Contrasting Adversarial Perturbations: The Space of Harmless Perturbations

Lu Chen[1, 2] **Shaofeng Li**[*3], **Benhao Huang**[1], **Fan Yang**[1], **Zheng Li**[1], **Jie Li**[1, 2], **Yuan Luo**[*1, 2]

[1]Shanghai Jiao Tong University, China
[2]Shanghai Jiao Tong University (Wuxi) Blockchain Advanced Research Center, China
[3]Southeast University, China
lu.chen@sjtu.edu.cn, shaofengli@seu.edu.cn, hbh001098hbh@sjtu.edu.cn, fan-yang@sjtu.edu.cn,
zli89@lsu.edu, lijiecs@sjtu.edu.cn, luoyuan@cs.sjtu.edu.cn

Abstract

Existing works have extensively studied adversarial examples, which are minimal perturbations that can mislead the output of deep neural networks (DNNs) while remaining imperceptible to humans. However, in this work, we reveal the existence of a harmless perturbation space, in which perturbations drawn from this space, regardless of their magnitudes, leave the network output unchanged when applied to inputs. Essentially, the harmless perturbation space emerges from the usage of non-injective functions (linear or non-linear layers) within DNNs, enabling multiple distinct inputs to be mapped to the same output. For linear layers with input dimensions exceeding output dimensions, any linear combination of the orthogonal bases of the nullspace of the parameter consistently yields no change in their output. For non-linear layers, the harmless perturbation space may expand, depending on the properties of the layers and input samples. Inspired by this property of DNNs, we solve for a family of general perturbation spaces that are redundant for the DNN's decision, and can be used to hide sensitive data and serve as a means of model identification. Our work highlights the distinctive robustness of DNNs (*i.e.*, consistency under large magnitude perturbations) in contrast to adversarial examples (vulnerability for small noises).

Code — https://github.com/csluchen/harmless-perturbations
Extended version — https://arxiv.org/pdf/2402.02095

Introduction

The robustness of Deep Neural Networks (DNNs) against structured and unstructured perturbations has attracted significant attention in recent years (Szegedy et al. 2014; Nguyen, Yosinski, and Clune 2015; Fawzi, Moosavi-Dezfooli, and Frossard 2016; Salman et al. 2021). In particular, deep learning models are shown highly vulnerable to adversarial perturbations (Szegedy et al. 2014). These well-crafted perturbations, which are imperceptibly small to the human eye, cause DNNs to misclassify with high confidence (Carlini and Wagner 2017; Madry et al. 2018; Croce and Hein 2020). Naturally, an inquiry arises:

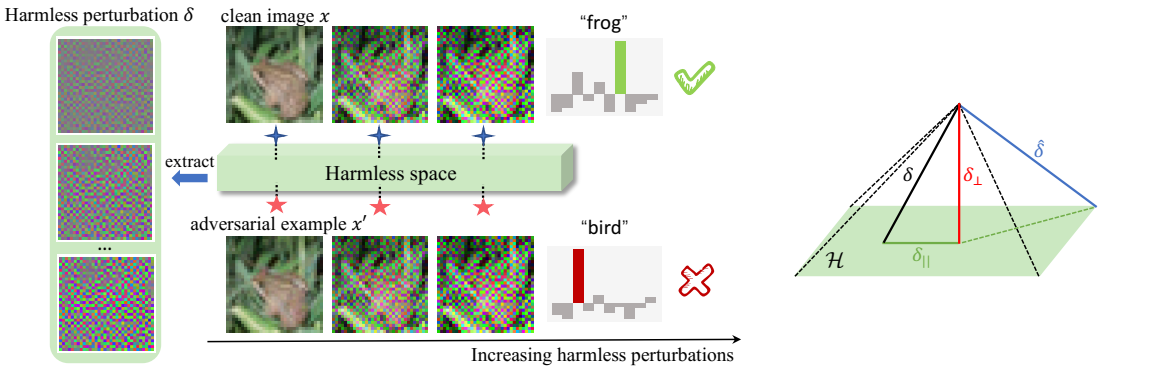
Are there perturbations within the input space capable of preserving network output invariance?

Unlike vulnerability against adversarial examples, in this paper, we reveal the robustness of DNNs to specific perturbations that render the network output mathematically strictly invariant. We demonstrate the existence of such *harmless* perturbations that, when introduced onto natural images or embeddings, regardless of their magnitude, will not affect the discrimination of the DNN. Such harmless perturbations arising from the linear layers are universal, as they are instance-independent and solely determined by the parameter space of the DNN. These harmless perturbations span a continuous harmless subspace, embedded within the high-dimensional feature space. The surprising existence of harmless perturbations reveals a distinctive view of DNN robustness.

For the linear layers of DNNs, we find that when its input dimension n exceeds the output dimension m, the harmless perturbation subspace of this layer can be derived by computing the *nullspace* of its parameter matrix A, *i.e.*, $N(A) = \{v \in \mathbb{R}^n | Av = \mathbf{0}\}$. To this end, the harmless subspace exhibits a dimension of $(n - m)$ and is embedded within an n-dimensional feature space. Furthermore, the harmless perturbation space *may* expand when involving non-linear layers, depending on the specific non-linear functions and input samples. Inspired by the harmless subspace of linear layers, we further investigate the robustness of DNNs against more general perturbations, *i.e.,* random noises or adversarial perturbations. We find that a family of those general perturbations, irrespective of their magnitude, identically influence the DNN's output. This phenomenon stems from the decomposition of arbitrary perturbations into the sum of any harmless and harmful components. Consequently, the network output for general perturbations becomes equivalent to that of harmful perturbations, particularly aligning with that of components orthogonal to the harmless perturbation subspace (Figure 1(b)).

The existence of harmless perturbations and their space promotes several potential benefits. First, capitalizing on the disparity between DNNs and human perception, *i.e.*, significant perturbations perceivable by the human eye may not affect the recognition of DNNs, we delve into the application of harmless perturbations to privacy-preserving data and model fingerprints. Additionally, as demonstrated in Figure 1(a), there exist equivalent adversarial spaces, ensuring equal attacking capabilities for adversarial perturbations regardless of their magnitude. In other words, the perturbation

*Corresponding author.



