Denoised Internal Models: a Brain-Inspired Autoencoder against Adversarial Attacks

Anonymous authors
Paper under double-blind review

Abstract

Despite its great success, deep learning severely suffers from robustness; that is, deep neural networks are very vulnerable to adversarial attacks, even the simplest ones. Inspired by recent advances in brain science, we propose the Denoised Internal Models (DIM), a novel generative autoencoder-based model to tackle this challenge. Simulating the pipeline in the human brain for visual signal processing, DIM adopts a two-stage approach. In the first stage, DIM uses a denoiser to reduce the noise and the dimensions of inputs, reflecting the information pre-processing in the thalamus. Inspired from the sparse coding of memory-related traces in the primary visual cortex, the second stage produces a set of internal models, one for each category. We evaluate DIM over 42 adversarial attacks, showing that DIM effectively defenses against all the attacks and outperforms the SOTA on the overall robustness.

1 Introduction

The great advances in deep learning (DL) techniques bring us a large number of sophisticated models that approach human-level performance in a broad spectrum of tasks, such as image classification (LeCun et al., 1989; He et al., 2016; Krizhevsky et al., 2012; Szegedy et al., 2016), speech recognition (Amodei et al., 2016; Xiong et al., 2016), and natural language processing (Vaswani et al., 2017; Devlin et al., 2019; Yang et al., 2019b; Gu et al., 2018). Despite its success, deep neural network (DNN) models are vulnerable to adversarial attacks (Szegedy et al., 2014; Biggio et al., 2013; Goodfellow et al., 2014a). Even with adding human-unrecognizable perturbations, the predictions of the underlying network model could be completely altered (Biggio & Roli, 2018; Goodfellow et al., 2014b; Moosavi-Dezfooli et al., 2016; Athalye & Carlini, 2018). On the other hand, the human brain, treated as an information processing system, enjoys remarkably high robustness (Xu & Vaziri-Pashkam, 2021; Athalye et al., 2018b). A question naturally arises whether knowledge about the working mechanism of the human brain can help us improve the adversarial robustness of DNN models (Casamassima et al., 2021; Huang et al., 2019).

Biological systems keep being an illuminating source of human engineering design. Two famous relevant examples are the perceptron model (Rosenblatt, 1958) and the Rectified Linear Unit (ReLU) (Agarap, 2019) activation function. Further, the recurrent neural network (RNN) (Elman, 1990) architecture also has its origin in the study of how to process time-series data, like natural language. In this work, we draw inspiration from the visual signal processing paradigm of the human brain and propose a novel model to address the robustness issue in the image classification task.

With the recent progress in neuroscience, we now better understand the information processing pipeline in the human brain's visual system. Two brain areas are involved in this pipeline: the thalamus and the primary visual cortex (Cudeiro & Sillito, 2006). Visual signals from the retina will travel to the Lateral Ganglion Nucleus (LGN) of the thalamus before reaching the primary visual cortex (Derrington et al., 1984). The LGN is specialized in handling visual information, helping to process different kinds of stimuli. In addition, some vertebrates, like zebrafish, have no visual cortex but still have some neural structure similar to the hypothalamus to receive and process visual signals (O'Connor et al., 2002). This fact highlights the importance of such an information pre-processing module in the biological visual signal processing system. The primary visual cortex is one of the best-studied brain areas, which displays a complex 6-layers structure and provides excellent pattern recognition capacities. An important finding (Xie et al., 2014) reveals that Layer 2/3 of
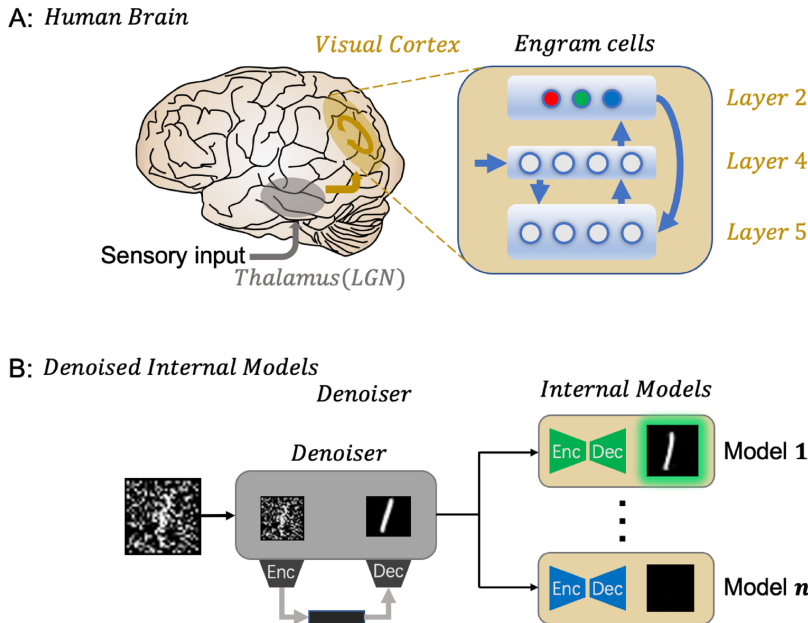
Figure 1: From the visual signal processing in human brain (A) to the Denoised Internal Models (B).

the primary visual cortex contains the so-called engram cells (Tonegawa et al., 2015), which only activate for specific stimuli and related ones, such as Jennifer Aniston's pictures (Quiroga et al., 2005). In other words, the concepts corresponding to those stimuli are encoded sparsely through the engram cells (McGaugh, 2000; Guan et al., 2016). Furthermore, artificial activation of the engram cells induces corresponding memory retrieval (Liu et al., 2012; 2014). Those discoveries suggest there could be internal generative models for different kinds of concepts in the human brain.

