Feature Compression May Be the Root Cause of Adversarial Fragility in Neural Network Classifiers (Student Abstract)

Jingchao Gao[2*], Ziqing Lu[1*], Raghu Mudumbai[1], Xiaodong Wu[1], Jirong Yi[1], Myung Cho[1], Catherine Xu[1], Hui Xie[1], Weiyu Xu[1]

[1]University of Iowa,
[2]Minnesota State University

Abstract

In this paper, we study the adversarial robustness of deep neural networks (DNN) for classification against optimal classifiers. We look at the smallest magnitude of possible additive perturbations that can change a classifier's output. We provide a novel matrix-theoretic explanation of the adversarial fragility of DNNs for classification. In particular, our theoretical results show that the adversarial robustness of a neural network can degrade as the input dimension d increases. Analytically, we show that the adversarial robustness of NN can be only $1/\sqrt{d}$ of the best possible adversarial robustness of optimal classifiers. Our theories match remarkably well with empirical results. The matrix-theoretic explanation aligns with an earlier information-theoretic feature-compression-based explanation for the adversarial fragility of neural networks.

Introduction

Neural network (NN) based classifiers achieve high accuracy across various tasks, but they are also universally vulnerable to adversarial perturbations and show poor robustness. There is no clear consensus on the reason for the adversarial fragility. This paper uniquely compares the *worst-case performances* of NN-based classifiers against *worst-case performances* of optimal classifiers, while previous works explained the adversarial fragility purely as the gap between the average-case performance and the worst-case performance (Goodfellow, Shlens, and Szegedy 2014). Our matrix-theoretic analysis shows that as the input dimension d increases, the adversarial robustness of neural networks can be only $1/\sqrt{d}$ of the best possible adversarial robustness.

We attribute this phenomenon to **feature compression**: NNs only use compressed features for classification, so an adversary only needs to perturb these compressed features. Figure 1 illustrates this concept. This observation aligns with information-theoretic hypotheses that adversarial fragility arises from feature compressions (Xie et al. 2019). Unlike this prior work, which gave a higher-level explanation based on the feature compression hypothesis and high-dimensional geometric analysis, we offer a detailed, architecture-specific analysis. Our results cover both linear networks and nonlinear networks, involving an exponential number (in d) of data points.

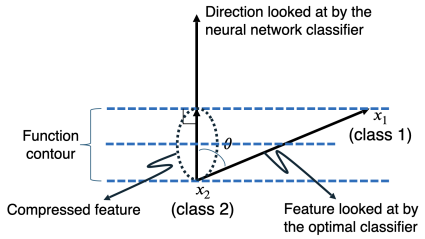

Figure 1: The true input $\mathbf{x}_2$ belongs to Class 2. Ideally, the optimal classifier separates classes by examining the direction $\mathbf{x}_1 - \mathbf{x}_2$, where $\mathbf{x}_1$ is the closest point in Class 1, since this direction represents the "weakest" separation between Class 1 and $\mathbf{x}_2$. However, the NN classifier relies on a compressed feature direction having angle θ with $\mathbf{x}_1 - \mathbf{x}_2$. Therefore, to change the label from Class 2 to Class 1, the attacker only needs to perturb $\mathbf{x}_2$ along the compressed feature direction.

Feature Compression

Data setup: We construct 2^d data points $(\mathbf{x}_i, y_i)$, with $\mathbf{x}_i = A\mathbf{z}_i, \mathbf{z}_i \in \{-1,+1\}^d, A_{ik} \sim \mathcal{N}(0,1)$. The label y_i is determined by the last component of $\mathbf{z}_i$: $y_i = +1$ if $\mathbf{z}_i(d) = +1$; $y_i = -1$ if $\mathbf{z}_i(d) = -1$. Let C_{+1} and C_{-1} denote the sets of inputs with labels $y_i = +1$ and $y_i = -1$.

