ON THE INTERACTION OF COMPRESSIBILITY AND ADVERSARIAL ROBUSTNESS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern neural networks are expected to simultaneously satisfy a host of desirable properties: accurate fitting to training data, generalization to unseen inputs, parameter and computational efficiency, and robustness to adversarial perturbations. While compressibility and robustness have each been studied extensively, a unified understanding of their interaction still remains elusive. In this work, we develop a principled framework to analyze how different forms of compressibility - such as neuron-level sparsity and spectral compressibility - affect adversarial robustness. We show that these forms of compression can induce a small number of highly sensitive directions in the representation space, which adversaries can exploit to construct effective perturbations. Our analysis yields a simple yet instructive robustness bound, revealing how neuron and spectral compressibility impact ℓ_∞ and ℓ_2 robustness via their effects on the learned representations. Crucially, the vulnerabilities we identify arise irrespective of how compression is achieved - whether via regularization, architectural bias, or implicit learning dynamics. Through empirical evaluations across synthetic and realistic tasks, we confirm our theoretical predictions, and further demonstrate that these vulnerabilities persist under adversarial training and transfer learning, and contribute to the emergence of universal adversarial perturbations. Our findings show a fundamental tension between structured compressibility and robustness and highlight new pathways for designing models that are both efficient and safe.

1 INTRODUCTION

Machine learning (ML) systems are increasingly deployed in high-stakes domains such as healthcare (Rajpurkar et al., 2022) and autonomous driving (Hussain & Zeadally, 2019), where reliability is paramount. With their growing social impact, modern neural networks are now expected to meet a suite of often conflicting demands: they must fit the data (explain observations), generalize to unseen inputs, remain efficient in storage and inference, *i.e.*, be compressible, and exhibit robustness against adversarial perturbations, as well as other distribution shifts. While each of these desiderata has been studied extensively in isolation, a mature and unified understanding of how they interact - and in particular, how compressibility shapes robustness - remains elusive.

As desirable as adversarial robustness and compressibility both are, the research has been equivocal regarding whether/when/how their simultaneous achievement is possible (Guo et al., 2018; Balda et al., 2020; Li et al., 2020a; Merkle et al., 2022; Liao et al., 2022; Piras et al., 2024). However, recent work has started to provide mechanism-based explanations for the relationship between the two, highlighting how compressibility impacts models' vulnerability to adversarial noise. For example, Savostianova et al. (2023) demonstrate that low-rank parameterizations may inadvertently amplify local Lipschitz constants, increasing sensitivity to perturbations. Nern et al. (2023) connect adversarial transferability to layer-wise operator norms and their impact on representation geometry. Feng et al. (2025) further shows that while moderate sparsity can enhance robustness, excessive sparsity causes ill-conditioning that reintroduces fragility and vulnerability. These results hint at a delicate, regime-dependent relationship between compressibility and robustness - but a principled and general framework is still lacking.

In this work, we develop a framework to investigate the effect of structured sparsity on adversarial robustness through its effect on parameter operator norms and network's Lipschitz constant. We

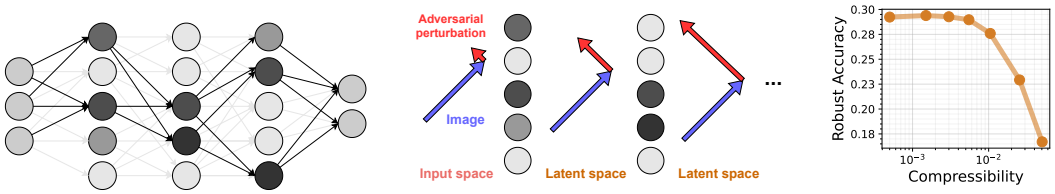

Figure 1: A visual preview of our findings. (Left) Sparsification expedites compression but creates sensitive latent directions. (Center) Adversaries exploit these sensitive directions to increase their potency. (Right) This leads to decreased adversarial robustness.

jointly study how different forms of compressibility - particularly neuron-level sparsity and spectral compression - affect adversarial robustness. Our central result is an intuitive and instructive adversarial robustness bound that reveals how compressibility can induce a small set of highly sensitive directions in the representation space. These "adversarial directions" dramatically amplify perturbations and are readily exploited by adversaries. Empirically, we confirm that these axes are not merely theoretical constructs: adversarial attacks reliably identify and exploit them across architectures, datasets, and attack models. Figure 1 provides a visual preview of our findings. Previous research tightly links compressibility to generalization (Arora et al., 2018; Barsbey et al., 2021); however, our findings imply that the very mechanisms that promote generalization can also introduce structural weaknesses. In summary, our contributions are:

1. We provide an **adversarial robustness bound** that decomposes into analytically interpretable terms, and predicts that neuron and spectral compressibility create adversarial vulnerability against ℓ_∞ and ℓ_2 attacks, through their effects on networks' Lipschitz constants.
2. Utilizing various compressibility-inducing interventions, we empirically validate our predictions regarding the **emergence of adversarial vulnerability under structured compressibility** with various datasets and models, including commonly used modern encoder architectures.
3. We demonstrate that the **detrimental effects of compressibility persist under adversarial training and transfer learning**, and contribute to the appearance of universal adversarial examples.
4. We demonstrate and discuss our findings' implications for compression in practice, and highlight promising paths for **designing models that reconcile efficiency and safety**.

We will make our implementation publicly available upon publication.

2 Setup

Notation. We denote scalars by lower case italic (k), vectors with lower case bold ($\boldsymbol{x}$), and matrices with upper case bold ($\mathbf{W}$) characters respectively. Vector ℓ_p norms are denoted by $\|\boldsymbol{x}\|_p$. For matrices, $\|\mathbf{W}\|_F, \|\mathbf{W}\|_2, \|\mathbf{W}\|_\infty$ correspond to Frobenius, spectral, and ℓ_∞ operator norms, respectively. We denote the i^{th} element of a vector $\boldsymbol{x}$ with x_i, and row i of a matrix $\mathbf{W}$ with $\mathbf{w}_i$. Elements of a sequence of matrices (e.g. layer matrices) are referred to by $\mathbf{W}^l, l \in [\lambda]$. For an integer n, we use $[n] := (1, \ldots, n)$.

Unless otherwise specified, we will be focusing on supervised classification problems, which will involve the input $\boldsymbol{x} \in \mathcal{X}$ and label $y \in \mathcal{Y}$. A predictor $g : \mathcal{X} \to \mathbb{R}^{|\mathcal{Y}|}$, parametrized by $\boldsymbol{\theta} \in \Theta$ produces output logits $\boldsymbol{s} = g(\boldsymbol{x}, \boldsymbol{\theta})$, the maximum of which is the predicted label $\hat{y} = \arg\max_{i \in |\mathcal{Y}|} s_i$. Predictions are evaluated by a loss function $\ell : \mathbb{R}^{|\mathcal{Y}|} \times \mathcal{Y} \to \mathbb{R}_+$. For brevity, we define the composite loss function $f(\boldsymbol{x}, \boldsymbol{\theta}) := \ell(g(\boldsymbol{x}, \boldsymbol{\theta}), y)$.

