HIERARCHICAL BINDING IN CONVOLUTIONAL NEURAL NETWORKS CONFERS ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

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 a small 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. We further provide evidence, through the controlled manipulation of a key hyperparameter, synthetic data-sets, and ablation analyses, that this robustness is dependent on the introduction of the hierarchical binding representations.

1 INTRODUCTION

Adversarial examples are images modified by small (L_p norm constrained) perturbations that cause machine vision systems to catastrophically misclassify objects (Szegedy et al., 2014). Since their discovery, various efforts have been made at both explaining their existence, and conferring resistance to them (Goodfellow et al., 2015; Madry et al., 2018; Gilmer et al., 2018; Schott et al., 2019; Ilyas et al., 2019; Stutz et al., 2019). This includes arguments that adversarial examples represent perturbations of the input off of the class manifold (Tanay & Griffin, 2016; Khoury & Hadfield-Menell, 2018; Stutz et al., 2019), and in particular that models with a high co-dimension (difference in the dimension between the embedding space and the representational manifold), will have many such directions in which it can be attacked (Khoury & Hadfield-Menell, 2018) (Figure 1a).

We argue that part of the phenomenon of adversarial examples is that model representations often assume a manifold of the object class that is too low-dimensional. This assumption entails the loss of low-level spatial details of the input in order to achieve invariance and linear separability of classes. Thus many adversarial examples appear off-manifold because the model does not represent these low-level features, even when many of them may represent perceptually visible and class preserving changes to the object. Without an explicit representation of these features, the decision boundary can lie close to the low-dimensional manifold (Tanay & Griffin, 2016). We propose that with an explicit representation, many adversaries could be viewed as on-manifold adversarial examples, after which robustness is a case of standard generalization (Gilmer et al., 2018; Stutz et al., 2019). We introduce a novel architecture that, through hierarchical feature binding (defined below), captures such information *alongside* invariant representations. The difficulty then is learning a useful decision boundary in such a high-dimensional space. Our contributions are as follows:

- We present a mechanism to preserve low-level information about an object representation through hierarchical binding; these representations are inspired by those predicted to exist within the primate brain, and offer an explanation for the apparent sensitivity of humans to adversarial examples under appropriate conditions.
- We present empirical results showing the robustness of these augmented networks to adversarial examples, following the use of techniques to ensure a useful decision boundary.
- We use iterative adjustments of a key hyperparameter (the γ-proportion), a synthetic dataset, and ablation analyses to provide evidence that the performance boost seen is indeed a consequence of the low-level representations preserved by hierarchical binding.

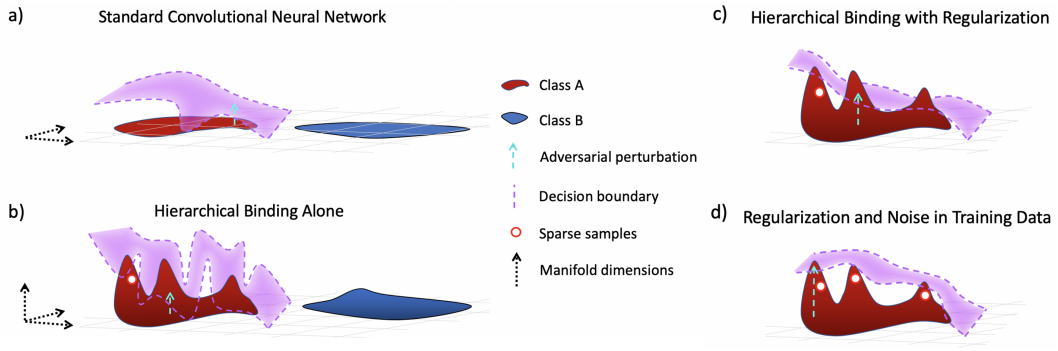

Figure 1: *The effect of binding on the decision boundary.* Red and blue represent two different object manifolds (e.g. cats and dogs). Adversarial perturbations (light-blue arrows) for the red class move the input beyond the decision boundary into a region where it is classified as blue. a) A common 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. 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, 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 than in a basic model.

2 Related Work

Binding Feature binding describes the brain's ability to jointly represent and reason about features that are encoded separately (such as the colour and shape that jointly describe a yellow triangle) (Treisman, 1998; Von Der Malsburg, 1999; Gray, 1999). For example, when we look at a cat, we see not only that it is a cat, but also the particular spatial features of that feline, from edges to the possible presence of a scarred eye, or the absence of an ear. Similarly, when we look at a letter T, we can see the vertical and horizontal bars that comprise the letter as distinct elements, as well as the fact that these constituent elements are part of the letter T itself. Such *hierarchical binding* captures the causal relations between multiple scales of abstraction, e.g. that an eye is part of a cat and not a dog nearby. These features carry meaningful spatial information due to their small receptive fields and high-dimension, and thereby can encode class-preserving transformations of the object - note in particular in our example that it is not simply the concept of an eye that is bound to the animal, but the specific representation localised in space, along with its own hierarchical components such as edges. We emphasize that we do not discount the importance of invariant features; rather the goal is to capture a continuum of abstraction jointly. Previous work has been done in encoding binding-like representations (Reichert & Serre, 2014; Greff et al., 2016; Schlag et al., 2019; Burgess et al., 2019;

Locatello et al., 2020; Bear et al., 2020; Whittington et al., 2020), albeit using different mechanisms to those discussed here, and without an investigation of its relevance to adversarial robustness.

Eguchi et al. (2018) proposed a mechanism by which the brain might capture hierarchical binding, encoding the relations between low-level and high-level features throughout the visual processing stream. Such binding representations would be consistent with experimentally observed neurons such as border-ownership cells, which have a small classical receptive field, but whose response is modulated by what object they form an edge of (Zhou et al., 2000). While Eguchi et al. (2018) predicted that both the temporal coincidence detection afforded by the spike-timing of biological neurons, as well as the lateral and top-down connectivity observed in the brain would be essential for implementing this binding mechanism, we use a non-local algorithm to capture such representations at the computational level (Marr, 1982), enabling us to explore the significance of these for robust object classification. We discuss the relationship of our work to the neuroscience and psychology literature in greater detail in Appendix A.1, including its relevance to the apparent sensitivity of humans to adversarial examples under appropriate conditions.

Preserving Low-Level Information Many methods exist to preserve low-level information in deep neural networks (Srivastava et al., 2015; Ronneberger et al., 2015; He et al., 2016; Huang et al., 2017; Jacobsen et al., 2018), spatially enrich feature representations (Sabour et al., 2017; Hinton et al., 2018) or encourage the encoding of additional factors of variation (Cheung et al., 2015). Our architecture is novel in that it captures which low-level neurons causally drove high-level representations, and explicitly encodes such information as layers in their own right for classification. This explicit encoding is important for down-stream read-out of the representations – architectures using skip connections, for example, combine information from low and high level layers in an operation that can obscure their respective contributions and makes classifier read-out of the low-level details more challenging. Our approach is related to the motivation for (and biological evidence of) disentangled/untangled representations (DiCarlo & Cox, 2007; Gáspár et al., 2019; Higgins et al., 2020) - our architecture is biased so as to "disentangle as many factors as possible, discarding as little information about the data as is practical" (Bengio et al., 2013).

Adversarial Examples While there is a large literature on adversarial examples (see e.g. Yuan et al. (2019) for a review), we focus on those papers that are most relevant to the current work. The concept of adversarial examples as manifold failures has inspired several defenses (Jalal et al., 2017; Samangouei et al., 2018; Song et al., 2018; Schott et al., 2019; Jang et al., 2020). Stutz et al. (2019) showed that typical adversarial examples move orthogonal to the manifold, and Khoury & Hadfield-Menell (2018) provided evidence in synthetic data-sets that a greater number of directions normal to the manifold (which can be quantified by the co-dimension) is associated with increasing vulnerability. This appears to be because the decision boundary can be arbitrary off of the manifold (Khoury & Hadfield-Menell, 2018), and may indeed lie very close to it (Tanay & Griffin, 2016). Finally, previous work on extracting hierarchical interpretations for the predictions of neural networks has shown that these interpretations themselves can be resistant to adversarial attacks (Singh et al., 2019), although this does not address the issue of robust classification.

3 Model Description

Implementing Hierarchical Binding To capture which low-level features causally drove max-pooled representations, we use the operation known as unpooling, which projects the max-pooled values into the equivalent positions they occupied in the previous layer, and sets the activations of all other neurons to zero (Zeiler & Fergus, 2014; Badrinarayanan et al., 2017) (Figure 2a). A modified version of unpooling, termed 'ratio unpooling', has previously been employed in a mixed bottom-up and top-down network as a means of preserving spatial information (Xu et al., 2019). Importantly however, this modified form of unpooling does not capture what lower-level features causally drove higher-level features (i.e. hierarchical binding), consistent with this not being the motivation.

To capture which simple features contributed to abstract representations, we introduce what we term 'gradient unpooling' (Figure 2a). For this operation, the gradient of the activations of the max-pooled neurons is taken with respect to each neuron in a lower-level layer. Specifically let $a_i^{(1)}$ be the activations of the lower-level layer $L^{(1)}$, and $a_j^{(2)}$ the activations of the higher-level layer $L^{(2)}$. The

gradients are a tensor of the same dimension as $L^{(1)}$, where each unit's value is the sum $\sum_j \frac{\partial a_j^{(2)}}{\partial a_i^{(1)}}$ over the activations $a_j^{(2)}$ in $L^{(2)}$. As the gradient is taken over the entire max-pooling layer, a proportion γ of the largest gradients are then selected with the intent of capturing the most important driving neurons. Here γ is a hyperparameter between 0 and 1. The winning gradients are used to generate a Boolean mask applied to the activations of the low-level layer, and this representation is up-projected. This captures the neurons that contributed to the distributed representation in the max-pooled layer, although it is only an approximation of the actual causal relations between them. Better measures of the importance of a low-level neuron to a higher-level representation exist (e.g. Dhamdhere et al. (2019)), but we use this method due to its computational efficiency. In Figure 2b, we show how these operations relate to hierarchical feature binding, as described in Eguchi et al. (2018) and Isbister et al. (2018). In our networks, we concatenate the results of unpooling and gradient-unpooling along-side the max-pooled activations in the feed-forward stream. This serves to provide both invariant and spatially detailed representations jointly.