Simulating the pipeline mentioned above, we proposed the Denoised Internal Models (DIM) (see Figure 1 B), which consists of a global denoising network (a.k.a. denoiser) and a set of generative autoencoders, one for each category. The denoiser helps pre-process the input data, similar to what LGN does. The autoencoders can be regarded as internal models for specific concepts mimicking the function of engram cells in the primary visual cortex. In order to have a comprehensive evaluation of DIM's robustness, we conduct our experiments on MNIST (Lecun et al., 1998), using DIM against 42 attacks in the foolbox v3.2.1 package (Rauber et al., 2017) and comparing its performance to SOTA models. The results show that DIM outperforms the SOTA models on the overall robustness and has the most stable performance across the 42 attacks.

2 RELATED WORKS

2.1 ADVERSARIAL ATTACKS

A large number of adversarial attacks have been proposed recently (Tramèr et al., 2018; Rony et al., 2019; Rauber & Bethge, 2020; Hosseini et al., 2017; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016; Brendel et al., 2020). From the viewpoint of the attacker's knowledge about the models, these attacks can be divided into white-box ones and black-box ones. The former (Rony et al., 2019; Rauber & Bethge, 2020; Moosavi-Dezfooli et al., 2016; Brendel et al., 2020) knows all model information, including the network architecture, parameters, and learning mechanisms. The latter (Brendel et al., 2018) only knows limited or zero knowledge about the models, but it can interact with the model through inputs and outputs. Typically, white-box attacks are harder for defense.

Viewing from the norm types, i.e., the distance measure of the adversarial perturbations, major existing adversarial attacks fall into four categories: L_0 attacks (Schott et al., 2019), L_1 attacks (Hosseini et al., 2017; Brendel et al., 2020), L_2 attacks (Rony et al., 2019; Rauber & Bethge, 2020; Carlini
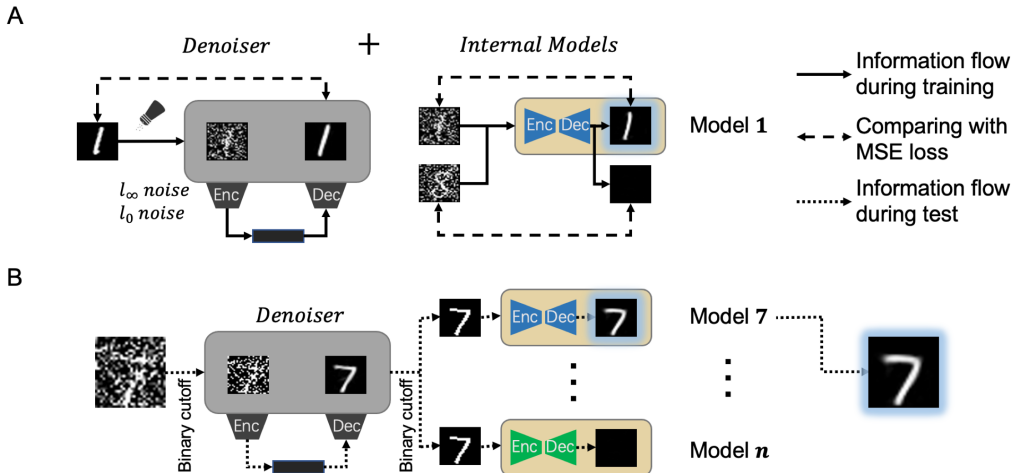
Figure 2: The training phase (A) and the inference phase (B) of DIM.

& Wagner, 2017; Moosavi-Dezfooli et al., 2016), and L_∞ attacks (Moosavi-Dezfooli et al., 2016; Brendel et al., 2020).

2.2 DEFENSE METHODS

Many defense approaches have been proposed to tackle the robustness challenge in deep learning (Yin et al., 2019; Pang et al., 2019b; Hu et al., 2019; Verma & Swami, 2019; Bafna et al., 2018; Pang et al., 2019a). Roughly, there are four major types.

2.2.1 ADVERSARIAL TRAINING

Proposed by Madry et al. (2018), it is one of the most popular defense methods that can withstand strong attacks. It follows the simple idea of training the model on the generated adversarial samples.