(a) Harmless perturbations (b) Decomposition of general perturbations

Figure 1: (a) Harmless perturbations added to images completely do not change the network output of the images, regardless of the magnitude of these harmless perturbations. (b) Illustration of the equivalent effect of any perturbation on the network output. Given any linear layer with a harmless subspace $\mathcal{H}$, the network outputs of any perturbations δ and $\hat{\delta}$ are equivalent to those of their components $\delta_\perp$ orthogonal to the harmless subspace.

magnitude is not a decisive factor in attacking the network. Instead, focusing on the attack utility of the "effective component" of the perturbation facilitates a deeper understanding of the robustness of DNNs. In summary, this paper makes the following contributions:

- We demonstrate for the first time the concept of "harmless perturbations" and show the existence of a harmless perturbation space for DNNs. For any linear layer with the input dimension n exceeding the output dimension m, there exists a continuous harmless perturbation subspace of dimension $(n - m)$. The harmless perturbation space *may* expand when considering non-linear layers, depending on the properties of the layers and input samples.
- We present a novel perspective to decompose any general perturbation (*i.e.,* random noises or adversarial perturbations) into its harmful and harmless counterparts. Given any linear layer with a harmless subspace, the network output solely depends on its orthogonal (harmful) component, irrespective of its magnitude (innocuous) part.
- We reveal the difference between DNNs and human perception, *i.e.*, significant perturbations captured by humans may not affect DNNs, which highlights a distinctive aspect of DNN robustness. Based on this insight, we employ the proposed harmless perturbations with a large magnitude to hide the sensitive image data for DNN usage. As harmless perturbations are usually not transferable across different DNNs, they can also serve as model fingerprints.

Related Work

Adversarial Examples and Adversarial Robustness. Existing literature extensively explored the impact of adversarial perturbations (Szegedy et al. 2014; Goodfellow, Shlens, and Szegedy 2015; Papernot et al. 2017; Moosavi-Dezfooli et al. 2017) on the robustness of DNNs. Many defenses against these adversarial perturbations have been proposed but they were susceptible to being broken by more powerful or adapted attacks (Carlini and Wagner 2017; Athalye, Carlini, and Wagner 2018). Amongst them, adversarial training (Madry et al. 2018) and its variant (Zhang et al. 2019) still indicated their relatively reliable robustness against more powerful attack (Croce and Hein 2020).

Adversarial Space. Previous studies have delved into the vulnerability of DNNs from the perspective of high-dimensional input spaces. Goodfellow, Shlens, and Szegedy (2015) argued that the "highly linear" of DNNs explained their instability to adversarial perturbations. Fawzi, Moosavi-Dezfooli, and Frossard (2016) quantified the robustness of classifiers from the dimensionality of subspaces within the semi-random noise regime. Gilmer et al. (2018) suggested that adversarial perturbations arised from the high-dimensional geometry of data manifolds. Tramèr et al. (2017) stated that adversarial transferability arised from the intersection of high-dimensional adversarial subspaces from different models. Shafahi et al. (2019) empirically discussed that how dimensionality affected the robustness of classifiers to adversarial perturbations. Jetley, Lord, and Torr (2018) identified that the directions in the input space most vulnerable to attacks overlap with those used to achieve classification performance.

Unrecognizable Features. A series of prior works (Geirhos et al. 2019; Ilyas et al. 2019; Tsipras et al. 2019; Jacobsen et al. 2019a; Yin et al. 2019; Wang et al. 2020) have demonstrated that humans and DNNs tend to utilize different features to make decisions. Besides, producing totally unrecognizable images (Nguyen, Yosinski, and Clune 2015) or introducing visually perceptible patches to images (Salman et al. 2021; Wang et al. 2022; Si et al. 2023) may not alter the classification categories of DNNs. In constrast, Tao et al. (2022) proposed that slight modifications to a misclassified sample can lead to correct classification.

"Harmless" is a concept similar to Definition 1 in (Jacobsen et al. 2019b), which introduced invariant perturbations for networks and optimized generated images with semantically meaningful variations from natural images. Unlike this work, we define and derive exact solutions for the harmless perturbation space that keep logits invariant for any DNN.

The Space of Harmless Perturbations

We develop a framework to rigorously define "harmless" and "harmful" perturbations *w.r.t.* the network output. In particular, we formally define and solve for the subspace for harmless perturbations in any linear layer of a given DNN. Subsequently, the definitions and solutions are extended to non-linear layers by analyzing the properties of the functions.

Harmless Perturbations for a Linear Layer

Consider a function mapping $\mathcal{L} : \mathbb{R}^n \mapsto \mathbb{R}^m$ on an input sample $x \in \mathbb{R}^n$, the goal is to find a set of input perturbations $\delta \in \mathbb{R}^n$ that rigorously do not change the output of the function. To this end, we define harmless perturbations.

Definition 1 (Harmless perturbations)**.** *The set of harmless perturbations for a function $\mathcal{L}$ is defined as $\mathcal{S} \coloneqq \{\delta | \mathcal{L}(x+\delta) = \mathcal{L}(x)\}$ subject to $\|\delta\|_p < \xi$, $\xi > 0$.*

Definition 1 denotes a set of input perturbations that *thoroughly* do not affect the function output. Then, the set of harmless perturbations for a linear function $\mathcal{L}(x) = Ax$, where $A \in \mathbb{R}^{m \times n}$ is the parameter matrix, can be formulated as $\mathcal{S} = \{\delta | A(x+\delta) = Ax\} = \{\delta | A\delta = \mathbf{0}\}$. It indicates that the set of harmless perturbations for a single linear layer $\mathcal{L}$ is equivalent to the *nullspace* of the parameter matrix A, *i.e.*, $\mathcal{S} = N(A) = \{v \in \mathbb{R}^n | Av = \mathbf{0}\}$. ξ is application-specific.

Theorem 1 (Dimension of harmless perturbation subspace, proven by the Rank-Nullity Theorem)**.** *Given a linear layer $\mathcal{L}(x) = Ax \in \mathbb{R}^m$ and an input sample $x \in \mathbb{R}^n$, where the parameter matrix $A \in \mathbb{R}^{m \times n}$. The dimension of the subspace for harmless perturbations is $dim(\mathcal{S}) = n - rank(A)$.*

Theorem 1 shows that the subspace for harmless perturbations is the span of $dim(\mathcal{S})$ linearly independent vectors $U \subset \mathcal{S}$, *i.e.*, $\mathcal{S} = span(U) = \{\sum_{i=1}^{dim(\mathcal{S})} c_i u_i | c_i \in \mathbb{R}, u_i \in U\}$. As a special case, the parameter matrix A of a linear layer in DNNs learned through an optimization algorithm (*e.g.*, SGD) starting from an arbitrary initialization, usually possesses linearly independent (row) vectors (Feng and Zhang 2007). So the dimension of the harmless perturbation subspace for a linear layer $\mathcal{L}(x) = Ax \in \mathbb{R}^m$ is $dim(\mathcal{S}) = n - m$.

Remark 1 (proven in Appendix A)**.** *Consider the case that the input dimension of the linear layer is less than or equal to the output dimension, i.e., $n \leq m$. In this case, if the column vectors of the parameter matrix A are linearly independent, then the dimension of the subspace for harmless perturbations is $dim(\mathcal{S}) = 0$.*

Remark 1 states that there exists *no* (non-zero) harmless perturbation that does not affect the output of the linear layer when $n \leq m$ and $rank(A) = n$.