Risk and adversarial robustness. Assuming a data distribution π on $\mathcal{X} \times \mathcal{Y}$, we define the population and empirical risks accordingly: $F(\boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{x},y\sim\pi}[f(\boldsymbol{x}, \boldsymbol{\theta})]$, and $\widehat{F}(\boldsymbol{\theta}, S) := \frac{1}{n}\sum_{i=1}^n f(\boldsymbol{x}_i, \boldsymbol{\theta})$, where $(\boldsymbol{x}_i, y_i)_{i=1}^n$ denotes a set of i.i.d. samples from π. Adversarial attacks are minimal perturbations to input that dramatically disrupt a model's predictions (Szegedy et al., 2014). In this paper, we focus on bounded p-norm attacks, which we define as

$$\boldsymbol{a}^* = \arg\max_{\|\boldsymbol{a}\|_p \leq \delta} f(\boldsymbol{x} + \boldsymbol{a}, \boldsymbol{\theta}). \tag{1}$$

Given the adversarial loss $f_p^{\text{adv}}(\boldsymbol{x}, \boldsymbol{\theta}; \delta) := f(\boldsymbol{x} + \boldsymbol{a}^*, \boldsymbol{\theta})$, we define adversarial risk and empirical adversarial risk as $F_p^{\text{adv}}(\boldsymbol{\theta}; \delta) := \mathbb{E}_{\boldsymbol{x}\sim\pi}[f_p^{\text{adv}}(\boldsymbol{x}, \boldsymbol{\theta}; \delta)]$ and $\widehat{F}_p^{\text{adv}}(\boldsymbol{\theta}, S; \delta) := \frac{1}{n}\sum_{i=1}^n f_p^{\text{adv}}(\boldsymbol{x}_i, \boldsymbol{\theta}; \delta)$,

respectively. The type of the selected *attack norm* p for the *attack budget* δ, determines the type of adversarial attack in question, with $p = 2$ and $p = \infty$ as the most common choices. In this paper, we are primarily interested in what we call the *adversarial robustness gap*: $\Delta_p^{\text{adv}} := F_p^{\text{adv}}(\boldsymbol{\theta}, \delta) - F(\boldsymbol{\theta})$. A model with small Δ_p^{adv} is considered *adversarially robust*.

Neural networks. Our analyses will focus on neural networks under classification. We define a fully connected neural network (FCN) with λ hidden layers of h units as below:

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = \mathbf{C}\phi(\mathbf{W}^{\lambda}\phi(\dots \mathbf{W}^{1}\boldsymbol{x})), \tag{2}$$

where $\boldsymbol{\theta} := (\mathbf{C}, \mathbf{W}^1, \dots, \mathbf{W}^\lambda)$, ϕ is elementwise ReLU activation function. We can write g as the composition of two functions, a linear classifier head $c : \mathbb{R}^h \to \mathbb{R}^{|\mathcal{Y}|}$, and a feature encoder $\Phi : \mathcal{X} \to \mathbb{R}^h$, such that $g(\boldsymbol{x}, \boldsymbol{\theta}) := c(\cdot, \mathbf{C}) \circ \Phi(\cdot, \mathbf{W}^1 \dots \mathbf{W}^\lambda)(\boldsymbol{x})$. To avoid notational clutter and without loss of generality, throughout our analyses we assume that $\boldsymbol{x} \in \mathbb{R}^h$, and omit bias parameters.

Lipschitz continuity. Given two L^p spaces $\mathcal{X}$ and $\mathcal{Y}$, a function $g : \mathcal{X} \to \mathcal{Y}$ is called Lipschitz continuous if there exists a constant K_p such that $\|g(\boldsymbol{x}^1) - g(\boldsymbol{x}^2)\|_p \leq K_p \|\boldsymbol{x}^1 - \boldsymbol{x}^2\|_p, \forall\, \boldsymbol{x}^1, \boldsymbol{x}^2 \in \mathcal{X}$. Said K_p is called the (global) Lipschitz constant. Any $\bar{K}_p$ that is valid for a subspace $\mathcal{U} \subset \mathcal{X}$ is called a local Lipschitz constant. Although its computation is NP-hard for even the simplest neural networks (Scaman & Virmaux, 2018); as a notion of input-based volatility, estimation, utilization, and regularization of the Lipschitz constant have been a staple of robustness research (Cisse et al., 2017; Bubeck et al., 2020; Muthukumar & Sulam, 2023; Grishina et al., 2025). Note that the FCN as defined in Eq (2) is Lipschitz continuous in ℓ_p for $p \in [1, \infty]$, along with other commonly used architectures such as convolutional neural networks (CNN) (Zühlke & Kudenko, 2025).

Compressibility. Various prominent approaches to neural network compression exist, such as pruning, quantization, distillation, and conditional computing, (O'Neill, 2020). Here we focus on pruning and low-rank approximation, two of the most commonly used and researched forms of compression (Hohman et al., 2024). More specifically, we focus on inherent properties of network parameters that make them amenable to pruning or low-rank approximation, i.e. their *compressibility*. We will first present a formal definition of a *compressible* vector, and then will show how this definition can be utilized to describe both structured prunability and low-rankness.

Definition 2.1 ((q, k, ϵ)-compressibility)**.** *Given a vector $\boldsymbol{\theta} \in \mathbb{R}^d$ and a non-negative integer $k \leq d$, let $\boldsymbol{\theta}_k$ denote the compressed vector which contains the largest (in magnitude) k elements of $\boldsymbol{\theta}$ with all the other elements set to 0. Then, $\boldsymbol{\theta}$ is (q, k, ϵ)-compressible if and only if*

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_k\|_q \,/\, \|\boldsymbol{\theta}\|_q \;\leq\; \epsilon. \tag{3}$$

In the case of equality, we call $\boldsymbol{\theta}$ to be strictly (q, k, ϵ)-compressible. Complementarily, the spread variable $\beta \in [0, 1]$ can be used to characterize the dispersion of top-k terms, such that $|\boldsymbol{\theta}_{m_k}| = (1 - \beta)|\boldsymbol{\theta}_{m_1}|$, where m_i indexes the i'th largest magnitude elements in the vector.

Moving forward we will assume any vector denoted as compressible is strictly compressible, unless otherwise noted. See the Appendix for a more in-depth discussion of our compressibility definition and how it relates to other notions of approximate sparsity.

Structured compressibility. Importantly, given that the $\boldsymbol{\theta}$ can be any vector, the above definition can be used flexibly to describe different notions of compressibility, including those of structured compressibility, where particular substructures in the model dominate the rest. More specifically, given a layer parameter matrix $\mathbf{W} \in \mathbb{R}^{h \times h}$ from Eq (2), let $\boldsymbol{\nu} := (\|\mathbf{w}_1\|_1, \dots, \|\mathbf{w}_h\|_1)$ denote ℓ_1 norms of rows of the matrix $\mathbf{W}$. The compressibility of $\boldsymbol{\nu}$ would correspond to *row/neuron compressibility*, which is a desirable property for neural network parameters as it expedites pruning of whole neurons, with tangible computational gains. Note that this also would correspond to filter compressibility/pruning in CNNs with a matricization of the convolution tensor. Similarly, let $\boldsymbol{\sigma} := (\sigma_1, \sigma_2, \dots)$ denote the singular values of matrix $\mathbf{W}$. Compressibility of $\boldsymbol{\sigma}$ would correspond to *spectral compressibility*, closely related to the notion of approximate/numerical low-rankness.

3 Norm-based adversarial robustness bounds

Motivating hypothesis. Our analysis relies on a simple intuition: Although structured (neuron, spectral) compressibility is desirable from a computational perspective, it also focuses the total energy of the parameter on a few dominant terms (rows/filters, singular values).