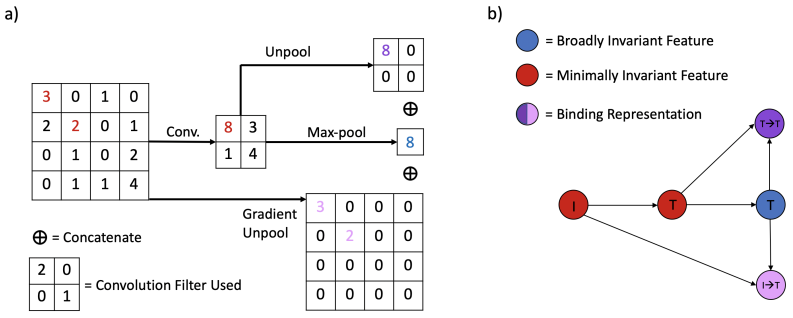

Figure 2: *Implementing hierarchical binding in a convolutional neural network.* a) 'Conv.' is a convolution operation with stride of 2. Our depicted representation of gradient unpooling is simplified for the sake of intuition, as the max-pooling layer in the figure consists of only a single neuron; in reality, we take the gradient of each low-level activation w.r.t. the entire max-pooled layer and only use the proportion γ of the largest gradients to apply a Boolean mask to the activations. b) A toy diagram to demonstrate the connection to hierarchical binding. The desire is to capture which low-level features (such as a vertical bar or a minimally invariant 'T' neuron) causally drove the more invariant representation of a 'T'. These hierarchical binding representations are then made available, alongside the invariant representations, to higher layers.

Model Architectures Both the unpooling and gradient unpooling computations can be introduced into standard CNN architectures, potentially at multiple levels. This work covers its use for both the MNIST (LeCun et al., 1998), Fashion-MNIST (FMNIST) (Xiao et al., 2017), and CIFAR-10 (Krizhevsky, 2009) data-sets. The models used for MNIST and FMNIST are based on the LeNet-5 architecture (Lecun et al., 1998). In our Hierarchical Binding CNN (HB-CNN), the LeNet-5 architecture is augmented with one unpooling and one gradient unpooling layer (see Appendix A.8). The models used for CIFAR-10 are based on a VGG-like architecture (Simonyan & Zisserman, 2015). For each architectural variant (including the non-binding control models), hyperparameter tuning for adversarial robustness was performed on a hold-out data-set (10k examples). To provide an unbiased measure of the effect on robustness, 30 randomly generated networks for each hyperparameter variant were then trained on the full training data-set and evaluated on the test data-set, with the median performance reported in all following results. All our code will be made available at publication.

To regularize the networks we use label-smoothing, a method that punishes over-confident predictions by replacing the typical one-hot label vector with a 'one-warm' vector (Szegedy et al., 2016; Pereyra et al., 2019). Specifically, the target probability of the correct label is assigned as $1 - \delta$, while the probability mass δ is uniformly distributed among the other classes. On synthetic data-sets such as two 'crescent moons', introducing label smoothing results in visibly smoother decision boundaries (Goibert & Dohmatob, 2019), and its use on computer vision tasks has been found to improve adversarial robustness (Warde-Farley & Goodfellow, 2016; Summers & Dinneen, 2019) (although used alone, it can also enhance vulnerability to certain attacks (Shafahi et al., 2019)).

4 ADVERSARIAL ATTACKS

Inspired by the thorough evaluation in Schott et al. (2019), we evaluate our model against a broad range of black-box and white-box methods, covering L_0, L_2 and L_∞ norm measured attacks (where the norm is used to quantify the distance between the original and the perturbed image). All attacks were evaluated using FoolBox v2.4 (Rauber et al., 2017), with hyperparameters specified in Appendix A.9. As in Schott et al. (2019), our main result is the median distance of adversaries, as this is less affected by outliers, and unlike when reporting accuracy, is not vulnerable to over-fitting on an arbitrary threshold; for completeness we also report bounded accuracy. All results presented are based on adversaries generated from a subset of 512 images from the test data-sets.

Gradient-Based Attacks Intuitively, a basic gradient-based attack can use knowledge of the model to perform gradient *ascent* of the loss with respect to the input *pixels*; contrast this with training where one performs gradient descent of the loss w.r.t. the weights. Given this privileged access to the model, these are known as white-box attacks. Various methods have been developed within this class that can be used to minimize both the L_2 and L_∞ norm; as in the Schott et al. (2019) evaluation protocol, we use the Fast Gradient Method (FGM), Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015), L_2 and L_∞ Basic Iterative Method (BIM) (Kurakin et al., 2019), L_2 and L_∞ DeepFool (Moosavi-Dezfooli et al., 2016), and Momentum Iterative Method (MIM) (Dong et al., 2018). For MNIST and FMNIST, these attacks are also repeated using the method for numerically estimated gradients used in Schott et al. (2019). Finally, for MNIST we also include Projected Gradient Descent (PGD, closely related to BIM), using multiple random starts.

Decision-Based Attacks These rely only on the decision output of the network, and are therefore a form of black-box attack. A particularly powerful method is the Boundary Attack (Brendel et al., 2018); intuitively, an image is first perturbed by noise until it is misclassified, after which the Boundary Attack iteratively moves the adversary closer to the original image while ensuring it remains misclassified. By taking sufficiently small steps, it can treat the decision boundary as approximately linear and move along it. As in Schott et al. (2019), we include the Boundary Attack, their L_0 and L_2 Pointwise attack, the Salt&Pepper noise attack and the Gaussian noise attack.

Transfer Attacks Transfer attacks leverage the empirical observation that adversaries generated for one network can often transfer to other networks, even those with different architectures (Papernot et al., 2016). Together with decision-based attacks, greater resistance to these methods provides evidence that a model's robustness is not simply a result of gradient masking (Papernot et al., 2017; Athalye et al., 2018), a common but less interesting method of resisting white-box methods. Our main addition to the assessment protocol in Schott et al. (2019) is that we derive transfer attacks from both a standard substitute network, and one with our proposed architecture. This is to ensure that transfer attack robustness isn't simply a result of the HB-CNN having an exotic architecture, while still being vulnerable to transfer attacks derived from other HB-CNNs. Specifically, adversaries are generated using the FGSM, L_∞ BIM, FGM, and L_2 BIM attacks against both a standard substitute network and one augmented with binding (see Appendix A.8). When attacking a network, all of these adversarial candidates are leveraged. Using the same line-search from Schott et al. (2019), images are iteratively perturbed from the baseline image until they are misclassified, and the minimally perturbed, successful transfer image is used in all distance and accuracy measures.

5 EXPERIMENTS

Hierarchical Binding Alone and the Decision Boundary Our opening assumption is that the input is high-dimensional, but that the true manifold representing object transformations is also reasonably high-dimensional, and that capturing this would improve robustness. In Appendix A.2, we demonstrate our main premise with a synthetic data-set that builds on the results of Khoury & Hadfield-Menell (2018) to model this assumption; in particular, representing additional dimensions of variation can enhance model robustness, even if the features do not perfectly separate the classes. For an image-based data-set, the HB-CNN architecture affords the possibility of learning decision boundaries along these additional dimensions of variation. Importantly however, decision boundaries in high dimension are challenging to learn due to the sampling complexity (i.e. the chance of sampling along a particular dimension becomes vanishingly small). What then is the effect of introducing binding representations? Rather than attempting to visualise high-dimensional decision

Table 1: Boundary Attack and Non-Linear Decision Boundaries

	LeNet	HB-CNN	HB-CNN + S	LeNet + Noisy Logits
$\boldsymbol{L_2}$**-metric** ($\epsilon = 1.5$)	1.6 (57%)	7.2 (98%)	5.3 (96%)	3.5 (95%)

Shown is the median L_2 distance of an adversary (median across 30 networks) for each condition, followed by the median bounded accuracy (at the given ϵ) across 30 networks. S = label smoothing.

boundaries directly, we can use the model's performance on the Boundary Attack as a proxy. Given the sparse sampling along the hierarchical binding dimensions, we would expect the augmented model to have a more non-linear (and not particularly useful) decision boundary (Figure 1b). Consistent with this, there is no general improvement to robustness, and in fact some attacks are more successful (Table 4 in Appendix A.4); clearly the decision boundaries are not very useful and can in fact be worse than those of a standard model. However, as a result of both the new feature dimensions and the actual operations introduced by gradient-unpooling and unpooling, we observe a considerable increase in resistance to the Boundary Attack (Table 1). In particular, as the Boundary Attack attempts to navigate the decision boundary, it frequently becomes stuck in local minima and encounters regions where it must decrease its step-size for the boundary to behave linearly. We exclude the possibility of stochastic elements explaining this result with a control included in Table 1 and described in Appendix A.3.

Hierarchical Binding with Regularization To improve the decision boundary, we regularize with label-smoothing, with the hope that sparse sampling in the high-dimensional space can still inform useful decision boundaries (Figure 1c). We also apply label-smoothing to the control network so as to ensure a fair comparison to a robust model. As expected, we observe an improvement on a range of attacks (Table 2), although the model also appears more vulnerable to others; without noisy training data (our next step), the regularized decision boundary can sit close to clean examples and be worse than a standard boundary. Introducing label smoothing, one would also predict that the decision boundary would behave more linearly, and consistent with this, it reduces the resistance of the HB-CNN to the Boundary Attack (Table 1).

Hierarchical Binding with Regularization and Noisy Training Data Given the spectrum of robustness following label-smoothing, a major component of the effect could be explained by gradient masking, and the proposed defense is limited. To better sample the new high-dimensional space, we introduce Gaussian (STD=0.3) and salt-and-pepper (120/784) noise during training (Figure 1d). As shown in Table 2, this creates a network with enhanced robustness to virtually all attacks, and the wide range of black and white-box attacks for which robustness is improved suggests this result cannot be explained as gradient masking alone.

