HIERARCHICAL BINDING IN CONVOLUTIONAL NEURAL NETWORKS: MAKING ADVERSARIAL ATTACKS GEOMETRICALLY CHALLENGING

Niels Leadholm & Simon Stringer
Department of Experimental Psychology
University of Oxford
Oxford, OX2 6GG, UK
{niels.leadholm}@seh.ox.ac.uk

ABSTRACT

We approach the issue of robust machine vision by presenting a novel deep-learning architecture, inspired by work in theoretical neuroscience on how the primate brain performs visual feature binding. Feature binding describes how separately represented features are encoded in a relationally meaningful way, such as an edge composing part of the larger contour of an object, or the ear of a cat forming part of its head representation. We propose that the absence of such representations from current models such as convolutional neural networks might partly explain their vulnerability to small, often humanly-imperceptible changes to images known as adversarial examples. It has been proposed that adversarial examples are a result of 'off-manifold' perturbations of images, as the decision boundary is often unpredictable in these directions. Our novel architecture is designed to capture hierarchical feature binding, providing representations in these otherwise vulnerable directions. Having introduced these representations into convolutional neural networks, we provide empirical evidence of enhanced robustness against a broad range of L_0, L_2 and L_∞ attacks in both the black-box and white-box setting on MNIST, Fashion-MNIST, and CIFAR-10. While we eventually report that the model remains vulnerable to a sufficiently powerful attacker, we demonstrate that our main results cannot be accounted for by trivial sources of false robustness. Through the controlled manipulation of a key hyperparameter, we also provide evidence that this robustness is dependent on the introduction of the hierarchical binding representations. Finally, we propose how such representations relate to the observation that, under appropriate viewing conditions, humans show sensitivity to adversarial examples.

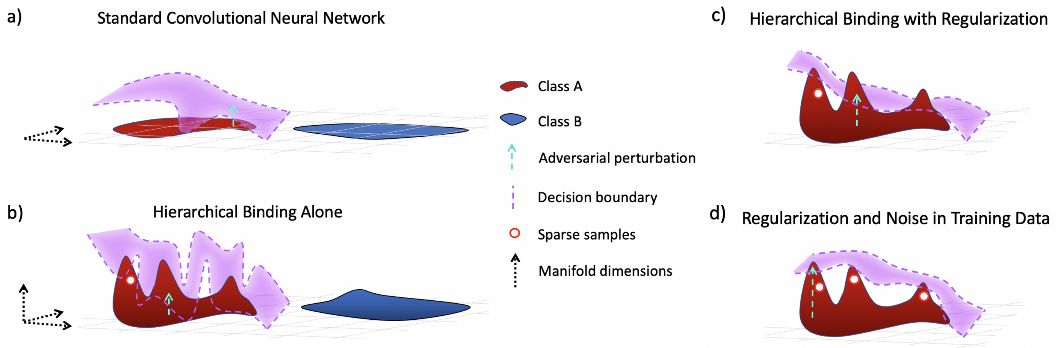
a)
Standard Convolutional Neural Network
b)
Hierarchical Binding Alone
Class A
Class B
Adversarial perturbation
Decision boundary
Sparse samples
Manifold dimensions
c)
Hierarchical Binding with Regularization
d)
Regularization and Noise in Training Data

Figure 1: *The ability of hierarchical binding to support an improved decision boundary.* Red and blue represent two different object manifolds (e.g. cats and dogs). Adversarial perturbations (light-blue arrows) for the red class move the input beyond the decision boundary into a region where it is classified as blue. a) A common assumption for classification is to represent object classes in low-dimensions, which enables linear decision boundaries that accurately separate them. Unfortunately, the learned decision boundary can be unpredictable off the manifold (here the manifold is represented as a 2D surface). Given the high-dimensional embedding space (e.g. pixel-space), there may be many such directions vulnerable to small perturbations. b) We argue that there are additional, class-preserving dimensions of variation to the underlying object manifold (here depicted as a 3D solid), but that these are difficult to model with typical convolutional neural network (CNN) architectures. Adding hierarchical binding enables the network to explicitly represent these features alongside the more abstract dimensions, but due to the sparsity of samples in high-dimensions, further steps are required for a robust decision boundary. c) Introducing regularization such as label smoothing means that even sparse data points can inform a more useful decision boundary. d) Complimenting label smoothing with noise during training helps further address the sampling problem, providing a more robust decision boundary against a variety of adversarial attacks.