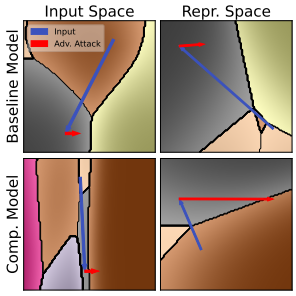

Figure 2: Decision boundaries under compressibility.

This in turn creates a few, potent directions in the latent space and increases the operator norms of the parameters (ℓ_∞, ℓ_2 operator norms respectively). This increases their sensitivity to worst-case perturbations: adversarial attacks exploiting these directions are amplified in the representation space, and can more easily disrupt the predictions of the model. Taken from an experiment presented in full detail in Section 4, Figure 2 visualizes the input image, adversarial perturbation, and decision boundaries for a single sample under a baseline vs. compressible (low-rank) model, obtained through a PCA of the representations. The top row visualizes the baseline model, where the minuscule adversarial perturbation fails to move the perturbed image across class boundaries. The bottom row however, illustrates the compressible model under attack. Here, although attack budget is identical in the input space, the adversarial perturbation is dramatically amplified in the representation space, leading to a successful adversarial attack. Note that the decision boundaries in compressible model's input space is much more contracted to reflect this vulnerability.

Compressibility-based Lipschitz bounds. Our theory will relate structured compressibility to robustness through its effect on the network's operator norms and Lipschitz constants. However, this brings about a particular conceptual challenge. Our notion of (q, k, ϵ)-compressibility, like others' (Diao et al., 2023), is a *scale-independent* measure. Therefore, any direct relation between compressibility and Lipschitz constants would be rendered void by the arbitrary scaling of the parameters. Therefore, we characterize ℓ_∞ and ℓ_2 operator norms of the parameters by an upper bound that decomposes into (compressibility $\times$ Frobenius norm) terms. This "structure vs. scale" decomposition allows us to meaningfully relate compressibility and robustness, and also allows us to develop concrete hypotheses regarding the effect of various interventions in neural network training.

Theorem 3.1. *The following statements relate operator norms and structured compresibility.*

(a) Neuron compressibility (i.e. row-sparsity): *Let $\mathbf{w}_i, i \in [h]$ denote the rows of the matrix $\mathbf{W}$, and let $\boldsymbol{\nu} := (\|\mathbf{w}_1\|_1, \dots, \|\mathbf{w}_h\|_1)$ denote ℓ_1 norms of its rows. Assuming $\boldsymbol{\nu}$ is $(1, k_{\boldsymbol{\nu}}, \epsilon_{\boldsymbol{\nu}})$ compressible and each row $\mathbf{w}_i$ is $(2, k_r, \epsilon_r)$ compressible implies:*

$$\|\mathbf{W}\|_\infty \leq \frac{(1-\epsilon_{\boldsymbol{\nu}})}{(1-\beta_{\boldsymbol{\nu}})} \left(\frac{\sqrt{hk_r} + h\epsilon_r}{k_{\boldsymbol{\nu}}} \right) \|\mathbf{W}\|_F. \tag{4}$$

(b) Spectral compressibility (i.e. low-rankness): *Let $\boldsymbol{\sigma} := (\sigma_1, \sigma_2, \dots)$ denote the singular values of matrix $\mathbf{W}$. Assuming $\boldsymbol{\sigma}$ is $(1, k_{\boldsymbol{\sigma}}, \epsilon_{\boldsymbol{\sigma}})$ compressible implies:*

$$\|\mathbf{W}\|_2 \leq \frac{(1-\epsilon_{\boldsymbol{\sigma}})}{(1-\beta_{\boldsymbol{\sigma}})} \left(\frac{\sqrt{h}}{k_{\boldsymbol{\sigma}}} \right) \|\mathbf{W}\|_F. \tag{5}$$

Intuitively, Theorem 3.1 describes how increasing compressibility affects layer operator norms: Neuron compressibility, *i.e.* a small number of rows dominating the matrix increases ℓ_∞ operator norm of the matrix, especially if the spread within these dominant rows are high. Similarly, increased spectral compressibility and spread increases the ℓ_2 operator norm. Note that the latter result is closely related to results from the literature that connect stable rank or condition number to robustness (Savostianova et al., 2023; Feng et al., 2025). We highlight that although Theorem 3.1 directly relates neuron and spectral compressibility to perturbations defined in ℓ_∞ and ℓ_2 norms, we do not claim that relationships across attack and operator norms do not hold. Indeed in our Appendix, we show that the two operator norms are likely to move together under compressibility, connecting structured compressibility to a broader notion of adversarial vulnerability. Lastly, while we utilize the upper bounds for our following theoretical results, additional theoretical results in the Appendix characterize lower bounds on the operator norm with similar implications.

As we move on to characterizing layers within a neural network, $\mathbf{W}_k^l$ will be used to denote the *compressed* version of the parameter matrix of layer l. In the case of row compression, this will correspond to keeping the k dominant rows as is, and setting the $h - k$ trailing rows to $\mathbf{0}$. In the case of spectral compression, given the singular value decomposition (SVD), $\mathbf{W}^l = \mathbf{U}^l \mathbf{\Sigma}^l \mathbf{V}^{l^T}$, the compressed matrix would correspond to $\mathbf{W}_k^l := \mathbf{U}_k^l \mathbf{\Sigma}_k^l \mathbf{V}_k^{l^T}$, where the $h-k$ smallest singular values are truncated.

Note that the sensitivity of the network not only relies on the characteristics of layer parameters, but also on the interactions between them. As an informative extreme case, assume that layer $\mathbf{W}^l$ greatly amplifies the input in the direction $\mathbf{u}_1$, due to spectral compressibility producing a large σ_1. Ignoring nonlinearities for now, if $\mathbf{u}_1$ is in the null space of $\mathbf{W}^{l+1}$, this amplification will have no effect on the sensitivity of the overall network. Thus, potent attack directions in the network are determined not only through layers' inherent properties, but how well the dominant directions in consecutive layers "align", in consideration with the nonlinearities between them. We will characterize this crucial interaction with the *interlayer alignment terms* A^*_∞ and A^*_2. With $\mathcal{D}$ as the set of all diagonal binary matrices, standing for all possible ReLU activation patterns, these are defined as:

$$A^*_\infty(\mathbf{W}^{l+1}_k, \mathbf{W}^l_k) \triangleq \max_{\mathbf{D}\in\mathcal{D}} \frac{\|\mathbf{W}^{l+1}_k \mathbf{D} \mathbf{W}^l_k\|_\infty}{\|\mathbf{W}^{l+1}\|_\infty \|\mathbf{W}^l\|_\infty} + R_\infty(\epsilon) \tag{6}$$

$$A^*_2(\mathbf{W}^{l+1}_k, \mathbf{W}^l_k) \triangleq \max_{\mathbf{D}\in\mathcal{D}} \frac{\|\sqrt{\Sigma^{l+1}_k}\mathbf{V}^{l+1^T}_k \mathbf{D} \mathbf{U}^l_k \sqrt{\Sigma^l_k}\|_2}{\sqrt{\|\mathbf{W}^{l+1}\|_2 \|\mathbf{W}^l\|_2}} + R_2(\epsilon), \tag{7}$$

where $R_\infty(\epsilon) := \nu^l_{k+1}/\nu^l_1 + \nu^{l+1}_{k+1}/\nu^{l+1}_1 + \nu^l_{k+1}\nu^{l+1}_{k+1}/\nu^l_1\nu^{l+1}_1$ is a remainder alignment term and likewise, $R_2(\epsilon) := \sqrt{\sigma^l_k/\sigma^l_1} + \sqrt{\sigma^{l+1}_{k+1}/\sigma^{l+1}_1} + \sqrt{\sigma^l_{k+1}\sigma^{l+1}_{k+1}/\sigma^l_1\sigma^{l+1}_1}$. In the Appendix, we show that for $p \in \{2, \infty\}$, $R_p(\epsilon) \to 0$ as $\epsilon \to 0$. There, we also show that for all layers $A^*_p \leq 1$; alignment terms can therefore be interpreted to act as a normalized function that corrects the worst-case bound based on the dominant terms' misalignment. Next theorem will use Theorem 3.1 and Eq. (6) and (7) to provide an upper bound to the Lipschitz constant of the network.