We note of course that the result that both the control and HB-CNN models are resistant to some gradient-based L_∞ attacks above a perturbation of 0.5 indicates that some gradient masking is clearly occurring. We emphasize that the interesting result is the broad range of enhanced robustness seen, in particular to black-box attacks, and across multiple L_p norms. To control for the possibility that the greater number of parameters in the HB-CNN simply enables it to fit the noisy training data better, we also train a larger CNN with an equivalent number of parameters to the HB-CNN (results in Appendix A.4). Its performance is not comparable, and in many cases simply adding parameters appears to make the model more vulnerable. Finally, we compare our model to adversarial training (Madry et al., 2018). We highlight that our model beats adversarial training on all black-box attacks, the All L_2 metric, and the All L_0 metric, but at a smaller cost in clean classification accuracy.

We do not include the robust Analysis by Synthesis (ABS) model (Schott et al., 2019) due to the computational resources required to run it (around three orders of magnitude more time for a forward pass on our GPUs in comparison to the HB-CNN). Based on the results in Table 1 of Schott et al. (2019), our network outperforms their non-binary ABS model on 13 of 17 attacks (not considering the latent descent attack), and beats their state-of-the-art (SOTA) All L_2 result (median distance 2.3 for ABS vs median distance 2.7 for ours). Our results are thus in keeping with SOTA robustness on MNIST, however we are hesitant to make a strong claim due to the difficulties of a fair, head-to-head comparison - beyond controlling for the number of samples and transfer images used, it is unclear for example what effect Gaussian/salt-and-pepper noise in training or non-linear boundary behaviour might have on the performance of the ABS model. Indeed our basic, non-binding model with noisy training data and label smoothing already performs comparable to ABS and adversarial

Table 2: MNIST Results

	CNN+S	HBCNN+S	CNN+S+N	HBCNN+S+N	CNN+AT
Clean accuracy	99.17%	99.08%	99.10%	99.05%	98.40%
L_2-metric ($\epsilon = 1.5$)					
Transfer	4.4 (92%)	4.6 (94%)	5.8 (97%)	**6.6** (98%)	5.0 (97%)
Gaussian Noise	6.8 (98%)	5.5 (98%)	10.0 (99%)	**10.5** (99%)	5.3 (97%)
Boundary	1.5 (50%)	5.3 (96%)	2.5 (88%)	**9.2** (98%)	1.4 (42%)
Pointwise	3.4 (95%)	2.7 (93%)	4.4 (96%)	**4.5** (97%)	1.9 (73%)
FGM	8.9 (91%)	9.1 (92%)	8.9 (95%)	9.6 (96%)	9.0 (98%)
FGM w/GE	8.8 (93%)	8.6 (94%)	8.8 (95%)	9.3 (96%)	∞ (97%)
DeepFool	4.5 (82%)	5.5 (84%)	7.3 (91%)	**8.1** (93%)	9.4 (94%)
DeepFool w/GE	6.5 (88%)	4.8 (86%)	7.3 (93%)	**7.7** (94%)	9.5 (94%)
BIM	2.4 (67%)	3.2 (73%)	3.6 (84%)	**4.0** (86%)	4.9 (93%)
BIM w/GE	2.4 (67%)	3.1 (77%)	3.6 (84%)	**4.0** (87%)	4.5 (93%)
PGD	1.3 (41%)	1.6 (54%)	2.5 (77%)	**2.8** (79%)	2.8 (86%)
All L_2	1.1 (29%)	1.6 (53%)	2.1 (75%)	**2.7** (79%)	1.4 (39%)
L_∞-metric ($\epsilon = 0.3$)					
Transfer	0.33 (58%)	0.33 (55%)	0.41 (76%)	**0.52** (88%)	0.44 (96%)
FGSM	0.46 (70%)	0.46 (73%)	0.48 (76%)	**0.62** (82%)	0.44 (95%)
FGSM w/GE	0.46 (72%)	0.46 (74%)	0.50 (78%)	**0.63** (83%)	∞ (95%)
DeepFool	0.37 (60%)	0.43 (67%)	0.83 (81%)	**1.0** (85%)	0.46 (94%)
DeepFool w/GE	0.59 (75%)	0.44 (65%)	**1.0** (86%)	**1.0** (86%)	0.71 (94%)
BIM	0.19 (37%)	0.29 (48%)	0.34 (56%)	**0.44** (66%)	0.36 (93%)
BIM w/GE	0.18 (35%)	0.29 (48%)	0.34 (57%)	**0.45** (67%)	0.63 (93%)
MIM	0.21 (37%)	0.30 (50%)	0.32 (54%)	**0.42** (66%)	0.34 (93%)
MIM w/GE	0.21 (37%)	0.30 (50%)	0.35 (56%)	**0.44** (68%)	0.44 (94%)
PGD	0.09 (6%)	0.12 (7%)	0.22 (28%)	**0.24** (33%)	0.33 (91%)
All L_∞	0.09 (5%)	0.12 (6%)	0.21 (24%)	**0.23** (29%)	0.33 (91%)
L_0-metric ($\epsilon = 12$)					
Pointwise	13 (50%)	9 (25%)	22 (75%)	**23** (79%)	5 (5%)
Salt&Pepper Noise	48 (92%)	17 (67%)	142 (97%)	135 (97%)	14 (57%)
All L_0	13 (50%)	9 (25%)	22 (75%)	**23** (79%)	5 (5%)

Shown is the median L_p distance of a successful adversary for the different attacks (rows), provided as the median performance across 30 networks for each model condition (columns). In parentheses is the median accuracy (at the given ϵ) across 30 networks. The All-L_0, All-L_2 and All-L_∞ distances show the minimal adversarial distance across all attacks for each image. Bold indicates the best performance between the CNN+S+N, Size-Controlled CNN+S+N (Appendix A.4) and HBCNN+S+N; blue indicates the best performance across all networks. AT=adversarial training; S=label smoothing; N=Gaussian and Salt-and-pepper noise during training; GE=Gradient Estimation.

training on several metrics, a marked improvement that has been separately reported when these methods are combined (Shafahi et al., 2019). It is unsurprising that our control model would benefit from the boundary improvements introduced by label smoothing and noisy training data - the key difference however is that without hierarchical binding, it lacks the expressive power to realise the same magnitude of improvement.

To confirm that the observed effects generalize to a more complex setting, we apply the same architectures to FMNIST, albeit with $\gamma = 0.3$ rather than 0.4 (Appendix A.5), and implement a VGG style HB-CNN for the CIFAR-10 data-set with $\gamma = 0.1$ (Table 3, model details in appendix). While there are a few attacks where we fail to generalize the effect of the HB-CNN being stronger than the robust control (FMNIST: Gaussian noise, Pointwise L_2, DeepFool attacks, All L_∞; CIFAR-10: Pointwise L_2, MIM, All L_0), the trend of enhanced robustness above the control model and across multiple attacks is observed. Interestingly, the HB-CNN surpasses the clean accuracy of the control model on FMNIST. This is hardly a fair comparison, as the HB-CNN has more parameters, but it supports the proposal that when binding representations are included, adversarial robustness becomes a question of on-manifold generalization error, making robustness compatible with clean classification accu-

Table 3: CIFAR-10 Results

	Vanilla-VGG	VGG+S+N	HB-VGG+S+N	ResNet+AT
Clean	86.51%	**86.43%**	86.07%	87.25%
$\boldsymbol{L_2}$**-metric** ($\epsilon = 1.5$)				
Transfer	0.68 (26%)	0.99 (36%)	**1.03** (38%)	6.23 (80%)
Gaussian Noise	2.31 (69%)	3.89 (80%)	**4.12** (79%)	7.50 (82%)
Boundary	0.27 (3%)	0.42 (5%)	**4.81** (84%)	1.11 (36%)
Pointwise	1.60 (53%)	**1.91** (60%)	1.88 (60%)	2.11 (62%)
FGM	0.37 (19%)	0.58 (28%)	**0.67** (32%)	2.01 (59%)
DeepFool	0.17 (0%)	0.44 (14%)	**0.49** (14%)	0.97 (35%)
BIM	0.14 (1%)	0.22 (1%)	**0.23** (2%)	0.66 (21%)
All $\boldsymbol{L_2}$	0.14 (0%)	0.22 (0%)	**0.23** (1%)	0.66 (20%)
$\boldsymbol{L_\infty}$**-metric** ($\epsilon = 8/255$)				
Transfer	0.19 (34%)	0.030 (48%)	**0.031** (50%)	0.153 (82%)
FGSM	0.07 (17%)	0.010 (26%)	**0.012** (30%)	0.037 (54%)
DeepFool	0.005 (1%)	0.013 (28%)	**0.015** (28%)	0.039 (56%)
BIM	0.004 (1%)	0.006 (4%)	**0.007** (5%)	0.028 (46%)
MIM	0.004 (1%)	**0.007** (4%)	**0.007** (6%)	0.029 (47%)
All $\boldsymbol{L_\infty}$	0.004 (0%)	0.006 (3%)	**0.007** (5%)	0.028 (46%)
$\boldsymbol{L_0}$**-metric** ($\epsilon = 12$)				
Pointwise	7 (35%)	9 (42%)	**10** (43%)	11 (45%)
Salt&Pepper Noise	15 (53%)	**21** (57%)	**21** (57%)	27 (57%)
All $\boldsymbol{L_0}$	7 (34%)	**9** (42%)	**9** (43%)	11 (45%)

racy (Stutz et al., 2019). Finally, we note that on CIFAR-10, while the observed improvement is broad-spectrum, it does not match the highly established method of adversarial training.

We conclude by noting that it is possible that an approach such as even more random starts to PGD or an advanced attack method would find better adversaries than the attacks we've leveraged - our proposal does not guarantee the absence of vulnerable spaces in the decision boundary, it rather aims to reduce the number of directions in which these are found. Consistent with this, our results show that the model is more robust across a variety of methods and distance measures, in particular to transfer attacks, i.e. when using directions that are successful against other models, supporting the notion of fewer vulnerable dimensions (Tramèr et al., 2017).

