. Spatiotemporal sequence memory for prediction using deep sparse coding. In *Proceedings of the 7th Annual Neuro-inspired Computational Elements Workshop*, pages 1–7, 2019.

[28] Edward Kim, Jessica Yarnall, Priya Shah, and Garrett T Kenyon. A neuromorphic sparse coding defense to adversarial images. In *Proceedings of the International Conference on Neuromorphic Systems*, page 12. ACM, 2019.

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[30] Lars Muckli and Lucy S Petro. Network interactions: Nongeniculate input to v1. *Current Opinion in Neurobiology*, 23(2):195–201, 2013.

[31] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.0.1. *CoRR*, 1807.01069, 2018.

[32] Aude Oliva, Antonio Torralba, and Philippe G Schyns. Hybrid images. *ACM Transactions on Graphics (TOG)*, 25(3):527–532, 2006.

[33] Bruno A Olshausen and David J Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision research*, 37(23):3311–3325, 1997.

[34] Bruno A Olshausen and David J Field. How close are we to understanding v1? *Neural computation*, 17(8):1665–1699, 2005.

[35] Kuntal Kumar Pal and KS Sudeep. Preprocessing for image classification by convolutional neural networks. In *2016 IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*, pages 1778–1781. IEEE, 2016.

[36] Günther Palm. Neural associative memories and sparse coding. *Neural Networks*, 37:165–171, 2013.

[37] Lucy S Petro, Luca Vizioli, and Lars Muckli. Contributions of cortical feedback to sensory processing in primary visual cortex. *Frontiers in psychology*, 5, 2014.

[38] Albert R Powers III, Megan Kelley, and Philip R Corlett. Hallucinations as top-down effects on perception. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 1(5):393–400, 2016.

[39] Hettie Roebuck, Patrick Bourke, and Kun Guo. Role of lateral and feedback connections in primary visual cortex in the processing of spatiotemporal regularity- a tms study. *Neuroscience*, 263:231–239, 2014.

[40] Christopher Rozell, Don Johnson, Richard Baraniuk, and Bruno Olshausen. Locally competitive algorithms for sparse approximation. In *Image Processing, 2007. ICIP 2007. IEEE International Conference on*, volume 4, pages IV–169. IEEE, 2007.

[41] Peter H Schiller, Julie H Sandell, and John HR Maunsell. Functions of the on and off channels of the visual system. *Nature*, 322(6082):824–825, 1986.

[42] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[43] Eero P Simoncelli and Bruno A Olshausen. Natural image statistics and neural representation. *Annual review of neuroscience*, 24(1):1193–1216, 2001.

[44] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[45] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017.

[46] Yijing Watkins, Austin Thresher, David Mascarenas, and Garrett T Kenyon. Sparse coding enables the reconstruction of high-fidelity images and video from retinal spike trains. In *Proceedings of the International Conference on Neuromorphic Systems*, page 8. ACM, 2018.

[47] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

[48] Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature communications*, 10(1):1334, 2019.