2.2.2 RANDOMIZATION

This approach (Vaishnavi et al., 2020; Vincent et al., 2010) randomizes the input layer or some intermediate layers to neutralize the adversarial perturbations and help to protect the underlying model.

2.2.3 GRADIENT MASKING

This approach mainly defends against gradient-based attacks by building a model with no useful gradient (Xiao et al., 2019); that is, the gradients of the model outputs with respect to its inputs are almost zero. However, it turns out such a method fails to work in practice (Athalye et al., 2018a).

2.2.4 GENERATIVE MODELS

This approach exploits a generative model, normally GAN (Samangouei et al., 2018) or autoencoder (Cintas et al., 2020; Meng & Chen, 2017), to project the high-dimensional inputs into a low-dimensional manifold. It is generally believed that such a way can reduce the risk of overfitting and improve the adversarial robustness (Jang et al., 2019).

Among this type, we would like to mention the ABS (Schott et al., 2019) model. From a different starting point, it also arrived at the design that uses an individual generative model for each category in the dataset.

3 MODEL

3.1 BIOLOGICAL INSPIRTION

In this subsection, we give a closer look at the visual signals processing pipeline in the human brain. As depicted in Figure 1 (A), visual perceptual information streams are firstly received and processed by the LGN in the thalamus before projecting to the primary visual cortex (O'Connor et al., 2002).

The cell bodies in the LGN arrange to form a 6-layer structure, where the inner two layers are called the magnocellular layers, while the outer four are called parvocellular layers (Brodal, 2004). Previous studies (White et al., 2009) reveal that parvocellular layers are sensitive to color and perceive a high level of detail. On the other hand, magnocellular layers are highly sensitive to motion while insensitive to color and detail. In this way, the LGN pre-processes different kinds of visual information.

The primary visual cortex has a complex hierarchical structure of six layers. Roughly speaking, Layer 4 handles the input signals from the LGN, and Layer 5 sends outputs to other regions in the brain. Upon receiving inputs, Layer 4 sent strong signals directly to Layer 2/3 for processing (Markram et al., 2015). Recent advances in neuroscience surprisingly find that contextual information and memory components are sparsely encoded in Layer 2, namely, only a distinct population of neurons in Layer 2 respond to a specific kind of context, and these populations are spatially separated (Xie et al., 2014). Such spatial sparsity reflects typical engram-cell behavior. As mentioned in the introduction, artificial activation of the engram cells induces memory retrieval, suggesting that internal models operate inside the primary visual cortex.

It is worth emphasizing that we only consider the functional level analogy to the brain's visual signal processing system rather than repeating its precise connections and structure. We recognize the LGN as a pre-processor that distills inputs signals, separating different kinds of information. On the other hand, the primary visual cortex corresponds to a set of internal generative models. In this way, we abstract the human visual signal processing system as a two-stage model and will lay our model design based on it.

3.2 DENOISED INTERNAL MODELS

Based on the above abstraction, we proposed a two-stage model Denoised Internal Models (DIM). The corresponding schematic diagram is shown in Figure 1 (B). In the first stage, we seek a global denoiser that helps filter the "true" signals out of the input images, which is analogous to the function of LGN in the thalamus. The basic idea is that adversarial perturbations are generally semantically meaningless and can be effectively treated as noise in the raw images. The second stage consists of a set of internal generative models, which operate in a dichotomous sense. Each internal model only accepts images from a distinct category and will reject images from other categories. Upon acceptance, the internal model will output a reconstructed image, while it returns a black image if the input is rejected. In this way, our model reflects the engram-cell behavior in the primary visual cortex.

One of the main targets of this paper is to evaluate whether a functional level analogy to the brain's visual signal processing system helps improve the adversarial robustness rather than focusing on specific algorithms. Hence, we have kept our network architecture simple to avoid complexities during evaluation. Details about the architecture and parameter settings can be found in the appendix.

3.2.1 DENOISER

There exist different methods to filter the raw image out of a noisy one (Yang et al., 2019a; Candès & Recht, 2009; Chatterjee, 2015; Chen & Chi, 2018). We adopt a simple autoencoder as the denoise network model. In the training phase, we add noise to the images in the training dataset. Those noisy images serve as the inputs to the denoiser and the original ones as the learning targets. The model is trained on the mean-square error (MSE) loss to minimize the mean reconstruction error.

$$\mathcal{L}_D = \mathbb{E}\left[\|x - D(x + \epsilon)\|_2^2\right], \tag{1}$$

where $D(\cdot)$ refers to the function of the denoiser, and ϵ indicates the added noise. $\|\cdot\|_p$ denotes the $L_p, p = 0, 1, 2, \infty$ norm.

