Sparse coding protects against adversarial attacks

Dylan M Paiton, Joel Bowen, Jasmine Collins, Charles Frye, Alex Terekhov, Bruno Olshausen

Summary: Lateral and feedback connections abound in biological neural networks [1], but are absent from the feedforward, convolutional architectures popular in artificial neural networks. Artificial neural networks also exhibit adversarial examples [2]: small perturbations to the input can produce large changes to the output. Recent work [3] has identified adversarial examples that transfer to human observers, but only when those observers are time-limited. This suggests that the slower lateral and feedback computations help protect biological neural networks from adversarial examples. In this work, we demonstrate that a classifier with lateral connectivity in its hidden layer is more robust to adversarial examples than a feedforward classifier, requiring larger input perturbations to achieve the same change in output. The hidden layer uses the Locally-Competitive Algorithm (LCA) [4], a leaky integrator network implementation of sparse coding [5]. We hypothesize that the curved geometry of neuron responses with population nonlinearities (via lateral connections) [6] improves selectivity and provides this protection. We analyze this geometry in the context of adversarial examples.

Additional Details: Pointwise nonlinearities are a function of only a single neuron within a layer and include rectification and sigmoid. Population nonlinearities represent an alternative class, where the nonlinear output is a function of multiple neurons in a set. Examples include softmax, divisive normalization [7,8], and the network nonlinearity present in sparse coding [4,5]. Biological neurons are highly interconnected and exhibit strong population nonlinear effects, but nearly all work in neuron modeling uses pointwise nonlinearities due to the ease in interpretation and implementation. However, population nonlinearities have been used to great effect to explain non-classical visual receptive fields, which are nonlinear neuron response properties that cannot be explained by thresholding or saturation alone [5,7,8,9].

To better understand the difference between these classes of nonlinearity, we visualize the input-output maps of model neurons in the form of iso-response contours (Figs 1,2). The iso-response contours of linear neurons are linear: any input perturbation that is orthogonal to the weight vector will result in equal activation. For pointwise nonlinearities, this remains true: because the nonlinearity is performed after a linear projection, the output must also produce straight iso-response contours. For a population nonlinearity, the gradient of the activation function with respect to the a small perturbation in the input is a function of all other neurons. Thus, for a perturbation that is orthogonal to a target neuron, it is highly likely that an alternative neuron will have a non-orthogonal weight vector, which will result in a net change in all neuron outputs.

Adversarial examples are closely tied to neuron iso-response contours. While an iso-response contour represents a perturbation direction in stimulus space that produces no change in the output, an adversarial example is a perturbation direction that produces a maximal change in the output, which will be orthogonal to the iso-response surface. To test how population nonlinearities affect a network's susceptibility to adversarial attacks, we trained three networks on the MNIST classification dataset. The first network is an MLP that has one hidden layer with 768 rectified linear units. The next network is a combination of LCA with 768 outputs and a single layer perceptron (SLP) classifier. LCA is a recurrent network of leaky-integrator neurons that converges to minimize the sparse coding objective function [4] and is trained without supervised labels. The final network is LISTA [10] and an SLP. LISTA is a pointwise nonlinear feedforward network with 768 outputs that is trained to produce a code with a small L_2 distance to that produced by LCA. The SLPs and MLP were all trained with a supervised cross-entropy loss. We then produced adversarial attacks following the method in [2]. We show that for a fixed adversarial confidence, the attacks on the LCA+SLP network are 1) larger and 2) closer to the space of possible MNIST digits than for the other two networks (Fig 3).

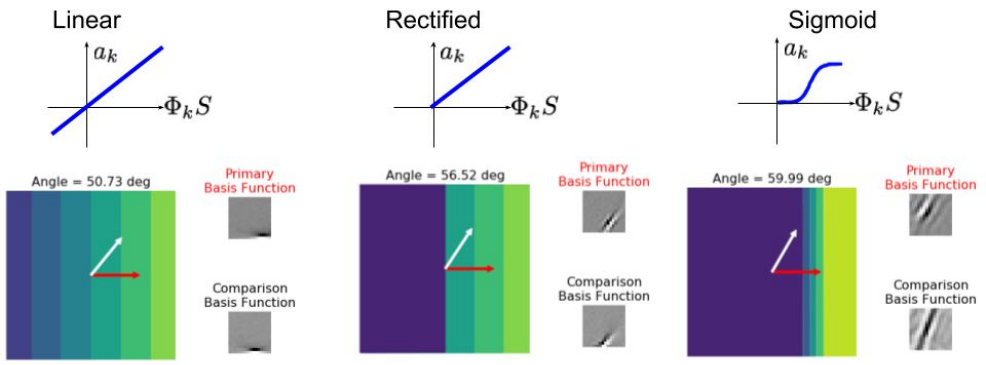

Figure 1: Pointwise nonlinear neurons always produce straight iso-response contours. Here our neuron's output (contour colors), a_k, is computed by applying a pointwise nonlinear function, $g(.)$, to a linear projection, $\Phi_k S$, where Φ_k is a neuron's weight vector (red arrow) and S is the input image. The white arrow indicates a comparison neuron that has an inner product within this stimulus space and therefore would also respond to inputs. The 2D plot represents a plane of points that live in the 256D image space.