The Space of Harmless Perturbations for DNNs

Extending harmless perturbations from a single linear layer to the entire DNN is challenging. Consider a DNN $f : \mathbb{R}^{n_{\text{in}}} \mapsto \mathbb{R}^{n_{\text{out}}}$ on an input sample $x \in \mathbb{R}^{n_{\text{in}}}$, the goal now is to identify a set of harmless input perturbations $\delta \in \mathbb{R}^{n_{\text{in}}}$ which ultimately do not alter the network output. Notice that harmless perturbations solved for the intermediate layers do not influence subsequent layers. Therefore, we can formally define the set of harmless perturbations layer by layer for a DNN.

Definition 2 (Set of harmless perturbations for DNNs)**.** *The set of harmless perturbations on the $(l+1)$-th layer of a DNN f is defined as $\mathcal{H}^{(l)} \coloneqq \{\delta^{(l)} | f^{(l+1)}(z^{(l)} + \delta^{(l)}) = f^{(l+1)}(z^{(l)})\}$.*

$z^{(l)} \in \mathbb{R}^{n^{(l)}}$ represents the l-th intermediate-layer features of the input sample x, and $\delta^{(l)}$ denotes the perturbations added to the features $z^{(l)}$. Definition 2 shows that if the set of harmless perturbations on the features can be found, these perturbations leave the network output unaffected. Furthermore, if we identify a set of perturbations on the input $\mathcal{P}^{(l)} \coloneqq \{\delta | z^{(l)} + \delta^{(l)} = (f^{(l)} \circ \cdots \circ f^{(1)})(x+\delta), \exists \delta^{(l)} \in \mathcal{H}^{(l)}\}$ such that $\delta^{(l)} \in \mathcal{H}^{(l)}$, then $\mathcal{P}^{(l)}$ do not alter the network output.

Lemma 1 (proven in Appendix C)**.** *The set of harmless perturbations on the input for a DNN f with L layers is derived as $\mathcal{P} = \bigcup_{l=0}^{L-1} \mathcal{P}^{(l)}$, $\mathcal{P}^{(0)} \coloneqq \mathcal{H}^{(0)}$, $\mathcal{P} \subset \mathbb{R}^{n_{in}}$.*

Lemma 1 suggests that the set of harmless input perturbations for the entire DNN $\mathcal{P}$ is the union of the corresponding set of harmless input perturbations $\mathcal{P}^{(l)}$ on each layer. Besides, $\mathcal{P}^{(l)} \subseteq \mathcal{P}^{(l+1)}$ (proven in Appendix) shows that the set of harmless input perturbations grows monotonically with increasing layer l. Theoretically, Lemma 1 does not restrict whether any layer in the DNN is linear or nonlinear, *i.e.*, given any layer, if $\mathcal{H}^{(l)}$ and $\mathcal{P}^{(l)}$ can be evaluated, then harmless input perturbations for this layer can still be obtained. Based on Lemma 1, we further investigate the effect of a single layer of nonlinearity on the harmless perturbation space. In scenarios involving non-linear layers, the harmless perturbation space *may* expand, depending on the specific non-linear functions and input samples.

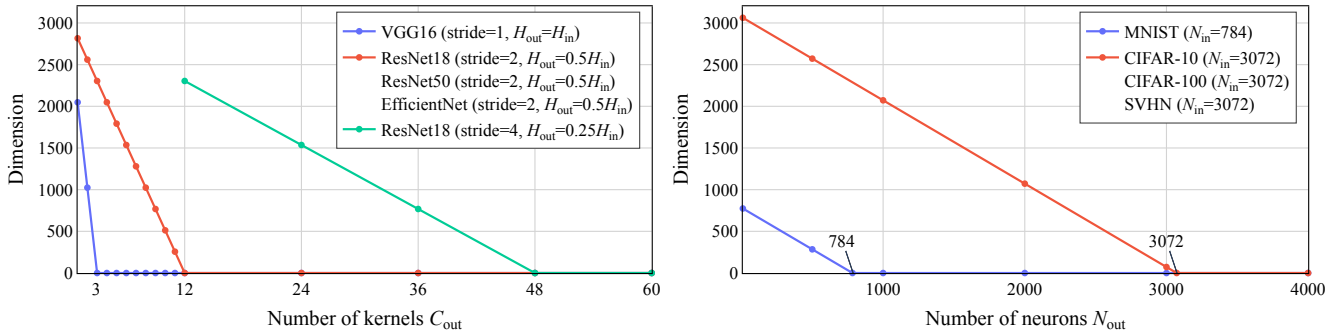
Lemma 1.1 (Harmless perturbations for injective functions, proven in Appendix C)**.** *If the layer $f^{(l+1)}$ is an injective function, $\mathcal{H}^{(l)} = \{\mathbf{0}\}$. Otherwise, $\mathcal{H}^{(l)} \neq \{\mathbf{0}\}$.*

Lemma 1.2 (Harmless perturbations for ReLU layers, proven in Appendix C)**.** *Suppose $f^{(l+1)}$ is the* ReLU *layer, $\mathcal{H}^{(l)} = \{\delta^{(l)} | \forall i, \delta_i^{(l)} = \begin{cases} 0, & z_i^{(l)} > 0 \\ t(\forall t \leq -z_i^{(l)}), & z_i^{(l)} \leq 0 \end{cases}\}$, which is determined by features $z^{(l)}$ and hence the input sample x.*

Lemma 1.3 (Harmless perturbations for Softmax layers, proven in Appendix C)**.** *Suppose $f^{(l+1)}$ is the* Softmax *layer, $\mathcal{H}^{(l)} = \{c \cdot \mathbf{1}, c \in \mathbb{R}\}$.*

Lemma 1.4 (Harmless perturbations for Average Pooling layers, proven in Appendix C)**.** *Suppose $f^{(l+1)}$ is the* Average Pooling *layer, $\mathcal{H}^{(l)} = N(A_{\text{avg}})$. A_{avg} is a coefficient matrix determined by the constraints that must be satisfied by the perturbations within each averaging region.*

Lemma 1.5 (Harmless perturbations for Max Pooling layers, proven in Appendix C)**.** *Suppose $f^{(l+1)}$ is the* Max Pooling *layer, $\mathcal{H}^{(l)} = \{\forall p, i, \delta_{p,i}^{(l)} \leq c_p - z_{p,i}^{(l)}\} \cap \{\forall p, \prod_{j=1}^{k \times k} (\delta_{p,j}^{(l)} - c_p + z_{p,j}^{(l)}) = 0\}$. $c_p \coloneqq \max\{z_{p,1}^{(l)}, z_{p,2}^{(l)}, \cdots, z_{p,k \times k}^{(l)}\}$ is the maximum value of features within the $k \times k$ region of the p-th patch. $\mathcal{H}^{(l)}$ is determined by intermediate-layer features $z^{(l)}$ and hence the input sample x.*



(a) Convolutional layer (b) Fully-connected layer

Figure 2: Dimension of harmless perturbation subspace for (a) convolutional layers and (b) fully-connected layers. When the input dimension n of the linear layer is larger than the output dimension m, the dimension of the harmless subspace is $(n-m)$. Otherwise, the dimension is 0.