3.2.2 INTERNAL MODELS

We train an autoencoder as the internal generative model for each category in the dataset. Ideally, the input, the bottleneck, and the output of the autoencoders are analogous to the roles of neurons in Layer 4, Layer 2/3, and Layer 5, respectively, in the primary visual cortex. The internal models receive inputs from the outputs of the denoiser. Then we add noise to the inputs to reflect the randomness in the brain's neural activities. The i-th autoencoder in the internal models is trained on the following loss

$$\mathcal{L}_{IM,i} = \mathbb{E}\left[\|F_i(x + \epsilon) - x * I_i(x)\|_2^2\right],\tag{2}$$

where indicator $I_i(x) = 0$ if x belongs to category i, and 0 otherwise. $F_i(\cdot)$ denotes the function that corresponds to the i-th autoencoder, and ϵ indicates the added noise. We choose this loss function to encourage the engram-cell behavior of the autoencoders. As a result of this behavior, it is natural to perform inference based on the relative output intensities from different autoencoders. Specifically, for each input image x, we estimate its relative intensity from the i-th autoencoder for as

$$\widetilde{P}(x|i) = \|F_i(D(x))\|_1 / \|D(x)\|_1.\tag{3}$$

The prediction on x by DIM will be

$$p(x) = \arg\max_{i \in [K]} \widetilde{P}(x|i),\tag{4}$$

where $[K] := \{0, 1, \dots, K-1\}$ and K is the number of categories.

Finally, we also consider a variation of the DIM model in the inference phase, which includes two binarization operations, one applied to the input images and the other applied to the outputs of the denoiser. We refer to this variation as biDIM hereafter.

4 EXPERIMENTS

To evaluate the adversarial robustness of our model, we compare DIM and biDIM against two SOTA methods: the adversarial training (Madry et al., 2018), a SOTA L_∞ defense method, and the analysis by synthesis (ABS) model as well as its variation with binarization inputs (biABS) (Schott et al., 2019). We also include a vanilla convolutional neural network (CNN) model and its variation with input binarization (biCNN) as the baseline models. The DIM models are implemented using relatively simple neural networks. Consequently, their clean accuracies are 96%, while the other models reach 99%. Despite this disadvantage, biDIM still beats the SOTA on the overall robustness. We will discuss more details in the next section.

In our experiments, we applied almost all attacks available in foolbox v3.2.1 against all models. Those attacks consist of 22 L_2 attacks, 12 L_∞ attacks, 6 L_1 attacks and 2 L_0 attacks [1]. The most effective ones are those based on model gradients and those based on the prediction boundary. More specifically, the gradient-based attacks include the DeepFool Attack (Moosavi-Dezfooli et al., 2016), Basic Iterative Method (BIM) Attack (Kurakin et al., 2016), and the Carlini&Wagner Attack (Carlini & Wagner, 2017). They exploit the gradients at the raw input images to find directions leading to wrong predictions. The boundary attacks rely on the model decision. Starting from adversarial samples with relative large perturbation size, these attacks search towards the corresponding raw input images along the boundary between the adversarial and non-adversarial regions. Within this type, there are white-box attacks L_2 Boundary Attack (Brendel et al., 2018) and L_0 BrendelBethge Attack as well as black-box attacks L_1, L_2, and L_∞ BrendelBethge Attack (Brendel et al., 2020). Our experiments also cover other types of attacks, such as the additive random noise attacks, including the Gaussian Noise Attack and the Uniform Noise Attack and their variations.

In practice, the overall robustness is more important than the robustness under a single attack, since the adversaries will not restrict themselves to any specific attack. To reflect the overall robustness, we summarize the experimental results within each L_p, $p = 0, 1, 2, \infty$ norm and leave the full results for individual attacks in the appendix. More concretely, we count a sample as successfully attacked as long as one attack finds the adversarial image. Furthermore, the corresponding perturbation size on the sample is computed by minimizing across all successful attacks.

---

[1] We use foolbox v2.4.0 for the Pointwise attack since it is not available in foolbox v3.2.1.

Table 1: Results for different kinds of models under defferent adversarial attacks, arranged according to distance metrics. Each entry shows the accuracy of the model for the threshold of $\epsilon_{L_0} = 12$, $\epsilon_{L_1} = 8$, $\epsilon_{L_2} = 1.5$, and $\epsilon_{L_\infty} = 0.3$. For each $L_p, p = 0, 1, 2, \infty$ norm, we also summarize all attacks in the type, calculating both the median adversarial distance (left value) between all samples and the overall accuracy (right value). The last row shows the minimal accuracy of each model across all the attacks. The best results for the overall performance are shown in bold. Due to the space limit, we only display 15 important attacks out of the total 42 in this table and leave the full results in the appendix.

