Why Robust Generalization in Deep Learning is Difficult: Perspective of Expressive Power

Advisor: Liwei Wang
Speaker: Binghui Li

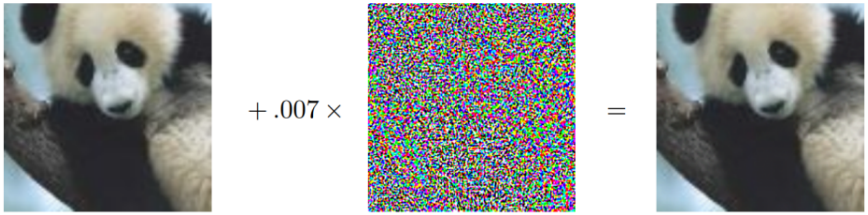
Turing Class
Peking University

Outline

Table of Contents

Adversarial Examples

- Although deep neural networks have achieved remarkable success in practice, it is well-known that modern neural networks are vulnerable to adversarial examples.
- Specifically, for a given image $\boldsymbol{x}$, an indistinguishable small but adversarial perturbation $\boldsymbol{\delta}$ is chosen to fool the classifier f to produce a wrong class using $f(\boldsymbol{x} + \delta)$.



Approaches to Achieve Adversarial Robustness

- To mitigate this problem, a series of robust learning algorithms have been proposed.
- A common approach is to design adversarial training algorithms by using adversarial examples as training data [MMS+17, TKP+18, SNG+19], which centrally considers the min-max optimization problem as follow,

$$\min_{\theta\in\Theta}\left\{\mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\max_{\|x'-x\|\leq\delta}\text{Ł}(f_\theta(x'),y)\right]\right\},$$

  where $\Theta, \mathcal{D}, \delta, \text{Ł}(\cdot)$ denote parameter-space, data distribution, perturbation radius and loss function, respectively.
- Another line of works proposes some provably robust models to tackle this problem, such as randomized smoothing [CRK19] and $\ell_\infty-$dist Net [ZCL+21].

Robust Generalization Gap is Large!

- However, while the state-of-the-art adversarial training methods can achieve high robust training accuracy (e.g. nearly 100% on CIFAR-10 [RXY+19]), all existing methods suffer from large robust test error.

	Standard training	**Adversarial training**
Robust test	3.5%	45.8%
Robust train	-	100%
Standard test	95.2%	87.3%
Standard train	100%	100%

- Therefore, it is natural to ask what is the cause for such a large generalization gap in the context of robust learning.

Question

Why is robust generalization in deep learning difficult?
Can we provide a **theoretical understanding** of this puzzling phenomenon?

Binghui Li*, Jikai Jin*, Han Zhong, John E. Hopcroft, Liwei Wang

Our paper [LJZ+22] can be found at:
https://arxiv.org/abs/2205.13863

Key Empirical Observation: Data are Far from Each Other

- In fact, it is observed that for real data sets, different classes tend to be well-separated [YRZ+20].

	adversarial perturbation ε	minimum Train-Train separation	minimum Test-Train separation
MNIST	0.1	0.737	0.812
CIFAR-10	0.031	0.212	0.220
SVHN	0.031	0.094	0.110
ResImageNet	0.005	0.180	0.224

- Moreover, the typical perturbation radius is often much smaller than the separation distance.

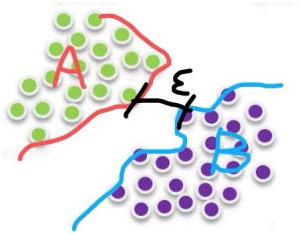
Well-separated Data

Definition 1.1 (Separated Data)

Suppose that $A, B \subset \mathbb{R}^d$ and $\epsilon > 0$. We say that A, B are ϵ-separated under ℓ_p norm $(1 \leq p \leq +\infty)$ if

$$\|\boldsymbol{x}_A - \boldsymbol{x}_B\|_p \geq \epsilon, \quad \forall \boldsymbol{x}_A \in A, \boldsymbol{x}_B \in B.$$

Indeed, this assumption is necessary to ensure the existence of a robust classifier. Without this separated condition, it is clear that there is no robust classifier even if a non-robust classifier always exists.