The Importance of Binding Dimension and Causality We have claimed that a) the dimension of the binding representations contributes to robustness by augmenting the low-dimensional representations already present in a standard network, and that b) these representations are meaningful because they capture how low-level features causally drive higher-level features. An alternative possibility, however, is that the novel architecture described here simply makes the gradients of the network more challenging to use. For example, gradient unpooling has similarities to k-Winner-Take-All applied directly to activations, which is associated with enhanced robustness to gradient-based attacks (Xiao et al., 2020). In Figure 3, we show a benefit even without masking, the importance of the dimension of the binding representations, and the significance of their causal role in the network activity, thereby arguing against this possibility. Specifically, as the proportion of largest gradients used is increased, the network shows a rapid rise in its robustness, presumably as the key causally important dimensions are captured. Increasing it further appears to eventually cause the model robustness to decrease slightly. As γ approaches 1.0, we lose the sparse representation of the low-level features that were causally involved (by our approximate measure) in the high-level representation – the classifier should be influenced by dimensions that describe the object, not noise contributed by other low-level feature detectors that may be co-active (Guo et al., 2018; Ahmad & Scheinkman, 2019), although alternative explanations for the modest drop in robustness are possible, such as removing discontinuities in the network (Xiao et al., 2020). Using a higher γ-proportion also reduces the non-linearity of the model's decision boundary, making it more vulnerable to the Boundary Attack. Importantly however, the model is still more robust than the baseline. Using the same algorithm but with the γ-smallest gradients to derive the binding information confers no robustness to the BIM attacks until virtually all activations are up-projected (i.e. finally capturing the bottom portion of the important features), suggesting that the operation itself is not the source of robustness.

In Appendix A.6, we also provide ablation analyses showing the importance of all three representational layers (max-pooling, unpooling, and gradient unpooling) for robustness, supporting that it is the combined high-dimensional representation that is beneficial. If the majority of the effect observed was a result of gradient masking, one would predict that inactivating the max-pooling representations would actually enhance robustness. Finally, our earlier observation that there is no enhanced resistance to gradient-based attacks prior to regularization and noise further supports that our main effect cannot be explained by previous defences such as Xiao et al. (2020).

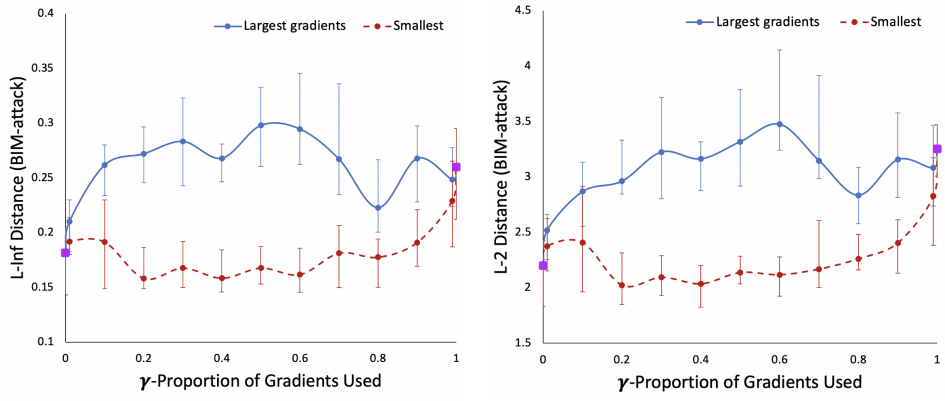

Figure 3: *The effect of hierarchical binding dimension and causal role on robustness.* We systematically vary the γ-proportion of gradients used to mask the gradient unpooling representations along the x-axis; results are shown from using both the largest (blue-solid) and smallest (red-dashed) γ gradients. 0.0 is equivalent to the LeNet control model we study, while 1.0 means no masking is applied and all low-level activations are up-projected (indicated in purple). Each point represents the median distance of a successful adversary, provided as the median performance across 30 networks trained on MNIST without unpooling. Error bars show the 95% confidence interval of the median.

6 Discussion

We have demonstrated the implementation of a novel CNN architecture, inspired by recent work in theoretical neuroscience (Eguchi et al., 2018; Isbister et al., 2018). This architecture seeks to capture hierarchical binding representations that encode the causal relations between lower-level and higher-level visual features of an object. Within the framework of adversarial examples as off-manifold perturbations, we provided empirical evidence of enhanced robustness following the introduction of these representations. Through principled hyper-parameter scaling, a synthetic data-set, and ablation analyses, we demonstrated that the introduction of these representations is fundamental to the observed adversarial robustness. A key strength of the proposed approach is the broad range of attack types against which adversarial robustness is enhanced, including L_0, L_∞ and L_2 norms, all within both black box and white box settings. In of itself, the method does not cause a considerable drop in clean classification accuracy, and these benefits come at relative computational efficiency. Finally, in Appendix A.1, we detailed how hierarchical binding representations could help explain the apparent sensitivity of humans to adversarial examples under specific experimental conditions.

Nevertheless, we must highlight some notable limitations of this work. The proposed architecture adds a significant number of additional parameters, in particular for larger models, as well as a hyperparameter (the γ-proportion) that for optimal performance requires tuning. In general, a more complex data-set appears to benefit from a smaller proportion (MNIST=0.4, FMNIST=0.3, CIFAR-10=0.1). In the setting of CIFAR-10, the effect size vs. the control was more subtle, it was more sensitive to the hyper-parameters used, and the robustness was not comparable to adversarial training. It is possible that unsupervised pre-training (Hénaff et al., 2019; Chen et al., 2020) to make use of unlabelled data would help with some of the core challenges of the proposed architecture, such as the sampling complexity. Finally, the adversarial examples that fool the HB-CNN do not appear, on-average, to be more meaningful than for the robust control model (Appendix A.10). Notwithstanding these limitations, our analysis indicates that vulnerability to adversarial attacks is at least partly due to the absence of the hierarchical representations described herein.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: A system for large-scale machine learning. In *Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation, OSDI 2016*, 2016. ISBN 9781931971331.

Subutai Ahmad and Luiz Scheinkman. How Can We Be So Dense? The Robustness of Highly Sparse Representations. *ICML 2019 Workshop on Uncertainty and Robustness in Deep Learning*, 2019.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *35th International Conference on Machine Learning*. MIT, 2018.

Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(12), 2017. ISSN 01628828. doi: 10.1109/TPAMI.2016.2644615.

Daniel M. Bear, Chaofei Fan, Damian Mrowca, Yunzhu Li, Seth Alter, Aran Nayebi, Jeremy Schwartz, Li Fei-Fei, Jiajun Wu, Joshua B. Tenenbaum, and Daniel L. K. Yamins. Learning Physical Graph Representations from Visual Scenes. *arXiv preprint arXiv:2006.12373*, 2020. URL http://arxiv.org/abs/2006.12373.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2013. ISSN 01628828. doi: 10.1109/TPAMI.2013.50.

Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Tom Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MONet: Unsupervised Scene Decomposition and Representation. *arXiv preprint arXiv:1901.11390*, 1 2019. URL http://arxiv.org/abs/1901.11390.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A Simple Framework for Contrastive Learning of Visual Representations. *Proceedings of the 37th International Conference on Machine Learning*, 2 2020. URL http://arxiv.org/abs/2002.05709.

Brian Cheung, Jesse A. Livezey, Arjun K. Bansal, and Bruno A. Olshausen. Discovering hidden factors of variation in deep networks. In *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*, 2015.

Francesco Croce, Maksym Andriushchenko, and Matthias Hein. Provable robustness of relu networks via maximization of linear regions. In *AISTATS 2019 - 22nd International Conference on Artificial Intelligence and Statistics*, 2020.

Joel Dapello, Tiago Marques, Martin Schrimpf, Franziska Geiger, David Cox, and James DiCarlo. Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Kedar Dhamdhere, Qiqi Yan, and Mukund Sundararajan. How important is a neuron? In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

James J. DiCarlo and David D. Cox. Untangling invariant object recognition. *Trends in Cognitive Sciences*, 2007. ISSN 13646613. doi: 10.1016/j.tics.2007.06.010.

Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting Adversarial Attacks with Momentum. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018. ISBN 9781538664209. doi: 10.1109/CVPR.2018.00957.

Marin Dujmović, Gaurav Malhotra, and Jeffrey S. Bowers. What do adversarial images tell us about human vision? *eLife*, 9, 2020. ISSN 2050084X. doi: 10.7554/ELIFE.55978.

Akihiro Eguchi, James B. Isbister, Nasir Ahmad, and Simon Stringer. The emergence of polychronization and feature binding in a spiking neural network model of the primate ventral visual system. *Psychological Review*, 2018. ISSN 0033295X. doi: 10.1037/rev0000103.

Gamaleldin F. Elsayed, Nicolas Papernot, Shreya Shankar, Alexey Kurakin, Brian Cheung, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. In *Advances in Neural Information Processing Systems*, volume 2018-December, 2018.

Merse E. Gáspár, Pierre Olivier Polack, Peyman Golshani, M. Lengyel, and Gergő Orbán. Representational untangling by the firing rate nonlinearity in V1 simple cells. *eLife*, 8, 2019. ISSN 2050084X. doi: 10.7554/eLife.43625.

Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, Ian Goodfellow, and Google Brain. The Relationship Between High-Dimensional Geometry and Adversarial Examples. *arXiv:1801.00634*, 2018.

Morgane Goibert and Elvis Dohmatob. Adversarial Robustness via Label-Smoothing. *arXiv preprint arXiv:1906.11567*, 6 2019. URL http://arxiv.org/abs/1906.11567.

Ian Goodfellow, Jonathon Schlens, and Christian Szegedy. Explaining and Harnessing Adversarial Examples. In *International Conference on Learning Representations*, 2015.

C M Gray. The temporal correlation hypothsis of visual feature integration: Still alive and well. *Neuron*, 24:31–47, 1999.

Klaus Greff, Rupesh Kumar Srivastava, and Jürgen Schmidhuber. Binding via Reconstruction Clustering. *4th International Conference on Learning Representations, ICLR 2016*, 11 2016. URL http://arxiv.org/abs/1511.06418.