	CNN	biCNN	Madry	biABS	ABS	biDIM	DIM
L_2-metric ($\epsilon = 1.5$)							
L_2 DDNAttack	15%	71%	94%	85%	84%	92%	93%
L_2 PGD	30%	76%	96%	86%	88%	93%	94%
L_2 BasicIterativeAttack	17%	67%	95%	83%	83%	93%	94%
L_2 FastGradientAttack (FGM)	55%	92%	97%	94%	86%	94%	95%
L_2 DeepFoolAttack	21%	21%	95%	49%	83%	75%	89%
L_2 CarliniWagnerAttack	13%	10%	83%	45%	84%	51%	74%
L_2 BrendelBethgeAttack	12%	8%	50%	48%	93%	57%	71%
L_2 BoundaryAttack	19%	62%	54%	93%	90%	80%	80%
All L_2 attacks	1.1/9%	0.9/7%	1.4/41%	1.3/41%	**2.2/83**%	1.4/45%	1.9/66%
L_∞-metric ($\epsilon = 0.3$)							
L_∞ PGD	0%	73%	95%	88%	11%	89%	85%
L_∞ BasicIterativeAttack	0%	70%	96%	83%	8%	89%	82%
L_∞ FastGradientAttack (FGSM)	7%	78%	96%	86%	38%	90%	89%
L_∞ DeepFoolAttack	0%	83%	95%	86%	7%	91%	78%
L_∞ BrendelBethgeAttack	2%	81%	94%	89%	11%	88%	9%
All L_∞ attacks	0.08/0%	0.36/69%	0.34/**93**%	0.42/82%	0.22/3%	**0.49**/78%	0.2/8%
L_0-metric ($\epsilon = 12$)							
Pointwise $\times 10$	25%	43%	2%	82%	76%	53%	59%
All L_0 attacks	8/25%	11/43%	4/2%	**26/82**%	19/76%	13/53%	14/59%
L_1-metric ($\epsilon = 8$)							
L_1 BrendelBethgeAttack	11%	4%	16%	48%	89%	65%	65%
All L_1 attacks	5/11%	3/4%	4/16%	8/47%	**19/89**%	13/65%	11/65%
Minimal Accuracy	0%	4%	2%	41%	3%	**45%**	8%

Table 1 reports the model's accuracy within given bounds of perturbations, i.e., $\epsilon_{L_0} = 0.3$, $\epsilon_{L_1} = 8.$, $\epsilon_{L_2} = 1.5$, and $\epsilon_{L_\infty} = 12..$ It has been recognized that the model's accuracy on bounded adversarial perturbations is often biased (Schott et al., 2019); nonetheless, we reported it for completeness. On the other hand, the median adversarial perturbation size reflects the perturbation with which the model achieves 50% accuracy. It is hardly affected by the outliers; hence, it can help summarize the distribution of adversarial perturbations better. We also report the median perturbation size for each model in all four $L_p, p = 0, 1, 2, \infty$ norm cases (values before the slash). Note that clean samples that are already misclassified are counted as adversarial samples with a perturbation size of 0, and failed attacks are assigned a perturbation size of ∞.

For a better understanding of how each component in DIM affects the adversarial robustness, we further carry out an ablation study as well as investigate the latent representations of autoencoders in the internal models.

The ablation study involves six models. Starting from the vanilla CNN as the baseline, we first consider two ablations of DIM: the stand-alone Internal Models (IM) without the denoiser and the single-head internal model (single-IM). The single-head internal model combines a single encoder with a set of decoders, each of which corresponds to a category in the dataset. Then, we extend the single-head internal model by adding the denoiser to it. At last, we compare the performance of those ablations to DIM and biDIM. The results are summarized in Table 2.

Table 2: Results for ablation study, including six models: the vanilla CNN, the single-head Internal Model (single-IM), the Internal Model without denoiser, the single-head Internal Model with denoiser (Dn-singleIM), the DIM, and the biDIM. The rest settings are the same as those in Table 1.

	CNN	singleIM	Internal Models	Dn-singleIM	DIM	biDIM
L_2-metric ($\epsilon = 1.5$)						
L_2 DDNAttack	15%	83%	91%	87%	93%	92%
L_2 PGDAttack	30%	89%	95%	89%	94%	93%
L_2 BasicIterativeAttack	17%	88%	94%	90%	94%	93%
L_2 FastGradientAttack (FGM)	55%	89%	95%	90%	95%	94%
L_2 DeepFoolAttack	21%	71%	83%	82%	89%	75%
L_2 CarliniWagnerAttack	13%	54%	66%	68%	74%	51%
L_2 BrendelBethgeAttack	12%	61%	58%	70%	71%	57%
L_2 BoundaryAttack	19%	65%	67%	75%	80%	80%
All L_2 attacks	9%	52%	51%	65%	**66%**	45%
L_∞-metric ($\epsilon = 0.3$)						
L_∞ PGDAttack	0%	49%	70%	72%	85%	89%
L_∞ BasicIterativeAttack	0%	54%	61%	72%	82%	89%
L_∞ FastGradientAttack (FGSM)	7%	64%	78%	79%	89%	90%
L_∞ DeepFoolAttack	0%	44%	61%	66%	78%	91%
L_∞ BrendelBethgeAttack	2%	2%	1%	6%	9%	88%
All L_∞ Attacks	0%	2%	0%	6%	8%	**78%**
L_0-metric ($\epsilon = 12$)						
Pointwise $\times 10$	25%	54%	50%	58%	59%	53%
All L_0 attacks	25%	54%	50%	58%	**59%**	53%
L_1-metric ($\epsilon = 8$)						
L_1 BrendelBethgeAttack	11%	61%	57%	65%	65%	65%
All L_1 attacks	11%	61%	57%	65%	**65%**	65%

In Figure 3, we visualize the clustering of latent representations in all the ten latent spaces by applying the tSNE method to reduce the dimension of the latent representations.