Robust Binary Classification Problems with Well-separated Data

- In our work, we consider robust binary classification problems with well-separated data. Formally, let $A, B \subset [0,1]^d$ be two disjoint sets that are $2\epsilon-$separated, where points in A have label $+1$ and points in B have label -1, and $\delta > 0$ be the perturbation radius that satisfies $\delta < \epsilon$.
- We are mainly interested in the following questions:
  - In **robust training** setting, given a $N-$sample data set $\mathcal{D}$ from arbitrary $2\epsilon-$separated A and B, then how many parameters are enough to achieve zero $\delta-$robust training error on $\mathcal{D}$ for ReLU nets?
  - In **robust generalization** setting, how many parameters are enough to $\delta-$robustly classify arbitrary $2\epsilon-$separated A and B for ReLU nets?

Table of Contents

Robust Training of Well-separated Data

- Suppose that $\mathcal{D} \subset \{\boldsymbol{x} \in \mathbb{R}^d | \|\boldsymbol{x}\|_p \leq 1\}$ with $p \in \{2, +\infty\}$ consists of N data, and the two classes in $\mathcal{D}$ are 2ϵ-separated, where $\epsilon \in \left(0, \frac{1}{2}\right)$ is a constant.
- With access to only finite amount of data, a common practice for learning a robust classifier is to minimize the *robust training error* defined as

$$\hat{\mathrm{L}}_{\mathcal{D}}^{p,\delta}(f) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}\left\{\exists \boldsymbol{x}', \|\boldsymbol{x}' - \boldsymbol{x}_i\|_p \leq \delta, \mathrm{sgn}(f(\boldsymbol{x}')) \neq y_i\right\}.$$

where $\delta \geq 0$ is the adversarial perturbation radius. When $\delta = 0$, the definition coincides with the standard training error.

$\tilde{\mathcal{O}}(Nd)$ Parameters are Enough to Achieve Zero Robust Training Error

Theorem 2.1 (Upper Bound for the Size of Networks of Robust Training)

Let robustness radius $\delta < \frac{1}{2}\epsilon$, then there exists a classifier f represented by a ReLU network with at most

$$\mathcal{O}\left(Nd \log\left(\delta^{-1}d\right) + N \cdot \mathrm{polylog}(\delta^{-1}N)\right)$$

parameters, such that $\hat{\mathfrak{L}}_{\mathcal{D}}^{p,\delta}(f) = 0$.

- This theorem is our main result in this section, which states that for binary classification problems, a ReLU net with $\tilde{\mathcal{O}}(Nd)$ weights can robustly classify a data set of size N. It implies that over-parameterization is sufficient to achieve zero robust training error.
- While optimal (non-robust) memorization of N data points only needs constant width [VYS21], our construction in Theorem 2.1 has width $\tilde{\mathcal{O}}(Nd)$. Therefore, if our upper bound is nearly tight, then it can probably explain why increasing the network width can benefit robust training [MMS+17].

Table of Contents

Robust Generalization of Well-separated Data

- In the previous section, we give an upper bound on the size of ReLU networks to robustly classify finite training data. However, it says nothing about *robust generalization*.
- To evaluate the robust test performance, for a given probability measure P on $\mathbb{R}^d \times \{-1, +1\}$ and a robust radius $\delta \geq 0$, the robust test error of a classifier $f : \mathbb{R}^d \to \mathbb{R}$ w.r.t P and δ under ℓ_p norm is defined as

$$Ł_P^{p,\delta}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim P}\left[\max_{\|\boldsymbol{x}'-\boldsymbol{x}\|_p \leq \delta} \mathbb{I}\{y \neq \operatorname{sgn}(f(\boldsymbol{x}'))\}\right].$$

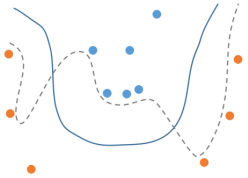
- In contrast with the training set which only consists of finite data points, when studying generalization, we must consider potentially infinite points in the classes that we need to classify.

There Exists a Robust Classifier

Proposition 3.1 (There Exists a Lipschitz and Robust Classifier)

For 2ϵ-separated $A, B \subset [0,1]^d$ under ℓ_p norm with $p \in \{2, +\infty\}$, the classifier $f^*(\boldsymbol{x}) := \frac{d_p(\boldsymbol{x},B)-d_p(\boldsymbol{x},A)}{d_p(\boldsymbol{x},A)+d_p(\boldsymbol{x},B)}$ is ϵ^{-1}-Lipschitz continuous, and satisfies $\text{Ł}_P^{p,\epsilon}(f^*) = 0$ for any probability distribution P on $A \cup B$, where $d_p(\boldsymbol{q}, S) := \inf_{\boldsymbol{q}' \in S} \|\boldsymbol{q} - \boldsymbol{q}'\|_p$.