Theorem 2 (Harmless perturbations for two-layer neural networks, proven in Appendix A)**.** *Given a two-layer neural network $f(x) = \sigma(Ax)$, where σ represents any function. If σ is an injective function, the set of harmless perturbations on the input $\mathcal{P}$ for f is $\mathcal{P} = \mathcal{P}^{(0)}$. Otherwise, $\mathcal{P} = \mathcal{P}^{(0)} \cup \mathcal{P}^{(1)} \supsetneq \mathcal{P}^{(0)}$. Here, $\mathcal{P}^{(1)} = \{\delta | A\delta = \delta^{(1)}, \exists \delta^{(1)} \in \mathcal{H}^{(1)} \cap C(A)\}$*[1] *is determined by the function σ and the input sample x.*

Theorem 2 suggests that the property of the function σ determines whether the set of harmless perturbations for Ax may expand. For instance, if σ is an injective function, such as Sigmoid, Tanh, leaky ReLU (Maas, Hannun, and Ng 2013), exponential linear unit (ELU) (Clevert, Unterthiner, and Hochreiter 2016) and scaled exponential linear unit (SeLU) (Klambauer et al. 2017) activation functions, and the linear Batch Normalization (BN) layers at inference time (Ioffe and Szegedy 2015), the set of harmless perturbations on the input $\mathcal{P}$ remains unchanged, compared to that of Ax. Conversely, if σ is a non-injective function, such as ReLU (Nair and Hinton 2010), Softmax, Average Pooling (LeCun et al. 1990), and Max Pooling layers (Scherer, Muller, and Behnke 2010) (see Lemmas 1.2 to 1.5 and Theorem 2 for their $\mathcal{P}$, respectively), the set of harmless perturbations on the input $\mathcal{P}$ *may* expand $\mathcal{P} \supseteq \mathcal{P}^{(0)}$, depending on the specific functions and input samples. Note that, in the above non-linear layers, the harmless perturbation space for the ReLU layer is determined by the input sample x. In an extreme case, if every element of Ax is positive, then its harmless perturbation subspace $\mathcal{P} = \mathcal{P}^{(0)}$. Otherwise, if every element of Ax is not positive, $\mathcal{P} = \mathcal{P}^{(0)} \cup \mathcal{P}^{(1)} \supsetneq \mathcal{P}^{(0)}$. (For more details, please refer to Lemma 1.2 in Appendix C). In summary, the harmless perturbation space on the input *does* expand $\mathcal{P} \supsetneq \mathcal{P}^{(0)}$ if there exists at least one harmless perturbation $\delta^{(1)} \in \mathcal{H}^{(1)} \cap C(A) (\delta^{(1)} \neq \mathbf{0})$ for the non-injective function σ.

Lemma 1.6 (Harmless perturbations for two-layer linear networks, proven in Appendix C)**.** *Given a two-layer linear network $f(x) = A_2 A_1 x$, $\mathcal{P} = \mathcal{P}^{(0)} \cup \mathcal{P}^{(1)} \supseteq \mathcal{P}^{(0)}$. Here, $\mathcal{P}^{(0)} = N(A_1)$ and $\mathcal{P}^{(1)} = \{\delta | A_1\delta = \delta^{(1)}, \exists \delta^{(1)} \in N(A_2) \cap C(A_1)\}$.*

Furthermore, Lemma 1.6 illustrates the expansion of harmless perturbations on the input $\mathcal{P}$ solely depends on the dimensions of those two linear layers. For two common scenarios in DNNs, where given $A_1 \in \mathbb{R}^{d \times n}$ and $A_2 \in \mathbb{R}^{m \times d}$, when $n, m > d$, $\mathcal{P} = \mathcal{P}^{(0)}$. Otherwise, when $n, m < d$, $\mathcal{P} = \mathcal{P}^{(1)}$. (Please see Lemma 1.6 in Appendix C for the details.)

The Subspace of Harmless Perturbations for Linear Layers in DNNs

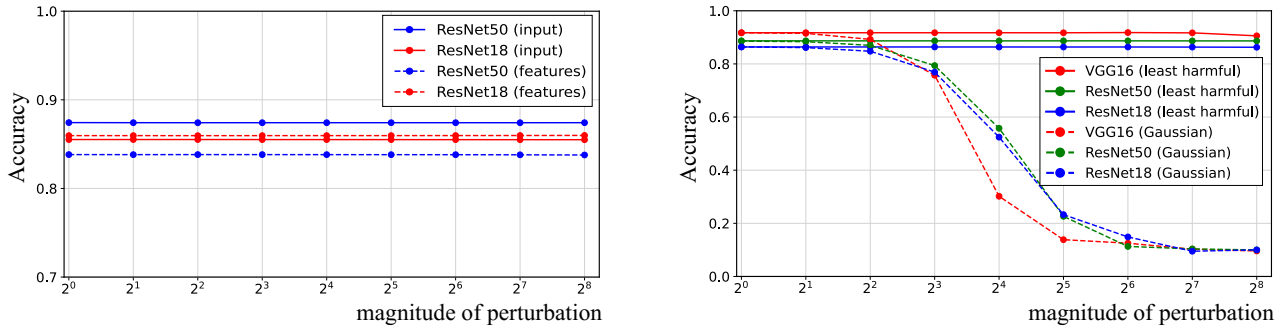
Nevertheless, in this section, we focus on the set of harmless perturbations for two classical linear layers in DNNs, *i.e.*, convolutional layers and fully-connected layers.