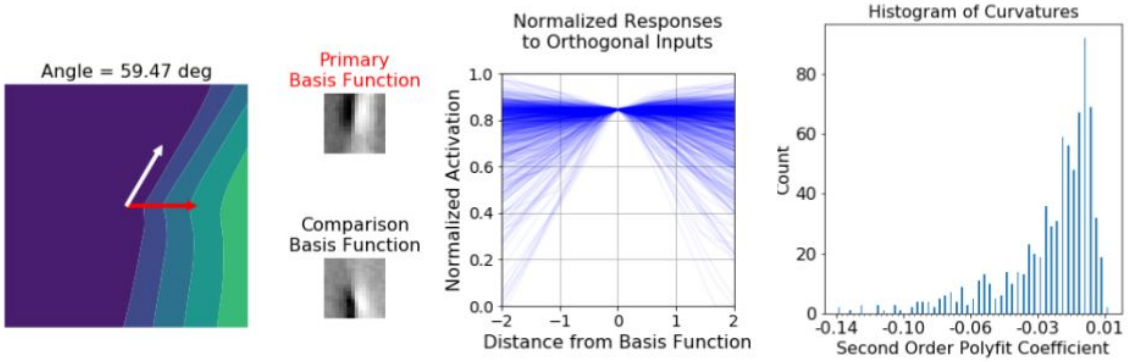

Figure 2: Population nonlinear neurons can have bent iso-response contours. Here we use LCA to demonstrate that population nonlinearities can produce curved iso-response contours. The left plot is constructed as in Fig 1. The middle plot shows the activation for 768 planes, where each is chosen with respect to a different comparison neuron. The right plot is a histogram of second order polynomial fit coefficients for the curves in the middle plot. Nearly all of the contours are flat (coefficient=0) or have exo-origin curvature (coefficient<0), indicating that this neuron has a high degree of selectivity for many orthogonal directions.

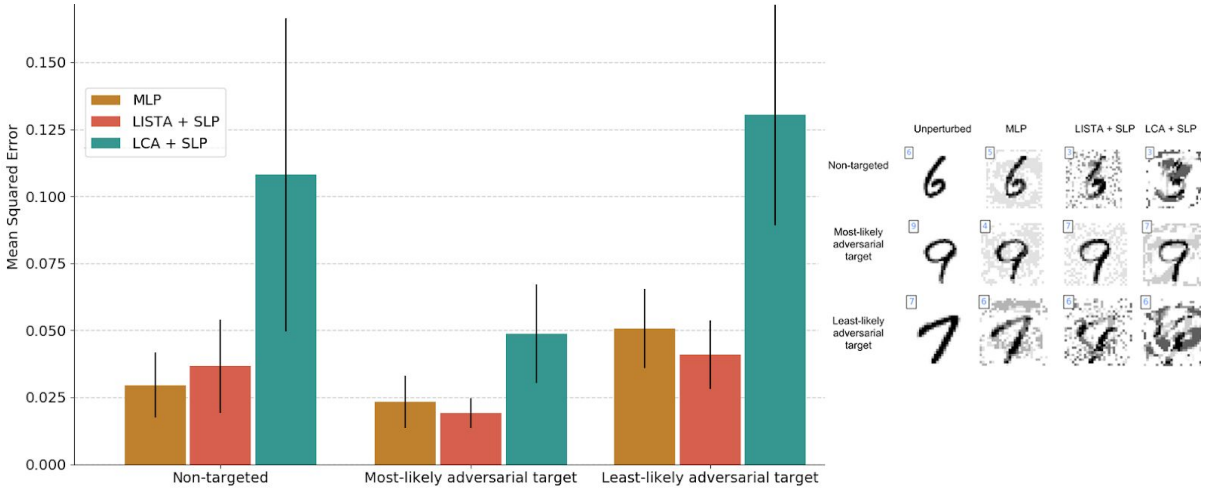

Figure 3: LCA+SLP is more protected against adversarial attacks. We trained a classifier on MNIST digits and show that using LCA as a preprocessing step protects the classifier against adversarial attacks. All models were trained to approximately equal test accuracy (98.2%, 97.15%, 97.76% for MLP, LISTA, LCA). All attacks were controlled for equal adversarial confidence across models. Error bars indicate standard deviation for 100 example inputs.

References:

[1] Xu, X., et al. "Primary visual cortex shows laminar-specific and balanced circuit organization of excitatory and inhibitory synaptic connectivity." *The Journal of physiology* (2016)

[2] Szegedy, C., et al. "Intriguing properties of neural networks." *arXiv:1312.6199* (2013)

[3] Elsayed, G.F., et al. "Adversarial examples that fool both human and computer vision." *arXiv:1802.08195* (2018)

[4] Rozell, C.J., et al. "Sparse coding via thresholding and local competition in neural circuits." *Neural Comp.* (2008)

[5] Olshausen, B.A., et al. "Sparse coding with an overcomplete basis set: A strategy employed by V1?." *Vision research* (1997)

[6] Golden, J.R., et al. "Conjectures regarding the nonlinear geometry of visual neurons." *Vision research* (2016)

[7] Carandini, M., et al. "Normalization as a canonical neural computation." *Nat Rev Neurosci* (2012)

[8] Schwartz, O. et al. "Modeling surround suppression in V1 neurons with a statistically-derived normalization model." *NIPS*, (1998)

[9] Zhu, M., et al. "Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system." *PLoS Comp Bio* (2013)

[10] Gregor, Karol, and Yann LeCun. "Learning fast approximations of sparse coding." *ICML*. (2010)