Theorem 1 (Robustness of Linear Networks) *Consider an l-layer linear neural network with the hidden layers' output $\mathbf{o} = H_{l-1}...H_1\mathbf{x}, H_i \in \mathbb{R}^{n_{i+1}\times n_i}, n_1 = d$. For each class C_{+1} or C_{-1}, suppose that the two output neurons are $f_{+1}(\mathbf{x}) = \mathbf{w}_{+1}^T\mathbf{o}$ and $f_{-1}(\mathbf{x}) = \mathbf{w}_{-1}^T\mathbf{o}$. Suppose that the neural network makes perfect classification on the constructed dataset:*

$$f_{+1}(\mathbf{x}_i) = \begin{cases} +1, \text{if } \mathbf{z}_i(d) = +1, \\ -1, \text{if } \mathbf{z}_i(d) = -1. \end{cases} \quad f_{-1}(\mathbf{x}_i) = \begin{cases} +1, \text{if } \mathbf{z}_i(d) = -1, \\ -1, \text{if } \mathbf{z}_i(d) = +1. \end{cases}$$

Denote the last element of $\mathbf{z}_i$ corresponding to the ground-truth input $\mathbf{x}_i$ by 'bit'. Then, with high probability we have: ***Class separation****: There exists a constant $\alpha > 0$ such that the minimum distance between any data point in C_{+1} and any data point in C_{-1} satisfies $\min_{\mathbf{x}_i \in C_{+1},\ \mathbf{x}_j \in C_{-1}} \|\mathbf{x}_i - \mathbf{x}_j\|_2 \geq \alpha\sqrt{d}$.*

*These authors contributed equally.

Experiment No.	1	2	3	4	5	6	7	8	9	10
$\cos(\theta_1)$	-0.1970	-0.1907	-0.6017	-0.2119	-0.2449	-0.5054	-0.7794	-0.5868	-0.1655	-0.4739
$\cos(\theta_2)$	-0.9992	-0.9992	-0.9984	-0.9994	-0.9955	-0.9988	-0.0795	-0.9972	-0.9993	-0.9942
ϕ	0.1812	0.1870	0.5888	0.2048	0.2032	0.4985	0.0738	0.5895	0.1480	0.4497

Table 1: Cosine of angles of trained models with training accuracy equal to 1, $d = 12$: $|\cos(\theta_1)|$ is the actual feature compression ratio, indicating actual needed attack perturbation size. Our theory predicts that $|\cos(\theta_1)|$ should be close to the theoretical feature compression ratio ϕ, when $|\cos(\theta_2)|$ is close to 1. All these properties are verified in the 10 experiments.

***Existence of small adversarial perturbations**: For any $\mathbf{x} = \mathbf{x}_i$, there exists a perturbation $\mathbf{e}$ with $\|\mathbf{e}\|_2 \leq D$ such that $f_{-bit}(\mathbf{x}_i+\mathbf{e}) > f_{bit}(\mathbf{x}_i+\mathbf{e})$, which means that the classifier's decision can be flipped by a small perturbation.*

As we can see in the proof of Theorem 1 (See Appendix E), because the NN classifier bases its decision on the compressed features aligned with $Q_{:,d}$ (the last column of Q from QR decomposition of A), attacks along this direction require a much smaller perturbation. In the following theorem, we extend our results to nonlinear NN-based classifiers, showing that successful attacks only need to change the input along the direction of "compression" imposed on the input data. (See proof of Theorem 2 in Appendix F).

Theorem 2 (Robustness of nonlinear networks) *Consider a multi-layer neural network classifier with an arbitrary input $\mathbf{x} \in \mathbb{R}^d$, and let $f_i(\mathbf{x})$ denote the output for class i with gradient $\nabla f_i(\mathbf{x})$. For each class i, let $\mathbf{x} + \mathbf{x}_i$ be the closest point in that class to $\mathbf{x}$, and fix a small $\epsilon > 0$. Assume that the first-order approximation error of f_i near $\mathbf{x}$ is of the order $o(\epsilon)$.*

For points $\mathbf{x} + \epsilon\mathbf{x}_1$ and $\mathbf{x} + \epsilon\mathbf{x}_2$, there exits a perturbation $\mathbf{e}$ such that

$$f_1(\mathbf{x} + \epsilon\mathbf{x}_1 + \mathbf{e}) \doteq f_1(\mathbf{x} + \epsilon\mathbf{x}_2) \quad \textit{and} \quad f_2(\mathbf{x} + \epsilon\mathbf{x}_1 + \mathbf{e}) \doteq f_2(\mathbf{x} + \epsilon\mathbf{x}_2),$$

where $\doteq$ denotes the equality up to first-order approximation of ϵ, namely $o(\epsilon)$. Moreover,

$$\|\mathbf{e}\|_2 \leq \epsilon \|P_{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x})}(\mathbf{x}_1 - \mathbf{x}_2)\|_2,$$

with $P_{\nabla f_1(\mathbf{x}), \nabla f_2(\mathbf{x})}$ the projection onto the subspace spanned by $\nabla f_1(\mathbf{x})$ and $\nabla f_2(\mathbf{x})$.