Corollary 1 (Harmless subspace for convolutional layers, proven in Appendix B)**.** *Given a convolutional layer $f^{(l+1)}$ with linearly independent vectorized kernels whose kernel size is not smaller than the stride, $z^{(l+1)} = f^{(l+1)}(z^{(l)}) \in \mathbb{R}^{C_{out} \times H_{out} \times W_{out}}$ and $z^{(l)} \in \mathbb{R}^{C_{in} \times H_{in} \times W_{in}}$. If the input dimension is greater than the output dimension, then the dimension of the subspace for harmless perturbations is $dim(\mathcal{H}^{(l)}) = C_{in}H_{in}W_{in} - C_{out}H_{out}W_{out}$. Otherwise, $\mathcal{H}^{(l)} = \{\mathbf{0}\}$.*

Corollary 1 demonstrates the subspace for harmless perturbations in a convolutional layer is the span of $dim(\mathcal{H}^{(l)})$ linearly independent vectors $U \subset \mathcal{H}^{(l)}$. Specifically, $\mathcal{H}^{(l)}$ can be obtained by computing the nullspace of a matrix $A \in \mathbb{R}^{(C_{out}H_{out}W_{out}) \times (C_{in}H_{in}W_{in})}$. In practice, A is affected by the padding and the stride of the convolutional layer (see Appendix E for details). Similarly, given a fully-connected layer $z^{(l+1)} = W^\top z^{(l)} \in \mathbb{R}^{N_{out}}$ and $z^{(l)} \in \mathbb{R}^{N_{in}}$, the harmless subspace is the span of $dim(\mathcal{H}^{(l)}) = N_{in} - N_{out}$ linearly independent vectors $U \subset \mathcal{H}^{(l)}$ (see Corollary 2 in Appendix B). Here, $\mathcal{H}^{(l)}$ is the nullspace of a matrix $A = W^\top$.

Experiments on various DNNs verify Corollaries 1 and 2. In Figure 2, the dimension of the harmless perturbation

[1]Note that the equation $A\delta = \delta^{(1)} (\delta \neq \mathbf{0})$ has a solution (meaning at least one solution) if and only if $\delta^{(1)}$ is in the column space of A, *i.e.*, $\delta^{(1)} \in C(A)$.



(a) Harmless perturbation (b) Least harmful perturbation

Figure 3: The effect of perturbation magnitude on the performance of the network. We trained the CIFAR-10 dataset on various networks and tested the effect of varying magnitudes on (a) harmless perturbations and (b) the least harmful perturbations.

	ϵ	2ϵ	4ϵ	8ϵ	16ϵ	32ϵ
Gaussian noise	0.1226	0.3154	0.8112	1.7871	3.3921	5.0009
Adversarial perturbation (Madry et al. 2018)	6.1994	6.3225	5.5410	5.6122	6.6585	11.2747
Harmless perturbation	**3.63e-15**	**3.70e-15**	**3.77e-15**	**4.20e-15**	**5.38e-15**	**8.55e-15**
Least harmful perturbation	0.0003	0.0007	0.0013	0.0027	0.0053	0.0105

Table 1: Root mean squared errors between the network outputs of the perturbed and original images on the CIFAR-10 dataset.

subspace $dim(\mathcal{H}^{(l)})$ decreased as the output dimension increased. When the output dimension exceeds the input dimension, $dim(\mathcal{H}^{(l)})$ becomes 0. Specifically, we verified the dimension of the harmless subspace for convolutional layers using various DNNs, including ResNet-18/50 (He et al. 2016), VGG-16 (Simonyan and Zisserman 2014) and EfficientNet (Tan and Le 2019), on the CIFAR-10 dataset (Krizhevsky, Hinton et al. 2009). Here, we modified the feature size of the output of the first convolutional layer by setting different strides (see Appendix F.1). Furthermore, we verified the dimension of the harmless perturbation subspace for fully-connected layers using the MLP-5 on various datasets, including the MNIST dataset (LeCun and Cortes 2010), the CIFAR-10/100 dataset (Krizhevsky, Hinton et al. 2009) and the SHVN dataset (Netzer et al. 2011), to compare the dimension of the subspace under different input dimensions.

Conversely, there exists *no* (non-zero) perturbation making the network output invariant, if the input dimension of a given linear layer is not greater than the output dimension. However, the least harmful perturbation can be solved for such that the layer output is minimally affected, *i.e.*, given the matrix A with equivalent effect of a linear layer, the least harmful perturbation is $(\delta^{(l)})^* = \text{argmin}_{\delta^{(l)}} \|A\delta^{(l)}\|_2$, *s.t.*, $\|\delta^{(l)}\|_2 = 1$. Hence, the least harmful perturbation $(\delta^{(l)})^*$ is the eigenvector corresponding to the smallest eigenvalue of the matrix $A^\top A$ (see Lemma 2 in Appendix C).

We validated the impact of harmless perturbations and the least harmful perturbations on network performance across varying perturbation magnitudes. In Figure 3(a),harmless perturbations, regardless of their magnitude, do not affect the discrimination of DNNs (see Appendix F.2 for details). For the least harmful perturbations in Figure 3(b), they also have negligible effects on the network performance, compared with the Gaussian noise $\mathcal{N}(0, 1)$ added to each pixel. Furthermore, we evaluated the root mean squared error (RMSE) between the network outputs of the perturbed images $\hat{y}_x$ and the network outputs of natural images y_x on the ResNet-50, *i.e.*, RMSE=$\mathbb{E}_x[\frac{1}{\sqrt{n}}||\hat{y}_x - y_x||]$. Table 1 further demonstrates that compared to adversarial perturbations and Gaussian noise, harmless perturbations completely did not change the network output with negligible errors, and the least harmful perturbation had a weak impact on the network output as the perturbation magnitude increased.

Projection onto the Harmless Subspace

Inspired by the harmless subspace of linear layers, we can decompose any given perturbation (*e.g.,* random noise, adversarial perturbations) into its two orthogonal counterparts, namely, harmful and harmless components. This section extends the harmless subspace to any given perturbations and investigates the projections of these perturbations onto their corresponding harmless subspaces.

Theorem 3 (Arbitrary decomposition of perturbations, proven in Appendix A)**.** *Given the $(l+1)$-th linear layer with harmless subspace $\mathcal{H}^{(l)} \neq \{\mathbf{0}\}$ and any perturbation $\forall \delta^{(l)} \notin \mathcal{H}^{(l)}$, it can be arbitrarily decomposed into the sum of a harmless perturbation and a harmful perturbation, i.e., $\delta^{(l)} = \delta_a^{(l)} + \delta_b^{(l)}, \forall \delta_a^{(l)} \in \mathcal{H}^{(l)}$ and $\delta_b^{(l)} \notin \mathcal{H}^{(l)}$. Then, $f^{(l+1)}(\delta^{(l)}) = f^{(l+1)}(\delta_b^{(l)})$.*

Theorem 3 indicates that the network output of any perturbation $\delta^{(l)} \notin \mathcal{H}^{(l)}$ is equivalent to that of its corresponding harmful component $\delta_b^{(l)} := (\delta^{(l)} - \delta_a^{(l)}) \notin \mathcal{H}^{(l)}, \forall \delta_a^{(l)} \in \mathcal{H}^{(l)}$,

no matter how large the ℓ_p norm of harmful component is. According to Theorem 3, an infinite number of perturbations, regardless of their magnitude, will induce the equivalence of a continuous harmful space[2]. What is the extent of these perturbations concerning a given DNN? Theorem 4 extends the argument by establishing the existence of a *unique* perturbation characterized by the smallest ℓ_2 norm. This perturbation is orthogonal to the harmless subspace, and exhibits network output consistent with the infinite number of perturbations embedded in the continuous harmful space (Figure 1(b)).