Theorem 3.2. *Let L^p_Φ be the Lipschitz constant of the encoder Φ defined following Eq* (2)*. Let $\mathcal{D}$ denote the set of all diagonal binary matrices, corresponding to ReLU activation layers. Then:*

(a) Row/neuron compressibility: *The ℓ_∞ Lipschitz constant of Φ can be upper bounded by:*

$$L^\infty_\Phi \leq \hat{L}^\infty_\Phi := \prod_{l=1}^{\lambda} \frac{(1-\epsilon_{\boldsymbol{\nu}})}{(1-\beta_{\boldsymbol{\nu}})} \left(\frac{\sqrt{hk_r} + h\epsilon_r}{k_{\boldsymbol{\nu}}}\right) \|\mathbf{W}\|_F \prod_{l=1}^{\lambda-1} \tilde{A}^*_\infty(\mathbf{W}^{\{l+1\}}_k, \mathbf{W}^l_k), \tag{8}$$

where $\tilde{A}^*_\infty(\mathbf{W}^{\{l+1\}}_k, \mathbf{W}^l_k) = A^*_\infty(\mathbf{W}^{\{l+1\}}_k, \mathbf{W}^l_k)$ if $l \in S_{opt}$, and 1 otherwise. $S_{opt} \subseteq \{1, 2, \ldots, L-1\}$ is the optimal alignment partition set (See Definition A.4) that can be determined in $O(\lambda)$ time.

(b) Spectral compressibility: *The ℓ_2 Lipschitz constant of Φ can be upper bounded by:*

$$L^2_\Phi \leq \hat{L}^2_\Phi := \prod_{l=1}^{\lambda} \frac{(1-\epsilon_{\boldsymbol{\sigma}})}{(1-\beta_{\boldsymbol{\sigma}})} \left(\frac{\sqrt{h}}{k_{\boldsymbol{\sigma}}}\right) \|\mathbf{W}\|_F \prod_{l=1}^{\lambda-1} A^*_2(\mathbf{W}^{\{l+1\}}_k, \mathbf{W}^l_k). \tag{9}$$

We note that this upper bound can be directly used in conjunction with other results from the literature (Ribeiro et al., 2023) to characterize adversarial robustness gap:

Corollary 3.3. *Under a binary classification task with cross-entropy loss, $\ell(y, \boldsymbol{x}^\top\boldsymbol{\theta}) = \ell(y, \hat{y}) = \log\left(1 + e^{-y\hat{y}}\right)$, given a neural network classifier as described in* (2)*, under the same assumptions with* (8)*, $F^{\text{adv}}_\infty(\boldsymbol{\theta}; \delta) \leq F(\boldsymbol{\theta}) + \delta\hat{L}^\infty_\Phi\|\boldsymbol{\theta}\|_1$. Similarly, under the assumptions of $F^{\text{adv}}_2(\boldsymbol{\theta}; \delta) \leq F(\boldsymbol{\theta}) + \delta\hat{L}^2_\Phi\|\boldsymbol{\theta}\|_2$.*

Note that although bounds provided in Theorem 3.2 are tighter than the pessimistic "product-of-norms" bounds, it deliberately *trades off* some tightness by utilizing Theorem 3.1. However, in return, this results in a bound that decomposes into analytically interpretable and actionable terms. Such bounds have proven valuable in analyzing adversarial robustness in deep learning (Wen et al., 2020). Regardless, Figure 3 demonstrates the close correlation our bound shows with the empirical robustness gap ($\rho = 0.947$), in a 2-hidden-layer neural network with varying spectral compressibility (layer rank). We provide full details in the Appendix, where we also show that as the global Lipschitz constant increases, empirically estimated local Lipschitz constants scale accordingly. There, we also explore the alignment terms' empirical behavior and estimation techniques, although a detailed analysis thereof lies beyond our primary focus. We now translate these theoretical insights into concrete hypotheses and test them through experiments.

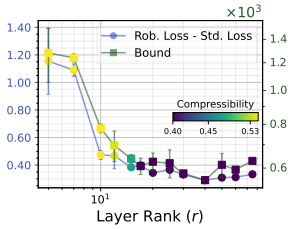

Figure 3: Theorem 3.2 vs. empirical robustness gap.

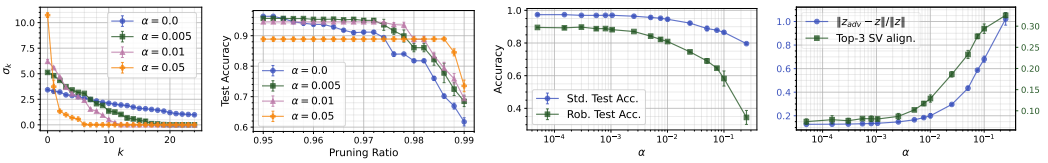

Figure 4: Model statistics under increasing strength of nuclear norm regularization (α).

4 Experimental Evaluation

We now validate our theoretical findings through specific experimentation. We first validate our *motivating hypothesis* and then empirically show that (i) neuron and spectral compressibility-inducing interventions will reduce adversarial robustness against ℓ_∞ and ℓ_2 adversarial attacks; (ii) the negative effects of compressibility to persist under adversarial training, (iii) the compressibility-related vulnerabilities being baked into the learned representations during pretraining, will impact any downstream task in transfer learning; (iv) increasing compressibility creates vulnerable directions in the latent space, further enabling universal adversarial examples (UAEs), while increasing Frobenius norm will create vulnerability without leading to UAEs; and (v) compressed models will inherit the vulnerability of the original models, and conducting compression based on (q, k, ϵ)-compressibility and reducing the spread of the dominant terms will improve robustness.

Datasets, architectures, and training. We conduct our experiments in the most commonly used datasets and architectures in the literature on adversarial robustness under pruning (Piras et al., 2024). Datasets we use include MNIST (Deng, 2012), CIFAR-10, CIFAR-100 (Krizhevsky & Hinton, 2009), SVHN (Netzer et al., 2011), Flickr30k (Young et al., 2014), and ImageNet-1k (Deng et al., 2009). Architectures we utilize include fully connected networks (FCN), ResNet18 (He et al., 2016), VGG16 (Simonyan & Zisserman, 2014), WideResNet-101-2 (Zagoruyko & Komodakis, 2016), vision transformer (ViT) - both as a standalone classifier (Dosovitskiy et al., 2021) and as part of a CLIP encoder (Radford et al., 2021), and Swin Transformer (Liu et al., 2021). Unless otherwise noted, we use softmax cross-entropy loss, the AdamW optimizer with a weight decay of 0.01, a learning rate of 0.001, and use a validation set based model selection for early stopping.