To make the classifier misclassify $\mathbf{x} + \epsilon\mathbf{x}_1$ as $\mathbf{x} + \epsilon\mathbf{x}_2$ in the two output neurons, it is sufficient to add a small perturbation $\boldsymbol{e}$ rather than the full perturbation $\epsilon(\mathbf{x}_2 - \mathbf{x}_1)$, due to the compression of $\mathbf{x}_2 - \mathbf{x}_1$ along $\nabla f_1(\mathbf{x})$ and $\nabla f_2(\mathbf{x})$.

Global analysis: We extend our results from the local analysis around the input $\mathbf{x}$ in the previous section to the "global" analysis. Consider the same setting as described in Figure 1.Consider the direct path from $\mathbf{x}_2$ to $\mathbf{x}_1$, and define $g(\mathbf{x}) = f_2(\mathbf{x}) - f_1(\mathbf{x})$, where $f_2(\cdot)$ and $f_1(\cdot)$ are the neuron outputs for Class 2 and Class 1. Along this path, $g(\mathbf{x})$ changes by $D = g(\mathbf{x}_1) - g(\mathbf{x}_2)$. The path has length $\|\mathbf{x}_1 - \mathbf{x}_2\|$, which we parameterize by the length parameter $0 \leq \gamma \leq \|\mathbf{x}_1 - \mathbf{x}_2\|$.

Following this path, we can write D in another way: $D = \int_0^{\|\mathbf{x}_1 - \mathbf{x}_2\|} \|\nabla g\left(\mathbf{x}_{\gamma,\mathbf{x}_1,\mathbf{x}_2}\right)\| \times \cos(\theta_\gamma)\, d\gamma$, where $\mathbf{x}_{\gamma,\mathbf{x}_1,\mathbf{x}_2} = \frac{\gamma}{\|\mathbf{x}_1 - \mathbf{x}_2\|}\mathbf{x}_1 + (1 - \frac{\gamma}{\|\mathbf{x}_1 - \mathbf{x}_2\|})\mathbf{x}_2$ is the point at distance γ along the segment, ∇g is the gradient of g, and θ_γ is the angle between $\nabla g(\mathbf{x})$ at the point and the direction $\mathbf{x}_1 - \mathbf{x}_2$.

Suppose that an attacker starting at $\mathbf{x}_2$ takes infinitely small steps along the negative direction of the gradient of $g(\mathbf{x})$. Let the attack path have length z, and parameterize the path by $0 \leq \gamma \leq z$, with $\mathbf{x}_\gamma$ the point reached after the attacker has traveled length γ. If the attack produces the same net change D in g, then $D = \int_0^z -\|\nabla g(\mathbf{x}_\gamma)\|\, d\gamma$.

Due to the **compression factor** "$\cos(\theta_\gamma)$" (often small, sometimes even negative), the length of the direct path $\|\mathbf{x}_1 - \mathbf{x}_2\|$ must generally be much bigger than the attack path z to produce the same D, assuming $\|g(\mathbf{x})\|$ to be comparable at locations of interest.

Numerical Results

We present numerical results in Table 1 that confirm the theoretical predictions of adversarial fragility, focusing on the linear network setting of Theorem 1. Feature compression results for nonlinear networks trained on MNIST and ImageNet also confirm similar results. (See Appendix H for details) For higher input dimensions $d = 7, 8, 9, 10, 12, 14, 15, 16, 17$, the averaged "fraction" ϕ, and the compressed feature $\cos(\theta_1)$ are shown in Figure 2. They match really well.

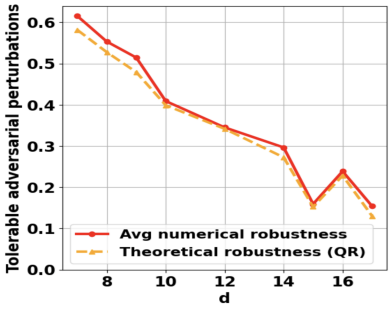

Figure 2: Theoretical predictions ϕ matches practically-trained NN's $\cos(\theta_1)$. ImageNet experiments in appendix.

References

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2014. Explaining and harnessing adversarial examples. *arXiv:1412.6572*.

Xie, H.; Yi, J.; Xu, W.; and Mudumbai, R. 2019. An Information-Theoretic Explanation for the Adversarial Fragility of AI Classifiers. In *2019 IEEE International Symposium on Information Theory (ISIT)*, 1977–1981.