Theorem 4 (Orthogonal decomposition of perturbations, proven in Appendix A)**.** *Given the $(l+1)$-th linear layer with harmless subspace and any perturbation $\forall \delta^{(l)} \notin \mathcal{H}^{(l)}$, it has a unique decomposition $\delta^{(l)} = \delta_{\parallel}^{(l)} + \delta_{\perp}^{(l)}$ with the parallel component $\delta_{\parallel}^{(l)} = P\delta^{(l)} \in \mathcal{H}^{(l)}$ and the orthogonal component $\delta_{\perp}^{(l)} = (I - P)\delta^{(l)} \notin \mathcal{H}^{(l)}$. Then, $f^{(l+1)}(\delta_{\parallel}^{(l)}) = \mathbf{0}$ and $f^{(l+1)}(\delta^{(l)}) = f^{(l+1)}(\delta_{\perp}^{(l)})$.*

$P = U(U^\top U)^{-1}U^\top$ represents the projection matrix onto the harmless subspace $\mathcal{H}^{(l)} \subset \mathbb{R}^{n^{(l)}}$, and $U \in \mathbb{R}^{n^{(l)} \times dim(\mathcal{H}^{(l)})}$denotes a set of $dim(\mathcal{H}^{(l)})$ orthogonal bases for the subspace $\mathcal{H}^{(l)}$.

As a special case of Theorem 3, Theorem 4 demonstrates that the network output of a family of features/perturbations is equivalent to that of the component of this perturbation family, which is orthogonal to the subspace. Then, a collection of perturbations can be categorized as a perturbation family with identical impact on the network output, if their orthogonal components are the same, *i.e.*, $\delta_{\perp}^{(l)} = \hat{\delta}_{\perp}^{(l)}$, then $f^{(l+1)}(\delta^{(l)}) = f^{(l+1)}(\hat{\delta}^{(l)})$ (see Lemma 3 and Figure 6 in Appendix). We believe that the perturbation decomposition approach allows us to re-examine the properties of adversarial examples by decomposing the perturbations into their harmful and harmless counterparts.

Applications of Harmless Perturbations

Privacy Protection

We consider a scenario where users employ a pre-trained model on a third-party server to analyze data containing sensitive information (Schick et al. 2023; Shen et al. 2023; Wu et al. 2023; Liang et al. 2023). Specifically, either the third-party server or the user provides a pre-trained model, enabling the user to access the network parameters. Subsequently, the user locally generates privacy-preserving data using the parameters, and then deploys the protected data, along with the network, to the third-party server. To alleviate information leakage from sensitive data, harmless perturbations with sufficiently large magnitudes can be added to original samples. This process renders the generated samples unrecognizable to humans, effectively obscuring sensitive information within the images, without compromising network performance.

To be specific, our goal is to generate a visually unrecognizable image, denoted as $\hat{x} \in \mathbb{R}^{n_{\text{in}}}$, to substitute the original image x, ensuring that its network output is identical with that of the original image x. Specifically, given a DNN with a harmless perturbation subspace $\mathcal{H}^{(0)} \subset \mathbb{R}^{n_{\text{in}}}$ in its first linear layer, and a set of orthonormal bases $\{u_1, u_2, \cdots, u_d\}$ of the subspace $\mathcal{H}^{(0)}$ ($d = dim(\mathcal{H}^{(0)})$), visually unrecognizable harmless perturbations can simply be generated by maximizing the dissimilarity between the original image x and the generated image $\hat{x} := x + \sum_{i=1}^{d} c_i u_i, c_i \in \mathbb{R}, u_i \in \mathbb{R}^{n_{\text{in}}}$. Without loss of generality, we quantify the difference between the two images using the Mean Squared Error (MSE), *i.e.*, $\max_{\{c_1,c_2,\cdots,c_d\}} \frac{1}{n_{\text{in}}}\|\hat{x}-x\|_2^2$. To make the pixels of the generated image in the range $[0, 1]$, we add two penalties on the pixels out of bounds, *i.e.*, $\sum_i |\mathbb{1}(\hat{x}_i < 0) \cdot \hat{x}_i| + |\mathbb{1}(\hat{x}_i > 1) \cdot \hat{x}_i|$, as shown in Figure 1(a). Additional results on various datasets are presented in Appendix G. Algorithm 1 in Appendix also shows the pseudo-code of generating harmless perturbations.

Recovering Original Images. Reconstructing the original image x from the generated image $\hat{x}$ is a challenging task even if the attacker can access network parameters. Since the parameter matrix A uniquely determine the harmless perturbation subspace, it is equivalent to specifying the subspace $\mathcal{H}^{(0)}$. However, according to Theorem 3, the generated image $\hat{x} \notin \mathcal{H}^{(0)}$ can be decomposed into the sum of an *infinite* number harmless components $\hat{\delta} \in \mathcal{H}^{(0)}$ (Figure 1(b)) and reconstructed images $x^{\text{recon}} := \hat{x} - \hat{\delta}, \forall \hat{\delta} \in \mathcal{H}^{(0)}$. Therefore, the original image cannot be determined when the magnitude and direction of the harmless perturbation are unknown.

Visual Indistinguishability. To quantify the capability of the generated images in preserving privacy for human perception, we evaluated the perceptual similarity using two similarity metrics, *i.e.*, the Structural Similarity Index (SSIM) (Wang et al. 2004) and the Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018) metrics. Besides, we evaluated the degradation in classification accuracy of the generated images, compared to the original images. We compared the privacy-preserving capability of the generated harmless perturbations with the Gaussian noise $\mathcal{N}(0, 0.1^2)$ added to each pixel on the ResNet-50. Table 2 shows that the generated harmless perturbations achieved a similar level of privacy preservation as Gaussian noises, but harmless perturbations completely did not change the network outputs.

	Harmless perturbation	Gaussian noise
SSIM (↓)	**0.4719**	0.6825
LPIPS (↑)	0.2031	**0.3007**
Δaccuracy (↓)	**0.00%**	32.46%

Table 2: Perceptual similarity between the perturbed images and the original images on the CIFAR-10 dataset.

Model Fingerprint

Harmless perturbations also can be used for model fingerprints (Finlayson, Ren, and Swayamdipta 2024; Zeng et al. 2024) to faithfully reflect the model's changes, as they are determined by the parameter space of the DNN. We consider

[2]Note that the harmful space is not a linear subspace of $\mathbb{R}^{n^{(l)}}$, since it does not contain $\mathbf{0} \in \mathbb{R}^{n^{(l)}}$.