Evaluating and training for adversarial robustness. When evaluating adversarial robustness, we utilize AutoPGD as the primary adversarial attack algorithm for evaluation (Croce & Hein, 2020), through its implementation by Nicolae et al. (2018). When training for adversarial robustness, we utilize a PGD attack to generate adversarial samples at every iteration (Madry et al., 2018). Unless otherwise noted, we use a ratio of 0.5 for adversarial samples in a training minibatch. We use $\epsilon = 8/255$ and $\epsilon = 0.5$ for ℓ_∞ and ℓ_2 attacks respectively for end-to-end adversarially trained models. We use $0.25\times$ of these budgets for standard trained or adversarially fine-tuned models to allow a visible comparison (See Appendix for qualitatively identical results under different budgets). By default, we present results for ℓ_∞ and ℓ_2 attacks when evaluating robustness under neuron and spectral compressibility respectively, and defer the cross-norm results to the supplementary material, which also includes further details on our experiment settings and implementation.

4.1 Results

Testing the motivating hypothesis. We start our empirical analysis with a demonstrative experiment to visually investigate the implications of our initial hypothesis. For this, we train a single 400-width hidden layer FCN with ReLU activations on the MNIST dataset. We use nuclear norm regularization (NNR) to encourage singular value (SV) compressibility, adding the term $\alpha\|\boldsymbol{\sigma}\|_1$ to the training objective, with α as a hyperparameter. To avoid confounding by NNR decreasing overall parameter norms, we apply Frobenius norm normalization to $\mathbf{W}^1$ at every iteration (Miyato et al., 2018).

In Figure 4 (left) we validate that our intervention indeed increases spectral norm compressibility. As expected, Figure 4 (center left) shows that SV compressibility actually allows pruning: the more compressible models retain their performance under stronger pruning. Figure 4 (center right) shows that increased compressibility comes at the cost of adversarial robustness: as α increases, adversarial accuracy dramatically falls. We further investigate whether this fall is due to our hypothesized mechanism. Let $\boldsymbol{z} = \Phi(\boldsymbol{x})$ and $\boldsymbol{z}_{\text{adv}} = \Phi(\boldsymbol{x} + \boldsymbol{a}^*)$ denote the learned representations of clean and perturbed input images. If the adversarial attacks are taking advantage of the potent directions created by compressibility, then as compressibility increases: (1) The perturbations $\boldsymbol{a}^*$ should align more

with the dominant singular directions, *i.e.*, $\mathbf{v}_i^\intercal \boldsymbol{a}^* \gg \mathbf{v}_j^\intercal \boldsymbol{a}^* \; \forall i \in [k], j \notin [k]$, (2) representations of adversarial perturbations should grow stronger in relation to the original image's representation, *i.e.* $\|\boldsymbol{z}_{\text{adv}} - \boldsymbol{z}\|_2 / \|\boldsymbol{z}\|_2$ should increase. Figure 4 (right) confirms both predictions. Lastly, the previously presented Figure 2 visualizes the effect of compressibility in the input and representation space.

Adversarial robustness and compressibility under standard training. For implications of our analysis under more realistic settings, we start by investigating the effects of compressibility on adversarial robustness in fully connected networks (FCN). We induce neuron and spectral compressibility through group lasso regularization[1] and low-rank factorization, respectively (latter avoids the excessive cost of nuclear norm regularization). As above, we conduct Frobenius norm normalization at every iteration. Figure 5 (top) presents the results of these experiments: The reduction in adversarial robustness as a function of increasing compressibility is clear in both cases, confirming our main hypothesis. Note that we present robust accuracy / standard accuracy ratio alongside robust accuracy to highlight that the obtained results are not due to baseline standard accuracy being lower under compressibility.

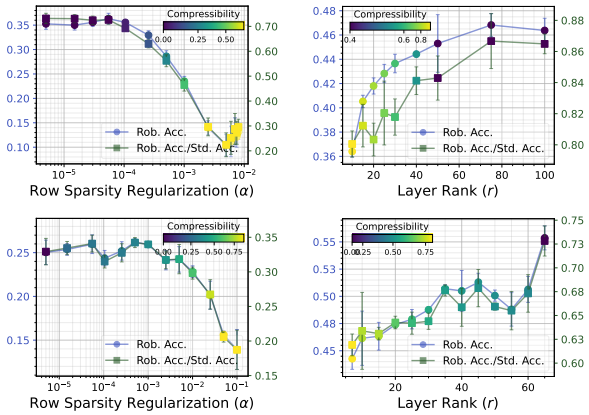

Figure 5: Results with FCN (top) and ResNet18 (bottom) trained on CIFAR-10 dataset.

We then investigate whether our hypotheses apply beyond the context of our theory, starting with convolutional neural networks (CNNs). We first test our predictions in ResNet18 models trained on CIFAR-10 datasets. Here we eschew Frobenius norm normalization for standard weight decay. However, to prevent confounding from group lasso's effect on general parameter scales, we create a scale-invariant version that regularizes row norms' ℓ_1/ℓ_2 norm ratio.[2] Figure 5 (bottom) demonstrates that the effects described above clearly translate to this setting as well, further solidifying the relationship between structured compressibility and adversarial robustness. We present similar results on two other architectures (VGG16, WideResNet-101) and two other datasets (CIFAR-100, SVHN) in the Appendix. Going forward, for brevity we will focus on neuron compressibility results, and defer corresponding spectral compressibility results to the Appendix, where we also discuss unstructured compressibility and inductive-bias based emergent compressibility.

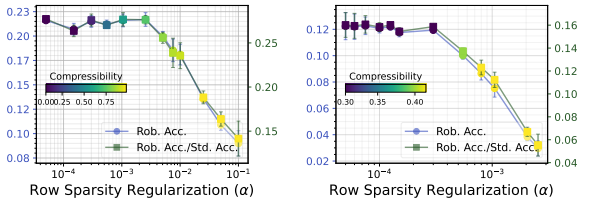

Figure 6: Results with ViT (left) and CLIP (right).

Experiments with transformers. We next test our hypotheses under transformer architectures. Figure 6 (left) replicates our results under a ViT classifier model trained on CIFAR-10 dataset. Further, to test whether our hypothesis holds under a zero-shot classification setting, we fine-tune a pre-trained CLIP model on Flickr30k dataset under varying degrees of sparsification regularization, and conduct standard and adversarial zero-shot classification using ImageNet-1k dataset. We find that our results (Figure 6, right) replicate here as well. That simply fine-tuning with sparsification can create this vulnerability with commonly repurposed encoder backbones highlights the safety implications of our results. See Appendix for further details and findings under other training settings.

Effects of compressibility on robustness under adversarial training. Given that adversarial training is the primary method for obtaining models that are robust against adversaries, we next investigate whether the effects we have observed will persist under this regime. To make this setting

[1] Group lasso regularization penalizes the ℓ_1 norm of row ℓ_2 norms of each layer, promoting row-sparsity.

[2] In the Appendix, we show that standard group lasso creates a "tug-of-war" between increasing compressibility and decreasing parameter scales; the former eventually wins, resulting in decreased robustness.