Y Guo, C Zhang, C Zhang, and Y Chen. Sparse DNNs with Improved Adversarial Robustness. In *32nd Conference on Neural Information Processing Systems*, Montréal, 2018.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, pp. 770–778. IEEE Computer Society, 12 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.90.

Martin N. Hebart, Charles Y. Zheng, Francisco Pereira, and Chris I. Baker. Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, 4(11), 2020. ISSN 23973374. doi: 10.1038/s41562-020-00951-3.

Olivier J. Hénaff, Aravind Srinivas, Jeffrey De Fauw, Ali Razavi, Carl Doersch, S. M. Ali Eslami, and Aaron van den Oord. Data-Efficient Image Recognition with Contrastive Predictive Coding. *arXiv preprint arXiv:1905.09272*, 5 2019. URL http://arxiv.org/abs/1905.09272.

Irina Higgins, Le Chang, Victoria Langston, Demis Hassabis, Christopher Summerfield, Doris Tsao, and Matthew Botvinick. Unsupervised deep learning identifies semantic disentanglement in single inferotemporal neurons. *arXiv preprint arXiv:2006.14304*, 2020.

Geoffrey Hinton, Sara Sabour, and Nicholas Frosst. Matrix capsules with EM routing. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Shaul Hochstein and Merav Ahissar. View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 2002. ISSN 08966273. doi: 10.1016/S0896-6273(02)01091-7.

Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017. ISBN 9781538604571. doi: 10.1109/CVPR.2017.243.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander M Adry Mit. Adversarial Examples are not Bugs, they are Features. In *Neural Information Processing Systems (NIPS)*, 2019.

James B Isbister, Akihiro Eguchi, Nasir Ahmad, Juan Galeazzi, Mark Buckley, and Simon Stringer. A new approach to solving the feature binding problem in primate vision. *Interface Focus*, 8:–, 2018.

Jörn Henrik Jacobsen, Arnold Smeulders, and Edouard Oyallon. I-RevNet: Deep invertible networks. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G. Dimakis. The Robust Manifold Defense: Adversarial Training using Generative Models. *arXiv preprint arXiv:1712.09196*, 12 2017. URL http://arxiv.org/abs/1712.09196.

Uyeong Jang, Somesh Jah, and Susmit Jah. On the Need for Topology-Aware Generative Models for Manifold-Based Defenses. *International Conference on Learning Representations*, 2020.

Marc Khoury and Dylan Hadfield-Menell. On the Geometry of Adversarial Examples. *arXiv preprint arXiv:1811.00525*, 11 2018. URL http://arxiv.org/abs/1811.00525.

Sung Ho Kim and Jacob Feldman. Globally inconsistent figure/ground relations induced by a negative part. *Journal of Vision*, 9(10), 2009. ISSN 15347362. doi: 10.1167/9.10.8.

Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Alex Krizhevsky. Learning Multiple Layers of Features from Tiny Images. *Science Department, University of Toronto, Tech.*, 2009. ISSN 1098-6596. doi: 10.1.1.222.9220.

Jonas Kubilius, Martin Schrimpf, Kohitij Kar, Rishi Rajalingham, Ha Hong, Najib J. Majaj, Elias B. Issa, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Aran Nayebi, Daniel Bear, Daniel L.K. Yamins, and James J. DiCarlo. Brain-like object recognition with high-performing shallow recurrent ANNs. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2019.

Yann Lecun, Leon Bottou, Yoshua Bengio, and Patrick Ha. LeNet. In *Proceedings of the IEEE*, number November, pp. 1–46, 1998. ISBN 0018-9219. doi: 10.1109/5.726791.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2323, 1998. ISSN 00189219. doi: 10.1109/5.726791.

Sidney R. Lehky, Roozbeh Kiani, Hossein Esteky, and Keiji Tanaka. Dimensionality of object representations in monkey inferotemporal cortex. *Neural Computation*, 26(10), 2014. ISSN 1530888X. doi: 10.1162/NECO{_}a{_}00648.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. *arXiv preprint arXiv:2006.15055*, 6 2020. URL http://arxiv.org/abs/2006.15055.

Yiliang Lu, Jiapeng Yin, Zheyuan Chen, Hongliang Gong, Ye Liu, Liling Qian, Xiaohong Li, Rui Liu, Ian Max Andolina, and Wei Wang. Revealing Detail along the Visual Hierarchy: Neural Clustering Preserves Acuity from V1 to V4. *Neuron*, 98(2), 2018. ISSN 10974199. doi: 10.1016/j.neuron.2018.03.009.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

David Marr. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. *MIT Press*, 1982. URL http://books.google.com/books?id=EehUQwAACAAJ&printsec=frontcover%5Cnpapers2://publication/uuid/FBD15E5F-E503-450B-B059-4C15D54099CE.

Seyed Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2016. ISBN 9781467388504. doi: 10.1109/CVPR.2016.282.

Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve Restricted Boltzmann machines. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 2010. ISBN 9781605589077.

Andrew Y. Ng. Feature selection, L1 vs. L2 regularization, and rotational invariance. In *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, 2004. ISBN 1581138385. doi: 10.1145/1015330.1015435.

Nicolas Papernot, Patrick D McDaniel, and Ian J Goodfellow. Transferability in Machine Learning: from Phenomena to Black-Box Attacks using Adversarial Samples. *arXiv preprint arXiv:1605.07277*, 2016.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security - ASIA CCS '17*, 2017. ISBN 9781450349444. doi: 10.1145/3052973.3053009.

Gabriel Pereyra, George Tucker, Jan Chorowski, Łukasz Kaiser, and Geoffrey Hinton. Regularizing neural networks by penalizing confident output distributions. In *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2019.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*, 2017. URL http://arxiv.org/abs/1707.04131.

David P. Reichert and Thomas Serre. Neuronal synchrony in complex-valued deep networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2015. ISBN 9783319245737. doi: 10.1007/978-3-319-24574-4{_}28.

Sara Sabour, Nicholas Frosst, and Geoffrey E. Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, 2017.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-Gan: Protecting classifiers against adversarial attacks using generative models. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. Enhancing the Transformer with Explicit Relational Encoding for Math Problem Solving. *arXiv preprint arXiv:1910.06611*, 10 2019. URL http://arxiv.org/abs/1910.06611.

Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on MnIsT. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Ali Shafahi, Amin Ghiasi, Furong Huang, and Tom Goldstein. Label Smoothing and Logit Squeezing: A Replacement for Adversarial Training? *arXiv preprint arXiv:1910.11585*, 10 2019. URL http://arxiv.org/abs/1910.11585.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

Chandan Singh, Bin Yu, and W. James Murdoch. Hierarchical interpretations for neural network predictions. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

Yang Song, Sebastian Nowozin, Nate Kushman, Taesup Kim, and Stefano Ermon. PixelDefend: Leveraging generative models to understand and defend against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

Rupesh Kumar Srivastava, Klaus Greff, and Jürgen Schmidhuber. Training very deep networks. In *Advances in Neural Information Processing Systems*, 2015.

David Stutz, Matthias Hein, and Bernt Schiele. Disentangling adversarial robustness and generalization. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2019. ISBN 9781728132938. doi: 10.1109/CVPR.2019.00714.

Cecilia Summers and Michael J. Dinneen. Improved Adversarial Robustness via Logit Regularization Methods. *arXiv preprint arXiv:1906.03749*, 6 2019. URL http://arxiv.org/abs/1906.03749.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, 2016. doi: 10.1109/CVPR.2016.308.

Timothy Tadros, Ramyaa Ramyaa, Giri P Krishnan, and Maxim Bazhenov. Biologically inspired sleep algorithm for increased generalization and adversarial robustnes in deep neural networks. *ICLR*, 2020.

Thomas Tanay and Lewis Griffin. A Boundary Tilting Persepective on the Phenomenon of Adversarial Examples. *arXiv preprint arXiv:1608.07690*, 8 2016. URL http://arxiv.org/abs/1608.07690.

Florian Tramèr, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. The Space of Transferable Adversarial Examples. *arXiv preprint arXiv:1704.03453*, 2017.

Anne Treisman. Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 1998. ISSN 09628436. doi: 10.1098/rstb.1998.0284.

C. Von Der Malsburg. The what and why of binding: The modeler's perspective. *Neuron*, 24(1): 95–104, 1999. ISSN 08966273. doi: 10.1016/S0896-6273(00)80825-9.

Manish Vuyyuru, Andrzej Banburski, Nishka Pant, and Tomaso Poggio. Biologically Inspired Mechanisms for Adversarial Robustness. *34th Conference on Neural Information Processing Systems (NeurIPS 2020)*, 2020.

Thomas Sa Wallis, Christina M. Funke, Alexander S. Ecker, Leon A. Gatys, Felix A. Wichmann, and Matthias Bethge. Image content is more important than Bouma's law for scene metamers. *eLife*, 8, 2019. ISSN 2050084X. doi: 10.7554/eLife.42512.

David Warde-Farley and Ian Goodfellow. 11 adversarial perturbations of deep neural networks. In *Perturbations, Optimization, and Statistics*, pp. 311. 2016.

James C.R. Whittington, Timothy H. Muller, Shirley Mark, Guifen Chen, Caswell Barry, Neil Burgess, and Timothy E.J. Behrens. The Tolman-Eichenbaum Machine: Unifying Space and Relational Memory through Generalization in the Hippocampal Formation. *Cell*, 183, 11 2020. ISSN 00928674. doi: 10.1016/j.cell.2020.10.024.

Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing Adversarial Defense by k-Winners-Take-All. *International Conference on Learning Representations 2020*, 2020.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. *arXiv preprint arXiv:1708.07747*, 8 2017. URL `http://arxiv.org/abs/1708.07747`.

Chunyan Xu, Jian Yang, Hanjiang Lai, Junbin Gao, Linlin Shen, and Shuicheng Yan. UP-CNN: Un-pooling augmented convolutional neural network. *Pattern Recognition Letters*, 2019. ISSN 01678655. doi: 10.1016/j.patrec.2017.08.007.

Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. Adversarial Examples: Attacks and Defenses for Deep Learning. *IEEE transactions on neural networks and learning systems*, 2019. ISSN 21622388. doi: 10.1109/TNNLS.2018.2886017.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference 2016, BMVC 2016*, volume 2016-September, 2016. doi: 10.5244/C.30.87.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2014. ISBN 9783319105895. doi: 10.1007/978-3-319-10590-1-53.

H Zhou, H S Friedman, and Rüdiger von der Heydt. Coding of border ownership in monkey visual cortex. *The Journal of Neuroscience*, 20(17):6594–6611, 2000. ISSN 0270-6474. doi: 10.1523/JNEUROSCI.2797-12.2013.

Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature Communications*, 10(1), 2019. ISSN 20411723. doi: 10.1038/s41467-019-08931-6.

A Appendix

A.1 Relevance to Neuroscience and Adversarial Examples in Humans

We have argued that the primate brain's high-dimensional representations of objects may be important for robustness to adversarial examples, yet the notion of object representations as high-dimensional may seem counter-intuitive to some traditional concepts of vision. It is clear that many aspects of primate vision rely primarily on abstract, low-dimensional representations (Hebart et al., 2020), which is consistent with our proposal here, and that vision can be impoverished e.g. outside of attention. What we have described however is the primate perception of "objects along with their detailed features" (Lu et al., 2018). For example, even in the periphery, humans are sensitive to low-level image changes when these impact scene-like content as opposed to textures (Wallis et al., 2019). Primates have access to such low/mid-level information in higher processing, such as local border ownership (Kim & Feldman, 2009). Thus the representations we've described are consistent with "vision with scrutiny", rather than coarse object recognition (Hochstein & Ahissar, 2002).

Previous work measuring low-dimensional object representations in primates, such as Lehky et al. (2014), is compatible with our proposal here, as they measured neural activity in anterior inferotemporal cortex. Based on current evidence, it is likely that the neural binding representations would be predominantly distributed in lower-levels of the visual cortex (Zhou et al., 2000; Lu et al., 2018).

Our arguments re. the dimensionality of primate object representations due to hierarchical binding can then be connected to the perception of adversarial examples. Humans without time-constraints do not appear to be sensitive to adversarial perturbations (see e.g. recent limitations identified by Dujmović et al. (2020) regarding Zhou & Firestone (2019)), and as such the only experimental evidence for human sensitivity to adversarial examples is from Elsayed et al. (2018). While the effect on classification was not comparable to the dramatic shift seen in machine vision systems, they measured significant drops in accuracy when humans were constrained to view adversarial images for a very short duration (around 60-70ms) followed by masking intended to limit recurrent and top-down processing. However, the introduction of recurrent activity in a CNN designed to better match the primate ventral stream (CORNet-S) (Kubilius et al., 2019) is not associated with enhanced robustness – rather the base CORNet-S architecture appears to be more vulnerable than a standard CNN such as AlexNet to adversarial examples (Dapello et al., 2020). Given that recurrence alone in a CNN is not sufficient to enable robustness, this raises the question of what specific computation was actually hindered experimentally in Elsayed et al. (2018) to explain their results. As noted in our introduction, it has been predicted that top-down and lateral activity in a spiking neural network would be needed to implement the proposed hierarchical binding algorithm in a biological system, and so the disruption of such processing provides an alternative hypothesis for the observed effect.

There have been some valuable contributions to explain human robustness to adversarial examples, but we feel these still leave important questions unanswered. Vuyyuru et al. (2020) examined the effect of non-uniform retinal sampling and varying receptive field sizes with eccentricity, which could certainly contribute to human robustness, but their effect was specific to small perturbations, and the evaluation of black-box attacks did not assess whether transfer attack robustness was greater for the proposed model vs. an undefended one. Tadros et al. (2020) examined the effect of sleep-like algorithms on robustness, and again while this might account for some of the difference between humans and artificial systems, their method actually increased the vulnerability of the model on MNIST to the Boundary Attack (see their Table 1), with no evaluation of other black-box methods such as transfer attacks or the Pointwise attack. Finally, in Dapello et al. (2020), various aspects of early primate visual processing were explored as defence methods, but the primary benefit was accounted for by V1 stochasticity, and again there was no analysis of transfer attacks. In summary, we believe our work helps address an explanatory gap re. human robustness to adversarial examples, in particular to transfer attacks – the one method that has been leveraged against humans, and a threat setting where our method performs consistently.

A.2 Partially Informative Features and Co-Dimension

Our opening assumption was that a higher-dimensional manifold would be helpful for robustness, even if features do not perfectly separate the classes. To model this in a synthetic setting, we created a data-set with 32 dimensions, of which the first two features are Gaussians with virtually no overlap between the two classes, analogous to the invariant, low-dimensional representations in a CNN (Figure 4a; means 0 and 1, STD=0.15). For the other 30 dimensions, these either take the form of a feature with no information about the class (like those used in Khoury & Hadfield-Menell (2018), and which therefore increase the co-dimension), or features that carry some, albeit imperfect information about the class (represented with partially overlapping Gaussians; means 0 and 0.5, STD=0.15). These latter features are roughly analogous to the binding representations. We model the effect of introducing hierarchical binding by replacing uninformative features with partially informative ones, while maintaining a constant total input dimension. For each variant of the input, we train multiple multi-layer perceptrons with two 100-unit hidden layers and label smoothing (0.25) on the synthetic data-set, and evaluate their robustness to the L_2 BIM attack. Our results show that as the dimension of the class manifold vs. the input increases, so too does the robustness of the model (Figure 4b). This builds on Khoury & Hadfield-Menell (2018) by replicating their result with the introduction of features that are imperfect for class discrimination.

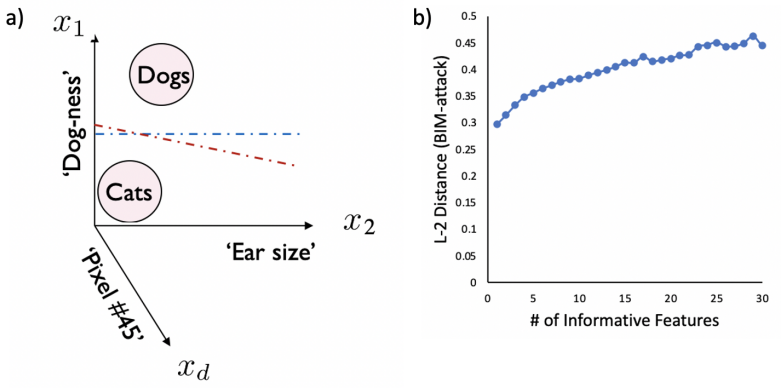

Figure 4: *The effect of additional features on robustness in a synthetic data-set.* a) In our synthetic data-set of Gaussians, two dimensions separate the data well (like the 'Dog-ness' dimension), while others carry either no information (like 'Pixel-45'), or some information about class identity (like 'Ear-size'). The proposal is that explicitly including features such as 'Ear-size' can help inform a better decision boundary in otherwise vulnerable directions. b) By keeping the total input dimension fixed, but having a greater number of partially informative features (varied along the x-axis), we see an improved robustness to L_2 BIM attacks as the co-dimension decreases (shown is the median L_2 distance of an adversary, given as the median performance across 100 networks for each condition).

A.3 Boundary Attack Resistance and Stochasticity

As has been pointed out elsewhere (Brown et al., 2018), an uninteresting means of resistance to the Boundary Attack would be the presence of noise in a model's predictions. No stochastic elements were introduced into the HB-CNN model, and indeed we analysed the logits from multiple runs of the network to confirm no numerical instability or other forms of noise were inadvertently present. As a further control, we trained a normal LeNet-5 model where Gaussian noise of mean 0 and standard-deviation 0.01 is added to the logits (orders of magnitude more noise than would fail our own checks for numerical instability), but it does not show comparable robustness (Table 1).

A.4 Additional MNIST Results

We provide our results from the size-controlled CNN with label smoothing and noisy training data in Table 4, along with 'vanilla' LeNet and HB-CNN models (i.e. without label smoothing or noise).

A.5 Fashion-MNIST and CIFAR-10

The protocol for FMNIST is largely the same as that used for MNIST (see Appendix A.8 for details), including the addition of Gaussian (STD=0.3) and salt-and-pepper (120/784) noise. Results are shown in Table 5.

In order to preserve clean classification accuracy on CIFAR-10, our noise-augmented training regime uses only Gaussian noise an order of magnitude smaller (STD=0.03 vs 0.3). With the exception of the 'vanilla' model, we also use weight decay (Ng, 2004) in the VGG models to further regularize the decision boundaries (see hyper-parameter details in Appendix A.8). We note that the optimal weight-decay value was tuned for robustness for both the control and the binding-augmented models separately.

A.6 Ablation Analyses

To elucidate the relative influence of the two forms of binding, we performed ablation analyses. Trained HB-CNNs were taken and the activations of either the unpooling, gradient unpooling, or both layer's activations set to 0 on all forward passes. No further training was applied, and the networks were analysed for both robustness and classification accuracy on the clean MNIST test set

Table 4: Additional MNIST Results

	Vanilla-CNN	Vanilla-HBCNN	Size-Controlled CNN+S+N
Clean	99.25	99.13	**99.43%**
L_2**-metric** ($\epsilon = 1.5$)			
Transfer	2.8 (88%)	2.9 (90%)	5.5 (98%)
Gaussian Noise	7.1 (98%)	5.9 (98%)	9.7 (99%)
Boundary	1.6 (57%)	7.2 (98%)	2.1 (84%)
Pointwise	3.4 (96%)	2.9 (93%)	4.2 (97%)
FGM	3.7 (90%)	3.3 (88%)	**10.0** (97%)
FGM w/GE	4.0 (92%)	5.2 (92%)	**9.8** (97%)
DeepFool	1.4 (46%)	1.4 (39%)	3.6 (89%)
DeepFool w/GE	1.7 (59%)	1.6 (54%)	4.6 (91%)
BIM	1.3 (37%)	1.3 (34%)	3.0 (85%)
BIM w/GE	1.2 (34%)	1.5 (51%)	3.1 (84%)
PGD	1.0 (13%)	1.0 (12%)	1.9 (71%)
All L_2	1.0 (12%)	1.0 (12%)	1.8 (69%)
L_∞**-metric** ($\epsilon = 0.3$)			
Transfer	0.23 (21%)	0.23 (18%)	0.39 (65%)
FGSM	0.18 (15%)	0.18 (20%)	0.50 (76%)
FGSM w/GE	0.23 (38%)	0.40 (54%)	0.51 (76%)
DeepFool	0.12 (0%)	0.12 (0%)	0.32 (54%)
DeepFool w/GE	0.14 (0%)	0.13 (1%)	0.45 (73%)
BIM	0.10 (0%)	0.10 (0%)	0.25 (31%)
BIM w/GE	0.10 (9%)	0.11 (29%)	0.25 (31%)
MIM	0.10 (0%)	0.10 (0%)	0.25 (33%)
MIM w/GE	0.10 (9%)	0.13 (27%)	0.26 (38%)
PGD	0.08 (0%)	0.08 (0%)	0.15 (0%)
All L_∞	0.08 (0%)	0.08 (0%)	0.15 (0%)
L_∞**-metric** ($\epsilon = 12$)			
Pointwise	13 (55%)	10 (36%)	21 (76%)
Salt&Pepper Noise	56 (93%)	22 (73%)	**156** (98%)
All L_0	13 (55%)	10 (36%)	21 (76%)

Bold indicates the best performance between CNN+S+N, Size-Controlled CNN+S+N and HBCNN+S+N; blue indicates the best performance across all networks.