- It turns out that, the 2ϵ-separated condition ensures the existence of such a classifier. Moreover, it can be realized by a Lipschitz function.
- However, it remains unclear whether the Lipschitz function constructed in Proposition 3.1 can actually be efficiently approximated by neural networks.



Warm Up: Robust Generalization Requires Exponential Parameters

The following theorem shows that ReLU networks with exponential size is sufficient for as robust classification.

Theorem 3.2 (Upper Bound for the Size of Networks of Robust Generalization)

For any two 2ϵ-separated $A, B \subset [0,1]^d$ under ℓ_p norm with $p \in \{2, +\infty\}$, distribution P on the supporting set $S = A \cup B$ and robust radius $c \in (0,1)$, there exists a ReLU network f with at most

$$\tilde{\mathcal{O}}\left(((1-c)\epsilon)^{-d}\right)$$

parameters, such that $Ł_P^{p,c\epsilon}(f) = 0$.

Indeed, it is well known that without additional assumptions, an exponentially large number of parameters is also *necessary* for approximating a Lipschitz function [DHM89, SYZ22], which motivates us to consider the lower bound for the size of networks in the same setting.

Warm Up: Robust Generalization Requires Exponential Parameters

Theorem 3.3 (Lower Bound for the Size of Networks of Robust Generalization)

Let $\mathcal{F}_n$ be the set of functions represented by ReLU networks with at most n parameters. Suppose that for any 2ϵ-separated sets $A, B \subset [0,1]^d$ under ℓ_p norm with $p \in \{2, +\infty\}$, there exists $f \in \mathcal{F}_n$ that can classify A, B with zero (standard) test error, then it must hold that

$$n = \Omega\left((2\epsilon)^{-\frac{d}{2}} \left(d \log\left(1/2\epsilon\right)\right)^{-\frac{1}{2}}\right).$$

- This result shows that even *without* requiring robustness, neural networks need to be exponentially large to correctly classify A and B.
- It implies that mere separability of data sets is insufficient to guarantee that they can be classified by ReLU networks, unless the network size is exponentially large.

Finer Data Structures Should be Taken into Consideration

- However, one should be careful when interpreting the conclusion of Theorem 3.3, since real-world data sets may possess additional structural properties.
- Specifically, the joint distribution of data can be decomposed as

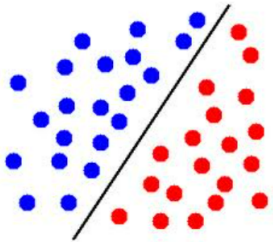
$$\mathcal{P}(X, Y) = \underbrace{\mathcal{P}(Y \mid X)}_{\text{labeling mapping}} \underbrace{\mathcal{P}(X)}_{\text{input}},$$

  where $\mathcal{P}(X, Y), \mathcal{P}(Y \mid X)$, and $\mathcal{P}(X)$ denote the joint, conditional and marginal distributions, respectively.
- In subsequent sections, we consider two well-known properties of data sets that correspond to the labeling mapping structure and the input structure, respectively, and study whether they can bring improvement to neural networks' efficiency for robust classification.

Table of Contents

Linear Separable Data

- As we have seen for separated data, if no other structural properties are taken into consideration, even standard generalization requires exponentially large neural networks.
- This motivates us to consider the following question: assuming that there exists a simple classifier that achieves zero standard test error on the data such as **the arguably simplest setting where the given data is linear separable and well-separated**, is it guaranteed that neural networks with reasonable size can also achieve high *robust* test accuracy?



Our Main Result: EXP Lower Bound Still Holds for Linear Separable Data

The following theorem is the main result of our paper.

Theorem 4.1 (Lower Bound for Linear Separable Data)

Let $\epsilon \in (0,1)$ be a small constant, $p \in \{2, +\infty\}$ and $\mathcal{F}_n$ be the set of functions represented by ReLU networks with at most n parameters. There exists a sequence $N_d = \Omega\left((2\epsilon)^{-\frac{d-1}{6}}\right), d \geq 1$ and a universal constant $C_1 > 0$ such that the following holds: for any $c \in (0,1)$, there exists two linear separable sets $A, B \subset [0,1]^d$ that are 2ϵ-separated under ℓ_p norm, such that for any μ_0-balanced distribution P on the supporting set $S = A \cup B$ and robust radius $c\epsilon$ we have

$$\inf\left\{Ł_P^{p,c\epsilon}(f) : f \in \mathcal{F}_{N_d}\right\} \geq C_1\mu_0.$$

- Theorem 4.1 states that the robust test error is lower-bounded by a positive constant $\alpha = C_1\mu_0$ unless the ReLU network has size larger than $\exp(\Omega(d))$.
- On the contrary, if we do not require robustness, then the data can be classified by a simple linear function.