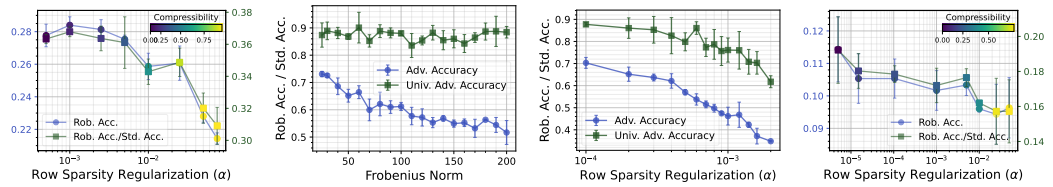

Figure 7: (Left) Effects of compressibility under adversarial training. UAEs under increasing (center left) compressibility vs. (center right) parameter scale. (Right) Robustness under transfer learning.

as close to practice as possible, we also include a learning rate annealing schedule (Cosine annealing) and basic data augmentation (random horizontal flip and crops). The results almost identically replicate our observations under standard training (Figure 7, left). Although adversarial training increases adversarial robustness overall, the relative effect of compressibility remains as it is.

Universal adversarial examples. Examining the terms in Theorem 3.2, we predict that while both compressibility and Frobenius norm are likely to increase vulnerability, only the former is likely to lead to universal adversarial examples (UAEs) (Moosavi-Dezfooli et al., 2017), due to the global vulnerable directions it creates. To test our hypothesis, we modify the setting of FCN experiments presented above: As an alternative to increasing row sparsity regularization under a fixed Frobenius norm, in an alternative set of experiments we systematically increase the constant to which Frobenius norm of the layers is fixed, without any row sparsity regularization. We utilize a FGSM-based (Goodfellow et al., 2015) UAE computation to develop adversarial samples. Figure 7 (center left, center right) confirms our hypothesis: while increasing Frobenius norm only decreases standard adversarial robustness, increasing compressibility additionally creates vulnerability to UAEs.

Adversarial vulnerability under transfer learning. Next, we investigate our hypothesis that the effects of compressibility should persist under transfer learning due to the structural effects created on representations. We train a ResNet18 model on CIFAR-100 dataset with increasing row sparsity regularization. After the training is complete, we train a linear classifier head for prediction on CIFAR-10 dataset and evaluate the robustness of the resulting model. Figure 7 (right) shows that the effects of compressibility observed above directly translate to the context of transfer learning, where increased compressibility in pretraining affects robustness performance in the downstream task, for which the network is fine-tuned.

Compression and robustness. We now investigate the behavior of models under layerwise filter pruning. Using the ResNet18 and CIFAR-10 combination under adversarial training, in Figure 8 (left), we compare the baseline model ($\alpha = 0.0$) to a model regularized to be compressible ($\alpha = 0.1$). We see that at no point the compressed model surpasses the baseline model's uncompressed performance in terms of standard and robust accuracy. However, as pruning ratio increases, the baseline model fails to retain its standard and robust performance, whereas the compressible model does considerably better, demonstrating the fundamental tension between robustness and compressibility.

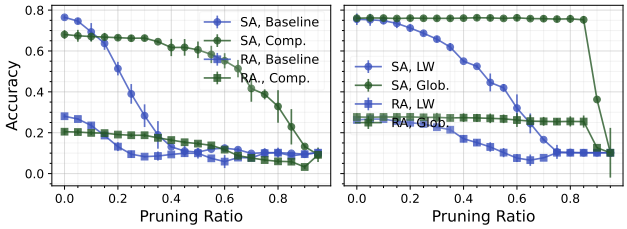

Figure 8: Robustness under compression. SA/RA: Standard/Robust Acc. LW/Glob.: Layerwise vs. global pruning.

In Figure 8 (right), we show that conducting pruning based on two simple interventions inspired by our bounds results in tangible improvements in standard and robust performance under pruning. Given the fact that layerwise pruning is known to produce harmful bottlenecks that lead to layer collapse (Blalock et al., 2020), instead of targeting a pruning ratio and pruning each layer accordingly, we set a target ϵ for each layer, and for each compute k that satisfies this ϵ level. Given a target global pruning ratio, we scan over different levels of ϵ and determine the level that gets closest to the target ratio. Moreover, during training we control the spread of the dominant terms, β, which our analyses show to be harmful for robustness, without decreasing compressibility. We accomplish this through regularizing the variance of the top 0.05 of each layer's filters' norms. Figure 8 (right) demonstrates that our interventions create a tangible improvement in performance retention. However, as useful as such interventions can be, we also highlight the fundamental dangers of concentrating parameter energy in very few substructures that our findings reveal. Therefore, while pruning and low-rank

approximation remain valuable compression methods, combining intermediate levels thereof with other compression methods such as quantization or knowledge distillation seems to be the most promising approach in reconciling safety and robustness, which is in line with other recent findings in the literature (Kuzmin et al., 2023; Hong et al., 2024).

5 Related work

Adversarial robustness. The susceptibility of the neural network models to adversarial examples created through small perturbations (Szegedy et al., 2014) engendered a lot of research investigating the issue (Madry et al., 2018). To this day adversarial robustness remains one of the most important topics in machine learning security (Malik et al., 2024). The literature ranges from the development of new attacks and defenses (Moosavi-Dezfooli et al., 2016; Abdollahpoorrostam et al., 2024), to investigating sources/mechanisms of adversarial vulnerability, to implications of AEs for the inductive biases of modern machine learning architectures (Ilyas et al., 2019; Ortiz-Jimenez et al., 2021; Xu et al., 2024), to developing strategies to retain model expressivity and generalization while defending against adversarial attacks (Tsipras et al., 2019; Zhang et al., 2024).

Compressibility and pruning. Prominent compression approaches include pruning, quantization, distillation, conditional computing, and efficient architecture development (O'Neill, 2020). Out of these, pruning remains among the most actively researched compression approaches due to its versatility (Cheng et al., 2024). Inducing compressibility / sparsity at training time is the easiest way to obtain prunable models (Hohman et al., 2024). Compressibility across different substructures, a.k.a group sparsity (Li et al., 2020b), allows for structured pruning (e.g. neuron/row, filter/channel, kernel pruning), which is computationally efficient (Yang et al., 2018), yet lead to a sharp reduction in network connectivity, threatening performance (Blalock et al., 2020). Lastly, spectral compressibility relaxes the notion of low-rankness, utilized for approximating large matrices with appealing theoretical properties (Suzuki et al., 2020; Schotthöfer et al., 2022).

Compressibility and robustness. Whereas some research argues that compressibility is beneficial for adversarial robustness (Guo et al., 2018; Balda et al., 2020; Liao et al., 2022), others indicate the relation is *at best* highly dependent on the degree and type of compressibility, as well as attack type (Li et al., 2020a; Merkle et al., 2022; Savostianova et al., 2023; Feng et al., 2025). While a stream of new methods that incorporate adversarial robustness in novel ways to pruning, newly emerging systematic benchmarks reveal at best marginal benefits for such methods compared to weight-based pruning (Lee et al., 2020; Piras et al., 2024). Whereas some methods demonstrate benefits of adversarial training-aware sparsification (Gui et al., 2019; Sehwag et al., 2020; Pavlitska et al., 2023), infamous problems adversarial training (AT) poses for standard generalization, transferability, as well as computational feasibility especially for larger models still plague such methods (Tsipras et al., 2019; Wen et al., 2020; Yang et al., 2024).

6 Conclusion and future work

In this paper, we present a unified theoretical and empirical treatment of how structured compressibility shapes adversarial robustness. Via a novel analysis of neuron-level and spectral compressibility, we uncover a fundamental mechanism: compression concentrates sensitivity along a small number of directions in representation space, rendering models more vulnerable—even under adversarial training and transfer learning. Our norm-based robustness bounds offer interpretable decompositions that predict both standard and universal adversarial vulnerability, and shed light on the trade-offs between efficiency and security in modern neural networks. Empirically, we validate these insights across datasets, architectures, and training regimes, showing how both compressibility and its spread determines adversarial susceptibility. While these vulnerabilities can be mitigated through targeted strategies, combining these methods with other approaches is likely the most promising approach.