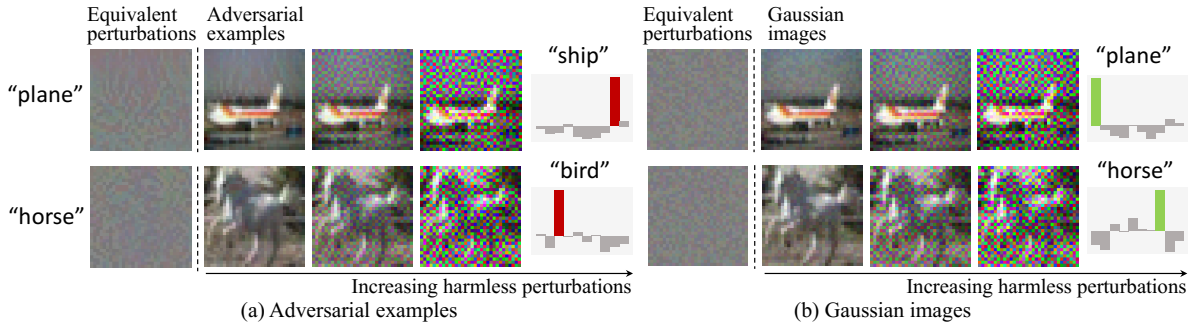


Figure 4: Adding Harmless perturbations to images with different noises ((a) adversarial perturbations (b) Gaussian noises) completely do not change the network output of the perturbed images, regardless of the magnitude of these harmless perturbations. Perturbed images incorporating harmless perturbations of arbitrary magnitude, drawn from the equivalent perturbation space, exhibit an effect on the network output that is equivalent to the impact of images with equivalent (orthogonal) perturbations.

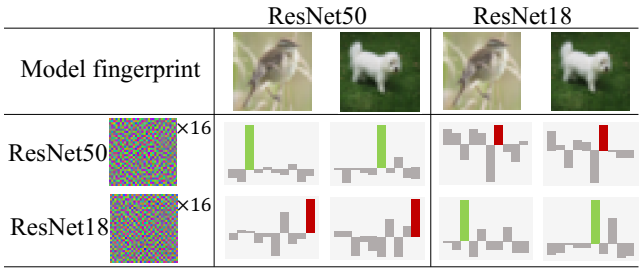


Figure 5: Harmless perturbations (magnified by 16x) can serve as identity fingerprint for models, allowing for tracking changes in closed-source models.

the simple case of establishing identity fingerprints for two DNNs. Figure 5 illustrates the network's output when adding two harmless perturbations extracted from two models on input images. Incorporating significant harmless perturbations generated by one model into various input samples preserves the outputs of that model, while applying them to input samples of another model leads to significant changes in the outputs of another model.

Transferability of Harmless Perturbations. Typically, given two DNNs with different parameters, their harmless perturbation spaces are not equal, *i.e.*, $\mathcal{P}_1 \neq \mathcal{P}_2$. However, there may exist few harmless perturbations that are transferable and serve as harmless perturbations for both DNNs, *i.e.*, $\delta \in \mathcal{P}_1 \cap \mathcal{P}_2$. To avoid choosing those rare transferable harmless perturbations as model fingerprints, we constraint the sampling of model fingerprints solely from the non-intersecting harmless perturbation spaces of the two DNNs, satisfying $\delta_1 \in \mathcal{P}_1 - (\mathcal{P}_1 \cap \mathcal{P}_2)$ and $\delta_2 \in \mathcal{P}_2 - (\mathcal{P}_1 \cap \mathcal{P}_2)$.

Intriguing Properties from Harmless Perturbations

Seeing Is Not Always Believing. Surprisingly, we find that distances within the feature space may exhibit considerable variation between DNNs and human perception. *DNNs tend to disregard the magnitude of features/perturbations.* DNNs are completely unaffected by harmless perturbations, highlighting a distinctive aspect of DNN robustness. Furthermore, harmless perturbations invalidate distance-based similarity metrics, such as the widely used Euclidean and cosine distances (Mensink et al. 2013; Zhang et al. 2018). For example, two vectors sampled from harmless/equivalent space, regardless of their magnitude or direction, may deemed dissimilar through these similarity metrics, yet deep networks still regard them as identical. Consequently, there arises a necessity to reassess whether these similarity metrics faithfully reflect the true modelling of similarity by deep networks.

Equivalent Adversarial Spaces. Theorems 3 and 4 demonstrate that *infinitely large, infinitely numerous* features/perturbations are equivalent to their components orthogonal to the harmless subspace. Therefore, for any perturbation, *there exist equivalent (adversarial) perturbation spaces*, ensuring equal attacking capabilities for perturbations. Compared to the Gaussian noises in Figure 4(b), the adversarial perturbations which have similar perturbation magnitudes in Figure 4(a), lead to completely different attack utilities. Interestingly, all adversarial perturbations for each sample in Figure 4(a) have the same attack utility, irrespective of their magnitudes. The equivalent adversarial spaces imply that: 1) *the perturbation magnitude is not a decisive factor attacking the network*, and 2) attention should be paid to the "effective components" of perturbations, *i.e.,* we can further decompose the perturbation in a more fine-grained way. We believe that further exploration of the equivalent space helps to understand the robustness of DNNs.

Conclusion

In this paper, we first show harmless perturbations, regardless of their magnitude, render the network output completely unaltered. For any linear layer where the input dimension n exceeds the output dimension m, there exists a continuous harmless subspace with a dimension of $(n - m)$. We further show that the perturbation space characterized by identical orthogonal components consistently affect the network output. Besides, the harmless perturbation space may expand when involving non-linear layers. Harmless perturbations can be used for hiding sensitive data and model fingerprints.

Acknowledgments

This work was supported in part by the National Key R&D Program of China under Grant 2022YFA1005000, the Start-up Research Fund of Southeast University (No. RF1028624178), the National Science and Technology Major Project (2021ZD0111602), the National Nature Science Foundation of China (92370115, 62276165), the National Natural Science Foundation of China (NSFC) under Grant 62402313. Jie Li was supported in part by the National Key R&D Program of China under Grant 2020YFB1710900 and Grant 2020YFB1806700, in part by NSFC under Grant 61932014 and Grant 62232011.

References

Athalye, A.; Carlini, N.; and Wagner, D. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *International Conference on Machine Learning (ICML)*.

Carlini, N.; and Wagner, D. 2017. Towards Evaluating the Robustness of Neural Networks. *IEEE Symposium on Security and Privacy (SP)*.

Clevert, D.-A.; Unterthiner, T.; and Hochreiter, S. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). *International Conference on Learning Representations (ICLR)*.

Croce, F.; and Hein, M. 2020. Reliable Evaluation of Adversarial Robustness with an Ensemble of Diverse Parameter-free Attacks. *International Conference on Machine Learning (ICML)*.

Fawzi, A.; Moosavi-Dezfooli, S.-M.; and Frossard, P. 2016. Robustness of classifiers: from adversarial to random noise. *Neural Information Processing Systems (NeurIPS)*.