(Figure A.6). Both binding layers and the max-pooling layer are important to the robustness; only with this combined representation is the network more resistant than the LeNet model, and ablating any one layer causes a considerable drop in performance.

A.7 Performance Distributions

To visualise the distribution of performance across the 30 networks we trained for each configuration, we include histograms for several black-box attacks, as well as the all-attacks measure for each distance metric (Figure 6).

A.8 Model Details

All models were implemented in TensorFlow 1.14 (Abadi et al., 2016), and used the ReLU activation function (Nair & Hinton, 2010). The LeNet-5 model used as the vanilla and label smoothing control for MNIST consists of two convolution (6 and 16 channels) and two max-pooling layers, followed by two fully connected layers of size 120 and 84. When we introduced noise in training, we used two fully connected layers of dimension 256 and 128. For the HB-CNN, the unpooling layer is applied to the last max-pooling layer, while the gradient-unpooling layer is between the last max-pooling layer and the pre-convolution activations proceeding it (Figure 7a). For the size-controlled CNN for MNIST (877,440 parameters vs the 850,070 in the HB-CNN - see Table 6), the two convolutions had

Table 5: Fashion-MNIST Results

	Vanilla-CNN	CNN+S+N	HBCNN+S+N	CNN+AT
Clean	90.87%	87.38%	**88.26%**	88.46%
L_2-metric ($\epsilon = 1.5$)				
Transfer	1.7 (55%)	3.7 (78%)	**4.2** (81%)	3.1 (80%)
Gaussian Noise	2.9 (76%)	**8.4** (86%)	8.1 (86%)	3.9 (87%)
Boundary	0.4 (7%)	1.6 (52%)	**4.5** (84%)	0.9 (23%)
Pointwise	2.3 (70%)	**4.0** (83%)	**4.0** (82%)	1.7 (59%)
FGM	1.0 (37%)	3.0 (67%)	**3.4** (69%)	4.7 (84%)
FGM w/GE	1.1 (41%)	3.0 (68%)	**3.6** (71%)	5.4 (83%)
DeepFool	0.4 (12%)	**1.6** (52%)	**1.6** (52%)	2.3 (66%)
DeepFool w/GE	0.5 (1%)	**2.1** (58%)	1.9 (56%)	2.3 (65%)
BIM	0.3 (1%)	1.0 (36%)	**1.1** (39%)	1.8 (58%)
BIM w/GE	0.3 (8%)	1.0 (35%)	**1.3** (46%)	1.8 (57%)
All L_2	0.3 (0%)	1.0 (32%)	**1.1** (38%)	0.9 (21%)
L_∞-metric ($\epsilon = 0.1$)				
Transfer	0.09 (45%)	0.22 (77%)	**0.23** (78%)	0.18 (87%)
FGSM	0.04 (24%)	0.13 (57%)	**0.15** (60%)	0.24 (80%)
FGSM w/GE	0.05 (30%)	0.13 (58)	**0.16** (62%)	0.32 (80%)
DeepFool	0.03 (2%)	**0.13** (58%)	0.12 (56%)	0.26 (79%)
DeepFool w/GE	0.03 (2%)	**0.16** (62%)	0.13 (57%)	0.27 (79%)
BIM	0.02 (1%)	0.07 (37%)	**0.08** (41%)	0.15 (76%)
BIM w/GE	0.02 (9%)	0.07 (36%)	**0.08** (42%)	0.16 (76%)
MIM	0.02 (1%)	0.07 (38%)	**0.08** (41%)	0.15 (76%)
MIM w/GE	0.02 (8%)	0.08 (38%)	**0.09** (46%)	0.16 (76%)
All L_∞	0.02 (0%)	**0.07** (35%)	**0.07** (39%)	0.15 (76%)
L_0-metric ($\epsilon = 12$)				
Pointwise	8 (32%)	24 (69%)	**26** (68%)	4 (17%)
Salt&Pepper Noise	23 (63%)	**167** (85%)	133 (84%)	16 (54%)
All L_0	8 (32%)	24 (69%)	**26** (68%)	4 (17%)

Bold indicates the best performance between CNN+S+N and HBCNN+S+N; blue indicates the best performance across all networks. S = label smoothing; N = Gaussian noise during training; AT = adversarial training.

Table 6: Number of Trainable Parameters in HB-CNN+S+N vs Size-Controlled CNN+S+N

	HB-CNN		Size-Controlled CNN	
	# of Parameters	# input/output	# of Parameters	# input/output
1st Conv.	156	1/6 c	832	1/32 c
2nd Conv.	2416	6/16 c	51264	32/64 c
1st FC	813312	400+1600+1176/256 u	819712	1600/512 u
2nd FC	32896	256/128 u	5632	512/10 u
3rd FC	1290	128/10 u	0	n/a
Total	**850,070**		**877,440**	

Abbreviations: Conv. = convolution; FC = fully connected; c = channels; u = units

channel sizes 32 and 64, and the fully connected component consisted of a single layer of dimension 512. For FMNIST, we used the same architectures with the larger fully connected layers.

The adversarially trained models for MNIST and CIFAR-10 were loaded from the MadryLab Challenge repositories (`https://github.com/MadryLab/mnist_challenge`) and (`https://github.com/MadryLab/cifar10_challenge`), the latter of which is based on a Wide ResNet (Zagoruyko & Komodakis, 2016). These were generated by the original authors of Madry et al. (2018), and use the adversarial training with PGD described therein. The adversarially trained model for FMNIST was loaded from the repository from Croce et al. (2020)) (`https://github.`

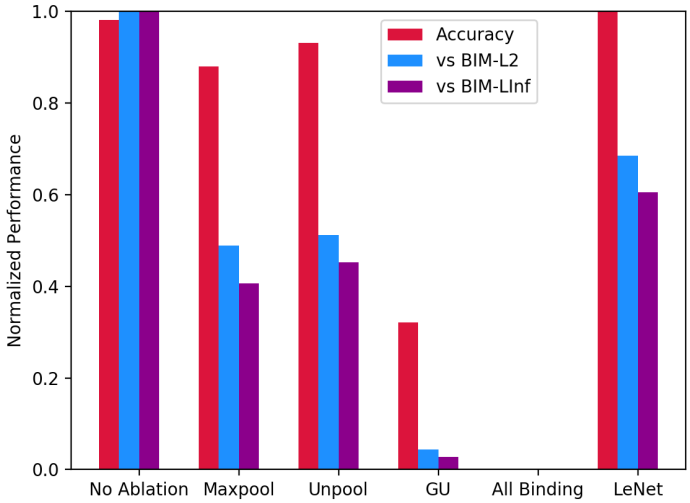

Figure 5: *The effect of layer-wise ablations.* We normalize the values of the three different metrics so they can be compared side by side (median clean accuracy and median distances of adversaries for BIM-L_2 and BIM-L_∞ attacks across 30 networks). For any given metric, the best performance corresponds to 1 and the worst to 0. For comparison to other results in this paper, the performance of the un-ablated HB-CNN and LeNet model is as shown in Table 2 (no noise in training data). The worst performing ablation (all binding) corresponds to a clean accuracy of 94.43%, BIM L_∞ of 0.05, and BIM-L_2 of 0.6. GU = gradient unpooling.

`com/max-andr/provable-robustness-max-linear-regions`); these were trained using PGD attacks (40 iterations), with 50% adversarial images, and 50% clean images in each batch, for 100 epochs.

For the VGG architectures (Figure 7b), we used three blocks, each containing two convolutions and one max-pooling (channel dimensions 32, 32, 64, 64, 128, 128), followed by two fully connected layers of dimension 120 and 84. For the HB-CNN VGG variant we used two un-pooling layers (corresponding to the 2nd and 3rd blocks), and two gradient unpooling layers (from the 3rd max-pooling layer to the 1st and 2nd max-pooling layers respectively).

All of our models used dropout of 0.25, label smoothing of 0.1, and batch-size of 128. We used the Adam optimizer (Kingma & Ba, 2015) with a learning rate of 0.001 (MNIST and FMNIST) and 0.0005 (CIFAR-10). Training was performed for 45 epochs (vanilla MNIST model, and LeNet-based models with smoothing), 90 epochs (all other LeNet-based models), and 500 epochs (all VGG models). For VGG models with smoothing (used as surrogates), we used an L_2 regularization of 10^{-5} on the weights from the final max-pooled layer (standard and binding-augmented), and 10^{-4} for the binding-representations. For VGG models with smoothing and noise in the training data, we used L_2 regularization of 10^{-3} (standard) or 10^{-5} (HB-CNN) for the weights from the final max-pooled layer, and 10^{-3} for the binding-representations. For VGG models, we also augmented the training data with random shifts and horizontal flipping. Surrogates for all three data-set transfer attacks were based on the respective standard and HB-CNN architectures with label smoothing, except for the vanilla CNN and HB-CNN models on MNIST, where vanilla surrogates were used.