5 REUSLTS AND DISCUSSION

The last row of Table 1 shows the minimal accuracy of a model against all the 42 attacks. Higher minimal accuracy indicates more stable performance in robustness under different types of adversarial attacks. From the table, we find biDIM achieves the highest minimal accuracy.

Our model closely simulates the brain's visual signal processing pipeline while been implemented with relatively simple neural networks in the denoiser and internal models. Its stable robustness suggests that drawing inspiration from bio-systems could be a promising direction to explore when tackling the robustness issue in deep learning.

For a more detailed comparison with the SOTA methods:

- For L_2 attacks, ABS has the highest accuracy while biDIM outperforms biABS in both accuracy and median perturbations size. Madry achieves good performance except for two boundary attacks: the L_2 boundary attack and L_2 BBA, which considerably degrade its overall robustness.

- For L_∞ attacks, Madry has the best accuracy since their adversarial training is based on the L_∞ norm. We find that the overall accuracy of ABS and DIM both decrease rapidly. The individual accuracy of DIM only drops for the L_∞ BBA case, while the performance of ABS deteriorates on all five listed attacks. On the other hand, both biDIM and biABS

retain decent robustness under L_∞ attacks with the help of the input binarization. It is worth noting that biDIM has the largest median adversarial perturbation size.

- Under L_0 attacks, Madry suffers a significant decrease in accuracy, becoming even worse than the baseline methods. On the contrary, DIM/biDIM and ABS/biABS still show moderate performance, especially the biABS, which has the highest accuracy.

- For L_1 attacks, no model performs particularly poorly. ABS stands out in this case, while biDIM outperforms biABS and both of them are much better than Madry.

In summary, biDIM has the most stable performance over all kinds of attacks in our experiments, even though it may be inferior to other SOTA methods under specific circumstances. More importantly, biDIM achieves the highest minimal accuracy, indicating that it is not only stable but also a competitive defense method against all types of adversarial attacks.

We would like to also remark that the inference using DIM (biDIM) is much faster than the ABS (biABS), which may take several seconds for a single forward pass. This heavy time cost highly restricts the extensibility of the ABS model.
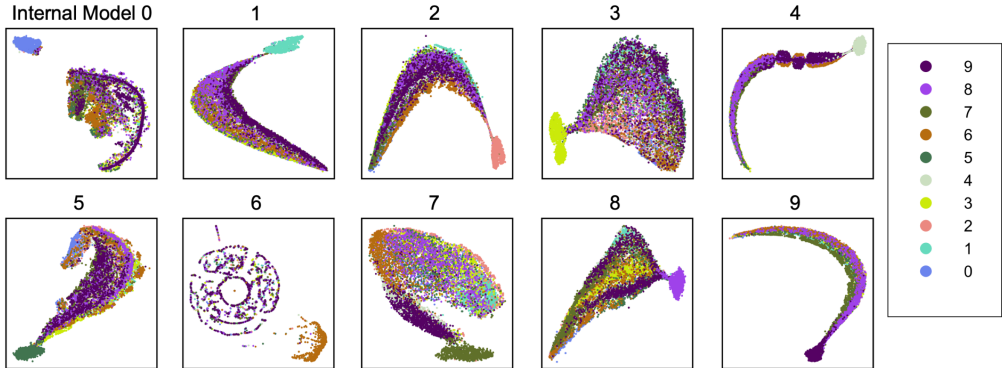

Figure 3: The clustering of the latent representations of the ten autoencoders in the Internal Models. Different colors corresponds to different categories (digits) in the MNIST dataset.

In Table 2, we first compared our DIM model with its three ablations: the single-head Internal Model (singleIM), the stand-alone Internal Models, and the singleIM with denoiser (Dn-singleIM). A vanilla CNN is included as the baseline. We found that extending from the singleIM to the Internal Models generally increases the accuracy, indicating better robustness. Further, models with the denoiser outperform the ones without in almost all cases. More interestingly, we note that the increase in accuracy is more evident for L_∞ attacks. The comparison between DIM and biDIM shows that input binarization is a very effective method against L_∞ attacks. However, our results suggest that it often degrades the performance under L_2 attacks.

At last, we studied the clustering of the latent representations, whose knowledge might provide us clues about how the internal models work. For the sake of visualization, we apply the tSNE algorithm (van der Maaten & Hinton, 2008) to map the original 10-dimensional latent representations into 2-dimensional ones. The results are shown in Figure 3.

From Figure 3, we saw that, for the i-th autoencoder, the distribution of the representations in the i-th category is well centralized and stands distinguished from the others. In addition, the representations of other categories are scattered over the latent space and show no apparent pattern. This behavior reflects how we train internal models, i.e., we require each autoencoder to only react to images from the corresponding category and return a black image otherwise.