Our work provides a novel insight into the relationship between structured compressibility and adversarial vulnerability. A limitation is our theory's reliance on global Lipschitz constants to characterize network performance: future work should focus on providing a unified view that incorporates both structural/global weaknesses, as well as the localization of sensitivity in the input space. Moreover, while the simple interventions suggested by our theory provides cost-effective improvements to the compressibility-robustness trade-off, these insights should be combined with novel compression methods to improve the frontiers of robust compression.

Reproducibility statement. We enable the reproduction of our work through a detailed description of our methods in the main paper and the Appendix, as well as the source code for our main experiments provided as supplementary material.

REFERENCES

Alireza Abdollahpoorrostam, Mahed Abroshan, and Seyed-Mohsen Moosavi-Dezfooli. SuperDeepFool: A new fast and accurate minimal adversarial attack. *Advances in Neural Information Processing Systems*, 37:98537–98562, December 2024.

Arash Amini, Michael Unser, and Farokh Marvasti. Compressibility of Deterministic and Random Infinite Sequences. *IEEE Transactions on Signal Processing*, 59(11):5193–5201, November 2011. ISSN 1941-0476. doi: 10/cznsth.

Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. In *International Conference on Machine Learning*, pp. 254–263. PMLR, 2018.

Emilio Balda, Niklas Koep, Arash Behboodi, and Rudolf Mathar. Adversarial Risk Bounds through Sparsity based Compression. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, pp. 3816–3825. PMLR, June 2020.

Melih Barsbey, Milad Sefidgaran, Murat A. Erdogdu, Gaël Richard, and Umut Şimşekli. Heavy Tails in SGD and Compressibility of Overparametrized Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34. Curran Associates, Inc., 2021.

Davis Blalock, Jose Javier Gonzalez Ortiz, Jonathan Frankle, and John Guttag. What is the State of Neural Network Pruning? *arXiv:2003.03033 [cs, stat]*, March 2020.

Sébastien Bubeck, Yuanzhi Li, and Dheeraj Nagaraj. A law of robustness for two-layers neural networks. *arXiv:2009.14444 [cs, stat]*, November 2020.

Hongrong Cheng, Miao Zhang, and Javen Qinfeng Shi. A Survey on Deep Neural Network Pruning: Taxonomy, Comparison, Analysis, and Recommendations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10558–10578, December 2024.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval Networks: Improving Robustness to Adversarial Examples, May 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *ICML'20*, pp. 2206–2216. JMLR.org, July 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, June 2009. doi: 10.1109/CVPR.2009.5206848.

Li Deng. The MNIST Database of Handwritten Digit Images for Machine Learning Research [Best of the Web]. *IEEE Signal Processing Magazine*, 29(6):141–142, November 2012. ISSN 1558-0792. doi: 10.1109/MSP.2012.2211477.

Enmao Diao, Ganghua Wang, Jiawei Zhan, Yuhong Yang, Jie Ding, and Vahid Tarokh. Pruning Deep Neural Networks from a Sparsity Perspective, August 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*, October 2021.

Yangqi Feng, Shing-Ho J Lin, Baoyuan Gao, and Xian Wei. Lipschitz constant meets condition number: Learning robust and compact deep neural networks. *arXiv preprint arXiv:2503.20454*, 2025.

A. Frank. Some Polynomial Algorithms for Certain Graphs and Hypergraphs. *Utilitas Mathematica*, 1976. ISSN ,.

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

Rémi Gribonval, Volkan Cevher, and Mike E. Davies. Compressible Distributions for High-Dimensional Statistics. *IEEE Transactions on Information Theory*, 58(8):5016–5034, August 2012. ISSN 1557-9654. doi: 10/f3585p.

Ekaterina Grishina, Mikhail Gorbunov, and Maxim Rakhuba. Tight and Efficient Upper Bound on Spectral Norm of Convolutional Layers. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 19–34, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-73024-5. doi: 10.1007/978-3-031-73024-5_2.

Shupeng Gui, Haotao N Wang, Haichuan Yang, Chen Yu, Zhangyang Wang, and Ji Liu. Model Compression with Adversarial Robustness: A Unified Optimization Framework. In *NeurIPS*, pp. 12, 2019.

Yiwen Guo, Chao Zhang, Changshui Zhang, and Yurong Chen. Sparse DNNs with Improved Adversarial Robustness. In *NeurIPS*, pp. 10, 2018.

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 (CVPR)*, pp. 770–778, 2016.

Fred Hohman, Mary Beth Kery, Donghao Ren, and Dominik Moritz. Model Compression in Practice: Lessons Learned from Practitioners Creating On-device Machine Learning Experiences. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, pp. 1–18, New York, NY, USA, May 2024. Association for Computing Machinery. ISBN 9798400703300. doi: 10.1145/3613904.3642109.

Junyuan Hong, Jinhao Duan, Chenhui Zhang, Zhangheng Li, Chulin Xie, Kelsey Lieberman, James Diffenderfer, Brian Bartoldson, Ajay Jaiswal, Kaidi Xu, Bhavya Kailkhura, Dan Hendrycks, Dawn Song, Zhangyang Wang, and Bo Li. Decoding Compressed Trust: Scrutinizing the Trustworthiness of Efficient LLMs Under Compression, June 2024.

Rasheed Hussain and Sherali Zeadally. Autonomous Cars: Research Results, Issues, and Future Challenges. *IEEE Communications Surveys & Tutorials*, 21(2):1275–1313, 2019. ISSN 1553-877X. doi: 10.1109/COMST.2018.2869360. Conference Name: IEEE Communications Surveys & Tutorials.

Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. URL `https://doi.org/10.5281/zenodo.5143773`. If you use this software, please cite it as below.

Andrew Ilyas, Logan Engstrom, Shibani Santurkar, Brandon Tran, Dimitris Tsipras, and Aleksander Ma. Adversarial Examples are not Bugs, they are Features. In *NeurIPS*, pp. 12, 2019.

Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario, 2009. URL `https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf`.

Andrey Kuzmin, Markus Nagel, Mart van Baalen, Arash Behboodi, and Tijmen Blankevoort. Pruning vs Quantization: Which is Better? *Advances in Neural Information Processing Systems*, 36: 62414–62427, December 2023.

Jaeho Lee, Sejun Park, Sangwoo Mo, Sungsoo Ahn, and Jinwoo Shin. Layer-adaptive Sparsity for the Magnitude-based Pruning. In *International Conference on Learning Representations*, 2020.

Fuwei Li, Lifeng Lai, and Shuguang Cui. On the Adversarial Robustness of Feature Selection Using LASSO. In *2020 IEEE 30th International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6, September 2020a. doi: 10.1109/MLSP49062.2020.9231631.

Yawei Li, Shuhang Gu, Christoph Mayer, Luc Van Gool, and Radu Timofte. Group Sparsity: The Hinge Between Filter Pruning and Decomposition for Network Compression. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8015–8024, Seattle, WA, USA, June 2020b. IEEE. ISBN 978-1-72817-168-5. doi: 10.1109/CVPR42600.2020.00804.

Ningyi Liao, Shufan Wang, Liyao Xiang, Nanyang Ye, Shuo Shao, and Pengzhi Chu. Achieving adversarial robustness via sparsity. *Machine Learning*, 111(2):685–711, February 2022. ISSN 1573-0565. doi: 10.1007/s10994-021-06049-9.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, Montreal, QC, Canada, October 2021. IEEE. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00986.