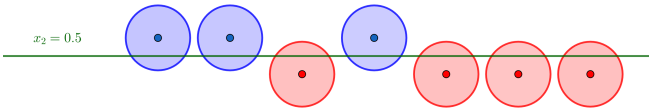
An Exponential Separation of Neural Network Size

The practical implication of Theorem 4.1 is two-fold:

- First, by comparing with non-robust linear classifiers, one can conclude that robust classification may require exponentially more parameters than the non-robust case, which is consistent with the common practice that larger models are used for adversarial robust training.
- Second, together with our upper bound in Theorem 2.1, Theorem 4.1 implies an *exponential* separation of neural network size for achieving high robust training and test accuracy.

Proof Sketch of Theorem 4.1

Let $K = [\frac{1}{2\epsilon}]$, and $\phi : \{1, 2, \cdots, K\}^{d-1} \to \{-1, +1\}$ be an arbitrary mapping, we define $S_\phi = \left\{ \left(\frac{i_1}{K}, \frac{i_2}{K}, \cdots, \frac{i_{d-1}}{K}, \frac{1}{2} + \epsilon_0 \cdot \phi(i_1, i_2, \cdots, i_{d-1}) \right) : 1 \le i_1, i_2, \cdots, i_{d-1} \le K \right\}$, where ϵ_0 is an arbitrarily small constant. The hyperplane $x^{(n)} = \frac{1}{2}$ partitions S_ϕ into two subsets, which we denote by A_ϕ and B_ϕ. We can check that A_ϕ and B_ϕ satisfies all the required conditions.



The remain of proof is to show that there exists some choice of ϕ such that robust classification is hard. By estimating growth function of $\{\phi\}$, we can derive the lower bound for the VC-dimension of $\mathcal{F}_n$ i.e.

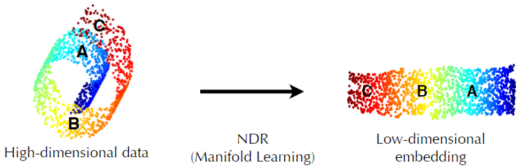
$$\text{VC-Dim}(\mathcal{F}_n) = \exp(\Omega(d)).$$

Finally, by applying the relation between the VC-dimension and the number of parameters appeared in [BHLM19], we prove the lower bound for the size of networks in Theorem 4.1 .

Table of Contents

Real-life Data Lies on a Low-dimensional Manifold

- A common belief of real-life data such as images is that the data points lie on a low-dimensional manifold.



e.g. some empirical work shows that the $28 \times 28 = 784$ dimensional image from MNIST can be reduced to nearly 10 dimensional representations [WYZ16].

- Motivated by this, we assume that data lies on a low-dimensional manifold $\mathcal{M}$ embedded in $[0,1]^d$ with the intrinsic dimension k ($k \ll d$) i.e. $\text{supp}(X) \subset \mathcal{M} \subset [0,1]^d$. And we extend robust classification to the version of manifold as

$$Ł_{\mathcal{M},P}^{p,\delta}(f) = \mathbb{E}_{(x,y)\sim P}\left[\max_{x'\in\mathcal{M},\|x'-x\|_p\leq\delta} \mathbb{I}\{y \neq f(x')\}\right].$$

Improved Upper Bound for Low-dimensional-manifold Data

Now, we present our main result in this section, which establishes an improved upper bound for size that is mainly exponential in the intrinsic dimension k instead of the ambient data dimension d.

Theorem 5.1 (Improved Upper Bound for Low-dimensional-manifold Data)

Let $\mathcal{M} \subset [0,1]^d$ be a $k-$dimensional compact poly-partitionable Riemannian manifold with the condition number $\tau > 0$. For any two $2\epsilon-$ separated $A, B \subset \mathcal{M}$ under ℓ_∞ norm, distribution P on the supporting set $S = A \cup B$ and robust radius $c \in (0,1)$, there exists a ReLU network f with at most

$$\tilde{\mathcal{O}}\left(\left((1-c)\,\epsilon/\sqrt{d}\right)^{-\tilde{k}}\right)$$

parameters, such that $\text{Ł}_{\mathcal{M},P}^{\infty,c\epsilon}(f) = 0$, where $\tilde{k} = \mathcal{O}(k \log d)$ is almost linear with respect to the intrinsic dimension k, only up to a logarithmic factor.