A.9 Hyperparameters for Adversarial Attacks

For PGD we used 20 random starts (selecting the best outcome for each image), 250 iterations, and an initial step-size of 0.01. For BIM we used 10 iterations with an initial step-size of 0.05. MIM was applied with 10 iterations, an initial step-size of 0.06, and a decay factor of 1.0. For PGD, BIM, and MIM, the step-size and epsilon were automatically adapted in Foolbox using a binary search.

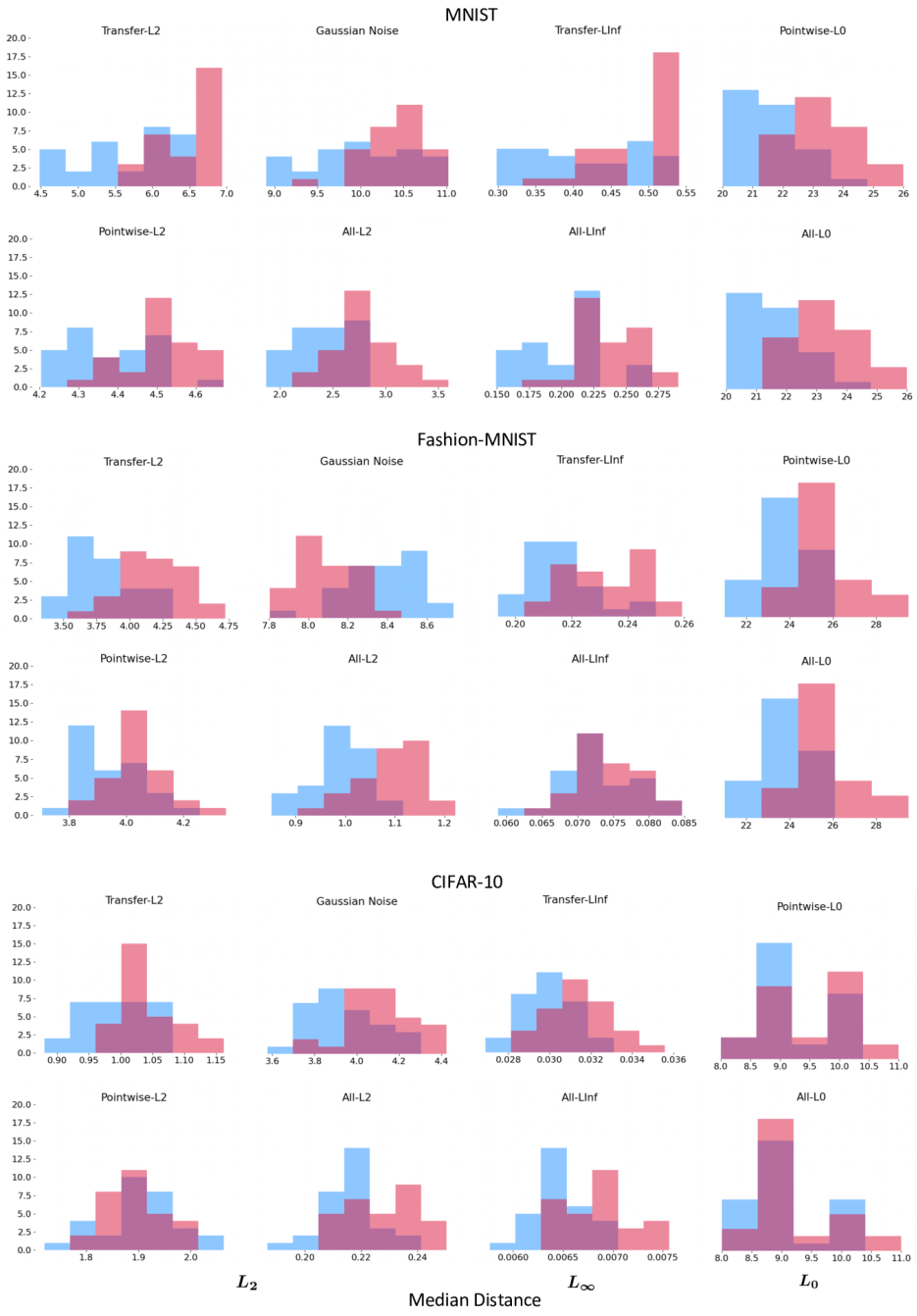

Figure 6: *Distributions of the Median Performance of Networks*. Red is the HB-CNN/VGG+S+N, blue is the standard CNN/VGG+S+N. Note that for some of these attacks we reported no improvement in the main text, such as the Pointwise L_2 attack metric on Fashion-MNIST.

For DeepFool we used 100 iterations. For the Boundary attack we used 1,000 iterations, a step-adaptation size of 1.5, and initial adversaries generated with the Blended Uniform Noise attack; we checked the performance against one of the HB-CNN+S+N networks using 25,000, 100,000, and 1,000,000 iterations (tuning the step adaptation size for each), but this did not significantly improve its performance, and it remained uncompetitive against the HB-CNN model in comparison to other attacks.

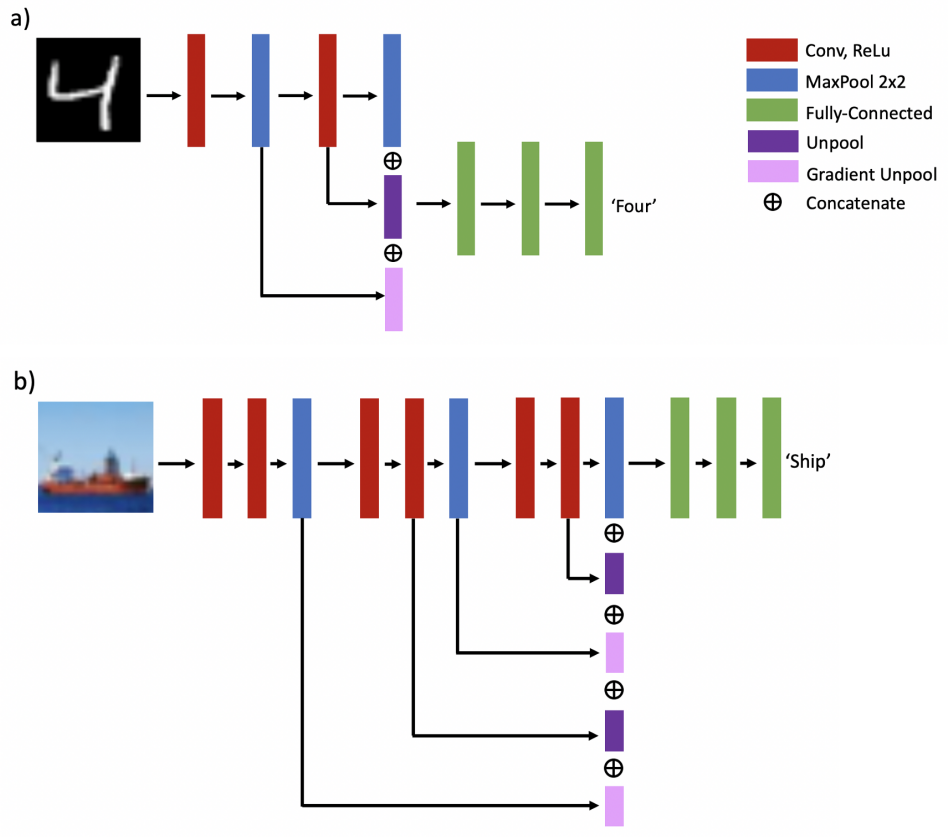

Figure 7: *Integration of Unpooling and Gradient Unpooling in a Hierarchical Binding CNN.* The diagram shows the architecture for a) MNIST and Fashion-MNIST and b) CIFAR-10. Note that it is not necessary to use unpooling or gradient-unpooling representations from every layer, thus deeper architectures can select intermittent representations to use for unpooling or gradient-unpooling, avoiding an excessive growth in parameters.

A.10 Cherry Picked Examples

From among 300 MNIST adversarial examples for the 30 models of the LeNet and HB-CNN networks with label smoothing and noisy training data, we selected what we thought to be (in our potentially biased view) the 10 most semantically convincing adversaries generated by the BIM L_2 attack for each model (Figure 8). We feel there are convincing examples for both models, with no obvious systematic difference.

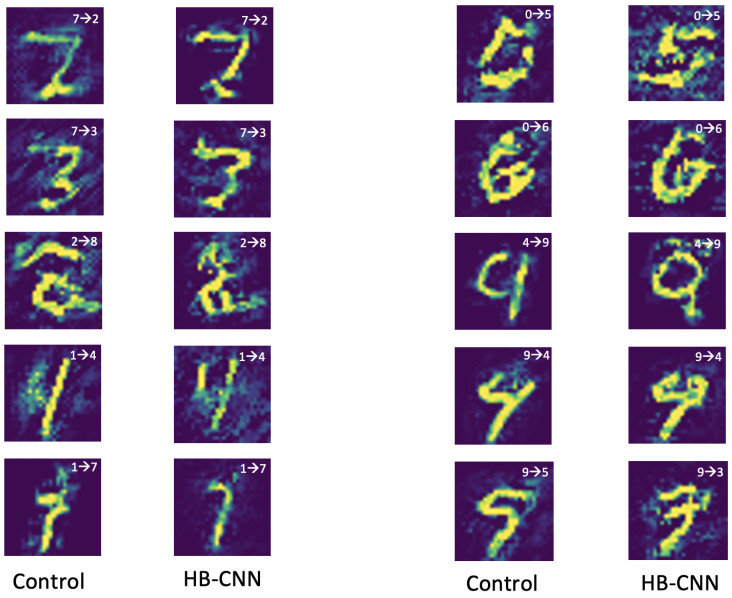

Figure 8: *Cherry-picked, semantically meaningful adversaries for the control and HB-CNN models.* The annotation shows the original class followed by the prediction of the network following the adversarial perturbation.