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*, February 2018.

TorchVision maintainers and contributors. Torchvision: Pytorch's computer vision library. `https://github.com/pytorch/vision`, 2016.

Jasmita Malik, Raja Muthalagu, and Pranav M. Pawar. A Systematic Review of Adversarial Machine Learning Attacks, Defensive Controls, and Technologies. *IEEE Access*, 12:99382–99421, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3423323.

Florian Merkle, Maximilian Samsinger, and Pascal Schöttle. Pruning in the Face of Adversaries. In Stan Sclaroff, Cosimo Distante, Marco Leo, Giovanni M. Farinella, and Federico Tombari (eds.), *Image Analysis and Processing – ICIAP 2022*, pp. 658–669, Cham, 2022. Springer International Publishing. ISBN 978-3-031-06427-2. doi: 10.1007/978-3-031-06427-2_55.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral Normalization for Generative Adversarial Networks. In *International Conference on Learning Representations*, February 2018.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2574–2582, Las Vegas, NV, USA, June 2016. IEEE. doi: 10.1109/cvpr.2016.282.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1765–1773, 2017.

Ramchandran Muthukumar and Jeremias Sulam. Adversarial Robustness of Sparse Local Lipschitz Predictors. *SIAM Journal on Mathematics of Data Science*, 5(4):920–948, December 2023. doi: 10.1137/22M1478835.

Laura F Nern, Harsh Raj, Maurice André Georgi, and Yash Sharma. On transfer of adversarial robustness from pretraining to downstream tasks. In *Advances in neural information processing systems*, volume 36, pp. 59206–59226, 2023.

Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011. URL `http://ufldl.stanford.edu/housenumbers/`.

Maria-Irina Nicolae, Mathieu Sinn, Minh Tran, Beat Buesser, Anish Rawat, Martin Wistuba, Valerio Zantedeschi, Nathalie Baracaldo, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.0.0. *arXiv preprint arXiv:1807.01069*, 2018.

James O'Neill. An Overview of Neural Network Compression. *arXiv:2006.03669*, August 2020.

Guillermo Ortiz-Jimenez, Apostolos Modas, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Optimism in the Face of Adversity: Understanding and Improving Deep Learning Through Adversarial Robustness. *Proceedings of the IEEE*, 109(5):635–659, May 2021. ISSN 0018-9219, 1558-2256. doi: 10.1109/JPROC.2021.3050042.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, December 2019.

Svetlana Pavlitska, Hannes Grolig, and J. Marius Zollner. Relationship between Model Compression and Adversarial Robustness: A Review of Current Evidence. In *2023 IEEE Symposium Series on Computational Intelligence (SSCI)*, pp. 671–676, December 2023.

Giorgio Piras, Maura Pintor, Ambra Demontis, Battista Biggio, Giorgio Giacinto, and Fabio Roli. Adversarial Pruning: A Survey and Benchmark of Pruning Methods for Adversarial Robustness, September 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *International Conference on Machine Learning*, 2021.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J. Topol. AI in health and medicine. *Nature Medicine*, 28:31–38, January 2022. ISSN 1546-170X. doi: 10.1038/s41591-021-01614-0. URL https://www.nature.com/articles/s41591-021-01614-0. Bandiera_abtest: a Cg_type: Nature Research Journals Primary_atype: Reviews Publisher: Nature Publishing Group Subject_term: Computational biology and bioinformatics;Medical research Subject_term_id: computational-biology-and-bioinformatics;medical-research.

Antonio Ribeiro, Dave Zachariah, Francis Bach, and Thomas Schön. Regularization properties of adversarially-trained linear regression. In *Advances in Neural Information Processing Systems*, volume 36, pp. 23658–23670, December 2023.

Dayana Savostianova, Emanuele Zangrando, Gianluca Ceruti, and Francesco Tudisco. Robust low-rank training via approximate orthonormal constraints. In *Advances in Neural Information Processing Systems*, volume 36, pp. 66064–66083, 2023.

Kevin Scaman and Aladin Virmaux. Lipschitz regularity of deep neural networks: Analysis and efficient estimation. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3839–3848. Curran Associates Inc., December 2018.

Steffen Schotthöfer, Emanuele Zangrando, Jonas Kusch, Gianluca Ceruti, and Francesco Tudisco. Low-rank lottery tickets: Finding efficient low-rank neural networks via matrix differential equations. In *Advances in Neural Information Processing Systems*, October 2022.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19655–19666, 2020.

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

Taiji Suzuki, Hiroshi Abe, Tomoya Murata, Shingo Horiuchi, Kotaro Ito, Tokuma Wachi, So Hirai, Masatoshi Yukishima, and Tomoaki Nishimura. Spectral Pruning: Compressing Deep Neural Networks via Spectral Analysis and its Generalization Error. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, pp. 2839–2846, Yokohama, Japan, July 2020. International Joint Conferences on Artificial Intelligence Organization.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *arXiv:1312.6199 [Cs]*, February 2014.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness May Be at Odds with Accuracy. *arXiv:1805.12152 [cs, stat]*, September 2019.

Yijun Wan, Melih Barsbey, Abdellatif Zaidi, and Umut Simsekli. Implicit Compressibility of Overparametrized Neural Networks Trained with Heavy-Tailed SGD. In *Forty-First International Conference on Machine Learning*, June 2024.

Yuxin Wen, Shuai Li, and Kui Jia. Towards understanding the regularization of adversarial robustness on neural networks. In *International Conference on Machine Learning*. PMLR, 2020.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020.

Tianlong Xu, Chen Wang, Gaoyang Liu, Yang Yang, Kai Peng, and Wei Liu. United We Stand, Divided We Fall: Fingerprinting Deep Neural Networks via Adversarial Trajectories. *Advances in Neural Information Processing Systems*, 37:69299–69328, December 2024.

Carl Yang, Aydın Buluç, and John D. Owens. Design Principles for Sparse Matrix Multiplication on the GPU. In *Euro-Par 2018: Parallel Processing: 24th International Conference on Parallel and Distributed Computing, Turin, Italy, August 27 - 31, 2018, Proceedings*, pp. 672–687, Berlin, Heidelberg, August 2018. Springer-Verlag. ISBN 978-3-319-96982-4. doi: 10.1007/978-3-319-96983-1_48.

Sheng Yang, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Spectral regularization for adversarially-robust representation learning, May 2024.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint 1605.07146*, 2016.

Kaibo Zhang, Yunjuan Wang, and Raman Arora. Stability and Generalization of Adversarial Training for Shallow Neural Networks with Smooth Activation. *Advances in Neural Information Processing Systems*, 37:16160–16193, December 2024.

Shaochen (Henry) Zhong, Zaichuan You, Jiamu Zhang, Sebastian Zhao, Zachary LeClaire, Zirui Liu, Daochen Zha, Vipin Chaudhary, Shuai Xu, and Xia Hu. One Less Reason for Filter Pruning: Gaining Free Adversarial Robustness with Structured Grouped Kernel Pruning. *Advances in Neural Information Processing Systems*, 36:62032–62061, December 2023.

Monty-Maximilian Zühlke and Daniel Kudenko. Adversarial Robustness of Neural Networks from the Perspective of Lipschitz Calculus: A Survey. *ACM Comput. Surv.*, 57(6):142:1–142:41, February 2025. ISSN 0360-0300. doi: 10.1145/3648351.