6 CONCLUSION

In this work, we proposed the Denoised Internal Models (DIM), a novel two-stage model closely following the human brain's visual signal processing paradigm. The model is carefully designed to

mimic the cooperation between the thalamus and the primary visual cortex. Moreover, instead of a single complex model, we adopt a set of simple internal generative models to encode different categories in the dataset. This reflects the sparse coding that is based on the engram cells in Layer 2/3 of the primary visual cortex. Recent progress in neuroscience suggests that the engram-based generative models may serve as the base of a robust cognitive function in the human brain.

In order to comprehensively evaluate the robustness of our model, we conducted extensive experiments across a broad spectrum of adversarial attacks. The results (see Table 1) demonstrate that DIM, especially its variation biDIM, achieves a stable and competitive performance over all kinds of attacks. biDIM beats the SOTA methods adversarial training and ABS on the overall performance across the 42 attacks in our experiments. Further investigations show that the clusters corresponding to different categories are well separated in the latent spaces of the internal models, which provides some clues about the good robustness of DIM. The present work is an initial attempt to integrate the brain's working mechanism with the model design in deep learning. We will explore more sophisticated realizations of the internal models and extend our model to real-world datasets in future work. May the brain guides our way.

REFERENCES

Abien Fred Agarap. Deep learning using rectified linear units (relu), 2019.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*, pp. 173–182. PMLR, 2016.

Anish Athalye and Nicholas Carlini. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pp. 274–283. PMLR, 2018a.

Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018b.

Mitali Bafna, Jack Murtagh, and Nikhil Vyas. Thwarting adversarial examples: an l 0-robust sparse fourier transform. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 10096–10106, 2018.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

W Brendel, J Rauber, M Kümmerer, I Ustyuzhaninov, and M Bethge. Accurate, reliable and fast robustness evaluation. In *Thirty-third Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 12817–12827. Curran, 2020.

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

Per Brodal. *The central nervous system: structure and function.* oxford university Press, 2004.

Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717, 2009. doi: 10.1007/s10208-009-9045-5. URL https://doi.org/10.1007/s10208-009-9045-5.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Enzo Casamassima, Andrew Herbert, and Cory Merkel. Exploring cnn features in the context of adversarial robustness and human perception. In *Applications of Machine Learning 2021*, volume 11843, pp. 1184313. International Society for Optics and Photonics, 2021.

Sourav Chatterjee. Matrix estimation by universal singular value thresholding. *The Annals of Statistics*, 43(1), Feb 2015. ISSN 0090-5364. doi: 10.1214/14-aos1272. URL http://dx.doi.org/10.1214/14-AOS1272.

Yudong Chen and Yuejie Chi. Harnessing structures in big data via guaranteed low-rank matrix estimation: Recent theory and fast algorithms via convex and nonconvex optimization. *IEEE Signal Processing Magazine*, 35(4):14–31, 2018.

Celia Cintas, Skyler Speakman, Victor Akinwande, William Ogallo, Komminist Weldemariam, Srihari Sridharan, and Edward McFowland. Detecting adversarial attacks via subset scanning of autoencoder activations and reconstruction error. IJCAI, 2020.

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

Andrew M Derrington, John Krauskopf, and Peter Lennie. Chromatic mechanisms in lateral geniculate nucleus of macaque. *The Journal of physiology*, 357(1):241–265, 1984.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990. doi: https://doi.org/10.1207/s15516709cog1402_1. URL https://onlinelibrary.wiley.com/doi/abs/10.1207/s15516709cog1402_1.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014a.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014b.

Jiuxiang Gu, Zhenhua Wang, Jason Kuen, Lianyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern Recognition*, 77:354–377, 2018.

Ji-Song Guan, Jun Jiang, Hong Xie, and Kai-Yuan Liu. How does the sparse memory "engram" neurons encode the memory of a spatial–temporal event? *Frontiers in neural circuits*, 10:61, 2016.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. On the limitation of convolutional neural networks in recognizing negative images. In *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 352–358. IEEE, 2017.

Shengyuan Hu, Tao Yu, Chuan Guo, Wei-Lun Chao, and Kilian Q Weinberger. A new defense against adversarial images: Turning a weakness into a strength. *Advances in Neural Information Processing Systems*, 32, 2019.

Yujia Huang, Sihui Dai, Tan Nguyen, Pinglei Bao, Doris Y Tsao, Richard G Baraniuk, and Anima Anandkumar. Brain-inspired robust vision using convolutional neural networks with feedback. 2019.

Uyeong Jang, Susmit Jha, and Somesh Jha. On the need for topology-aware generative models for manifold-based defenses. In *International Conference on Learning Representations*, 2019.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world, 2016.

Y LeCun, B Boser, JS Denker, D Henderson, RE Howard, W Hubbard, and LD Jackel. Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1(4):541–551, 1989.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

Xu Liu, Steve Ramirez, Petti T Pang, Corey B Puryear, Arvind Govindarajan, Karl Deisseroth, and Susumu Tonegawa. Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature*, 484(7394):381–385, 2012.