The Curse of Dimensionality is Inevitable

Although we have shown that robust classification will be more efficient when data lies on a low-dimensional manifold, there is also a curse of dimensionality

Theorem 5.2 (Lower Bound for Low-dimensional-manifold Data)

Let $\epsilon \in (0,1)$ be a small constant. There exists a sequence $\{N_k\}_{k\geq 1}$ that satisfies $N_k = \Omega\left((2\epsilon\sqrt{d/k})^{-\frac{k}{2}}\right)$. and a universal constant $C_1 > 0$ such that the following holds: let $\mathcal{M} \subset [0,1]^d$ be a complete and compact $k-$dimensional Riemannian manifold with non-negative Ricci curvature , then there exists two 2ϵ-separated sets $A, B \subset \mathcal{M}$ under ℓ_∞ norm, such that for any μ_0-balanced distribution P on the supporting set $S = A \cup B$ and robust radius $c \in (0,1)$, we have

$$\inf\left\{Ł_P^{\infty, c\epsilon}(f) : f \in F_{N_k}\right\} \geq C_1\mu_0.$$

In other words, the robust test error is lower-bounded by a positive constant $\alpha = C_1\mu_0$ unless the neural network has size larger than $\exp(\Omega(k))$.

Table of Contents

Conclusion

- In our work, we show that there exists an *exponential* separation between the required size of neural networks for achieving low robust training and test error.

Table 1: Summary of our main results.

Params	Setting			
	Robust Training	Robust Generalization		
		General Case	Linear Separable	$k-$dim Manifold
Upper Bound	$\mathcal{O}(Nd)$ (Thm 2.2)	$\exp(\mathcal{O}(d))$ (Thm 3.3)		$\exp(\mathcal{O}(k))$ (Thm 5.5)
Lower Bound	$\Omega(\sqrt{Nd})$ (Thm 2.3)	$\exp(\Omega(d))$ (Thm 3.4)	$\exp(\Omega(d))$ (Thm 4.3)	$\exp(\Omega(k))$ (Thm 5.8)

- Based on our results, we conjecture that the widely observed drop of robust test accuracy is not due to limitations of existing algorithms – rather, it is a more fundamental issue originating from the expressive power of neural networks.

北京智源大会 BAAI CONFERENCE 2022.5.31-6.02

THANKS

BAAI CONFERENCE

Reference I

Peter L Bartlett, Nick Harvey, Christopher Liaw, and Abbas Mehrabian.
Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks.
The Journal of Machine Learning Research, 20(1):2285–2301, 2019.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter.
Certified adversarial robustness via randomized smoothing.
In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Ronald A DeVore, Ralph Howard, and Charles Micchelli.
Optimal nonlinear approximation.
Manuscripta mathematica, 63(4):469–478, 1989.

Binghui Li, Jikai Jin, Han Zhong, John E Hopcroft, and Liwei Wang.
Why robust generalization in deep learning is difficult: Perspective of expressive power.
arXiv preprint arXiv:2205.13863, 2022.

Reference II

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.
Towards deep learning models resistant to adversarial attacks.
arXiv preprint arXiv:1706.06083, 2017.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang.
Adversarial training can hurt generalization.
arXiv preprint arXiv:1906.06032, 2019.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein.
Adversarial training for free!
Advances in Neural Information Processing Systems, 32, 2019.

Reference III


Zuowei Shen, Haizhao Yang, and Shijun Zhang.
Optimal approximation rate of relu networks in terms of width and depth.
Journal de Mathématiques Pures et Appliquées, 157:101–135, 2022.


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


Gal Vardi, Gilad Yehudai, and Ohad Shamir.
On the optimal memorization power of relu neural networks.
arXiv preprint arXiv:2110.03187, 2021.


Yasi Wang, Hongxun Yao, and Sicheng Zhao.
Auto-encoder based dimensionality reduction.
Neurocomputing, 184:232–242, 2016.

Reference IV

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri.
A closer look at accuracy vs. robustness.
Advances in neural information processing systems, 33:8588–8601, 2020.

Bohang Zhang, Tianle Cai, Zhou Lu, Di He, and Liwei Wang.
Towards certifying l-infinity robustness using neural networks with l-inf-dist neurons.
In *International Conference on Machine Learning*, pages 12368–12379. PMLR, 2021.