Feng, X.; and Zhang, Z. 2007. The rank of a random matrix. *Applied mathematics and computation*, 185(1): 689–694.

Finlayson, M.; Ren, X.; and Swayamdipta, S. 2024. Logits of API-Protected LLMs Leak Proprietary Information. *arXiv preprint arXiv:2403.09539*.

Geirhos, R.; Rubisch, P.; Michaelis, C.; Bethge, M.; Wichmann, F. A.; and Brendel, W. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations (ICLR)*.

Gilmer, J.; Metz, L.; Faghri, F.; Schoenholz, S. S.; Raghu, M.; Wattenberg, M.; and Goodfellow, I. 2018. Adversarial Spheres. *Workshop of International Conference on Learning Representations (ICLR)*.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. *International Conference on Learning Representations (ICLR)*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ilyas, A.; Santurkar, S.; Tsipras, D.; Engstrom, L.; Tran, B.; and Madry, A. 2019. Adversarial Examples Are Not Bugs, They Are Features. *Neural Information Processing Systems (NeurIPS)*.

Ioffe, S.; and Szegedy, C. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *International Conference on Machine Learning (ICML)*.

Jacobsen, J.-H.; Behrmann, J.; Zemel, R.; and Bethge, M. 2019a. Excessive Invariance Causes Adversarial Vulnerability. *International Conference on Learning Representations (ICLR)*.

Jacobsen, J.-H.; Behrmann, J.; Zemel, R.; and Bethge, M. 2019b. Excessive invariance causes adversarial vulnerability. *International Conference on Learning Representations (ICLR)*.

Jetley, S.; Lord, N.; and Torr, P. 2018. With friends like these, who needs adversaries? *Advances in neural information processing systems*, 31.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-Normalizing Neural Networks. *Neural Information Processing Systems (NeurIPS)*.

Krizhevsky, A.; Hinton, G.; et al. 2009. Learning multiple layers of features from tiny images.

LeCun, Y.; Boser, B.; Denker, J.; Henderson, D.; Howard, R.; Hubbard, W.; and Jackel, L. 1990. Handwritten digit recognition with a back-propagation network. *Neural Information Processing Systems (NeurIPS)*.

LeCun, Y.; and Cortes, C. 2010. MNIST handwritten digit database.

Liang, Y.; Wu, C.; Song, T.; Wu, W.; Xia, Y.; Liu, Y.; Ou, Y.; Lu, S.; Ji, L.; Mao, S.; Wang, Y.; Shou, L.; Gong, M.; and Duan, N. 2023. TaskMatrix.AI: Completing Tasks by Connecting Foundation Models with Millions of APIs. *arXiv preprint arXiv:2303.16434*.

Maas, A. L.; Hannun, A. Y.; and Ng, A. Y. 2013. Rectifier Nonlinearities Improve Neural Network Acoustic Models. *International Conference on Machine Learning (ICML)*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. *International Conference on Learning Representations (ICLR)*.

Mensink, T.; Verbeek, J.; Perronnin, F.; and Csurka, G. 2013. Distance-Based Image Classification: Generalizing to New Classes at Near-Zero Cost. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 35(11): 2624–2637.

Moosavi-Dezfooli, S.-M.; Fawzi, A.; Fawzi, O.; and Frossard, P. 2017. Universal adversarial perturbations. *Computer Vision and Pattern Recognition (CVPR)*.

Nair, V.; and Hinton, G. E. 2010. Rectified Linear Units Improve Restricted Boltzmann Machines. *International Conference on Machine Learning (ICML)*.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng., A. Y. 2011. Reading digits in natural images with unsupervised feature learning. *In Neural Information Processing Systems (NeurIPS) Workshop on Deep Learning and Unsupervised Feature Learning*.

Nguyen, A.; Yosinski, J.; and Clune, J. 2015. Deep Neural Networks are Easily Fooled: High Confidence Predictions

for Unrecognizable Images. *Computer Vision and Pattern Recognition (CVPR)*.

Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z. B.; and Swami, A. 2017. Practical Black-Box Attacks against Machine Learning. *The ACM ASIA Conference on Computer and Communications Security (ACM ASIACCS)*.

Salman, H.; Ilyas, A.; Engstrom, L.; Vemprala, S.; Madry, A.; and Kapoor, A. 2021. Unadversarial Examples: Designing Objects for Robust Vision. *Neural Information Processing Systems (NeurIPS)*.

Scherer, D.; Muller, A.; and Behnke, S. 2010. Evaluation of pooling operations in convolutional architectures for object recognition. *International Conference on Artificial Neural Networks (ICANN)*.

Schick, T.; Dwivedi-Yu, J.; Dessì, R.; Raileanu, R.; Lomeli, M.; Zettlemoyer, L.; Cancedda, N.; and Scialom, T. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. *Neural Information Processing Systems (NeurIPS)*.

Shafahi, A.; Huang, W. R.; Studer, C.; Feizi, S.; and Goldstein, T. 2019. Are adversarial examples inevitable? *International Conference on Learning Representations (ICLR)*.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. *Neural Information Processing Systems (NeurIPS)*.

Si, W.; Li, S.; Park, S.; Lee, I.; and Bastani, O. 2023. Angelic Patches for Improving Third-Party Object Detector Performance. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Simonyan, K.; and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. *International Conference on Learning Representations (ICLR)*.

Tan, M.; and Le, Q. V. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *International Conference on Machine Learning (ICML)*.

Tao, L.; Feng, L.; Yi, J.; and Chen, S. 2022. With False Friends Like These, Who Can Notice Mistakes? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 8458–8466.

Tramèr, F.; Papernot, N.; Goodfellow, I.; Boneh, D.; and McDaniel, P. 2017. The Space of Transferable Adversarial Examples. *ArXiv preprint arXiv:1704.03453*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. *International Conference on Learning Representations (ICLR)*.

Wang, H.; Wu, X.; Huang, Z.; and Xing, E. P. 2020. High Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. *Computer Vision and Pattern Recognition (CVPR)*.

Wang, J.; Yin, Z.; Hu, P.; Liu, A.; Tao, R.; Qin, H.; Liu, X.; and Tao, D. 2022. Defensive Patches for Robust Recognition in the Physical World. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing (TIP)*.

Wu, C.; Yin, S.-K.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *arXiv preprint arXiv:2303.04671*.

Yin, D.; Lopes, R. G.; Shlens, J.; Cubuk, E. D.; and Justin, G. 2019. A Fourier Perspective on Model Robustness in Computer Vision. *Neural Information Processing Systems (NeurIPS)*.

Zeng, B.; Zhou, C.; Wang, X.; and Lin, Z. 2024. Human-Readable Fingerprint for Large Language Models. *arXiv preprint arXiv:2312.04828*.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. *International Conference on Machine Learning (ICML)*.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.