Xu Liu, Steve Ramirez, and Susumu Tonegawa. Inception of a false memory by optogenetic manipulation of a hippocampal memory engram. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1633):20130142, 2014.

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

Henry Markram, Eilif Muller, Srikanth Ramaswamy, Michael W Reimann, Marwan Abdellah, Carlos Aguado Sanchez, Anastasia Ailamaki, Lidia Alonso-Nanclares, Nicolas Antille, Selim Arsever, et al. Reconstruction and simulation of neocortical microcircuitry. *Cell*, 163(2):456–492, 2015.

James L McGaugh. Memory–a century of consolidation. *Science*, 287(5451):248–251, 2000.

Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pp. 135–147, 2017.

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 conference on computer vision and pattern recognition*, pp. 2574–2582, 2016.

Daniel H O'Connor, Miki M Fukui, Mark A Pinsk, and Sabine Kastner. Attention modulates responses in the human lateral geniculate nucleus. *Nature neuroscience*, 5(11):1203–1209, 2002.

Tianyu Pang, Kun Xu, Yinpeng Dong, Chao Du, Ning Chen, and Jun Zhu. Rethinking softmax crossentropy loss for adversarial robustness. In *International Conference on Learning Representations*, 2019a.

Tianyu Pang, Kun Xu, Chao Du, Ning Chen, and Jun Zhu. Improving adversarial robustness via promoting ensemble diversity. In *International Conference on Machine Learning*, pp. 4970–4979. PMLR, 2019b.

R Quian Quiroga, Leila Reddy, Gabriel Kreiman, Christof Koch, and Itzhak Fried. Invariant visual representation by single neurons in the human brain. *Nature*, 435(7045):1102–1107, 2005.

Jonas Rauber and Matthias Bethge. Fast differentiable clipping-aware normalization and rescaling. *arXiv preprint arXiv:2007.07677*, 2020.

Jonas Rauber, Wieland Brendel, and Matthias Bethge. Foolbox: A python toolbox to benchmark the robustness of machine learning models. *arXiv preprint arXiv:1707.04131*, 2017.

Jérôme Rony, Luiz G Hafemann, Luiz S Oliveira, Ismail Ben Ayed, Robert Sabourin, and Eric Granger. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

F Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386—408, November 1958. ISSN 0033-295X. doi: 10.1037/h0042519. URL https://doi.org/10.1037/h0042519.

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

L Schott, J Rauber, M Bethge, and W Brendel. Towards the first adversarially robust neural network model on mnist. In *Seventh International Conference on Learning Representations (ICLR 2019)*, pp. 1–16, 2019.

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*, 2014.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

Susumu Tonegawa, Xu Liu, Steve Ramirez, and Roger Redondo. Memory engram cells have come of age. *Neuron*, 87(5):918–931, 2015.

Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. In *International Conference on Learning Representations*, 2018.

Pratik Vaishnavi, Tianji Cong, Kevin Eykholt, Atul Prakash, and Amir Rahmati. Can attention masks improve adversarial robustness? In *International Workshop on Engineering Dependable and Secure Machine Learning Systems*, pp. 14–22. Springer, 2020.

L.J.P. van der Maaten and G.E. Hinton. Visualizing high-dimensional data using t-sne. 2008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pp. 5998–6008, 2017.

Gunjan Verma and Ananthram Swami. Error correcting output codes improve probability estimation and adversarial robustness of deep neural networks. *Advances in Neural Information Processing Systems*, 32:8646–8656, 2019.

Pascal Vincent, Hugo Larochelle, Isabelle Lajoie, Yoshua Bengio, Pierre-Antoine Manzagol, and Léon Bottou. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(12), 2010.

Brian J White, Susan E Boehnke, Robert A Marino, Laurent Itti, and Douglas P Munoz. Color-related signals in the primate superior colliculus. *Journal of Neuroscience*, 29(39):12159–12166, 2009.

Chang Xiao, Peilin Zhong, and Changxi Zheng. Enhancing adversarial defense by k-winners-take-all. *arXiv preprint arXiv:1905.10510*, 2019.

Hong Xie, Yu Liu, Youzhi Zhu, Xinlu Ding, Yuhao Yang, and Ji-Song Guan. In vivo imaging of immediate early gene expression reveals layer-specific memory traces in the mammalian brain. *Proceedings of the National Academy of Sciences*, 111(7):2788–2793, 2014.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Mike Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. Achieving human parity in conversational speech recognition. *arXiv preprint arXiv:1610.05256*, 2016.

Yaoda Xu and Maryam Vaziri-Pashkam. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nature communications*, 12(1):1–16, 2021.

Yuzhe Yang, Guo Zhang, Zhi Xu, and Dina Katabi. Me-net: Towards effective adversarial robustness with matrix estimation. In *ICML*, 2019a.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019b.

Xuwang Yin, Soheil Kolouri, and Gustavo K Rohde. Adversarial example detection and classification with asymmetrical adversarial training. *arXiv preprint arXiv:1905.11475*, 2019.